ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ – ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

Πρόγραμμα Μεταπτυχιακών Σπουδών

«Κυβερνοασφάλεια και Επιστήμη Δεδομένων»

Μεταπτυχιακή Διατριβή

Τίτλος Διατριβής	**Πρόβλεψη αφίξεων στον Διεθνή Αερολιμένα Αθηνών με χρήση μοντέλων Μηχανικής Μάθησης.** **Prediction of arrivals at Athens International Airport using Machine Learning models.**
Ονοματεπώνυμο Φοιτητή	**Εμμανουήλ Βουβάκης**
Πατρώνυμο	**Κωνσταντίνος**
Αριθμός Μητρώου	**ΜΠΚΕΔ21005**
Επιβλέπων	**Ιωάννης Θεοδωρίδης, Καθηγητής**

Ημερομηνία Παράδοσης **Μάρτιος 2023**

Τριμελής Εξεταστική Επιτροπή

Ιωάννης Θεοδωρίδης Καθηγητής	Νικόλαος Πελέκης Αναπλ. Καθηγητής	Άγγελος Πικράκης Επίκ. Καθηγητής

Περιεχόμενα

Σύνοψη :

Στη μελέτη αυτή γίνεται πρόβλεψη του αριθμού αφίξεων για τον Διεθνή Αερολιμένα Αθηνών με διάφορα μοντέλα μηχανικής μάθησης. Πρώτος σκοπός της έρευνας είναι η ανακάλυψη της κατηγορίας μοντέλων που επιτυγχάνουν καλύτερα αποτελέσματα με βάση τα δεδομένα αυτά. Επιπροσθέτως, στόχος είναι η ανακάλυψη του καλύτερου τρόπου διαχείρισης των δεδομένων αλλά και των εξωγενών μεταβλητών για τη βελτίωση των προβλέψεων. Ένα από τα κύρια συμπεράσματα που θα παρουσιαστούν στη συνέχεια είναι ότι τα ensemble μοντέλα απέδωσαν τα καλύτερα αποτελέσματα. Συνεχίζοντας, αξίζει να αναφερθεί ότι η μετατροπή των δεδομένων σε στάσιμα δεν επέφερε κάποια σημαντική βελτίωση στις προβλέψεις των μοντέλων. Κλείνοντας, σημαντικό είναι το γεγονός ότι συμπεριλαμβάνοντας εξωγενείς μεταβλητές, όπως τον χρόνο, τις καραντίνες και τον αριθμό των εμβολιασμένων πολιτών κατά της Covid-19 δεν βελτιώθηκαν οι προβλέψεις.

Abstract :

This study predicts the number of arrivals at the Athens International Airport using various machine learning models. The primary goal of the research is to discover which category of models achieves better results based on this data. Additionally, the objective is to discover the best way to manage the data as well as the exogenous variables to improve predictions. One of the main conclusions that will be presented later is that ensemble models achieved the best results. Furthermore, it is worth mentioning that converting the data to stationary did not significantly improve the models' predictions. Finally, it is important to note that including exogenous variables such as time, quarantine measures, and the number of vaccinated citizens against Covid-19 did not improve the predictions.

1 Εισαγωγή

Η αεροπορία συμβάλλει σημαντικά στις οικονομίες των χωρών σε παγκόσμιο επίπεδο. Ο τομέας υποστηρίζει σημαντικό αριθμό θέσεων εργασίας και παράγει υψηλά έσοδα. [1] Τα κέρδη αυτά βασίζονται στην λειτουργία των αεροπορικών εταιρειών, την κατασκευή αεροσκαφών και εξαρτημάτων αλλά και την παροχή υπηρεσιών συντήρησης επισκευής αεροσκαφών (MRO). Διαδραματίζει επίσης ζωτικό ρόλο στη σύνδεση των χωρών μεταξύ τους και στη διευκόλυνση του εμπορίου και των πολιτιστικών ανταλλαγών. Εκτός από τις άμεσες συνεισφορές της στην οικονομία, συμβάλει και έμμεσα, αφού υποστηρίζει και άλλους τομείς, όπως ο τουρισμός. Ο κλάδος της αεροπορίας υποστηρίζεται από ένα δίκτυο αεροδρομίων και άλλων υποδομών και επιτρέπει την αποτελεσματική και ασφαλή μετακίνηση ανθρώπων και αγαθών τόσο εντός όσο και μεταξύ των χωρών.

Δυστυχώς, ιστορικά μπορεί να παρατηρηθεί ότι ο κλάδος των αερομεταφορών είναι αρκετά ευάλωτος σε μια ποικιλία διαφορετικών προκλήσεων και γεγονότων. Γεγονότα όπως η οικονομική ύφεση, οι φυσικές καταστροφές και οι κρίσεις υγείας (εξάρσεις ιών) μπορούν να έχουν σημαντικό αντίκτυπο. Αυτού του είδους τα γεγονότα μπορεί να οδηγήσουν σε μείωση της ζήτησης για αεροπορικά ταξίδια και, συνεπώς, να έχουν αρνητικό αντίκτυπο στα έσοδα των αεροπορικών εταιρειών και άλλων επιχειρήσεων στον τομέα των αερομεταφορών. Ο κλάδος υπόκειται επίσης σε άλλους τύπους κινδύνων, όπως γεωπολιτικές εντάσεις, αλλαγές στις προτιμήσεις των καταναλωτών και τεχνολογικές διαταραχές, που μπορούν επίσης να επηρεάσουν την απόδοσή του. Παρά αυτές τις προκλήσεις, η αεροπορική βιομηχανία έχει δείξει ιστορικά μια ισχυρή ικανότητα προσαρμογής και ανάκαμψης από δυσμενή γεγονότα και αναμένεται να συνεχίσει να διαδραματίζει ζωτικό ρόλο στην παγκόσμια οικονομία.

Παρατηρώντας ιστορικά γεγονότα κρίσεων υγείας, όπως την έξαρση του SARS (Severe Acute Respiratory Syndrome) το έτος 2003 ή την έξαρση του Ebola το 2014, ο κλάδος των αερομεταφορών δεν είχε επηρεαστεί ποτέ ξανά σε τόσο μεγάλο βαθμό συγκριτικά με αυτό που έμελλε να αντιμετωπίσει το 2020. [2]

Με την εμφάνιση της Covid-19 τον Φεβρουάριο του 2020, η καθημερινότητα ανατράπηκε. Παγκοσμίως δημιουργήθηκαν απρόσμενες επιδράσεις σε όλες τις οικονομίες με τεράστιο αντίκτυπο. Η πανδημία είχε βαθιές επιπτώσεις στον τομέα των αερομεταφορών παγκοσμίως. Σημειώθηκε σημαντική μείωση της ζήτησης για αεροπορικά ταξίδια εξαιτίας του φόβου για ταξίδια λόγω του μεγάλου αριθμού κρουσμάτων. Επιπροσθέτως, πολλές χώρες εφάρμοσαν καθολικές απαγορεύσεις κυκλοφορίας (lockdown) και ταξιδιωτικούς περιορισμούς για τον έλεγχο της εξάπλωσης του ιού.

Ο κλάδος της αεροπορίας στην Ελλάδα, ο οποίος συμβάλλει σημαντικά στην εξαρτώμενη από τον τουρισμό οικονομία της χώρας, επηρεάστηκε ιδιαίτερα από την πανδημία εφόσον παρουσιάστηκε σημαντική μείωση στη ζήτηση για αεροπορικά ταξίδια. Λόγω της σημασίας του κλάδου για την ελληνική οικονομία και του αριθμού των θέσεων εργασίας που υποστηρίζει, σημειώθηκε μείωση στα έσοδα των αεροπορικών εταιρειών καθώς και παρουσιάστηκαν ευρύτερες οικονομικές επιπτώσεις.

Την περίοδο Ιανουαρίου-Δεκεμβρίου 2021 καταγράφηκαν 12,6 εκατ. διεθνείς αεροπορικές αφίξεις έναντι 21,5 εκατ. της περιόδου Ιανουαρίου-Δεκεμβρίου 2019. Παρουσιάστηκε δηλαδή μείωση κατά -41,2% / -8,9 εκατ. αφίξεις. [3]

Έχοντας υπόψιν όλα τα παραπάνω, γίνεται κατανοητή η σημασία του κλάδου της αεροπορίας στις οικονομίες, καθώς και οι προκλήσεις που αντιμετωπίζει. Εφόσον διανύουμε μια περίοδο όπου παρατηρείται ανάκαμψη του κλάδου, επιτακτική είναι η ανάγκη για πρόβλεψη του αριθμού αφίξεων στα αεροδρόμια. Η μελέτη αυτή μπορεί να χρησιμοποιηθεί από αεροδρόμια, αεροπορικές εταιρείες, τοπικές κοινότητες, αλλά και από την κυβέρνηση για εξαγωγή κρίσιμων πληροφοριών, οι οποίες, δυνητικά, θα μπορούσαν να ληφθούν υπόψη κατά τη διαδικασία λήψης αποφάσεων.

Αρχικά, όσον αναφορά τις αεροπορικές εταιρείες, η μελέτη αυτή μπορεί να φανεί χρήσιμη για τη βελτιστοποίηση των δρομολογίων και την καλύτερη αξιοποίηση των πόρων. Για παράδειγμα, εάν μια αεροπορική εταιρεία γνωρίζει ότι ένας συγκεκριμένος αριθμός αεροπλάνων θα φτάσει σε ένα συγκεκριμένο αεροδρόμιο, μπορεί να προγραμματίσει και να διαθέσει ανάλογα τους απαραίτητους πόρους, όπως πλήρωμα εδάφους και εξοπλισμό, για να διασφαλίσει ότι τα αεροπλάνα υπόκεινται στις απαραίτητες και αποτελεσματικές διεργασίες ομαλής λειτουργίας κατά την άφιξή τους. Συνεπώς, μπορεί να βοηθήσει τις αεροπορικές εταιρείες να βελτιώσουν την αποτελεσματικότητά τους, να μειώσουν το κόστος αλλά και να βελτιώσουν την εμπειρία των επιβατών.

Επιπροσθέτως, η πρόβλεψη του αριθμού των αφίξεων μπορεί να βοηθήσει τα αεροδρόμια να σχεδιάσουν και να κατανείμουν τους πόρους τους πιο αποτελεσματικά. Το γεγονός αυτό μπορεί να βοηθήσει στη μείωση των καθυστερήσεων και στη βελτίωση της συνολικής αποτελεσματικότητας του αεροδρομίου. Εάν ένα αεροδρόμιο αναμένει μεγάλο όγκο αφίξεων, μπορεί να ειδοποιήσει τους τοπικούς ανταποκριτές έκτακτης ανάγκης και να συνεργαστεί μαζί τους για να διασφαλίσει ότι υπάρχουν επαρκείς πόροι, όπως, για παράδειγμα, διαθεσιμότητα σε χώρους πύλης και πλήρωμα εδάφους, για να εξυπηρετηθεί η εισροή ταξιδιωτών.

Η πρόβλεψη των αφίξεων μπορεί να είναι επίσης επωφελής και για τις τοπικές εταιρείες με διάφορους τρόπους. Εάν ένα ξενοδοχείο γνωρίζει ότι πολλά αεροπλάνα έρχονται στο αεροδρόμιο, μπορεί να αποφασίσει να προσλάβει περισσότερο προσωπικό για προετοιμασία μιας αύξησης της ζήτησης δωματίων. Παρόμοια με τα ξενοδοχεία, τα εστιατόρια μπορεί να αποφασίσουν να προσθέσουν περισσότερο προσωπικό ή να αλλάξουν το μενού τους μπροστά στην αύξηση των επισκεπτών. Γενικά, η πρόβλεψη των αφίξεων στα αεροδρόμια μπορεί να βοηθήσει τις τοπικές εταιρείες και τους κοινοτικούς οργανισμούς να προετοιμαστούν καλύτερα για αλλαγές στη ζήτηση των καταναλωτών, που μπορεί να οδηγήσει σε βελτιωμένη εξυπηρέτηση πελατών και πιθανώς υψηλότερες πωλήσεις.

Οι κυβερνήσεις σε περίπτωση αύξησης των αφίξεων μπορούν να συνεργαστούν με αεροδρόμια και με τοπικούς ανταποκριτές έκτακτης ανάγκης για να βεβαιωθούν ότι υπάρχουν αρκετοί πόροι. Με τον τρόπο αυτό γίνεται εύρεση τυχόν τρωτών σημείων ή αδυναμιών που θα πρέπει να αντιμετωπιστούν για να αυξηθεί η ασφάλεια των ταξιδιωτών αλλά και η ασφάλεια του αεροδρομίου. Το προσωπικό ασφαλείας ή οι προμήθειες έκτακτης ανάγκης είναι παραδείγματα τέτοιων πόρων, οι οποίοι μπορούν να χρησιμοποιηθούν για τη διαχείριση οποιασδήποτε πιθανής κρίσης. Αντιθέτως, εφόσον προβλεφθεί χαμηλός αριθμός πτήσεων για μεγάλο χρονικό διάστημα, οι κυβερνήσεις μπορούν να κρίνουν εάν χρήζει ανάγκης η δημιουργία κάποιας προσφοράς (π.χ. παραστατικού – voucher) για την ενίσχυση της τοπικής οικονομίας και τον δελεασμό περισσότερων ταξιδιωτών.Έχοντας υπόψιν όλα τα παραπάνω, γίνεται καλύτερα αντιληπτή η σημασία του κλάδου της αεροπορίας στις οικονομίες αλλά και η βαρύτητα της πρόβλεψης των αφίξεων.

Στα επόμενα δύο κεφάλαια θα γίνει ενημέρωση για την απαραίτητη θεωρητική γνώση των χρονοσειρών και της μηχανικής μάθησης. Συνεχίζοντας, στο κεφάλαιο 4, θα γίνει ανάλυση του θεωρητικού υποβάθρου των υπό εξέταση μοντέλων αλλά και του τρόπου κατηγοριοποίησης τους. Επιπλέον, στο κεφάλαιο 5 θα γίνει παρουσίαση κρίσιμων υπερπαραμέτρων των μοντέλων αλλά και μετρικών αξιολόγησης των προβλέψεων τους. Στο κεφάλαιο που ακολουθεί, θα γίνει αναφορά για την επιλεγμένη γλώσσα προγραμματισμού και για τις σημαντικότερες βιβλιοθήκες που χρησιμοποιήθηκαν. Στο κεφάλαιο 7, θα γίνει ανάλυση της διαδικασίας που υλοποιήθηκε αλλά και παρουσίαση των αποτελεσμάτων που αποκτήθηκαν. Επιπροσθέτως, στο κεφάλαιο 8, θα γίνει παρουσίαση της εφαρμογής που αναπτύχθηκε με στόχο τη συνοπτική παρουσίαση της μελέτης. Κλείνοντας την παρούσα μελέτη, θα γίνει παρουσίαση συμπερασμάτων αλλά και προτάσεων που θα μπορούσαν να υλοποιηθούν για περαιτέρω επέκταση.

2 Χρονοσειρές

2.1 Ορισμός

Μια χρονοσειρά είναι μια ακολουθία σημείων δεδομένων, τα οποία χρησιμοποιούνται για την παρακολούθηση μιας συγκεκριμένης μεταβλητής με την πάροδο του χρόνου.

Ειδικότερα, η χρονοσειρά αποτελείται από ένα σύνολο παρατηρήσεων μιας μεταβλητής, οι τιμές της οποίας είναι ιεραρχημένες με βάση τη χρονική περίοδο στην οποία αναφέρονται (π.χ. έτος, τρίμηνο, μήνας κ.ά.). Αυτά τα σημεία δεδομένων μπορεί να συλλέγονται σε τακτά ή τυχαία χρονικά διαστήματα και να είναι αριθμητικά, όπως τιμές μετοχών ή μετρήσεις θερμοκρασίας, ή κατηγορικά, όπως ο αριθμός των επισκεπτών του ιστότοπου ανά ημέρα της εβδομάδας. Τα δεδομένα χρονοσειρών μπορούν να χρησιμοποιηθούν για διάφορους σκοπούς, συμπεριλαμβανομένης της πρόβλεψης μελλοντικών τιμών, του προσδιορισμού προτύπων και τάσεων αλλά και για τον εντοπισμό ανωμαλιών ή αλλαγών στην υποκείμενη διαδικασία δημιουργίας – επεξεργασίας δεδομένων.

Η ανάλυση χρονοσειρών είναι η διαδικασία χρήσης στατιστικών τεχνικών και τεχνικών μηχανικής μάθησης για τη μοντελοποίηση και την κατανόηση των δεδομένων. Αυτό μπορεί να περιλαμβάνει εργασίες όπως η αποσύνθεση και μελέτη μιας χρονοσειράς στις συνιστώσες: τάση (T), εποχικότητα (S), κυκλικότητα (C) και τυχαιότητα (I).

- Το στοιχείο τάσης αντιπροσωπεύει το υποκείμενο μοτίβο των δεδομένων με την πάροδο του χρόνου, όπως μια συνολική άνοδος ή πτώση. Θα μπορούσε να είναι αλλά και να μην είναι γραμμικό. Μια μη γραμμική τάση μπορεί να περιγραφεί χρησιμοποιώντας μια πολυωνυμική παλινδρόμηση ή άλλα μη γραμμικά μοντέλα, ενώ μια γραμμική τάση μπορεί να χαρακτηριστεί χρησιμοποιώντας ένα μοντέλο γραμμικής παλινδρόμησης.
- Η εποχική συνιστώσα αντιπροσωπεύει τα μοτίβα που επαναλαμβάνονται σε ένα καθορισμένο χρονικό διάστημα, όπως καθημερινά, εβδομαδιαία ή ετήσια. Για παράδειγμα, την περίοδο των εορτών συχνά παρατηρείται αύξηση στις λιανικές πωλήσεις.
- Εκτός από αυτά τα τρία κύρια στοιχεία, ορισμένες χρονοσειρές μπορεί επίσης να έχουν πρόσθετα στοιχεία, όπως μια συνιστώσα κύκλου που αντιπροσωπεύει τα μοτίβα που επαναλαμβάνονται σε μια περίοδο μεγαλύτερη από ένα έτος, όπως ένας επιχειρηματικός κύκλος.
- Η ακανόνιστη συνιστώσα είναι η τυχαία διακύμανση των δεδομένων που δεν μπορεί να εξηγηθεί από την τάση ή τις εποχικές συνιστώσες. Ο «θόρυβος» στα δεδομένα είναι ένα άλλο όνομα για αυτό. Είναι σημαντικό να σημειωθεί ότι τα υπολείμματα πρέπει να είναι τυχαία και άσχετα μεταξύ τους.

2.2 Μαθηματική περιγραφή

Αξίζει να αναφερθεί ότι υπάρχουν δύο μοντέλα τα οποία βοηθούν στη μαθηματική περιγραφή μιας χρονοσειράς. Τα μοντέλα αυτά φανερώνουν δηλαδή τον τρόπο με τον οποίο οι παρατηρήσεις προσδιορίζονται από τις τέσσερις συνιστώσες. Τα μοντέλα αυτά είναι το προσθετικό μοντέλο (additive model) και το πολλαπλασιαστικό μοντέλο (multiplicative model).

Στο προσθετικό μοντέλο (additive) οι τιμές της χρονοσειράς για κάθε περίοδο θεωρούνται ως το άθροισμα των τεσσάρων συνιστωσών στο ίδιο σύστημα μονάδων μέτρησης και δημιουργούνται με τον ακόλουθο τρόπο:

$$Y_t = T_t + S_t + C_t + I_t$$

Στο πολλαπλασιαστικό μοντέλο οι τιμές της χρονοσειράς προσδιορίζονται από το γινόμενο των τεσσάρων συνιστωσών ανεξαρτήτου συστήματος μονάδων μέτρησης ως ακολούθως:

$$Y_t = T_t * S_t * C_t * I_t$$

2.3 Στασιμότητα

Στο σημείο αυτό αξίζει να γίνει αναφορά στην στασιμότητα των χρονοσειρών. Αναλυτικότερα, θα ακολουθήσει σύντομη περιγραφή του ορισμού αλλά και των επιπτώσεων που έχει στη διαδικασία μελέτης των χρονοσειρών.

Μια χρονοσειρά λέγεται ότι είναι στάσιμη εάν οι στατιστικές ιδιότητες των δεδομένων, όπως ο μέσος όρος και η διακύμανση, δεν αλλάζουν με την πάροδο του χρόνου. Με άλλα λόγια, μια σταθερή χρονική σειρά είναι εκείνη όπου η υποκείμενη κατανομή πιθανοτήτων των δεδομένων παραμένει σταθερή με την πάροδο του χρόνου. Είναι σημαντικό να σημειωθεί ότι τα περισσότερα δεδομένα χρονοσειρών που συναντώνται είναι μη στάσιμα. Τα στάσιμα δεδομένα είναι πολύ πιο εύκολο να μοντελοποιηθούν και να προβλεφθούν από τα μη στάσιμα δεδομένα. Ως εκ τούτου, είναι συχνά απαραίτητο με τη χρήση κατάλληλων τεχνικών η χρονοσειρά να μετατραπεί σε στάσιμη πριν από την εφαρμογή οποιωνδήποτε μοντέλων πρόβλεψης ή στατιστικής ανάλυσης.

2.4 Πρόβλεψη

Η πρόβλεψη χρονοσειρών είναι η διαδικασία αξιοποίησης ιστορικών δεδομένων για την πρόβλεψη μελλοντικών τιμών μιας χρονοσειράς. Αναλυτικότερα, ο στόχος της πρόβλεψης χρονοσειρών είναι να χρησιμοποιήσει τα ιστορικά δεδομένα για να δημιουργήσει ένα μοντέλο που μπορεί να χρησιμοποιηθεί για να γίνουν ακριβείς προβλέψεις σχετικά με τις μελλοντικές τιμές της χρονοσειράς. Αυτό γίνεται, συνήθως, αναλύοντας τα κρυμμένα μοτίβα και τάσεις στα ιστορικά δεδομένα, χρησιμοποιώντας τα για να γίνουν προβλέψεις σχετικά με τις μελλοντικές τιμές.

Υπάρχουν πολλές διαφορετικές τεχνικές και μοντέλα που μπορούν να χρησιμοποιηθούν για την πρόβλεψη χρονοσειρών, όπως:

- Στατιστικά μοντέλα, π.χ ARIMA (AutoRegressive Integrated Moving Average)
- Μοντέλα μηχανικής μάθησης, π.χ νευρωνικά δίκτυα
- Υβριδικά μοντέλα, που συνδυάζουν τεχνικές στατιστικής και μηχανικής μάθησης

Η πρόβλεψη χρονοσειρών είναι ένας ενεργός τομέας έρευνας, όπου νέες τεχνικές και μοντέλα αναπτύσσονται τακτικά. Υπάρχουν πολλές διαφορετικές τεχνικές και μοντέλα που μπορούν να χρησιμοποιηθούν για την πρόβλεψη χρονοσειρών. Όμως η επιλογή της τεχνικής ή του μοντέλου θα εξαρτηθεί από τα ειδικά χαρακτηριστικά των δεδομένων και το πρόβλημα πρόβλεψης.

2.5 Εφαρμογές

Η πρόβλεψη χρονοσειρών είναι ένα βασικό εργαλείο σε πολλούς τομείς που επιτρέπει σε οργανισμούς και άτομα να λαμβάνουν καλύτερες αποφάσεις παρέχοντάς τους πληροφορίες για μελλοντικές τάσεις και μοτίβα στα δεδομένα τους. Με την ανάλυση ιστορικών δεδομένων και τον εντοπισμό μοτίβων, η πρόβλεψη χρονοσειρών επιτρέπει στους οργανισμούς να προβλέπουν μελλοντικές τάσεις και να λαμβάνουν τεκμηριωμένες αποφάσεις. Με τη βοήθεια των νέων τεχνολογιών και των μεγάλων δεδομένων, το πεδίο της πρόβλεψης χρονοσειρών εξελίσσεται συνεχώς και παρέχει πιο ακριβείς προβλέψεις.

Ένα από τα σημαντικότερα πλεονεκτήματα της πρόβλεψης χρονοσειρών είναι ότι βοηθά τους οργανισμούς να προγραμματίσουν το μέλλον παρέχοντας μια σαφή κατανόηση του τι είναι πιθανό να συμβεί.

Για παράδειγμα, οι επιχειρήσεις μπορούν να χρησιμοποιήσουν την πρόβλεψη χρονοσειρών για να προβλέψουν τη ζήτηση για τα προϊόντα και τις υπηρεσίες τους, επιτρέποντάς τους να σχεδιάσουν ανάλογα χρονοδιαγράμματα παραγωγής, επίπεδα αποθεμάτων και προσωπικό. Αυτό μπορεί να βοηθήσει τις εταιρείες να βελτιστοποιήσουν τις δραστηριότητές τους, να μειώσουν το κόστος και να βελτιώσουν την ικανοποίηση των πελατών.

Στον τομέα των οικονομικών, η πρόβλεψη χρονοσειρών χρησιμοποιείται για την πρόβλεψη τιμών μετοχών, συναλλαγματικών ισοτιμιών και άλλων χρηματοοικονομικών μεταβλητών. Αυτό μπορεί να βοηθήσει τους επενδυτές και τους εμπόρους να λάβουν καλύτερες επενδυτικές αποφάσεις, καθώς και τα χρηματοπιστωτικά ιδρύματα να προβλέπουν τη μελλοντική τους οικονομική απόδοση.

Στην πρόβλεψη καιρού, η πρόβλεψη χρονοσειρών χρησιμοποιείται για την πρόβλεψη καιρικών συνθηκών, η οποία μπορεί να βοηθήσει τους ανθρώπους να προετοιμαστούν για ακραία καιρικά φαινόμενα, όπως καταιγίδες και κύματα καύσωνα. Αυτό μπορεί να βοηθήσει τις κυβερνήσεις και τους οργανισμούς να προβούν στις κατάλληλες ενέργειες για να ελαχιστοποιήσουν τον αντίκτυπο τέτοιων γεγονότων στην κοινωνία και την οικονομία.

Στη διαχείριση πόρων, η πρόβλεψη χρονοσειρών μπορεί να χρησιμοποιηθεί για την πρόβλεψη της κατανάλωσης πόρων όπως η ενέργεια και το νερό, κάτι που μπορεί να βοηθήσει τους οργανισμούς να διαχειρίζονται αυτούς τους πόρους πιο αποτελεσματικά. Αυτό μπορεί να βοηθήσει στην ελαχιστοποίηση των απορριμμάτων και στη μείωση του κόστους.

Όσο αναφορά τον ποιοτικό έλεγχο, η πρόβλεψη χρονοσειρών μπορεί να χρησιμοποιηθεί στη μεταποίηση και σε άλλους κλάδους για την πρόβλεψη της ποιότητας των προϊόντων και την έγκαιρη ανίχνευση ελαττωμάτων, γεγονός που μπορεί να βοηθήσει τους οργανισμούς να μειώσουν το κόστος και να βελτιώσουν την ικανοποίηση των πελατών.

Συνοπτικά, η πρόβλεψη χρονοσειρών είναι ένα ισχυρό εργαλείο και είναι κρίσιμη για τους οργανισμούς και τα άτομα ώστε να λαμβάνουν καλύτερες αποφάσεις. Μπορεί να χρησιμοποιηθεί σε πολλούς τομείς βοηθώντας στη βελτιστοποίηση των λειτουργιών, στην ελαχιστοποίηση της σπατάλης, στη μείωση του κόστους και στη βελτίωση της ικανοποίησης των πελατών.

3 Μηχανική Μάθηση

Η λήψη πληροφοριών από τα δεδομένα είναι ιδιαίτερα σημαντική στην εποχή των μεγάλων δεδομένων. Η τεχνολογική βάση για την εξόρυξη δεδομένων παρέχεται από τη Μηχανική Μάθηση (ML). Χρησιμοποιείται για την εξαγωγή πληροφοριών από ακατέργαστα δεδομένα βάσεων δεδομένων, που μπορούν να ερμηνευτούν με κατανοητό τρόπο και να χρησιμοποιηθούν για μια ποικιλία εφαρμογών.

