

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

Σχολή Πολιτικών Μηχανικών

Τομέας Μεταφορών και Συγκοινωνιακής Υποδομής



Πρόβλεψη χρόνου άφιξης λεωφορείων σε στάσεις με
χρήση μοντέλων Μηχανικής Μάθησης

Διπλωματική Εργασία

Ελένη Μαρία Νησιώτη

Επιβλέπουσα Καθηγήτρια: Ελένη Βλαχογιάννη,

Καθηγήτρια Σχολής Πολιτικών Μηχανικών ΕΜΠ

Αθήνα, Νοέμβριος 2025

Ευχαριστίες

Θα ήθελα να ευχαριστήσω την κυρία Ελένη Βλαχογιάννη, Καθηγήτρια της Σχολής Πολιτικών Μηχανικών Ε.Μ.Π. για την ανάθεση και επίβλεψη της παρούσας Διπλωματικής Εργασίας και την ευκαιρία που μου έδωσε να ασχοληθώ με ένα τόσο ενδιαφέρον θέμα. Είμαι ευγνώμων για την άψογη συνεργασία και την καθοδήγησή της.

Ιδιαίτερες ευχαριστίες θα ήθελα να εκφράσω στον Κωνσταντίνο Κατζηλιέρη, υποψήφιο Διδάκτορα της Σχολής Πολιτικών Μηχανικών Ε.Μ.Π. για τον εκτενή χρόνο που μου αφιέρωσε, την πολύτιμη και διαρκή βοήθειά του.

Ευχαριστώ πολύ για τις συμβουλές, τις γνώσεις και το άψογο κλίμα που μου προσέφεραν.

Τέλος, θα ήθελα να ευχαριστήσω την οικογένειά μου που με στήριξε καθ' όλη τη διάρκεια των σπουδών μου και έκαναν αυτό το ταξίδι δυνατό και ευχάριστο. Επίσης ευχαριστώ από καρδιάς όλους μου τους φίλους για την ενθάρρυνσή, την πίστη και την έμπνευση που μου προσέφεραν κατά τη διάρκεια αυτής της πορείας.

Τίτλος: Πρόβλεψη εκτιμώμενου χρόνου άφιξης λεωφορείων σε στάσεις με χρήση μοντέλων Μηχανικής Μάθησης

Ελένη Μαρία Νησιώτη

Επιβλέπουσα Καθηγήτρια: Ελένη Ι. Βλαχογιάννη

Σύνοψη

Σκοπός της παρούσας διπλωματικής εργασίας είναι η ανάπτυξη μοντέλων πρόβλεψης χρόνου άφιξης των λεωφορείων στις στάσεις στο οδικό δίκτυο της Αθήνας με την εφαρμογή μεθόδων μηχανικής μάθησης. Για το σκοπό αυτό αντλήθηκαν ιστορικά δεδομένα δρομολογίων, επικυρώσεις εισιτηρίων, καιρικών συνθηκών, αργιών και δεδομένων κυκλοφοριακού φόρτου από αισθητήρες κίνησης. Στη συνέχεια, έγινε επεξεργασία των δεδομένων για να χρησιμοποιηθούν στην εκπαίδευση των αλγόριθμων μηχανικής μάθησης. Από τα μοντέλα που δημιουργήθηκαν, πιο αποτελεσματικό κρίθηκε αυτό του XGBoost το οποίο φάνηκε να παρουσιάζει σταθερότητα και εύστοχες προβλέψεις. Η περαιτέρω ανάλυση των αποτελεσμάτων δείχνει πως τα μοντέλα ερμηνεύουν καλύτερα τις διακυμάνσεις και περιορίζουν τα σφάλματα σημαντικά με την προσθήκη πολλών διαφορετικών δεδομένων. Επιπλέον, η εκπαίδευση των μοντέλων απέδωσε καλύτερα με το διαχωρισμό των δεδομένων σε χρονικές ζώνες (πρωινές και βραδινές ώρες), λόγω της πολύ διαφορετικής κατανομής και συχνότητας των δρομολογίων.

Λέξεις κλειδιά: Χρόνος Άφιξης Λεωφορείου, Μοντέλα Πρόβλεψης, Μηχανική Μάθηση, Τυχαία Δάση, Ενισχυμένα Δέντρα Απόφασης

Title: Prediction of Bus Arrival Times at Stops Using Machine Learning Models

Eleni Maria Nisioti

Supervisor: Dr. Eleni I. Vlachogianni

Abstract

This thesis aims at the development of models for predicting bus arrival times at stops within the road network of Athens using machine learning methods. For this purpose, historical data were collected, including route information, ticket validations, weather conditions, holidays, and traffic volume data from motion sensors. Subsequently, the data were processed to be used for training the machine learning algorithms. Among the developed models, the XGBoost model proved to be the most effective, demonstrating stability and accurate predictions. Further analysis of the results indicates that the models interpret variations more effectively and significantly reduce errors with the inclusion of diverse data sources. Moreover, model training performed better when the data were separated into time zones (morning and night hours), due to the markedly different distribution and frequency of bus routes during these periods.

Keywords: Bus Arrival Time, Predictive Models, Machine Learning, Random Forest, Gradient Boosted Decision Trees

Περίληψη

Η πρόβλεψη του χρόνου άφιξης λεωφορείου στη στάση αποτελεί κρίσιμο παράγοντα για την αποτελεσματική λειτουργία των Μέσων Μαζικής Μεταφοράς και τη σωστή εξυπηρέτηση του επιβατικού κοινού. Για αρκετό καιρό επικρατούσαν οι απλές στατιστικές μέθοδοι, βασισμένες σε ιστορικά γεγονότα, παρόλα αυτά, όσο αυξάνεται η πολυπλοκότητα και το πλήθος των δεδομένων, κρίνονται ανεπαρκείς και αναποτελεσματικές. Με την ανάπτυξη της τεχνολογίας, οι μέθοδοι Μηχανικής Μάθησης έχουν εφαρμοστεί σε μεγάλο βαθμό σε προβλήματα προβλεψιμότητας, καθώς είναι αρκετά ικανά να ερμηνεύσουν τις πολύπλοκες σχέσεις μεταξύ πλήθους διαφορετικών δεδομένων.

Η παρούσα διπλωματική αποσκοπεί στην ανάπτυξη ενός μοντέλου μηχανικής μάθησης για την πρόβλεψη του χρόνου άφιξης λεωφορείου σε στάση. Εφαρμόστηκαν οι αλγόριθμοι Linear Regression, Random Forest και XGBoost με σκοπό να δημιουργηθούν μοντέλα και να βρεθεί ο εκείνο με την ακριβέστερη πρόβλεψη.

Η εφαρμογή της μεθοδολογίας πραγματοποιείται σε γραμμή αστικού λεωφορείου της Αθήνας και αξιοποιούνται ιστορικά δεδομένα δρομολογίων από των ΟΑΣΑ, επικυρώσεις εισιτηρίων από το gov.gr και καιρικές συνθήκες από το meteostat που αναφέρονται στη χρονική περίοδο από Νοέμβριο του 2024 έως Απρίλιο του 2025. Εξετάζεται κατά πόσο η ενίσχυση των δεδομένων βελτιώνει την ακρίβεια των μοντέλων, χωρίζεται η ανάλυση σε ζώνη ημέρας και νύχτας και επιλέγεται το πιο κατάλληλο μοντέλο. Πιο συγκεκριμένα, πραγματοποιήθηκε κατάλληλη επεξεργασία και μορφοποίηση των δεδομένων και έπειτα εκπαιδεύτηκε το κάθε μοντέλο και στις δύο χρονικές ζώνες. Η ίδια ανάλυση πραγματοποιήθηκε για αρχή στα χρονολογικά και γεωγραφικά δεδομένα πρώτα και έπειτα προστέθηκαν τα υπόλοιπα, ώστε να υπάρχει καλύτερη εικόνα στην απόδοση των αλγορίθμων. Σε κάθε περίπτωση έγινε σύγκριση μέσω μετρικών αξιολόγησης και απεικονίσεων της πρόβλεψης με τα πραγματικά δεδομένα.

Μέσω της έρευνας, προέκυψε ότι το βέλτιστο μοντέλο είναι εκείνο του XGBoost, για το οποίο παρατέθηκαν τα αναλυτικά αποτελέσματα. Συγκεκριμένα, τα αποτελέσματα δείχνουν ότι διατήρησε μία σταθερή και ικανοποιητική απόδοση και στις δύο κατηγορίες δεδομένων, γεγονός που επισημαίνει την αξιοπιστία του. Παρουσίασε πολύ καλές τιμές στις μετρικές αξιολόγησης, που αντιστοιχούν σε καλή επεξήγηση της διακύμανσης και χαμηλά σφάλματα στις προβλέψεις.

Από την ανάλυση των αποτελεσμάτων προέκυψαν σημαντικά συμπεράσματα. Αρχικά, επισημάνθηκε η σημασία της επιλογής των κατάλληλων μεταβλητών, αλλά και της σωστής επεξεργασίας τους για τη βέλτιστη απόδοση. Η συγκριτική αξιολόγηση δείχνει τη σημασία χρήσης των μοντέλων Μηχανικής Μάθησης σε σύγχρονα συγκοινωνιακά προβλήματα.

Η παρούσα διπλωματική αναφέρεται στο οδικό δίκτυο της Αθήνας, γεγονός που δεν επιβεβαιώνει ότι τα συμπεράσματά που προέκυψαν ισχύουν και για διαφορετικά δίκτυα.

Παρά τα ενδεικτικά αποτελέσματα που προέκυψαν, παρουσιάστηκαν κάποιοι περιορισμοί στην πορεία της έρευνας. Υπήρξε μία δυσκολία στη συγκέντρωση έγκυρων δεδομένων που να αντιστοιχούν στην ίδια χρονική περίοδο και για αυτό το λόγο, επιλέχθηκε σχετικά μικρό διάστημα, της τάξης των έξι μηνών. Επιπλέον, μία σημαντική πληροφορία που δεν αντλήθηκε, ήταν εκείνη του κυκλοφοριακού φόρτου στους δρόμους που αφορούν το εξεταζόμενο δρομολόγιο, καθώς πολλοί φωρατές βρίσκονται εκτός λειτουργίας.

Τα αποτελέσματα που προέκυψαν, μπορούν να αποτελέσουν την αρχή για μία πιο εμπειριστατωμένη έρευνα. Μέσα στα επόμενα χρόνια, θα έχει συγκεντρωθεί ένας πολύ μεγαλύτερος όγκος δεδομένων, που θα δίνει τη δυνατότητα ανάλυσης παραπάνω λεωφορειακών γραμμών και μέσων για μεγάλα χρονικά διαστήματα. Ακόμα, περαιτέρω μεταβλητές όπως ο κυκλοφοριακός φόρτος, τα στατιστικά ατυχημάτων και περιπτώσεις αποκλεισμών δρόμων, θα έδιναν πιο άγκυρα αποτελέσματα. Σημαντικά αποτελέσματα θα έδινε και η χρήση επιπλέον μοντέλων Βαθιάς Μάθησης και υβριδικών μοντέλων.

Περιεχόμενα

1.Εισαγωγή	1
1.1. Ορισμός του προβλήματος.....	1
1.2.Σκοπός	1
1.3.Διάρθρωση Διπλωματικής	2
2.Βιβλιογραφική Ανασκόπηση	3
2.1.Χρόνος Άφιξης σε στάση.....	3
2.2.Μοντέλα Εκτίμησης Χρόνου Άφιξης σε Στάση	5
2.3.Βασικά Συμπεράσματα.....	11
3.Μεθοδολογική Προσέγγιση.....	12
3.1. Ροή Εργασιών	12
3.2.Θεωρητικό Υπόβαθρο	15
3.2.1.Γραμμική Παλινδρόμηση	15
3.2.2.Τυχαία Δάση	15
3.2.3.Ενισχυμένη Κλιμακωτή Ενίσχυση	16
3.3.Εκπαίδευση μοντέλων.....	17
3.4.Αξιολόγηση μοντέλων	17
3.7.Σημασία Χαρακτηριστικών	19
4.Εφαρμογή Μεθοδολογίας και Αποτελέσματα	19
4.1. Συλλογή και επεξεργασία δεδομένων	19
4.1.1.Καθορισμός και μορφοποίηση δεδομένων	20
4.1.2.Διαχωρισμός Δεδομένων ανά Χρονική Ζώνη	21
4.1.3.Προσθήκη εξωτερικών δεδομένων.....	22
4.1.4.Διαγράμματα πραγματικής μεταβολής.....	23
4.2. Ανάπτυξη Μοντέλων και Αποτελέσματα	25
4.2.1.Παραμετροποίηση μοντέλων	25
4.2.2.Γενίκευση Μοντέλου	25
4.2.3.Συνολικά Αποτελέσματα	27
4.3. Συγκριτική Αξιολόγηση.....	32
4.4. Ανάλυση Αποτελεσμάτων ΧGBoost.....	34
5.Συμπεράσματα	39
5.1.Βασικά Συμπεράσματα.....	39
5.2.Προτάσεις για Περαιτέρω Έρευνα.....	41

Ευρετήριο Εικόνων

Εικόνα 1: Απεικόνιση στάσεων της διαδρομής του 608.....	13
Εικόνα 2: Απεικόνιση φωρατών που αντιστοιχούν στη διαδρομή του 608	13
Εικόνα 3: Διάγραμμα Ροής Εργασιών	14

Ευρετήριο Διαγραμμάτων

Διάγραμμα 1:Διάγραμμα σύγκρισης πρόβλεψης και πραγματικών τιμών για Τυχαία Δάση - ΗΜΕΡΑ.....	23
Διάγραμμα 2:Διάγραμμα σύγκρισης πρόβλεψης και πραγματικών τιμών για Γραμμική Παλινδρόμηση- ΗΜΕΡΑ	24
Διάγραμμα 3:Διάγραμμα σύγκρισης πρόβλεψης και πραγματικών τιμών για ΧGBoost - ΗΜΕΡΑ.....	24
Διάγραμμα 4:Απεικόνιση Σημασίας Χαρακτηριστικών για ΧGBoost κατά τη διάρκεια της ημέρας	35
Διάγραμμα 5:Απεικόνιση Σημασίας Χαρακτηριστικών για ΧGBoost κατά τη διάρκεια της νύχτας.....	35
Διάγραμμα 6:Διαγράμματα προβλεπόμενων και πραγματικών χρόνων άφιξης για ΧGBoost (Ημέρα)	37
Διάγραμμα 7:Διαγράμματα προβλεπόμενων και πραγματικών χρόνων άφιξης για ΧGBoost (Νύχτα)	38

Ευρετήριο Πινάκων

Πίνακας 1: Συστήματα πρόβλεψης χρόνου άφιξης σε στάση ευρωπαϊκών πόλεων	4
Πίνακας 2: Συνολικά αποτελέσματα κατά τη διάρκεια της ημέρας.....	27
Πίνακας 3: Συνολικά αποτελέσματα κατά τη διάρκεια της νύχτας.....	28
Πίνακας 4: Αποτελέσματα R^2 κατά τη διάρκεια της ημέρας	28
Πίνακας 5: Αποτελέσματα R^2 κατά τη διάρκεια της νύχτας	29
Πίνακας 6: Αποτελέσματα MAE κατά τη διάρκεια της ημέρας	29
Πίνακας 7: Αποτελέσματα MAE κατά τη διάρκεια της νύχτας	30
Πίνακας 8: Αποτελέσματα MSE κατά τη διάρκεια της ημέρας	30
Πίνακας 9: Αποτελέσματα MSE κατά τη διάρκεια της νύχτας	31
Πίνακας 10: Αποτελέσματα μετρικών Αξιολόγησης για κάθε μοντέλο κατά τη διάρκεια της ημέρας.....	31
Πίνακας 11: Αποτελέσματα μετρικών Αξιολόγησης για κάθε μοντέλο κατά τη διάρκεια της ημέρας	32
Πίνακας 12: Σύγκριση μοντέλων και γενικές παρατηρήσεις.....	33
Πίνακας 13: Σύγκριση ημέρας και νύχτας βάσει διαγραμμάτων	34
Πίνακας 14: Αξιολόγηση κόστους – οφέλους στην προσθήκη δεδομένων.....	40

1.Εισαγωγή

1.1. Ορισμός του προβλήματος

Στον τομέα των δημόσιων συγκοινωνιών, τα τελευταία χρόνια, πραγματοποιούνται μεγάλα βήματα εκσυγχρονισμού για τη βελτιστοποίηση της κινητικότητας (*Public Transport*, 2014). Σε αστικά δίκτυα, όπως είναι η Αθήνα, παρουσιάζονται αρκετά ζητήματα στη διαδικασία αυτή, λόγω της υψηλής πυκνότητας της κυκλοφορίας και της άναρχης χωροταξίας. Παρατηρούνται έντονες διακυμάνσεις και συχνές καθυστερήσεις, λόγω απρόβλεπτων παραγόντων, όπως καιρικά φαινόμενα, εκδηλώσεις, έργα οδοποιίας και αυξομειώσεις στη ζήτηση μετακινήσεων. Ιδιαίτερα οι λεωφορειακές γραμμές, που αποτελούν βασικό μέσο μεταφοράς χιλιάδων επιβατών καθημερινά, αντιμετωπίζουν προκλήσεις στην αξιοπιστία των δρομολογίων, με αποτέλεσμα μειωμένη ικανοποίηση των χρηστών και οικονομικές συνέπειες για τον φορέα λειτουργίας (Gaschi-Uciecha, 2023).

Στις μέρες μας, ο σωστός σχεδιασμός και λειτουργία ενός συστήματος μαζικών μεταφορών είναι μείζονος σημασίας, καθώς η εύρυθμη λειτουργία των συγκοινωνιακών υποδομών απειλείται σημαντικά με τη ραγδαία αύξηση των ιδιωτικών οχημάτων ΙΧ. Προκειμένου να αποφευχθεί ο υπερκορεσμός, η μόνη λύση, είναι η δημιουργία και η ρύθμιση ενός λειτουργικού και αποτελεσματικού δικτύου που θα ικανοποιεί τις ανάγκες της πόλης και θα αποτελεί αποτρεπτικό παράγοντα χρήσης των ΙΧ, ενώ θα ρυθμίζει την κυκλοφορία στους δρόμους (Cleland et al., 2023).

Υπό αυτές τις συνθήκες, η ανάγκη πρόβλεψης με υψηλή ακρίβεια του χρόνου άφιξης λεωφορείων αυξάνεται σημαντικά (Jeong et al., 2024). Η πρόβλεψη αυτή θα ενισχύσει την αυτοματοποιημένη ενημέρωση του χρόνου άφιξης σε στάσεις και θα συμβάλει στη δημιουργία και εφαρμογών που αναμένεται να βελτιώσουν σε μεγάλο βαθμό τη μετακίνηση των πολιτών και τη λειτουργία της πόλης γενικότερα, όπως έχει συμβεί σε μεγάλες ευρωπαϊκές πόλεις. Με ακριβείς προβλέψεις υπάρχει η δυνατότητα βελτιστοποίησης της κατανομής του στόλου, ενισχύοντας την κινητικότητα και επιφέροντας έτσι περιβαλλοντικά και οικονομικά οφέλη.

