ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΗΛΕΚΤΡΙΚΩΝ ΒΙΟΜΗΧΑΝΙΚΩΝ ΔΙΑΤΑΞΕΩΝ ΚΑΙ ΣΥΣΤΗΜΑΤΩΝ ΑΠΟΦΑΣΕΩΝ

ΜΟΝΑΔΑ ΠΡΟΒΛΕΨΕΩΝ ΚΑΙ ΣΤΡΑΤΗΓΙΚΗΣ

Πιθανοτικές προβλέψεις με χρήση ντετερμινιστικών μεταβλητών και μοντέλων μηχανικής μάθησης

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

ΕΥΑΓΓΕΛΟΥ ΜΑΡΓΩΝΗ

Επιβλέπων: Βασίλειος Ασημακόπουλος

Ομότιμος Καθηγητής Ε.Μ.Π.

Αθήνα, Φεβρουάριος 2025

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΗΛΕΚΤΡΙΚΩΝ ΒΙΟΜΗΧΑΝΙΚΩΝ ΔΙΑΤΑΞΕΩΝ ΚΑΙ ΣΥΣΤΗΜΑΤΩΝ ΑΠΟΦΑΣΕΩΝ

ΜΟΝΑΔΑ ΠΡΟΒΛΕΨΕΩΝ ΚΑΙ ΣΤΡΑΤΗΓΙΚΗΣ

Πιθανοτικές προβλέψεις με χρήση ντετερμινιστικών μεταβλητών και μοντέλων μηχανικής μάθησης

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

ΕΥΑΓΓΕΛΟΥ ΜΑΡΓΩΝΗ

Επιβλέπων: Βασίλειος Ασημακόπουλος

Ομότιμος Καθηγητής Ε.Μ.Π.

Υπεύθυνος: Ευάγγελος Σπηλιώτης

Επίκουρος Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 24η Φεβρουαρίου 2025.

(Υπογραφή)	(Υπογραφή)	(Υπογραφή)
…….	…….	…….
Βασίλειος Ασημακόπουλος	Δημήτριος Ασκούνης	Χρυσόστομος Δούκας
Ομότιμος Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.

Αθήνα, Φεβρουάριος 2025

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΗΛΕΚΤΡΙΚΩΝ ΒΙΟΜΗΧΑΝΙΚΩΝ ΔΙΑΤΑΞΕΩΝ ΚΑΙ ΣΥΣΤΗΜΑΤΩΝ ΑΠΟΦΑΣΕΩΝ

ΜΟΝΑΔΑ ΠΡΟΒΛΕΨΕΩΝ ΚΑΙ ΣΤΡΑΤΗΓΙΚΗΣ

ΔΗΛΩΣΗ ΜΗ ΛΟΓΟΚΛΟΠΗΣ ΚΑΙ ΑΝΑΛΗΨΗΣ ΠΡΟΣΩΠΙΚΗΣ ΕΥΘΥΝΗΣ

Με πλήρη επίγνωση των συνεπειών του νόμου περί πνευματικών δικαιωμάτων, δηλώνω ενυπογράφως ότι είμαι αποκλειστικός συγγραφέας της παρούσας Πτυχιακής Εργασίας, για την ολοκλήρωση της οποίας κάθε βοήθεια είναι πλήρως αναγνωρισμένη και αναφέρεται λεπτομερώς στην εργασία αυτή. Έχω αναφέρει πλήρως και με σαφείς αναφορές, όλες τις πηγές χρήσης δεδομένων, απόψεων και προτάσεων, ιδεών και λεκτικών αναφορών, είτε κατά κυριολεξία είτε βάσει επιστημονικής παράφρασης. Αναλαμβάνω την προσωπική και ατομική ευθύνη ότι σε περίπτωση αποτυχίας στην υλοποίηση των ανωτέρω δηλωθέντων στοιχείων, είμαι υπόλογος έναντι λογοκλοπής, γεγονός που σημαίνει αποτυχία στην Πτυχιακή μου Εργασία και κατά συνέπεια αποτυχία απόκτησης του Τίτλου Σπουδών, πέρα των λοιπών συνεπειών του νόμου περί πνευματικών δικαιωμάτων. Δηλώνω, συνεπώς, ότι αυτή η Πτυχιακή Εργασία προετοιμάστηκε και ολοκληρώθηκε από εμένα προσωπικά και αποκλειστικά και ότι, αναλαμβάνω πλήρως όλες τις συνέπειες του νόμου στην περίπτωση κατά την οποία αποδειχθεί, διαχρονικά, ότι η εργασία αυτή ή τμήμα της δεν μου ανήκει διότι είναι προϊόν λογοκλοπής άλλης πνευματικής ιδιοκτησίας.

(Υπογραφή)

…….

Ευάγγελος Μαργώνης

Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

24 Φεβρουαρίου 2025

Περίληψη

Ένα πρόβλημα που αντιμετωπίζει διαχρονικά ο κλάδος του λιανεμπορίου, είναι η αποτελεσματική αναπλήρωση του αποθέματος προϊόντων. Οι παράγοντες που επηρεάζουν τη ζήτηση και ως εκ τούτου τη διαδικασία αναπλήρωσης, περιλαμβάνουν μεταξύ άλλων τις προωθητικές ενέργειες του ίδιου του καταστήματος, την εποχιακότητα και τα ειδικά γεγονότα. Επομένως, κάθε πρόβλεψη ζήτησης οφείλει να λαμβάνει υπόψη τους συγκεκριμένους παράγοντες και να τους αξιοποιεί βέλτιστα για τη μείωση του σφάλματος. Ταυτόχρονα, η πρόοδος των υπολογιστικών συστημάτων, έχει οδηγήσει στην κατακόρυφη άνοδο της τεχνητής νοημοσύνης και εδραιώσει τα μοντέλα μηχανικής μάθησης στον τομέα των προβλέψεων χρονοσειρών.

Στην παρούσα διπλωματική εργασία εξετάζεται η συνεισφορά ενδεικτικών ντετερμινιστικών μεταβλητών, όπως οι παράγοντες που αναφέρθηκαν παραπάνω, στην ακρίβεια πρόβλεψης ζήτησης και διερευνάται αν πράγματι αυτές ωφελούν την τελική πρόβλεψη ή εισάγουν περιττή πληροφορία, μειώνοντας έτσι την απόδοση των μοντέλων. Η ανάλυση επικεντρώνεται στην αξιολόγηση πιθανοτικών προβλέψεων θεωρώντας διάστημα εμπιστοσύνης 95%, το οποίο επιλέγεται συχνά από τις επιχειρήσεις για την εξασφάλιση υψηλής διαθεσιμότητας προϊόντων στις αποθήκες, χρησιμοποιώντας παράλληλα το σύνολο δεδομένων του διαγωνισμού προβλέψεων M5, που αφορά στις πωλήσεις προϊόντων σε δέκα καταστήματα της Walmart.

Ξεκινώντας από κλασικές μεθόδους που χρησιμοποιούνται κατά κόρον για την πρόβλεψη της ζήτησης στον κλάδου του λιανεμπορίου, όπως η απλή μέθοδος εκθετικής εξομάλυνσης και τα μοντέλα ARIMA με εξωγενείς μεταβλητές, καταλήγουμε σε μοντέλα μηχανικής μάθησης και πιο συγκεκριμένα σε δέντρα παλινδρόμησης αξιοποιώντας τον αλγόριθμο LightGBM. Με αυτό τον τρόπο εξετάζονται και μη γραμμικά μοντέλα πρόβλεψης τα οποία δεν υποθέτουν κάποια συγκεκριμένη κατανομή για τη ζήτηση αλλά την εκτιμούν εμπειρικά με βάση τα διαθέσιμα δεδομένα.

Τα αποτελέσματα της εργασίας δείχνουν ότι η χρήση των μοντέλων μηχανικής μάθησης παρουσιάζει αρκετά βελτιωμένα αποτελέσματα στην ακρίβεια πρόβλεψη ζήτησης σε σχέση με τις κλασικές μεθόδους και πως η χρήση ντετερμινιστικών μεταβλητών είναι απαραίτητη για την καλύτερη παραμετροποίηση των μοντέλων πρόβλεψης.

Λέξεις Κλειδιά

Χρονοσειρά, Τεχνικές Προβλέψεων, Πιθανοτικές Προβλέψεις, M5, Δέντρα Αποφάσεων, LightGBM

Abstract

A problem that the retail industry has been facing over time is the inefficient replenishment of product inventory. Factors that affect demand and therefore the replenishment process include, among others, the store's own promotions, seasonality and special events. Therefore, any demand forecast must take into account these specific factors and make optimal use of them to reduce the error. At the same time, the advancement of computing systems has led to the meteoric rise of artificial intelligence and established machine learning models in the field of time series forecasting.

This thesis examines the contribution of indicative deterministic variables, such as the factors mentioned above, to the accuracy of demand forecasting and it is investigated whether they actually benefit the final forecast or introduce unnecessary information, thus reducing the performance of the models. The analysis focuses on the evaluation of probabilistic forecasts considering a 95% confidence interval, which is often chosen by companies to ensure high product availability in warehouses, using in parallel the M5 forecast competition data set, concerning product sales in ten Walmart stores.

Starting from classic methods commonly used to forecast demand in the retail industry, such as the simple exponential smoothing method and ARIMA models with exogenous variables, we end up with machine learning models and more specifically with regression trees utilizing the LightGBM algorithm. In this way, non-linear forecasting models are also examined which do not assume a specific distribution for the demand but estimate it empirically based on the available data.

The results of the thesis show that the use of machine learning models shows significantly improved results in demand forecasting accuracy compared to classical methods and that the use of deterministic variables is necessary for optimized parameterization of the forecasting models.

Keywords

Time Series, Forecasting Techniques, Probabilistic Forecasting, M5, Decision Trees, LightGBM

Ευχαριστίες

Θα ήθελα να ευχαριστήσω τον Καθηγητή κ. Βασίλη Ασημακόπουλο που μου έδωσε την ευκαιρία να εκπονήσω αυτήν την διπλωματική εργασία στη Μονάδα Προβλέψεων και Στρατηγικής και να ασχοληθώ με την επιστήμη των προβλέψεων. Επίσης, θα ήθελα να ευχαριστήσω τον Καθηγητή κ. Δημήτριο Ασκούνη και τον Καθηγητή κ. Χρυσόστομο Δούκα για την συμμετοχή τους στην εξεταστική επιτροπή της εργασίας.

Ευχαριστώ ιδιαίτερα τον Επίκουρο Καθηγητή κ. Ευάγγελο Σπηλιώτη, που υπήρξε ο Υπεύθυνος και η αιτία της ενασχόλησής μου με τον τομέα των προβλέψεων, μέσω του μαθήματος που διδάσκει.

Τέλος, θα ήθελα να ευχαριστήσω τους γονείς μου και τους φίλους μου για την στήριξή τους.

Αθήνα, Φεβρουάριος 2025

Ευάγγελος Μαργώνης

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

1. Ευρεία Περίληψη

1.1 Εισαγωγή

Ο κλάδος του λιανικού εμπορίου γνωρίζει ραγδαίες εξελίξεις τόσο στη δομή, με την ανάπτυξη των διαδικτυακών επιχειρήσεων, όσο και στο ανταγωνιστικό περιβάλλον που αντιμετωπίζουν οι εταιρίες. Ο μελλοντικός σχεδιασμός των επιχειρήσεων αυτού του κλάδου, εξαρτάται εν μέρει, από τις προβλέψεις ζήτησης, οι οποίες παρέχονται μέσω μεθόδων και διαδικασιών, ενσωματωμένων σε ένα σύστημα υποστήριξης προβλέψεων. Η πρόβλεψη ζήτησης υψηλής ακρίβειας, έχει άμεσο αντίκτυπο στην απόδοση της κάθε επιχείρησης, καθώς βελτιώνει πολλές διαδικασίες κατά μήκος της εφοδιαστικής αλυσίδας (Fildes, Ma & Kolassa, 2019).

Τα προβλήματα που θέλουν να λύσουν όσοι απασχολούνται στον συγκεκριμένο κλάδο, είτε αναφερόμαστε σε μεγαλύτερες εταιρίες είτε σε μικρότερες επιχειρήσεις, απαιτούν την ανάπτυξη στρατηγικής. Ένας από τους βασικούς τομείς του κλάδου για τον οποίο απαιτείται στρατηγική και λήψη αποφάσεων, είναι τα μακροχρόνια πλάνα της κάθε επιχείρησης. Για παράδειγμα, αν μία εταιρία θελήσει να αναπτύξει την διαδικτυακή της παρουσία, θα αναπτύξει στρατηγική που θα έχει έναν σχετικά μεγάλο χρονικό ορίζοντα για να πραγματοποιηθεί, αλλά θα στηρίζεται κατά κύριο λόγο σε προβλέψεις για αρκετά συντομότερο διάστημα, οι οποίες θα επηρεάζουν άμεσα το δίκτυο διανομής, το είδος των προϊόντων, τις τιμές και γενικότερα αποφάσεις που θα πρέπει να ληφθούν αρκετά νωρίς στην όλη διαδικασία, καθώς αν καθυστερήσουν, τα κόστη για τυχόν αλλαγές θα είναι υπέρογκα και γενικότερα η ευελιξία θα είναι χαμηλή. Οι μικρότερες επιχειρήσεις αντιμετωπίζουν την ίδια αβεβαιότητα στην λήψη αποφάσεων, απλώς στην περίπτωσή τους αλλάζουν οι τομείς που υπάρχει αυτή η αβεβαιότητα, με πιο συχνό την κοινωνική ομάδα που πρέπει να έχουν ως στόχο, ώστε να αυξήσουν το κέρδος τους.

Ένας άλλος τομέας στον οποίο απαιτείται ιδιαίτερη στρατηγική σε αυτόν τον κλάδο, ανεξάρτητα από το μέγεθος της επιχείρησης, είναι οι καθημερινές αποφάσεις για την εύρυθμη λειτουργία της επιχείρησης. Αυτές περιλαμβάνουν την διαφήμιση των προϊόντων τους, τις κατηγορίες των προϊόντων που διαθέτουν στο κοινό και την ποικιλία της κάθε κατηγορίας. Σε οποιαδήποτε περίπτωση, ο βασικός στόχος των αποφάσεων αυτών είναι η μεγιστοποίηση του κέρδους. Ο τρόπος χρησιμοποίησης των προωθητικών εργαλείων, η τιμολόγηση του κάθε προϊόντος και η συχνότητα των προωθητικών ενεργειών είναι μερικοί από τους παράγοντες που μπορούν να οδηγήσουν στην μεγιστοποίηση του κέρδους, εφόσον πάντα γίνει η καλύτερη πρόβλεψη για αυτές τις ενέργειες και ως ακόλουθο η βέλτιστη τελική απόφαση.

Η διαθεσιμότητα των προϊόντων της επιχείρησης στα ράφια της, εξαρτάται σημαντικά από τη σχέση μεταξύ του αποθέματος στις αποθήκες και του συστήματος διανομής που χρησιμοποιείται. Όταν προβλέπεται σωστά η μελλοντική ζήτηση των προϊόντων, γίνονται ευκολότερα τόσο η αναπλήρωση των ελλείψεων που υπάρχουν όσο και ο προγραμματισμός του εργατικού δυναμικού που θα χρειαστεί για κάθε μέρα λειτουργίας. Το σύστημα διανομής, είτε αυτό ανήκει στην ίδια την επιχείρηση, είτε έχει ανατεθεί εξωτερικά, εξοικονομεί εργατικές ώρες και κόστη, αν υπάρχει σωστός προγραμματισμός των διαδρομών του και εξυπηρετεί κάθε κατάστημα με βάση τις ανάγκες που προκύπτουν καθημερινά. Το εργατικό δυναμικό σε περιόδους που αναμένονται είτε

αυξημένες είτε μειωμένες πωλήσεις, είναι σε θέση να μεταβάλλεται και να μην παραμένει σταθερό, δημιουργώντας επιπλέον κόστη.

Είναι ξεκάθαρο, λοιπόν, ότι για να είναι επιτυχημένη μία επιχείρηση στους τομείς που αναφέρθηκαν, πρέπει να διαχειριστεί τη ζήτηση των προϊόντων της και τις διαδικασίες ανεφοδιασμού αυτών, ώστε να αποφευχθούν θέματα στην εξυπηρέτηση των πελατών, άσκοπο απόθεμα προϊόντων και περιττές εργατικές ώρες. Τα θέματα αυτά είναι πολύπλοκα, αν λάβει κανείς υπόψιν τις συνεχείς διακυμάνσεις στα δεδομένα ζήτησης, την παρουσία σημαντικού αριθμού μεσαζόντων στις διαδικασίες, την ποικιλία των προϊόντων προς πώληση και τις απαιτήσεις των καταναλωτών για ποιότητα στις υπηρεσίες προς αυτούς. Επομένως, η ακρίβεια της πρόβλεψης της ζήτησης είναι ζωτικής σημασίας για την οργάνωση και τον προγραμματισμό, την αγορά, την διανομή και το εργατικό δυναμικό, καθώς και για παρεχόμενες υπηρεσίες μετά την πώληση. Η ικανότητα των καταστημάτων λιανικής να εκτιμήσουν την αναμενόμενη ποσότητα πωλήσεων, οδηγεί σε βελτιωμένη εξυπηρέτηση πελατών, μειωμένη σπατάλη πόρων, αύξηση εσόδων από πωλήσεις και αποτελεσματικότερη και αποδοτικότερη διανομή.

Οι κλασικές μέθοδοι πρόβλεψης παρέχουν μία σημειακή εκτίμηση, η οποία μπορεί να είναι χρήσιμη, αλλά δεν είναι σε θέση να λάβει υπόψιν τις αβεβαιότητες και τη μεταβλητότα που κυριαρχεί στα δεδομένα του πραγματικού κόσμου. Εδώ φαίνεται ξεκάθαρη η σημασία της πιθανοτικής πρόβλεψης, η οποία ποσοτικοποιεί την αβεβαιότητα για το μέλλον και περιέχει ένα σύνολο πιθανοτήτων που σχετίζονται με όλα τα πιθανά μελλοντικά αποτελέσματα. Με αυτό τον τρόπο, οι επιχειρήσεις είναι σε θέση να καλύψουν όποιο ποσοστό της ζήτησης απαιτούν οι ίδιες. Ωστόσο, η αποτελεσματική εκτίμηση της αβεβαιότητας είναι ένα δύσκολο πρόβλημα λόγω του γεγονότος ότι οι εξεταζόμενες χρονοσειρές πολύ συχνά είναι διακοπτόμενες, δηλαδή ένα μεγάλο ποσοστό των διαθέσιμων δεδομένων, είναι μηδέν.

Ο τρόπος που γίνεται η πρόβλεψη στον κλάδο της λιανικής διαφέρει ανάλογα με το επίπεδο που προβλέπεται. Η πρόβλεψη συνολικών πωλήσεων λιανικής χωρίζεται στα επίπεδα της αγοράς, της αλυσίδας και του καταστήματος.

Η ανάλυση σε επίπεδο αγοράς μπορεί να αναφέρεται σε συνολικές πωλήσεις μίας ολόκληρης κατηγορίας προϊόντων ή μίας συγκεκριμένης βιομηχανίας σε χώρα ή περιοχή. Οι προβλέψεις αυτές είναι απαραίτητες για να κατανοηθούν οι μεταβαλλόμενες συνθήκες της αγοράς και την επίδραση που μπορεί να έχουν αυτές σε κάθε μία επιχείρηση. Βοηθούν στον εντοπισμό των δυνατοτήτων ανάπτυξης διαφορετικών επιχειρηματικών μοντέλων και ενισχύουν την ανάπτυξη νέων στρατηγικών ώστε οι επιχειρήσεις να διατηρήσουν την θέση τους στην αγορά. Εμφανίζουν ισχυρές τάσεις, εποχιακές διακυμάνσεις και μη αναμενόμενες μεταβολές, καθώς οι μεγάλες περίοδοι δεδομένων μπορεί να περιλαμβάνουν οικονομική ανάπτυξη, πληθωρισμό και απροσδόκητα γεγονότα (Geurts & Kelly, 1986). Στο συγκεκριμένο επίπεδο δεν έχει αποδειχθεί μέχρι στιγμής ότι τα μοντέλα μηχανικής μάθησης είναι σε θέση να κάνουν καλύτερη πρόβλεψη από τις κλασικές μεθόδους ή από οικονομετρικά μοντέλα ή από τον συνδυασμό αυτών των δύο (Bechter & Rutner, 1978).

Η ανάλυση σε επίπεδο αλυσίδας είναι σημαντική για τις επιχειρήσεις, καθώς απαιτούνται ακριβείς προβλέψεις λιανικών πωλήσεων για την οικονομική διαχείρισή τους και για να βοηθηθούν οι αποφάσεις χρηματοοικονομικών επενδύσεών τους. Σε γενικές

γραμμές, ότι ισχύει για το επίπεδο αγοράς, ισχύει και για το επίπεδο αλυσίδας, όσον αφορά τις μεθόδους πρόβλεψης που χρησιμοποιούνται. Όμως, η χρήση κάποιων ενδογενών μεταβλητών όπως είναι τα δεδομένα αποθέματος ή δεδομένα μικτού περιθωρίου (ο λόγος των εσόδων από τις πωλήσεις ως προς το κόστος των πωληθέντων αγαθών), μπορούν να βοηθήσουν σημαντικά την τελική πρόβλεψη.

Η ανάλυση σε επίπεδο καταστήματος επηρεάζεται άμεσα από την τοποθεσία, την τοπική οικονομία, τα δημογραφικά στοιχεία των καταναλωτών, προσφορές του καταστήματος ή ανταγωνιστών, καιρό, εποχή και τοπικές εκδηλώσεις (Alon, Qi & Sadowski, 2000). Χωρίζεται σε δύο περιπτώσεις, με την πρώτη να είναι η πρόβλεψη πωλήσεων υπάρχοντος καταστήματος για τη διανομή, τη βιωσιμότητα και τον προγραμματισμό του εργατικού δυναμικού και με την δεύτερη να είναι η πρόβλεψη πιθανών πωλήσεων νέου καταστήματος που βοηθάει στην τελική επιλογή της τοποθεσίας. Για την πρώτη περίπτωση, οι κλασσικές μέθοδοι με τη χρήση εξωτερικής πληροφορίας όπως δείκτης ανεργίας, μέσο εβδομαδιαίο εισόδημα και αργίες παρέχουν καλύτερα αποτελέσματα σε σχέση με τα οικονομετρικά μοντέλα ή μοντέλα που λαμβάνουν υπόψιν την κρίση κάποιου. Η πρόβλεψη για τις πωλήσεις ενός καινούριου καταστήματος παρουσιάζει αρκετές δυσκολίες, όμως είναι βασικό για την επιτυχία κάθε εταιρίας λιανικής. Δεν έχει αποδειχθεί ότι συγκεκριμένος τύπος μοντέλων βοηθάει περισσότερο την τελική απόφαση, η οποία στηρίζεται σε μεγάλο βαθμό στην κρίση ενός υπεύθυνου.

Για τις συνολικές πωλήσεις στα επίπεδα που είδαμε μέχρι στιγμής, απαιτούταν μακροπρόθεσμες προβλέψεις με σχετικά μεγάλο ορίζοντα. Οι προβλέψεις λιανικής στο επίπεδο προϊόντος γίνονται για μικρότερο ορίζοντα πρόβλεψης, για την επόμενη ημέρα ή την επόμενη εβδομάδα κατά κύριο λόγο και μπορεί να εκτείνεται μέχρι μήνα. Η δυνατότητα πρόβλεψης της ζήτησης για κάθε προϊόν ενός καταστήματος είναι κρίσιμη για την επιβίωση και την ανάπτυξή του και οι βασικότεροι λόγοι για αυτό έχουν ήδη αναφερθεί. Τα σφάλματα στην πρόβλεψη οδηγούν σε χειρότερη εξυπηρέτηση πελατών και υψηλότερο κόστος, καθώς οι αποφάσεις για τις μελλοντικές παραγγελίες πρέπει να διασφαλίζουν ότι το επίπεδο του αποθέματος δεν είναι ούτε αρκετά υψηλό, που έχει ως αποτέλεσμα υψηλό κόστος στην διατήρησή του, ούτε αρκετά χαμηλό, που έχει ως αποτέλεσμα χαμένες πιθανές πωλήσεις. Επομένως, η πρόβλεψη για τον έλεγχο των αποθεμάτων θα πρέπει να δίνει μεγαλύτερη έμφαση στις πιθανοτικές προβλέψεις και στα διαστήματα πρόβλεψης.

Για τις προβλέψεις σε επίπεδο προϊόντος, σημαντικό ρόλο παίζει η ιεραρχική δομή όπου υπάρχουν τα εξής τρία επίπεδα: επίπεδο SKU (Stock Keeping Unit ή Μονάδα Διατήρησης Αποθεμάτων), επίπεδο επωνυμίας και επίπεδο κατηγορίας. Το πρώτο επίπεδο είναι το βασικότερο, καθώς αποτελεί την βασική λειτουργική μονάδα για τον προγραμματισμό ημερήσιας αναπλήρωσης και διανομής αποθεμάτων και οι προβλέψεις για αυτό το επίπεδο γίνονται σε ημερήσια ή εβδομαδιαία βάση. Σε μία αλυσίδα, ο αριθμός των SKU μπορεί να είναι της τάξης του εκατομμυρίου, οπότε η διαχείριση και πρόβλεψή τους δεν είναι εύκολη υπόθεση. Η ιεραρχική πρόβλεψη των προϊόντων καθιστά ευκολότερο αυτό το κομμάτι και συνήθως μέθοδοι παλινδρόμησης είναι που παράγουν τα καλύτερα αποτελέσματα.

Ένα βασικό πρόβλημα που αντιμετωπίζει ο τομέας της πρόβλεψης για την αναπλήρωση αποθεμάτων, είναι ότι τυπικά θεωρείται κανονική κατανομή και σταθερή ζήτηση, κάτι

που δεν είναι ρεαλιστικό, ειδικά όταν η ζήτηση των προϊόντων είναι διακοπτόμενη και επηρεάζεται συνήθως από διάφορες προωθητικές ενέργειες. Η λύση σε αυτό το πρόβλημα είναι η εισαγωγή περισσότερων μη γραμμικών μοντέλων πρόβλεψης, που παράγουν κατανομή βασισμένη στα δεδομένα που έχουν ήδη στη διάθεσή τους. Τέτοια μοντέλα συνήθως χρησιμοποιούν και κάποιον αλγόριθμο μηχανικής μάθησης, ο οποίος είναι κατάλληλος και για να χρησιμοποιεί εξωγενείς μεταβλητές, που βοηθάνε στην αναγνώριση πολυπλοκότερων μοτίβων.

Η σωστή αξιολόγηση των προβλέψεων στο επίπεδο του καταστήματος είναι ένα ακόμα συχνό πρόβλημα που αντιμετωπίζει όποιος απασχολείται στον συγκεκριμένο τομέα. Οι συχνότεροι δείκτες ακρίβειας που χρησιμοποιούνται κατά κόρον στις περισσότερες εφαρμογές της πρόβλεψης, δεν τυγχάνουν της ίδιας αναγνώρισης στον κλάδο της λιανικής. Πρέπει να χρησιμοποιούνται κανονικοποιημένα σφάλματα, τα οποία είναι σε θέση να αξιολογήσουν αν οι προβλέψεις για μία σειρά που παρουσιάζει αρκετά μηδενικά στα δεδομένα της είναι αξιόλογες ή όχι. Ειδικά όταν οι προβλέψεις χρησιμοποιούνται για την αναπλήρωση των αποθεμάτων, δεν αρκούν μόνο οι σημειακές προβλέψεις, αλλά απαιτούνται και οι πιθανοτικές, οι οποίες θα επιβεβαιώσουν σε τι ποσοστό θα γίνει η σωστή αναπλήρωση.

Γενικότερα, ο τομέας της λιανικής βιώνει μεγάλες αλλαγές. Οι κλασικές μέθοδοι πρόβλεψης δεν αρκούν για να εξηγήσουν τις συνεχείς μεταβολές της αγοράς. Η χρήση εξωτερικής πληροφορίας για την κατανόηση μοτίβων στις πωλήσεις προϊόντων είναι επιτακτική και δεν περιορίζεται μόνο στις προωθητικές ενέργειες και τις εκπτώσεις του κάθε καταστήματος, αλλά απαιτείται και η γνώση του τι κάνει ο ανταγωνισμός και η γενικότερη αγορά. Η χρήση μοντέλων μηχανικής μάθησης που χρησιμοποιούν κατανομές με βάση μόνο τα δεδομένα που τους παρέχονται είναι ένα μεγάλο βήμα προς την σωστή κατεύθυνση, αλλά δεν αρκεί μόνο αυτό. Ειδικά για κολοσσούς του κλάδου, τα υπολογιστικά κόστη είναι δυσθεώρητα ακόμα, από τη στιγμή που απαιτούνται καθημερινοί υπολογισμοί για την αναπλήρωση των αποθεμάτων εκατομμυρίων διαφορετικών προϊόντων.

1.2 Στόχος Εργασίας

Στόχος της παρούσας διπλωματικής εργασίας, είναι η δημιουργία ενός πλαισίου παραγωγής πιθανοτικών προβλέψεων, το οποίο θα βασίζεται αποκλειστικά στην εμπειρική εκτίμηση της κατανομής των δεδομένων και των σφαλμάτων των σημειακών προβλέψεων, καθώς στην πραγματική ζωή, τα δεδομένα σπανίως ακολουθούν πιστά κάποια στατιστική θεωρητική κατανομή και οι πιθανοτικές προβλέψεις που προκύπτουν από μία κλασική προσέγγιση είναι συνήθως υποδεέστερης ποιότητας. Επιπλέον, θα εξεταστεί η δυνατότητα χρήσης διάφορων ντετερμινιστικών μεταβλητών (όπως ημέρα εβδομάδας, μήνας έτους ή κάποιο ειδικό γεγονός), για τον εμπλουτισμό της πληροφορίας που λαμβάνουν υπόψη τα μοντέλα πρόβλεψης. Συγκεκριμένα, το πλαίσιο θα εφαρμοστεί σε δεδομένα πωλήσεων προϊόντων και με τη χρήση στατιστικών μοντέλων, αλλά και με τη χρήση μεθόδων μηχανικής μάθησης.

Θα χρησιμοποιηθούν δεδομένα πωλήσεων από τον διαγωνισμό M5, για τα οποία θα πραγματοποιήσουμε αρχικά πιθανοτικές προβλέψεις των πιο κλασικών μεθόδων που χρησιμοποιούνται για την παραγωγή προβλέψεων στον κλάδο της λιανικής. Στη συνέχεια θα χρησιμοποιηθεί συγκεκριμένος αλγόριθμος μηχανικής μάθησης, με τον οποίο

θα αξιολογηθεί η σημασία που έχει κάθε εξωγενής μεταβλητή που θα ληφθεί υπόψιν στην βελτίωση του τελικού αποτελέσματος, ώστε τελικά να γίνει η επιλογή ποιες προσφέρουν ωφέλιμη πληροφορία και ποιες όχι. Θα απαιτηθεί συγκεκριμένο διάστημα πρόβλεψης, ώστε να είναι το μοντέλο μας πιο αντιπροσωπευτικό των μεθόδων που χρησιμοποιούνται συνηθέστερα στον τομέα του λιανικού εμπορίου για την αναπλήρωση των αποθεμάτων.

1.3 Δομή Εργασίας

Στο κεφάλαιο 2, παρουσιάζονται οι μέθοδοι που θα χρησιμοποιηθούν για την πρόβλεψη. Αρχικά, αναλύεται η έννοια της χρονοσειράς. Στη συνέχεια, γίνεται αναφορά στις πιο κλασικές μεθόδους που υπάρχουν και που είτε θα χρησιμοποιηθούν για να δημιουργηθεί ένα μοντέλο ορόσημο, είτε αποτελούν τη βάση για μία πολυπλοκότερη μέθοδο. Αναλύεται η έννοια της πιθανοτικής πρόβλεψης και παρουσιάζεται ο τρόπος ένταξης των εξωγενών μεταβλητών στην πρόβλεψη. Καταλήγουμε στην ανάλυση της πρόβλεψης με μηχανική μάθηση και πιο συγκεκριμένα με δέντρα αποφάσεων, τα οποία θα χρησιμοποιηθούν στο τελικό μοντέλο.

Στο κεφάλαιο 2, επίσης, αναλύονται οι δείκτες ακρίβειας που χρησιμοποιούνται συνηθέστερα για να αξιολογήσουν τις μεθόδους πρόβλεψης. Παρουσιάζονται επιγραμματικά οι πιο συνήθεις δείκτες και δίνεται έμφαση στον δείκτη που θα χρησιμοποιήσουμε για να αξιολογήσουμε τις πιθανοτικές προβλέψεις μας. Παρουσιάζεται ακόμα ένας στατιστικός δείκτης, αυτός της σχετικής συχνότητας, ο οποίος είναι σημαντικός για την επίτευξη του στόχου της παρούσας διπλωματικής εργασίας.

Στο κεφάλαιο 3, αναλύεται σε βάθος ο διαγωνισμός προβλέψεων M5, από τον οποίο προέρχονται τα σύνολα δεδομένων που θα χρησιμοποιηθούν στην εργασία. Γίνεται μία επεξηγηματική ανάλυση των δεδομένων του διαγωνισμού και επικεντρωνόμαστε σε αυτά που παρουσιάζουν το μεγαλύτερο ενδιαφέρον. Στη συνέχεια, παρουσιάζεται η πειραματική διάταξη της εργασίας και πως από την δημιουργία ενός απλού μοντέλου πρόβλεψης ώστε να χρησιμοποιηθεί σαν ορόσημο για τα επόμενα, καταλήγουμε στο τελικό μοντέλο μηχανικής μάθησης βασισμένο στα δέντρα αποφάσεων. Για κάθε μοντέλο που χρησιμοποιήσαμε αναφέρεται η πειραματική διάταξή του και σε τι διαφέρει από τα προηγούμενα και επικεντρωνόμαστε κυρίως στη μηχανική μάθηση που έχει τις περισσότερες παραμέτρους που πρέπει να οριστούν κατάλληλα για την επίτευξη του βέλτιστου δυνατού αποτελέσματος.

Στο κεφάλαιο 4, παρουσιάζονται τα αποτελέσματα των πειραματικών διατάξεων που αναλύθηκαν στο κεφάλαιο 3. Πρώτα αναλύονται τα αποτελέσματα των πιο κλασικών μεθόδων πρόβλεψης και επικεντρωνόμαστε σε αυτά των δέντρων αποφάσεων, όπου φαίνεται ο ρόλος που έπαιξε κάθε επεξηγηματική μεταβλητή στο τελικό μοντέλο, γιατί χρησιμοποιήθηκε η κάθε μία και ποιες τελικά δεν χρειάστηκαν, καθώς δεν συνείσφεραν στον επιθυμητό βαθμό στην τελική βελτίωση του μοντέλου. Στο τέλος του κεφαλαίου αυτού, πραγματοποιείται και ένας συνδυασμός των αλγορίθμων των δέντρων αποφάσεων με στόχο την περεταίρω βελτίωση του τελικού αποτελέσματος.

Στο κεφάλαιο 5, παρουσιάζονται συνοπτικά τα συμπεράσματα που προέκυψαν από το κεφάλαιο 4 αλλά και ολόκληρη την εργασία και γίνεται αναφορά στις μελλοντικές προεκτάσεις που μπορεί να λάβει η συγκεκριμένη διπλωματική.

2. Μέθοδοι Πρόβλεψης και Δείκτες Ακρίβειας

Η πρόβλεψη είναι η διαδικασία κατά την οποία χρησιμοποιούνται ιστορικά δεδομένα για την λήψη αποφάσεων για μελλοντικά γεγονότα ή καταστάσεις. Είναι αναπόσπαστο κομμάτι της καθημερινότητάς μας και πολύ συχνά συμβαίνει, χωρίς οι ίδιοι να το καταλαβαίνουμε. Για παράδειγμα, η απόφαση για το αν θα πάρεις ομπρέλα και για να ντυθείς ανάλογα σε περίπτωση βροχής, είναι ένας συνδυασμός της προσωπικής κρίσης του κάθε ανθρώπου με βάση τις συνθήκες που επικρατούν και τις μετεωρολογικές προβλέψεις.

Η σημασία της πρόβλεψης γίνεται πιο ξεκάθαρη στις στρατηγικές αποφάσεις που πρέπει να λάβει μία επιχείρηση, είτε αυτή είναι μικρή ή μεγαλύτερη. Ανάλογα την επιχείρηση, μπορεί να χρειάζονται προβλέψεις τόσο σε καθημερινό επίπεδο, όσο και μακροπρόθεσμα, οι οποίες οδηγούν σε στρατηγικές αποφάσεις με άμεσο αποτέλεσμα το βέλτιστο κέρδος. Για μία αλυσίδα καταστημάτων όπως είναι ένα σούπερ-μάρκετ, απαιτούνται προβλέψεις για τις πωλήσεις σε καθημερινή βάση αλλά και επενδυτικές αποφάσεις για τα επόμενα έτη, που αφορούν την επέκταση. Αν οι ημερήσιες πωλήσεις είναι σημαντικά περισσότερες από τις προβλέψεις που έγιναν, αυτό θα οδηγήσει σε μεγαλύτερο κόστος παραγωγής και υλικών και τελικά σε αδιάθετο εμπόρευμα. Αν είναι σημαντικά λιγότερες, θα οδηγήσει σε άδεια ράφια, με αποτέλεσμα την μην ικανοποίηση της πελατείας, γεγονός που μπορεί να έχει και μακροπρόθεσμες επιπτώσεις. Από την άλλη, αν οι επενδυτικές αποφάσεις ληφθούν με προβλέψεις που δεν λαμβάνουν υπόψιν όλες τις απαραίτητες παραμέτρους, ίσως οδηγηθεί σε σπατάλη εκατομμυρίων.

Η ταχεία πρόοδος των υπολογιστών και πιο συγκεκριμένα, της υπολογιστικής δυνατότητας, έχει καταστήσει ικανή την ανάλυση μεγαλύτερων και πιο περίπλοκων συνόλων δεδομένων και έχει διεγείρει το ενδιαφέρον για την ανάλυση δεδομένων. Αυτό έχει ως αποτέλεσμα, να διευρυνθούν τα εργαλεία και οι μέθοδοι για πρόβλεψη. Στις συνήθεις μεθόδους πρόβλεψης που υπάρχουν εδώ και χρόνια, έχουν προστεθεί τελευταία αυτές των νευρωνικών δικτύων και άλλοι τύποι μηχανικής μάθησης, όπως είναι τα δέντρα αποφάσεων. Για αρκετό καιρό οι συγκεκριμένες μέθοδοι ήταν σημείο διαφωνίας στον τομέα των προβλέψεων, με αρκετούς να αμφιβάλλουν για τη χρησιμότητά τους και να θεωρούν ότι δεν μπορούν να αντικαταστήσουν τις ήδη υπάρχουσες. Όμως, πλέον, με την ανάπτυξη των υπολογιστικών μέσων, η συνεισφορά τους είναι αδιαμφισβήτητη.

Ο τομέας της πρόβλεψης, ανά καιρούς δέχεται κριτική για το ότι δεν είναι ακριβής τις χρονικές περιόδους που πραγματοποιούνται απότομες κοινωνικές και πολιτικές μεταβολές. Αυτό συμβαίνει, επειδή δεν μπορεί να γίνει υποκατάστατο της προφητείας. Σε πρωτόγνωρες καταστάσεις, προφανώς και δεν μπορεί κάποιος να χρησιμοποιήσει κάποια υπάρχουσα μέθοδο και να προβλέψει ακριβώς τι θα συμβεί. Η ύπαρξη σφαλμάτων είναι αναπόφευκτη, αλλά το ζήτημα είναι μειωθούν στο όριο του εφικτού τα λάθη. Εδώ είναι που υπεισέρχεται η πρακτική ενασχόληση και η εμπειρία χρόνων. Η λεγόμενη κριτική πρόβλεψη, είναι αυτή που ξεχωρίζει σε τέτοιους καιρούς και συνήθως βασίζεται σε κάποια υπάρχουσα μέθοδο, που τροποποιείται ανάλογα με τις συνθήκες που επικρατούν.

Η θεωρία της πρόβλεψης βασίζεται στην υπόθεση ότι τρέχουσα και προηγούμενη γνώση μπορεί να χρησιμοποιηθεί για την παραγωγή προβλέψεων για το μέλλον.

Συγκεκριμένα για τις χρονοσειρές, θεωρούμε ότι είναι δυνατό να προσδιοριστούν μοτίβα στις ιστορικές τιμές και να εφαρμοστούν επιτυχώς στη διαδικασία πρόβλεψης των μελλοντικών τιμών. Ωστόσο, δεν αναμένεται ο ακριβής προσδιορισμός της μελλοντικής τιμής και ανάμεσα από τις επιλογές για την πρόβλεψη μίας χρονοσειράς σε συγκεκριμένη χρονική στιγμή είναι η αναμενόμενη τιμή (γνωστή και ως σημειακή πρόβλεψη), το διάστημα πρόβλεψης, το εκατοστημόριο (percentile) και ολόκληρη η κατανομή της πρόβλεψης.

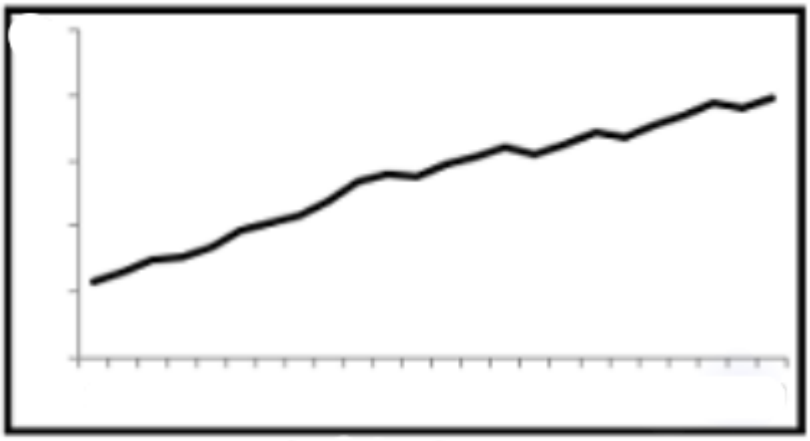
Στο παρόν κεφάλαιο, θα παρουσιάσουμε τα χαρακτηριστικά που μπορεί να εμφανίζει μία χρονοσειρά, θα αναφερθούμε στις σημαντικότερες μεθόδους πρόβλεψης και θα επικεντρωθούμε σε αυτές που χρησιμοποιήσαμε στο πειραματικό κομμάτι της παρούσας εργασίας (De Gooijer & Hyndman, 2006).

