# Αρχιτεκτονικές 5G, Τεχνολογίες, Εφαρμογές και Βασικοί Δείκτες Απόδοσης

Εαρινό Εξάμηνο 2025
19 Μαΐου 2025

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΑΤΡΩΝ
UNIVERSITY OF PATRAS

Τμήμα Μηχανικών Η / Υ και Πληροφορικής
**Πολυτεχνική Σχολή**

**Τομέας Υλικού και Αρχιτεκτονικής των Υπολογιστών**
**Επιλεγόμενο Μάθημα – CEID_NE577**

**Στοιχεία Ομάδας**
**Όνομα:** Χρυσαυγή, Μηλτιάδης
**Επώνυμο:** Πατέλη**,** Μαντές
**Α.Μ.:** 1084513, 1084661
**E – mail:** up1084513@ac.upatras.gr, up1084661@ac.upatras.gr
**Εξάμηνο:** 10ο, 10ο

**Διδάσκων**: Χρήστος Βερυκούκης
**Επιβλέπων**: Παναγιώτης Μαράντης

# ΘΕΜΑ: Πρόβλεψη Downlink Bitrate (Throughput) Με Βάση Μετρικές Ποιότητας

https://github.com/miltiadiss/CEID_NE577-5G-Architectures-Technologies-Applications-and-Key-Performance-Indicators

# ΠΕΡΙΕΧΟΜΕΝΑ

# 0 ΕΙΣΑΓΩΓΗ

Χρησιμοποιείται το dataset με τίτλο **"4G LTE Speed Dataset and Bandwidth"**, το οποίο αποτελεί ένα πλούσιο σύνολο **πραγματικών μετρήσεων πεδίου**, που έχουν συλλεχθεί μέσω της εφαρμογής **G-NetTrack Pro** σε συσκευές Android χωρίς πρόσβαση root. Οι μετρήσεις πραγματοποιούνται **ανά δευτερόλεπτο**, παρέχοντας λεπτομερή χρονική απεικόνιση της συμπεριφοράς του 4G/LTE δικτύου υπό πραγματικές συνθήκες χρήσης.

Το σύνολο δεδομένων περιλαμβάνει **135 διακριτές καταγραφές (sessions)**, καθεμία εκ των οποίων έχει διάρκεια περίπου **15 λεπτών**, με συνολικό αριθμό δειγμάτων που υπερβαίνει τις **100.000 χρονικές στιγμές**. Οι καταγραφές καλύπτουν **ποικιλία σεναρίων κινητικότητας**, όπως στατική θέση, κίνηση με τα πόδια, επιβατικά οχήματα, λεωφορεία και τρένα, αλλά και **διαφορετικά γεωγραφικά σημεία**. Αυτό το χαρακτηριστικό καθιστά το dataset ιδανικό για μελέτη της επίδρασης της κινητικότητας, του χώρου και του τύπου σύνδεσης στη συμπεριφορά του δικτύου.

Κάθε εγγραφή (χρονική στιγμή) περιλαμβάνει τα εξής:

- **Μετρικές ποιότητας καναλιού (radio signal metrics):**
  - **RSRP** (*Reference Signal Received Power*): δείχνει την ισχύ του λαμβανόμενου σήματος από τον σταθμό βάσης.
  - **RSRQ** (*Reference Signal Received Quality*): δείκτης ποιότητας σήματος, βασισμένος στη σχέση ισχύος προς παρεμβολή.
  - **RSSI** (*Received Signal Strength Indicator*): συνολική ισχύς του λαμβανόμενου σήματος, περιλαμβανομένων παρεμβολών.
  - **SNR** (*Signal-to-Noise Ratio*): λόγος σήματος προς θόρυβο, σημαντικός για το modulation και την απόδοση.
  - **CQI** (*Channel Quality Indicator*): εκτίμηση της ποιότητας του καναλιού, με βάση την ανάλυση του σταθμού βάσης.
- **Μετρικές χρήσης δικτύου (network utilization):**
  - **UL_bitrate**: ταχύτητα αποστολής δεδομένων (uplink) σε kbps.
  - **DL_bitrate**: ταχύτητα λήψης δεδομένων (downlink), που αποτελεί και τη **μεταβλητή-στόχο (target)** στην πρόβλεψη.
- **Χρονικά, γεωγραφικά και περιβαλλοντικά δεδομένα:**
  - **Time**: χρονική σήμανση κάθε καταγραφής.
  - **Latitude, Longitude**: συντεταγμένες θέσης του χρήστη.
  - **Speed, Altitude**: ταχύτητα κίνησης και υψόμετρο.
  - **CellID, NetworkType, Technology**: τεχνικά στοιχεία για τον τύπο του κελιού και της σύνδεσης.

Το μέγιστο καταγεγραμμένο **DL bitrate** αγγίζει τα **~173 Mbps**, ενώ παρατηρούνται σημαντικές διακυμάνσεις που εξαρτώνται από τη γεωγραφική θέση, την πυκνότητα του δικτύου, την ταχύτητα του χρήστη, και εξωτερικούς περιβαλλοντικούς παράγοντες. Αυτή η έντονη **ετερογένεια** καθιστά το πρόβλημα πρόβλεψης ιδιαίτερα ενδιαφέρον και απαιτητικό.

Ο στόχος είναι η **πρόβλεψη της στιγμιαίας τιμής του DL_bitrate**, δηλαδή της ταχύτητας λήψης δεδομένων σε συγκεκριμένο χρονικό σημείο, χρησιμοποιώντας ως είσοδο:

- τις μετρικές ποιότητας σήματος,
- τις τιμές του uplink (UL),
- χωρικά και χρονικά χαρακτηριστικά,
- καθώς και **ιστορικές τιμές** του DL/UL bitrate (μέσω lagged ή rolling features).

Πρόκειται για πρόβλημα **χρονικής παλινδρόμησης (time series regression)**, το οποίο ενσωματώνει στοιχεία από **προβλήματα πρόβλεψης multivariate time series** και **προγνωστικής μοντελοποίησης σε κινητά δίκτυα**.

**Προκλήσεις και χαρακτηριστικά του προβλήματος:**

- **Μη γραμμικότητα**: Η σχέση μεταξύ των εισόδων και του **DL_bitrate** δεν είναι γραμμική, απαιτώντας μοντέλα που συλλαμβάνουν πολύπλοκες συσχετίσεις.

- **Υψηλή διακύμανση & θόρυβος**: Η ταχύτητα λήψης επηρεάζεται από πλήθος αστάθμητων παραγόντων, καθιστώντας την προβλεψιμότητα δύσκολη.
- **Αυτοσυσχέτιση**: Παλιότερες τιμές bitrate μεταφέρουν πληροφορία για την τρέχουσα απόδοση.
- **Ετερογένεια εισόδων**: Οι τιμές σήματος ποικίλλουν δραστικά ανά περιοχή, συσκευή και ώρα.
- **Εποχικότητα**: Υπάρχουν καθημερινά ή ωριαία μοτίβα (π.χ. κυκλοφοριακή αιχμή).

Η αντιμετώπιση του προβλήματος απαιτεί **σύγχρονες τεχνικές μηχανικής μάθησης και βαθιάς μάθησης**, σε συνδυασμό με **feature engineering** (rolling windows, normalization, encoding) και τεχνικές **ερμηνευσιμότητας μοντέλων** (π.χ. SHAP).

Στην επόμενη ενότητα παρουσιάζεται πιο αναλυτικά η σημασία πρόβλεψης του DL_bitrate όσον αφορά τα σύγχρονα 5G δίκτυα.

# 1 ΣΗΜΑΣΙΑ ΤΗΣ ΠΡΟΒΛΕΨΗΣ ΤΟΥ DOWNLINK BITRATE ΣΥΣΚΕΥΩΝ ΣΤΑ ΣΥΓΧΡΟΝΑ ΔΙΚΤΥΑ 5G

Η πρόβλεψη του throughput (ρυθμού μετάδοσης καθοδικής ζεύξης) στις υποδομές 5G αποτελεί καθοριστικό παράγοντα για την αποδοτική λειτουργία των τηλεπικοινωνιακών δικτύων, τη βελτιστοποίηση της ποιότητας υπηρεσίας (QoS) και τη δυναμική διαχείριση των διαθέσιμων πόρων. Με τη ραγδαία αύξηση των απαιτήσεων για αξιόπιστες και υψηλής ταχύτητας συνδέσεις, η ικανότητα ενός δικτύου να προβλέπει και να προσαρμόζεται σε μεταβαλλόμενες συνθήκες είναι ζωτικής σημασίας.

1. **Βελτιστοποίηση της κατανομής πόρων και αποφυγή συμφόρησης**

Η ακριβής πρόβλεψη του throughput επιτρέπει στα δίκτυα να κατανέμουν δυναμικά το διαθέσιμο εύρος ζώνης και τους ραδιοπόρους με αποδοτικό τρόπο, εξισορροπώντας τα φορτία και αποφεύγοντας τη συμφόρηση. Με αυτόν τον τρόπο, μειώνεται η πιθανότητα εμφάνισης προβλημάτων όπως η μείωση της ταχύτητας μετάδοσης δεδομένων και το αυξημένο latency, που μπορεί να επηρεάσουν κρίσιμες εφαρμογές.

2. **Βελτίωση εμπειρίας χρήστη και ποιότητας υπηρεσίας (QoS)**

Η πρόβλεψη των απαιτήσεων μετάδοσης επιτρέπει στο δίκτυο να προσαρμόζει δυναμικά τις παραμέτρους του, μειώνοντας τη λανθάνουσα κατάσταση (latency) και εξασφαλίζοντας σταθερή και αξιόπιστη σύνδεση. Αυτό είναι ιδιαίτερα σημαντικό για εφαρμογές που απαιτούν υψηλή διαθεσιμότητα και ελάχιστες καθυστερήσεις, όπως το cloud gaming, η εικονική/επαυξημένη πραγματικότητα (VR/AR), η τηλεϊατρική και τα αυτόνομα οχήματα.

3. **Υποστήριξη προηγμένων εφαρμογών και δικτυακών τεχνολογιών**

Η ακριβής πρόβλεψη του throughput αποτελεί βασικό συστατικό για τη λειτουργία καινοτόμων τεχνολογιών, όπως:

- **Network Slicing**: Δυνατότητα δημιουργίας εξειδικευμένων υποδικτύων που εξυπηρετούν διαφορετικές ανάγκες, με το throughput να προσαρμόζεται στις απαιτήσεις κάθε slice.
- **Edge Computing**: Προσαρμογή της κατανομής υπολογιστικών πόρων βάσει των αναγκών throughput, βελτιώνοντας την απόδοση των αποκεντρωμένων συστημάτων.
- **Διαχείριση Massive IoT**: Η μαζική διασύνδεση εκατομμυρίων συσκευών απαιτεί αξιόπιστες προβλέψεις throughput ώστε να διασφαλίζεται η σταθερότητα του δικτύου.

4. **Οικονομικά και ενεργειακά οφέλη**

Η σωστή διαχείριση του throughput μειώνει το λειτουργικό κόστος των παρόχων δικτύου, επιτρέποντας καλύτερο σχεδιασμό επενδύσεων σε υποδομές. Επιπλέον, συμβάλλει στην ενεργειακή αποδοτικότητα, καθώς η δυναμική προσαρμογή της ισχύος μετάδοσης και η βελτιστοποίηση της χρήσης των πόρων μειώνουν την κατανάλωση ενέργειας, τόσο στις ίδιες τις τηλεπικοινωνιακές υποδομές όσο και στις συσκευές των χρηστών.