1.2.Σκοπός

Σκοπός της παρούσας διπλωματικής είναι η ανάπτυξη μοντέλων πρόβλεψης του χρόνου άφιξης των Μέσων Μαζικής Μεταφοράς στη στάση. Συγκεκριμένα, η έρευνα εστιάζει στις λεωφορειακές γραμμές, οι οποίες αποτελούν το πιο πολύπλοκο μέσο λόγω της εμπλοκής τους στις πραγματικές κυκλοφοριακές συνθήκες της πόλης, καθιστώντας τα δεδομένα που θα χρησιμοποιηθούν πολυπληθή και περίπλοκα στην επεξεργασία. Ακόμα, γίνεται προσπάθεια ενσωμάτωσης και άλλων πληροφοριών στα μοντέλα πρόβλεψης, όπως καιρικές και κυκλοφοριακές συνθήκες στο οδικό δίκτυο. Διερευνώντας τεχνολογίες πρόβλεψης που βασίζονται στη Μηχανική μάθηση, όπως Linear Regression, Random

Forest Regression, XGBoost και άλλων υβριδικών συστημάτων που μπορούν να επεξεργαστούν πολύπλοκες αλληλεπιδράσεις μεταξύ χωρικών, χρονικών και εξωτερικών μεταβλητών. Τέτοιες τεχνολογίες προσφέρουν αυξημένη προσαρμοστικότητα, ακρίβεια και δυνατότητα πρόβλεψης σε μεταβαλλόμενες συνθήκες, για παράδειγμα, αλλαγές στην κυκλοφορία λόγω καιρού ή αργιών.

Στην παρούσα χρησιμοποιούνται εργαλεία Μηχανικής Μάθησης σε περιβάλλον Python, αξιοποιώντας έναν όγκο διαφορετικών δεδομένων όπως:

- **Δεδομένα δρομολογίων:** Χρονικά στίγματα άφιξης σε στάσεις και γεωγραφική απόσταση μεταξύ αυτών.
- **Καιρικές συνθήκες:** Θερμοκρασία και βροχόπτωση
- **Ζήτηση επιβατών:** Επικυρώσεις εισιτηρίων σε πραγματικό χρόνο.
- **Παρουσία Λεωφορειολωρίδας:** Αποκλειστική κίνηση λεωφορείου σε ειδικά σχεδιασμένη λωρίδα.

1.3.Διάρθρωση Διπλωματικής

Η διάρθρωση εννοιών της παρούσας διπλωματικής έχει ως εξής:

Στο δεύτερο κεφάλαιο παρουσιάζεται βιβλιογραφική ανασκόπηση που αφορά την πρόβλεψη χρόνου άφιξης Μέσων Μαζικής Μεταφοράς σε ένα συγκεκριμένο σημείο. Πιο συγκεκριμένα, αναλύονται σχετικοί μέθοδοι και τα δεδομένα που χρησιμοποιήθηκαν. Περιγράφεται η χρήση στατιστικών μεθόδων, Μηχανικής Μάθησης και Νευρωνικών Δικτύων σε διάφορες περιπτώσεις Μέσων Μεταφοράς και αναπτύσσονται τα αποτελέσματα, αλλά και τα συμπεράσματα. Επιπλέον, επισημαίνονται τεχνολογίες και μέθοδοι που εφαρμόζονται σε διάφορες ευρωπαϊκές πόλεις σήμερα.

Στο τρίτο κεφάλαιο, αναλύεται η μεθοδολογική προσέγγιση που ακολουθήθηκε για την πρόβλεψη χρόνου άφιξης λεωφορειακής γραμμής σε στάση. Διατυπώνεται το πρόβλημα, αλλά και τα διαθέσιμα δεδομένα που πρόκειται να χρησιμοποιηθούν στην παρούσα διπλωματική. Απεικονίζεται η ροή εργασιών και γίνεται μία γενική περιγραφή της πορείας διαδικασιών που θα ακολουθήσει. Στη συνέχεια παρουσιάζεται το θεωρητικό υπόβαθρο των μεθόδων και των εργαλείων που θα χρησιμοποιηθούν. Πιο συγκεκριμένα, παρουσιάζονται οι ιδιότητες του κάθε αλγόριθμου και η διαδικασία εκπαίδευσης, αλλά και αξιολόγησης των μοντέλων.

Το τέταρτο κεφάλαιο παραθέτει αναλυτικά τα αποτελέσματα που προέκυψαν. Περιγράφει λεπτομερώς τη διαδικασία συλλογής και επεξεργασίας δεδομένων και τη δημιουργία γραφημάτων. Έπειτα, εστιάζει στην ανάπτυξη των μοντέλων και στα αποτελέσματα που προέκυψαν μέσω των μετρικών αξιολόγησης. Τέλος, παρουσιάζεται το βέλτιστο μοντέλο, σύμφωνα με τα αποτελέσματα.

Στο πέμπτο κεφάλαιο αναφέρονται τα συμπεράσματα που προκύπτουν από τα αποτελέσματα, αλλά και τρόποι βελτιστοποίησης και εναλλακτικές λύσεις.

2.Βιβλιογραφική Ανασκόπηση

2.1.Χρόνος Άφιξης σε στάση

Η πρόβλεψη του χρόνου άφιξης (Estimated Time Arrival - ETA) σε στάση αποτελεί ένα πρόβλημα που έχει συγκεντρώσει το ενδιαφέρον αρκετών ερευνητών. Έχουν πραγματοποιηθεί πολλές έρευνες από τις οποίες έχουν προκύψει διάφορες μέθοδοι, που εξελίσσονται και ενισχύονται όσο η τεχνολογία προοδεύει. Μέσω των δεδομένων πραγματικού χρόνου που προσφέρουν τα συστήματα AVL/GPS, έχουν αναπτυχθεί αλγόριθμοι που μπορούν να προβλέψουν με μεγαλύτερη ακρίβεια τις αφίξεις μειώνοντας την αβεβαιότητα (Turay et al., 2025). Πλέον, αντιμετωπίζεται ως ένα πολυπαραγοντικό και δυναμικό πρόβλημα, που επηρεάζεται άμεσα από εξωτερικούς παράγοντες όπως είναι οι κυκλοφοριακές και καιρικές συνθήκες, η ζήτηση και τα χαρακτηριστικά μιας πόλη (N. Shanthi et al., 2022)

Η βιβλιογραφία των τελευταίων δεκαετιών ξεφεύγει από τις απλές στατιστικές προσεγγίσεις, όπως είναι οι μέσοι ιστορικοί χρόνοι και τα γραμμικά μοντέλα παλινδρόμησης και εστιάζει σε πιο σύνθετες μεθοδολογίες μηχανικής μάθησης (Machine Learning) και βαθιάς μάθησης (Deep Learning). Μοντέλα όπως τα φίλτρα Kalman, οι χρονοσειρές ARIMA, τα δέντρα αποφάσεων, οι μέθοδοι boosting, τα νευρωνικά δίκτυα (LSTM, GRU) και τα γραφικά νευρωνικά δίκτυα έχουν εφαρμοστεί σε διάφορα περιβάλλοντα, με στόχο την αύξηση της ακρίβειας και της ανθεκτικότητας σε απρόβλεπτες συνθήκες (Reich et al., 2019). Ωστόσο, η ετερογένεια των δεδομένων, η έλλειψη τυποποιημένων μετρικών και η απουσία κοινών συνόλων δεδομένων εξακολουθούν να αποτελούν προκλήσεις, καθιστώντας απαραίτητη μια συστηματική βιβλιογραφική ανασκόπηση που να χαρτογραφεί τις υφιστάμενες μεθόδους, τα πλεονεκτήματα και τους περιορισμούς τους (Alexandre et al., 2023).

Στον **Πίνακα 1** παρουσιάζονται τα συστήματα που χρησιμοποιούνται σε διάφορες ευρωπαϊκές πόλεις για την πρόβλεψη χρόνου άφιξης σε στάση:

Πίνακας 1: Συστήματα πρόβλεψης χρόνου άφιξης σε στάση ευρωπαϊκών πόλεων

Πόλη	Σύστημα	Τεχνολογίες	Μοντέλο/Αλγόριθμος	Παράμετροι	Ανοιχτά Δεδομένα
Βερολίνο-BVG, Μοναχό-MVG	Ενσωματωμένα συστήματα RTPI (Real-Time Passenger Information)	Χρήση GPS + δίκτυο 4G/5G για live παρακολούθηση.	Machine Learning μοντέλα	Πραγματικός χρόνος άφιξης Χρονικά μοτίβα κυκλοφορίας Δεδομένα από Google Maps/HERE Maps	Παροχή δεδομένων μέσω APIs
Παρίσι-RATP	Συστήματα πρόβλεψης AVL (Automatic Vehicle Location) και ITS (Intelligent Transport Systems)	Χρήση Real-time GPS και detectors	Hybrid αλγόριθμος με στατιστικά και Machine Learning μοντέλα	Ιστορικά δεδομένα Πραγματικός χρόνος άφιξης Δεδομένα καυσίμου, θερμοκρασίας και πληρότητας επιβατών Πρόβλεψη συμφόρησης	Παροχή δεδομένων μέσω APIs πλήρους πρόσβασης για developers (Citymapper, Transit App, Google maps κτλ)
Άμστερνταμ-GVB	Προηγμένα predictive AVL (Automatic Vehicle Location) συστήματα που χρησιμοποιούν ακόμα και αισθητήρες στις στάσεις	Χρήση Τεχνητής Νοημοσύνης	Deep Learning Μοντέλα από συνεργασία με ερευνητικά ινστιτούτα	Θέση οχήματος Προγραμματισμένο δρομολόγιο Τρέχουσα ταχύτητα Ιστορικά δεδομένα κίνησης Καιρικές συνθήκες Συμβάντα στο δίκτυο	Παροχή δεδομένων μέσω APIs πλήρους πρόσβασης για developers
Λονδίνο-TfL	Προηγμένο σύστημα iBus	Χρήση Real-time GPS και ασύρματου modem	Bayesian μοντέλα πρόβλεψης	Ανάλυση διαδρομής Καθυστερήσεις Μεταβλητότητα χρόνου στάσεων Θέση οχήματος Ιστορικά δεδομένα	Ανοιχτά δεδομένα μέσω TfL Unified API
Στοκχόλμη-SL	Συστήματα πρόβλεψης AVL	Χρήση Real Time GPS στέλνουν συνεχή δεδομένα στο κέντρο ελέγχου	Hybrid μοντέλα που χρησιμοποιούν στατιστικά μοτίβα και real-time δεδομένα συνδυάζοντας forecasting και adaptive learning	Προβλέψεις καιρού Crowd analytics	Ανοιχτά δεδομένα API για ανάπτυξη εφαρμογών

2.2. Μοντέλα Εκτίμησης Χρόνου Άφιξης σε Στάση

Οι YU Bo et. al. (2011) παρουσίασαν ένα προσαρμοστικό μοντέλο πρόβλεψης χρόνου άφιξης λεωφορείων στη στάση, το οποίο συνδυάζει απλά γραμμικά μοντέλα παλινδρόμησης με ένα μηχανισμό συνεχούς διόρθωσης σε πραγματικό χρόνο. Πιο συγκεκριμένα, η έρευνα αποτελείται από:

- Τα γραμμικά μοντέλα παλινδρόμησης τα οποία χρησιμοποιούνται για την εκτίμηση των βασικών χρόνων διαδρομής μεταξύ των στάσεων με βάση τα ιστορικά δεδομένα (baseline)
- Τον προσαρμοστικό αλγόριθμο, που ενσωματώνει τις πιο πρόσφατες αφίξεις λεωφορείων ώστε να διορθώνει σε πραγματικό χρόνο τα σφάλματα

Η προτεινόμενη προσέγγιση εφαρμόστηκε στη λεωφορειακή γραμμή No. 23 στην πόλη Dalian της Κίνας, αποτελούμενη από 19 στάσεις και 14,5 km. Συλλέχθηκαν 520 έγκυρα δρομολόγια μέσα σε ένα μήνα, που αντιστοιχούν σε 7.800 μετρήσεις χρόνων διαδρομής ανά τμήμα.

Μετά την ανάλυση προέκυψε ότι τα γραμμικά μοντέλα παλινδρόμησης είχαν ήδη καλύτερη επίδοση από τα επίσημα χρονοδιαγράμματα και το προσαρμοστικό μοντέλο μείωσε περαιτέρω τα σφάλματα σύμφωνα με τις ενδείξεις του RMSE. Οι προβλέψεις ήταν πιο σταθερές και αξιόπιστες, παρόλα αυτά, σε ορισμένα τμήματα, η προσαρμοστική μέθοδος εμφάνισε μεγαλύτερα σφάλματα λόγω συσσώρευσης λαθών από προηγούμενα τμήματα. Επιπλέον, τα αποτελέσματα δείχνουν ότι η ενσωμάτωση των πιο πρόσφατων δεδομένων βελτιώνει σημαντικά την ακρίβεια πρόβλεψης σε σχέση με τη χρήση αποκλειστικά ιστορικών δεδομένων. Τέλος, το προτεινόμενο σχήμα είναι υπολογιστικά αποδοτικό, μπορεί να λειτουργήσει online και είναι κατάλληλο για εφαρμογές σε συστήματα πληροφόρησης επιβατών σε πραγματικό χρόνο.

Σύμφωνα με την έρευνα, ο συνδυασμός της γραμμικής παλινδρόμησης με τους προσαρμοστικούς αλγόριθμους σε πραγματικό χρόνο βελτιώνει την προβλεψιμότητα του χρόνου άφιξης σε σχέση με τα απλά στατιστικά μοντέλα και χρονοδιαγράμματα, ιδιαίτερα σε περιπτώσεις κυκλοφοριακής αβεβαιότητας.

Μία διαφορετική προσέγγιση πραγματοποιήθηκε από τους Altinkaya & Zontul et al. (2013), η οποία εξετάζει διαφορετικές προσεγγίσεις πρόβλεψης. Αρχικά, μοντέλα βασισμένα σε ιστορικά δεδομένα (μέσοι χρόνοι, μέσες ταχύτητες, συνδυασμοί με πραγματικά GPS). Στατιστικά μοντέλα (χρονικές σειρές, παλινδρόμηση), μοντέλα Kalman Filtering, που συνδυάζουν δεδομένα πραγματικού χρόνου και ιστορικά, προσφέροντας δυνατότητα δυναμικής ενημέρωσης, μοντέλα Μηχανικής Μάθησης (Artificial Neural Networks, Support Vector Machines) και υβριδικά μοντέλα, δηλαδή πολλαπλές προσεγγίσεις

Από αυτή την έρευνα προκύπτει πως τα ιστορικά μοντέλα, παρόλο που είναι απλά και μπορούν να εφαρμοστούν εύκολα, είναι αρκετά αναξιόπιστα σε μεγάλες πόλεις με

πολύπλοκα δεδομένα, καθώς κάνουν υποθέσεις για σταθερότητα που δεν ισχύουν στην πραγματικότητα. Τα στατιστικά μοντέλα είναι μία βελτιωμένη μέθοδος που περιγράφει καλύτερα τη μεταβλητότητα των συνθηκών, αλλά με αρκετά περιορισμένη ακρίβεια, καθώς αδυνατεί να περιγράψει συσχέτιση των μεταβλητών. Τα Kalman Filtering χρησιμοποιούνται στην online πρόβλεψη και υπάρχει η δυνατότητα συνεχούς βελτίωσης καθώς ανανεώνονται τα δεδομένα. Λειτουργούν με μεγαλύτερη ακρίβεια από τα ιστορικά και τα γραμμικά μοντέλα. Τα ANN, σε συνδυασμό με δεδομένα σε πραγματικό χρόνο, δίνουν υψηλή ακρίβεια και μπορούν να επεξεργαστούν μη γραμμικές σχέσεις. Χρειάζονται όμως ιδιαίτερη προσοχή στην εκπαίδευση με κατάλληλα datasets. Τα SVM παρουσιάζουν ανθεκτικότητα σε overfitting και οι προβλέψεις είναι αρκετά αξιόπιστες, αλλά είναι πολύ βαριά υπολογιστικά σε μεγάλα σύνολα δεδομένων. Τέλος, τα υβριδικά μοντέλα χρησιμοποιούνται όλο και πιο ευρέως και συνδυάζουν χρήσιμες λειτουργίες των διαφορετικών μεθόδων.

Ανάλογα με τις ανάγκες της κάθε περίπτωσης και τη φύση των δεδομένων, επιλέγεται το κατάλληλο μοντέλο, χωρίς να υφίσταται η έννοια της καλύτερης μεθόδου. Τα υβριδικά μοντέλα αποτελούν μία αρκετά συχνή λύση, λόγω της ευελιξίας τους.

Ιδιαίτερη έμφαση δίνεται στη Μηχανική Μάθηση (Machine Learning) με ανάλυση παλινδρόμησης όσον αφορά τους χρόνους άφιξης των ΜΜΜ τα τελευταία χρόνια, με αρκετές έρευνες να εστιάζουν γύρω από τις διάφορες εκδοχές του.

Η έρευνα των Md. Miraz Kabir et al. (2018) που πραγματοποιήθηκε στο Μπαγλαντές εξετάζει διάφορες περιπτώσεις με χρήση διαφορετικών μοντέλων. Πιο συγκεκριμένα χρησιμοποιούνται η Γραμμική Παλινδρόμηση (Linear Regression), η Παλινδρόμηση με Νευρωνικά Δίκτυα (Neural Network Regression), η Παλινδρόμηση Poisson (Poisson Regression) και η Ταξινομημένη Παλινδρόμηση (Ordinal Regression).

Η αξιολόγησή τους πραγματοποιείται με κατάλληλες μετρικές σφάλματος (MAE, RMSE, RAE, RSE), μέσω των οποίων προκύπτουν κάποια αποτελέσματα για τα μοντέλα. Η Linear Regression αποδίδει την καλύτερη ακρίβεια σύμφωνα με τις δοκιμές. Η Neural Regression είναι κατάλληλη για μεγάλα datasets και δίνει πολύ καλή ακρίβεια. Η Poisson Regression είναι αρκετά σταθερή, αλλά παρουσιάζει μεγαλύτερες αποκλίσεις από τα πραγματικά δεδομένα. Τέλος, η Ordinal Regression, είναι ικανοποιητική αλλά επηρεάζεται πολύ από ακραίες τιμές.

Το συμπέρασμα που προκύπτει από την έρευνα είναι ότι για τη συγκεκριμένη τοποθεσία και τα δεδομένα που επεξεργάστηκαν, η Linear Regression φάνηκε να έδωσε τα πιο αξιόπιστα αποτελέσματα με τη Neural Regression να έχει τη δυνατότητα να επεξεργαστεί γρηγορότερα πολύπλοκες βάσεις δεδομένων.

Στη Μελβούρνη πραγματοποιήθηκε αντίστοιχη έρευνα από τους Malzoumi E. et al., (2012) που αφορά τη λεωφορειακή γραμμή 246 και αποσκοπεί στην πρόβλεψη των χρόνων διαδρομής των λεωφορείων. Η γραμμή αυτή είναι 8 χλμ. και χωρίστηκε σε τέσσερα

ισομήκη τμήματα και έχουν ορισθεί πέντε χρονικά σημεία. Για την ανάλυση συγκεντρώθηκαν πραγματικά δεδομένα από συσκευές GPS.

Χρησιμοποιήθηκαν αρκετές μέθοδοι, τόσο για την ταξινόμηση των μεταβλητών, όσο και για την πρόβλεψη, όπως Regression Analysis για την επιρροή κάθε μεταβλητής, Artificial Neural Network (ANN) models για την εκπαίδευση των μοντέλων με τα δεδομένα που αντλήθηκαν από το GPS και Traffic Flow Data-Based Model, Historical Data-Based Model, Timetable-Based Model για την πρόβλεψη του χρόνου άφιξης.