2.1 Χρονοσειρά

Η χρονοσειρά αποτελεί ένα σύνολο διαδοχικών παρατηρήσεων της τιμής κάποιου φυσικού ή άλλου μεγέθους (Πετρόπουλος & Ασημακόπουλος, 2013). Οι παρατηρήσεις αυτές δεν είναι ανεξάρτητες μεταξύ τους και οι μελλοντικές τιμές μπορούν να προσδιοριστούν από τις προηγούμενες. Αυτή η διαδικασία ονομάζεται ντετερμινιστική. Στην πραγματικότητα, όμως, δεν ισχύει πλήρως, αφού το μέλλον καθορίζεται μέχρι έναν βαθμό από το παρελθόν και όχι πλήρως και στα περισσότερα μεγέθη υπεισέρχεται ο τυχαίος παράγοντας, ο οποίος αντιπροσωπεύει μία στατιστική μεταβλητή. Τότε αυτά τα μοντέλα ονομάζονται στοχαστικά.

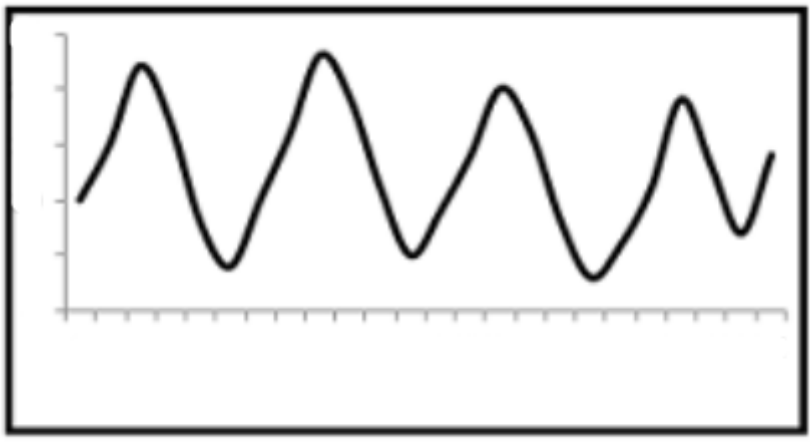
Η ανάλυση μίας χρονοσειράς μπορεί να δείξει πως αλλάζουν οι μεταβλητές με την πάροδο του χρόνου και πως τα δεδομένα προσαρμόζονται σε αυτές τις αλλαγές. Συνήθως η ανάλυση αυτή, απαιτεί μεγάλο αριθμό δεδομένων για να εξασφαλιστεί η συνοχή και η αξιοπιστία. Ένα εκτενές σύνολο δεδομένων εξασφαλίζει την ύπαρξη αντιπροσωπευτικού δείγματος και ότι θα μπορεί να αναγνωριστεί ο υπάρχων θόρυβος. Η πλειοψηφία των παραδοσιακών μεθόδων ανάλυσης και πρόβλεψης μίας χρονοσειράς, βασίζονται στην αποσύνθεση της διακύμανσής της. Από την αποσύνθεση αυτή, προκύπτουν τα βασικά ποιοτικά χαρακτηριστικά μίας χρονοσειράς, η τάση, η κυκλικότητα, η εποχιακότητα και οι ασυνέχειες ή μη κανονικές διακυμάνσεις, οι οποίες αναλύονται στη συνέχεια.

- Τάση, η οποία ορίζεται ως η γενική κατεύθυνση ή ως μακροπρόθεσμη μεταβολή του μέσου επιπέδου των τιμών της χρονοσειράς. Μπορεί να αυξάνεται, να μειώνεται ή να παραμένει σταθερή με την πάροδο του χρόνου. Ένα ζήτημα που προκύπτει από την μακροπρόθεσμη μεταβολή, είναι η περίοδος παρατήρησης, όπου ανάλογα το μέγεθος, μπορεί να προκύψει σύγχυση. Πιο συγκεκριμένα, αν κατά τη διάρκεια των παρατηρήσεων υπάρχει κυκλικότητα, η οποία θα αναλυθεί αμέσως μετά, αλλά το σύνολο δεδομένων διακόπτεται πριν επαναληφθεί, τότε υπάρχει περίπτωση να εκτιμηθεί λανθασμένα ως τάση. Επομένως, απαιτείται ικανός αριθμός παρατηρήσεων για κάθε χρονοσειρά, ώστε να αποφευχθούν τέτοια προβλήματα.



Σχήμα 2.1 : Παράδειγμα τάσης σε χρονοσειρά.

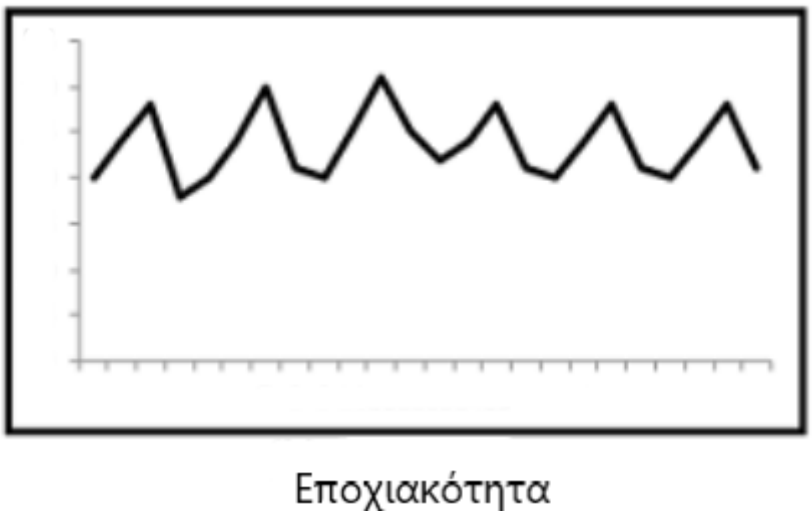
- Κυκλικότητα, η οποία αντιπροσωπεύει μια κυματοειδή μεταβολή που οφείλεται σε ειδικές εξωγενείς συνθήκες και εμφανίζεται κατά περιόδους. Οι διακυμάνσεις της δεν διαρκούν αρκετά για να χαρακτηριστούν ως τάση και δεν έχουν συγκεκριμένη διάρκεια. Το βασικότερο παράδειγμα χρονοσειρών που παρουσιάζουν κυκλικότητα, είναι οικονομικά μεγέθη, όπως οι τιμές των μετοχών και το Ακαθόριστο Εθνικό Προϊόν (Α.Ε.Π.), που είναι αποτέλεσμα οικονομικών συνθηκών, οι οποίες χαρακτηρίζονται από διαδοχικές ανόδους και υφέσεις, γνωστές ως επιχειρηματικός κύκλος.



Σχήμα 2.2 : Παράδειγμα κυκλικότητας σε χρονοσειρά.

- Η εποχιακότητα, η οποία είναι μια περιοδική διακύμανση με μήκος που είναι σταθερό και μικρότερο του έτους. Από τα χαρακτηριστικά μίας χρονοσειράς είναι η πιο ξεκάθαρη και προβλέψιμη. Οι διακυμάνσεις λόγω της εποχιακότητας είναι φανερές σε πλάτος, κατεύθυνση και χρονική περίοδο. Παρατηρούνται τόσο ανά μήνες όσο και ανά εβδομάδες. Αν τα δεδομένα που έχουμε στη διάθεσή μας εκτείνονται σε αρκετά έτη, τότε κάθε χρόνο την ίδια περίοδο, μπορούμε να δούμε συγκεκριμένα σε ποιους μήνες αυξάνονται ή μειώνονται

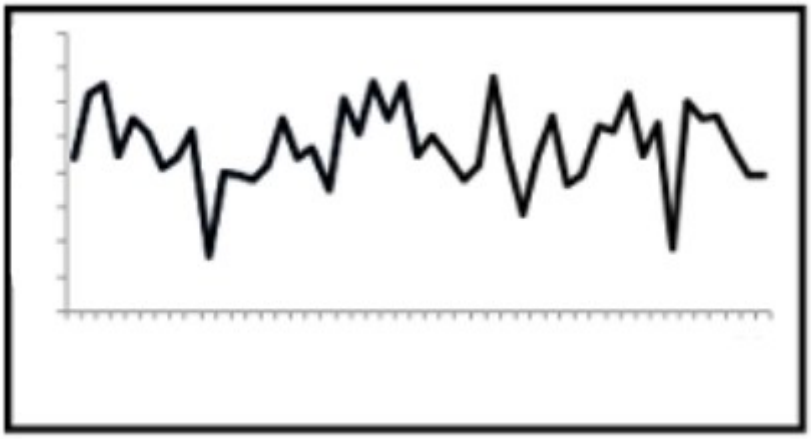
αντίστοιχα. Για σύνολα δεδομένων που εκτείνονται σε μικρότερες χρονικές περιόδους, μπορούμε να δούμε κορυφές και βυθίσεις για ημέρες της εβδομάδας. Για παράδειγμα, οι πωλήσεις ενός σούπερ-μάρκετ παρουσιάζουν κατακόρυφη αύξηση στο τέλος της εβδομάδας. Αντίστοιχο παράδειγμα για εκτενέστερα σύνολα, είναι ο αριθμός βημάτων που κάνει κάποιος άνθρωπος μέσα στην ημέρα, όπου τους μήνες της Άνοιξης και του Φθινοπώρου που δεν έχει ούτε πολλή ζέστη ούτε πολύ κρύο, θα αυξάνονται σημαντικά. Επειδή είναι τόσο εύκολο να αναλυθούν, απομονώνονται και προκύπτουν τα αποεποχικοποιημένα δεδομένα, για καλύτερη ανάλυση της χρονοσειράς.



Σχήμα 2.3 : Παράδειγμα εποχιακότητας σε χρονοσειρά.

- Οι ασυνέχειες, οι οποίες είναι οι απομονωμένες παρατηρήσεις που εμφανίζονται κατά τη διάρκεια μίας χρονοσειράς ως απότομες αλλαγές σε σχέση με το τι θα αναμενόταν να συμβεί και δεν υπήρχε τρόπος να προβλεφθούν με βάση τα προηγούμενα ιστορικά δεδομένα. Περιλαμβάνουν αλλοπρόσαλλες και απρόβλεπτες αποκλίσεις, αφού έχει γίνει έλεγχος για τάση, κυκλικότητα και εποχιακότητα. Οι αλλαγές αυτές μπορεί να έχουν παροδικό ή μόνιμο χαρακτήρα. Στην πρώτη περίπτωση, ονομάζονται ειδικά γεγονότα (special events) ή ακραίες τιμές (outliers) και η επίδραση που μπορεί να έχουν στην χρονοσειρά, έχει μικρή χρονική διάρκεια. Κάτι τέτοιο, μπορεί να αντιπροσωπεύει ένα ιδιαίτερο ή απρόβλεπτο γεγονός, που θα προκαλέσει ιδιαίτερη πτώση ή άνοδο στις παρατηρήσεις για εκείνη την ημέρα. Ένα τέτοιο γεγονός είναι τα Χριστούγεννα, όπου ακριβώς πριν από αυτό οι πωλήσεις ενός προϊόντος μπορεί να αυξηθούν κατακόρυφα, όμως την ημέρα των Χριστουγέννων που μπορεί να είναι και κλειστά τα περισσότερα καταστήματα, να είναι σχεδόν μηδενικές. Στην δεύτερη περίπτωση, ονομάζονται level-shifts και προκαλούν απότομες αλλαγές στο μέσο επίπεδο των τιμών μίας χρονοσειράς. Ένα τέτοιο γεγονός μπορεί να είναι το άνοιγμα ή το κλείσιμο ενός καταστήματος από μεγάλη αλυσίδα, όπου αντίστοιχα θα παρατηρηθεί είτε σημαντική αύξηση είτε σημαντική μείωση στις συνολικές πωλήσεις και τελικά το μέσο επίπεδο θα σταθεροποιηθεί είτε σε υψηλότερο είτε σε χαμηλότερο σημείο. Υπάρχουν και οι μη κανονικές διακυμάνσεις, οι οποίες μπορεί να αντιπροσωπεύουν μία εντελώς τυχαία μεταβλητή, που εκφράζει τον τυχαίο παράγοντα μίας στοχαστικής διαδικασίας.

Γενικεύοντας, ότι δεν μπορεί να εξηγηθεί μέσω των τριών πρώτων χαρακτηριστικών, θεωρείται μη κανονική διακύμανση.



Σχήμα 2.4 : Παράδειγμα ασυνεχειών σε χρονοσειρά.

Σε χρονοσειρές διακοπτόμενης ζήτησης, είναι πολύ σημαντικός ο καθορισμός των δύο παρακάτω βασικών στατιστικών δεικτών:

- Μέση τιμή διαστήματος μεταξύ ζητήσεων (Average Demand Interval ή ADI), που εκφράζει τη μέση τιμή των αποστάσεων διαδοχικών περιόδων με μη μηδενική τιμή. Αν η τιμή του δείκτη προκύψει ίση με τη μονάδα, τότε αναφερόμαστε σε χρονοσειρά συνεχούς ζήτησης, ενώ στην περίπτωση διακοπτόμενης ζήτησης, παίρνει τιμές μεγαλύτερες της μονάδας. Μεγάλη τιμή του δείκτη συνεπάγεται μεγάλα, κατά μέσο όρο, μεσοδιαστήματα μεταξύ μη μηδενικών παρατηρήσεων.
- Συντελεστής μεταβλητότητας (Coefficient of Variation ή CV), που είναι ένα κανονικοποιημένο μέτρο της διασποράς των παρατηρήσεων ενός δείγματος ή ενός πληθυσμού. Σε σύγκριση με την απλή τυπική απόκλιση, έχει το πλεονέκτημα ότι είναι απαλλαγμένο από την επίδραση του επιπέδου των παρατηρήσεων, αλλά δεν μπορεί να υπολογισθεί όταν η μέση τιμή είναι μηδέν.

2.2 Αποσύνθεση Χρονοσειρών

Οι μέθοδοι αποσύνθεσης εφαρμόζουν απλές μαθηματικές σχέσεις, ώστε να απομονώσουν τα τέσσερα βασικά χαρακτηριστικά των χρονοσειρών που αναφέραμε παραπάνω, δηλαδή τάση, κυκλικότητα, εποχιακότητα και τυχαιότητα (ασυνέχειες σε συνδυασμό με μη κανονικές διακυμάνσεις). Η μαθηματική διατύπωση της αποσύνθεσης είναι η παρακάτω:

$$Y_t = f(S_t, T_t, C_t, R_t)$$

όπου Y_t η παρατήρηση κατά τη χρονική περίοδο t, S_t η συνιστώσα της εποχιακότητας, T_t η συνιστώσα της τάσης, C_t η συνιστώσα του κύκλου και R_t η συνιστώσα της τυχαιότητας.

Οι πιο απλές συναρτησιακές μορφές της μαθηματικής διατύπωσης είναι η προσθετική και η πολλαπλασιαστική.

Το προσθετικό μοντέλο έχει την εξής μορφή:

$$Y_t = S_t + T_t + C_t + R_t$$

Το πολλαπλασιαστικό μοντέλο έχει την εξής μορφή:

$$Y_t = S_t * T_t * C_t * R_t$$

Το προσθετικό μοντέλο χρησιμοποιείται όταν οι μεταβολές γύρω από την τάση που παρουσιάζει η χρονοσειρά δεν μεταβάλλεται με βάση το επίπεδό της, ενώ το πολλαπλασιαστικό είναι καταλληλότερο όταν η τάση είναι ανάλογη του επιπέδου της χρονοσειράς.

Συχνά η τάση και η κυκλικότητα ομαδοποιούνται στη σειρά τάσης-κύκλου, εξαλείφοντας τις συνιστώσες της εποχιακότητας και της τυχαιότητας. Σε αυτή τη σειρά εφαρμόζεται στατιστική πρόβλεψη και στη συνέχεια εποχικοποιείται με βάση τα αποτελέσματα της αποσύνθεσης. Για να προκύψει το τελικό αποτέλεσμα, υπεισέρχεται η κριτική πρόβλεψη με βάση την εμπειρία και την διαίσθηση του ατόμου που προβλέπει και έτσι καταλήγουμε στην τελική στατιστική πρόβλεψη.

Όσον αφορά τον υπολογισμό της σειράς τάσης-κύκλου, απαιτείται εξομάλυνση ή αποσύνθεση των δεδομένων. Η πιο απλή μέθοδος εξομάλυνσης είναι η χρήση κινητών μέσων όρων, η οποία θα αναλυθεί στη συνέχεια.

Η εφαρμογή κατάλληλων κινητών μέσων όρων είναι ο πιο απλός τρόπος εξομάλυνσης και στόχος τους είναι η εκτίμηση της σειράς τάσης-κύκλου για κάθε παρατήρηση. Παρακάτω θα δούμε επιγραμματικά τέσσερις από αυτούς, τον απλό κινητό μέσο όρο, τον σταθμισμένο κινητό μέσο όρο, τον κεντρικό κινητό μέσο όρο και τον διπλό κινητό μέσο όρο.

- Για τον απλό κινητό μέσο όρο (ΚΜΟ), ο υπολογισμός του γίνεται για τον μέσο όρο n τιμών της χρονοσειράς, γύρω από την παρατήρηση για την οποία πρέπει να υπολογιστεί η σειρά τάσης-κύκλου. Για την διατήρηση της συμμετρίας στους υπολογισμούς, ο αριθμός n πρέπει να είναι περιττός. Η μαθηματική εξίσωση που περιγράφει την συγκεκριμένη διαδικασία είναι η εξής:

$$TC_t = KMO(n)_t = \frac{1}{n} \sum_{i=-(n \bmod 2)}^{(n \bmod 2)} Y_{t+i} \Longrightarrow$$

$$TC_t = \frac{1}{n}(Y_{t-(n \bmod 2)} + \cdots + Y_{t-1} + Y_t + Y_{t+1} + \cdots + Y_{t+(n \bmod 2)})$$

όπου, ΤC η σειρά τάσης κύκλου, Y η αρχική χρονοσειρά, t η τρέχουσα χρονική παρατήρηση, n το μήκος του απλού κινητού μέσου όρου εξομάλυνσης και mod η πράξη τη ακέραιας διαίρεσης.

Σε αυτή την μέθοδο, το σημαντικότερο σημείο είναι η επιλογή του μήκους n. Όσο μεγαλύτερο το n, τόσο μεγαλύτερη η εξομάλυνση, καθώς περισσότερες

παρατηρήσεις συμμετέχουν στον υπολογισμό της κάθε μίας παρατήρησης. Αν όμως είναι πολύ μεγάλη η τιμή του n, τότε υπάρχει ο κίνδυνος να χαθεί η διάκριση της τάσης και της κυκλικότητας. Για έντονη εποχιακότητα, συνίσταται η εφαρμογή κινητού μέσου όρου ίσου ή μεγαλύτερου από το μήκος της περιοδικότητας.

- Ο σταθμισμένος κινητός μέσος όρος (ΣΚΜΟ), είναι μία παραλλαγή του απλού κινητού μέσου όρου, όπου οι γειτονικές παρατηρήσεις συμμετέχουν στον υπολογισμό της τάσης-κύκλου με άνισα βάρη. Αυτά τα βάρη έχουν άθροισμα την μονάδα και συνήθως δίνεται μεγαλύτερο βάρος στις παρατηρήσεις που είναι πιο κοντά στην τρέχουσα και επιλέγονται ώστε να είναι συμμετρικά ως προς αυτή.

  Η μαθηματική εξίσωση που περιγράφει την συγκεκριμένη διαδικασία είναι η εξής:

$$TC_t = \Sigma KMO(n)_t = \sum_{i=-(n\ mod\ 2)}^{(n\ mod\ 2)} a_i Y_{t+i} \text{ , όπου } \sum_{i=-(n\ mod\ 2)}^{(n\ mod\ 2)} a_i = 1$$

  όπου α_i είναι τα βάρη για κάθε γειτονική παρατήρηση. Το τελικό αποτέλεσμα του σταθμισμένου είναι μία πιο εξομαλυμένη σειρά τάσης-κύκλου.

- Ο διπλός κινητός μέσος όρος (ΔΚΜΟ), είναι πρακτικά διπλή εφαρμογή του απλού με ίσα ή άνισα μήκη. Το τελικό αποτέλεσμα είναι διπλή εξομάλυνση και λειτουργεί όπως ο σταθμισμένος, καθώς ο υπολογισμός της σειράς τάσης κύκλου γίνεται με άνισα βάρη.

- Ο κεντρικός κινητός μέσος όρος (ΚΚΜΟ), είναι ουσιαστικά συνδυασμός του απλού και του διπλού. Η διαφορά τους έγκειται στο γεγονός ότι οι προηγούμενοι απαιτούσαν n αριθμό παρατηρήσεων, ενώ αυτός άρτιο. Το τελικό αποτέλεσμα και σε αυτή την περίπτωση είναι ένα είδος σταθμισμένου με άνισα βάρη.

Η ακριβής μεθοδολογία για την κλασσική μέθοδο αποσύνθεσης δεν θα μελετηθεί στην παρούσα εργασία, καθώς δεν είναι αντικείμενό της. Μέχρι τώρα αναφέρθηκαν κατά κύριο λόγο οι κινητοί μέσοι όροι για να εστιάσουμε στην τάση, στην εποχιακότητα και στην τυχαιότητα και πως αυτά τα χαρακτηριστικά των χρονοσειρών μπορούν να αποτυπωθούν για την καλύτερη κατανόηση τους.

2.3 Πιθανοτικές Προβλέψεις

Οι πιθανοτικές προβλέψεις παρέχουν μία πλήρη εικόνα της αβεβαιότητας που σχετίζεται με τις μελλοντικές τιμές, προβλέποντας ολόκληρη την κατανομή των πιθανών αποτελεσμάτων, αντί να πραγματοποιούν εκτίμηση ενός μόνο σημείου. Αυτή η προσέγγιση, είναι ζωτικής σημασίας στη διαδικασία λήψης αποφάσεων, όπου η κατανόηση του εύρους των πιθανών αποτελεσμάτων και των σχετικών πιθανοτήτων είναι πολυτιμότερη από μία μεμονωμένη ντετερμινιστική πρόβλεψη.

Αν οι μεταβλητές για τις οποίες γίνεται η πρόβλεψη είναι διακριτές, τότε οι πιθανοτικές προβλέψεις ονομάζονται προβλέψεις πιθανοτήτων, ενώ αν είναι συνεχείς, τότε υπάρχουν διάφορα είδη προβλέψεων και παρακάτω αναλύονται τα σημαντικότερα.

- Προβλέψεις Συνόλου (Ensemble Forecasts), όπου παράγονται πολλαπλές προσομοιώσεις του μοντέλου που χρησιμοποιούμε, καταλήγοντας τελικά σε μία πρόβλεψη συνόλου. Το σύνολο αυτό, παρέχει διαφορετικά αλλά εξίσου πιθανά σενάρια δεδομένων των αρχικών συνθηκών. Διαφορετικά μέρη του συνόλου επιχειρούν να εκφράσουν της εξάπλωση της αβεβαιότητας στη διαδικασία της μοντελοποίησης, η οποία προκύπτει από εκλιπούσες ή λανθασμένες παρατηρήσεις, παραμετρική αβεβαιότητα και σφάλμα προτύπου.
- Προβλέψεις Εκατοστημορίων (Quantile Forecasts), για τις οποίες χρησιμοποιείται μέθοδος παλινδρόμησης, η οποία αναλύεται στη συνέχεια της εργασίας. Η πρόβλεψη εκατοστημορίου είναι η τιμή που η παρατήρηση έχει προκαθορισμένη πιθανότητα να είναι μικρότερη ή ίση από αυτή. Πιο αναλυτικά, υπολογίζεται ένας πεπερασμένος αριθμός από εκατοστημόρια της πιθανολογικής κατανομής του μεγέθους που προβλέπουμε. Το εκατοστημόριο θ, ορίζεται ως η τιμή που η πιθανότητα το μέγεθος να είναι μικρότερο από αυτό, ισούται με θ. Δεδομένου ότι η F_t είναι μία γνησίως αύξουσα συνάρτηση, το εκατοστημόριο ${q_t}^{(a)}$ της τυχαίας μεταβλητής P_t(με $a \epsilon [0,1]$), ορίζεται μονοσήμαντα ως η τιμή του τέτοια ώστε:

  $$P(P_t < x) = a$$

  Ή ισοδύναμα:

  $${q_t}^{(a)} = {F_t}^{-1}(a)$$

- Διαστήματα Πρόβλεψης (Prediction Intervals), τα οποία χρησιμοποιούνται για να εκφράσουν ένα πιθανό εύρος τιμών μέσα στο οποίο το πραγματικό γεγονός P_t, αναμένεται να βρίσκεται με μία συγκεκριμένη πιθανότητα. Είναι δηλαδή, το εύρος των τιμών εντός του οποίου, με βεβαιότητα p%, αναμένουμε να πέσουν οι μελλοντικές τιμές.

Στη συνέχεια θα επικεντρωθούμε αποκλειστικά στις μεθόδους προβλέψεων, ξεκινώντας από τις βασικότερες και καταλήγοντας σε αυτές που θα μας απασχολήσουν περισσότερο στο πειραματικό στάδιο της παρούσας εργασίας.

2.4 Κλασικές Μέθοδοι Πρόβλεψης

2.4.1 Αφελής Μέθοδος

Η Αφελής μέθοδος (Naive), αποτελεί την πιο απλή στατιστική μέθοδο. Σύμφωνα με αυτή, η πρόβλεψη που πραγματοποιεί για μία χρονική στιγμή, είναι ίδια με την παρατήρηση της ακριβώς προηγούμενης χρονικής στιγμής. Η μαθηματική εξίσωση που την περιγράφει είναι η εξής:

$$F_t = Y_{t-1}$$

όπου t η χρονική στιγμή, F η πρόβλεψη και Y η παρατήρηση της χρονοσειράς.

Η χρήση αυτής της προσέγγισης ακούγεται πραγματικά αφελής στην αρχή, αλλά σε πολλές περιπτώσεις είναι αρκετά δύσκολο να παράγεις καλύτερο αποτέλεσμα και για αυτό το λόγο πολύ συχνά χρησιμοποιείται σαν ορόσημο που πρέπει να ξεπεράσουν οι επόμενες μέθοδοι που θα χρησιμοποιηθούν. Αν πάρουμε ως παράδειγμα την πρόβλεψη για το τι θερμοκρασία θα έχει δέκα λεπτά αργότερα, είναι η μέθοδος που θα έχει το πιο κοντινό αποτέλεσμα, καθώς απλά προβλέποντας την ίδια θερμοκρασία θα είναι ακριβής.

Το στατιστικό μοντέλο που αποτελεί τη βάση της Αφελής μεθόδου είναι αυτό του Τυχαίου Περιπάτου (Random Walk), όπου περιγράφει μία ακολουθία διακριτών βημάτων συγκεκριμένου μήκους προς τυχαία κατεύθυνση και περιγράφεται από την μαθηματική εξίσωση:

$$Y_t = Y_{t-1} + e_t$$

Η μεταβλητότητα του e_t επηρεάζει το πόσο γρήγορα αλλάζουν τα δεδομένα και όσο μεγαλύτερο, τόσο ταχύτερη είναι αυτή η αλλαγή.

Η συγκεκριμένη μέθοδος απαντάται πολύ συχνά σε οικονομικές και χρηματοοικονομικές χρονοσειρές, καθώς σε απρόσμενες μεταβολές, αλλάζει αμέσως το επίπεδο του, λαμβάνοντας υπόψιν μόνο την τελευταία παρατήρηση.

Αν συναντάται εποχιακότητα σε κάποια χρονοσειρά, μπορούμε με παρόμοια λογική να χρησιμοποιήσουμε την μέθοδο της εποχιακής Αφελής μεθόδου. Όπως και νωρίτερα, βασίζεται μόνο σε μία παρατήρηση, αλλά αντί να χρησιμοποιηθεί η τελευταία διαθέσιμη, λαμβάνουμε αυτή από την τελευταία περίοδο που παρουσιάζει ίδια εποχιακότητα. Για παράδειγμα, αν παρατηρείται εβδομαδιαία εποχιακότητα σε χρονοσειρά πωλήσεων και θέλουμε να προβλέψουμε την τιμή που θα έχει την ημέρα της Παρασκευής, αυτή θα ισούται με την παρατήρηση της προηγούμενης Παρασκευής. Η μαθηματική διατύπωση είναι και αυτή παρόμοια:

$$F_t = Y_{t-m}$$

όπου m η εποχιακή συχνότητα της χρονοσειράς.

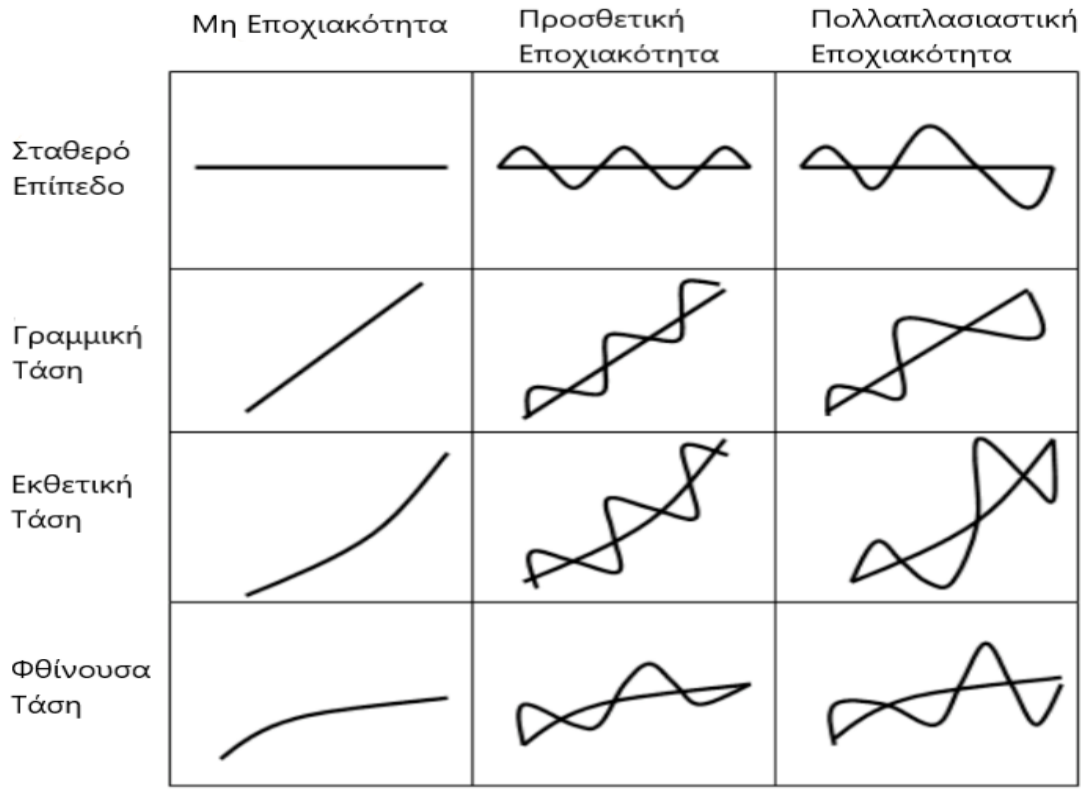
Παρομοίως με την απλή Αφελή, η εποχιακή στηρίζεται στο στατιστικό μοντέλο του Εποχιακού Τυχαίου Περιπάτου (Seasonal Random Walk).

2.4.2 Μοντέλο Απλής Εκθετικής Εξομάλυνσης Σταθερού Επιπέδου

Από τις πιο δημοφιλείς απλές μεθόδους πρόβλεψης, είναι αυτές της εκθετικής εξομάλυνσης, καθώς απαιτούν ελάχιστο υπολογιστικό χρόνο και λιγοστές παρατηρήσεις για την παραγωγή προβλέψεων (Hyndman et al, 2008). Είναι κατάλληλες για βραχυπρόθεσμες και μεσοπρόθεσμες προβλέψεις μεγάλου όγκου χρονοσειρών και αποδίδουν βέλτιστα σε δεδομένα που παρουσιάζουν στασιμότητα ή μικρό ρυθμό ανάπτυξης. Αυτού του είδους οι μέθοδοι, αποδίδουν άνισα βάρη στα ιστορικά δεδομένα, τα οποία φθίνουν εκθετικά, ξεκινώντας από την πιο πρόσφατη τιμή και φτάνοντας στην πιο μακρινή. Στηρίζονται στην ιδέα ότι υπάρχει ένα λανθάνον πρότυπο συμπεριφοράς (underlying pattern), το οποίο ακολουθούν οι τιμές των μεταβλητών πρόβλεψης και τα ιστορικά

δεδομένα αντιπροσωπεύουν το συγκεκριμένο πρότυπο σε συνδυασμό με τυχαίες διακυμάνσεις. Έχουν ως βασικό στόχο την αναγνώριση του προτύπου και των τυχαίων αποκλίσεων, το οποίο το πετυχαίνουν εξομαλύνοντας αυτά τα δεδομένα και τελικά, περιορίζουν την τυχαιότητα όσο περισσότερο μπορούν και η πρόβλεψή τους βασίζεται στο εξομαλυμένο πρότυπο συμπεριφοράς των δεδομένων.

Η κατηγοριοποίηση των μοντέλων εξομάλυνσης πραγματοποιείται σύμφωνα με το παρακάτω σχήμα, όπου φαίνεται η γραφική αναπαράσταση των ιστορικών δεδομένων με οριζόντιο άξονα τον χρόνο.



Σχήμα 2.5 : Τύποι μοντέλων εξομάλυνσης.

Εδώ, τέσσερα μοτίβα τάσης συνδυάζονται με τρία εποχιακά μοτίβα. Το μοτίβο σταθερού επιπέδου χρησιμοποιείται σε όσες χρονοσειρές θεωρούμε ότι κινούνται γύρω από έναν σχετικά σταθερό μέσο όρο και η επόμενη πρόβλεψη παράγεται απλώς από την προέκταση μίας οριζόντιας ευθείας γραμμής. Η χρήση του περιορίζεται όπου απαιτείται πρόβλεψη ενός βήματος ή επικρατεί θόρυβος και τυχαιότητα στην εξεταζόμενη χρονοσειρά.

Το πιο δημοφιλές από τα τέσσερα μοτίβα τάσης είναι το γραμμικό, όπου και εδώ η πρόβλεψη παράγεται από την προέκταση μία ευθείας γραμμής, αλλά στην συγκεκριμένη περίπτωση δεν είναι οριζόντια. Όμως, αν έχουμε μεσοπρόθεσμο ορίζοντα πρόβλεψης, οι τελικές προβλέψεις θα μεγαλώσουν αρκετά και δεν θα μπορέσει να αναπαριστά λογικά την πραγματική πορεία της χρονοσειράς. Σε αυτή την περίπτωση, το μοτίβο εκθετικής τάσης θα παρουσιάζει καλύτερα αποτελέσματα, καθώς δεν αυξάνεται η πρόβλεψη με την ίδια ταχύτητα. Φτάνοντας σε μακροπρόθεσμο ορίζοντα πρόβλεψης, το μοτίβο που θα λειτουργήσει καλύτερα από όσα έχουμε ήδη αναφέρει, είναι αυτό της

φθίνουσας τάσης, όπου υπάρχει βαθμιαία μείωση του μεγέθους με το οποίο αυξάνονται οι τιμές της χρονοσειράς μέσα σε μία χρονική περίοδο.

Όσον αφορά τα μοτίβα εποχιακότητας, στο προσθετικό, το εύρος που έχουν οι εποχιακές διακυμάνσεις, θεωρείται σταθερό και ανεξάρτητο από τυχούσες αυξήσεις στις τιμές της χρονοσειράς, σε αντίθεση με το πολλαπλασιαστικό, όπου το εύρος αυτό είναι ανάλογο του ύψους τιμών των δεδομένων. Αυτό σημαίνει ότι, όσο αυξάνεται η τάση της χρονοσειράς, τόσο αυξάνονται και οι εποχιακές διακυμάνσεις. Από τα δύο αυτά μοντέλα εποχιακότητας, αυτό που συναντάται περισσότερο είναι το πολλαπλασιαστικό.

Στην παρούσα εργασία, θα επικεντρωθούμε στην απλή εκθετική εξομάλυνση σταθερού επιπέδου (Simple Exponential Smoothing ή SES), ώστε να το χρησιμοποιήσουμε ως ορόσημο που θα πρέπει να ξεπεραστεί από τα επόμενα μοντέλα πρόβλεψης. Η επιλογή του συγκεκριμένου μοντέλου έγινε, επειδή είναι μία μέθοδος που μπορεί να προεκτείνει στοιχεία του συνόλου δεδομένων, όπως τάσεις και εποχιακούς κύκλους, στο μέλλον, κάτι πολύ σημαντικό με βάση το σύνολο δεδομένων που θα χρησιμοποιηθεί.

Το μοντέλο σταθερού επιπέδου, περιγράφεται από τις παρακάτω εξισώσεις:

$$e_t = Y_t - F_t$$

$$S_t = S_{t-1} + a * e_t$$

$$F_{t+1} = S_t$$

όπου, e είναι το σφάλμα, δηλαδή η απόκλιση της πραγματικής τιμής από την πρόβλεψη, S το επίπεδο, F η πρόβλεψη, t η χρονική περίοδος και α ο συντελεστής εξομάλυνσης της μεθόδου που λαμβάνει τιμές στο διάστημα [0,1].

Αρχικά, θα επικεντρωθούμε στο επίπεδο S και στο συντελεστή εξομάλυνσης α, καθώς είναι οι δύο μεταβλητές, των οποίων η επιλογή της τιμής τους δεν είναι ξεκάθαρη. Το αρχικό επίπεδο S_0 είναι αναγκαίο να οριστεί προκειμένου να ξεκινήσει η όλη διαδικασία της πρόβλεψης. Οι πιο συχνές επιλογές για αρχικό επίπεδο είναι οι παρακάτω:

- Ο μέσος όρος όλων των διαθέσιμων παρατηρήσεων.
- Ο μέσος όρος των ν πρώτων παρατηρήσεων.
- Η πρώτη παρατήρηση.
- Το σταθερό επίπεδο του μοντέλου της απλής γραμμικής παλινδρόμησης.

Η επιλογή γίνεται έχοντας ως γνώμονα τα ποιοτικά χαρακτηριστικά της χρονοσειράς για την οποία θέλουμε να παράγουμε προβλέψεις. Είναι φανερό πόσο σημαντική είναι η επιλογή του αρχικού επιπέδου, καθώς επηρεάζει ολόκληρη την διαδικασία, τις τιμές των προβλέψεων αλλά και την επιλογή του συντελεστή εξομάλυνσης. Ουσιαστικά, το αρχικό επίπεδο είναι και υποχρεωτικά η πρώτη πρόβλεψη που θα γίνει.

Η επιλογή του συντελεστή εξομάλυνσης γίνεται με βάση δύο παράγοντες. Ο πρώτος είναι ο θόρυβος που επικρατεί στη χρονοσειρά και είναι δυσανάλογος με τον συντελεστή, καθώς όσο αυξάνεται ο θόρυβος, τόσο πιο χαμηλή πρέπει να είναι η τιμή του συντελεστή, ώστε να αποφεύγεται η υπερβολική αντίδραση σε αυτόν. Ο δεύτερος παράγοντας είναι η σταθερότητα του μέσου όρου που παρουσιάζει η χρονοσειρά. Αν επικρατεί σταθερότητα, ο συντελεστής παραμένει μικρός, ενώ σε αντίθετη περίπτωση,

πρέπει να είναι μεγάλος, ώστε να παρακολουθούνται οι μεταβολές που παρουσιάζουν τα δεδομένα.

Πλέον ο υπολογισμός του βέλτιστου συντελεστή εξομάλυνσης γίνεται πολύ εύκολα από ταχύτατα υπολογιστικά συστήματα που χρησιμοποιούν ως μέθοδο την ελαχιστοποίηση κάποιου δείκτη ακριβείας σε επίπεδο προσαρμογής, για τους οποίους θα αναφερθούμε αναλυτικά στο επόμενο κεφάλαιο.

2.4.3 Μοντέλα Γραμμικής Παλινδρόμησης

Στον τομέα της στατιστικής, η ανάλυση της παλινδρόμησης, μελετάει τη σχέση μεταξύ μίας εξαρτημένης μεταβλητής με συγκεκριμένες ανεξάρτητες μεταβλητές, οι οποίες αποκαλούνται και επεξηγηματικές. Ανάλογα με το πλήθος των επεξηγηματικών μεταβλητών, διακρίνονται σε απλά και πολλαπλά και θεωρούνται κατάλληλα για την παραγωγή μακροπρόθεσμων προβλέψεων.

Η χρήση της ανάλυσης της παλινδρόμησης είναι αρκετά συνήθης, όταν θέλουμε να υπολογίσουμε τη μέση τιμή της προσδοκώμενης εξαρτημένης μεταβλητής, υπό την προϋπόθεση ότι οι ανεξάρτητες διατηρούνται σταθερές. Η μαθηματική συνάρτηση που προβλέπει την τιμή της εξαρτημένης, καλείται εξίσωση της παλινδρόμησης. Κατά την ανάλυση αυτή, ενδιαφέρον παρουσιάζει η διακύμανσή της γύρω από μία γραμμική εξίσωση και η οποία μπορεί να περιγραφεί από μία πιθανοτική κατανομή.

Οι πιο διαδεδομένες τεχνικές για τον προσδιορισμό της συσχέτισης μεταξύ των μεταβλητών είναι η γραμμική παλινδρόμηση και η απλή ευθεία ελάχιστων τετραγώνων.

