ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ Μ/Υ
ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ
ΣΧΟΛΗ ΝΑΥΤΙΛΙΑΣ ΚΑΙ ΒΙΟΜΗΧΑΝΙΑΣ
ΤΜΗΜΑΤΟΣ ΒΙΟΜΗΧΑΝΙΚΗΣ ΔΙΟΙΚΗΣΗΣ & ΤΕΧΝΟΛΟΓΙΑΣ
ΔΙΑΠΑΝΕΠΙΣΤΗΜΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
«ΤΕΧΝΟ-ΟΙΚΟΝΟΜΙΚΑ ΣΥΣΤΗΜΑΤΑ»

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΠΡΟΒΛΕΨΗ ΑΦΙΞΕΩΝ ΕΠΙΒΑΤΩΝ ΣΕ ΠΤΗΣΕΙΣ ΕΣΩΤΕΡΙΚΟΥ ΜΕ ΧΡΗΣΗ ΜΟΝΤΕΛΩΝ ΧΡΟΝΟΣΕΙΡΩΝ ΚΑΙ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ

ΟΛΓΑ ΚΡΙΝΤΑ

ΕΠΙΒΛΕΠΩΝ: ΒΑΣΙΛΕΙΟΣ ΑΣΗΜΑΚΟΠΟΥΛΟΣ
ΟΜΟΤΙΜΟΣ ΚΑΘΗΓΗΤΗΣ Ε.Μ.Π.

Αθήνα, Φεβρουάριος 2024

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ Μ/Υ
ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ
ΣΧΟΛΗ ΝΑΥΤΙΛΙΑΣ ΚΑΙ ΒΙΟΜΗΧΑΝΙΑΣ
ΤΜΗΜΑΤΟΣ ΒΙΟΜΗΧΑΝΙΚΗΣ ΔΙΟΙΚΗΣΗΣ & ΤΕΧΝΟΛΟΓΙΑΣ
ΔΙΑΠΑΝΕΠΙΣΤΗΜΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
«ΤΕΧΝΟ-ΟΙΚΟΝΟΜΙΚΑ ΣΥΣΤΗΜΑΤΑ»

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΠΡΟΒΛΕΨΗ ΑΦΙΞΕΩΝ ΕΠΙΒΑΤΩΝ ΣΕ ΠΤΗΣΕΙΣ ΕΣΩΤΕΡΙΚΟΥ ΜΕ ΧΡΗΣΗ ΜΟΝΤΕΛΩΝ ΧΡΟΝΟΣΕΙΡΩΝ ΚΑΙ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ

ΟΛΓΑ ΚΡΙΝΤΑ

ΕΠΙΒΛΕΠΩΝ: ΒΑΣΙΛΕΙΟΣ ΑΣΗΜΑΚΟΠΟΥΛΟΣ
ΟΜΟΤΙΜΟΣ ΚΑΘΗΓΗΤΗΣ Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 14η Φεβρουαρίου 2024.

.....	.....	.....
Βασίλειος Ασημακόπουλος	Δημήτριος Ασκούνης	Ευάγγελος Μαρινάκης
Ομότιμος Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Επίκουρος Καθηγητής Ε.Μ.Π.

Αθήνα, Φεβρουάριος 2024

(Υπογραφή)

……
Όλγα Κρίντα
Διπλωματούχος στο Διαπανεπιστημιακό πρόγραμμα μεταπτυχιακών σπουδών «Τεχνοοικονομικά Συστήματα»

Περίληψη

Η πρόβλεψη της ροής των επιβατών στα αεροδρόμια είναι σημαντική για την προετοιμασία των υποδομών υποδοχής τους στους αερολιμένες, τις επιχειρήσεις εστίασης και παροχής καταλυμάτων αλλά και τον κρατικό μηχανισμό εν γένει, και κατ' επέκταση για την εκτίμηση των εσόδων που προέρχονται από τουρίστες και επισκέπτες ανά περιοχή. Σε αυτό το πλαίσιο, στην παρούσα εργασία μελετήθηκαν προσεγγίσεις για την ακριβή πρόβλεψη του αριθμού επιβατών σε πτήσεις εσωτερικού χρησιμοποιώντας δεδομένα από 20 ενδεικτικά αεροδρόμια των ΗΠΑ.

Σε επίπεδο μεθοδολογίας, αρχικά συγκεντρώθηκαν ιστορικά δεδομένα για τις πτήσεις εσωτερικού στις ΗΠΑ, οι οποίες αφορούσαν μηνιαίες εγγραφές για τα έτη 1990-2009. Στη συνέχεια, τα δεδομένα επεξεργάστηκαν με τη χρήση εργαλείων ανοιχτού λογισμικού ώστε να δημιουργηθούν οι αντίστοιχες χρονοσειρές. Κατόπιν, από το σύνολο των επιλεγμένων δεδομένων, τα έτη 2004-2007 χρησιμοποιήθηκαν ως σύνολο εκπαίδευσης, το έτος 2008 ως σύνολο αξιολόγησης, ενώ το έτος 2009 ως σύνολο ελέγχου. Επόμενο βήμα ήταν ο σχεδιασμός και η εφαρμογή διαφορετικών μοντέλων πρόβλεψης που βασίζονταν στους αλγορίθμους της γραμμικής παλινδρόμησης (Linear Regression), των τυχαίων δασών (Random Forests) και των δέντρων σταδιακής ενίσχυσης (Gradient Boosting Trees). Η ακρίβεια των προβλέψεων των μοντέλων αξιολογήθηκε μέσω του συμμετρικού μέσου απόλυτου ποσοστιαίου σφάλματος.

Τα αποτελέσματα της μελέτης έδειξαν ότι το σφάλμα πρόβλεψης κυμαίνεται κατά μέσο όρο στο 5%, γεγονός που αποδεικνύει ότι η πρόβλεψη αφίξεων τουριστών είναι εφικτή με υψηλή ακρίβεια. Επίσης συνιστούν ότι και οι τρεις μέθοδοι πρόβλεψης που εξετάστηκαν, παράγουν παρόμοια αποτελέσματα, γεγονός που δεν παροτρύνει αναγκαστικά τη χρήση αλγορίθμων μηχανικής μάθησης έναντι κλασικών στατιστικών προσεγγίσεων. Τέλος, έδειξαν ότι τα παραπάνω ευρήματα γενικεύονται για όλα τα αεροδρόμια που μελετήθηκαν, με τις μεταβλητές που αφορούν την εποχικότητα, την υστέρηση κατά ένα έτος και το πλήθος των θέσεων να είναι από τις πλέον κρίσιμες για τη βελτίωση της προβλεπτικής ακρίβειας των μοντέλων.

Η συγκεκριμένη εργασία μπορεί να χρησιμοποιηθεί ως εφαλτήριο για τη περαιτέρω βελτίωση των προβλέψεων αφίξεων σε αεροδρόμια με τη χρήση π.χ. επιπλέον εξωγενών μεταβλητών ή πιο εξελιγμένων αλγορίθμων μηχανικής μάθησης.

Λέξεις κλειδιά: αεροδρόμιο άφιξης, πρόβλεψη αριθμού επιβατών, γραμμική παλινδρόμηση, τυχαία δάση, δέντρα σταδιακής ενίσχυσης

Abstract

The prediction of passenger flow at airports is important for the preparation of their reception infrastructure at airports, catering and accommodation businesses, as well as the overall government mechanism. Additionally, it is crucial for estimating the revenues derived from tourists and visitors in each region. In this context, this study examined approaches for the accurate prediction of the number of passengers on domestic flights using data from 20 representative airports in the United States.

In terms of methodology, historical data for domestic flights in the U.S. were initially collected, covering monthly records for the years 1990-2009. Subsequently, the data was processed using open-source tools to create corresponding time series. Then, from the selected dataset, the years 2004-2007 were used as a training set, the year 2008 as a validation set, and the year 2009 as a test set. The next step involved designing and implementing different prediction models based on linear regression, random forests, and gradient boosting trees algorithms. The accuracy of the forecasts was evaluated using the symmetric mean absolute percentage error.

The results of the study indicate that the prediction error averages around 5%, demonstrating that forecasting tourist arrivals is feasible with high accuracy. Furthermore, all three forecasting methods produced similar results, suggesting that there is not necessarily a preference for the use of machine learning algorithms over classical statistical approaches. Additionally, they demonstrated that the above findings generalize to all airports studied, considering variables related to seasonality, a one-year lag, and the number of available seats to be among the most critical for improving the forecasting accuracy of the models.

This specific study can serve as a starting point for further improving accuracy in airport arrival applications by using additional exogenous variables or more advanced machine learning algorithms, among others.

Keywords: arrival airport, prediction of number of passengers, linear regression, random forests, gradient boosting trees

Ευχαριστίες

Θα ήθελα να ευχαριστήσω τον Ομότιμο Καθηγητή κ. Βασίλειο Ασημακόπουλο για την ευκαιρία που μου έδωσε να ασχοληθώ με τον συναρπαστικό κόσμο της ανάλυσης των δεδομένων και των προβλέψεων.

Επίσης θα ήθελα να ευχαριστήσω ιδιαίτερα τον Μεταδιδακτορικό Ερευνητή κ. Ευάγγελο Σπηλιώτη, ο οποίος με καθοδήγησε καθ' όλη τη διάρκεια της εκπόνησης της διπλωματικής μου εργασίας, με τις πολύτιμες συμβουλές του και τις παρατηρήσεις του, που με βοήθησαν να την ολοκληρώσω με επιτυχία.

Τέλος θα ήθελα να ευχαριστήσω την οικογένεια μου και τους φίλους που με στηρίζουν σε κάθε νέο μου πόνημα.

Πίνακας περιεχομένων

Ευρετήριο Πινάκων

Ευρετήριο Διαγραμμάτων

Συντομογραφίες

ARIMAX: AutoRegressive Integrated Moving Average with eXogenous variables

GBT: Gradient Boosting Trees

IATA: International Air Transport Association

LR: Linear Regression

MAE: Mean Absolute Error

MASE: Mean Absolute Scaled Error

ML: Machine Learning

MOOC: Massive Open Online Course

NN: Neural Networks

RF: Random Forests

RMSE: Root Mean Square Error

SARIMA: Seasonal AutoRegressive Integrated Moving Average

sMAPE: Symmetrical Mean Absolute Percentage Error

SQL: Structured Query Language

SVM: Support Vector Machines

SVR: Support Vector-machine Regression

TSR: Temporal Super Resolution

Κεφάλαιο 1. Εισαγωγή

Στο παρόν κεφάλαιο παρουσιάζεται αρχικά το αντικείμενο της εργασίας, το οποίο αφορά την πρόβλεψη των αφίξεων επιβατών σε πτήσεις εσωτερικού, χρησιμοποιώντας ως μελέτη περίπτωσης ενδεικτικά αεροδρόμια των ΗΠΑ. Στη συνέχεια γίνεται αναφορά σε πρόσφατες εργασίες που σχετίζονται με τις προβλέψεις στον τομέα των εναέριων μετακινήσεων και τα αποτελέσματα αυτών. Τέλος, παρουσιάζεται ο τρόπος οργάνωσης της διπλωματικής εργασίας, δηλαδή η δομή και περιληπτικά το περιεχόμενο του κάθε κεφαλαίου.

1.1 Αντικείμενο Εργασίας

Η πρόβλεψη, ως αντικείμενο, εκτός του ότι συναρπάζει γιατί δίνει τη δυνατότητα στην παροχή απαντήσεων για μελλοντικά γεγονότα, είναι αναπόσπαστο κομμάτι της καθημερινότητας καθώς εξυπηρετεί από τον πιο απλό πολίτη (π.χ. πρόβλεψη καιρού), το κράτος (π.χ. πρόβλεψη εσόδων), τους ερευνητές (π.χ. πρόβλεψη ασθενειών) μέχρι και τους μεγαλύτερους επιχειρηματικούς κολοσσούς (π.χ πρόβλεψη πωλήσεων).

Η αύξηση των υπολογιστικών δυνατοτήτων στην εποχή μας δίνει τη δυνατότητα στην υλοποίηση αλγορίθμων πρόβλεψης, που ήταν αδύνατον να υλοποιηθούν στο παρελθόν λόγω τεχνολογικών περιορισμών, καθώς και την ανάπτυξη εξελιγμένων τεχνικών πρόβλεψης χρησιμοποιώντας εκτός από τις κλασσικές μεθόδους, μοντέλα μηχανικής και βαθιάς μάθησης.

Σήμερα, περισσότερο από ποτέ άλλοτε, υπάρχει επιτακτική ανάγκη να επιτευχθεί η πρόβλεψη με όσο το δυνατόν μεγαλύτερη ακρίβεια, κυρίως για να βοηθήσει στη λήψη αποφάσεων σε κρίσιμους τομείς όπως είναι η υγεία, η οικονομία, η παιδεία κ.ά. Σημαντικό μέρος της οικονομίας αποτελεί και ο τουρισμός που απασχολεί τον επιχειρηματικό κόσμο αλλά και το κάθε κράτος. Είναι σημαντικό να μπορεί να προβλεφθεί ο αριθμός των ανθρώπων που αφικνούνται σε μία πόλη ή/και χώρα, διότι σε όλες τις χώρες του κόσμου ο τουρισμός αποτελεί πηγή σημαντικών εσόδων για μία οικονομία.

Η μετακίνηση των ανθρώπων μπορεί να γίνει με τη χρήση επίγειων, ενάλιων ή εναέριων μέσων. Σημαντικό μέρος των τουριστών φθάνει σε έναν προορισμό μέσω των αεροδρομίων, ειδικά για τους ανθρώπους που βρίσκονται σε άλλη χώρα, οπότε είναι σημαντικό να μπορεί να γίνει πρόβλεψη, με όσο πιο ακριβή τρόπο, της άφιξης των επιβατών με εναέρια μέσα. Γι' αυτόν τον λόγο, το πρόβλημα που επιλέχθηκε να

αντιμετωπιστεί σε αυτή την εργασία είναι ο προσδιορισμός του μελλοντικού αριθμού των επιβατών που αφικνούνται σε ένα αεροδρόμιο. Η γνώση των προβλέψεων της μετακίνησης των ανθρώπων με τη χρήση εναέριων μέσων για τουρισμό ή για άλλους λόγους, αποτελεί έναν σημαντικό παράγοντα λήψης αποφάσεων των αεροπορικών εταιρειών, των αερολιμένων, των επιχειρηματιών που δραστηριοποιούνται στην εστίαση και στην παροχή καταλυμάτων και γενικότερα το κράτος.

Οι λόγοι που οι προβλέψεις των ροών των επιβατών αφορούν όλους του παραπάνω φορείς δεν είναι μόνο οικονομικοί, δηλαδή ο υπολογισμός των εσόδων από τις εν λόγω δραστηριότητες αλλά και ζητήματα που αφορούν την προετοιμασία τους στην υποδοχή των τουριστών.

Συγκεκριμένα οι αεροπορικές εταιρείες βασιζόμενες στις προβλέψεις των αφίξεων μπορούν να αυξήσουν τις ημερήσιες πτήσεις προς ένα συγκεκριμένο προορισμό που φαίνεται να αυξάνει την επισκεψιμότητα του και τη μείωση κάποιου άλλου στον οποίον συμβαίνει το αντίθετο.

Οι αερολιμένες μπορούν να χρησιμοποιήσουν τις προβλέψεις των αφίξεων για να διαχειριστούν καλύτερα τις ροές των επιβατών που φθάνουν στα αεροδρόμια και να ρυθμίσουν τη λειτουργία των αεροδιαδρόμων τους για να αποφεύγονται οι καθυστερήσεις και τα παράπονα.

Επίσης μπορούν να αυξήσουν τα καταστήματα που λειτουργούν και γενικότερα τις υπηρεσίες που παρέχονται εντός του αερολιμένα αλλά και να κάνουν προσλήψεις προσωπικού στον τομέα της ασφάλειας, των μεταφορών κ.τ.λ.

Οι επιχειρηματίες στους χώρους εστίασης π.χ. μπορούν να αποφασίσουν αν θα επεκταθούν λαμβάνοντας υπόψη τον αριθμό των ατόμων που καταφθάνουν στην πόλη τους. Επίσης οι εταιρείες ενδυμάτων, εστίασης κ.ά θα μπορούν με αρκετή βεβαιότητα να λάβουν απόφαση αν θα ανοίξουν κατάστημα σε κάποιο αεροδρόμιο.

Οι επιχειρηματίες στον χώρο των κατασκευών μπορούν να αποφασίσουν αν είναι βιώσιμη η επένδυση κατασκευής ξενοδοχείων ή άλλων καταλυμάτων κοντά στο αεροδρόμιο για την εξυπηρέτηση των διερχόμενων (transit) επιβατών.

Ένα κράτος μπορεί να λάβει μία απόφαση για τη δημιουργία περισσότερων του ενός αεροδρομίων σε μία πόλη. Να φροντίσει ώστε να υπάρχει συχνή και εύκολα προσβάσιμη διασύνδεση μεταξύ αεροδρομίου και των υπόλοιπων μέσων μαζικής μεταφοράς ώστε να εξυπηρετούν τη ροή των επιβατών. Επιπλέον αυτών, οι ιδιοκτήτες

ταξί ή οι εταιρείες μεταφορών μπορούν να αποφασίσουν αν είναι κερδοφόρο μία συγκεκριμένη περίοδο, μέρα ή/και ώρα να περιμένουν για πελάτες στο αεροδρόμιο.

Η λίστα των αποφάσεων που μπορούν να παρθούν μόνο και μόνο γνωρίζοντας τη ροή άφιξης των επιβατών είναι ανεξάντλητη και είναι ένα δείγμα του πόσο σημαντικό είναι να γίνεται πρόβλεψη στις εναέριες μεταφορές.

Με αφορμή τα παραπάνω, η συγκεκριμένη εργασία έχει ως αντικείμενο την πρόβλεψη του πλήθους των επιβατών σε αεροπορικές πτήσεις στις ΗΠΑ με τη χρήση μοντέλων χρονοσειρών και μηχανικής μάθησης. Στα πλαίσια αυτού του συγκεκριμένου σκοπού ακολουθείται η ενδεδειγμένη μεθοδολογία για την ανάλυση και επεξεργασία των χρονοσειρών, την ανάπτυξη των κατάλληλων προβλεπτικών αλγορίθμων (ορισμός μεταβλητών και παραμέτρων) και την εξαγωγή και ερμηνεία των αποτελεσμάτων.

Χρησιμοποιούνται λοιπόν τα ιστορικά δεδομένα από τις εσωτερικές πτήσεις με προορισμό είκοσι συγκεκριμένα αεροδρόμια των ΗΠΑ και επιλέγοντας τρία μοντέλα, τον αλγόριθμο της πολλαπλής γραμμικής παλινδρόμησης (Linear Regression – LR) και τους αλγορίθμους των τυχαίων δασών (Random Forest – RF) και των δέντρων σταδιακής ενίσχυσης (Gradient Boosting Trees – GBT), που ανήκουν στους αλγορίθμους δέντρων απόφασης (Decision Trees - DT), γίνεται πρόβλεψη των αφίξεων στα επιλεγμένα αεροδρόμια. Βραχυπρόθεσμος στόχος είναι η επιλογή αυτού του μοντέλου, με τη χρήση του κατάλληλου συνδυασμού των μεταβλητών, που δίνει την ακριβέστερη πρόβλεψη. Μακροπρόθεσμος στόχος είναι η δημιουργία μοντέλων προβλέψεων, προσθέτοντας ενδεχομένως επιπλέον παραμέτρους, που θα μπορούσαν να γενικευθούν και να χρησιμοποιηθούν για την πρόβλεψη αφίξεων επιβατών σε οποιοδήποτε αεροδρόμιο του κόσμου.

1.2 Σχετικές Μελέτες

Οι προβλέψεις στον τομέα του τουρισμού κεντρίζει συνεχώς το ενδιαφέρον των ερευνητών, οι οποίοι ανάλογα τα σύνολα ιστορικών δεδομένων στα οποία μπορούν να έχουν πρόσβαση προσπαθούν να προβλέψουν την τουριστική ζήτηση εξετάζοντας τις αφίξεις, τον αριθμό κλινών και διανυκτερεύσεων, επισκεπτών κ.τ.λ. Ιδιαίτερα σε συνάρτηση με τις μεταφορές και δη τις εναέριες, υπάρχει έντονο ερευνητικό ενδιαφέρον το οποίο μπορεί να διαπιστωθεί από την ενδεικτική παρουσίαση, των σκοπών και των αποτελεσμάτων, επιλεγμένων συναφών μελετών παρακάτω:

- Το 2014 στο άρθρο *Forecasting of Hong Kong airport's passenger throughput* (Tsui, Balli, Gilbey, & Gow, 2014) παρουσιάστηκε μία μελέτη που χρησιμοποίησε τα μοντέλα Box–Jenkins Seasonal ARIMA (SARIMA) και ARIMAX για την πρόβλεψη της κίνησης επιβατών στον Διεθνή Αερολιμένα του Χονγκ Κονγκ. Σκοπός ήταν η χρήση της πρόβλεψης για τον σχεδιασμό και τη λήψη αποφάσεων όσον αφορά τις εγκαταστάσεις και τα δίκτυα πτήσεων. Το αποτέλεσμα έδειξε σταθερή άνοδο της κίνησης έως το 2015.

- Σε άρθρο του 2018 στο IEEE με τίτλο *Hybrid forecasting model to predict air passenger and cargo in Indonesia* (Sulistyowati, Kuswanto, & Astuti, 2018), παρουσιάζεται μία έρευνα που αφορά τις προβλέψεις επιβατών και φορτίων σε τρεις πολυσύχναστους αερολιμένες της Ινδονησίας (Soekarno Hatta, I Gusti Ngurah Rai, Juanda). Με τη συνδυαστική χρήση γραμμικών και μη γραμμικών μοντέλων επιτυγχάνονται προβλέψεις με μεγάλη ακρίβεια βάσει του κριτηρίου του Μέσου Απόλυτου Ποσοστιαίου Σφάλματος (MAPE). Συγκεκριμένα τα υβριδικά μοντέλα ARIMAX-NN και TSR-NN προσφέρουν πιο ακριβείς προβλέψεις σε σχέση με τα υβριδικά TSR-SVR και ARIMAX-SVR.

- Με το άρθρο *Machine learning approach to predict aircraft boarding* (Schultz & Reitmann, 2019), παρουσιάζεται ένα εργαλείο πρόβλεψης της επιβίβασης με τη χρήση Νευρωνικών Δικτύων για την αξιολόγηση και τη βελτίωση της αποτελεσματικότητας της επιβίβασης στις αερομεταφορές. Χρησιμοποιείται ένα μοντέλο Long Short-Term Memory (LSTM), εκπαιδευμένο σε ένα περιβάλλον προσομοίωσης επιβίβασης στο οποίο με τη συμπερίληψη των αλληλοεπιδράσεων μεταξύ των επιβατών αυξάνεται η προβλεπτική ικανότητα του μοντέλου.

- Το άρθρο Deep learning models for forecasting aviation demand time series (Kanavos, Kounelis, Iliadis, & Makris, 2021), το οποίο εκδόθηκε το 2021, αναφέρει ότι η δυναμική ζήτηση των επιβατών έχει πολύ σημαντικές επιπτώσεις στην διαχείριση και λειτουργία στο σύνολο της αεροπορικής βιομηχανίας. Συγκεκριμένα χρησιμοποιώντας κλασσικές μεθόδους ARIMA, SARIMA αλλά και μηχανικής μάθησης όπως βαθιάς μάθησης Νευρωνικών Δικτύων ανέπτυξαν μοντέλα εκτίμησης και πρόβλεψης της ζήτησης αεροπορικών ταξιδιών.

- Στο πρόσφατο άρθρο *Forecasting airport transfer passenger flow using real-time data and machine learning* (Guo, Grushka-Cockayne, & De Reyck, 2022), παρουσιάζεται μία μελέτη που αφορά έναν προβλεπτικό μηχανισμό, βασισμένο στη μηχανική μάθηση, για να προβλέψει τη ροή επιβατών σε πραγματικό χρόνο. Η πρόβλεψη του χρόνου σύνδεσης και τη ροής των επιβατών επιτρέπει τη βελτιστοποίηση της διαδικασίας με αποτέλεσμα την ικανοποίηση των επιβατών και τη μείωση του κόστους για τον αερολιμένα.

Υπάρχουν πληθώρα άρθρων που ασχολούνται με τις εναέριες μεταφορές για διαφορετικούς σκοπούς, είτε αυτοί αφορούν τον τουρισμό είτε την οργάνωση της λειτουργίας και του κόστους ενός αερολιμένα. Στη δική μας μελέτη, όπως θα φανεί και παρακάτω, σκοπός είναι να επιλεγεί η ακριβέστερη προβλεπτική μέθοδος άφιξης επιβατών για να γενικευθεί και να εφαρμοστεί σε όσους περισσότερους δυνατούς τομείς.

1.3 Δομή Εργασίας

Η διπλωματική εργασία χωρίζεται σε πέντε μέρη. Στόχος είναι, ξεκινώντας από το θεωρητικό υπόβαθρο, να παρουσιαστεί ο τρόπος ανάλυσης και επεξεργασίας των δεδομένων, στη συνέχεια η πειραματική διαδικασία και τελικά, παίρνοντας σκυτάλη από τα αποτελέσματα, να εξαχθούν συμπεράσματα και να προταθούν τρόποι επέκτασης της μελέτης αυτής.

Στο δεύτερο κεφάλαιο γίνεται μία εισαγωγική παρουσίαση της σημαντικότερης θεωρίας που αφορά στις προβλέψεις, τις χρονοσειρές, τη μηχανική μάθηση, τα μοντέλα προβλέψεων, τα κριτήρια μέτρησης της ακρίβειας των προβλέψεων και ενδεικτικά παραδείγματα προβλέψεων με τη χρήση χρονοσειρών. Έμφαση δίνεται στην ανάλυση των μεθόδων που χρησιμοποιούνται στην πειραματική διαδικασία, ήτοι στις προβλέψεις με τη χρήση γραμμικής παλινδρόμησης και από τις μεθόδους επιβλεπόμενης μάθησης με τη χρήση δέντρων αποφάσεων, το τυχαίο δάσος (Random Forest) και τα δέντρα σταδιακής ενίσχυσης (Gradient Boosting Trees).

Στο τρίτο κεφάλαιο γίνεται μία αναλυτική περιγραφή σχετικά με το πώς έγινε η συλλογή των δεδομένων, η προετοιμασία των χρονοσειρών, δηλαδή έλεγχος εποχικότητας, τάσης, προτύπων, ακραίων ή μηδενικών τιμών κ.τ.λ., η απεικόνιση των δεδομένων, η προσαρμογή τους και τέλος η ανάλυση τους. Τέλος, παρουσιάζεται συνοπτικά πώς από την περιγραφόμενη διαδικασία προκύπτουν οι κατάλληλα

επεξεργασμένες χρονοσειρές από το σύνολο των ιστορικών δεδομένων, που χρησιμοποιούνται στην πειραματική διαδικασία μέσω των επιλεγμένων μοντέλων πρόβλεψης.

Στο τέταρτο κεφάλαιο παρουσιάζεται με λεπτομέρεια η πειραματική διαδικασία και τα ποσοτικά αποτελέσματα που προκύπτουν από αυτή τη διαδικασία, δηλαδή υπολογίζονται οι δείκτες ακρίβειας για την κάθε μία από τις εφαρμοζόμενες μεθόδους και συγκρίνονται μεταξύ τους με στόχο τον εντοπισμό της βέλτιστης προβλεπτικής μεθόδου.

Τέλος, στο πέμπτο κεφάλαιο εξάγονται τα συμπεράσματα από την πειραματική μελέτη και γίνονται διάφορες προτάσεις για προεκτάσεις πάνω στην συγκεκριμένη μελέτη.

Κεφάλαιο 2. Προβλέψεις

2.1 Εισαγωγή στις Προβλέψεις

Οι προβλέψεις με τη χρήση χρονοσειρών αποτελούν μια πρόκληση. Είναι σημαντικό χρησιμοποιώντας τα ιστορικά δεδομένα να προβλέπονται οι παρατηρήσεις στο μέλλον. Οι διαδικασίες και τα μοντέλα που αναπτύσσονται χρησιμοποιούν τις ιστορικές πληροφορίες για να μπορέσουν να προβλέψουν μελλοντικές συμπεριφορές με αναπόφευκτο αποτέλεσμα να υπάρχουν σφάλματα. Αυτό από μόνο του σημαίνει ότι μπαίνουν όρια και περιορισμοί στο πόσο αποδοτική μπορεί να είναι μια πρόβλεψη σε σχέση με ένα τυχαίο γεγονός που μπορεί να διαταράξει την ισορροπία αυτής της πρόβλεψης (Πετρόπουλος & Ασημακόπουλος, 2011) όπως π.χ. η επιδημία covid-19.

Πρέπει να γίνει κατανοητό ότι η πρόβλεψη δεν είναι συνώνυμη της προφητείας, υπόκειται σε κάποιους επιστημονικούς περιορισμούς, παρόλα αυτά διακρίνεται από κάποια σημαντικά πλεονεκτήματα όπως:

- Συμβολή στη λήψη αποφάσεων (π.χ. εταιρείες, κράτη κ.ά)
- Συμβολή στο σχεδιασμό
- Μεγάλη ακρίβεια με τη χρήση δεδομένων υψηλής ποιότητας

Και φυσικά κάποια μειονεκτήματα που συνδέονται με τα παρακάτω:

- Μη ακριβή και αναξιόπιστα δεδομένα
- Πολύ παλιά δεδομένα
- Μη προβλέψιμοι εξωγενείς παράγοντες (π.χ. οικονομική κρίση, covid-19 κ.ά.)

Πάραυτα με τα χρόνια παρατηρείται μία εξέλιξη στην πορεία των μεθόδων, ώστε οι προβλέψεις να γίνονται ακριβέστερες ακόμα και κάτω από ιδιαίτερες συνθήκες αβεβαιότητας.

Εξαιτίας αυτού και των μεγαλύτερων δυνατοτήτων των υπολογιστικών συστημάτων σε αποθήκευση, επεξεργασία κ.τ.λ. παρατηρείται μία αύξηση των τομέων στους οποίους χρησιμοποιούνται οι προβλέψεις.

Παρακάτω περιγράφουμε τις βασικές κατηγορίες στις οποίες διακρίνονται οι προβλέψεις εστιάζοντας στις ποσοτικές μεθόδους και συγκεκριμένα στα μοντέλα χρονοσειρών.

2.2 Κατηγορίες Μεθόδων Πρόβλεψης

Οι μέθοδοι τις οποίες χρησιμοποιούμε για να κάνουμε προβλέψεις ανήκουν σε δύο βασικές κατηγορίες (Πετρόπουλος & Ασημακόπουλος, 2011):

- Ποσοτικές. Οι μέθοδοι αυτές χρησιμοποιούν ιστορικά δεδομένα για την πρόβλεψη των μελλοντικών παρατηρήσεων. Οι μεταβλητές που χρησιμοποιούν είναι αριθμητικές, δηλαδή η πληροφορία είναι ποσοτικοποιημένη και υπάρχει ισχυρή πεποίθηση ότι τα πρότυπα που παρατηρούνται θα συνεχιστούν και μελλοντικά
- Κριτικές: Οι μέθοδοι αυτές βασίζονται περισσότερο σε ψυχολογικούς και γνωστικούς παράγοντες όπως είναι η γνώση, η εμπειρία, η διαίσθηση και η κριτική ικανότητα για την μελλοντική πορεία μιας κατάστασης

Οι ποσοτικές μέθοδοι όπως περιγράφηκαν περιληπτικά παραπάνω αναλύονται περαιτέρω σε δύο κατηγορίες:

- Μοντέλο Χρονοσειρών: το μοντέλο αυτό χρησιμοποιεί τα ιστορικά δεδομένα σαν είσοδο Χ και σαν έξοδο Υ είναι η πρόβλεψη που προκύπτει από τη συνάρτηση της με την είσοδο. Το μοτίβο συμπεριφοράς του μοντέλου αυτού παραμένει το ίδιο όταν οι εξωτερικές συνθήκες παραμένουν σταθερές. Τα X_1, X_2 κ.τ.λ. είναι τα ιστορικά δεδομένα τις αντίστοιχες i χρονικές περιόδους

$$Y = f(X1, X2, \dots, Xi)$$

- Αιτιοκρατικό μοντέλο: στο μοντέλο αυτό θεωρούμε ότι υπάρχει μία σχέση μεταξύ της εξαρτημένης μεταβλητής Υ και κάποιων από τις ανεξάρτητες μεταβλητών Xi.

$$Y = f(X1, X2, \dots, Xi)$$

Ο τύπος των δύο παραπάνω μοντέλων φαίνεται να είναι ο ίδιος, η διαφορά τους έγκειται στο γεγονός ότι στο μοντέλο των χρονοσειρών η συνάρτηση μεταξύ ιστορικών δεδομένων και πρόβλεψης είναι προκαθορισμένη ενώ στο αιτιοκρατικό μοντέλο εξετάζονται οι σχέσεις μεταξύ της εξαρτημένης και των ανεξάρτητων μεταβλητών. Δεν είναι προκαθορισμένη η συνάρτηση και μελετάται η μορφή της συνεξάρτησης.

Επειδή στην πειραματική μέθοδο γίνεται ανάλυση και χρήση χρονοσειρών για τις προβλέψεις, παρακάτω αναλύονται οι βασικές έννοιες, ιδιότητες, χρήσεις κ.τ.λ. των χρονοσειρών.

2.2.1 Χρονοσειρές

Οι χρονολογικές σειρές αφορούν αριθμητικά δεδομένα τα οποία συλλέγονται σε σταθερά χρονικά πλαίσια π.χ. ωριαία, ημερήσια, εβδομαδιαία, μηνιαία κ.ο.κ. και αφορούν διάφορες μεταβλητές (π.χ. καιρός) (Berenson, Levine, & Szabat, 2018). Οι χρονοσειρές απαρτίζονται συνεπώς από ιστορικά δεδομένα, που χρησιμοποιούνται για την πρόβλεψη της ακολουθίας των παρατηρήσεων στο μέλλον. Έτσι, αποτελούν σημαντικό και αναπόσπαστο κομμάτι των επιστημών της στατιστικής και της ανάλυσης δεδομένων (Box, Jenkins, & Reinsel, 1994).