Τα ιστορικά δεδομένα αφορούν σε μεταβλητές, όπως η ώρα της ημέρας, η ημέρα της εβδομάδας, ο μήνας του χρόνου και κατά πόσο είναι εντός χρονοδιαγράμματος.

Αντίστοιχα το Timetable Model χρησιμοποιεί τους προγραμματισμένους χρόνους για να υπολογίσει τους χρόνους διαδρομής.

Για την πρόβλεψη ETA χρησιμοποιήθηκαν δεδομένα όπως χρόνοι άφιξης στη στάση σε σύγκριση με τους πραγματικούς χρόνους, βροχόπτωση (rainfall) μετρούμενη σε χιλιοστά για την ώρα της μέτρησης και κυκλοφοριακοί φόρτοι στο κομμάτι που αφορά τη διαδρομή της συγκεκριμένης λεωφορειακής γραμμής, οι οποίοι αντλήθηκαν από κυκλοφοριακούς φωρατές (inductive loop detectors). Οι κυκλοφοριακές συνθήκες και ο βαθμός κορεσμού μετρήθηκαν για τη μελετώμενη γραμμή και τις διασταυρούμενες οδούς.

Σύμφωνα με τα αποτελέσματα, το Timetable-Based Model έδωσε τα πιο ανακριβή αποτελέσματα και το Traffic Flow Data-Based Model τα πιο έγκυρα. Δεν παρατηρήθηκε μεγάλη απόκλιση μεταξύ του τελευταίου και του Historical Data-Based Model, γεγονός που υποδεικνύει πως οι χρονικές μεταβλητές μπορούν να οδηγήσουν σε αρκετά αξιόπιστες προβλέψεις.

Μια άλλη έρευνα που πραγματοποιήθηκε στη Μαλαισία (Noor et al., 2020) και επικεντρώνεται στην πρόβλεψη του χρόνου άφιξης λεωφορείων (ETA) αποσκοπεί στη βελτίωση της αξιοπιστίας των δημόσιων συγκοινωνιών με χρήση Support Vector Regression (SVR).

Τα δεδομένα αντλήθηκαν από τη γραμμή PJ03 στο Petaling Jaya (9,4 km, 23 στάσεις, 350 δρομολόγια μέσα σε μία εβδομάδα). Κατηγοριοποιήθηκαν σε τμήματα διαδρομής, διάρκεια ταξιδιού, απόσταση και ώρα. Επιπλέον τα δεδομένα εμπλουτίστηκαν με μετεωρολογικές πληροφορίες. Έγινε και χρήση RBF kernel και βελτιστοποίηση παραμέτρων,.

Το μοντέλο έδωσε καλή πολύ καλή ακρίβεια με μέσο σφάλμα (RMSE) περίπου 22 δευτερόλεπτα, που είναι καλύτερο από προηγούμενες έρευνες που παρατάθηκαν. Στα περισσότερα τμήματα παρουσιάστηκε σφάλμα κάτω από 45 δευτερόλεπτα, ενώ ο καιρός δεν αποδείχθηκε σημαντικός παράγοντας, καθώς η απόδοση του μοντέλου δεν επηρεάστηκε σημαντικά με την προσθήκη των μετεωρολογικών δεδομένων. Επιπλέον τα αποτελέσματα δεν είναι τόσο καθοριστικά, καθώς η εκπαίδευση περιορίστηκε σε δεδομένα μίας εβδομάδας.

Το συμπέρασμα που προέκυψε είναι ότι το SVR έχει τη δυνατότητα να γενικεύει, γεγονός που το καθιστά αρκετά αποδοτικό, ακόμα και σε δεδομένα μικρής χρονικής περιόδου. Παρόλα αυτά, πιο εκτενή datasets, θα βελτίωναν την ακρίβεια. Πιο συγκεκριμένα, παραπάνω δεδομένα για τις κυκλοφοριακές συνθήκες θα εμπλούτιζαν τα αποτελέσματα.

Παρατηρείται ότι γενικά στις συγκοινωνιακές αναλύσεις είναι απαραίτητη η πρόβλεψη χρόνων για την ομαλή λειτουργία ακόμα και του εμπορίου. Σε ανάλυση που έγινε από τον Balster et al. (2020) εξετάζεται η πρόβλεψη ETA σε διατροπικά δίκτυα μεταφορών με χρήση Machine Learning. Πιο συγκεκριμένα, στόχο έχει να προβλέψει το ETA των εμπορευματοκιβωτίων που μετακινούνται μέσω πολλαπλών μεταφορών (φορτηγά, τρένα, πλοία). Κάτι τέτοιο έχει πολύ μεγάλη σημασία, καθώς προσδίδει διαφάνεια στη διαδικασία του εφοδιασμού.

Το συγκεκριμένο πρόβλημα είναι αρκετά πιο σύνθετο από ETA Μέσων Μαζικής Μεταφοράς, για αυτό και εξετάζονται πάνω από ένα εκατομμύριο κινήσεις container, 35.000 δρομολόγια τρένων, 96.000 δρομολόγια φορτηγών και 33 εκατομμύρια μετεωρολογικές παρατηρήσεις σε διάστημα τριών ετών. Για την επεξεργασία των δεδομένων, το πρόβλημα σπάει σε επιμέρους προβλήματα.

Χρησιμοποιήθηκαν οι εξής τεχνικές:

- Linear Regression trees: Για χρόνους οδικής μεταφοράς
- Random Forests – Gradient Boosting: Για χρόνους σιδηροδρομικών μεταφορών
- Ordinal Forests: Για τη σύνδεση του container με το τρένο που θα το παραλάβει

Το Random Forest έδωσε την καλύτερη πρόβλεψη για χρόνους επεξεργασίας σε τερματικούς σταθμούς. Τα χαρακτηριστικά του τρένου και του τερματικού (χωρητικότητα, συχνότητα δρομολογίων κτλ.) αποδείχθηκαν πολύ πιο σημαντικά από τους περιβαλλοντικούς παράγοντες (καιρός, βάρος φορτίου). Το ζήτημα προσεγγίστηκε με διάφορους τρόπους και αποδείχθηκε ότι η πρόβλεψη για το ποιο τρένο θα παραλάβει ένα container αποδείχθηκε πολύ πιο ακριβής από την πρόβλεψη των χρόνων παραμονής. Η χρήση διαφορετικών μοντέλων για κάθε κομμάτι της διαδρομής βοήθησε στο να αποτυπωθούν οι μη γραμμικότητες των δεδομένων και να ερμηνευτεί η συσχέτιση των μεταβλητών.

Η έρευνα καταλήγει στο ότι η εφαρμογή Machine Learning σε ένα τόσο σύνθετο πρόβλημα αποδείχθηκε ιδιαίτερα αξιόπιστη και αποδοτική με τη σωστή χρήση των μοντέλων. Σημαντικό ρόλο αποτέλεσε και ο μεγάλος όγκος ακριβούς δεδομένων που προσφέρθηκε από όλα τα εμπλεκόμενα μέσα. Γενικά η πληροφορία ETA είναι πολύ σημαντική στις πλατφόρμες διατροπικών δικτύων, καθώς βελτιώνει τη διαχείριση πόρων, μειώνει τους κινδύνους και βοηθάει στον χωροταξικό σχεδιασμό, ενισχύοντας την επικοινωνία των μελών της εφοδιαστικής αλυσίδας.

Το ζήτημα της άφιξης εμπορικών τρένων αναλύει επίσης ο M.G. Vaessen στην έρευνά του με τίτλο “Predicting the ETA of cargo trains using AI models”. Εξετάζει την πρόβλεψη του αναμενόμενου χρόνου άφιξης με χρήση Machine Learning (M.G. Vaessen et al., 2021).

Ακολούθησε τη γνωστή πορεία συλλογής δεδομένων από GPS (AVL Data) από αμαξοστοιχία που εκτελεί το δρομολόγιο Ρότερνταμ-Γερμανία σε συνδυασμό με δεδομένα χρονοδιαγραμμάτων (Timetable data). Μετά την επεξεργασία και την κατηγοριοποίηση των δεδομένων, δοκιμάστηκαν διάφορα μοντέλα μηχανικής μάθησης:

- Γραμμική παλινδρόμηση (και εκδοχές Lasso, Ridge, Elastic Net)
- Support Vector Regression (SVR)
- Random Forest Regression (RFR)
- Gradient Boosting Regression (GBR)
- Multi-Layer Perceptron (MLP)

Χρησιμοποιήθηκε τριπλή διασταυρούμενη επικύρωση (3-fold CV) και ρύθμιση υπερπαραμέτρων μέσω Grid Search.

Με την εφαρμογή των παραπάνω προέκυψε ότι το baseline μοντέλο (χρονοδιάγραμμα) είχε πολύ υψηλά σφάλματα. Οι γραμμικές μέθοδοι βελτίωσαν την απόδοση, αλλά παρέμειναν περιορισμένες. Το MLP είχε χαμηλότερο RMSE, αλλά παρουσιάζει προβληματική συμπεριφορά στην παραγωγική χρήση. Το GBR έδωσε το χαμηλότερο σφάλμα με ικανοποιητική συμπεριφορά και κρίθηκε το πιο κατάλληλο και το SVR με το Random Forest είχαν τη χαμηλότερη απόδοση.

Σύμφωνα με τα ευρήματα, προκύπτει το συμπέρασμα ότι τα Gradient Boosting μοντέλα υπερέχουν έναντι άλλων στο συγκεκριμένο ζητούμενο προσφέροντας σημαντικά βελτιωμένες εκτιμήσεις και δυνατότητα εφαρμογής σε «έξυπνα» συστήματα logistics.

Η της Koutsadourou (2024) εξέτασε την πρόβλεψη στάσης αποβίβασης επιβατών σε αστικές συγκοινωνίες. Το ζήτημα προσεγγίζεται και σε αυτή την περίπτωση με Machine Learning και αποσκοπεί στη βελτίωση της εμπειρίας των επιβατών και στην κατάλληλη οργάνωση των δρομολογίων. Για την ανάλυση αυτή αξιοποιούνται δεδομένα από το σύστημα επικυρωμένων εισιτηρίων της Ρίγας.

Τα δεδομένα που χρησιμοποιήθηκαν είναι οι επικυρώσεις εισιτηρίων από δύο ημερομηνίες (7/9/2021 και 11/9/2021). Έγινε κατάλληλη επεξεργασία με κατηγοριοποίηση και κωδικοποίηση μεταβλητών για την εισαγωγή τους στα μοντέλα.

Αφού έγινε ο διαχωρισμός σε train/test sets, εφαρμόστηκαν τα εξής μοντέλα:

- Decision Tree: Απλή μέθοδος και με περιορισμένη ακρίβεια σε ευαισθησία και overfitting
- Random Forest: Βελτίωσε την ακρίβεια και υπερέχει έναντι του Bagging λόγω της μείωσης διακύμανσης

- Bagging: Βελτίωσε την ακρίβεια
- Gradient Boosting: Έδωσε κάποια από τα καλύτερα αποτελέσματα
- Kernel Ridge Regression: Αποδείχθηκε λιγότερο αποτελεσματικό σε σύγκριση με τα δέντρα
- XGBoost: Έδωσε επίσης κάποια από τα καλύτερα αποτελέσματα
- LightGBM: Από τα καλύτερα αποτελέσματα και επιπλέον επιτυγχάνει υψηλή ταχύτητα με μεγάλη ακρίβεια
- Multilayer Perceptron (Νευρωνικό Δίκτυο): Έχει ανταγωνιστική απόδοση, αλλά απαιτεί πού χρόνο και δεδομένα για βέλτιστη απόδοση

Τέλος η αξιολόγηση των αποτελεσμάτων πραγματοποιήθηκε με κατάλληλες μετρικές αξιολόγησης όπως accuracy, mean absolute error (MAE), mean squared error (MSE).

Το συμπέρασμα που προκύπτει και σε αυτή την περίπτωση είναι ότι δεν υπάρχει μία καθαρά βέλτιστη μέθοδος, αλλά η επιλογή εξαρτάται από τη φύση των δεδομένων και τις ανάγκες της κάθε μελέτης. Για τη συγκεκριμένη έρευνα, οι αλγόριθμοι boosting ήταν πιο ακριβείς και γρήγοροι. Επιπλέον η σωστή επεξεργασία των δεδομένων είναι εξίσου σημαντική με την επιλογή μοντέλου όταν ο στόχος είναι μια όσο το δυνατόν ακριβή πρόβλεψη. Τέλος αναδεικνύεται η [πρακτική σημασία τέτοιων αυτοματοποιημένων συστημάτων στα σύγχρονα αστικά κέντρα για τη ενίσχυση της βιώσιμης κινητικότητας.

Μία πιο επιστημονική και ακριβή προσέγγιση αποτελεί η εργασία των Chondrodima, et al. (2022), η οποία προτείνει ένα ολοκληρωμένο πλαίσιο χρόνου άφιξης δημόσιων συγκοινωνιών βασισμένο σε Νευρωνικά Δίκτυα. Πιο συγκεκριμένα, γίνεται χρήση του Radial Basis Function Neural Networks (RBF NNs), το οποίο εκπαιδεύεται με τη χρήση του αλγορίθμου Particle Swarm Optimization (PSO). Τα δεδομένα που αξιοποιούνται προέρχονται από το General Transit Feed Specification (GTFS). Η μελέτη εισάγει μία νέα προσέγγιση, βασισμένη σε δεδομένα, καθώς και ένα pipeline προεπεξεργασίας (CR-GTFS) για τον καθαρισμό και την ανασυγκρότηση δεδομένων GTFS, το οποίο υπερτερεί έναντι προηγούμενων χρόνων όσο και σε χρόνους υπολογισμού. Γενικά, η εργασία αυτή αποτελεί ένα σημαντικό εύρημα προς την ενίσχυση των «έξυπνων πόλεων» και προτείνεται η εφαρμογή σε περισσότερες πόλεις με την ενσωμάτωση πρόσθετων χαρακτηριστικών.

Άλλη μία έρευνα στα Νευρωνικά Δίκτυα αποτελεί αυτή των Derrow-Pinion, et al. (2021), οι οποίοι παρουσίασαν μία σημαντική πρόοδο στην πρόβλεψη χρόνου διαδρομής μέσω της εφαρμογής GNNs στο Google maps. Στην έρευνα αυτή, η εκτίμηση ETA βασίζεται αποκλειστικά σε GNN και το μοντέλο αναλύει τα τοπολογικά χαρακτηριστικά του οδικού δικτύου και προβλέπει τις μελλοντικές συνθήκες κυκλοφορίας, ξεπερνώντας τις προηγούμενες μεθόδους. Στο Σίδνεϊ κατάφερε να μειώσει τις αρνητικές εκτιμήσεις ETA κατά 40%. Η αρχιτεκτονική GNN, αν και αξιοποιεί τυπικά δομικά στοιχεία, ενσωματώνει μεθόδους προγραμματισμού εκπαίδευσης όπως τα MetaGradients, καθιστώντας το μοντέλο ανθεκτικό και κατάλληλο για εφαρμογή στον πραγματικό κόσμο. Το άρθρο αναλύει τις προκλήσεις που υπάρχουν στην αναπαράσταση των οδικών δικτύων για εφαρμογές μηχανικής μάθησης και τις λύσεις που υιοθετήθηκαν, συμπεριλαμβανομένης της χρήσης

τόσο δεδομένων πραγματικού χρόνου όσο και ιστορικών δεδομένων για την τροφοδότηση των προβλέψεων του GNN. Η εφαρμογή του συγκεκριμένου μοντέλου GNN στο Google Maps υπογραμμίζει τη χρησιμότητά του στον πραγματικό κόσμο, προσφέροντας βελτιωμένες εμπειρίες πλοήγησης στους χρήστες μέσω πιο ακριβών και αξιόπιστων εκτιμήσεων χρόνου διαδρομής.

Η έρευνα για το Google Maps συνεχίζεται και με τους Mehta et al. (2019) οι οποίοι αναβάθμισαν τις δυνατότητες πλοήγησης της εταιρίας, προσθέτοντας καινούριες λειτουργίες όπως το Street View και η εκτίμηση χρόνου άφιξης (ETA). Επιπλέον προστέθηκε η δυνατότητα εύρεσης της συντομότερης διαδρομής, μέσω της εφαρμογής των αλγορίθμων. Η μέθοδος αυτή εξελίχθηκε από τον Zafar et al., ο οποίος παρουσίασε ένα υβριδικό μοντέλο Deep Learning βασισμένο σε GRU-LSTM για την πρόβλεψη κυκλοφοριακής συμφόρησης σε «έξυπνες πόλεις» αξιοποιώντας δεδομένα από διάφορες πηγές. Το συγκεκριμένο μοντέλο έφτασε 95% ακρίβεια, γεγονός πολύ ικανοποιητικό.

Οι Chen et al., (2004), Ramakrishna et al. (2006) Jeong, 2004· Chien et al. (2002) Park et al. (2004) συνέβαλαν στην έρευνα των Τεχνητών Νευρωνικών Δικτύων (ANNs), καθώς έχουν την ικανότητα να επιλύουν πολύπλοκες μη γραμμικές σχέσεις. Τα ANN αποτελούν μία πολύ αποδοτική μέθοδο, καθώς δεν είναι απαραίτητος ο καθορισμός της μορφής της συνάρτησης, ενώ οι περιορισμοί που σχετίζονται με την πολυσυγγραμμικότητα των εξηγηματικών μεταβλητών μπορούν να παραλειφθούν. Ο Chien ανέπτυξε ένα βελτιωμένο μοντέλο Τεχνητού Νευρωνικού Δικτύου για την πρόβλεψη δυναμικού χρόνου άφιξης λεωφορείων χρησιμοποιώντας τον αλγόριθμο Back-Propagation (Chien et al., 2002).

2.3. Βασικά Συμπεράσματα

Γενικότερα, παρατηρείται ότι τα τελευταία χρόνια η πρόβλεψη το εκτιμώμενου χρόνου άφιξης (ETA) βασίζεται όλο και λιγότερο σε απλά στατιστικά μοντέλα και εστιάζει σε καινούριες μεθόδους που μπορούν να συγχρονιστούν με το πλήθος και την πολυπλοκότητα των δεδομένων που παρουσιάζουν τα μεγάλα αστικά κέντρα. Το Machine Learning και το Deep Learning, έχουν διαδοθεί σημαντικά και εφαρμόζονται πλέον σε ένα μεγάλο κομμάτι προβλημάτων μεταφορών και κυκλοφοριακής τεχνικής. Μέθοδοι όπως τα φίλτρα Kalman, τα Support Vector Machines, τα Τεχνητά Νευρωνικά Δίκτυα (ANNs), τα μοντέλα Boosting και τα Γραφικά Νευρωνικά Δίκτυα (GNNs) αποδεικνύονται ιδιαίτερα αποτελεσματικά στις εφαρμογές που εξετάστηκαν, με τα υβριδικά μοντέλα να αποτελούν την καλύτερη και πιο ευέλικτη λύση. Η ανάγκη εξέλιξης πάνω στο συγκεκριμένο ζήτημα είναι ακόμα αισθητή λόγω της ετερογένειας των δεδομένων και της δυσκολίας συσχέτισής τους σε ποικιλόμορφα datasets, για αυτό και σε κάθε πρόβλημα, η ποιότητα και η ποσότητα των δεδομένων παίζει καθοριστικό ρόλο στην επιλογή μοντέλου.