Κατά την μέθοδο της απλής παλινδρόμησης, υποθέτουμε αρχικά την ύπαρξη σχέσης μεταξύ της μεταβλητής πρόβλεψης και μίας άλλης μεταβλητής και ότι αυτή η σχέση είναι γραμμική, μία παραδοχή που προφανώς δεν ισχύει σε πολλές περιπτώσεις. Σκοπός της απλής γραμμικής παλινδρόμησης είναι να εκφράσει τη σχέση μεταξύ δύο μεταβλητών X και Y, με την εξίσωση ευθείας γραμμής:

$$\hat{Y}_i = a + bX_i$$

όπου, α η τεταγμένη του σημείου τομής της ευθείας με τον άξονα της εξαρτημένης μεταβλητής και b η κλίση τη ευθείας, δηλαδή ο ρυθμός μεταβολής του $\hat{Y}$ ανά μοναδιαία αύξηση του X. Για n δεδομένες πραγματικές τιμές της εξαρτημένης μεταβλητής, υπολογίζουμε τους συντελεστές α και b, με σκοπό την ελαχιστοποίηση του αθρόισματος των τετραγώνων των διαφορών των πραγματικών τιμών Y από τις τιμές $\hat{Y}$ που προκύπτουν από την εξίσωση παλινδρόμησης. Αυτή είναι η μέθοδος των ελάχιστων τετραγώνων και έχει ως λογική την ελαχιστοποίηση της απόστασης των πραγματικών παρατηρήσεων Y από τη βέλτιστη γραμμή παλινδρόμησης. Οι συντελεστές υπολογίζονται με τους παρακάτω τύπους:

$$b = \frac{\frac{\sum_{i=1}^{n} X_i \cdot Y_i}{n} - \bar{X}\bar{Y}}{\frac{\sum_{i=1}^{n} {X_i}^2}{n} - \bar{X}^2} = \frac{\sum_{i=1}^{n}[(X_i - \bar{X})(Y_i - \bar{Y})]}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

$$a = \bar{Y} - b\bar{X}$$

Η απλή γραμμική παλινδρόμηση μπορεί να εφαρμοστεί και σε περιπτώσεις που οι μεταβλητές δεν συνδέονται με κάποια γραμμική σχέση, αλλά είναι σε θέση να επιτευχθεί μετασχηματισμός για να γίνει. Δεν είναι εύκολη διαδικασία, αλλά όπου μπορεί να επιτευχθεί, διευρύνει τη δυνατότητα εφαρμογής της μεθόδου για πολλούς άλλους τύπους σχέσεων.

Ο συντελεστής γραμμικής συσχέτισης υπολογίζεται από τον τύπο:

$$r_{XY} = \frac{n \cdot \sum_{i=1}^{n}(X_i \cdot Y_i) - \sum_{i=1}^{n} X_i \cdot \sum_{i=1}^{n} Y_i}{\sqrt{n \cdot \sum_{i=1}^{n} {X_i}^2 - \left(\sum_{i=1}^{n} X_i\right)^2} \cdot \sqrt{n \cdot \sum_{i=1}^{n} {Y_i}^2 - \left(\sum_{i=1}^{n} Y_i\right)^2}}$$

Αποτελεί ένα μέτρο του βαθμού συσχέτισης που μπορεί να υπάρχει μεταξύ δύο μεταβλητών και λαμβάνει τιμές από 0 έως ±1. Τιμές κοντά στο 0 είναι ένδειξη ότι δεν υπάρχει συσχέτιση και όταν πλησιάζουν το -1 ή το 1, είναι ένδειξη απόλυτης συσχέτισης. Όταν ο συντελεστής είναι μεγαλύτερος του μηδενός, οι δύο μεταβλητές καλούνται θετικά συσχετισμένες και όταν είναι μικρότερος, αρνητικά συσχετισμένες.

Ο συντελεστής αυτός, μπορεί να δείχνει την κατεύθυνση της σχέσης ανάμεσα σε δύο μεταβλητές, δηλαδή αν οι τιμές τους αυξάνονται ή μειώνονται συγχρόνως (θετικά συσχετισμένες), ή όταν η τιμή της μίας αυξάνεται, τότε η άλλη μειώνεται (αρνητικά συσχετισμένες), ή αν μεταβάλλονται ανεξάρτητα η μία από την άλλη (ασυσχέτιστες). Επίσης, μπορεί να δείχνει πόσο ισχυρή είναι η συσχέτιση των δύο μεταβλητών.

Ο συντελεστής προσδιορισμού υπολογίζεται από τον τύπο:

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{Y} - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

Ο συντελεστής αυτός, είναι πάντα θετικός, παίρνει τιμές από 0 έως 1 και αντιπροσωπεύει το ποσοστό της διακύμανσης της μεταβλητής Y, η οποία ερμηνεύεται από την ευθεία παλινδρόμησης.

Συχνά, συναντώνται περισσότερες από μία ανεξάρτητες μεταβλητές και τότε το μοντέλο της απλής γραμμικής παλινδρόμησης, μπορεί να γενικευτεί μέσω την τεχνικής της πολλαπλής παλινδρόμησης, ώστε να συμπεριλάβει όλες τις μεταβλητές που επηρεάζουν την μεταβλητή πρόβλεψης. Η γενική μορφή της πολλαπλής γραμμικής παλινδρόμησης είναι:

$$Y = b_0 + b_1 \cdot X_1 + b_2 \cdot X_2 + \cdots + b_k \cdot X_k + e$$

όπου, η μεταβλητή Y εκφράζει την εξαρτημένη μεταβλητή, οι μεταβλητές $X_1, X_2 \ldots X_k$ τις ανεξάρτητες, οι συντελεστές $b_0, b_1, \ldots b_k$ είναι σταθερές παράμετροι και το e δηλώνει τον τυχαίο παράγοντα, ο οποίος θεωρείται κανονικά κατανεμημένος γύρω από το μηδέν.

Οι συντελεστές της παλινδρόμησης υπολογίζονται με τη μέθοδο των ελάχιστων τετραγώνων μέσω της σχέσης:

$$\sum_{i=1}^{n} {e_i}^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (b_0 - b_1 X_{1,i} - \ldots - b_k X_{k,i})^2$$

Τελικά, για την εύρεση των συντελεστών $b_0, b_1, \ldots, b_k$, οι οποίοι ελαχιστοποιούν την παραπάνω ποσότητα, απαιτείται ο υπολογισμός των μερικών παραγώγων αυτής, για κάθε έναν από τους συντελεστές, η θεώρηση των υπολογισμένων παραγώγων ίσων με το μηδέν και η λύση ενός γραμμικού συστήματος τριών εξισώσεων με τρεις αγνώστους.

Ο συντελεστής πολλαπλής συσχέτισης υπολογίζεται από τον τύπο:

$$R_{Y\hat{Y}} = \frac{n \cdot \sum_{i=1}^{n}(Y_i \cdot \hat{Y}_i) - (\sum_{i=1}^{n} Y_i)(\sum_{i=1}^{n} \hat{Y}_i)}{\sqrt{n \cdot \sum_{i=1}^{n} Y_i - (\sum_{i=1}^{n} Y_i)^2}\sqrt{n \cdot \sum_{i=1}^{n} \hat{Y}_i - (\sum_{i=1}^{n} \hat{Y}_i)^2}}$$

Εκφράζει τη συσχέτιση ανάμεσα στην εξαρτημένη μεταβλητή Y και την εκτίμηση της $\hat{Y}$ με βάση τις ανεξάρτητες μεταβλητές.

Για τον υπολογισμό του συντελεστή προσδιορισμού, υπολογίζεται με τον ίδιο τύπο με την απλή γραμμική παλινδρόμηση. Με αυτό τον τρόπο, όμως, δεν λαμβάνονται υπόψιν ο αριθμός των ανεξάρτητων μεταβλητών και ο αριθμός του συνόλου των παρατηρήσεων. Για αυτό το λόγο, υπολογίζεται ο διορθωμένος συντελεστής προσδιορισμού ως εξής:

$$\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-k-1}$$

Ο διορθωμένος συντελεστής πλέον, εκφράζει το ποσοστό της διασποράς της μεταβλητής πρόβλεψης που αιτιολογείται από τις ανεξάρτητες μεταβλητές. Η διαφορά n-1 εκφάζει τους συνολικούς βαθμούς ελευθερίας της συνολικής διακύμανσης του μοντέλου, ενώ η διαφορά n-k-1, τους βαθμούς ελευθερίας της ερμηνευθείσας διακύμανσης.

2.4.4 Ολοκληρωμένα Αυτοπαλινδρομικά Μοντέλα Κινητών Μέσων Όρων

Τα Ολοκληρωμένα Αυτοπαλινδρομικά Μοντέλα Κινητών Μέσων Όρων (AutoRegressive Integrated Moving Average ή ARIMA), είναι στοχαστικά μαθηματικά μοντέλα με τα οποία μπορούμε να περιγράψουμε την εξέλιξη κάποιου μεγέθους σε ορισμένο χρονικό διάστημα και ως εκ τούτου, να προβλέψουμε την τιμή του στο μέλλον.

Τα ντετερμινιστικά μοντέλα δίνουν την δυνατότητα να υπολογίσεις ακριβώς κάποιο μελλοντικό γεγονός, χωρίς αυτό να επηρεάζεται από τυχαιότητα. Για αυτά τα μοντέλα υπάρχουν όλα τα απαραίτητα δεδομένα για να γίνει με βεβαιότητα η πρόβλεψη. Τα στοχαστικά μοντέλα έχουν την δυνατότητα να διαχειριστούν αβεβαιότητες που εφαρμόζονται στα δεδομένα. Τα συγκεκριμένα μοντέλα έχουν έμφυτη την τυχαιότητα και ακόμα και αν οι αρχικές παράμετροι και συνθήκες παραμένουν ίδιες, μπορούν να προκαλέσουν διαφορετικό τελικό αποτέλεσμα. Στη συνέχεια θα ξεχωρίζουμε τα ντετερμινιστικά και τα στοχαστικά μοντέλα αναφερόμενοι στις προβλέψεις που παράγουν, οι οποίες είναι σημειακές και πιθανοτικές αντίστοιχα.

Το πόσο σημαντικές είναι οι πιθανοτικές προβλέψεις, γίνεται ξεκάθαρο αν αναλογιστεί κανείς ότι τα ντετερμινιστικά μοντέλα δεν μπορούν να περιγράψουν τα περισσότερα φυσικά μεγέθη, καθώς αυτά εξαρτώνται συνήθως από μη ντετερμινιστικούς παράγοντες και τυχαία γεγονότα. Επομένως, γίνεται πολύ δύσκολη η ακριβής πρόβλεψη και

ένα στοχαστικό μοντέλο μπορεί να υπολογίσει την πιθανότητα να βρεθεί η τιμή ενός τέτοιου μεγέθους σε κάποιο διάστημα.

Τα μοντέλα ARIMA βρίσκουν εφαρμογή στη μελέτη πολλών μεγεθών και δίνουν καλή εικόνα για τη διαχρονική συμπεριφορά τους και ικανοποιητικά αποτελέσματα στην πρόβλεψη των μελλοντικών τιμών τους. Αυτό συμβαίνει, καθώς δεν υποθέτουν από την αρχή τον μηχανισμό με τον οποίο εξελίσσεται το μέγεθος υπό μελέτη, αλλά επιλέγει μέσα από μία πληθώρα διαθέσιμων μηχανισμών, με βάση ποιος έχει την μεγαλύτερη πιθανότητα να ανακαλύψει τα μοτίβα που παρουσιάζουν οι παρατηρήσεις σε σχέση με τις προηγούμενες. Μελετήθηκαν εκτεταμένα από τους Box και Jenkins τη δεκαετία του ΄90 (Box & Jenkins, 1994).

Πριν την ανάλυση των τριών βασικών παραμέτρων ενός μοντέλου ARIMA, είναι σημαντικό να επικεντρωθούμε στην αυτοσυσχέτιση.

Η αυτοσυσχέτιση αναφέρεται στον βαθμό συσχέτισης παρατηρήσεων της ίδιας μεταβλητής μίας χρονοσειράς με χρονική υστέρηση k χρονικές περιόδους. Ο υπολογισμός του συντελεστή αυτοσυσχέτισης μπορεί να δώσει απάντηση στο αν τα δεδομένα μίας χρονοσειράς επιδεικνύουν τυχαιότητα και πόσο σχετίζεται στατιστικά μία παρατήρηση με την αμέσως επόμενή της. Λαμβάνει τιμές στο διάστημα [0,1] και όσο μικρότερη η τιμή του, τόσο λιγότερο συσχετίζονται οι εξεταζόμενες παρατηρήσεις και αντίστοιχα, όσο πιο κοντά στη μονάδα, τόσο μεγαλύτερη συσχέτιση υπάρχει. Είναι επίσης γνωστή ως σειριακή συσχέτιση. Βοηθάει τον αλγεβρικό εντοπισμό της εποχιακής συμπεριφοράς μίας χρονοσειράς και υπολογίζεται μέσω του παρακάτω τύπου:

$$ACF_k = \frac{\sum_{i=1+k}^{n}[(Y_i - \mu)(Y_{i-k} - \mu)]}{\sum_{i=1}^{n}(Y_i - \mu)^2}$$

όπου, ACF η συνάρτηση του συντελεστή αυτοσυσχέτισης, n το μήκος της χρονοσειράς, k οι χρονικές περίοδοι ανάμεσα στις παρατηρήσεις, Y η τιμή της χρονοσειράς και μ η μέση τιμής της χρονοσειράς.

Ιδιαίτερα χρήσιμός είναι και ο μερικός συντελεστής αυτοσυσχέτισης (PACF), κατά τον οποίο επαναλαμβάνεται η ίδια διαδικασία για την εύρεση της συσχέτισης μεταξύ συγκεκριμένων παρατηρήσεων, όμως, στην περίπτωση αυτή, μπορεί να υπολογιστεί για δύο παρατηρήσεις, αγνοώντας τελείως την επίδραση που μπορεί να έχουν οι τιμές που παρεμβάλλονται.

Αυτή η δυνατότητα της μερικής αυτοσυσχέτισης, είναι ο λόγος για τον οποίο στα μοντέλα ARIMA προτιμάται η χρήση της για την κατανόηση της τάσης, της στασιμότητας και της εποχιακότητας μίας χρονοσειράς.

Κάθε μοντέλο ARIMA μπορεί να εκφραστεί ως ο γραμμικός συνδυασμός των παρακάτω μοντέλων και για αυτό τον λόγο ακολουθεί και η ανάλυση τους (Σπηλιώτης, 2020).

- Μοντέλο Αυτοπαλινδρόμησης (AR): Τα μοντέλα αυτοπαλινδρόμησης θεωρούν αποκλειστικά γραμμικές σχέσεις μεταξύ των παρατηρήσεων της χρονοσειράς και τις αξιοποιούν για την παραγωγή προβλέψεων. Για ένα τέτοιο μοντέλο, με

τάξη p, όπου p το πλήθος των παρελθοντικών παρατηρήσεων για την παραγωγή πρόβλεψης, έχουμε αλγεβρικά:

$$Y_t = c + \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + \ldots + \varphi_p Y_{t-p}$$

όπου, φ_i οι συντελεστές αυτοσυσχέτισης του μοντέλου για υστέρηση i, Y_t η τιμής της παρατήρησης την χρονική στιγμή t και c μία σταθερά, η οποία μπορεί να υπολογιστεί από τον τύπο:

$$c = \mu(1 - \varphi_1 - \varphi_2 - \ldots - \varphi_p)$$

όπου, μ η μέση τιμή των παρατηρήσεων της χρονοσειράς.

- Μοντέλο Κινητού Μέσου Όρου (MA): Χρησιμοποιεί τα παρελθοντικά σφάλματα πρόβλεψης, αντί να χρησιμοποιεί τις παλαιότερες τιμές πρόβλεψης της παλινδρόμησης για να παράγει τις προβλέψεις του. Ορίζεται ως q το μέγεθος του παραθύρου του κινητού μέσου όρου και ταυτόχρονα ως η τάξη ενός τέτοιου μοντέλου και παράγει τις προβλέψεις του, μέσω των σφαλμάτων σε αυτό το παράθυρο. Αντιλαμβάνεται την επίδραση που έχουν τα σφάλματα αυτά στην τελευταία παρατήρηση και είναι ιδιαίτερα χρήσιμο στην κατανόηση βραχυπρόθεσμων αλλαγών στα δεδομένα ή τυχαίων διακυμάνσεων. Είναι γνωστός και ως ο τυχαίος παράγοντας των μοντέλων. Πρακτικά, θεωρεί γραμμικές σχέσεις ανάμεσα στα σφάλματα και τις παρατηρήσεις και ένα τέτοιο μοντέλο, απεικονίζεται αλγεβρικά ως παρακάτω:

$$Y_t = c + \theta_1 e_{t-1} + \theta_2 e_{t-2} + \ldots + \theta_q e_{t-q}$$

όπου, θ_i οι συντελεστές μερικής αυτοσυσχέτισης του μοντέλου MA για υστέρηση i και e_t το σφάλμα πρόβλεψης την χρονική στιγμή t. Για τη σταθερά c ενός τέτοιου μοντέλου ισχύει:

$$c = \mu$$

- Μοντέλο Διαφόρισης (I) : Υπάρχει η απαίτηση η χρονοσειρά που αναλύουμε να είναι στάσιμη, δηλαδή ο μέσος όρος και η διακύμανσή της πρέπει να παραμείνουν σχετικά σταθερές μέσα σε μία χρονική περίοδο. Πρακτικά, γίνεται διαφόριση της χρονοσειράς και τα δεδομένα αντικαθίστανται αφαιρώντας την τρέχουσα τιμή του δεδομένου με την προηγούμενη. Με αυτό τον τρόπο, απελευθερωνόμαστε από την έννοια του χρόνου, αποκτούμε στασιμότητα και το μοντέλο τελικά έχει τα απαραίτητα δεδομένα χωρίς τον ανεπιθύμητο θόρυβο. Υπάρχουν και άλλοι τρόποι για την εξομάλυνση των διακυμάνσεων που παρουσιάζει κάποια χρονοσειρά, όπως είναι η χρήση μετασχηματισμών. Συνήθως αυτοί οι τρόποι εφαρμόζονται πριν την διαφόριση, η οποία αφαιρεί τάση και εποχιακότητα κυρίως και δύο από αυτούς τους μετασχηματισμούς είναι η λογαρίθμηση της χρονοσειράς και ο μετασχηματισμός Box-Cox. Επίσης, σε περίπτωση που η εποχιακότητα είναι αρκετά έντονη, μπορεί να χρησιμοποιηθεί η εποχιακή διαφόριση, όπου η παραγόμενη χρονοσειρά είναι το αποτέλεσμα της διαφοράς

μεταξύ των παρατηρήσεων της αρχικής και εκείνων των προηγούμενων αντίστοιχων εποχιακών περιόδων. Ιδιαίτερη προσοχή χρειάζεται, ώστε να μην φτάσουμε στο σημείο της υπερδιαφόρισης, κατά το οποίο ειδικά σε περιπτώσεις που δεν υπάρχουν αρκετές παρατηρήσεις, μειώνει σημαντικά την αυτοσυσχέτιση της χρονοσειράς, την οποία θα μελετήσουμε αμέσως μετά και αυξάνει ιδιαίτερα την τυχαιότητα του μοντέλου.

Για να αναλύσουμε και να προβλέψουμε στάσιμες χρονοσειρές, μπορούμε να συνδυάσουμε τα τρία παραπάνω μοντέλα και να έχουμε τελικά τα μοντέλα ARIMA(p,d,q) όπου p, d και q είναι αντίστοιχα η τάξη των AR, I και MA.

Το τελικό μοντέλο αναπαρίσταται παρακάτω με τη βοήθεια του τελεστή ολίσθησης B, ο οποίος δίνει τα δεδομένα για κάποια προηγούμενη χρονική περίοδο, η οποία ορίζεται από τον εκθέτη του:

$$\left(1-\varphi_1 B-\varphi_2 B^2-\cdots-\varphi_p B^p\right)(1-B)^n(1-B^m)^N Y_t = c+(\theta_1 B+\theta_2 B^2+\ldots+\theta_q B^q)e_t$$

Για n=N=0, η σταθερά c ισούται με:

$$c=\mu(1-\varphi_1-\varphi_2-\ldots-\varphi_p)$$

Σε οποιαδήποτε άλλη περίπτωση, η σταθερά είναι μηδενική, που είναι και το αναμενόμενο, καθώς ο μέσος όρος μίας διαφορισμένης και στάσιμης χρονοσειράς είναι ιδανικά μηδέν.

Στην πραγματικότητα, η χρονοσειρά που προκύπτει αφού γίνει η διαφόριση, δεν θα είναι ποτέ στάσιμη γύρω από το μηδέν. Επομένως, συνηθίζεται να προστίθεται μία σταθερά c ακόμα και στις διαφορισμένες χρονοσειρές. Θα αναφερθούμε στις πιο συνήθεις τιμές που μπορεί να πάρει ο δείκτης d και τι σημαίνει αυτό για τη σταθερά c σε κάθε μία περίπτωση.

- Για d=0, δηλαδή στην περίπτωση που δεν χρειάζεται διαφόριση, υπάρχει σταθερότητα στη χρονοσειρά. Ακόμη και τότε, η εισαγωγή μίας σταθεράς μπορεί να προσδιορίσει ακριβέστερα το πραγματικό επίπεδο της χρονοσειράς και να βελτιώσει την τελική πρόβλεψη. Αυτό συμβαίνει, καθώς η σταθερότητα αυτή μπορεί να μην έχει ως κέντρο της το μηδέν.
- Για d=1, δηλαδή στην περίπτωση που απαιτείται διαφόριση πρώτης τάξης, υπάρχει τάση στην αρχική χρονοσειρά, η οποία τελικά εξαλείφεται και οι προβλέψεις του μοντέλου κινούνται γύρω από ένα σταθερό επίπεδο. Χρησιμοποιώντας μία σταθερά στα συγκεκριμένα μοντέλα, μπορεί να επιστρέψει η τάση που προηγουμένως αφαιρέθηκε και να βελτιωθεί με αυτό τον τρόπο η τελική πρόβλεψη.
- Για d=2, δηλαδή στην περίπτωση που απαιτείται διαφόριση δεύτερης τάξης, υπάρχει τάση στην αρχική χρονοσειρά, η οποία είναι και χρονικά μεταβαλλόμενη. Σε αυτή την περίπτωση, δεν προτείνεται η εισαγωγή σταθεράς, εκτός εάν μπορεί να παρατηρηθεί κάποιο εκθετικό μοτίβο ανάπτυξης.

Υπάρχουν διάφοροι τρόποι αξιολόγησης των μοντέλων ARIMA, είτε οπτικοί με τη χρήσης διαγραμμάτων κατανομής πρόβλεψης και υπολειπόμενων σφαλμάτων, είτε μέσω κριτηρίων, τα οποία αξιολογούν την απόδοση συναρτήσει της πολυπλοκότητάς του. Στο πλαίσιο της παρούσας εργασίας θα επικεντρωθούμε σε δύο κριτήρια, τα οποία δείχνουν σε τι βαθμό πολυπλοκότητας μπορούμε να φτάσουμε για την παραγωγή ακριβέστερων προβλέψεων.

Γενικά, ισχύει ότι όσο αυξάνεται η πολυπλοκότητα ενός μοντέλου πρόβλεψης, τόσο μειώνεται η προκατάληψη των παραγόμενων προβλέψεων, αλλά ταυτόχρονα αυξάνονται οι διακυμάνσεις των παραγόμενων σφαλμάτων, το οποίο έχει ως άμεσο αποτέλεσμα χαμηλότερη ακρίβεια. Τα δύο κριτήρια που αναφέραμε, είναι τα Akaike's Information Criterion (AIC) και Bayesian Information Criterion (BIC).

Για το AIC, η τιμή του κριτηρίου υπολογίζεται από την παρακάτω σχέση:

$$AIC = -2\log L + 2(p + q + P + Q + k + 1)$$

όπου, L η μεγιστοποιημένη τιμή της συνάρτησης πιθανότητας των δεδομένων του μοντέλου και η παρένθεση είναι οι παράμετροι του μοντέλου με το k να είναι μηδέν αν η σταθερά του μοντέλου είναι μηδέν και σε αντίθετη περίπτωση είναι ίσο με τη μονάδα. Οι κεφαλαίες τιμές των παραμέτρων αναφέρονται σε τυχόν εποχιακές τιμές των μοντέλων που αντιστοιχούν.

Υπάρχει και η διορθωμένη εκδοχή του συγκεκριμένου κριτηρίου, η οποία δίνει μεγαλύτερο βάρος στην πολυπλοκότητα του μοντέλου και παρακάτω φαίνεται ο υπολογισμός του:

$$AIC_c = AIC + \frac{2(p + q + P + Q + k + 1)(p + q + P + Q + k + 2)}{n - p - q - P - Q - k - 2}$$

όπου, n το μέγεθος του δείγματος.

Το BIC λειτουργεί με παρόμοιο τρόπο, αλλά δίνει μεγάλο βάρος στην πολυπλοκότητα του μοντέλου ώστε να μειωθεί το φαινόμενο της υπερπροσαρμογής. Δίνεται από τον παρακάτω τύπο:

$$BIC = AIC + [\log(n) - 2](p + q + P + Q + k + 1)$$

Είναι ξεκάθαρο ότι και τα δύο κριτήρια δεν λαμβάνουν υπόψιν το αν υπάρχει διαφόριση ή όχι και πιο συγκεκριμένα τους παράγοντες d και D.

2.4.5 Μοντέλα ARIMA με Εξωγενείς Μεταβλητές

Πολλές φορές στον τομέα των προβλέψεων, προστίθεται εξωτερική πληροφορία σαν βοήθεια, ώστε το μοντέλο πρόβλεψης που χρησιμοποιείται, να κατανοήσει καλύτερα διάφορα μοτίβα της χρονοσειράς υπό μελέτη (Moslemi et al, 2023). Αν θέλουμε, για παράδειγμα, να προβλέψουμε πόσο ρεύμα θα καταναλώσει η Ελλάδα σαν χώρα σε μία ημέρα, θα χρειαστεί να παρέχουμε στο μοντέλο μας την πληροφορία για τον καιρό, καθώς είναι από τους παράγοντες που επηρεάζουν άμεσα την κατανάλωση ρεύματος. Σε περίπτωση κρύου, θα χρησιμοποιηθούν περισσότερο οι ηλεκτρικές συσκευές, με άμεσο αποτέλεσμα την αύξηση της κατανάλωσης. Για καλύτερο αποτέλεσμα, μπορούμε να παρέχουμε στο μοντέλο μας, την πληροφορία για τον καιρό για κάθε ημέρα

για την οποία υπάρχουν δεδομένα και παρελθοντικές παρατηρήσεις, έτσι ώστε να αντιληφθεί την σύνδεση του καιρού με την κατανάλωση ρεύματος και τα μοτίβα που δημιουργούνται λόγω αυτού.

Αυτή η εξωτερική πληροφορία ονομάζεται εξωγενής μεταβλητή και έχει σημαντικό ρόλο στη βελτίωση των μοντέλων πρόβλεψης. Το πρόβλημα είναι ότι τα απλά μοντέλα πρόβλεψης δεν έχουν την δυνατότητα να αξιοποιήσουν με κάποιο τρόπο αυτή την πληροφορία και έτσι παραμένουν περιορισμένα σε σχέση με αυτά που μπορούν. Τα μοντέλα ARIMA έχουν την δυνατότητα να το κάνουν και στην συγκεκριμένη περίπτωση ονομάζονται ARIMAX, δηλαδή Ολοκληρωμένα Αυτοπαλινδρομικά Μοντέλα Κινητών Μέσων Όρων με Εξωγενείς Μεταβλητές (AutoRegressive Integrated Moving Average with eXogenous variables).

Αλγεβρικά ένα τέτοιο μοντέλο αναπαρίσταται ως εξής:

$$Y_t = c + \sum_{i=1}^{p} \varphi_i Y_{t-i} + \sum_{j=1}^{q} \theta_j e_{t-j} + \sum_{k=1}^{m} \beta_k X_{t-k} + e_t$$

όπου, β οι συντελεστές των εξωγενών μεταβλητών με την επίδραση που έχουν στην αντίστοιχη μεταβλητή και X οι τιμές των εξωγενών μεταβλητών την χρονική στιγμή t.

Η πληροφορία μπορεί να είναι τιμή η οποία αλλάζει με τον χρόνο, όπως είναι το ποσοστό πληθωρισμού, κατηγορική μεταβλητή, όπως είναι οι μέρες της εβδομάδας, Boolean μεταβλητή, όπως είναι η ύπαρξη ή μη αργίας μία συγκεκριμένη ημέρα, ή ένας συνδυασμός των παραπάνω.

Τέτοιου είδους μοντέλα δεν παρουσιάζουν μόνο πλεονεκτήματα σε σχέση με απλούστερα, αλλά κρύβουν και κινδύνους για τους οποίους πρέπει να είμαστε αρκετά προσεκτικοί. Ένα από αυτά τα μειονεκτήματα, είναι η επιλογή των εξωγενών μεταβλητών, κατά την οποία πρέπει να είμαστε αρκετά σίγουροι ότι θα προσθέσουν ωφέλιμη πληροφορία στο μοντέλο μας, αντί να το γεμίσουν με περιττά δεδομένα, τα οποία θα το μπερδέψουν και το τελικό αποτέλεσμα καταλήξει ακόμα χειρότερο από ότι νωρίτερα. Αν η πληροφορία δεν έχει άμεση σχέση με τα αρχικά δεδομένα μας, θα ανακαλυφθούν μοτίβα στις χρονοσειρές μας που δεν σχετίζονται με την πραγματικότητα. Ένας άλλος βασικός περιορισμός που θα πρέπει να έχουμε υπόψιν στο χτίσιμο τέτοιων μοντέλων είναι ο αριθμός των εξωγενών μεταβλητών που θα χρησιμοποιήσουμε. Ακόμη και αν υπάρχει συσχέτιση μεταξύ αρκετών μεταβλητών και των δεδομένων μας, θα πρέπει να επιλέξουμε τις σημαντικότερες, δηλαδή αυτές που θεωρούμε ότι επηρεάζουν σε μεγαλύτερο βαθμό την τελική πρόβλεψη. Υπάρχει πιθανότητα μεγάλος αριθμός εξωτερικής πληροφορίας να οδηγήσει σε αρκετά καλό αποτέλεσμα στα δεδομένα που χρησιμοποιήθηκαν για εκπαίδευση αλλά σε πολύ φτωχό αποτέλεσμα σε καινούρια δεδομένα, πράγμα που δεν είναι το επιθυμητό, καθώς το μοντέλο μας δεν θα προσαρμόζεται ταχύτατα σε μη αναμενόμενες μελλοντικές αλλαγές. Τέλος, τα συγκεκριμένα μοντέλα υποθέτουν γραμμικότητα ανάμεσα στις εξωγενείς μεταβλητές και στα αρχικά δεδομένα μας και στην περίπτωση που δεν ισχύει αυτή η γραμμικότητα, το μοντέλο δεν θα αποδώσει στον αναμενόμενο βαθμό.

Συνοψίζοντας, τα μοντέλα ARIMAX είναι ένα ισχυρό εργαλείο για την παραγωγή προβλέψεων χρονοσειρών, προσφέροντας μία εξελιγμένη προσέγγιση με την ενσωμάτωση εξωγενών μεταβλητών. Αυτή η ικανότητά τους, τα καθιστά ιδιαίτερα πολύτιμα σε

διάφορους τομείς, από την οικονομία έως την περιβαλλοντική επιστήμη. Κατανοώντας τα συστατικά τους στοιχεία, τη μαθηματική τους βάση και τις εφαρμογές τους, οι ειδικοί μπορούν να τα αξιοποιήσουν για να αποκτήσουν βαθύτερη γνώση για μία χρονοσειρά και να κάνουν πιο ακριβείς προβλέψεις για αυτή, λαμβάνοντας υπόψιν τους περιορισμούς στη χρήση της εξωτερικής πληροφορίας.

2.5 Μοντέλα Μηχανικής Μάθησης

Η ταχύτατη πρόοδος της μηχανικής μάθησης τα τελευταία χρόνια είναι εκπληκτική και έχει βελτιώσει πράγματα και καταστάσεις στους περισσότερους τομείς της καθημερινής ζωής μας (Bojer, 2022). Ο τομέας της πρόβλεψης δεν θα μπορούσε να απουσιάζει και έχει επηρεαστεί σε μεγάλο βαθμό. Στο πρόσφατο παρελθόν αρκετοί είχαν αμφιβολίες για το κατά πόσο θα μπορούν να βοηθήσουν τα μοντέλα μηχανικής μάθησης στη διαδικασία της πρόβλεψης και συνέχιζαν να εμπιστεύονται τα απλούστερα μοντέλα σε συνδυασμό με τη δική τους κριτική πρόβλεψη. Θεωρούσαν την εμπειρία τους πιο καθοριστική και για καιρό τα αποτελέσματα των διαγωνισμών πρόβλεψης τους επιβεβαίωναν. Όμως, έχει γίνει πλέον αντιληπτό και ευρέως διαδεδομένο το πόσο σημαντικά είναι αυτά τα μοντέλα στην βελτίωση του τελικού αποτελέσματος της πρόβλεψης. Μάλιστα, όσο αυξάνεται η υπολογιστική ικανότητα, τόσο γρηγορότερα και πιο αξιόπιστα τελικά αποτελέσματα έχουμε. Η διαχείριση εκτενών συνόλων δεδομένων έχει γίνει αρκετά απλούστερη και έχουν λυθεί αρκετά προβλήματα που αντιμετωπίζονταν παλαιότερα.

Οι κλασικές μέθοδοι βασίζονται σε προκαθορισμένους κανόνες για την πρόβλεψη, σε αντίθεση με την μηχανική μάθηση, η οποία έχει την ικανότητα να αποκτάει γνώση και να προσαρμόζεται από ένα ευρύ φάσμα δεδομένων. Η βασική διαφορά των δύο, έγκειται στην διαδικασία ελαχιστοποίησης του σφάλματος. Οι μεν βασίζονται στην προδιαγραφή κάποιας σχέσης μεταξύ των δεδομένων, ενώ οι δε, δεν υποθέτουν κάποια σχέση στα δεδομένα.

Τα προβλήματα που μπορεί να αντιμετωπίσει κανείς με μεθόδους μηχανικής μάθησης είναι ο υπολογιστικός χρόνος, που είναι αρκετά μεγαλύτερος σε σχέση με τις απλούστερες μεθόδους και ότι η επεξήγηση και η ευκολία στην ερμηνεία τους δεν είναι η πιο εμφανής συνήθως. Βέβαια, ο υπολογιστικός χρόνος μειώνεται συνεχώς με την εξέλιξη της τεχνολογίας σε αυτόν τον τομέα και δεν αποτελεί πρακτικά πρόβλημα που δυσκολεύει την τελική πρόβλεψη. Η ερμηνεία γίνεται δύσκολη, καθώς απαιτείται προεπεξεργασία των δεδομένων, συντονισμός των υπερπαραμέτρων κάθε μεθόδου και δεν υπάρχει κοινή ορολογία για τα μέρη του σχεδιασμού της κάθε μεθόδου, ούτε και η πλήρη ανάλυση της.

Οι μέθοδοι πρόβλεψης με μηχανική μάθηση που έχουν χρησιμοποιηθεί με μεγάλη επιτυχία στους πιο πρόσφατους διαγωνισμούς πρόβλεψης, μπορούν να διαχωριστούν στις εξής δύο κατηγορίες:

- Μηχανική μάθηση που βασίζεται σε παλινδρόμηση.
- Μέθοδοι πρόβλεψης βασισμένοι σε νευρωνικά δίκτυα, οι οποίες χρησιμοποιούν αρχιτεκτονικές που επιτρέπουν την άμεση επεξεργασία χρονοσειρών και τη δημιουργία χρήσιμων αναπαραστάσεων από αυτές.

Στο πλαίσιο της παρούσας εργασίας, θα επικεντρωθούμε στην πρώτη κατηγορία και πιο συγκεκριμένα στη μηχανική μάθηση που είναι βασισμένη σε δέντρα (Spiliotis, 2022).

2.5.1 Μηχανική Μάθηση βασισμένη σε Δέντρα

Τα μοντέλα μηχανικής μάθησης βασισμένα σε δέντρα, είναι μία οικογένεια αλγορίθμων μηχανικής μάθησης που χρησιμοποιούν δέντρα αποφάσεων για την επίλυση προβλημάτων παλινδρόμησης και ταξινόμησης. Με τη χρήση επεξηγηματικών μεταβλητών, παράγουν πρόβλεψη για μία εξαρτημένη μεταβλητή. Οι τελικές προβλέψεις διαμορφώνονται από ένα σύνολο κανόνων που καθορίζουν το πως κατηγοριοποιούνται τα δεδομένα.

Ο τρόπος λειτουργίας τους βασίζεται στον αναδρομικό διαχωρισμό του χώρου των χαρακτηριστικών σε μικρότερες περιοχές και στην προσαρμογή απλών μοντέλων σε αυτές τις περιοχές. Τα συγκεκριμένα μοντέλα, είναι εύκολα στην ερμηνεία και ικανά να καταγράφουν πολύπλοκες, μη γραμμικές σχέσεις στα δεδομένα.

Ένα δέντρο απόφασης αποτελείται από τη ρίζα, τα κλαδιά και τα φύλλα. Η ρίζα είναι ο κόμβος που ξεκινάει το γράφημα του δέντρου και είναι η μεταβλητή που διαχωρίζει βέλτιστα το δέντρο μας, τουλάχιστον ως προς τον δείκτη ακριβείας που χρησιμοποιούμε για την αξιολόγησή του. Η ρίζα αυτή διαχωρίζεται σε δύο κλαδιά, το οποίο καθορίζεται αυτόματα από τον αλγόριθμο δημιουργίας του δέντρου με βάση κάποιον κανόνα ή κάποια συνθήκη. Τα δεδομένα μας ανατίθενται σε κάποιο από τα δύο κλαδιά, με βάση το αν πληρούν την συνθήκη που ορίστηκε ή όχι. Κάθε κλαδί μπορεί να διαχωριστεί ξανά σε δύο νέα κλαδιά με βάση κάποιον άλλο κανόνα ή να καταλήξει σε κάποιον τερματικό κόμβο, τον οποίο ονομάζουμε φύλλο και ο οποίος δεν επιδέχεται περεταίρω διάσπαση. Η συγκεκριμένη διαδικασία επαναλαμβάνεται μέχρι να επιτευχθεί μία προκαθορισμένη προϋπόθεση για τον τερματισμό της.

Οι συνθήκες για την δόμηση ενός δέντρου ονομάζονται υπερπαράμετροι και οι βασικότεροι είναι:

- Ο μέγιστος αριθμός φύλλων.
- Το μέγιστο βάθος του δέντρου, δηλαδή πόσους διαχωρισμούς μπορεί να κάνει ένα δέντρο από τη ρίζα του έως ένα φύλλο.
- Ο ελάχιστος αριθμός παρατηρήσεων που απαιτούνται για την δημιουργία ενός φύλλου.
- Ο ελάχιστος αριθμός παρατηρήσεων ενός κόμβου ώστε να υπάρξει προσπάθεια να γίνει διαχωρισμός.

Υπάρχουν τρεις μεγάλες κατηγορίες μοντέλων μηχανικής μάθησης που βασίζονται σε δέντρα και είναι οι εξής:

- Το Δέντρο Παλινδρόμησης (Regression Tree), το οποίο χωρίζει τα δεδομένα του σε υποσύνολα με βάση τις τιμές των χαρακτηριστικών και σχηματίζει με αυτό τον τρόπο μία δομή δέντρου, όπου κάθε εσωτερικός κόμβος αντιπροσωπεύει έναν κανόνα απόφασης, κάθε κλαδί αντιστοιχεί σε ένα αποτέλεσμα του

κανόνα αυτού και κάθε φύλλο αποτελεί μία πρόβλεψη. Χρησιμοποιούνται τόσο για ταξινόμηση, όσο και για παλινδρόμηση, αλλά στην συγκεκριμένη περίπτωση μας ενδιαφέρει μόνο η παλινδρόμηση. Το μοντέλο ξεκινάει από τη ρίζα και διαχωρίζει τα δεδομένα με βάση ποιο χαρακτηριστικό οδηγεί στην μεγαλύτερη μείωση της διακύμανσης της παλινδρόμησης. Συνεχίζει να χωρίζεται αναδρομικά, μέχρι να ικανοποιηθεί μία από τις συνθήκες που αναφέρθηκαν και νωρίτερα. Είναι αρκετά εύκολο να κατανοηθεί η λειτουργία του και δεν υπάρχει ανάγκη για κλιμάκωση των χαρακτηριστικών του. Όμως, υπάρχει μεγάλος κίνδυνος να προσαρμοστεί σε μεγάλο βαθμό στα δεδομένα στα οποία εκπαιδεύτηκε και να μην μπορεί να προσαρμοστεί σε τυχούσες απότομες αλλαγές στο μέλλον. Επίσης, μπορεί να είναι ασταθές, καθώς μικρές αλλαγές στα δεδομένα, ίσως οδηγήσουν σε διαφορετικούς διαχωρισμούς.

