ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΑΤΡΩΝ
ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ Η/Υ & ΠΛΗΡΟΦΟΡΙΚΗΣ

ΑΝΑΦΟΡΑ ΠΑΡΑΔΟΣΗΣ

11 Εργασία
Quality of Service 5G - Μοντελοποίηση παλινδρόμησης με XGBoost regressor και random forest regressor

ΜΑΘΗΜΑ:

Αρχιτεκτονικές 5G

ΣΤΕΦΑΝΟΣ Ι. ΜΕΤΖΙΔΑΚΗΣ
ΑΜ: 1070107

ΚΑΙ

ΕΜΜΑΝΟΥΕΛΑ ΞΕΝΟΥ
ΑΜ : 1054286

Πάτρα, 2024

ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

ΠΙΝΑΚΑΣ ΕΙΚΟΝΩΝ

1. ΘΕΩΡΙΑ

1.1. Τι είναι το 5G

Το 5G αντιπροσωπεύει τη νέα γενιά τηλεπικοινωνιακών τεχνολογιών, που προσφέρει αναβαθμισμένες δυνατότητες επικοινωνίας και σύνδεσης σε σύγκριση με τις προηγούμενες γενιές (Διεθνής Ένωση Τηλεπικοινωνιών (ITU), 2022). Το κύριο χαρακτηριστικό του 5G είναι η υψηλή ταχύτητα μετάδοσης δεδομένων, που επιτρέπει τη λήψη και τη μετάδοση μεγάλου όγκου πληροφοριών με εκπληκτική ταχύτητα (Διεθνής Ένωση Τηλεπικοινωνιών (ITU), 2022). Επιπλέον, το 5G χαρακτηρίζεται από χαμηλότερη καθυστέρηση στη μετάδοση δεδομένων, γεγονός που οδηγεί σε αποτελεσματικότερη και αμεσότερη επικοινωνία (Διεθνής Ένωση Τηλεπικοινωνιών (ITU), 2022). Τέλος, μια από τις κύριες επιδιώξεις του 5G είναι η επίτευξη μεγαλύτερης συνδεσιμότητας, που θα επιτρέψει στις συσκευές να είναι συνδεδεμένες μεταξύ τους με μεγαλύτερη αξιοπιστία και αποτελεσματικότητα (Διεθνής Ένωση Τηλεπικοινωνιών (ITU), 2022). Αυτά τα χαρακτηριστικά καθιστούν το 5G μια εξαιρετικά ελκυστική τεχνολογία με πληθώρα προοπτικών και δυνατοτήτων για τη μελλοντική ανάπτυξη και εξέλιξη των τηλεπικοινωνιακών δικτύων (Διεθνής Ένωση Τηλεπικοινωνιών (ITU), 2022).
Το 5G έχει σημαντική επίδραση και στην ίδια την ανάλυση δεδομένων και τις τεχνολογίες συλλογής δεδομένων λόγω των χαρακτηριστικών του όπως περιεγράφηκαν παραπάνω:

- **Υψηλή Ταχύτητα και Χωρητικότητα:** Το 5G παρέχει υψηλότερες ταχύτητες μετάδοσης δεδομένων και μεγαλύτερη χωρητικότητα σε σύγκριση με τις προηγούμενες γενιές δικτύων (Ένωση Βιομηχανιών Κινητής Τηλεφωνίας (GSMA)., n.d.). Αυτό επιτρέπει την ταχύτερη συλλογή, μετάδοση και επεξεργασία μεγάλων όγκων δεδομένων (Ένωση Βιομηχανιών Κινητής Τηλεφωνίας (GSMA)., n.d.).

- **Χαμηλή Καθυστέρηση:** Η χαμηλή καθυστέρηση στη μετάδοση δεδομένων που προσφέρει το 5G επιτρέπει την πραγματοποίηση πραγματικού χρόνου ανάλυσης δεδομένων και ανταπόκρισης σε πραγματικό χρόνο σε διάφορες εφαρμογές, όπως τα αυτόνομα οχήματα, οι υπηρεσίες υγείας και η βιομηχανία.

- **Μεγαλύτερη Συνδεσιμότητα:** Η αυξημένη συνδεσιμότητα που προσφέρει το 5G επιτρέπει τη σύνδεση μεγαλύτερου αριθμού συσκευών και αισθητήρων σε ένα δίκτυο, δημιουργώντας ένα πλήθος νέων πηγών δεδομένων (Ένωση Βιομηχανιών Κινητής Τηλεφωνίας (GSMA)., n.d.)..

- **Καινοτόμες Εφαρμογές:** Το 5G επιτρέπει τη δημιουργία καινοτόμων εφαρμογών που βασίζονται στην ανάλυση δεδομένων, όπως η επαυξημένη πραγματικότητα, οι ευφυείς πόλεις, οι ιατρικές τηλεματικές υπηρεσίες και η βιομηχανική αυτοματοποίηση (Ένωση Βιομηχανιών Κινητής Τηλεφωνίας (GSMA)., n.d.)..

Οι τεχνολογίες συλλογής δεδομένων είναι αναγκαίες για τη συλλογή, την αποθήκευση και την επεξεργασία των δεδομένων που παράγονται από τα δίκτυα 5G και τις συνδεδεμένες συσκευές [(M. Chiang, 2020) και (Y. Wu, 2021)]. Αυτές οι τεχνολογίες περιλαμβάνουν αισθητήρες IoT, κεντρικές μονάδες επεξεργασίας δεδομένων (CPU), συστήματα αποθήκευσης δεδομένων (όπως cloud και edge computing) και λογισμικό ανάλυσης δεδομένων. Η εξέλιξη αυτών των τεχνολογιών συμβάλλει στην ανάπτυξη καινοτόμων λύσεων ανάλυσης δεδομένων που εκμεταλλεύονται τα πλεονεκτήματα του 5G (X. Li, 2022).

1.2. Περιγραφή των Δεδομένων

Τα δεδομένα μας αποτελούν ένα σύνολο που περιέχει πληροφορίες σχετικά με την ποιότητα της εξυπηρέτησης σε ένα δίκτυο 5G. Κάθε εγγραφή αντιστοιχεί σε έναν χρήστη και περιλαμβάνει πληροφορίες σχετικά με το είδος της εφαρμογής που χρησιμοποιείται, την ισχύ του σήματος, το χρόνο καθυστέρησης στη μετάδοση δεδομένων, καθώς και πληροφορίες σχετικά με την απαιτούμενη και τη διατεθείσα εύρος ζώνης, καθώς και το ποσοστό των πόρων που έχουν εκχωρηθεί στην εφαρμογή.
Ας δούμε πώς η κάθε μεταβλητή σχετίζεται με τις δυνατότητες του 5G:

- **Application Type (Τύπος Εφαρμογής):** Αυτή η μεταβλητή περιγράφει το είδος της εφαρμογής που χρησιμοποιείται στο δίκτυο. Το 5G επιτρέπει τη λειτουργία ποικίλων εφαρμογών με διαφορετικές απαιτήσεις σε ταχύτητα, χωρητικότητα και χρόνο απόκρισης.

- **Signal Strength (Ισχύς Σήματος):** Η ισχύς του σήματος επηρεάζει την ποιότητα της επικοινωνίας σε ένα δίκτυο 5G (Singh, 2023). Ένα ισχυρό σήμα εξασφαλίζει σταθερή και γρήγορη σύνδεση, ενώ ένα ασθενές σήμα μπορεί να οδηγήσει σε απώλεια σύνδεσης ή χαμηλή ταχύτητα μετάδοσης δεδομένων (Διεθνής Ένωση Τηλεπικοινωνιών (ITU), 2022).

- **Latency (Καθυστέρηση):** Η καθυστέρηση, ο χρόνος που απαιτείται για τα πακέτα δεδομένων να ταξιδέψουν από το ένα σημείο στο άλλο στο δίκτυο, είναι κρίσιμο για εφαρμογές που απαιτούν ανταπόκριση σε πραγματικό χρόνο, όπως αυτόνομα οχήματα και εμπειρίες εικονικής πραγματικότητας (Chiang, 2020). Η καθυστέρηση στη μετάδοση δεδομένων είναι κρίσιμη για πολλές εφαρμογές, όπως τα αυτόνομα οχήματα και οι εικονικές πραγματικότητες. Το 5G προσφέρει χαμηλή καθυστέρηση, επιτρέποντας την πραγματοποίηση ανταπόκρισης σε πραγματικό χρόνο.

- **Required Bandwidth (Απαιτούμενη Εύρος Ζώνης):** Η απαίτηση εύρους ζώνης καθορίζει την απόδοση της εφαρμογής (Y. Wu, 2021). Το 5G παρέχει μεγαλύτερη χωρητικότητα στο δίκτυο, επιτρέποντας τη μετάδοση μεγάλων όγκων δεδομένων με υψηλές ταχύτητες.

- **Allocated Bandwidth (Εκχωρημένο Εύρος Ζώνης):** Η εκχώρηση εύρους ζώνης σε μια εφαρμογή επηρεάζει την ταχύτητα και την απόδοση της μετάδοσης δεδομένων σε αυτήν (X. Li, 2022).

- **Resource Allocation (Κατανομή Πόρων):** Η κατανομή πόρων σε μια εφαρμογή είναι σημαντική για την εξασφάλιση της αποτελεσματικής λειτουργίας της σε ένα δίκτυο 5G. Η ορθή κατανομή πόρων επιτρέπει τη βέλτιστη χρήση τους και την εξασφάλιση της απόδοσης της εφαρμογής (Giordani, 2020).

Συνολικά, οι μεταβλητές μας αντικατοπτρίζουν την ποιότητα της υπηρεσίας σε ένα δίκτυο 5G και τις δυνατότητες που παρέχει για τη συλλογή, μετάδοση και επεξεργασία δεδομένων με υψηλές ταχύτητες, χαμηλή καθυστέρηση και αποτελεσματική χρήση πόρων. Η ανάλυση και η

κατανόηση αυτών των δεδομένων μας επιτρέπει να λάβουμε σημαντικά συμπεράσματα και να προβλέψουμε την απόδοση της εφαρμογής σε διάφορες συνθήκες.

1.3. Διάγραμμα περίληψης SHAP (shap_summary_plot)

Το διάγραμμα περίληψης SHAP, το οποίο είναι ένας συγκεκριμένος τύπος διαγράμματος δύναμης που χρησιμοποιείται για να εξηγήσει την επίδραση διαφορετικών χαρακτηριστικών στην έξοδο ενός μοντέλου μηχανικής μάθησης (Lundberg, A unified approach to interpreting model predictions, 2020). Παρέχουν:

- πληροφορίες για τη σημαντικότητα και την επίδραση των προβλεπτικών παραγόντων (features) στην έξοδο του μοντέλου μηχανικής μάθησης.
- πολύτιμες πληροφορίες για την ερμηνεία του μοντέλου μηχανικής μάθησης.