Συμπερασματικά, η πρόβλεψη του downlink bitrate στα δίκτυα 5G δεν αποτελεί απλά μια τεχνική βελτιστοποίησης αλλά έναν θεμελιώδη παράγοντα για την αποδοτική λειτουργία των δικτύων επόμενης γενιάς. Συμβάλλει στη βελτίωση της εμπειρίας των χρηστών, τη μείωση του κόστους και της κατανάλωσης ενέργειας, ενώ παράλληλα υποστηρίζει τις απαιτήσεις προηγμένων τεχνολογιών, όπως το network slicing, το edge computing και η μαζική διασύνδεση IoT συσκευών. Μέσω έξυπνων αλγορίθμων και τεχνικών ανάλυσης δεδομένων, τα σύγχρονα δίκτυα μπορούν να προσαρμόζονται δυναμικά στις απαιτήσεις των χρηστών, διασφαλίζοντας μια γρήγορη, αξιόπιστη και αποδοτική σύνδεση.

## 2 ΒΙΒΛΙΟΘΗΚΕΣ

Οι βιβλιοθήκες που χρησιμοποιήθηκαν είναι οι εξής:

- **pandas:** χρησιμοποιεί DataFrames για την ανάλυση και διαχείριση δεδομένων.
- **numpy:** χρησιμοποιείται για επιστημονικούς υπολογισμούς.
- **os:** χρησιμοποιείται για την διαχείριση αρχείων.
- **matplotlib:** χρησιμοποιείται για την δημιουργία γραφημάτων (συνάρτηση **pyplot**).
- **statsmodels:** παρέχει εργαλεία για στατιστική ανάλυση χρονοσειρών. Η συνάρτηση **seasonal_decompose** χρησιμοποιείται για την αποσύνθεση της χρονοσειράς σε trend, seasonality και residual. Επίσης, χρησιμοποιείται η συνάρτηση **plot_acf**, η οποία δημιουργεί ένα γράφημα που απεικονίζει την αυτοσυσχέτιση της χρονοσειράς για διάφορες τιμές lag.
- **seaborn:** χρησιμοποιείται για την δημιουργία στατιστικών γραφημάτων όπως heatmaps, boxplots.
- **scipy.stats:** περιλαμβάνει πολλές στατιστικές συναρτήσεις: υπολογισμό κατανομών (π.χ., normal distribution).
- **math**: χρησιμοποιείται για μαθηματικές πράξεις.
- **sklearn**: χρησιμοποιείται:
    - **pca** για την μείωση διαστατικότητας στα δεδομέναχρησιμοποιείται
    - **oneHotEncoder** για την μετατροπή κατηγορικών μεταβλητών σε αριθμητικά διανύσματα.
    - **MinMaxScaler** για την κανονικοποίηση δεδομένων
    - **train_test_split** για τη διαχωρισμό σε σύνολα εκπαίδευσης και δοκιμής,
    - συναρτήσεις **mean_squared_error**, **mean_absolute_error** για την αξιολόγηση της απόδοσης των μοντέλων.
- **xgboost**: χρησιμοποιείται για την εκπαίδευση μοντέλων μηχανικής μάθησης με gradient boosting trees.
- **itertools**: χρησιμοποιείται η συνάρτηση **product** για την παραγωγή όλων των δυνατών συνδυασμών παραμέτρων κατά την αναζήτηση υπερπαραμέτρων (**grid search**).
- **shap**: χρησιμοποιείται για την ερμηνεία των προβλέψεων των μοντέλων μέσω της ανάλυσης της σημασίας των χαρακτηριστικών (feature importance) με βάση τις τιμές SHAP (SHapley Additive exPlanations).
- **keras :** Η βιβλιοθήκη Keras χρησιμοποιείται για την κατασκευή και εκπαίδευση νευρωνικών δικτύων.
    - Το **Sequential** επιτρέπει τη δημιουργία σειριακών μοντέλων, όπου τα επίπεδα προστίθενται διαδοχικά.
    - Το **LSTM** χρησιμοποιείται για την ανάλυση χρονοσειρών.
    - Το **Dense** είναι ένα πλήρως συνδεδεμένο επίπεδο για την τελική πρόβλεψη.
    - Το **Dropout** χρησιμοποιείται για να μειώσει το overfitting.
    - Το **EarlyStopping** χρησιμοποιείται για να σταματά την εκπαίδευση όταν η απόδοση δεν βελτιώνεται

# 3 ΠΡΟΕΠΕΞΕΡΓΑΣΙΑ

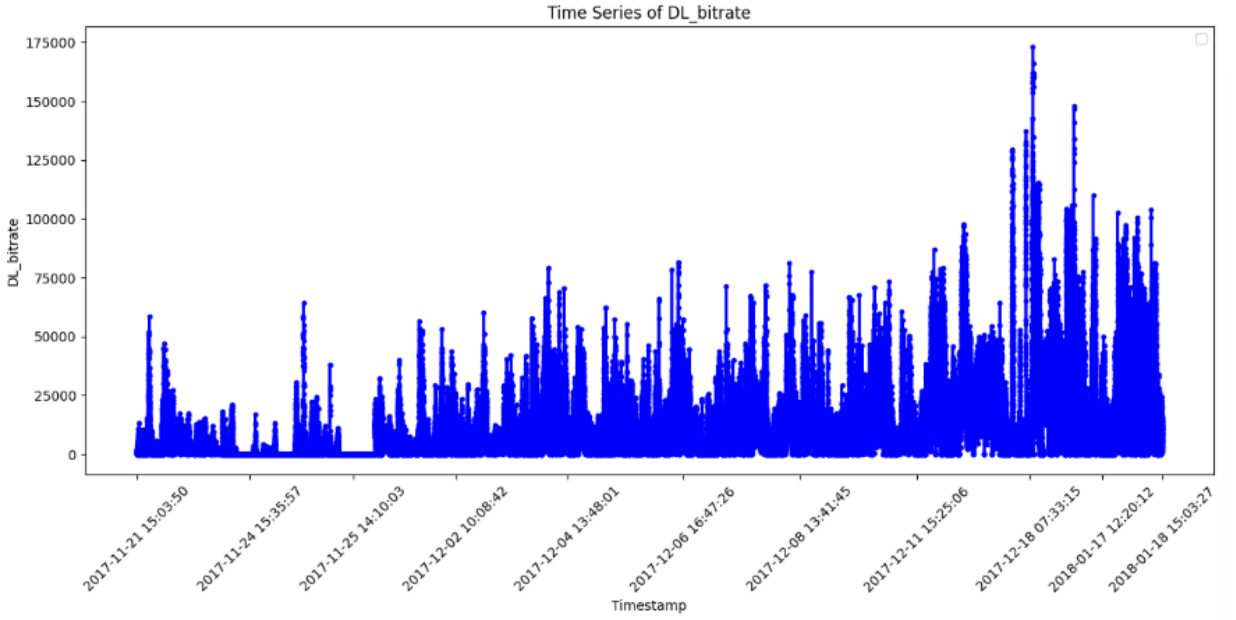
## 3.1 Ανάλυση Χρονοσειράς Target και Στατιστική Ανάλυση Χαρακτηριστικών

Το dataset περιέχει διαφορετικούς φακέλους με ονόματα train, pedestrian, static, car, bus τα οποία αποτελούν το σενάριο κίνησης. Κάθε φάκελος αποτελείται από διαφορετικό πλήθος csv αρχείων. Αρχικά, από κάθε CSV αρχείο εξαλείφθηκαν διπλότυπες εγγραφές χρησιμοποιώντας τη μέθοδο **drop_duplicates()**, ώστε να παραμείνουν μοναδικές εγγραφές σε όλα τα πεδία. Επίσης, με την μέθοδο **drop()** αφαιρέθηκαν στήλες όπως **Longitude**, **Latitude**, **ServingCell_Lon**, **ServingCell_Lat,** καθώς σχετίζονται με τις γεωγραφικές συντεταγμένες της κινητής συσκευής και του σταθμού βάσης και δεν περιέχουν χρήσιμες πληροφορίες για την πρόβλεψη του downlink bitrate. Άλλωστε η ίδια πληροφορία αποτυπώνεται και στη στήλη **ServingCell_Distance** που μετράει την απόσταση ανάμεσά τους. Επιπλέον, στις αριθμητικές στήλες οι κενές σειρές που σημειώνονται με "-" αντικαταστάθηκαν με NaN χρησιμοποιώντας την **replace()** και οι στήλες μετατράπηκαν σε αριθμητικό τύπο με την **to_numeric()**.

Στην συνέχεια, όλα τα επιμέρους DataFrames από τους φακέλους συνενώθηκαν με τη μέθοδο **pd.concat()** σε ένα ενιαίο DataFrame (**combined_data**). Προκειμένου να διατηρηθεί η πληροφορία που αντιπροσωπεύει τις συνθήκες καταγραφής, δεδομένου ότι το όνομα κάθε φακέλου υποδηλώνει το είδος της κίνησης δημιουργήθηκε μια νέα στήλη με την ονομασία **scenario**. Έπειτα, για την πραγματοποιήθηκε καθαρισμός της στήλης **Timestamp** με αντικατάσταση του χαρακτήρα "_" με το κενό. Επιπλέον, η στήλη μετατράπηκε σε τύπο **datetime** χρησιμοποιώντας τη μέθοδο **pd.to_datetime()** Τέλος, το συνολικό DataFrame ταξινομήθηκε σε αύξουσα χρονολογική σειρά βάσει της στήλης **Timestamp** χρησιμοποιώντας τη μέθοδο **sort_values()**.

Το σύνολο δεδομένων διαχωρίστηκε σε **train** (80%) και **test** (20%) πριν από την εφαρμογή οποιουδήποτε σταδίου επεξεργασίας, προκειμένου το test set να παραμείνει πλήρως ανεξάρτητο και ανεπηρέαστο. Η προεπεξεργασία πραγματοποιήθηκε ξεχωριστά σε κάθε υποσύνολο, ώστε να αποτραπεί η διαρροή πληροφορίας από το train set προς το test set και να διασφαλιστεί η αντικειμενικότητα της αξιολόγησης.

Για το train set πραγματοποιήθηκε απεικόνιση της χρονοσειράς του **DL_bitrate** σε συνάρτηση με το **Timestamp**, αντιμετωπίζοντας το timestamp ως κατηγορική μεταβλητή, δεδομένου ότι οι χρονικές στιγμές δεν είναι συνεχόμενες και τα διαστήματα μεταξύ τους μπορεί να διαφέρουν.

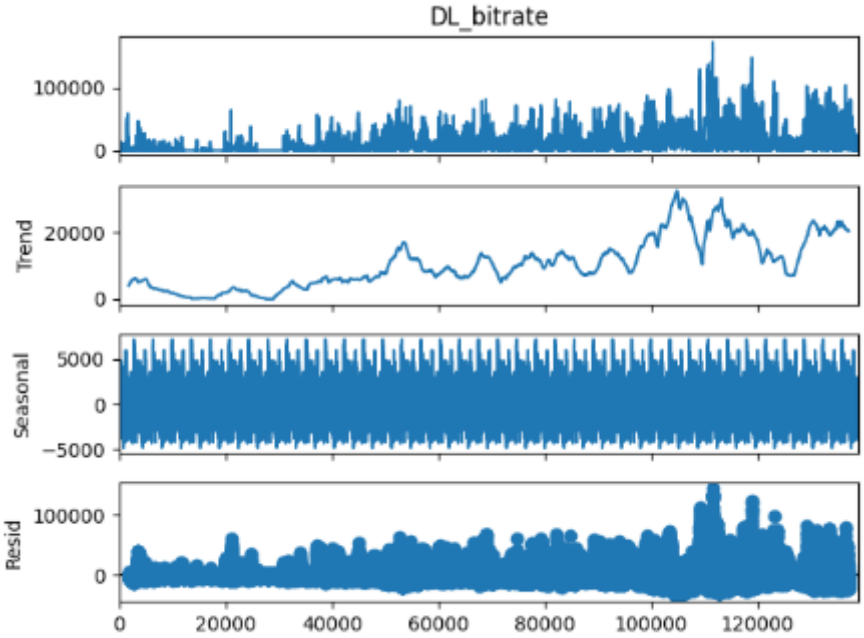



Παρατηρείται ότι το **DL_bitrate** εμφανίζει έντονες διακυμάνσεις κατά τη διάρκεια της χρονικής περιόδου από τα τέλη του 2017 έως τις αρχές του 2018. Στις αρχικές ημερομηνίες οι τιμές του **DL_bitrate** είναι σχετικά χαμηλές και εμφανίζουν μεγαλύτερη διασπορά. Με την πάροδο του χρόνου το **DL_bitrate** αυξάνεται σταδιακά, με πολύ

υψηλές τιμές στα τέλη Δεκεμβρίου. Στις αρχές του 2018 διακρίνεται μια σχετική πτώση και σταθεροποίηση των τιμών. Το γεγονός ότι υπάρχουν έντονες διακυμάνσεις δείχνει ότι το δίκτυο παρουσίαζε περιόδους αιχμής όπου το downlink bitrate αυξανόταν απότομα, πιθανώς λόγω μεταβολών στις συνθήκες κίνησης, στη χρήση δεδομένων ή στις παραμέτρους του δικτύου.