2.2.2 Ποιοτικά Χαρακτηριστικά Χρονοσειρών

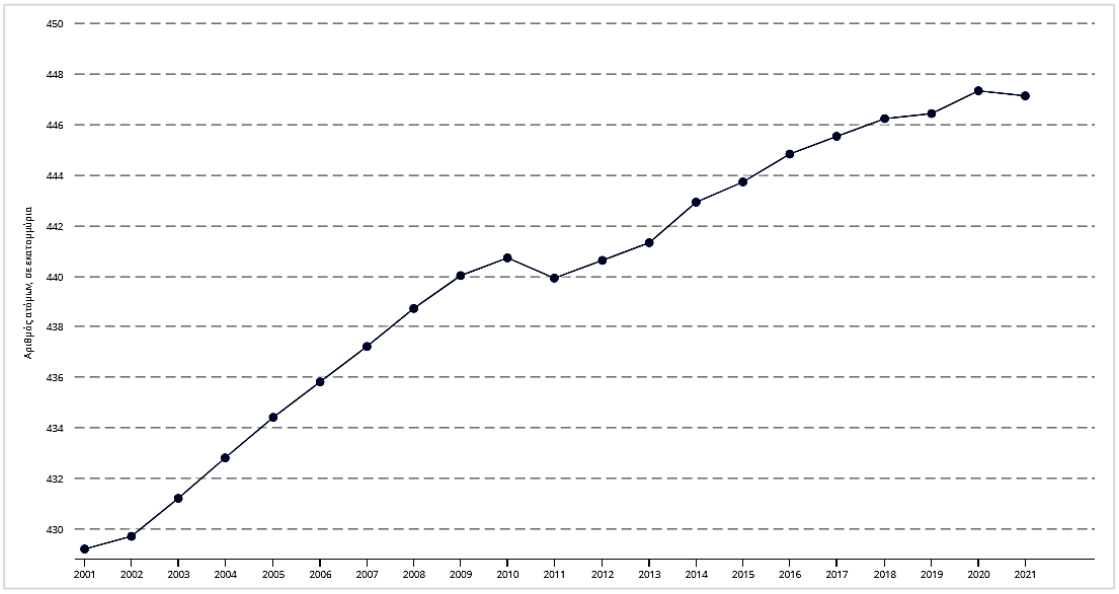
Οι χρονολογικές σειρές αναλύονται σε βασικά ποιοτικά χαρακτηριστικά, τα οποία ονομάζονται συνιστώσες. Οι τέσσερις βασικές συνιστώσες των χρονολογικών σειρών είναι:

- Η τάση
- Η κυκλική επίδραση
- Η εποχική επίδραση και
- Η ακανόνιστη τυχαία επίδραση

Τα ποιοτικά χαρακτηριστικά μιας χρονοσειράς μπορούν να εντοπιστούν μέσα από την παρατήρηση των γραφικών παραστάσεων των χρονοσειρών, για αυτόν τον λόγο στην ανάλυση κάθε ποιοτικού χαρακτηριστικού παρακάτω δίνεται και η απεικόνισή του μέσα από μία γραφική παράσταση. Επιπλέον η γραφική παράσταση λειτουργεί ως εργαλείο για τον εντοπισμό λανθασμένων και ακραίων τιμών (outliers) καθώς επίσης και για την διευκόλυνση του επιστήμονα/ερευνητή στην επιλογή της κατάλληλης μεθόδου πρόβλεψης.

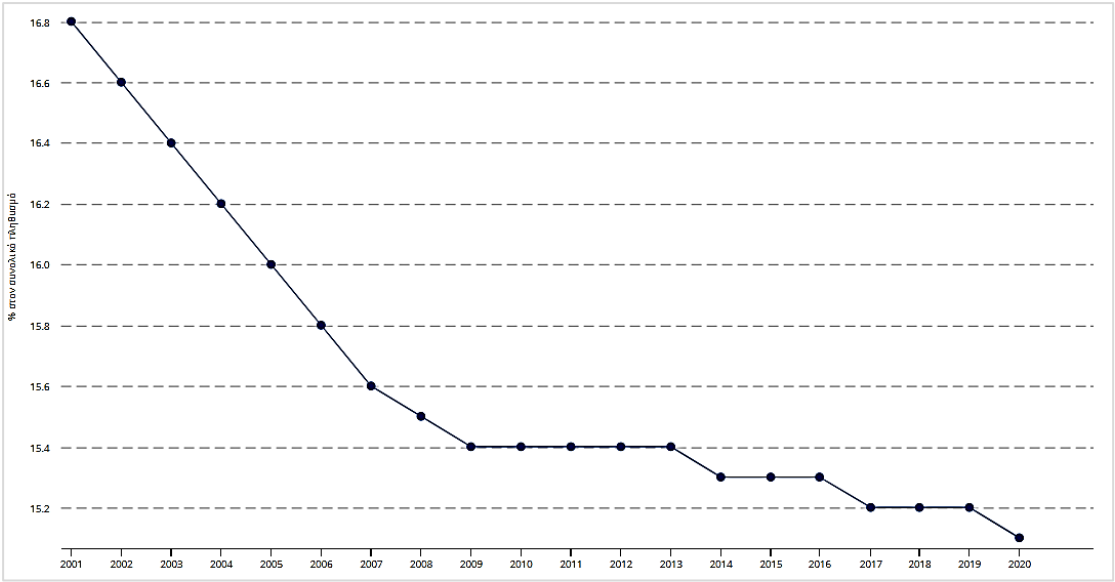
Τάση ονομάζεται το φαινόμενο κατά το οποίο με την πάροδο του χρόνου εμφανίζεται μια ανοδική ή καθοδική πορεία στο επίπεδο των χρονοσειρών. Είναι σημαντικό να ανιχνεύεται στα χρονολογικά δεδομένα ώστε να μπορούν να εξαχθούν συμπεράσματα για την μελλοντική πορεία των αντικειμένων που πραγματεύονται π.χ. οικονομία κ.ά.

Διάγραμμα 2.1 Γραμμική Τάση - Πληθυσμός της Ευρωπαϊκής Ένωσης το διάστημα 2001-2021 (σε εκατομμύρια)



Πηγή: ΕΛΣΤΑΤ - Συνολικός Πληθυσμός

Διάγραμμα 2.2 Μη γραμμική Τάση - Ποσοστό πληθυσμού κάτω των 15 ετών % στο σύνολο του πληθυσμού στην Ευρωπαϊκή Ένωση το διάστημα 2001-2020



Πηγή: ΕΛΣΤΑΤ - Ποσοστό πληθυσμού κάτω των 15 ετών

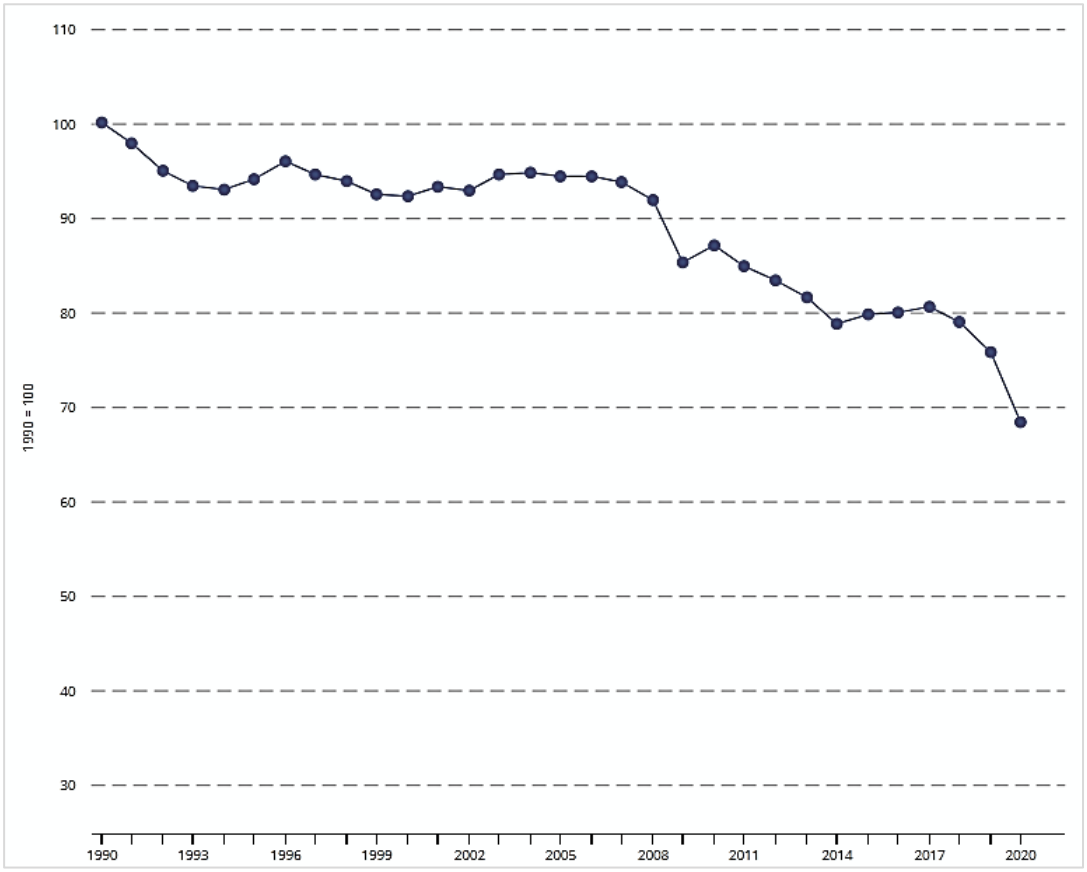
Η τάση στα δεδομένα μπορεί να έχει κυρίως δύο μορφές (Hyndman & Athanasopoulos, 2018):

- Γραμμική: οπότε στις περιόδους που εμφανίζεται παρουσιάζει σταθερή αύξηση ή μείωση και μπορεί να εντοπισθεί πάνω στο γράφημα ως γραμμή με συγκεκριμένη κλίση, αρνητική ή θετική (Διάγραμμα 2.1).

- Μη Γραμμική: Μερικές φορές η αύξηση ή η μείωση του επιπέδου της χρονοσειράς δεν ακολουθεί τη συμπεριφορά μιας ευθείας γραμμής με θετική ή αρνητική κλίση αλλά μπορεί να ακολουθεί για παράδειγμα εκθετική αύξηση ή μείωση (Διάγραμμα 2.2). Σε αυτήν την περίπτωση η τάση θα μπορούσε να αναλυθεί λογαριθμίζοντας τη συνάρτηση ώστε να οδηγηθούμε σε μία λογαριθμική γραμμική εξίσωση παλινδρόμησης, όπου και η εκτίμηση της σταθεράς και της κλίσης καθίσταται εφικτή.

Κυκλική επίδραση εμφανίζεται όταν υπάρχει μία κυκλική διακύμανση σε σχέση με τη συμπεριφορά των δεδομένων (βλέπε παρακάτω Διάγραμμα 2.3) και η διάρκεια τους είναι μεγαλύτερη του έτους (Berenson, Levine, & Szabat, 2018). Υφίσταται σε μεγάλο αριθμό κοινωνικο-οικονομικών αλλά και περιβαλλοντικών χρονοσειρών (π.χ. εκπομπές αερίων CO2 στην ατμόσφαιρα κ.ά.).

Διάγραμμα 2.3 Κυκλική επίδραση - Εκπομπές αερίων (φαινόμενο θερμοκηπίου) στην Ευρωπαϊκή Ένωση για το διάστημα 1990-2020 (δείκτης 1990=100)



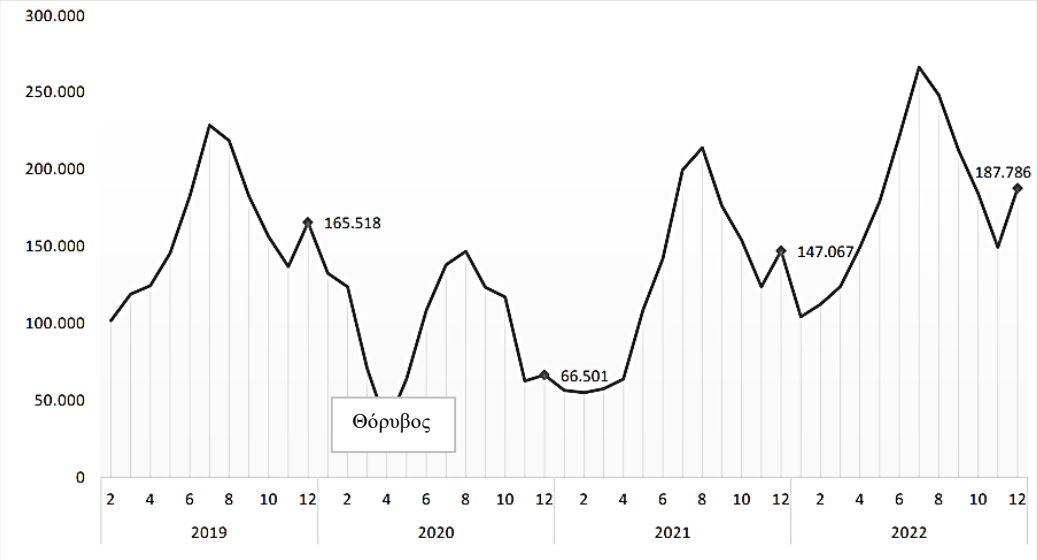
Πηγή: Eurostat - Εκπομπή Ρύπων

Ακανόνιστη ή τυχαία επίδραση είναι παράγοντες που μπορεί να συμβούν ξαφνικά, δεν ακολουθούν κάποιο συγκεκριμένο μοτίβο και δεν μπορεί εύκολα κάποιος να τους

προβλέψει (π.χ. πανδημία covid-19) οδηγώντας τις χρονοσειρές σε μη προβλέψιμες συμπεριφορές (Berenson, Levine, & Szabat, 2018). Σύμφωνα με τους (Hyndman & Athanasopoulos, 2018) η συνιστώσα αυτή ονομάζεται τυχαίος θόρυβος (Διάγραμμα 2.4).

Εποχικότητα ονομάζεται η διακύμανση των δεδομένων λόγω εποχιακών παραγόντων σε συγκεκριμένα διαστήματα μέσα στο έτος (π.χ. τουρισμός το καλοκαίρι). Η εποχική συνιστώσα αφορά μια κυκλική διακύμανση γύρω από την τάση μέσα όμως στο πλαίσιο ενός έτους (βλέπε Διάγραμμα 2.4). Αυτή συνδέεται με περιόδους εορτών, διακοπών και κλιματολογικών συνθηκών και αφορά μέρες, εβδομάδες, μήνες ή εποχές του έτους (Brockwell & Davis, 2016). Είναι σημαντικό να κατανοήσουμε την εποχικότητα για να μπορούμε να αναγνωρίσουμε συγκεκριμένες συμπεριφορές (μοτίβα) που επηρεάζουν τη χρονοσειρά και να τα αντιμετωπίσουμε ώστε να μπορέσουν να μελετήσουμε τα υπόλοιπα χαρακτηριστικά της χρονοσειράς (Hyndman & Athanasopoulos, 2018).

Διάγραμμα 2.4 Εποχικότητα και τυχαίος θόρυβος - Κύκλος εργασιών (σε χιλ. €) επιχειρήσεων στον κλάδο των Υπηρεσιών Εστίασης για το διάστημα 2019-2022



Πηγή: ΕΛΣΤΑΤ - Εξέλιξη κύκλου Εργασιών Επιχειρήσεων Παροχής Εστίασης

Στο παραπάνω σχήμα (Διάγραμμα 2.4) βλέπουμε την επίδραση του covid-19 στην εστίαση μεταξύ Φεβρουαρίου και Απριλίου του 2020 σε σύγκριση με την συμπεριφορά για το ίδιο διάστημα τα προηγούμενα έτη.

2.2.3 Βασικά Βήματα Διαδικασίας Πρόβλεψης

Τα βασικά βήματα της διαδικασίας μίας πρόβλεψης είναι τα πέντε παρακάτω (Makridakis, Wheelwright, & Hyndman, Forecasting: methods and applications, 1998):

- **Καθορισμός Προβλήματος**: είναι από τα βασικότερα σημεία κατά τη διαδικασία της πρόβλεψης καθώς πρέπει να καθοριστεί το τι θα προβλεφθεί, ποιο σκοπό επιτελεί αυτή η διαδικασία και ποιους θα ωφελήσει η συγκεκριμένη πρόβλεψη.
- **Συγκέντρωση Πληροφοριών**: η συγκέντρωση των δεδομένων αποτελεί μία επίσης ουσιώδη εργασία καθώς πρέπει να γίνει από άτομα τα οποία έχουν εμπειρία στις συγκεκριμένες κατηγορίες δεδομένων και γνωρίζουν πως να συντηρούν, επεξεργάζονται και να ενημερώνουν τα δεδομένα. Για να είναι τα δεδομένα αξιόπιστα θα πρέπει τα άτομα που αναλαμβάνουν τη συγκέντρωσή τους να έχουν εξειδικευμένες γνώσεις στα πλαίσια της εργασίας τους.
- **Προετοιμασία Χρονοσειρών**: η προετοιμασία των ιστορικών δεδομένων είναι μία χρονοβόρα διαδικασία που περιλαμβάνει αρκετές ενέργειες. Η πρώτη ματιά στα δεδομένα δίνει σε αδρές γραμμές τη συμπεριφορά τους, δηλαδή αν παρατηρούνται ακραίες τιμές, εποχικότητα, μοτίβα συμπεριφορών κ.τ.λ. Πιο φανερά γίνονται αυτά τα στοιχεία όταν γίνεται απεικόνιση των δεδομένων. Όταν για παράδειγμα εντοπιστούν κενές ή ακραίες τιμές στα δεδομένα τότε επιλέγεται η κατάλληλη μέθοδος για προσαρμογή των δεδομένων. Το επόμενο βήμα είναι η ανάλυση των δεδομένων που επίσης περιλαμβάνει μία σειρά ενεργειών όπως είναι η στατιστική ανάλυση, ο υπολογισμός στατιστικών δεικτών κ.ά. Η προετοιμασία των χρονοσειρών θα βοηθήσει τον ερευνητή στην επιλογή της σωστής μεθόδου πρόβλεψης.
- **Επιλογή και Προσαρμογή μοντέλου**: Σε αυτό το σημείο και εφόσον έχουν γίνει προσεκτικά και διεξοδικά όλα τα προηγούμενα επιλέγεται η μέθοδος πρόβλεψης που θεωρείται ότι θα δώσει με μεγαλύτερη ακρίβεια τις μελλοντικές παρατηρήσεις. Επιπλέον καθορίζονται οι μεταβλητές και οι παράμετροι που θα χρησιμοποιηθούν στο μοντέλο και θα παίξουν καθοριστικό ρόλο στην ακριβή πρόβλεψη.
- **Χρήση και Αποτίμηση του μοντέλου**: Τέλος γίνεται χρήση του μοντέλου που έχει επιλεχθεί και εξετάζονται οι προβλέψεις του. Το πόσο ικανοποιητικό είναι το μοντέλο είναι μία απάντηση που δίνεται συν τω χρόνω καθώς μπορεί να χρειαστεί να βελτιστοποιηθεί η διαδικασία μέσα από την επανάληψη για να είναι αποδοτικότερο το μοντέλο.

2.2.4 Πεδία και Εφαρμογές Προβλέψεων

Οι προβλέψεις πλέον έχουν εφαρμογή σε όλους τους τομείς της καθημερινότητας, πόσο μάλλον οι προβλέψεις με τη χρήση χρονοσειρών. Στη συνέχεια αναφέρουμε τους βασικούς τομείς στους οποίους χρησιμοποιούμε τις προβλέψεις και παραθέτουμε και αντίστοιχα αντιπροσωπευτικά παραδείγματα μέσα από τη βιβλιογραφία.

- **Υγεία**: η διαρκής εξέλιξη των μεθόδων πρόβλεψης αλλά και των τεχνολογικών εξελίξεων έχει συμβάλλει θετικά στο πολύ σημαντικό τομέα της υγείας. Τα τελευταία χρόνια χάρη στις προηγμένες μεθόδους πρόβλεψης έχει γίνει πιο εύκολη η διαχείριση ζητημάτων υγείας. Παράδειγμα, σχετικά πρόσφατο, αποτελεί η εργασία των (Flaxman, και συν., 2020) όπου εκτιμάται ο αντίκτυπος των μη φαρμακευτικών παρεμβάσεων στον Covid-19 στην Ευρώπη.
- **Οικονομία**: ένας άλλος τομέας που βασίζεται σήμερα σε μεγάλο βαθμό στις προβλέψεις είναι και η οικονομία. Σε κάθε πτυχή της οικονομίας μπορούν να χρησιμοποιηθούν οι προβλέψεις όπως στις μελλοντικές επενδύσεις των εταιρειών, στην αγορά μετοχών, στη λήψη αποφάσεων για τις κυβερνητικές και παγκόσμιες πολιτικές στον τομέα της οικονομίας κ.ά. Στην εργασία των (Faust & Wright, 2013) παρουσιάζεται μια από αυτές τις πτυχές που είναι η χρήση των προβλέψεων στον πληθωρισμό.
- **Εφοδιαστική Αλυσίδα**: τα τελευταία χρόνια γίνεται πολύς λόγος για τη μείωση του κόστους της διαχείρισης των αποθεμάτων αλλά και της διανομής των παραγγελιών, η λεγόμενη εφοδιαστική αλυσίδα (logistics). Και στον τομέα αυτόν χρησιμοποιούνται μοντέλα προβλέψεων για τη βελτιστοποίηση χρόνων παράδοσης, βέλτιστης τιμής ελάχιστου αποθέματος κ.ο.κ. Στο συγκεκριμένο παράδειγμα (Mahya & Fereshteh, 2020) χρησιμοποιείται μεγάλος όγκος δεδομένων για να προβλεφθεί η ζήτηση στην αλυσίδα εφοδιασμού.
- **Ενέργεια**: σήμερα περισσότερο από ποτέ άλλοτε είναι αναγκαία η σωστή διαχείριση της ενέργειας. Σε αυτό το δύσκολο έργο συμβάλλουν οι προβλέψεις με τη χρήση χρονολογικών σειρών. Χαρακτηριστικά αναφέρουμε την εργασία των (Deb, Zhang, Yang, Lee, & Shah, 2017) στην οποία χρησιμοποιούνται οι χρονοσειρές στα πλαίσια διαφόρων μεθόδων μηχανικής μάθησης για την πρόβλεψη της κατανάλωσης ενέργειας.

- **Καιρικές Συνθήκες**: με την πάροδο των χρόνων γινόμαστε μάρτυρες ακραίων φαινομένων ακόμα και στη χώρα μας. Είναι επιτακτική ανάγκη να μπορούμε να προβλέψουμε τον καιρό ώστε να προφυλάσσουμε το περιβάλλον και τις ανθρώπινες ζωές. Στη μελέτη (Jiang, Yang, & Heng, 2019) παρουσιάζεται ένα συνδυαστικό μοντέλο προβλέψεων της ταχύτητας του αέρα με τη χρήση χρονοσειρών.
- **Διαδίκτυο**: σήμερα διακινούνται τεράστιες ποσότητες δεδομένων μέσα από το διαδίκτυο. Οι μελλοντικές προβλέψεις θα μπορούσαν να αφορούν τον αριθμό των «πελατών» του διαδικτύου, την κίνηση ενός δικτύου κτλ. Η συγκεκριμένη πρόταση (Vinayakumar, Soman, & Poornachandran, 2017) αναφέρεται στην πρόβλεψη της κίνησης δικτύου με τη χρήση τη βαθιάς μάθησης.
- **Παιδεία/Εκπαίδευση**: όπως όλοι οι τομείς της ζωής αλλάζουν ραγδαία έτσι αλλάζει και η παιδεία και η εκπαίδευση. Στην εκπαίδευση θα μπορούσαν να χρησιμοποιηθούν οι προβλέψεις για να σχεδιαστούν προγράμματα σύμφωνα με τη μελλοντική ζήτηση. Ένα παράδειγμα χρήσης των προβλέψεων στην εκπαίδευση είναι η εργασία των (Yang, Brinton, Joe-Wong, & Chiang, 2017), η οποία μελετάει την πρόβλεψη των βαθμών στα MOOC’s σε συνάρτηση με την συμπεριφορά, με τη χρήση χρονοσειρών στα νευρωνικά δίκτυα.
- **Μετακίνηση/Μεταφορά**: όταν κάποιος μιλάει για μετακίνηση αναφέρεται σε όλες τις διαστάσεις αυτής (π.χ. τόπος, χώρος, χρόνος) και με όλες τις μορφές που μπορεί να υπάρξει, δηλαδή επίγεια, θαλάσσια και εναέρια. Γίνεται εστίαση στον τομέα αυτό μιας και το θέμα της συγκεκριμένης εργασίας αφορά προβλέψεις σε σχέση με την εναέρια μετακίνηση. Όσον αφορά στις εναέριες μετακινήσεις λοιπόν, υπάρχουν πολλές χρήσεις των προβλέψεων όπως η χρήση χρονοσειρών για την πρόβλεψη καθυστερήσεων των πτήσεων (Gui, Liu, Yang, Zhou, & Zhao, 2020), η συνδυαστική μελέτη διαφόρων μεθόδων για την πρόβλεψη στατιστικών δεικτών οι οποίοι μπορούν να χρησιμοποιηθούν στη διαχείριση της χωρητικότητάς και του σχεδιασμού στην βιομηχανία των εναέριων μετακινήσεων (Xu, Chan, & Zhang, 2019), η δημιουργία μεθόδων για την πρόβλεψη της ενεργής ζωής των κινητήρων των αεροπλάνων όπως απεικονίζεται στη μελέτη (Ordóñez, Sánchez Lasheras, Roca-Pardiñas, & Javier de Cos Juez, 2019) και πάρα πολλές άλλες χρήσεις των προβλέψεων στον τομέα των εναέριων μεταφορών/μετακινήσεων που μπορεί κάποιος να βρει

στην παγκόσμια βιβλιογραφία και σκοπός τους είναι να βοηθήσουν στη βελτίωση των εκάστοτε συνθηκών, δηλαδή στη δημιουργία μικρότερου αποτυπώματος στο περιβάλλον, στην καλύτερη εξυπηρέτηση του επιβατικού κοινού, στην ασφαλέστερη μετακίνηση και εν γένει σε όλες τις διαστάσεις της εναέριας κυκλοφορίας.

Είναι πάρα πολλοί οι τομείς στους οποίους οι προβλέψεις παίζουν καθοριστικό ρόλο, για την ακρίβεια οτιδήποτε μπορεί να φανταστεί ο ανθρώπινος νους. Παραπάνω γίνεται μια μικρή αναφορά σε κάποιες από τις βασικές συνιστώσες της ανθρώπινης ύπαρξης χωρίς να σημαίνει ότι η έρευνα περιορίζεται σε αυτές.

2.3 Μέθοδοι πρόβλεψης

Στο κεφάλαιο αυτό θα αναπτυχθούν θεωρητικά οι περισσότερο γνωστές μέθοδοι πρόβλεψης. Στόχος είναι, πρώτον να αναλυθούν οι διαφορετικές κατηγορίες μεθόδων πρόβλεψης και να γίνει μία προσπάθεια καλύτερου διαχωρισμού των μεθόδων, δεύτερον να γίνει κατανοητό και να συνδεθεί με την πράξη το θεωρητικό υπόβαθρο των τριών μεθόδων που χρησιμοποιούνται στην πειραματική μας διαδικασία, δηλαδή της γραμμικής παλινδρόμησης (Linear Regression - LR), των μεθόδων τυχαίου δάσους (Random Forest - RF) και των δέντρων σταδιακής ενίσχυσης (Gradient Boosting Trees - GBT) όσον αφορά τη μηχανική μάθηση.

Ο διαχωρισμός των μεθόδων σε στατιστικές και μηχανικής μάθησης (Machine Learning - ML) μερικές φορές είναι δυσδιάκριτος, οπότε ως ML λογίζονται οι μέθοδοι στις οποίες δεν υπάρχουν γραμμικές εξισώσεις, επίσης δεν είναι γνωστό πως παράγεται το αποτέλεσμα, δηλαδή δεν υπάρχει κάποια εξίσωση ή συγκεκριμένη δομή πίσω από αυτό (Petropoulos, και συν., 2022). Από την άλλη οι στατιστικές μέθοδοι έχουν εξισώσεις και δομημένη μορφή.

Σύμφωνα λοιπόν με τον πιο πάνω διαχωρισμό κατατάσσουμε τις μεθόδους πρόβλεψης ως εξής:

- **Στατιστικές**: σε αυτήν την κατηγορία εντάσσουμε μεθόδους πρόβλεψης οι οποίες χρησιμοποιούν τη διαθέσιμη ιστορική πληροφορία, την οποία εισάγουν σε ένα σύστημα εξισώσεων και μαθηματικών σχέσεων και ως αποτέλεσμα προκύπτει η πρόβλεψη της εξόδου όπως
    - Γραμμική Παλινδρόμηση

- Λογιστική Παλινδρόμηση

- Μηχανικής Μάθησης: στην περίπτωση αυτή δεν είναι γνωστός ο τρόπος που λειτουργούν τα συγκεκριμένα μοντέλα. Εισάγονται τα διάφορα δεδομένα ως είσοδος για να προβλεφθεί η έξοδος. Οι διαδικασίες που επιτελούνται από το μοντέλο της πρόβλεψης λειτουργούν ως «μαύρο κουτί». Αυτό που μας δίνει ένα μέτρο για το αν οι προβλέψεις είναι σωστές, είναι ο έλεγχος των εξόδων σε συνάρτηση με τι αναμενόμενες εξόδους. Ενδεικτικά μοντέλα ML είναι τα ακόλουθα.
    - Νευρωνικά Δίκτυα (Neural Network)
    - Τυχαίο Δάσος (Random Forest)
    - Υποστήριξη Μηχανών Διανυσμάτων (Support Vector Machines)
    - Δέντρα Παλινδρόμησης (Reggression Trees)
    - Δέντρα Σταδιακής Ενίσχυσης (Gradient Boosting Trees)

2.3.1 Μηχανική Μάθηση

Στην παράγραφο αυτή θα εξηγήσουμε τι σημαίνει μηχανική μάθηση, από ποιες υποκατηγορίες αποτελείται και σε ποιους τομείς της καθημερινής ζωής έχει εφαρμογή.

Η μηχανική μάθηση ως πεδίο ανήκει στην τεχνητή νοημοσύνη και σκοπός της είναι η ανάπτυξη συστημάτων που εκπαιδεύονται στο να παράγουν προβλέψεις που προκύπτουν από διαθέσιμα δεδομένα. Δεν ορίζουμε κάποιους κανόνες ή συγκεκριμένη λογική αλλά η παραγωγή των προβλέψεων προκύπτει από τα ίδια τα δεδομένα (Goodfellow, Bengio, & Courville, 2016). Επίσης, το μοντέλο μαθαίνει από τις διαδικασίες που ακολουθεί και στην ουσία εκπαιδεύεται μόνο του και εκ νέου για να αυτό-βελτιωθεί.

Η χρήση των αλγορίθμων μηχανικής μάθησης σε εφαρμογές της τεχνητής νοημοσύνης έχει κυρίως τρεις μορφές:

- Περιγραφική (Descriptive): μας περιγράφει τι συμβαίνει
- Προτεινόμενη (Prescriptive): παράγει αλγορίθμους που μας προτείνουν τι να κάνουμε
- Προβλεπτική (Predictive): γίνεται πρόβλεψη του γεγονότος, του τι δηλαδή αναμένεται να συμβεί

Κάποιες ενδεικτικές εφαρμογές ανά τομέα της μηχανικής μάθησης παρουσιάζονται παρακάτω (Sarker, 2021):

- Προγνωστική ανάλυση και έξυπνη λήψη αποφάσεων
- Πληροφορίες για την ασφάλεια στον κυβερνοχώρο και τις απειλές
- Διαδίκτυο των πραγμάτων (IoT) και έξυπνες πόλεις
- Πρόβλεψη κίνησης και μεταφοράς
- Υγειονομική περίθαλψη και πανδημία COVID-19
- Ηλεκτρονικό εμπόριο και προτάσεις προϊόντων
- Επεξεργασία φυσικής γλώσσας και ανάλυση συναισθήματος
- Αναγνώριση εικόνας, ομιλίας και μοτίβων
- Βιώσιμη γεωργία
- Αναλύσεις συμπεριφοράς χρήστη και εφαρμογή smartphone με επίγνωση του περιβάλλοντος

2.3.1.1 Επιβλεπόμενη Μάθηση (supervised learning)

Η μηχανική μάθηση σε αυτή την περίπτωση χρησιμοποιεί δεδομένα εισόδου για τα οποία είναι γνωστές οι έξοδοι (labeled δεδομένα) με σκοπό να δημιουργήσει ένα μοντέλο και να το εκπαιδεύσει κατάλληλα ώστε όταν χρησιμοποιηθούν νέες τιμές εισόδου, να μπορούν να προβλεφθούν οι νέες άγνωστες τιμές εξόδου. Οι πιο γνωστές λειτουργίες της επιβλεπόμενης μάθησης είναι η ταξινόμηση, η οποία κάνει χρήση αριθμητικών ή κατηγορικών μεταβλητών και ταξινομεί/κατατάσσει τα δεδομένα, και η παλινδρόμηση, η οποία με τη χρήση διαφόρων μεθόδων μηχανικής μάθησης προβλέπει μία συνεχή αριθμητική μεταβλητή, η οποία βασίζεται στις τιμές μίας ή περισσότερων προβλεπτικών μεταβλητών (Sarker, 2021).

Από τους αλγορίθμους που θα απασχολήσουν στην παρούσα εργασία, σε αυτή την κατηγορία ανήκουν οι πιο κάτω μέθοδοι πρόβλεψης:

- Τυχαία Δάση
- Δέντρα Σταδιακής Ενίσχυσης
    - LightGBM (Light Gradient Boosting Machine)

- XGBoost (Extreme GB)
- Ada Boost (Adaptive Boosting)

2.4 Περιγραφή Μεθόδων Πρόβλεψης

Μετά από μία γενική περιγραφή των μεθόδων πρόβλεψης και την κατηγοριοποίηση αυτών, παρουσιάζουμε παρακάτω την κάθε μέθοδο ξεχωριστά εστιάζοντας κυρίως στη Γραμμική Παλινδρόμηση, στα Τυχαία Δάση και στα Δέντρα Σταδιακής Ενίσχυσης που χρησιμοποιούμε στο κεφάλαιο 4, όπου αναλύεται η πειραματική διαδικασία.

2.4.1 Γραμμική Παλινδρόμηση

Όταν μιλάμε για παλινδρόμηση μιλάμε για σχέσεις μεταξύ μεταβλητών. Πως δηλαδή μπορεί μία ή περισσότερες μεταβλητές να επηρεάσουν τη μεταβολή της τιμής μίας άλλης μεταβλητής. Οι πρώτες μεν λέγονται ανεξάρτητες μεταβλητές και η άλλη εξαρτημένη γιατί η τιμή της εξαρτάται από τη μεταβολή των υπολοίπων.

Οι σχέσεις αυτές μοντελοποιούνται μαθηματικά και ανάλογα το μαθηματικό μοντέλο που τις εκφράζει κατατάσσονται σε ορισμένες κατηγορίες.

Όταν η σχέση είναι μεταξύ της εξαρτημένης και μίας ανεξάρτητης μεταβλητής είναι γραμμική τότε μιλάμε για απλή γραμμική παλινδρόμηση. Αν η εξαρτημένη μεταβλητή είναι μία και οι ανεξάρτητες πολλές τότε μιλάμε για πολλαπλή γραμμική παλινδρόμηση (Berenson, Levine, & Szabat, 2018).

Ο πιο απλός τύπος γραμμικής παλινδρόμησης στην πρόβλεψη όπως φαίνεται παρακάτω συσχετίζει τις τιμές μίας εισόδου X με τις τιμές της εξόδου Y :

$$Y = b_0 + b_1\ X + e$$

Όταν ο αριθμός των ανεξάρτητων μεταβλητών που χρησιμοποιούνται στο μοντέλο για να προβλέψουν την έξοδο είναι δύο και περισσότερες, τότε η πολλαπλή γραμμική παλινδρόμηση έχει τον εξής τύπο:

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

Το e συμβολίζει το σφάλμα που προστίθεται στην κάθε παλινδρόμηση κατά την πρόβλεψη. Ακόμα και το τελειότερο μοντέλο είναι αδύνατον να προβλέψει ακριβώς τις τιμές της εξόδου, πάντα θα υπάρχει ένα σφάλμα λόγω της μη χρήσης όλων των δυνατών ανεξάρτητων μεταβλητών στον υπολογισμό της εξαρτημένης αλλά και άλλων μη προβλέψιμων παραγόντων.