- Το Τυχαίο Δάσος (Random Forest), το οποίο είναι ένα σύνολο πολλαπλών δέντρων αποφάσεων, όπου κάθε δέντρο εκπαιδεύεται σε ένα τυχαίο υποσύνολο δεδομένων και για ένα τυχαίο υποσύνολο χαρακτηριστικών. Για την παλινδρόμηση, που είναι αυτή που μας ενδιαφέρει, οι προβλέψεις γίνονται με τον μέσο όρο των εξόδων όλων των δέντρων. Η δημιουργία των υποσυνόλων γίνεται με τυχαίο τρόπο, είτε είναι υποσύνολο δεδομένων, είτε επεξηγηματικών μεταβλητών και σε κάθε δέντρο είναι διαφορετικά στις περισσότερες περιπτώσεις. Αυτό το μοντέλο μειώνει την πιθανότητα υπερπροσαρμογής που αναλύσαμε στο προηγούμενο μοντέλο και είναι σε θέση να διαχειριστεί μεγάλα σύνολα δεδομένων και χώρους υψηλών διαστάσεων. Τα αρνητικά του, είναι ότι είναι ακριβό σε υπολογιστικούς όρους και ότι δεν είναι το ίδιο εύκολα ερμηνεύσιμο όπως ένα απλό δέντρο απόφασης. Κατά την χρησιμοποίηση του Τυχαίου Δάσους, απαιτείται να προσδιοριστούν ο συνολικός αριθμός δέντρων των οποίων οι προβλέψεις θα δημιουργήσουν τον τελικό μέσο όρο, ο αριθμός των χαρακτηριστικών που θα χρησιμοποιηθούν τυχαία στην εκπαίδευση κάθε δέντρου και όπως και νωρίτερα, οι τιμές των υπερπαραμέτρων που ελέγχουν την ανάπτυξη κάθε δέντρου. Ένας από τους τρόπους να μειωθεί το υπολογιστικό κόστος, είναι η δημιουργία σχετικών ρηχών δέντρων, τα οποία αποκαλούνται και αδύναμοι μαθητές (weak learners), καθώς το μικρό τους βάθος, περιορίζει τη μαθησιακή τους ικανότητα και την εξειδίκευσή τους.
- Η Ενίσχυση Κλίσης (Gradient Boosting), χτίζει δέντρα διαδοχικά, με κάθε νέο δέντρο να διορθώνει τα σφάλματα του προηγούμενου. Με αυτό τον τρόπο, ελαχιστοποιεί την συνάρτηση απώλειας. Συνήθως συνδυάζει τους λεγόμενους αδύναμους μαθητές, ώστε να δημιουργήσει έναν πιο δυνατό συνδυασμό. Ακολουθεί την ίδια λογική με το Τυχαίο Δάσος, δηλαδή το αποτέλεσμά του προκύπτει από τον μέσο όρο όλων των δέντρων που δημιουργήθηκαν. Όμως, από την στιγμή που αυτά τα δέντρα δημιουργούνται διαδοχικά και καθώς το ένα μαθαίνει από το προηγούμενο, παράγονται πιο εξειδικευμένα αποτελέσματα και λαμβάνει υπόψιν περισσότερα χαρακτηριστικά, καθώς και πιο περίπλοκες σχέσεις μεταξύ των δεδομένων. Τα συγκεκριμένα μοντέλα, έχουν το πλεονέκτημα να μπορούν να διαχειριστούν ιδιαίτερα επιτυχημένα μη γραμμικές σχέσεις μεταξύ των δεδομένων και των εξωγενών μεταβλητών και έχουν πολύ υψηλή προγνωστική ακρίβεια. Η ύπαρξη πολλών μικρών δέντρων βοηθάει στην αποφυγή της υπερπροσαρμογής. Όπως και το Τυχαίο Δάσος, είναι πιο περίπλοκο από ένα

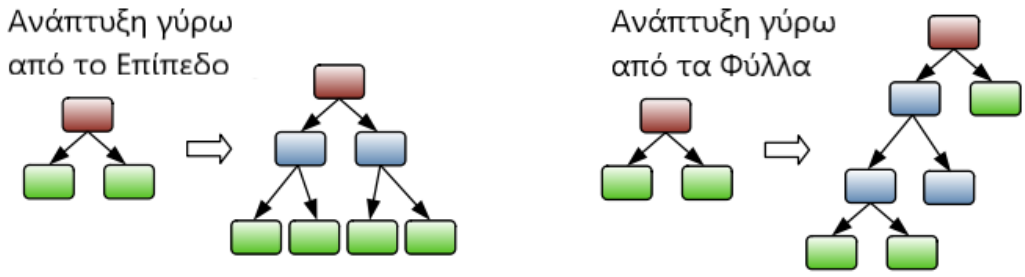
απλό δέντρο και απαιτεί ιδιαίτερη προσοχή στον καθορισμό των υπερπαραμέτρων του.

Από τις τρεις κατηγορίες που αναλύσαμε ήδη, θα εστιάσουμε περισσότερο στην τελευταία, καθώς από αυτή προέρχεται το βασικό μοντέλο που θα χρησιμοποιήσουμε στο πειραματικό μέρος της παρούσας εργασίας.

Οι συγκεκριμένοι αλγόριθμοι ακολουθούν μία από τις δύο επόμενες προσεγγίσεις, της ανάπτυξης γύρω από τα φύλλα και της ανάπτυξης γύρω από το επίπεδο.

Η ανάπτυξη γύρω από τα φύλλα, ξεκινάει με την εκπαίδευση ενός μόνου δέντρου, για το οποίο ο αλγόριθμος αξιολογεί την ακρίβεια του και προσδιορίζει ποιο από τα φύλλα του ευθύνεται σε μεγαλύτερο βαθμό για το συνολικό του σφάλμα πρόβλεψης. Στη συνέχεια, εστιάζει στο προβληματικό αυτό φύλλο και επεκτείνει τη λύση του, μοντελοποιώντας τα δεδομένα του σε κάποιο άλλο δέντρο. Πιο συγκεκριμένα, το επόμενο δέντρο χρησιμοποιεί τα δεδομένα, τα οποία απέτυχε να προβλέψει με ακρίβεια προηγουμένως. Με αυτό τον τρόπο, ο αλγόριθμος, αντί να προσπαθεί να προβλέψει ακριβώς τον στόχο που είχε, δημιουργεί ένα καινούριο δέντρο, το οποίο επιχειρεί να μοντελοποιήσει το υπολειπόμενο σφάλμα των προβλέψεων του προηγούμενου. Έτσι, αποσυνθέτει τη διαδικασία ακριβής στόχευσης της μεταβλητής που προβλέπει σε πολλές απλούστερες υποδιεργασίες και καθιστά την τελική εκτίμησή της πιο διαχειρίσιμη. Σε κάθε επακόλουθο βήμα, οι προβλέψεις του αλγορίθμου, θα είναι ένας σταθμισμένος μέσος όρος που παρέχεται από το αρχικό και το ακόλουθο δέντρο που δημιουργείται και αυτή η διαδικασία είναι η ενίσχυση. Συνεχίζεται η ίδια διαδικασία μέχρι τα κέρδη στον δείκτη ακριβείας που χρησιμοποιούμε να είναι ελάχιστα σε σχέση με την πολυπλοκότητα που προστίθεται τελικά στο μοντέλο μας.

Η ενίσχυση μπορεί να εφαρμοστεί και σε ανάπτυξη γύρω από το επίπεδο, όπως έχουμε ήδη αναφέρει. Σε αυτή την περίπτωση, όλα τα φύλλα του προηγούμενου επιπέδου υπόκεινται σε ενίσχυση, χωρίς να δίνεται προτεραιότητα σε μεμονωμένα φύλλα. Η συγκεκριμένη προσέγγιση, είναι σε θέση να παράγει προβλέψεις λιγότερο επιρρεπείς στο φαινόμενο της υπερπροσαρμογής και λειτουργούν καλύτερα σε μικρότερα σύνολα δεδομένων.



Σχήμα 2.6 : Απεικόνιση ανάπτυξης επιπέδου και ανάπτυξης φύλλου.

Στο παραπάνω σχήμα, απεικονίζεται πως λειτουργεί η κάθε ανάπτυξη που αναλύσαμε νωρίτερα. Η ανάπτυξη γύρω από τα φύλλα επικεντρώνεται σε μόνο ένα φύλλο του δέντρο σε κάθε επανάληψη, ενώ αυτή του επιπέδου σε όλα τα διαθεσίμα φύλλα του

δέντρου. Ο κόμβος της ρίζα είναι καφέ, ενώ οι κόμβοι των φύλλων είναι πράσινοι. Οι υπόλοιποι κανονικοί κόμβοι είναι μπλε (Choudhury, Mondal & Sarkar, 2024).

Οι βασικές συνθήκες που πρέπει να προσέξει κάποιος για να εφαρμόσει αλγορίθμους Ενίσχυσης Κλίσης, είναι η ρύθμιση των υπερπαραμέτρων του συνολικού αριθμού δέντρων που θα χρησιμοποιηθούν για την διόρθωση του υπολειπόμενου σφάλματος και το ποσοστό εκμάθησης, το οποίο εκφράζει πόσο γρήγορα απαιτείται να διορθωθεί το υπολειπόμενο σφάλμα από το ένα δέντρο στο επόμενο. Το συγκεκριμένο ποσοστό ελέγχει το μέγεθος της αλλαγής που συνεισφέρει κάθε προστιθέμενο δέντρο στην τελική πρόβλεψη. Χαμηλή τιμή του, σημαίνει πιο αργή εκμάθηση και πιο ακριβείς προσαρμογές σε κάθε βήμα, αλλά ταυτόχρονα προσθέτει υπολογιστικό κόστος, καθώς ίσως χρειαστούν περισσότερα δέντρα για αποτελεσματικότερη εκπαίδευση του μοντέλου.

2.5.2 Αλγόριθμος LightGBM

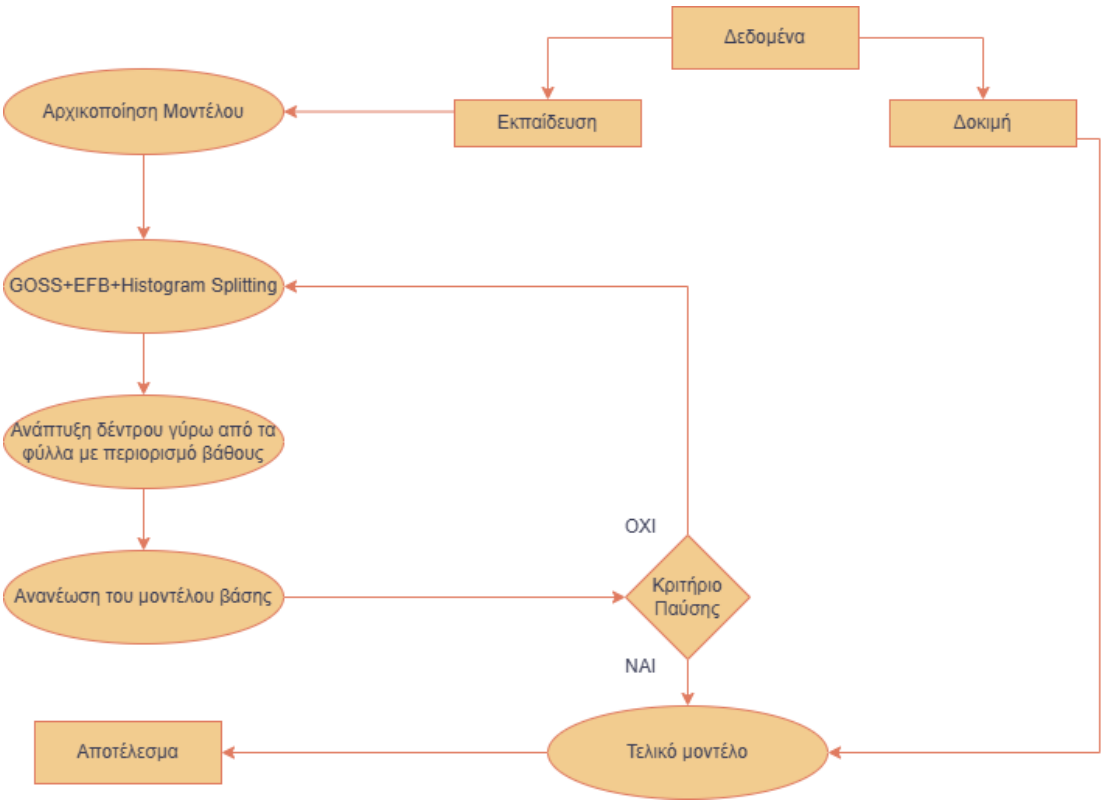
Στην κατηγορία της Ενίσχυσης Κλίσης ανήκει ο αλγόριθμος Light Gradient Boosting Machine και συγκεκριμένα στην ανάπτυξη γύρω από τα φύλλα (Ke et al, 2017). Στηρίζεται σε τρεις καινοτόμες τεχνικές, οι οποίες αναλύονται στη συνέχεια:

- Μονόπλευρη Δειγματοληψία βάσει Κλίσης (Gradient-based One-Side Sampling ή GOSS). Γενικά, στην κατηγορία Ενίσχυσης Κλίσης, όσο μικρότερη είναι η κλίση, τόσο καλύτερη γίνεται η εφαρμογή του δέντρου στο μοντέλο. Σε αυτή την τεχνική, όμως, δημιουργείται ένα υποσύνολο δεδομένων που διατηρεί κυρίως τα δέντρα που έχουν τη μεγαλύτερη κλίση και απορρίπτει, μέσω τυχαίας δειγματοληψίας, αυτά με τη μικρότερη. Πρακτικά, διατηρεί ένα ποσοστό $(a\ x\ 100\%)$ με την μεγαλύτερη κλίση και στη συνέχεια, επιλέγει τυχαία ένα άλλο ποσοστό $(b\ x\ 100\%)$ από τα υπόλοιπα δεδομένα και δημιουργεί ένα νέο υποσύνολο. Ισχύει ότι $a, b\ \epsilon\ [0,1]$. Καθώς υπολογίζει το κέρδος πληροφορίας, ο αλγόριθμος πολλαπλασιάζει με τον παράγοντα $\frac{1-a}{b}$ τις περιπτώσεις με την μικρότερη διαβάθμιση. Με αυτό τον τρόπο, οι περιπτώσεις που βρίσκονται υπό εκπαίδευση, λαμβάνουν περισσότερη προσοχή και ταυτόχρονα, το αρχικό σύνολο δεδομένων δεν αποκτάει κάποιο βάρος.
- Διαίρεση Ιστογράμματος (Histogram Splitting). Για την δημιουργία ενός κλαδιού, ο συγκεκριμένος αλγόριθμος τοποθετεί τις τιμές του χαρακτηριστικού σε δύο ή περισσότερα σύνολα, αντί να αξιολογήσει όλα τα πιθανά σημεία διαχωρισμού για κάθε χαρακτηριστικό. Για καλύτερη κατανόηση, θα δώσουμε ένα παράδειγμα. Ας υποθέσουμε ότι για ένα συγκεκριμένο σύνολο δεδομένων, υπάρχει το χαρακτηριστικό της ηλικίας και οι τιμές είναι 40,41,45,47,54,55,65 και 78. Η προσέγγιση που θα είχε ένας απλός αλγόριθμος της κατηγορίας της Ενίσχυσης Κλίσης, θα ήταν να δημιουργήσει ένα δέντρο απόφασης, ταξινομώντας την τιμή για την οποία θα πετύχαινε την καλύτερη πρόβλεψη. Στην περίπτωση της Διαίρεσης Ιστογράμματος, όμως, ο αλγόριθμος διαχωρίζει το σύνολο δεδομένων σε τουλάχιστον δύο σύνολα, ας πούμε τα 40-60 και 61-80 και συνεχίζει με την ανάλυσή του, πράγμα που κάνει τον αλγόριθμο αρκετά πιο γρήγορο και γίνεται ιδιαίτερα χρήσιμο σε πολύ εκτενέστερα σύνολα δεδομένων.
- Αποκλειστική Ομαδοποίηση Χαρακτηριστικών (Exclusive Feature Bundling ή EFB). Η πολυπλοκότητα ενός αλγορίθμου αυξάνεται, όσο προστίθενται χαρακτηριστικά. Για την βελτίωση της αποτελεσματικότητάς του, αναζητά

αποκλειστικά χαρακτηριστικά στο σύνολο δεδομένων του, δηλαδή χαρακτηριστικά που δεν λαμβάνουν μηδενική τιμή ταυτόχρονα. Ένα παράδειγμα τέτοιων χαρακτηριστικών σε ένα σύνολο δεδομένων, είναι η ύπαρξη έκπτωσης ή όχι σε ένα προϊόν. Αν υπάρχει έκπτωση, θα λαμβάνει η μία στήλη θα λαμβάνει την τιμή 1 και η άλλη στήλη την τιμή 0 και σε αντίθετη περίπτωση, θα γίνεται το αντίστροφο. Οπότε δεν μπορούν να έχουν την ίδια τιμή ταυτόχρονα. Η τεχνική αυτή, συνδυάζει προσεκτικά τα αποκλειστικά αυτά χαρακτηριστικά και δημιουργεί ένα νέο πακέτο, το οποίο μειώνει αποτελεσματικά τα χαρακτηριστικά του συνόλου δεδομένων, χωρίς να θυσιάζεται η ακρίβεια.

Το πρόβλημα που μπορεί να αντιμετωπίσει ένας LightGBM αλγόριθμος, έγκειται στο γεγονός ότι αναπτύσσεται γύρω από τα φύλλα του, πράγμα που μπορεί να οδηγήσει σε πιο περίπλοκες δομές δέντρων και στο φαινόμενο της υπερπροσαρμογής (Januschowski et al, 2021).

Για βαθύτερη κατανόηση του αλγορίθμου, φαίνεται παρακάτω το διάγραμμα ροής του. Γίνονται λίγο πιο ξεκάθαρα τα βήματα και η λογική που ακολουθεί για να δώσει το βέλτιστο αποτέλεσμα. Το κριτήριο παύσης που αναφέρεται, μπορεί να διαφέρει κάθε φορά ή να είναι ένα σύνολο κριτηρίων που ορίζουμε στην αρχή, ανάλογα τι ζητάμε εμείς από τον αλγόριθμο.



Σχήμα 2.7 : Διάγραμμα ροής αλγορίθμου LightGBM.

2.6 Δείκτες Ακρίβειας

Το σημαντικότερο κομμάτι για την παραγωγή καλών και αξιόπιστων προβλέψεων, είναι η δυνατότητα σύγκρισης των διαφόρων μοντέλων πρόβλεψης που χρησιμοποιήθηκαν κυρίως ως προς την ακρίβειά τους, αλλά και συχνά ως προς την αμεροληψία τους.

Η εφαρμογή μίας μεθόδου πρόβλεψης σε μία χρονοσειρά όπου είναι διαθέσιμες οι παρατηρήσεις Y_1 έως Y_n, έχει ως αποτέλεσμα τον υπολογισμό των τιμών F_1 έως F_{n+h}. Με αυτό τον τρόπο, το διάνυσμα F δηλώνει τις προβλέψεις πλήθους $n + h$, με n το πλήθος των τιμών της εξεταζόμενης χρονοσειράς και h, ο ζητούμενος χρονικός ορίζοντας πρόβλεψης.

Το σύνολο των τιμών πρόβλεψης, αποτελείται από δύο μέρη, αυτό που περιέχει τις n παρατηρήσεις και για τις οποίες υπάρχουν διαθέσιμα και τα αντίστοιχα πραγματικά δεδομένα και το οποίο καλείται προσαρμογή του μοντέλου πρόβλεψης (forecast model fitting) και από αυτό που περιέχει τις μελλοντικές προβλέψεις.

Με βάση τις τιμές των διανυσμάτων Y και F, μπορούν να υπολογιστούν οι βασικότεροι δείκτες ακριβείας, οι οποίοι βασίζονται στον παρακάτω ορισμό του σφάλματος, δηλαδή της διαφοράς μεταξύ της πραγματικής τιμής και της πρόβλεψης για μία περίοδο:

$$e_i = Y_i - F_i$$

Είναι ξεκάθαρο ότι, ο υπολογισμός της τιμής του σφάλματος μπορεί να γίνει μόνο για τις τιμές που έχουμε υπολογίσει μοντέλο πρόβλεψης (in-sample error), δηλαδή μέχρι την παρατήρηση Y_n της εξεταζόμενης χρονοσειράς και μελλοντικά, όταν γίνουν διαθέσιμα τα πραγματικά δεδομένα της χρονοσειράς, δηλαδή οι παρατηρήσεις Y_{n+1} έως Y_{n+h}, είναι σε θέση να υπολογιστεί τόσο το σφάλμα για το μοντέλο πρόβλεψης, όσο και το πραγματικό σφάλμα (out-of-sample error).

Χρονική Περίοδος	**Δεδομένα**	**Πρόβλεψη**	
1	Y1	F1	
2	Y2	F2	
…	…	…	
n	Yn	Fn	in sample
n+1	Yn+1	Fn+1	
…	…	…	out of
n+h	Yn+h	Fn+h	sample

Πίνακας 2.1 : Δεδομένα και προβλέψεις χρονοσειράς.

Στη συνέχεια του παρόντος κεφαλαίου θα αναφερθούμε επιγραμματικά στους βασικότερους δείκτες σφαλμάτων που χρησιμοποιούνται στην αξιολόγηση των προβλέψεων και θα επικεντρωθούμε σε αυτόν τον δείκτη ακριβείας που θα χρησιμοποιήσουμε για την αξιολόγηση των δικών μας μοντέλων.

2.6.1 Σφάλματα Εξαρτώμενα από Κλίμακα

Έχοντας ορίσει προηγουμενως το σφάλμα πρόβλεψης, προκύπτουν άμεσα τα εξαρτώμενα από την κλίμακα σφάλματα μοντέλου ή πρόβλεψης ως προς τη χρονοσειρά ενδιαφέροντος, τα οποία προκύπτουν από τον μέσο όρο των σφαλμάτων ή των παραλλαγών αυτών, όπως είναι η απόλυτη τιμή τους ή ο τετραγωνισμός τους. Είναι ιδιαιτέρως χρήσιμα, καθώς εκφράζονται στις μονάδες της εκάστοτε χρονοσειράς, αλλά παρουσιάζουν αδυναμία στη σύγκριση προβλέψεων χρονοσειρών με διαφορετική κλίμακα.

2.6.1.1 Μέσο Σφάλμα

Το μέσο σφάλμα (mean error), υπολογίζεται από τον παρακάτω τύπο:

$$ME = \frac{1}{n}\sum_{i=1}^{n}(Y_i - F_i)$$

Εκφράζει ένα μέτρο συστηματικότητας του σφάλματος και αν η τιμή του είναι κοντά στο μηδέν, αυτό δείχνει ότι τα σφάλματα είναι τυχαία και όχι συστηματικά. Όταν παίρνει θετικές τιμές, δηλώνει απαισιοδοξία στις προβλέψεις, καθώς αυτό σημαίνει ότι οι προβλέψεις ήταν κατά μέσο όρο μικρότερες των πραγματικών τιμών, ενώ όταν παίρνει αρνητικές τιμές, δηλώνει αντιστοίχως αισιοδοξία. Είναι γνωστός και ως δείκτης προκατάληψης.

2.6.1.2 Μέσο Απόλυτο Σφάλμα

Το μέσο απόλυτο σφάλμα (mean absolute error), υπολογίζεται από τον παρακάτω τύπο:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|Y_i - F_i|$$

Η βασική διαφορά με τον προηγούμενο, είναι ότι υποδηλώνει ένα μέσο μέτρο αστοχίας της πρόβλεψης, χωρίς να δίνει έμφαση στην κατεύθυνση της.

2.6.1.3 Μέσο Τετραγωνικό Σφάλμα

Το μέσο τετραγωνικό σφάλμα (mean squared error), υπολογίζεται από τον παρακάτω τύπο:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - F_i)^2$$

Το συγκεκριμένο σφάλμα δίνει πολύ μεγαλύτερο βάρος στα μεγάλα σφάλματα μίας πρόβλεψης, δεδομένου ότι αυτά τετραγωνίζονται. Χρησιμοποιείται ευρέως για τον υπολογισμό των βέλτιστων παραμέτρων εξομάλυνσης.

2.6.1.4 Ρίζα Μέσου Τετραγωνικού Σφάλματος

Η ρίζα μέσου τετραγωνικού σφάλματος (root mean squared error), υπολογίζεται από τον παρακάτω τύπο:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - F_i)^2}$$

Έχει παρόμοια χρήση με το MSE, αλλά το συγκεκριμένο είναι εκφρασμένο στις αρχικές μονάδες της χρονοσειράς.

2.6.2 Ποσοστιαία Σφάλματα

Τα σφάλματα που εξαρτώνται από την κλίμακα, δεν δίνουν πληροφορία για την αναλογία τους σφάλματος ως προς το επίπεδο της στάθμης των παρατηρήσεων. Για αυτό τον λόγο, κρίνεται απαραίτητος ο υπολογισμός των σφαλμάτων σε ποσοστιαία μορφή. Αυτός ο τρόπος είναι ιδιαίτερα χρήσιμος, όταν επιθυμούμε να συγκρίνουμε την ακρίβεια μίας μεθόδου πρόβλεψης, η οποία εφαρμόζεται σε περισσότερες από μία χρονοσειρές διαφορετικού επιπέδου μέσης τιμής ή όταν οι πραγματικές τιμές μίας χρονοσειράς είναι αρκετά υψηλές.

2.6.2.1 Μέσο Απόλυτο Ποσοστιαίο Σφάλμα

Το μέσο απόλυτο ποσοστιαίο σφάλμα (mean absolute percentage error), υπολογίζεται από τον παρακάτω τύπο:

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{Y_i - F_i}{Y_i}\right| \cdot 100(\%)$$

Όσο μικρότερη η τιμή του, τόσο καλύτερη η απόδοση του μοντέλου. Είναι ξεκάθαρο, όμως, ότι δεν μπορεί να εφαρμοστεί σε δεδομένα διακοπτόμενης ζήτησης.

2.6.2.2 Συμμετρικό Μέσο Απόλυτο Ποσοστιαίο Σφάλμα

Το συμμετρικό μέσο απόλυτο ποσοστιαίο σφάλμα (symmetric mean absolute percentage error), υπολογίζεται από τον παρακάτω τύπο:

$$sMAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{Y_i - F_i}{(\frac{Y_i + F_i}{2})}\right| \cdot 100(\%)$$

Το πρόβλημα του συγκεκριμένου δείκτη είναι ότι, δεν είναι όσο συμμετρικός δηλώνει το όνομά του και δεν υπάρχει πραγματικός διαχωρισμός στις αισιόδοξες και απαισιόδοξες προβλέψεις.

2.6.3 Σχετικά Σφάλματα

Τα σφάλματα πρόβλεψης που έχουμε ήδη αναφέρει, μπορούν να κανονικοποιηθούν ως προς το σφάλμα μίας μεθόδου αναφοράς (benchmark). Ο σκοπός είναι να εξακριβωθεί αν η μέθοδος πρόβλεψης που χρησιμοποιείται, οδηγεί σε βελτίωση της προβλεπτικής ικανότητας.

2.6.3.1 Μέσο Απόλυτο Κανονικοποιημένο Σφάλμα

Το μέσο απόλυτο κανονικοποιημένο σφάλμα (mean absolute scaled error), υπολογίζεται από τον παρακάτω τύπο:

$$MAsE = \frac{\frac{1}{n}\sum_{i=1}^{n}|Y_i - F_i|}{\frac{1}{n-1}\sum_{i=2}^{n}|Y_i - Y_{i-1}|}$$

Η κανονικοποίηση του συγκεκριμένου δείκτη γίνεται με την απλοϊκή μέθοδο. Δίνει την ίδια βαρύτητα σε μικρά και μεγάλα σφάλματα και αντιμετωπίζει τις περιπτώσεις

απροσδιοριστίας που προκύπτουν με τα ποσοστιαία σφάλματα που αναλύσαμε προηγουμένως.

2.6.3.2 Ρίζα Μέσου Τετραγωνικού Κανονικοποιημένου Σφάλματος

Η ρίζα μέσου τετραγωνικού κανονικοποιημένου σφάλματος (root mean squared error), υπολογίζεται από τον παρακάτω τύπο:

$$RMSSE = \sqrt{\frac{\frac{1}{n}\sum_{i=1}^{n}(Y_i - F_{i)}{}^2}{\frac{1}{n-1}\sum_{i=2}^{n}(Y_i - Y_{i-1})^2}}$$

Επιδεικνύει παρόμοια χαρακτηριστικά με τον MAsE, με βασική διαφορά το μεγαλύτερο βάρος που δίνει στα μεγαλύτερα σφάλματα λόγω των τετραγώνων των διαφορών πρόβλεψης και παρατήρησης.

2.6.4 Δείκτες Σφάλματος Πιθανοτικών Προβλέψεων

Με την αύξηση των πιθανοτικών μοντέλων πρόβλεψης, καθίσταται επιτακτική η ανάγκη κατάταξής τους, για να γίνει η τελική επιλογή αυτού που επιφέρει τα καλύτερα αποτελέσματα. Η επιλογή μοντέλου για τις σημειακές προβλέψεις είναι αρκετά πιο ξεκάθαρη διαδικασία, όμως τα πράγματα γίνονται αρκετά πιο περίπλοκα για τις πιθανοτικές προβλέψεις (Long et al, 2024).

2.6.4.1 Συνάρτηση Απώλειας Εκατοστημορίων

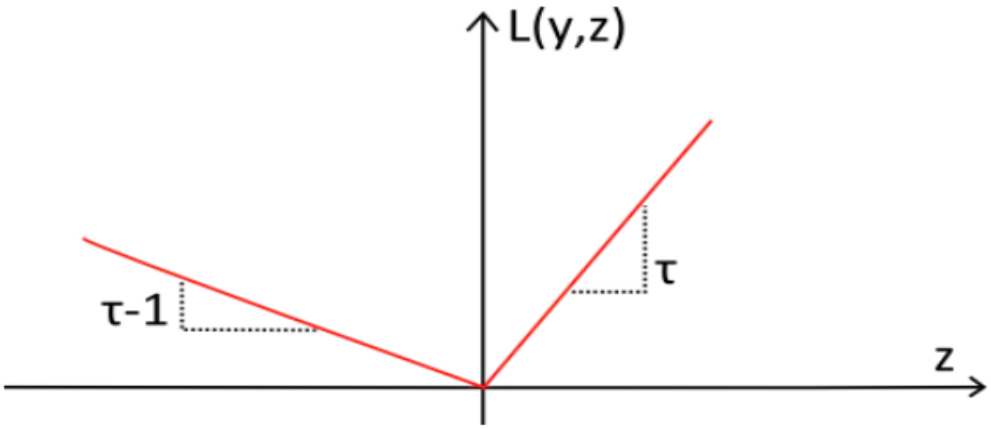
Η συνάρτηση απώλειας εκατοστημορίων (quantile loss function) ή αλλιώς Pinball Loss, αξιολογεί την απόδοση ενός μοντέλου πρόβλεψης, μετρώντας την απόκλιση μεταξύ των προβλεπόμενων εκατοστημορίων και των πραγματικών παρατηρήσεων. Η συγκεκριμένη συνάρτηση είναι ασύμμετρη και τιμωρεί την αισιοδοξία και την απαισιοδοξία των προβλέψεων διαφορετικά, με βάση το επιλεγμένο εκατοστημόριο.

Μαθηματικά εκφράζεται ως:

$$L_\tau(y, z) = (y - z)\tau, \alpha\nu\ y \geq z$$
$$= (z - y)(1 - \tau), \alpha\nu\ y < z$$

όπου, τ το ζητούμενο εκατοστημόριο, y η πραγματική τιμή και z η πρόβλεψη εκατοστημορίου (Vermorel, 2012).

Η ονομασία Pinball Loss προέρχεται από το σχήμα που έχει η συγκεκριμένη συνάρτηση και η οποία θυμίζει την τροχιά που μπορεί να πάρει η μπάλα στο παιχνίδι φλίπερ. Γίνεται πιο κατανοητό με την παρακάτω απεικόνιση:



Σχήμα 2.8 : Απεικόνιση του Pinball Loss.

Η συνάρτηση είναι πάντα θετική και όσο πιο μακριά είναι από τον στόχο y, τόσο πιο μεγάλη η τιμή της. Η κλίση απεικονίζει την επιθυμητή ανισορροπία κατά την πρόβλεψη εκατοστημορίου. Για τη σύγκριση δύο μοντέλων με το συγκεκριμένο δείκτη, αρκεί να υπολογιστεί ο μέσος όρος της συνάρτησης κάθε μοντέλου για έναν αξιόλογο αριθμό χρονοσειρών, ώστε να είναι σίγουρο ότι η παρατηρούμενη διαφορά, είναι στατιστικά σημαντική.

Για καλύτερη κατανόηση της ασυμμετρίας της συνάρτησης αυτής, θα παραθέσουμε ένα παράδειγμα. Αν θέλουμε το εκατοστημόριο να πάρει την τιμή 0.5, δηλαδή αυτή του ενδιαμέσου, τότε η συνάρτηση γίνεται συμμετρική και είτε η τελική πρόβλεψη είναι αισιόδοξη ή απαισιόδοξη, συμβάλλει το ίδιο στο τελικό αποτέλεσμα. Αν όμως θελήσουμε η τιμή του εκατοστημορίου να είναι 0.9, τότε οι προβλέψεις που βρίσκονται κάτω από την πραγματική τιμή, τιμωρούνται 9 φορές πιο βαριά από το να βρίσκονταν παραπάνω από την πραγματική τιμή. Γίνεται εύκολα κατανοητό η σημασία που έχει η χρήση του συγκεκριμένου δείκτη στις πιθανοτικές προβλέψεις και πιο συγκεκριμένα, στην εφοδιαστική αλυσίδα, όπου η πρόβλεψη ανωτέρων ή κατωτέρων ορίων ζήτησης, βελτιστοποιούν την λήψη αποφάσεων.

2.6.4.2 Κανονικοποιημένη Συνάρτηση Απώλειας Εκατοστημορίων

Είδαμε νωρίτερα ότι για να συγκρίνουμε χρονοσειρές που έχουν διαφορετικά επίπεδα, χρειάζεται να γίνει κανονικοποίηση κατά τον υπολογισμό του δείκτη σφάλματος. Έτσι προκύπτει και η κανονικοποιημένη συνάρτηση απώλειας εκατοστημορίων, η οποία ταυτίζεται με το Scale Pinball Loss (SPL). Η κανονικοποίηση, όπως και συνήθως, γίνεται με τον παρονομαστή να αποτελεί το μέσο απόλυτο σφάλμα του μοντέλου πρόβλεψης, όταν ως μέθοδος πρόβλεψης έχει επιλεχθεί η αφελής μέθοδος (naive), η οποία δίνει ως πρόβλεψη για την επόμενη περίοδο την πραγματική τιμή της τρέχουσας περιόδου.

Επομένως, λαμβάνει την εξής μαθηματική μορφή:

$$\boldsymbol{SPL} = \frac{1}{h}\frac{\sum_{i=n+1}^{n+h}(y_i - z_i)\tau\mathbf{1}\{z_i \le y_i\} + (z_i - y_i)(1-\tau)\mathbf{1}\{z_i > y_i\}}{\frac{1}{n-1}\sum_{i=2}^{n}|y_i - y_{i-1}|}$$

2.7 Σχετική Συχνότητα

Στον τομέα της στατιστικής, η σχετική συχνότητα αναφέρεται στην αναλογία των εμφανίσεων ενός συγκεκριμένου γεγονότος σε σχέση με τον συνολικό αριθμό γεγονότων που παρατηρούνται σε ένα σύνολο δεδομένων.

$$Σχετική\ Συχνότητα = \frac{Αριθμός\ εμφανίσεων\ ενός\ γεγονότος}{Συνολικός\ αριθμός\ γεγονότων}$$

Στο πλαίσιο ενός διακριτού συνόλου δεδομένων, τα εκατοστημόρια μπορούν να ερμηνευτούν χρησιμοποιώντας την έννοια της σχετικής συχνότητας. Αν για παράδειγμα επιζητούμε διάστημα πρόβλεψης 95%, που σημαίνει ότι το ανώτερο όριο του εκατοστημορίου είναι το 0.975, τότε απαιτούμε οι προβλέψεις μας να υπερβαίνουν τις πραγματικές παρατηρήσεις του συνόλου δεδομένων κατά ένα ποσοστό γύρω από το 97.5%.

3.Μελέτη Περίπτωσης

Στην παρούσα διπλωματική εργασία, θα ελέγξουμε πόσο σημαντική είναι η χρήση εμπειρικών κατανομών για την παραγωγή πιθανοτικών προβλέψεων, με τελικό στόχο την ανάλυση του ρίσκου και την υποστήριξη των αποφάσεων. Ταυτόχρονα, για τις ντετερμινιστικές μεταβλητές (π.χ. ημέρα και μήνας του έτους, ειδικά γεγονότα), θα ελεγχθεί σε τι βαθμό βελτιώνουν την παραμετροποίηση των μοντέλων πρόβλεψης. Αυτοί είναι οι βασικοί πυλώνες πάνω στους οποίους θα στηριχθεί η παρούσα διπλωματική εργασία.

3.1 Σύνολο Δεδομένων

Το σύνολο δεδομένων (dataset) που χρησιμοποιήθηκε στα πλαίσια της εργασίας, προέρχεται από τον διαγωνισμό προβλέψεων M5 και περιλαμβάνει δεδομένα πωλήσεων της Walmart, που παραχωρήθηκαν από την ίδια την εταιρεία (Makridakis, Spiliotis, Assimakopoulos, 2021). Εμπεριέχει μονάδες πωλήσεων (unit sales) διαφόρων προϊόντων σε διαφορετικά καταστήματα των ΗΠΑ, οργανωμένες με τη μορφή ομαδοποιημένων χρονοσειρών. Πιο συγκεκριμένα, δόθηκαν δεδομένα πωλήσεων για 3049 προϊόντα, ταξινομημένα σε τρεις κατηγορίες (Hobbies, Household και Foods) και δέκα καταστήματα που βρίσκονται σε τρεις Πολιτείες των ΗΠΑ (4 καταστήματα στην California, 3 καταστήματα στο Texas και 3 καταστήματα στο Wisconsin). Μάλιστα, οι τρεις κατηγορίες χωρίζονται και σε υποκατηγορίες (Hobbies-1, Hobbies-2, Household-1, Household-2, Foods-1, Foods-2 και Foods-3). Στον παρακάτω πίνακα φαίνεται η ιεραρχία που χρησιμοποιήθηκε για τον διαγωνισμό:

Επίπεδο	Επίπεδο Συνάθροισης	Αριθμός Χρονοσειρών
1	Συνολικές Πωλήσεις Μονάδων (Π.Μ)	1
2	Π.Μ. ανά Πολιτεία	3
3	Π.Μ. ανά Κατάστημα	10
4	Π.Μ. ανά Κατηγορία Προϊόντων	3
5	Π.Μ. ανά Υποκατηγορία Προϊόντων	7
6	Π.Μ. ανά Πολιτεία και Κατηγορία	9
7	Π.Μ. ανά Πολιτεία και Υποκατηγορία	21
8	Π.Μ. ανά Κατάστημα και Κατηγορία	30
9	Π.Μ. ανά Κατάστημα και Υποκατηγορία	70
10	Π.Μ. ανά Προϊόν	3049
11	Π.Μ. ανά Προϊόν και Πολιτεία	9147
12	Π.Μ. ανά Προϊόν και Κατάστημα	30490
Σύνολο		42840

Πίνακας 3.1 : Ιεραρχία επιπέδων συνόλου δεδομένων M5.

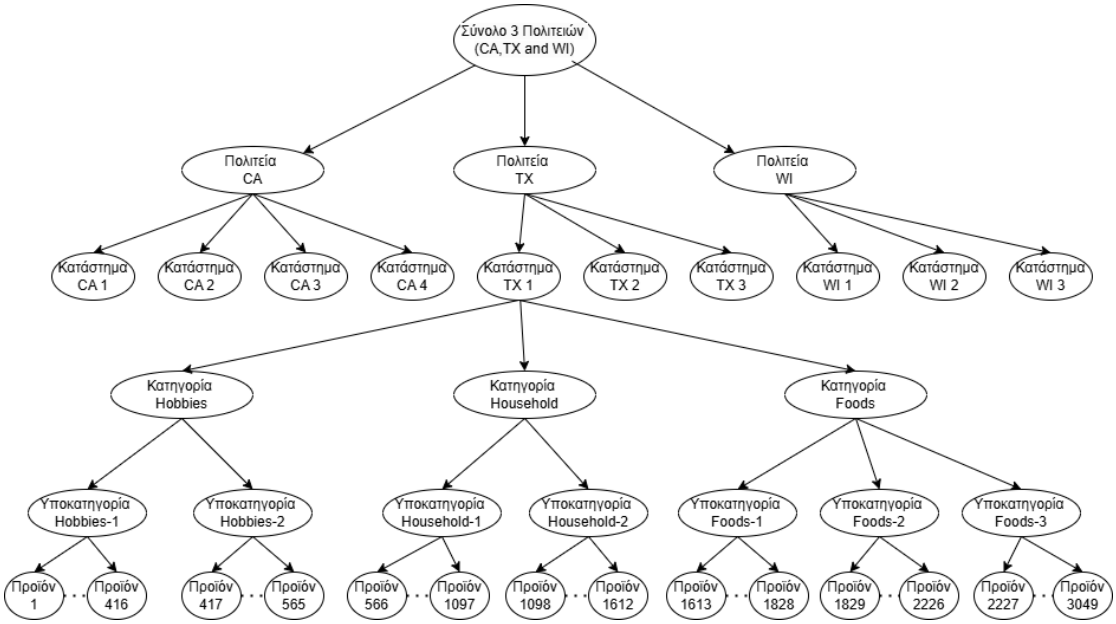
Στο Σχήμα 1, παρουσιάζεται μία επισκόπηση της ιεραρχικής οργάνωσης των δεδομένων του διαγωνισμού.

Τα ιστορικά δεδομένα εκτείνονται από 29/01/2011 έως και 19/06/2016. Επομένως, κάθε προϊόν μπορεί να έχει μέχρι και 1.941 μέρες ιστορικών παρατηρήσεων ή 5.4 χρόνια. Οι οργανωτές του διαγωνισμού χρησιμοποίησαν τις τελευταίες 28 ημέρες του διαθέσιμου συνόλου δεδομένων για τη φάση της επικύρωσης, δηλαδή έδωσαν τη δυνατότητα στους συμμετέχοντες να επανεκτιμήσουν τα μοντέλα τους με βάση τα

αποτελέσματα που είχαν σε αυτό το διάστημα, πριν υποβάλλουν τα τελικά τους μοντέλα στην τελική φάση του διαγωνισμού, στη λεγόμενη φάση αξιολόγησης (διάστημα 28 ημερών μετά το τέλος του συνόλου δεδομένων που δόθηκε).

Επιπλέον, το σύνολο δεδομένων αποτελείται και από τα ακόλουθα τρία csv αρχεία:

- calendar.csv, το οποίο παρέχει την πληροφορία για τις ημερομηνίες τις οποίες πωλήθηκε κάθε προϊόν (ημέρα εβδομάδας, μήνας, έτος). Σε αυτό το αρχείο περιέχονται επιπλέον πληροφορίες για μέχρι δύο ειδικά γεγονότα ή αργίες στις ΗΠΑ (Independence Day, Super Bowl, Christmas Day). Επεξηγεί, επίσης, σε τι κατηγορία ανήκει το ειδικό γεγονός ή η αργία (National, Sporting, Religious αντίστοιχα). Τέλος, δίνεται σε δυαδική μορφή (0 ή 1) η πληροφορία αν υπάρχει SNAP ή όχι σε κάθε πολιτεία την συγκεκριμένη ημέρα, όπου SNAP (Supplement Nutrition Assistance Program) είναι ένα πρόγραμμα παροχής εκπτωτικών κουπονιών σε άτομα με οικονομικά προβλήματα από την κυβέρνηση.
- sell_prices.csv, το οποίο παρέχει την πληροφορία για την τιμή που πωλήθηκε το κάθε προϊόν, για κάθε εβδομάδα (μέσος όρος της τιμής για τις εφτά ημέρες) και για κάθε ένα από τα δέκα καταστήματα που πωλήθηκε.
- sales_train_validation.csv, το οποίο παρέχει την πληροφορία για τις πωλήσεις των 30490 προϊόντων (3049 προϊόντα σε κάθε ένα από τα δέκα καταστήματα) για τις 1913 ημέρες (λείπουν οι τελευταίες 28 ημέρες που διήρκησε η φάση της επικύρωσης).