Βασικά χαρακτηριστικά:

- **Σημαντικότητα Χαρακτηριστικών:** Ο άξονας x συνήθως δείχνει τα ονόματα ή τις τιμές των χαρακτηριστικών και η θέση κάθε χαρακτηριστικού στον άξονα υποδεικνύει τη σημαντικότητά του για τις προβλέψεις του μοντέλου (Lundberg, Understanding the local structure of decision trees, 2020). Τα χαρακτηριστικά με μεγαλύτερες θετικές τιμές SHAP συμβάλλουν περισσότερο σε υψηλές εξόδους μοντέλου, ενώ τα χαρακτηριστικά με μεγαλύτερες αρνητικές τιμές SHAP συμβάλλουν περισσότερο σε χαμηλές εξόδους μοντέλου.
- **Κατανομή Επιπτώσεων:** Το χρώμα και το μέγεθος των γραμμών γύρω από τις τιμές των χαρακτηριστικών αντιπροσωπεύουν το μέγεθος και την κατεύθυνση της επίδρασης του χαρακτηριστικού στην έξοδο του μοντέλου (Ribeiro, 2016). Τα πιο σκούρα χρώματα υποδεικνύουν μεγαλύτερες τιμές SHAP (είτε θετικές είτε αρνητικές). Η εξάπλωση της κατανομής χρώματος γύρω από μια τιμή χαρακτηριστικού υποδεικνύει τη μεταβλητότητα της επίδρασης του χαρακτηριστικού στην έξοδο του μοντέλου. Μια κατανομή με μεγάλη διασπορά υποδηλώνει ότι το χαρακτηριστικό μπορεί να έχει διάφορες επιπτώσεις στην έξοδο του μοντέλου ανάλογα με την συγκεκριμένη περίπτωση.
- **Εξήγηση Μοντέλου:** Τα διαγράμματα περίληψης SHAP είναι ένα πολύτιμο εργαλείο για την ερμηνεία μοντέλων μηχανικής μάθησης, καθώς μπορούν να βοηθήσουν στον εντοπισμό των χαρακτηριστικών που είναι πιο σημαντικά για τις προβλέψεις του μοντέλου και στον τρόπο με τον οποίο αυτά τα χαρακτηριστικά αλληλεπιδρούν για να επηρεάσουν την έξοδο του μοντέλου (Molnar, 2022).

2. ΜΕΘΟΔΟΛΟΓΙΑ

2.1. Δεδομένα

Παρασχέθηκε το σύνολο δεδομένων μέσω αρχείου Microsoft Excel με όνομα «Quality of Service 5G», και αφορά σε ένα πρόβλημα σχετιζόμενο με quality of service στο εν λόγω 5G δίκτυο. Περιείχε ένα σύνολο από 400 εγγραφές (ισάριθμο πλήθος χρηστών), και 7 μεταβλητές σχετικές με το πρόβλημα. Με εξαίρεση την πρώτη στήλη-μεταβλητή, που είναι το user ID, οι υπόλοιπες στήλες-μεταβλητές χρησιμοποιήθηκαν στη μοντελοποίηση και υλοποίηση ενός προβλήματος μηχανικής μάθησης.
Αναλυτικά, η δεύτερη (και πρώτη χρήσιμη) μεταβλητή καλείται application type και είναι ποιοτική πολλών επιπέδων (περιέχει το είδος της εκάστοτε εφαρμογής). Οι υπόλοιπες

μεταβλητές είναι όλες ποσοτικές, και ονομάζονται signal strength (dBm), latency (msec), required bandwidth (Mbps), allocated bandwidth (Mbps), και resource allocation (αποτελεί % ποσοστό ή το αντίστοιχό του κλάσμα).

2.2. Προεπεξεργασία δεδομένων

Η ποιοτική μεταβλητή application type μετατράπηκε σε πλήθος από 0-1 dummies, κι όλες οι μεταβλητές ταυτόχρονα έγιναν min-max scaled (normalized [0, 1]), εκτός φυσικά από τη μεταβλητή που είναι target. Οι missing values στα δεδομένα είναι 0.

2.3. Μοντελοποίηση παλινδρόμησης

Η Μηχανική Μάθηση μπορεί να εφαρμοστεί στα δεδομένα μας με σκοπό την πρόβλεψη της καθυστέρησης (Latency) στο δίκτυο 5G με βάση τις διαφορετικές μεταβλητές που διαθέτουμε. Μέσω μηχανικών μοντέλων μπορούμε να εξάγουμε πολύπλοκες σχέσεις από τα δεδομένα και να προβλέψουμε τις τιμές της καθυστέρησης για νέες καταστάσεις.
Το XGBoost (Extreme Gradient Boosting) και το random forest regressor είναι προηγμένες μέθοδοι μηχανικής μάθησης που χρησιμοποιούνται ευρέως για προβλήματα παλινδρόμησης και ταξινόμησης. Αποτελούν αποδοτικούς αλγορίθμους μηχανικής μάθησης και είναι εξαιρετικά δημοφιλείς. Ειδικά ο XGBoost είναι δημοφιλής λόγω της ικανότητάς του να παρέχει υψηλή ακρίβεια και να αντιμετωπίζει αποτελεσματικά την υπερεκπαίδευση.
Για την εφαρμογή τους στα δεδομένα μας, δημιουργήσαμε ένα μοντέλο παλινδρόμησης όπου η μεταβλητή target θα είναι η καθυστέρηση (Latency), ενώ οι υπόλοιπες μεταβλητές θα είναι οι προβλεπόμενοι predictors. Στη συνέχεια εκπαιδεύσαμε το μοντέλο μας χρησιμοποιώντας τον XGBoost regressor και τον random forest regressor, ρυθμίζοντας τα υπερπαραμέτρους του αλγορίθμου με την τεχνική του cross-validation, προκειμένου να επιτύχουμε την καλύτερη δυνατή ακρίβεια και γενικευμένη απόδοση του μοντέλου μας.
Συγκεκριμένα η εκπαίδευση του μοντέλων ακολουθεί τα παρακάτω βήματα και με αυτόν τον τρόπο θα προχωρήσουμε την εργασία μας:

- **Ορισμός των Παραμέτρων Του Μοντέλου**: Αυτό περιλαμβάνει την επιλογή των υπερπαραμέτρων του μοντέλου, όπως οι μέγιστες βάθος των δέντρων αποφάσεων, η ταχύτητα μάθησης (learning rate), και ο αριθμός των δέντρων στο σύνολο (number of trees).

- **Εκπαίδευση του Μοντέλου**: Κατά την εκπαίδευση, το μοντέλο προσαρμόζεται στα δεδομένα εκπαίδευσης με σκοπό να μάθει τη σχέση μεταξύ των προβλεπόμενων παραμέτρων και της μεταβλητής "latency".

- **Προσαρμογή Παραμέτρων Μοντέλου**: Κατά τη διάρκεια της εκπαίδευσης, οι υπερπαράμετροι του μοντέλου μπορεί να προσαρμοστούν προκειμένου να βελτιστοποιηθεί η απόδοσή του.

- **Αξιολόγηση του Μοντέλου**: Μετά την εκπαίδευση, το μοντέλο αξιολογείται χρησιμοποιώντας τα δεδομένα ελέγχου. Αυτό γίνεται για να εκτιμηθεί η απόδοσή του μοντέλου σε ανεξάρτητα δεδομένα.

- **Πρόβλεψη**: Το εκπαιδευμένο μοντέλο χρησιμοποιείται για να προβλέψει τις τιμές της μεταβλητής "latency" για νέα δεδομένα που δεν έχουν χρησιμοποιηθεί κατά την εκπαίδευση.

2.4. Αποτίμηση, απεικόνιση, και γενικά παραδοτέα

Η αποτίμηση των αλγόριθμων έγινε με χρήση γνωστών μετρικών, όπως είναι τα root mean squared error (RMSE) και mean absolute percentage error (MAPE). Τα γραφήματα που παραχθηκαν είναι τα γνωστά SHAP diagrams (importance & scatter plot).. Εγινε μια γενικότερη σύγκριση και συζήτηση των αποτελεσμάτων, και εξάχθηκαν συμπεράσματα.

2.4.1. Διάγραμμα Σημαντικότητας (SHAP summary plot):

- Με αυτό το διάγραμμα αποτυπώθηκε η συνολική επίδραση κάθε χαρακτηριστικού στην έξοδο του μοντέλου.
- Η σειρά των χαρακτηριστικών στον κατακόρυφο άξονα αντικατοπτρίζει τη σημαντικότητά τους, με τα πιο σημαντικά χαρακτηριστικά στην κορυφή.
- Το χρώμα κάθε σημείου δείχνει την τιμή του χαρακτηριστικού (κόκκινο για υψηλές τιμές, μπλε για χαμηλές τιμές).
- Η οριζόντια θέση κάθε σημείου δείχνει την τιμή SHAP, η οποία αντιπροσωπεύει την επίδραση του χαρακτηριστικού στην έξοδο του μοντέλου.

2.4.2. Διάγραμμα Εξάρτησης (SHAP Feature importance):

- Αυτό το διάγραμμα δείχνει πώς η τιμή SHAP ενός χαρακτηριστικού ποικίλλει ανάλογα με την τιμή του ίδιου του χαρακτηριστικού, αλλά και σε συνάρτηση με άλλα χαρακτηριστικά.
- Κάθε γραμμή αντιστοιχεί σε ένα χαρακτηριστικό.
- Η οριζόντια θέση κάθε σημείου δείχνει την τιμή SHAP, ενώ η κατακόρυφη θέση δείχνει την τιμή του χαρακτηριστικού.
- Το χρώμα κάθε σημείου δείχνει την τιμή ενός άλλου χαρακτηριστικού που αλληλεπιδρά.

3. Σύντομη Περιγραφή του Κώδικα

Χρησιμοποιήθηκε η έκδοση 3.12.1 64bit της python

Ο κώδικας Python, που αναπτύχθηκε, αποτελείται από 2 python scripts (main.py και Data Visualization.py) . Το main.py έχει στόχο αρχικά την ανάγνωση δεδομένων από ένα csv αρχείο (QualityOfService5GDataset-3.csv), το οποίο βρίσκεται αποθηκευμένο στο φάκελο (data) . Στη συνέχεια δημιουργεί μια βάση δεδομένων MySQL στην οποία αποθηκεύονται τα αποτελέσματα της εκτέλεσης του XGBoost και του Random Forest πάνω στα δεδομένα του csv αρχείου. Επίσης αποθηκεύει όλα τα παραγόμενα γραφήματα (τόσο της στατιστικής ανάλυσης όσο και το ραβδόγραμμα SHAP Feature importance όσο και το SHAP summary plot) στο φάκελο graphs. Το Data Visualization.py διαβάζει τη βάση δεδομένων MySQL και δημιουργεί 3 νέα γραφήματα για σύγκριση δύο μοντέλων μηχανικής μάθησης, του XGBoost και του Random Forest.

Βιβλιοθήκες:

Ο κώδικας χρησιμοποιεί τις ακόλουθες βασικές βιβλιοθήκες:

- **`os`**: Για χειρισμό αρχείων και φακέλων.
- **`pandas`**: Για επεξεργασία δεδομένων με τη χρήση DataFrames.
- **`numpy`**: Για αριθμητικές πράξεις με πίνακες.
- **`matplotlib.pyplot`**: Για δημιουργία γραφημάτων.

- **`mysql.connector`**: Για σύνδεση και διαχείριση βάσης δεδομένων MySQL.
- **`shap`**: Για ανάλυση SHAP (SHapley Additive exPlanations) και απεικόνιση της σημασίας των χαρακτηριστικών.
- **`json`**: Για μετατροπή δεδομένων σε μορφή JSON.

Λειτουργικότητα:

1. **Σύνδεση με Βάση Δεδομένων:** Ο κώδικας συνδέεται με τη βάση δεδομένων MySQL "5Gpredictivemodeler".
2. **Ανάκτηση Δεδομένων:** Αντλεί δεδομένα από τον πίνακα `model_results` για τα μοντέλα XGBoost και Random Forest, συμπεριλαμβανομένων των δεδομένων των γραφημάτων.
3. **Δημιουργία Γραφημάτων:**
   - **SHAP Bar Plot:** Δημιουργεί ένα συνδυαστικό SHAP bar plot, εμφανίζοντας τη σημασία των χαρακτηριστικών και για τα δύο μοντέλα.
   - **SHAP Summary Plot:** Δημιουργεί ένα συνδυαστικό SHAP summary plot, εμφανίζοντας την κατανομή των τιμών SHAP για κάθε χαρακτηριστικό, ομαδοποιημένα ανά μοντέλο.
   - **Σύγκριση Απόδοσης Μοντέλων:** Δημιουργεί ένα γράφημα σύγκρισης μετρικών απόδοσης (RMSE, MSE, R^2) για τα δύο μοντέλα.
4. **Αποθήκευση Γραφημάτων:** Αποθηκεύει τα γραφήματα στον φάκελο "graphs", με ονόματα αρχείων που περιλαμβάνουν χρονική σήμανση.