Τέλος, παρά το γεγονός ότι παρατηρείται μεγάλη έρευνα στη διεθνή βιβλιογραφία, στην Ελλάδα και ιδιαίτερα στην Αθήνα, η έρευνα είναι ελλιπής. Εδώ και πολλά χρόνια είναι σε λειτουργία η υπηρεσιακή τηλεματική με ενημέρωση σε πραγματικό χρόνο, ενώ δεν έχει

πραγματοποιηθεί έρευνα για εφαρμογή Μηχανικής και Βαθιάς Μάθησης με ανάλυση χαρακτηριστικών και παρουσία μετρικών ακρίβειας. Λόγω του σύνθετου αστικού περιβάλλοντος, η Αθήνα θα μπορούσε να αποτελέσει σημαντικό πεδίο έρευνας με ενδιαφέροντα αποτελέσματα.

3.Μεθοδολογική Προσέγγιση

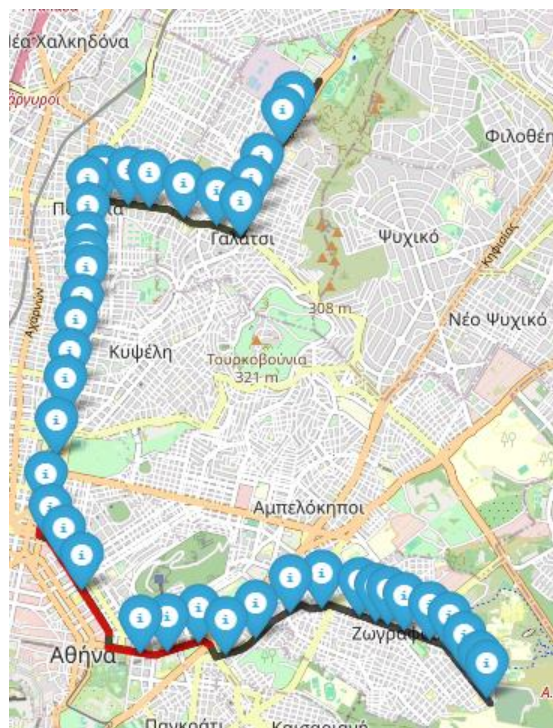
3.1. Ροή Εργασιών

Η παρούσα μεθοδολογία αποσκοπεί στην πρόβλεψη του χρόνου άφιξης λεωφορείου σε επόμενες στάσεις, χρησιμοποιώντας πραγματικά δεδομένα από ιστορικά αρχεία AVL/GPS και αναλυτικά χαρακτηριστικά διαδρομών. Η προσέγγιση βασίζεται σε εποπτευόμενη μάθηση (supervised learning) και εφαρμόζει αλγόριθμους Random Forest Regression, Linear Regression και XGBoost ως μεθόδους πρόβλεψης.

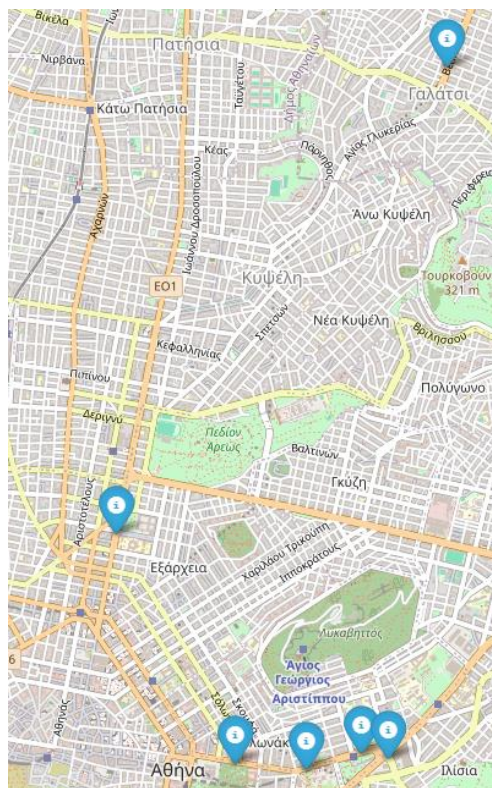
Με τη χρήση των παραπάνω αναπτύσσονται μοντέλα που εκπαιδεύονται σε δεδομένα δρομολογίων από τον ΟΑΣΑ, καθώς και επιπλέον δεδομένων που αφορούν καιρικές συνθήκες, γεωγραφικά στοιχεία, επικυρώσεις εισιτηρίων και αργίες και σαββατοκύριακα.

Για αυτό το σκοπό, αντλήθηκαν όσο το δυνατόν περισσότερα δεδομένα έτσι ώστε το μοντέλο να μπορέσει να συνυπολογίσει τους εξωτερικούς παράγοντες που επηρεάζουν την πορεία του λεωφορείου. Αρχικά έγινε ανάλυση των γεωγραφικών δεδομένων με εντοπισμό των συντεταγμένων των στάσεων και απεικόνιση στο πρόγραμμα GIS μέσω της Python (**Εικόνα 1**). Με χρήση των συντεταγμένων πραγματοποιήθηκε προσεγγιστικός υπολογισμός των αποστάσεων των στάσεων. Στη συνέχεια αντλήθηκαν ιστορικά δεδομένα από τις καταγραφές του ΟΑΣΑ για χρόνους άφιξης και διαδρομής. Σημαντική προσθήκη αποτέλεσαν οι επικυρώσεις εισιτηρίων από το gov.gr και καιρικές συνθήκες από meteostat. Επιπλέον δεδομένα παρείχαν και οι φωρατές κυκλοφορίας που είναι τοποθετημένοι σε διάφορα σημεία της πόλης. Εντοπίστηκαν εκείνοι που αφορούν τη συγκεκριμένη λεωφορειακή γραμμή με τη σωστή κατεύθυνση και απεικονίστηκαν στο χάρτη (**Εικόνα 2**). Παρόλα αυτά, προέκυψε ότι αρκετοί από αυτούς δε λειτουργούσαν και επομένως δεν ήταν εφικτό να αντιστοιχηθεί ο κατάλληλος κυκλοφοριακός φόρτος σε κάθε στάση.

Αφού έγινε η συλλογή δεδομένων και η επεξεργασία, προέκυψαν 326.430 δείγματα για εκπαίδευση των μοντέλων, που αναφέρονται στο διάστημα από Νοέμβριο του 2024 έως Απρίλιο του 2025.

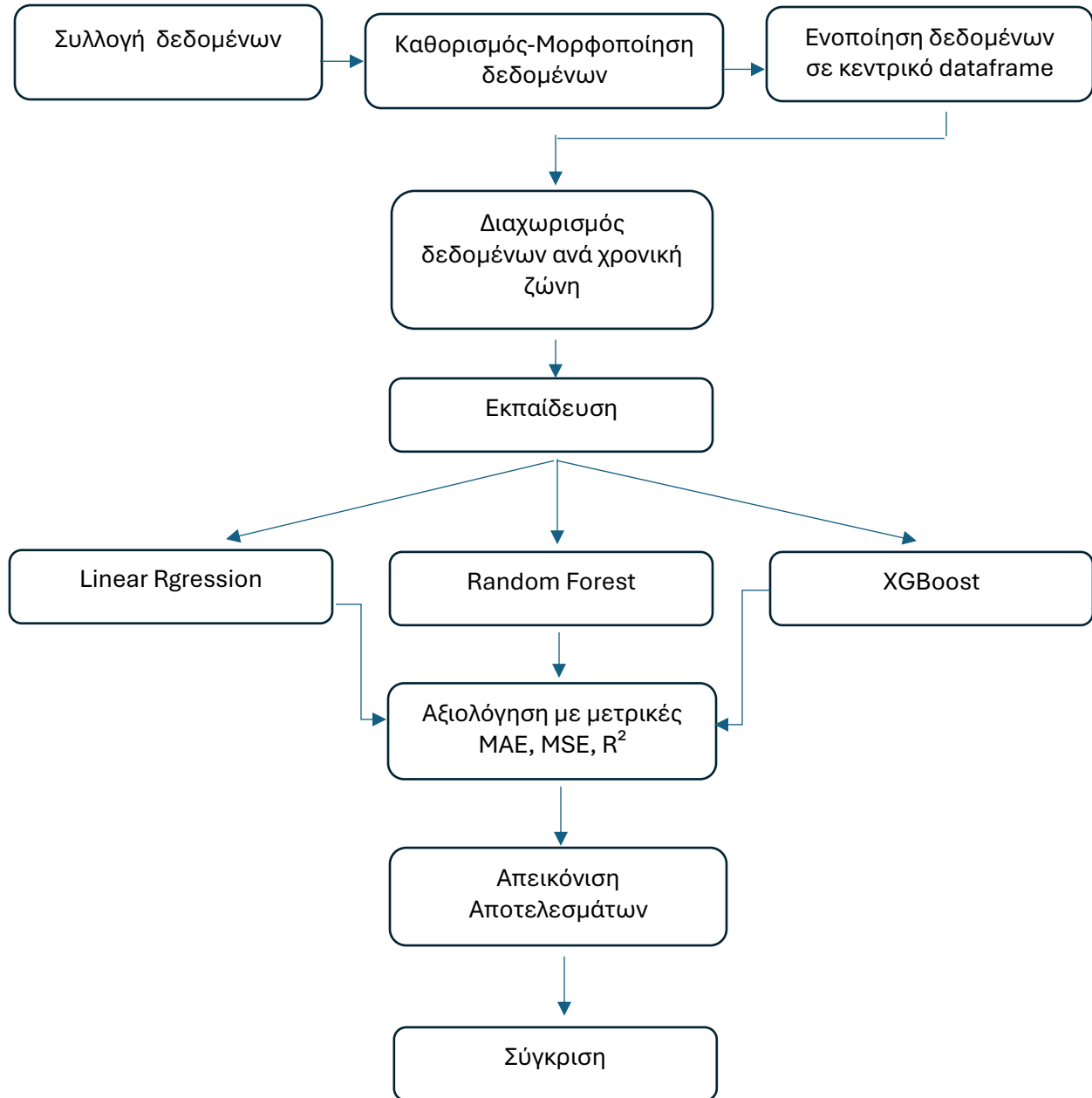


Εικόνα 1: Απεικόνιση στάσεων της διαδρομής του 608



Εικόνα 2: Απεικόνιση φωρατών που αντιστοιχούν στη διαδρομή του 608

Στην **Εικόνα 3** αποτυπώνεται η ροή των εργασιών που ακολουθήθηκε έτσι ώστε να δημιουργηθούν τρία μοντέλα πρόβλεψης χρόνου άφιξης λεωφορείου σε στάση.



Εικόνα 3: Διάγραμμα Ροής Εργασιών

3.2.Θεωρητικό Υπόβαθρο

3.2.1.Γραμμική Παλινδρόμηση

Η γραμμική παλινδρόμηση (Linear Regression) είναι μία μέθοδος supervised learning που χρησιμοποιείται για την πρόβλεψη συνεχούς αριθμητικής τιμής που βασίζεται στην υπόθεση ότι υπάρχει γραμμική σχέση μεταξύ της ανεξάρτητης και της εξαρτημένης μεταβλητής. Στόχος της είναι να κάνει μία, όσο το δυνατόν καλύτερη πρόβλεψη και να ελαχιστοποιήσει το σφάλμα. Αποτελεί μία από τις πιο ευρέως χρησιμοποιούμενες τεχνικές και παρά την απλή της μορφή αποτελεί μια ισχυρή μέθοδο πρόβλεψης.

Η βασική μορφή της εξίσωσης είναι :

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n + \epsilon \quad (1)$$

y: η μεταβλητή στόχος (π.χ. τιμή σπιτιού)

x_1, x_2, \dots, x_n : τα χαρακτηριστικά (features)

β_0 : το intercept (σταθερός όρος)

β_i : οι συντελεστές (weights) που προσδιορίζουν πόσο επηρεάζει κάθε μεταβλητή την έξοδο

ϵ : το σφάλμα (error)

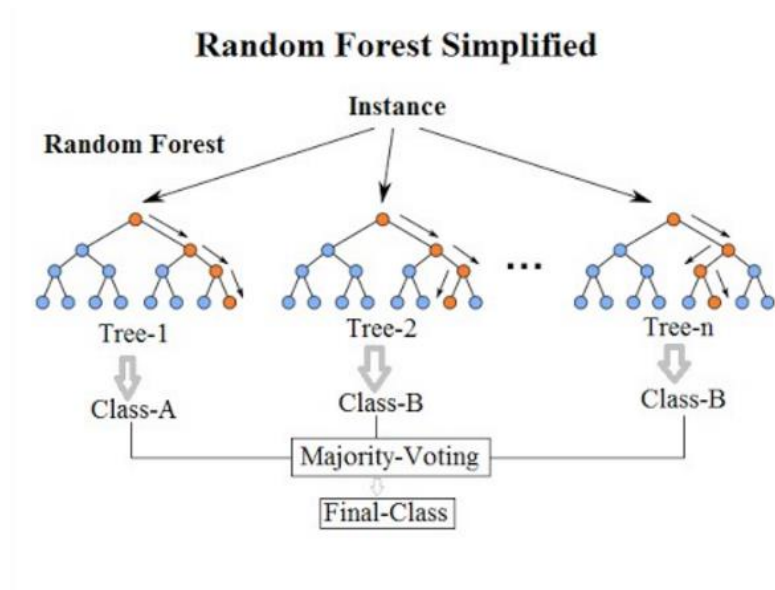
Ο στόχος του μοντέλου είναι να βρει τις τιμές των β_i που ελαχιστοποιούν το συνολικό σφάλμα μεταξύ των προβλεπόμενων και των πραγματικών τιμών. Ο λόγος που αποτελεί ακόμα ένα αρκετά διαδεδομένο μοντέλο, είναι επειδή είναι απλό και γρήγορο. Σε μη σύνθετα προβλήματα της καθημερινότητας είναι μία εύκολα ερμηνεύσιμη λύση με αποτελεσματικότητα. Ωστόσο, υποθέτει γραμμική σχέση, γεγονός που ισχύει σπάνια στην πράξη, καθώς επικρατεί πολυσυσχετισμός μεταξύ των χαρακτηριστικών. Επιπλέον επηρεάζεται από ακραίες τιμές (outliers) και προκαλούνται ανακρίβειες στα αποτελέσματα.

3.2.2.Τυχαία Δάση

Το **Random Forest Regression** είναι μια τεχνική μηχανικής μάθησης που χρησιμοποιείται για προβλήματα παλινδρόμησης, δηλαδή για την πρόβλεψη μιας συνεχούς αριθμητικής τιμής. Βασίζεται στο συνδυασμό πολλών μοντέλων, συγκεκριμένα πολλών δέντρων απόφασης (decision trees) για να μειωθεί ο κίνδυνος υπερπροσαρμογής. Δημιουργεί ένα «δάσος», δηλαδή ένα σύνολο από πολλά δέντρα, όπου κάθε ένα εκπαιδεύεται με ελαφρώς διαφορετικά δεδομένα (**Σχήμα 1**). Το τελικό αποτέλεσμα είναι ο μέσος όρος όλων των προβλέψεων.

Τα βασικά πλεονεκτήματα είναι τα εξής: i. Ανθεκτικότητα στην υπερεκπαίδευση (overfitting), δηλαδή μπορεί να γενικεύσει σε νέα και άγνωστα δεδομένα, ii. καλή ακρίβεια

σε πολλά προβλήματα παλινδρόμησης, iii. καλή ακρίβεια σε πολλά προβλήματα παλινδρόμησης και iv. δε χρειάζεται προεπεξεργασία των δεδομένων (feature scaling).



Σχήμα 1: Απλοποιημένη απεικόνιση μοντέλου Τυχαία Δάση

3.2.3. Ενισχυμένη Κλιμακωτή Ενίσχυση

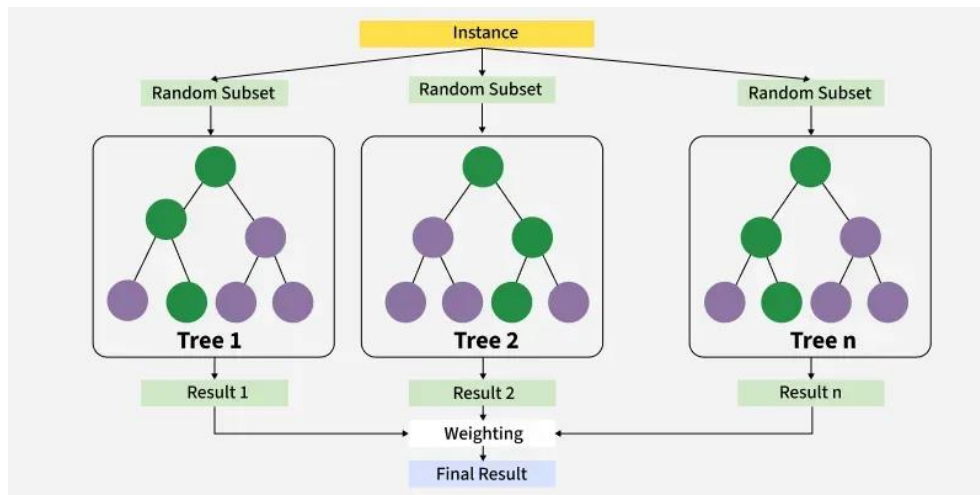
Το **XGBoost** (Extreme Gradient Boosting) είναι μια δημοφιλής και εξαιρετικά αποδοτική βιβλιοθήκη μηχανικής μάθησης, η οποία βασίζεται στη μέθοδο Gradient Boosted Decision Trees (GBDT). Χρησιμοποιείται κυρίως για προβλήματα παλινδρόμησης, ταξινόμησης και κατάταξης, και είναι γνωστό για την υψηλή του ακρίβεια και την υπολογιστική του ταχύτητα. Στην ουσία αποτελεί μία βελτιστοποιημένη υλοποίηση της μεθόδου gradient boosting και ανήκει στις μεθόδους συνένωσης (ensemble learning). Χρησιμοποιεί τα decision trees ως «βασικούς μαθητές» (base learners) και τα προσθέτει ακολουθιακά, ώστε το νέο δέντρο να διορθώνει τα σφάλματα των προηγούμενων (Σχήμα 2).

Μπορεί να διαχειριστεί πολύπλοκα προβλήματα με μεγάλο όγκο δεδομένων, αλλά και να υποστηρίξει παράλληλες επεξεργασίες, κάνοντας τις αναλύσεις ταχύτερες. Μέσω των υπερπαραμέτρων, παρέχει ευελιξία ρύθμισης για βελτιστοποίηση του μοντέλου και πραγματοποιεί μετρήσεις σημαντικότητας χαρακτηριστικών (Feature importance) για ερμηνεία.

Το XGBoost είναι μια βελτιστοποιημένη και πιο γρήγορη υλοποίηση του GBDT. Χτίζει δέντρα σειριακά, όχι παράλληλα. Αντί να κατασκευάζει πολλά ανεξάρτητα δέντρα και να παίρνει το μέσο όρο, τα μοντέλα εκπαιδεύονται διαδοχικά. Κάθε νέο δέντρο προσπαθεί να

διορθώσει τα λάθη του προηγούμενου και η πρόβλεψη γίνεται ως άθροισμα όλων των δέντρων. Ακόμα παρουσιάζει τα παρακάτω χαρακτηριστικά:

- Χρησιμοποιεί level-wise στρατηγική ανάπτυξης δέντρων.
- Βελτιώνει την ακρίβεια με έξυπνο έλεγχο διαχωρισμών.
- Είναι επεκτάσιμο και καταναμημένο, κατάλληλο για μεγάλες βάσεις δεδομένων.
- Ενσωματώνει regularization (ποινές πολυπλοκότητας) για την αποφυγή της υπερεκπαίδευσης (overfitting).