2.4.1.1 Συντελεστές Γραμμικής Παλινδρόμησης

Στην παράγραφο αυτή περιγράφονται οι συντελεστές της απλής γραμμικής παλινδρόμησης (Petropoulos, και συν., 2022). Για να εντοπιστεί η σχέση μεταξύ δύο μεταβλητών, δηλαδή το αν η τιμή της μίας μεταβλητής επηρεάζει την τιμή της άλλης χρησιμοποιούμε τον **συντελεστή συσχέτισης r**. Υπάρχουν δύο βασικές ερμηνείες για τον συγκεκριμένο συντελεστή:

- Προσδιορίζει την κατεύθυνση της σχέσης των μεταβλητών, δηλαδή θετική υποδεικνύοντας αύξηση της μίας όταν αυξάνεται η άλλη, αρνητική όταν συμβαίνει το αντίθετο ή ανεξαρτησία μεταξύ των μεταβλητών
- Προσδιορίζει το βαθμό συσχέτισης των μεταβλητών, δηλαδή ορίζει ως ισχυρότερη τη συσχέτιση όσο η τιμή του συντελεστή απομακρύνεται από το μηδέν

Ο συντελεστή συσχέτισης παίρνει τιμές στο διάστημα [-1,1] και δίνεται από τον παρακάτω τύπο:

$$r_{XY} = \frac{Cov_{XY}}{\sqrt{Cov_{YY}Cov_{XX}}} = \frac{Cov_{XY}}{S_Y S_X}$$

Όπου:

Συνδιακύμανση των X και Y: $Cov_{XY} = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{n}$

Διακύμανση του X: $Cov_{XX} = \frac{\sum(X_i - \bar{X})^2}{n} = Var_X = S_X^2$, όπου S η τυπική απόκλιση του X

Διακύμανση του Y: $Cov_{YY} = \frac{\sum(Y_i - \bar{Y})^2}{n} = Var_Y = S_Y^2$, όπου S η τυπική απόκλιση του Y

Ένας άλλος συντελεστής που χρησιμοποιείται κατά κόρον είναι ο **συντελεστής $\mathbf{R^2}$**, ο οποίος στην ουσία είναι ίσος με το τετράγωνο του συντελεστή συσχέτισης r και εκφράζει το ποσοστό της διακύμανσης της μεταβλητής μελέτης που ερμηνεύει η διακύμανση των ανεξάρτητων μεταβλητών. Ο συντελεστής αυτός είναι θετικός με τιμές μεταξύ 0 και 1. Όσο πιο κοντά είναι στην τιμή 1, τόσο καλύτερα ερμηνεύεται το ποσοστό της διακύμανσης της μεταβλητής Y από την ευθεία της γραμμικής παλινδρόμησης.

Σημειώνεται ότι όταν αναφερόμαστε σε πολλαπλή γραμμική παλινδρόμηση ο συντελεστής R^2 εκφράζει το ποσοστό της μεταβλητότητας της εξαρτημένης

μεταβλητής Y που εξηγείται από το σύνολο των ανεξάρτητων μεταβλητών και όχι από τη μία μεταβλητή X όπως ισχύει στην απλή γραμμική παλινδρόμηση (Berenson, Levine, & Szabat, 2018).

2.4.1.2 Στατιστικοί δείκτες

Οι στατιστικοί δείκτες χρησιμοποιούνται στη γραμμική παλινδρόμηση για να προσδιοριστούν κάποια βασικά στοιχεία της παλινδρόμησης που θα οδηγήσουν σε συμπεράσματα όσον αφορά την ακρίβεια και την αξιοπιστία του μοντέλου. Οι βασικοί δείκτες λοιπόν μας δίνουν εκτιμήσεις στα παρακάτω (Πετρόπουλος & Ασημακόπουλος, 2011):

- Της πιθανότητας να υπάρχει διαφορά μεταξύ των προβλεπόμενων και των μελλοντικών τιμών της εξαρτημένης μεταβλητής κατά μία συγκεκριμένη ποσότητα
- Της αξιοπιστίας όσον αφορά στον υπολογισμό της ευθείας της παλινδρόμησης
- Της ακρίβειας των συντελεστών b_0 και b_1

Στη συνέχεια αναλύουμε δύο συγκεκριμένους στατιστικούς δείκτες που αφορούν και στην απλή και στην πολλαπλή γραμμική παλινδρόμηση

Στατιστικός Δείκτης F: με το κριτήριο αυτό ελέγχουμε αν υπάρχει στατιστικά σημαντική σχέση μεταξύ του συνόλου των ανεξάρτητων μεταβλητών και της εξαρτημένης μεταβλητής. Ο τύπος υπολογισμού του F φαίνεται παρακάτω.

$$F = \frac{\frac{\sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2}{k}}{\frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{n-k-1}}$$

ή

$$F = \frac{\frac{R^2}{k}}{\frac{1-R^2}{n-k-1}}$$

Αν ο παρονομαστής είναι μεγάλος τότε σημαίνει ότι η διακύμανση των σφαλμάτων είναι μεγάλη με αποτέλεσμα το μοντέλο παλινδρόμησης να είναι αποτυχημένο. Αντίθετα όταν ο αριθμητής, δηλαδή η ερμηνευθείσα διακύμανση, είναι μεγαλύτερος

του παρονομαστή τότε ο F είναι μεγάλος με αποτέλεσμα ένα πιο επεξηγηματικό μοντέλο.

Στατιστικός δείκτης t: μετά τη χρήση του δείκτη F για την εξέταση της σημαντικότητας του μοντέλου στο σύνολο, πολλές φορές είναι απαραίτητο να εξεταστεί η σημαντικότητα καθενός από τους συντελεστές παλινδρόμησης. Μέτρο της σημαντικότητας ή μη των συντελεστών αυτών είναι ο δείκτης t.

Για κάθε έναν συντελεστή παλινδρόμησης b_k ορίζεται ένα τυπικό σφάλμα (Standard Error – SE) και έτσι υπολογίζεται ο δείκτης t, με προϋπόθεση την υπόθεση της κανονικότητας του μοντέλου παλινδρόμησης. Ο δείκτης t ακολουθεί την t-κατανομή με (n-k-1) βαθμούς ελευθερίας.

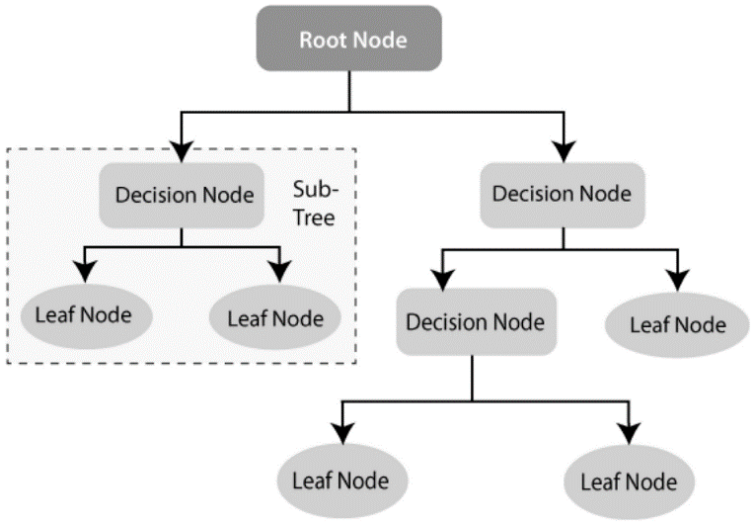
$$t_{b_j} = \frac{b_j}{SE_{b_j}}$$

2.4.2 Δέντρα Αποφάσεων

Τα δέντρα αποφάσεων παίρνουν το όνομά τους από τη μορφή τους, μορφή ανεστραμμένου δέντρου όπως φαίνεται στην παρακάτω εικόνα (Διάγραμμα 2.5).

Τα DT αποτελούν μοντέλα μηχανικής μάθησης βασισμένα σε κανόνες που μπορούν να χρησιμοποιηθούν για προβλήματα παλινδρόμησης και ταξινόμησης. Είναι πάρα πολύ δημοφιλή τα τελευταία χρόνια στον τομέα των προβλέψεων και ιδιαίτερα στην πρόβλεψη πωλήσεων.

Διάγραμμα 2.5 Παράδειγμα Δομής Δέντρων Αποφάσεων



Πηγή: (Sarker, 2021)

Τα δέντρα αποφάσεων για να προβλέψουν την εξαρτημένη μεταβλητή χρησιμοποιούν μεταβλητές, οι οποίες ονομάζονται επεξηγηματικές. Χρησιμοποιούνται συγκεκριμένοι κανόνες των οποίων ο τύπος και ο αριθμός ορίζεται από έναν αλγόριθμο για την ανάπτυξη του δέντρου. Αυτοί οι κανόνες υπόκεινται στους περιορισμούς που θέτει ο χρήστης μέσα από την εισαγωγή συγκεκριμένων δεδομένων εκπαίδευσης μέσω συγκεκριμένων μεταβλητών.

2.4.2.1 Ρίζα, Κλαδιά και Φύλλα των Δέντρων Απόφασης

Το δέντρο αποτελείται από τη ρίζα (root), τα κλαδιά (branches) και τα φύλλα (leaves). Το αρχικό σετ δεδομένων που βρίσκεται στη ρίζα διασπάται σε πρώτη φάση σε δύο επιμέρους υποσύνολα δεδομένων (κλαδιά). Ο διαχωρισμός αυτός γίνεται βασισμένος στη «βέλτιστη» μεταβλητή, αυτή που έχει δηλαδή την καλύτερη τιμή στο κριτήριο αξιολόγησης που έχει οριστεί. Η διάσπαση στα επιμέρους κλαδιά γίνεται με βάση τον κανόνα που δημιουργείται αυτόματα από το τον αλγόριθμο που χρησιμοποιεί το δέντρο. Σε κάθε κόμβο γίνεται έλεγχος ποια μεταβλητή ακολουθεί τον κανόνα. Τότε τα δεδομένα αυτής της μεταβλητής ανατίθενται στο ένα κλαδί και όλα τα υπόλοιπα στο άλλο. Η διαδικασία αυτή συνεχίζεται και τερματίζει βάσει της προκαθορισμένης συνθήκης τερματισμού που έχει οριστεί και καταλήγει στο φύλλο (leaf). Κάθε φύλλο αποτελεί και μια λύση (Spiliotis, 2022).

2.4.2.2 Υπερπαράμετροι

Για να δομηθεί ένα δέντρο χρησιμοποιούνται οι λεγόμενες υπερπαράμετροι (hyper-parameters). Οι οποίοι συνοπτικά παρουσιάζονται παρακάτω:

- Μέγιστος αριθμός φύλλων
- Το μέγιστο βάθος που είναι ο μέγιστος αριθμός των διασπάσεων που θα υποστεί το δέντρο από την ρίζα μέχρι τα φύλλα
- Ελάχιστος αριθμός παρατηρήσεων που χρειάζονται για να δημιουργηθεί ένα φύλλο
- Ελάχιστος αριθμός παρατηρήσεων για να προχωρήσει ο κόμβος σε διάσπαση

2.4.2.3 Δέντρα Παλινδρόμησης

Επειδή στην συγκεκριμένη εργασία γίνεται πρόβλεψη με τη χρήση χρονοσειρών θα γίνει μία συνοπτική αναφορά στη χρήση των δέντρων παλινδρόμησης και μόνο,

παραλείποντας τη χρήση τους στην ταξινόμηση. Επίσης είναι μια σημαντική μη γραμμική εναλλακτική έναντι της γραμμικής παλινδρόμησης για την πρόβλεψη χρονολογικών σειρών.

Τα δέντρα αυτά δεν χρησιμοποιούν κατηγορικές μεταβλητές αλλά αριθμητικές. Επίσης ο διαχωρισμός στον κόμβο γίνεται σύμφωνα με το ποιες παρατηρήσεις συμφωνούν περισσότερο με τη μεταβλητή στόχο, δηλαδή με την ακρίβεια των προβλεπόμενων δεδομένων βάσει του κριτηρίου αξιολόγησης που έχει τεθεί.

Στην ουσία η προσέγγιση που γίνεται από τα δέντρα παλινδρόμησης είναι τμηματική. Κάθε φορά ελέγχεται πόσο η επεξηγηματική μεταβλητή προσεγγίζει την τιμή στόχο (Spiliotis, 2022).

Σε αυτό το σημείο πρέπει να γίνει η εξής παρατήρηση, δέντρα που είναι πολύ «ρηχά» μπορεί να αποτύχουν να μοντελοποιήσουν καίριες σχέσεις δεδομένων ενώ δέντρα που είναι πολύ «βαθιά» μπορεί να υπερεκτιμούν τα δεδομένα προκαλώντας τυχαίες μοντελοποιήσεις. Όταν ο μέγιστος αριθμός φύλλων και το μέγιστο βάθος οριστούν στο άπειρο τότε μπορεί να γίνεται ελαχιστοποίηση του σφάλματος αλλά υπάρχει κίνδυνος υπερπροσαρμογής, άρα και αδύναμης απόδοσης στην πρόβλεψη.

2.4.2.4 Ο αλγόριθμος

Το πλεονέκτημα που έχουν οι αλγόριθμοι των δέντρων παλινδρόμησης είναι ότι δεν χρειάζεται το άτομο που κάνει την πρόβλεψη να αποφασίσει πώς θα επιλέγεται η μεταβλητή για το διαχωρισμό της ρίζας και κάθε κλαδιού, ούτε να ορίσει τους κανόνες με τους οποίους γίνεται ο διαχωρισμός σε κάθε επιλεγμένη μεταβλητή, ούτε το πότε θα σταματήσει ο διαχωρισμός. Όλες αυτές οι λειτουργίες γίνονται αυτόματα από τον αλγόριθμο.

Ο δημοφιλέστερος αλγόριθμος για την εκπαίδευση ενός δέντρου είναι ο «άπληστος» αλγόριθμος στον οποίο η επιλογή των μεταβλητών για το διαχωρισμό και οι κανόνες που ορίζουν πώς γίνεται αυτός ο διαχωρισμός καθορίζονται από την ελαχιστοποίηση του σφάλματος στα «κατά τόπους» δεδομένα. Εφόσον υπάρχουν N χαρακτηριστικά, ο συγκεκριμένος αλγόριθμος θα εξετάσει όλα αυτά τα χαρακτηριστικά ώστε να διαχωριστούν με κριτήριο την μεγιστοποίηση της ακρίβειας πρόβλεψης του συγκεκριμένου δείγματος.

Ο αλγόριθμος δεν προσφέρει την καλύτερη δυνατή αποτύπωση των προβλεπόμενων δεδομένων αλλά επιτελεί τον καλύτερο διαχωρισμό των μεταβλητών αφού τις εξετάσει

σειριακά, αποφασίζοντας ξεχωριστά σε κάθε βήμα χωρίς να λαμβάνει υπόψη του τον τελικό αντίκτυπο στο μοντέλο (Spiliotis, 2022).

Τα δέντρα αποφάσεων χρησιμοποιούνται κυρίως:

- Για την υποστήριξη αποφάσεων
- Κατανόηση χαρακτηριστικών που κάνουν μια παραγωγική διαδικασία επιτυχημένη (π.χ. προϊόν, διαφήμιση)
- Αιτιοκρατική παραγωγή προβλέψεων

Όπως αναφέραμε και στην προηγούμενη παράγραφο για την επιβλεπόμενη μάθηση υπάρχουν πολλές μέθοδοι που ανήκουν στην κατηγορία των τυχαίων δέντρων με σημαντικότερες τις παρακάτω:

- Τυχαία Δάση (Random Forests – RF)
- Δέντρα Σταδιακής Ενίσχυσης (Gradient Boosting Trees – GBT)
- Ada Boost

Στις επόμενες παραγράφους αναλύουμε τους τρεις παραπάνω αλγορίθμους.

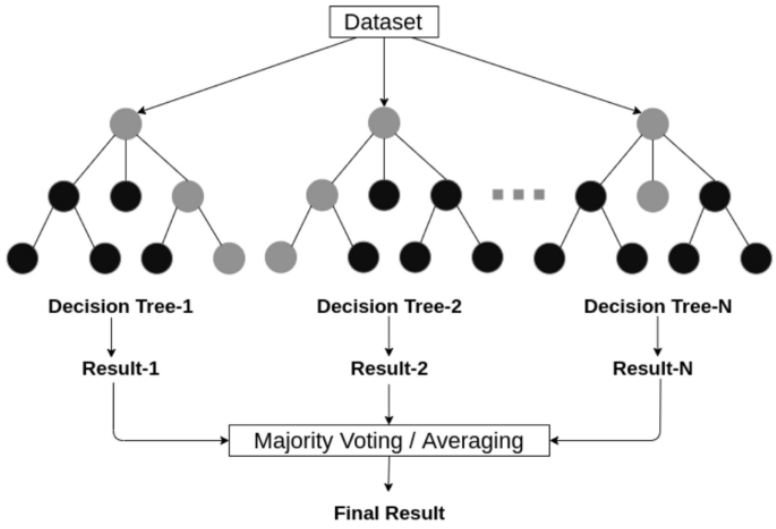
2.4.2.5 Τυχαία Δάση (RF)

Είναι μία συνδυαστική μέθοδος δέντρων πρόβλεψης. Τα αρχικά δεδομένα χωρίζονται τυχαία σε πολλαπλές ομάδες δεδομένων του ίδιου μεγέθους (Διάγραμμα 2.6). Το πρόβλημα που μπορεί να προκύψει εδώ είναι, ότι κατά τη δημιουργία των τυχαίων υποομάδων δεδομένων, κάποιες να περιλαμβάνουν τα ίδια δεδομένα πολλές φορές ενώ κάποια δεδομένα μπορεί να μην χρησιμοποιηθούν ποτέ.

Τυχαία επιλογή γίνεται και στα χαρακτηριστικά των υποσυνόλων ένας δάσους. Με αυτό τον τρόπο κάθε δέντρο του δάσους εκπαιδεύεται μέσω του δικού του σετ δεδομένων.

Σύμφωνα με τα παραπάνω προκύπτουν διαφορετικές προβλέψεις μεταξύ τους λόγω των ανομοιογενών σετ δεδομένων. Έτσι χρησιμοποιείται ο μέσο όρος των προβλέψεων, εξαλείφοντας σε μεγάλο βαθμό τα προβλήματα που έχει ένα μόνο δέντρο, όπως είναι η μη αξιοποίηση του συνόλου των δεδομένων, η ευαισθησία στις ακραίες τιμές και η πιθανότητα προκατάληψης στο αποτέλεσμα.

Διάγραμμα 2.6 Παράδειγμα Δομής Τυχαίου Δάσους συμπεριλαμβάνοντας πολλά δέντρα αποφάσεων



Πηγή: (Sarker, 2021)

Η χρήση των τυχαίων δασών προϋποθέτει τον καθορισμό των παρακάτω παραγόντων:

- Συνολικός αριθμός δέντρων, από τα οποία θα προκύψει ο μέσος όρος των προβλέψεων. Μία τυπική τιμή είναι της τάξης του 30 με 500 δέντρα.
- Αριθμός των χαρακτηριστικών που θα χρησιμοποιηθούν από κάθε δέντρο. Συνήθως επιλέγεται το 1/3 των μεταβλητών που χρησιμοποιούνται στο μοντέλο.
- Οι τιμές των υπερπαραμέτρων που καθορίζουν την ανάπτυξη του κάθε δέντρου.

Ορμώμενοι από το ζήτημα του χρόνου υπολογισμού και του κόστους, παρόλο που οι διαδικασίες γίνονται παράλληλα στα δέντρα, στο τυχαίο δάσος, οι ερευνητές προτείνουν την επιλογή «ρηχών» δέντρων ώστε να μειωθεί το κόστος. Τα δέντρα αυτά ονομάζονται «αδύναμοι μαθητές» διότι λόγω χαμηλού βάθους του δέντρου δεν εκπαιδεύονται επαρκώς από τα δεδομένα.

Όπως προκύπτει από τα αποτελέσματα στη βιβλιογραφία ο μέσος όρος των προβλέψεων που προκύπτει από αυτά τα διαφορετικά δέντρα έχει μεγαλύτερη προβλεπτική ακρίβεια (Spiliotis, 2022).

2.4.2.6 Δέντρα Σταδιακής Ενίσχυσης (GBT)

Η τεχνική των δέντρων σταδιακής ενίσχυσης στηρίζεται σε μία κάπως διαφορετική κατασκευαστικά λογική. Παρόλο που η λογική είναι παρόμοια με τα RF όσον αφορά

τη χρήση πολλαπλών δέντρων, η βασική ιδέα είναι να συνδυαστούν «αδύναμοι μαθητές» για να δημιουργήσουν ένα ισχυρό προβλεπτικό μοντέλο. Η μέθοδος αυτή χρησιμοποιεί πολλαπλά δέντρα αποφάσεων στα οποία η είσοδος προκύπτει από τα σφάλματα που παρήχθησαν από το προηγούμενο δέντρο. Σε κάθε επανάληψη προστίθεται ένα καινούργιο μοντέλο «βασικής γνώσης», το οποίο λογίζεται ως «αδύναμος μαθητής» που εκπαιδεύεται με βάση το σφάλμα του εξεταζόμενου μέχρι τώρα συνολικού μοντέλου (Natekin & Knoll, 2013).

Το GBT ουσιαστικά συνδυάζει τις προβλέψεις των επιμέρους δέντρων, αλλά η διαδικασία γίνεται σειριακά ώστε κάθε δέντρο να βελτιώνει την προβλεπτική του ακρίβεια σε σχέση με τα προηγούμενα.

Επίσης τα GBT εφαρμόζουν τεχνικές «ενίσχυσης» (boosting) που σημαίνει ότι οι προβλέψεις προκύπτουν από τον σταθμισμένο μέσο όρο των προβλέψεων του αρχικού δέντρου και αυτών που προστέθηκαν στο μοντέλο στη συνέχεια.

Οι αλγόριθμοι σταδιακής ενίσχυσης ακολουθούν δύο τύπους συμπεριφορών:

- Ανάπτυξη με έμφαση στα φύλλα (leaf-wise growth): Ξεκινάει με την εκπαίδευση ενός δέντρου και στη συνέχεια προσδιορίζει ποια φύλλα έχουν το μεγαλύτερο σφάλμα πρόβλεψης και εστιάζει σε αυτά, τα οποία και μοντελοποιεί με κάποιο άλλο δέντρο. Κάθε δέντρο προσπαθεί να μοντελοποίησει το σφάλμα που προκύπτει από το προηγούμενο δέντρο. Η διαφορά με τα τυχαία δέντρα είναι ότι δεν χωρίζονται σε δύο κλαδιά σε κάθε κόμβο αλλά μοντελοποιούνται με άλλα δέντρα με πολλά κλαδιά και κόμβους.
- Ανάπτυξη με έμφαση στο επίπεδο (level-wise growth): Σε αυτή την περίπτωση όλα τα φύλλα του προηγούμενου επιπέδου ενισχύονται χωρίς να υπάρχει διαχωρισμός μεταξύ των φύλλων. Αυτός ο αλγόριθμος αποδίδει καλύτερα συνήθως σε μικρά σετ δεδομένων.

Το μεγάλο μειονέκτημα αυτής της μεθόδου, παρόλο που χρησιμοποιείται για υλοποιήσεις βελτιστοποίησης δεδομένων στον τομέα της μηχανικής μάθησης, είναι ότι η αποδοτικότητα και η επεκτασιμότητα δεν είναι αρκετά ικανοποιητικές όταν μιλάμε για μεγάλο όγκο δεδομένων με μεγάλες διαστάσεις δέντρων. Αυτό συμβαίνει γιατί κάθε χαρακτηριστικό ελέγχει όλα τα στιγμιότυπα των δεδομένων (Ke, et al., 2017).

Παρακάτω παρουσιάζονται τρεις βασικές παραλλαγές των δέντρων σταδιακής ενίσχυσης εξίσου δημοφιλείς στην κοινότητα των προβλέψεων (Spiliotis, 2022).

- LightGBM: είναι μία παραλλαγή του αλγορίθμου των Δέντρων Σταδιακής Ενίσχυσης που κερδίζει έδαφος συνεχώς στον τομέα των προβλέψεων διότι μπορεί και διαχειρίζεται μεγάλο όγκο δεδομένων με τη χρήση μικρότερης μνήμης και με μικρότερο χρόνο εκπαίδευσης. Η μέθοδος αυτή δίνει έμφαση στα φύλλα κατά την εκπαίδευση των δεδομένων. Έχει την ικανότητα να προβλέπει πολλαπλές χρονοσειρές με διαφορετικά μοτίβα και χαρακτηριστικά και επίσης έχει την ικανότητα να οδηγείται στον βέλτιστο διαχωρισμό με μια τεχνική τυχαίας δειγματοληψίας εξασφαλίζοντας ισορροπία μεταξύ ταχύτητας και ακρίβειας.
- XGBoost: με τη μέθοδο αυτή αποφεύγεται η υπερπροσαρμογή καθώς βάζει όρια. Επιπλέον επιταχύνει τη διαδικασία της πρόβλεψης αφού λειτουργεί με παράλληλη επεξεργασία και εξοικονομεί πόρους στη μνήμη μέσω βελτιωμένων δομών δεδομένων. Βασίζεται κυρίως στο επίπεδο των δέντρων αλλά υποστηρίζει και την εκπαίδευση με έμφαση στα φύλλα.
- Ada Boost: Η λογική πίσω από αυτόν τον αλγόριθμο μηχανικής μάθησης βασίζεται στον αλγόριθμο της σταδιακής ενίσχυσης των δέντρων που μαθαίνει από τα σφάλματα του παρελθόντος. Με αυτή τη μέθοδο χρησιμοποιείται συνδυασμός «αδύναμων» δέντρων για τη δημιουργία ενός κανόνα υψηλής προβλεπτικής ικανότητας (Schapire, 2013). Σε αυτή τη μέθοδο χρησιμοποιούνται δέντρα αποφάσεων στα οποία υπάρχουν βάρη σε κάθε βήμα μοντελοποίησης για να τιμωρηθούν οι λιγότερο ακριβείς προβλέψεις. Τα βάρη συνεχώς αλλάζουν καθώς προχωράει η προβλεπτική διαδικασία μέχρι το συνολικό σφάλμα να ελαχιστοποιηθεί. Η διαφορά με τις άλλες μεθόδους σταδιακής ενίσχυσης είναι ότι ο συνδυασμός των βαρών γίνεται βάσει των παρατηρήσεων που δεν προβλέπονται με ακρίβεια παρά από τα κατάλοιπα που προκύπτουν από την πρόβλεψη για αυτές τις παρατηρήσεις. Το μειονέκτημα του αλγορίθμου αυτού είναι ότι είναι ευαίσθητο στο θόρυβο και στις ακραίες τιμές (Sarker, 2021).

2.5 Αξιολόγηση Ακρίβειας Προβλέψεων

Για να γίνει μέτρηση της ακρίβειας μιας πρόβλεψης χρησιμοποιούνται διάφορες μέθοδοι οι οποίοι μετρούν με διαφορετικό τρόπο τη διαφορά ανάμεσα στο πραγματικό αποτέλεσμα και την πρόβλεψη του αποτελέσματος, το οποίο ονομάζουμε σφάλμα της πρόβλεψης. Η βιβλιογραφία παρουσιάζει διάφορες μεθόδους για την αξιολόγηση των προβλέψεων με τη χρήση χρονολογικών σειρών. Παρακάτω παρουσιάζουμε τις βασικές και συχνότερα χρησιμοποιούμενες μετρικές δίνοντας έμφαση σε αυτή που χρησιμοποιούμε στη δική μας μελέτη, την sMAPE. Ο βασικός διαχωρισμός μεταξύ των μεθόδων RMSE, MAE και MASE, sMAPE, τις οποίες αναλύουμε πιο κάτω, είναι ότι οι πρώτες μέθοδοι μπορούν να συγκριθούν μεταξύ τους όταν αναφερόμαστε στο ίδιο σετ δεδομένων, όταν όμως συγκρίνουμε τις μεθόδους αυτές μεταξύ διαφορετικών σετ δεδομένων διαφορετικής κλίμακας και όγκου τότε τα αποτελέσματα δεν είναι αξιόπιστα (Hyndman & Koehler, 2006).

- **Ρίζα του μέσου τετραγωνικού σφάλματος (Root Mean Square Error - RMSE)**: υπολογίζει την ρίζα του μέσου τετραγωνικού σφάλματος . Κατά τους (Hyndman & Koehler, 2006) το κριτήριο RMSE έχει δύο βασικά πλεονεκτήματα. Πρώτον ο τρόπος υπολογισμού του εξασφαλίζει ότι είναι στην ίδια κλίμακα με τα δεδομένα, το οποίο προκύπτει από τη χρήση της τετραγωνικής ρίζας και δεύτερον το κριτήριο αυτό έχει θεωρητική συνάφειά με τη στατιστική μοντελοποίηση.

$$RMSE = \sqrt{\frac{1}{k}\sum_{t=1}^{k}\left(Y_t - \hat{Y}_t\right)^2}$$

- **Μέσο απόλυτο σφάλμα (Mean Absolute Error - MAE)**:Η μέθοδος αυτή εστιάζει στον υπολογισμό της μέσης απόλυτης τιμής των σφαλμάτων πρόβλεψης μεταξύ πραγματικής και προβλεπόμενης τιμής και είναι πολύ λιγότερο ευαίσθητη μέθοδος στις ακραίες τιμές σε σχέση με την RMSE (Hyndman & Koehler, 2006). Ο τύπος της MAE φαίνεται στη συνέχεια:

$$MAE = \frac{1}{k}\sum_{t=1}^{k}\left|Y_t - \hat{Y}_t\right|$$

- **Μέσο απόλυτο κλιμακωτό σφάλμα (Mean Absolute Scaled Error - MASE)**: Η διαφορά που παρουσιάζει η συγκεκριμένη μέθοδος σε σχέση με την προηγούμενη (MAE) είναι ότι χρησιμοποιεί όπως φαίνεται από τον παρακάτω τύπο τον μέσο όρο της μεταβολής της απόλυτης τιμής μεταξύ δύο διαδοχικών τιμών της μεταβλητής προσπαθώντας να ανεξαρτητοποιήσει το αποτέλεσμα από τον όγκο των δεδομένων (Hyndman & Koehler, 2006).

$$MASE = \frac{1}{k}\frac{\sum_{t=1}^{k}\left|Y_t - \hat{Y}_t\right|}{\frac{1}{n-m}\sum_{t=m+1}^{n}|Y_t - Y_{t-m}|}$$

Στην ουσία αν παρατηρήσουμε τον παραπάνω τύπο θα διαπιστώσουμε ότι στον αριθμητή έχουμε το MAE και στον παρονομαστή ορίζουμε ένα συντελεστή ο οποίος βοηθάει στο να απαλλαγούμε από την επιρροή της κλίμακας των δεδομένων (Makridakis, Spiliotis, & Assimakopoulos, Statistical and Machine Learning forecasting methods: Concerns and ways forward, 2018)

- **Συμμετρικό μέσο απόλυτο ποσοστό σφάλματος (Symmetrical Mean Absolute Percentage Error – sMAPE)**: Η μέθοδος αυτή προτάθηκε από τον (Makridakis, Accuracy measures: Theoretical and practical concerns, 1993) η οποία στην ουσία διορθώνει το πρόβλημα που ανακύπτει στις μεθόδους MAPE και MdAPE, στις οποίες υπάρχει μεγαλύτερο πέναλτι στα θετικά από ότι στα αρνητικά σφάλματα. Στην ουσία υπολογίζει την ποσοστιαία διαφορά μεταξύ πραγματικών και προβλεπόμενων τιμών λαμβάνοντας υπόψη τη συχνότητα των παρατηρήσεων (Makridakis, Spiliotis, & Assimakopoulos, Statistical and Machine Learning forecasting methods: Concerns and ways forward, 2018). Η μέθοδος αυτή όμως φαίνεται να έχει ένα πρόβλημα ως προς τη συμμετρία μεταξύ θετικών και αρνητικών σφαλμάτων· όταν η απόλυτη τιμή των σφαλμάτων είναι πολύ μεγάλη, το sMAPE δίνει μεγαλύτερη τιμή στα θετικά σφάλματα από ότι στα αρνητικά (Goodwin & Lawton, 1999).

$$sMAPE = \frac{2}{k}\sum_{t=1}^{k}\frac{\left|Y_t - \hat{Y}_t\right|}{|Y_t| + \left|\hat{Y}_t\right|}$$

Σε όλες τις παραπάνω μεθόδους υπολογισμού σφαλμάτων ισχύει:

Y_t: Πραγματική τιμή στο χρόνο t

$\hat{Y}_t$: Τιμή πρόβλεψης στο χρόνο t

k: Ορίζοντας των προβλέψεων

n : ο αριθμός των διαθέσιμων ιστορικών παρατηρήσεων

m: Συχνότητα των χρονοσειρών

2.6 Επαλήθευση Μοντέλων

Είναι σημαντικό, όποιο κριτήριο αξιολόγησης της πρόβλεψης κι αν χρησιμοποιηθεί να μεταχειριστεί κανείς τα ιστορικά δεδομένα με τρόπο που θα μειώσει τα σφάλματα στις μελλοντικές του προβλέψεις.

Όταν χρησιμοποιηθούν ιστορικά δεδομένα για να προβλεφθεί το μέλλον μπορεί να δοκιμαστεί η εν λόγω πρόβλεψη χωρίζοντας τα δεδομένα σε δεδομένα εκπαίδευσης (train set) του μοντέλου και σε τεστ δεδομένα (test set). Οπότε στη συνέχεια θα υπολοστιστή το σφάλμα εκπαίδευσης των δεδομένων.

Η εργασία όμως δεν πρέπει να σταματάει εκεί υπό την έννοια ότι η παρακολούθηση μόνο του συγκεκριμένου σφάλματος μπορεί να οδηγήσει σε παραπλανητικά ή και εσφαλμένα αποτελέσματα.

Θα πρέπει να γίνεται χρήση μιας κάποιας τεχνικής επαλήθευσης του μοντέλου, cross validation όπως λέγεται στη βιβλιογραφία, για να εξασφαλιστεί η αντικειμενικότερη αξιολόγηση των μοντέλων (Πετρόπουλος & Ασημακόπουλος, 2011).

Στο επόμενο κεφάλαιο ακολουθούνται ένα-ένα τα βήματα για ανάλυση και προετοιμασία των χρονοσειρών που χρησιμοποιούνται στα μοντέλα που παρουσιάζονται στο πειραματικό μέρος.

Κεφάλαιο 3. Προετοιμασία και Ανάλυση Χρονοσειρών

3.1 Συλλογή Δεδομένων

Στο κεφάλαιο αυτό παρουσιάζεται η πηγή προέλευσης των δεδομένων, η επιλογή των δεδομένων από το σύνολο που αντλήθηκε, η διαχείριση των ακραίων και μηδενικών τιμών, η ανάλυση των δεδομένων (π.χ εποχικότητα, τάση κτλ.) και η περιγραφή των πρωτογενών δεδομένων και των ιδιοτήτων τους ώστε στο επόμενο κεφάλαιο να επιλεγούν οι κατάλληλες μεταβλητές που θα χρησιμοποιηθούν στα μοντέλα της πρόβλεψης.