Στη συνέχεια, εφαρμόστηκε εποχιακή αποσύνθεση της χρονοσειράς του **DL_bitrate** με τη μέθοδο **seasonal_decompose()**. Ορίστηκε περίοδος 3600 δευτερολέπτων, ώστε να ανιχνευθούν επαναλαμβανόμενα μοτίβα σε ωριαία βάση. Μέσω της αποσύνθεσης, η χρονοσειρά χωρίστηκε σε:

- **Trend (τάση):** η μακροπρόθεσμη κατεύθυνση του DL_bitrate,
- **Seasonality (εποχικότητα):** περιοδικά μοτίβα που επαναλαμβάνονται σε βάθος ώρας,
- **Residual (υπόλοιπο):** τυχαίες διακυμάνσεις ή θόρυβος.




Στην παραπάνω εικόνα απεικονίζεται η εποχιακή αποσύνθεση της χρονοσειράς του **DL_bitrate**

- Στο **Trend**, φαίνεται η μακροπρόθεσμη τάση του σήματος. Παρατηρείται ότι υπάρχει μια ανοδική τάση μέχρι περίπου το μέσο της περιόδου και έπειτα μικρές διακυμάνσεις με περιοδικές αυξομειώσεις.
- Το **Seasonality**, απεικονίζει ένα καθαρό περιοδικό μοτίβο, σχεδόν συμμετρικό, με υψηλές και χαμηλές τιμές που επαναλαμβάνονται σε σταθερό χρονικό διάστημα. Αυτό υποδεικνύει ισχυρή εποχικότητα στη χρήση ή στη συμπεριφορά του δικτύου, πιθανότατα λόγω κυκλικών ημερησίων μοτίβων (π.χ., ώρες αιχμής/ηρεμίας).
- Τα **Residuals**, έχουν μεγάλη διασπορά και υψηλή μεταβλητότητα ανα περιόδους, γεγονός που υποδηλώνει ότι υπάρχουν απρόβλεπτες, τυχαίες μεταβολές που επηρεάζουν το DL bitrate.

Συνολικά, η αποσύνθεση αποκαλύπτει ότι το σήμα του **DL_bitrate** έχει και σαφή **τάση** και ισχυρή **εποχικότητα**, αλλά επηρεάζεται και από σημαντικό **τυχαίο θόρυβο**.

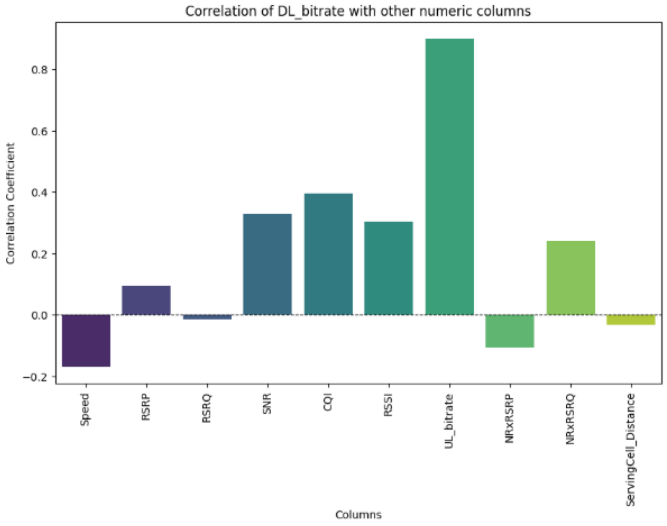
Υπολογίστηκε ο Συντελεστής Μεταβλητότητας (Coefficient of Variation-CV) του **DL_bitrate** πάνω στο training set ο οποίος δίνεται από τη σχέση: $CV = \frac{\sigma}{\mu}$. Στο **mean_dl** αποθηκεύτηκε ο μέσος όρος του DL_bitrate και στο **std_dl** η τυπική απόκλιση που εκφράζει το πόσο διαφέρουν οι τιμές του DL_bitrate από τον μέσο όρο.

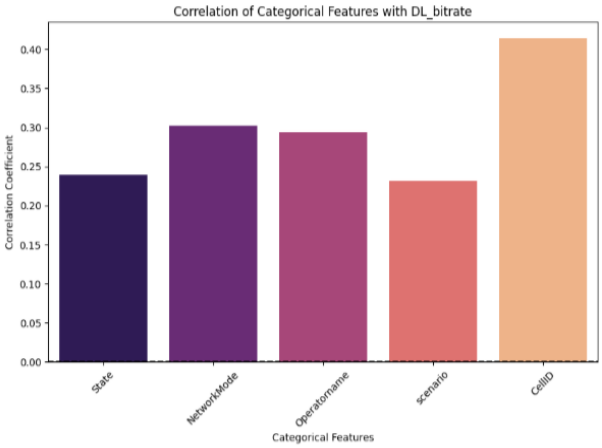
Το αποτέλεσμα που προέκυψε ήταν:

```
Coefficient of Variation for DL_bitrate: 1.3319
```

Η τιμή αυτή είναι **αρκετά υψηλή**, καθώς σημαίνει ότι η τυπική απόκλιση είναι σχεδόν 140% της μέσης τιμής. Γεγονός που υποδηλώνει ότι η πρόβλεψη του DL_bitrate θα είναι δύσκολη.

Προκειμένου να εντοπιστούν τα χαρακτηριστικά (**features**) που συμβάλλουν περισσότερο στην πρόβλεψη του **DL_bitrate**, πραγματοποιήθηκε στο train set ανάλυση συσχέτισης μεταξύ των εισόδων και της μεταβλητής στόχου. Οι μη αριθμητικές στήλες (**Timestamp**, **State**, **NetworkMode**, **Operatorname**, **scenario**, **CellID**) αφαιρέθηκαν και υπολογίστηκε ο συντελεστής συσχέτισης μεταξύ των αριθμητικών χαρακτηριστικών χρησιμοποιώντας τη μετρική **Pearson**. Για τα κατηγορικά χαρακτηριστικά (**State**, **NetworkMode**, **Operatorname**, **scenario**, **CellID**) εφαρμόστηκε **target encoding**, όπου σε κάθε εγγραφή ανατέθηκε ο μέσος όρος του DL_bitrate για την αντίστοιχη κατηγορία. Με αυτόν τον τρόπο κατέστη δυνατός ο υπολογισμός της **pearson correlation** μεταξύ των κατηγορικών χαρακτηριστικών και του DL_bitrate.



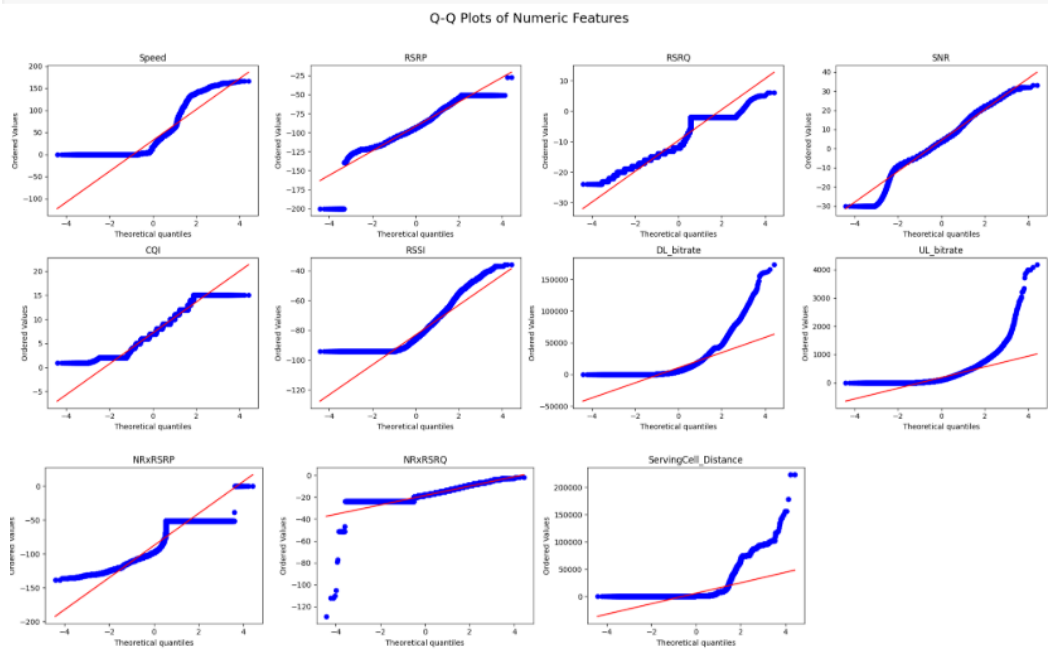




Παρατηρείται ότι το **DL_bitrate** παρουσιάζει ισχυρή θετική συσχέτιση με το **UL_bitrate**, γεγονός που υποδηλώνει ότι οι συνθήκες που ευνοούν τη μετάδοση δεδομένων προς το δίκτυο επηρεάζουν αντίστοιχα και τη λήψη δεδομένων. Επίσης, οι μεταβλητές **CQI** και **SNR**, που εκφράζουν την ποιότητα του καναλιού, σχετίζονται θετικά με το DL_bitrate. Αντίθετα, η **ταχύτητα κίνησης** παρουσιάζει αρνητική συσχέτιση, επιβεβαιώνοντας ότι η αυξημένη κινητικότητα επηρεάζει αρνητικά την απόδοση του δικτύου. Όσον αφορά τις κατηγορικές μεταβλητές, το **CellID** αναδείχθηκε ως το χαρακτηριστικό με τη μεγαλύτερη επιρροή, γεγονός που φανερώνει ότι η γεωγραφική θέση και τα χαρακτηριστικά του σταθμού βάσης παίζουν σημαντικό ρόλο στη διακύμανση του DL_bitrate. Επιπλέον, το **NetworkMode** και ο **Operatorname** φαίνεται να επηρεάζουν επίσης ουσιαστικά την ποιότητα της σύνδεσης.