Σχήμα 2: Απλοποιημένη απεικόνιση μοντέλου Ενισχυμένης Κλιμακωτής Ενίσχυσης

3.3. Εκπαίδευση μοντέλων

Η εκπαίδευση πραγματοποιείται σε δύο διαφορετικές περιπτώσεις. Στην πρώτη, τα μοντέλα εκπαιδεύονται στα αρχικά δεδομένα του ΟΑΣΑ, χωρίς προσθήκη των επιπλέον χαρακτηριστικών. Στην επόμενη, η εκπαίδευση γίνεται με την προσθήκη των προαναφερθέντων εξωτερικών χαρακτηριστικών. Αυτό γίνεται για να αξιολογηθεί η απόδοση της ενίσχυσης αυτής και να προκύψουν συμπεράσματα για τη σταθερότητα των μοντέλων. Γίνεται η κατάλληλη παραμετροποίηση σε κάθε μοντέλο και έπειτα εκτελείται η εκπαίδευση για βραδινές και πρωινές ώρες.

3.4. Αξιολόγηση μοντέλων

Έπειτα, τα μοντέλα αξιολογούνται με κάποιες κατάλληλες μετρικές όπως είναι το R^2 , MAE και MSE.

Ο συντελεστής προσδιορισμού R^2 (**Goodness of Fit**) είναι ένα μέτρο που χρησιμοποιείται κυρίως στην παλινδρόμηση για να δείξει πόσο καλά το μοντέλο εξηγεί τη μεταβλητότητα των δεδομένων. Πιο συγκεκριμένα, μετράει το ποσοστό διακύμανσης της εξαρτημένης μεταβλητής που εξηγείται από το μοντέλο.

$$R^2 = 1 - \frac{\sum_{i=1}^m (X_i - Y_i)^2}{\sum_{i=1}^m (\bar{Y} - Y_i)^2} \quad (2)$$

- $R^2 = 1$ → τέλεια πρόβλεψη, το μοντέλο εξηγεί το 100% της διακύμανσης
- $R^2 = 0$ → δεν εξηγεί καθόλου τη διακύμανση
- $R^2 < 0$ → θα ήτα προτιμότερο να υπολογίσουμε απλώς το μέσο όρο

Το Μέσο Απόλυτο Σφάλμα **MAE** (Mean Absolute Error) μετράει το μέσο όρο της απόλυτης διαφοράς ανάμεσα στις πραγματικές τιμές και τις προβλεπόμενες και χρησιμοποιείται για να μετρήσει την ακρίβεια ενός μοντέλου. Ορίζεται ως ο μέσος όρος των απόλυτων διαφορών ανάμεσα στις πραγματικές τιμές και στις προβλέψεις.

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (3)$$

Το MAE παίρνει πάντα θετικές τιμές και όσο μικρότερη τιμή έχει, τόσο καλύτερη είναι η πρόβλεψη, καθώς εκφράζει μικρή απόκλιση από την πραγματική τιμή.

Το Μέσο Τετραγωνικό Σφάλμα **MSE** (Mean Squared Error) αποτελεί ένα μέτρο σφάλματος που χρησιμοποιείται στην παλινδρόμηση για να υπολογίσει πόσο μακριά βρίσκονται οι προβλέψεις από τις πραγματικές τιμές, εστιάζοντας στα μεγάλα σφάλματα. Ορίζεται ως ο μέσος όρος των τετραγώνων των διαφορών μεταξύ πραγματικών τιμών και προβλέψεων.

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (4)$$

Το MSE έχει πάντα θετικές τιμές, καθώς εκφράζει το σφάλμα. Επιπλέον όσο μικρότερη τιμή έχει, τόσο καλύτερη απόδοση έχει το μοντέλο, ενώ δίνει μεγαλύτερη ποινή στα μεγάλα σφάλματα, λόγω της ύψωσης στο τετράγωνο.

3.7.Σημασία Χαρακτηριστικών

Σημαντικό βήμα είναι η εντολή permutation importance που χρησιμοποιείται για να μετρήσει τη σχετική σημασία κάθε χαρακτηριστικού (feature) σε κάθε μοντέλο Random Forest Regression και XGBoost ξεχωριστά. Πιο συγκεκριμένα, για κάθε χαρακτηριστικό ανακατεύει τυχαία τις τιμές κάθε στήλης στο test set και υπολογίζει για κάθε παραλλαγή την πτώση της απόδοσης του μοντέλου, μέσω των δεικτών αξιολόγησης. Όσο μεγαλύτερη είναι η πτώση της απόδοσης, τόσο πιο σημαντικό είναι το εκάστοτε feature για το μοντέλο.

Έπειτα, πραγματοποιείται η δημιουργία γραφήματος, που με φθίνουσα σειρά απεικονίζει τη σημαντικότητα της κάθε μεταβλητής βάσει τα αποτελέσματα της παραπάνω ανάλυσης.

4.Εφαρμογή Μεθοδολογίας και Αποτελέσματα

4.1. Συλλογή και επεξεργασία δεδομένων

Στην παρούσα διπλωματική γίνεται ανάλυση της λεωφορειακής γραμμής 608 “Γαλάτσι-Ακαδημία-Νεκρ. Ζωγράφου” με κατεύθυνση προς το Νεκροταφείου Ζωγράφου και αφορά την επεξεργασία τηλεματικών δεδομένων, την εξαγωγή και την ανάλυση χρήσιμων χαρακτηριστικών, καθώς και τη δημιουργία μοντέλου πρόβλεψης του χρόνου άφιξης, με χρήση μοντέλων μηχανικής μάθησης. Η συγκεκριμένη γραμμή επιλέχθηκε ως ενδεικτική, καθώς προσομοιώνει τις συνθήκες στον αστικό κλοιό.

Αρχικά, γίνεται εισαγωγή των δεδομένων τηλεματικής του ΟΑΣΑ, που αφορούν τις ακριβείς ώρες άφιξης των λεωφορείων της γραμμής σε κάθε στάση, οργανωμένα σε φακέλους ανά ημερομηνία. Τα δεδομένα αυτά περιλαμβάνουν καταγραφές γεωτοπισμού (GPS), γεγονότα αφίξεων και αναχωρήσεων σε στάσεις ή τέρματα και μοναδικά IDs για το όχημα και το δρομολόγιο. Με χρήση της βιβλιοθήκης pandas, όλα αυτά τα αρχεία ενσωματώθηκαν σε ένα ενιαίο Data Frame, δηλαδή πίνακα δεδομένων. Έπειτα, διατηρούνται μόνο οι εγγραφές που σχετίζονται με τα γεγονότα «Άφιξη σε στάση» ή «Αναχώρηση από τέρμα» ώστε να χρησιμοποιηθούν μόνο τα λειτουργικά σημεία της διαδρομής.

Παράλληλα, φιλτράρονται αυτά τα δεδομένα έτσι ώστε να παραμείνει η πιο συχνή αλληλουχία του λεωφορείου, δηλαδή η χρήση κάθε στάσης χωρίς κάποια παράλειψη.

Τελικώς, εντοπίστηκαν χειροκίνητα πάνω στο χάρτη οι συντεταγμένες του γεωγραφικού πλάτους και μήκους των στάσεων, οι οποίες και μετατράπηκαν σε αριθμητική μορφή και δημιουργήθηκε ένα μοναδικό αναγνωριστικό ανά όχημα και δρομολόγιο.

4.1.1.Καθορισμός και μορφοποίηση δεδομένων

Στη συνέχεια, πραγματοποιείται κυκλική κωδικοποίηση του χρόνου με αναγωγή σε ημίτονο και συνημίτονο για τους εξής λόγους :

Όταν μοντελοποιούνται χρονικά δεδομένα, όπως στιγμές άφιξης ή αναχώρησης λεωφορείων, η μεταβλητή «ώρα της ημέρας» έχει κυκλική φύση: μετά τις 23:59 ακολουθεί ξανά η 00:00. Εάν την κωδικοποιήσουμε γραμμικά (π.χ. 0–24), το μοντέλο δεν αντιλαμβάνεται ότι οι τιμές πλησιέστερες σε 0 και 24 είναι στην πραγματικότητα κοντινές. Αυτό μπορεί να οδηγήσει σε ψευδή συμπεράσματα και ανακριβείς προβλέψεις.

Για να ξεπεραστεί αυτό το πρόβλημα, εφαρμόζεται κυκλική κωδικοποίηση. Πιο συγκεκριμένα, αντιστοιχίζεται κάθε ώρα σε δύο νέες μεταβλητές, οι οποίες προσομοιώνουν τη μονάδα κύκλο (unit circle) στο επίπεδο. Με αυτόν τον τρόπο η απόσταση μεταξύ 23:00 και 01:00 γίνεται μικρή (στον κύκλο), κι όχι μεγάλη (στην ευθεία). Συνεπώς, το μοντέλο που κατασκευάζεται μπορεί να επεξεργαστεί καλύτερα τα δεδομένα του χρόνου. Η ίδια διαδικασία ακολουθείται και στις μέρες της εβδομάδας.

Άλλο ένα δεδομένο που χρησιμοποιείται είναι η ύπαρξη της λεωφορειολωρίδας στο συγκεκριμένο δρομολόγιο. Υπολογίζεται και η απόσταση που διανύει το λεωφορείο μέσα στη λωρίδα για κάθε στάση και η κατηγορική μεταβλητή ύπαρξης ή μη της προαναφερθείσας σε κάθε στάση (ένδειξη 0 και 1).

Επισημαίνεται και ο αριθμός του κάθε οχήματος (veh_no), αλλά και ο κωδικός της τρέχουσας και της προηγούμενης στάσης (stop_id), ώστε να γίνουν οι απαραίτητοι υπολογισμοί των χρόνων και των αποστάσεων μεταξύ τους.

Εκτελώντας αυτά τα βήματα, έχει δημιουργηθεί ένας πίνακας δεδομένων (Data Frame) που χρησιμοποιείται ως βάση για την μετέπειτα επεξεργασία και προσθήκη δεδομένων και περιέχει:

- event_datetime: ώρα και ημερομηνία
- veh_no: κωδικός οχήματος
- lon, lat: συντεταγμένες στάσεων
- Event: Άφιξη σε στάση το οποίο και είναι το μόνο που έχουμε κρατήσει μετά από φιλτράρισμα
- stop_id: κωδικός στάσης
- route_code: κατεύθυνση δρομολογίου όπου εξετάζουμε την ίδια σε όλο το script
- time_of_day_sin: ημίτονο ανηγμένης ώρας
- time_of_day_cos: Σινημίτονο ανηγμένης ώρας
- day_of_week_sin: ημίτονο ανηγμένης μέρας
- day_of_week_cos: συνημίτονο ανηγμένης μέρας
- previous_stop_id: κωδικός προηγούμενης στάσης
- altered_hour: ώρα στρογγυλοποιημένη προς τα πίσω

- `distance_between_stops`: απόσταση μεταξύ στάσεων
- `distance_in_bus_lane`: μήκος λεωφορειολωρίδας που διανύει το λεωφορείο από τη μία στάση στην άλλη
- `bus_lane`: ένδειξη ύπαρξης ή μη λεωφορειωρίδας

4.1.2. Διαχωρισμός Δεδομένων ανά Χρονική Ζώνη

Για την αξιόπιστη εκπαίδευση και αξιολόγηση των μοντέλων πρόβλεψης, κρίθηκε σκόπιμος ο προκαταρκτικός χρονικός διαχωρισμός των δεδομένων με βάση την ώρα εκτέλεσης του δρομολογίου. Ο λόγος για τον διαχωρισμό αυτό προκύπτει από την παρατήρηση ότι τα βραδινά δρομολόγια (22:00–07:00) παρουσιάζουν σημαντικά διαφορετικά χαρακτηριστικά και ασυνέπειες σε σχέση με τα ημερήσια δρομολόγια (07:00–22:00).

Πιο συγκεκριμένα, τα βραδινά δρομολόγια παρουσιάζουν μειωμένη συχνότητα, μικρότερο όγκο επιβατών και πιο ασταθή καταγραφή στάσεων. Αυτό οδηγεί σε μικρότερο αριθμό παρατηρήσεων και σε υποεκπροσώπηση των βραδινών τιμών στο συνολικό όγκο δεδομένων (dataset). Η εκπαίδευση των μοντέλων σε ενοποιημένο dataset ημέρας και νύχτας οδηγεί σε υπερβολική προσαρμογή στα ημερήσια μοτίβα και χειρότερη γενίκευση στις σπάνιες βραδινές τιμές.

Για την αντιμετώπιση αυτών των προβλημάτων, εφαρμόστηκε ο εξής χρονικός διαχωρισμός:

Κατηγοριοποίηση:

- Ημερήσια δεδομένα: 07:00 έως 21:59
- Νυχτερινά δεδομένα: 22:00 έως 06:59

Η κατηγοριοποίηση αυτή υλοποιήθηκε με βάση τη στήλη `event_datetime` ή ισοδύναμο χρονικό πεδίο, χρησιμοποιώντας συναρτήσεις εξαγωγής ώρας (`.dt.hour`) και λογικό φιλτράρισμα των παρατηρήσεων.

Η βάση δεδομένων διαιρέθηκε σε δύο υποσύνολα:

- `df_day`: δρομολόγια εντός του χρονικού πλαισίου 07:00–21:59
- `df_night`: δρομολόγια εκτός του παραπάνω (22:00–06:59)

Η προεπεξεργασία και εκπαίδευση των μοντέλων (Random Forest, XGBoost, Linear Regression) εφαρμόστηκε ξεχωριστά σε κάθε υποσύνολο.

Η αξιολόγηση έγινε ανεξάρτητα για κάθε χρονική κατηγορία, επιτρέποντας τη στοχευμένη ερμηνεία της απόδοσης των μοντέλων σε διαφορετικές συνθήκες λειτουργίας (π.χ. χαμηλή κυκλοφορία, μειωμένος αριθμός στάσεων κ.λπ.).

4.1.3. Προσθήκη εξωτερικών δεδομένων

Αφού έχουν επεξεργαστεί αυτά τα δεδομένα από τα telematics του ΟΑΣΑ, διερευνώνται περαιτέρω εξωτερικοί παράγοντες που μπορούν να συμβάλουν στη διαδρομή του λεωφορείου.

Σημαντική πηγή πληροφοριών αποτελεί το gov.gr, το οποίο διαθέτει δεδομένα, τόσο για τις επικυρώσεις των εισιτηρίων ανά δρομολόγιο, όσο και για την κίνηση της Αθήνας.

Για τις επικυρώσεις των εισιτηρίων χρειάστηκε να γίνει επεξεργασία των δεδομένων, έτσι ώστε να αντιστοιχούν σε επικυρώσεις ανά όχημα και ανά ώρα για ακριβέστερη ανάλυση. Αυτό επιτεύχθηκε με υπολογισμό της μέσης τιμής των εισιτηρίων και έπειτα τη διαίρεση του με τον αριθμό των οχημάτων που έχουμε σε διάστημα μίας ώρας.

Για την ενίσχυση της πρόβλεψης εξετάζεται η προσθήκη δεδομένων για τον κυκλοφοριακό φόρτο στους δρόμους της Αθήνας. Συγκεκριμένα, πόσα αμάξια πέρασαν και εντοπίστηκαν από το φωρατή του εκάστοτε δρόμου σε ένα συγκεκριμένο χρονικό διάστημα και με τι μέση ταχύτητα. Για να γίνει σωστή αντιστοίχιση, εντοπίστηκαν χειροκίνητα οι δρόμοι από τους οποίους περνάει η προς μελέτη γραμμή και αντιστοιχήθηκαν με τους κατάλληλους φωρατές. Ανάλογα την ώρα, όλα αυτά τα δεδομένα, επεξεργάζονται και προστίθενται κατάλληλα στην αρχική πληροφορία έτσι ώστε να αποτελέσουν μεταβλητές.

Να σημειωθεί εδώ ότι, στα csv (αρχείο τιμών) που αντλούνται και μετατρέπονται σε πίνακα περιεχομένων (dataframe) υπάρχουν και πληροφορίες, οι οποίες δεν ωφελούν στη δημιουργία του μοντέλου, οπότε με την εντολή filter, γίνεται συνεχώς αφαίρεση των άχρηστων πληροφοριών.

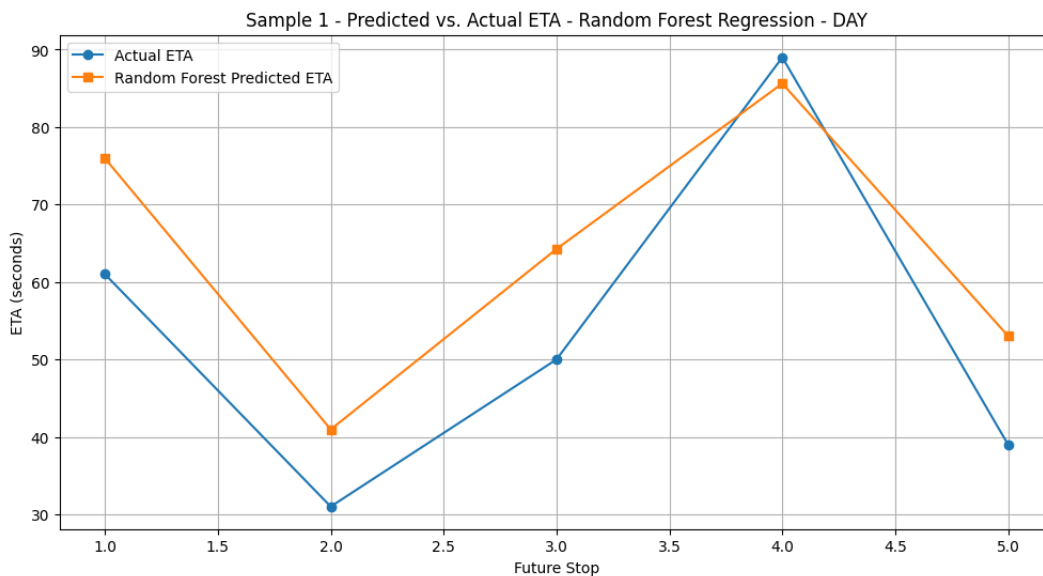
Επόμενο βήμα είναι η προσθήκη του καιρού, η οποία έγινε από την ιστοσελίδα της meteostat για το αεροδρόμιο του Ελευθέριου Βενιζέλου, καθώς ήταν η κοντινότερη τοποθεσία με ακριβείς πληροφορίες για τον καιρό, ανά ώρα. Από αυτά τα csv χρησιμοποιήθηκε η θερμοκρασία και η βροχή. Συγκεκριμένα, στη βροχή έγινε επεξεργασία ώστε να αποτελεί κατηγορική μεταβλητή και να δίνει 0 και 1 ανάλογα με το αν βρέχει το συγκεκριμένο διάστημα μίας ώρας. Στήλες όπως η πίεση του αέρα ή η ύπαρξη χιονιού υπεξαιρέθηκαν, διότι στην περίπτωση της Αθήνας δε θα ασκούσαν κάποια σημαντική επιρροή.

Άλλη μία πληροφορία που αποσκοπεί στην εκπαίδευση του μοντέλου όσο το δυνατόν καλύτερα, είναι η προσθήκη των αργιών και των Σαββατοκύριακων.