Η πρώτη σκέψη ήταν να αναζητηθούν ποσοτικά δεδομένα που να αφορούν αφίξεις επιβατών σε συγκεκριμένα αεροδρόμια, τα οποία δεδομένα όμως θα έπρεπε να είναι αξιόπιστα όσον αφορά τη δομή, τη χρονική διάταξη και τη συνέπεια καθώς επίσης να είναι σχετικά πρόσφατα και να καλύπτουν αρκετά έτη, ώστε να μπορούν να είναι επαρκή για να γίνει με ακρίβεια η πρόβλεψη των μελλοντικών τιμών.

Η αναζήτηση των δεδομένων υπήρξε μία κοπιαστική και χρονοβόρα διαδικασία καθώς αντλήθηκαν δεδομένα από άλλη πηγή[1], η οποία παρείχε συνθετικά δεδομένα (παραγόμενα) για πτήσεις. Αφού φορτώθηκαν τα δεδομένα αυτά σε μια βάση δεδομένων και μετά από στοιχειώδεις ελέγχους χαρακτηρίστηκαν μη αξιοποιήσιμα καθότι παρουσίαζαν σφάλματα στις εγγραφές, δηλαδή ήταν χρονικά ασυνεπή, είχαν μη ρεαλιστική συμπεριφορά η οποία αποτυπωνόταν ξεκάθαρα με γραφικό τρόπο και επίσης δεν παρείχαν τον επιθυμητό αριθμό χρονοσειρών για τουλάχιστον 20 αεροδρόμια αφίξεων.

Εν τέλει, τα δεδομένα αντλήθηκαν από το Kaggle[2], το οποίο είναι θυγατρική της google και αποτελεί μια διαδικτυακή κοινότητα επιστημόνων δεδομένων και μηχανικών που ασχολούνται με τη εφαρμογή αλγορίθμων τεχνητής νοημοσύνης. Μέσα από αυτή την πλατφόρμα δίνεται η δυνατότητα στους χρήστες να βρίσκουν σύνολα δεδομένων, να αναρτούν σύνολα δεδομένων, να συμμετέχουν σε συνεργασίες για την επεξεργασία των δεδομένων καθώς επίσης και να συμμετέχουν σε διαγωνισμούς για την επίλυση προβλημάτων που αφορούν στην επιστήμη των δεδομένων.

3.2 Διαχείριση Δεδομένων

[1] Πηγή: https://developers.amadeus.com/blog/free-fake-pnr-sample-data
[2] Πηγή: https://www.kaggle.com/

Στην εργασία αυτή χρησιμοποιούνται δεδομένα για τις Εσωτερικές Πτήσεις στις Ηνωμένες Πολιτείες Αμερικής[3]. Το σύνολο των δεδομένων που αντλήθηκε, αποτελείται από 3.599.849 εγγραφές και αφορά στην περίοδο 01/1990 έως και 12/2009. Τα δεδομένα είναι καταγεγραμμένα σε μηνιαία βάση και το περιεχόμενό τους έχει ως εξής:

- Origin_airport: κωδικός τριών γραμμάτων που υποδεικνύει το αεροδρόμιο αναχώρησης
- Destination_airport: κωδικός τριών γραμμάτων που υποδεικνύει το αεροδρόμιο άφιξης
- Destination_city: όνομα πόλης άφιξης
- Passengers: αριθμός επιβατών που μεταφέρονται από το αεροδρόμιο αναχώρησης στο αεροδρόμιο άφιξης. Ουσιαστικά αποτελεί τη μεταβλητή που καλούμαστε να προβλέψουμε στην παρούσα εργασία.
- Seats: αριθμός διαθέσιμων θέσεων των πτήσεων
- Flights: αριθμός πτήσεων σε μηνιαία βάση μεταξύ αεροδρομίου αναχώρησης και άφιξης
- Distance: απόσταση μεταξύ αεροδρομίου αναχώρησης και άφιξης
- Fly_date: ο μήνας και ο χρόνος της πτήσης
- Origin_population: ο πληθυσμός της πόλης αναχώρησης
- Destination_population: ο πληθυσμός της πόλης άφιξης

Επίσης, από την ίδια πηγή αντλήθηκαν δεδομένα για τους κωδικούς των αεροδρομίων[4] ώστε να συνδυαστούν οι προορισμοί στο αρχικό σετ με τα επίσημα ονόματα των αεροδρομίων σύμφωνα με την IATA ώστε παρακάτω να επιλεχθούν οι εγγραφές στις οποίες το όνομα των αεροδρομίων ταυτίζεται με το επίσημο όνομα.

3.3 Επεξεργασία Δεδομένων

Στη συνέχεια παρουσιάζεται βήμα-βήμα η επεξεργασία των δεδομένων, δηλαδή με τη χρήση των λογισμικών Rstudio και Mysql, πως έγινε η εισαγωγή των δεδομένων, η

[3] Πηγή: https://www.kaggle.com/datasets/flashgordon/usa-airport-dataset
[4] Πηγή: https://www.kaggle.com/datasets/mike90/airport-codes

επιλογή των συγκεκριμένων αεροδρομίων, η διαχείριση των κενών και μηδενικών τιμών και η επιλογή των μεταβλητών και των χρονοσειρών που θα χρησιμοποιηθούν στην πειραματική διαδικασία.

3.3.1 Λογισμικό Επεξεργασίας και Πρόβλεψης Χρονοσειρών

Τη σημερινή εποχή υπάρχει ένας μεγάλος αριθμός εργαλείων που μπορούν να χρησιμοποιήσουν οι ερευνητές για να προβλέψουν μελλοντικά αποτελέσματα με τη χρήση ιστορικών δεδομένων. Στον κλάδο της επιστήμης που ασχολείται με την πρόβλεψη των μελλοντικών δεδομένων, το ενδιαφέρον στρέφεται κυρίως σε δύο τύπους πακέτων:

- Τα πακέτα που ασχολούνται με την ανάλυση της παλινδρόμησης, ανάλυση χρονοσειρών και άλλων τεχνικών και
- Τα πακέτα που ασχολούνται με την πρόβλεψη

Ένας άλλος διαχωρισμός που γίνεται μεταξύ των προγραμμάτων είναι αν αποτελούν ελεύθερο λογισμικό ή εμπορικό λογισμικό.

Κάποια πολύ γνωστά λογισμικά τα οποία χρησιμοποιούνται για όλες τις ανωτέρω λειτουργίες είναι η γλώσσα R, το Eviews και το SPSS.

Στην πειραματική διαδικασία που περιγράφεται παρακάτω όλες οι προβλέψεις έχουν γίνει με τη χρήση της R καθώς επίσης και οι σχετικές γραφικές παραστάσεις που απεικονίζουν τη συμπεριφορά των χρονοσειρών.

3.3.1.1 Το λογισμικό πακέτο της R - R studio

Η R ανήκει στην κατηγορία των ανοικτών λογισμικών προγραμματισμού και διαθέτει πάρα πολλές βιβλιοθήκες. Μέσα σε αυτά τα πακέτα των βιβλιοθηκών ανήκουν και τα εργαλεία για ανάλυση δεδομένων και την πρόβλεψή τους. Το συγκεκριμένο πρόγραμμα έχει εκδόσεις για όλα τα γνωστά λειτουργικά συστήματα όπως Windows, Linux και MacOs. Πίσω από αυτή τη γλώσσα βρίσκεται μία τεράστια κοινότητα προγραμματιστών που την υποστηρίζουν και την εξελίσσουν.

Ένα από τα μεγάλα της πλεονεκτήματα είναι ότι παρέχει πολλές τεχνικές στατιστικής και γραφικής απεικόνισης και επιπλέον είναι εύκολα επεκτάσιμη. Υπάρχουν ειδικά διαμορφωμένα πακέτα τα οποία υποστηρίζουν διαφορετικές μεθόδους προβλέψεων. Εκτός από τις συναρτήσεις που έχει ενσωματωμένες στο πακέτο της R, υπάρχουν πάρα

πολλές που κυκλοφορούν χωρίς κόστος και μπορούν να ενσωματωθούν, καθώς επίσης υπάρχει η δυνατότητα ανάπτυξης και συγγραφής δικών μας συναρτήσεων.

Για να διευκολυνθεί η εργασία των ερευνητών με τη χρήση της R χρησιμοποιείται το εργαλείο ανοιχτού κώδικα Rstudio, το οποίο κάνει εύκολη την οπτικοποίηση των διαδικασιών, της διαχείρισης των δεδομένων και των αποτελεσμάτων.

3.3.1.2 Το λογισμικό MySQL

Είναι μία από τις πιο δημοφιλείς βάσεις δεδομένων ανοιχτού κώδικα στον κόσμο. Κατατάσσεται δεύτερη μετά την Oracle Database και χρησιμοποιείται από κολοσσούς διαδικτυακών εφαρμογών όπως είναι το Facebook, Netflix κ.ά.

Οι βάσεις δεδομένων είναι στην ουσία πίνακες στους οποίους αποθηκεύονται τα δεδομένα. Στη συγκριμένη περίπτωση η MySQL είναι μία σχεσιακή βάση δεδομένων, δηλαδή αυτός ο τύπος συστήματος διαχείρισης βάσεων δεδομένων (ΣΔΒΔ) οργανώνει και αποθηκεύει πληροφορίες με έναν δομημένο τρόπο, χρησιμοποιώντας ένα σύνολο πινάκων με στήλες και γραμμές. Οι σχέσεις μεταξύ των πινάκων είναι προκαθορισμένες, παρέχοντας έτσι ένα σταθερό πλαίσιο για την οργάνωση και τη διαχείριση των δεδομένων. Επίσης χρησιμοποιεί την γλώσσα SQL (Structured Query Language) για να διαχειριστεί τα δεδομένα. Οι κανόνες που επιβάλει βοηθούν στη διατήρηση της συνοχής των δεδομένων.

Όπως αναφέρθηκε προηγουμένως είναι λογισμικό ανοιχτού κώδικα οπότε επιτρέπει στον οποιοδήποτε να το χρησιμοποιήσει χωρίς κόστος. Επειδή έχει υψηλή απόδοση, ευκολία χρήσης και αξιοπιστία επιλέγεται από τους περισσότερους προγραμματιστές.

3.3.2 Χρήση των λογισμικών

Το πρώτο βήμα που έγινε ήταν να φορτωθούν τα δύο σύνολα δεδομένων, όπως περιγράφονται προηγουμένως, σε δύο πίνακες στη MySQL, δηλαδή στον έναν οι 3.599.849 εγγραφές των εσωτερικών πτήσεων των ΗΠΑ και στον δεύτερο στοιχεία των αεροδρομίων κατά IATA, όπως ο τριψήφιος κωδικός, η πόλη στην οποία βρίσκεται το αεροδρόμιο, η χώρα, το γεωγραφικό μήκος και πλάτος του αεροδρομίου.

Στο λογισμικό της MySQL έγιναν οι αρχικοί έλεγχοι των ιστορικών δεδομένων για μηδενικές και κενές τιμές, ο έλεγχος ταύτισης των ονομάτων των προορισμών με αυτούς της IATA, έγινε η συνάθροιση των δεδομένων χρονικά και γεωγραφικά,

δηλαδή αθροίστηκαν τα δεδομένα ανά αεροδρόμιο προορισμού και ανά μήνα/έτος άφιξης και η επιλογή των χρονοσειρών όπως θα περιγραφεί αναλυτικά πιο κάτω.

Στη συνέχεια οι επιλεγμένες χρονοσειρές φορτώθηκαν στο Rstudio, στο οποίο έγινε περαιτέρω η περιγραφική ανάλυση, δηλαδή η παρουσίαση στατιστικών δεικτών και γραφημάτων και η προκαταρκτική ανάλυσή τους.

Τέλος, όπως θα περιγράφει στο επόμενο κεφάλαιο, με το Rstudio έγινε η επιλογή και χρήση των μοντέλων πρόβλεψης και οι υπολογισμοί των σφαλμάτων πρόβλεψης των δεδομένων.

3.3.3 Διαχείριση Μηδενικών και Κενών Τιμών

Είναι πολύ σημαντικό πριν χρησιμοποιηθούν τα δεδομένα να ελέγχονται για την ύπαρξη, κενών ή μηδενικών τιμών. Στην περίπτωση λοιπόν του πρώτου πίνακα που φορτώθηκε στη MySQL έγιναν «ερωτήματα» ώστε να ελεγχθούν δύο βασικές μεταβλητές, οι «Seats» και «Flights», δηλαδή ο αριθμός των θέσεων και των πτήσεων αντίστοιχα, γιατί αυτές οι μεταβλητές έχουν άμεση σχέση με τον αριθμό των επιβατών σε μία πτήση και επίσης είναι δύο στοιχεία που χαρακτηρίζουν την πτήση. Αν οι τιμές σε αυτές τις μεταβλητές είναι κενές ή μηδενικές δεν έχει νόημα να προστεθεί αυτή η εγγραφή στο σύνολο των δεδομένων. Άρα η συγκεκριμένη εγγραφή δεν μπορεί να συναθροιστεί με τις υπόλοιπες εγγραφές με προορισμό το ίδιο αεροδρόμιο, τον ίδιο μήνα και έτος.

Με την παραπάνω λογική, από το σύνολο των δεδομένων αποφασίστηκε να εξαιρεθούν οι εγγραφές που είναι μηδενικές ή κενές σε αυτές τις δύο μεταβλητές, διότι αν έχουμε μηδενικές/κενές τιμές στις πτήσεις, τις θεωρούμε ως μη γενόμενες ή ότι υπάρχει σφάλμα στη συγκεκριμένη εγγραφή. Κατά αντιστοιχία μηδενική/κενή τιμή στις θέσεις σημαίνει σφάλμα στην εγγραφή διότι εφόσον υπάρχει πτήση και αεροπλάνο δεν μπορεί να μην υπάρχει αριθμός θέσεων.

Ωστόσο διατηρήθηκαν εγγραφές που είχαν μηδέν στην μεταβλητή «Passengers», με την οποία απεικονίζεται ο αριθμός των επιβατών ανά εγγραφή, με το σκεπτικό ότι μπορεί να μην πέταξε κανείς τη συγκεκριμένη μέρα με τη συγκεκριμένη πτήση (ίσως σπάνιο ως φαινόμενο, αλλά ρεαλιστικό σε ειδικές συνθήκες π.χ. επιστροφή στον αερολιμένα προέλευσης για επισκευή κ.ά.) ή να μην πραγματοποιήθηκε η πτήση λόγω κακοκαιρίας ή άλλων συνθηκών.

3.3.4 Επιλογή Δεδομένων Ενδιαφέροντος

Στην επόμενη φάση, και πάντα με το βλέμμα στην επίλυση του συγκεκριμένου προβλήματος, δηλαδή την πρόβλεψη της ροής των επιβατών σε κάποια συγκεκριμένα αεροδρόμια, τα οποία θα επιλεγούν στη συνέχεια μετά την οριστικοποίηση της επιλογής των μεταβλητών, εξετάζεται ο αριθμός των επιβατών που αφικνούνται σε ένα αεροδρόμιο ανεξάρτητα από το ποια περιοχή προέρχονται. Με αυτήν τη λογική εξαιρέθηκαν οι μεταβλητές «Origin airport» και «Origin_population» από το σύνολο των δεδομένων.

Επίσης επειδή οι μεταβλητές που αναφέρονται στον πληθυσμό της πόλης άφιξης και στην απόσταση μεταξύ αεροδρομίου αναχώρησης και άφιξης θεωρήθηκαν μη σχετικές με το προς εξέταση φαινόμενο, εξαιρέθηκαν κι αυτές, δηλαδή οι μεταβλητές «Destination_population» και «Distance».

Επιπλέον σε αυτό το στάδιο επιλέχτηκε να δημιουργηθούν δύο χρήσιμες μεταβλητές από την «Fly_date», η οποία ήταν της μορφής π.χ. 2009-01-01. Οι δύο μεταβλητές που δημιουργήθηκαν είναι το «month» και το «year», που όπως υποδηλώνει και το όνομά τους αναφέρονται στο μήνα και το έτος της άφιξης στον επιλεγμένο προορισμό. Ο διαχωρισμός αυτός έγινε με τη λογική ότι θα χρησιμοποιηθούν ως εξωγενείς μεταβλητές στην πειραματική διαδικασία, όπως θα εξηγηθεί παρακάτω.

Σε αυτό το σημείο έγινε ένας επιπλέον έλεγχος των δεδομένων των εσωτερικών πτήσεων των ΗΠΑ σε σχέση με τον πίνακα με τα στοιχεία των αεροδρομίων, που όπως αναφέρθηκε παραπάνω περιέχει τα επίσημα στοιχεία από την IATA. Με τη χρήση του εργαλείου της MySQL έγινε συνδυασμός των δύο πινάκων και διαπιστώθηκε ότι πολλές εγγραφές είτε είχαν κενή τιμή στη μεταβλητή του αεροδρομίου άφιξης είτε το όνομα δεν ταυτιζόταν με αυτό της IATA, οπότε και αυτές οι εγγραφές εξαιρέθηκαν από το σύνολο των δεδομένων.

3.3.5 Συνάθροιση Δεδομένων

Το επόμενο βήμα, μετά την εξαίρεση των παραπάνω μεταβλητών και εφόσον παρατηρήθηκε, όπως περιγράφηκε προηγουμένως, ότι στα δεδομένα που αφορούσαν τα ίδια αεροδρόμια άφιξης υπήρχαν αρκετές εγγραφές τον ίδιο μήνα και χρόνο, τόσο από τα ίδια αεροδρόμια αναχώρησης όσο και από διαφορετικά και επειδή στην εργασία αυτή το ενδιαφέρον εστιάζεται στον αριθμό των επιβατών που φτάνουν σε ένα

αεροδρόμιο χωρίς να υπάρχει ενδιαφέρον για την προέλευση τους, αποφασίστηκε να αθροιστούν οι εγγραφές ανά αεροδρόμιο άφιξης, έτος και μήνα.

Το αποτέλεσμα αυτής της συνάθροισης των εγγραφών σε μία μοναδική εγγραφή ανά αεροδρόμιο άφιξης τον συγκεκριμένο μήνα και χρόνο, οδήγησε στην δημιουργία 26022 μηνιαίων εγγραφών, οι οποίες αφορούν 280 αεροδρόμια άφιξης για τα έτη 1990 έως και 2009, δηλαδή 280 χρονοσειρών.

Οι μεταβλητές, μετά από αυτή την ενέργεια, τροποποιήθηκαν ως εξής:

destination: κωδικός τριών γραμμάτων IATA ο οποίος αφορά το αεροδρόμιο άφιξης

city: η πόλη άφιξης

sum_passengers: ο συνολικός αριθμός των επιβατών για το συγκεκριμένο προορισμό ανά μήνα και έτος

sum_seats: ο συνολικός αριθμός των θέσεων των πτήσεων για τον συγκεκριμένο προορισμό ανά μήνα και έτος

sum_flights: ο συνολικός αριθμός των πτήσεων για τον συγκεκριμένο προορισμό ανά μήνα και έτος

month: ο μήνας διενέργειας των πτήσεων

year: το έτος διενέργειας των πτήσεων

3.3.6 Κριτήρια Επιλογής Αεροδρομίων και Διαστήματος Δεδομένων

Επειδή το ενδιαφέρον της συγκεκριμένης εργασίας εστιάζεται στην πρόβλεψη του αριθμού των αφίξεων των επιβατών, η επιλογή των δεδομένων από το παραπάνω σύνολο έγινε με βάση τα εξής κριτήρια:

- Πιο πρόσφατες χρονικά εγγραφές. Ο λόγος είναι, όπως αναφέραμε και στη θεωρία, όσο πιο πρόσφατα τα δεδομένα τόσο πιο κοντά στη σημερινή πραγματικότητα θα είναι η πρόβλεψη των μελλοντικών τιμών
- Αεροδρόμια άφιξης τα οποία έχουν εγγραφές σε όλους τους μήνες και για όλα τα επιλεγμένα έτη, δηλαδή από 01/2002 έως και τον 12/2009
- Δείγμα αεροδρομίων που περιλαμβάνονται στα μεγάλα αεροδρόμια σύμφωνα με τον ορισμό της κατηγορίας των μεγεθών των αεροδρομίων του Σεπτεμβρίου

του 2008, της επίσημης ιστοσελίδας[5] της Ομοσπονδιακής Διοίκησης Αεροπορίας των ΗΠΑ.

- Τα ιστορικά δεδομένα πρέπει να είναι τουλάχιστον 4πλάσια από τα δεδομένα που προβλέπονται ώστε να μπορεί το μοντέλο να εκπαιδευτεί επαρκώς για να δώσει αξιόπιστα αποτελέσματα πρόβλεψης
- Αριθμός αεροδρομίων που να είναι σχετικά εύκολα διαχειρίσιμος ως προς την ανάλυση και την παρουσίαση των δεδομένων τους, δηλαδή ο αριθμός δεδομένων να εξασφαλίζει robust αποτελέσματα και παράλληλα να διευκολύνει την παρουσίαση τους

3.3.7 Επιλογή Δεδομένων

Λαμβάνοντας υπόψη όλες τις παραπάνω προϋποθέσεις που τέθηκαν για τα δεδομένα, από το σύνολο των ιστορικών δεδομένων, επιλέχθηκαν 20 χρονοσειρές που αντιστοιχούν σε 20 αεροδρόμια άφιξης για τα έτη 2002-2009.

Ο διαχωρισμός του συνόλου των δεδομένων βάσει των κριτηρίων της προηγούμενης παραγράφου έγινε ως εξής:

Δεδομένα εκπαίδευσης (Train set): επιλέχθηκαν τα έτη 2002-2007 (20 χρονοσειρές και 72 μήνες-παρατηρήσεις ανά χρονοσειρά)

Δεδομένα επαλήθευσης (Validation set): επιλέχθηκε το έτος 2008 (20 χρονοσειρές και 12 μήνες-παρατηρήσεις ανά χρονοσειρά)

Δεδομένα αξιολόγησης (Test set): επιλέχθηκε το έτος 2009 (20 χρονοσειρές και 12 μήνες-παρατηρήσεις ανά χρονοσειρά)

Στον παρακάτω πίνακα (Πίνακας 3.1) φαίνονται αναλυτικά τα στοιχεία για τα επιλεγμένα αεροδρόμια, όπως είναι ο τριψήφιος κωδικός τους κατά IATA, το αναλυτικό όνομα του αεροδρομίου, η πόλη στην οποία βρίσκεται και το μέσο ετήσιο πλήθος αφίξεων πτήσεων και επιβατών για τα έτη 2002-2007, που αντιστοιχεί στα δεδομένα εκπαίδευσης.

Παρατηρώντας κανείς τον παρακάτω πίνακα διαπιστώνει πως οι μέσες ετήσιες αφίξεις επιβατών κυμαίνονται από 7.318.193 στο IAD μέχρι 33.788.455 στο αεροδρόμιο ATL και του αριθμού των ετήσιων πτήσεων από 113.686,7 στο SFO έως 379.721 στο ATL.

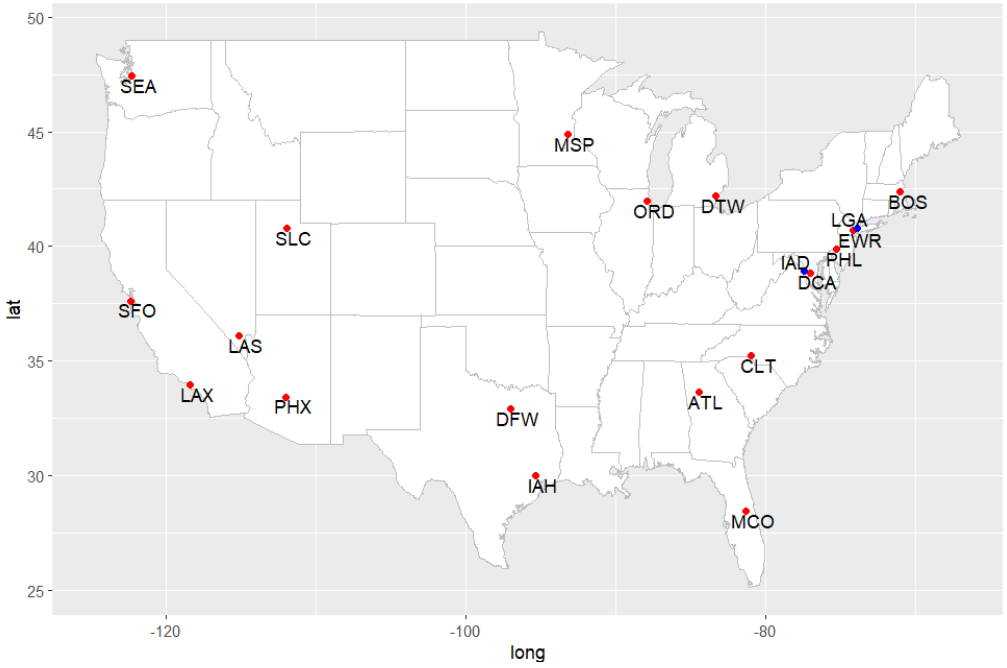
[5] Πηγή: https://www.faa.gov/airports/planning_capacity/npias/current/historical

Πίνακας 3.1 Αεροδρόμια άφιξης υπό εξέταση και μέσος αριθμός επιβατών και πτήσεων για τα έτη 2002-2007

	destination	city	avg_passengers	avg_flights
1	ATL	Atlanta, GA	33788455	379721
2	BOS	Boston, MA	9855498	132617,67
3	CLT	Charlotte, NC	11683510,83	173177
4	DCA	Washington, DC	7498611,83	116017,33
5	DFW	Dallas, TX	23069441,17	297914,67
6	DTW	Detroit, MI	14255969,83	202387,67
7	EWR	Newark, NJ	10841702,33	145381,33
8	IAD	Washington, DC	7318193	128387,33
9	IAH	Houston, TX	13992541,33	195807,83
10	LAS	Las Vegas, NV	17617715	155872
11	LAX	Los Angeles, CA	19740548,5	208711
12	LGA	New York, NY	10866825,5	160073,83
13	MCO	Orlando, FL	13510311,83	126323,17
14	MSP	Minneapolis, MN	14558048,67	191475,33
15	ORD	Chicago, IL	28119648,5	373969,83
16	PHL	Philadelphia, PA	11446228,33	166833,33
17	PHX	Phoenix, AZ	17382664,17	184916,67
18	SEA	Seattle, WA	11862321	129830
19	SFO	San Francisco, CA	10893032,17	113686,67
20	SLC	Salt Lake City, UT	8416272,33	115670,50

Πηγή: Kaggle και επεξεργασία στο Rstudio

Διάγραμμα 3.1 Υπό μελέτη αεροδρόμια στο χάρτη Ηνωμένων Πολιτειών Αμερικής



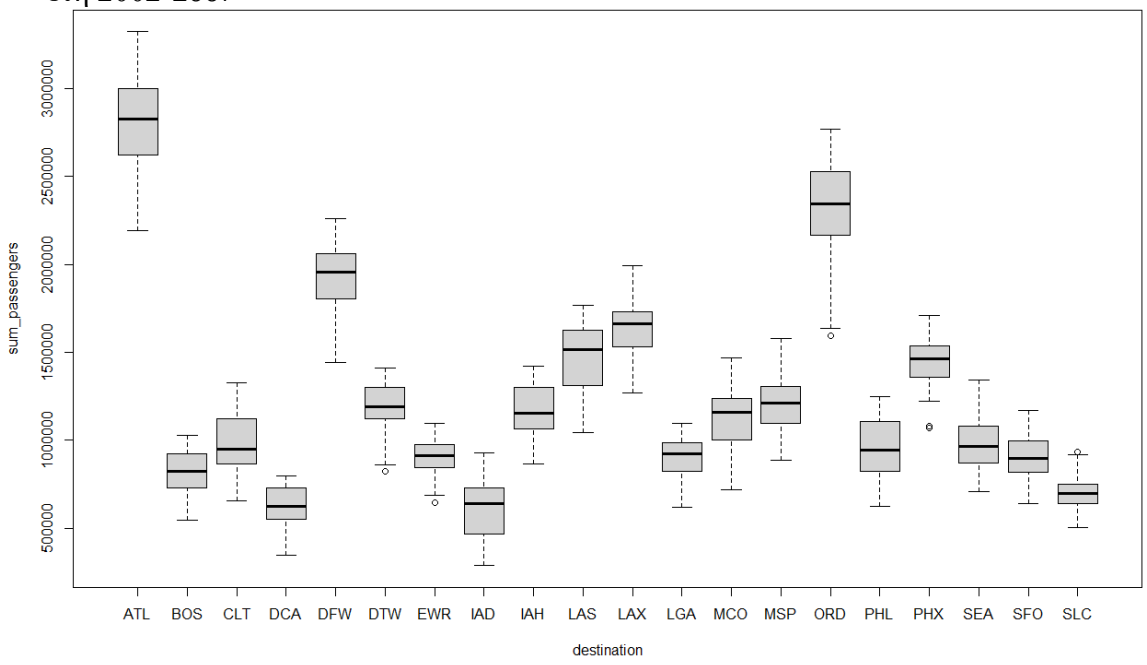
Πηγή: Επεξεργασία στο Rstudio

Επιπλέον της αναφοράς στα βασικά στοιχεία των αεροδρομίων που επιλέχθηκαν παρουσιάζεται και η τοποθεσία των αεροδρομίων στο χάρτη των Ηνωμένων Πολιτείων Αμερικής ώστε να μπορεί κάποιος να δει που βρίσκεται χωροταξικά το κάθε αεροδρόμιο και πως ενδεχομένως μπορεί να συνδέεται η κίνηση των αεροδρομίων σε σχέση με τις καιρικές συνθήκες ανάλογα την εποχή του έτους, κάτι που θα αναφερθεί εκτενέστερα στη συνέχεια.

3.3.8 Διαχείριση Ακραίων Τιμών

Σε αυτό το σημείο γίνεται αναφορά στην ύπαρξη ή όχι ακραίων τιμών. Εφόσον αποφασίστηκε να χρησιμοποιηθούν τα ιστορικά δεδομένα των ετών 2002-2007 για την πρόβλεψη των ετών 2008 και 2009 αντίστοιχα, θα ελεγχθεί αν τα δεδομένα αυτά φέρουν ακραίες τιμές στον αριθμό των επιβατών για το κάθε αεροδρόμιο και πως θα αντιμετωπιστεί το φαινόμενο. Το συμπέρασμα αυτό εξάγεται από τα boxplot, που απεικονίζονται στο διάγραμμα, στο οποίο φαίνεται ο αριθμός των αφίξεων επιβατών για τα έτη 2002-2007 ανά αεροδρόμιο και οι ακραίες τιμές τους, στο πάνω άκρο για το αεροδρόμιο SLC και στο κάτω άκρο για τα αεροδρόμια DTW, EWR, ORD και PHX.

Διάγραμμα 3.2 Boxplot του αριθμού των αφίξεων επιβατών ανά αεροδρόμιο για τα έτη 2002-2007



Πηγή: Επεξεργασία στο Rstudio

Μετά και από τον οπτικό έλεγχο για ακραίες τιμές επιλέγεται να μην αλλοιωθούν οι πραγματικές τιμές των ιστορικών δεδομένων και να χρησιμοποιηθούν ως έχουν για την πρόβλεψη των αφίξεων για τα επόμενα έτη, καθότι θεωρείται ότι δεν υπάρχει σφάλμα

στη καταγραφή των τιμών αλλά αποτυπώνουν μία πραγματική συνθήκη σε κάποιους συγκεκριμένους μήνες, που οδήγησε σε τόσο μικρές τιμές σε σχέση με τους μέσους όρους αφίξεων επιβατών στα συγκεκριμένα αεροδρόμια και θα βοηθήσει την πρόβλεψη των δεδομένων.

Στην προκαταρκτική ανάλυση των δεδομένων παρουσιάζονται αναλυτικά τα δεδομένα εκπαίδευσης καθότι αυτά θεωρούνται ιστορικά δεδομένα για τη διαδικασία της πρόβλεψης όπως αναφέρθηκε και πιο πάνω.

3.4 Προκαταρτική Ανάλυση Χρονοσειρών

Η πρώτη προσέγγιση μιας χρονοσειράς γίνεται μέσω της περιγραφικής στατιστικής όπου υπολογίζονται κάποιοι βασικοί στατιστικοί δείκτες όπως είναι η μέση τιμή, η τυπική απόκλιση κ.τ.λ.. Στα πλαίσια αυτής της στατιστικής ανάλυσης είναι και η οπτικοποίηση της συμπεριφοράς της χρονοσειράς μέσω της γραφικής της αναπαράστασης. Με μία ματιά μπορούν να εξαχθούν πάρα πολλά χρήσιμα συμπεράσματα για τη συμπεριφορά της, για παράδειγμα αν υπάρχει τάση, εποχικότητα, θόρυβος, κυκλικότητα, ακραίες τιμες κ.ά.

Στόχος είναι να εντοπιστούν αν υπάρχουν μη αναμενόμενες συμπεριφορές και πως μπορούν αυτές να αντιμετωπιστούν με τον βέλτιστο τρόπο κατά περίπτωση. Η ανάλυση αυτή οδηγεί σε μία πιο σωστή επιλογή μεταβλητών και εν γένει της μεθοδολογίας πρόβλεψης και του συγκεκριμένου προβλεπτικού μηχανισμού.

3.4.1 Στατιστική Ανάλυση

Οι μεταβλητές όπως περιγράφονται παραπάνω χωρίζονται σε δύο κατηγορίες, ποιοτικές/κατηγορικές μεταβλητές και ποσοτικές μεταβλητές. Στην πρώτη κατηγορία ανήκουν:

destination: περιλαμβάνει τα ονόματα των 20 αεροδρομίων άφιξης

month: μήνας διενέργειας των πτήσεων. Τοποθετείται εδώ με τη σημείωση ότι με τη χρήση της κατάλληλης εντολής στην R, τη factor, δημιουργούνται 12 στάθμες. Η επεξηγηματική αυτή μεταβλητή χρησιμοποιείται ανεξάρτητα από τον χρόνο για να δείξει σε ποια στάθμη βρίσκεται η κάθε τιμή των ποσοτικών μεταβλητών.

Στις ποσοτικές μεταβλητές, οι οποίες είναι διακριτές μεταβλητές και όχι συνεχόμενες διότι οι τιμές τους είναι ανά μήνα/έτος και προορισμό, κατατάσσονται οι παρακάτω:

sum_passengers: ο αριθμός των επιβατών ανά μήνα/έτος και αεροδρόμιο άφιξης

sum_seats: ο αριθμός των θέσεων ανά μήνα/έτος και αεροδρόμιο άφιξης

sum_flights: ο αριθμός των πτήσεων ανά μήνα/έτος και αεροδρόμιο άφιξης

year: έτος διενέργειας των πτήσεων. Είναι μία επεξηγηματική μεταβλητή για τις υπόλοιπες μεταβλητές διότι τις τοποθετεί χρονικά και με αυτό τον τρόπο μπορεί κάποιος να τις συγκρίνει ανά έτος

Στόχος είναι να μελετηθεί η συμπεριφορά των ποιοτικών και ποσοτικών μεταβλητών μέσα από τα ιστορικά δεδομένα, για να ελεγχθεί στη συνέχεια πως η κάθε μεταβλητή μπορεί να επηρεάσει την επικείμενη πρόβλεψη της εξαρτημένης μεταβλητής, που στην περίπτωση της συγκεκριμένης εργασίας είναι ο αριθμός αφίξεων επιβατών στα συγκεκριμένα 20 αεροδρόμια.