## 3.2 Διαχείριση αριθμητικών χαρακτηριστικών

Δεδομένου ότι ορισμένα αριθμητικά χαρακτηριστικά του dataset περιείχαν ελλιπείς τιμές **Null**, εφαρμόστηκε η μέθοδος **interpolate()** για τη συμπλήρωση αυτών των κενών. Συγκεκριμένα, χρησιμοποιήθηκε γραμμική παρεμβολή (**linear interpolation**), η οποία εκτιμά τις ελλιπείς τιμές υπολογίζοντας ενδιάμεσες τιμές με βάση τα γειτονικά υπάρχοντα δεδομένα της ίδιας στήλης. Με την επιλογή **limit_direction='both'**, η μέθοδος επιτρέπει την παρεμβολή τόσο προς τα εμπρός όσο και προς τα πίσω μέσα στη χρονοσειρά, διασφαλίζοντας ότι οι ελλείπεις τιμές μπορούν να αντικατασταθούν ανεξάρτητα από τη θέση τους. Η διαδικασία παρεμβολής εφαρμόστηκε ανεξάρτητα τόσο στο **train** όσο και στο **test** set, ώστε να διατηρηθεί η αυστηρή διάκριση μεταξύ των δύο συνόλων κατά την προεπεξεργασία.

Για την κατανόηση της στατιστικής κατανομής των αριθμητικών χαρακτηριστικών, δημιουργήθηκαν **Q-Q Plots** (Quantile-Quantile Plots). Τα Q-Q plots συγκρίνουν την κατανομή των παρατηρούμενων δεδομένων κάθε χαρακτηριστικού με την ιδανική κανονική κατανομή (Gaussian distribution).




Στα γραφήματα, αν τα σημεία ακολουθούν κατά μήκος της διαγωνίου γραμμής, τότε το χαρακτηριστικό προσεγγίζει την κανονική κατανομή. Αντίθετα, σημαντικές αποκλίσεις από τη γραμμή υποδεικνύουν την ύπαρξη ασυμμετρίας (skewness). Εφόσον τα περισσότερα features δεν ακολουθούν κανονική κατανομή επιλέχθηκε τα outliers να αντιμετωπιστούν χρησιμοποιώντας την **IQR Method**.

Πιο συγκεκριμένα, για την ανίχνευση και τον χειρισμό ακραίων τιμών (**outliers**) στα αριθμητικά χαρακτηριστικά, δημιουργήθηκαν αρχικά **Density Plots** και **Boxplots**, ώστε να οπτικοποιηθεί η αρχική κατανομή των δεδομένων. Ο εντοπισμός των outliers βασίστηκε στον κανόνα του **Interquartile Range (IQR)**, όπου ορίστηκαν τα κάτω και άνω όρια των **whiskers** στα **boxplots** ως:

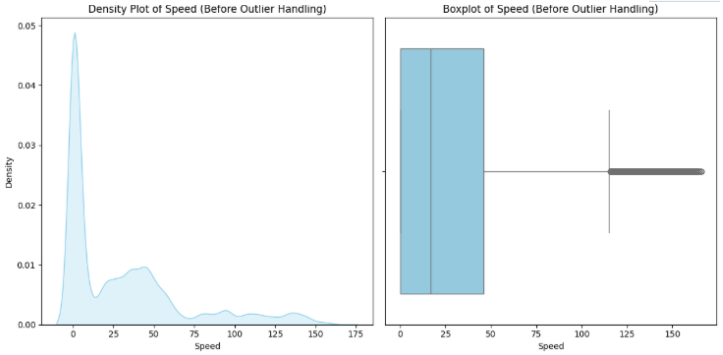
- **low = Q1 – 1.5 × IQR**
- **high = Q3 + 1.5 × IQR**

Η διαδικασία εντοπισμού και χειρισμού των outliers εφαρμόστηκε αποκλειστικά στο **train set**, προκειμένου να αποφευχθεί η μεταφορά πληροφορίας από το train στο test set και να διατηρηθεί η αντικειμενικότητα της αξιολόγησης.

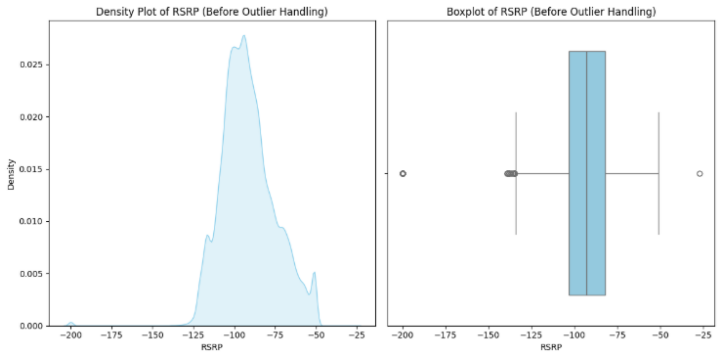
Οποιαδήποτε τιμή εκτός αυτών των ορίων θεωρήθηκε ακραία. Για τη διόρθωση των outliers εφαρμόστηκε **capping**, όπου οι τιμές κάτω από το κατώτερο όριο αντικαταστάθηκαν με το κατώτερο όριο και οι τιμές πάνω από το ανώτερο όριο αντικαταστάθηκαν με το ανώτερο όριο. Μετά το capping, αναπαραστάθηκαν εκ νέου τα Density Plots και τα Boxplots για κάθε χαρακτηριστικό, ώστε να επιβεβαιωθεί η μείωση των ακραίων τιμών.

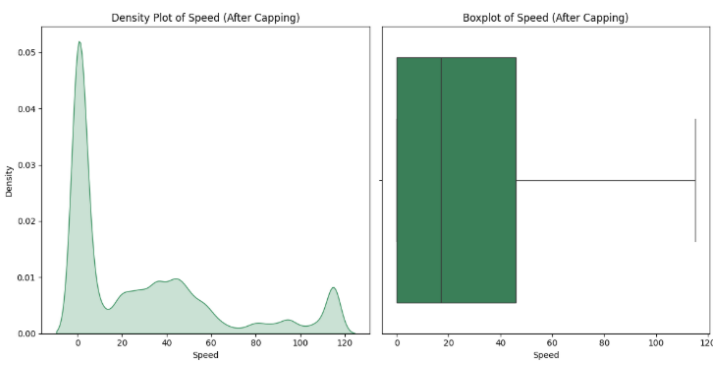
Παρατίθενται ενδεικτικά δύο παραδείγματα των χαρακτηριστικών **speed**, **RSRP**.

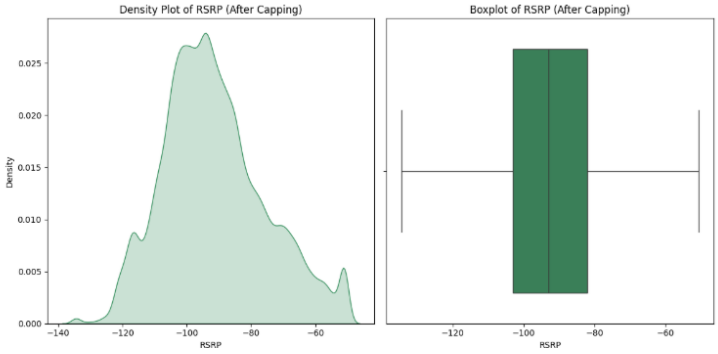
```
For column Speed:
Q1 = 1.0, Q3 = 47.0, IQR = 46.0
Lower Bound = -68.0, Upper Bound = 116.0
Number of Outliers: 8993
```



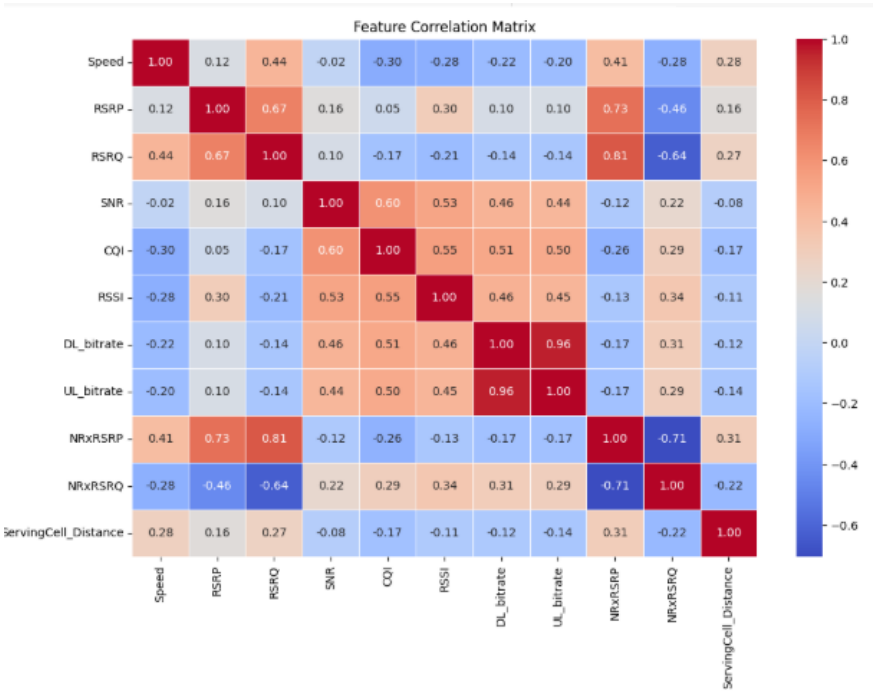

```
For column RSRP:
Q1 = -103.0, Q3 = -82.0, IQR = 21.0
Lower Bound = -134.5, Upper Bound = -50.5
Number of Outliers: 110
```






Για την ανάλυση της συσχέτισης μεταξύ των αριθμητικών χαρακτηριστικών του dataset, υπολογίστηκε το correlation matrix με την μέθοδο **corr()** χρησιμοποιώντας τη μετρική **Pearson**. Κατόπιν, η μήτρα συσχέτισης απεικονίστηκε μέσω **heatmap**.




Παρατηρείται ότι το **DL_bitrate** έχει ισχυρή θετική συσχέτιση με το **UL_bitrate** (0.96), γεγονός αναμενόμενο, καθώς οι συνθήκες του δικτύου που ευνοούν την άνοδο του downlink bitrate ευνοούν αντίστοιχα και το uplink bitrate. Επιπλέον, τα χαρακτηριστικά όπως **CQI**, **RSSI** και **SNR**, τα οποία αποτυπώνουν την ποιότητα του καναλιού, παρουσιάζουν θετική συσχέτιση με το **DL_bitrate** (~0.51, ~0.46 και ~0.46 αντίστοιχα), επιβεβαιώνοντας ότι καλύτερες συνθήκες επικοινωνίας οδηγούν σε υψηλότερες ταχύτητες. Αντίθετα, χαρακτηριστικά όπως η **Speed** και το **NRxRSRQ** δείχνουν ελαφρώς αρνητικές συσχετίσεις με το DL_bitrate, υποδηλώνοντας ότι η κίνηση και οι χαμηλές ποιοτικές ενδείξεις δικτύου επηρεάζουν αρνητικά τη λήψη δεδομένων. Τέλος, τα χαρακτηριστικά ισχύος σήματος (**RSRP**, **RSRQ**, **NRxRSRP**, **NRxRSRQ**) έχουν μεταξύ τους υψηλές συσχετίσεις.

## 3.3 Διαχείριση κατηγορικών χαρακτηριστικών

Για τη μείωση της διάστασης και της πολυπλοκότητας του dataset όλες οι κατηγορικές μεταβλητές συνδυάστηκαν σε μία νέα μεταβλητή και στα δύο σύνολα (train, test). Αρχικά, , οι μεταβλητές **State**, **NetworkMode**, **Operatorname**, **CellID** και **scenario** κωδικοποιήθηκαν με την μέθοδο **OneHotEncoder(),** όπου κάθε κατηγορία μετατράπηκε σε ξεχωριστό δυαδικό χαρακτηριστικό. Δεδομένου ότι η One-Hot κωδικοποίηση αυξάνει σημαντικά το πλήθος των χαρακτηριστικών, εφαρμόστηκε **Principal Component Analysis (PCA)** με στόχο τη μείωση των διαστάσεων και διατηρηθηκε μόνο η πρώτη κύρια συνιστώσα, η οποία εξηγεί τη μέγιστη

διακύμανση των κωδικοποιημένων δεδομένων. Τέλος, οι αρχικές κατηγορικές μεταβλητές αντικαταστάθηκαν με τη νέα στήλη "**Categorical_Impact**", η οποία περιέχει όλη τη σημαντική πληροφορία συμπυκνωμένη.