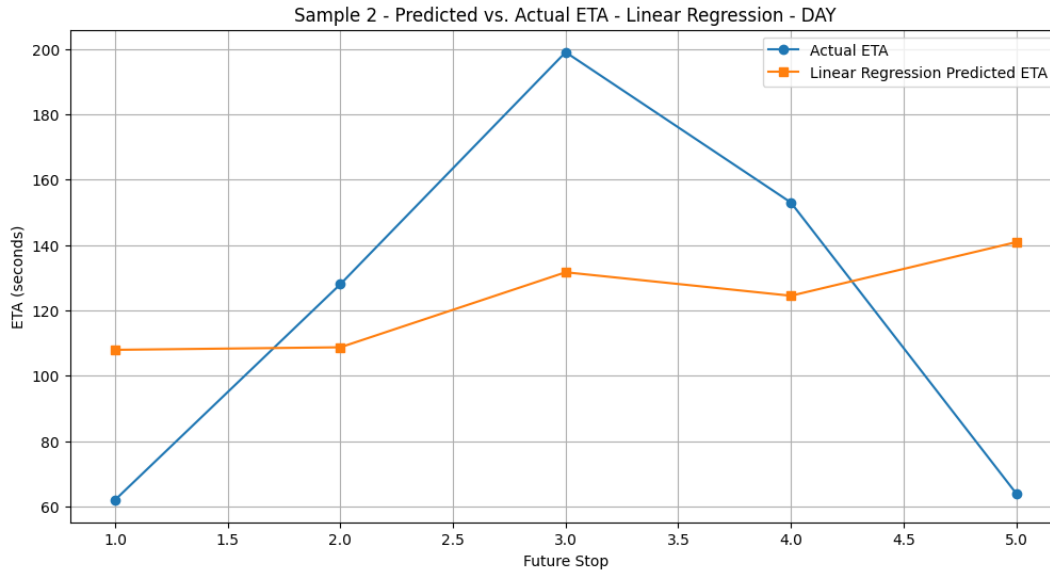
Σε κάθε νέο δεδομένο που προστίθεται, γίνεται στρογγυλοποίηση της ώρας προς την προηγούμενη και μετατρέπεται η ημερομηνία και η ώρα σε event_datetime, ώστε να μπορέσει να προστεθεί στο αρχικό DataFrame που έχει δημιουργηθεί.

4.1.4. Διαγράμματα πραγματικής μεταβολής

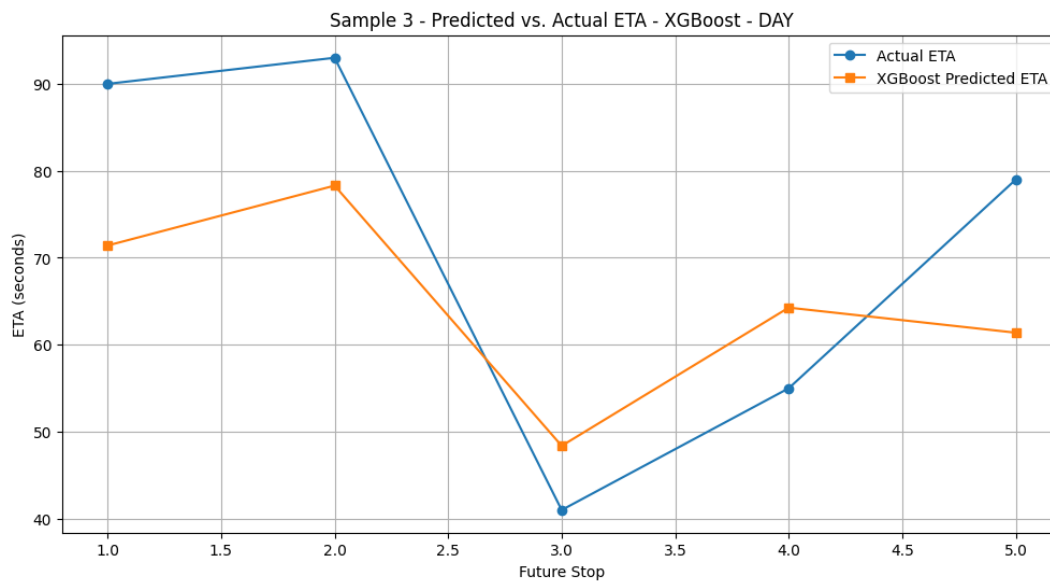
Για να πραγματοποιηθεί καλύτερη σύγκριση και αξιολόγηση των μοντέλων, δημιουργούνται τα διαγράμματα σύγκρισης προβλέψεων και πραγματικών τιμών, που για πέντε τυχαία δρομολόγια απεικονίζεται η πρόβλεψη του κάθε μοντέλου μαζί με την πραγματική τιμή για την άφιξή του σε κάθε μελλοντική στάση. Το διάγραμμα αυτό, δείχνει καθαρά τη σύγκριση και επισημαίνει πόσο σωστά γίνεται η πρόβλεψη. Παρακάτω φαίνονται τα **Διάγραμμα 1**, **Διάγραμμα 2** και **Διάγραμμα 3**, που παρουσιάζουν αυτή τη διαδικασία για το Random Forest, το Linear Regression και το XGBoost αντίστοιχα, κατά τη διάρκεια της ημέρας.



Διάγραμμα 1: Διάγραμμα σύγκρισης πρόβλεψης και πραγματικών τιμών για Τυχαία Δάση - ΗΜΕΡΑ



Διάγραμμα 2: Διάγραμμα σύγκρισης πρόβλεψης και πραγματικών τιμών για Γραμμική Παλινδρόμηση- ΗΜΕΡΑ



Διάγραμμα 3: Διάγραμμα σύγκρισης πρόβλεψης και πραγματικών τιμών για XGBoost - ΗΜΕΡΑ

Σύμφωνα με αυτά τα διαγράμματα προκύπτουν κάποια γενικά συμπεράσματα για την απόδοσή τους στην πρόβλεψη, την ικανότητα τους να ακολουθούν τη γενική τάση, να εντοπίζουν τις μεγάλες διακυμάνσεις και να διατηρούν μικρές αποκλίσεις.

4.2. Ανάπτυξη Μοντέλων και Αποτελέσματα

Σε αυτό το σημείο καθορίζονται τα τρία μοντέλα σε κάθε ομάδα δεδομένων και προκύπτουν οι μετρικές αξιολόγησης R^2 , MAE και MSE. Έπειτα, δημιουργούνται πέντε γραφικές παραστάσεις για το καθένα που απεικονίζουν την πρόβλεψη του μοντέλου σε σύγκριση με τις πραγματικές τιμές και με αυτό τον τρόπο προκύπτουν κάποια δεδομένα για την αποδοτικότητα σε κάθε περίπτωση.

4.2.1. Παραμετροποίηση μοντέλων

Στο πλαίσιο της παρούσας διπλωματικής πραγματοποιήθηκε παραμετροποίηση των μοντέλων Μηχανικής Μάθησης με στόχο τη βελτιστοποίηση της ακρίβειας πρόβλεψης των χρόνων άφιξης της λεωφορειακής γραμμής. Πιο συγκεκριμένα, η παραμετροποίηση αφορά τον καθορισμό των υπερπαραμέτρων κάθε μοντέλου, οι οποίες ελέγχουν τη διαδικασία εκπαίδευσης και τη γενίκευση του μοντέλου στα νέα δεδομένα.

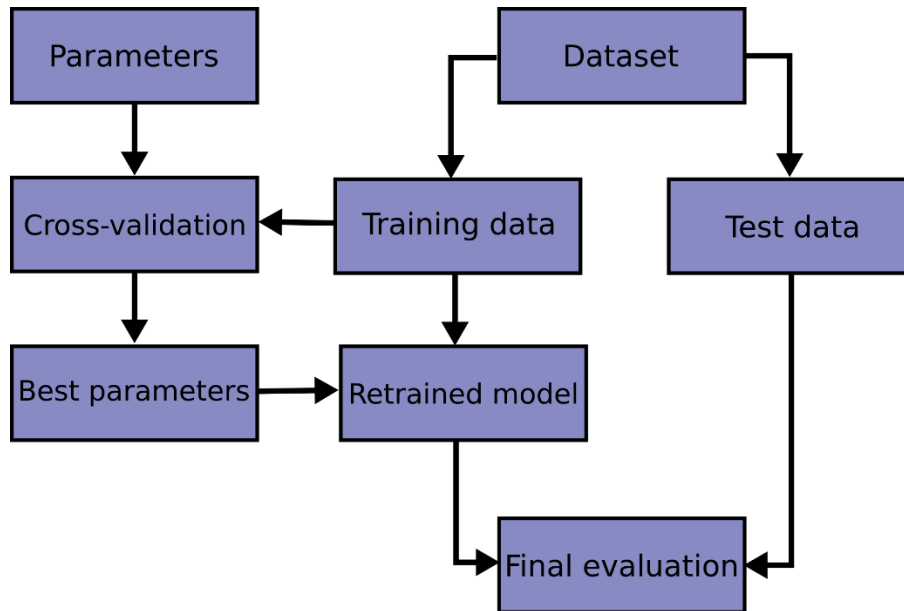
Για το XGBoost χρησιμοποιήθηκαν ως βασικοί υπερπαραμέτροι ο αριθμός των δέντρων ($n_estimators=100$) και η σταθερά τυχαιότητας ($random\ state=42$), η οποία καθορίζεται ώστε να παραχθεί το ίδιο σύνολο ψευδοτυχαίων αριθμών κάθε φορά που τρέχει το πρόγραμμα. Οι υπόλοιπες παράμετροι διατηρήθηκαν στις προεπιλεγμένες τιμές της βιβλιοθήκης xgboost, καθώς η απόδοση κρίθηκε ικανοποιητική.

Στο Random Forest επίσης επιλέχθηκε αντίστοιχη σταθερά τυχαιότητας ($random\ state=42$) και η επιλογή των υπόλοιπων υπερπαραμέτρων βασίστηκε σε εμπειρική αξιολόγηση και δοκιμές, λαμβάνοντας υπόψη την απόδοση και την υπολογιστική αποδοτικότητα.

Στη Linear Regression δεν καθορίστηκαν συγκεκριμένες υπερπαραμέτροι, διότι η επιλογή τους γίνεται αυτόματα από την επεξεργασία των δεδομένων.

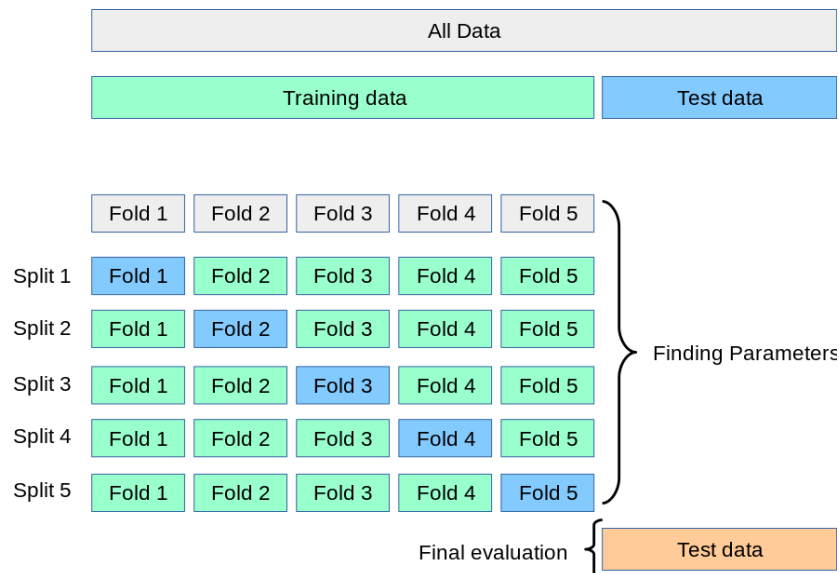
4.2.2. Γενίκευση Μοντέλου

Αφού όλα τα δεδομένα έχουν υποστεί κατάλληλη επεξεργασία ώστε να είναι συμβατά μεταξύ τους, γίνεται γενίκευση με δεδομένα εκτός του δείγματος που χρησιμοποιήθηκε για εκπαίδευση (out of sample testing). Αυτό συμβαίνει, διότι η εκμάθηση των παραμέτρων μίας συνάρτησης πρόβλεψης και ο έλεγχός της στα ίδια δεδομένα αποτελεί μεθοδολογικό λάθος και οδηγεί σε υπερπροσαρμογή (overfitting) ή υποπροσαρμογή (underfitting). Για να αποφευχθεί κάτι τέτοιο, διαχωρίζεται ένα μέρος των διαθέσιμων δεδομένων ως σετ ελέγχου (test set), δηλαδή X_test , y_test . Στο **Σχήμα 3** παρουσιάζεται ένα διάγραμμα ροής της τυπικής διαδικασίας διαχείρισης των δεδομένων κατά την εκπαίδευση ενός μοντέλου:



Σχήμα 3: Διάγραμμα ροής της τυπικής διαδικασίας Πλέγματος Υπερπαραμέτρων

Με τη διασταυρούμενη επικύρωση (cross validation) προκύπτει μία πιο αξιόπιστη εκτίμηση της γενίκευσης σε πολλά υποσύνολα. Στη βασική της μορφή, που ονομάζεται k-fold cross validation, το μοντέλο χωρίζεται σε μικρότερα k υποσύνολα. Έτσι, εκπαιδεύεται σε k-1 υποσύνολα (folds) και τα υπόλοιπα χρησιμοποιούνται ως test set για να υπολογιστεί ένα μέτρο επίδοσης. Η τελική επίδοση είναι ο μέσος όρος των μετρήσεων που προκύπτουν σε όλους τους κύκλους (**Σχήμα 4**).



Σχήμα 4: Απεικόνιση διαδικασίας k-fold διασταυρούμενης επικύρωσης

Στην παρούσα διπλωματική χρησιμοποιήθηκαν 5 υποσύνολα, ως παράμετρος.

4.2.3.Συνολικά Αποτελέσματα

Στον **Πίνακα 2** παρουσιάζονται τα συνολικά αποτελέσματα που προέκυψαν από την εκπαίδευση των μοντέλων κατά τη διάρκεια της ημέρας. Από τις παραπάνω τιμές, προκύπτει ότι μόνο τα μοντέλα του Random Forest και του XGBoost είναι ικανά να ερμηνεύσουν το συγκεκριμένο τύπο και πλήθος δεδομένων και να δώσουν ικανοποιητικά αποτελέσματα. Επιπλέον, οι μετρικές αξιολόγησης βελτιώθηκαν αισθητά με την προσθήκη επιπλέον διαθέσιμων δεδομένων.

Πίνακας 2: Συνολικά αποτελέσματα κατά τη διάρκεια της ημέρας

RESULTS	Linear Rgression		Random Forest		XGBoost	
DAY	Initial Data	Enhanced Data	Initial Data	Enhanced Data	Initial Data	Enhanced Data
R ²	0.170	0,173	0.591	0.699	0.543	0.680
MAE	38.703	38.661	25.708	22.731	17.408	22.805
MSE	3263.53	3248.72	1607.47	1182.24	1022.02	1258.12

Τα αποτελέσματα που προέκυψαν κατά τη διάρκεια της νύχτας απεικονίζονται στον **Πίνακα 3**. Τις βραδινές ώρες εξακολουθούν να υπερισχύουν τα μοντέλα Random Forest και XGBoost με χαμηλότερα σφάλματα, λόγω των πιο σταθερών συνθηκών. Οι διαφορές μεταξύ των αρχικών και των ενισχυμένων δεδομένων, καθιστούν το XGBoost το πιο κατάλληλο. Το R² παρουσιάζει γενικής μείωση, καθώς η σχέση μεταξύ των χαρακτηριστικών και του χρόνου άφιξης στη στάση γίνεια πιο ασταθής λόγω των τυχαίων προτύπων κυκλοφορίας και του μειωμένου όγκου δεδομένων.

Πίνακας 3: Συνολικά αποτελέσματα κατά τη διάρκεια της νύχτας

RESULTS	Linear Rgression		Random Forest		XGBoost	
	Initial Data	Enhanced Data	Initial Data	Enhanced Data	Initial Data	Enhanced Data
R ²	0.185	0.184	0.526	0.568	0.543	0.562
MAE	28.906	28.935	18.597	17.836	17.408	17.521
MSE	1721.04	1724.17	1052.28	952.21	1022.02	984.95

Το πρώτο αποτέλεσμα που προκύπτει είναι εκείνο του R² που επιδεικνύει την ικανότητα περιγραφής του μοντέλου. Στον **Πίνακα 4** και **Πίνακα 5** απεικονίζονται αυτά τα αποτελέσματα στις δύο κατηγορίες δεδομένων για ημέρα και νύχτα αντίστοιχα για ημέρα και νύχτα αντίστοιχα:

Πίνακας 4: Αποτελέσματα R² κατά τη διάρκεια της ημέρας

R ² -DAY	Initial Data	Enhanced Data
Linear Regression	0.170	0,173
Random Forest	0.591	0.699
XGBoost	0.543	0.680

Παρατηρούνται εξαιρετικά χαμηλές τιμές στο μοντέλου του Linear Regression που επιδεικνύουν ότι μπορεί να εξηγήσει ένα πολύ χαμηλό ποσοστό της διακύμανσης και επομένως δεν είναι κατάλληλο για χρήση. Το Random Forest Regression και το XGBoost δίνουν ικανοποιητικά και πολύ κοντινά αποτελέσματα, ιδιαίτερα αφότου έχουν προστεθεί τα παραπάνω δεδομένα.

Πίνακας 5: Αποτελέσματα R^2 κατά τη διάρκεια της νύχτας

R^2 -NIGHT	Initial Data	Enhanced Data
Linear Regression	0.185	0.184
Random Forest	0.526	0.568
Boost	0.543	0.562

Όσον αφορά στο μοντέλο της Linear Regression και στην περίπτωση της νύχτας κρίνεται ακατάλληλο, αλλά ελαφρώς καλύτερο σε σύγκριση με την ημέρα, λόγω μείωσης της πολυπλοκότητας των δεδομένων. Οι επιδόσεις στο Random Forest είναι χαμηλότερες σε σύγκριση με την ημέρα, καθώς τα δεδομένα είναι πιο αραιά και πιο ασυνεχή. Ακόμη, πτώση παρουσιάζεται και στο XGBoost για τον ίδιο λόγο, παρόλα αυτά παρουσιάζει μία σταθερότητα.

Το R^2 είναι αισθητά υψηλότερο τη μέρα σχεδόν σε όλα τα μοντέλα. Τα δεδομένα κατά τη διάρκεια τις ημέρας παρέχουν και πιο πολύπλοκη πληροφορία και πολύ περισσότερη, γεγονός που εκπαιδεύει το μοντέλο σε μοτίβα και διακυμάνσεις πολύ πιο αποτελεσματικά.