Η μηχανική μάθηση είναι ένα υποπεδίο της Τεχνητής Νοημοσύνης που ασχολείται με την ανάπτυξη αλγορίθμων και στατιστικών μοντέλων, μέσω των οποίων καθίσταται εφικτή στους υπολογιστές η «μάθηση» απευθείας από τα δεδομένα. Η πρόβλεψη χρονοσειρών είναι ένας σημαντικός τομέας της μηχανικής μάθησης που συχνά παραμελείται. Είναι σημαντικό γιατί υπάρχουν τόσα πολλά προβλήματα πρόβλεψης που περιλαμβάνουν μια συνιστώσα χρόνου. Αυτά τα προβλήματα παραμελούνται επειδή είναι η χρονική συνιστώσα κάνει πιο δύσκολο τον χειρισμό των προβλημάτων αυτών. Η χρήση μεθόδων μηχανικής μάθησης για την πρόβλεψη χρονοσειρών έχει επίσης αποκτήσει αυξανόμενο ενδιαφέρον, καθώς τέτοιες μέθοδοι συχνά παρέχουν καλύτερα αποτελέσματα σε σχέση με τα παραδοσιακά μοντέλα ARIMA από τη στατιστική βιβλιογραφία.

Με απλά λόγια, η μηχανική μάθηση είναι η διαδικασία εκπαίδευσης ενός μοντέλου ώστε να αναγνωρίζει μοτίβα και να λαμβάνει αποφάσεις ή να προβαίνει σε προβλέψεις βάσει δεδομένων. Χρησιμοποιείται για την εξαγωγή χρήσιμων πληροφοριών, προβλέψεων ή ταξινομήσεων από δεδομένα, τα οποία μπορούν να χρησιμοποιηθούν για τη λήψη αποφάσεων.

3.1 Ιστορική αναδρομή

Η προέλευση της μηχανικής μάθησης (ML) μπορεί να εντοπιστεί στη δεκαετία του 1950 και του 1960, όταν οι ερευνητές στην τεχνητή νοημοσύνη άρχισαν να αναπτύσσουν αλγόριθμους που μπορούσαν να μάθουν από δεδομένα. Το πεδίο της ML όπως το ξέρουμε σήμερα, ωστόσο, άρχισε να διαμορφώνεται στις δεκαετίες του 1980 και του 1990 με την εμφάνιση ισχυρών υπολογιστών και τη διαθεσιμότητα μεγάλων ποσοτήτων δεδομένων.

Η πρώτη ιδέα ενός αλγορίθμου μηχανικής μάθησης είναι το perceptron, ένα απλό μοντέλο νευρωνικών δικτύων, που επινοήθηκε το 1943 από τον Frank Rosenblatt. [4] Η πρώτη εφαρμογή του ήταν μια μηχανή που κατασκευάστηκε το 1958 στο Cornell Aeronautical Laboratory από τον Frank Rosenblatt, που χρηματοδοτήθηκε από το Γραφείο Ναυτικών Ερευνών των Ηνωμένων Πολιτειών

Τη δεκαετία του 1960, οι ερευνητές ανέπτυξαν περαιτέρω το πεδίο της ML εισάγοντας νέους αλγόριθμους όπως τα δέντρα απόφασης. Τις δεκαετίες του 1970 και 1980, ο τομέας της ML άρχισε να κερδίζει περισσότερη προσοχή και δημοτικότητα, με την εισαγωγή νέων τεχνικών όπως ο αλγόριθμος k-Nearest Neighbors (k-NN) [5] και τα δίκτυα Bayes. [6] Επιπλέον, την εποχή αυτή εμφανίστηκε ο αλγόριθμος της οπίσθιας διάδοσης του σφάλματος (Backpropagation algorithm), που επέτρεψε στα νευρωνικά δίκτυα να μάθουν πιο σύνθετες εργασίες. [7] Στα τέλη της δεκαετίας του 1990, ο τομέας της ML γνώρισε μια αναζωπύρωση του ενδιαφέροντος με τη διαθεσιμότητα μεγάλων ποσοτήτων δεδομένων και την ανάπτυξη νέων αλγορίθμων όπως τα μοντέλα Support Vector Machines (SVMs) [8] και Random Forest.[9]

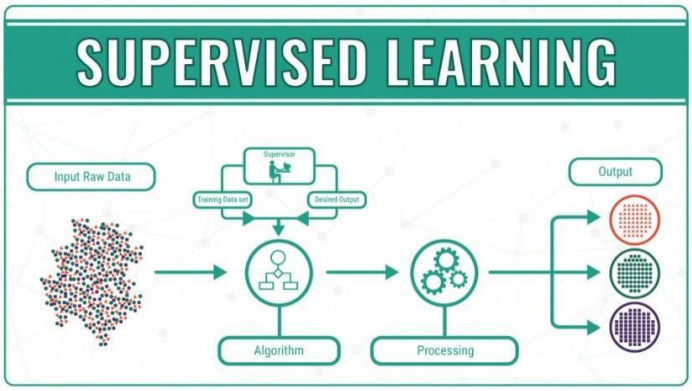
Τα τελευταία χρόνια, η μηχανική μάθηση (ML) έχει γίνει ακόμη πιο δημοφιλής με την εμφάνιση των μεγάλων δεδομένων, του υπολογιστικού νέφους και της αυξημένης χρήσης τεχνικών βαθιάς μάθησης. Με αυτούς τους τρόπους επιτρέπουν στις μηχανές να μαθαίνουν από μεγάλες ποσότητες δεδομένων και να εκτελούν σύνθετες εργασίες όπως η αναγνώριση εικόνας και ομιλίας. Συμπερασματικά, το πεδίο της μηχανικής μάθησης εξελίσσεται εδώ και αρκετές δεκαετίες, με βασικά ορόσημα και προόδους να καταγράφονται όλα αυτά τα χρόνια.

3.2 Κατηγορίες Μηχανικής Μάθησης

Οι κύριοι τύποι μηχανικής μάθησης είναι η εποπτευόμενη (Supervised), η μη εποπτευόμενη (Unsupervised) και η ενισχυτική (Reinforcement) μάθηση.

Η εποπτευόμενη μάθηση (Supervised Learning) είναι ένας τύπος μηχανικής μάθησης όπου ένα μοντέλο εκπαιδεύεται να προβλέπει μια έξοδο με βάση ένα σύνολο εισροών. Ο στόχος της εποπτευόμενης μάθησης είναι να μάθει μια αντιστοίχιση εισόδων-εξόδων από ένα σύνολο δεδομένων εκπαίδευσης.

Το σύνολο δεδομένων εκπαίδευσης αποτελείται από ζεύγη εισόδου-εξόδου, όπου η είσοδος αντιπροσωπεύει τα χαρακτηριστικά των δεδομένων και η έξοδος αντιπροσωπεύει την ετικέτα ή τη μεταβλητή στόχο. [10] Στο πλαίσιο των χρονοσειρών, είναι μια διαδικασία όπου ένα μοντέλο εκπαιδεύεται χρησιμοποιώντας ιστορικά δεδομένα με ετικέτες για να προβλέψει τις μελλοντικές τιμές μιας χρονοσειράς.

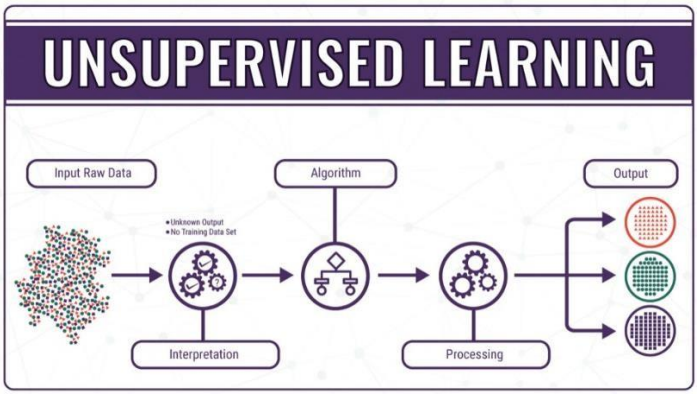


Σχήμα 1: Εποπτευόμενη Μάθηση [76]

Οι κύριες εργασίες της εποπτευόμενη μάθησης, είναι: [11]

- Ταξινόμηση ή κατηγοριοποίηση (Classification), όπου από το σύστημα αναμένεται η πρόβλεψη για την ταξινόμηση των δεδομένων σε διάφορες —εκ των προτέρων γνωστές— κατηγορίες.
- Παλινδρόμηση (Regression), όπου από το σύστημα αναμένεται η πρόβλεψη μιας τιμής.

Η μάθηση χωρίς επίβλεψη (Unsupervised Learning) είναι ένας τύπος μηχανικής μάθησης όπου ένα μοντέλο εκπαιδεύεται να αποκαλύπτει μοτίβα και σχέσεις στα δεδομένα χωρίς τη χρήση δεδομένων με ετικέτα. Ο στόχος της μάθησης χωρίς επίβλεψη είναι ο εντοπισμός υποκείμενων δομών και χαρακτηριστικών των δεδομένων, όπως μοτίβα, συστάδες ή ανωμαλίες, χωρίς να παρέχονται προκαθορισμένες ετικέτες ή μεταβλητές-στόχοι.

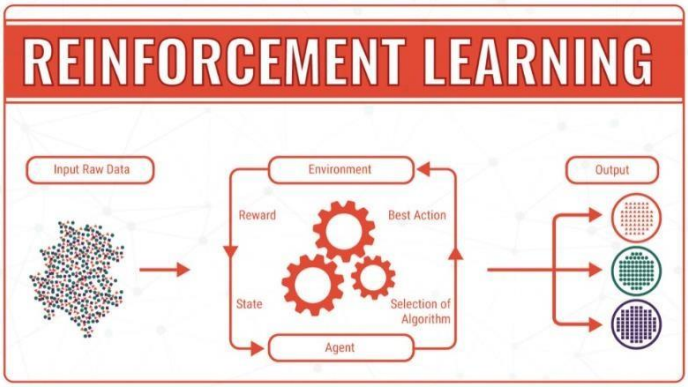


Σχήμα 2: Μη εποπτευόμενη Μάθηση [76]

Οι πιο συνηθισμένες εργασίες μη εποπτευόμενης μάθησης είναι:

- Η συσταδοποίηση (Clustering), όπου από το σύστημα αναμένεται ο διαχωρισμός των δεδομένων σε ομοειδείς συστάδες. Σημαντικοί αλγόριθμοι αυτής της κατηγορίας είναι ο k-Means και ο Hierarchical Cluster Analysis (HCA).
- Η μείωση διαστάσεων (Dimensionality reduction), όπου από το σύστημα αναμένεται η εξαγωγή χρήσιμων ιδιοτήτων των δεδομένων με απλοποίηση των δεδομένων χωρίς να χαθεί όμως χρήσιμη πληροφορία. Σημαντικοί αλγόριθμοι αυτής της κατηγορίας είναι η ανάλυση κυρίων συνιστωσών (Principal Component Analysis - PCA), η μέθοδος kernel PCA που αποτελεί επέκταση της PCA, και η τοπική-γραμμική ενσωμάτωση (Locally-Linear Embedding - LLE).

Η ενισχυτική μάθηση (Reinforcement Learning) είναι ένας τύπος μηχανικής μάθησης που επιτρέπει σε έναν πράκτορα να μάθει πώς να λαμβάνει αποφάσεις αλληλεπιδρώντας με το περιβάλλον του. Ο πράκτορας μαθαίνει να λαμβάνει αποφάσεις λαμβάνοντας ανταμοιβές ή ποινές για τις πράξεις του.



Σχήμα 3: Ενισχυτική Μάθηση [76]

Αναλυτικότερα, με την ορθή ενέργεια επιβραβεύεται και στην αντίθετη περίπτωση τιμωρείται. Σε αυτή την περίπτωση ο πράκτορας πρέπει να μαθαίνει μόνος του και ο στόχος είναι η συνεχής επιβράβευση.

Δηλαδή στόχος της ενισχυτικής μάθησης είναι η μεγιστοποίηση της σωρευτικής ανταμοιβής με την πάροδο του χρόνου.

Κλείνοντας, αξίζει να τονιστεί ότι στις κατηγορίες της μηχανικής μάθησης εντάσσεται και η ημι-επιβλεπόμενη μάθηση. Η κατηγορία αυτή είναι ένας συνδυασμός επιβλεπόμενης και μη επιβλεπόμενης μάθησης, όπου το σύστημα καλείται να εκπαιδευτεί σε σύνολο δεδομένων που περιέχει δομημένα παραδείγματα, αλλά και αδόμητα.

Στη συνέχεια της μελέτης θα παρουσιαστούν και θα αναλυθούν τα μοντέλα που χρησιμοποιήθηκαν. Έχοντας υπόψη τις παραπάνω πληροφορίες για τις μεθόδους μηχανικής μάθησης, χρήσιμη είναι η ενημέρωση ότι όλα τα μοντέλα που θα ακολουθήσουν ανήκουν στην κατηγορία της επιβλεπόμενης μάθησης.

4 Θεωρητικό υπόβαθρο μοντέλων

Στο κεφάλαιο αυτό θα δοθούν πληροφορίες για τα μοντέλα που επιλέχθηκαν να χρησιμοποιηθούν στη μελέτη. Αρχικά, αξίζει να αναφερθεί ότι οι μέθοδοι που χρησιμοποιήθηκαν μπορούν να κατηγοριοποιηθούν ως στατιστικά μοντέλα (statistical models), μοντέλα νευρωνικών δικτύων (ANN) και μέθοδοι συνόλου (ensemble methods). Αναλυτικότερα, το στατιστικό μοντέλο που επιλέχθηκε είναι το AutoArima. Ως μοντέλα νευρωνικών δικτύων επελέγησαν τα MLPRegressor, BDLSTM και TCN μοντέλα. Τέλος, τα μοντέλα XGBoostRegressor, AdaBoostRegressor και HistGradientBoostingRegressor αποτελούν την κατηγορία των ensemble μοντέλων. Στη συνέχεια θα γίνει μια πρώτη σύντομη περιγραφή των τριών κατηγοριών. Ακολούθως, θα γίνει αναλυτική παρουσίαση των εν λόγω μοντέλων, καθώς και των προτερημάτων - μειονεκτημάτων αυτών.

- Στατιστικά μοντέλα:

Τα στατιστικά μοντέλα είναι ένας τύπος μαθηματικού μοντέλου που περιγράφει τις υποκείμενες κατανομές πιθανοτήτων και τις σχέσεις μεταξύ μεταβλητών σε ένα σύνολο δεδομένων. Χρησιμοποιούνται συνήθως στην πρόβλεψη χρονοσειρών για την πρόβλεψη μελλοντικών τιμών με βάση ιστορικά δεδομένα.

- Νευρωνικά Δίκτυα:

Τα νευρωνικά δίκτυα είναι ένας τύπος μοντέλου μηχανικής μάθησης που εμπνέονται από τη δομή και τη λειτουργία του ανθρώπινου εγκεφάλου. Αποτελούνται από στρώματα διασυνδεδεμένων κόμβων, που ονομάζονται νευρώνες, οι οποίοι επεξεργάζονται και μεταδίδουν πληροφορίες. Κάθε νευρώνας λαμβάνει είσοδο από άλλους νευρώνες, εκτελεί έναν απλό υπολογισμό στην είσοδο και στέλνει την έξοδο σε άλλους νευρώνες στο επόμενο στρώμα. Αυτή η διαδικασία επαναλαμβάνεται μέσω πολλαπλών επιπέδων, επιτρέποντας στο δίκτυο να μάθει και να εξάγει σύνθετα χαρακτηριστικά από τα δεδομένα εισόδου. Όταν πρόκειται για πρόβλεψη χρονοσειρών, τα νευρωνικά δίκτυα είναι ιδιαίτερα κατάλληλα, εφόσον είναι σε θέση να καταγράφουν μοτίβα και τάσεις στα δεδομένα και να χρησιμοποιούν αυτές τις πληροφορίες για να ενημερώνουν τις προβλέψεις τους. Σύμφωνα με τη μελέτη των Sheela και Deepa (2013) [12] δεν υπάρχει γενικά αποδεκτή θεωρία για τον ορισμό του πλήθους των κρυμμένων νευρώνων. Στη μελέτη που ακολουθήσει, για τα μοντέλα MLPRegressor και BDLSTM επιλέχθηκαν 2 στρώματα νευρώνων με 14 και 7 νευρώνες, αντίστοιχα. Όσον αναφορά το μοντέλο TCN, επιλέχθηκε η χρήση 2 συνελικτικών στρωμάτων με 30 και 7 φίλτρα, αντίστοιχα.

- Μέθοδοι συνόλου - Ensemble:

Οι μέθοδοι συνόλου (ensemble) είναι ένας τύπος μηχανικής μάθησης που συνδυάζει τις προβλέψεις πολλαπλών μοντέλων για τη βελτίωση της συνολικής ακρίβειας και ευρωστίας των προβλέψεων. Στο πλαίσιο της πρόβλεψης χρονοσειρών, οι μέθοδοι συνόλου μπορούν να χρησιμοποιηθούν για να συνδυάσουν τις προβλέψεις πολλαπλών μοντέλων νευρωνικών δικτύων. Συνδυάζοντας τις προβλέψεις πολλαπλών μοντέλων, οι μέθοδοι συνόλου μπορούν να μειώσουν τη διακύμανση και να αυξήσουν τη σταθερότητα των προβλέψεων.

4.1 Arima

Το μοντέλο ARIMA βασίζεται στις τρεις υπερπαραμέτρους p, d, q. Το "p" προέρχεται από το αυτοπαλίνδρομο τμήμα (AR) και είναι ο αριθμός των καθυστερημένων παρατηρήσεων που χρησιμοποιούνται για την πρόβλεψη της τρέχουσας τιμής της χρονοσειράς. Επιπλέον, το "d" αντιπροσωπεύει τη σειρά διαφοροποίησης, η οποία χρησιμοποιείται για να κάνει τη χρονοσειρά στάσιμη. Μια στάσιμη χρονοσειρά έχει σταθερό μέσο όρο και διακύμανση. Αυτή η παράμετρος βοηθά στην επίτευξη στασιμότητας λαμβάνοντας τη διαφορά των δεδομένων (differencing), για την αφαίρεση τάσεων και εποχικότητας από τις χρονοσειρές. Συνεχίζοντας, το "q" αντιπροσωπεύει τον βαθμό του κινητού μέσου όρου (MA). Δηλαδή είναι ο αριθμός των προηγούμενων σφαλμάτων που χρησιμοποιήθηκαν για την πρόβλεψη της τρέχουσας τιμής της χρονοσειράς. [13]

Επιπροσθέτως, αξίζει να αναφερθεί το μοντέλο SARIMAX (Seasonal AutoRegressive Integrated Moving Average with eXogenous variables), το οποίο είναι επέκταση του ARIMA. Περιλαμβάνει τις ίδιες τρεις παραμέτρους με το αρχικό μοντέλο και προσθέτει τρεις αντίστοιχες παραμέτρους για να

ληφθεί υπόψη η εποχικότητα. Κλείνοντας, να αναφερθεί ότι περιλαμβάνει ακόμα την παράμετρο m, η οποία αφορά τον αριθμό των περιόδων για κάθε εποχή.

4.2 Auto Arima

Το Auto ARIMA είναι μια μέθοδος που χρησιμοποιείται στην ανάλυση χρονοσειρών για την αποτελεσματική πρόβλεψη μελλοντικών τιμών. Είναι μια αυτοματοποιημένη έκδοση του μοντέλου ARIMA (AutoRegressive Integrated Moving Average), η οποία περιλαμβάνει την εξέταση ιστορικών δεδομένων χρονοσειρών και στη συνέχεια την προσαρμογή μοντέλων αυτοπαλίνδρομου και κινούμενου μέσου όρου.

Λόγοι για την αυτοματοποιημένη μοντελοποίηση ARIMA: (α) η μέθοδος για την κατασκευή ενός μοντέλου ARIMA είναι περίπλοκη και απαιτεί πλήρη κατανόηση της μεθόδου. (β) Η κατασκευή ενός μοντέλου ARIMA απαιτεί ενδελεχή εκπαίδευση στη στατιστική ανάλυση (γ) Το πλήθος των χρονοσειρών και των διαφόρων συνδυασμών παραμέτρων είναι μεγάλο. [14]

Η μέθοδος Auto ARIMA εφαρμόζει την προσέγγιση Box-Jenkins. Για τον εντοπισμό, την προσαρμογή και τη διάγνωση μοντέλων ARIMA, οι George Box και Gwilym Jenkins δημιούργησαν την τεχνική Box-Jenkins τη δεκαετία του 1970. Η τεχνική αυτή συνάγεται από μια σειρά διαδικασιών, όπως ο καθορισμός του είδους της τάσης και της εποχικότητας στα δεδομένα, η επιλογή της σειράς του μοντέλου (δηλ. ο αριθμός των συστατικών αυτοπαλίνδρομου και κινούμενου μέσου όρου) και η αξιολόγηση της προσαρμογής του μοντέλου χρησιμοποιώντας στατιστικές δοκιμές και διαγνωστικά διαγράμματα. [15]

Το Auto ARIMA επιτρέπει τη σύγκριση διαφορετικών μοντέλων και εξετάζει τις παραμέτρους τους για να προσδιορίσει το καταλληλότερο μοντέλο για την πρόβλεψη μελλοντικών τιμών. Αυτή η αυτοματοποιημένη μέθοδος λαμβάνει υπόψη τις καθυστερημένες (lagged) τιμές των δεδομένων και την περιοδικότητα των σημείων δεδομένων. [16]

Θετικά:

- Μπορεί να χειριστεί μια ποικιλία καθημερινών πρακτικών σεναρίων καθώς και περίπλοκα μοτίβα, όπως εποχικότητα, κυκλικότητα, τάσεις, εξωγενείς ή εξωτερικές παρεμβάσεις, ακραίες και τυχαίες τιμές που προκαλούνται από άλλες αιτίες. [17, 18]
- Προσφέρει τη δυνατότητα να αποκαλύψει τη μη στάσιμη συμπεριφορά σε δεδομένα χρονοσειρών, βοηθώντας στην πρόβλεψη μελλοντικών τάσεων και στην ανάπτυξη χρήσιμων μοντέλων.
- Αναζητώντας δηλαδή πολλαπλές παραλλαγές των παραμέτρων για τη μοντελοποίηση χρονοσειρών, εξυπηρετεί τους χρήστες στη γρήγορη εύρεση των βέλτιστων.
- Ένα τελευταίο πλεονέκτημα αυτής της τεχνικής είναι ότι εξαλείφει την απαίτηση για χειροκίνητη επιλογή παραμέτρων, ιδιότητα που συμβάλει στην ελαχιστοποίηση της πιθανότητας υπερπροσαρμογής.

Αρνητικά:

- Στην περίπτωση όπου τα δεδομένα χρονοσειρών εμφανίζουν περίπλοκα μοτίβα ή χαρακτηριστικά, μπορεί να μην επιλέγει πάντα το καλύτερο μοντέλο εφόσον το μοντέλο ARIMA δεν είναι σε θέση να ανιχνεύσει. Συνεπώς, ακόμα και με την αναζήτηση του μοντέλου που ταιριάζει καλύτερα, δεν μπορεί να βασιστεί κανείς σε αυτή τη μέθοδο για να ανακαλύψει το γνήσιο βέλτιστο μοντέλο.
- Προκειμένου να επιλεγεί το καλύτερο μοντέλο, η συνάρτηση πρέπει να αξιολογήσει την προσαρμογή πολλαπλών μοντέλων ARIMA με διάφορους συνδυασμούς, καθώς απαιτείται η βελτιστοποίηση των παραμέτρων. Αυτή η διαδικασία μπορεί να είναι υπολογιστικά απαιτητική, ιδιαίτερα για μεγάλα σύνολα δεδομένων χρονοσειρών.
- Επιπροσθέτως, η αυτόματη επιλογή του βέλτιστου μοντέλου μπορεί να δυσκολέψει τους χρήστες να κατανοήσουν πώς επιλέχθηκε το μοντέλο και πώς ταιριάζει στα δεδομένα. Αποκτώντας υπερβολικά πολύπλοκα μοντέλα με περισσότερες παραμέτρους από τις απαιτούμενες, δυσκολεύει του χρήστες να κατανοήσουν καλύτερα τη διαδικασία μοντελοποίησης και τις υποθέσεις πίσω από το επιλεγμένο μοντέλο.
- Τέλος, ένα ακόμη μειονέκτημα αυτής της προσέγγισης είναι ότι δεν λαμβάνει υπόψη τις τιμές σημαντικότητας (p-values) των όρων του μοντέλου κατά την επιλογή του βέλτιστου μοντέλου. Η τιμή p είναι ένα μέτρο της στατιστικής σημασίας ενός όρου του μοντέλου και χρησιμοποιείται

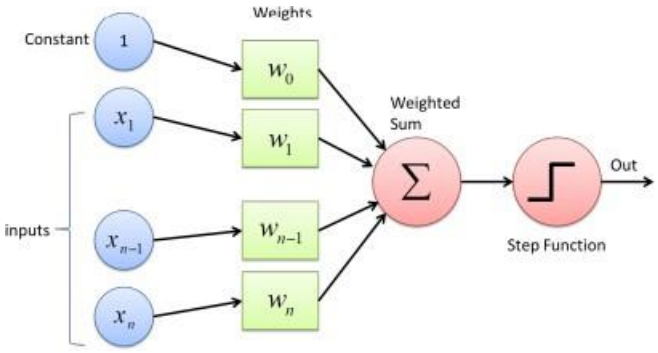
για να προσδιοριστεί εάν ο όρος συμβάλλει σημαντικά στην προσαρμογή του μοντέλου. Μη λαμβάνοντας υπόψη τις τιμές αυτές, το βέλτιστο μοντέλο μπορεί να περιλαμβάνει μη στατιστικά σημαντικούς όρους, γεγονός που οδηγεί σε υπερβολική προσαρμογή (overfitting) ή ανακριβείς προβλέψεις.

Συμπερασματικά, το AutoArima χρησιμοποιεί έναν συνδυασμό μοντέλων στατιστικής και μηχανικής μάθησης για να βρει τις καλύτερες παραμέτρους για ένα μοντέλο ARIMA. Απλοποιεί τη διαδικασία επιλογής των σωστών παραμέτρων για ένα μοντέλο ARIMA και μπορεί να βελτιώσει την απόδοση πρόβλεψης. Ωστόσο, είναι πάντα σημαντικό η απόδοση του μοντέλου να αξιολογείται χρησιμοποιώντας κατάλληλες μετρήσεις και συγκρίνοντάς το με άλλα μοντέλα.