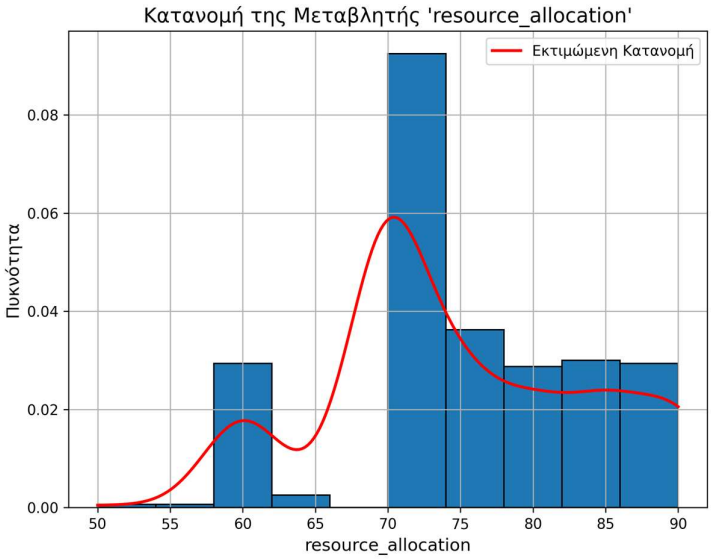
Σκοπός:

Ο κώδικας διευκολύνει την απεικόνιση και σύγκριση δύο μοντέλων μηχανικής μάθησης που έχουν εκπαιδευτεί στο ίδιο σύνολο δεδομένων, αξιοποιώντας τα αποθηκευμένα δεδομένα και τα γραφήματα SHAP.

Αποθετήριο:

https://github.com/biodeveloper/5Gpredictivemodeler

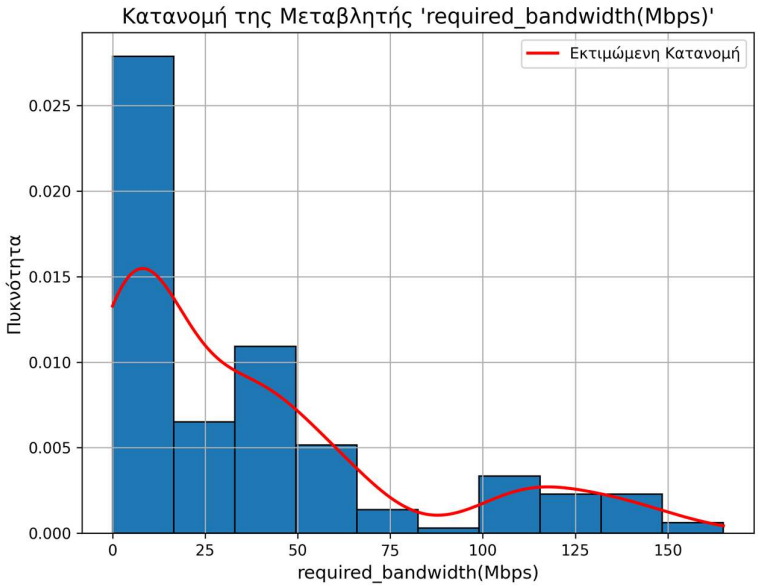
4. Στατιστική ανάλυση των δεδομενων



Εικόνα 1: πυκνότητα vs resource allocation

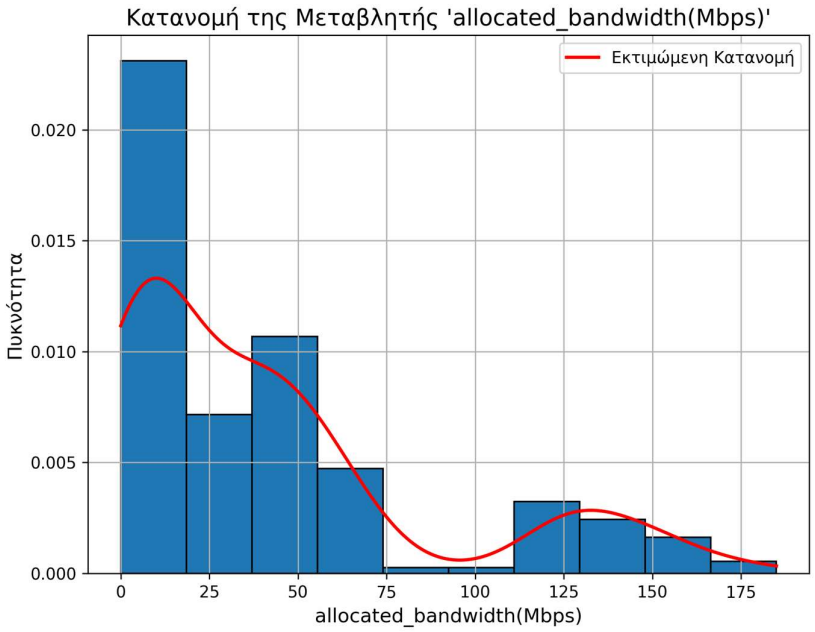
Παρατηρήσεις από τα Ιστογράμματα:

- **resource_allocation (**Εικόνα 1**)**: Η κατανομή είναι ασύμμετρη δεξιά, με την πλειοψηφία των τιμών να συγκεντρώνεται γύρω στο 70%. Αυτό υποδηλώνει ότι οι περισσότεροι χρήστες λαμβάνουν παρόμοια κατανομή πόρων. Η ασύμμετρη δεξιά κατανομή υποδηλώνει ότι λιγότερες περιπτώσεις χρήζουν πολύ υψηλής κατανομής πόρων. Η κεντρική τάση των τιμών στο 70% μπορεί να υποδηλώνει μία κοινή βάση για τις ανάγκες των περισσότερων εφαρμογών. Αυτό μπορεί να είναι χρήσιμο στην πρόβλεψη της latency, ειδικά σε συνδυασμό με τις απαιτήσεις bandwidth.



Εικόνα 2: Πυκνότητα vs required bandwidth (Mbps)

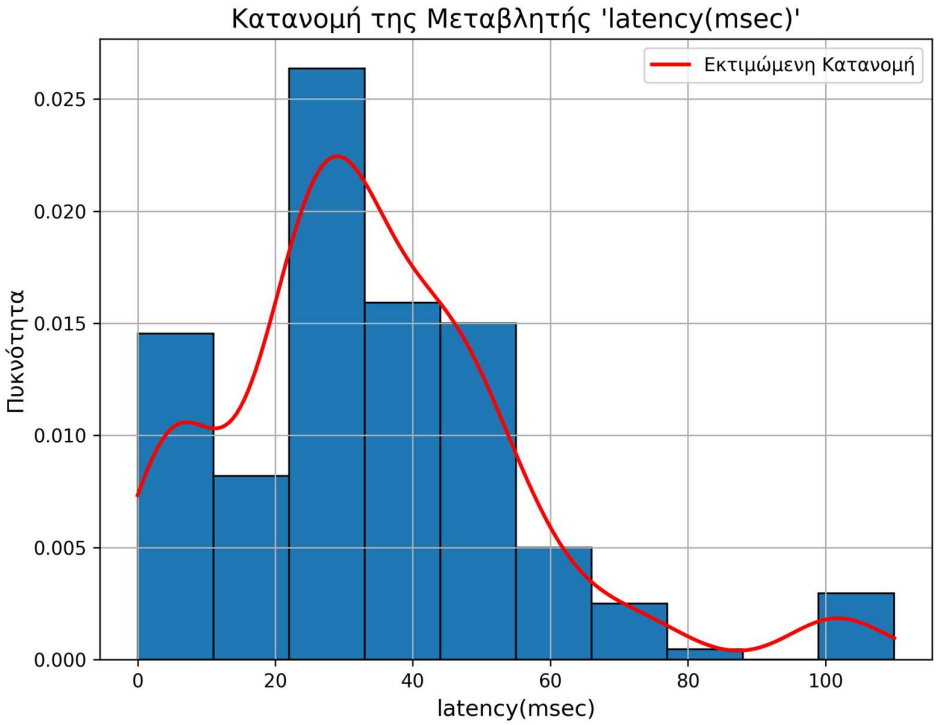
- **required_bandwidth (Mbps) (**Εικόνα 2**)**: Η κατανομή είναι ασύμμετρη δεξιά, με λίγες εφαρμογές να απαιτούν μεγάλο εύρος ζώνης.



Εικόνα 3: Πυκνότητα vs allocated bandwidth (Mbps)

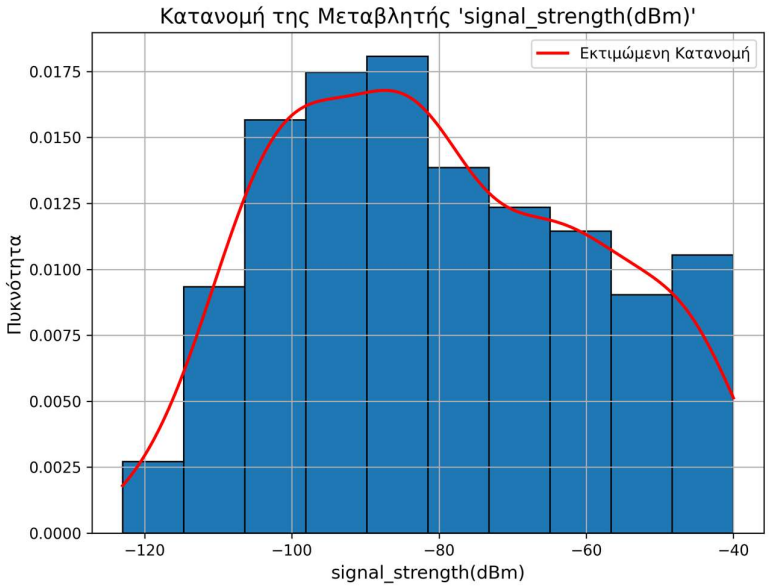
- **allocated_bandwidth (Mbps) (**Εικόνα 3**)**: Η κατανομή είναι παρόμοια με το required_bandwidth, υποδηλώνοντας μια συσχέτιση μεταξύ απαιτούμενου και εκχωρημένου εύρους ζώνης.

  Η ασυμμετρία δεξιά και η παρόμοια κατανομή ανάμεσα σε αυτές τις δύο μεταβλητές υποδηλώνει πως το υψηλότερο required bandwidth μπορεί να οδηγεί στην αντίστοιχη αύξηση του allocated bandwidth. Αυτό θα μπορούσε να έχει απτές επιπτώσεις στην latency.