Σχήμα 3.1 : Επισκόπηση ιεραρχικής οργάνωσης των δεδομένων του διαγωνισμού M5.

Ο διαγωνισμός M5 περιείχε ισάξια βραβεία τόσο για σημειακές όσο και για πιθανοτικές προβλέψεις. Για το πρώτο κομμάτι, οι διαγωνιζόμενοι χρειαζόταν να υποβάλλουν σημειακές προβλέψεις 28 ημερών για όλες τις χρονοσειρές του διαγωνισμού (σύνολο 42840). Για το δεύτερο κομμάτι, απαιτήθηκαν προβλέψεις τόσο για το ενδιάμεσο, όσο και για τα διαστήματα εμπιστοσύνης 50%, 67%, 95% και 99%.

Οι βασικές διαφορές του M5 σε σχέση με τους προηγούμενους διαγωνισμούς είναι οι εξής:

- Χρησιμοποιεί ομαδοποιημένα δεδομένα πωλήσεων και επίπεδα ιεράρχησης για τις χρονοσειρές του, ξεκινώντας από προϊόντα ανά κατάστημα και συνεχίζει σε υποκατηγορίες, κατηγορίες, καταστήματα και Πολιτείες.
- Περιλαμβάνει επεξηγηματικές μεταβλητές όπως προσφορές, ειδικά γεγονότα, και εκπτώσεις, που μπορούν να βελτιώσουν τις προβλέψεις.
- Πέρα από σημειακές προβλέψεις, χρησιμοποιεί και πιθανοτικές για διάφορα διαστήματα εμπιστοσύνης, πράγμα που βοηθάει αρκετά σε επιχειρήσεις όπως η Walmart. Μάλιστα, εφαρμόζει το ίδιο σύνολο δεδομένων και για τα δύο είδη προβλέψεων.
- Τέλος, η πλειοψηφία των χρονοσειρών έχουν σποραδική ζήτηση και αρκετά μηδενικά διαστήματα ζήτησης για πρώτη φορά σε έναν τέτοιο διαγωνισμό.

Για την αξιολόγηση των σημειακών προβλέψεων χρησιμοποιήθηκε ο δείκτης RMSSE (Root Mean Squared Scaled Error), ο οποίος υπολογίζεται από τον παρακάτω τύπο:

$$\boldsymbol{RMSSE} = \sqrt{\frac{1}{h}\frac{\sum_{t=n+1}^{n+h}(Y_t - \widehat{Y}_t)}{\frac{1}{n-1}\sum_{t=2}^{n}(Y_t - Y_{t-1})^2}},$$

όπου, Y_t η πραγματική μελλοντική τιμή της εξεταζόμενης χρονοσειράς τη χρονική στιγμή t, $\widehat{Y}_t$ η πρόβλεψη, n το μήκος του συνόλου δεδομένων για το οποίο εκπαιδεύτηκε η χρονοσειρά και h ο ορίζοντας πρόβλεψης (Makridakis, Spiliotis & Assimakopoulos, 2022).

Μετά τον υπολογισμό του RMMSE και για τις 42840 χρονοσειρές του διαγωνισμού, οι μέθοδοι των συμμετεχόντων θα αξιολογηθούν με βάση το Weighted RMSSE, που έχει τον ακόλουθο τύπο:

$$\boldsymbol{WRMSSE} = \sum_{i=1}^{42840} w_i * RMSSE,$$

όπου, w_i το βάρος της i-οστης χρονοσειράς. Τα βάρη κάθε χρονοσειράς υπολογίζονται με βάση τις τελευταίες 28 παρατηρήσεις του training του συνόλου δεδομένων, περίοδο στην οποία θα υπολογιστούν οι συνολικές πωλήσεις σε δολάρια (σύνολο μονάδων πωλήσεων πολλαπλασιασμένες με την τιμή κάθε μονάδας).

Η επιλογή του συγκεκριμένου δείκτη έγινε για τους εξής λόγους:

- Το κομμάτι της ακρίβειας (accuracy) του διαγωνισμού M5, δηλαδή οι σημειακές προβλέψεις , είναι να προβλεφθεί ακριβώς η μέση ζήτηση, πράγμα που επιτυγχάνεται καλύτερα με σφάλματα τετραγώνων. Αν ο στόχος του κομματιού αυτού ήταν η βέλτιστη διαχείριση των μηδενικών τιμών που υπάρχουν στις περισσότερες χρονοσειρές, θα επιλέγαμε έναν δείκτη ακριβείας, ο οποίος θα περιείχε σφάλματα απολύτων τιμών.
- Ο δείκτης αυτός είναι κανονικοποιημένος, που σημαίνει ότι μπορεί να συγκρίνει προβλέψεις για τελείως διαφορετικές χρονοσειρές.

- Ο υπολογισμός είναι αρκετά απλός και δεν κρύβει προβλήματα, καθώς ο παρονομαστής δεν μπορεί να είναι κοντά στο 0 και να διαφοροποιήσει σημαντικά το τελικό αποτέλεσμα.
- Ο δείκτης είναι συμμετρικός, δηλαδή το σφάλμα του παραμένει ίδιο, είτε η πρόβλεψη υπερβαίνει την τελική πώληση, είτε υπολείπεται αυτής.

Όσον αφορά τις πιθανοτικές προβλέψεις του διαγωνισμού, δηλαδή το uncertainty (αβεβαιότητα) μέρος, χρησιμοποιήθηκε ο δείκτης SPL (Scaled Pinball Loss), για τον οποίο έχουμε μιλήσει αναλυτικά στο Κεφάλαιο 3 της παρούσας εργασίας (Makridakis et al, 2021). Για τον συγκεκριμένο διαγωνισμό, υπολογίζεται ως εξής:

$$\boldsymbol{SPL(u)} = \frac{1}{h}\frac{\sum_{t=n+1}^{n+h}\left(Y_t - Q_t(u)\right)u\mathbf{1}\{Q_t(u) \le Y_t\} + (Q_t(u) - Y_t)(1-u)\mathbf{1}\{Q_t(u) > Y_t\}}{\frac{1}{n-1}\sum_{t=2}^{n}|Y_t - Y_{t-1}|},$$

όπου, Y_t η πραγματική μελλοντική τιμή της εξεταζόμενης χρονοσειράς τη χρονική στιγμή t, $Q_t(u)$ η πρόβλεψη για u εκατοστημόριο (quantile), h ο ορίζοντας πρόβλεψης, n το μήκος του συνόλου δεδομένων για το οποίο εκπαιδεύτηκε η χρονοσειρά και **1** η ένδειξη ανάλογα αν η πρόβλεψη μας ήταν μικρότερη ή μεγαλύτερη από την πραγματική μελλοντική τιμή.

Όπως για τον δείκτη RMSSE παραπάνω, ο παρονομαστής υπολογίζεται μόνο για τις περιόδους που κάθε προϊόν πωλείται ενεργά, δηλαδή για τα χρονικά διαστήματα που ακολουθούν την πρώτη μη μηδενική τιμή της υπό εξέταση χρονοσειράς.

Από τη στιγμή που ο διαγωνισμός απαιτεί προβλέψεις για τα διαστήματα πρόβλεψης 50%,67%,95%,99% καθώς και για το μέσο, οι τιμές που παίρνει το u είναι u_1=0.005, u_2=0.025, u_3=0.165, u_4=0.25, u_5=0.5, u_6=0.75, u_7=0.835, u_8=0.975 και u_9=0.995. Οι μικρότερες τιμές του u ανταποκρίνονται στην αριστερή πλευρά της κατανομής, ενώ οι μεγαλύτερες στην δεξιά πλευρά. Το μέσο διάστημα πρόβλεψης μαζί με τα 50% και 67% δίνουν μία καλή αίσθηση του μέσου της κατανομής ενώ τα διαστήματα πρόβλεψης 95% και 99% παρέχουν σημαντική πληροφορία όσον αφορά το ρίσκο να υπάρξουν αρκετά υψηλά ή αρκετά χαμηλά αποτελέσματα.

Ομοίως, μετά τον υπολογισμό του δείκτη SPL για τις 42840 χρονοσειρές και για όλα τα απαιτούμενα εκατοστημόρια, η κατάταξη των συμμετεχόντων θα γίνει με το Weighted SPL (WSPL), μέσω του παρακάτω τύπου:

$$\boldsymbol{WSPL} = \sum_{i=1}^{42840} w_i * \frac{1}{9}\sum_{j=1}^{9} SPL(u_j),$$

όπου, w_i είναι το βάρος της i-οστής χρονοσειράς και u_j το j-οστό από τα εξεταζόμενα εκατοστημόρια. Προφανώς, όσο μικρότερο το WSPL τόσο καλύτερο το τελικό αποτέλεσμα. Τα βάρη κάθε χρονοσειράς υπολογίζονται με βάση τις τελευταίες 28 παρατηρήσεις της εκπαίδευσης του συνόλου δεδομένων, περίοδο στην οποία θα υπολογιστούν οι συνολικές πωλήσεις σε δολάρια (σύνολο μονάδων πωλήσεων πολλαπλασιασμένες με την τιμή κάθε μονάδας) και για τον συγκεκριμένο δείκτη.

Η επιλογή του δείκτη SPL για τον διαγωνισμό M5, έγινε για τους εξής λόγους:

- Υπολογίζεται με ασφάλεια, καθώς δεν βασίζεται σε διαιρέσεις με τιμές που θα μπορούσαν να είναι μηδέν.
- Ο δείκτης αυτός, όπως και ο RMSSE είναι κανονικοποιημένος, που σημαίνει ότι μπορεί να συγκρίνει προβλέψεις για τελείως διαφορετικές χρονοσειρές.
- Ο διαγωνισμός M5 ούτε επικεντρώνεται σε συγκεκριμένο πρόβλημα λήψης αποφάσεων, ούτε καθορίζει τις ακριβείς παραμέτρους ενός τέτοιου προβλήματος. Αυτό σημαίνει ότι όλα τα εκατοστημόρια είναι εν δυνάμει χρήσιμα. Επομένως, από τη στιγμή που μας ενδιαφέρει τι συμβαίνει στην έκταση ολόκληρης της κατανομής, δεν υπάρχει λόγος να υπάρξουν βάρη για ξεχωριστά εκατοστημόρια.

Σε αντίθεση με τους προηγούμενους M διαγωνισμούς, ο συγκεκριμένος χρησιμοποιεί διαχωρισμό σε βάρη (weighting) στα προϊόντα της. Αυτό συμβαίνει επειδή, για μία επιχείρηση όπως η Walmart, είναι σημαντικό να χρησιμοποιεί κάποιο μοντέλο που προβλέπει καλύτερα τις χρονοσειρές υψηλότερης σημασίας, δηλαδή τα προϊόντα της τα οποία έχουν μεγαλύτερη τιμή και πωλούνται και αρκετά συχνά.

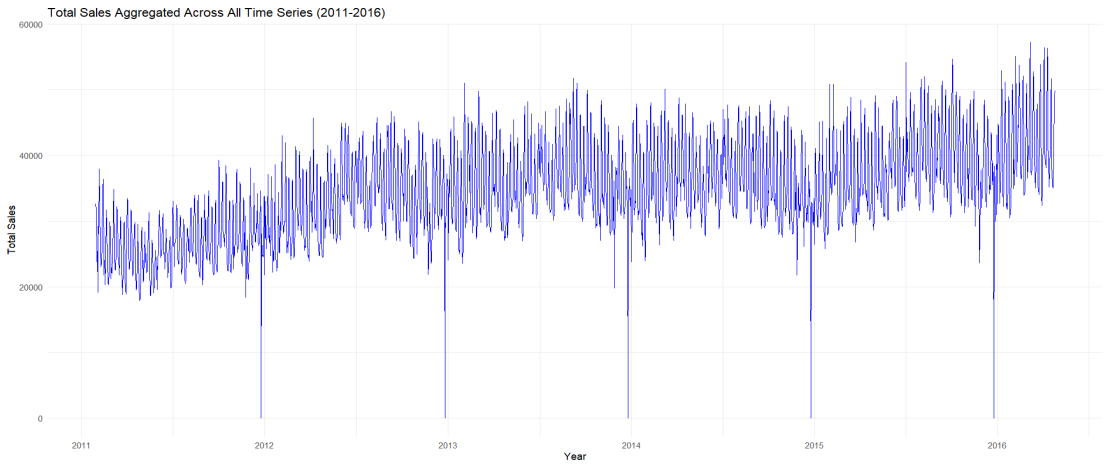
Στην παρούσα διπλωματική εργασία, η ανάλυση θα πραγματοποιηθεί μόνο για το κομμάτι της αβεβαιότητας του διαγωνισμού, δηλαδή μόνο για τις πιθανοτικές προβλέψεις και πιο συγκεκριμένα, για το εκατοστημόριο του 0.975. Αυτή η επιλογή έγινε με γνώμονα το είδος των καταστημάτων, τα οποία προτιμούν να έχουν απόθεμα στα προϊόντα τους από το να παρουσιάσουν έλλειψη σε κάποιο και για αυτό τον λόγο, απαιτείται η τελική πρόβλεψη να είναι σε ένα ποσοστό 95% των περιπτώσεων, ανώτερη της τελικής πώλησης. Ο δείκτης ακριβείας που θα χρησιμοποιηθεί θα είναι το SPL και η αξιολόγηση δεν θα γίνει στο σύνολο της ιεραρχίας, αλλά μόνο για το δέκατο επίπεδο ιεραρχίας του διαγωνισμού M5, το οποίο αθροίζει τις πωλήσεις κάθε προϊόντος για τα δέκα καταστήματα. Η επιλογή έγινε για να έχουμε σαφή εικόνα για το κάθε προϊόν, αλλά ταυτόχρονα δεν αναζητήσαμε μεγαλύτερη λεπτομέρεια για κάθε πολιτεία και κατάστημα.

3.2 Επεξηγηματική Ανάλυση Δεδομένων

Το σύνολο δεδομένων που χρησιμοποιήθηκε για τον διαγωνισμό, παρουσιάζει αρκετά ενδιαφέροντα στοιχεία, τα οποία πρέπει να αναλυθούν πριν γίνει οποιαδήποτε προσπάθεια πρόβλεψης (Makridakis, Petropoulos & Spiliotis, 2022). Η εβδομαδιαία και μηνιαία εποχιακότητα, οι αρκετές μηδενικές πωλήσεις προϊόντων την περίοδο ενεργής πώλησης τους, η τάση με την πάροδο των πέντε χρόνων καθώς και οι διαφορετικές χρονικοί περίοδοι πώλησης του κάθε προϊόντος είναι κάποια από αυτά.

3.2.1 Συνολικές Πωλήσεις

Στο παρακάτω σχήμα φαίνεται η χρονοσειρά με τις συνολικές πωλήσεις και στα δέκα καταστήματα της Walmart για την χρονική περίοδο από 29/01/2011 μέχρι και 25/04/2016, δηλαδή οι μέρες που δόθηκαν για την περίοδο της εκπαίδευσης (training set) και δεν περιλαμβάνει τις επόμενες είκοσι οκτώ (28) ημέρες που χρησιμοποιήθηκαν ως επικύρωση, ούτε και τις επόμενες είκοσι οκτώ (28) που χρησιμοποιήθηκαν για την τελική αξιολόγηση των υποβολών.

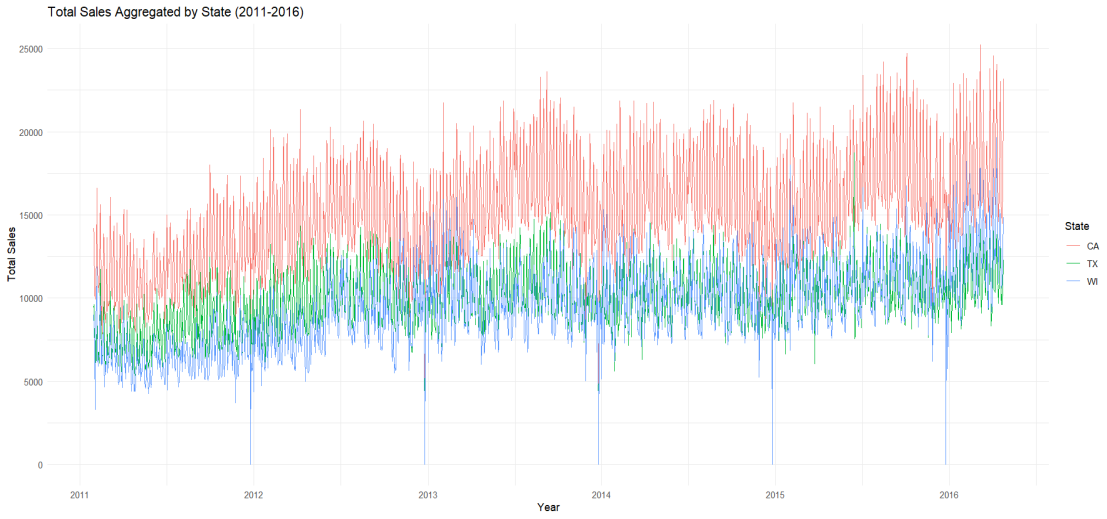


Σχήμα 3.2 : Συνολικές πωλήσεις στα 10 καταστήματα για τις 1913 ημέρες του συνόλου δεδομένων εκπαίδευσης.

Οι αιχμές που παρατηρούνται προς το 0 κάθε χρόνο οφείλονται στη μέρα των Χριστουγέννων, όπου οι πωλήσεις είναι ελάχιστες, σχεδόν μηδενικές.

Γενικά, μπορούμε εύκολα να παρατηρήσουμε μια ανοδική τάση στις πωλήσεις κάθε έτους, την οποία όμως δεν μπορούμε να την αποδώσουμε αποκλειστικά στο ότι κάθε χρόνο αυξάνονται οι ανάγκες και κατά συνέπεια οι πωλήσεις. Αυτό συμβαίνει γιατί πολλά προϊόντα, δεν ξεκίνησαν να πωλούνται από την πρώτη διαθέσιμη ημέρα του συνόλου δεδομένων, αλλά σε κάποια χρονική στιγμή στη διάρκεια των 5,4 ετών.

3.2.2 Συνολικές Πωλήσεις ανά Πολιτεία

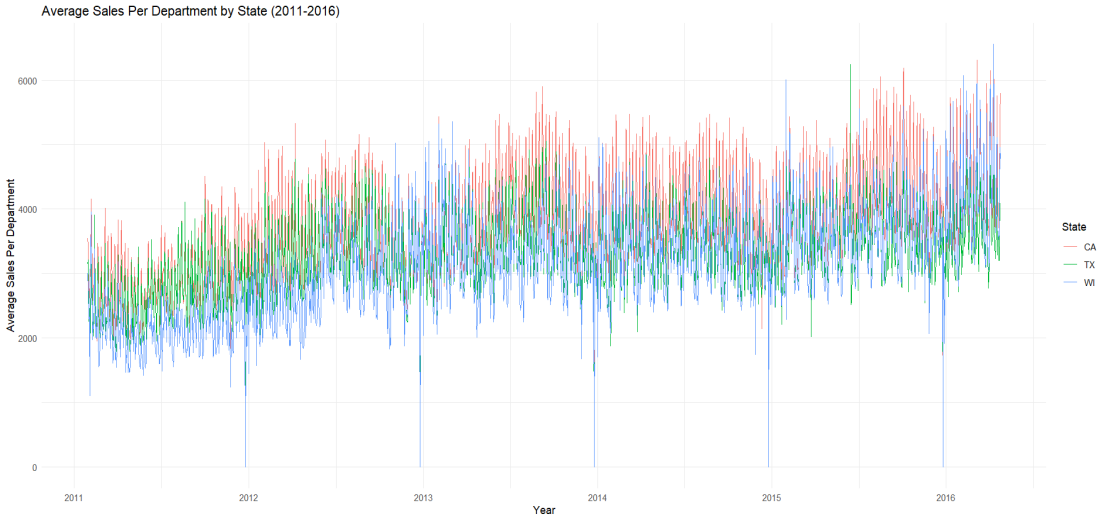


Σχήμα 3.3 : Συνολικές πωλήσεις ανά πολιτεία για τις 1913 ημέρες του συνόλου δεδομένων εκπαίδευσης.

Στο παραπάνω σχήμα παρουσιάζονται οι τρεις χρονοσειρές για τις πωλήσεις που έγιναν σε κάθε πολιτεία στο διάστημα του συνόλου δεδομένων. Οι τιμές για την πολιτεία της Καλιφόρνια (CA) είναι συνεχώς περισσότερες από αυτές του Τέξας (TX) και του Ουισκόνσιν (WI), επειδή υπάρχει ένα περισσότερο κατάστημα εκεί σε σχέση με τις άλλες. Επιπλέον, μπορούμε να παρατηρήσουμε ότι τα πρώτα έτη οι πωλήσεις στο Τέξας ήταν σχετικά περισσότερες σε σχέση με το Ουισκόνσιν, όμως στη συνέχεια και

συγκεκριμένα από το 2013 και μετά, φαίνεται να είναι ισοδύναμες και μάλιστα προς το τέλος φαίνεται να υπολείπονται. Δεν μπορούμε να εξηγήσουμε με σιγουριά γιατί έγινε αυτό, αλλά μπορούμε να κάνουμε διάφορες υποθέσεις. Είναι πιθανό να αναπτύχθηκαν οικονομικά οι γειτονιές του Ουισκόνσιν γύρω από αυτά τα καταστήματα ή να έκλεισαν παρόμοια μαγαζιά που υπήρχαν τριγύρω. Είναι, επίσης, φανερό ότι και οι τρεις χρονοσειρές παρουσιάζουν εποχιακότητες και κάποια spikes είτε προς τα πάνω είτε προς τα κάτω, ανά περιόδους, αλλά θα αναλύσουμε την εποχιακότητα τους στη συνέχεια.

Για να μπορέσουμε να συγκρίνουμε αναλυτικότερα τις πωλήσεις ανά πολιτεία, μιας και η πολιτεία της Καλιφόρνια έχει ένα κατάστημα περισσότερο, χρησιμοποιούμε στο παρακάτω σχήμα τον μέσο όρο πωλήσεων στα καταστήματα κάθε πολιτείας. Φαίνεται ότι η Καλιφόρνια υπερέχει των υπόλοιπων δύο πολιτειών. Αυτό μπορεί να οφείλεται στο γεγονός ότι τα συγκεκριμένα καταστήματα απευθύνονται σε μεγαλύτερες πληθυσμιακές και αγοραστικές ομάδες.



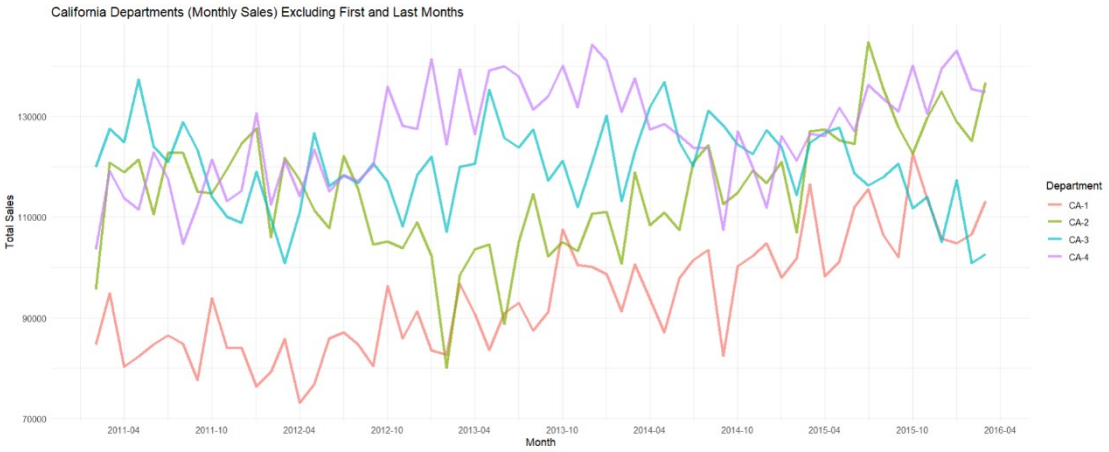
Σχήμα 3.4 : Μέσος όρος καταστήματος ανά πολιτεία για τις 1913 ημέρες του συνόλου δεδομένων εκπαίδευσης.

3.2.3 Συνολικές Πωλήσεις ανά Κατάστημα

Στη συνέχεια διαχωρίσαμε τις πωλήσεις ανά κατάστημα σε κάθε πολιτεία για να έχουμε μία πιο σφαιρική εικόνα των δεδομένων μας. Χρησιμοποιήσαμε μηνιαίες παρατηρήσεις στα παρακάτω γραφήματα για να είναι πιο διακριτές οι μεταβολές σε κάθε διαφορετικό κατάστημα. Επιπλέον δεν χρησιμοποιήσαμε σε κάθε γράφημα τον πρώτο μήνα του συνόλου δεδομένων, ο οποίος αποτελείται μόνο από τρεις ημέρες, επομένως η τιμή του θα ήταν πολύ μικρή και δεν θα προσέδιδε κάτι για τη συγκεκριμένη παρατήρηση.

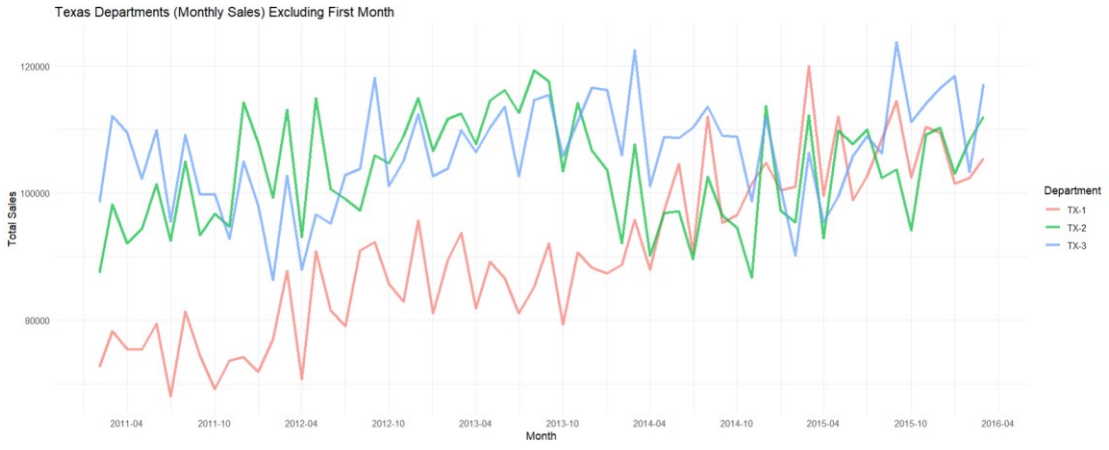
Ξεκινώντας από την Καλιφόρνια, μπορούμε να παρατηρήσουμε ότι δεν υπάρχει κάποια σταθερότητα ανά κατάστημα. Το CA-1 παρουσιάζει μία αύξουσα τάση σε όλη τη διάρκειά του, πράγμα που είναι και το αναμενόμενο, αφού προστίθενται συνεχώς καινούρια προϊόντα προς πώληση. Από την άλλη, το CA-2 έχει αρκετά σκαμπανεβάσματα. Το πιο αξιοσημείωτο είναι το CA-3, το οποίο τελικά έχει λιγότερες μηνιαίες πωλήσεις σε σχέση με την αρχή του dataset. Ο κοινός παρονομαστής και των τεσσάρων καταστημάτων είναι οι κορυφές και οι βυθίσεις που παρουσιάζουν κατά τις ίδιες χρονικές περιόδους, που λογικά οφείλεται στα Χριστούγεννα όπου και αυξάνεται η ζήτηση και

αντίστοιχα σε κάποια περίοδο που είναι μειωμένες οι ανάγκες της εκάστοτε πληθυσμιακής ομάδας. Αν εστιάσουμε στους μήνες του Ιουνίου και του Ιουλίου κάθε έτους, παρατηρούνται αυξήσεις και κορυφές στις πωλήσεις των προϊόντων, πράγμα που σημαίνει ότι όταν φτάνει το καλοκαίρι, ο κόσμος αυξάνει τις αγορές του.



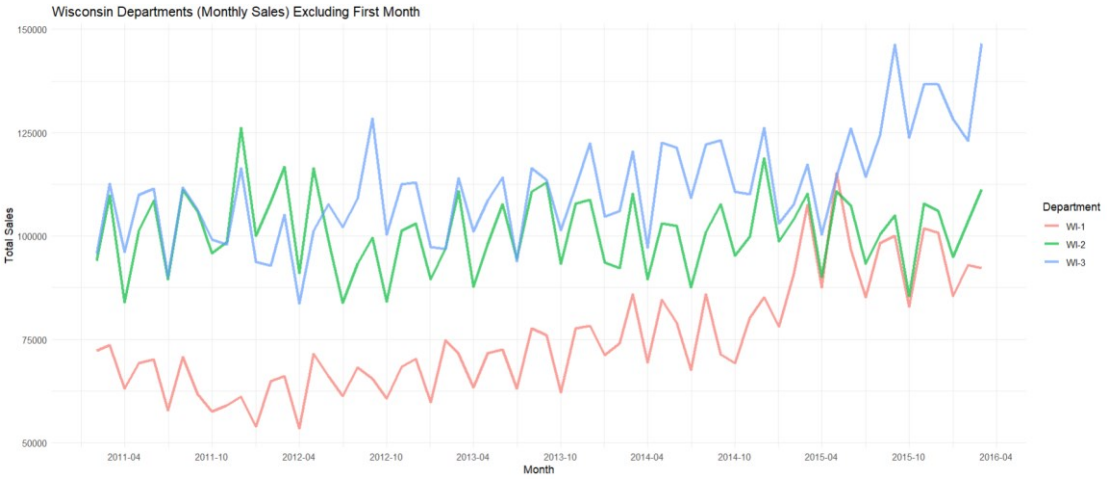
Σχήμα 3.5 : Συνολικές πωλήσεις ανά κατάστημα στην Καλιφόρνια για τις 1913 ημέρες του συνόλου δεδομένων εκπαίδευσης.

Για το Τέξας, τα πράγματα είναι πιο απλά και πιο ξεκάθαρα. Τα TX-2 και TX-3 είναι αρκετά παρόμοια σε πωλήσεις και εποχιακότητα. Το TX-1 παρουσιάζει και πολύ ισχυρή ανοδική τάση με την πάροδο του χρόνου και φτάνει στα επίπεδα πωλήσεων των άλλων δύο καταστημάτων.



Σχήμα 3.6 : Συνολικές πωλήσεις ανά κατάστημα στο Τέξας για τις 1913 ημέρες του συνόλου δεδομένων εκπαίδευσης.

Τέλος, όπως και στο Τέξας, η κατάσταση στο Ουισκόνσιν είναι πιο ξεκάθαρη, καθώς τα τρία καταστήματα φαίνεται να ακολουθούν ίδια μοτίβα κατά κύριο λόγο. Παρόμοιες τάσεις και εποχιακότητες, με εξαίρεση το WI-2 που παρουσιάζει μια στασιμότητα. Και εδώ το WI-1 ξεκινάει από αρκετά λιγότερες πωλήσεις, πλησιάζει αρκετά όμως τα άλλα καταστήματα με την πάροδο του χρόνου.

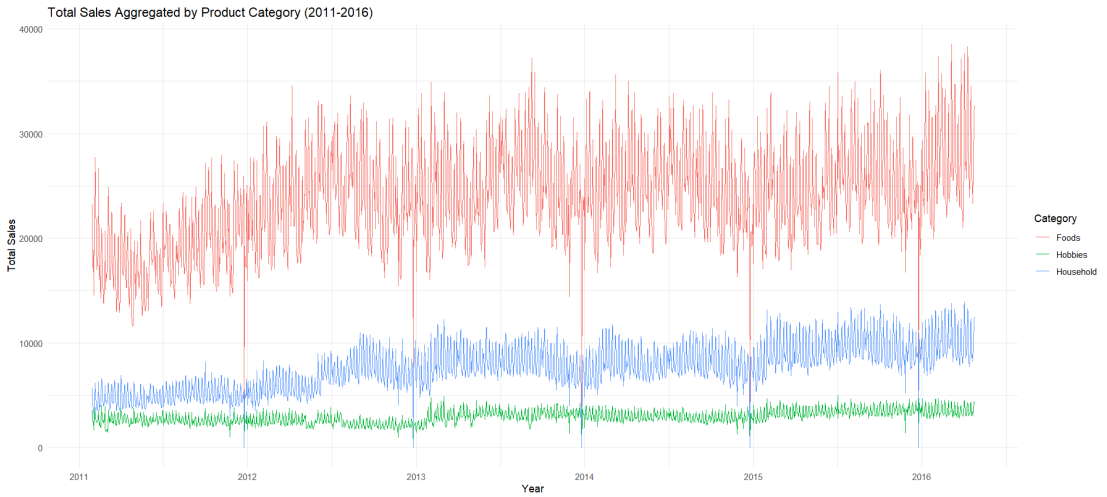


Σχήμα 3.7 : Συνολικές πωλήσεις ανά κατάστημα στο Ουισκόνσιν για τις 1913 ημέρες του συνόλου δεδομένων εκπαίδευσης.

Σαν συνολική εικόνα, μία βασική διαφορά που μπορεί να παρατηρηθεί ανάμεσα στα τρία παραπάνω γραφήματα, είναι ότι το μέσο επίπεδο κάθε καταστήματος στο Ουισκόνσιν είναι πολύ πιο σταθερό σε σχέση με αυτό των καταστημάτων των άλλων δύο πολιτειών.

3.2.4 Συνολικές Πωλήσεις ανά Κατηγορία Προϊόντος

Στη συνέχεια, διαχωρίζουμε τις συνολικές πωλήσεις σε κάθε κατηγορία προϊόντος. Οι τρεις κατηγορίες, όπως έχουμε αναφέρει είναι Hobbies, Household και Foods. Φαίνεται πως ανοδική τάση έχουν όλες. Το επίπεδο των πωλήσεων διαφέρει αρκετά καθώς κάθε κατηγορία έχει διαφορετικό αριθμό προϊόντων και πιο συγκεκριμένα τα Φαγητά αποτελούν το 47,2% των συνολικών χρονοσειρών, τα Οικιακά το 34,3% και τα Χόμπι το 18,5%. Η διαφορά μεταξύ των φαγητών και των οικιακών είναι αναλογικά αρκετά μεγαλύτερη σε σχέση με τη διαφορά των οικιακών με τα χόμπι, πράγμα που σημαίνει ότι πωλούνται περισσότερες μονάδες ανά προϊόν στην κατηγορία των φαγητών σε σχέση με τις υπόλοιπες, το οποίο είναι και το λογικό να συμβαίνει.



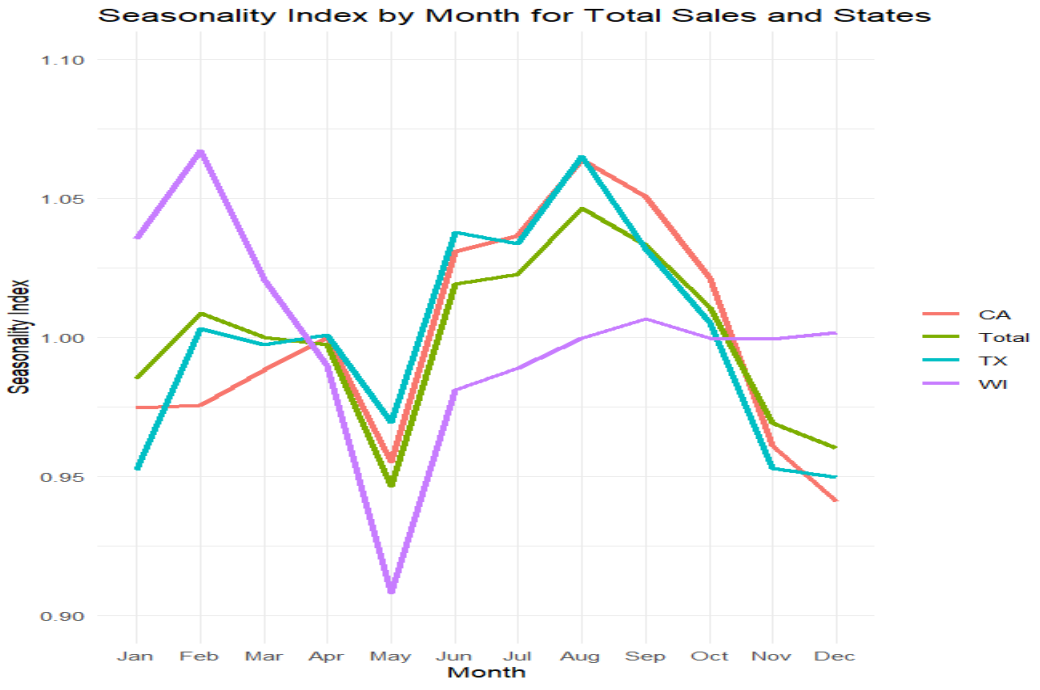
Σχήμα 3.8 : Συνολικές πωλήσεις ανά κατηγορία για τις 1913 ημέρες του συνόλου δεδομένων εκπαίδευσης.

Φαίνεται, επίσης, ότι τα Hobbies δεν παρουσιάζουν τόσο έντονες κορυφές, όπως τα Foods, πράγμα που ίσως οφείλεται στο SNAP, το οποίο εξ ορισμού επηρεάζει κυρίως τα Φαγητά, σε σχέση με τις υπόλοιπες κατηγορίες.

3.2.5 Μηνιαία και Εβδομαδιαία Εποχιακότητα

Ο δείκτης εποχιακότητας είναι από τα πιο ενδιαφέροντα στοιχεία που μπορεί να παρατηρήσει κάποιος σχετικά με τον διαγωνισμό. Δείχνει το μέγεθος που μπορεί να έχουν εποχιακές επιδράσεις σε μία χρονοσειρά και υπολογίζεται αφαιρώντας την τάση της και υπολογίζοντας τον μέσο όρο των δεδομένων για κάθε εποχή ή μήνα ή ημέρα, ανάλογα τι είδους εποχιακότητα θέλουμε να εξετάσουμε. Αν η τιμής ενός μήνα είναι 1.05 για παράδειγμα, αυτό σημαίνει ότι υπάρχει αύξηση της τάξης του 5% σε σχέση με έναν τυπικό μήνα του έτους.

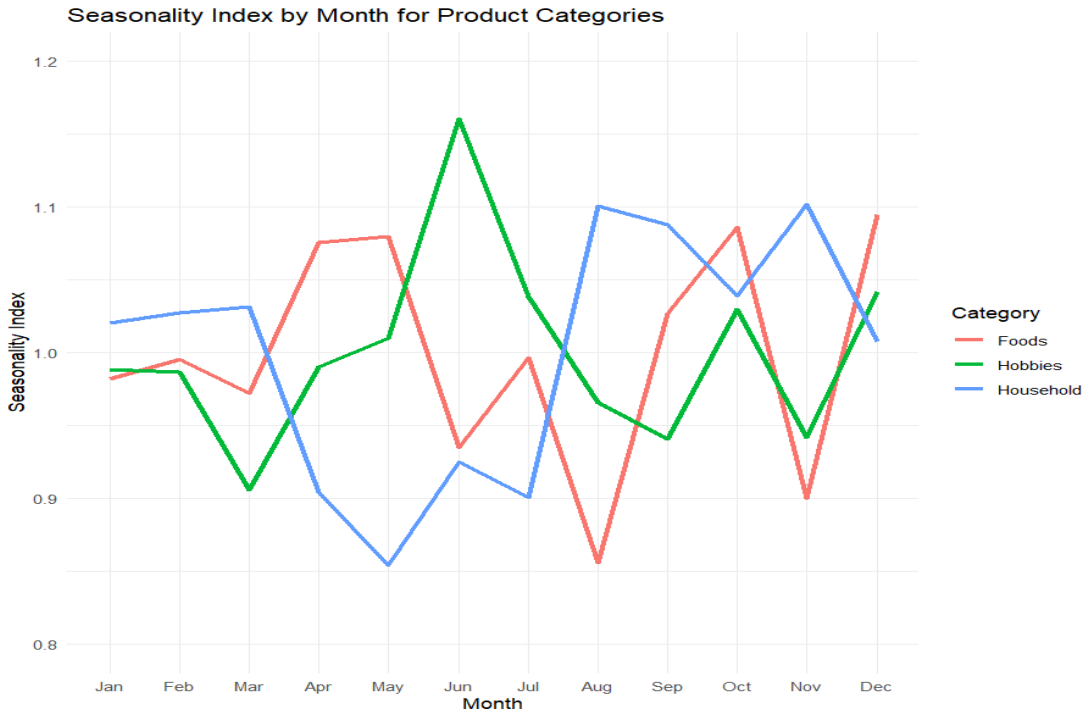
Στο παρακάτω σχήμα απεικονίζεται η μηνιαία εποχιακότητα που παρουσιάζουν οι συνολικές πωλήσεις, καθώς και οι πωλήσεις σε κάθε πολιτεία ξεχωριστά. Παρόμοια συμπεριφορά φαίνεται να έχουν οι συνολικές μαζί με της Καλιφόρνια και του Τέξας. Εν αντιθέσει, το Ουισκόνσιν δεν παρουσιάζει καθόλου εποχιακότητα κατά τους μήνες του καλοκαιριού και αυξάνεται αρκετά στους χειμερινούς μήνες. Αυτό, λογικά οφείλεται στο αρκετά διαφορετικό κλίμα που έχει, με κρύους και χιονισμένους χειμώνες, ενώ οι άλλες δύο πολιτείες, διακρίνονται για τα ζεστά τους καλοκαίρια. Για αυτό άλλωστε παρουσιάζουν αιχμή εποχιακότητας τον μήνα Αύγουστο. Κοινός παρονομαστής όλων είναι ο Μάιος, που ο δείκτης βρίσκεται στις χαμηλότερες τιμές του παντού.