Στον πίνακα (Πίνακας 3.2) παρουσιάζονται περιληπτικά οι στατιστικοί δείκτες για όλες τις μεταβλητές ποσοτικές και ποιοτικές. Οι βασικοί δείκτες είναι η μέση τιμή (Mean), η τυπική απόκλιση (Std.Dev.), η ελάχιστη τιμή (Min), η τιμή για το πρώτο τεταρτημόριο (Pctl.25), η τιμή για το 3^{ο} τεταρτημόριο (Pctl. 75) και η μέγιστη τιμή (Max).

Πίνακας 3.2 Περίληψη στατιστικών δεικτών για το σύνολο των δεδομένων

Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
destination	1440						
sum_passengers	1440	1236323	582082	286268	847154	1455798	3323942
sum_seats	1440	1688629	742465	508939	1220519	1949010	4264570
sum_flights	1440	15412	6641	4137	10645	16916	35585
year	1440	2004	1.7	2002	2003	2006	2007

Πηγή: Επεξεργασία στο Rstudio-Excel

Σε αυτό το σημείο να σημειωθεί ότι οι υπολογισμοί αφορούν μηνιαία δεδομένα, δηλαδή όλοι οι στατιστικοί δείκτες προκύπτουν από τις τιμές των μεταβλητών σε μηνιαία κλίμακα π.χ. η μέγιστη τιμή στον αριθμό των αφίξεων επιβατών αφορά το αεροδρόμιο της Ατλάντα (ATL), τον μήνα Ιούλιο του 2017.

Πίνακας 3.3 Συνοπτικά στατιστικά στοιχεία ανά αεροδρόμιο για τα έτη 2002-2007

Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max	Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
destination: ATL								destination: LAX							
sum_passengers	72	2815705	264524	2193234	2623606	2995229	3323942	sum_passengers	72	1645046	171074	1267048	1534455	1729904	1995373
sum_seats	72	3783788	202873	3219229	3668354	3905189	4264570	sum_seats	72	2205910	111650	1910054	2162462	2275735	2405522
sum_flights	72	31643	2117	26050	30343	33297	35585	sum_flights	72	17393	1187	13368	16885	18135	19707
year	72	2004	1.7	2002	2003	2006	2007	year	72	2004	1.7	2002	2003	2006	2007
destination: BOS								destination: LGA							
sum_passengers	72	821292	118784	547076	728346	922754	1027894	sum_passengers	72	905569	107238	617183	827611	981728	1098162
sum_seats	72	1187889	78562	938448	1119366	1243640	1336616	sum_seats	72	1325042	81456	1070756	1296913	1379263	1466433
sum_flights	72	11051	812	9039	10378	11680	12464	sum_flights	72	13339	1635	8586	13196	14302	15030
year	72	2004	1.7	2002	2003	2006	2007	year	72	2004	1.7	2002	2003	2006	2007
destination: CLT								destination: MCO							
sum_passengers	72	973626	169956	653454	865918	1118827	1327797	sum_passengers	72	1125859	162641	716897	1001055	1240155	1469450
sum_seats	72	1392557	137665	1072369	1285591	1512563	1644674	sum_seats	72	1428035	151330	1104508	1291622	1549321	1703082
sum_flights	72	14431	2296	8908	13508	16250	17301	sum_flights	72	10527	1248	8041	9476	11599	12712
year	72	2004	1.7	2002	2003	2006	2007	year	72	2004	1.7	2002	2003	2006	2007
destination: DCA								destination: MSP							
sum_passengers	72	624884	110908	343989	550878	719769	795480	sum_passengers	72	1213171	153487	886014	1099060	1301452	1580735
sum_seats	72	947889	101944	659279	894218	1025270	1085252	sum_seats	72	1693204	126012	1410072	1607643	1778397	1968574
sum_flights	72	9668	1474	5244	9654	10498	11033	sum_flights	72	15956	1393	12792	14866	16924	19027
year	72	2004	1.7	2002	2003	2006	2007	year	72	2004	1.7	2002	2003	2006	2007
destination: DFW								destination: ORD							
sum_passengers	72	1922453	195325	1442859	1802802	2057315	2260627	sum_passengers	72	2343304	269627	1593825	2171197	2527553	2772441
sum_seats	72	2592307	136557	2230802	2529432	2666871	2827979	sum_seats	72	3149318	186803	2620890	3056184	3261257	3555405
sum_flights	72	24826	1927	20466	23483	26124	28686	sum_flights	72	31164	2213	23829	30393	32724	34890
year	72	2004	1.7	2002	2003	2006	2007	year	72	2004	1.7	2002	2003	2006	2007

destination: DTW							
sum_passengers	72	1187997	131119	822722	1121030	1302345	1411368
sum_seats	72	1677213	101113	1412990	1615264	1764949	1849828
sum_flights	72	16866	1159	13952	16238	17651	19237
year	72	2004	1.7	2002	2003	2006	2007
destination: EWR							
sum_passengers	72	903475	103352	645623	847210	973375	1096492
sum_seats	72	1258051	62135	1022273	1234108	1295837	1379497
sum_flights	72	12115	703	9633	11706	12627	13385
year	72	2004	1.7	2002	2003	2006	2007
destination: IAD							
sum_passengers	72	609849	168334	286268	470105	726748	929144
sum_seats	72	866650	208908	508939	719184	993775	1338459
sum_flights	72	10699	3834	4137	8853	11785	19854
year	72	2004	1.7	2002	2003	2006	2007
destination: IAH							
sum_passengers	72	1166045	150112	867544	1064227	1296893	1421455
sum_seats	72	1543799	104876	1277353	1469982	1633222	1719164
sum_flights	72	16317	1989	12445	14566	18105	19621
Year	72	2004	1.7	2002	2003	2006	2007
destination: LAS							
sum_passengers	72	1468143	186165	1044885	1315973	1626456	1769650
sum_seats	72	1888346	181147	1498002	1725286	2038754	2185563
sum_flights	72	12989	1449	9941	11653	14287	15272
Year	72	2004	1.7	2002	2003	2006	2007

destination: PHL							
sum_passengers	72	953852	167880	622454	823202	1102033	1250273
sum_seats	72	1406103	158644	1028630	1261580	1533024	1691996
sum_flights	72	13903	2298	8587	12587	15713	17128
year	72	2004	1.7	2002	2003	2006	2007
destination: PHX							
sum_passengers	72	1448555	140045	1072574	1357595	1536654	1711593
sum_seats	72	1944439	93354	1637505	1895029	2014809	2109511
sum_flights	72	15410	1060	11570	15156	16061	16637
year	72	2004	1.7	2002	2003	2006	2007
destination: SEA							
sum_passengers	72	988527	165689	705604	872875	1079329	1344640
sum_seats	72	1306988	129757	1065198	1209371	1392475	1568223
sum_flights	72	10819	895	9001	10246	11315	12688
year	72	2004	1.7	2002	2003	2006	2007
destination: SFO							
sum_passengers	72	907753	122617	640508	820977	994092	1172150
sum_seats	72	1226165	84645	1023377	1167858	1288375	1395548
sum_flights	72	9474	840	6888	9090	9963	11194
year	72	2004	1.7	2002	2003	2006	2007
destination: SLC							
sum_passengers	72	701356	94844	502645	640330	750382	932332
sum_seats	72	948880	77586	774900	907598	985504	1144702
sum_flights	72	9639	1898	4919	9282	10408	12711
year	72	2004	1.7	2002	2003	2006	2007

Πηγή: Επεξεργασία στο RStudio-Excel

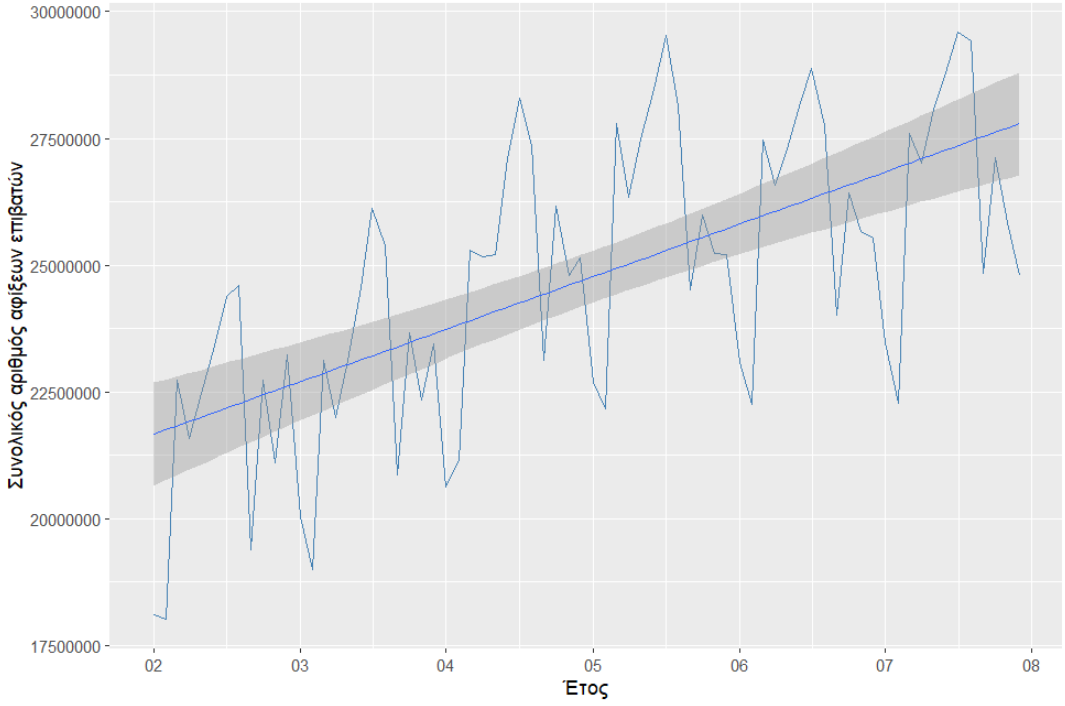
Από τον παραπάνω πίνακα (Πίνακας 3.2) λοιπόν προκύπτει όσον αφορά τις ποσοτικές μεταβλητές ότι για παράδειγμα για τον αριθμό των επιβατών (sum_passengers) ο συνολικός αριθμός εγγραφών είναι 1440, η μέση τιμή τους είναι 1.236.323, η τυπική απόκλιση είναι 582.082, η ελάχιστη τιμή 286.268, στο 1ο τεταρτημόριο η τιμή είναι 847.154, στο 3ο τεταρτημόριο 1.455.798 και η μέγιστη τιμή 3.323.942.

Εκτός από τα συγκεντρωτικά δεδομένα, δημιουργήθηκε πίνακας (**Σφάλμα! Το αρχείο προέλευσης της αναφοράς δεν βρέθηκε.**) όπου παρουσιάζονται τα συνοπτικά στατιστικά στοιχεία για κάθε ένα από τα 20 επιλεγμένα αεροδρόμια (destination) για να μπορεί με εύκολο τρόπο να δει κανείς τα μεγέθη που αφορούν το κάθε αεροδρόμιο χωριστά.

3.4.2 Γραφική Αναπαράσταση

Όπως εξηγήθηκε παραπάνω, για να εξαχθούν τα πρώτα συμπεράσματα για τη συμπεριφορά των χρονολογικών μας σειρών γίνεται απεικόνιση τους.

Διάγραμμα 3.3 Συνολικός αριθμός αφίξεων επιβατών για τα έτη 2002-2007



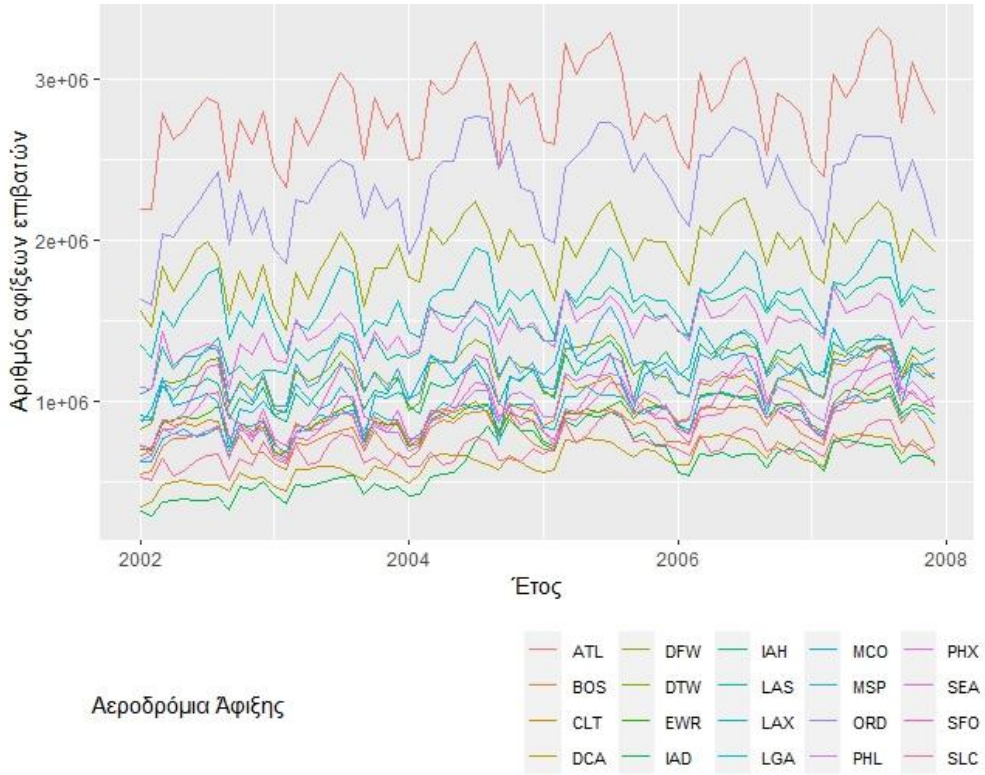
Πηγή: Επεξεργασία δική μου στο Rstudio

Αρχικά χρησιμοποιώντας το σύνολο των χρονοσειρών αφίξεων επιβατών για τα 20 αυτά αεροδρόμια γίνεται μία συνάθροιση του αριθμού των αφίξεων των επιβατών για το σύνολο των επιλεγμένων αεροδρομίων για να οπτικοποιηθεί η συνολική συμπεριφορά των ιστορικών δεδομένων. Από το σχήμα (Διάγραμμα 3.3) εξάγεται το

συμπέρασμα ότι τα συνολικά δεδομένα έχουν μία ανοδική τάση και από ότι διακρίνεται και θα εκτιμηθεί πιο σωστά παρακάτω, εποχικότητα, καθώς τα δεδομένα φαίνεται να ακολουθούν τις ίδιες διακυμάνσεις κάθε έτος στην ίδια περίοδο. Ο συνολικός αριθμός αφίξεων των επιβατών για όλα τα αεροδρόμια εμφανίζει χαμηλότερη τιμή της τάξης των 18.000.000 το έτος 2002 και φτάνει περίπου τα 29.000.000 το έτος 2007.

Στην επόμενη εικόνα (Διάγραμμα 3.4) αναπαρίστανται οι συνολικές αφίξεις των επιβατών ανά αεροδρόμιο και ανά μήνα για τα έτη 2002-2007 με τη χρήση κυματομορφών. Η απεικόνιση με διαφορετικά χρώματα δείχνει την αναλογία μεταξύ του αριθμού αφίξεων των επιβατών των διαφορετικών αεροδρομίων που ανήκουν όλα στην κατηγορία «μεγάλα αεροδρόμια».

Διάγραμμα 3.4 Διακύμανση αριθμού αφίξεων επιβατών ανά αεροδρόμιο 2002-2007



Πηγή: Επεξεργασία στο Rstudio

Στο Διάγραμμα 3.4 φαίνεται πως στα αεροδρόμια υπό εξέταση, η μηνιαία ροή επιβατών κυμαίνεται από περίπου 280.000 στο αεροδρόμιο IAD το 2002 μέχρι και 3.300.000 στο ATL το 2007, σε συμφωνία με τη στατιστική ανάλυση που έγινε προηγουμένως και απεικονίζεται στους παραπάνω πίνακες. Μία άλλη παρατήρηση που προκύπτει είναι ότι υπάρχει μια περιοδικότητα στα δεδομένα και μία αύξηση με την πάροδο των ετών, η οποία μας δημιουργεί υποψία για την ύπαρξη εποχικότητας και τάσης σε όλα τα υπό εξέταση αεροδρόμια.

3.4.3 Τάση Χρονοσειρών

Στην παράγραφο αυτή θα εξεταστεί η ύπαρξη τάσης στα δεδομένα ανά εξεταζόμενο αεροδρόμιο. Παρατηρώντας το σχήμα (Διάγραμμα 3.5), στο οποίο φαίνεται η απεικόνιση στον άξονα y του αριθμού των αφίξεων επιβατών ανά αεροδρόμιο και στον άξονα x το έτος για τα ιστορικά δεδομένα 2002-2007, σε συνδυασμό με τον πίνακα 3.4 όπου φαίνεται η κλίση των γραμμών και η στατιστική σημαντικότητα των παραμέτρων, διαπιστώνεται ξεκάθαρα πως στα δεδομένα υπάρχει μία γενική ανοδική τάση.

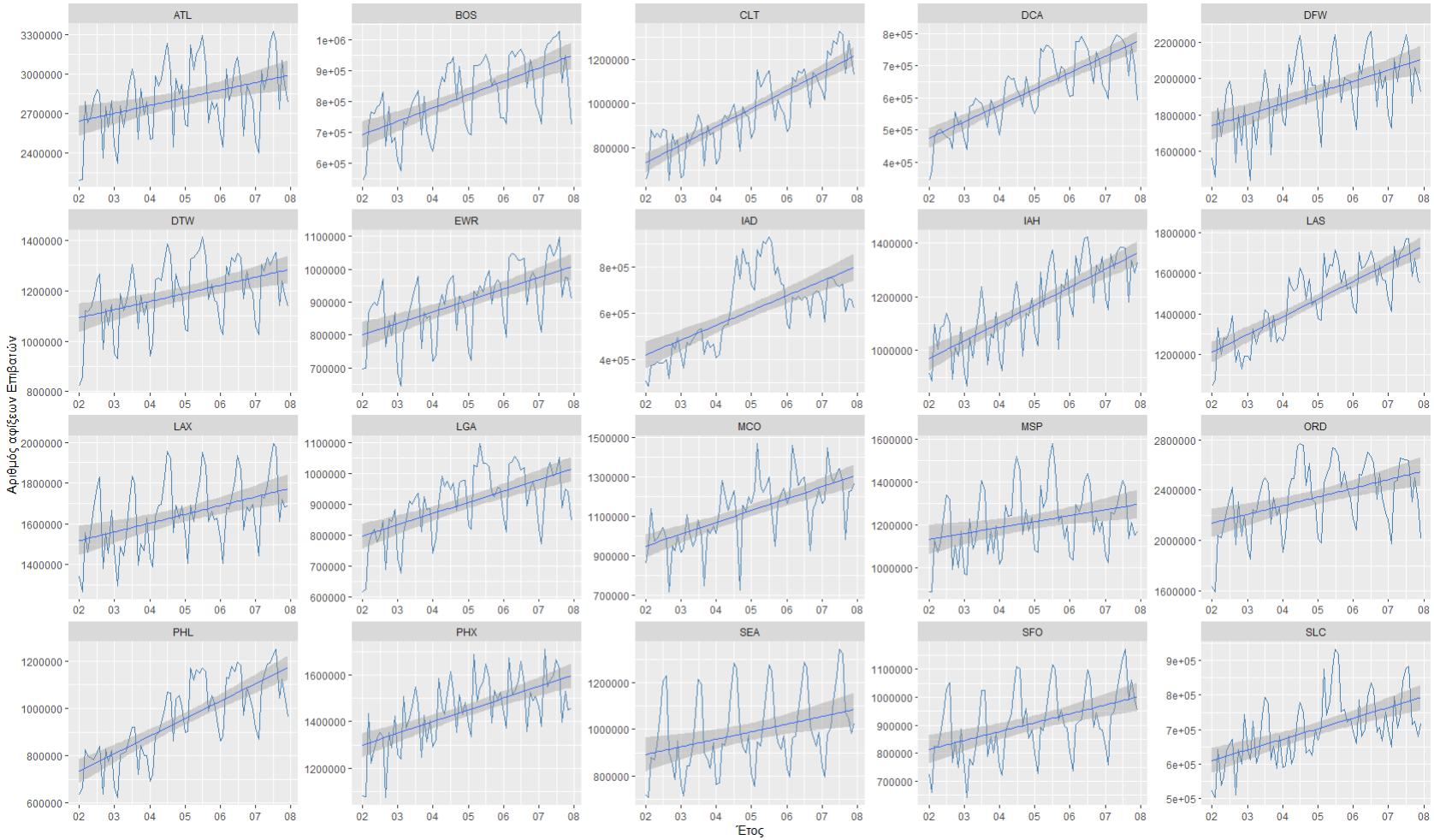
Πίνακας 3.4 Τιμές κλίσης των γραμμικών παλινδρομήσεων και σημαντικότητας των παραμέτρων

	destination	slope	p-value
1	ATL	160,17	0,00082
2	BOS	117,93	2,54E-09
3	CLT	224,28	2,60E-20
4	DCA	139,48	2,95E-17
5	DFW	168,17	6,13E-07
6	DTW	88,07	0,00018
7	EWR	95,36	5,71E-08
8	IAD	175,17	2,25E-10
9	IAH	182,2	1,71E-15
10	LAS	**238,30**	2,81E-18
11	LAX	117,56	0,000121
12	LGA	100,23	3,47E-08
13	MCO	164,09	1,16E-09
14	MSP	**75,81**	0,007112
15	ORD	188,63	8,75E-05
16	PHL	202,34	3,61E-15
17	PHX	137,72	3,97E-09
18	SEA	89,48	0,00309
19	SFO	86,75	7,12E-05
20	SLC	83,87	2,59E-07

Πηγή: Επεξεργασία στο Rstudio και στο Excel

Σε κάποια αεροδρόμια η τάση είναι μικρότερη όπως για παράδειγμα στο αεροδρόμιο SLC (κλίση 83,87) του Salt Lake City και σε κάποια άλλα μεγαλύτερη, όπως φαίνεται στο αεροδρόμιο του Chicago, ORD (κλίση 188,63). Η μεγαλύτερη κλίση είναι στο αεροδρόμιο LAS 238,30 και η μικρότερη στο MSP που είναι 75,81, υποδεικνύοντας πως σε όλα τα αεροδρόμια υπάρχει ανοδική τάση κατά το διάστημα 2002-2007.

Διάγραμμα 3.5 Αριθμός αφίξεων επιβατών ανά αεροδρόμιο για τα έτη 2002-2007



Πηγή: Επεξεργασία στο Rstudio

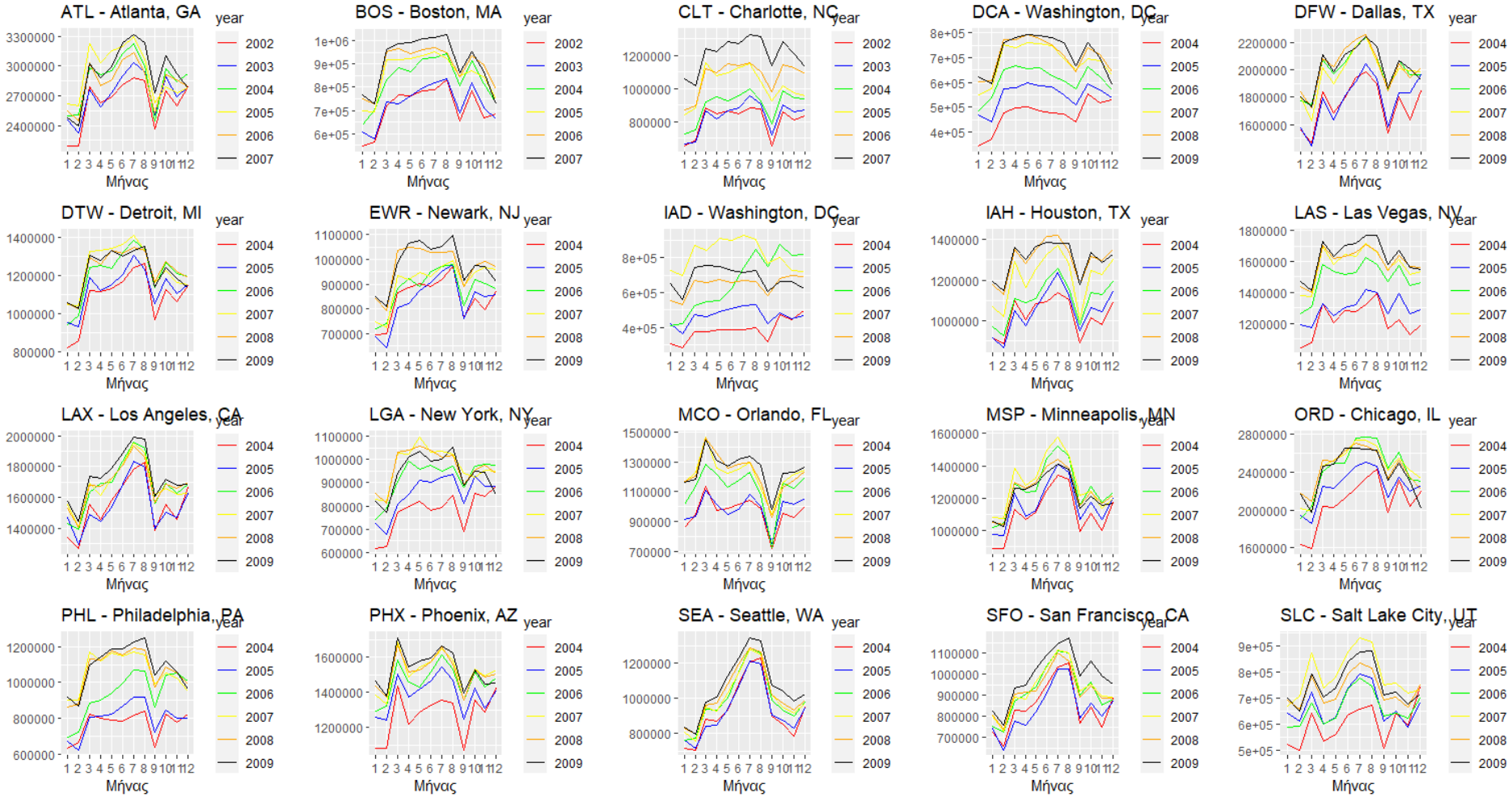
3.4.4 Εποχικότητα Χρονοσειρών

Σε αυτό το σημείο θα εξεταστεί η συνιστώσα της εποχικότητας για κάθε ένα από τα υπό εξέταση αεροδρόμια. Στην εικόνα παρακάτω (Διάγραμμα 3.6) απεικονίζονται ο αριθμός αφίξεων επιβατών ανά μήνα για κάθε έτος. Σε κάθε αεροδρόμιο παρουσιάζεται με διαφορετικό χρώμα ο αριθμός των επιβατών ανά έτος. Τα 20 αεροδρόμια εμφανίζονται ανά πέντε σε αλφαβητική σειρά και το κάθε ένα με τη δική του κλίμακα μέτρησης.

Η ύπαρξη εποχικότητας στα δεδομένα για το κάθε αεροδρόμιο είναι αισθητή με απλή παρατήρηση στο Διάγραμμα 3.6 . Οι διαφορές μεταξύ των αεροδρομίων έγκεινται πρώτον στην περίοδο που εμφανίζει το κάθε ένα από αυτά εποχικότητα, αν και στα περισσότερα ταυτίζεται, και στην ένταση του φαινομένου. Σε κάποια από τα αεροδρόμια η αύξηση ή η μείωση του αριθμού των αφίξεων των επιβατών είναι πολύ μικρή σε σχέση με τη συνηθισμένη κίνηση των αεροδρομίων π.χ. BOS, CLT, DCA κ.ά και σε κάποια άλλα είναι έκδηλη αυτή η διαφοροποίηση του αριθμού των επιβατών ανά εποχή π.χ. ATL, ORD, DFW, LAX, τα οποία είναι και τα αεροδρόμια με την μεγαλύτερη ροή επιβατών.

Εκτός από τη γενική παρατήρηση ότι σε όλα τα αεροδρόμια παρουσιάζεται εποχικότητα σε μεγαλύτερο ή μικρότερο βαθμό, επίσης αν παρατηρήσει κανείς τα παρακάτω διαγράμματα μπορεί να δει ότι σε όλα τα αεροδρόμια τον Σεπτέμβριο για όλα τα εξεταζόμενα έτη υπάρχει πτώση στις αφίξεις επιβατών. Επιπλέον φαίνεται να υπάρχουν ομοιότητες στις μεταβολές του αριθμού των αφίξεων των επιβατών σε κάποια από τα υπό μελέτη αεροδρόμια όπως π.χ. LAS - IAH, LGW - EWR, SEA - SFO κ.ά., που πολύ πιθανό να συνδέονται με τη θέση των αεροδρομίων στο χάρτη, δηλαδή σε σχέση με τις κλιματολογικές συνθήκες.

Διάγραμμα 3.6 Συνιστώσα εποχικότητας ανά αεροδρόμιο για τα έτη 2002-2007



Πηγή: Επεξεργασία Rstudio

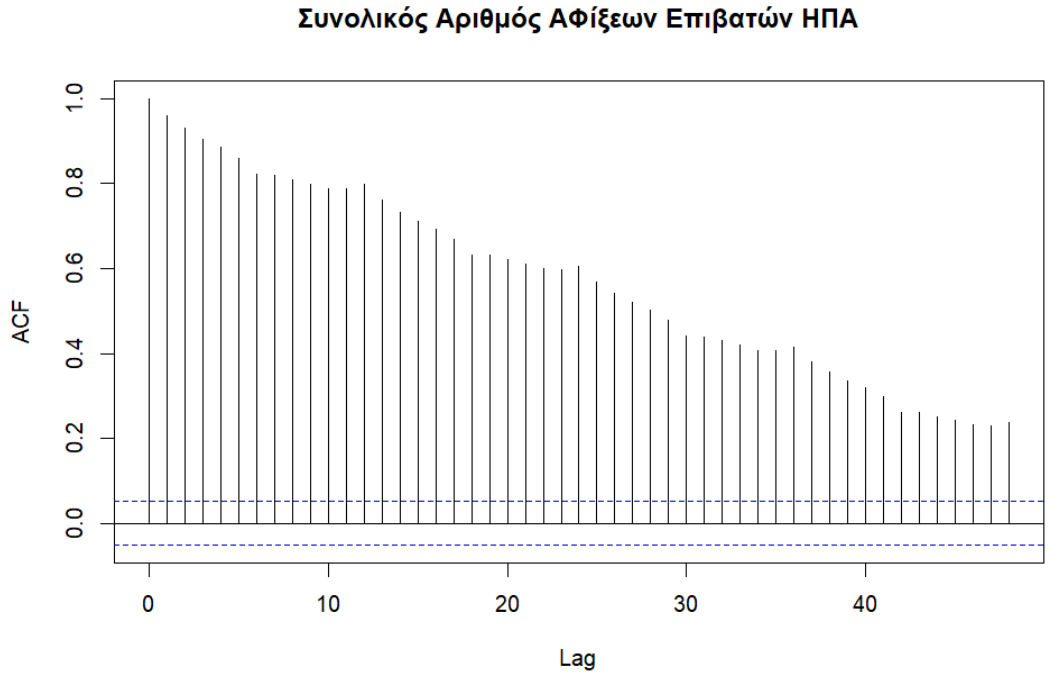
3.4.5 Τελική Επιλογή Εξωγενών Μεταβλητών

Μετά την ανάλυση που έγινε παραπάνω και με εμφανή την τάση και την περιοδικότητα στις χρονοσειρές αποφασίστηκε η χρήση επιπλέον μεταβλητών, οι οποίες στη λογική της αυτοσυσχέτισης, περιλαμβάνουν υστερήσεις του αριθμού των αφίξεων των επιβατών.

Όταν στα δεδομένα παρουσιάζεται τάση, τότε οι αυτοσυσχετίσεις (Autocorrelation Function – ACF) τείνουν να είναι μεγάλες και θετικές. Όταν τα δεδομένα παρουσιάζουν εποχικότητα τότε πάλι οι αυτοσυσχετίσεις θα είναι μεγαλύτερες για τις εποχιακές υστερήσεις. Όταν έχουμε όμως συνδυασμό και των δύο φαινομένων τότε οι αυτοσυσχετίσεις θα είναι θετικές και τις κοντινές χρονικές στιγμές και στα πολλαπλάσια της εποχιακής συχνότητας (Hyndman & Athanasopoulos, 2018).

Στο διάγραμμα παρακάτω παρουσιάζεται η αυτοσυσχέτιση της ροής των επιβατών που φτάνουν και στα 20 αεροδρόμια που εξετάζονται σε αυτή την εργασία σε σχέση με υστέρηση 48 μηνών. Η αργή μείωση που παρατηρείται στο διάγραμμα οφείλεται στην τάση ενώ η ελαφρώς κυματιστή όψη στην εποχικότητα

Διάγραμμα 3.7 Αυτοσυσχέτιση ροής επιβατών με υστερήσεις (max 48 μήνες)



Πηγή: Επεξεργασία στο Rstudio

Σύμφωνα λοιπόν με τα παραπάνω προστίθενται δύο επιπλέον μεταβλητές. Η πρώτη αφορά την υστέρηση κατά ένα χρόνο (12 μήνες) και η άλλη κατά δύο χρόνια (24 μήνες). Η λογική πίσω από την επιλογή αυτή είναι ότι αυτές οι δύο μεταβλητές θα

βοηθήσουν τα επιλεγόμενα μοντέλα να κάνουν ακριβέστερες προβλέψεις του αριθμού των επιβατών καθώς θα εμπλέξουν τη συσχέτιση των προβλέψεων των αφίξεων με τα ιστορικά δεδομένα των αφίξεων των προηγούμενων δύο ετών.