4.3 Perceptron

Το perceptron είναι η απλούστερη μορφή νευρωνικού δικτύου και αποτελεί το βασικό δομικό στοιχείο του δικτύου MLP. Είναι ένας αλγόριθμος δυαδικής ταξινόμησης που λειτουργεί όπως η λογιστική παλινδρόμηση. Αποτελείται, ουσιαστικά, από έναν νευρώνα που λαμβάνει ως είσοδο ένα διάνυσμα της μορφής $x = [x_1, \ldots, x_n]$ και η έξοδός του καθορίζεται από το αποτέλεσμα του σταθισμένου αθροίσματος των εισόδων με βάρη (weights) $w = [x_1, \ldots, x_n]$ συν μία σταθερά b που ονομάζεται πόλωση (bias). Πιο συγκεκριμένα:

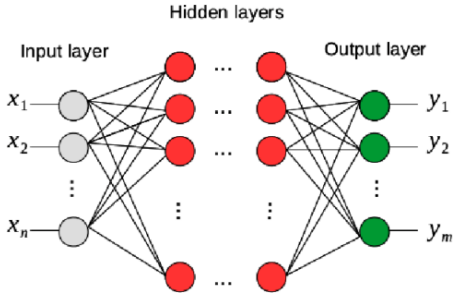
$$f(x; w, b) = \begin{cases} 1, & w \cdot x + b > 0 \\ 0, & διαφορετικά \end{cases}$$



Σχήμα 4: Perceptron [77]

4.4 Multilayer Perceptron

Το πολυεπίπεδο perceptron (Multilayer Perceptron, MLP) αποτελεί μια γενίκευση του απλού perceptron. Το μοντέλο MLPRegressor είναι ένας τύπος εμπροσθοτροφοδοτούμενου νευρωνικού δικτύου που χρησιμοποιείται για την πρόβλεψη χρονοσειρών και άλλες εργασίες παλινδρόμησης. Είναι ένα εποπτευόμενο μοντέλο μηχανικής μάθησης το οποίο απαιτεί δεδομένα εκπαίδευσης με ετικέτες για να μάθει πώς να κάνει προβλέψεις. Αποτελείται από πλήθος νευρώνων οργανωμένων σε διαδοχικά επίπεδα (layers), στα οποία κάθε νευρώνας σε κάποιο επίπεδο συνδέεται με κάθε νευρώνα του προηγούμενου επιπέδου. Όπως όλα τα νευρωνικά δίκτυα, αποτελείται από ένα επίπεδο εισόδου, ένα ή περισσότερα κρυφά επίπεδα και ένα επίπεδο εξόδου.[19]



Σχήμα 5: Εμπροσθο-τροφοδοτούμενο Τεχνητό Νευρωνικό Δίκτυο [78]

Το επίπεδο εισόδου λαμβάνει τα δεδομένα εισόδου και τα περνάει στο κρυφό επίπεδο, το οποίο επεξεργάζεται τα δεδομένα χρησιμοποιώντας ένα σύνολο βαρών και πόλωσης. Το κρυφό επίπεδο μπορεί να έχει πολλαπλές μονάδες. Τα βάρη και η πόλωση μαθαίνονται κατά τη διάρκεια της εκπαιδευτικής διαδικασίας για την εξαγωγή σχετικών χαρακτηριστικών από τα δεδομένα εισόδου. Το επίπεδο εξόδου λαμβάνει την έξοδο από το κρυφό επίπεδο και παράγει την τελική πρόβλεψη για την έξοδο στόχο. Η διαδικασία υπολογισμού της εξόδου των κρυφών επιπέδων και των επιπέδων εξόδου αποτελεί μία γενίκευση της διαδικασίας υπολογισμού της εξόδου του απλού perceptron. Για την πρόβλεψη χρονοσειρών, με βάση τα δεδομένα των προηγούμενων χρονικών σημείων, η πρόβλεψη είναι η προβλεπόμενη τιμή για το επόμενο χρονικό σημείο δεδομένων. [20]

Αναλυτικότερα, η εκπαίδευση σημαίνει βελτιστοποίηση των βαρών και της πόλωσης, συσχετίζοντας τα διανύσματα εισόδου και εξόδου αλλά και ελαχιστοποιώντας τα σφάλματα μεταξύ των προβλεπόμενων και των παρατηρούμενων τιμών εξόδου. Αυτό το μοντέλο βελτιστοποιεί το τετραγωνικό σφάλμα (squared error) χρησιμοποιώντας LBFGS ή stochastic gradient descent.

Επιπρόσθετα, αξίζει να σημειωθεί ότι οι επιλεγμένες συναρτήσεις ενεργοποίησης είναι η ReLU (Rectified Linear Unit Activation Function) και η Tanh ενώ ο αλγόριθμος ελαχιστοποίησης σφάλματος είναι ο Adam (Zhang et al.2019). [21]

Θετικά:

- Το δίκτυο μπορεί να μάθει να προσομοιώνει τις αλληλεπιδράσεις μεταξύ των χαρακτηριστικών εισόδου. Σε αντίθεση με το να λαμβάνει υπόψη μόνο κάθε χαρακτηριστικό ξεχωριστά, το δίκτυο μπορεί να μάθει να αναγνωρίζει μοτίβα στα δεδομένα που εξαρτώνται από τον συνδυασμό πολλών χαρακτηριστικών εισόδου. Αναλυτικότερα, μπορεί να μάθει να αναγνωρίζει και να εκφράζει αυτές τις συνδέσεις χρησιμοποιώντας τα κρυφά επίπεδα του δικτύου. Συνεπώς, είναι πιθανό να επιφέρει καλύτερο αποτέλεσμα από ένα μοντέλο που απλώς λαμβάνει κάθε χαρακτηριστικό εισόδου σε ξεχωριστά.

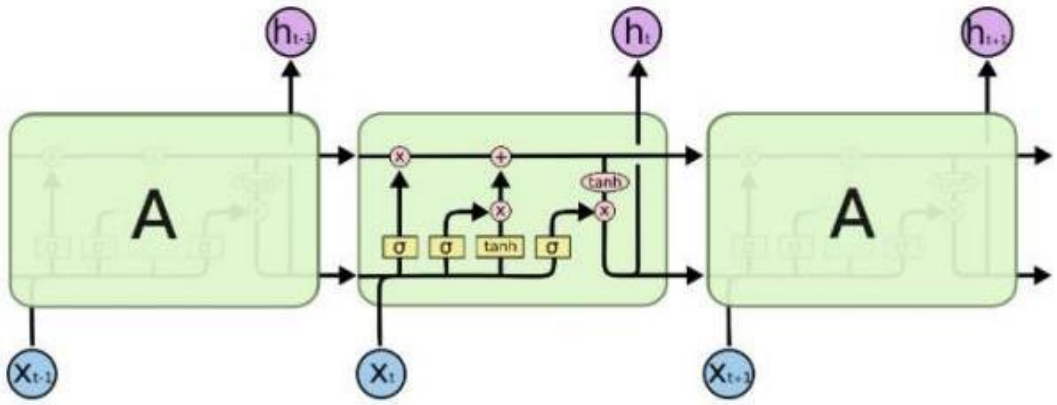
Αρνητικά:

- Τα MLP απαιτούν μεγάλο αριθμό παραμέτρων, γεγονός που μπορεί να τα καταστήσει επιρρεπή σε υπερπροσαρμογή. Το φαινόμενο της υπερπροσαρμογής παρατηρείται όταν ένα μοντέλο πλησιάζει πολύ στα δεδομένα εκπαίδευσης, κάτι που το καθιστά περίπλοκο και ως εκ τούτου οδηγεί σε χαμηλής ποιότητας γενίκευση για νέα δεδομένα Είναι σημαντικό να εφαρμόζονται μέθοδοι όπως η weight decay ή η dropout και να βελτιστοποιείται το μοντέλο χρησιμοποιώντας μεθόδους όπως η cross-validation, προκειμένου να μειωθεί ο κίνδυνος υπερβολικής προσαρμογής.
- Επιπροσθέτως, είναι ευαίσθητο στα αρχικά βάρη και τις προκαταλήψεις (bias) του δικτύου, γεγονός που μπορεί να επηρεάσει τη σύγκλιση του μοντέλου κατά τη διάρκεια της εκπαίδευσης. Εάν τα βάρη και οι προκαταλήψεις δεν αρχικοποιηθούν σωστά, το μοντέλο μπορεί να μην είναι ικανό να μάθει και υπάρχει η πιθανότητα να μην συγκλίνει. Για την αντιμετώπιση αυτού του ζητήματος, είναι σημαντικό να χρησιμοποιηθούν κατάλληλες τεχνικές για την αρχικοποίηση των βαρών και να παρακολουθείται η σύγκλιση του μοντέλου κατά τη διάρκεια της εκπαίδευσης.[22,23]
- Ακολούθως, τα MLP, και ειδικότερα το MLPRegressor, αναφέρονται μερικές φορές ως μοντέλα "μαύρου κουτιού" λόγω των προβλημάτων ερμηνείας τους. Αυτό οφείλεται στην ικανότητα των κρυφών επιπέδων του δικτύου να μαθαίνουν και να αναπαριστούν περίπλοκες συναρτήσεις. Αυτό το γεγονός μπορεί να δυσκολέψει την ακριβή κατανόηση του τρόπου με τον οποίο το μοντέλο δημιουργεί προβλέψεις. Ειδικότερα, είναι αρκετά δύσκολο να γίνει προσδιορισμός των σημαντικών για το μοντέλο χαρακτηριστικών ή να γίνει κατανόηση του τρόπου χρήσης των χαρακτηριστικών εισόδου για την παραγωγή προβλέψεων. Άλλες μορφές μοντέλων, όπως τα δέντρα γραμμικής παλινδρόμησης ή απόφασης, είναι πιο διαφανείς ως προς την κατανόηση μέθοδοι. [24]

Συμπερασματικά, το μοντέλο MLPRegressor είναι ένα ισχυρό και ευέλικτο εργαλείο για την πρόβλεψη συνεχών αποτελεσμάτων σε ένα ευρύ φάσμα εφαρμογών. Η ικανότητά του να μαθαίνει πολύπλοκες μη γραμμικές σχέσεις, να χειρίζεται υψηλές διαστάσεις αλλά και η ικανότητά του να βελτιώνεται με την πάροδο του χρόνου το καθιστούν πολύτιμο πλεονέκτημα για κάθε επιστήμονα δεδομένων. Απαιτεί καλή κατανόηση της υποκείμενης αρχιτεκτονικής των νευρωνικών δικτύων για την επίτευξη βέλτιστης απόδοσης.

4.5 LSTM block

Μια μονάδα LSTM (Long Short-Term Memory), η οποία είναι το βασικό δομικό στοιχείο του δικτύου BDLSTM (Deep Bidirectional LSTM Network), αποτελείται από κελιά μνήμης (cell), πύλες εισόδου, πύλες εξόδου και πύλες επιλεκτικής συγκράτησης (forget gate). [25]



Σχήμα 6: LSTM Block [79]

Το κελί μνήμης είναι υπεύθυνο για την παρακολούθηση των πληροφοριών στην ακολουθία, με το να συνδυάζει την προηγούμενη κατάσταση, την τρέχουσα μνήμη και την είσοδο. Η πύλη επιλεκτικής συγκρατησης (forget gate), πρόκειται για σιγμοειδές επίπεδο που αποφασίζει εάν η τιμή εξόδου της προηγούμενης κατάστασης θα διατηρηθεί ή όχι. Το επόμενο βήμα είναι να ληφθεί η απόφαση σχετικά με το ποια νέα πληροφορία πρόκειται να καταγραφεί στο κελί μνήμης. Αυτό υλοποιείται μέσω της πύλης εισόδου με σιγμοειδές επίπεδο, όπου παράγονται υποψήφιες τιμές που θα μπορούσαν να εισαχθούν στην κατάσταση. Ο συνδυασμός τους πρόκειται να ενημερώσει την κατάσταση του κελιού μνήμης.

4.6 Bidirectional Long Short-Term Memory

Το μοντέλο Bidirectional Long Short-Term Memory (BDLSTM) είναι ένας τύπος ανατροφοδοτούμενου νευρωνικού δικτύου (RNN). Η ιδέα προέρχεται από το αμφίδρομο RNN των Schuster και Paliwal (1997). [26] Το μπλοκ LSTM μπορεί να διατηρήσει πληροφορίες για μεγάλο χρονικό διάστημα και ξεπερνάει το πρόβλημα των κλίσεων (vanishing exploding gradients) που έχουν τα RNN. [27] Επιπλέον, επεξεργάζεται τις ακολουθίες εισόδου τόσο προς τα εμπρός όσο και προς τα πίσω. Σύμφωνα με τους Box et al. (2015), η ανάλυση της περιοδικότητας χρονοσειρών και από τις δύο χρονικές προοπτικές ενισχύει την προγνωστική απόδοση. [28]

Σε ένα αμφίδρομο LSTM, χρησιμοποιούνται δύο ξεχωριστά δίκτυα LSTM: το ένα επεξεργάζεται την ακολουθία εισόδου προς τα εμπρός (από αριστερά προς τα δεξιά) και το άλλο επεξεργάζεται την ακολουθία εισόδου προς τα πίσω (από δεξιά προς τα αριστερά). Η έξοδος από τα δύο διαφορετικών προοπτικών μοντέλα ενώνεται και περνά μέσα από ένα πλήρως συνδεδεμένο στρώμα για να παραχθεί η τελική έξοδος.

Αξίζει να τονιστεί ότι επιλέχθηκε η χρήση της συνάρτησης Huber ως συνάρτηση απώλειας, εφόσον χρησιμοποιείται σε ισχυρά προβλήματα παλινδρόμησης. Είναι ένας συνδυασμός των συναρτήσεων απώλειας μέσου τετραγωνικού σφάλματος (MSE) και μέσου απόλυτου σφάλματος (MAE). Η συνάρτηση απώλειας Huber είναι λιγότερο ευαίσθητη σε ακραίες τιμές στα δεδομένα από τη συνάρτηση απώλειας μέσου τετραγωνικού σφάλματος, γεγονός που την καθιστά πιο ανθεκτική στην παρουσία θορύβου στα δεδομένα. [29] Επιπλέον, στο μοντέλο χρησιμοποιήθηκε ο αλγόριθμος ελαχιστοποίησης σφαλμάτων RMSProp.

Το BidirectionalLSTM συμπεριλαμβάνεται κυρίως χάριν πληρότητας των εργαλείων που δοκιμάστηκαν και σημειώνουμε ότι λύνει ένα μικρότερο πρόβλημα, δηλαδή αυτό της εκτίμησης της επόμενης χρονικής στιγμής (t+1), δοθέντων των μετρήσεων παρελθόντος (έως και τη χρονική στιγμή t) και μέλλοντος (μετά τη χρονική στιγμή t+1). Πρόκειται δηλαδή εν προκειμένω για μέθοδο συμπλήρωσης τιμής.

Θετικά:

- Με την επεξεργασία της ακολουθίας εισόδου τόσο προς τα εμπρός όσο και προς τα πίσω, ένα αμφίδρομο LSTM μπορεί να λάβει υπόψη τόσο το παρελθόν όσο και το μελλοντικό πλαίσιο όταν κάνει προβλέψεις. Αυτό επιτρέπει στο μοντέλο να μάθει πιο ακριβείς χρονικές εξαρτήσεις στα δεδομένα, γεγονός που μπορεί να οδηγήσει σε πιο ακριβείς προβλέψεις.
- Τα δεδομένα χρονοσειρών συχνά εμφανίζουν μη γραμμικές εξαρτήσεις και τα παραδοσιακά μονοκατευθυντικά LSTM μπορεί να δυσκολεύονται να μοντελοποιήσουν αυτές τις εξαρτήσεις. Ένα αμφίδρομο LSTM, μπορεί να χειριστεί καλύτερα μη γραμμικές εξαρτήσεις εφόσον επεξεργάζεται τα δεδομένα από δύο κατευθύνσεις και συνδυάζει τις πληροφορίες από τα δύο LSTM.

Αρνητικά

- Θεωρείται μοντέλο μαύρου κουτιού επειδή οι εσωτερικές τους λειτουργίες είναι πολύπλοκες και δυσνόητες. Αυτό οφείλεται κυρίως στο γεγονός ότι το μοντέλο εκπαιδεύεται χρησιμοποιώντας μεγάλο αριθμό παραμέτρων αλλά και επειδή οι σχέσεις μεταξύ των εισόδων και των εξόδων του μοντέλου είναι μη γραμμικές και δύσκολα ερμηνεύσιμες.
- Η έλλειψη ερμηνείας του μπορεί να κάνει δύσκολη την κατανόηση του τρόπου με τον οποίο το μοντέλο κάνει τις προβλέψεις του. Πιο συγκεκριμένα, είναι δύσκολος ο προσδιορισμός των χαρακτηριστικών εισόδου που είναι πιο σημαντικά για την πραγματοποίηση μιας συγκεκριμένης πρόβλεψης. Επίσης γίνεται δύσκολη η κατανόηση των υποκείμενων μοτίβων τα οποία ανιχνεύει το μοντέλο. Αυτή η έλλειψη διαφάνειας μπορεί να δυσκολέψει τον εντοπισμό σφαλμάτων του μοντέλου και την κατανόησή του, ιδιαίτερα όταν οι προβλέψεις δεν είναι ακριβείς. Ακόμα, το χαρακτηριστικό του αυτό μπορεί να σταθεί εμπόδιο στην εξήγηση των προβλέψεων σε εμπειρογνώμονες ή επιχειρηματίες οι οποίοι θέλουν να βασιστούν σε αυτό για την λήψη διαφόρων αποφάσεων.

4.7 TCN

Ένα TCN, ή Temporal Convolutional Network, είναι ένας τύπος μοντέλου βαθιάς μάθησης που έχει σχεδιαστεί για την επεξεργασία διαδοχικών δεδομένων, όπως χρονοσειρές ή φυσική γλώσσα. Αξίζει να τονιστεί ότι η αρχιτεκτονική TCN ενσωματώνει τα πλεονεκτήματα του recurrent neural network (RNN) αλλά και του convolutional neural network (CNN). [30] Είναι μια παραλλαγή ενός Συνελικτικού Νευρωνικού Δικτύου (CNN), η οποία προτάθηκε για πρώτη φορά το 2018, αλλά με μερικές βασικές διαφορές. [31, 32] Μια βασική διαφορά είναι ότι χρησιμοποιεί αιτιακές συνελίξεις, γεγονός που συνεπάγεται ότι ότι η έξοδος σε κάθε χρονικό βήμα εξαρτάται μόνο από την είσοδο μέχρι αυτό το χρονικό βήμα και όχι από ολόκληρη την ακολουθία. [33] Αυτό επιτρέπει στο μοντέλο να μαθαίνει αποτελεσματικά τις μακροπρόθεσμες εξαρτήσεις στα δεδομένα. Επιπλέον, τα TCN χρησιμοποιούν συχνά διευρυμένες συνελίξεις, οι οποίες αυξάνουν το δεκτικό πεδίο κάθε νευρώνα στο δίκτυο, επιτρέποντας στο μοντέλο να εξετάσει ένα μεγαλύτερο πλαίσιο για κάθε χρονικό βήμα. [34]

Μια αρχιτεκτονική TCN αποτελείται συνήθως από πολλαπλά στρώματα / επίπεδα διευρυμένης αιτιολογικής συνέλιξης, ακολουθούμενα από μη γραμμική συνάρτηση ενεργοποίησης, όπως το ReLU, και στη συνέχεια ένα τελικό πυκνό στρώμα που παράγει την έξοδο. [35] Ο αριθμός των στρώσεων και ο παράγοντας διαστολής που χρησιμοποιούνται μπορούν να προσαρμοστούν ώστε να ταιριάζουν στη συγκεκριμένη εργασία και στην πολυπλοκότητα των δεδομένων. Η αρχιτεκτονική TCN μπορεί να εκπαιδευτεί χρησιμοποιώντας τυπικούς αλγόριθμους backpropagation και βελτιστοποίησης. Αφού εκπαιδευτεί, το μοντέλο μπορεί να χρησιμοποιηθεί για την πρόβλεψη μελλοντικών τιμών στη χρονοσειρά τροφοδοτώντας μια ακολουθία ιστορικών τιμών.

Ωστόσο, ένα TCN, αντί να έχει σταθερό αριθμό νευρώνων, η αρχιτεκτονική χρησιμοποιεί συνελικτικά στρώματα, τα οποία αποτελούνται από ένα σύνολο φίλτρων. [36] Κάθε φίλτρο είναι μια μικρή μήτρα βαρών που χρησιμοποιείται για την εξαγωγή συγκεκριμένων χαρακτηριστικών από τα δεδομένα εισόδου. Το φίλτρο ολισθαίνει πάνω από την είσοδο και σε κάθε θέση λαμβάνεται το γινόμενο μεταξύ των βαρών του φίλτρου και της αντίστοιχης περιοχής της εισόδου. Αυτό έχει ως αποτέλεσμα έναν νέο χάρτη χαρακτηριστικών, ο οποίος υπογραμμίζει ορισμένα χαρακτηριστικά των δεδομένων εισόδου.

Ο αριθμός των φίλτρων σε κάθε επίπεδο είναι μια υπερπαράμετρος που μπορεί να επιλέξει ο χρήστης. Καθορίζει τον αριθμό των «νευρώνων» σε αυτό το στρώμα. Όσο περισσότερα φίλτρα σε ένα επίπεδο τόσο πιο περίπλοκο μπορεί να γίνει το μοντέλο. Ωστόσο, η αύξηση του αριθμού των φίλτρων αυξάνει επίσης τον αριθμό των παραμέτρων που λαμβάνονται υπόψη κατά τη διαδικασία της εκπαίδευσης και μπορεί να οδηγήσει σε υπερπροσαρμογή. Συνοπτικά, αντί να έχει σταθερό αριθμό νευρώνων σε κάθε στρώμα, ένα TCN χρησιμοποιεί συνελικτικά στρώματα που αποτελούνται από ένα σύνολο φίλτρων.

Θετικά:

- Αρχικά, αξίζει να αναφερθεί το γεγονός ότι το μοντέλο αυτό μπορεί να λάβει υπόψη περισσότερα ιστορικά δεδομένα όταν κάνει προβλέψεις, κάτι που είναι ιδιαίτερα χρήσιμο για δεδομένα χρονοσειρών που έχουν μακροπρόθεσμες εξαρτήσεις.

Αρνητικά:

- Λαμβάνοντας υπόψη το γεγονός ότι το μοντέλο θεωρείται μαύρο κουτί, τα αποτελέσματά του μπορεί να είναι δύσκολο να ερμηνευτούν, γεγονός που μπορεί να δυσκολέψει την κατανόηση του συλλογισμού πίσω από τις προβλέψεις του μοντέλου.
- Η χρήση και η εκπαίδευση TCN μπορεί να είναι υπολογιστικά δαπανηρή, ειδικά όταν το πλήθος των δεδομένων εισόδου είναι μεγάλο ή ο σχεδιασμός είναι πολύπλοκος. Αυτό συμβαίνει επειδή υπάρχουν μεγάλο πλήθος παραμέτρων που πρέπει να ληφθούν υπόψη κατά τη διαδικασία της εκπαίδευσης.
- Τα TCN ενδέχεται να είναι ευάλωτα στην υπερπροσαρμογή, καθώς περιέχουν πολλές παραμέτρους, ειδικά εάν δεν υπάρχουν πολλά διαθέσιμα δεδομένα εκπαίδευσης. Αυτό ισχύει ιδιαίτερα όταν το μοντέλο είναι πολύπλοκο, δεδομένου ότι ο αριθμός των παραμέτρων αυξάνεται εκθετικά με τον αριθμό των επιπέδων.

Συμπερασματικά, το μοντέλο TCN με τα ισχυρά συνελικτικά στρώματά του, είναι μια αρχιτεκτονική αιχμής βαθιάς μάθησης που κατακλύζει τον κόσμο της πρόβλεψης χρονοσειρών. Ωστόσο, αξίζει να σημειωθεί ότι όπως και κάθε άλλο μοντέλο, μπορεί να έχει κάποιους περιορισμούς ανάλογα με τη συγκεκριμένη εργασία και τα δεδομένα στα οποία χρησιμοποιείται.

4.8 XGBoost Regressor

Το μοντέλο XGBoost ή αλλιώς Extreme Gradient Boosting, είναι ένα ισχυρό και ευέλικτο εργαλείο για τη δημιουργία μοντέλων παλινδρόμησης. Έχει γίνει αρκετά δημοφιλές στον τομέα της μηχανικής μάθησης εξαιτίας της αποτελεσματικότητας και της ακρίβειάς του. Ως βασικό αλγόριθμο εκμάθησης χρησιμοποιεί δέντρα απόφασεων, τα οποία είναι επίσης γνωστά ως Classification and Regression Trees (CART). Ωστόσο, δεν είναι ένας απλός αλγόριθμος CART. Αποτελεί μια βελτιστοποιημένη εκδοχή του, ονομαζόμενη ως gradient boosting, η οποία χρησιμοποιεί κλίσεις ή μερικές παραγώγους, για να βελτιστοποιήσει τις παραμέτρους των δέντρων και, συνεπώς, να βελτιώσει την απόδοσή του. [37, 38] Συνοπτικά, χρησιμοποιείται ένας συνδυασμός άπληστων αλγορίθμων για την ανάπτυξη των δέντρων αποφάσεων, κάτι που βοηθά στη βελτίωση της ακρίβειας και της αποτελεσματικότητας του μοντέλου.

Κατά τη διάρκεια της εκπαίδευσης ο παλινδρομητής χρησιμοποιεί τον αλγόριθμο Gradient-based One-Side Sampling (GOSS). Ο άπληστος αυτός αλγόριθμος επιλέγει περιπτώσεις, μία κάθε φορά με βάση μια ευρετική. Η ευρετική που χρησιμοποιείται δίνει προτεραιότητα σε παρατηρήσεις που έχουν μεγαλύτερες κλίσεις, οι οποίες είναι οι περιπτώσεις δεδομένων που ταξινομούνται πιο εσφαλμένα από τα προηγούμενα δέντρα. Η διαίσθηση πίσω από αυτό είναι ότι οι περιπτώσεις με μεγαλύτερες κλίσεις είναι πιο ενημερωτικές και θα οδηγήσουν σε μεγαλύτερη μείωση της συνάρτησης απώλειας όταν χρησιμοποιούνται για την εκπαίδευση του επόμενου δέντρου. [39]

Χρησιμοποιεί επίσης έναν αλγόριθμο που ονομάζεται Exclusive Feature Bundling (EFB) για να επιλέξει τα χαρακτηριστικά για κάθε δέντρο. Ο αλγόριθμος αυτός είναι επίσης άπληστος και επιλέγει το πιο σημαντικό υποσύνολο χαρακτηριστικών σε κάθε επίπεδο του δέντρου. [40]

Θετικά:

- Αρχικά, ο παλινδρομητής XGBoost είναι γνωστός για την αποδοτικότητά του. Οι εφαρμοσμένες τεχνικές επιτρέπουν στο XGBoost να χειρίζεται αποτελεσματικά μεγάλα σύνολα δεδομένων και

χώρους χαρακτηριστικών υψηλών διαστάσεων, καθιστώντας το μια καλή επιλογή για προβλήματα πρόβλεψης χρονοσειρών που περιλαμβάνουν πολλά δεδομένα. [41]

- Επιπροσθέτως, που μπορεί να αποφέρει τη σημαντικότητα των χαρακτηριστικών (feature importance), οι οποίες μπορούν να χρησιμοποιηθούν για τον εντοπισμό των πιο κρίσιμων στοιχείων που επηρεάζουν τις προβλέψεις. Αυτό μπορεί να βοηθήσει στον εντοπισμό των πιο σημαντικών χρονικών παραγόντων ή καθυστερήσεων σε καταστάσεις που περιλαμβάνουν πρόβλεψη χρονοσειρών. [42]
- Συνεχίζοντας, αξίζει να αναφερθεί η ικανότητα του μοντέλου να διαχειριστεί δεδομένα τα οποία δεν είναι στάσιμα. Η πρόβλεψη χρονοσειρών απαιτεί συχνά το μοντέλο να είναι σε θέση να προσαρμοστεί στις αλλαγές στην υποκείμενη κατανομή των δεδομένων, η οποία είναι γνωστή ως μη σταθερότητα. Το XGBoost, με τη μη παραμετρική του προσέγγιση, μπορεί να προσαρμοστεί σε μια τέτοια μη σταθερότητα και να αυξήσει την ευελιξία του μοντέλου. [43]

Αρνητικά:

- Το XGBoost είναι ένα μοντέλο που βασίζεται σε δέντρα και παρόλο που μπορεί να καταγράψει μη γραμμικές σχέσεις, δεν μοντελοποιεί ρητά αλληλεπιδράσεις ή μη γραμμικά χαρακτηριστικά μεταξύ των χαρακτηριστικών εισόδου. Επομένως, ενδέχεται να μην έχει καλή απόδοση σε σύνολα δεδομένων με εξαιρετικά μη γραμμικές σχέσεις.
- Ο XGBoost είναι ένας πολύπλοκος αλγόριθμος και μπορεί να είναι δύσκολο να ερμηνευθούν οι λόγοι πίσω από τις προβλέψεις του. Αυτό μπορεί να είναι ένα μειονέκτημα όταν κάποιος προσπαθεί να λάβει επιχειρηματικές αποφάσεις με βάση τα αποτελέσματα.
- Πριν από την εκπαίδευση, ο αλγόριθμος XGBoost απαιτεί τη διαμόρφωση μιας ποικιλίας υπερπαραμέτρων, συμπεριλαμβανομένου του αριθμού των δέντρων, του ρυθμού εκμάθησης, του βάθους δέντρου, κ.λπ. Η επιλογή του ιδανικού συνόλου υπερπαραμέτρων μπορεί να είναι δύσκολη και χρονοβόρα διαδικασία, ιδιαίτερα αν ο χρήστης δεν είναι εξοικειωμένος με το μοντέλο. [44]

Συμπερασματικά, το μοντέλο XGBoost Regressor είναι ένας ισχυρός αλγόριθμος μηχανικής μάθησης που έχει χρησιμοποιηθεί ευρέως για προβλήματα παλινδρόμησης. Είναι ένα ισχυρό και αποτελεσματικό εργαλείο για εργασίες παλινδρόμησης και έχει χρησιμοποιηθεί για την επίτευξη αποτελεσμάτων τελευταίας τεχνολογίας σε διάφορους τομείς.