Σχήμα 3.9 : Μηνιαίοι δείκτες εποχιακότητας ανά πολιτεία και συνολικά.

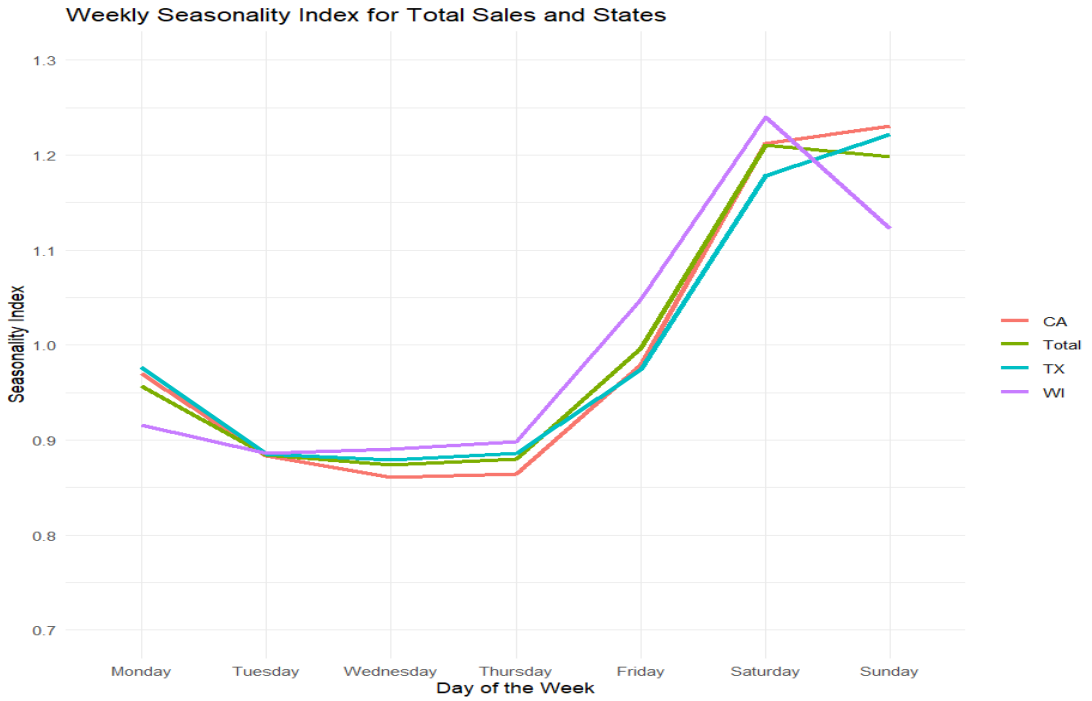
Παρακολουθώντας τον δείκτη εποχιακότητας για κάθε κατηγορία προϊόντος παρακάτω, δεν φαίνεται να έχουν ίδια μοτίβα. Ενδιαφέρον ότι τα χόμπι αυξάνονται κατακόρυφα τον Ιούνιο, που μάλλον σημαίνει ότι ο κόσμος προετοιμάζεται και κάνει αγορές για τις καλοκαιρινές του διακοπές. Τα οικιακά αυξάνονται πολύ τον Αύγουστο και τους φθινοπωρινούς μήνες, που ίσως συμπίπτει με την έναρξη του καινούριου σχολικού

έτους, οπότε οι οικογένειες προετοιμάζονται κατάλληλα. Τους μήνες από Απρίλιο μέχρι Ιούλιο, τα οικιακά απασχολούν σε πολύ μικρό βαθμό τους αγοραστές. Τα φαγητά δεν φαίνεται να ακολουθούν συγκεκριμένη λογική, εκτός από τον Δεκέμβριο που ο κόσμος αγοράζει περισσότερο από κάθε άλλο μήνα, καθώς πλησιάζουν τα Χριστούγεννα και κάθε οικογένεια θα κάνει το τραπέζι της και θα φιλοξενήσει κόσμο.



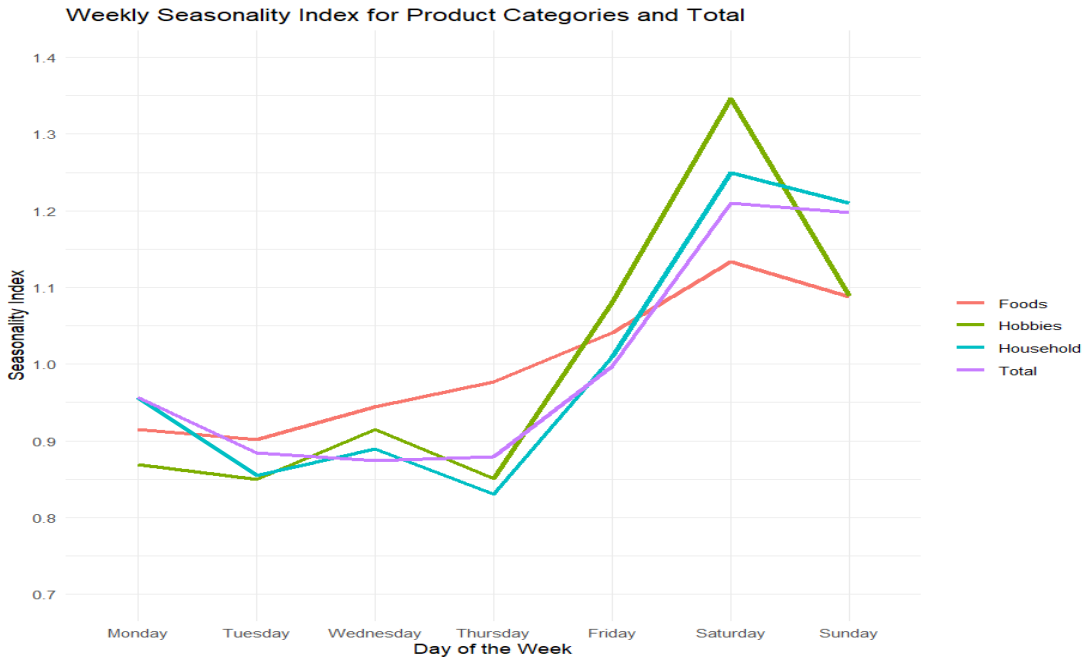
Σχήμα 3.10 : Μηνιαίοι δείκτες εποχιακότητας ανά κατηγορία.

Για να εμβαθύνουμε στα δεδομένα του διαγωνισμού, απαιτείται η μελέτη των πωλήσεων σε εβδομαδιαία βάση. Σε γενικές γραμμές, οι πωλήσεις κάθε μέρα της εβδομάδας ακολουθούν το ίδιο μοτίβο σε επίπεδο πολιτείας και ως φυσικό επακόλουθο σε επίπεδο συνολικών πωλήσεων. Κατά τη διάρκεια της εβδομάδας οι πωλήσεις παραμένουν σε χαμηλό επίπεδο, χωρίς ιδιαίτερη κινητικότητα στα καταστήματα. Όμως, μόλις πλησιάζει το τέλος της εβδομάδας, οι πωλήσεις αυξάνονται κατακόρυφα και μάλιστα φτάνουν στο μέγιστο την Κυριακή στις πολιτείες της Καλιφόρνια και του Τέξας. Η μόνη διαφορά των τριών πολιτειών, είναι ότι στο Ουισκόνσιν, η ζήτηση κορυφώνεται το Σάββατο. Αυτό που συμβαίνει είναι και το λογικό, καθώς μεσοβδόμαδα ο κόσμος εργάζεται και δεν έχει τον ίδιο χρόνο και την ίδια όρεξη να ψωνίσει τα απαραίτητα. Στις ΗΠΑ συγκεκριμένα, τα καταστήματα της Walmart όπως και άλλες παρόμοιες αλυσίδες, δουλεύουν από Δευτέρα έως Κυριακή με το ωράριό τους να διαρκεί από τις 6:00 το πρωί έως και τις 23:00 το βράδυ. Είναι φανερό, ότι η εποχιακότητα, είτε αυτή είναι μηνιαία είτε εβδομαδιαία, είναι από τους σημαντικότερους παράγοντες που θα βοηθήσουν στην βελτίωση των μεθόδων πρόβλεψης, ειδικά αν χρησιμοποιηθεί στα διάφορα επίπεδα ιεράρχησης του διαγωνισμού.



Σχήμα 3.11 : Εβδομαδιαίοι δείκτες εποχιακότητας ανά πολιτεία και συνολικά.

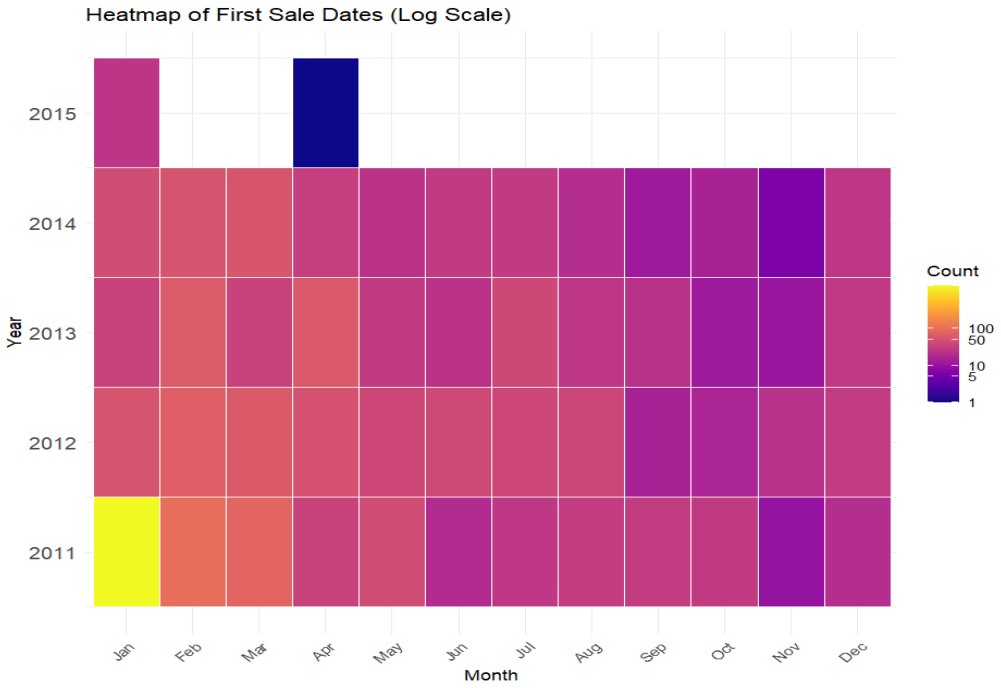
Εξίσου ενδιαφέρουσα είναι η παρακολούθηση της εβδομαδιαία εποχιακότητας σε επίπεδο κατηγορίας προϊόντος, όπου και οι τρεις κατηγορίες παρουσιάζουν παρόμοια συμπεριφορά, αλλά θα ήθελα να επικεντρωθούμε περισσότερα στα φαγητά, για τα οποία καθ' όλη τη διάρκεια της εβδομάδας αυξάνεται η ζήτηση γραμμικά μέχρι και το Σάββατο. Από την άλλη, τα χόμπι και τα οικιακά, ενώ όλη την εβδομάδα παραμένουν χαμηλά, ξαφνικά παρουσιάζουν ραγδαία αύξηση στο τέλος της εβδομάδας. Αυτό συμβαίνει, επειδή η ανάγκη για φαγητό υπάρχει όλη την εβδομάδα, ενώ η ανάγκη για τις άλλες δύο κατηγορίες προκύπτει κυρίως όταν υπάρχει ελεύθερος χρόνος για αγορές.



Σχήμα 3.12 : Εβδομαδιαίοι δείκτες εποχιακότητας ανά πολιτεία και συνολικά.

3.2.6 Κατανομές Αφετηρίας Παρατηρήσεων

Ένα ακόμη σημαντικό χαρακτηριστικό των χρονοσειρών του συνόλου δεδομένων είναι η διασπορά των πρώτων και των τελευταίων μηδενικών παρατηρήσεων κάθε προϊόντος. Στο επόμενο σχήμα, απεικονίζεται σε heatmap το πότε ξεκίνησαν να πωλούνται οι μονάδες. Παρατηρούμε ότι στην αρχή του dataset, πωλούταν περίπου το 50% του συνόλου. Η τελευταία καταχώρηση που έγινε ήταν τον Απρίλιο του 2015 και σαν συνολικό συμπέρασμα, προς το τέλος κάθε έτους δεν προστίθενται νέοι κωδικοί, σε αντίθεση με την αρχή που υπάρχει η μεγαλύτερη άφιξη νέων προϊόντων. Βέβαια, κάθε Δεκέμβρη και εξαιτίας των επικείμενων εορτών υπάρχει αύξηση σε σχέση με τους φθινοπωρινούς μήνες. Το πότε ξεκίνησε η πώληση κάθε μονάδας, επιβεβαιώνει και την ανοδική τάση που παρατηρούσαμε στις συνολικές πωλήσεις. Οι μήνες του Φεβρουαρίου και του Μαρτίου είναι κενοί, επειδή δεν παρατηρήθηκε πρώτη παρατήρηση κάποιου προϊόντος για αυτούς τους μήνες.



Σχήμα 3.13 : Heatmap πρώτων παρατηρήσεων προϊόντων ανά μήνα και έτος.

3.3 Πειραματική Διάταξη

Το πείραμά μας, θα επικεντρωθεί στο σχεδιασμό ενός πλαισίου πιθανοτικών προβλεψεων, το οποίο θα βασίζεται αποκλειστικά στην εμπειρική εκτίμηση της κατανομής των δεδομένων και των σφαλμάτων των σημειακών προβλέψεων. Επιπλέον, θα εξεταστεί η επιρροή που έχουν οι ντετερμινιστικές μεταβλητές (π.χ. ημέρα, μήνας, ειδικό γεγονός) στην καλύτερη παραμετροποίηση των μοντέλων πρόβλεψης.

3.3.1 Επίπεδο Ιεράρχησης Δεδομένων

Για το πειραματικό μέρος, χρησιμοποιήθηκε το δέκατο (10^o) επίπεδο ιεράρχησης του διαγωνισμού M5, δηλαδή οι πωλήσεις μονάδων ανά προϊόν. Πρακτικά, αθροίζονται οι πωλήσεις κάθε μονάδας σε όλα τα καταστήματα κάθε πολιτείας και προκύπτουν 3049 χρονοσειρές, από τις αρχικές 30490 που έχει το σύνολο δεδομένων που τέθηκε στη

διαθεσιμότητα από τη Walmart. Η επιλογή αυτού του επιπέδου έγινε επειδή ο σκοπός του πειράματος αυτού δεν είναι να συγκρίνουμε τι έγινε σε κάθε διαφορετική πολιτεία και σε κάθε διαφορετικό κατάστημα που ήταν διαθέσιμο το κάθε προϊόν, αλλά να δούμε τη συνολική εικόνα και πως κάθε ένας από τους εξωτερικούς παράγοντες που θα λάβουμε υπόψιν, επηρέασε τις πωλήσεις είτε θετικά είτε αρνητικά. Για καλύτερη κατανόηση του συνόλου δεδομένων, στο παρακάτω σχήμα παρουσιάζεται μία επισκόπηση των δεδομένων των πωλήσεων που δόθηκαν στο κοινό, ώστε να υπάρχει μία πιο σαφής εικόνα του πως έγινε η άθροιση στις τελικές 3049 χρονοσειρές.

Filter | Cols: « < 1 - 50 > »

	item_id	dept_id	cat_id	store_id	state_id	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8	d_9	d_10
1	HOBBIES_1_001	HOBBIES_1	HOBBIES	CA_1	CA	0	0	0	0	0	0	0	0	0	0
2	HOBBIES_1_002	HOBBIES_1	HOBBIES	CA_1	CA	0	0	0	0	0	0	0	0	0	0
3	HOBBIES_1_003	HOBBIES_1	HOBBIES	CA_1	CA	0	0	0	0	0	0	0	0	0	0
4	HOBBIES_1_004	HOBBIES_1	HOBBIES	CA_1	CA	0	0	0	0	0	0	0	0	0	0
5	HOBBIES_1_005	HOBBIES_1	HOBBIES	CA_1	CA	0	0	0	0	0	0	0	0	0	0
6	HOBBIES_1_006	HOBBIES_1	HOBBIES	CA_1	CA	0	0	0	0	0	0	0	0	0	0
7	HOBBIES_1_007	HOBBIES_1	HOBBIES	CA_1	CA	0	0	0	0	0	0	0	0	0	0
8	HOBBIES_1_008	HOBBIES_1	HOBBIES	CA_1	CA	12	15	0	0	0	4	6	5	7	0
9	HOBBIES_1_009	HOBBIES_1	HOBBIES	CA_1	CA	2	0	7	3	0	2	3	9	0	0
10	HOBBIES_1_010	HOBBIES_1	HOBBIES	CA_1	CA	0	0	1	0	0	0	0	0	0	0

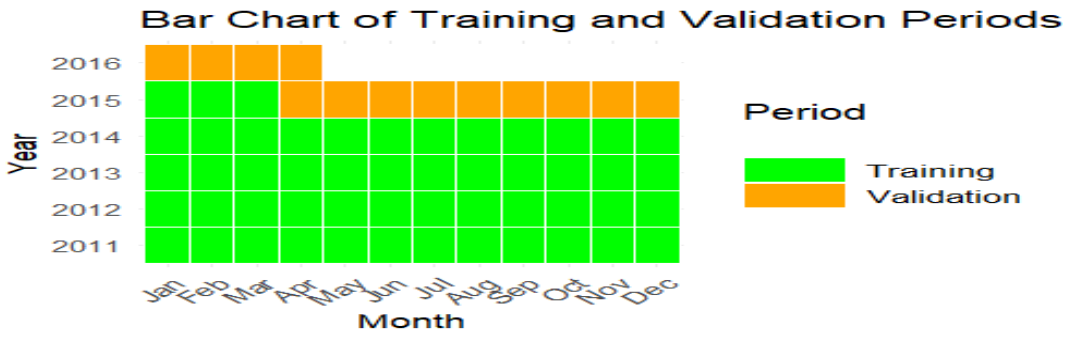
Σχήμα 3.14 : Σύνολο δεδομένων πωλήσεων από Walmart.

Το item_id περιγράφει τον κωδικό του κάθε προϊόντος, το dept_id την υποκατηγορία του, το cat_id την κατηγορία του, το store_id το κατάστημα που πωλείται, το state_id την πολιτεία και οι υπόλοιπες στήλες που εκτείνονται από d_1 έως και d_1913, αποτελούν τις ημέρες του συνόλου δεδομένων. Να σημειωθεί ότι στο πείραμά μας, δεν χρησιμοποιήσαμε ούτε το validation set ούτε και το test set, παρά μόνο αυτές τις 1913 διαθέσιμες ημέρες. Η πληροφορία που αντλήσαμε για αρχή από το παραπάνω είναι ο κωδικός, ο οποίος επαναλαμβάνεται κάθε 3049 σειρές και οι πωλήσεις κάθε διάθεσιμης ημέρας που αθροίστηκαν για κάθε διαφορετικό κωδικό.

3.3.2 Μέθοδοι Αναφοράς

3.3.2.1 Πειραματική Διάταξη Απλής Εκθετικής Εξομάλυνσης

Ξεκινώντας το πείραμα, αποφασίσαμε να χρησιμοποιήσουμε ως πρώτη μέθοδο πρόβλεψης μία απλή Simple Exponential Smoothing (SES), η οποία μελλοντικά θα χρησιμοποιούταν σαν ορόσημο (benchmark). Δηλαδή, κάθε μέθοδος πρόβλεψης που θα επιλέξουμε αργότερα, θα πρέπει να παρουσιάζει βελτιωμένα τελικά αποτελέσματα σε σχέση με αυτήν.

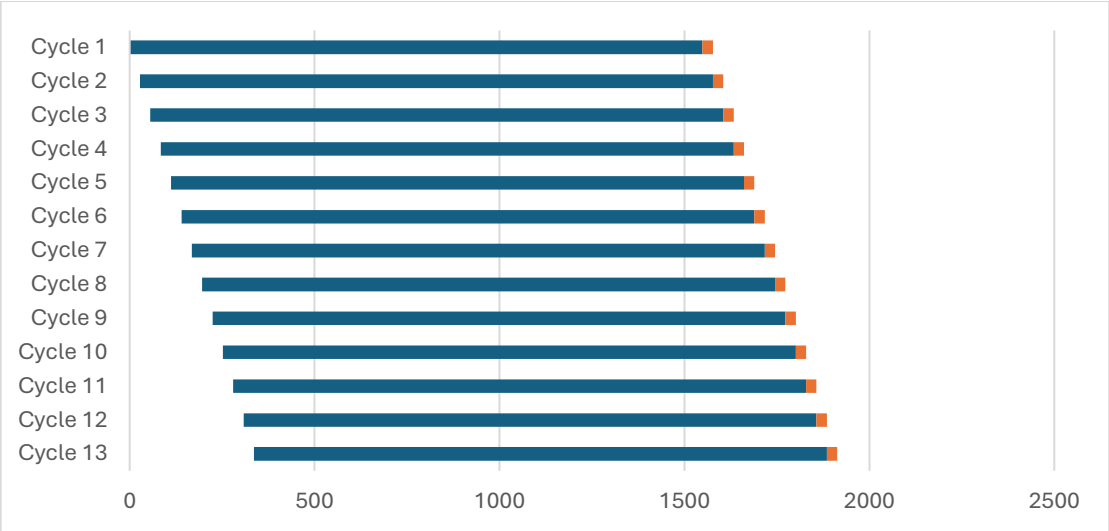


Σχήμα 3.15 : Περίοδοι εκπαίδευσης και δοκιμής του πειράματος.

Για την προετοιμασία των δεδομένων μας, επιλέξαμε ως σετ δοκιμής, το τελευταίο έτος του συνόλου δεδομένων και ακριβέστερα τις τελευταίες 364 ημέρες παρατηρήσεων από τις 1913 διαθέσιμες.. Το σύνολο των δεδομένων που έχουμε στη διάθεσή μας είναι αρκετά μεγάλο, επομένως για την πλειοψηφία των προϊόντων υπάρχουν αρκετές παρελθοντικές παρατηρήσεις για να προβλεφθεί σωστά κάθε χρονοσειρά.

Αν πάρουμε μία από τις χρονοσειρές σαν παράδειγμα, αυτό σημαίνει ότι στην αρχή θα κάνει μία πρόβλεψη με βάση τα ιστορικά δεδομένα που έχει, για τις επόμενες 28 ημέρες. Όμως, στη συνέχεια θα χρησιμοποιήσει τα ιστορικά δεδομένα που έχει μειωμένα κατά τις 28 πρώτες παρατηρήσεις, ενώ ταυτόχρονα θα προσθέσει τις 28 προβλέψεις που έκανε το ίδιο το μοντέλο. Ας υποθέσουμε ότι το προϊόν άρχισε να πωλείται την πρώτη διαθέσιμη ημέρα του συνόλου δεδομένων, επομένως σαν ιστορικά δεδομένα έχει 1549 παρατηρήσεις. Για την επόμενη του πρόβλεψη, πάλι 28 ημερών, πλέον θα χρησιμοποιήσει σαν ιστορικά δεδομένα και τις 28 προβλέψεις που έκανε στον προηγούμενο κύκλο προβλέψεων και από τις 1549 αρχικές παρατηρήσεις, θα χρησιμοποιήσει 1549-28=1521. Αυτό θα συνεχιστεί και για τους 13 κύκλους προβλέψεων ώστε τελικά να έχουμε 13*28=364 ημέρες προβλέψεων. Αναλυτικότερα, φαίνεται η διαδικασία στο παρακάτω σχήμα.

Το κυριότερο πλεονέκτημα που προσφέρει η χρήση του κυλιόμενου παραθύρου είναι η ευελιξία, καθώς ένα τέτοιο μοντέλο προσαρμόζεται ευκολότερα στις αλλαγές που μπορεί να συμβαίνουν στην αγορά και εξασφαλίζει ότι χρησιμοποιούνται τα πιο πρόσφατα διαθέσιμα στοιχεία. Δίνει την δυνατότητα να τροποποιήσεις το μοντέλο σου ανάλογα με τις καινούριες απαιτήσεις που εμφανίζονται διαρκώς.



Σχήμα 3.16 : Απεικόνιση κυλιόμενου παραθύρου 28 ημερών.

Στην υλοποίηση της SES, όπως και σε κάθε μοντέλο που θα εξετάσουμε στη συνέχεια, χρησιμοποιούμε κυλιόμενο παράθυρο 28 ημερών και απαιτούμε διάστημα πρόβλεψης 95%. Αυτό σημαίνει ότι το εκατοστημόριο λαμβάνει την τιμή 0.975. Πρακτικά, με αυτή την τιμή της παραμέτρου, απαιτούμε το 97.5% των προβλέψεων μας να ξεπερνούν τις πραγματικές πωλήσεις. Με απλά λόγια, λόγω της φύσης της Walmart σαν επιχείρηση, προτιμούμε οι προβλέψεις μας να υπερβαίνουν τις πωλήσεις, ώστε η εταιρία να μην βρεθεί προ εκπλήξεως και δεν έχει απόθεμα για τα προϊόντα της.

Για την τιμή της παραμέτρου α, δεν ορίζουμε εμείς κάποια συγκεκριμένη τιμή μέσα στο διάστημα [0.1,0.3], αλλά αφήνουμε τον αλγόριθμο να επιλέγει αυτόματα την βέλτιστη τιμή για κάθε χρονοσειρά. Την περιορίζουμε σε αυτό το διάστημα, καθώς λόγω της μηνιαίας εποχιακότητας που επικρατεί στο σύνολο δεδομένων που έχουμε, επιθυμούμε να δίνει μεγαλύτερο βάρος σε παρελθοντικές τιμές, δηλαδή κοντά στον χρόνο αντί για πρόσφατες, που μπορεί να ήταν μέσα στον ίδιο μήνα.

3.3.2.2 Πειραματική Διάταξη ARIMAX

Μετά τη SES, σαν μέθοδο πρόβλεψης δοκιμάσαμε την Autoregressive Integrated Moving Average with Explanatory Variable (ARIMAX). Επομένως, πριν από οποιαδήποτε προσπάθεια, έπρεπε να ετοιμάσουμε τις επεξηγηματικές μεταβλητές που θα χρησιμοποιούσαμε στο μοντέλο. Την πληροφορία για τις μεταβλητές αυτές τις πήραμε από το Calendar που δόθηκε από την Walmart. Ένα στιγμιότυπο για να κατανοηθεί καλύτερα τι περιέχει, δίνεται παρακάτω.

Filter

	date	wm_yr_wk	weekday	wday	month	year	event_name_1	event_type_1	event_name_2	event_type_2	snap_CA	snap_TX	snap_WI
1	0029-01-20	11101	Saturday	1	1	2011	NA	NA	NA	NA	0	0	0
2	0030-01-20	11101	Sunday	2	1	2011	NA	NA	NA	NA	0	0	0
3	0031-01-20	11101	Monday	3	1	2011	NA	NA	NA	NA	0	0	0
4	0001-02-20	11101	Tuesday	4	2	2011	NA	NA	NA	NA	1	1	0
5	0002-02-20	11101	Wednesday	5	2	2011	NA	NA	NA	NA	1	0	1
6	0003-02-20	11101	Thursday	6	2	2011	NA	NA	NA	NA	1	1	1
7	0004-02-20	11101	Friday	7	2	2011	NA	NA	NA	NA	1	0	0
8	0005-02-20	11102	Saturday	1	2	2011	NA	NA	NA	NA	1	1	1
9	0006-02-20	11102	Sunday	2	2	2011	SuperBowl	Sporting	NA	NA	1	1	1
10	0007-02-20	11102	Monday	3	2	2011	NA	NA	NA	NA	1	1	0
11	0008-02-20	11102	Tuesday	4	2	2011	NA	NA	NA	NA	1	0	1
12	0009-02-20	11102	Wednesday	5	2	2011	NA	NA	NA	NA	1	1	1
13	0010-02-20	11102	Thursday	6	2	2011	NA	NA	NA	NA	1	0	0
14	0011-02-20	11102	Friday	7	2	2011	NA	NA	NA	NA	0	1	1

Showing 1 to 14 of 1,913 entries, 13 total columns

Σχήμα 3.17 : Ημερολόγιο με πληροφορίες για τις 1913 ημέρες του συνόλου δεδομένων εκπαίδευσης.

Η σημαντική πληροφορία που μπορούμε να λάβουμε από εδώ είναι το date που είναι η ημερομηνία, το weekday που είναι η μέρα της εβδομάδας, το month που είναι ο μήνας, το year που είναι το έτος, το event_name_1/2 και event_type_1/2 που είναι το όνομα του ειδικού γεγονότος και ο τύπος του αντίστοιχα και τα snap_CA, snap_TX, snap_WI που δείχνουν αν εκείνη την ημέρα είχε έκπτωση SNAP ή όχι σε κάθε πολιτεία. Μάλιστα, στα ειδικά γεγονότα μπορεί να τυχαίνει να υπάρχουν δύο διαφορετικά την ίδια ημέρα, όπως συνέβη στις 15 Ιουνίου 2014, όπου ήταν οι τελικοί του πρωταθλήματος καλαθοσφαίρισης των ΗΠΑ που θεωρείται αθλητικό γεγονός και η ημέρα του Πατέρα που θεωρείται πολιτιστικό γεγονός.

Επειδή τα μοντέλα ARIMAX είναι αρκετά χρονοβόρα, έπρεπε να διαλέξουμε προσεκτικά ποιες μεταβλητές θα χρησιμοποιήσουμε, ώστε να μην γεμίσουμε με περιττή πληροφορία. Καταλήξαμε, μέσα από ελέγχους, στην έκπτωση SNAP για αδύναμες οικονομικά οικογένειες σε όλες τις πολιτείες, στην ημέρα της εβδομάδας για να υπάρχει ένα είδος εβδομαδιαίας εποχιακότητας στο μοντέλο μας και τα ειδικά γεγονότα που τα ονομάσαμε αργία (holiday) καθώς στα συγκεκριμένα μοντέλα χρησιμοποιήσαμε μόνο τις Εθνικές Εορτές των ΗΠΑ (Bank Holidays) και τις δύο ημέρες που προηγούνται κάθε Εθνικής Εορτής, ώστε να γνωρίζει ότι, αν παρατηρήσει μεγαλύτερη κινητικότητα

εκείνες τις ημέρες, αποδίδεται στην επικείμενη αργία. Για την οπτικοποίηση όσων είπαμε, υπάρχει το επόμενο σχήμα.

Filter

	snap_CA	snap_WI	snap_TX	holiday	day
17	0	1	0	0	3
18	0	1	1	0	4
19	0	0	0	0	5
20	0	0	0	0	6
21	0	0	0	0	7
22	0	0	0	3	1
23	0	0	0	2	2
24	0	0	0	1	3
25	0	0	0	0	4
26	0	0	0	0	5
27	0	0	0	0	6
28	0	0	0	0	7

Showing 16 to 29 of 1,913 entries, 5 total columns

Σχήμα 3.18 : Επεξηγηματικές μεταβλητές για ARIMAX για τις 1913 ημέρες του συνόλου δεδομένων εκπαίδευσης.

3.3.2 Πειραματική Διάταξη LightGBM

Μετά την υλοποίηση του μοντέλου ARIMAX, περνάμε στην εφαρμογή της μηχανικής μάθησης και πιο συγκεκριμένα στα δένδρα αποφάσεων. Το μοντέλο που χρησιμοποιήσαμε είναι το Light Gradient-Boosting Machine (LightGBM). Για τα δέντρα παλινδρόμησης έχουμε μιλήσει αναλυτικά στο θεωρητικό κομμάτι της παρούσας εργασίας, επομένως δεν υπάρχει λόγος να επαναληφθούμε.

Πριν αναφερθούμε στις επεξηγηματικές μεταβλητές που χρησιμοποιήσαμε για να εκπαιδεύσουμε το μοντέλο μας, είναι σημαντικό να τονίσουμε τον ρόλο που παίζουν οι παρελθοντικές πωλήσεις της κάθε χρονοσειράς. Αρχικά, για να δημιουργήσουμε μία μέθοδο ορόσημο (benchmark) για τα LightGBM μοντέλα, εκπαιδεύσαμε τα δέντρα αποφάσεων με τις τιμές που είχαμε στα δεδομένα μας για την προηγούμενη ημέρα πριν την πρόβλεψη, για μία εβδομάδα πριν, για δύο εβδομάδες πριν, για τέσσερις εβδομάδες πριν και έναν μέσο όρο των δεδομένων των πωλήσεων των τελευταίων είκοσι οκτώ ημερών. Τα ονομάσαμε αντίστοιχα lag1, lag7, lag14, lag28 και mean_last_28_days. Οι τιμές αυτές βοηθάνε το μοντέλο να καταλάβει την εποχιακότητα που υπάρχει σε κάθε χρονοσειρά, κυρίως στην περίπτωση της εβδομαδιαίας. Για ακόμα βελτιωμένη απόδοση του μοντέλου μας, θα μπορούσαμε να προσθέσουμε περισσότερες παρόμοιες μεταβλητές, όπως ένας μέσος όρος των πωλήσεων του προϊόντος για κάθε μήνα, αλλά ο σκοπός μας είναι να εξετάσουμε κυρίως τις επεξηγηματικές μεταβλητές και όχι να παρουσιάσουμε το καλύτερο δυνατό αποτέλεσμα.

Filter

	sales_data	lag_1	lag_7	lag_14	lag_28	mean_last_28_days
61	8	21	7	22	10	9.535714
62	15	8	1	3	4	9.928571
63	7	15	8	12	12	9.750000
64	7	7	15	6	18	9.357143
65	9	7	6	10	8	9.392857
66	5	9	14	12	3	9.464286
67	20	5	21	5	10	9.821429
68	13	20	8	7	5	10.107143
69	9	13	15	1	12	10.000000
70	6	9	7	8	8	9.928571
71	6	6	7	15	24	9.285714
72	5	6	9	6	4	9.321429
73	1	5	5	14	4	9.214286
74	7	1	20	21	5	9.285714
75	13	7	13	8	22	8.964286
76	20	13	9	15	3	9.571429

Σχήμα 3.19 : Παράδειγμα lagged features για τυχαία χρονοσειρά.

Για καλύτερη κατανόηση του μοντέλου LightGBM, θεωρούμε σημαντικό να παρουσιάσουμε αναλυτικά τις παραμέτρους με τις οποίες εκπαιδεύσαμε τα δέντρα αποφάσεων και να εξηγήσουμε πως έγινε η επιλογή αυτών. Πριν τις σχολιάσουμε, φαίνεται παρακάτω συνοπτικά η τιμή που αναθέσαμε σε κάθε παράμετρο που χρειαστήκαμε.

Name	Type	Value
params	list [13]	List of length 13
objective	character [1]	'quantile'
metric	character [1]	'quantile'
alpha	double [1]	0.975
boosting_type	character [1]	'gbdt'
subsample	double [1]	0.5
subsample_freq	double [1]	1
learning_rate	double [1]	0.015
num_leaves	double [1]	2047
min_data_in_leaf	double [1]	4095
feature_fraction	double [1]	0.5
max_bin	double [1]	100
n_estimators	double [1]	3000
boost_from_average	logical [1]	FALSE

Σχήμα 3.20 : Παράμετροι του LightGBM μοντέλου.

- Ο σκοπός (objective) είναι παλινδρόμηση εκατοστημορίου (quantile regression), που είναι μία στατιστική τεχνική και χρησιμοποιείται για την εκτίμηση συγκεκριμένων διαστημάτων πρόβλεψης μίας μεταβλητής, δεδομένου ενός συνόλου ανεξαρτήτων μεταβλητών.
- Ο δείκτης αξιολόγησης (metric) που θα χρησιμοποιηθεί για τα δέντρα αποφάσεων θα είναι δείκτης ακριβείας με χρήση εκατοστημορίου (quantile), το οποίο

σημαίνει ότι η αξιολόγηση θα γίνει με βάση συγκεκριμένο εκατοστημόριο, αντί να χρησιμοποιηθεί το μέσο διάστημα πρόβλεψης ή αυτό του ενδιαμέσου, που είναι οι συνηθέστερες επιλογές. Στην περίπτωσή μας είναι το Scaled Pinball Loss (SPL).

- Το alpha που χρησιμοποιούμε είναι το 0.975, το οποίο είναι το καθορισμένο εκατοστημόριο για να έχουμε δείκτη αξιολόγησης με χρήση εκατοστημορίου και επιλέγεται επειδή θέλουμε το διάστημα πρόβλεψης να είναι 95%.
- Το boosting_type επιλέχθηκε gbdt, δηλαδή gradient boosting decision tree, που είναι ο τύπος των δέντρων αποφάσεων που έχουμε επιλέξει για το σκοπό αυτής της διπλωματικής εργασίας.
- Το subsample τέθηκε 0.5, που σημαίνει ότι το πρώτο δέντρο που δημιουργείται, χρησιμοποιεί ένα τυχαίο 50% των διαθέσιμων δεδομένων και όλα τα επόμενα δέντρα εστιάζουν να διορθώσουν το σφάλμα. Μία τέτοια τιμή χρησιμεύει ιδιαίτερα για να αποφύγουμε το overfitting, δηλαδή το μοντέλο μας να δίνει πολύ καλά αποτελέσματα για τα υπάρχοντα δεδομένα, αλλά να μην μπορεί να ανταπεξέλθει σε δεδομένα που θα προστεθούν στο μέλλον. Επίσης, μειώνει αρκετά τον χρόνο που χρειάζεται για να εκπαιδευτεί το μοντέλο μας.
- Το subsample_freq καθορίζει την συχνότητα με την οποία γίνεται το subsample που μόλις εξηγήσαμε και εμείς το έχουμε θέσει 1, που σημαίνει ότι σε κάθε boosting iteration (κάθε ενίσχυση του δέντρου μας), χρησιμοποιεί το ίδιο ποσοστό των διαθέσιμων δεδομένων, το οποίο είναι τυχαίο και διαφορετικό κάθε φορά.
- Το learning_rate (ποσοστό μάθησης), το οποίο καθορίζει τον βαθμό που συνεισφέρει το καινούριο δέντρο που δημιουργείται κάθε φορά στο συνολικό μοντέλο. Πιο συγκεκριμένα, ποσοστικοποιεί τις προβλέψεις κάθε νέου δέντρου, πριν τις προσθέσει στο υπάρχον μοντέλο. Ένα χαμηλό ποσοστό μάθησης, μειώνει την επιρροή που έχει κάθε δέντρο, κάνοντας το μοντέλο να μαθαίνει περισσότερο σταδιακά. Εμείς επιλέξαμε την τιμή 0.015, επομένως κάθε δέντρο συνεισφέρει μόλις το 1.5% της προβλεπτικής του δύναμης στο τελικό μοντέλο. Χρειάζεται περισσότερος υπολογιστικός χρόνος για την εκπαίδευση αλλά πάλι αποφεύγεται το overfitting.
- Το num_leaves είναι ο μέγιστος αριθμός φύλλων που μπορεί να έχει ένα δέντρο. Στην περίπτωσή μας, το έχουμε θέσει ίσο με $2^{11} - 1 = 2047$. Ένας τόσο υψηλός αριθμός φύλλων, δίνει την δυνατότητα στο μοντέλο να καταλάβει πιο περίπλοκα μοτίβα στα δεδομένα. Όμως, αυξάνει τον κίνδυνο για overfitting και τον υπολογιστικό χρόνο, για τα οποία κάνουμε υποχωρήσεις σε άλλες παραμέτρους και τα ισορροπούμε. Για μεγάλα σύνολα δεδομένων, όπως στην περίπτωση μας, απαιτείται για να βελτιώσει την ακρίβεια των προβλέψεών μας.
- Το min_data_in_leaf είναι ο ελάχιστος αριθμός δεδομένων που μπορεί να έχει ένα φύλλο. Στην περίπτωσή μας, το έχουμε θέσει ίσο με $2^{12} - 1 = 4095$. Αν δημιουργηθεί κάποιο φύλλο με λιγότερα δεδομένα, θα απορριφθεί αμέσως, ακόμα και στην περίπτωση που βελτιώνει το μοντέλο μας. Συνδυάζεται με την προηγούμενη παράμετρο, ώστε να αποφευχθούν διασπάσεις σε φύλλα με πολύ λίγα δεδομένα και πάρουν τη θέση φύλλων με σημαντικότερη συνεισφορά, μιας και έχουμε και μέγιστο όριο στον αριθμό τους. Επομένως, είναι αρκετά σημαντικό

να συνδυαστούν για να έχουμε καλύτερο αποτέλεσμα και να ισορροπήσουμε το overfitting και την κατανόηση πιο περίπλοκων μοτίβων.