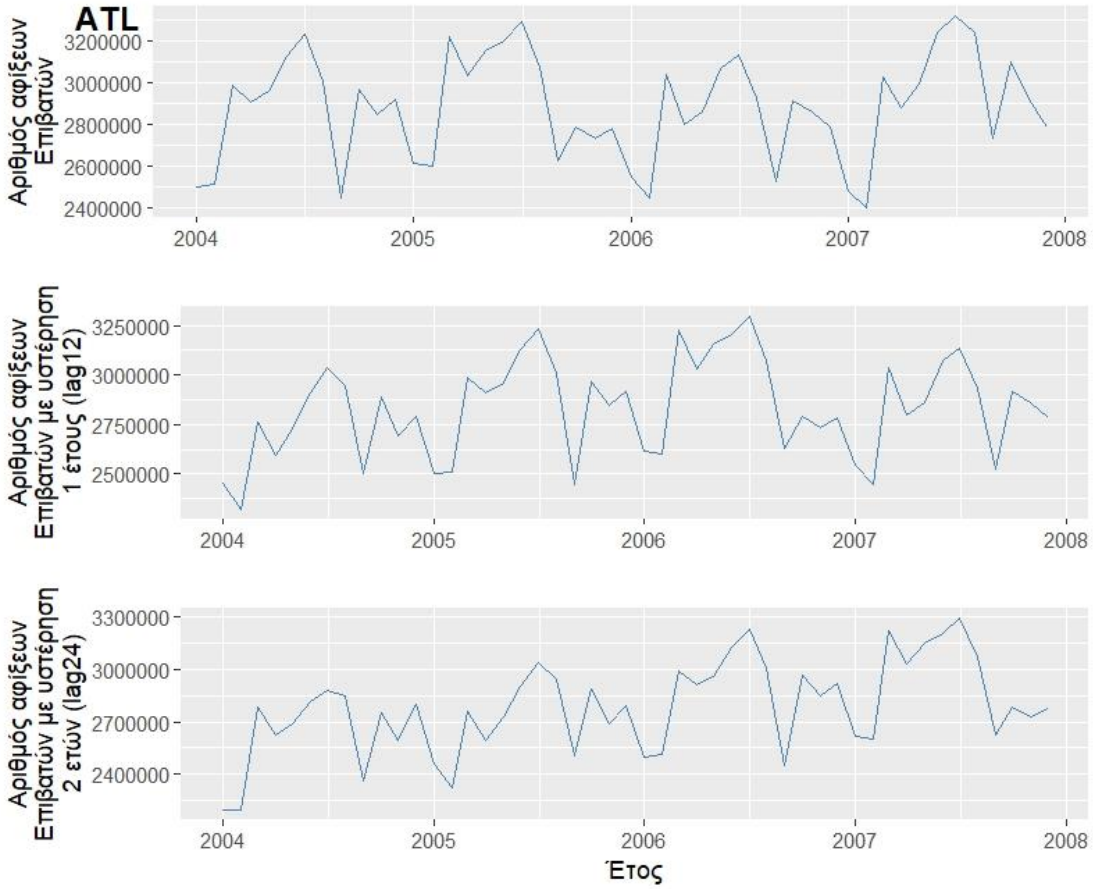
Οι επιπλέον μεταβλητές που δημιουργούνται συνεπώς είναι οι εξής:

Lag12: υστέρηση δεδομένων ενός έτους στον αριθμό των αφίξεων επιβατών

Lag24: υστέρηση δεδομένων δύο ετών στον αριθμό των αφίξεων επιβατών

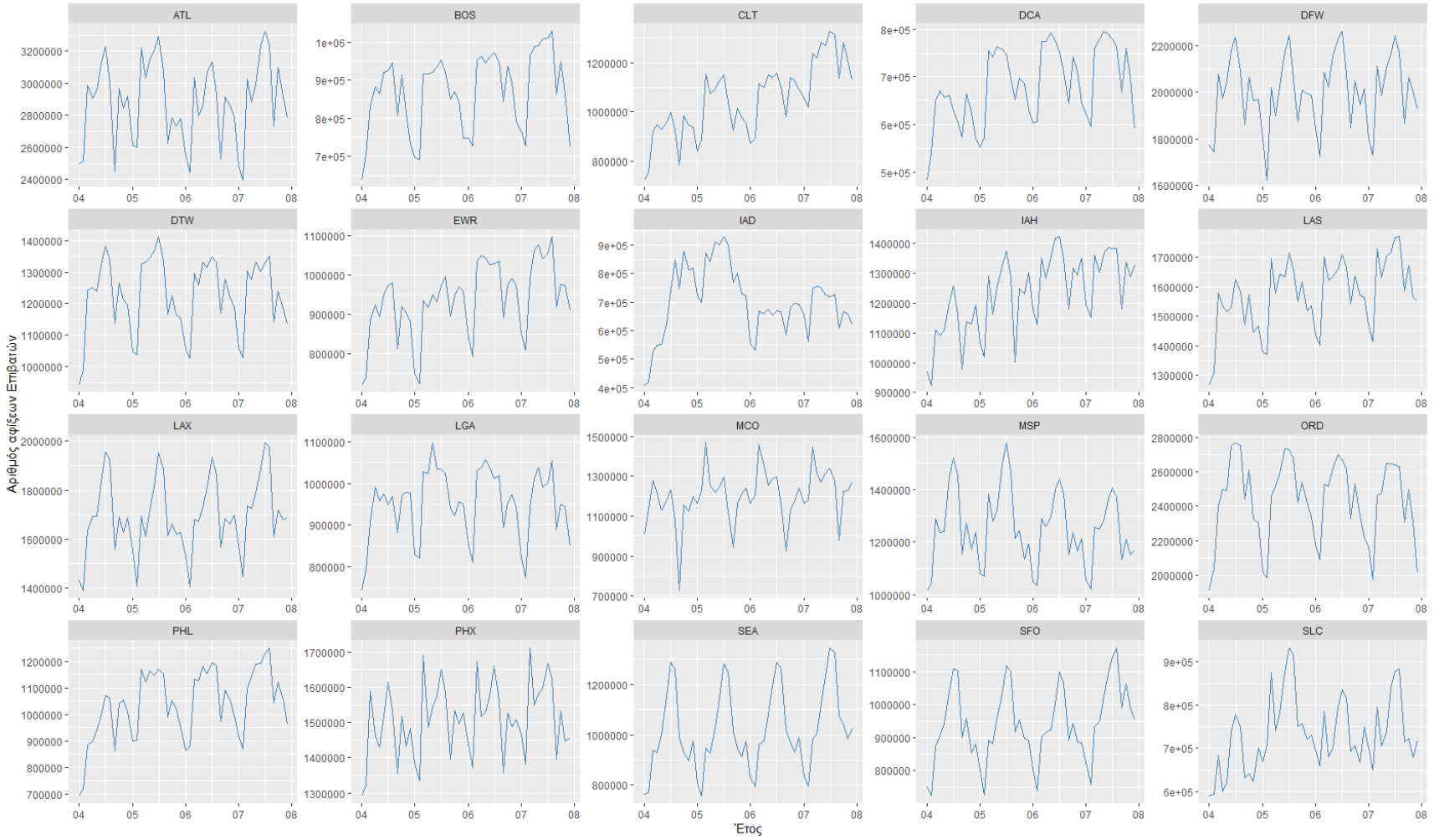
Η προσθήκη των δεδομένων αυτών των δύο μεταβλητών στο σύνολο των δεδομένων που χρησιμοποιούνται δημιουργεί κενές τιμές για το 2002 και 2003 για κάθε αεροδρόμιο οπότε επιλέγεται η εξαίρεση αυτών των δύο ετών από το σετ εκπαίδευσης (train set) οπότε ο τελικός αριθμός των μηνιαίων παρατηρήσεων που χρησιμοποιούνται στην πειραματική διαδικασία είναι 1440 και αφορούν τα έτη 2004-2009.

Διάγραμμα 3.8 Αριθμός των αφίξεων των επιβατών, της υστέρησης κατά ένα έτος και κατά δύο έτη στο αεροδρόμιο της Ατλάντα (ATL) για τα έτη 2004-2007



Πηγη: Επεξεργασία στο Rstudio

Διάγραμμα 3.9 Γραφική αναπαράσταση των ιστορικών δεδομένων ανά αεροδρόμιο για τα έτη 2004-2007



Πηγή: Επεξεργασία στο Rstudio

Στο σχήμα παραπάνω (Διάγραμμα 3.8) φαίνεται αυτή η σχέση του αρχικού αριθμού αφίξεων σε συνάρτηση με τις υστερήσεις ενός και δυο ετών (lag12 και lag24) για το αεροδρόμιο της Ατλάντα. Από το σχήμα είναι προφανές γιατί επιλέχθηκαν οι υστερήσεις να χρησιμοποιηθούν ως μεταβλητές στα μοντέλα πρόβλεψης.

Η εμπειρική συνθήκη ότι τα ιστορικά δεδομένα είναι τετραπλάσια της πρόβλεψης εξακολουθεί να ισχύει ώστε να εξασφαλιστεί ότι το μοντέλο θα μπορέσει να μάθει τη σχέση των εποχικοτήτων.

Οπότε στο διάγραμμα παραπάνω (Διάγραμμα 3.9) απεικονίζεται το τελικό σύνολο δεδομένων, που απαρτίζεται από 20 αεροδρόμια με μηνιαία δεδομένα για 4 έτη (2004-2007), ήτοι 20 χρονοσειρές με 960 μηνιαίες παρατηρήσεις, που θα χρησιμοποιηθούν στην πειραματική διαδικασία όπως περιγράφεται στο επόμενο κεφάλαιο, δηλαδή ο συνολικός αριθμός των αφίξεων επιβατών ανά αεροδρόμιο για το διάστημα 2004 έως και 2007.

Κεφάλαιο 4. Πειραματική Διαδικασία

Στο κεφάλαιο αυτό περιγράφεται η διαδικασία επιλογής των μοντέλων πρόβλεψης βασιζόμενη στα χαρακτηριστικά των χρονοσειρών που επιλέχθηκαν στο κεφάλαιο 3. Επίσης γίνεται εφαρμογή των μοντέλων πρόβλεψης για όλες τις μεταβλητές παλινδρόμησης και των μοντέλων πρόβλεψης μηχανικής μάθησης και παρουσίαση των αποτελεσμάτων για κάθε ένα ξεχωριστά. Τέλος, γίνεται μία συγκριτική αξιολόγηση των μεθόδων.

4.1 Μεθοδολογική Προσέγγιση

Η επιλογή των μοντέλων είναι μια διαδικασία η οποία απαιτεί κάποιου είδους ανάλυση. Βασίζεται συχνά στα χαρακτηριστικά των χρονοσειρών που εξετάζονται, ούτως ώστε οι προβλέψεις των μελλοντικών αποτελεσμάτων να είναι ακριβείς, δηλαδή να παρουσιάζουν το ελάχιστο δυνατό σφάλμα.

Όπως αναφέρθηκε στο κεφάλαιο 2 υπάρχουν πολλοί τρόποι υπολογισμού του σφάλματος ανάλογα το μέγεθος των δεδομένων, το είδος των δεδομένων κ.τ.λ., όμως όποιο μέτρο κι αν επιλεχθεί στόχος είναι η δημιουργία ή η επιλογή εκείνου του μοντέλου που θα το κάνει ελάχιστο.

Στην βιβλιογραφία παρουσιάζεται η προσπάθεια των ερευνητών να ορίσουν κάποιες παραμέτρους ώστε να επιλέγεται η καλύτερη μέθοδος πρόβλεψης για κάθε σύνολο δεδομένων ή πιο συγκεκριμένα για κάθε χρονοσειρά. Ιδανικότερα θα ήταν να δημιουργηθεί μία γενικότερη μεθοδολογία, η οποία λαμβάνοντας υπόψιν όλες τις παραμέτρους των συνόλων των δεδομένων, να παρέχει το κατάλληλο εργαλείο πρόβλεψης μειώνοντας το «εκτός δείγματος» σφάλμα ώστε να αποτελέσει σημαντικό εργαλείο στις προβλέψεις και κατ’ επέκταση στη λήψη αποφάσεων.

4.1.1 Επιλογή Μοντέλων Πρόβλεψης

Βάσει λοιπόν της πιο πάνω θεωρίας παρουσιάζονται παρακάτω τα χαρακτηριστικά των χρονοσειρών που θα χρησιμοποιηθούν στην πρόβλεψη της ροής των επιβατών στα επιλεγμένα αεροδρόμια-προορισμούς, ώστε να αιτιολογηθεί η επιλογή των μοντέλων πρόβλεψης που αναλύονται παρακάτω.

Σκοπός όπως έχει γίνει φανερό μέχρι τώρα είναι η πρόβλεψη του αριθμού των επιβατών που φτάνουν σε συγκεκριμένα αεροδρόμια (Πίνακας 3.1 Αεροδρόμια άφιξης υπό

εξέταση και μέσος αριθμός επιβατών και πτήσεων για τα έτη 2002-2007Πίνακας 3.1). Συνοψίζοντας τα στοιχεία που έχουν μελετηθεί μέχρι τώρα, τα ιστορικά δεδομένα είναι μηνιαία και καλύπτουν τα έτη 2004-2007. Όπως διαπιστώθηκε στο προηγούμενο κεφάλαιο χαρακτηρίζονται από εποχικότητα και τάση, με πιο έντονο στο στοιχείο της εποχικότητας. Επίσης η συχνότητα της πρόβλεψης είναι ετήσια, δηλαδή η πρόβλεψη είναι μεσοπρόθεσμη.

Σύμφωνα με εμπειρικές μελέτες και ορμώμενοι από την επιθυμία της χρήσης τουλάχιστον μίας στατιστικής μεθόδου και μίας τουλάχιστον μηχανικής μάθησης, διότι πέρα από τον στόχο της ακρίβειας στην πρόβλεψη, ερευνητικό ενδιαφέρον έχει και η σύγκριση των δύο κατηγοριών αλγορίθμων, επιλέχθηκαν για τα δεδομένα οι τρεις παρακάτω αλγόριθμοι πρόβλεψης:

- Γραμμική Παλινδρόμηση (LR)
- Τυχαία Δάση (RF)
- Σταδιακή Ενίσχυση Δέντρων (GBT)

4.1.2 Επιλογή Εξωγενών Μεταβλητών Πρόβλεψης

Όπως αναφέρουν οι (Hyndman & Athanasopoulos, 2018) η επιλογή των παραγόντων πρόβλεψης απαιτεί αρκετή σκέψη και έναν στρατηγικό σχεδιασμό. Υπάρχουν κάποιες λογικές που θα πρέπει να αποφευχθούν και κάποιες που θα πρέπει να υιοθετηθούν έτσι ώστε να εξασφαλιστεί η καλύτερη δυνατή επιλογή των παραμέτρων πρόβλεψης για τα μοντέλα που χρησιμοποιούνται στην πειραματική διαδικασία.

Κάποιες από τις λογικές που δεν συνίσταται αναφέρονται περιληπτικά παρακάτω:

- Οι παράγοντες πρόβλεψης δεν θα πρέπει να εξαιρούνται από τα μοντέλα επειδή ατομικά δεν έχουν αξιοσημείωτη σχέση ως προς τη μεταβλητή πρόβλεψης, διότι δεν μπορεί να ελεγχθεί πάντα η επίδραση που μπορεί να έχει αυτός ο παράγοντας στις υπόλοιπες μεταβλητές του μοντέλου.
- Σε μία προγνωστική διαδικασία πολλαπλής γραμμικής παλινδρόμησης δεν είναι σωστό να εξαιρούνται κάποιες μεταβλητές επειδή δεν είναι στατιστικά σημαντικές, δηλαδή το $p>0.05$, διότι η στατιστική σημαντικότητα δεν δηλώνει πάντα προγνωστική αξία.

Επίσης κάποιες λογικές που θα πρέπει να υιοθετηθούν παρουσιάζονται παρακάτω:

- Προτείνεται η χρήση της προσαρμοσμένης R^2 ως μέτρο της επεξηγηματικότατας των μεταβλητών, δηλαδή κατά πόσο μπορεί να ερμηνεύσει η κάθε ανεξάρτητη μεταβλητή το φαινόμενο.
- Προτείνεται πάντα η διασταυρούμενη επικύρωση, δηλαδή η χρήση των δεδομένων εκπαίδευσης για την εκτίμηση οποιωνδήποτε παραγόντων μιας μεθόδου πρόβλεψης και των δεδομένων αξιολόγησης για την αξιολόγηση της ακρίβειας της.

Πίνακας 4.1 Παράμετροι και χαρακτηριστικά δεδομένων

ΜΕΤΑΒΛΗΤΕΣ				
month	**year**	**lag12**	**lag24**	**sum_seats**
ΔΕΔΟΜΕΝΑ				
dataset	years	frequency	airports	timeseries
data	2004-2009	monthly	20	20
train set	2004-2007			
validation set	2008			
test set	2009			

Πηγή: Επεξεργασία στο excel

Οι παραπάνω μεταβλητές χρησιμοποιούνται στα μοντέλα απλής γραμμικής παλινδρόμησης, πέντε συνδυασμοί αυτών παρουσιάζονται στην πολλαπλή γραμμική παλινδρόμηση, όπως θα εξηγηθεί παρακάτω. Στα μοντέλα μηχανικής μάθησης χρησιμοποιούνται και οι πέντε παράμετροι για τη δημιουργία των δυο αλγορίθμων RF και GBT και σε συνδυασμό με τον καθορισμό των αντίστοιχων υπερπαραμέτρων σε κάθε μοντέλο, γίνεται προσπάθεια να ενισχυθεί η προβλεπτική ικανότητα του μοντέλου.

4.1.3 Επιλογή Μετρικής Αξιολόγησης Προβλέψεων

Από τα κριτήρια αξιολόγησης της ακρίβειας της πρόβλεψης που αναλύθηκαν στο θεωρητικό μέρος επιλέγεται το συμμετρικό μέσο απόλυτο ποσοστό σφάλματος (sMAPE), καθώς έχει το πλεονέκτημα της μέτρησης της ποσοστιαίας διαφορά μεταξύ πραγματικών και προβλεπόμενων τιμών, κάτι που βοηθάει στην σύγκριση σφαλμάτων πρόβλεψης αεροδρομίων με διαφορετικές τάξεις μεγέθους στα δεδομένα.

4.2 Μοντέλα Γραμμικής Παλινδρόμησης

Μία από τις πιο διαδεδομένες και αποτελεσματικές στατιστικές μεθόδους πρόβλεψης, όπως παρουσιάστηκε στο κεφάλαιο 2, είναι η Γραμμική Παλινδρόμηση.

Στην πειραματική διαδικασία, αρχικά, αναλύονται τα μοντέλα των πέντε απλών γραμμικών παλινδρομήσεων που χρησιμοποιούν ως μεταβλητές εισόδου αυτές που αναφέρονται στον παραπάνω πίνακα και πέντε παλινδρομήσεων με συνδυασμό των παραπάνω παραγόντων, με στόχο την πρόβλεψη του αριθμού των αφίξεων επιβατών ανά αεροδρόμιο.

Στη συνέχεια παρουσιάζονται τα αποτελέσματα των μοντέλων πρόβλεψης που επιλέχθηκαν για τη γραμμική παλινδρόμηση (απλή και πολλαπλή) και γίνεται αξιολόγηση των μοντέλων πρόβλεψης της συνολικής ροής των επιβατών λαμβάνοντας υπόψη το σφάλμα για το κάθε αεροδρόμιο χωριστά και τον μέσο όρο των σφαλμάτων όλων των αεροδρομίων ανά παράμετρο πρόβλεψης.

4.2.1 Απλή Γραμμική Παλινδρόμηση

Το βασικό μοντέλο της γραμμικής παλινδρόμησης παρουσιάστηκε στο δεύτερο κεφάλαιο και πάνω σε αυτό σχεδιάστηκαν τα μοντέλα πρόβλεψης που αναλύονται παρακάτω. Οι συναρτήσεις που δημιουργήθηκαν είναι:

$$sum\_passengers = b_0 + b_1\, month + e \quad (4.1)$$

$$sum\_passengers = b_0 + b_1\, year + e \quad (4.2)$$

$$sum\_passengers = b_0 + b_1\, lag12 + e \quad (4.3)$$

$$sum\_passengers = b_0 + b_1\, lag24 + e \quad (4.4)$$

$$sum\_passengers = b_0 + b_1\, sum\_seats + e \quad (4.5)$$

Με τις παραπάνω συναρτήσεις ελέγχονται όλες οι παράμετροι παλινδρόμησης των απλών γραμμικών μοντέλων για κάθε ένα από τα 20 επιλεγμένα αεροδρόμια. Οι μεταβλητές αυτές είναι:

- Μεταβλητή μήνα (month): η μεταβλητή αυτή είναι κατηγορική και οι τιμές που μπορεί να πάρει είναι 12 πιθανές στάθμες, που αντιπροσωπεύουν τους μήνες του έτους.

- Μεταβλητή έτους (year): η μεταβλητή αυτή είναι ποσοτική και έχει τέσσερις διακριτές τιμές για τα έτη 2004-2007.
- Μεταβλητή υστέρησης ενός έτους (lag12): είναι η μεταβλητή που βάζει τον παράγοντα αυτοσυσχέτισης στο μοντέλο κατά ένα έτος, εισάγει δηλαδή ως ανεξάρτητη μεταβλητή στην παλινδρόμηση τον αριθμό των επιβατών του προηγούμενου έτους τον ίδιο μήνα για το ίδιο αεροδρόμιο.
- Μεταβλητή υστέρησης δύο ετών (lag24): είναι η μεταβλητή που βάζει τον παράγοντα αυτοσυσχέτισης στο μοντέλο κατά δύο έτη, εισάγει δηλαδή ως ανεξάρτητη μεταβλητή στην παλινδρόμηση τον αριθμό των επιβατών του προ προηγούμενου έτους τον ίδιο μήνα για το ίδιο αεροδρόμιο.

Πίνακας 4.2 Τιμή προσαρμοσμένου R^2 τις απλές γραμμικές παλινδρομήσεις

	ΜΕΤΑΒΛΗΤΕΣ				
Τιμή Προσαρμοσμένου R2 ανά αεροδρόμιο άφιξης	**month**	**year**	**lag12**	**lag24**	**sum_seats**
Ατλάντα ATL	**0,84**	-0,02	0,66	0,56	0,48
Βοστώνη BOS	**0,83**	0,07	0,81	0,79	0,63
Σαρλοτ CLT	0,20	0,55	**0,83**	**0,83**	0,60
Ουάσιγκτον Ρόναλντ Ρέιγκαν DCA	0,57	0,22	**0,82**	0,69	0,60
Ντάλας DFW	**0,94**	-0,02	0,62	0,66	0,30
Ντιτρόιτ DTW	**0,92**	-0,02	0,81	0,71	0,32
Νιού Αρκ EWR	0,65	0,16	**0,83**	0,77	0,38
Ουάσιγκτον IAD	0,08	-0,01	-0,01	-0,02	0,78
Χιούστον IAH	0,42	0,35	0,87	0,76	**0,90**
Λας Βέγκας LAS	**0,71**	0,16	0,76	0,63	0,70
Λος Άντζελες LAX	**0,94**	0,00	0,80	0,83	0,73
Νιού Γιορκ LGA	**0,78**	-0,02	0,63	0,42	0,40
Ορλάντο MCO	0,75	0,09	0,76	0,63	**0,78**
Μιννεάπολις MSP	**0,92**	-0,01	0,80	0,59	0,45
Σικάγο ORD	**0,90**	-0,02	0,75	0,51	0,48
Φιλαδέλφεια PHL	0,57	0,14	**0,67**	0,51	0,48
Φοίνιξ PHX	**0,84**	0,04	0,86	0,68	0,40
Σιάτλ SEA	0,95	0,01	**0,97**	**0,97**	0,90
Σαν Φρανσίσκο SFO	**0,89**	0,02	0,81	0,85	0,71
Σολτ Λέικ SLC	0,55	0,06	0,36	0,56	**0,65**
AVERAGE	**0,71**	**0,09**	**0,72**	**0,65**	**0,58**

Πηγή: Επεξεργασία στο Rstudio και στο Excel

- Μεταβλητή του αριθμού των θέσεων του αεροπλάνου (sum_seats): είναι ποσοτική μεταβλητή και συνδέεται άμεσα με τον μέγιστο αριθμό επιβατών που μπορούν να ταξιδέψουν σε ένα αεροπλάνο, δίνοντας και τη χωρητικότητα του.

Η εφαρμογή αρχικά της απλής γραμμικής παλινδρόμησης έχει διττή σημασία. Πρώτον με αυτόν τον τρόπο δημιουργήθηκε ένας δείκτης σημαντικότητας για την ακρίβεια της πρόβλεψης για κάθε ανεξάρτητη μεταβλητή χωριστά και δεύτερον, ως συνέπεια του πρώτου, αποτέλεσε οδηγό για τη δημιουργία συνδυασμού ανεξάρτητων μεταβλητών, που ενδεχομένως, θα μπορούσαν να βελτιώσουν την προβλεπτική ικανότητα των νέων μοντέλων.

Ένα μέτρο, που όπως εξηγήθηκε, βοηθάει στην επιλογή του καλύτερου προγνωστικού παράγοντα είναι ο δείκτης του προσαρμοσμένου R^2. Μελετώντας τον παραπάνω πίνακα (Πίνακας 4.2) σε σχέση με τον δείκτη διαπιστώνεται ότι κατά μέσο όρο υψηλότερο δείκτη έχει η μεταβλητή lag12 και η month, στη συνέχεια η lag24 και η sum_seats.

Η μεταβλητή του έτους φαίνεται να μην είναι καλός προγνωστικός παράγοντας, τουλάχιστον όχι από μόνος του. Σε αντίθεση με R^2 το προσαρμοσμένο R^2 μπορεί να πάρει αρνητικές τιμές και αυτό συμβαίνει όταν η τιμή του R^2 είναι κοντά στο μηδέν και δείχνει ότι ο επιλεγμένος παράγοντας δεν συνεισφέρει στην πρόβλεψη. Η εξήγηση για την περιορισμένη προβλεπτική ικανότητα του παράγοντα year ίσως είναι ο μικρός αριθμός τιμών (4 ανά αεροδρόμιο) και όπως θα φανεί στη συνέχεια ο συνδυασμός του με τον μήνα θα δώσει πολύ καλύτερες τιμές της μεταβλητής. Επίσης η λογική λέει ότι η προγνωστική ικανότητα του sum_seats θα έπρεπε να είναι ακόμα μεγαλύτερη αλλά προφανώς στα αεροδρόμια με έντονο το φαινόμενο της εποχικότητας από μόνος του ο παράγοντας έχει απλά μία μέτρια προβλεπτική ικανότητα.

Στη συνέχεια κατασκευάζονται τα μοντέλα πρόβλεψης που χρησιμοποιούν τις παραπάνω εξισώσεις παλινδρόμησης για να προβλέψουν τις τιμές για το έτος 2008, το οποίο όπως ειπώθηκε, χρησιμοποιείται ως μέτρο σύγκρισης της προβλεπτικής ικανότητας του μοντέλου (validation set).

Με μια ματιά (Πίνακας 4.3) αυτό που διαπιστώνεται είναι πως κατά μέσο όρο το μοντέλο που χρησιμοποιεί τον αριθμό των θέσεων των επιβατών ως ανεξάρτητη μεταβλητή εμφανίζει το μικρότερο σφάλμα πρόβλεψης, του οποίου η τιμή είναι

περίπου 4,8%, στη συνέχεια ακολουθεί η παράμετρος month και τέλος η υστέρηση κατά ένα έτος lag12.

Πίνακας 4.3 Σφάλματα πρόβλεψης αφίξεων επιβατών για το έτος 2008

ΑΕΡΟΔΡΟΜΙΑ ΑΦΙΞΗΣ	ΜΕΤΑΒΛΗΤΕΣ				
	month	year	lag12	lag24	sum_seats
	sMAPE1	sMAPE2	sMAPE3	sMAPE4	sMAPE5
Ατλάντα ATL	**2,93**	6,35	1,69	2,98	5,88
Βοστώνη BOS	**3,94**	12,42	10,95	13,40	6,30
Σαρλοτ CLT	17,78	5,97	3,83	**3,77**	5,91
Ουάσιγκτον Ρόναλντ Ρέιγκαν DCA	**2,47**	11,31	8,03	11,51	4,18
Ντάλας DFW	5,36	8,23	7,30	9,52	**4,37**
Ντιτρόιτ DTW	**4,58**	8,79	4,96	7,57	6,03
Νιού Αρκ EWR	**4,20**	10,73	8,08	10,82	4,30
Ουάσιγκτον IAD	8,42	10,10	10,17	10,56	**5,90**
Χιούστον IAH	6,02	12,87	8,69	13,39	**3,94**
Λας Βέγκας LAS	6,44	10,59	9,39	10,18	**3,00**
Λος Άντζελες LAX	**4,57**	9,79	7,47	6,68	4,88
Νιού Γιορκ LGA	9,79	11,19	10,58	14,46	**5,03**
Ορλάντο MCO	4,56	10,41	7,77	7,87	**4,16**
Μιννεάπολις MSP	6,76	9,87	**5,73**	8,98	5,81
Σικάγο ORD	10,30	11,02	10,46	14,01	**3,98**
Φιλαδέλφεια PHL	**5,12**	10,14	5,07	7,03	7,84
Φοίνιξ PHX	5,42	9,23	8,01	8,84	**3,63**
Σιάτλ SEA	5,93	12,69	4,12	**2,64**	4,03
Σαν Φρανσίσκο SFO	12,76	11,87	5,83	9,04	**3,18**
Σολτ Λέικ SLC	5,23	12,57	7,21	8,49	**3,35**
AVERAGE	**6,63**	**10,31**	**7,27**	**9,09**	**4,79**

Πηγή: Επεξεργασία στο Rstudio και στο Excel

Επίσης διαπιστώνεται ότι το σφάλμα για το sum_seats, ανεξάρτητα από το αεροδρόμιο αναφοράς, είναι σχετικά σταθερό και κυμαίνεται από 3%-7.8% σε αντίθεση με τις υπόλοιπες μεταβλητές. Επειδή το R^2 δεν δουλεύει καλά με διαφορετικής κλίμακας δεδομένα, είναι πολύ πιθανό για αυτόν τον λόγο η επεξηγηματική ικανότητα του sum_seats να είναι αρκετά μικρότερη σε σχέση με το month ή τα lag12 και lag24. Ενώ στο σφάλμα της πρόβλεψης των μοντέλων που χρησιμοποιείται, το sMAPE, το οποίο

είναι ανεξάρτητο της κλίμακας των εξεταζόμενων δεδομένων, το month ηγείται στην προβλεπτική ικανότητα και έπονται τα month, lag12 και lag24.

Το επόμενο κρίσιμο βήμα στη διαδικασία των προβλέψεων είναι η διασταυρούμενη επικύρωση (cross validation), για αυτόν τον σκοπό χρησιμοποιείται το σύνολο δεδομένων του έτους 2009 ως δεδομένα επαλήθευσης.

Πίνακας 4.4 Σφάλματα πρόβλεψης αφίξεων επιβατών για το έτος 2009

ΑΕΡΟΔΡΟΜΙΑ ΑΦΙΞΗΣ	**ΜΕΤΑΒΛΗΤΕΣ**				
	month	**year**	**lag12**	**lag24**	**sum_seats**
	sMAPE1	**sMAPE2**	**sMAPE3**	**sMAPE4**	**sMAPE5**
Ατλάντα ATL	**2,53**	4,09	2,39	2,54	3,84
Βοστώνη BOS	**6,93**	12,43	10,78	13,55	7,54
Σαρλοτ CLT	19,14	5,77	3,53	**3,46**	5,71
Ουάσιγκτον Ρόναλντ Ρέιγκαν DCA	**2,58**	9,61	5,99	9,84	3,02
Ντάλας DFW	5,64	7,69	7,02	8,72	**4,95**
Ντιτρόιτ DTW	3,99	2,56	3,65	**2,33**	2,94
Νιού Αρκ EWR	3,70	9,60	4,73	8,01	**3,68**
Ουάσιγκτον IAD	6,56	7,79	7,84	8,18	**5,20**
Χιούστον IAH	6,38	9,93	7,05	10,42	**6,18**
Λας Βέγκας LAS	5,35	5,23	**5,03**	5,15	6,52
Λος Άντζελες LAX	**6,30**	8,27	7,09	6,81	6,32
Νιού Γιορκ LGA	9,08	9,69	9,37	12,15	**7,82**
Ορλάντο MCO	4,65	5,27	**4,46**	**4,46**	4,79
Μιννεάπολις MSP	4,02	6,24	**3,74**	5,58	**3,74**
Σικάγο ORD	**4,92**	5,32	5,01	7,75	5,09
Φιλαδέλφεια PHL	2,62	6,36	**2,61**	3,61	4,19
Φοίνιξ PHX	**4,33**	5,44	4,93	5,23	4,58
Σιάτλ SEA	4,18	10,71	3,38	3,49	**3,38**
Σαν Φρανσίσκο SFO	16,59	15,59	8,99	12,44	**6,34**
Σολτ Λέικ SLC	7,17	12,85	8,22	8,92	**6,23**
AVERAGE	**6,33**	**8,02**	**5,79**	**7,13**	**5,10**

Πηγή: Επεξεργασία στο Rstudio και στο Excel

Στην ουσία χρησιμοποιείται το σφάλμα που δημιουργήθηκε από την πρόβλεψη του αριθμού αφίξεων επιβατών του μοντέλου σε σύγκριση τις πραγματικές τιμές του έτους 2008. Στη συνέχεια πολλαπλασιάζεται η διαφορά του σφάλματος από τη μονάδα με τις τιμές του 2008 και η τιμή που προκύπτει συγκρίνεται με αυτή του 2009. Στον παραπάνω πίνακα (Πίνακας 4.4) παρουσιάζονται συγκεντρωτικά τα σφάλματα της

πρόβλεψης του αριθμού των αφίξεων για το έτος 2009 ανά παράγοντα πρόβλεψης και ανά αεροδρόμιο καθώς επίσης και ο μέσος όρος αυτών.

Από την επικύρωση του μοντέλου πάλι προκύπτει πως το μικρότερο μέσο σφάλμα έχει ο προβλεπτικός παράγοντας sum_seats και στη συνέχεια όπως και πριν ο month και η υστέρηση κατά ένα έτος.

Όπως είναι γνωστό εμπειρικά, ένα φαινόμενο συνήθως δεν καθορίζεται από έναν μόνο παράγοντα όπως στην παραπάνω περίπτωση. Έτσι, παρόλο που το σφάλμα που προκύπτει από την μεταβλητή αριθμός θέσεων είναι αρκετά ικανοποιητικό, θα εξεταστούν παρακάτω με βάση τις πιο πάνω μετρήσεις τα αποτελέσματα των προβλέψεων από πέντε διαφορετικούς συνδυασμούς των πιο πάνω μεταβλητών.