Ακόμη μία μετρική αξιολόγησης αποτελεί το MAE, το οποίο θα πρέπει να διατηρείται σε χαμηλές τιμές, ώστε να μην υπάρχει μεγάλο σφάλμα στις προβλέψεις. Στον **Πίνακα 6** και **Πίνακα 7** παραθέτονται τα αποτελέσματα MAE στις δύο περιπτώσεις δεδομένων για ημέρα και νύχτα αντίστοιχα:

Πίνακας 6: Αποτελέσματα MAE κατά τη διάρκεια της ημέρας

MAE-DAY	Initial Data	Enhanced Data
Linear Regression	38.703	38.661
Random Forest	25.708	22.731
XGBoost	17.408	22.805

Οι τιμές του MAE επιβεβαιώνουν την αδυναμία του Linear Regression και όσον αφορά το σφάλμα. Παρουσιάζει τις πιο υψηλές τιμές σε σύγκριση με τα άλλα μοντέλα, γεγονός που καθιστά την πρόβλεψη αδύναμη. Το Random Forest δίνει μία καλή απόδοση όσον αφορά το σφάλμα που θα μπορούσε να θεωρηθεί ικανοποιητική. Το XGBoost παρουσιάζει την καλύτερη γενική εικόνα, επομένως αποτελεί την πιο κατάλληλη επιλογή.

Πίνακας 7: Αποτελέσματα MAE κατά τη διάρκεια της νύχτας

MAE-NIGHT	Initial Data	Enhanced Data
Linear Regression	28.906	28.935
Random Forest	18.597	17.836
XGBoost	17.408	17.521

Το Linear Regression πάλι παρουσιάζει τις υψηλότερες τιμές, επομένως και τη νύχτα αποτελεί το λιγότερο αποδοτικό μοντέλο. Το Random Forest είναι αποτελεσματικό με το XGBoost να βρίσκεται σε ελαφρώς καλύτερη θέση. Γενικότερα, παρατηρείται μείωση του MAE καθώς η κυκλοφορία είναι πιο σταθερή και οι ταχύτητες πιο ομοιόμορφες, χωρίς πολλούς εξωτερικούς παράγοντες.

Γενικά, το MAE είναι χαμηλότερο κατά τη νύχτα σε όλες τις περιπτώσεις. Τις βραδινές ώρες τα δρομολόγια είναι πολύ λιγότερα, πιο αραιά και οι διαδρομές πιο σύντομες. Τα λεωφορεία είναι λιγότερο συμφωρημένα και η επιβατική κίνηση μικρότερη, γεγονός που εξασθενεί την επιρροή των εξωτερικών παραγόντων. Επιπλέον, σπάνια ένα λεωφορείο θα σταματήσει σε όλες τις στάσεις, επομένως απορρίπτονται αρκετές αλληλουχίες στάσεων και τα δεδομένα που μπορεί να επεξεργαστεί το μοντέλο μειώνονται.

Το Mean Squared Error εκφράζει το σφάλμα με μία ευαισθησία στις μεγάλες αποκλίσεις λόγω του τετραγώνου στο οποίο υψώνεται. Στόχος είναι να διατηρείται όσο το δυνατόν πιο χαμηλά. Στον **Πίνακα 8** και **Πίνακα 9** φαίνονται οι τιμές του MSE που προέκυψαν για ημέρα και νύχτα αντίστοιχα σε όλες τις περιπτώσεις δεδομένων:

Πίνακας 8: Αποτελέσματα MSE κατά τη διάρκεια της ημέρας

MSE -DAY	Initial Data	Enhanced Data
Linear Regression	3263.53	3248.72
Random Forest	1607.47	1182.24
XGBoost	1022.02	1258.12

Το Linear Regression παρουσιάζει με μεγάλη διαφορά τα μεγαλύτερα αποτελέσματα, επομένως υπάρχει μεγάλο σφάλμα στην πρόβλεψη και στις δύο κατηγορίες δεδομένων. Το Random Forest παρουσιάζει αισθητά βελτιωμένες και αποδεκτές τιμές με το XGBoost να

υπερτερεί στα γενικά αποτελέσματα και να φαίνεται το πιο αξιόπιστο. Παρακάτω απεικονίζονται τα αποτελέσματα σε γράφημα:

Στον **Πίνακα 9** φαίνονται τα αποτελέσματα του MSE κατά τη διάρκεια της νύχτας. Κατά τις βραδινές ώρες, το Linear Regression παρουσιάζει μεγάλες τιμές MSE και συγκριτικά πολύ μεγαλύτερες από τα άλλα δύο μοντέλα, γεγονός που το καθιστά μη ικανό να περιγράψει την πολυπλοκότητα των δεδομένων. Το Random Forest φαίνεται να επωφελήθηκε αρκετά από την ενίσχυση, ενώ το XGBoost φαίνεται να διατηρεί σταθερή απόδοση και να έχει καλή προσαρμοστικότητα.

Πίνακας 9: Αποτελέσματα MSE κατά τη διάρκεια της νύχτας

MSE -NIGHT	Initial Data	Enhanced Data
Linear Regression	1721.04	1724.17
Random Forest	1052.28	952.21
XGBoost	1022.02	984.95

Γενικά, υπάρχει μείωση και στο MSE, αλλά λιγότερο σταθερή. Δεν παρουσιάζονται μεγάλες αποκλίσεις λόγω της ευστάθειας των νυχτερινών δρομολογίων. Αυτό συμβαίνει διότι δεν υπάρχουν μεγάλες διακυμάνσεις σε εξωτερικά δεδομένα. Ωστόσο, ο αριθμός των δειγμάτων είναι μικρότερος και κάποια σφάλματα μπορούν να έχουν μεγαλύτερη επίδραση.

Σε όλες τις περιπτώσεις, τα επιπλέον δεδομένα ενίσχυσαν τα μοντέλα. Μείωσαν τα σφάλματα και αύξησαν το ποσοστό περιγραφής της διακύμανσης. Τα αναλυτικά αποτελέσματα των ενισχυμένων δεδομένων, εμφανίζονται στον *Πίνακα 10* και *Πίνακα 11*:

Πίνακας 10: Αποτελέσματα μετρικών Αξιολόγησης για κάθε μοντέλο κατά τη διάρκεια της ημέρας

DAY	R ²	MAE	MSE
Linear Regression	0,173	38.661	3248.72
Random Forest	0.699	22.731	1182.24
XGBoost	0.680	22.805	1258.12

Πίνακας 11: Αποτελέσματα μετρικών Αξιολόγησης για κάθε μοντέλο κατά τη διάρκεια της ημέρας

NIGHT	R ²	MAE	MSE
Linear Regression	0.184	28.935	1724.17
Random Forest	0.568	17.836	952.21
XGBoost	0.562	17.521	984.95

Από τα παραπάνω αποτελέσματα μπορούν να αντληθούν τα εξής συμπεράσματα:

Το μοντέλο Linear Regression δυσκολεύεται να αποτυπώσει τις έντονες διακυμάνσεις στους χρόνους άφιξης, ειδικά όταν υπάρχουν απότομες αυξομειώσεις. Η απόδοσή του είναι συγκριτικά χαμηλότερη από πιο πολύπλοκα μοντέλα όπως το Random Forest ή το XGBoost, επειδή κάνει γραμμικές υποθέσεις. Οι γραμμές του μοντέλου έχουν γενικά ομαλότερη καμπύλη και δεν προσαρμόζονται καλά στα αληθινά δεδομένα, που συχνά είναι πιο αυξομιούμενα. Η ακρίβεια είναι πολύ χαμηλή σύμφωνα με τους δείκτες αξιολόγησης και παρουσιάζει πολύ χαμηλό R², που δείχνει αδυναμία εξήγησης μεταβλητότητας των δεδομένων. Γενικότερα είναι ακατάλληλο για την πρόβλεψη ETA, διότι απλουστεύει υπερβολικά το πρόβλημα και δεν αποδίδει. Μπορεί να χρησιμοποιηθεί μόνο ως γραμμή σύγκρισης.

Το μοντέλο παρουσιάζει υψηλή ακρίβεια με χαμηλές τιμές MAE και MSE (ελαφρώς υψηλότερες από XGBoost) και πολύ καλή επεξήγηση της διακύμανσης R² (λίγο υψηλότερη από XGBoost). Έχει αρκετά καλή απόδοση και προσαρμόζεται στην πολυπλοκότητα των μη γραμμικών δεδομένων. Μπορεί και εξηγεί σε ικανοποιητικό βαθμό τη διακύμανση των δεδομένων. Είναι επίσης ευπροσάρμοστο σε δεδομένα με θόρυβο. Συμπερασματικά είναι αρκετά αποδοτικό, σταθερό και αξιόπιστο μοντέλο, κατάλληλο για συγκοινωνιολογικά συστήματα.

Το μοντέλο XGBoost παρουσιάζει πολύ υψηλή ακρίβεια με χαμηλές τιμές MAE και MSE. Το R² είναι πολύ καλό και δείχνει ότι το μοντέλο εξηγεί πολύ καλά τη συσχέτιση μεταξύ των χαρακτηριστικών. Γενικά, αποτελεί ένα από τα πιο ισχυρά και αποδοτικά μοντέλα μηχανικής μάθησης για προβλήματα παλινδρόμησης. Συνδυάζει πολλές αδύναμες προβλέψεις σε ένα ισχυρό μοντέλο. Οι αποδόσεις του είναι εξαιρετικά ικανοποιητικές για εφαρμογές συγκοινωνιακού χαρακτήρα.

4.3. Συγκριτική Αξιολόγηση

Βάσει των αποτελεσμάτων των μετρικών αξιολόγησης στο κάθε μοντέλο προκύπτουν κάποια γενικά συμπεράσματα για τη συμπεριφορά των μοντέλων και την αποδοτικότητά τους στην πρόβλεψη, τα οποία παραθέτονται στον **Πίνακα 12**.

Πίνακας 12: Σύγκριση μοντέλων και γενικές παρατηρήσεις

Κριτήριο	XGBoost	Random Forest	Linear Regression
ΜΑΕ (Σφάλμα)	Χαμηλό (17–23 sec)	Χαμηλό, Συγκρίσιμο	Υψηλό (30–39 sec)
MSE (Διασπορά σφαλμάτων)	Σταθερό & χαμηλό	Παρόμοιο ή ελαφρώς καλύτερο	Πολύ υψηλό
R^2	Πολύ υψηλό	Ικανοποιητικά υψηλό	Χαμηλό έως κακό (<0.2)
Μη-γραμμικές Σχέσεις	Αντιμετωπίζονται πλήρως	Αντιμετωπίζονται καλά	Δεν τις καλύπτει
Αντοχή σε outliers	Πολύ καλή	Πολύ καλή	Ευάλωτη
Ερμηνευσιμότητα μεταβλητών	Καλή	Καλή	Περιορισμένη
Αξιοπιστία σε διαφορετικά σενάρια	Πολύ καλή	Καλή	Μη ικανοποιητική
Υπολογιστική Αποδοτικότητα	Πολύ καλή	Πολύ καλή	Άριστη, αλλά με χαμηλή ακρίβεια

Από την απεικόνιση των αποτελεσμάτων προέκυψαν κάποια ευρήματα όσον αφορά το διαχωρισμό ημέρα και νύχτας, τα οποία παρουσιάζονται στον **Πίνακα 13**:

Πίνακας 13: Σύγκριση ημέρας και νύχτας βάσει διαγραμμάτων

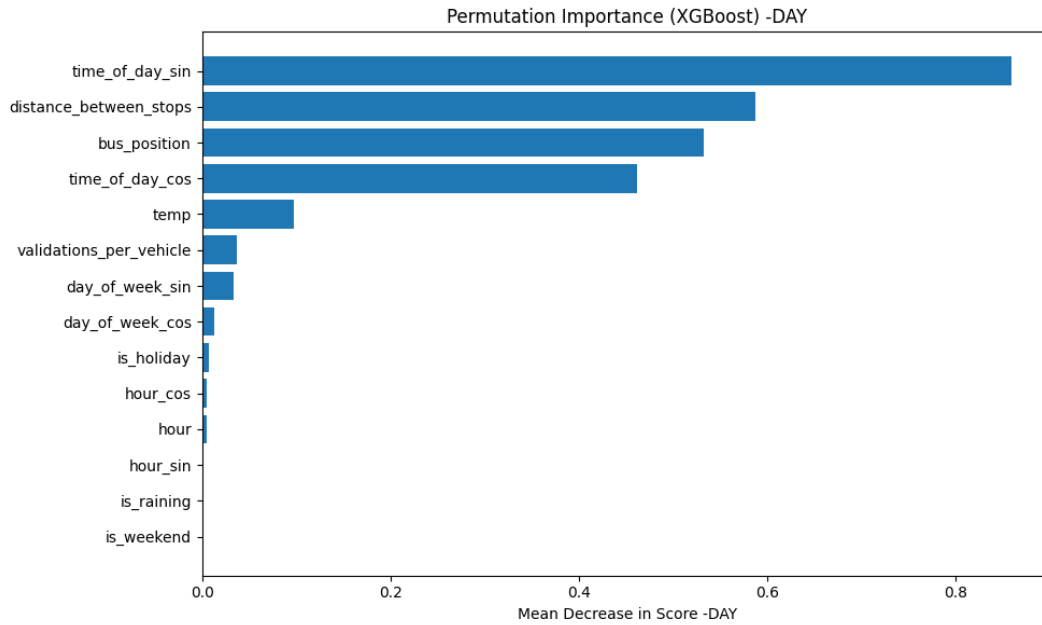
Κριτήριο Αξιολόγησης	Ημέρα (07:00–21:59)	Νύχτα (22:00–06:59)
Ακρίβεια πρόβλεψης	Μέτρια – καλή στις περισσότερες περιπτώσεις	Πολύ καλή και σταθερή ακρίβεια
Ευθυγράμμιση με πραγματική καμπύλη	Συχνά αποκλίνει, ειδικά σε στάσεις με αιχμές	Γενικά ακολουθεί καλά την πραγματική καμπύλη
Αντίδραση σε αιχμές ETA	Καθυστερημένη ή μη αντιπροσωπευτική	Ήπια αντίδραση – συχνά αγνοεί τις αιχμές
Υποεκτιμήσεις	Περιορισμένες – εμφανίζονται σποραδικά	Παρουσιάζονται ήπια σε στάσεις μέσης διαδρομής
Υπερεκτιμήσεις	Εμφανείς – κυρίως σε αρχικές στάσεις ή σε ώρες αιχμής	Λιγότερες – πιο κοντά στις πραγματικές τιμές
Προβλεπτική συνέπεια	Κυμαίνεται – σε ορισμένες διαδρομές είναι ασταθής	Υψηλή – η σταθερότητα της κυκλοφορίας βοηθά
Απόδοση σε μεταβατικά σημεία	Δυσκολία να συλλάβει μεταβολές ανάμεσα σε στάσεις	Περιορισμένη μεταβλητότητα – λιγότερα λάθη
Ομαλότητα γραμμής πρόβλεψης	Εμφανίζει εξομάλυνση, αλλά όχι πάντα προς τη σωστή κατεύθυνση	Πιο λεία καμπύλη, κοντά στην πραγματικότητα
Γενική εικόνα συμπεριφοράς μοντέλου	Αντιμετωπίζει προκλήσεις λόγω κυκλοφορίας και αιχμών	Εμφανίζει υψηλή σταθερότητα και χαμηλή απόκλιση

Γενικά ο διαχωρισμός αυτός βελτίωσε την ποιότητα των προβλέψεων, καθώς αναλύει ξεχωριστά χρονικές ζώνες με πολύ διαφορετικές συνθήκες. Τα μη γραμμικά μοντέλα (XGBoost, Random Forest), αποδίδουν καλύτερα στην επεξήγηση της πολύπλοκης σχέσης μεταξύ των μεταβλητών. Αντίθετα η γραμμική παλινδρόμηση αποτυγχάνει τόσο τη μέρα, αλλά ακόμα περισσότερο τη νύχτα.

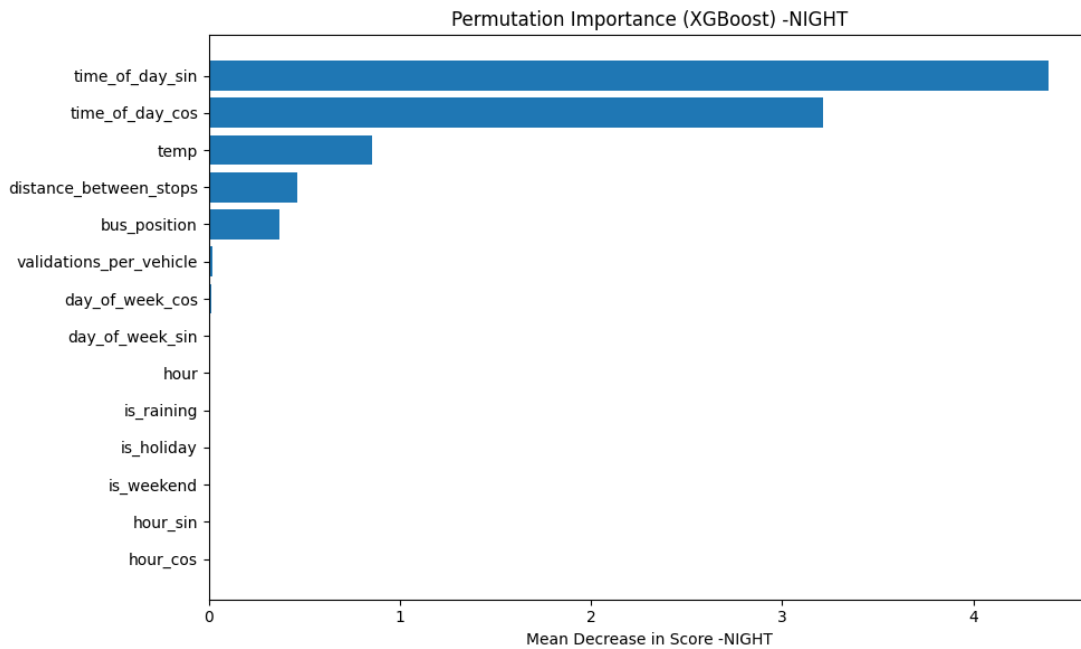
4.4. Ανάλυση Αποτελεσμάτων XGBoost

Το XGBoost κρίθηκε το πιο κατάλληλο μοντέλο, διότι μπορεί να περιγράψει καλύτερα τα δεδομένα. Είναι σταθερό και αποτελεσματικό στη συγκεκριμένη περίπτωση, καθώς αναλύει και συσχετίζει πολύπλοκα δεδομένα μεταξύ τους. Αυτό γίνεται αντιληπτό από τις μετρικές αξιολόγησης, οι οποίες διατηρούν υψηλές τιμές χωρίς μεγάλη απόκλιση στις δύο περιπτώσεις δεδομένων.