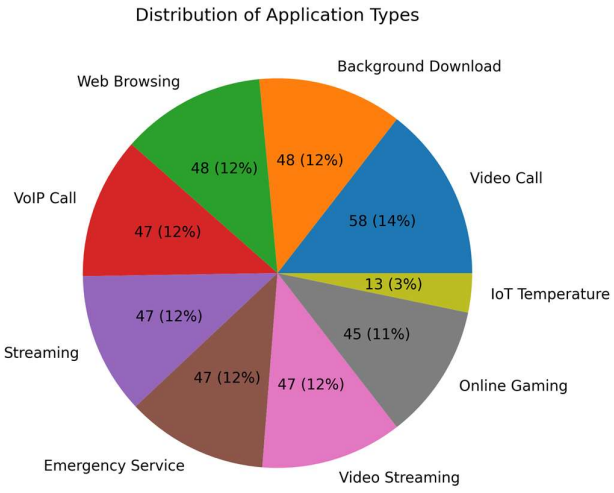
Εικόνα 4 : Πυκνότητα vs Latency (msec)

- **latency (msec) (**Εικόνα 4**)**: Η κατανομή είναι ελαφρώς ασύμμετρη δεξιά, με την πλειοψηφία των τιμών να κυμαίνεται μεταξύ 0 και 40 ms. Υπάρχουν όμως και κάποιες περιπτώσεις με μεγαλύτερη καθυστέρηση. Η κατανομή της latency υποδηλώνει ότι η πλειοψηφία των ερωτημάτων επεξεργάζεται με σχετικά χαμηλή καθυστέρηση, αλλά υπάρχουν και κάποιες ακραίες περιπτώσεις. Αυτές οι ακραίες τιμές θα μπορούσαν να είναι σημαντικές για την πρόβλεψη και μπορεί να απαιτούν ειδική ανάλυση ή μεταχείριση.



Εικόνα 5 : Πυκνότητα vs signal strength (dBm)

- **signal_strength (dBm) (**Εικόνα 5**)**: Η κατανομή μοιάζει κανονική (Gaussian) με τις περισσότερες τιμές να συγκεντρώνονται γύρω στο -80 dBm. Η κανονική κατανομή της ισχύος του σήματος υποδηλώνει ότι οι τιμές του σήματος είναι σχετικά ομοιόμορφα διασπαρμένες. Αυτό προσδίδει ευκολία στην ερμηνεία και μπορεί να είναι ένας σταθερός προβλεπτικός παράγοντας για την καθυστέρηση ή άλλες μεταβλητές.



Εικόνα 6 : Distribution of application types

Παρατηρήσεις από το Γράφημα Πίτας (Εικόνα 6**):**

- Οι τύποι εφαρμογών με τη μεγαλύτερη συχνότητα είναι "Web Browsing", "Background Download" και "Video Call".
- Οι τύποι εφαρμογών "IoT Temperature" και "Emergency Service" έχουν τη χαμηλότερη συχνότητα.
- Η κατανομή των τύπων εφαρμογών μπορεί να δώσει επιπλέον πληροφορίες για την απόδοση δικτύου, καθώς διαφορετικοί τύποι εφαρμογών έχουν διαφορετικές ανάγκες για bandwidth και latency. Για παράδειγμα, οι video calls απαιτούν υψηλότερη ποιότητα σήματος και ευρύτερο bandwidth σε σύγκριση με το web browsing ή τις background downloads.

Η κατανόηση της κατανομής και των συσχετίσεων μεταξύ των μεταβλητών είναι κρίσιμη για την επιλογή και την βελτιστοποίηση του μοντέλου μηχανικής μάθησης. Είναι πιθανό ο τύπος της εφαρμογής, η ισχύς του σήματος και το εκχωρημένο εύρος ζώνης να είναι σημαντικοί προβλεπτικοί παράγοντες για την καθυστέρηση. Περαιτέρω ανάλυση, όπως η εξέταση συσχετίσεων και η εφαρμογή τεχνικών μείωσης διαστάσεων, μπορεί να βελτιώσει την απόδοση του μοντέλου.
Η διεξοδική ανάλυση των μεταβλητών και η κατανόηση των συσχετίσεών τους είναι κρίσιμα σημεία για την επιλογή του κατάλληλου μοντέλου μηχανικής μάθησης. Η εξέταση συσχετίσεων μέσω scatter plots ή correlation matrices και η χρήση τεχνικών μείωσης διαστάσεων όπως PCA μπορεί να βοηθήσει στην αποφυγή πολυπλοκότητας και στη βελτίωση της προβλεπτικής ακρίβειας του τελικού μοντέλου.

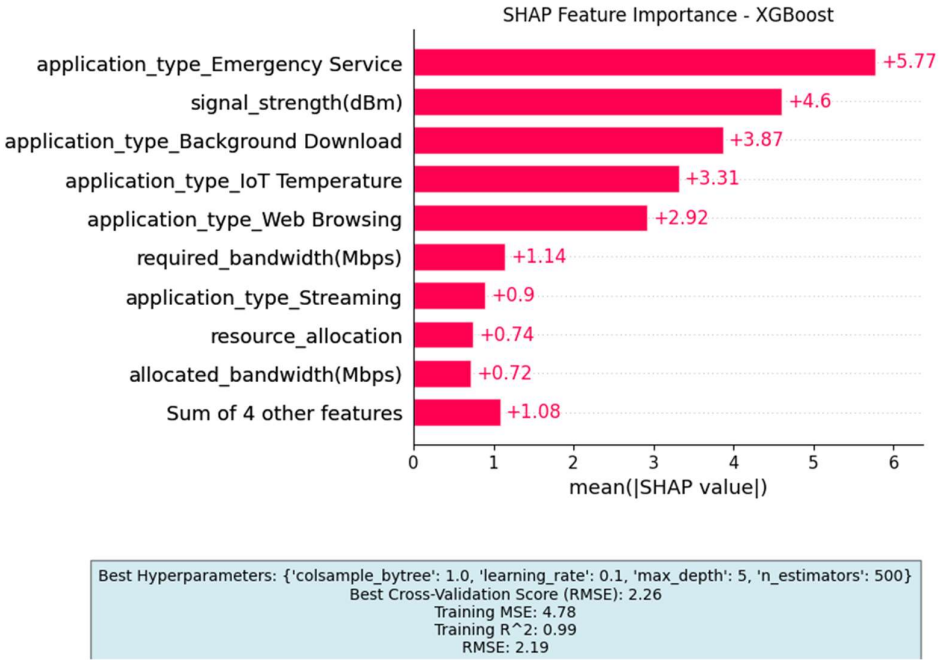
5. XGBoost regressor

5.1. SHAP Feature importance

Τα βασικά εξαγόμενα από το διάγραμμα (Εικόνα 7) είναι:

5.1.1. Προτεραιότητα Ισχυρής Ισχύος Σήματος:

Ένα ξεκάθαρο συμπέρασμα από το διάγραμμα είναι η σημαντική έμφαση που δίνεται στο "signal_strength(dBm)". Αυτό το χαρακτηριστικό έχει την υψηλότερη θετική τιμή SHAP, υποδεικνύοντας ότι το XGBoost δίνει προτεραιότητα σε μια ισχυρή ισχύ σήματος κατά τη διάθεση εύρους ζώνης. Αυτή η προτεραιοποίηση είναι λογική, καθώς ένα ισχυρό σήμα είναι απαραίτητο για αξιόπιστη μετάδοση δεδομένων. Εφαρμογές που απαιτούν αδιάκοπη ροή δεδομένων, όπως η τηλεδιάσκεψη ή τα online παιχνίδια, θα ωφεληθούν από αυτό.



Εικόνα 7: SHAP feature importance for XGBOOST

5.1.2. Δικαιοσύνη στη Διάθεση Εύρους Ζώνης:

Μια άλλη ενδιαφέρουσα πληροφορία προέρχεται από την αρνητική επίδραση του "allocated_bandwidth(Mbps)". Αυτό υποδηλώνει ότι το XGBoost μπορεί να ευνοήσει μια πιο δίκαιη διάθεση εύρους ζώνης. Με άλλα λόγια, οι εφαρμογές που έχουν ήδη λάβει σημαντική ποσότητα εύρους ζώνης είναι λιγότερο πιθανό να λάβουν ακόμη περισσότερο. Αυτό μπορεί να βοηθήσει στη διασφάλιση ενός βασικού επιπέδου υπηρεσίας για όλους τους χρήστες και να αποτρέψει τις εφαρμογές από το να μονοπωλούν πόρους εύρους ζώνης.

5.1.3. Οι Υπηρεσίες Έκτακτης Ανάγκης και οι Λήψεις Παρασκηνίου Έχουν Προτεραιότητα:

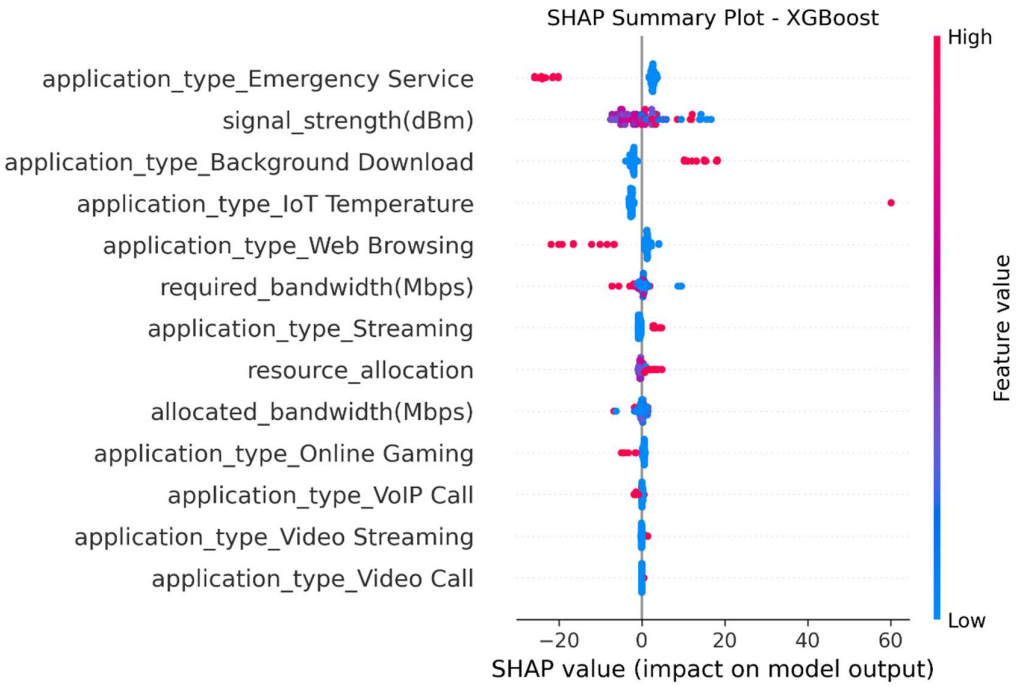
Οι θετικές τιμές SHAP για "application_type_Emergency Service" και "application_type_Background Download" επισημαίνουν την προτεραιοποίηση του μοντέλου σε αυτούς τους τύπους εφαρμογών. Οι υπηρεσίες έκτακτης ανάγκης λαμβάνουν τον υψηλότερο θετικό αντίκτυπο μετά την ισχύ σήματος, αντικατοπτρίζοντας την κρίσιμη φύση τους για τη δημόσια ασφάλεια και την έγκαιρη ανταπόκριση. Οι λήψεις παρασκηνίου, οι οποίες συνήθως δεν απαιτούν άμεση ανταπόκριση σε πραγματικό χρόνο αλλά ωφελούνται από τις αδιάκοπες συνδέσεις, ευνοούνται επίσης από το μοντέλο.

5.1.4. Περιήγηση Ιστού και Απαιτούμενο Εύρος Ζώνης:

Οι αρνητικές τιμές SHAP για "application_type_Web Browsing" και "required_bandwidth(Mbps)" υποδηλώνουν ότι οι εφαρμογές περιήγησης στο διαδίκτυο και εκείνες που απαιτούν χαμηλότερο εύρος ζώνης λαμβάνουν λιγότερο εύρος ζώνης από το μοντέλο XGBoost. Αυτή η προτεραιοποίηση μπορεί να αποδοθεί στο γεγονός ότι η περιήγηση στο διαδίκτυο συχνά ανέχεται μικρές

καθυστερήσεις ή buffering, και οι χρήστες μπορεί να έχουν διαφορετικές απαιτήσεις εύρους ζώνης ανάλογα με το περιεχόμενο της ιστοσελίδας.