4.2.2 Πολλαπλή Γραμμική Παλινδρόμηση

Λαμβάνοντας υπόψη τις μετρήσεις του δείκτη R^2 στα γραμμικά μοντέλα που παρουσιάστηκαν και τα σφάλματα πρόβλεψης των μοντέλων, δημιουργήθηκαν οι παρακάτω συναρτήσεις πολλαπλής γραμμικής παλινδρόμησης, που χρησιμοποιούν ως ανεξάρτητες μεταβλητές για την πρόβλεψη του αριθμού των αφίξεων επιβατών, γραμμικό συνδυασμό των μεταβλητών της προηγούμενης παραγράφου.

$$sum\_passengers = b_0 + b_1\, month + b_2\, year + e \quad (4.6)$$

$$sum\_passengers = b_0 + b_1\, lag12 + b_2\, lag24 + e \quad (4.7)$$

$$sum\_passengers = b_0 + b_1\, month + b_2\, year + b_3\, sum\_seats + e \quad (4.8)$$

$$sum\_passengers = b_0 + b_1\, lag12 + b_2 * lag24 + b_3\, sum_{seats} + e \quad (4.9)$$

$$sum\_passengers = b_0 + b_1\, month + b_2\, year + b_3\, lag12 + b_4\, lag24 + b_5\, sum\_seats + e \quad (4.10)$$

Οι παράγοντες προβλέψεις που επιλέχθηκαν για το κάθε μοντέλο είναι:

- month, year: δύο μεταβλητές που έχουν ερμηνευτική σχέση μεταξύ τους
- lag12, lag24: δύο μεταβλητές που συσχετίζουν τις παρούσες τιμές με το παρελθόν
- month, year, sum_seats: με την προσθήκη του αριθμού των θέσεων στο μοντέλο, για ισχυροποίηση του μοντέλου πρόβλεψης που έχει σχέση με την τάση και την εποχικότητα

- lag12, lag24, sum_seats: με την ίδια λογική όπως στο προηγούμενο μοντέλο για την ισχυροποίηση του δεύτερου μοντέλου
- month, year, lag12, lag24, sum_seats: συνδυασμός όλων των ανεξάρτητων μεταβλητών που από την απλή γραμμική παλινδρόμηση φαίνεται να προβλέπουν από ικανοποιητικά μέχρι πολύ καλά το μοντέλο (5%-8% κατά μέσο όρο)

Πριν αξιολογήσουμε τα παραπάνω μοντέλα ως προς το σφάλμα πρόβλεψης, θα ελεγχθεί πρώτα αν οι προβλεπτικοί παράγοντες που επιλέχθηκαν για κάθε μοντέλο έχουν ικανοποιητική τιμή στο δείκτη του προσαρμοσμένου R^2.

Πίνακας 4.5 Τιμή προσαρμοσμένου R^2 στις πολλαπλές γραμμικές παλινδρομήσεις

	ΜΕΤΑΒΛΗΤΕΣ				
Τιμή Προσαρμοσμένου R2 ανά αεροδρόμιο άφιξης	**month+year**	**lag12+lag24**	**month+year+sum_seats**	**lag12+lag24+sum_seats**	**month+year+lag12+lag24+sum_seats**
Ατλάντα ATL	0,84	0,67	0,94	0,90	**0,94**
Βοστώνη BOS	0,94	0,83	**0,97**	0,92	**0,97**
Σαρλοτ CLT	0,92	0,87	**0,96**	0,94	**0,96**
Ουάσιγκτον Ρόναλντ Ρέιγκαν DCA	0,87	0,82	0,97	0,88	**0,98**
Ντάλας DFW	0,94	0,70	0,95	0,89	**0,96**
Ντιτρόιτ DTW	0,92	0,80	0,95	0,96	**0,97**
Νιού Αρκ EWR	0,89	0,85	**0,95**	0,88	**0,95**
Ουάσιγκτον IAD	0,07	-0,03	0,94	0,88	**0,96**
Χιούστον IAH	0,90	0,87	0,96	0,94	**0,98**
Λας Βέγκας LAS	0,95	0,75	0,95	0,80	**0,96**
Λος Άντζελες LAX	0,96	0,86	**0,97**	0,92	**0,97**
Νιού Γιορκ LGA	0,78	0,64	0,94	0,81	0,94
Ορλάντο MCO	0,89	0,76	**0,97**	0,85	**0,97**
Μιννεάπολις MSP	0,93	0,81	0,96	0,96	**0,98**
Σικάγο ORD	0,90	0,75	0,94	0,91	**0,95**
Φιλαδέλφεια PHL	0,78	0,67	**0,97**	0,86	**0,97**
Φοίνιξ PHX	0,91	0,86	0,95	0,87	0,95
Σιάτλ SEA	0,99	0,98	**0,99**	0,99	**0,99**
Σαν Φρανσίσκο SFO	0,94	0,87	**0,97**	0,94	**0,97**
Σολτ Λέικ SLC	0,64	0,59	0,96	0,94	**0,97**
AVERAGE	**0,85**	**0,75**	**0,96**	**0,90**	**0,96**

Πηγή: Επεξεργασία στο Rstudio και στο Excel

Η παρατήρηση των τιμών στον παραπάνω πίνακα οδηγεί στο συμπέρασμα πως εφόσον οι τιμές των δεικτών κυμαίνονται κατά μέσο πάνω από το 0,7, οι προβλέψεις των μοντέλων αναμένονται να είναι αρκετά ακριβείς. Επίσης, παρατηρούμε ότι η ανεξάρτητη μεταβλητή sum_seats όπου χρησιμοποιείται οδηγεί σε δείκτες πάνω από το 0,9 που δείχνει και πάλι την ισχυρή επίδραση στα μοντέλα.

Αρχικά, χρησιμοποιούνται τα ιστορικά δεδομένα για να προβλεφθεί η ροή των επιβατών για το έτος 2008. Στον παρακάτω πίνακα τα σφάλματα πρόβλεψης είναι αρκετά ικανοποιητικά καθώς κυμαίνονται κατά μέσο όρο από 3,9% μέχρι και 7,7%.

Πίνακας 4.6 Σφάλματα πρόβλεψης αφίξεων επιβατών για το έτος 2008

ΑΕΡΟΔΡΟΜΙΑ ΑΦΙΞΗΣ	**ΜΕΤΑΒΛΗΤΕΣ ΑΘΡΟΙΣΤΙΚΑ**				
	month+year	**lag12+lag24**	**month+year+sum_seats**	**lag12+lag24+sum_seats**	**month+year+lag12+lag24+sum_seats**
	sMAPE6	**sMAPE7**	**sMAPE8**	**sMAPE9**	**sMAPE10**
Ατλάντα ATL	2,55	**1,63**	1,91	3,56	2,08
Βοστώνη BOS	10,66	12,43	7,97	**6,20**	7,70
Σαρλοτ CLT	3,19	3,42	3,00	3,68	**2,96**
Ουάσιγκτον Ρόναλντ Ρέιγκαν DCA	10,96	6,89	7,41	**5,49**	5,77
Ντάλας DFW	6,29	8,90	6,07	**4,50**	5,79
Ντιτρόιτ DTW	5,01	5,36	4,45	**1,96**	2,29
Νιού Αρκ EWR	10,08	9,36	8,99	**6,20**	7,84
Ουάσιγκτον IAD	7,47	10,18	7,03	**3,73**	5,15
Χιούστον IAH	12,66	8,22	2,76	**2,46**	3,67
Λας Βέγκας LAS	10,27	9,49	6,70	**5,50**	9,50
Λος Άντζελες LAX	5,88	7,21	4,19	**3,29**	5,01
Νιού Γιορκ LGA	10,85	9,07	7,80	**5,48**	6,41
Ορλάντο MCO	7,14	7,72	**2,61**	4,11	2,61
Μιννεάπολις MSP	4,92	4,78	4,58	**1,74**	2,79
Σικάγο ORD	8,84	9,10	7,26	**3,70**	4,55
Φιλαδέλφεια PHL	7,34	5,42	3,92	**2,53**	3,31
Φοίνιξ PHX	8,02	7,76	**3,55**	6,06	3,62
Σιάτλ SEA	3,17	3,09	**2,91**	3,28	3,28
Σαν Φρανσίσκο SFO	7,72	7,58	2,29	**1,66**	1,69
Σολτ Λέικ SLC	10,84	8,78	7,42	**3,29**	5,90
AVERAGE	**7,69**	**7,32**	**5,14**	**3,92**	**4,60**

Πηγή: Επεξεργασία στο Rstudio και στο Excel

Στη συνέχεια για την επικύρωση των αποτελεσμάτων συγκρίνονται οι υπολογισθείσες τιμές του αριθμού των αφίξεων επιβατών για το έτος 2009 με τις πραγματικές τιμές, με τη χρήση του δείκτη sMAPE και όπως προκύπτει από τον πιο κάτω πίνακα οι τιμές των σφαλμάτων κυμαίνονται από 5%-6%.

Επίσης και πάλι διαπιστώνεται η σημαντική συμβολή του παράγοντα sum_seats καθώς ο συνδυασμός του με τις άλλες ανεξάρτητες μεταβλητές μειώνει το σφάλμα κατά 2.5%-3,7%.

Πίνακας 4.7 Σφάλματα πρόβλεψης αφίξεων επιβατών για το έτος 2009

ΑΕΡΟΔΡΟΜΙΑ ΑΦΙΞΗΣ	**ΜΕΤΑΒΛΗΤΕΣ ΑΘΡΟΙΣΤΙΚΑ**				
	month+year	**lag12+lag24**	**month+year+sum_seats**	**lag12+lag24+sum_seats**	**month+year+lag12+lag24+sum_seats**
	sMAPE6	**sMAPE7**	**sMAPE8**	**sMAPE9**	**sMAPE10**
Ατλάντα ATL	2,46	2,40	**2,36**	2,63	2,38
Βοστώνη BOS	10,45	12,44	8,34	**7,51**	8,19
Σαρλοτ CLT	2,86	3,09	2,67	3,37	**2,64**
Ουάσιγκτον Ρόναλντ Ρέιγκαν DCA	9,23	4,86	9,23	**3,72**	3,91
Ντάλας DFW	6,30	8,18	6,15	**5,04**	5,95
Ντιτρόιτ DTW	3,60	**3,33**	4,13	6,69	6,36
Νιού Αρκ EWR	5,83	5,43	5,23	**3,75**	4,60
Ουάσιγκτον IAD	6,04	7,85	5,81	**4,26**	4,81
Χιούστον IAH	9,76	6,88	6,39	6,44	**6,23**
Λας Βέγκας LAS	5,17	**5,03**	5,31	5,65	**5,03**
Λος Άντζελες LAX	6,52	7,00	**6,30**	**6,30**	6,35
Νιού Γιορκ LGA	9,50	8,81	8,35	**7,84**	8,01
Ορλάντο MCO	**4,46**	**4,46**	5,53	4,81	5,53
Μιννεάπολις MSP	**3,88**	3,90	3,94	4,48	4,25
Σικάγο ORD	4,22	4,30	**4,06**	5,29	4,69
Φιλαδέλφεια PHL	3,83	2,70	**2,59**	3,14	2,80
Φοίνιξ PHX	4,93	4,84	4,60	**4,44**	4,58
Σιάτλ SEA	3,40	3,41	3,44	**3,38**	**3,38**
Σαν Φρανσίσκο SFO	11,01	10,86	5,58	**5,05**	5,07
Σολτ Λέικ SLC	11,04	9,14	8,33	**6,21**	7,52
AVERAGE	**6,22**	**5,95**	**5,42**	**5,00**	**5,11**

Πηγή: Επεξεργασία στο Rstudio και στο Excel

Στη συνέχεια γίνεται μία αξιολόγηση των μοντέλων παλινδρόμησης με τη χρήση των συγκεντρωτικών αποτελεσμάτων και η δημιουργία ενός μοντέλου μέσα από τη

σύγκριση των βέλτιστων προβλέψεων (ελάχιστων σφαλμάτων) του κάθε αεροδρομίου και ανάλυση αυτού.

Πίνακας 4.8 Σφάλμα πρόβλεψης αριθμού αφίξεων επιβατών 2009 με συνδυασμό μοντέλων παλινδρόμησης με τα μικρότερα σφάλματα πρόβλεψης ανά αεροδρόμιο το έτος 2008

ΑΕΡΟΔΡΟΜΙΑ ΑΦΙΞΗΣ	ΜΕΤΑΒΛΗΤΕΣ			ΜΕΤΑΒΛΗΤΕΣ ΑΘΡΟΙΣΤΙΚΑ				ΠΡΟΒΛΕΨΗ 2009
	month	lag24	sum_seats	lag12+lag24	month+year+sum_seats	lag12+lag24+sum_seats	month+year+lag12+lag24+sum_seats	
	sMAPE1	sMAPE4	sMAPE5	sMAPE7	sMAPE8	sMAPE9	sMAPE10	
Ατλάντα ATL				**1,63**				2,40
Βοστώνη BOS	**3,94**							6,93
Σαρλοτ CLT							**2,96**	2,64
Ουάσιγκτον Ρόναλντ Ρέιγκαν DCA	**2,47**							2,58
Ντάλας DFW			**4,37**					4,95
Ντιτρόιτ DTW						**1,96**		6,69
Νιού Αρκ EWR	**4,20**							3,70
Ουάσιγκτον IAD						**3,73**		4,26
Χιούστον IAH						**2,46**		6,44
Λας Βέγκας LAS			**3,00**					6,52
Λος Άντζελες LAX						**3,29**		6,30
Νιού Γιορκ LGA			**5,03**					7,82
Ορλάντο MCO					**2,61**			5,53
Μιννεάπολις MSP						**1,74**		4,48
Σικάγο ORD						**3,70**		4,69
Φιλαδέλφεια PHL						**2,53**		3,14
Φοίνιξ PHX					**3,55**			4,60
Σιάτλ SEA		**2,64**						3,38
Σαν Φρανσίσκο SFO						**1,66**		5,05
Σολτ Λέικ SLC						**3,29**		6,21
AVERAGE								**4,92**

Πηγή: Επεξεργασία στο Rstudio και στο Excel

Από την μελέτη των παραπάνω πινάκων (βλέπε Πίνακας 4.3, Πίνακας 4.6) εξάγεται ένας πίνακας (Πίνακας 4.8) που περιέχει τα ελάχιστα σφάλματα πρόβλεψης ανά

αεροδρόμιο για τον αριθμό αφίξεων των επιβατών σε σχέση με το έτος 2008. Κατόπιν χρησιμοποιούνται αυτά τα σφάλματα για τον υπολογισμό του αριθμού επιβατών του 2009 και από τη σύγκριση τους με τα πραγματικά δεδομένα προκύπτει το σφάλμα πρόβλεψης με τη χρήση του δείκτη sMAPE για το έτος 2009.

Όπως προκύπτει από τον πίνακα αυτόν ο συνδυασμός των ελάχιστων σφαλμάτων πρόβλεψης δίνει ένα μέσο σφάλματα της τάξης του 4,9%. Η διαπίστωση είναι πως η χρήση των βέλτιστων μοντέλων ανά αεροδρόμιο δεν βελτιώνει σημαντικά το σφάλμα, δηλαδή γενικεύοντας μπορεί να ειπωθεί με σχετική ασφάλεια πως η χρήση των τριών μοντέλων που χρησιμοποιούν τις συναρτήσεις των παλινδρομήσεων 4.8, 4.9 και 4.10 παράγουν κατά μέσο όρο τις καλύτερες προβλέψεις της ροής των αφίξεων των επιβατών, καθώς το σφάλμα τους κυμαίνεται κατά μέσο όρο στο 5-5.4% (Πίνακας 4.7).

Αν όμως ως σημείο αναφοράς χρησιμοποιηθεί το κάθε αεροδρόμιο χωριστά και συγκρίνοντας μεταξύ τους πίνακες των σφαλμάτων για τις προβλέψεις των ετών 2008 (Πίνακας 4.3 και Πίνακας 4.6) και 2009 (Πίνακας 4.4 και Πίνακας 4.7) στην απλή και πολλαπλή γραμμική παλινδρόμηση αντίστοιχα, προκύπτουν αποτελέσματα με μικρά σφάλματα όταν σε κάποια χρησιμοποιείται ένας παλινδρομητής και σε άλλα όταν χρησιμοποιούνται περισσότεροι. Επομένως δεν μπορεί να ισχυριστεί κανείς με σιγουριά ότι τα μοντέλα με τη χρήση περισσότερων ανεξάρτητων μεταβλητών είναι αποδοτικότερα ως προς την πρόβλεψη, τουλάχιστον για την συγκεκριμένη περίπτωση. Άρα εδώ παίζει ρόλο αν θα εξεταστεί το φαινόμενο συνολικά, δηλαδή το ενδιαφέρον εστιάζεται στον ρυθμό αφίξεων των αεροδρομίων σε μία χώρα ή περιοχή, ή ειδικά, δηλαδή εξετάζοντας τις ιδιαιτερότητες του κάθε αεροδρομίου χωριστά.

Επειδή στην εξέταση του φαινομένου με μοντέλα μη γραμμικής παλινδρόμησης εξετάζεται η πρόβλεψη για το σύνολο των αεροδρομίων, θα παρουσιαστεί και στην γραμμική παλινδρόμηση ένα μοντέλο που αφορά την πρόβλεψη του συνόλου των αφίξεων των επιβατών στα 20 υπό μελέτη αεροδρόμια.

Αρχικά επιλέχθηκε το καλύτερο μοντέλο από την προηγούμενη ανάλυση που είναι το μοντέλο, όπως φαίνεται από τους πίνακες, που χρησιμοποιεί τις μεταβλητές υστέρησης ενός και δύο ετών και του συνολικού αριθμού των θέσεων και για το 2008 και το 2009. Επίσης σε αυτό το μοντέλο προστίθεται και η μεταβλητή του αεροδρομίου άφιξης (destination), η οποία είναι κατηγορική και δίνει σε 20 διαφορετικές στάθμες τα 20

αεροδρόμια άφιξης. Ο λόγος που ενσωματώνεται στο μοντέλο η μεταβλητή αυτή είναι να παίξει το ρόλο του identifier, για να κάνει το μοντέλο πιο εξειδικευμένο ώστε να λαμβάνει υπόψη του τις ιδιαιτερότητες και την κλίμακα κάθε αεροδρομίου.

Σύμφωνα λοιπόν με αυτή την περιγραφή, οι τιμές των σφαλμάτων που προκύπτουν για το συγκεκριμένο μοντέλο φαίνονται στον πίνακα παρακάτω (Πίνακας 4.9). Κάνοντας μία σύγκριση των μέσων όρων που εμφανίζονται στους προηγούμενους πίνακες και στις τιμές αυτού του πίνακα διαπιστώνεται πως η προσέγγιση του μέσου όρου είναι ενδεικτική της πραγματικής τιμής για την πρόβλεψη του συνολικού αριθμού άφιξης επιβατών στα εν λόγω αεροδρόμια.

Πίνακας 4.9 Σφάλμα πρόβλεψης για τον συνολικό αριθμό άφιξης επιβατών

Πρόβλεψη	**LR**
	lag12+lag24+sum_seats+destination
	sMAPE
Validation set (2008)	4,19
Test set (2009)	4,92

Πηγή: Επεξεργασία στο Rstudio και στο Excel

Όπως φαίνεται από τον παραπάνω πίνακα το global μοντέλο, δηλαδή αυτό που κάνει χρήση όλων των δεδομένων από το σύνολο των επιλεγμένων αεροδρομίων έχει ελαφρώς καλύτερη προβλεπτική ικανότητα οπότε ως βέλτιστο θα χρησιμοποιηθεί στη σύγκριση με τους επόμενους αλγορίθμους. Επίσης κάτι που πρέπει να επισημανθεί είναι ότι είναι πιο απλό ως προσέγγιση αφού γίνεται χρήση ενός μοντέλου αντί 20 διαφορετικών.

4.3 Μοντέλα Μη Γραμμικής Παλινδρόμησης

Τα τελευταία χρόνια τα μοντέλα μηχανικής μάθησης έχουν μπει δυναμικά στον χώρο των προβλέψεων και υπάρχουν πάρα πολλές μελέτες που εξετάζουν την προβλεπτική τους ικανότητα πάνω στα διάφορα πεδία της επιστήμης και της καθημερινότητας. Λόγω αυτού, στο πειραματικό μέρος της συγκεκριμένης εργασίας παρουσιάζονται αλγόριθμοι που ανήκουν στην κατηγορία της μηχανικής μάθησης και συγκεκριμένα από τις μεθόδους των δέντρων απόφασης χρησιμοποιούνται δύο δημοφιλή μοντέλα, τα τυχαία δάση και τα δέντρα σταδιακής ενίσχυσης.

Οι λόγοι της επιλογής αυτών των μοντέλων είναι βασικά δύο. Πρώτον για να ελεγχθεί η ακρίβεια των προβλέψεων και με τη χρήση νέων μεθόδων και να συγκριθεί η αποδοτικότητά τους σε σύγκριση με τις κλασσικές στατιστικές μεθόδους και δεύτερον για να εξαχθεί ένα πρώτο συμπέρασμα για την προβλεπτική τους ακρίβεια στον τομέα των εναέριων μετακινήσεων και ιδιαίτερα των αφίξεων των επιβατών σε επιλεγμένα αεροδρόμια. Ειδικά όσον αφορά τον δεύτερο λόγο, η έμπνευση για την επιλογή των συγκεκριμένων μεθόδων προέκυψε και από τη χρήση τους σε δύο έρευνες που αφορούσαν, η πρώτη την πρόβλεψη των καθυστερήσεων αεροσκαφών όπου γίνεται χρήση των τυχαίων δασών (Rebollo & Balakrishnan, 2014) και η δεύτερη την πρόβλεψη καθυστερήσεων αφίξεων και αναχωρήσεων όπου χρησιμοποιείται ο αλγόριθμος της σταδιακής ενίσχυσης δέντρων (Manna, et al., 2017).

Οι μέθοδοι μηχανικής μάθησης χρησιμοποιούν ως μεταβλητές εισόδου αυτές που επιλέχθηκαν και στις γραμμικές μεθόδους (Πίνακας 4.1). Επιπρόσθετα ορίζεται ως κατηγορική μεταβλητή (dummy) το αεροδρόμιο (destination) διότι στις δύο αυτές μεθόδους θα χρησιμοποιηθεί το σύνολο των δεδομένων για την πρόβλεψη και δεν θα γίνει ξεχωριστά για κάθε μια από τις 20 χρονοσειρές καθώς απαιτούν αρκετά δεδομένα για να εκπαιδευτούν και η συγκεκριμένη προσέγγιση δεν θα έδινε καλά αποτελέσματα, ανεξαρτήτως αν τα μοντέλα είναι αποτελεσματικά ή όχι στη βάση τους. Επιπλέον ορίζονται και οι κανόνες δημιουργίας των δέντρων μέσα από τις υπερπαραμέτρους, όπως εξηγήθηκε και στη θεωρία. Οι υπερπαράμετροι που χρησιμοποιούνται στην πειραματική διαδικασία είναι κυρίως:

- Το πλήθος των δέντρων που χρησιμοποιείται για να διορθώσει το υπολειπόμενο σφάλμα, η οποία χρησιμοποιείται και στα RF και στα GBT
- Ο ρυθμός μάθησης, ο οποίος καθορίζει πόσο γρήγορα θα πρέπει να διορθώνεται το σφάλμα από το ένα δέντρο στο άλλο, η οποία χρησιμοποιείται στα GBT και
- Το πλήθος των μεταβλητών που χρησιμοποιείται για δειγματοληψία ανά δέντρο, η οποία χρησιμοποιείται στα RF

Μία άλλη υπερπαράμετρος που μπορεί να ρυθμιστεί είναι το βάθος του κάθε δέντρου, αλλά μετά από έλεγχο στο πρόβλημα της πρόβλεψης που αντιμετωπίζεται σε αυτήν την εργασία, φαίνεται να είναι μικρότερης σημασίας σε σχέση με τις υπόλοιπες, οπότε δεν χρησιμοποιήθηκε εν τέλει στην πειραματική διαδικασία που παρουσιάζεται παρακάτω.

4.3.1 Τυχαία Δάση (RF)

Τα τυχαία δάση είναι ένας αλγόριθμος που χρησιμοποιείται ευρέως στις προβλέψεις. Έχει αρκετά πλεονεκτήματα όπως ότι μπορεί να χειριστεί μεγάλο αριθμό μεταβλητών και λειτουργεί καλύτερα όταν υπάρχουν αρκετές μεταβλητές, παρέχει δυνατότητα μέτρησης της σημασίας της κάθε παραμέτρου ώστε να ελεγχθεί η επιρροή της κάθε μίας στο μοντέλο και επίσης έχει καλή προσαρμοστικότητα σε κατηγορικές και αριθμητικές μεταβλητές.

Επίσης για την υλοποίηση των τυχαίων δασών πολύ σημαντικό ρόλο παίζει η επιλογή των ιστορικών δεδομένων τα οποία θα πρέπει να είναι αντιπροσωπευτικά του προβλήματος που προσπαθούν να επιλύσουν και να έχουν αρκετή μεταβλητότητα ώστε να καταφέρει ο αλγόριθμος να εντοπίσει επιτυχώς τα μοτίβα συμπεριφοράς και σχέσεων.

Σύμφωνα λοιπόν με τα παραπάνω στο μοντέλο χρησιμοποιήθηκαν και οι πέντε μεταβλητές (Πίνακας 4.1) και οι παρακάτω δύο υπερπαράμετροι:

- num_trees: που είναι το πλήθος των δέντρων, που στη συγκεκριμένη περίπτωση έχει δοκιμαστεί για τις τιμές 25, 50, 75 και 100.
- num_features: που είναι ο αριθμός των παραμέτρων που χρησιμοποιούνται όταν γίνεται δειγματοληψία σε κάθε δέντρο και παίρνει τις τιμές 2,3,4 και 5.

Στον πιο κάτω πίνακα (βλέπε Πίνακας 4.10) φαίνονται οι τιμές των σφαλμάτων πρόβλεψης μετά την εκπαίδευση του μοντέλου για τις αφίξεις επιβατών το έτος 2008 και τα σφάλματα που προκύπτουν από την πιστοποίηση του μοντέλου για το έτος 2009.

Παρατηρώντας τον πίνακα δύο βασικές επισημάνσεις μπορούν να γίνουν. Πρώτον η μικρότερη τιμή σφάλματος για το validation set είναι στο 5,10% και για το test set στο 5.05% και προκύπτει για αριθμό δέντρων 50 και αριθμό παραμέτρων για τη δειγματοληψία σε κάθε δέντρο 5 μεταβλητών εισόδου, άρα δουλεύει καλύτερα όταν τις χρησιμοποιεί σχεδόν όλες σε σχέση με το να παίρνει μόνο κάποιες, το οποίο είναι αναμενόμενο δεδομένου ότι το πλήθος θέσεων είναι καθοριστική μεταβλητή και ως εκ τούτου βοηθά να συμπεριλαμβάνεται σε όλα τα δέντρα.

Προς επικύρωση αυτής της παρατήρησης για όλους τους αριθμούς δέντρων που χρησιμοποιήθηκαν στο μοντέλο, μικρότερο σφάλμα εμφανίζουν αυτά στα οποία

γίνεται χρήση και των 5 μεταβλητών εισόδου. Δεύτερον το συγκεκριμένο μοντέλο εμφανίζει μία σταθερότητα, δηλαδή η χρήση των διαφορετικών παραμέτρων δεν επηρεάζει δραματικά τη μείωση του σφάλματος καθώς όπως φαίνεται κυμαίνεται για το 2008 από 5,1 έως 5,4% και για το 2009 από 5 έως 5,1% περίπου.

Πίνακας 4.10 Σφάλμα πρόβλεψης της άφιξης επιβατών για τα έτη 2008 και 2009

num_trees	num_features	month+year+lag12 +lag24+sum_seats +destination	month+year+lag12 +lag24+sum_seats+ destination
		sMAPE (2008)	sMAPE (2009)
25	2	5,26	5,09
	3	5,40	5,12
	4	5,21	5,07
	5	5,10	5,05
50	2	5,23	5,08
	3	5,37	5,11
	4	5,12	5,05
	5	5,11	5,05
75	2	5,26	5,09
	3	5,31	5,10
	4	5,19	5,07
	5	5,15	5,06
100	2	5,19	5,07
	3	5,28	5,09
	4	5,23	5,08
	5	5,16	5,06
MINIMUM		**5,10**	**5,05**
MAXIMUM		**5,40**	**5,12**

Πηγή: Επεξεργασία στο Rstudio και στο Excel

4.3.2 Σταδιακή Ενίσχυση Δέντρων

Ο αλγόριθμος της σταδιακής ενίσχυσης δέντρων έχει αρκετές υπερπαραμέτρους, οι οποίες συνδυαστικά μπορούν να βελτιώσουν την προβλεπτική ικανότητα του μοντέλου. Παρακάτω παρουσιάζονται αυτές που χρησιμοποιήθηκαν στο μοντέλο και φάνηκαν να επηρεάζουν περισσότερο το αποτέλεσμα:

- num_trees: που είναι το πλήθος των δέντρων
- Learning rate (ή shrinkage fit): που είναι ο ρυθμός μάθησης

Όσον αφορά τις παραπάνω υπερπαραμέτρους, η επιλογή βασίστηκε σε κάποιες από τις δημοσιευμένες εργασίες του Friedman, (Friedman J. H., 2001) και (Friedman J. H., 2002) για τον καλύτερο συνδυασμό των υπερπαραμέτρων ώστε να βελτιστοποιηθεί το αποτέλεσμα της πρόβλεψης.

Η πρότασή του είναι να ξεκινήσει κανείς την πειραματική διαδικασία με έναν αρκετά μεγάλο αριθμό δέντρων και στη συνέχεια να τροποποιεί τον ρυθμό μάθησης. Με βάση αυτή τη λογική επιλέχθηκαν 4 τιμές για τον αριθμό των δέντρων 100, 300, 500 και 700. Όσον αφορά στις τιμές του ρυθμού μάθησης επιλέχθηκαν τιμές μικρότερες - ίσες του 0.1 και συγκεκριμένα 0.02, 0.05, 0.07 και 0.1.

Πίνακας 4.11 Σφάλμα μεθόδου GBT για τα δεδομένα επαλήθευσης (2008) και τα πειραματικά δεδομένα (2009)

num_trees	**learning rate**	**month+year+lag12 +lag24+sum_seats +destination**	**month+year+lag12 +lag24+sum_seats+ destination**
		sMAPE (2008)	**sMAPE (2009)**
100	0,02	11,92	9,21
	0,05	6,48	5,50
	0,07	5,33	5,10
	0,1	5,06	5,04
300	0,02	5,78	5,24
	0,05	4,65	4,97
	0,07	4,71	4,98
	0,1	4,74	4,98
500	0,02	4,81	4,99
	0,05	4,61	4,96
	0,07	4,61	4,96
	0,1	4,70	4,98
700	0,02	4,66	4,97
	0,05	4,60	4,96
	0,07	4,59	4,95
	0,1	4,74	4,96
MINIMUM		**4,59**	**4,95**
MAXIMUM		**11,92**	**9,21**

Πηγή: Επεξεργασία στο Rstudio και το Excel

Οι παρατηρήσεις στην προκειμένη περίπτωση (Πίνακας 4.11) είναι οι εξής. Αρχικά η μικρότερη τιμή σφάλματος είναι 4,48% το 2018 και 4,95% το 2019 και παρέχεται για 700 δέντρα και ρυθμό μάθησης 0.07. Επιπλέον το συγκεκριμένο μοντέλο φαίνεται να επηρεάζεται αρκετά και θετικά ως προς τις υπεραπαραμέτρους, δηλαδή φαίνεται ότι

κυρίως η αύξηση του αριθμού των δέντρων μικραίνει το σφάλμα της πρόβλεψης σε συνδυασμό φυσικά με τον ρυθμό μάθησης ο οποίος όμως παίζει μικρότερο ρόλο.

Σε αυτό το σημείο θα πρέπει γίνει ένας συμβιβασμός μεταξύ της ταχύτητας απόκρισης του μοντέλου άρα και του κόστους χρήσης των πόρων και του αποτελέσματος, εννοώντας ότι όταν ο αριθμός των δέντρων γίνεται από 300 και πάνω φαίνεται μία σταδιακή μείωση του σφάλματος από την τιμή 4,7% σε 4,5% στα 700 δέντρα αλλά δεν είναι τόσο μεγάλη ώστε να προτιμήσουμε ένα πιο αργό και δαπανηρό μοντέλο. Επίσης η χρήση μεγάλου αριθμού δέντρων ενέχει και τον κίνδυνο της υπερπροσαρμογής (overfitting) του μοντέλου κάτι το οποίο δίνει εσφαλμένα θετικά αποτελέσματα.

4.4 Συγκριτική Αξιολόγηση Μοντέλων

Συγκεντρώνοντας τις προηγούμενες αναλύσεις σε έναν συγκριτικό πίνακα παρακάτω (Πίνακας 4.12) διαφαίνεται πως οι διαφορές στο σφάλμα πρόβλεψης είναι μικρές μεταξύ των μεθόδων γραμμικής παλινδρόμησης και των μεθόδων μηχανικής μάθησης όσον αφορά τα πειραματικά δεδομένα.

Πίνακας 4.12 Σφάλμα του καλύτερου μοντέλου και από τις τρεις μεθόδους

Πρόβλεψη	**LR**	**RF**	**GBT**
	lag12+lag24+sum_seats+destination	**month+year+lag12+lag24+sum_seats+destination**	**month+year+lag12+lag24+sum_seats+destination**
	sMAPE	**sMAPE**	**sMAPE**
Validation set (2008)	4,19	5,06	4,48
Test set (2009)	4,92	5,04	4,95

Πηγή: Επεξεργασία στο Rstudio και στο Excel

Συγκεκριμένα όσον αφορά το σφάλμα που προκύπτει από την πρόβλεψη με τη χρήση ιστορικών δεδομένων σε σχέση με τα δεδομένα επαλήθευσης από το έτος 2008 για τον αριθμό αφίξεων επιβατών στο σύνολο των αεροδρομίων λαμβάνοντας υπόψη το βέλτιστο μοντέλο για τη γραμμική παλινδρόμηση, το επικρατέστερο μοντέλο είναι αυτό της μεθόδου LR με μικρή διαφορά από το δεύτερο GBT με μεταβλητές την υστέρηση κατά ένα και δύο χρόνια και τον αριθμό των θέσεων επιβατών. Και τέλος με μικρή διαφορά στο προβλεπτικό σφάλμα έρχεται ο αλγόριθμος RF.

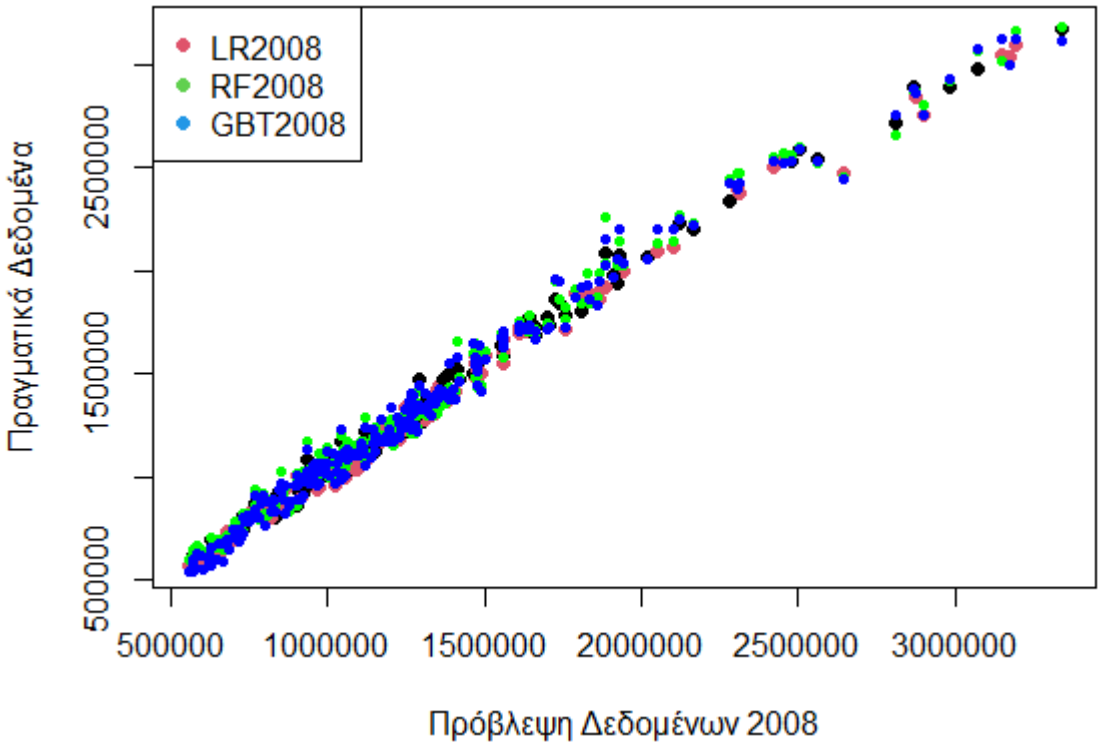
Μια παρατήρηση που μπορεί να γίνει σε αυτό το σημείο είναι πως τα δέντρα είναι λιγότερα αποδοτικά ίσως λόγω του μικρού όγκου δεδομένων που χρησιμοποιείται και

των σχετικά απλών μοτίβων, που ακόμα και η γραμμική παλινδρόμηση μπορεί να προεκτείνει αποτελεσματικά.