Αρχικά, παρουσιάζονται τα γραφήματα Permutation Importance για το μοντέλο του XGBoost για τις πρωινές και τις βραδινές ώρες **Διάγραμμα 4** και **Διάγραμμα 5**:



Διάγραμμα 4: Απεικόνιση Σημασίας Χαρακτηριστικών για XGBoost κατά τη διάρκεια της ημέρας



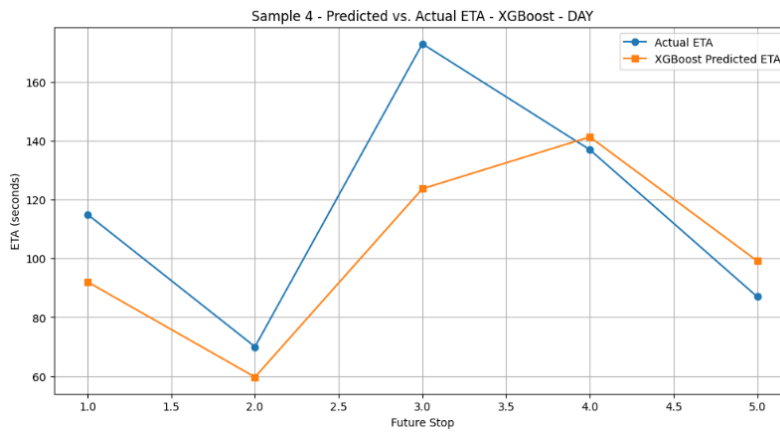
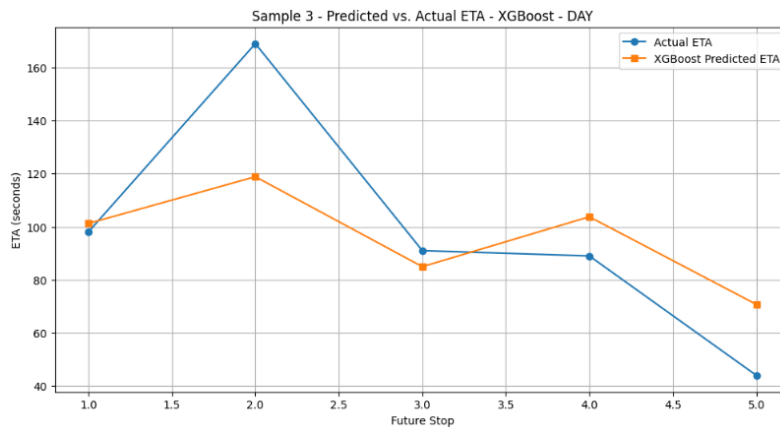
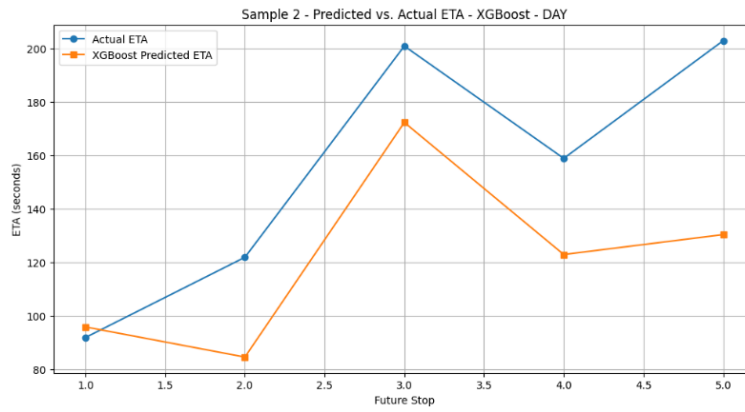
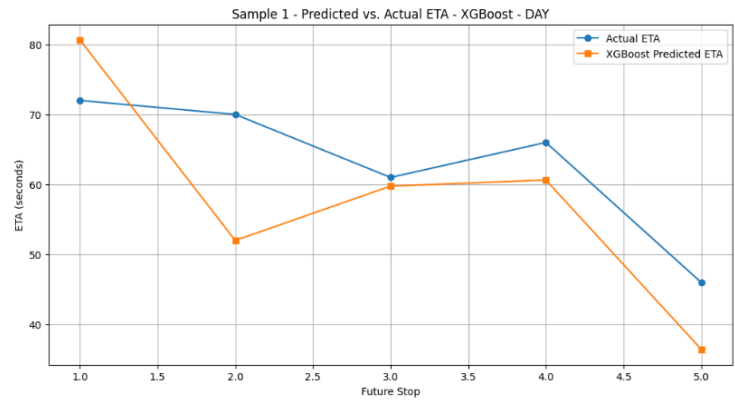
Διάγραμμα 5: Απεικόνιση Σημασίας Χαρακτηριστικών για XGBoost κατά τη διάρκεια της νύχτας

Κύρια χαρακτηριστικά:

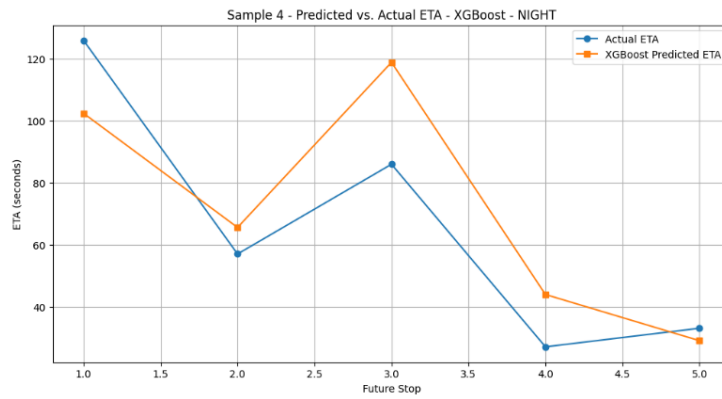
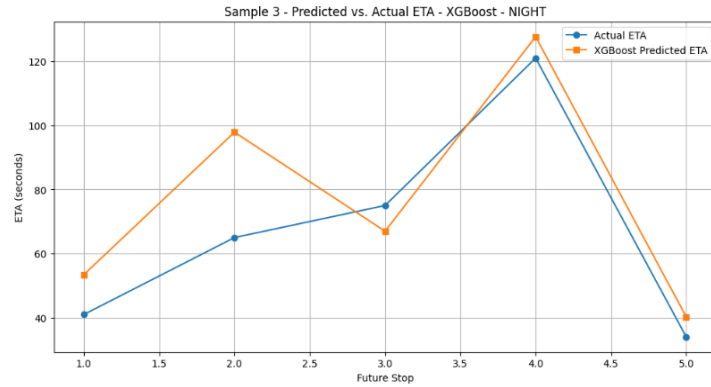
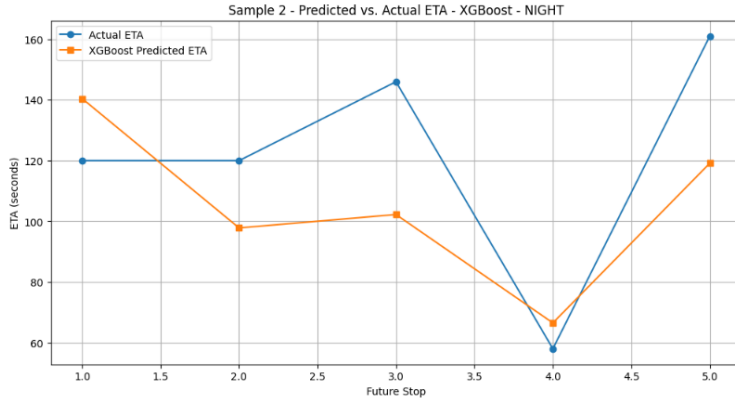
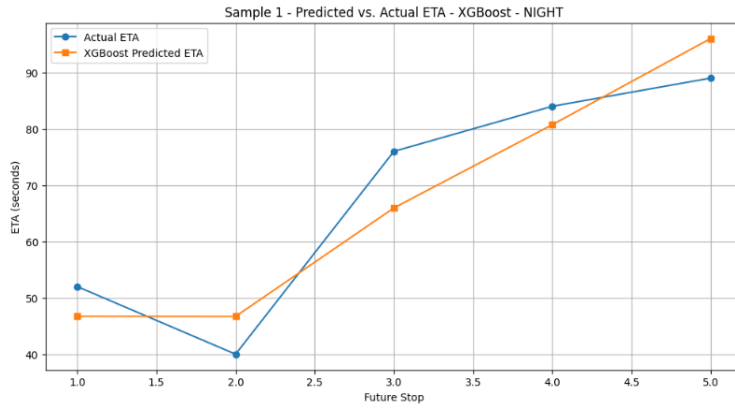
- `time_of_day_sin`: Αποτελεί πολύ ισχυρό δείκτη του χρόνου άφιξης
- `distance_between_stops`: Σε εξίσου μεγάλο βαθμό επηρεάζει και η απόσταση, καθώς όσο μεγαλύτερη απόσταση, τόσο μεγαλύτερος χρόνος άφιξης.
- `bus_position`: Επίσης κρίσιμο, καθώς καθορίζει πόσο έχει προχωρήσει το λεωφορείο στη διαδρομή.
- `time_of_day_cos`: Συμπληρώνει το `time_of_day_sin`, περιγράφει την ώρα ως διάνυσμα.
- `Temp`: Μικρή, αλλά υπαρκτή επίδραση που σχετίζεται με την εποχικότητα και τις συνθήκες κίνησης.
- `validations_per_vehicle`: Ελάχιστη επίδραση, σχετική μόνο σε πολύ γεμάτα ή άδεια δρομολόγια.
- `is_raining`, `is_weekend`, `is_holiday`: Χαμηλή σημασία, πιθανόν επειδή το μοντέλο δεν βρίσκει συστηματική επίδραση ή δεν υπάρχει αρκετή ποικιλία στα δεδομένα.

Άλλος ένας λόγος για τον οποίο υπερτερεί έναντι του Random Forest είναι ότι περισσότερα χαρακτηριστικά έχουν σημαντικό ρόλο στην ανάλυση και τα αποτελέσματα, γεγονός που το καθιστά πιο αξιόπιστο.

Παρακάτω, παρουσιάζονται τέσσερεις δοκιμές εφαρμογής του XGBoost, για κάθε χρονική περίοδο. Στα **Διάγραμμα 6** και **Διάγραμμα 7** απεικονίζονται σε διαγράμματα οι προβλεπόμενες τιμές σε σύγκριση με τις πραγματικές και μπορούν να φανούν καθαρά οι ικανότητες πρόβλεψης του μοντέλου κατά τη διάρκεια της ημέρας και της νύχτας αντίστοιχα.



Διάγραμμα 6: Διαγράμματα προβλεπόμενων και πραγματικών χρόνων άφιξης για XGBoost (Ημέρα)



Διάγραμμα 7: Διαγράμματα προβλεπόμενων και πραγματικών χρόνων άφιξης για XGBoost (Νύχτα)

Στο **Διάγραμμα 6** παρατηρούνται πολύ καλές προβλέψεις, με σχεδόν παράλληλες καμπύλες, ενώ οι διαφορές είναι μικρές και σταθερές. Επικρατεί μία υποτίμηση των τιμών, χωρίς να είναι μεγάλο το σφάλμα.

Στο **Διάγραμμα 7** το μοντέλο κάνει αρκετά καλές προβλέψεις σε όλες τις δοκιμές με μία μικρή δυσκολία σε μερικές περιπτώσεις στις πολύ μεγάλες αιχμές, Παρόλα αυτά, ακολουθεί ικανοποιητικά τη γενική τάση με μικρές αποκλίσεις.

Το XGBoost τη νύχτα αποδίδει καλύτερα σε σταθερές συνθήκες, καθώς φαίνεται να κάνει πιο συνεπείς προβλέψεις (επιβεβαιώνεται και από το χαμηλότερο MAE). Την ημέρα, το μοντέλο αποτυγχάνει όταν υπάρχουν έντονες αυξομειώσεις λόγω μεγαλύτερης πολυπλοκότητας (κυκλοφορία, πληρότητα κ.λπ.). Η ποιότητα της πρόβλεψης είναι καλύτερη τη νύχτα, παρά το ελαφρώς χαμηλότερο R^2 , αυτό δείχνει ότι προβλέπει πιο σταθερά έστω και με λιγότερη εξήγηση της διακύμανσης.

Γενικά ακολουθείται η γενική τάση και οι πτώσεις και οι κορυφώσεις προβλέπονται κοντά στις πραγματικές τιμές. Οι αποκλίσεις διατηρούνται σε αποδεκτές τιμές, ενώ ακόμα και κατά τις νυχτερινές ώρες, η προβλεψιμότητα είναι σε αρκετά καλό επίπεδο.

Συμπερασματικά, το XGBoost είναι το πιο κατάλληλο μοντέλο διότι:

1. Ισορροπεί ακρίβεια και γενίκευση:
 - Διατηρεί χαμηλά σφάλματα (MAE/MSE) χωρίς να υπερεκπαιδεύεται.
 - Το R^2 υποδηλώνει ότι κατανοεί σε βάθος τη σχέση μεταξύ των χαρακτηριστικών.
2. Αντιμετωπίζει πολύπλοκα, μη γραμμικά μοτίβα:
 - Στο πρόβλημα ETA, παράγοντες όπως η ώρα, η θέση και η απόσταση συσχετίζονται μη γραμμικά, κάτι που το XGBoost χειρίζεται με εξαιρετική ακρίβεια.
3. Ανθεκτικότητα σε σφάλματα και outliers:
 - Δεν παρασύρεται από ακραίες τιμές, ειδικά σε περιπτώσεις μεταβολών λόγω συνθηκών (π.χ. κυκλοφορία, βράδυ).
4. Ερμηνευσιμότητα και σημασία χαρακτηριστικών:
 - Το XGBoost επιτρέπει αξιολόγηση της "feature importance", κάτι χρήσιμο για περαιτέρω βελτιώσεις ή επεξηγήσεις.

5. Συμπεράσματα

5.1. Βασικά Συμπεράσματα

Η παρούσα διπλωματική εργασία εστίασε στην αξιολόγηση και σύγκριση τριών διαφορετικών μοντέλων μηχανικής μάθησης για την πρόβλεψη του εκτιμώμενου χρόνου άφιξης λεωφορείων σε στάση, τόσο κατά τη διάρκεια της μέρας όσο και της νύχτας. Η

ανάλυση έγινε μόνο για τα δεδομένα τηλεματικής του ΟΑΣΑ αρχικά και έπειτα με επιπρόσθετα εξωτερικά δεδομένα.

Στην αρχή οι αναλύσεις πραγματοποιήθηκαν στα στοιχεία τηλεματικής, αποκλειστικά για χωρικά και χρονικά δεδομένα. Έπειτα έγινε προσθήκη δεδομένων καιρού (θερμοκρασία και ύπαρξη βροχής), επικυρώσεων εισιτηρίων και αργιών. Σκοπός των δύο αναλύσεων είναι να αξιολογηθεί αν η προσθήκη δεδομένων βελτιώνει σημαντικά το μοντέλο.

Από τις απεικονίσεις που προέκυψαν παρατηρήθηκαν βελτιώσεις των αποτελεσμάτων στα Μοντέλα του Random Forest και του XGBoost, αλλά όχι στο Linear Regression λόγω της αδυναμίας του να συσχετίσει τόσα πολύμορφα δεδομένα.

Στα διαγράμματα Permutation Importance, χαρακτηριστικά όπως bus_position, distance_between_stops, hour_sin, hour_cos κυριάρχησαν σε σημαντικότητα, παρόλα αυτά και η μικρή συμβολή των περαιτέρω δεδομένων διαφοροποίησε τα αποτελέσματα.

Στον **Πίνακα 14** γίνεται αξιολόγηση κόστους – οφέλους στην προσθήκη δεδομένων

Πίνακας 14: Αξιολόγηση κόστους – οφέλους στην προσθήκη δεδομένων

Παράγοντας	Αξιολόγηση
Χρόνος επεξεργασίας	Υψηλός
Ωφέλεια στην ακρίβεια	Μέτρια, αλλά σημαντική
Αύξηση πολυπλοκότητας	Σημαντική
Ερμηνευσιμότητα	Αυξάνεται σε RF και XGBoost

Η προσθήκη εξωτερικών δεδομένων οδήγησε σε ουσιαστική βελτίωση των μοντέλων, αλλά αύξησε την πολυπλοκότητα, και επομένως το χρόνο επεξεργασίας. Τα χρονικά και χωρικά δεδομένα είναι τα πιο αποτελεσματικά για εφαρμογές πρόβλεψης ETA. Εξωτερικά δεδομένα θα μπορούσαν να ενισχύσουν την πρόβλεψη ακόμα περισσότερο εάν είχαν μεγάλο χρονικό βάθος ή εάν ήταν πιο αξιόπιστα. Για παράδειγμα, για την πόλη της Αθήνας, πιο ακριβή δεδομένα για τις επικυρώσεις εισιτηρίων και για τον κυκλοφοριακό φόρτο στους κεντρικούς δρόμους θα είχαν σημαντική συνεισφορά στο μοντέλο. Ακόμα, σημαντική πληροφορία θα ήταν πόσο συχνά κλείνουν δρόμοι λόγω διαδηλώσεων ή σημαντικών γεγονότων, φαινόμενο στην καθημερινότητα του μέσου Αθηναίου.

Παρόλα αυτά, λόγω της πολυπλοκότητας των δεδομένων, η καλύτερη λύση θα ήταν ένα υβριδικό μοντέλο Machine και Deep Learning, το οποίο εφαρμόζεται σε πολλές χώρες με πολύ ικανοποιητικές αποδόσεις.

Ο διαχωρισμός αναλύσεων μεταξύ ημέρας και νύκτας έγινε εξαρχής διότι τις βραδινές ώρες, οι συνθήκες είναι πολύ διαφορετικές με αποτέλεσμα η διακύμανση να επηρεάζεται σημαντικά και να μη βγαίνουν αληθή αποτελέσματα για ένα συνεχόμενο 24ωρο.

Η συγκριτική ανάλυση των δεικτών αξιολόγησης των μοντέλων έδειξε κάποιες γενικές τάσεις. Σχεδόν σε όλες τις περιπτώσεις τα σφάλματα MAE και MSE μειώθηκαν με την προσθήκη επιπλέον δεδομένων, ενώ το R^2 αυξήθηκε, όπως και η ικανότητα περιγραφής της διακύμανσης. Ανάμεσα σε πρωινές και νυχτερινές ώρες, τα σφάλματα παρουσιάζουν μείωση κατά το βράδυ λόγω του λιγότερου πλήθους δεδομένων και επομένως ευκολία στην πρόβλεψη

5.2.Προτάσεις για Περαιτέρω Έρευνα

Η συγκεκριμένη διπλωματική εστίασε στο οδικό δίκτυο της Αθήνας, επομένως δεν επιβεβαιώνει ότι οι προβλεπτικές ικανότητες των μοντέλων μπορούν να εφαρμοστούν και σε διαφορετικά δίκτυα.

Υπήρξαν αρκετά εμπόδια όσον αφορά στην εύρεση έγκυρων δεδομένων που να αφορούν την ίδια χρονική περίοδο, επομένως η έρευνα δεν πραγματοποιήθηκε σε μεγάλο όγκο δεδομένων, ενώ δεν μπόρεσαν να αντληθούν δεδομένα τα οποία θα αποτελούσαν κρίσιμο παράγοντα για την ακρίβεια της πρόβλεψης. Επιπλέον, η Αθήνα διακρίνεται από μεγάλη ποικιλομορφία στη γεωμετρία της, γεγονός που αυξάνει την πολυπλοκότητα του προβλήματος.

Αρχικά, αιτείται η επαναξιολόγηση των μοντέλων πρόβλεψης με μεγαλύτερο όγκο έγκυρων δεδομένων, που να επεκτείνονται σε πολύ παραπάνω λεωφορειακές γραμμές, αλλά και σε παραπάνω Μέσα Μαζικής Μεταφοράς.

Ακόμα, η αύξηση των μεταβλητών (features) θα οδηγούσε την έρευνα ένα βήμα παρακάτω. Πληροφορίες όπως ο κυκλοφοριακός φόρτος, στατιστικά ατυχημάτων, περιπτώσεις αποκλεισμών δρόμων θα ενίσχυαν την ακρίβεια των αποτελεσμάτων.

Σημαντική θα ήταν και η εφαρμογή μοντέλων Βαθιάς Μάθησης (Deep Learning) που παρουσιάζουν αντίστοιχη ανάπτυξη και θα έδιναν μία ενδιαφέρουσα σύγκριση. Σε περαιτέρω έρευνα θα μπορούσε να εξεταστεί και η εφαρμογή ενός υβριδικού μοντέλου.