5.2. SHAP summary plot



Best Hyperparameters: {'colsample_bytree': 1.0, 'learning_rate': 0.1, 'max_depth': 5, 'n_estimators': 500}
Best Cross-Validation Score (RMSE): 2.26
Training MSE: 4.78
Training R^2: 0.99
RMSE: 2.19

Εικόνα 8 : SHAP summary plot for XGBoost

Το διάγραμμα (Εικόνα 8) είναι ένα SHAP summary plot για το μοντέλο XGBoost. Σε αυτήν την περίπτωση, το μοντέλο προσπαθεί να προβλέψει το πόσο εύρος ζώνης θα διατεθεί (allocated_bandwidth), με βάση διάφορα χαρακτηριστικά, όπως ο τύπος της εφαρμογής που χρησιμοποιείται (application_type) και η ένταση του σήματος (signal_strength).

Τα βασικά εξαγόμενα από το διάγραμμα είναι:

- **Οι υπηρεσίες έκτακτης ανάγκης έχουν την υψηλότερη προτεραιότητα:** Το χαρακτηριστικό με τη μεγαλύτερη θετική επίδραση στην έξοδο του μοντέλου είναι "application_type_Emergency Service". Αυτό σημαίνει ότι το μοντέλο προβλέπει ότι στις εφαρμογές που σχετίζονται με τις υπηρεσίες έκτακτης ανάγκης θα διατεθεί το μεγαλύτερο εύρος ζώνης.
- **Η περιήγηση στο διαδίκτυο έχει την χαμηλότερη προτεραιότητα:** Αντίστροφα, το "application_type_Web Browsing" έχει τη μεγαλύτερη αρνητική επίδραση στην έξοδο του

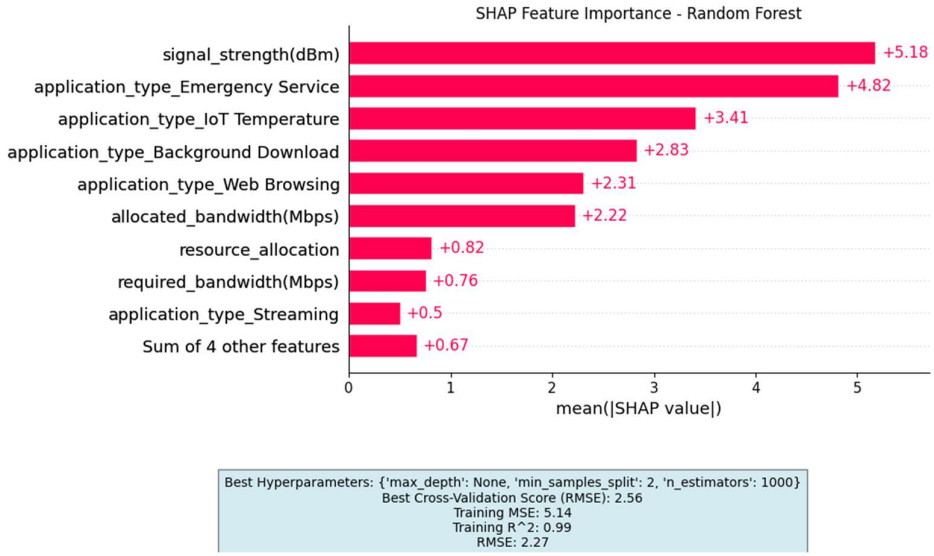
μοντέλου. Αυτό σημαίνει ότι το μοντέλο προβλέπει ότι στις εφαρμογές περιήγησης στο διαδίκτυο θα διατεθεί το λιγότερο εύρος ζώνης.

Τα υπόλοιπα χαρακτηριστικά βρίσκονται μεταξύ αυτών των δύο άκρων. Για παράδειγμα, τα "application_type_Background Download" και "application_type_IoT Temperature" έχουν θετική επίδραση στην έξοδο του μοντέλου, αλλά όχι τόσο όσο το "application_type_Emergency Service". Από την άλλη πλευρά, τα "application_type_Online Gaming" και "application_type_VoIP Call" έχουν αρνητική επίδραση στην έξοδο του μοντέλου, αλλά όχι τόσο όσο το "application_type_Web Browsing".

Είναι σημαντικό να σημειωθεί ότι αυτή είναι απλώς μια πρόβλεψη του μοντέλου και η πραγματική κατανομή εύρους ζώνης μπορεί να διαφέρει ανάλογα με άλλους παράγοντες. Ωστόσο, αυτό το διάγραμμα μας δίνει μια καλή ιδέα για το πώς το μοντέλο σταθμίζει τα διαφορετικά χαρακτηριστικά όταν κάνει τις προβλέψεις του.

1. Random forest

1.1. SHAP Feature importance



Εικόνα 9: SHAP feature importance for Random Forest

Αυτό το γράφημα SHAP (Εικόνα 9) οπτικοποιεί τον αντίκτυπο διαφόρων χαρακτηριστικών στις προβλέψεις που κάνει το μοντέλο Random Forest.

1.1.1. Κυρίαρχα Χαρακτηριστικά:

- **`signal_strength(dBm)` (Ισχύς Σήματος):** Αυτό το χαρακτηριστικό ξεχωρίζει ως το πιο σημαντικό, με μέση απόλυτη τιμή SHAP να υπερβαίνει το 5. Αυτό δείχνει ότι η ισχύς του

σήματος στο δίκτυο 5G έχει ισχυρή θετική σχέση με την πρόβλεψη του μοντέλου (πιθανώς την καθυστέρηση). Ένα ισχυρότερο σήμα συνδέεται συνήθως με χαμηλότερη καθυστέρηση.

- **`application_type_Emergency Service` (Εφαρμογή Υπηρεσίας Έκτακτης Ανάγκης):** Αυτό το χαρακτηριστικό, που αντιπροσωπεύει εφαρμογές έκτακτης ανάγκης, έχει τη δεύτερη μεγαλύτερη επίδραση. Η θετική τιμή SHAP (περίπου 4,8) δείχνει ότι οι εφαρμογές έκτακτης ανάγκης είναι πιθανότερο να αντιμετωπίσουν υψηλότερη καθυστέρηση σε σύγκριση με άλλους τύπους εφαρμογών. Αυτό είναι λογικό, καθώς αυτές οι εφαρμογές συχνά απαιτούν προτεραιοποίηση και μπορεί να χρησιμοποιούν διαφορετικούς πόρους δικτύου.
- **`application_type_IoT Temperature` (Εφαρμογή Παρακολούθησης Θερμοκρασίας IoT):** Εφαρμογές που σχετίζονται με την παρακολούθηση θερμοκρασίας μέσω IoT εμφανίζουν επίσης αξιοσημείωτη επίδραση στην καθυστέρηση, με μέση τιμή SHAP περίπου 3,4. Η θετική επίδραση υποδηλώνει ότι αυτές οι εφαρμογές, ίσως λόγω των συχνών μεταδόσεων δεδομένων, τείνουν να συμβάλλουν σε υψηλότερη καθυστέρηση.

1.1.2. Χαρακτηριστικά Μέτριας Επίδρασης:

application_type_Background Download (Λήψη στο Φόντο) και application_type_Web Browsing (Περιήγηση στο Web): Αυτοί οι τύποι εφαρμογών έχουν μέτρια και σχετικά παρόμοια επιρροή στην καθυστέρηση, με τιμές SHAP περίπου 2,8 και 2,3, αντίστοιχα. Αυτό υποδηλώνει ότι οι λήψεις στο φόντο και η περιήγηση στο web έχουν αισθητή αλλά λιγότερο έντονη επίδραση σε σύγκριση με τα τρία κορυφαία χαρακτηριστικά.

1.1.3. Χαρακτηριστικά με Περιορισμένη Επίδραση:

allocated_bandwidth(Mbps) (Εκχωρημένο Εύρος Ζώνης), resource_allocation (Κατανομή Πόρων), required_bandwidth(Mbps) (Απαιτούμενο Εύρος Ζώνης), application_type_Streaming (Ροή), και other features (Άλλα Χαρακτηριστικά): Αυτά τα χαρακτηριστικά έχουν σχετικά μικρές τιμές SHAP, υποδεικνύοντας περιορισμένη συμβολή στις προβλέψεις του μοντέλου. Ενώ μπορεί να παίζουν κάποιο ρόλο, η επίδρασή τους στην καθυστέρηση είναι λιγότερο σημαντική σε σύγκριση με τα κυρίαρχα χαρακτηριστικά.

1.1.4. Αλληλεπιδράσεις Χαρακτηριστικών και Πολυπλοκότητα:

Είναι σημαντικό να θυμόμαστε ότι οι τιμές SHAP αντιπροσωπεύουν τη **μέση** επίδραση ενός χαρακτηριστικού. Η πραγματική επιρροή ενός χαρακτηριστικού μπορεί να ποικίλλει ανάλογα με τις τιμές άλλων χαρακτηριστικών. Το γράφημα δεν αποκαλύπτει ρητώς πολύπλοκες αλληλεπιδράσεις μεταξύ χαρακτηριστικών, οι οποίες μπορεί να υπάρχουν.

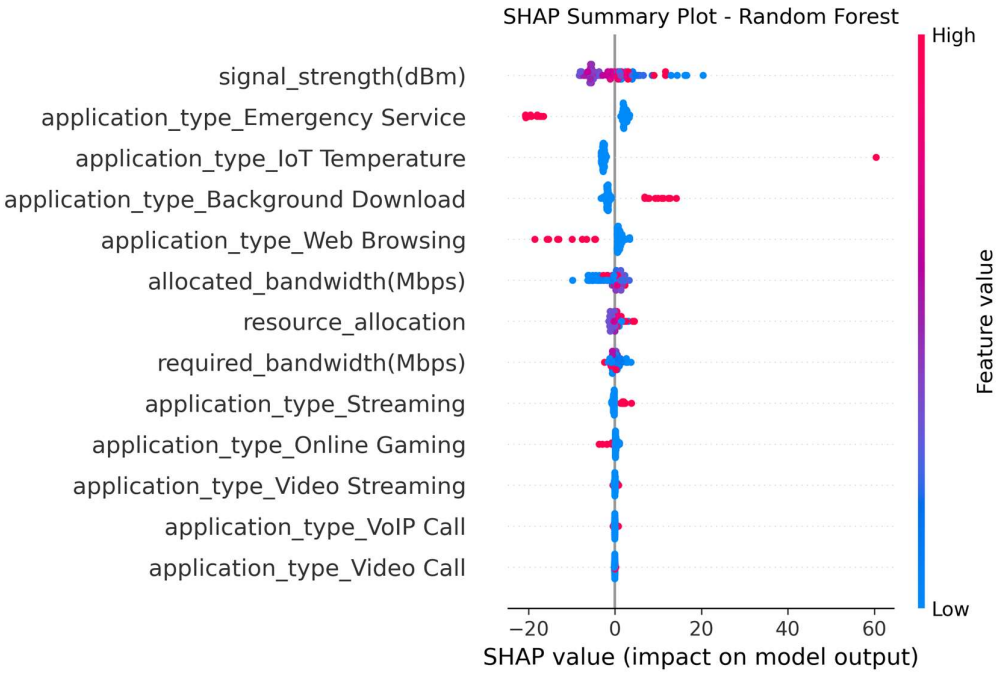
1.1.5. Απόδοση Μοντέλου και Γενίκευση:

Το μοντέλο εμφανίζει ισχυρή απόδοση τόσο στα δεδομένα εκπαίδευσης όσο και στα δεδομένα δοκιμής, όπως αποδεικνύεται από την υψηλή τιμή R^2 (0,99) και τις χαμηλές τιμές RMSE (2,56 για διασταυρούμενη επικύρωση και 2,27 για τα δεδομένα δοκιμής). Αυτό υποδηλώνει ότι το μοντέλο γενικεύει καλά σε αόρατα δεδομένα και δεν υπερπροσαρμόζεται στα δεδομένα εκπαίδευσης.