4.9 AdaBoostRegressor

Ο AdaBoost (Adaptive Boosting) είναι ένας ensemble αλγόριθμος εκμάθησης που μπορεί να χρησιμοποιηθεί για να συνδυάσει τις προβλέψεις πολλαπλών βασικών μοντέλων. Τα βασικά μοντέλα που χρησιμοποιούνται είναι συνήθως απλά μοντέλα, όπως δέντρα αποφάσεων, αλλά μπορούν να χρησιμοποιηθούν με οποιοδήποτε τύπο μοντέλου που μπορεί να εκπαιδευτεί για να κάνει προβλέψεις. [45]

Ο αρχικός αλγόριθμος AdaBoost, ο οποίος εισήχθη στα μέσα της δεκαετίας του 1990, χρησιμοποίησε δέντρα απόφασης ως βασικά μοντέλα. [46] Τα δέντρα αποφάσεων είναι απλά και εύκολα στην εκπαίδευση μοντέλα τα οποία μπορούν να χρησιμοποιηθούν για ένα ευρύ φάσμα προβλημάτων, συμπεριλαμβανομένης της ταξινόμησης και της παλινδρόμησης. Η απλότητα των δέντρων αποφάσεων τα καθιστά μια καλή επιλογή για τα βασικά μοντέλα στον AdaBoost, καθώς μπορούν να εκπαιδεύονται γρήγορα και να χρησιμοποιηθούν για να κάνουν προβλέψεις με υψηλό βαθμό ακρίβειας.

Για να λύσουν προβλήματα παλινδρόμησης, οι Freu και Schapire δημιούργησαν το μοντέλο AdaboostR. [47] Στη συνέχεια, ο Drucker ανέπτυξε το Adaboost.R2 το οποίο μοντέλο χρησιμοποιείται τώρα στη βιβλιοθήκη sklearn. [48]

Η βασική ιδέα πίσω από το AdaBoost είναι η επαναληπτική εκπαίδευση μιας σειράς βασικών μοντέλων, κάθε φορά εστιάζοντας στις περιπτώσεις εκπαίδευσης που είχαν ταξινομηθεί εσφαλμένα από τα προηγούμενα μοντέλα.

Στην πρώτη επανάληψη, σε όλα τα στιγμιότυπα εκπαίδευσης δίνεται το ίδιο βάρος και ένα βασικό μοντέλο εκπαιδεύεται στα δεδομένα. Στη συνέχεια, το μοντέλο χρησιμοποιείται για να κάνει προβλέψεις σχετικά με τα δεδομένα εκπαίδευσης και οι περιπτώσεις που ταξινομούνται εσφαλμένα λαμβάνουν μεγαλύτερη βαρύτητα στην επόμενη επανάληψη. Αυτή η διαδικασία επαναλαμβάνεται πολλές φορές, με κάθε βασικό μοντέλο να εστιάζει στα σημεία δεδομένων που είχαν ταξινομηθεί εσφαλμένα από τα προηγούμενα μοντέλα. Στο τέλος, οι προβλέψεις όλων των βασικών μοντέλων

συνδυάζονται, συνήθως με σταθμισμένη πλειοψηφία, για να γίνει η τελική πρόβλεψη. Και γι' αυτό ονομάζεται Adaptive Boosting, επειδή ο αλγόριθμος προσαρμόζεται στα δεδομένα εκπαίδευσης σε κάθε επανάληψη.

Ο AdaBoost.R2, γνωστός και ως Adaptive Boosting with R Squared είναι μια επέκταση του τυπικού αλγόριθμου AdaBoost που έχει σχεδιαστεί ειδικά για την πρόβλεψη χρονοσειρών. Όπως και ο τυπικός AdaBoost, χρησιμοποιεί έναν συνδυασμό αδύναμων βασικών μοντέλων. Ωστόσο, αντί να χρησιμοποιεί το ποσοστό εσφαλμένης ταξινόμησης (ή το ποσοστό σφάλματος) ως κριτήριο για την ενημέρωση των βαρών του δείγματος, χρησιμοποιεί τη στατιστική R-squared, ένα μέτρο που δείχνει πόσο καλά ταιριάζουν οι προβλέψεις ενός μοντέλου στις πραγματικές παρατηρήσεις. Αναλυτικότερα, ο αλγόριθμος AdaBoost.R2 εκπαιδεύει ένα βασικό μοντέλο χρησιμοποιώντας ολόκληρο το σύνολο δεδομένων και, στη συνέχεια, υπολογίζει την τιμή του R-squared για το τρέχον μοντέλο. Στη συνέχεια, ενημερώνει τα βάρη για την επόμενη επανάληψη και επαναλαμβάνει αυτή τη διαδικασία για έναν καθορισμένο αριθμό επαναλήψεων ή μέχρι να ικανοποιηθεί κάποιο κριτήριο διακοπής. Μετά από αυτό, συνδυάζει τις προβλέψεις από όλα τα βασικά μοντέλα σε ένα τελικό μοντέλο ensemble με βάρη.

Θετικά:

- Ένα πλεονέκτημα του Adaboost.R2 είναι η ικανότητά του να χειρίζεται δεδομένα υψηλών διαστάσεων. Ως αλγόριθμος που βασίζεται σε δέντρα, ο Adaboost μπορεί να χειριστεί μεγάλο αριθμό λειτουργιών χωρίς την ανάγκη μείωσης διαστάσεων. Αυτό τον καθιστά χρήσιμο αλγόριθμο για την πρόβλεψη χρονοσειρών, καθώς τα δεδομένα χρονοσειρών έχουν συχνά πολλούς πιθανούς προγνωστικούς παράγοντες.
- Ο αλγόριθμος αυτός θεωρείται γενικά ένας αρκετά ακριβής για την πρόβλεψη χρονοσειρών αλγόριθμος, ο οποίος καταγράφει μη γραμμικές σχέσεις μεταξύ των χαρακτηριστικών και της μεταβλητής στόχου. Επιπλέον, συνδυάζοντας τον αλγόριθμο Adaboost με τη στατιστική R2, ο Adaboost.R2 είναι σε θέση να εντοπίσει τα πιο σημαντικά χαρακτηριστικά για την πρόβλεψη. [49]

Αρνητικά:

Ο Adaboost.R2 είναι ένας ισχυρός αλγόριθμος για την πρόβλεψη χρονοσειρών, αλλά έχει ορισμένα μειονεκτήματα που πρέπει να ληφθούν υπόψη.

- Ένα από τα κύρια μειονεκτήματα του Adaboost.R2 για την πρόβλεψη χρονοσειρών είναι η ευαισθησία του σε θορυβώδη δεδομένα. Ο αλγόριθμος Adaboost βασίζεται σε δέντρα αποφάσεων, τα οποία μπορεί να είναι ευαίσθητα σε ακραίες τιμές και θόρυβο στα δεδομένα. Αυτό σημαίνει ότι εάν τα δεδομένα είναι θορυβώδη, ο αλγόριθμος ενδέχεται να μην έχει καλή απόδοση. Επιπλέον, η στατιστική R2 είναι επίσης ευαίσθητη σε ακραίες τιμές και μπορεί να μην είναι η καλύτερη μέτρηση για την αξιολόγηση της απόδοσης του αλγορίθμου παρουσία θορύβου.
- Ένα άλλο μειονέκτημα του Adaboost.R2 είναι η έλλειψη ερμηνείας του. Ο αλγόριθμος Adaboost δημιουργεί ένα σύνθετο μοντέλο συνδυάζοντας πολλαπλά δέντρα απόφασης. Αυτό μπορεί να καταστήσει δύσκολη την κατανόηση των υποκείμενων σχέσεων μεταξύ των χαρακτηριστικών και της μεταβλητής στόχου. Επιπλέον, το στατιστικό R2 δεν είναι καλό μέτρο ερμηνείας του μοντέλου, καθώς δεν παρέχει πληροφορίες για τις σχέσεις μεταξύ των χαρακτηριστικών και της μεταβλητής στόχου.
- Επιπλέον, ο Adaboost.R2 μπορεί να είναι ευαίσθητος στην επιλογή των παραμέτρων, ιδιαίτερα στον ρυθμό εκμάθησης και στον αριθμό των δέντρων που χρησιμοποιούνται. Εάν ο ρυθμός εκμάθησης είναι πολύ υψηλός, ο αλγόριθμος μπορεί να προσαρμοστεί υπερβολικά στα δεδομένα εκπαίδευσης, οδηγώντας σε κακή απόδοση σε νέα δεδομένα. Εάν ο αριθμός των δέντρων είναι πολύ χαμηλός, ο αλγόριθμος ενδέχεται να μην καταγράψει όλες τις σχετικές σχέσεις στα δεδομένα. Αυτό καθιστά σημαντικό τον προσεκτικό συντονισμό των παραμέτρων του αλγορίθμου για την επίτευξη της βέλτιστης απόδοσης.
- Τέλος, το μοντέλο αυτό μπορεί να μην είναι κατάλληλο για χρονοσειρές με πολλαπλά εποχιακά μοτίβα ή να έχει πολύπλοκες τάσεις. Ενώ ο αλγόριθμος Adaboost καταγράφει τη μη γραμμικότητα, ενδέχεται να μην είναι σε θέση να χειριστεί την πολυπλοκότητα πολλαπλών εποχιακών μοτίβων ή πολύπλοκων τάσεων που υπάρχουν σε πολλά δεδομένα χρονοσειρών.

Συνοπτικά, ο Adaboost.R2 είναι ένας ισχυρός αλγόριθμος για την πρόβλεψη χρονοσειρών, αλλά μπορεί να είναι ευαίσθητος σε θορυβώδη δεδομένα. Χαρακτηρίζεται από έλλειψη ερμηνευσιμότητας και ευαισθησία στην επιλογή παραμέτρων. Αυτοί οι παράγοντες θα πρέπει να λαμβάνονται υπόψη όταν αποφασίζεται εάν θα χρησιμοποιηθεί ο Adaboost.R2 για ένα συγκεκριμένο πρόβλημα πρόβλεψης χρονοσειρών.

4.10 HistGradientBoostingRegressor

Το HistGradientBoostingRegressor είναι ένας αλγόριθμος gradient Boosting για προβλήματα παλινδρόμησης και είναι εμπνευσμένος από τη μέθοδο LightGBM. [50,51] Είναι ensemble μοντέλο που συνδυάζει πολλά αδύναμα μοντέλα (όπως δέντρα αποφάσεων) για να δημιουργήσει ένα πιο ισχυρό μοντέλο. Αυτός ο παλινδρομητής είναι πολύ πιο γρήγορος από το GradientBoostingRegressor για μεγάλα σύνολα δεδομένων (n_samples ≥ 10.000). Ειδικότερα, είναι σχεδιασμένος για αποτελεσματικό χειρισμό μεγάλων συνόλων δεδομένων και για πρόβλεψη χρονοσειρών.

Ο αλγόριθμος ξεκινά με ένα μόνο δέντρο που έχει εκπαιδευτεί στα δεδομένα. Σε κάθε επανάληψη, ένα νέο δέντρο προστίθεται στο σύνολο. Το νέο δέντρο είναι εκπαιδευμένο να μειώνει τα υπολειπόμενα λάθη που, συνολικά, έχουν παρατηρηθεί μέχρι εκείνη τη στιγμή. Στη συνέχεια, το νέο δέντρο προστίθεται στο σύνολο και η διαδικασία επαναλαμβάνεται. Το σύνολο των δέντρων συνδυάζεται λαμβάνοντας τον σταθμισμένο μέσο όρο των προβλέψεων όλων των δέντρων. Το βάρος κάθε δέντρου καθορίζεται από την απόδοσή του στο σύνολο επικύρωσης.

Το βασικό χαρακτηριστικό του HistGradientBoostingRegressor είναι η χρήση μιας προσέγγισης βασισμένης σε ιστογράμματα. Αυτό περιλαμβάνει τη δέσμευση των δεδομένων και τον κατά προσέγγιση υπολογισμό των κλίσεων χρησιμοποιώντας μόνο τα δεσμευμένα δεδομένα, αντί για τη χρήση του πλήρους συνόλου δεδομένων. Αυτό μπορεί να μειώσει σημαντικά τη μνήμη και τις υπολογιστικές απαιτήσεις του αλγορίθμου, καθιστώντας τον πιο κατάλληλο για μεγάλα σύνολα δεδομένων. [52]

Θετικά:

- Ο HistGradientBoostingRegressor μπορεί να αντιμετωπίσει τιμές που λείπουν και κατηγορικές μεταβλητές, γεγονός που τον καθιστά εξαιρετική επιλογή για πρόβλεψη χρονοσειρών.
- Επίσης, έχει υψηλή ακρίβεια πρόβλεψης, καθώς δημιουργεί ένα σύνολο δέντρων αποφάσεων και κάθε δέντρο εκπαιδεύεται να διορθώνει τα λάθη που έγιναν από τα προηγούμενα δέντρα του συνόλου. Αυτό επιτρέπει στο μοντέλο να μαθαίνει από τα λάθη του και να βελτιώνει τις προβλέψεις του με την πάροδο του χρόνου.

Αρνητικά:

- Ο HistGradientBoostingRegressor είναι ένας πολύπλοκος αλγόριθμος που απαιτεί καλή κατανόηση των δέντρων αποφάσεων για αποτελεσματική χρήση. Απαιτεί επίσης λεπτομερή ρύθμιση πολλών παραμέτρων, όπως ο αριθμός των δέντρων, το βάθος των δέντρων και ο ρυθμός εκμάθησης. Αυτό το γεγονός μπορεί να δυσκολέψει τη χρήση του.
- Όπως κάθε ensemble μοντέλο, ο HistGradientBoostingRegressor κινδυνεύει να υπερπροσαρμοστεί εάν ο αριθμός των δέντρων στο σύνολο είναι πολύ μεγάλος ή το βάθος των δέντρων είναι πολύ μεγάλο. Αυτό μπορεί να οδηγήσει σε κακή απόδοση γενίκευσης σε νέα δεδομένα.
- Επιπροσθέτως, εφόσον ο HistGradientBoostingRegressor συνίσταται σε ένα σύνολο δέντρων αποφάσεων, μπορεί να είναι δύσκολο να ερμηνευτούν τα μεμονωμένα δέντρα αποφάσεων και να γίνει κατανοητό πώς το μοντέλο κάνει τις προβλέψεις του.
- Κλείνοντας, ο HistGradientBoostingRegressor ενδέχεται να είναι υπολογιστικά «ακριβός» στην εκπαίδευση, ιδιαίτερα όταν ο αριθμός των δέντρων στο σύνολο είναι μεγάλος ή το βάθος των δέντρων είναι μεγάλο. Αυτό μπορεί να καταστήσει χρονοβόρα την εκπαίδευση ενός μοντέλου σε μεγάλα σύνολα δεδομένων.

Συμπερασματικά, ο HistGradientBoostingRegressor είναι ένας ευέλικτος και ισχυρός αλγόριθμος για την πρόβλεψη χρονοσειρών. Συνδυάζει τα πλεονεκτήματα του gradient boosting με μια προσέγγιση που βασίζεται σε ιστογράμματα, γεγονός που τον καθιστά αποτελεσματικό και κατάλληλο για μεγάλα σύνολα δεδομένων. Αν και μπορεί να απαιτεί ένα ορισμένο επίπεδο τεχνογνωσίας και λεπτομέρειας για να χρησιμοποιηθεί αποτελεσματικά, μπορεί να είναι ένα πολύτιμο εργαλείο για την επίτευξη ακριβών και αξιόπιστων προβλέψεων όταν χρησιμοποιείται κατάλληλα.

5 Υπερπαράμετροι

Στη μηχανική μάθηση, ένα μοντέλο έχει δύο τύπους παραμέτρων: αυτές που μαθαίνονται από τα δεδομένα και υπερπαραμέτρους που ορίζονται από τον επαγγελματία. Οι υπερπαράμετροι είναι τιμές που ορίζονται πριν από την εκπαίδευση του μοντέλου και ελέγχουν διάφορες πτυχές της εκπαιδευτικής διαδικασίας και της συμπεριφοράς του μοντέλου. Παραδείγματα υπερπαραμέτρων περιλαμβάνουν τον ρυθμό εκμάθησης αλλά και τον αριθμό των κρυφών στρωμάτων σε ένα νευρωνικό δίκτυο. Ο συντονισμός των υπερπαραμέτρων είναι ένα ουσιαστικό βήμα στη διαδικασία εκπαίδευσης ενός μοντέλου μηχανικής μάθησης, καθώς η επιλογή των υπερπαραμέτρων μπορεί να έχει σημαντικό αντίκτυπο στην απόδοση του εκπαιδευμένου μοντέλου.

5.1 Συναρτήσεις Ενεργοποίησης

Μια συνάρτηση ενεργοποίησης είναι μια μαθηματική συνάρτηση που εφαρμόζεται στην έξοδο ενός νευρώνα, η οποία καθορίζει την έξοδο δεδομένης μιας εισόδου ή ενός συνόλου εισόδων.

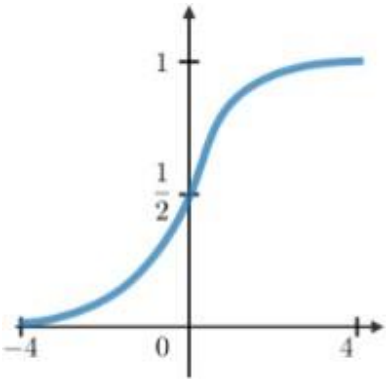
Η συνάρτηση ενεργοποίησης είναι ένα σημαντικό στοιχείο ενός νευρωνικού δικτύου, επειδή επιτρέπει στο δίκτυο να μαθαίνει και να δημιουργεί περίπλοκα, μη γραμμικά όρια απόφασης. Οι συναρτήσεις ενεργοποίησης είναι σημαντικές επειδή εισάγουν μη γραμμικότητα στην έξοδο ενός νευρώνα. Αυτή η μη γραμμικότητα επιτρέπει στα νευρωνικά δίκτυα να μοντελοποιούν πολύπλοκες σχέσεις μεταξύ εισόδων και εξόδων και να μαθαίνουν πιο εκφραστικά μοντέλα.

Είναι σημαντικό να σημειωθεί ότι διαφορετικές συναρτήσεις ενεργοποίησης είναι πιο κατάλληλες για διαφορετικούς τύπους εργασιών και αρχιτεκτονικών. Επίσης, η επιλογή της συνάρτησης ενεργοποίησης μπορεί να έχει μεγάλο αντίκτυπο στην απόδοση του μοντέλου. Επομένως, είναι σημαντικό να γίνει πειραματισμός με διαφορετικές επιλογές, με σκοπό την εύρεση της καλύτερης επιλογής για ένα δεδομένο πρόβλημα.

5.1.1 Sigmoid

Η σιγμοειδής συνάρτηση ή αλλιώς λογιστική (Logistic), είναι μια ευρέως χρησιμοποιούμενη συνάρτηση ενεργοποίησης σε νευρωνικά δίκτυα. Πρόκειται για μία από τις πρώτες συναρτήσεις ενεργοποίησης που χρησιμοποιήθηκε ευρέως, καθώς η παράγωγός της είναι πολύ εύκολα υπολογίσιμη. Η συνάρτηση αντιστοιχίζει οποιαδήποτε τιμή εισόδου σε μια τιμή μεταξύ 0 και 1, η οποία μπορεί να ερμηνευτεί ως πιθανότητα. Η σιγμοειδής συνάρτηση ορίζεται ως: [53]

$$sigmoid(x) = \frac{1}{1 + e^{-x}}$$



Σχήμα 7: Sigmoid [80]

Όπου x είναι η είσοδος στον νευρώνα και e η βάση του φυσικού λογάριθμου. Αξίζει να αναφερθεί ότι η συνάρτηση έχει σχήμα S. Καθώς η είσοδος x γίνεται μεγαλύτερη σε μέγεθος, η έξοδος της συνάρτησης πλησιάζει το 0 ή το 1. Αυτό σημαίνει ότι για μεγάλες θετικές τιμές εισόδου, η έξοδος θα είναι κοντά στο 1 και για μεγάλες αρνητικές τιμές εισόδου, η έξοδος θα είναι κοντά στο 0.

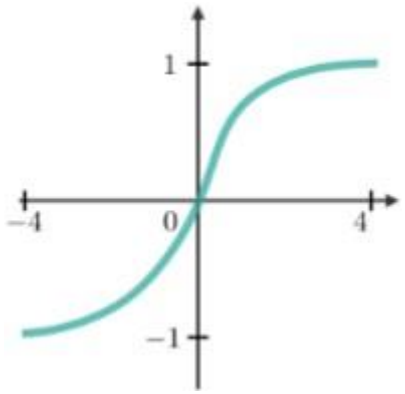
Το κύριο πλεονέκτημα της χρήσης της σιγμοειδούς συνάρτησης είναι ότι επιτρέπει στο νευρωνικό δίκτυο να εξάγει μια τιμή πιθανότητας, η οποία μπορεί να χρησιμοποιηθεί για τη λήψη μιας απόφασης δυαδικής ταξινόμησης. [54] Ωστόσο, η σιγμοειδής λειτουργία έχει επίσης ορισμένα μειονεκτήματα. Πρώτο είναι το πρόβλημα των εξαφανιζόμενων κλίσεων (vanishing gradients problem), όπου οι διαβαθμίσεις της συνάρτησης γίνονται πολύ μικρές, γεγονός που επιβραδύνει τη διαδικασία εκμάθησης. Ένα άλλο πρόβλημα είναι ότι μπορεί να προκαλέσει πολύ μεγάλη ενημέρωση των βαρών.

5.1.2 Tanh

Η συνάρτηση ενεργοποίησης υπερβολικής εφαπτομένης (tanh) είναι μια μη γραμμική συνάρτηση που χρησιμοποιείται συνήθως στα νευρωνικά δίκτυα. Είναι μια παραλλαγή της συνάρτησης ενεργοποίησης σιγμοειδούς και ορίζεται μαθηματικά ως: [55]

$$tanh(x) = \frac{2}{1 + e^{-2x}} - 1$$

Η συνάρτηση αντιστοιχίζει οποιαδήποτε τιμή εισόδου x σε μια τιμή μεταξύ -1 και 1. Αυτό την καθιστά ιδιαίτερα χρήσιμη για αρχιτεκτονικές νευρωνικών δικτύων που απαιτούν οριοθέτηση τιμών εξόδου. Η συνάρτηση tanh έχει παρόμοιο σχήμα με τη σιγμοειδή συνάρτηση, δηλαδή είναι σχήματος S, αλλά έχει ως άξονα συμμετρίας τον άξονα x. Αυτό επιτρέπει στη συνάρτηση να εξάγει θετικές και αρνητικές τιμές, γεγονός που την καθιστά χρήσιμη για τη μοντελοποίηση δεδομένων που μπορεί να έχουν θετικές και αρνητικές τιμές.



Σχήμα 8: Tanh [80]

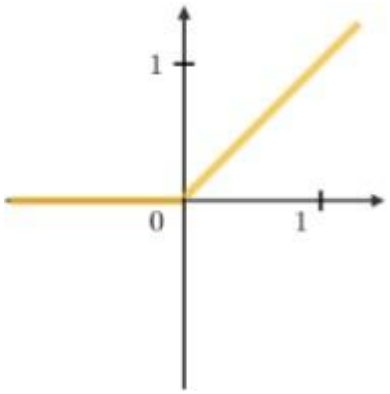
Ένα από τα βασικά πλεονεκτήματα της συνάρτησης tanh έναντι της σιγμοειδούς συνάρτησης είναι ότι οι τιμές εξόδου της επικεντρώνονται γύρω από το μηδέν, γεγονός που μπορεί να βοηθήσει στην αντιμετώπιση του προβλήματος εξαφανιζόμενων κλίσεων (vanishing gradients problem).

5.1.3 ReLU

Η συνάρτηση ενεργοποίησης Rectified Linear Unit (ReLU) είναι μια μη γραμμική συνάρτηση. Ορίζεται μαθηματικά ως: [56]

$$relu(x) = \max(0, x)$$

Η συνάρτηση αντιστοιχίζει οποιαδήποτε τιμή εισόδου, x, σε μια τιμή μεταξύ 0 και x. Εάν το x είναι μεγαλύτερο από 0, θα βγάζει x, διαφορετικά θα επιστρέφει 0. Πιο συνοπτικά, διατηρεί τις θετικές εισόδους και μηδενίζει τις αρνητικές. Αυτό την καθιστά ιδιαίτερα χρήσιμη για αρχιτεκτονικές νευρωνικών δικτύων που απαιτούν οι τιμές εξόδου να είναι μη αρνητικές.



Σχήμα 9: ReLU [80]

Η ReLU είναι υπολογιστικά αποδοτική και δεν απαιτεί πολύπλοκες μαθηματικές πράξεις όπως εκθετικές, επομένως είναι ταχύτερη στον υπολογισμό σε σύγκριση με άλλες συναρτήσεις ενεργοποίησης, όπως η σιγμοειδής ή η tanh. Επιπλέον, βοηθά στην επίλυση του προβλήματος των εξαφανιζόμενων κλίσεων (vanishing gradients problem).

Αξίζει να αναφερθεί ότι χρησιμοποιείται εκτενώς σε βαθιά νευρωνικά δίκτυα επειδή είναι υπολογιστικά αποδοτική, εφόσον επιτρέπει ταχύτερη εκπαίδευση. Είναι ιδιαίτερα χρήσιμη στα κρυφά στρώματα ενός νευρωνικού δικτύου, όπου βοηθά στην εκμάθηση των μη γραμμικών ορίων απόφασης.

5.1.4 Softmax

Η συνάρτηση softmax είναι μια δημοφιλής συνάρτηση ενεργοποίησης που χρησιμοποιείται συχνά στο επίπεδο εξόδου ενός νευρωνικού δικτύου για προβλήματα ταξινόμησης πολλαπλών κλάσεων. Ορίζεται μαθηματικά ως:

$$Softmax(x) = \frac{e^{x_i}}{\sum_{j=1}^{N} e^{x_j}}$$

όπου x_i είναι το i-οστό στοιχείο του διανύσματος εισόδου x και ο παρονομαστής είναι το άθροισμα της εκθετικής τιμής όλων των στοιχείων στο διάνυσμα εισόδου.

Η συνάρτηση softmax αντιστοιχίζει οποιοδήποτε διάνυσμα εισόδου πραγματικών αριθμών σε μια κατανομή πιθανότητας στις K διαφορετικές κλάσεις, όπου K είναι ο αριθμός των κλάσεων. Αυτό γίνεται λαμβάνοντας την εκθετική τιμή για κάθε είσοδο, που έχει ως αποτέλεσμα θετικές τιμές, και στη συνέχεια διαιρώντας κάθε τιμή με το άθροισμα όλων των εκθετικών τιμών.

Οι κύριες ιδιότητες της συνάρτησης softmax είναι:

- Η έξοδος της συνάρτησης είναι μια κατανομή πιθανότητας στις K διαφορετικές κλάσεις, που σημαίνει ότι το άθροισμα των τιμών εξόδου είναι ίσο με 1.
- Κάθε τιμή εξόδου είναι στην περιοχή [0, 1], η οποία μπορεί να ερμηνευθεί ως η πιθανότητα το διάνυσμα εισόδου να ανήκει στην αντίστοιχη κλάση.
- Η συνάρτηση είναι διαφοροποιήσιμη, γεγονός που της επιτρέπει να χρησιμοποιείται σε αλγόριθμους βελτιστοποίησης που βασίζονται σε κλίση.

Εξαιτίας αυτών των ιδιοτήτων, η συνάρτηση softmax χρησιμοποιείται συχνά στο επίπεδο εξόδου ενός νευρωνικού δικτύου για προβλήματα ταξινόμησης πολλαπλών κλάσεων, όπου ο στόχος είναι να παραχθεί μια κατανομή πιθανότητας στις κλάσεις. Η κλάση με τη μεγαλύτερη πιθανότητα θεωρείται η τελική έξοδος ή πρόβλεψη του δικτύου.