- Το feature_fraction προσδιορίζει το ποσοστό των μεταβλητών που θα χρησιμοποιήσει κάθε δέντρο για να εκπαιδευτεί. Λειτουργεί με την ίδια λογική με το subsample, δηλαδή σε κάθε δέντρο επιλέγει τυχαία την αναλογία που έχουμε ορίσει των διαθέσιμων χαρακτηριστικών (features), τα οποία είναι οι μεταβλητές που χρησιμοποιούμε. Με την τιμή 0.5, επιλέγει, κάθε φορά που δημιουργεί ένα δέντρο, το 50% των features. Αυτή η επιλογή δεν είναι η ίδια κάθε φορά, καθώς γίνεται τυχαία. Πρακτικά, βοηθάει σε όσα θέλουμε να πετύχουμε ή να αποφύγουμε, τα οποία έχουν ήδη αναφερθεί αρκετά στις προηγούμενες παραμέτρους. Επιπλέον, βοηθάει το μοντέλο να μην επικεντρώνεται σε όποιο feature θεωρεί σημαντικότερο και να προσπαθεί να καταλάβει και άλλα περίπλοκα μοτίβα.
- Η παράμετρος max_bin καθορίζει τον μέγιστο αριθμό ¨κάδων¨ (bins) στους οποίους μπορούν να διακριθούν τα features κατά την εκπαίδευση του μοντέλου. Το LightGBM μοντέλο μετατρέπει τα συνεχή αριθμητικά features σε διακριτά bins για να μειώσει τον υπολογιστικό χρόνο και τη χρήση μνήμης. Το εύρος κάθε χαρακτηριστικού διαιρείται σε διαστήματα και κάθε διαθέσιμο δεδομένο αντιστοιχίζεται σε ένα από αυτά τα διαστήματα με βάση την τιμή του. Στην περίπτωσή μας, έχουμε επιλέξει ως μέγιστο αριθμό το 100, που είναι μία καλή ισορροπία ανάμεσα στο καλύτερο αποτέλεσμα και τον μεγαλύτερο χρόνο εκπαίδευσης του μοντέλου.
- Η παράμετρος n_estimators προσδιορίζει τον αριθμό των επαναλήψεων ενίσχυσης, δηλαδή τον αριθμό των δέντρων αποφάσεων που θα εκπαιδευτούν ακολούθως στο μοντέλο. Οι επαναλήψεις αυτές δημιουργούν δέντρα που σκοπός τους είναι να διορθώσουν τα λάθη των προηγούμενων. Η τιμή που επιλέξαμε είναι 3000, η οποία είναι σχετικά υψηλή και μπορεί να οδηγήσει σε overfitting, κίνδυνος που μειώνεται σε συνδυασμό με την παράμετρο min_leaf_in_data. Έχοντας και μικρή τιμή για το learning_rate, το τελικό αποτέλεσμα είναι η αργή αλλά σταθερή βελτίωση του μοντέλου.
- Το boost_from_average καθορίζει εάν το μοντέλο ξεκινάει τις προβλέψεις του για την πρώτη επανάληψη με βάση τη μέση τιμή που έχει ως στόχο ή όχι. Από τη στιγμή που είναι ψευδής αυτή η παράμετρος στο μοντέλο μας, κάθε αρχική πρόβλεψη ξεκινάει από μηδενική τιμή και κάθε μεταγενέστερο δέντρο χτίζει πάνω σε αυτή την τιμή. Συνήθως τα LightGBM μοντέλα ξεκινούν τις προβλέψεις τους χρησιμοποιώντας τον μέσο όρο του στόχου για regression. Παραλείποντας αυτό το βήμα, δεν έχει την πληροφορία αυτή και επεξεργάζεται αμέσως τα δεδομένα που έχει για κάθε χρονοσειρά. Είναι καλή επιλογή να διατηρηθεί ψευδής η παράμετρος όταν υπάρχει ασύμμετρη κατανομή ή απαιτούμε συνάρτηση απώλειας πέρα από τις σύνηθες.

Οι επιλογές των συγκεκριμένων τιμών έγινε με βάση το LightGBM μοντέλο που νίκησε στο κομμάτι της αβεβαιότητας του διαγωνισμού M5. Αφού μελετήσαμε αναλυτικά τις παραμέτρους του LightGBM μοντέλου και εξηγήσαμε για ποιο λόγο έγινε η επιλογή των τιμών τους, μπορούμε να αναφερθούμε στις επεξηγηματικές μεταβλητές που χρησιμοποιήσαμε για την περαιτέρω βελτίωση του. Κάποιες παρέμειναν ίδιες με αυτές που

είχαμε στο ARIMAX, κάποιες τροποποιήθηκαν ως προς τη χρήση τους στην εκπαίδευση και κάποιες προστέθηκαν. Φαίνονται αναλυτικά παρακάτω:

	snap_CA	snap_WI	snap_TX	day	month	holiday
274	0	0	0	1	10	0
275	0	0	0	2	10	0
276	0	0	0	3	10	Cultural
277	1	0	1	4	11	0
278	1	1	0	5	11	0
279	1	1	1	6	11	0
280	1	0	0	7	11	0
281	1	1	1	1	11	0
282	1	1	1	2	11	0
283	1	0	1	3	11	Religious
284	1	1	0	4	11	0
285	1	1	1	5	11	0
286	1	0	0	6	11	0
287	0	1	1	7	11	National
288	0	1	1	1	11	0
289	0	0	1	2	11	0

Πίνακας 3.21 : Επεξηγηματικές μεταβλητές του LightGBM μοντέλου.

Οι τιμές του SNAP ανά πολιτεία και το τι μέρα είναι, παρέμειναν ίδιες και σε αυτό το μοντέλο. Τα ειδικά γεγονότα, τα οποία τα ονομάζουμε holiday για να υπάρχει αντιστοίχιση με το μοντέλο ARIMAX, στο οποίο χρησιμοποιήσαμε μόνο τις αργίες, τροποποιήθηκαν. Πλέον, ούτε λαμβάνουμε τις Εθνικές Εορτές μόνο, αλλά οποιαδήποτε ημέρα υπήρχε κάποιο γεγονός Πολιτιστικό, Θρησκευτικό ή Αθλητικό, το οποίο ίσως επηρέασε τις πωλήσεις των προϊόντων. Ακόμη, δεν ενημερώνουμε το μοντέλο μας για τις αμέσως προηγούμενες ημέρες των ειδικών γεγονότων και χρησιμοποιούμε τον τύπο κάθε ενός γεγονότος, καθώς το μοντέλο των δέντρων αποφάσεων είναι ικανό να τα κατηγοριοποιήσει κατά είδος για να βελτιωθεί. Επιπλέον, προστέθηκε η πληροφορία για το τι μήνας είναι με σκοπό να καταλαβαίνει καλύτερα το μοντέλο την μηνιαία εποχιακότητα.

Στη συνέχεια, ακολουθήσαμε την ίδια λογική με το ARIMAX μοντέλο, δηλαδή παράγουμε τις προβλέψεις για τις επόμενες 364 ημέρες, έχοντας πάντα ως rolling window το διάστημα των 28 ημερών.

Σε αυτό το σημείο, υπολογίζουμε το importance του μοντέλου, δηλαδή την σπουδαιότητα που είχανε οι μεταβλητές κατά την εκπαίδευσή του ως προς τρεις δείκτες, οι οποίοι αναλύονται άμεσα. Η αναλυτική απεικόνιση αυτών για κάθε μοντέλο που θεωρούμε σημαντικό θα γίνει στο επόμενο κεφάλαιο, αυτό των Αποτελεσμάτων.

- Το gain, δηλαδή το κέρδος, το οποίο δείχνει τη συνολική βελτίωση που επέφερε ένα από τα χαρακτηριστικά (features) σε όλες τις διαιρέσεις στις οποίες χρησιμοποιήθηκε. Πιο συγκεκριμένα, δείχνει πόσο συνείσφερε στη μείωση του λάθους και όσο μεγαλύτερη η τιμή του, τόσο μεγαλύτερη η επίδραση του feature

στο μοντέλο. Θεωρείται ο πιο σημαντικός δείκτης από τους τρεις, καθώς δείχνει άμεσα την προβλεπτική δύναμη που είχε το καθένα.

- Το cover, το οποίο μετράει την σχετική αναλογία των παρατηρήσεων που αντιστοιχίζονται στις διαιρέσεις σε σχέση με κάθε feature. Δηλαδή, αντιπροσωπεύει πόσο συχνά αυτό το χαρακτηριστικό χρησιμοποιήθηκε για να διαχωρίσει τα δεδομένα κατά την διαδικασία του χτισίματος του δέντρου. Υψηλή τιμή του δείκτη αυτού για κάποιο feature, υποδεικνύει ότι μπορεί να εφαρμοστεί ευρέως σε πολλές παρατηρήσεις, χωρίς αυτό να υποδηλώνει ισχυρή προβλεπτική δύναμη.
- Το frequency, δηλαδή η συχνότητα, η οποία μετράει τον αριθμό που κάποιο feature χρησιμοποιείται σε οποιαδήποτε διαίρεση σε όλα τα δέντρα του μοντέλου. Ο δείκτης βοηθάει να καταλάβεις τη συνολική χρήση ενός χαρακτηριστικού και ταυτόχρονα υψηλή τιμή του δεν σημαίνει το ίδιο και για τους δύο προηγούμενους δείκτες.

Οι τρεις αυτοί δείκτες είναι σαν ποσοστά και κάθε ένας ξεχωριστά υπολογίζεται σαν μονάδα, δηλαδή αν συνολικά έχουμε έντεκα features στο μοντέλο μας, τότε αθροισμένες οι τιμές του καθενός, για παράδειγμα για το gain, δίνουν την μονάδα.

Κάτι για το οποίο έχουμε παραλείψει να μιλήσουμε μέχρι στιγμής, είναι ο τρόπος αξιολόγησης των μοντέλων μας. Αυτό έχει συμβεί, διότι επιλέξαμε τον δείκτη Scaled Pinball Loss για κάθε ένα από τα μοντέλα μας, τον οποίο έχουμε ήδη αναλύσει αρκετά στην έκταση αυτής της εργασίας. Σε αυτό το σημείο θα υπενθυμίσουμε απλώς τον τύπο, για τον οποίο το u είναι 0.975 :

$$\boldsymbol{SPL(u)} = \frac{1}{h}\frac{\sum_{t=n+1}^{n+h}\big(Y_t - Q_t(u)\big)u\mathbf{1}\{Q_t(u) \le Y_t\} + (Q_t(u) - Y_t)(1-u)\mathbf{1}\{Q_t(u) > Y_t\}}{\frac{1}{n-1}\sum_{t=2}^{n}|Y_t - Y_{t-1}|},$$

Μετά τον υπολογισμό του SPL, υπολογίζουμε και τη σχετική συχνότητα κάθε μοντέλου, για την οποία έχουμε μιλήσει αρκετά στο προηγούμενο Κεφάλαιο, των Δεικτών Ακρίβειας.

Η παρουσίαση και η ανάλυση των Αποτελεσμάτων των μοντέλων για τα οποία συζητήσαμε, θα γίνει στο αμέσως επόμενο Κεφάλαιο.

4. Αποτελέσματα

Στο κεφάλαιο αυτό, παρουσιάζονται τα αποτελέσματα των μοντέλων που χρησιμοποιήσαμε για την πρόβλεψη του δέκατου ($10^{ου}$) επιπέδου ιεράρχησης του διαγωνισμού M5. Θα ξεκινήσουμε πρώτα με τα αποτελέσματα των μοντέλων SES (Simple Exponential Smoothing) και ARIMAX (Autoregressive Integrated Moving Average with Explanatory Variable). Για αυτά τα δύο μοντέλα θα παρουσιάσουμε μόνο τους δείκτες ακρίβειάς του, δηλαδή το υπολογισμένο SPL (Scaled Pinball Loss) των προβλεπόμενων χρονοσειρών.

Στη συνέχεια, θα επικεντρωθούμε στα μοντέλα LightGBM (Light Gradient-Boosting Machine), στα οποία δοκιμάζουμε αρκετές εναλλακτικές για να κατανοήσουμε καλύτερα ποιες επεξηγηματικές μεταβλητές συνεισφέρουν περισσότερο στην καλύτερη εκπαίδευση των δέντρων αποφάσεων ή δεν προσφέρουν σε αυτά, έχοντας ίσως και αρνητικό αντίκτυπο από αυτόν που αναμέναμε. Μιας και τα συγκεκριμένα μοντέλα είναι αυτά που απασχολούν τον σκοπό της παρούσας διπλωματικής εργασίας, θα εμβαθύνουμε αρκετά. Πέρα από τον δείκτη ακρίβειας SPL, θα αναλύσουμε και την σπουδαιότητα των features (χαρακτηριστικών) που χρησιμοποιούμε σε κάποιους από τους συνδυασμούς που κάνουμε. Αυτό θα γίνει μέσω του importance, το οποίο εξετάσαμε σε μεγάλο βαθμό στο προηγούμενο κεφάλαιο. Επιπλέον, θα επικεντρωθούμε και στο αν τα μοντέλα μας τηρήσαν τους περιορισμούς που τους θέσαμε, δηλαδή αν οι προβλέψεις πραγματοποιήθηκαν για διάστημα πρόβλεψης 95%. Αυτό θα συμβεί με το relative frequency (σχετική συχνότητα), η οποία πρέπει να έχει τιμή γύρω στο 0.975.

Αφού παρουσιαστούν και τα αποτελέσματα για τα διάφορα LightGBM μοντέλα μας, θα κάνουμε μία προσπάθεια να βελτιώσουμε το χαμηλότερο δυνατό SPL. Πριν από αυτό, θα αναλύσουμε τα μοντέλα με το χαμηλότερο δείκτη ακρίβειας στις μέρες που υπάρχει SNAP ή όχι, στις μέρες που υπάρχει ειδικό γεγονός ή όχι, καθώς και τι γίνεται σε κάθε κατηγορία προϊόντος (Hobbies, Household και Foods), σε κάθε υποκατηγορία προϊόντος(Hobbies-1, Hobbies-2, Household-1, Household-2, Foods-1, Foods-2 και Foods-3) και τον συνδυασμό τις ημέρες του SNAP/No SNAP, ειδικό γεγονός/όχι ειδικό γεγονός μαζί με τις τρεις κατηγορίες προϊόντων. Με αυτό τον τρόπο, θα λάβουμε τις χαμηλότερες τιμές του δείκτη ακρίβειας και θα τις συνδυάσουμε για να δούμε πόσο βελτιωμένο παρουσιάζεται το τελικό μοντέλο μας και πως θα μπορούσαμε τελικά να χρησιμοποιούμε διαφορετικό μοντέλο ανά διαφορετικές συνθήκες, ανάλογα με το τι μέρα είναι, αν υπάρχει ειδικό γεγονός ή κάποια έκπτωση και να μειώνουμε το λάθος των τελικών προβλέψεων μας.

Μάλιστα, θα προσπαθήσουμε να πάμε ένα βήμα παραπέρα και να δοκιμάσουμε μήπως μπορούμε να βελτιώσουμε το αποτέλεσμα μας με τη μέθοδο της ομαδοποίησης. Αρχικά, αυτό σημαίνει, όπως και νωρίτερα, ότι θα διαλέξουμε τα μοντέλα με το χαμηλότερο δείκτη ακρίβειας στις επεξηγηματικές μεταβλητές που θα χρησιμοποιήσουμε, ανάλογα και την κατηγορία των προϊόντων. Στη συνέχεια, θα εκπαιδεύσουμε εκ νέου αυτά τα μοντέλα στο σύνολο δεδομένων που φαίνεται πως λειτουργούν καλύτερα. Κατά τη διάρκεια των προβλέψεων, κάθε μοντέλο θα χρησιμοποιηθεί συγκεκριμένα για τα δεδομένα για τα οποία εκπαιδεύτηκε και πρακτικά θα παράγει αποτελέσματα μόνο στις ημέρες και υπό τις συνθήκες που χρησιμοποιήθηκαν στην εκπαίδευση του. Σε κάθε άλλη περίπτωση, θα χρησιμοποιηθεί αυτό που αναλογεί. Τελικά, θα παραχθούν

προβλέψεις για τις 3049 χρονοσειρές που θέλουμε. Για την καλύτερη κατανόηση του αναγνώστη, θα εξηγήσουμε αναλυτικά την μέθοδο αυτή αργότερα στο κεφάλαιο, όταν θα έχουμε κάποια αποτελέσματα και θα είναι ξεκάθαρο πως και για ποιες ημέρες έγινε η επιλογή κάθε μοντέλου. Η διαφορά που έχει με τον συνδυασμό αποτελεσμάτων για τον οποίο μιλήσαμε νωρίτερα, έγκειται στο γεγονός ότι τώρα συνδυάζεται η εκπαίδευση των μοντέλων σε αντίθεση με πριν που συνδυάζαμε τα αποτελέσματα των μοντέλων.

4.1 Αποτελέσματα SES και ARIMAX

Τα αποτελέσματα της SES δεν είναι ιδιαίτερα σημαντικά για τον σκοπό της παρούσας εργασίας, όμως τα χρησιμοποιούμε σαν βασικό ορόσημο των μοντέλων μας. Αν για οποιονδήποτε λόγο κάποιο μοντέλο παρουσιάσει χειρότερο δείκτη ακρίβειας από αυτόν, τότε κάτι πάει πολύ στραβά. Από την άλλη, τα αποτελέσματα του ARIMAX είναι σημαντικότερα, καθώς έχουν την πληροφορία κάποιων επεξηγηματικών μεταβλητών που χρησιμοποιούμε και στα μοντέλα των δέντρων αποφάσεων.

Μοντέλο	Συνολικό SPL
Simple Exponential Smoothing (SES)	0,2065

Πίνακας 4.1: Ακρίβεια πρόβλεψης (SPL) μοντέλου SES.

Για αρχή, παρουσιάζουμε την τιμή της SES με τέσσερα δεκαδικά ψηφία και αμέσως μετά συγκρίνουμε ποσοστιαία την τιμή του ARIMAX με αυτή. Αυτό συμβαίνει, ώστε να διαβάζονται καλύτερα τα αποτελέσματα και να είναι προφανής η βελτίωση που παρουσιάζει κάθε μοντέλο. Το ίδιο θα γίνει και στη συνέχεια με το LightGBM, όμως θα συγκρίνονται με το αντίστοιχο ορόσημο που θα χρησιμοποιήσουμε τότε.

Μοντέλο	Συνολικό SPL	Βελτίωση %
Simple Exponential Smoothing (SES)	0,2065	0,00
ARIMAX	0,1888	8,60

Πίνακας 4.2 : Ακρίβεια πρόβλεψης (SPL) μοντέλου ARIMAX.

Τα αποτελέσματα του ARIMAX είναι βελτιωμένα σε σχέση με της SES, γεγονός το οποίο το περιμέναμε, καθώς υπεισέρχονται επεξηγηματικές μεταβλητές που καθιστούν το μοντέλο ικανότερο να καταλάβει συγκεκριμένα μοτίβα και να παράγει ανάλογες προβλέψεις. Βέβαια, θα περιμέναμε αρκετά μεγαλύτερη βελτίωση σε σχέση με το περίπου 10% που βλέπουμε, καθώς υπάρχουν μεταβλητές για την ημέρα και αντιλαμβάνεται η εποχιακότητα που υπάρχει, κάτι που δεν υφίσταται στη SES.

4.2 Αποτελέσματα LightGBM

Ήρθε η στιγμή να παρουσιάσουμε και να αναλύσουμε τα αποτελέσματα των LightGBM μοντέλων που δημιουργήσαμε για την πρόβλεψη των 3049 χρονοσειρών. Αρχικά, είναι σημαντικό να επισημάνουμε ποιο είναι το μοντέλο που θα χρησιμοποιήσουμε σαν ορόσημο, γύρω από το οποίο θα μπορούμε να παρακολουθήσουμε τη βελτίωση όλων των άλλων. Αυτό είναι ένα απλό LightGBM που εκπαιδεύεται με τις παραμέτρους που αναφέραμε στο προηγούμενο κεφάλαιο και έχει ως χαρακτηριστικά (features) του, τις πωλήσεις μονάδων της προηγούμενης ημέρας, της προηγούμενης

εβδομάδας, δύο εβδομάδων νωρίτερα, τεσσάρων εβδομάδων νωρίτερα και έναν μέσο όρο των πωλήσεων των τελευταίων είκοσι οκτώ ημερών. Θα παρουσιάσουμε την τιμή του δείκτη ακρίβειάς του, άμεσα συγκρινόμενη με την αντίστοιχη τιμή της SES, ώστε να έχουμε μια πλήρη εικόνα της βελτίωσης από τα πιο απλά μοντέλα, σε αυτά που περιέχουν δέντρα αποφάσεων.

Μοντέλο	Συνολικό SPL	Βελτίωση %
Simple Exponential Smoothing (SES)	0,2065	0,00
Benchmark LightGBM	0,0852	58,75

Πίνακας 4.3 : Ακρίβεια πρόβλεψης (SPL) μοντέλου benchmark LightGBM.

Η διαφορά ανάμεσα στα δύο μοντέλα είναι χαοτική. Η ποσοστιαία διαφορά είναι σχεδόν 60% και δεν έχουν υπεισέλθει ακόμη οι επεξηγηματικές μεταβλητές που μας απασχολούν κυρίως. Παρατηρώντας αυτή την διαφορά, φαίνεται ακόμα πιο περίεργη η σχετικά μικρή βελτίωση που παρατηρήσαμε ανάμεσα στα δύο μοντέλα νωρίτερα. Βέβαια, είναι σημαντικό να σχολιάσουμε ότι οι μεταβλητές του ARIMAX δεν είναι αυτές που έχει το ορόσημο, οπότε ίσως να μην επηρεάζουν στον ίδιο βαθμό.

Από εδώ και στο εξής, όλες οι ποσοστιαίες διαφορές είναι γύρω από το ορόσημο του LightGBM μοντέλου και όλοι οι επόμενοι συνδυασμοί συγκρίνονται απ' ευθείας με αυτό.

Στη συνέχεια κάναμε αρκετούς συνδυασμούς των επεξηγηματικών μεταβλητών, με σκοπό να κατανοήσουμε όσο καλύτερα μπορούμε τι προσφέρει η κάθε μία και ποιος συνδυασμός προσφέρει το καλύτερο αποτέλεσμα. Οι διάφορες μεταβλητές είναι η ημέρα, ο μήνας, αν έχει SNAP κάθε μία από τις τρεις διαφορετικές πολιτείες και αν εκείνη την ημέρα υπάρχει κάποιο ειδικό γεγονός. Στον παρακάτω πίνακα, αναφέρονται ως day, month, SNAP και special events αντίστοιχα. Μάλιστα, υπάρχει και το πρώτο μοντέλο, το οποίο έχει όλα τα features που έχει το ορόσημο, εκτός από τον μέσο όρο των τελευταίων είκοσι οκτώ ημερών. Αυτό έγινε για να δούμε και πόση σημασία έχει η ύπαρξη αυτού του μέσου όρου ή είναι περιττός.

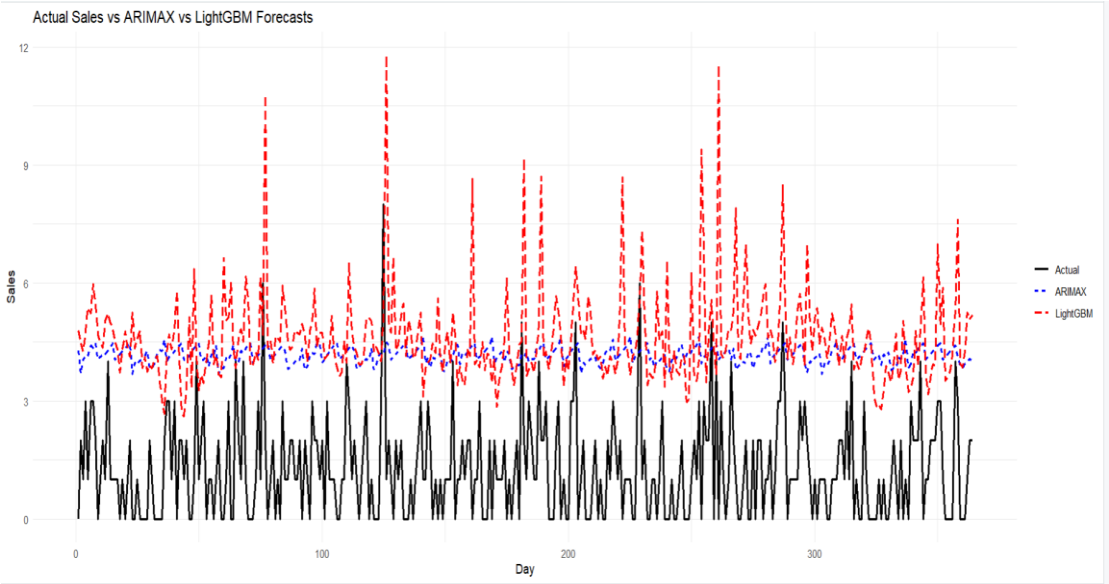
Μοντέλο	Βελτίωση %
LightGBM with only lags	-11,48
Benchmark LightGBM	0,00
Benchmark LightGBM+SNAP+Day+Special Events	2,67
Benchmark LightGBM+SNAP	0,06
Benchmark LightGBM+SNAP+Special Events	-0,21
Benchmark LightGBM+SNAP+Day	2,71
Benchmark LightGBM+SNAP+Day+Special Events+Month	3,13
Benchmark LightGBM+SNAP+Day+Month	3,17
Benchmark LightGBM+Day+Month	3,10

Πίνακας 4.4 : Ακρίβεια πρόβλεψης (SPL) LightGBM μοντέλων.

Η πρώτη παρατήρηση που πρέπει να γίνει είναι ότι ο μέσος όρος είναι ιδιαίτερα σημαντικός και χρειάζεται να συμπεριληφθεί στο ορόσημο. Στη συνέχεια, ας δούμε μία προς μία τις επεξηγηματικές μεταβλητές. Το Day φαίνεται να βελτιώνει σε μεγαλύτερο βαθμό το μοντέλο μας σε σχέση με τα υπόλοιπα, της τάξεως του 2.5% τουλάχιστον.

Αυτό είναι λογικό, καθώς ενισχύει την κατανόηση του μοντέλου σχετικά με την εβδομαδιαία εποχιακότητα και μπορεί να βρει μοτίβα που στηρίζονται κυρίως στο τι ημέρα της εβδομάδας έχουμε. Στη συνέχεια, η μεταβλητή που είμαστε βέβαιοι πως βοηθάει αρκετά είναι το Month, της τάξεως του 0.4%. Παρομοίως με την ημέρα, ενισχύει την ικανότητα των δέντρων αποφάσεως να αναγνωρίσουν την μηνιαία εποχιακότητα. Αυτές οι δύο μεταβλητές είναι ξεκάθαρο πως βελτιώνουν τα μοντέλα μας. Τώρα, όσον αφορά το SNAP, το πρόγραμμα κουπονιών της Αμερικανικής Κυβέρνησης προς τις ασθενέστερες οικονομικά οικογένειες, φαίνεται πως βοηθάνε αλλά σε μικρό βαθμό, της τάξεως του 0.05%. Αργότερα, που θα αναλύσουμε σε κατηγορίες και υποκατηγορίες προϊόντων, θα είμαστε σε θέση να αποφασίσουμε με μεγαλύτερη βεβαιότητα αν μας είναι χρήσιμη αυτή η επεξηγηματική μεταβλητή. Τέλος, τα Special Events, φαίνεται να μπερδεύουν και να αποπροσανατολίζουν σε κάποιο βαθμό το κάθε μοντέλο. Η τάξη που το επηρεάζει αλλάζει ανάλογα με το ποιες μεταβλητές το συνοδεύουν, αλλά είναι φανερό ότι δεν δίνει ιδιαίτερη αξία και βρίσκεται στο όριο του τυχαίου.

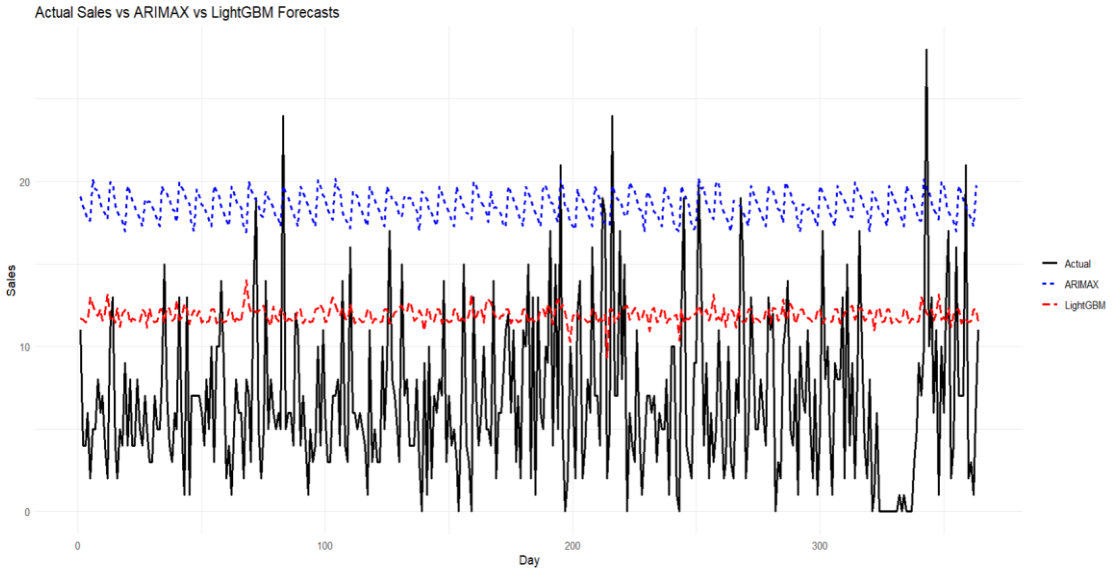
Πριν συνεχίσουμε την ανάλυση των αποτελεσμάτων, είναι ένα καλό σημείο για να συγκρίνουμε τα μοντέλα ARIMAX και LightGBM. Για τα δέντρα αποφάσεων, χρησιμοποιούμε το μοντέλο με το χαμηλότερο SPL από τον πίνακα παραπάνω. Διαλέξαμε τυχαία τρεις χρονοσειρές προβλέψεων από τις 3049 διαθέσιμες. Μαζί με τις προβλέψεις, υπάρχουν και οι πραγματικές πωλήσεις για τις τελευταίες 364 ημέρες.



Σχήμα 4.1 : Πρώτη σύγκριση μοντέλων ARIMAX, LightGBM και πραγματικών πωλήσεων.

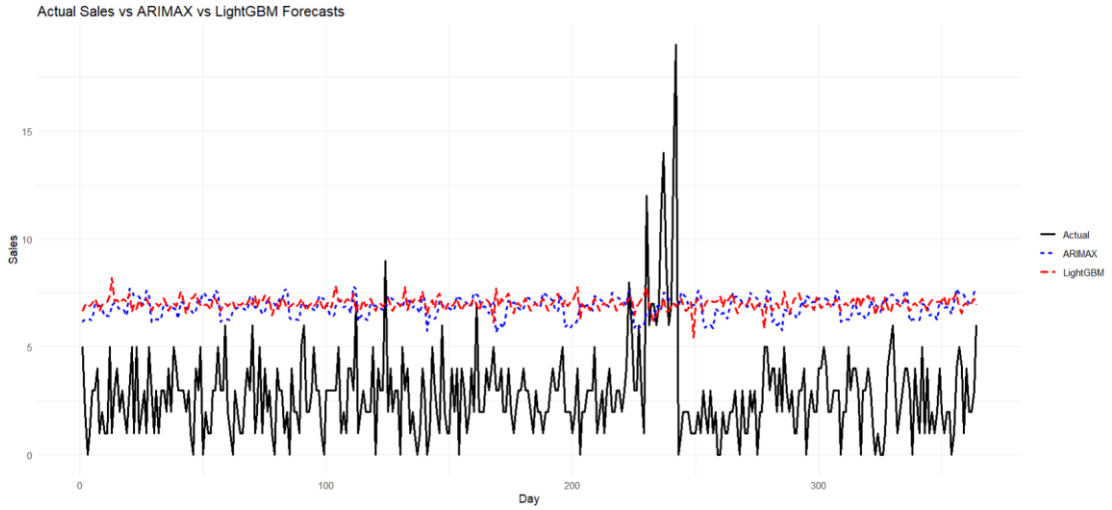
Το πιο σημαντικό που μπορούμε να αποφανθούμε βλέποντας το παραπάνω σχήμα, είναι ότι οι προβλέψεις μας είναι πολύ κοντά στο να υπερτερούν συνεχώς των πραγματικών πωλήσεων, το οποίο είναι και από τα βασικά μας ζητούμενα. Έπειτα, οι προβλέψεις του ARIMAX είναι σχετικά σταθερές κατά την διάρκεια του προβλεπόμενου έτους. Αυτό σημαίνει ότι προβλέπει με ασφάλεια και δεν ακολουθεί τα μοτίβα που παρουσιάζουν στην πραγματικότητα οι πωλήσεις. Φαίνεται να ακολουθεί μια μικρή εποχιακότητα αλλά οι τιμές κυμαίνονται από το 4 έως το 5. Από την άλλη, το LightGBM ακολουθεί αρκετά τα μοτίβα των πωλήσεων μονάδων, είτε αυτά παρουσιάζουν αιχμές

είτε βυθίσματα. Να υπενθυμίσουμε ότι το μοντέλο αυτό έχει ως μεταβλητές ημέρα, μήνα και SNAP. Τα μοτίβα της μηνιαίας και της εβδομαδιαίας εποχιακότητας ακολουθούνται πιστά, επομένως μπορούμε να συμπεράνουμε ότι το μοντέλο λειτουργεί όπως θέλουμε. Οι τιμές του είναι αρκετά παραπάνω από τις πραγματικές, το οποίο συμβαίνει λόγω του διαστήματος πρόβλεψης που έχουμε ζητήσει. Για μικρότερο διάστημα, οι τιμές θα ήταν αρκετά πιο κοντά και τα μοτίβα θα παρέμειναν παρόμοια.



Σχήμα 4.2 : Δεύτερη σύγκριση μοντέλων ARIMAX, LightGBM και πραγματικών πωλήσεων.

Στην συγκεκριμένη χρονοσειρά δεν επικρατεί η ίδια εικόνα με παραπάνω. Οι προβλέψεις του ARIMAX μοντέλου παραμένουν σταθερές γύρω από το επίπεδο τους, όμως εδώ οι προβλέψεις του LightGBM είναι χαμηλότερες και δεν φαίνεται να καταλαβαίνουν πλήρως τα μοτίβα που επικρατούν. Το επίπεδό τους, μάλιστα, είναι αρκετά σταθερό σε σχέση με την προηγούμενη χρονοσειρά.



Σχήμα 4.3 : Τρίτη σύγκριση μοντέλων ARIMAX, LightGBM και πραγματικών πωλήσεων.

Σε αυτή τη σύγκριση, οι προβλέψεις των δύο μοντέλων κυμαίνονται στο ίδιο επίπεδο και παρουσιάζουν πολύ μικρές διαφορές μεταξύ τους. Το πρόβλημα είναι ότι δεν αναμένουν την κορυφή που δημιουργείται.

Σε γενικές γραμμές, οι προβλέψεις κάθε χρονοσειράς αλλάζουν από τη μία στην άλλη και δεν είναι εξίσου καλές για κάθε προϊόν.

Στον παρακάτω πίνακα παρουσιάζονται τα αποτελέσματα της σχετικής συχνότητας (relative frequency) των διάφορων LightGBM μοντέλων:

Μοντέλο	Σχετική Συχνότητα
LightGBM with only lags	0,976
Benchmark LightGBM	0,975
Benchmark LightGBM+SNAP+Day+Special Events	0,974
Benchmark LightGBM+SNAP	0,975
Benchmark LightGBM+SNAP+Special Events	0,975
Benchmark LightGBM+SNAP+Day	0,974
Benchmark LightGBM+SNAP+Day+Special Events+Month	0,974
Benchmark LightGBM+SNAP+Day+Month	0,974
Benchmark LightGBM+Day+Month	0,974
Simple Exponential Smoothing (SES)	0,982
ARIMAX	0,978

Πίνακας 4.5 : Σχετική συχνότητα LightGBM μοντέλων.

Είναι σημαντικό ότι όλες οι τιμές είναι γύρω από το 0.975 που είναι ο στόχος που θέσαμε, επιλέγοντας το quantile να έχει αυτή την τιμή. Τα τελευταία LightGBM μοντέλα, όπως είδαμε και νωρίτερα είναι αυτά που έχουν την χαμηλότερη τιμή SPL. Παρατηρούμε ότι όσο φτάνουμε σε αυτά τα μοντέλα μικραίνει ο δείκτης της σχετικής συχνότητας, αν και ανεπαίσθητα. Δεν μπορούμε να πούμε με σιγουριά ότι αυτοί οι δύο παράγοντες συνδυάζονται, αλλά ίσως με ποιοτικότερη πληροφορία και πετυχαίνοντας καλύτερο αποτέλεσμα, το μοντέλο κάνει τις προβλέψεις του λίγο πιο χαμηλές. Όπως και να έχει, δεν επηρεάζεται πραγματικά το μοντέλο από αυτό.

Θα σχολιάσουμε τώρα την σημασία που έχουν τα χαρακτηριστικά (features) που χρησιμοποιήσαμε για την εκπαίδευση και ακολούθως για την πρόβλεψη των LightGBM μοντέλων μας. Αρχικά, παρουσιάζουμε τα αποτελέσματα του importance για το ορόσημο.

	Feature	Gain	Cover	Frequency
1	mean_last_28_days	0.51354875	0.2969228	0.2630279
2	lag_1	0.26960098	0.1877925	0.1922081
3	lag_7	0.13565079	0.1726444	0.1842152
4	lag_14	0.04157778	0.1763612	0.1837575
5	lag_28	0.03962170	0.1662791	0.1767913

Σχήμα 4.4 : Σημαντικότητα μεταβλητών του benchmark LightGBM μοντέλου.

Το μεγαλύτερο κέρδος παρουσιάζει ο μέσος όρος των τελευταίων είκοσι οκτώ ημερών, πράγμα που σημαίνει ότι συνείσφερε πάνω από 50% στη μείωση του λάθους των δέντρων αποφάσεων σε σχέση με τις υπόλοιπες μεταβλητές. Οπότε ήταν λογικό στα αποτελέσματα νωρίτερα, πόσο χειρότερο γινόταν το ορόσημο όταν απουσίαζε η συγκεκριμένη μεταβλητή. Οι άλλοι δύο δείκτες, cover και frequency παρουσιάζουν και αυτοί την μεγαλύτερη τιμή τους για το ίδιο feature. Επομένως, αυτό χρησιμοποιήθηκε αρκετά συχνότερα στον διαχωρισμό των δεδομένων κατά τη διάρκεια του χτισίματος του δέντρου και γενικότερα σε όλη τη διαδικασία. Το ότι ένα χαρακτηριστικό έχει την υψηλότερη τιμή σε έναν από τους τρεις δείκτες, δεν σημαίνει απαραίτητα ότι θα έχει την υψηλότερη και για τους υπόλοιπους δύο. Στη συνέχεια που θα αναλύσουμε και κάποιους άλλους συνδυασμούς χαρακτηριστικών, θα σχολιάσουμε μόνο το gain, το οποίο είναι και ο πιο αντιπροσωπευτικός δείκτης.

Το επόμενο LightGBM μοντέλο για το οποίο θεωρούμε σημαντικό να παρουσιαστεί και να σχολιαστεί το importance του, είναι αυτό που περιέχει όλες τις επεξηγηματικές μεταβλητές που χρησιμοποιήσαμε.

	Feature	Gain	Cover	Frequency
1	mean_last_28_days	0.4758133811	0.21765639	0.186424311
2	lag_1	0.2326021314	0.15605810	0.131406683
3	lag_7	0.1597629987	0.13984552	0.127475518
4	lag_14	0.0514105889	0.13683500	0.122925118
5	lag_28	0.0411826885	0.13124999	0.122576603
6	month	0.0187642556	0.08675200	0.134197913
7	day	0.0134263678	0.06407878	0.096390279
8	holiday	0.0042605697	0.02153332	0.008376803
9	snap_CA	0.0014245417	0.01628572	0.025106551
10	snap_TX	0.0008040324	0.01475260	0.022333991
11	snap_WI	0.0005484441	0.01495257	0.022786231

Σχήμα 4.5 : Σημαντικότητα μεταβλητών LightGBM με όλες τις επεξηγηματικές μεταβλητές.

Εδώ δεν αλλάζει η κατάσταση όσον αφορά τα features που υπήρχαν στο ορόσημο, απλώς μειώνονται τα ποσοστά των δύο σημαντικότερων, πράγμα που είναι λογικό, καθώς έχουν προστεθεί άλλες έξι επεξηγηματικές μεταβλητές. Παρατηρούμε όμως, ότι οι άλλες τρεις μεταβλητές lag αυξάνουν το ποσοστό τους σε σχέση με νωρίτερα, κάτι που δεν είναι το αναμενόμενο και φαίνεται πως τα χαρακτηριστικά που υπεισήλθαν στην εκπαίδευση του μοντέλου, τα καθιστούν αυτόματα πιο σημαντικά. Λογικά, το μοντέλο μας κατανοεί καλύτερα κάποια μοτίβα που δημιουργούνται λόγω του συνδυασμού των features μεταξύ τους. Τώρα, όσον αφορά τις υπόλοιπες μεταβλητές, ο μήνας είναι σημαντικότερος στη διόρθωση του λάθους σε σχέση με την ημέρα, ενώ στα αποτελέσματα

του δείκτη ακρίβειας συμβαίνει το αντίθετο. Επομένως, η μηνιαία εποχιακότητα συμβάλλει περισσότερο στη διόρθωση του λάθους κατά το χτίσιμο των δέντρων αποφάσεων ενώ η εβδομαδιαία, στην τελική πρόβλεψη του μοντέλου. Τα ειδικά γεγονότα, που τα έχουμε ονομάσει holiday, τουλάχιστον σε αυτό το δείκτη, είναι σημαντικότερα και από τα τρία SNAP για κάθε πολιτεία αθροισμένα μαζί. Βέβαια, αυτό δεν μεταφράζεται σε καλύτερο αποτέλεσμα τελικά, καθώς, όπως είδαμε, μπερδεύει το μοντέλο και δεν συνεισφέρει στη μείωση του SPL. Από την άλλη το SNAP φαίνεται να συμβάλλει ελάχιστα σε αυτόν τον δείκτη, αλλά συνολικά χρησιμεύει ιδιαίτερα στα τελικά αποτελέσματα. Το γεγονός ότι, για την Καλιφόρνια, το ποσοστό είναι αρκετά μεγαλύτερο σε σχέση με τα άλλα, οφείλεται στις περισσότερες πωλήσεις και στο ένα παραπάνω κατάστημα που έχει. Τον λόγο που το Γουισκόνσιν είναι αρκετά χαμηλότερα από το Τέξας δεν μπορούμε να το αποδώσουμε κάπου.

Τέλος, θα παρουσιάσουμε το importance του μοντέλου με το μικρότερο SPL, ώστε να το κατανοήσουμε καλύτερα.