Συνοπτικά, το γράφημα σημασίας χαρακτηριστικών SHAP υπογραμμίζει ότι η ισχύς του σήματος και ο τύπος εφαρμογής παίζουν τους πιο σημαντικούς ρόλους στον καθορισμό της καθυστέρησης, ενώ άλλοι παράγοντες όπως το εύρος ζώνης και η κατανομή πόρων έχουν μικρότερη επίδραση. Το

γράφημα υποδηλώνει επίσης ότι το μοντέλο Random Forest είναι αποτελεσματικό στην πρόβλεψη της καθυστέρησης σε αυτό το σενάριο δικτύου 5G.

1.2. SHAP summary plot



Best Hyperparameters: {'max_depth': None, 'min_samples_split': 2, 'n_estimators': 1000}
Best Cross-Validation Score (RMSE): 2.56
Training MSE: 5.14
Training R^2: 0.99
RMSE: 2.27

Εικόνα 10 : SHAP summary plot for Random Forest

Βασικά ευρήματα:

- **Η ισχύς σήματος είναι το πιο σημαντικό χαρακτηριστικό:** Το χαρακτηριστικό με τη μεγαλύτερη θετική επίδραση στην έξοδο του μοντέλου είναι "signal_strength(dBm)". Αυτό σημαίνει ότι το μοντέλο προβλέπει ότι μια ισχυρότερη ισχύς σήματος θα οδηγήσει σε μεγαλύτερη διάθεση εύρους ζώνης.
- **Οι υπηρεσίες έκτακτης ανάγκης και οι λήψεις παρασκηνίου έχουν προτεραιότητα:** Τα επόμενα πιο σημαντικά χαρακτηριστικά με θετικές επιπτώσεις είναι "application_type_Emergency Service" και "application_type_Background Download". Αυτό σημαίνει ότι το μοντέλο προβλέπει ότι στις εφαρμογές που σχετίζονται με τις υπηρεσίες έκτακτης ανάγκης και τις λήψεις παρασκηνίου θα διατεθεί περισσότερο εύρος ζώνης.
- **Πολλοί παράγοντες μειώνουν τη διάθεση εύρους ζώνης:** Υπάρχουν ορισμένα χαρακτηριστικά που έχουν αρνητικές επιπτώσεις στην έξοδο του μοντέλου, πράγμα που σημαίνει ότι μειώνουν την προβλεπόμενη διάθεση εύρους ζώνης. Αυτά περιλαμβάνουν "application_type_Web Browsing", "required_bandwidth(Mbps)" και "allocated_bandwidth(Mbps)". Είναι ενδιαφέρον ότι το "allocated_bandwidth" έχει αρνητική επίδραση, γεγονός που υποδηλώνει ότι το μοντέλο μπορεί να δίνει προτεραιότητα στη

δικαιοσύνη στη διάθεση εύρους ζώνης - οι εφαρμογές που έχουν ήδη λάβει εύρος ζώνης είναι λιγότερο πιθανό να λάβουν περισσότερο.

Συνολικά, το διάγραμμα υποδηλώνει ότι το μοντέλο προσπαθεί να βρει ισορροπία μεταξύ διαφόρων παραγόντων κατά τη διάθεση εύρους ζώνης. Η ισχύς σήματος είναι ο πιο σημαντικός παράγοντας, αλλά ο τύπος της εφαρμογής που χρησιμοποιείται και το ποσό εύρους ζώνης που έχει ήδη διατεθεί είναι επίσης σημαντικές παράμετροι.

Εδώ είναι μερικές πρόσθετες λεπτομέρειες από το διάγραμμα:

- Ο άξονας x απεικονίζει την τιμή SHAP για κάθε χαρακτηριστικό. Οι τιμές SHAP εξηγούν την επίδραση ενός χαρακτηριστικού σε μια συγκεκριμένη πρόβλεψη. Μια θετική τιμή SHAP σημαίνει ότι το χαρακτηριστικό αυξάνει την πρόβλεψη, ενώ μια αρνητική τιμή SHAP σημαίνει ότι το χαρακτηριστικό μειώνει την πρόβλεψη.
- Ο άξονας y απεικονίζει το όνομα του χαρακτηριστικού.
- Το χρώμα της γραμμής δείχνει το μέγεθος της τιμής SHAP. Τα πιο σκούρα χρώματα υποδεικνύουν μεγαλύτερες τιμές SHAP (είτε θετικές είτε αρνητικές).
- Η κάθετη γραμμή στο κέντρο της κατανομής δείχνει τη μέση τιμή SHAP για κάθε χαρακτηριστικό.

2. Ανάλυση Αποτελεσμάτων:

2.1. Σημαντικότητα Μεταβλητών

2.1.1. Οι πιο σημαντικές μεταβλητές

Τα SHAP diagrams των 2 μοντέλων καθώς και το συνδυαστικά διαγράμματα αυτών (**Σφάλμα! Το αρχείο προέλευσης της αναφοράς δεν βρέθηκε.**και Εικόνα 12) δείχνουν ότι οι πιο σημαντικές μεταβλητές για την πρόβλεψη του latency στο δεδομένο δίκτυο 5G σύμφωνα και με τα 2 μοντέλα είναι:

"signal_strength(dBm)":

- Αποτελεί τον πιο σημαντικό predictor και για τα δύο μοντέλα (XGBoost και Random Forest).
- Αυτό είναι λογικό, καθώς η ισχύς του σήματος έχει άμεση επίδραση στην ποιότητα της σύνδεσης και κατ' επέκταση στην latency.

"application_type_Emergency Service":

- Είναι η δεύτερη πιο σημαντική μεταβλητή, δείχνοντας ότι οι εφαρμογές έκτακτης ανάγκης έχουν ιδιαίτερες απαιτήσεις latency.
- Η υψηλή σημασία της υποδηλώνει ότι η latency είναι κρίσιμη σε αυτές τις εφαρμογές και πιθανώς επηρεάζεται από άλλους παράγοντες (π.χ., διαθέσιμο εύρος ζώνης).

"application_type_IoT Temperature":

- Ενδιαφέρον είναι ότι η "IoT Temperature" κατατάσσεται ως σημαντική μεταβλητή, υποδεικνύοντας την αυξανόμενη σημασία των IoT εφαρμογών.
- Πιθανώς, η latency σε αυτές τις εφαρμογές έχει σημαντικό αντίκτυπο στην ακρίβεια των μετρήσεων και στην σύνολο της λειτουργίας του συστήματος.

"application_type_Background Download":

- Η μεταβλητή "Background Download" βρίσκεται επίσης στο top 5, ίσως λόγω της απαιτούμενης ποσότητας δεδομένων.

- Ενδεχομένως, η latency επηρεάζεται από τον ρυθμό μεταφοράς δεδομένων, ο οποίος θα μπορούσε να εξαρτάται από το επίπεδο συνωστισμού στο δίκτυο.

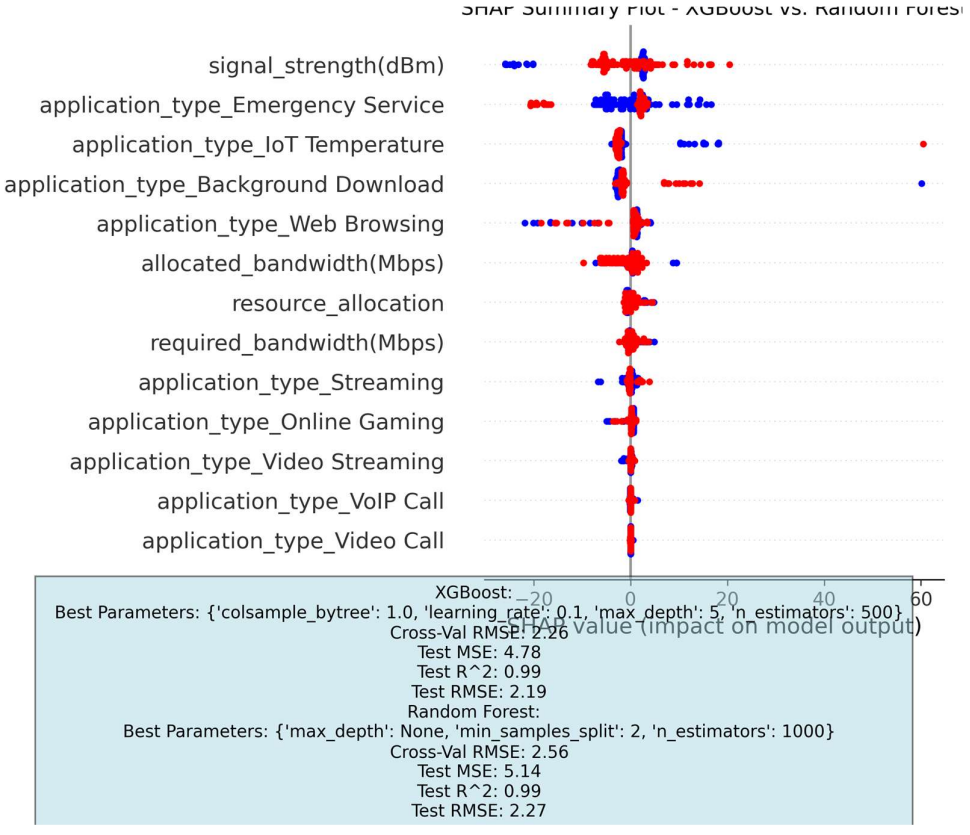
"application_type_Web Browsing":

- Η "Web Browsing" κατατάσσεται μεταξύ των πέντε πιο σημαντικών μεταβλητών, όπως είναι λογικό, δεδομένου ότι οι webbrowsing εφαρμογές έχουν υψηλές απαιτήσεις latency για ομαλή λειτουργία.

Σημαντικές Παρατηρήσεις:

- Η "signal_strength(dBm)" είναι κατά πολύ η πιο σημαντική μεταβλητή, υπογραμμίζοντας την σημασία της ποιότητας του σήματος για το latency.
- Οι εφαρμογές εκτάκτης ανάγκης και οι IoT εφαρμογές έδειξαν επίσης σημαντική επίδραση στο latency, υποδεικνύοντας την ανάγκη για ειδικές οπτιμοποιήσεις σε αυτά τα σενάρια.
- Η "allocated_bandwidth(Mbps)" και η "resource_allocation" έδειξαν χαμηλότερη σημαντικότητα, υποδεικνύοντας ότι η διαθέσιμη bandwidth δεν είναι ο πιο κρίσιμος παράγοντας για το latency, τουλάχιστον σε αυτό το σύνολο δεδομένων.
- Η "application_type_Background Download" και η "application_type_Web Browsing" επίσης επηρεάζουν την latency, πιθανώς λόγω της ποσότητας δεδομένων που μεταφέρονται.