Η τελική μορφή του training set είναι πλέον η ακόλουθη:

**Signal Quality Metrics** | **Bitrate-Related Features** | **Signal Quality Metrics** | **Categorical Features**

**Timestamp**

| | Timestamp | Speed | RSRP | RSRQ | SNR | CQI | RSSI | DL_bitrate | UL_bitrate | NRxRSRP | NRxRSRQ | ServingCell_Distance | Categorical_Impact |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 2017-11-21 15:03:50 | 0.0 | -95.0 | -13.0 | 4.0 | 10.0 | -80.0 | 0.0 | 0.0 | -106.0 | -19.0 | 551.370000 | 0.758840 |
| 1 | 2017-11-21 15:03:51 | 0.0 | -95.0 | -13.0 | 2.0 | 8.0 | -78.0 | 0.0 | 0.0 | -106.0 | -19.0 | 551.370000 | 0.599827 |
| 2 | 2017-11-21 15:03:52 | 0.0 | -95.0 | -13.0 | 13.0 | 9.0 | -80.0 | 0.0 | 0.0 | -106.0 | -19.0 | 553.430000 | 0.599827 |
| 3 | 2017-11-21 15:03:53 | 1.0 | -95.0 | -13.0 | 13.0 | 9.0 | -80.0 | 0.0 | 0.0 | -106.0 | -19.0 | 563.480000 | 0.599827 |
| 4 | 2017-11-21 15:03:54 | 1.0 | -97.0 | -13.0 | -2.0 | 9.0 | -80.0 | 0.0 | 0.0 | -106.0 | -19.0 | 563.480000 | 0.599827 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 138948 | 2018-01-18 15:03:26 | 0.0 | -95.0 | -14.0 | -5.0 | 8.0 | -88.0 | 6811.0 | 103.0 | -98.0 | -19.0 | 2911.320000 | 0.581546 |
| 138949 | 2018-01-18 15:03:26 | 46.0 | -104.0 | -15.0 | -2.0 | 7.0 | -88.0 | 14337.0 | 216.0 | -107.0 | -16.0 | 4965.960421 | 0.784428 |
| 138950 | 2018-01-18 15:03:27 | 0.0 | -95.0 | -14.0 | -5.0 | 8.0 | -88.0 | 8443.0 | 142.0 | -98.0 | -19.0 | 2911.320000 | 0.581546 |
| 138951 | 2018-01-18 15:03:28 | 46.0 | -104.0 | -15.0 | -2.0 | 7.0 | -88.0 | 15378.0 | 256.0 | -107.0 | -16.0 | 4965.960421 | 0.784428 |
| 138952 | 2018-01-18 15:03:28 | 0.0 | -95.0 | -15.0 | -5.0 | 8.0 | -88.0 | 8410.0 | 134.0 | -96.0 | -18.0 | 2911.320000 | 0.581546 |

138953 rows × 13 columns

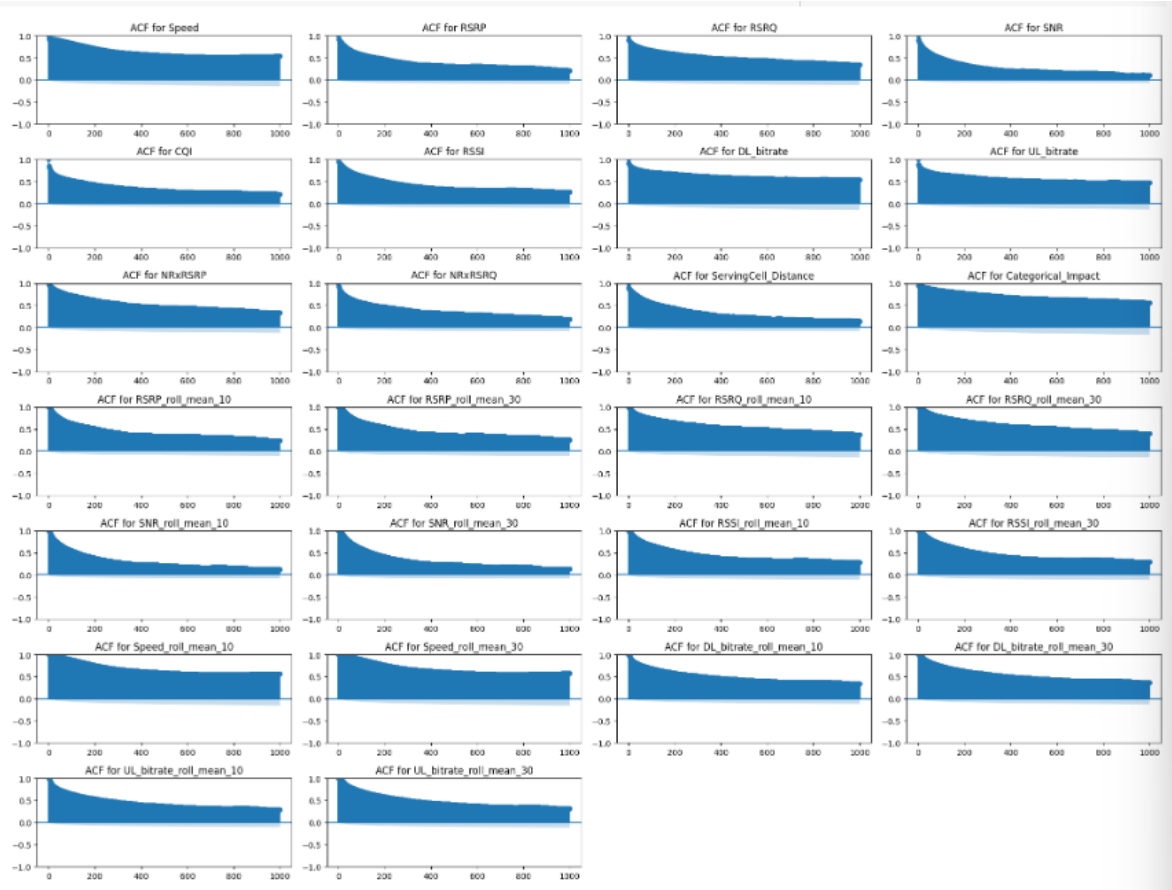
**Target**

# 4 ΥΛΟΠΟΙΗΣΗ XGBOOST REGRESSOR

## 4.1 Feature Engineering

Αρχικά, εξάγονται **χρονικά χαρακτηριστικά** από τη στήλη **Timestamp** (η ώρα της ημέρας (**Hour**), η ημέρα της εβδομάδας (**Day_of_week**) και μια Boolean μεταβλητή για το αν η καταγραφή έγινε Σαββατοκύριακο (**Is_weekend**)). Με αυτόν τον τρόπο το μοντέλο θα εντοπίσει ακόμα πιο συγκεκριμένα patterns τα οποία σχετίζονται με τη συμπεριφορά του DL_bitrate στον χρόνο. Επιπλέον, υπολογίζονται **rolling means** για επιλεγμένες μεταβλητές (**RSRP, RSRQ, SNR, RSSI, Speed, DL_bitrate, UL_bitrate**) σε χρονικά παράθυρα **10 και 30 λεπτών**, με σκοπό να αποτυπωθούν τοπικές τάσεις και ομαλές μεταβολές. Τέλος, εφαρμόζεται ο μετασχηματισμός **log1p** στις μεταβλητές **DL_bitrate** και **UL_bitrate** για τη μείωση της ασυμμετρίας και τη σύγκλιση των κλιμάκων τους με τα υπόλοιπα χαρακτηριστικά του μοντέλου.

Για την κατανόηση της χρονικής εξάρτησης των χαρακτηριστικών, σχεδιάστηκαν τα **Autocorrelation Functions (ACF)** για όλα τα αριθμητικά features του dataset. Το ACF κάθε χαρακτηριστικού υπολογίστηκε και απεικονίστηκε για έως και 1000 **lags**, επιτρέποντας την ανάλυση της συστηματικής συσχέτισης κάθε χαρακτηριστικού με παλαιότερες χρονικές του τιμές.




Από τα διαγράμματα ACF παρατηρείται ότι όλες οι μεταβλητές εμφανίζουν σημαντική αυτοσυσχέτιση για μεγάλα χρονικά διαστήματα, φτάνοντας έως και περίπου 800 δευτερόλεπτα στο παρελθόν. Αυτό υποδηλώνει ότι η ιστορική πληροφορία των μεταβλητών έχει ουσιαστική συνεισφορά και μπορεί να βελτιώσει την πρόβλεψη του **DL_bitrate**. Ιδιαίτερα χαρακτηριστικά όπως το **Speed**, το **ServingCell_Distance** και το **Categorical_Impact** παρουσιάζουν βραδεία φθορά στη συσχέτιση, γεγονός που ευνοεί τη χρήση μεγαλύτερων χρονικών παραθύρων (time windows) κατά τη δημιουργία χαρακτηριστικών. Με βάση αυτή την παρατήρηση, επιλέχθηκαν να δειγματοληπτηθούν τιμές lags ανά 1 λεπτό (κάθε 60 δευτερόλεπτα), καθώς η πληροφορία δεν αλλάζει ουσιαστικά σε μικρότερα χρονικά βήματα.

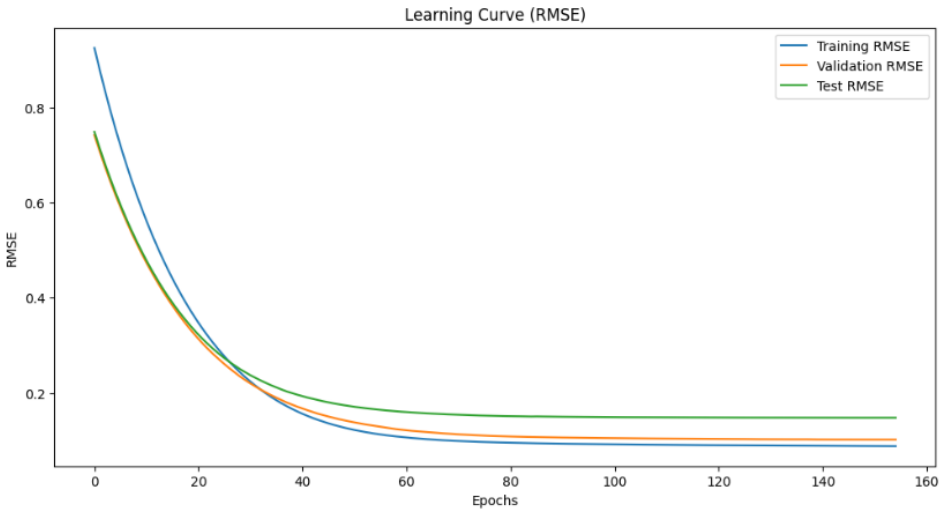
Για να εντοπιστεί ο βέλτιστος συνδυασμός υπερπαραμέτρων για το **XGBoost Regressor**, δημιουργούνται lags με χρονικό βήμα 60 δευτερολέπτων και δοκιμάζονται διαφορετικά πλήθη (5, 10, 20, 30), ώστε να ληφθεί υπόψη η χρονική εξάρτηση των μεταβλητών. Το σύνολο εκπαίδευσης διασπάται σε σύνολο εκπαίδευσης (60%) και επικύρωσης (20%). Τα δεδομένα όλων των συνόλων (train, validation, test) κανονικοποιούνται με χρήση **MinMaxScaler**. Στη συνέχεια πραγματοποιείται **grid search** σε διάφορους συνδυασμούς υπερπαραμέτρων (**n_estimators**, **max_depth**, **learning_rate**) για την εκπαίδευση του μοντέλου και αξιολογείται η επίδοση στο σύνολο επικύρωσης με την χρήση **Mean Absolute Error (MAE)**. Τέλος, καταγράφεται και εμφανίζεται ο βέλτιστος συνδυασμός παραμέτρων και αριθμού lags που οδηγεί στο χαμηλότερο σφάλμα πρόβλεψης.