Χρησιμοποιώντας τα παραπάνω βέλτιστα μοντέλα απεικονίζονται γραφικά τα προβλεπόμενα δεδομένα σε σχέση με τα πραγματικά και για τις τρεις μεθόδους πρόβλεψης.

Στις δύο γραφικές παραστάσεις (Διάγραμμα 4.1 και Διάγραμμα 4.2) φαίνεται και οπτικά ότι η πρόβλεψη είναι αρκετά επιτυχής και στις τρεις μεθόδους καθώς τα αποτελέσματα των προβλέψεων των μοντέλων σχεδόν ταυτίζονται με τα πραγματικά δεδομένα. Η διαπίστωση αυτή ισχύει και για το δεδομένα πιστοποίησης που φαίνονται στο πρώτο διάγραμμα, όπου παρουσιάζεται ο πραγματικός και ο προβλεπόμενος αριθμός των αφίξεων επιβατών για το 2008 όσο και στο δεύτερο διάγραμμα όπου απεικονίζεται η ίδια διαδικασία για τα πειραματικά δεδομένα που αφορούν το έτος 2009.

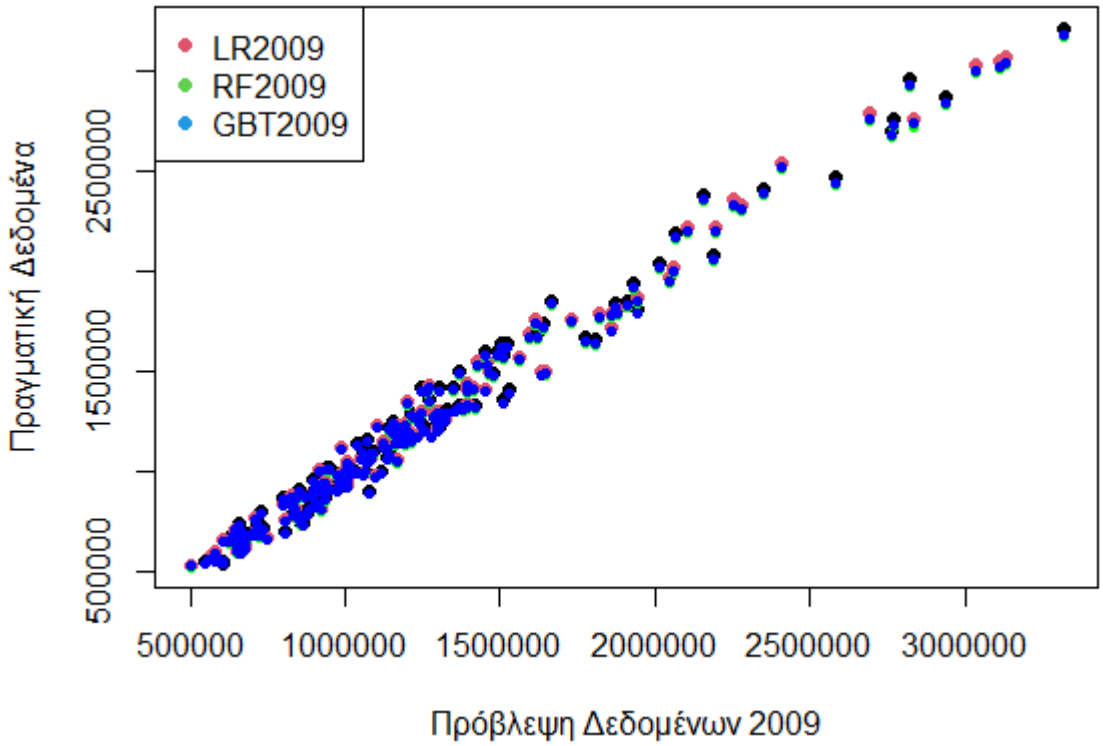
Διάγραμμα 4.1 Παρουσίαση των αποτελεσμάτων πρόβλεψης για το validation set των βέλτιστων μεθόδων σε σχέση με τα πραγματικά δεδομένα



Πηγή: Επεξεργασία στο Rstudio

Τόσο το Διάγραμμα 4.1, που παρουσιάζει τα πραγματικά δεδομένα (μαύρο χρώμα) σε σχέση με τη βέλτιστη πρόβλεψη των μοντέλων LR (κόκκινο χρώμα), RF (πράσινο χρώμα) και GBT (μπλε χρώμα) για το έτος 2008, όσο και το Διάγραμμα 4.2 που παρουσιάζει τα πραγματικά δεδομένα σε σχέση με τα προβλεπόμενα και για τα τρία μοντέλα το 2009, υποστηρίζουν το συμπέρασμα που προέκυψε από τη συγκριτική μελέτη, ότι δηλαδή και τα τρία μοντέλα έχουν μία αρκετά καλή προβλεπτική ικανότητα της οποίας το σφάλμα κινείται κατά μέσο όρο στο 5%.

Διάγραμμα 4.2 Παρουσίαση των αποτελεσμάτων πρόβλεψης για το το test set των βέλτιστων μεθόδων σε σχέση με τα πραγματικά δεδομένα



Πηγή: Επεξεργασία στο Rstudio

Παρακάτω θα αναλυθεί το βέλτιστο μοντέλο πρόβλεψης, δηλαδή αυτό της LR.

4.4.1 Ανάλυση του Βέλτιστου Μοντέλου Πρόβλεψης

Στην παράγραφο αυτή θα αναλυθεί το επικρατέστερο μοντέλο, που προέκυψε από την προηγούμενη πειραματική διαδικασία, δηλαδή αυτό της πολλαπλής γραμμικής παλινδρόμησης LR.

Μία σύνοψη των στατιστικών δεδομένων της συγκεκριμένης πολλαπλής παλινδρόμησης φαίνεται στον παρακάτω πίνακα (Πίνακας 4.13).

Από αυτόν τον πίνακα εξάγονται κάποια βασικά συμπεράσματα. Πρώτον όλες οι υπό μελέτη μεταβλητές είναι στατιστικά σημαντικές για το μοντέλο και επίσης ο δείκτης προβλεπτικής ικανότητας, το προσαρμοσμένο R^2 πάρα πολύ υψηλό (0.9934).

Πίνακας 4.13 Στατιστική ανάλυση της global Πολλαπλής Γραμμικής Παλινδρόμησης

```
lm(formula = sum_passengers ~ lag12 + lag24 + sum_seats + dest,
    data = train_all)

Residuals:
    Min      1Q  Median      3Q     Max
-189755  -27703     602   27700  199263

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -9.755e+05  5.174e+04  -18.85   <2e-16 ***
lag12        3.820e-01  2.078e-02   18.38   <2e-16 ***
lag24        2.138e-01  1.836e-02   11.64   <2e-16 ***
sum_seats    5.739e-01  1.474e-02   38.94   <2e-16 ***
destBOS      6.626e+05  3.637e+04   18.22   <2e-16 ***
destCLT      6.333e+05  3.346e+04   18.93   <2e-16 ***
destDCA      7.121e+05  3.929e+04   18.13   <2e-16 ***
destDFW      3.412e+05  1.896e+04   18.00   <2e-16 ***
destDTW      5.346e+05  3.046e+04   17.55   <2e-16 ***
destEWR      6.586e+05  3.574e+04   18.43   <2e-16 ***
destIAD      7.458e+05  3.963e+04   18.82   <2e-16 ***
destIAH      6.168e+05  3.159e+04   19.52   <2e-16 ***
destLAS      5.336e+05  2.635e+04   20.25   <2e-16 ***
destLAX      4.242e+05  2.352e+04   18.03   <2e-16 ***
destLGA      6.031e+05  3.460e+04   17.43   <2e-16 ***
destMCO      6.467e+05  3.247e+04   19.92   <2e-16 ***
destMSP      5.289e+05  3.017e+04   17.53   <2e-16 ***
destORD      1.895e+05  1.309e+04   14.47   <2e-16 ***
destPHL      5.975e+05  3.306e+04   18.07   <2e-16 ***
destPHX      4.812e+05  2.648e+04   18.17   <2e-16 ***
destSEA      6.629e+05  3.502e+04   18.93   <2e-16 ***
destSFO      6.795e+05  3.612e+04   18.81   <2e-16 ***
destSLC      7.363e+05  3.952e+04   18.63   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 47440 on 937 degrees of freedom
Multiple R-squared:  0.9936,    Adjusted R-squared:  0.9934
F-statistic:  6562 on 22 and 937 DF,  p-value: < 2.2e-16
```

Πηγή: Επεξεργασία στο Rstudio

Το αποτέλεσμα της πρόβλεψης της πειραματικής διαδικασίας φαίνεται να έρχεται σε συμφωνία με αυτή την ανάλυση.

Αν χρησιμοποιηθούν τα δεδομένα για ένα αεροδρόμιο π.χ. της Ατλάντα, αναλύοντας το μοντέλο για το συγκεκριμένο αεροδρόμιο διαπιστώνεται, παρατηρώντας τον

επόμενο πίνακα, ότι οι μεταβλητές είναι στατιστικά σημαντικές (p-value), το προσαρμοσμένο R^2 είναι ικανοποιητικό (0.8987) καθώς επίσης τα κατάλοιπα φαίνονται να είναι συμμετρικά.

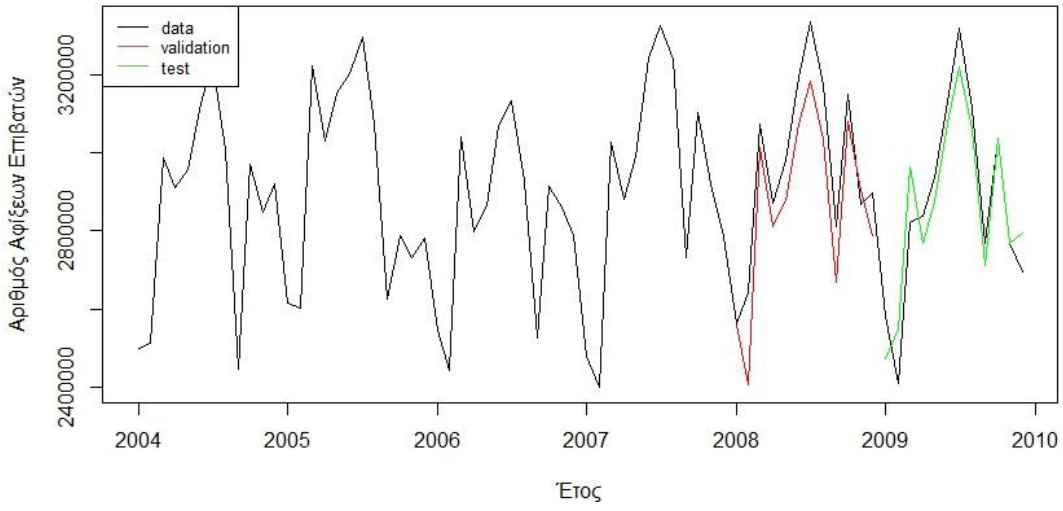
Αυτές οι παρατηρήσεις μας δίνουν μία αίσθηση ότι το προβλεπτικό μοντέλο θα επιφέρει ικανοποιητικά αποτελέσματα.

Πίνακας 4.14 Στατιστική ανάλυση της Πολλαπλής Γραμμικής Παλινδρόμησης για το αεροδρόμιο της Ατλάντα (ATL)

```
Call:
lm(formula = sum_passengers ~ lag12 + lag24 + sum_seats, data = train_ATL)

Residuals:
    Min      1Q  Median      3Q     Max
-165745  -52943     479   56653  136873

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.328e+06  2.215e+05  -5.998 3.39e-07 ***
lag12        1.944e-01  9.896e-02   1.964   0.0558 .
lag24        4.752e-01  8.499e-02   5.591 1.34e-06 ***
sum_seats    6.151e-01  6.015e-02  10.225 3.35e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 80540 on 44 degrees of freedom
Multiple R-squared:  0.9052,    Adjusted R-squared:  0.8987
F-statistic:   140 on 3 and 44 DF,  p-value: < 2.2e-16
```

Πηγή: Επεξεργασία στο Rstudio

Διάγραμμα 4.3 Αριθμός Αφίξεων Επιβατών στο αεροδρόμιο της Ατλάντα για τα έτη 2004-2009, πρόβλεψη δεδομένων (2008) και διασταύρωση δεδομένων (2009)



Πηγή: Επεξεργασία στο Rstudio

Έτσι παρουσιάζοντας οπτικά τα δεδομένα της πρόβλεψης (Διάγραμμα 4.3) με τις ίδιες μεταβλητές όπως και για το συνολικό μοντέλο που παρουσιάστηκε προηγουμένως προκύπτει πως τα αποτελέσματα της πρόβλεψης απεικονίζουν πολύ ικανοποιητικά την πραγματικότητα.

Στην παραπάνω γραφική παράσταση απεικονίζονται τα πραγματικά δεδομένα για τα έτη 2004-2009 και με κόκκινη γραμμή απεικονίζεται η πρόβλεψη των δεδομένων για το έτος 2008 και με πράσινη η πρόβλεψη για το έτος 2009 αντίστοιχα, του αριθμού άφιξης επιβατών στο αεροδρόμια της Ατλάντα (ATL).

Κεφάλαιο 5. Συμπεράσματα και Προεκτάσεις

Στο κεφάλαιο αυτό θα παρουσιαστούν συνοπτικά ο στόχος της εργασίας και της πειραματικής διαδικασίας, καθώς επίσης θα δοθούν αναλυτικά τα συμπεράσματα της παρούσας εργασίας και θα συζητηθεί πως αυτή μπορεί να αποτελέσει το έναυσμα για την περαιτέρω βελτίωση της προβλεπτικής ικανότητας των μοντέλων.

5.1 Σύνοψη Στόχων

Στην παρούσα εργασία ο βασικός στόχος ήταν η ακριβής πρόβλεψη του αριθμού επιβατών που αφικνούνται σε 20 ενδεικτικά αεροδρόμια των ΗΠΑ. Για να επιτευχθεί αυτός ο στόχος χρησιμοποιήθηκαν μηνιαία δεδομένα από το 2004-2009, τα οποία επεξεργάστηκαν και διαχωρίστηκαν σε τρεις ομάδες δεδομένων, ομάδα εκπαίδευσης, ομάδα ελέγχου και πειραματική ομάδα. Τα δεδομένα εκπαίδευσης χρησιμοποιήθηκαν στα μοντέλα πρόβλεψης που επιλέχθηκαν.

Στόχος της πειραματικής διαδικασίας ήταν η χρήση διαφορετικών μοντέλων πρόβλεψης τόσο γραμμικών αλγορίθμων, απλής και πολλαπλής γραμμικής παλινδρόμησης, όσο και μη γραμμικών μεθόδων, όπως τα μοντέλα μηχανικής μάθησης Τυχαίων Δασών και Σταδιακής Ενίσχυσης Δέντρων, για την πρόβλεψη του αριθμού των αφίξεων των επιβατών στα υπό μελέτη αεροδρόμια και στη συνέχεια σύγκριση των στατιστικών μεθόδων και των μεθόδων μηχανικής μάθησης μεταξύ τους.

Ως μέτρο σύγκρισης για την ανάδειξη του ακριβέστερου μοντέλου, όσον αφορά στα αποτελέσματα της πρόβλεψης, των επιλεχθέντων μοντέλων χρησιμοποιήθηκε το συμμετρικό μέσο απόλυτο ποσοστιαίο σφάλμα (sMAPE).

5.2 Συμπεράσματα

Από τα αποτελέσματα της πειραματικής διαδικασίας, και στηριζόμενοι στην ανάλυση που ακολούθησε, αλλά και στους στατιστικούς δείκτες που προέκυψαν ανά περίπτωση, μπορούν να εξαχθούν τα παρακάτω χρήσιμα συμπεράσματα:

- Το σχετικά μικρό σφάλμα που δίνουν και οι τρεις αλγόριθμοι τους καθιστά εξίσου κατάλληλους για την πρόβλεψη τέτοιου τύπου δεδομένων.
- Το γεγονός ότι δεν υπάρχει μεγάλη απόκλιση στο σφάλμα πρόβλεψης μεταξύ των αλγορίθμων αναδεικνύει την αξία των παραδοσιακών στατιστικών

μεθόδων. Θα πρέπει να επισημανθεί όμως, πως το μικρό πλήθος δεδομένων μπορεί να έπαιξε σε κάποιο βαθμό ανασταλτικό ρόλο στην προβλεπτική δυνατότητα των μεθόδων μηχανικής μάθησης.

- Τα global μοντέλα, δηλαδή αυτά που περιλαμβάνουν δεδομένα για όλα τα υπό εξέταση αεροδρόμια, δίνουν ακριβέστερες προβλέψεις, απλοποιώντας παράλληλα σε κάποιο βαθμό τη διαδικασία πρόβλεψης.
- Όσον αφορά την Γραμμική Παλινδρόμηση:
    - Οι παράγοντες πρόβλεψης που επηρεάζουν περισσότερο το μοντέλο κατά σειρά σημαντικότητας είναι η μεταβλητή month που εκφράζει την εποχικότητα, η υστέρηση κατά ένα και δύο έτη και ο συνολικός αριθμός των θέσεων του αεροπλάνου.
    - Αξιοσημείωτο είναι ότι σε κάποια αεροδρόμια αρκεί μία μεταβλητή εισόδου για να είναι αρκετά ακριβής η πρόβλεψη και σε άλλα περισσότερες. Επομένως δεν μπορεί να ισχυριστεί κανείς με σιγουριά ότι τα μοντέλα με τη χρήση περισσότερων ανεξάρτητων μεταβλητών είναι αποδοτικότερα ως προς την πρόβλεψη.
- Όσον αφορά τα Τυχαία Δάση, τα χρησιμοποιούμενα μοντέλα εμφανίζουν σχετικά σταθερά σφάλματα ανεξάρτητα από τη χρήση των υπερπαραμέτρων, το οποίο δείχνει ότι η πρόβλεψη δεν είναι ευαίσθητη στην επιλογή των τιμών των υπερπαραμέτρων.
- Σε σύγκριση με τα Τυχαία Δάση, στα Δέντρα Σταδιακής Ενίσχυσης είναι σε μεγαλύτερο βαθμό σημαντικές οι υπερπαράμετροι που επιλέγονται για να μειώσουν το σφάλμα πρόβλεψης, αλλά πρέπει να επιλέγονται με προσοχή για να μην οδηγηθεί το μοντέλο σε υπερπροσαρμογή

Επίσης ένα κοινό συμπέρασμα που προκύπτει και στα τρία μοντέλα είναι, τουλάχιστον για την συγκεκριμένη εργασία, πως η πρόβλεψη των αφίξεων κατά μέσο όρο είναι αρκετά επιτυχής ανεξάρτητα από το αεροδρόμιο πρόβλεψης, με την έννοια ότι μία πρόβλεψη μπορεί να γενικευτεί χρησιμοποιώντας αυτούς τους παράγοντες σε οποιοδήποτε αεροδρόμιο.

Συνοψίζοντας, τα μοντέλα που χρησιμοποιήθηκαν ήταν αρκετά αποδοτικά με τα πλεονεκτήματά και τα μειονεκτήματά τους το καθένα και ανέδειξαν τη σημασία των αξιόπιστων ιστορικών δεδομένων, την αξία των κλασικών στατιστικών μεθόδων έναντι των μεθόδων μηχανικής μάθησης και το γεγονός πως απλές μεταβλητές όπως ο αριθμός των θέσεων ενός αεροπλάνου, η υστέρηση ενός και δύο ετών και ο μήνας σε σχέση με το έτος μπορούν να προβλέψουν ικανοποιητικά, βραχυπρόθεσμα τουλάχιστον, τον αριθμό αφίξεων των επιβατών σε κάποιο αεροδρόμιο ή/και χώρα για να βοηθήσουν τα ενδιαφερόμενα μέρη για τη λήψη σημαντικών αποφάσεων.

5.3 Προεκτάσεις

Στη συγκεκριμένη εργασία οι προτεινόμενες μεθοδολογίες ήταν αρκετά επιτυχείς, καθότι, όπως περιεγράφηκε προηγουμένως σημείωσαν ένα σφάλμα της τάξης του 5% και στους τρεις αλγορίθμους, γεγονός που δείχνει συγκριτικά την αξιοπιστία και την καταλληλόλητα των μοντέλων. Όμως θα μπορούσε κανείς να χρησιμοποιήσει αυτή την εργασία ως εφαλτήριο για την βελτίωση του προβλεπτικού μηχανισμού. Παρακάτω παρουσιάζονται τέσσερις προτάσεις που ενδεχομένως θα μπορούσαν να βελτιώσουν την απόδοση των προβλέψεων.

5.3.1 Εισαγωγή Επιπλέον Μεταβλητών

Όπως παρουσιάστηκε στη συγκεκριμένη μελέτη χρησιμοποιήθηκε ένας αριθμός συγκεκριμένων μεταβλητών και για τα 20 αεροδρόμια στα οποία προβλέφθηκε ο αριθμός αφίξεων επιβατών.

Μία βελτίωση θα μπορούσε να είναι η διαφορετική αντιμετώπιση των αεροδρομίων βάσει χωροθέτησης, λαμβάνοντας δηλαδή υπόψη τις τοπικές καιρικές συνθήκες. Αν παρατηρήσουμε το χάρτη των ΗΠΑ στο κεφάλαιο 3 θα δούμε ότι το αεροδρόμιο του Σιατλ βρίσκεται στη δυτική ακτή σε σχέση με την Νέα Υόρκη που βρίσκεται στην Ανατολική.

Μια άλλη βελτίωση θα ήταν να αθροιστούν τα δεδομένα για αεροδρόμια που βρίσκονται στην ίδια πόλη, αντιμετωπίζοντας τη ροή των επιβατών ως ενιαία ανά πόλη πια και όχι αεροδρόμιο.

Μια τρίτη βελτίωση θα μπορούσε να είναι η περαιτέρω διερεύνηση των ειδικών συνθηκών της κάθε περιοχής όπως π.χ. κάποια συγκεκριμένη αργία που προσελκύει

πολύ κόσμο ή ένα συγκεκριμένο δρώμενο κάποια περίοδο του χρόνου που θα μπορούσε να αυξήσει τη ροή των επιβατών ή ακόμα ένα ακραίο φαινόμενο όπως ένας τυφώνας με περιοδική συμπεριφορά που θα μπορούσε να μειώσει την άφιξη επιβατών.

Η προσθήκη επιπλέον παραμέτρων θα μπορούσε να εξηγήσει καλύτερα τη συμπεριφορά των δεδομένων και να οδηγήσει σε βελτίωση των δεικτών πρόβλεψης των χρονοσειρών.

5.3.2 Συνδυαστική Πρόβλεψη

Ένας τρόπος που συνηθίζεται και θα μπορούσε να χρησιμοποιηθεί και στη συγκεκριμένη εργασία είναι να χρησιμοποιηθούν οι ίδιες χρονοσειρές για προβλέψεις με διαφορετικά μοντέλα και στη συνέχεια να υπολογιστεί ο απλός ή σταθμισμένος μέσος όρος των προβλέψεων.

Εδώ και δεκαετίες είναι γνωστή η χρησιμότητα του συνδυασμού προβλέψεων, τόσο σε μια δημοσίευση του 1969 (Bates & Granger, 1969) όσο και του 1989 (Clemen, 1989) επισημαίνεται πως ο συνδυασμός πολλαπλών προβλέψεων οδηγεί σε πολύ ικανοποιητικά αποτελέσματα στις προβλέψεις σε σύγκριση με την κάθε μία μέθοδο ξεχωριστά. Αποδεικνύεται ότι υπολογίζοντας απλά τον μέσο όρο βελτιώνεται σημαντικά η απόδοση στην πρόβλεψη.

Επίσης, παρόλο που έγινε προσπάθεια και με τη χρήση σταθμισμένων μέσων και πιο σύνθετων προσεγγίσεων, η χρήση του απλού μέσου όρου παραμένει συνήθως εξίσου αποτελεσματική (Hyndman & Athanasopoulos, 2018).

5.3.3 Μοντέλα Μηχανικής Μάθησης και Χρήση Δεδομένων από το Διαδίκτυο

Μία άλλη βελτίωση που θα μπορούσε να χρησιμοποιηθεί στη διαδικασία της πρόβλεψης είναι η προσθήκη «μη συμβατικών» παραμέτρων στις χρονοσειρές όπως π.χ. η αναζήτηση των χρηστών του διαδικτύου συγκεκριμένων προορισμών σε συγκεκριμένο χρονικό διάστημα. Πάνω σε αυτό, πολλές φορές παρατηρείται η προτίμηση σε κάποιους τουριστικούς προορισμούς ως μία μεταδοτική τάση μεταξύ των ατόμων, η οποία θα μπορούσε να αποτυπωθεί και να ενσωματωθεί στο μοντέλο πρόβλεψης.

Επίσης θα μπορούσαν να χρησιμοποιηθούν δεδομένα από τα κοινωνικά δίκτυα για την πρόβλεψη της ροής των επιβατών σε συγκεκριμένους προορισμούς. Ήδη υπάρχουν

αρκετές μελέτες που βασίζονται στα συγκεκριμένα δεδομένα για διαφόρων ειδών προβλέψεις. Τα κοινωνικά δίκτυα αποτελούν πηγή άντλησης εκατομμύρια δεδομένων που θα μπορούσαν να χρησιμοποιηθούν για την πρόβλεψη διάφορων διαστάσεων της ζωής, φυσικά με τη χρήση συγκεκριμένων κανόνων και οριοθετήσεων.

5.3.4 Παρουσίαση Δεδομένων σε Πραγματικό Χρόνο

Μια άλλη προέκταση της συγκεκριμένης εργασίας που δεν σχετίζεται με την βελτίωση της πρόβλεψης αλλά με την απεικόνιση αυτής είναι η δημιουργία, είτε με τη χρήση έτοιμων εργαλείων ή με προγραμματιστικό τρόπο, η παρουσίαση των δεδομένων σε πραγματικό χρόνο.

Θα μπορούσαν τα ιστορικά δεδομένα να τροφοδοτούν έναν μηχανισμό πρόβλεψης και καθώς ο χρόνος περνάει και αποκτώνται νέα δεδομένα η πρόβλεψη να μετατοπίζεται χρονικά π.χ. ανά ημέρα, μήνα, τρίμηνο, εξάμηνο κ.τ.λ. Το πρόγραμμα αυτό θα μπορούσε να αξιολογήσει τον προβλεπτικό μηχανισμό και να τον βελτιώνει χρησιμοποιώντας μεθόδους μηχανικής μάθησης για την ακριβέστερη μελλοντική πρόβλεψη της ροής των αφίξεων σε ένα αεροδρόμιο.

Η διαδικασία αυτή θα μπορούσε να είναι αυτοματοποιημένη αλλά παράλληλα να μπορεί να γίνεται ανθρώπινη παρέμβαση με προσθήκη έκτακτων παραμέτρων π.χ. μία επικείμενη χιονοθύελλα ή ένα καύσωνας.

Βιβλιογραφία

Bates, J. M., & Granger, C. W. (1969). The combination of forecasts. *Operational Research Quarterly, 20*(4), σσ. 451-468. doi:https://doi.org/10.1057/jors.1969.103

Berenson, M. L., Levine, D. M., & Szabat, K. A. (2018). *Πρόβλεψη Χρονολογικών Σειρών*. Cyprus: Broken Hill Publishers LTD.

Box, G. P., Jenkins, G. M., & Reinsel, G. C. (1994). *Time series Analysis.* Prentice Hall.

Brockwell, P. J., & Davis, R. A. (2016). *Introduction to Time Series and Forecasting.* Springer publication.

Clemen, R. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting, 5*(4), σσ. 559-583. doi:https://doi.org/10.1016/0169-2070(89)90012-5

Deb, C., Zhang, F., Yang, J., Lee, S., & Shah, K. (2017, 07). A review on time series forecasting techniques for building energy consumption. *Renewable and Sustainable Energy Reviews, 74*, σσ. 902-924. doi:https://doi.org/10.1016/j.rser.2017.02.085

Faust, J., & Wright, J. H. (2013). Forecasting Inflation. Στο *Handbook of Economic Forecasting* (Τόμ. 2 part A, σσ. 2-56). Graham Elliott, Allan Timmermann. doi:https://doi.org/10.1016/B978-0-444-53683-9.00001-3

Flaxman, S., Mishra, S., Gandy, A., Unwin, H. T., Mellan, T. A., Coupland, H., . . . Bhatt, S. (2020, 08 13). Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. *Nature, 584*, σσ. 257-261. doi:10.1038/s41586-020-2405-7

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, σσ. 1189-1232.

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis*, *38*, σσ. 367-378. doi:https://doi.org/10.1016/S0167-9473(01)00065-2

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press.

Goodwin, P., & Lawton, R. (1999). On the asymmetry of the symmetric MAPE. *International Journal of Forecasting, 15*, σσ. 405-408. doi:https://doi.org/10.1016/S0169-2070(99)00007-2

Gui, G., Liu, F., Yang, J., Zhou, Z., & Zhao, D. (2020, 01). Flight Delay Prediction Based on Aviation Big Data and Machine Learning. *Transactions on Vehicular Technology, 69*(1), σσ. 140-150. doi:10.1109/TVT.2019.2954094

Guo, X., Grushka-Cockayne, Y., & De Reyck, B. (2022). Forecasting Airport Transfer Passenger Flow Using Real-Time Data and Machine Learning. *Manufacturing & Service Operations Management,, 24*(6), σσ. 3193-3214.

Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting Principles and Practice.* OTexts.

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting, 22*, σσ. 679-688. doi:https://doi.org/10.1016/j.ijforecast.2006.03.001

Jiang, P., Yang, H., & Heng, J. (2019). A hybrid forecasting system based on fuzzy time series and multi-objective optimization for wind speed forecasting. *Applied Energy, 235*, σσ. 786-801. doi:https://doi.org/10.1016/j.apenergy.2018.11.012

Kanavos, A., Kounelis, F., Iliadis, L., & Makris, C. (2021). Deep learning models for forecasting aviation demand time series. *Neural Computing and Applications, 33*, σσ. 16329-16343. doi:https://doi.org/10.1007/s00521-021-06232-y

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., . . . Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tre. *Advances in neural information processing systems*.

Mahya, S., & Fereshteh, M. (2020, 07 25). Predictive big data analytics for supply chain demand forecasting: methods, applications, and research opportunities. *Journal of Big Data*.

Makridakis, S. (1993). Accuracy measures: Theoretical and practical concerns. *International journal of forecasting, 9*(4), σσ. 527-529.

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018, 03 27). Statistical and Machine Learning forecasting methods: Concerns and ways forward. *Plos One, 13*(3). doi:https://doi.org/10.1371/journal.pone.0194889

Makridakis, S., Wheelwright, S., & Hyndman, R. J. (1998). *Forecasting: methods and applications* (3rd Edition εκδ.). John Wiley & Sons.

Manna, S., Biswas, S., Kundu, R., Rakshit, S., Gupta, P., & Barman, S. (2017). A statistical approach to predict flight delay using gradient boosted decision tree. *2017 International Conference on Computational Intelligence in Data Science(ICCIDS)*, σσ. 1-5. doi:10.1109/ICCIDS.2017.8272656

Natekin, A., & Knoll, A. (2013, 12). Gradient boosting machines, a tutorial. *Frontiers in neurorobotics, 7*. doi:https://doi.org/10.3389/fnbot.2013.00021

Ordóñez, C., Sánchez Lasheras, F., Roca-Pardiñas, J., & Javier de Cos Juez, F. (2019). A hybrid ARIMA–SVM model for the study of the remaining useful life of aircraft engines. *Journal of Computational and Applied Mathematics, 346*, σσ. 184-191. doi:https://doi.org/10.1016/j.cam.2018.07.008

Petropoulos, F., Apiletti, D., Assimakopoulos, V., Zied Badai, M., Barrow, D. K., Ben Taieb, S., . . . Ziel, F. (2022, Ιούλιος-Σεπτέμβριος). Forecasting: theory and practice. *International Journal of Forecasting, 38*, σσ. 705-871.

Rebollo, J., & Balakrishnan, H. (2014). Characterization and prediction of air traffic delays. *Transportation Research Part C: Emerging Technologies, 44*. doi:https://doi.org/10.1016/j.trc.2014.04.007

Sarker, I. H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science, 2*(3). doi:https://doi.org/10.1007/s42979-021-00592-x

Schapire, R. E. (2013). Explaining adaboost. Στο *Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik* (σσ. 37-52). Berlin, Heidelberg: Springer Berlin Heidelberg.

Schultz, M., & Reitmann, S. (2019). Machine learning approach to predict aircraft boarding. *Transportation Research Part C: Emerging Technologies, 98*, σσ. 391-408. doi:https://doi.org/10.1016/j.trc.2018.09.007

Spiliotis, E. (2022). Decision trees for time-series forecasting. *Foresight, 1*, σσ. 30-44.

Sulistyowati, R., Kuswanto, H., & Astuti, E. (2018). Hybrid forecasting model to predict air passenger and cargo in Indonesia. *2018 international conference on information and communications technology (ICOIACT)*, σσ. 442-447.

Tsui, W., Balli, H., Gilbey, A., & Gow, H. (2014, 06). Forecasting of Hong Kong airport's passenger throughput. *Tourism Management, 42*, σσ. 62-76. doi:https://doi.org/10.1016/j.tourman.2013.10.008

Vinayakumar, R., Soman, K. P., & Poornachandran, P. (2017, 08 11). Applying deep learning approaches for network traffic prediction. *International Conference on Advances in Computing, Communications and Informatics (ICACCI)* (σσ. 2353-2358). Udupi: IEEE. doi:10.1109/ICACCI.2017.8126198

Xu, S., Chan, H., & Zhang, T. (2019). Forecasting the demand of the aviation industry using hybrid time series SARIMA-SVR approach. *Transportation Research Part E: Logistics and Transportation Review, 122*, σσ. 169-180. doi:https://doi.org/10.1016/j.tre.2018.12.005

Yang, T.-Y., Brinton, C. G., Joe-Wong, C., & Chiang, M. (2017, 08). Behavior-Based Grade Prediction for MOOCs Via Time Series Neural Networks. *Journal of Selected Topics in Signal Processing, 11*(5), σσ. 716-728. doi:10.1109/JSTSP.2017.2700227

Πετρόπουλος, Φ., & Ασημακόπουλος, Β. (2011). *Επιχειρησιακές Προβλέψεις.* Συμμετρία.