2.1.2. Σύγκριση shap_summary_plot_XGBoost vs shap_summary_plot_Random Forest



Εικόνα 11: SHAP summary plot comparing XGBoost vs Random Forest

Συγκρίση διαγράμματων SHAP summary plot για XGBoost και Random Forest (Εικόνα 11):

2.1.2.1. Κύριες Διαφορές στη Σημαντικότητα Χαρακτηριστικών:

- Ισχύς Σήματος: Το διάγραμμα XGBoost δείχνει σαφώς μεγαλύτερη έμφαση στο "signal_strength(dBm)" σε σύγκριση με το διάγραμμα Random Forest. Η σημαντικότητα του χαρακτηριστικού ισχύος σήματος στο XGBoost είναι σχεδόν διπλάσια από αυτή του Random Forest, υποδεικνύοντας ότι το XGBoost δίνει πολύ μεγαλύτερη προτεραιότητα σε μια ισχυρή ισχύ σήματος κατά τη διάθεση εύρους ζώνης.
- Τύποι Εφαρμογών: Υπάρχουν ορισμένες ενδιαφέρουσες διαφορές στον τρόπο που σταθμίζονται οι τύποι εφαρμογών μεταξύ των μοντέλων. Στο XGBoost, το "application_type_Emergency Service" έχει τον υψηλότερο θετικό αντίκτυπο, ακολουθούμενο από τα "application_type_Background Download" και "application_type_loT Temperature". Στο Random Forest, το "signal_strength(dBm)" είναι ο πιο σημαντικός παράγοντας, ακολουθούμενο από τα "application_type_Emergency Service" και "application_type_Background Download". Αυτό υποδηλώνει ότι ενώ και τα δύο μοντέλα δίνουν προτεραιότητα στις υπηρεσίες έκτακτης ανάγκης και τις λήψεις παρασκηνίου, το XGBoost δίνει μεγαλύτερη έμφαση στις λήψεις παρασκηνίου και τις εφαρμογές θερμοκρασίας loT σε σύγκριση με το Random Forest.
- Διατεθέν Εύρος Ζώνης:** Η επίδραση του "allocated_bandwidth(Mbps)" είναι αρνητική και στα δύο διαγράμματα, αλλά το αποτέλεσμα φαίνεται πιο έντονο στο XGBoost. Αυτό υποδηλώνει ότι το XGBoost μπορεί να είναι πιο πιθανό να ευνοήσει τη δίκαιη διάθεση εύρους ζώνης, δίνοντας λιγότερο πρόσθετο εύρος ζώνης σε εφαρμογές που έχουν ήδη λάβει σημαντική ποσότητα.

2.1.2.2. Ομοιότητες στη Σημαντικότητα Χαρακτηριστικών:

- Περιήγηση και Απαιτούμενο Εύρος Ζώνης: Και τα δύο μοντέλα δείχνουν αρνητικές επιπτώσεις για "application_type_Web Browsing" και "required_bandwidth(Mbps)". Αυτό σημαίνει ότι και τα δύο μοντέλα είναι πιθανό να διαθέσουν λιγότερο εύρος ζώνης σε εφαρμογές περιήγησης στο διαδίκτυο και σε εφαρμογές που απαιτούν γενικά χαμηλότερο εύρος ζώνης.
- Άλλες Εφαρμογές: Και τα δύο διαγράμματα δείχνουν ένα μείγμα θετικών και αρνητικών επιπτώσεων για διάφορους άλλους τύπους εφαρμογών, υποδηλώνοντας ότι τα μοντέλα λαμβάνουν υπόψη τον συγκεκριμένο τύπο εφαρμογής κατά τη διάθεση εύρους ζώνης.

2.1.2.3. Γενικά

- Προτεραιοποίηση: Το XGBoost δίνει μεγαλύτερη προτεραιότητα σε μια ισχυρή ισχύ σήματος από το Random Forest, ενώ και τα δύο μοντέλα δίνουν προτεραιότητα στις υπηρεσίες έκτακτης ανάγκης και τις λήψεις παρασκηνίου για τη διάθεση εύρους ζώνης.
- Fairness/Αντικειμενικότητα: Το XGBoost μπορεί να ευνοήσει τη δικαιότερη διάθεση εύρους ζώνης σε σύγκριση με το Random Forest.
- Διακυμάνσεις Χαρακτηριστικών: Το XGBoost μπορεί να συλλάβει πιο λεπτές λεπτομέρειες σχετικά με τη σημαντικότητα των χαρακτηριστικών, με πολλά χαρακτηριστικά να έχουν μέτρια επίδραση. Το Random Forest μπορεί να εντοπίσει έναν μικρότερο αριθμό βασικών χαρακτηριστικών με μεγαλύτερη επίδραση.

2.1.3. Σύγκριση shap_bar_plot_Random Forest vs shap_bar_plot_XGBoost

Σύγκριση ραβδογράμματος SHAP για τον XGBoost με το ραβδόγραμμα SHAP για τον Random Forest (Εικόνα 12):

2.1.3.1. Βασικές Διαφορές

- Προτεραιοποίηση της Ισχύος Σήματος: Το διάγραμμα XGBoost δίνει σημαντικά μεγαλύτερη βαρύτητα στο "signal_strength(dBm)" σε σύγκριση με το διάγραμμα Random Forest. Αυτό δείχνει ότι το XGBoost δίνει πολύ μεγαλύτερη προτεραιότητα σε μια ισχυρή ισχύ σήματος κατά τη διάθεση εύρους ζώνης.
- Κατανομή Σημαντικότητας Χαρακτηριστικών: Οι βαθμοί σημαντικότητας χαρακτηριστικών στο μοντέλο XGBoost φαίνεται να είναι πιο διασκορπισμένοι, με πολλά χαρακτηριστικά να έχουν αισθητή επίδραση. Αντίθετα, το μοντέλο Random Forest φαίνεται να έχει μικρότερο αριθμό κυρίαρχων χαρακτηριστικών, με πιο απότομη πτώση στη σημαντικότητα για τα υπόλοιπα χαρακτηριστικά.

2.1.3.2. Ομοιότητες:

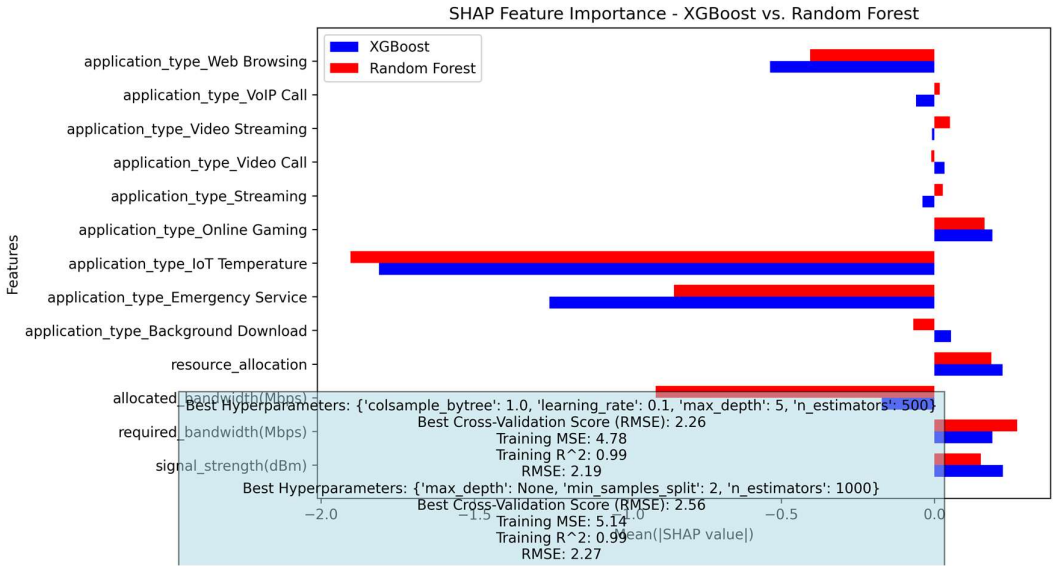
- Σημαντικότητα του Τύπου Εφαρμογής: Και τα δύο διαγράμματα δείχνουν ότι ο τύπος της εφαρμογής παίζει ρόλο στη διάθεση εύρους ζώνης. Και στα XGBoost και Random Forest, τα χαρακτηριστικά που σχετίζονται με τις υπηρεσίες έκτακτης ανάγκης ("application_type_Emergency Service") και τις λήψεις παρασκηνίου ("application_type_Background Download") έχουν θετικές επιπτώσεις, υποδηλώνοντας ότι αυτές οι εφαρμογές λαμβάνουν περισσότερο εύρος ζώνης.
- Αρνητική Επίδραση της Περιήγησης και του Απαιτούμενου Ε εύρους ζώνης: Και τα δύο μοντέλα αποδίδουν αρνητική σημαντικότητα στο "application_type_Web Browsing" και "required_bandwidth(Mbps)". Αυτό υποδηλώνει ότι στις εφαρμογές περιήγησης στο διαδίκτυο διατίθεται λιγότερο εύρος ζώνης και ότι οι εφαρμογές που απαιτούν λιγότερο εύρος ζώνης λαμβάνουν συνολικά λιγότερο.

2.1.3.3. Γενικά

- Το XGBoost δίνει προτεραιότητα σε μια ισχυρή ισχύ σήματος** για τη διάθεση εύρους ζώνης, ενώ το Random Forest φαίνεται λιγότερο εστιασμένο σε αυτόν τον συγκεκριμένο παράγοντα.
- Το XGBoost μπορεί να συλλάβει πιο λεπτές λεπτομέρειες σχετικά με τη σημαντικότητα των χαρακτηριστικών**, με πολλά χαρακτηριστικά να έχουν μέτρια επίδραση. Το Random Forest μπορεί να εντοπίσει έναν μικρότερο αριθμό βασικών χαρακτηριστικών με μεγαλύτερη επίδραση.
- Και τα δύο μοντέλα λαμβάνουν υπόψη τον τύπο εφαρμογής και το απαιτούμενο εύρος ζώνης** κατά τη διάθεση εύρους ζώνης.

2.1.3.4. Επιπρόσθετες επισημάνσεις:

Είναι σημαντικό να θυμόμαστε ότι αυτές είναι απλώς τιμές SHAP και η πραγματική επίδραση κάθε χαρακτηριστικού στη διάθεση εύρους ζώνης μπορεί να είναι πολύπλοκη.
Οι συγκεκριμένες υπερπαράμετροι που χρησιμοποιούνται για την εκπαίδευση των μοντέλων XGBoost και Random Forest μπορούν επίσης να επηρεάσουν τη σημαντικότητα των χαρακτηριστικών.



Εικόνα 12 : SHAP Feature Importance comparing XGBoost vs Random Forest

2.2. Επίδοση Μοντέλων

2.2.1. Prioritization of Strong Signal Strength:

Το πιο αξιοσημείωτο χαρακτηριστικό στο διάγραμμα είναι η κυριαρχία του "signal_strength(dBm)" με την υψηλότερη θετική τιμή SHAP. Αυτό δείχνει ότι το XGBoost δίνει προτεραιότητα σε μια ισχυρή ισχύ σήματος κατά τη διάθεση εύρους ζώνης. Αυτή η προτεραιοποίηση είναι λογική, καθώς η ισχυρή ισχύς σήματος είναι απαραίτητη για αξιόπιστη και υψηλής ποιότητας μετάδοση δεδομένων. Εφαρμογές που απαιτούν αδιάκοπη ροή δεδομένων, όπως η τηλεδιάσκεψη ή τα online παιχνίδια, θα ωφεληθούν από αυτήν την προτεραιοποίηση.

2.2.2. Fairness in Bandwidth Allocation:

Μια άλλη ενδιαφέρουσα πληροφορία προέρχεται από την αρνητική επίδραση του "allocated_bandwidth(Mbps)". Αυτό υποδηλώνει ότι το XGBoost μπορεί να ευνοήσει μια πιο δίκαιη διάθεση εύρους ζώνης. Με άλλα λόγια, οι εφαρμογές που έχουν ήδη λάβει εύρος ζώνης είναι λιγότερο πιθανό να λάβουν πρόσθετο εύρος ζώνης. Αυτό μπορεί να βοηθήσει στη διασφάλιση του ότι όλοι οι χρήστες έχουν ένα βασικό επίπεδο υπηρεσίας και μπορεί να αποτρέψει τις εφαρμογές από το να μονοπωλούν πόρους εύρους ζώνης.