	Feature	Gain	Cover	Frequency
1	mean_last_28_days	0.4075253999	0.20982973	0.17905796
2	lag_1	0.2916160772	0.15808633	0.13041676
3	lag_7	0.1549642572	0.14380973	0.12563985
4	lag_28	0.0649597010	0.13853823	0.12983714
5	lag_14	0.0494793651	0.14332322	0.13056193
6	month	0.0155812279	0.08756554	0.13145188
7	day	0.0140477418	0.06924936	0.09944730
8	snap_CA	0.0008407572	0.01776566	0.02630195
9	snap_WI	0.0005106843	0.01553038	0.02315346
10	snap_TX	0.0004747883	0.01630182	0.02413177

Σχήμα 4.6 : Σημαντικότητα μεταβλητών LightGBM με χαμηλότερο SPL.

Για το συγκεκριμένο μοντέλο δεν παρατηρούνται σημαντικές διαφορές σε σχέση με το προηγούμενο, αλλά υπάρχουν σημεία που πρέπει να σχολιαστούν. Αρχικά, η κατάταξη των features αλλάζει σε δύο περιπτώσεις. Πλέον είναι σημαντικότερη η μεταβλητή των είκοσι οκτώ ημερών νωρίτερα σε σχέση με αυτή των δεκατεσσάρων ημερών νωρίτερα και το SNAP του Γουισκόνσιν ξεπερνάει αυτό του Τέξας. Μειώνεται η διαφορά μεταξύ της ημέρας και του μήνα και πλέον είναι πολύ κοντά η μία στην άλλη. Τα SNAP είναι παρόμοια μεταξύ τους, ειδικά αν λάβουμε υπόψιν ότι η Καλιφόρνια έχει ένα περισσότερο κατάστημα. Η επίδραση του μέσου όρου μειώνεται αρκετά σε σχέση με πριν και αυξάνεται αρκετά η επίδραση που έχει η τιμή της προηγούμενης ημέρας. Συνολικά, μπορούμε να πούμε ότι όσο μειώνεται ο δείκτης ακρίβειας, που είναι και το ζητούμενο, τόσο μειώνεται η επίδραση, στο κομμάτι της εκπαίδευσης, του μέσου όρου.

Για να εμβαθύνουμε περισσότερο στα αποτελέσματα και τον ρόλο που παίζει κάθε μοντέλο, θα χρησιμοποιήσουμε τα τρία τελευταία, δηλαδή αυτά που παρουσιάζουν βελτίωση 3% και περισσότερο σε σχέση με το ορόσημο. Για αυτά τα μοντέλα, αρχικά, παρουσιάζουμε την βελτίωση που έχουν, πάντα σε σχέση με το ορόσημο, συγκεκριμένα τις ημέρες που υπάρχει και δεν υπάρχει SNAP και τις ημέρες που υπάρχουν ειδικά γεγονότα (special events) και αυτές που δεν υπάρχουν. Θα παρατηρήσουμε, δηλαδή, ποιο μοντέλο λειτουργεί καλύτερα σε συγκεκριμένες ημέρες.

Μοντέλο	SPL with SNAP	SPL without SNAP
Benchmark LightGBM+SNAP+Day+Special Events+Month	3,36	2,90
Benchmark LightGBM+SNAP+Day+Month	3,40	2,94
Benchmark LightGBM+Day+Month	3,05	3,14

Πίνακας 4.6 : Ακρίβεια πρόβλεψης (SPL) ημέρες με και χωρίς SNAP.

Πριν την ανάλυση του παραπάνω πίνακα, είναι σημαντικό να επισημανθεί ότι από τις 364 ημέρες που προβλέπουμε, οι 180 έχουν τουλάχιστον μία φορά SNAP σε κάποια από τις τρεις πολιτείες ενώ στις 184 δεν υπάρχει σε καμία. Επομένως, όσον αφορά την επεξηγηματική μεταβλητή SNAP, οι ημέρες είναι χωρισμένες στη μέση. Παρατηρούμε ότι, τα δύο μοντέλα που χρησιμοποιούν αυτή την μεταβλητή κατά την εκπαίδευσή τους, έχουν καλύτερο αποτέλεσμα τις ημέρες που υπάρχει το SNAP και είναι αρκετά κοντά μεταξύ τους. Όμως, τις υπόλοιπες ημέρες παρουσιάζει πολύ καλύτερα αποτελέσματα, το μοντέλο που δεν την περιέχει.

Μοντέλο	SPL with Special Events	SPL without Special Events
Benchmark LightGBM+SNAP+Day+Special Events+Month	2,40	3,19
Benchmark LightGBM+SNAP+Day+Month	3,00	3,18
Benchmark LightGBM+Day+Month	2,89	3,12

Πίνακας 4.7 : Ακρίβεια πρόβλεψης (SPL) ημέρες με και χωρίς ειδικά γεγονότα.

Από τις 364 ημέρες που προβλέπουμε, έχουμε μόνο 29 ημέρες που έχουν ειδικά γεγονότα, δηλαδή ένα ποσοστό της τάξεως του 8%. Οπότε, ακόμα και αν τα μοντέλα παρουσίαζαν μεγάλη διαφορά τις ημέρες που έχουμε, η τελική διαφορά στον δείκτη ακρίβειας θα ήταν αρκετά μικρή, καθώς δεν συμβάλλει σε μεγάλο ποσοστό. Σε αυτόν τον πίνακα παρατηρούμε κάτι μη αναμενόμενο. Τις ημέρες των ειδικών γεγονότων, το μοντέλο που έχει εκπαιδευτεί με αυτές είναι αρκετά χειρότερο από τα υπόλοιπα, ενώ στις υπόλοιπες είναι πρακτικά ισάξιο με τα άλλα δύο με πολύ μικρές διαφορές, Αυτό σημαίνει ότι η επεξηγηματική μεταβλητή των ειδικών γεγονότων μπερδεύει αρκετά το μοντέλο, τουλάχιστον για τις συγκεκριμένες ημέρες και δεν καταλαβαίνει τα περίπλοκα μοτίβα που δημιουργούνται υπό αυτές τις συνθήκες.

Στη συνέχεια, θα δούμε τα αποτελέσματα του δείκτη ακρίβειας σε κάθε κατηγορία προϊόντος, δηλαδή τι συμβαίνει στα Χόμπι, στα Οικιακά και στα Φαγητά κατά την διάρκεια των 364 ημερών πρόβλεψης.

Μοντέλο	HOBBIES	HOUSEHOLD	FOODS
Benchmark LightGBM+SNAP+Day+Special Events+Month	-0,60	4,24	3,79
Benchmark LightGBM+SNAP+Day+Month	-0,64	4,31	3,83
Benchmark LightGBM+Day+Month	-0,56	4,39	3,60

Πίνακας 4.8 : Ακρίβεια πρόβλεψης (SPL) για κάθε κατηγορία προϊόντος.

Παρατηρούμε από τον παραπάνω πίνακα, ότι αρχικά τα Χόμπι δείχνουν χειρότερο αποτέλεσμα σε σχέση με το συνολικό του οροσήμου. Λογικά, οι πωλήσεις των συγκεκριμένων προϊόντων δεν επηρεάζονται από το SNAP και από κάποιο ειδικό γεγονός, αλλά παρουσιάζουν μία σταθερή συμπεριφορά κατά την περίοδο του ενός έτους, η οποία γίνεται αντιληπτή από τις τιμές lag στις οποίες έχουμε ήδη αναφερθεί. Είναι σίγουρο ότι παίζει ρόλο τι είδος Χόμπι είναι και σε ποια υποκατηγορία υπόκεινται αλλά αυτό θα αναλυθεί αργότερα στην εργασία. Τα τρία μοντέλα, για αυτή την κατηγορία, φαίνεται να δίνουν παρόμοιο αποτέλεσμα, με καλύτερο αυτό που δεν έχει τις δύο επεξηγηματικές μεταβλητές που αναφέραμε. Από την άλλη, τα Οικιακά και τα Φαγητά δίνουν πολύ καλύτερο αποτέλεσμα για το SPL τους σε σχέση τόσο με το ορόσημο, όσο και με τα μοντέλα που εξετάζουμε, που έδιναν μία βελτίωση της τάξεως του 3%. Τα Οικιακά, συγκεκριμένα, φαίνεται να μην επηρεάζονται από το SNAP, σε αντίθεση με τα Φαγητά τα οποία δίνουν καλύτερο αποτέλεσμα όταν υπάρχει στην εκπαίδευση. Αυτό είναι κάτι το αναμενόμενο, καθώς τα εκπτωτικά αυτά κουπόνια είναι κυρίως για τρόφιμα και παρόμοια είδη, δηλαδή για προϊόντα καθημερινής κατανάλωσης και όχι τόσο για προϊόντα χρήσιμα για ένα σπίτι.

Είναι εξίσου σημαντικό να διερευνηθούν τα αποτελέσματα των υποκατηγοριών στις οποίες χωρίζεται κάθε κατηγορία προϊόντος. Για καλύτερη κατανόηση των αποτελεσμάτων, θεωρείται απαραίτητο να υπενθυμιστεί πόσους κωδικούς προϊόντων έχει κάθε υποκατηγορία, ώστε να εμβαθύνουμε σε αυτές που έχουν τον μεγαλύτερο αριθμό. Τα Hobbies-1 είναι 416, τα Hobbies-2 149, τα Household-1 532, τα Household-2 515, τα Foods-1 216, τα Foods-2 398 και τα Foods-3 823.

Μοντέλο	HOB-1	HOB-2	HOUS-1	HOUS-2	FOODS-1	FOODS-2	FOODS-3
Benchmark LightGBM+SNAP+Day+Special Events+Month	9,03	-27,48	0,00	7,02	-1,98	-2,34	8,26
Benchmark LightGBM+SNAP+Day+Month	8,97	-27,48	1,57	7,14	-1,99	-2,36	8,35
Benchmark LightGBM+Day+Month	9,09	-27,50	1,66	7,20	-1,86	-2,78	8,12

Πίνακας 4.9 : Ακρίβεια πρόβλεψης (SPL) για κάθε υποκατηγορία προϊόντος.

Το πρώτο σημείο του παραπάνω πίνακα που προσέχει κάποιος είναι το πόσο χειρότερα αποτελέσματα έχει η δεύτερη υποκατηγορία των Χόμπι σε σχέση με το ορόσημο και στα τρία διαθέσιμα μοντέλα, τα οποία συγκρινόμενα μεταξύ τους κυμαίνονται στα ίδια επίπεδα. Ο λόγος που συμβαίνει αυτό δεν είναι προφανής και ούτε θα εξηγηθεί στα πλαίσια της συγκεκριμένης εργασίας. Το μόνο που μπορούμε να σχολιάσουμε, είναι ότι λογικά δεν υπάρχει κάποιο ξεκάθαρο μοτίβο πωλήσεων στην συγκεκριμένη κατηγορία και ότι ιδιαίτερο ρόλο παίζει η τυχαιότητα, καθώς αν ακολουθούσε παρόμοιες πωλήσεις μονάδων σε κάθε μήνα, δεν θα ήταν τόσο χαμηλό το τελικό αποτέλεσμα. Το σημαντικό στη συγκεκριμένη περίπτωση είναι ότι αυτή η υποκατηγορία, έχει και τους λιγότερους κωδικούς προϊόντων, επομένως το συνολικό τελικό αποτέλεσμα δεν επηρεάζεται τόσο. Από την άλλη, η πρώτη κατηγορία των Χόμπι παρουσιάζει σημαντική

βελτίωση και στα τρία μοντέλα και ειδικά σε αυτό που εκπαιδεύτηκε μόνο με τον μήνα και την ημέρα, πράγμα που μας οδηγεί στο συμπέρασμα ότι δεν φταίει η κατηγορία συνολικά, αλλά το πως γίνεται ο διαχωρισμός των δύο. Μία εικασία που μπορεί να γίνει, είναι ότι τα προϊόντα της δεύτερης υποκατηγορίας πωλούνται σχετικά σπάνια και έχουν συνεχώς μηδενικές πωλήσεις, όμως, ζητώντας από το μοντέλο διάστημα πρόβλεψης 95%, αυτό προσπαθεί να προβλέπει πάντα περισσότερο από το μηδέν και έτσι να εξηγείται το -27.5%. Για τα υπόλοιπα αποτελέσματα, τα Οικιακά και στις δύο υποκατηγορίες έχουν ως καλύτερο αποτέλεσμα, αυτό του τελευταίου μοντέλου, καθώς όπως αναφέραμε νωρίτερα, δεν επηρεάζονται αρκετά οι πωλήσεις του από τις επεξηγηματικές μεταβλητές του SNAP και των Ειδικών Γεγονότων. Τα Φαγητά παρουσιάζουν μία ποικιλομορφία στα αποτελέσματά τους, καθώς κάθε μία από τις τρεις υποκατηγορίες, παρουσιάζει βέλτιστο αποτέλεσμα σε διαφορετικό μοντέλο. Μάλιστα, οι δύο πρώτες υποκατηγορίες είναι χειρότερες σε σχέση με το ορόσημο, ενώ η τελευταία είναι αρκετά καλύτερη του. Η συνολική εικόνα για τα τρία μοντέλα, είναι πως το τελευταίο παρουσιάζει το καλύτερο αποτέλεσμα στις περισσότερες υποκατηγορίες, ακολουθούμενο από το πρώτο και μετά το δεύτερο. Παρ' όλα αυτά, το δεύτερο μοντέλο έχει συνολικά τον μικρότερο δείκτη ακρίβειας και αυτό οφείλεται στο γεγονός ότι συνήθως στις υποκατηγορίες που είναι χειρότερο, είναι πάντα αρκετά κοντά στο καλύτερο αλλά κυρίως στο ότι είναι το βέλτιστο στην τρίτη υποκατηγορία των φαγητών, που έχει με διαφορά τους περισσότερους κωδικούς από οποιαδήποτε άλλη. Τέλος, η μόνη υποκατηγορία που υπάρχει μεγάλη διαφορά ανάμεσα σε κάποιο μοντέλο σε σχέση με τα υπόλοιπα δύο, είναι η πρώτη των Οικιακών, όπου φαίνεται ότι η ύπαρξη των Ειδικών Γεγονότων στην εκπαίδευση του πρώτου μοντέλου, το μπερδεύει αρκετά.

Ως τελευταία ανάλυση αποτελεσμάτων των μοντέλων που ήδη έχουμε συζητήσει, θεωρούμε σημαντικό να παρατηρήσουμε πως συμπεριφέρεται κάθε κατηγορία κάθε ενός μοντέλου τις ημέρες που υπάρχει SNAP και αυτές που δεν υπάρχει. Δεν κάνουμε το ίδιο για τα Ειδικά Γεγονότα, καθώς έχουμε ήδη καταλάβει ότι δεν προσφέρουν ιδιαίτερα στο τελικό αποτέλεσμα και οι μικρές βελτιώσεις που μπορούν να παρατηρηθούν, είναι στο όριο του τυχαίου.

Μοντέλο	HOBwSNAP	HOBwoSNAP	HOUSwSNAP	HOUSwoSNAP	FOODSwSNAP	FOODSwoSNAP
BenchmarkLightGBM+SNAP+Day+SpecialEvents+Month	0,12	-1,30	5,32	3,18	3,21	4,35
BenchmarkLightGBM+SNAP+Day+Month	0,07	-1,34	5,38	3,26	3,27	4,38
BenchmarkLightGBM+Day+Month	0,00	-1,11	5,25	3,55	2,65	4,52

Πίνακας 4.10 : Ακρίβεια πρόβλεψης (SPL) για κάθε κατηγορία προϊόντος με και χωρίς SNAP.

Εδώ η κατάσταση είναι πιο ξεκάθαρη και πιο κατανοητή με μία πρώτη ματιά. Αν εξαιρέσουμε την κατηγορία των Χόμπι τις ημέρες που υπάρχει SNAP, όπου το καλύτερο μοντέλο είναι το πρώτο αλλά με πολύ μικρή διαφορά από το δεύτερο, της τάξεως του 0.05%, τότε μπορούμε να συνοψίσουμε ότι το δεύτερο μοντέλο λειτουργεί καλύτερα για κάθε κατηγορία τις ημέρες που έστω μία από τις τρεις πολιτείες έχει SNAP, ενώ το τρίτο μοντέλο είναι το βέλτιστο για κάθε κατηγορία όταν δεν υπάρχει. Αξιοσημείωτο είναι ότι τα Χόμπι, χωρίς SNAP, είναι η μόνη κατηγορία που παρουσιάζει χειρότερο αποτέλεσμα σε σχέση με το ορόσημο, ενώ με SNAP, παρουσιάζει μία ελάχιστη βελτίωση. Επομένως, μπορούμε να πούμε με μεγαλύτερη σιγουριά πλέον, ότι οι

επεξηγηματικές μεταβλητές που έχουμε προσθέσει δίνουν πολύ καλύτερο αποτέλεσμα στα Οικιακά και στα Φαγητά, αλλά στα Χόμπι όχι τόσο. Να θυμίσουμε βέβαια ότι η σύγκριση γίνεται με το συνολικό αποτέλεσμα του μοντέλου ορόσημο που χρησιμοποιούμε, οπότε δεν σημαίνει ότι δεν βελτιώνεται το τελικό αποτέλεσμα για την συγκεκριμένη κατηγορία, αλλά δεν βελτιώνεται όσο αναμέναμε.

4.3 Συνδυασμός Μοντέλων LightGBM

Ολόκληρη η ανάλυση των αποτελεσμάτων που έγινε παραπάνω, είχε ως σκοπό την καλύτερη κατανόηση των μοντέλων και την προετοιμασία για τον συνδυασμό τους, ώστε τελικά να μπορέσουμε να παράγουμε ένα καλύτερο αποτέλεσμα, με έναν χαμηλότερο δείκτη ακρίβειας SPL. Με βάση το ποιο μοντέλο έδινε το καλύτερο αποτέλεσμα κάτω από τις συγκεκριμένες συνθήκες που θέταμε κάθε φορά, κάναμε τέσσερις συνδυασμούς, οι οποίοι τελικά δίνουν βελτίωση του SPL από το 3.17% που είχαμε ως καλύτερο.

Μοντέλο	Βελτίωση %
Βέλτιστο SPL για κάθε υποκατηγορία	3,22
Βέλτιστο SPL για κάθε κατηγορία	3,21
Βέλτιστος συνδυασμός με και χωρίς SNAP	3,27
Βέλτιστο SPL για κάθε κατηγορία με και χωρίς SNAP	3,28

Πίνακας 4.11 : Ακρίβεια πρόβλεψης (SPL) για κάθε συνδυασμό μοντέλου.

Πιο αναλυτικά, ο πρώτος συνδυασμός που βλέπουμε, δηλαδή ο χαμηλότερος δείκτης ακρίβειας για κάθε υποκατηγορία, έγινε με βάση τον Πίνακα 15, όπου επιλέξαμε το πρώτο μοντέλο για την δεύτερη υποκατηγορία των Χόμπι και των Φαγητών, το δεύτερο μοντέλο για την τρίτη υποκατηγορία των Φαγητών και το τρίτο μοντέλο για όλες τις υπόλοιπες υποκατηγορίες. Η βελτίωση σε σχέση με το δεύτερο από τα τρία μοντέλα στα οποία εμβαθύναμε, είναι της τάξεως του 0.05%.

Ο δεύτερος συνδυασμός που βλέπουμε, δηλαδή ο χαμηλότερος δείκτης ακρίβειας για κάθε κατηγορία, έγινε με βάση τον Πίνακα 14, όπου επιλέξαμε το δεύτερο μοντέλο για την κατηγορία των Φαγητών και το τρίτο μοντέλο για τις κατηγορίες των Χόμπι και των Οικιακών. Η βελτίωση σε σχέση με το δεύτερο μοντέλο είναι της τάξεως του 0.04% και σε σχέση με τον πρώτο συνδυασμό μειωμένο μόλις κατά 0.01%. Συμπεραίνουμε λοιπόν, ότι η υπερβολική ανάλυση στις υποκατηγορίες δίνει βελτίωση αλλά ανεπαίσθητη συνολικά.

Ο τρίτος συνδυασμός που βλέπουμε, δηλαδή ο χαμηλότερος δείκτης ακρίβειας για τις ημέρες που υπάρχει SNAP και αυτές που δεν υπάρχει, έγινε με βάση τον Πίνακα 12, όπου επιλέξαμε το δεύτερο μοντέλο για τις ημέρες που υπάρχει και το τρίτο για αυτές που δεν υπάρχει. Η βελτίωση σε σχέση με το δεύτερο μοντέλο είναι της τάξεως του 0.1%, που είναι αρκετά αξιόλογη. Παρατηρούμε ότι, χωρίς να χρειαστεί καν να διαχωρίσουμε σε κατηγορίες ή σε υποκατηγορίες τα αποτελέσματά μας, η επεξηγηματική μεταβλητή SNAP και η ύπαρξη της σε κάποιο μοντέλο ή απλώς η πληροφορία για το αν υπάρχει σε κάποια πολιτεία την συγκεκριμένη ημέρα, βελτιώνει σημαντικά το αποτέλεσμά μας.

Ο τέταρτος και τελευταίος συνδυασμός αυτού του είδους που βλέπουμε, δηλαδή ο χαμηλότερος δείκτης ακρίβειας για κάθε κατηγορία τις ημέρες που υπάρχει SNAP και αυτές που δεν υπάρχει, έγινε με βάση τον πίνακα 16, όπου επιλέξαμε το πρώτο μοντέλο για τα Χόμπι τις ημέρες που υπάρχει SNAP, το δεύτερο μοντέλο για τα Οικιακά και τα Φαγητά τις ίδιες ημέρες και το τρίτο μοντέλο για κάθε κατηγορία όταν δεν υπάρχει. Η βελτίωση σε σχέση με το δεύτερο μοντέλο είναι της τάξεως του 0.11%. Βέβαια, και πάλι μπορούμε να αποφανθούμε ότι η υπερβολική ανάλυση και ο διαχωρισμός σε κατηγορίες δεν βοήθησε όσο θα αναμέναμε, καθώς η διαφορά με τον αμέσως προηγούμενο συνδυασμό είναι της τάξεως του 0.01% μόλις.

Είδαμε, λοιπόν, πως και οι τέσσερις συνδυασμοί παρουσίασαν τελικά καλύτερο αποτέλεσμα, πράγμα αναμενόμενο αφού απλώς χειριστήκαμε κατάλληλα τις βέλτιστες τιμές κάθε προηγούμενου πίνακα. Καταλήγουμε στο συμπέρασμα ότι έπρεπε να γίνουν οι προηγούμενες αναλύσεις και οι διαχωρισμοί σε ημέρες SNAP και μη, όμως ο διαχωρισμός των αποτελεσμάτων σε κατηγορίες και υποκατηγορίες μπορούσε τελικά να αποφευχθεί, μιας και δεν υπήρξε η αντίστοιχη βελτίωση, στον βαθμό δηλαδή που να δικαιολογεί την περαιτέρω ανάλυση.

Τέλος, θα δοκιμάσουμε έναν πιο ιδιαίτερο συνδυασμό μοντέλων, ο οποίος θα γίνει με τη μέθοδο επιλογής μοντέλου (model selection) (In & Jung, 2021). Για αυτή τη μέθοδο, θα διαλέξουμε από τα μοντέλα που παρουσίασαν τον χαμηλότερο δείκτη ακρίβειας. Πιο συγκεκριμένα, με βάση τον Πίνακα 16, όπου έχουμε τα SPL για κάθε κατηγορία και χωρισμένα με το αν υπάρχει ή όχι SNAP εκείνη την ημέρα, θα χρησιμοποιήσουμε το δεύτερο και το τρίτο. Αυτή η επιλογή έγινε καθώς το τρίτο παρουσιάζει τα καλύτερα αποτελέσματα σε κάθε κατηγορία τις ημέρες που δεν υπάρχει SNAP, ενώ το δεύτερο παρουσιάζει τα καλύτερα αποτελέσματα στα Οικιακά και στα Φαγητά και το δεύτερο καλύτερο στα Χόμπι τις ημέρες που υπάρχει. Δεν λάβαμε καθόλου υπόψιν το πρώτο μοντέλο, για να διευκολύνουμε τον συνδυασμό που θα πραγματοποιήσουμε, καθώς η διαφορά της τάξεως του 0.05% που έχουν για τα Χόμπι, δεν δικαιολογεί την πολυπλοκότητα που θα προστεθεί στην εκπαίδευση του μοντέλου που θα δημιουργήσουμε.

Η μεγάλη διαφορά που έχει το μοντέλο στο οποίο αναφερόμαστε σε σχέση με τους προηγούμενους συνδυασμούς και ο λόγος που το ξεχωρίζουμε και δεν το σχολιάζουμε με τα υπόλοιπα, είναι ότι εκπαιδεύεται από την αρχή και δεν παίρνει απλώς τα καλύτερα αποτελέσματα του δείκτη ακρίβειας και τα συνδυάζει.

Η μέθοδος επιλογής μοντέλου που αναφέραμε, είναι πρακτικά ο διαχωρισμός του συνόλου δεδομένων που έχουμε στη διάθεσή μας. Ο τρόπος διαχωρισμού που θα χρησιμοποιήσουμε στην περίπτωσή μας, είναι οι ημέρες που υπάρχει SNAP και οι ημέρες που δεν υπάρχει. Επιλέξαμε την συγκεκριμένη επεξηγηματική μεταβλητή, καθώς, κατά την ανάλυση των αποτελεσμάτων, παρατηρήσαμε ότι, η ύπαρξη της ή αντίστοιχα η έλλειψη της, συμβάλλει αρκετά στην βελτίωση των μοντέλων μας. Η άλλη επιλογή ήταν τα Ειδικά Γεγονότα, τα οποία όμως δεν συμβάλλουν στον επιθυμητό βαθμό και η συνεισφορά στη βελτίωση κάποιου μοντέλου είναι στα όρια του τυχαίου.

Με αυτόν τον διαχωρισμό, μπορούμε να χρησιμοποιήσουμε κάθε ένα από τα δύο μοντέλα που επιλέξαμε στο επιθυμητό σύνολο δεδομένων. Δηλαδή, το μοντέλο που λειτουργεί καλύτερα κατά τις ημέρες που υπάρχει το SNAP, εκπαιδεύεται μόνο στις συγκεκριμένες ημέρες και κάνει προβλέψεις μόνο για τις αντίστοιχες ημέρες. Από την

άλλη, αυτό που λειτουργεί βέλτιστα όταν δεν υπάρχει SNAP, εκπαιδεύεται αποκλειστικά για τις συγκεκριμένες ημέρες και αντίστοιχα κάνει προβλέψεις μόνο για αυτές. Το τελικό αποτέλεσμα που λαμβάνουμε, είναι ο συνδυασμός των δύο, ανάλογα ποιες προϋποθέσεις πληρούνται.

Παρακάτω είναι το αποτέλεσμα του συνδυασμού μοντέλων με την μέθοδο επιλογής μοντέλου.

Μοντέλο	Βελτίωση%	Σχετική Συχνότητα
Benchmark LightGBM	0,00	0,975
Συνδυασμός με επιλογή μοντέλου για SNAP/No SNAP ημέρες	2,44	0,974

Πίνακας 4.12 : Ακρίβεια πρόβλεψης (SPL) και σχετική συχνότητα συνδυασμού με επιλογή μοντέλου.

Το τελικό αποτέλεσμα είναι προφανώς βελτιωμένο σε σχέση με το ορόσημο, αλλά καθόλου κοντά στους συνδυασμούς που κάναμε νωρίτερα και στα μοντέλα που εμβαθύναμε περισσότερο και τα οποία παρουσίαζαν το χαμηλότερο SPL. Είναι πιθανό, ο συνδυασμός αυτός, να μπερδεύει τελικά το μοντέλο μας και να καταλήγει να επικεντρώνεται σε σημεία που δεν υπάρχει ιδιαίτερη ανάγκη, να κάνει δηλαδή υπερπροσαρμογή και με αυτό τον τρόπο να μην καταλαβαίνει τα σημαντικά μοτίβα που παρουσιάζει η κάθε μία χρονοσειρά. Η σχετική συχνότητα κυμαίνεται στα ίδια επίπεδα με όλα τα μοντέλα, επομένως δεν υπάρχει λόγος σχολιασμού επ' αυτού.

5. Συμπεράσματα και Προεκτάσεις

Η παρούσα διπλωματική εργασία είχε ως σκοπό να εξετάσει την χρησιμότητα των εμπειρικών κατανομών στην παραγωγή προβλέψεων για τον κλάδο των λιανικών πωλήσεων, καθώς και την βελτίωση που μπορούν να προσφέρουν διάφορες επεξηγηματικές μεταβλητές, όπως τι μέρα είναι, τι μήνας, αν υπάρχει έκπτωση ή κάποιο ειδικό γεγονός εκείνη την ημέρα.

Στα πρώτα κεφάλαια επεξηγήθηκε το θεωρητικό υπόβαθρο στο οποίο βασίζεται η εργασία. Αναλύθηκε ο ρόλος της πρόβλεψης στον κλάδο της λιανικής και στη συνέχεια παρουσιάστηκαν οι μέθοδοι που χρησιμοποιούνται πιο συχνά για τις πωλήσεις προϊόντων. Μελετήθηκε η μηχανική μάθηση με τη χρήση δέντρων αποφάσεων και παρουσιάστηκαν οι βασικότεροι δείκτες ακρίβειας που χρησιμοποιούνται στον τομέα των προβλέψεων, καθώς και αυτός που χρησιμοποιήθηκε συγκεκριμένα για την αξιολόγηση των μοντέλων μας.

Έγινε επισκόπηση του διαγωνισμού M5, από τον οποίο αντλήθηκαν τα δεδομένα που χρησιμοποιήθηκαν και δόθηκε έμφαση στα χαρακτηριστικά του που παρουσίασαν το μεγαλύτερο ενδιαφέρον. Αναλύθηκαν εις βάθος τα μοντέλα πρόβλεψης που επιλέχθηκαν καθώς και οι παράμετροι τους. Τέλος, παρουσιάστηκαν τα αποτελέσματα όλων αυτών των μοντέλων, τα οποία κατέδειξαν ποιες μεταβλητές τελικά είναι ωφέλιμες και ποιες προσθέτουν περιττή πληροφορία.

5.1 Συμπεράσματα

Από τις μετρήσεις της απόδοσης των προβλέψεων για το 10^{ο} επίπεδο ιεράρχησης του συνόλου δεδομένων, προέκυψαν ορισμένα διάφορα ευρήματα, τα οποία εναρμονίζονται με τον αρχικό στόχο που είχε η παρούσα διπλωματική εργασία. Αναλυτικότερα:

- Η χρήση μοντέλων μηχανικής μάθησης, τα οποία δεν υποθέτουν κάποια κανονική κατανομή αλλά καθορίζουν αυτή την κατανομή με βάση τα δεδομένα που τους παρουσιάζονται, παρουσιάζει πολύ καλύτερα αποτελέσματα σε σχέση με τις κλασικές μεθόδους πρόβλεψης και τις ήδη υπάρχουσες στατιστικές κατανομές που αυτά χρησιμοποιούν.
- Η χρήση ντετερμινιστικών μεταβλητών θεωρείται απαραίτητη για την καλύτερη παραμετροποίηση των μοντέλων πρόβλεψης. Μπορεί να προσθέτουν υπολογιστικό κόστος αλλά το τελικό αποτέλεσμα είναι σημαντικά καλύτερο. Κάποιες επεξηγηματικές μεταβλητές όμως, ίσως κρύβουν την παγίδα να μην βελτιώνουν πραγματικά το μοντέλο. Εκεί, πρέπει να γίνει η επιλογή για το ποιες μεταβλητές κάνουν πραγματική διαφορά στο τελικό μοντέλο και ποιες προσθέτουν περιττή πληροφορία ή και πληροφορία που τελικά μπερδεύει το μοντέλο μας αντί να το βοηθάει. Στην περίπτωσή μας, οι επεξηγηματικές μεταβλητές της ημέρας και του μήνα, βοήθησαν πραγματικά στον εντοπισμό εβδομαδιαίας και μηνιαίας εποχιακότητας. Η μεταβλητή του SNAP, παρείχε βελτίωση στο μοντέλο, κυρίως στην κατηγορία των Φαγητών που κυριαρχεί των άλλων, ενώ η μεταβλητή των ειδικών γεγονότων παρείχε πληροφορία που δεν μπορούσε να αξιοποιήσει σωστά το τελικό μοντέλο, επομένως δεν επιλέχθηκε.
- Ο αλγόριθμος του LightGBM και επομένως τα δέντρα αποφάσεων, στηριζόμενα στη μηχανική μάθηση, παρουσιάζουν μεγάλη βελτίωση σε σχέση με τις

κλασικές μεθόδους πρόβλεψης. Ακόμα και το ARIMA με εξωγενείς μεταβλητές, δεν μπορεί να συγκριθεί με το τελικό μοντέλο που επιλέξαμε. Επομένως, φαίνεται ξεκάθαρο το πόσο σημαντικά είναι τα μοντέλα μηχανικής μάθησης κυρίως στον τομέα της λιανικής που μελετήσαμε στο πλαίσιο της παρούσας εργασίας αλλά και στους περισσότερους τομείς της καθημερινότητάς μας. Μάλιστα, με την συνεχή βελτίωσή τους και τη μείωση του υπολογιστικού τους κόστους, θα φανούν ακόμα πιο χρήσιμα στο σύντομο μέλλον.

- Το υπολογιστικό κόστος του αλγορίθμου LightGBM δεν είναι ιδιαίτερα μεγάλο, αν λάβει κανείς υπόψη την βελτίωση που παρουσίασε. Με βάση τα μοντέλα που χρησιμοποιήθηκαν, το ARIMAX ήταν αυτό που απαίτησε τους περισσότερους πόρους για να παράγει αποτέλεσμα.
- Ο συνδυασμός μοντέλων του αλγορίθμου LightGBM στο κομμάτι της εκπαίδευσής τους και στη συνέχεια στην πρόβλεψή τους, που έγινε στο τέλος, δεν απέδωσε καλύτερο αποτέλεσμα, επομένως μπορούμε να συμπεράνουμε ότι αυξάνοντας την πολυπλοκότητα, δεν σημαίνει ότι θα έχουμε και υποχρεωτικά καλύτερο αποτέλεσμα. Από την άλλη, λαμβάνοντας υπόψη απλώς το μοντέλο LightGBM με το βέλτιστο αποτέλεσμα κάθε κατηγορίας ή υποκατηγορίας προϊόντων, μπορούμε να έχουμε τελικά ένα αρκετά αξιόλογο αποτέλεσμα.

5.2 Μελλοντικές Προεκτάσεις

Η παρούσα εργασία εξέτασε την συνεισφορά που μπορεί να έχουν ντετερμινιστικές μεταβλητές στην παραγωγή πιθανοτικών προβλέψεων για τον κλάδο της λιανικής. Μελλοντικά θα μπορούσε να εξεταστεί η συνεισφορά της πληροφορίας για την τιμή των προϊόντων κάθε εβδομάδα, η οποία ήταν διαθέσιμη στα πλαίσια του διαγωνισμού αλλά δεν χρησιμοποιήθηκε στα πλαίσια της παρούσας εργασίας. Ακόμα, θα μπορούσαν να αναζητηθούν και άλλες πληροφορίες για τις ημέρες πωλήσεων των προϊόντων, όπως τι καιρό είχε εκείνη την ημέρα.

Επίσης, θα ήταν χρήσιμη η μελέτη περισσότερων συνδυασμών μοντέλων. Θα μπορούσε να εξεταστεί η χρήση του αλγορίθμου LightGBM σε συνδυασμό με το μοντέλο ARIMAX, που όπως είδαμε, σε κάποιες περιπτώσεις παρουσίαζε βελτιωμένο αποτέλεσμα. Να γίνει έλεγχος αν η αλλαγή των παραμέτρων του LightGBM προκαλεί καλύτερα αποτελέσματα σε συγκεκριμένες χρονοσειρές και για αυτές να χρησιμοποιείται μία παραλλαγή του μοντέλου που χρησιμοποιήθηκε. Να εξεταστεί με την αλλαγή της ιεράρχησης που χρησιμοποιήθηκε, αν θα παραμείνει εξίσου καλή η αποτελεσματικότητα του αλγορίθμου ή τα μοντέλα SES και ARIMAX θα παρουσιάσουν σημαντική βελτίωση.

Φαίνεται ακόμα πως είναι αναγκαία η έρευνα περισσότερων συνόλων δεδομένων πέρα από τον συγκεκριμένο κλάδο. Ενδιαφέρον θα παρουσίαζε η εφαρμογή σε δεδομένα ενέργειας και πρόβλεψης των ημερήσιων ενεργειακών αναγκών, καθώς η αβεβαιότητα που παρουσιάζει ο συγκεκριμένος τομέας, καθιστά τις πιθανοτικές προβλέψεις που δεν ακολουθούν συγκεκριμένη κατανομή ιδιαίτερα χρήσιμες. Πέρα από αυτό, μπορούν να ληφθούν αρκετές επεξηγηματικές μεταβλητές υπόψιν, αυξάνοντας την πολυπλοκότητα του συγκεκριμένου προβλήματος και ενισχύει το πλεονέκτημα των μοντέλων μηχανικής μάθησης έναντι των πιο κλασικών μεθόδων.

Βιβλιογραφία

[1] S. Makridakis et al., “The M5 uncertainty competition: Results, findings and conclusions,” *International Journal of Forecasting*, Dec. 2021, doi: 10.1016/j.ijforecast.2021.10.009

[2] Y. In and J. Jung, "Simple averaging of direct and recursive forecasts via partial pooling using machine learning," *International Journal of Forecasting*, vol. 38, no. 4, pp. 1386-1399, Oct.–Dec. 2022, doi: j.ijforecast.2021.11.007

[3] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T. Liu, "LightGBM: A highly efficient gradient boosting decision tree," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst. (NIPS’17)*, Dec. 2017, pp. 3149-3157.

[4] S. Makridakis, E. Spiliotis, and V. Assimakopoulos, "M5 accuracy competition: Results, findings and conclusions," *International Journal of Forecasting*, vol. 38, no. 4, pp. 1346-1364, Oct.–Dec. 2022, doi: j.ijforecast.2021.11.013

[5] S. Makridakis, E. Spiliotis, and V. Assimakopoulos, "The M5 competition: Background, organization and implementation," *International Journal of Forecasting*, vol. 38, no. 4, pp. 1325-1336, Oct.–Dec. 2022, doi: j.ijforecast.2021.07.007

[6] G. E. P. Box and G. M. Jenkins, *Time Series Analysis: Forecasting and Control*, 3rd ed. Englewood Cliffs, NJ, USA: Prentice Hall, 1994. ISBN: 978-1-349-35027-8

[7] E. Spiliotis, "Decision Trees for Time-Series Forecasting," *Foresight: The International Journal of Applied Forecasting*, no. 64, pp. 30-44, Q1 2022.

[8] C. S. Bojer, "Understanding machine learning-based forecasting methods: A decomposition framework and research opportunities," *International Journal of Forecasting*, vol. 38, no. 4, pp. 1555-1561, Oct.–Dec. 2022, doi: j.ijforecast.2021.11.003

[9] T. Januschowski, Y. Wang, K. Torkkola, T. Erkkilä, H. Hasson, and J. Gasthaus, "Forecasting with trees," *International Journal of Forecasting*, vol. 38, no. 4, pp. 1473-1481, Oct.–Dec. 2022, doi: j.ijforecast.2021.10.004

[10] S. Makridakis, F. Petropoulos, and E. Spiliotis, "Introduction to the M5 forecasting competition," *International Journal of Forecasting*, vol. 38, no. 4, pp. 1279-1282, Oct.–Dec. 2022, doi: j.ijforecast.2022.04.005

[11] R. Hyndman, A. B. Koehler, J. K. Ord, and R. D. Snyder, *Forecasting with Exponential Smoothing: The State Space Approach*, 2008. ISBN: 978-3-540-71918-2, doi: 978-3-540-71918-2

[12] Z. Moslemi et al., "Comprehensive forecasting of California’s energy consumption: A multi-source and sectoral analysis using ARIMA and ARIMAX models," *World Journal of Advanced Research and Reviews*, vol. 22, no. 2, pp. 484–497, 2024, doi: 10.30574/wjarr.2024.22.2.1367

[13] J. G. De Gooijer and R. Hyndman, "25 years of time series forecasting," *International Journal of Forecasting*, vol. 22, no. 3, pp. 443-473, 2006, doi: j.ijforecast.2006.01.001

[14] A. Choudhury, A. Mondal, and S. Sarkar, "Searches for the BSM scenarios at the LHC using decision tree-based machine learning algorithms: A comparative study and review of Random Forest, Adaboost, XGboost, and LightGBM frameworks," *The European Physical Journal Special Topics*, vol. 233, no. 2, Sep. 2024, doi: 10.1140/epjs/s11734-024-01308-x

[15] Φ. Πετρόπουλος και Β. Ασημακόπουλος, *Επιχειρησιακές Προβλέψεις*, Εκδόσεις Συμμετρία, 2013. ISBN: 978-960-266-333-2

[16] Ε. Σπηλιώτης, *Σημειώσεις μαθήματος Τεχνικές Προβλέψεων, Ολοκληρωμένα Αυτοπαλινδρομικά Μοντέλα Κινητού Μέσου Όρου (ARIMA)*, 2020. [Online]. Available: https://www.fsu.gr/el/component/jdownloads/finish/6/1715

[17] J. Vermorel, "Pinball Loss Function," 2012. [Online]. Available: https://www.lokad.com/pinball-loss-function-definition [Accessed: 23-Jan-2025].

[18] X. Long et al., "Scalable probabilistic forecasting in retail with gradient boosting trees: A practitioner's approach," *International Journal of Production Economics*, vol. 279, Jan. 2025, 109449, doi: j.ijpe.2024.109449

[19] R. Fildes, S. Ma, and S. Kolassa, "Retail forecasting: Research and practice," *International Journal of Forecasting*, vol. 38, no. 4, pp. 1283-1318, Oct.–Dec. 2022, doi: j.ijforecast.2019.06.004

[20] I. Alon, M. Qi, and R. J. Sadowski, "Forecasting aggregate retail sales: A comparison of artificial neural networks and traditional methods," *Journal of Retailing and Consumer Services*, vol. 8, no. 3, pp. 147-156, May 2001, doi: S0969-6989(00)00011-4

[21] D. M. Bechter and J. L. Rutner, "Forecasting with statistical models and a case study of retail sales," *Economic Review*, vol. 63, pp. 3-11, 1978.

[22] M. D. Geurts and J. P. Kelly, "Forecasting retail sales using alternative models," *International Journal of Forecasting*, vol. 2, no. 3, pp. 261-272, 1986, doi: 0169-2070(86)90046-4