5.2 Αλγόριθμοι Ελαχιστοποίησης Σφάλματος

Οι αλγόριθμοι αυτοί παίζουν σημαντικό ρόλο στην εκπαίδευση των νευρωνικών δικτύων. Είναι υπεύθυνοι για την ενημέρωση των παραμέτρων του μοντέλου με τρόπο που ελαχιστοποιεί το σφάλμα μεταξύ της προβλεπόμενης και της πραγματικής εξόδου. Υπάρχουν διάφοροι τύποι αλγόριθμων ελαχιστοποίησης σφάλματος, καθένας με τα δικά του δυνατά και αδύνατα σημεία. Η επιλογή του βελτιστοποιητή θα εξαρτηθεί από το συγκεκριμένο πρόβλημα και τα χαρακτηριστικά των δεδομένων. Στη συνέχεια, θα ακολουθήσει συνοπτική παρουσίαση των αλγορίθμων Gradient Descent, Stochastic Gradient Descent, RMSProp και Adam.

5.2.1 Gradient Descent

Ο στόχος του αλγορίθμου είναι να βρει το σύνολο των παραμέτρων που ελαχιστοποιούν τη συνάρτηση απώλειας. Αυτό γίνεται με την επαναληπτική ενημέρωση των παραμέτρων του μοντέλου προς την κατεύθυνση της αρνητικής κλίσης της συνάρτησης απώλειας σε σχέση με τις παραμέτρους.

Η βασική ιδέα είναι η εκκίνηση με ένα αρχικό σύνολο παραμέτρων και στη συνέχεια η επαναληπτική ενημέρωση των παραμέτρων.

Ο αλγόριθμος προχωρά ως εξής:

1. Αρχικοποιήση των παραμέτρων του μοντέλου με τυχαίες τιμές.
2. Υπολογισμός της συνάρτηση απώλειας για το τρέχον σύνολο παραμέτρων.
3. Υπολογισμός της κλίσης της συνάρτησης απώλειας σε σχέση με τις παραμέτρους.
4. Ενημέρωση των παραμέτρων αφαιρώντας μια μικρή τιμή της κλίσης από το τρέχον σύνολο παραμέτρων. Αυτό το βήμα είναι γνωστό ως κανόνας ενημέρωσης (update rule) και συνήθως έχει τη μορφή:
$$a_{n+1} = a_n - learning\ rate * gradient$$
5. Επανάληψη των βημάτων 2 έως 4 για σταθερό αριθμό επαναλήψεων ή έως ότου η συνάρτηση απώλειας συγκλίνει στο ελάχιστο.

Ο ρυθμός εκμάθησης (learning rate) είναι μια υπερπαράμετρος που ελέγχει το μέγεθος βήματος της ενημέρωσης. Εάν ο ρυθμός εκμάθησης είναι πολύ υψηλός, το μοντέλο μπορεί να υπερβεί τη βέλτιστη λύση, ενώ εάν είναι πολύ χαμηλό, το μοντέλο μπορεί να συγκλίνει πολύ αργά. Ο συγκεκριμένος αλγόριθμος είναι πολύ σημαντικός, καθώς περιγράφει τη βασική ιδέα της στρατηγικής που ακολουθείται και σε όλους τους υπόλοιπους αλγορίθμους βελτιστοποίησης που θα αναφερθούν.

5.2.2 Stochastic Gradient Descent

Ο Stochastic gradient descent (SGD) είναι μια παραλλαγή του αλγορίθμου gradient descent που χρησιμοποιείται ευρέως στην εκπαίδευση νευρωνικών δικτύων. Η βασική ιδέα πίσω από τον SGD είναι η ίδια με του Gradient Descent. Συγκεκριμένα, ο σκοπός, είναι να βρεθεί το σύνολο των παραμέτρων που ελαχιστοποιούν τη συνάρτηση απώλειας. Ωστόσο, αντί να υπολογίζει την κλίση της συνάρτησης απώλειας σε σχέση με τις παραμέτρους που χρησιμοποιούν ολόκληρο το σύνολο δεδομένων, ο SGD χρησιμοποιεί ένα μόνο παράδειγμα εκπαίδευσης (ή έναν μικρό αριθμό παραδειγμάτων) για να ενημερώσει τις παραμέτρους.

Ο αλγόριθμος προχωρά ως εξής:

1. Αρχικοποίηση των παραμέτρων του μοντέλου με τυχαίες τιμές.
2. Επιλογή τυχαίου παραδείγματος εκπαίδευσης από το σύνολο δεδομένων.
3. Υπολογισμός της συνάρτησης απώλειας για το τρέχον σύνολο παραμέτρων και το επιλεγμένο παράδειγμα εκπαίδευσης.
4. Υπολογισμός της κλίσης της συνάρτησης απώλειας σε σχέση με τις παραμέτρους για το επιλεγμένο παράδειγμα εκπαίδευσης.
5. Ενημέρωση των παραμέτρων αφαιρώντας μια μικρή τιμή της κλίσης από το τρέχον σύνολο παραμέτρων.
6. Επανάληψη των βημάτων 2 έως 5 για σταθερό αριθμό επαναλήψεων ή έως ότου η συνάρτηση απώλειας συγκλίνει στο ελάχιστο.

Ωστόσο, οι ενημερώσεις βασίζονται σε ένα μόνο παράδειγμα εκπαίδευσης, το οποίο εισάγει πολύ θόρυβο στη διαδικασία βελτιστοποίησης. Αυτός ο θόρυβος μπορεί να προκαλέσει ταλάντωση του αλγόριθμου και να χάσει τη βέλτιστη λύση. Για να μετριαστεί αυτό το πρόβλημα, ο ρυθμός μάθησης συνήθως μειώνεται με την πάροδο του χρόνου, σε μια διαδικασία που ονομάζεται learning rate decay.

5.2.3 RMSProp

Είναι μια επέκταση του αλγορίθμου gradient descent, που βοηθά στην επιτάχυνση της σύγκλισης της διαδικασίας βελτιστοποίησης και στην αποτροπή της υπέρβασης της βέλτιστης λύσης. Η βασική ιδέα πίσω από τον RMSprop (Root Mean Square Propagation) είναι η προσαρμογή του ρυθμού εκμάθησής του με βάση τις πληροφορίες ιστορικής κλίσης. Χρησιμοποιεί το ριζικό μέσο τετράγωνο (RMS) της διαβάθμισης με την πάροδο του χρόνου για να κλιμακώσει τον ρυθμό εκμάθησης. [57]

Ο αλγόριθμος προχωρά ως εξής:

1. Αρχικοποίηση των παραμέτρων του μοντέλου με τυχαίες τιμές.
2. Αρχικοποίηση μιας κρυφής μνήμης για κάθε παράμετρο για αποθήκευση των πληροφορίων κλίσης.
3. Για κάθε επανάληψη εκπαίδευσης, υπολογισμός της κλίσης της συνάρτησης απώλειας σε σχέση με τις παραμέτρους για το τρέχον παράδειγμα εκπαίδευσης.
4. Ενημέρωση της μνήμης για κάθε παράμετρο υπολογίζοντας τον κινητό μέσο όρο του τετραγώνου της κλίσης.
5. Υπολογισμός του κανόνα ενημέρωσης για κάθε παράμετρο, διαιρώντας την κλίση με την τετραγωνική ρίζα της μνήμης.
6. Επανάληψη των βημάτων 3 έως 5 για σταθερό αριθμό επαναλήψεων ή έως ότου η συνάρτηση απώλειας συγκλίνει στο ελάχιστο.

Ένα από τα βασικά πλεονεκτήματα του RMSprop σε σχέση με άλλους αλγόριθμους βελτιστοποίησης είναι ότι μπορεί να χειριστεί τις διαφορετικές κλίμακες και κλίσεις των παραμέτρων. Επιπλέον, ο αλγόριθμος μπορεί να προσαρμόσει αυτόματα τον ρυθμό εκμάθησης για κάθε παράμετρο, κάτι που μπορεί να βοηθήσει στη σύγκλιση σε μια καλύτερη λύση.

5.2.4 Adam

Ο Adam (Adaptive Moment Estimation) είναι ένας αλγόριθμος βελτιστοποίησης που χρησιμοποιείται ευρέως στην εκπαίδευση νευρωνικών δικτύων. Είναι μια επέκταση του αλγορίθμου Stochastic Gradient Descent (SGD) που ενσωματώνει πληροφορίες σχετικά με τις προηγούμενες κλίσεις για να προσαρμόσει τον ρυθμό εκμάθησης. [58] Επιχειρεί να συνδυάσει τα πλεονεκτήματα των δύο προγενέστερων αλγορίθμων, του Adagrad και του RMSProp.

Ο αλγόριθμος προχωρά ως εξής

1. Αρχικοποίηση των παραμέτρων του μοντέλου με τυχαίες τιμές.
2. Αρχικοποίηση των δύο κινητών μέσων όρων, έναν για την κλίση και έναν για την τετραγωνική κλίση.
3. Για κάθε επανάληψη εκπαίδευσης, υπολογισμός της κλίσης της συνάρτησης απώλειας για το τρέχον παράδειγμα εκπαίδευσης.
4. Ενημέρωση του κινούμενου μέσου όρου κλίσης και της τετραγωνικής κλίσης.
5. Διόρθωση της προκατάληψη των κινητών μέσων όρων υπολογίζοντας τις διορθωμένες εκδόσεις.
6. Υπολογισμός του κανόνα ενημέρωσης για κάθε παράμετρο.
7. Επανάληψη των βημάτων τα βήματα 3 έως 7 για σταθερό αριθμό επαναλήψεων ή έως ότου η συνάρτηση απώλειας συγκλίνει στο ελάχιστο.

Ο Adam είναι γνωστό ότι λειτουργεί καλά στην πράξη και χρησιμοποιείται συχνά σε ένα ευρύ φάσμα εργασιών βαθιάς μάθησης. Ένα από τα βασικά πλεονεκτήματά του έναντι άλλων αλγορίθμων βελτιστοποίησης είναι ότι μπορεί να χειριστεί τις διαφορετικές κλίμακες των παραμέτρων και των κλίσεων που εμφανίζονται συχνά σε βαθιά νευρωνικά δίκτυα. Επιπλέον, ο αλγόριθμος μπορεί να προσαρμόσει αυτόματα τον ρυθμό εκμάθησης για κάθε παράμετρο, κάτι που μπορεί να βοηθήσει στη σύγκλιση σε μια καλύτερη λύση.

5.3 Βελτιστοποίηση

Στο σημείο αυτό, εφόσον έγινε αναφορά σε μερικές υπερπαραμέτρους, σημαντικό είναι να γίνει αναφορά στον τρόπο βελτιστοποίησής τους. Ο στόχος της βελτιστοποίησης (hyperparameter tuning) είναι να βρεθεί ο συνδυασμός των υπερπαραμέτρων που έχουν ως αποτέλεσμα την καλύτερη απόδοση του μοντέλου σε ένα σύνολο επικύρωσης ή σύνολο δοκιμής.

Η βελτιστοποίηση υπερπαραμέτρων εκτελείται συνήθως μέσω ενός συνδυασμού δοκιμής και λάθους αλλά και κατανόησης του υποκείμενου μοντέλου και του προβλήματος. Μπορεί να είναι μια χρονοβόρα διαδικασία, καθώς απαιτεί εκπαίδευση του μοντέλου πολλές φορές με διαφορετικές ρυθμίσεις υπερπαραμέτρων.

Υπάρχουν πολλές διαφορετικές τεχνικές συντονισμού υπερπαραμέτρων, καθεμία με τα δικά της πλεονεκτήματα και μειονεκτήματα. Στη συνέχεια, θα ακολουθήσει περιγραφή για μερικές από τις πιο διαδεδομένες τεχνικές.

5.3.1 Grid Search

Η αναζήτηση πλέγματος (Grid Search) είναι μια απλή και ευρέως χρησιμοποιούμενη μέθοδος βελτιστοποίησης των υπερπαραμέτρων στη μηχανική μάθηση. Η ιδέα πίσω από την αναζήτηση πλέγματος είναι να καθοριστεί ένα εύρος τιμών για κάθε υπερπαράμετρο και στη συνέχεια να εκπαιδευτεί το μοντέλο με όλους τους πιθανούς συνδυασμούς υπερπαραμέτρων. Ο αλγόριθμος προχωρά ως εξής:

1. Ορισμός πλέγματος υπερπαραμέτρων, το οποίο περιλαμβάνει το εύρος τιμών για κάθε μια.
2. Εκπαίδευση του μοντέλου με όλους τους πιθανούς συνδυασμούς υπερπαραμέτρων.
3. Για κάθε συνδυασμό υπερπαραμέτρων, αξιολόγηση της απόδοσης του μοντέλου σε ένα σύνολο επικύρωσης χρησιμοποιώντας μια μέτρηση απόδοσης όπως η ακρίβεια (precision), το F1-score ή μέθοδος AUC.
4. Τέλος, επιλογή του συνδυασμού των υπερπαραμέτρων που έχουν ως αποτέλεσμα την καλύτερη απόδοση στο σύνολο επικύρωσης.

Ένα από τα κύρια πλεονεκτήματα της αναζήτησης πλέγματος είναι ότι η μέθοδος αυτή είναι απλή και εύκολη στην εφαρμογή. Ωστόσο, μπορεί να είναι υπολογιστικά ακριβή, ειδικά όταν ο αριθμός των υπερπαραμέτρων ή το εύρος των τιμών είναι μεγάλο. Επιπλέον, η αναζήτηση πλέγματος μπορεί να είναι επιρρεπής σε υπερβολική προσαρμογή, ιδιαίτερα εάν ο αριθμός των συνδυασμών είναι μεγάλος και η μέτρηση απόδοσης είναι θορυβώδης.

5.3.2 Random Search

Η τυχαία αναζήτηση (Random Search) είναι μια μέθοδος βελτιστοποίησης των υπερπαραμέτρων που είναι παρόμοια με την αναζήτηση πλέγματος. Όμως, αντί να γίνεται δοκιμή όλων των πιθανών συνδυασμών, λαμβάνονται τυχαία δείγματα υπερπαραμέτρων από μια προκαθορισμένη κατανομή.

Η τυχαία αναζήτηση έχει πολλά πλεονεκτήματα σε σχέση με την αναζήτηση πλέγματος. Πρώτον, μπορεί να είναι πιο αποδοτική υπολογιστικά, ειδικά όταν ο αριθμός των υπερπαραμέτρων ή το εύρος τιμών είναι μεγάλο. Δεύτερον, μπορεί να εξερευνήσει τον χώρο των υπερπαραμέτρων πιο αποτελεσματικά, ιδιαίτερα όταν η κατανομή των βέλτιστων υπερπαραμέτρων δεν είναι γνωστή. Τρίτον, μπορεί να μειώσει τον κίνδυνο υπερβολικής προσαρμογής, καθώς το μοντέλο εκπαιδεύεται με ποικίλες υπερπαραμέτρους.

Ωστόσο, η τυχαία αναζήτηση έχει επίσης ορισμένα μειονεκτήματα. Μπορεί να είναι υπολογιστικά ακριβή εάν ο αριθμός των επαναλήψεων είναι μεγάλος. Επιπλέον, μπορεί να μην βρει το συνολικό βέλτιστο εάν ο αριθμός των επαναλήψεων δεν είναι επαρκής. Κλείνοντας, η επιλογή των κατανομών για κάθε υπερπαράμετρο μπορεί να έχει σημαντικό αντίκτυπο στα αποτελέσματα και μπορεί να είναι δύσκολο να προσδιοριστούν οι βέλτιστες κατανομές.

5.4 Μετρικές

Στο κομμάτι αυτό θα γίνει παρουσίαση των μετρικών που χρησιμοποιήθηκαν για την αξιολόγηση των αποτελεσμάτων.

Η γνώση των καλύτερων μετρικών για την αξιολόγηση ενός μοντέλου είναι ζωτικής σημασίας για τη δημιουργία ακριβών και αποτελεσματικών μοντέλων μηχανικής εκμάθησης. Επιτρέπει μια δίκαιη σύγκριση διαφορετικών μοντέλων και, επιπλέον, βοηθάει στον εντοπισμό των αδυναμιών ενός μοντέλου. Εφόσον το υπό εξέταση πρόβλημα είναι η πρόβλεψη χρονοσειράς, επιλέχθηκαν οι μετρικές RMSE, MAPE και MAE.

Στις παρακάτω μαθηματικές εξισώσεις των μετρικών θεωρούνται οι εξής μεταβλητές:

Σύμβολο	Μεταβλητή
$\hat{y}$	Προβλεπόμενη τιμή
y	Πραγματική τιμή
n	Πλήθος τιμών

5.4.1 RMSE

Αναλυτικότερα, το RMSE (Root Mean Squared Error) είναι ένα ευρέως χρησιμοποιούμενο μέτρο της διαφοράς μεταξύ προβλεπόμενων και πραγματικών τιμών. Η έννοια του RMSE, καθώς και η μαθηματική του διατύπωση, είναι καλά εδραιωμένη και μπορεί να βρεθεί σε πολλά εγχειρίδια στατιστικής και ερευνητικές εργασίες. Είναι ένα μέτρο του πόσο καλά ένα μοντέλο είναι σε θέση να προβλέψει τη μεταβλητή στόχο.

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{n}}$$

Το RMSE είναι μια δημοφιλής μέτρηση αξιολόγησης για προβλήματα παλινδρόμησης, καθώς περιορίζει τα μεγάλα σφάλματα. [59] Αυτό το καθιστά μια καλή επιλογή για την αξιολόγηση της απόδοσης των μοντέλων που χρησιμοποιούνται για εργασίες πρόβλεψης.

Είναι σημαντικό να σημειωθεί ότι όσο χαμηλότερο είναι το RMSE, τόσο καλύτερα ανταποκρίνεται το μοντέλο στην πρόβλεψη της μεταβλητής στόχου. Ένα μοντέλο με χαμηλό RMSE έχει μια μικρή διαφορά μεταξύ των προβλεπόμενων και των πραγματικών τιμών και το αντίστροφο. Οι τιμές που παίρνει η μετρική αυτή είναι πάντα μη αρνητικές και μια τιμή 0 υποδηλώνει τέλεια προσαρμογή στα δεδομένα.

5.4.2 MAPE

Το μέσο απόλυτο ποσοστό σφάλματος (Mean Absolute Percentage Error - MAPE) είναι μια ευρέως χρησιμοποιούμενη μέτρηση αξιολόγησης για προβλήματα παλινδρόμησης, ιδιαίτερα στον τομέα των οικονομικών και των οικονομικών. Το MAPE χρησιμοποιείται ευρέως σε διάφορους τομείς και βιομηχανίες και έχει χρησιμοποιηθεί σε πολλές ερευνητικές εργασίες και εγχειρίδια.

Είναι σημαντικό να αναφέρουμε ότι το MAPE είναι ένα σχετικό μέτρο σφάλματος, δίνει το ποσοστό του σφάλματος μεταξύ των προβλεπόμενων και των πραγματικών τιμών. Το γεγονός αυτό είναι χρήσιμο όταν η κλίμακα της μεταβλητής δεν είναι ίδια σε όλες τις παρατηρήσεις.

$$MAPE = \frac{100\%}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right|$$

Αξίζει να τονιστεί το γεγονός ότι η μετρική αυτή δεν ορίζεται όταν η πραγματική τιμή (y) είναι μηδέν. Το γεγονός αυτό μπορεί να καταστήσει δύσκολη τη χρήση για προβλήματα όπου η μεταβλητή στόχος μπορεί να λάβει την τιμή μηδέν. Σε άλλες περιπτώσεις πρέπει να ληφθεί αυτό υπόψιν κατά την προεπεξεργασία των δεδομένων.

5.4.3 MAE

Το μέσο απόλυτο σφάλμα (Mean Absolute Error - MAE) είναι μια κοινή μέτρηση που χρησιμοποιείται για τη μέτρηση της απόδοσης ενός μοντέλου σε προβλήματα παλινδρόμησης. Είναι η μέση διαφορά μεταξύ των προβλεπόμενων και των πραγματικών τιμών. Όσο μικρότερο είναι το MAE, τόσο καλύτερη είναι η απόδοση του μοντέλου.

$$MAE = \frac{\sum_{i=1}^{n} |\hat{y}_i - y_i|}{n}$$

Επειδή χρησιμοποιεί απόλυτες τιμές, το MAE δεν επηρεάζεται από ακραίες τιμές τόσο όσο άλλες μετρήσεις. Είναι μια ερμηνεύσιμη μέτρηση και είναι εύκολο να υπολογιστεί. Η μετρική αυτή είναι μια καλή επιλογή όταν τα δεδομένα έχουν ακραίες τιμές και ο στόχος είναι να γίνει κατανόηση του μεγέθους των σφαλμάτων, ανεξάρτητα από την κατεύθυνση.

6 Περιβάλλον Υλοποίησης

Η διαδικασία υλοποίησης της μελέτης είναι το αντικείμενο αυτού του κεφαλαίου. Αρχικά, θα παρουσιαστούν πληροφορίες για την γλώσσα προγραμματισμού που επιλέχθηκε. Στη συνέχεια, θα γίνει μια σύντομη περιγραφή των βασικών βιβλιοθηκών που χρησιμοποιήθηκαν.

6.1 Γλώσσα προγραμματισμού

Η γλώσσα προγραμματισμού που επιλέχθηκε να χρησιμοποιηθεί είναι η Python. Η Python είναι μια εξαιρετικά ευέλικτη και δημοφιλής γλώσσα προγραμματισμού στην κοινότητα της επιστήμης δεδομένων. Η γλώσσα αυτή κυκλοφόρησε για πρώτη φορά τον Φεβρουάριου του 1991 από τον Guido van Rossum. [60] Είναι μια διερμηνευμένη, υψηλού επιπέδου, γενικής χρήσης γλώσσα προγραμματισμού. Η σχεδιαστική φιλοσοφία της Python δίνει έμφαση στην αναγνωσιμότητα κώδικα με την αξιοσημείωτη χρήση σημαντικού κενού χώρου. Οι γλωσσικές κατασκευές και η αντικειμενοστραφής προσέγγισή της στοχεύουν να βοηθήσουν τους προγραμματιστές να γράφουν σαφή, λογικό κώδικα για έργα μικρής και μεγάλης κλίμακας. Η Python χρησιμοποιείται ευρέως στη βιομηχανία και τον ακαδημαϊκό χώρο, έχει φιλική προς τον χρήστη σύνταξη και είναι εύκολο να την μάθει κανείς, καθιστώντας την εξαιρετική επιλογή τόσο για αρχάριους όσο και για έμπειρους επαγγελματίες του χώρου. Έχει, επίσης, μια μεγάλη κοινότητα χρηστών που συμβάλλουν στην ανάπτυξη και την υποστήριξή της, κάτι που μπορεί να είναι χρήσιμο για την αντιμετώπιση προβλημάτων και την εύρεση λύσεων σε προβλήματα.

Η Python, ως μια ερμηνευμένη γλώσσα και έχοντας ένα τεράστιο οικοσύστημα βιβλιοθηκών και πλαισίων, την καθιστά ιδανική γλώσσα για πολλούς επαγγελματίες της επιστήμης δεδομένων. Η δημοτικότητα της Python και το ευρύ φάσμα βιβλιοθηκών και πλαισίων της την καθιστούν πολύτιμο εργαλείο για τους επαγγελματίες της επιστήμης δεδομένων.

Η έκδοση που χρησιμοποιήθηκε είναι η 3.10.6 και εκτελέστηκε σε περιβάλλον VScode. Έγινε δημιουργία δύο αρχείων, ένα αρχείο όπου περιλαμβάνει τις απαραίτητες δημιουργημένες κλάσεις (Utils.py) και ένα διαφορετικό αρχείο για την εκτέλεση αυτών (Execution.py).

6.2 Βιβλιοθήκες

Οι βιβλιοθήκες στον προγραμματισμό είναι συλλογές προ-γραμμένου κώδικα που μπορούν να χρησιμοποιηθούν για την εκτέλεση κοινών εργασιών. Αυτές οι βιβλιοθήκες έχουν σχεδιαστεί για να κάνουν τη διαδικασία ανάπτυξης πιο αποτελεσματική, παρέχοντας επαναχρησιμοποιήσιμο κώδικα που μπορεί εύκολα να ενσωματωθεί σε ένα έργο.

Η σημασία των βιβλιοθηκών στην επιστήμη των δεδομένων έγκειται στο γεγονός ότι παρέχουν δοκιμασμένο και βελτιστοποιημένο κώδικα που μπορεί να εκτελέσει σύνθετες εργασίες με λίγες μόνο γραμμές κώδικα. Αυτό εξοικονομεί πολύ χρόνο και προσπάθεια στον επιστήμονα δεδομένων, επιτρέποντάς του να επικεντρωθεί στις πιο σημαντικές πτυχές του έργου του, όπως η ανάλυση δεδομένων και η μοντελοποίηση.

Επιπλέον, οι βιβλιοθήκες έχουν συχνά μεγάλες κοινότητες χρηστών που συμβάλλουν στην ανάπτυξη και την υποστήριξή τους, κάτι που μπορεί να είναι χρήσιμο για την αντιμετώπιση προβλημάτων και την εύρεση λύσεων σε προβλήματα.

Συνολικά, οι βιβλιοθήκες είναι ένα ουσιαστικό εργαλείο για τους επαγγελματίες της επιστήμης δεδομένων και διαδραματίζουν κρίσιμο ρόλο στην διαδικασία ανάπτυξης και στην αύξηση της αποτελεσματικότητάς της.