2.2.3. Οι Υπηρεσίες Έκτακτης Ανάγκης και οι Λήψεις Παρασκηνίου Έχουν Προτεραιότητα

Οι θετικές τιμές SHAP για "application_type_Emergency Service" και "application_type_Background Download" επισημαίνουν την προτεραιοποίηση του μοντέλου σε αυτούς τους τύπους εφαρμογών. Οι υπηρεσίες έκτακτης ανάγκης λαμβάνουν τον υψηλότερο θετικό αντίκτυπο μετά την ισχύ σήματος, αντικατοπτρίζοντας την κρίσιμη φύση αυτών των εφαρμογών για τη δημόσια ασφάλεια και την έγκαιρη ανταπόκριση. Οι λήψεις παρασκηνίου, οι οποίες συνήθως δεν απαιτούν άμεση ανταπόκριση σε πραγματικό χρόνο αλλά ωφελούνται από τις αδιάκοπες συνδέσεις, ευνοούνται επίσης από το μοντέλο.

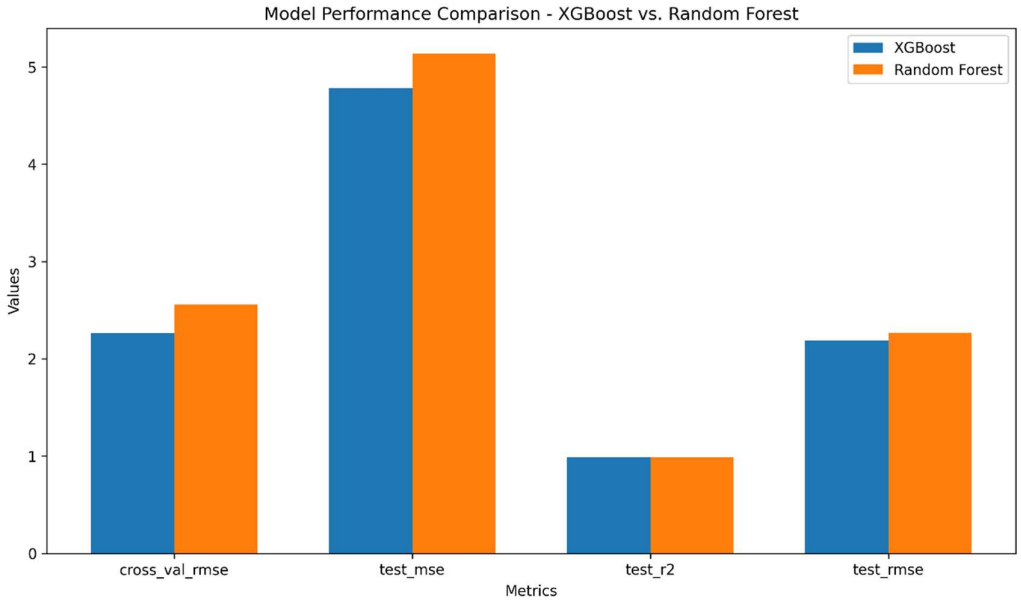
2.2.4. Περιήγηση Ιστού και Απαιτούμενο Εύρος Ζώνης:

Οι αρνητικές τιμές SHAP για "application_type_Web Browsing" και "required_bandwidth(Mbps)" υποδηλώνουν ότι οι εφαρμογές περιήγησης στο διαδίκτυο και εκείνες που απαιτούν χαμηλότερο εύρος ζώνης λαμβάνουν λιγότερο εύρος ζώνης από το μοντέλο XGBoost. Αυτή η προτεραιοποίηση μπορεί να αποδοθεί στο γεγονός ότι η περιήγηση στο διαδίκτυο συχνά ανέχεται μικρές καθυστερήσεις ή buffering, και οι χρήστες μπορεί να έχουν διαφορετικές απαιτήσεις εύρους ζώνης ανάλογα με το περιεχόμενο της ιστοσελίδας.

2.2.5. Επιπρόσθετα σημεία:

- Οι τιμές SHAP αντιπροσωπεύουν την επίδραση ενός χαρακτηριστικού σε μια μόνο πρόβλεψη. Η πραγματική διάθεση εύρους ζώνης μπορεί να επηρεαστεί από τη συνεργασία πολλών χαρακτηριστικών.
- Οι συγκεκριμένες υπερπαράμετροι που χρησιμοποιούνται για την εκπαίδευση του μοντέλου XGBoost μπορούν επίσης να επηρεάσουν τη σχετική σημαντικότητα διαφορετικών χαρακτηριστικών.
- Αναλύοντας τη συνοπτική γραφική παράσταση SHAP, αποκτούμε πολύτιμες πληροφορίες για το πώς το μοντέλο XGBoost λαμβάνει αποφάσεις σχετικά με την κατανομή εύρους ζώνης. Το μοντέλο δίνει προτεραιότητα στο strong signal strength, διασφαλίζει δικαιοσύνη στην κατανομή (fairness in allocation) και λαμβάνει υπόψη την κρίσιμη φύση των εφαρμογών υπηρεσιών έκτακτης ανάγκης, ενώ παράλληλα βελτιστοποιεί για λήψεις στο παρασκήνιο. Ωστόσο, η περιήγηση στον ιστό και οι εφαρμογές χαμηλού εύρους ζώνης ενδέχεται να έχουν χαμηλότερη προτεραιότητα. Είναι σημαντικό να θυμάστε ότι αυτές είναι οι τάσεις του μοντέλου και η πραγματική κατανομή μπορεί να επηρεαστεί από διάφορες συνθήκες δικτύου σε πραγματικό χρόνο.

Η ανάλυση των δύο μοντέλων παλινδρόμησης, XGBoost και Random Forest, αποκαλύπτει μια διαφορά στην επίδοση, με το XGBoost να υπερέχει σε όλες τις μετρικές (Εικόνα 13):



Εικόνα 13 : Cross values RMSE, test MSE, test R^2 and test RMSE – comparison between XGBoost vs Random Forest

XGBoost:

- **Cross-Validation RMSE:** 2.26
- **Test RMSE:** 2.19
- **Test R^2:** 0.99

Random Forest:

- **Cross-Validation RMSE:** 2.56
- **Test RMSE:** 2.27
- **Test R^2:** 0.99

Συνοψίζοντας:

- Το XGBoost παρουσιάζει χαμηλότερο RMSE τόσο στην cross-validation όσο και στο test set, δείχνοντας καλύτερη γενικότερη επίδοση.
- Η διαφορά στο test R^2 είναι μικρή, υποδηλώνοντας ότι και τα δύο μοντέλα είναι αρκετά καλά στην εξήγηση της μεταβλητότητας των δεδομένων.

Ερμηνεία:

- Η βελτιωμένη επίδοση του XGBoost πιθανώς οφείλεται στην τεχνική boosting, η οποία αντιμετωπίζει διαδοχικά τα αδύναμα decision trees με στόχο να δημιουργήσει ένα ισχυρό μοντέλο.
- Ο Random Forest, από την άλλη, χρησιμοποιεί bagging, δημιουργώντας πολλά decision trees και λαμβάνοντας την μέση τιμή των προβλέψεων.
- Η διαφορά στα δύο μοντέλα δείχνει ότι η boosting μπορεί να είναι περισσότερο αποτελεσματική στην αντιμετώπιση των συγκεκριμένων περιπτώσεων των δεδομένων 5G latency.

Σημαντική Σημείωση:

- Η αξιολόγηση των δύο μοντέλων βασίζεται στο συγκεκριμένο σύνολο δεδομένων που χρησιμοποιήθηκε στην ανάλυση.
- Για διαφορετικά σύνολα δεδομένων, η επίδοση των δύο μοντέλων μπορεί να διαφέρει σημαντικά.

Συνολικά:

- Το XGBoost απέδωσε καλύτερα (υπερέχει ελαφρώς), με χαμηλότερο cross-validation και test RMSE.
- Η διαφορά στην επίδοση δεν είναι τεράστια, υποδηλώνοντας ότι και τα δύο μοντέλα είναι αρκετά καλά.
- Αμφότερα τα μοντέλα δείχνουν καλή επίδοση στην πρόβλεψη της latency.

3. Συμπεράσματα

- Η latency στο δίκτυο 5G επηρεάζεται από την ισχύ του σήματος και τον τύπο της εφαρμογής.
- Η boosting τεχνική του XGBoost μπορεί να είναι περισσότερο αποτελεσματική στην αντιμετώπιση των συγκεκριμένων περιπτώσεων των δεδομένων 5G latency.
- Η χρήση των decision-tree-ensemble μοντέλων παρέχει ικανοποιητική πρόβλεψη της latency.
- Η ανάλυση SHAP επιβεβαιώνει την εξαιρετικά σημαντική επίδραση του "signal_strength(dBm)" στην latency. Η "signal_strength(dBm)" και οι συγκεκριμένες εφαρμογές (Emergency Service, IoT Temperature, Background Download, Web Browsing) αποτελούν επομένως key predictors.
- Τα διαγράμματα SHAP μπορούν να χρησιμοποιηθούν για τον εντοπισμό πιθανών προβλημάτων στο μοντέλο, όπως υπερπροσαρμογή (overfitting) ή μεροληψία (bias).
- Οι εφαρμογές εκτάκτης ανάγκης και οι IoT εφαρμογές απαιτούν ιδιαίτερη προσοχή στην βελτιστοποίση του latency για να εξασφαλιστεί η ομαλή λειτουργία τους.
- Συνολικά, η ανάλυση SHAP προσφέρει τις απαραίτητες ενδείξεις για να βελτιωθεί η πρόβλεψη latency και να εξασφαλιστεί η βέλτιστη λειτουργία του δικτύου 5G.

Βιβλιογραφία

1. Chiang, M. Z. (2020). *5G and Data Analytics: A Survey. IEEE Future Networks.*
2. Giordani, M. &. (2020). *Resource Allocation in 5G Networks: A Survey and Challenges.* IEEE Access.
3. Lundberg, S. M. (2020). A unified approach to interpreting model predictions. *In Advances in neural information processing systems*, σσ. pp. 4765-4774.
4. Lundberg, S. M. (2020). Understanding the local structure of decision trees. *Journal of Artificial Intelligence Research*, σσ. 75, 1339-1370.
5. M. Chiang, T. Z. (2020). "5G and Data Analytics: A Survey," 2020. Ανάκτηση από https://futurenetworks.ieee.org/images/files/pdf/applications/Data-Analytics-in-5G-Applications030518.pdf
6. Molnar, C. (2022). *Interpretable Machine Learning.* Packt Publishing.
7. Ribeiro, M. T. (2016). Why do some features have more influence than others? A unified approach to feature importance. *Proceedings of the 22nd ACM international conference on information & knowledge management.* ACM.
8. Singh, S. &. (2023). *5G and Beyond: Wireless Communications for a Smart World.* CRC Press.
9. X. Li, X. W. (2022). "The Impact of 5G on Data Analysis and Machine Learning: A Survey and Outlook,". Ανάκτηση από https://futurenetworks.ieee.org/images/files/pdf/applications/Data-Analytics-in-5G-Applications030518.pdf
10. Y. Wu, M. M. (2021). "5G for Data Analytics: Challenges and Opportunities,". Ανάκτηση από https://www.sciencedirect.com/science/article/abs/pii/S1389128620300827
11. Διεθνής Ένωση Τηλεπικοινωνιών (ITU). (2022, April). *https://www.itu.int*. Ανάκτηση από https://www.itu.int:

https://www.itu.int/en/mediacentre/backgrounders/Pages/5G-fifth-generation-of-mobile-technologies.aspx

12. Ένωση Βιομηχανιών Κινητής Τηλεφωνίας (GSMA). (χ.χ.). *"The 5G Impact Report" ()* . Ανάκτηση από www.gsma.com : https://www.gsma.com/futurenetworks/volte-related-news/understanding-5g-perspectives-on-future-technological-advancements-in-mobile-gsmai-report/