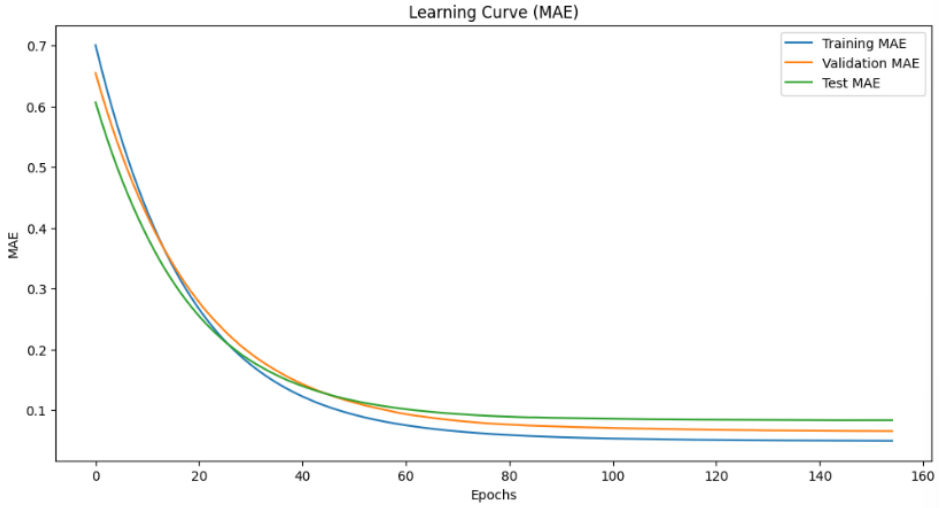
**Αποτελέσματα Grid Serch**

```
--- Best Configuration Based on Validation MAE ---
Num Lags     : 30.0
n_estimators  : 300.0
max_depth    : 5
learning_rate : 0.05
Validation MAE: 0.21
```

Στην συνέχεια, μετά την εκτέλεση του **Grid Search**, προστίθεται το βέλτιστο πλήθος **χρονικών καθυστερήσεων (lags)** στα σύνολα εκπαίδευσης και δοκιμής. Ο βέλτιστος αριθμός lag που προέκυψε ήταν ίσος με 30 και, δεδομένου ότι ο ρυθμός δειγματοληψίας (**lag_interval**) είναι **60 δευτερόλεπτα**, αυτό αντιστοιχεί σε 30 λεπτά παρελθοντικής πληροφορίας. Για κάθε χαρακτηριστικό και για κάθε lag, δημιουργούνται νέες στήλες οι οποίες περιέχουν την αντίστοιχη τιμή του χαρακτηριστικού σε παρελθοντικό χρόνο. Μετά την προσθήκη των lagged χαρακτηριστικών αφαιρούνται οι γραμμές που περιέχουν **NaN** τιμές, οι οποίες προκύπτουν λόγω του χρονικού μετατοπισμού (**shift**) κυρίως στις αρχικές γραμμές όπου δεν υπάρχουν διαθέσιμες παρελθοντικές τιμές

Ύστερα, αφού έχουν εντοπιστεί οι βέλτιστες υπερπαράμετροι και έχει ολοκληρωθεί το feature engineering υλοποιείται το τελικό μοντέλο XGBoost Regressor. Τα χαρακτηριστικά διαχωρίζονται στην είσοδο **X_full** η οποία περιλαμβάνει όλα τα πεδία εκτός από το **Timestamp** και το **DL_bitrate**, και στον στόχο **y_full_raw** που αντιστοιχεί στη μεταβλητή-στόχο **DL_bitrate**. Το σύνολο εκπαίδευσης χωρίζεται σε **training (60%)** και **validation (20%)**, ενώ το **test set** (20%) φορτώνεται ξεχωριστά. Εφαρμόζεται κανονικοποίηση με **MinMaxScaler**, τόσο στα χαρακτηριστικά εσόδου όσο και στην έξοδο, ώστε οι τιμές τους να ανήκουν στο διάστημα **[0,1]** και να διασφαλιστεί ότι τα μοντέλα δεν θα επηρεαστούν από διαφορετικές κλίμακες των χαρακτηριστικών και θα συγκλίνουν πιο γρήγορα κατά την εκπαίδευση. Η κανονικοποίηση γίνεται με **fit** μόνο στο training set και **transform** στα υπόλοιπα, για να αποφευχθεί data leakage. Ο αλγόριθμος εκπαιδεύεται με την υλοποίηση του **XGBoost μέσω DMatrix**, με τις βέλτιστες υπερπαραμέτρους **(max_depth=3, learning_rate=0.05, n_estimators=300**) και χρησιμοποιείται **early stopping** για να αποτραπεί το overfitting. Κατά τη διάρκεια της εκπαίδευσης αποθηκεύονται και απεικονίζονται οι μετρικές **RMSE** και **MAE** μέσω των **learning curves.** Μετά την εκπαίδευση, το μοντέλο πραγματοποιεί προβλέψεις με την μέθοδο **predict()** στα σύνολα training, validation, test. Οι προβλεπόμενες και οι πραγματικές τιμές επαναμετασχηματίζονται στην αρχική τους κλίμακα μέσω **inverse transform και expm1**, ώστε οι τελικές τιμές να βρίσκονται στην πραγματική τους κλίμακα. Τέλος, οι πραγματικές και οι προβλεπόμενες τιμές συγκεντρώνονται στο DataFrame **predictions_df**.






Από τα learning curves παρατηρείται ότι το σφάλμα μειώνεται απότομα κατά τις πρώτες 20 εποχές εκπαίδευσης και στη συνέχεια σταθεροποιείται σε πολύ χαμηλά επίπεδα, κοντά στο μηδέν. Το γεγονός ότι οι καμπύλες του **training** και του **validation error** σχεδόν ταυτίζονται υποδηλώνει ότι **δεν εμφανίζεται overfitting** και ότι το μοντέλο **γενικεύει αποτελεσματικά**. Επιπλέον, το **test error** διατηρείται επίσης πολύ κοντά στο validation error, παρουσιάζοντας μόνο ελαφρώς μεγαλύτερες τιμές, κάτι που είναι αναμενόμενο και αποδεκτό. Συνολικά, τα αποτελέσματα καταδεικνύουν ότι το μοντέλο έχει εκπαιδευτεί σωστά και έχει ισχυρή γενίκευση σε άγνωστα δεδομένα.

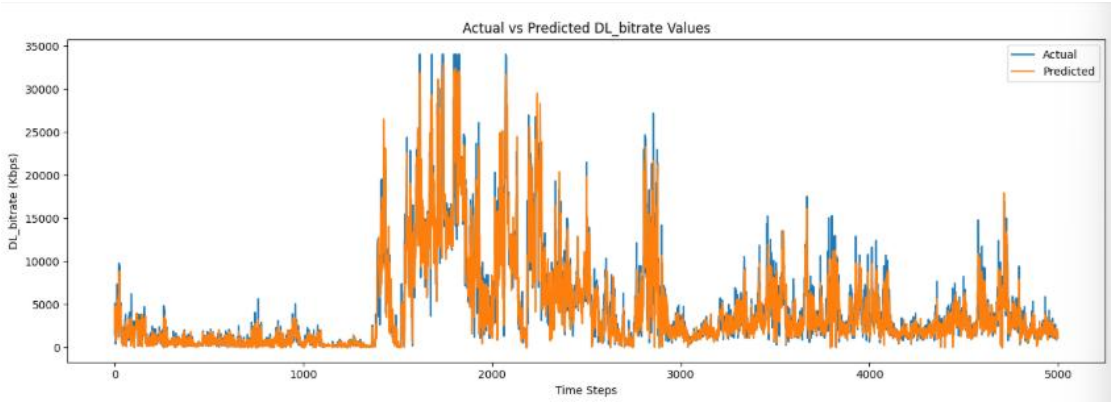
**Μετρικές σφάλματος (kbps)**

```
Train RMSE: 1432.33
Train MAE : 781.86
----------------------------------------
Validation RMSE: 4472.76
Validation MAE : 2815.53
----------------------------------------
Test RMSE: 6810.02
Test MAE : 2871.10

Mean actual DL_bitrate in test set: 11951.01 Kbps
Relative MAE: 24.02%
Relative RMSE: 56.98%
```

Ένα αποδεκτό ποσοστό απόκλισης του **MAE** για τη πρόβλεψη του **DL_bitrate** μετά από έρευνα που πραγματοποιήσαμε σε αντίστοιχο paper θεωρείται ότι είναι **κοντά στο 25%-30%**, καθώς σαν feature παρουσιάζει αρκετά υψηλό variance και παρουσιάζει σχετικά απρόβλεπτες συμπεριφορές.

Τέλος, για την οπτικοποίηση των αποτελεσμάτων του μοντέλου, απεικονίζονται οι **πρώτες 5000 τιμές** του συνόλου δοκιμής (**test set**), συγκρίνοντας τις **πραγματικές** με τις **προβλεπόμενες** τιμές του **DL_bitrate**.

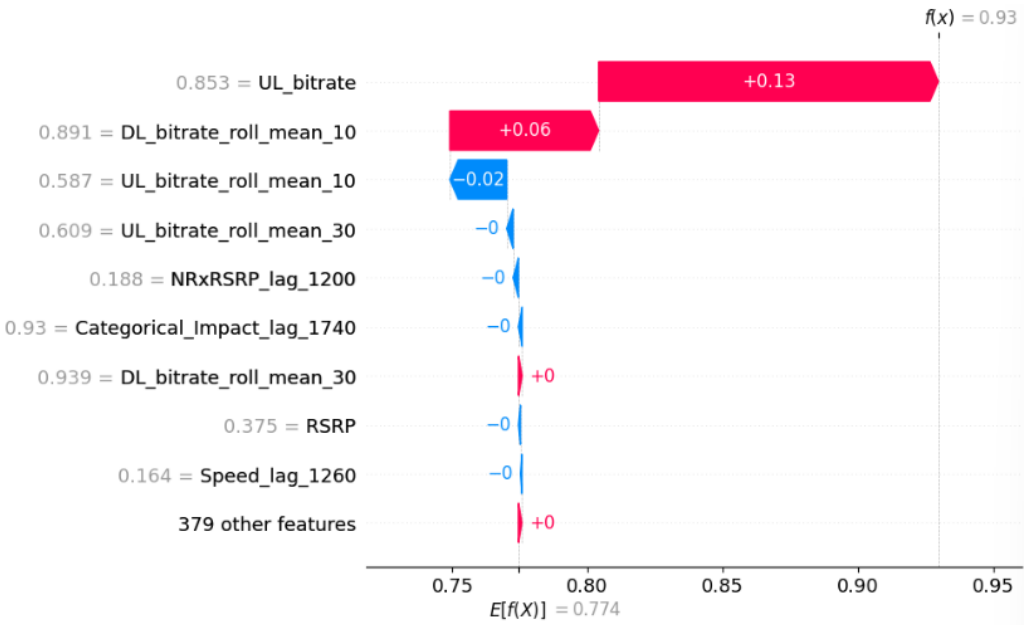

Actual vs Predicted DL_bitrate Values
Actual
Predicted
DL_bitrate (Kbps)
Time Steps
0
5000
10000
15000
20000
25000
30000
35000
0
1000
2000
3000
4000
5000