BeautifulSoup: Χρησιμοποιείται για την απόξεση ιστού (web scraping) και την ανάλυση εγγράφων HTML και XML. Παρέχει έναν απλό και βολικό τρόπο εξαγωγής δεδομένων από ιστοσελίδες, καθιστώντας το ιδανικό εργαλείο για την εξαγωγή δεδομένων. Συνολικά, το Beautiful Soup είναι μια ισχυρή βιβλιοθήκη που διευκολύνει την εξαγωγή δεδομένων από ιστοσελίδες, καθιστώντας την ένα πολύτιμο εργαλείο για επαγγελματίες της επιστήμης δεδομένων και ερευνητές που χρειάζονται συλλογή δεδομένων από τον Παγκόσμιο Ιστό. [61]

Pandas: Παρέχει δομές δεδομένων και εργαλεία ανάλυσης δεδομένων για αριθμητικούς πίνακες και δεδομένα χρονοσειρών. Υποστηρίζει επίσης την ανάγνωση και εγγραφή δεδομένων από διάφορες μορφές αρχείων, όπως βάσεις δεδομένων, αρχεία CSV, Excel και SQL. Το Pandas χρησιμοποιείται ευρέως στην επιστήμη δεδομένων και είναι απαραίτητο εργαλείο για τον καθαρισμό, την προετοιμασία και την εξερεύνηση δεδομένων. Συνολικά, το Pandas είναι μια ισχυρή και ευέλικτη βιβλιοθήκη για χειρισμό και ανάλυση δεδομένων, παρέχει έναν απλό και βολικό τρόπο χειρισμού δεδομένων σε διάφορες μορφές και χρησιμοποιείται ευρέως στην επιστήμη δεδομένων. [62]

Numpy: Χρησιμοποιείται για επιστημονικούς υπολογισμούς και χειρισμό δεδομένων. Παρέχει ένα ισχυρό αντικείμενο πίνακα N-διάστάσεων, μαζί με μια μεγάλη συλλογή μαθηματικών συναρτήσεων για λειτουργία σε αυτούς τους πίνακες. Το NumPy χρησιμοποιείται ευρέως στην επιστήμη δεδομένων και τη μηχανική μάθηση, καθώς επιτρέπει τον αποτελεσματικό χειρισμό μεγάλων σειρών δεδομένων. Παρέχει επίσης εργαλεία για γραμμική άλγεβρα, μετασχηματισμό Fourier και δημιουργία τυχαίων αριθμών. Συνολικά, η NumPy είναι μια βασική βιβλιοθήκη για την εργασία με αριθμητικά δεδομένα. [63]

Matplotlib: Παρέχει μια διεπαφή υψηλού επιπέδου για τη σχεδίαση ελκυστικών και ενημερωτικών στατιστικών γραφικών. Μπορεί επίσης να χρησιμοποιηθεί για την απεικόνιση τρισδιάστατων γραφικών παραστάσεων. Χρησιμοποιείται ευρέως στην οπτικοποίηση δεδομένων, την επιστημονική έρευνα και, γενικά, τη μηχανική μάθηση για τη δημιουργία στατικών, κινούμενων και διαδραστικών απεικονίσεων. Επιτρέπει την προσαρμογή γραφικών παραστάσεων μέσω ρυθμίσεων και θεμάτων και υποστηρίζει διάφορες μορφές αρχείων για αποθήκευση και εξαγωγή γραφημάτων. Συνολικά, το Matplotlib είναι μια ισχυρή και ευέλικτη βιβλιοθήκη για τη δημιουργία στατικών, κινούμενων και διαδραστικών απεικονίσεων. [64]

Seaborn: Είναι μια βιβλιοθήκη οπτικοποίησης δεδομένων, χτισμένη πάνω από τη βιβλιοθήκη οπτικοποίησης δεδομένων Matplotlib. Παρέχει μια διεπαφή υψηλού επιπέδου για τη δημιουργία ελκυστικών και ενημερωτικών στατιστικών γραφικών. Ειδικότερα, είναι κατάλληλη για την οπτικοποίηση πολύπλοκων συνόλων δεδομένων με πολλαπλές μεταβλητές. Το Seaborn παρέχει λειτουργίες που διευκολύνουν την οπτικοποίηση στατιστικών μοντέλων.[65]

Keras: Το Keras είναι μια βιβλιοθήκη βαθιάς εκμάθησης ανοιχτού κώδικα γραμμένη σε Python. Έχει σχεδιαστεί για να είναι φιλικό προς το χρήστη και εύκολο να επεκταθεί. Το Keras αναπτύχθηκε για να κάνει τη δημιουργία μοντέλων βαθιάς μάθησης όσο το δυνατόν πιο απλή. Η βιβλιοθήκη αυτή είναι χτισμένη πάνω από άλλες βιβλιοθήκες βαθιάς μάθησης, όπως το TensorFlow, το Theano και το CNTK. [66]

Sklearn: Το Scikit-learn, γνωστό και ως sklearn, είναι μια βιβλιοθήκη μηχανικής εκμάθησης ανοιχτού κώδικα. Παρέχει ένα ευρύ φάσμα εργαλείων για εξόρυξη και ανάλυση δεδομένων. Συμπεριλαμβάνει εποπτευόμενους και μη εποπτευόμενους αλγόριθμους μάθησης, προεπεξεργασίας, επιλογής και αξιολόγησης μοντέλων. Είναι χτισμένο πάνω σε άλλες δημοφιλείς βιβλιοθήκες Python, όπως το NumPy, Pandas, και ενσωματώνεται καλά με άλλες βιβλιοθήκες όπως το Matplotlib για σκοπούς οπτικοποίησης. [67]

Sktime: Είναι βασισμένο στο scikit-learn και παρέχει ένα συνεπές, υψηλού επιπέδου API για τη δημιουργία, εκπαίδευση και την αξιολόγηση μοντέλων χρονοσειρών. Περιλαμβάνει ένα ευρύ φάσμα εργαλείων για προεπεξεργασία, επιλογή μοντέλων και αξιολόγηση. [68]

Streamlit: Το Streamlit είναι μια βιβλιοθήκη ανοιχτού κώδικα για τη δημιουργία εφαρμογών που βασίζονται στον ιστό. Επιτρέπει στους προγραμματιστές να δημιουργούν διαδραστικές, φιλικές προς το χρήστη εφαρμογές γρήγορα και εύκολα. Χρησιμοποιεί μια απλή σύνταξη και ένα ενσωματωμένο "αντιδραστικό" μοντέλο προγραμματισμού, το οποίο ενημερώνει αυτόματα την εφαρμογή ως απόκριση στις αλληλεπιδράσεις των χρηστών. Περιλαμβάνει επίσης δυνατότητες όπως υποστήριξη για πολλαπλές εισόδους και εξόδους, ενσωματωμένο στυλ και θέματα, καθώς και τη δυνατότητα προσθήκης προσαρμοσμένου κώδικα JavaScript και CSS.[69]

Το συνολικό περιβάλλον που είναι απαραίτητο για την μελέτη αυτή μπορεί να ληφθεί από εδώ:

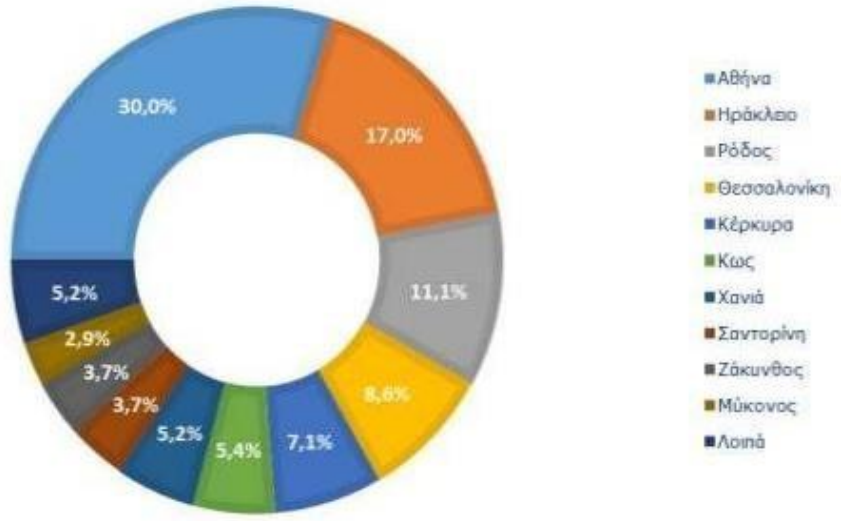
ΛΗΨΗ

7 Δεδομένα

Η επιλογή των δεδομένων είναι ένα σημαντικό βήμα στην επιστήμη των δεδομένων, διότι βοηθά στην εστίαση στις σχετικές πληροφορίες και στη βελτίωση της αποτελεσματικότητας της ανάλυσης. Επιλέγοντας τα κατάλληλα δεδομένα για την ανάλυση, οι επιστήμονες δεδομένων μπορούν να διασφαλίσουν ότι έχουν τις πιο σχετικές και χρήσιμες πληροφορίες για το πρόβλημα. Αυτό μπορεί να βοηθήσει στη βελτίωση της ακρίβειας και της αξιοπιστίας των αποτελεσμάτων, εστιάζοντας στα δεδομένα που είναι πιο πιθανό να παρέχουν πολύτιμες πληροφορίες. Μπορεί επίσης να βοηθήσει στη μείωση του όγκου των δεδομένων που πρέπει να υποβληθούν σε επεξεργασία, καθιστώντας την ανάλυση πιο αποτελεσματική και διαχειρίσιμη.

7.1 Πηγή

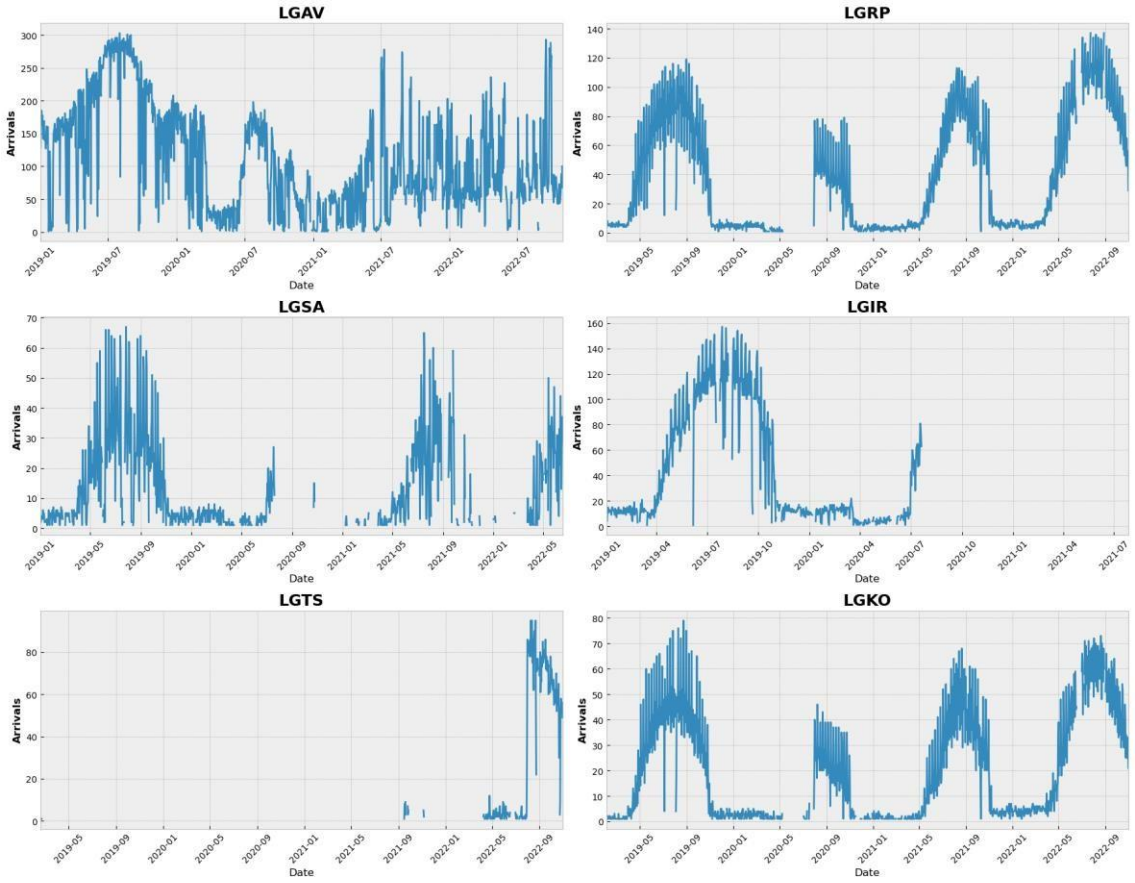
Πηγή δεδομένων για την υλοποίηση της μελέτης αποτελεί το OpenSky Network 2020. [70] Από την πηγή αυτή γίνεται απόκτηση δεδομένων από την πρώτη ημέρα του 2019 έως και τον Οκτώβρη του 2022, για πτήσεις σε όλο τον κόσμο. Με τη χρήση της προγραμματιστικής γλώσσας Python, εστιάζουμε στις μεταβλητές day και destination, μέσω των οποίων παρέχεται η δυνατότητα να μετρήσουμε τις αφίξεις σε καθημερινή βάση.



Σχήμα 10: Μερίδιο αγοράς Αεροδρομίων [71]

Συνεπώς, αποκτήθηκαν δεδομένα αφίξεων για έξι από τα μεγαλύτερα αεροδρόμια της Ελλάδας. Στη συνέχεια, ακολουθεί η οπτικοποίηση των εν λόγω δεδομένων.

ICAO	Αεροδρόμιο
LGAV	Διεθνής Αερολιμένας Αθηνών
LGRP	Κρατικός Αερολιμένας Ρόδου Διαγόρας
LGSA	Διεθνές Αεροδρόμιο Χανίων Ιωάννης Δασκαλογιάννης
LGIR	Κρατικός Αερολιμένας Ηρακλείου Ν. Καζαντζάκης
LGTS	Διεθνής Κρατικός Αερολιμένας Θεσσαλονίκης Μακεδονία
LGKO	Κρατικός Αερολιμένας Κω Ιπποκράτης



Σχήμα 11: Αφίξεις σε διαφορετικά αεροδρόμια

7.2 Επιλεγμένο Αεροδρόμιο

Το πιο πολυσύχναστο αεροδρόμιο στην Ελλάδα είναι το Διεθνές Αεροδρόμιο Αθηνών, γνωστό με τον κωδικό LGAV, το οποίο και κατέχει την πρωτιά με ποσοστό 30% επί του συνόλου, ως προς το μερίδιο αγοράς των αεροδρομίων της Ελλάδας.

Ο ρόλος του εν λόγω αεροδρομίου είναι νευραλγικός τόσο για τον τουρισμό όσο και, γενικά, για την οικονομία της χώρας. Αυτό προκύπτει ως αποτέλεσμα του ρόλου του ως κόμβος τόσο για την εσωτερική όσο και για τη διεθνή αεροπορική κίνηση στην Ελλάδα. Το αεροδρόμιο λειτουργεί ως κρίσιμος σύνδεσμος μεταξύ απομονωμένων και μη απομονωμένων περιοχών και του υπόλοιπου έθνους. Συνεπώς, αποτελεί αναπόσπαστο στοιχείο του ελληνικού συστήματος μεταφορών. Εκτός από τα άμεσα οφέλη για την τοπική οικονομία, το αεροδρόμιο έχει επίσης έμμεσες επιδράσεις στις επιχειρήσεις των βιομηχανιών ταξιδιών, της φιλοξενίας και του λιανικού εμπορίου.

Σύμφωνα με μια μελέτη του FOUNDATION FOR ECONOMIC & INDUSTRIAL RESEARCH για τις οικονομικές επιπτώσεις του Διεθνούς Αεροδρομίου Αθηνών παρουσιάστηκαν τα εξής: [72]

- Η συνεισφορά του στο ΑΕΠ εκτιμάται σε 7,9 δισ. ευρώ το 2017 (4,4% του ΑΕΠ).
- Ο αντίκτυπος στην απασχόληση από τη λειτουργία του LGAV υπολογίζεται σε 181.000 θέσεις εργασίας το 2017 (4,8% της συνολικής εγχώριας απασχόλησης).

Γενικά, ο Διεθνής Αερολιμένας Αθηνών έχει πολύ σημαντικό αντίκτυπο στην ελληνική οικονομία και στην κατεύθυνση της οικονομίας της χώρας. Λόγω της τεράστιας αυτής σημασίας αλλά και με γνώμονα την ποιότητα των δεδομένων, επιλέχθηκε το Διεθνές Αεροδρόμιο Αθηνών ως το υπό μελέτη αεροδρόμιο για πρόβλεψη του αριθμού αφίξεων.

7.3 Καθαρισμός Δεδομένων

Η φάση καθαρισμού είναι ένα σημαντικό βήμα στην επιστήμη των δεδομένων, επειδή διασφαλίζει ότι τα δεδομένα είναι ακριβή, συνεπή και έτοιμα για ανάλυση. Βοηθά στην αφαίρεση ή τη διόρθωση τυχόν σφαλμάτων ή ασυνεπειών στα δεδομένα, τα οποία μπορεί να οδηγήσουν σε μεροληπτικά ή παραπλανητικά αποτελέσματα εάν δεν αντιμετωπιστούν. Για παράδειγμα, οι ανακρίβειες ή τα ελλιπή δεδομένα μπορεί να παραμορφώσουν τα αποτελέσματα της ανάλυσης και να δυσκολέψουν την εξαγωγή ποιοτικών συμπερασμάτων. Ομοίως, οι ασυνέπειες στη μορφή δεδομένων μπορεί να δυσκολέψουν την εργασία με τα δεδομένα και να δημιουργήσουν προβλήματα κατά την προσπάθεια συνδυασμού δεδομένων από διαφορετικές πηγές.

Επιπλέον, κατά τη φάση καθαρισμού μπορούν να ανιχνευθούν οι ακραίες τιμές και να αντιμετωπιστούν, γεγονός που μπορεί επίσης να βελτιώσει την ακρίβεια των αποτελεσμάτων. Η διαδικασία καθαρισμού των δεδομένων μπορεί επίσης να βοηθήσει στη μείωση του απαιτούμενου χώρου αποθήκευσης, κάτι που μπορεί να είναι σημαντικό στα μεγάλα σύνολα δεδομένων. Συνοπτικά, η φάση καθαρισμού είναι ένα ουσιαστικό βήμα στην επιστήμη των δεδομένων που συμβάλλει στη βελτίωση της ποιότητας και της αξιοπιστίας των αποτελεσμάτων αφαιρώντας ανακρίβειες, ασυνέπειες και άσχετες πληροφορίες.

7.3.1 Missing Values

Στα δεδομένα που αποκτήθηκαν παρατηρήθηκε ότι για τα περισσότερα αεροδρόμια υπάρχει υψηλή εμφάνιση missing values. Αναλυτικότερα :

ICAO	Αεροδρόμιο	Missing Values
LGAV	Διεθνής Αερολιμένας Αθηνών	2.57%
LGRP	Κρατικός Αερολιμένας Ρόδου Διαγόρας	11.14%
LGKO	Κρατικός Αερολιμένας Κω Ιπποκράτης	15.00%
LGSA	Διεθνές Αεροδρόμιο Χανίων Ιωάννης Δασκαλογιάννης	42.93%
LGIR	Κρατικός Αερολιμένας Ηρακλείου Ν. Καζαντζάκης	63.14%
LGTS	Διεθνής Κρατικός Αερολιμένας Θεσσαλονίκης Μακεδονία	86.93%

Εφόσον επελέγη ως περίπτωση μελέτης ο Διεθνής Αερολιμένας Αθηνών, θα εστιάσουμε αποκλειστικά στην επεξεργασία των συγκεκριμένων δεδομένων.

Το πρώτο πρόβλημα που χρήζει επίλυσης είναι οι ελλείπουσες τιμές. Μία αιτία που μπορεί να προκαλεί το φαινόμενο αυτό μπορεί να είναι ενημερώσεις του συστήματος ή αντιμετώπιση σφαλμάτων. Κατά τη διάρκεια τέτοιων περιστατικών, το σύστημα που λαμβάνει τα σήματα των αεροπλάνων και τα αποθηκεύει μπορεί να παραμένει κλειστό και, συνεπώς, να παρουσιάζονται ημέρες χωρίς καθόλου δεδομένα. Τα ελλείποντα δεδομένα διορθώθηκαν με την χρήση της γραμμικής παρεμβολής (linear interpolation). Τι είναι όμως γραμμική παρεμβολή;

Αναλυτικότερα, η γραμμική παρεμβολή είναι μια μέθοδος εκτίμησης της τιμής μιας συνάρτησης μεταξύ δύο γνωστών σημείων. Χρησιμοποιείται συνήθως στα μαθηματικά, τη μηχανική και τα γραφικά υπολογιστών για την ομαλή μετάβαση μεταξύ των τιμών. Στην επιστήμη δεδομένων, η γραμμική παρεμβολή είναι μια μέθοδος που χρησιμοποιείται για τη συμπλήρωση σημείων - τιμών, που, πιθανώς, λείπουν από ένα σύνολο δεδομένων. Λειτουργεί υποθέτοντας ότι τα σημεία - τιμές που λείπουν έχουν μια γραμμική σχέση με τα γύρω γνωστά σημεία - τιμές. Αυτή η μέθοδος χρησιμοποιείται συνήθως όταν μελετώνται δεδομένα χρονοσειρών, όπου τα δεδομένα μπορεί να συλλέγονται σε ακανόνιστα διαστήματα ή / και με κενά.

Δίνονται δύο σημεία (x_0, y_0) και (x_1, y_1) και ένα άγνωστο σημείο (x, y), η γραμμική παρεμβολή βρίσκει την τιμή του y στο σημείο x παίρνοντας μια ευθεία γραμμή μεταξύ των δύο γνωστών σημείων και υπολογίζοντας την τιμή y στο σημείο x. Αυτό γίνεται βρίσκοντας την κλίση της ευθείας μεταξύ των δύο σημείων και στη συνέχεια χρησιμοποιώντας αυτήν την κλίση για να βρεθεί η τιμή y στο x. [73]

Η εξίσωση που χρησιμοποιείται είναι:

$$y = y_0 + (x - x_0)\frac{y_1 - y_0}{x_1 - x_0}$$

7.3.2 Outliers

Πέρα από τις 36 ελλείπουσες τιμές, παρατηρήθηκε ότι υπήρχαν αρκετές ημέρες όπου στο LGAV σημειώθηκε πολύ χαμηλός αριθμός αφίξεων. Αναλυτικότερα, περισσότερες πληροφορίες παρουσιάζονται στον παρακάτω πίνακα.

	2019	2020	2021	2022	TOTAL
ARRIVALS LGAV < 10	12	33	56	10	111
ARRIVALS LGAV < 20	17	61	82	17	177

Όπως παρουσιάστηκε, σημειώνονται 111 ημέρες όπου στο αεροδρόμιο LGAV κατέφθασαν λιγότερες από 10 πτήσεις. Επιπλέον, σημειώνονται 177 ημέρες όπου στο αεροδρόμιο LGAV κατέφθασαν λιγότερες από 20 πτήσεις.

Δεδομένου του μεγέθους του αεροδρομίου LGAV, εγείρονται βάσιμες υποψίες ότι υπάρχει σφάλμα στα δεδομένα. Συμπληρωματικά στην παραπάνω παρατήρηση και λαμβάνοντας επίσης υπόψη ότι και σε άλλα αεροδρόμια παρατηρείται το ίδιο φαινόμενο (μικρός αριθμός αφίξεων), συμπεραίνουμε ότι, πιθανώς, το πρόβλημα εντοπίζεται στη διαδικασία λήψης / καταγραφής των δεδομένων.

Μια πιθανή ερμηνεία ενδέχεται να είναι ότι δεν συνεργάζονται όλες οι αεροπορικές εταιρείες με την πηγή που ερευνάται, δηλαδή το OpenSky Network. Μη στέλνοντας δεδομένα για όλες τις πτήσεις, εμφανίζεται χαμηλότερος αριθμός πτήσεων για κάποιες ημέρες.

Για την αντιμετώπιση του προβλήματος των outliers δοκιμάστηκε σε πρώτη φάση η απαλοιφή τους με την χρήση του κανόνα 3 τυπικών αποκλίσεων (3σ rule). [74] Με αυτή την αντιμετώπιση δεν επιτεύχθηκε καμία διόρθωση στο σύνολο δεδομένων. Συνεπώς, χρειάστηκε επινόηση διαφορετικού τρόπου για την αντιμετώπιση του προβλήματος.

Αναλυτικότερα, στην περίπτωση αυτή υπολογίζεται η ποσοστιαία μεταβολή ανά ζεύγη ημερών. Αν αυτή η μεταβολή είναι μεγαλύτερη από το 80%, τότε γίνεται αφαίρεση της μικρότερης τιμής από τα δεδομένα. Οι τιμές που αφαιρούνται αντιμετωπίζονται ως missing values με τη γραμμική παρεμβολή (linear interpolation). Αυτή η διαδικασία γίνεται επαναληπτικά για την επίτευξη καλύτερων αποτελεσμάτων. Η διαδικασία τερματίζεται όταν δεν εντοπιστεί κάποια τιμή που να χαρακτηρίζεται ως outlier.

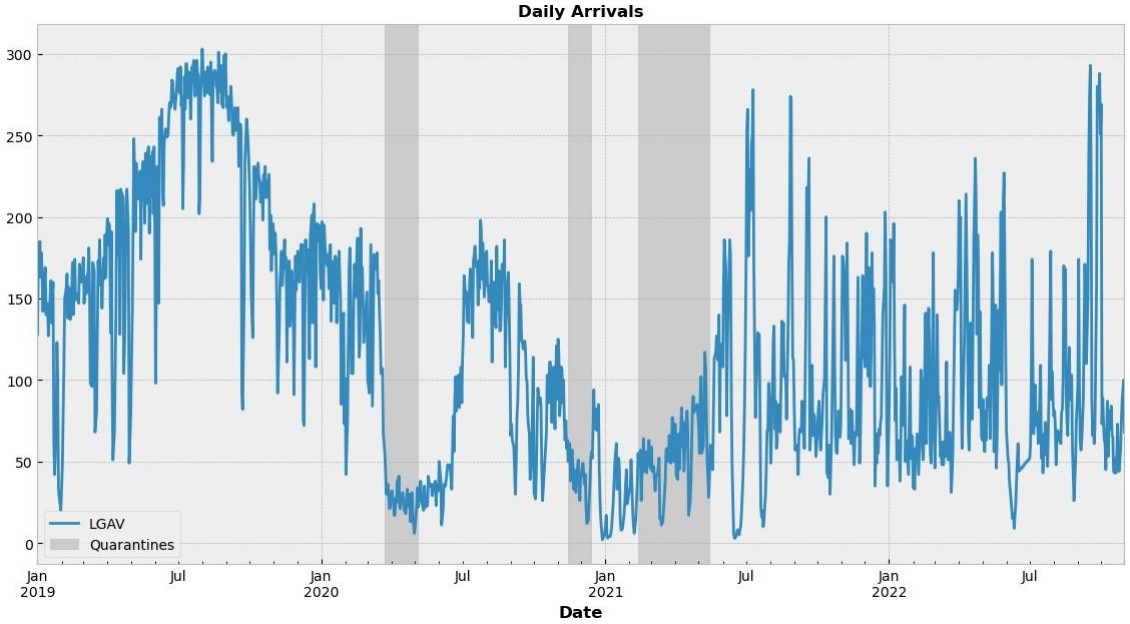
Ο τρόπος που παρουσιάστηκε μείωσε το πλήθος των χαμηλών τιμών αλλά δεν τις εξαφάνισε. Επίσης, βελτίωσε τις αποδόσεις των μοντέλων πρόβλεψης. Μετά τη διαδικασία διαχείρισης των ακραίων τιμών παρατηρούνται τα εξής:

	2019	2020	2021	2022	TOTAL
ARRIVALS LGAV < 10	0	6	19	1	26
ARRIVALS LGAV < 20	0	21	40	4	65

Συγκρίνοντας τους παραπάνω πίνακες γίνεται καλύτερα αντιληπτός ο βαθμός επεξεργασίας των δεδομένων. Στο σημείο αυτό αξίζει να αναφερθεί το γεγονός ότι κατά το πέρας και των δύο διαδικασιών καθαρισμού των δεδομένων, διορθώθηκαν συνολικά 270 παρατηρήσεις.

7.4 Exploratory analysis

Στο τμήμα αυτό θα γίνει διερευνητική ανάλυση (EDA) των δεδομένων που προέκυψαν μετά από τη διαδικασία καθαρισμού. Η EDA (Exploratory Data Analysis) είναι μια προσέγγιση για την ανάλυση και την κατανόηση δεδομένων μέσω συνοπτικών στατιστικών, οπτικοποιήσεων και άλλων μεθόδων. Στόχος είναι η αποκάλυψη μοτίβων σχέσεων και ιδεών στα δεδομένα που μπορεί να μην είναι άμεσα εμφανή. Είναι σημαντικό να σημειωθεί ότι το EDA είναι ένα κρίσιμο βήμα σε κάθε έργο ανάλυσης δεδομένων, καθώς παρέχει τη βάση για την ανάπτυξη πιο επίσημων στατιστικών μοντέλων και βοηθά τον ερευνητή να εντοπίσει τυχόν πιθανά προβλήματα με τα δεδομένα ή το ερευνητικό ερώτημα.



Σχήμα 12: Αφίξεις στο Διεθνή Αερολιμένα Αθηνών

Από την παραπάνω οπτικοποίηση αξίζει να τονιστεί η απότομη πτώση που παρατηρείται στην εποχή της πρώτης καραντίνας. Επιπρόσθετα, κατά τη διάρκεια της δεύτερης καραντίνας σημειώνεται μια επιπλέον μείωση, σε αντίθεση με την τελευταία καραντίνα. Τέλος, παρατηρώντας προσεκτικά το διάγραμμα φαίνεται μια εποχικότητα με μέγιστο τον μήνα Ιούλιο κάθε έτους. Η παρατήρηση αυτή όμως χρειάζεται περεταίρω ανάλυση για επιβεβαίωση.

Ο παρακάτω πίνακας περιλαμβάνει σημαντικά στατιστικές μετρικές για την χρονοσειρά.

	LGAV
COUNT	1400.00
MEAN	115.95
STD	75.16
MIN	2.00
25%	55.00
50%	98.00
75%	168.00
MAX	303.00
KSTESTSTATISTIC	1.00
P_VALUE	0.00
ADFULLERSTATISTIC	-2.50
P_VALUE	0.12

Όπως φαίνεται, η μέση τιμή των αφίξεων είναι 116 ημερήσιες πτήσεις. Η μέγιστη τιμή αφίξεων που έχει παρατηρηθεί είναι 303. Επιπρόσθετα, αξίζει να σημειωθεί ότι η υπό εξέταση χρονοσειρά δεν ακολουθάει την κανονική κατανομή και δεν είναι στάσιμη.

Αναλυτικότερα, ο έλεγχος για το αν η χρονοσειρά ακολουθεί την κανονική κατανομή ελέγχθηκε από το Kolmogorov-Smirnov test. Το τεστ αυτό έχει τις παρακάτω υποθέσεις:

H_0: Τα δείγματα ανήκουν στην ίδια κατανομή.

H_1: Τα δείγματα ανήκουν σε διαφορετική κατανομή.

Η τιμή p είναι χαμηλότερη από το όριο του 0,05. Επομένως, γίνεται απόρριψη της μηδενικής υπόθεσης υπέρ της εναλλακτικής. Δηλαδή τα δεδομένα δεν κατανέμονται σύμφωνα με την κανονική κατανομή.

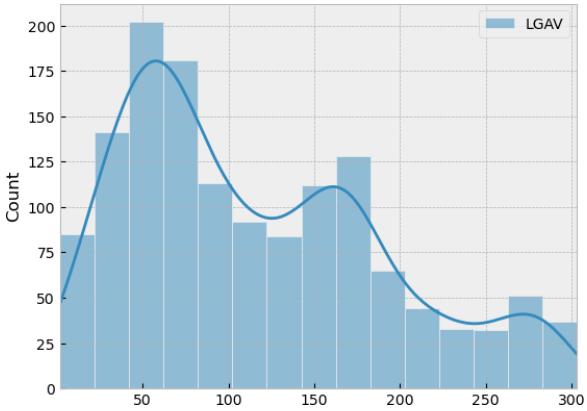
Επιπροσθέτως, ο έλεγχος στασιμότητας γίνεται με τη χρήση του Augmented Dickey-Fuller Test με τις παρακάτω υποθέσεις:

H_0: Η χρονοσειρά δεν είναι στάσιμη.

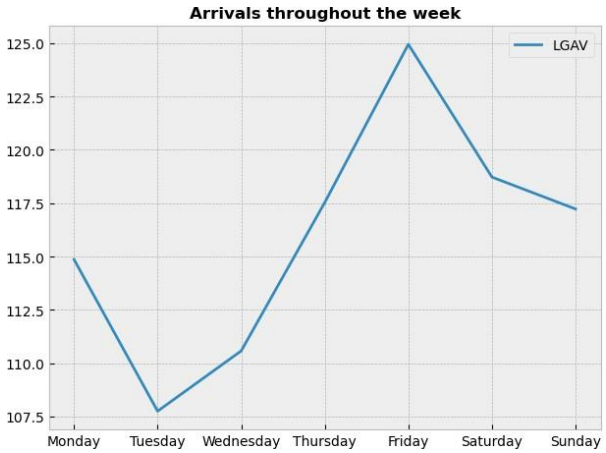
H_1: Η χρονοσειρά είναι στάσιμη.

Εάν η τιμή p από το τεστ είναι μικρότερη από το επίπεδο σημαντικότητας (α = 0,05), τότε γίνεται απόρριψη της μηδενικής υπόθεσης και συμπεραίνεται ότι η χρονοσειρά είναι στάσιμη. Όμως η τιμή p είναι μεγαλύτερη από το όριο. Επομένως, γίνεται αποδοχή της μηδενικής υπόθεσης και συμπεραίνεται ότι η χρονοσειρά δεν είναι στάσιμη.

Επιπροσθέτως από το παρακάτω σχήμα μπορεί να παρατηρηθεί ότι πιο συχνά υπάρχουν περίπου 50 ημερήσιες αφίξεις.

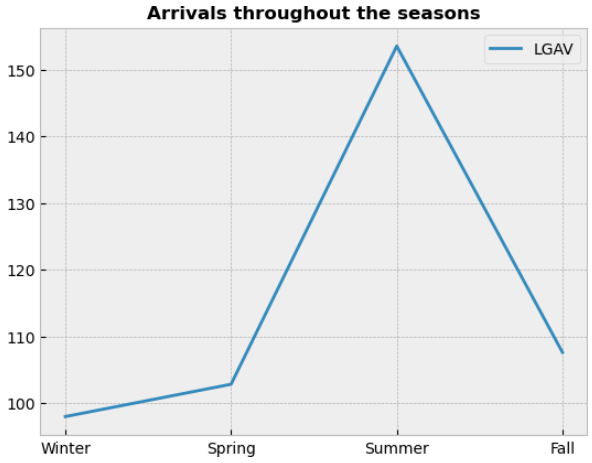


Σχήμα 13: Ιστόγραμμα συχνοτήτων



Σχήμα 14: Μέσος αριθμών πτήσεων για τις ημέρες της εβδομάδος

Την Παρασκευή σημειώνονται κατά μέσο όρο οι περισσότερες αφίξεις, ενώ την Τρίτη σημειώνονται οι λιγότερες.



Σχήμα 15: Μέσος αριθμών πτήσεων για τις εποχές

Κλείνοντας το κεφάλαιο αυτό, αξίζει να σχολιαστεί η εποχικότητα. Όπως παρατηρήθηκε και σε προηγούμενη οπτικοποίηση έτσι και εδώ επιβεβαιώνεται η ύπαρξη της εποχικότητας. Το γεγονός αυτό είναι λογικό εφόσον το καλοκαίρι γίνονται τα περισσότερα ταξίδια και υπάρχει η άνθηση του τουρισμού.

7.5 Εξωγενείς μεταβλητές

Οι εξωγενείς μεταβλητές, γνωστές και ως εξωτερικές ή επεξηγηματικές μεταβλητές, είναι παράγοντες που δεν περιλαμβάνονται στις χρονοσειρές που αναλύονται αλλά μπορεί να έχουν επίδραση σε αυτές. Η συμπερίληψη εξωγενών μεταβλητών στην πρόβλεψη χρονοσειρών μπορεί να είναι κρίσιμη επειδή επιτρέπει μια πιο ολοκληρωμένη κατανόηση των υποκείμενων παραγόντων που επηρεάζουν τις χρονοσειρές. Χωρίς να λαμβάνονται υπόψη εξωγενείς μεταβλητές, μια πρόβλεψη χρονοσειρών μπορεί να είναι λιγότερο ακριβής λόγω της παράλειψης σημαντικών πληροφοριών. Επιπλέον, με την ενσωμάτωση εξωγενών μεταβλητών, καθίσταται δυνατός ο εντοπισμός σχέσεων αιτίου-αποτελέσματος μεταξύ των μεταβλητών, οι οποίες μπορούν να βοηθήσουν στη λήψη κατάλληλων ενεργειών για τη βελτίωση της μελλοντικής απόδοσης. Συμπερασματικά, οι εξωγενείς μεταβλητές παίζουν σημαντικό ρόλο στην πρόβλεψη χρονοσειρών παρέχοντας πρόσθετες πληροφορίες που μπορούν να βελτιώσουν την ακρίβεια και την κατανόηση των προβλεπόμενων τάσεων.

Λαμβάνοντας τις παραπάνω πληροφορίες υπόψιν, δημιουργήθηκαν διαφορετικά σύνολα δεδομένων στα οποία συμπεριλαμβάνονται διαφορετικές εξωγενείς μεταβλητές. Οι εξωγενείς μεταβλητές που χρησιμοποιήθηκαν περιλαμβάνουν τον χρόνο, τις καραντίνες αλλά και των αριθμό των ατόμων που έχουν εμβολιαστεί. Αναλυτικότερα, πέρα από το σύνολο δεδομένων που περιλαμβάνει τη χρονοσειρά χωρίς εξωγενείς μεταβλητές, δημιουργήθηκαν 3 επιπρόσθετα σύνολα δεδομένων. Αρχικά, δημιουργήθηκε ένα πρώτο σύνολο δεδομένων το οποίο περιλαμβάνει χρονικές πληροφορίες ως καινούργια χαρακτηριστικά.

Η σημασία της προσθήκης του χρόνου ως εξωγενούς μεταβλητής στην πρόβλεψη χρονοσειρών έγκειται στην βελτίωση της ικανότητά του μοντέλου να λαμβάνει υπόψη τα πρότυπα και τις τάσεις στα δεδομένα. Η εποχικότητα και η ανάλυση τάσεων είναι μερικοί από τους βασικούς παράγοντες που μπορούν να αποτυπωθούν προσθέτοντας τον χρόνο ως εξωγενή μεταβλητή. Συμπεριλαμβάνοντας τον χρόνο ως εξωγενή μεταβλητή, είναι δυνατή η ανάλυση και η πρόβλεψη αυτών των τάσεων.

Τα χαρακτηριστικά που δημιουργήθηκαν είναι η μέρα της βδομάδας (Weekday) , το τρίμηνο αλλά και το έτος. Το χαρακτηριστικό Weekday είναι αριθμητικό και παίρνει τιμές από 0 έως 6. Το μηδέν αντιστοιχεί στην Δευτέρα και το έξι στην Κυριακή. Επιπλέον, το τρίμηνο είναι και αυτό αριθμητική μεταβλητή και παίρνει τιμές από 1 έως 4, όπου 1 είναι ο χειμώνας ενώ 4 το φθινόπωρο.

Όσον αναφορά την καραντίνα ως εξωγενές χαρακτηριστικό, δημιουργήθηκε μια Boolean μεταβλητή, όπου έχει την τιμή 0 όταν δεν υπήρχε καραντίνα, ενώ παίρνει την τιμή 1 σε αντίθετη περίπτωση.

Τέλος, συνεκτιμήθηκε η μεταβλητή που περιγράφει αριθμό των ατόμων που έχουν εμβολιαστεί. Το χαρακτηριστικό αυτό αποκτήθηκε από δημοσιευμένα δεδομένα που αφορούν την Covid-19. [75] Τα δεδομένα αυτά ελέγχθηκαν με ανάλυση συσχέτισης και επιλέχθηκε το χαρακτηριστικό των εμβολιασμών εφόσον είχε την μεγαλύτερη συσχέτιση με την χρονοσειρά των αφίξεων.

7.6 Διαχείριση Στασιμότητας

Όπως παρουσιάστηκε στο πρώτο κεφάλαιο, σημαντική είναι η αντιμετώπιση της μη στασιμότητας της χρονοσειράς. Ο μετασχηματισμός μιας μη στάσιμης χρονοσειράς σε στάσιμη είναι σημαντική επειδή πολλές τεχνικές ανάλυσης χρονοσειρών, όπως η πρόβλεψη και η στατιστική μοντελοποίηση, υποθέτουν ότι τα δεδομένα είναι ακίνητα. Τα μη στάσιμα δεδομένα μπορούν να παράγουν παραπλανητικά ή εσφαλμένα αποτελέσματα κατά τη χρήση αυτών των τεχνικών.

Παρατηρώντας ότι η χρονοσειρά δεν είναι στάσιμη, αξίζει να αναφερθεί ότι τα μοντέλα που χρησιμοποιήθηκαν δοκιμάστηκαν με τα διάφορα σύνολα δεδομένων αλλά και με τα σύνολα δεδομένων τροποποιημένα σε στάσιμα. Αναλυτικότερα, παίρνοντας τις πρώτες διαφορές (differencing) στη χρονοσειρά σημειώθηκε η μετατροπή της σε στάσιμη. Οι προβλέψεις που δημιουργήθηκαν από τις στάσιμες χρονοσειρές προβλήθηκαν εκ νέου στο αρχικό επίπεδο της χρονοσειράς πριν την αξιολόγηση με τη χρήση των μετρικών.

7.7 Πίνακας αποτελεσμάτων

Στον παρακάτω πίνακα παρατίθενται τα αποτελέσματα της μελέτης για την πρόβλεψη των αφίξεων στο Διεθνές Αεροδρόμιο Αθηνών. Αξίζει να τονιστεί ότι το διάστημα πρόβλεψης που επιλέχθηκε είναι 60 ημέρες.

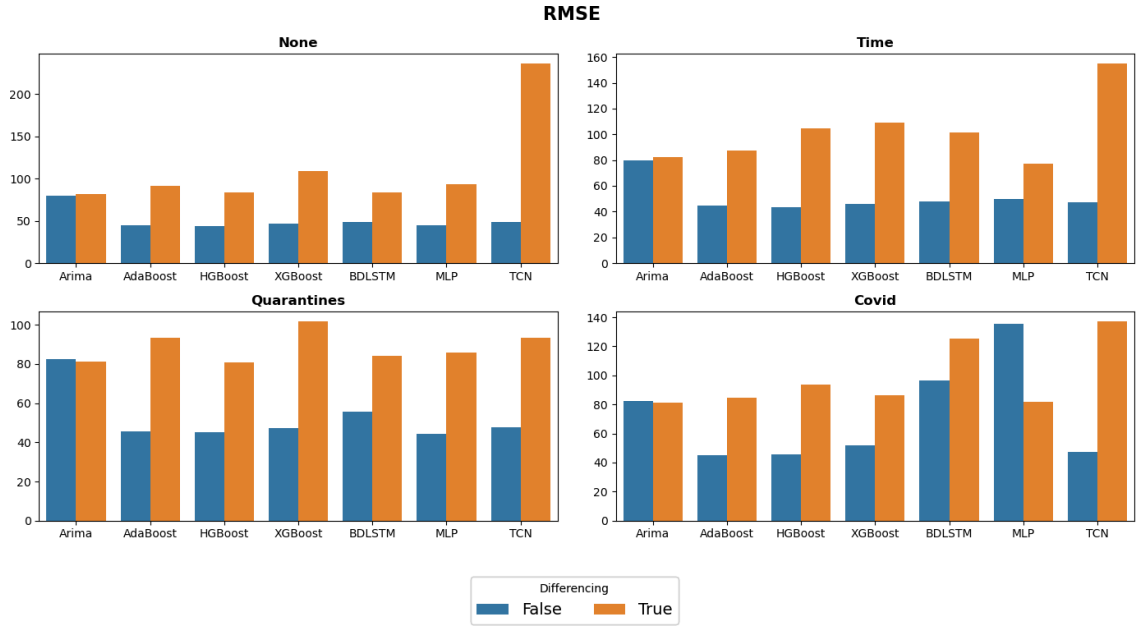
Χρειάζεται να αναφερθεί ότι στην στήλη Features βρίσκονται οι μεταβλητές που χρησιμοποιήθηκαν ως εξωγενείς. Σε προηγούμενη ενότητα έγινε αναλυτική αναφορά για τα υπό εξέταση εξωγενής χαρακτηριστικά.

MODEL	FEATURES	RMSE	MAPE	MAE	DIFFERENCING	TIME
AUTO ARIMA	Covid	81.34	0.39	51.73	True	0:01:06.410
	Covid	82.37	0.37	51.32	False	0:01:15.872
	None	81.24	0.39	51.7	True	0:00:44.425
	None	79.76	0.45	53.83	False	0:01:05.483
	Quarantines	81.24	0.39	51.7	True	0:01:34.822
	Quarantines	82.51	0.36	50.83	False	0:02:07.827
	Time	82.37	0.39	52.18	True	0:07:07.864
	Time	79.55	0.51	56.65	False	0:17:42.342
ADABOOST	Covid	84.52	0.39	52.17	True	0:00:00.205
	Covid	44.98	0.28	28.52	False	0:00:00.220
	None	91.89	0.4	57.92	True	0:00:00.151
	None	44.65	0.28	27.87	False	0:00:00.160
	Quarantines	93.31	0.37	56.77	True	0:00:00.161
	Quarantines	45.7	0.29	29.38	False	0:00:00.171
	Time	87.34	0.39	53.97	True	0:00:00.223
	Time	44.94	0.28	28.15	False	0:00:00.236

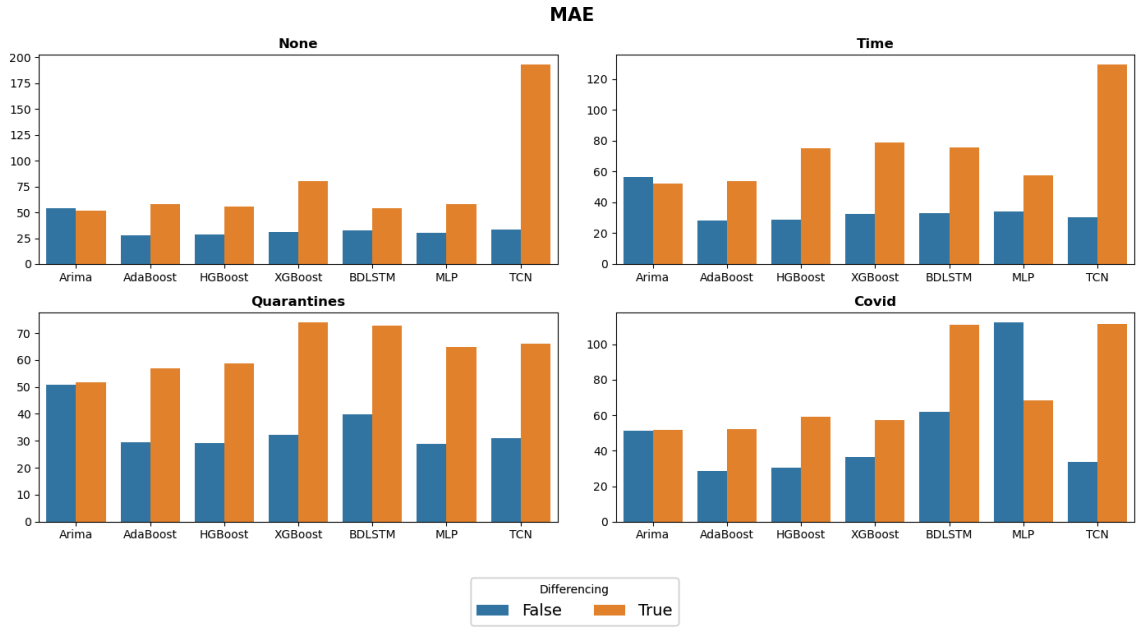
HGBOOST	Covid	93.66	0.44	59.23	True	0:00:00.553
	Covid	45.55	0.3	30.32	False	0:00:00.611
	None	83.55	0.49	55.73	True	0:00:00.465
	None	44.19	0.29	28.73	False	0:00:00.483
	Quarantines	80.8	0.58	58.73	True	0:00:00.537
	Quarantines	45.02	0.3	29.23	False	0:00:00.598
	Time	104.46	0.6	75.22	True	0:00:00.709
	Time	43.56	0.29	28.67	False	0:00:00.755
XGBOOST	Covid	86.19	0.48	57.2	True	0:00:00.082
	Covid	51.81	0.31	36.6	False	0:00:00.083
	None	108.67	0.67	80.25	True	0:00:00.070
	None	46.52	0.31	30.65	False	0:00:00.064
	Quarantines	101.67	0.65	73.93	True	0:00:00.065
	Quarantines	47.41	0.33	32.3	False	0:00:00.071
	Time	108.82	0.64	78.72	True	0:00:00.082
	Time	45.76	0.32	32.43	False	0:00:00.087
BDLSTM	Covid	125.17	1.54	110.87	True	0:00:05.338
	Covid	96.79	0.39	61.8	False	0:00:09.000
	None	83.71	0.38	53.93	True	0:00:07.257
	None	49.03	0.34	32.2	False	0:00:05.730
	Quarantines	84.22	0.82	72.83	True	0:00:06.780
	Quarantines	55.52	0.49	39.7	False	0:00:09.679
	Time	101.76	0.62	75.52	True	0:00:09.060
	Time	47.87	0.36	33.32	False	0:00:06.753
MLP	Covid	81.6	0.76	68.55	True	0:00:00.085
	Covid	135.44	1.0	112.3	False	0:00:00.168
	None	93.75	0.38	57.77	True	0:00:00.042
	None	44.49	0.31	30.22	False	0:00:00.091
	Quarantines	85.69	0.66	64.97	True	0:00:00.071
	Quarantines	44.48	0.28	28.85	False	0:00:00.095
	Time	77.49	0.55	57.72	True	0:00:00.041
	Time	49.52	0.34	34.32	False	0:00:00.237
TCN	Covid	137.13	1.26	111.38	True	0:01:28.562
	Covid	47.2	0.38	33.82	False	0:00:16.276
	None	235.96	2.95	192.8	True	0:00:55.868
	None	48.8	0.32	33.37	False	0:00:22.782
	Quarantines	93.41	0.5	65.92	True	0:01:06.691
	Quarantines	47.81	0.38	31.08	False	0:00:19.738
	Time	154.75	1.89	129.35	True	0:01:42.322
	Time	47.29	0.32	30.63	False	0:00:29.911

7.8 Οπτικοποίηση αποτελεσμάτων

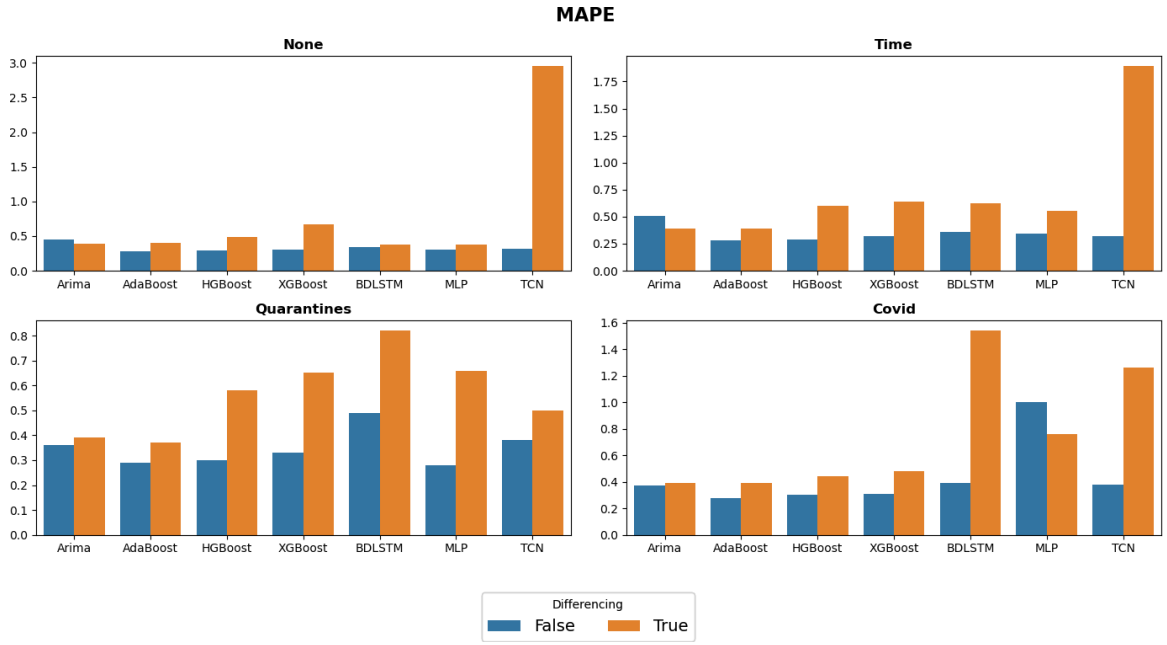
Στο κεφάλαιο αυτό θα γίνει οπτική παρουσίαση των αποτελεσμάτων, καθώς και αναφορά των συμπερασμάτων που προέκυψαν.



Σχήμα 16: Οπτικοποίηση RMSE



Σχήμα 17: Οπτικοποίηση MAE



Σχήμα 18: Οπτικοποίηση MAPE

Πρώτο και κύριο συμπέρασμα αποτελεί η μη αποτελεσματικότητα της μεθόδου πρώτων διαφορών (μετατροπή χρονοσειράς σε στάσιμη) στη βελτίωση των αποτελεσμάτων των μοντέλων. Με βάση όλες τις μετρικές, παρατηρείται ότι βελτίωσε σημαντικά τα αποτελέσματα μόνο του μοντέλου MLPRegressor στην περίπτωση όπου χρησιμοποιήθηκε σαν εξωγενής μεταβλητή ο αριθμό των ατόμων που έχουν εμβολιαστεί.

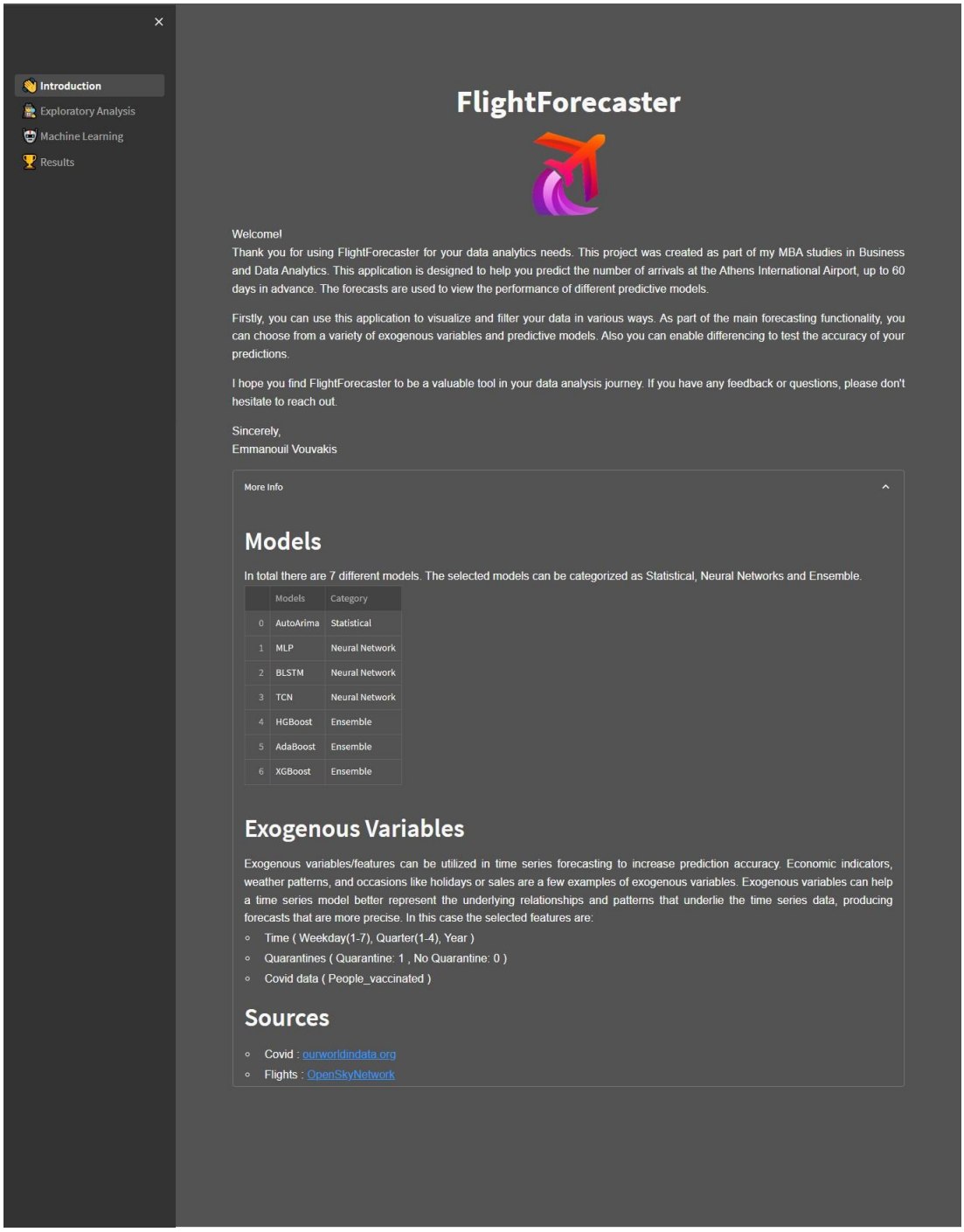
Επιπλέον, αξίζει να τονιστεί η σύγκριση των αποτελεσμάτων μεταξύ των κατηγοριών των μοντέλων. Λαμβάνοντας υπόψιν και τις τρεις μετρικές αξιολόγησης των προβλέψεων στα μη στάσιμα δεδομένα, παρατηρείται ότι τα ensemble και τα νευρωνικά μοντέλα παρείχαν σε ορισμένες περιπτώσεις καλύτερες προβλέψεις συγκριτικά με το στατιστικό μοντέλο. Αναλυτικότερα, στην απλή χρονοσειρά αλλά και στο σύνοδο δεδομένων με τον χρόνο σαν εξωγενή χαρακτηριστικά παρατηρείται ότι τα μοντέλα αυτά αξιολογήθηκαν καλύτερα από το AutoArima.