# 5 ΥΛΟΠΟΙΗΣΗ XAI

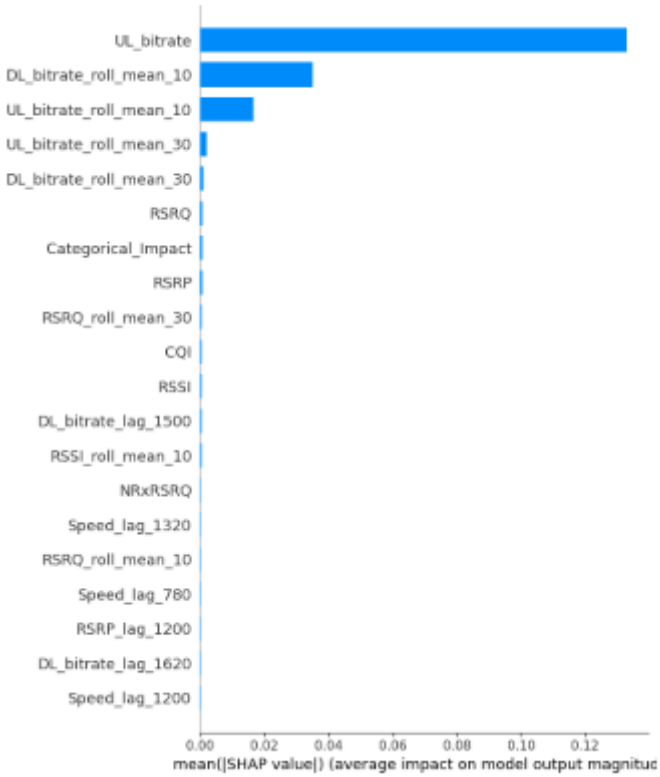
## 5.1 SHAP Plots

Εφαρμόστηκε SHAP για την ερμηνεία του XGBoost μοντέλου, προσφέροντας λεπτομερή ανάλυση του πώς κάθε χαρακτηριστικό συμβάλλει στην τελική πρόβλεψη. Αρχικά, το **X_test** μετατρέπεται σε DataFrame ώστε να διατηρηθούν τα ονόματα των χαρακτηριστικών, και στη συνέχεια δημιουργείται ένας **SHAP Explainer**, ο οποίος λαμβάνει ως είσοδο το εκπαιδευμένο μοντέλο (**bst**) και το **test set**. Ο Explainer υπολογίζει τα **SHAP values**, δηλαδή τις επιμέρους συνεισφορές κάθε χαρακτηριστικού στην έξοδο του μοντέλου για κάθε δείγμα. Τα αποτελέσματα οπτικοποιήθηκαν χρησιμοποιώντας το waterfall plot, το bar plot και το beeswarm plot.

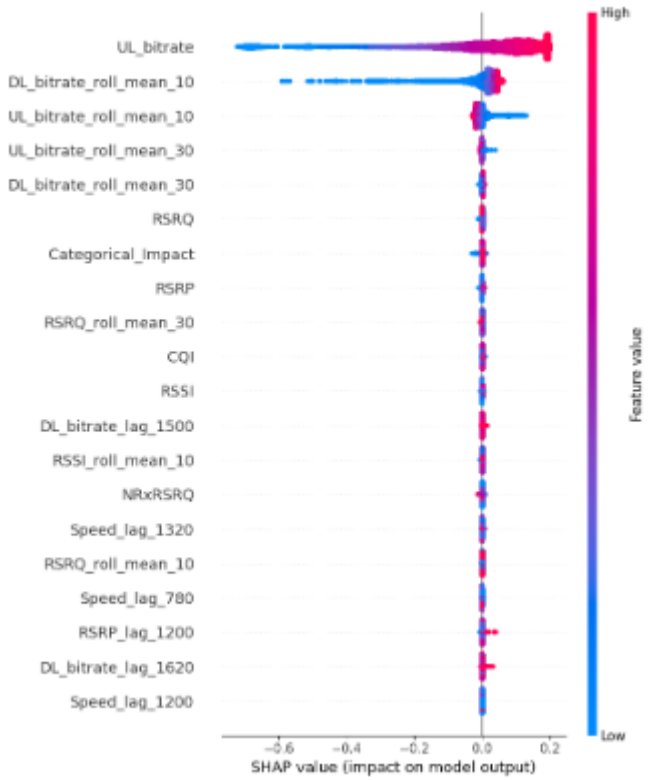
Το **Waterfall Plot** παρουσιάζει αναλυτικά τη συμβολή κάθε χαρακτηριστικού σε μία συγκεκριμένη πρόβλεψη του μοντέλου, ξεκινώντας από τη **μέση προβλεπόμενη τιμή** του XGBoost (**E[f(x)]**, baseline) και την αυξάνει ή την μειώνει για να καταλήξει στο τελικό output. Οι **κόκκινες μπάρες** αντιστοιχούν σε χαρακτηριστικά που αύξησαν την τιμή της πρόβλεψης, ενώ οι **μπλε** σε αυτά που τη μείωσαν. Από την ανάλυση, φαίνεται πως το μοντέλο στηρίζεται κυρίως στην **τρέχουσα τιμή του UL_bitrate**, γεγονός που είναι αναμενόμενο καθώς συχνά υπάρχει συσχέτιση μεταξύ uplink και downlink λόγω συνθηκών του καναλιού ή της απόστασης από τη βάση. Το **rolling mean του DL_bitrate (τελευταίων 10 λεπτών)** εμφανίζει μέτρια επίδραση, υποδηλώνοντας ότι το ιστορικό του DL_bitrate είναι χρήσιμο για την πρόβλεψη της τρέχουσας τιμής. Άλλα χαρακτηριστικά, όπως το **UL_bitrate_roll_mean_10** και το **UL_bitrate_roll_mean_30**, είχαν μικρότερη επίδραση. Αν και συνολικά υπήρχαν **379 πρόσθετα χαρακτηριστικά**, η συλλογική τους επίδραση ήταν αμελητέα σε αυτή την πρόβλεψη.




Το **Bar Plot** δείχνει ποια χαρακτηριστικά είχαν τη μεγαλύτερη συνολική επίδραση στις προβλέψεις του μοντέλου. Συγκεκριμένα, η **μέση απόλυτη τιμή SHAP (mean |SHAP value|)** υπολογίζεται για κάθε feature και αποτυπώνει το μέγεθος της επιρροής του χαρακτηριστικού ανεξαρτήτως κατεύθυνσης (θετική ή αρνητική). Το χαρακτηριστικό **UL_bitrate** αναδεικνύεται ως το σημαντικότερο, με μέση SHAP τιμή περίπου **+0.15**, υποδηλώνοντας ότι το μοντέλο στηρίζεται σε μεγάλο βαθμό σε αυτό για την πρόβλεψη του **DL_bitrate**. Αυτό είναι αναμενόμενο, καθώς το υψηλό bitrate στο uplink συχνά συνοδεύεται από καλές συνθήκες σύνδεσης. Ακολουθεί το **rolling mean του DL_bitrate** (τελευταίων 10 λεπτών), το οποίο δείχνει την ύπαρξη αυτοσυσχέτισης στο σήμα — δηλαδή ότι η πρόσφατη ιστορία του DL_bitrate αποτελεί ισχυρό δείκτη για την τρέχουσα τιμή. Αντίθετα, η πλειοψηφία των lagged, rolling και χρονικών χαρακτηριστικών φαίνεται να έχει χαμηλή σημασία, ενδεχομένως λόγω πλεονασμού πληροφορίας ή κάλυψής τους από θόρυβο. Συνολικά, τα **υπόλοιπα 379 χαρακτηριστικά** συνεισφέρουν ελάχιστα (συνολική SHAP επίδραση ~+0.01), γεγονός που δείχνει ότι το μοντέλο εστιάζει κυρίως σε λίγες κύριες μεταβλητές.

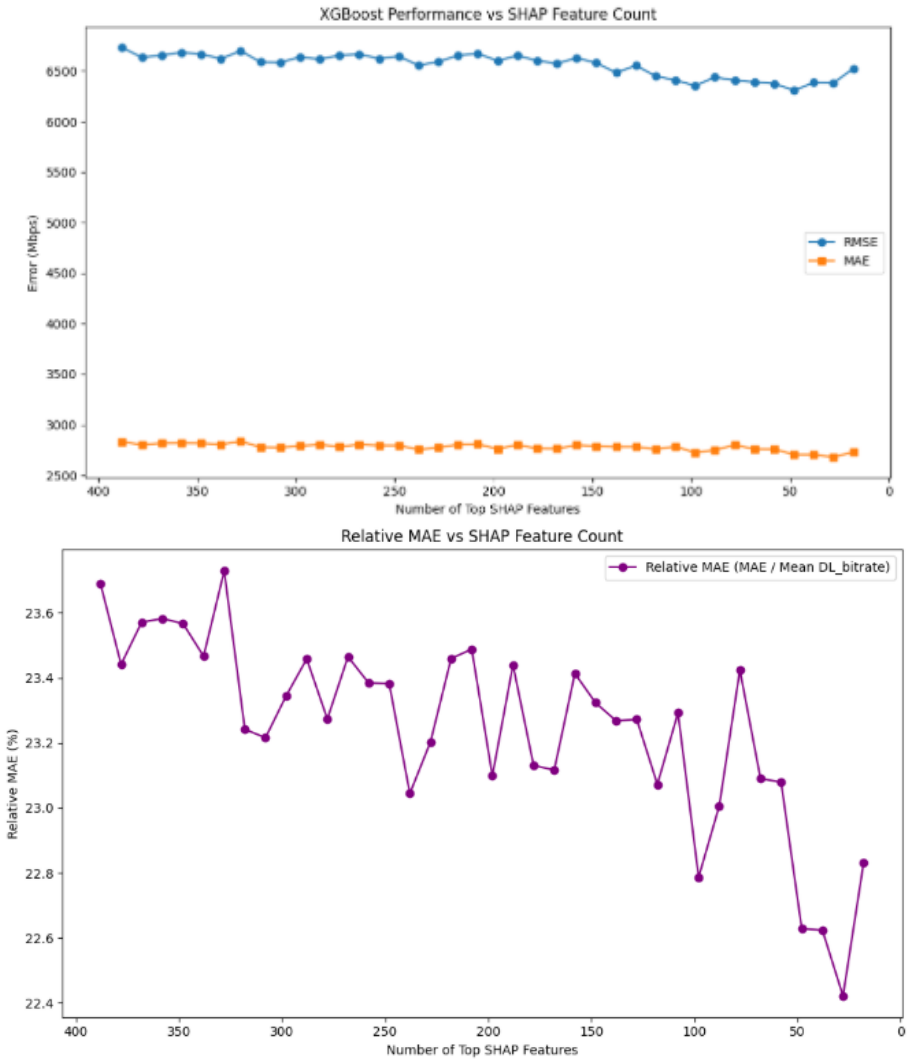



Το **Beeswarm Plot** παρουσιάζει τη διανομή των **SHAP values** για κάθε χαρακτηριστικό σε όλες τις παρατηρήσεις του dataset, απεικονίζοντας πώς οι διαφορετικές τιμές κάθε μεταβλητής επηρεάζουν τις προβλέψεις του μοντέλου. Στον άξονα Χ αποτυπώνεται το μέγεθος και η κατεύθυνση της επίδρασης κάθε χαρακτηριστικού: θετικές τιμές υποδηλώνουν ότι το χαρακτηριστικό αύξησε την πρόβλεψη του **DL_bitrate**, ενώ αρνητικές τιμές δείχνουν μείωση. Παράλληλα, η **χρωματική κωδικοποίηση** (κόκκινο = υψηλή τιμή, μπλε = χαμηλή τιμή) δείχνει πώς οι διαφορετικές τιμές του χαρακτηριστικού συσχετίζονται με την πρόβλεψη. Από το γράφημα προκύπτει ότι το μοντέλο στηρίζεται κυρίως στο **UL_bitrate** και στους **rolling μέσους όρους** του DL και του UL, καθώς αυτά τα χαρακτηριστικά εμφανίζουν τη μεγαλύτερη μεταβλητότητα και επιρροή στις προβλέψεις. Οι **rolling means** φαίνεται να παρέχουν σταθεροποιητική πληροφορία για την τάση του σήματος, ενώ τα υπόλοιπα 379 χαρακτηριστικά έχουν σχεδόν μηδενική συμβολή, γεγονός που υποδεικνύει ότι το μοντέλο τελικά βασίζεται σε **πολύ λίγες, ουσιαστικές μεταβλητές**.