Ειδικότερα, αξίζει να αναφερθεί περισσότερο και η αποδοτικότητα των μοντέλων ensemble στα μη στάσιμα δεδομένα. Αναλυτικότερα, με βάση και τις τρεις μετρικές αξιολόγησης τα μοντέλα αυτά έφεραν σταθερά καλύτερα αποτελέσματα σε όλα τα σύνολα δεδομένων συγκριτικά με το μοντέλο AutoArima. Ένα τελευταίο συμπέρασμα είναι το γεγονός ότι η εισαγωγή εξωγενών μεταβλητών δεν δημιούργησε κάποια σημαντική βελτίωση στα υπό εξέταση μοντέλα.

8 FlightForecaster

Στο κεφάλαιο αυτό θα γίνει αναφορά για την ανάπτυξη εφαρμογής FlightForecaster. Η εφαρμογή αυτή δημιουργήθηκε με σκοπό την σύμπτυξη των σταδίων οπτικοποίησης των δεδομένων, εκτέλεσης των μοντέλων αλλά και της παρουσίασης των ευρημάτων. Η διαδραστική αυτή εφαρμογή αναπτύχθηκε με την χρήση της βιβλιοθήκης Streamlit της Python.

Αναλυτικότερα, η εφαρμογή αυτή αποτελείται από 4 καρτέλες. Στην πρώτη καρτέλα περιλαμβάνεται μια σύνοψη των δυνατοτήτων της εφαρμογής, καθώς και άλλες πληροφορίες που χρήζουν αναφοράς.

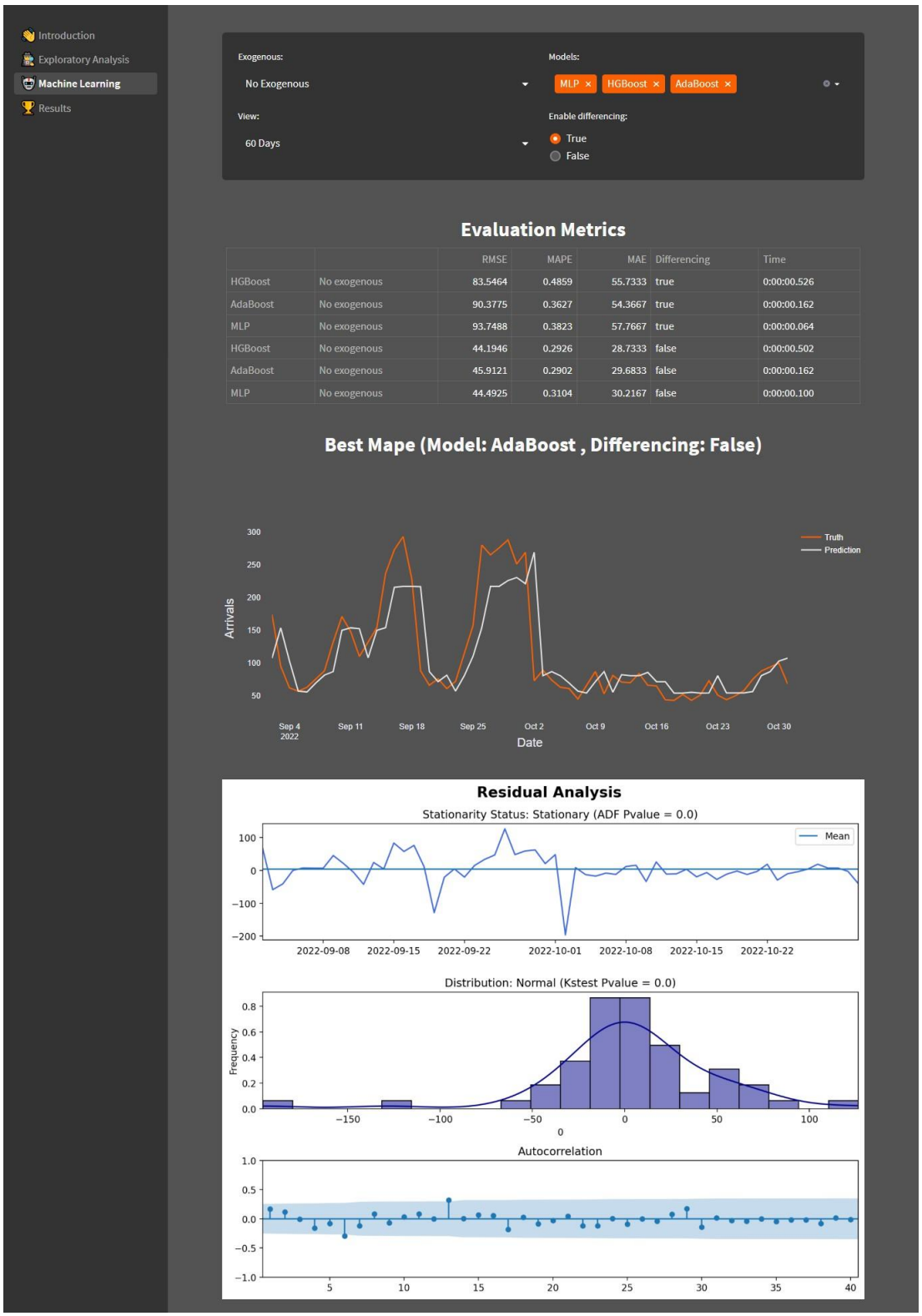


Σχήμα 19: Εισαγωγική καρτέλα

Συνεχίζοντας, στη δεύτερη καρτέλα, γίνεται η οπτικοποίηση των δεδομένων αφίξεων του υπο μελέτη αεροδρομίου με ποικίλους τρόπους.

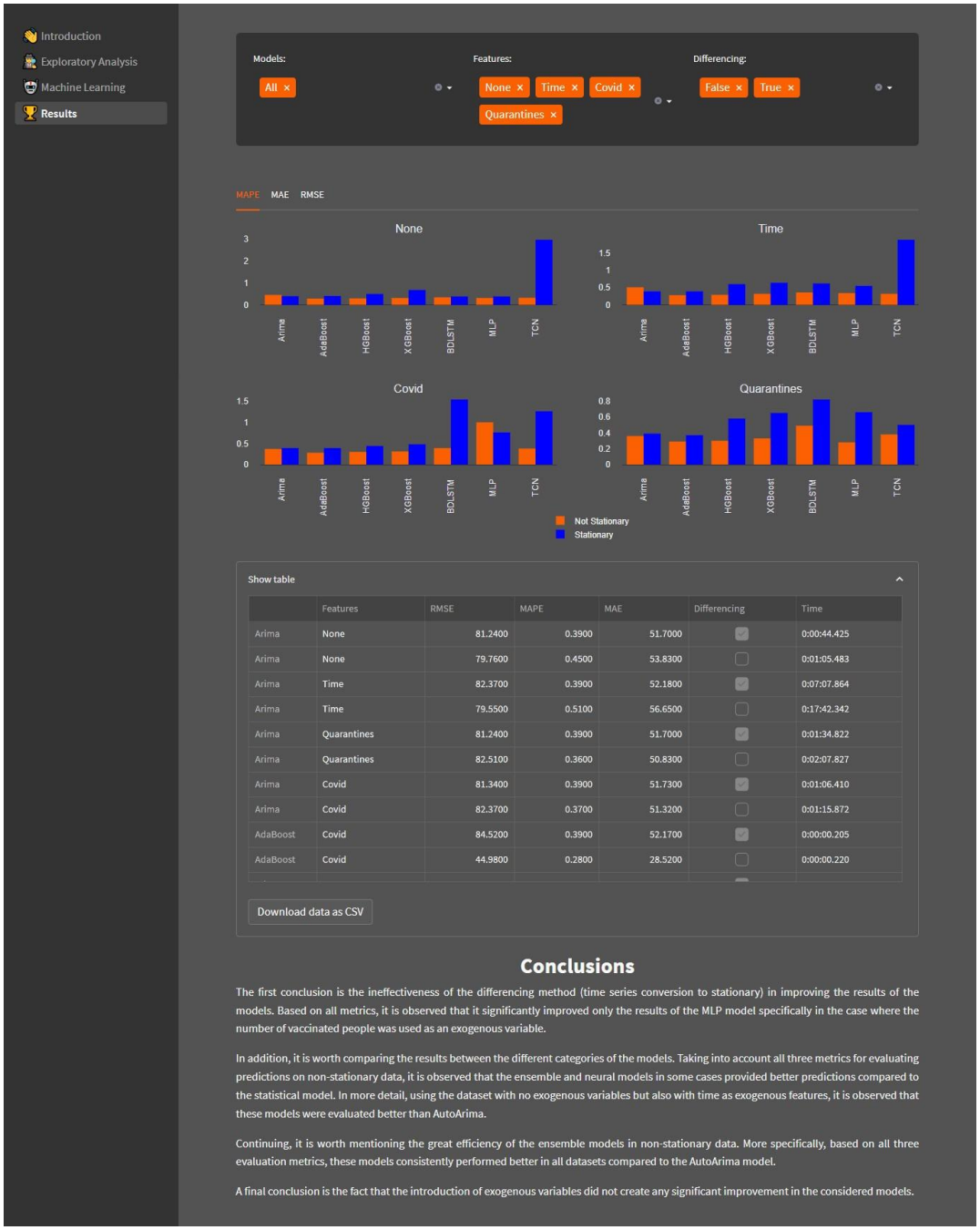
Σχήμα 20: Καρτέλα εξερεύνησης

Στην τρίτη καρτέλα δίνεται η δυνατότητα στον χρήστη να υλοποιήσει ο ίδιος τα μοντέλα πρόβλεψης σε πραγματικό χρόνο. Επιπλέον, γίνεται οπτικοποίηση των προβλέψεων και των σφαλμάτων του μοντέλου που απέδωσε το μικρότερο MAPE.



Σχήμα 21: Καρτέλα εκπαίδευσης μοντέλων

Στην τελευταία καρτέλα της εφαρμογής, γίνεται η οπτικοποίηση των αποτελεσμάτων που προέκυψαν στο πλαίσιο της συγκεκριμένης μελέτης. Αυτό αποφασίστηκε για να μην χρειάζεται ο χρήστης να περιμένει να τρέξουν τα μοντέλα για να δει τα αποτελέσματα. Επιπρόσθετα, αξίζει να σημειωθεί ότι τα αποτελέσματα αυτά μπορούν να ληφθούν επίσης από αυτή την καρτέλα. Κλείνοντας, ακολουθεί η καταγραφή συμπερασμάτων που προέκυψαν από τα αποτελέσματα που παρήχθησαν.



Σχήμα 22: Καρτέλα αποτελεσμάτων

Τέλος, για λόγους πληρότητας και καλύτερης παρουσίασης των δυνατοτήτων της εφαρμογής δημιουργήθηκε ένα σύντομο βίντεο. Το βίντεο αυτό μπορεί να βρεθεί εδώ: ΛΗΨΗ

9 Συμπεράσματα - Προτάσεις

Στο τελευταίο αυτό κεφάλαιο αξίζει να γίνει μια σύντομη ανακεφαλαίωση της μελέτης αλλά και παρουσίαση προτάσεων για την εξέλιξη της. Σε αρχικό πλαίσιο γίνεται περιγραφή της σημασίας του κλάδου της αεροπορίας στις οικονομίες, καθώς και των προκλήσεων που αντιμετωπίζει. Συνεχίζοντας, περιεγράφηκαν οι τρόποι με τους οποίους η πρόβλεψη του αριθμού αφίξεων στα αεροδρόμια μπορεί να χρησιμοποιηθεί από αεροδρόμια, αεροπορικές εταιρείες, τοπικές κοινότητες, αλλά και από την κυβέρνηση για εξαγωγή κρίσιμων πληροφοριών.

Για να γίνει πλήρως κατανοητός ο τρόπος πρόβλεψης χρονοσειράς με μηχανική μάθηση, απαραίτητη είναι η γνώση θεωρίας. Συνεπώς ακολούθησε θεωρητική ανάλυση για τις χρονοσειρές αλλά και για την μηχανική μάθηση. Συνεχίζοντας, έγινε αναφορά στο θεωρητικό υπόβαθρο των επιλεγμένων μοντέλων πρόβλεψης αλλά και στον τρόπο κατηγοριοποίησης τους σε στατιστικά, νευρωνικά και ensemble μοντέλα. Λόγω της πληθώρας των υπερπαραμέτρων αλλά και της μεγάλης σημασίας τους στα μοντέλα, παρουσιάστηκαν συναρτήσεις ενεργοποίησης αλλά και αλγόριθμοι ελαχιστοποίησης σφάλματος. Για την ολοκλήρωση του θεωρητικού υποβάθρου, αναπτύχθηκε η απαραίτητη θεωρία για τις μετρικές αξιολόγησης.

Στο επόμενο επίπεδο της μελέτης, έγινε παρουσίαση του περιβάλλοντος που υλοποιήθηκε αυτή αλλά και οι τεχνολογίες που χρησιμοποιήθηκαν. Στο σημείο αυτό επίσης αναφέρθηκε ο τρόπος διαχείρισης των δεδομένων από την πηγή αλλά και ο λόγος που επιλέχθηκε το Διεθνές Αεροδρόμιο Αθηνών ως το υπό μελέτη αεροδρόμιο. Επιπλέον έγινε εξήγηση του τρόπου επεξεργασίας των αποκτημένων δεδομένων αλλά και των επιλεγμένων εξωγενών μεταβλητών.

Σε τελικό επίπεδο έγινε οπτικοποίηση των αποτελεσμάτων αλλά και αναφορά στα συμπεράσματα που δημιουργήθηκαν. Αναλυτικότερα, παρουσιάστηκε ότι τα ensemble μοντέλα απέδωσαν τα καλύτερα αποτελέσματα αλλά και ότι η μετατροπή των δεδομένων σε στάσιμα δεν επέφερε κάποια σημαντική βελτίωση στις προβλέψεις των μοντέλων. Επιπροσθέτως, παρατηρήθηκε ότι συμπεριλαμβάνοντας εξωγενείς μεταβλητές, όπως τον χρόνο, τις καραντίνες και τον αριθμό των εμβολιασμένων πολιτών κατά της Covid-19 δεν βελτιώθηκαν οι προβλέψεις. Κλείνοντας, έγινε παρουσίαση της εφαρμογής FlightForecaster η οποία δημιουργήθηκε με σκοπό την σύμπτυξη των σταδίων οπτικοποίησης των δεδομένων, εκτέλεσης των μοντέλων αλλά και παρουσίασης των ευρημάτων.

Η παραπάνω μελέτη με τον τρόπο που έχει υλοποιηθεί προγραμματιστικά μπορεί με ευκολία να εφαρμοστεί σε οποιοδήποτε άλλο αεροδρόμιο. Επιπλέον, επιλέγοντας περισσότερα από ένα αεροδρόμια κατά τη διαδικασία σάρωσης των αρχείων της πηγής, και αθροίζοντας τα ανά ημέρα αποτελέσματα, γίνεται δυνατή η υλοποίηση της μελέτης σε ένα ευρύτερο γεωγραφικό χώρο (π.χ. χώρα). Μια ακόμα πιθανή παραλλαγή της μελέτης θα μπορούσε να υλοποιηθεί με την χρήση διαφορετικού τρόπου διαχείρισης της μη στασιμότητας των δεδομένων για σύγκριση των αποτελεσμάτων. Κλείνοντας, αξίζει να αναφερθεί ότι πολύ ενδιαφέρουσα θα ήταν η επέκταση της μελέτης αυτής με την ενσωμάτωση διαφορετικών εξωγενών μεταβλητών. Παραδείγματα τέτοιων μεταβλητών θα μπορούσαν να είναι ο πληθωρισμός της χώρας, η μέση θερμοκρασία ή ακόμα και η τιμή των καυσίμων των αεροπλάνων.

10 Πηγές

[1] Whitelegg, John. "AVIATION: the social, economic and environmental impact of flying." Ashden Trust, London (2000)

[2] Wilder-Smith, Annelies. "The severe acute respiratory syndrome: impact on travel and tourism." Travel medicine and infectious disease 4.2 (2006): 53-60

[3] https://insete.gr/wp-content/uploads/2022/03/Bulletin_2203.pdf

[4] McCulloch, Warren S. "A logical calculus of ideas imminent in nervous activity." Biol Math. Biophys (1943)

[5] Cover, Thomas, and Peter Hart. "Nearest neighbor pattern classification." IEEE transactions on information theory 13.1 (1967): 21-27.

[6] Pearl, Judea. "Bayesian netwcrks: A model cf self-activated memory for evidential reasoning." Proceedings of the 7th conference of the Cognitive Science Society, University of California, Irvine, CA, USA. 1985

[7] Rumelhart, David E., Geoffrey E. Hinton, and Ronald J. Williams. Learning internal representations by error propagation. California Univ San Diego La Jolla Inst for Cognitive Science, 1985

[8] Cortes, C., Vapnik, V. Support-vector networks. Mach Learn 20, 273–297 (1995). https://doi.org/10.1007/BF00994018

[9] Ho, Tin Kam. "Random decision forests." Proceedings of 3rd international conference on document analysis and recognition. Vol. 1. IEEE, 1995.

[10] R. van Loon, “Machine Learning Explained: Understanding Supervised, Unsupervised, and Reinforcement Learning,” 2018.https://www.linkedin.com/pulse/machine-learning-explained-understandingsupervised-ronald-van-loon (accessed Apr. 25, 2020).

[11] Javatpoint, “Regression vs. Classification in Machine Learning.” 2021, [Online]. Available: https://www.javatpoint.com/machine-learning.

[12] Sheela KG, Deepa SN (2013). Review on methods to fix number of hidden neurons in neural networks. Mathematical Problems in Engineering, 2013: 425740.

[13] R. R. Maaliw, M. A. Ballera, Z. P. Mabunga, A. T. Mahusay, D. A. Dejelo and M. P. Seño, "An Ensemble Machine Learning Approach For Time Series Forecasting of COVID-19 Cases," 2021 IEEE 12th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), Vancouver, BC, Canada, 2021, pp. 0633-0640, doi: 10.1109/IEMCON53756.2021.9623074.

[14] G. Mélard, J.-M. Pasteels, International Journal of Forecasting (2000) Automatic ARIMA modeling including interventions, using time series expert software https://www.sciencedirect.com/science/article/pii/S0169207000000674

[15] Box, G.E.P. and Jenkins, G.M. (1970) Time Series Analysis: Forecasting and Control. Holden-Day, San Francisco.

[16] Perone G. Using the SARIMA Model to Forecast the Fourth Global Wave of Cumulative Deaths from COVID-19: Evidence from 12 Hard-Hit Big Countries. Econometrics. 2022; 10(2):18. https://doi.org/10.3390/econometrics10020018

[17] Pack, David J. "In defense of ARIMA modeling." International Journal of Forecasting 6.2 (1990): 211-218.

[18] Barnett, Adrian G., and Annette J. Dobson. Analysing seasonal health data. Vol. 30. Berlin: Springer, 2010.

[19] https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPRegressor.html

[20] N. Harun, S. S. Dlay and W. L. Woo, "Performance of Keystroke Biometrics Authentication System Using Multilayer Perceptron Neural Network (MLP NN)", 7th International Symposium on communication Systems Networks & Digital Signal Processing (CSNDSP 2010), pp. 711-714, 2010.

[21] Zhang A, Lipton ZC, Li M, Smola AJ (2019c). Optimization algorithms. In: Dive into deep learning. Available at https://www.d2l.ai/.

[22] Xie, Xuetao, Yi-Fei Pu, and Jian Wang. "A fractional gradient descent algorithm robust to the initial weights of multilayer perceptron." Neural Networks 158 (2023): 154-170.

[23] Nielsen, Michael A. Neural networks and deep learning. Vol. 25. San Francisco, CA, USA: Determination press, 2015.

[24] Ljung, Lennart. "Black-box models from input-output measurements." IMTC 2001. Proceedings of the 18th IEEE instrumentation and measurement technology conference. Rediscovering measurement in the age of informatics (Cat. No. 01CH 37188). Vol. 1. IEEE, 2001.

[25] Hochreiter, Sepp, and Jürgen Schmidhuber. "Long short-term memory." Neural computation 9.8 (1997): 1735-1780.

[26] Schuster, Mike, and Kuldip K. Paliwal. "Bidirectional recurrent neural networks." IEEE transactions on Signal Processing 45.11 (1997): 2673-2681.

[27] Hochreiter S: The Vanishing Gradient Problem During Learning Recurrent Neural Netsand Problem Solutions. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 1998, 6(02):107-116.

[28] Box, George EP, et al. Time series analysis: forecasting and control. John Wiley & Sons, 2015.

[29] Meyer, Gregory P. "An alternative probabilistic interpretation of the huber loss." Proceedings of the ieee/cvf conference on computer vision and pattern recognition. 2021.

[30] Chun Wang, Weihua Zhang, Cong Wu, Heng Hu, Wenjia Zhu, "Combined Prediction Method of Short-Term Distance Headway Based on EB-GRA-TCN", Journal of Advanced Transportation, vol. 2022, Article ID 6456186, 12 pages, 2022. https://doi.org/10.1155/2022/6456186

[31] Bai, Shaojie, J. Zico Kolter, and Vladlen Koltun. "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling." arXiv preprint arXiv:1803.01271 (2018)

[32] Lemaire, Quentin, and Andre Holzapfel. "Temporal convolutional networks for speech and music detection in radio broadcast." 20th International Society for Music Information Retrieval

Conference, ISMIR 2019, 4-8 November 2019. International Society for Music Information Retrieval, 2019.

[33] Fan, J., Zhang, K., Huang, Y. et al. Parallel spatio-temporal attention-based TCN for multivariate time series prediction. Neural Comput & Applic (2021). https://doi.org/10.1007/s00521-021-05958-z

[34] Huaitao Shi, Yajun Shang, Xiaochen Zhang, Yinghan Tang, "Research on the Initial Fault Prediction Method of Rolling Bearings Based on DCAE-TCN Transfer Learning", Shock and Vibration, vol. 2021, Article ID 5587756, 15 pages, 2021. https://doi.org/10.1155/2021/5587756

[35] Jingxuan Hou, Guijin Wang, Xinghao Chen, Jing-Hao Xue, Rui Zhu, Huazhong Yang; Proceedings of the European Conference on Computer Vision (ECCV) Workshops, 2018, pp. 0-0

[36] Oord, Aaron van den, et al. "Wavenet: A generative model for raw audio." arXiv preprint arXiv:1609.03499 (2016).

[37] Chen, Tianqi, and Carlos Guestrin. "Xgboost: A scalable tree boosting system." Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016.

[38] https://xgboost.readthedocs.io/en/latest/python/python_api.html#xgboost.XGBRegressor

[39] Ke, Guolin, et al. "Lightgbm: A highly efficient gradient boosting decision tree." Advances in neural information processing systems 30 (2017).

[40] L. Yang and A. Shami, "A Lightweight Concept Drift Detection and Adaptation Framework for IoT Data Streams," in IEEE Internet of Things Magazine, vol. 4, no. 2, pp. 96-101, June 2021, doi: 10.1109/IOTM.0001.2100012.

[41] Dhaliwal, S.S.; Nahid, A.-A.; Abbas, R. Effective Intrusion Detection System Using XGBoost. Information 2018, 9, 149. https://doi.org/10.3390/info9070149

[42] H. Li, Y. Cao, S. Li, J. Zhao and Y. Sun, "XGBoost Model and Its Application to Personal Credit Evaluation," in IEEE Intelligent Systems, vol. 35, no. 3, pp. 52-61, 1 May-June 2020, doi: 10.1109/MIS.2020.2972533

[43] Fang Z, Yang S, Lv C, et alApplication of a data-driven XGBoost model for the prediction of COVID-19 in the USA: a time-series studyBMJ Open 2022;12:e056685. doi: 10.1136/bmjopen-2021-056685

[44] Abdu-Aljabar, Rana Dhia'A., and Osama A. Awad. "A Comparative analysis study of lung cancer detection and relapse prediction using XGBoost classifier." IOP conference series: materials science and engineering. Vol. 1076. No. 1. IOP Publishing, 2021

[45] Schapire, R.E. (2013). Explaining AdaBoost. In: Schölkopf, B., Luo, Z., Vovk, V. (eds) Empirical Inference. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-41136-6_5

[46] Freund, Y., Schapire, R.E. (1995). A desicion-theoretic generalization of on-line learning and an application to boosting. In: Vitányi, P. (eds) Computational Learning Theory. EuroCOLT 1995. Lecture Notes in Computer Science, vol 904. Springer, Berlin, Heidelberg. https://doi.org/10.1007/3-540-59119-2_166

[47] Freund, Yoav, and Robert E. Schapire. "A decision-theoretic generalization of on-line learning and an application to boosting." Journal of computer and system sciences 55.1 (1997): 119-139.

[48] Drucker, Harris. “Improving Regressors using Boosting Techniques.” International Conference on Machine Learning (1997).

[49] Oeing, Jonas, et al. "Flooding Prevention in Distillation and Extraction Columns with Aid of Machine Learning Approaches." Chemie Ingenieur Technik 93.12 (2021): 1917-1929.

[50] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.HistGradientBoostingRegressor.html

[51] Ke, Guolin, et al. "Lightgbm: A highly efficient gradient boosting decision tree." Advances in neural information processing systems 30 (2017).

[52] Gayathri, Rajakumaran, et al. "A Comparative Analysis of Machine Learning Models in Prediction of Mortar Compressive Strength." Processes 10.7 (2022): 1387.

[53] https://www.ml-science.com/sigmoid-activation-function?rq=sigmoid

[54] https://machinelearningmastery.com/a-gentle-introduction-to-sigmoid-function/

[55] https://wwzzzzw.ml-science.com/tanh-activation-function

[56] https://www.ml-science.com/rectifier-activation-function?rq=relu

[57] T. Tieleman and G. Hinton. Lecture 6.5—rmsprop: Divide the gradient by a running average of its recent magnitude. COURSERA:Neural Networks for Machine Learning, 2012.

[58] Kingma, Diederik P., and Jimmy Ba. "Adam: A method for stochastic optimization." arXiv preprint arXiv:1412.6980 (2014).

[59] Chai, Tianfeng, and Roland R. Draxler. "Root mean square error (RMSE) or mean absolute error (MAE)?–Arguments against avoiding RMSE in the literature." Geoscientific model development 7.3 (2014): 1247-1250.

[60] https://pythoninstitute.org/about-python

[61] https://beautiful-soup-4.readthedocs.io/en/latest/

[62] https://pandas.pydata.org/docs/

[63] https://numpy.org/doc/

[64] https://matplotlib.org/stable/index.html

[65] https://seaborn.pydata.org/

[66] https://keras.io/api/

[67] https://scikit-learn.org/stable/

[68] https://www.sktime.org/en/stable/

[69] https://docs.streamlit.io/

[70] https://zenodo.org/record/7323875#.Y8mH1HZBxHU

[71] https://insete.gr/wp-content/uploads/2022/03/Bulletin_2203.pdf

[72] https://www.aia.gr/media/aces/aces6/03_Contribution_of_AIA_to_the_Greek_economy_mrVrettas.pdf

[73] https://www.hellenicaworld.com/Science/Mathematics/gr/GrammikiParemvoli.html

[74] Dare, Peter. "Linear and Nonlinear Models. Fixed effects, random effects, and mixed models." Geomatica 60.4 (2006): 382-383.

[75] https://ourworldindata.org/coronavirus/country/greece

[76] https://datafloq.com/read/machine-learning-explained-understanding-learning/

[77] https://pianalytix.com/perceptronmultilayer-neural-network-algorithm/

[78] https://www.researchgate.net/figure/Deep-Feed-Forward-Neural-Network_fig1_319901002

[79] https://www.codeproject.com/Articles/1272354/ANNT-Recurrent-neural-networks

[80] https://studymachinelearning.com/activation-functions-in-neural-network/