## 5.2 Αφαίρεση features

Αρχικά, υπολογίζονται τα **SHAP values** για το test set και εξάγεται η **μέση απόλυτη τιμή** τους για κάθε χαρακτηριστικό, ώστε να προκύψει η κατάταξή τους με βάση τη συνολική επίδραση στο μοντέλο. Στη συνέχεια, εφαρμόζεται μια δοκιμαστική διαδικασία επανεκπαίδευσης: σε κάθε επανάληψη, επιλέγονται τα N πιο σημαντικά χαρακτηριστικά (ξεκινώντας από όλα και μειώνοντας ανά βήμα **step = 10**) και εκπαιδεύεται εκ νέου το μοντέλο μόνο με αυτά τα features. Ο διαχωρισμός των δεδομένων γίνεται σε training (60%) και validation (20%), ενώ το test set παραμένει σταθερό. Μετά από κάθε εκπαίδευση, το μοντέλο κάνει πρόβλεψη στο test set, και οι προβλεπόμενες τιμές αντιστέφονται από την κανονικοποίηση και τον λογαριθμικό μετασχηματισμό. Υπολογίζονται οι μετρικές απόδοσης **RMSE** και **MAE** για κάθε αριθμό χαρακτηριστικών και καταγράφονται. Η απόδοση του μοντέλου απεικονίζεται σε γράφημα, όπου φαίνεται η μεταβολή των σφαλμάτων ως προς το πλήθος των SHAP-selected χαρακτηριστικών.

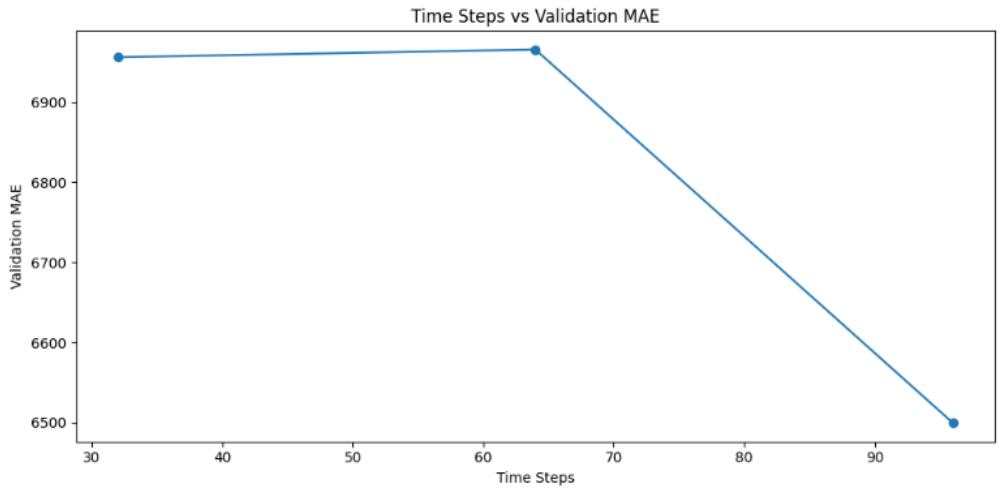



**Συμπεράσματα**:
Παρατηρείται ότι καθώς μειώνεται ο αριθμός των χαρακτηριστικών από περίπου 400 προς τα 50–30, το σφάλμα είτε παραμένει σταθερό είτε μειώνεται ελαφρώς, υποδεικνύοντας ότι πολλές από τις αρχικές μεταβλητές έχουν χαμηλή ή πλεονάζουσα πληροφορία. Τοπικά ελάχιστα των μετρικών RMSE και MAE εμφανίζονται όταν χρησιμοποιούνται περίπου **50 έως 70 χαρακτηριστικά**, γεγονός που υποδηλώνει ότι το μοντέλο μπορεί να επιτύχει υψηλή ακρίβεια ακόμη και με σημαντική μείωση της διαστατικότητας. Αντίθετα, όταν ο αριθμός των χαρακτηριστικών πέσει κάτω από τις 30 μεταβλητές, το RMSE αυξάνεται απότομα, δείχνοντας ότι αφαιρούνται κρίσιμες πληροφορίες. Από την άλλη πλευρά, όταν χρησιμοποιούνται περισσότερα από 300 χαρακτηριστικά, η απόδοση του μοντέλου δεν βελτιώνεται, γεγονός που ενδεχομένως οφείλεται σε υπερεκπαίδευση (overfitting**)** ή παρουσία περιττών χαρακτηριστικών. Συνολικά, το ιδανικό πλήθος χαρακτηριστικών για το μοντέλο φαίνεται να κυμαίνεται μεταξύ 30 και 70, εύρος στο οποίο το MAE είναι ελάχιστο και το RMSE διατηρείται σταθερά χαμηλό.

# 6 ΥΛΟΠΟΙΗΣΗ LSTM

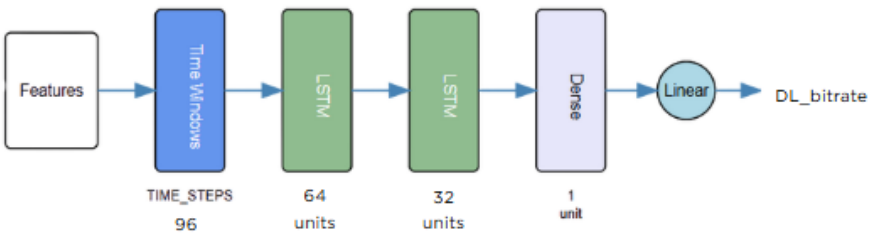
## 6.1 Fine Tunning

Αρχικά, χρησιμοποιείται η συνάρτηση **create_dataset()** για τη δημιουργία παραθύρων δεδομένων εισόδου και εξόδου. Τα χαρακτηριστικά διαχωρίζονται στην είσοδο **X_full** η οποία περιλαμβάνει όλα τα πεδία εκτός από το **Timestamp** και το **DL_bitrate**, και στον στόχο **y_full** που αντιστοιχεί στη μεταβλητή-στόχο **DL_bitrate**. Το σύνολο εκπαίδευσης χωρίζεται σε **training (60%)** και **validation (20%)**, ενώ το **test set** (20%) φορτώνεται ξεχωριστά. Εφαρμόζεται κανονικοποίηση με **MinMaxScaler**, τόσο στα χαρακτηριστικά εσόδου όσο και στην έξοδο, ώστε οι τιμές τους να ανήκουν στο διάστημα **[0,1]** και να διασφαλιστεί ότι τα μοντέλα δεν θα επηρεαστούν από διαφορετικές κλίμακες των χαρακτηριστικών και θα συγκλίνουν πιο γρήγορα κατά την εκπαίδευση. Η κανονικοποίηση γίνεται με **fit** μόνο στο training set και **transform** στα υπόλοιπα, για να αποφευχθεί data leakage. Επίσης, εφαρμόζεται ο μετασχηματισμός **log1p** στις μεταβλητές **DL_bitrate** και **UL_bitrate** για τη μείωση της ασυμμετρίας και τη σύγκλιση των κλιμάκων τους με τα υπόλοιπα χαρακτηριστικά. Το μοντέλο LSTM δύο επιπέδων (64 και 32 μονάδες αντίστοιχα), εκπαιδεύεται για διάφορα πλήθη χρονικών παραθύρων (32, 64, 96). Επιπλέον, χρησιμοποιείται **EarlyStopping** για αποφυγή υπερπροσαρμογής, και το μοντέλο αξιολογείται με το Mean Absolute Error (MAE) αφού γίνει επαναφορά στην αρχική κλίμακα τιμών. Από την παραπάνω διαδικασία θα προκύψει το βέλτιστο πλήθος χρονικών παραθύρων.



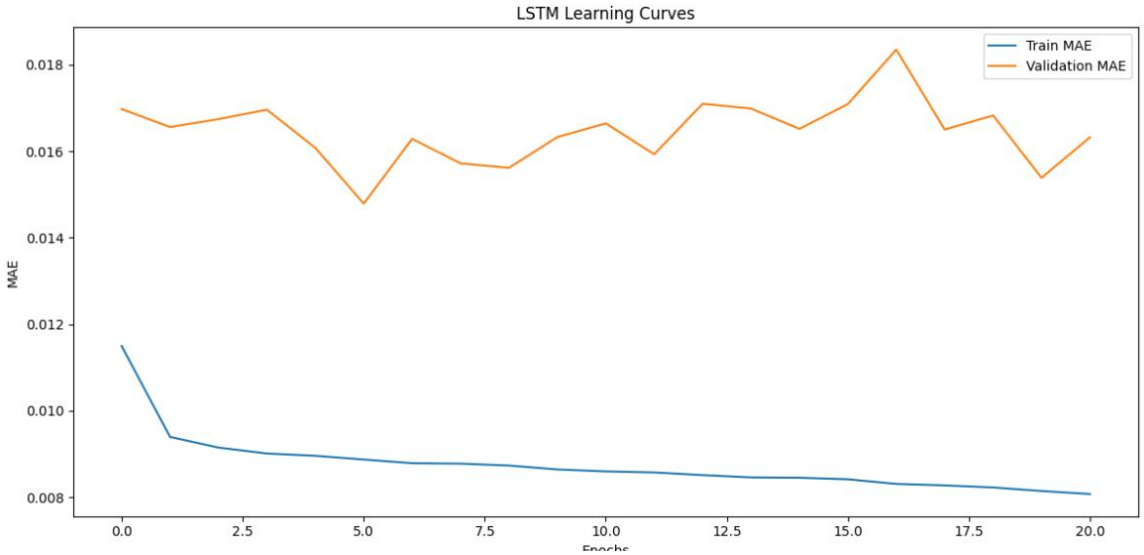

Παρατηρείται, ότι η χρήση **96 time steps** οδήγησε στη χαμηλότερη τιμή του σφάλματος μέσης απόλυτης τιμής (MAE) και γι'αυτό επιλέχθηκε ως η βέλτιστη τιμή για την τελική εκπαίδευση.

Η τελική αρχιτεκτονική του μοντέλου είναι:




Στην συνέχεια, πραγματοποιείται πάλι η εκπαίδευση του μοντέλου LSTM δύο επιπέδων (64 και 32 μονάδες αντίστοιχα) ακολουθώντας την διαδικασία που αναφέρθηκε παραπάνω για 96 χρονικά παράθυρα. Κατά τη διάρκεια της εκπαίδευσης αποθηκεύονται και απεικονίζονται οι μετρικές **RMSE** και **MAE** μέσω των **learning curves**. Μετά την εκπαίδευση που πραγματοποιείται με την μέθοδο **fit()**, το μοντέλο πραγματοποιεί προβλέψεις με την μέθοδο **predict()** στα σύνολα training, validation, test. Οι προβλεπόμενες και οι πραγματικές τιμές επαναμετασχηματίζονται στην αρχική τους κλίμακα μέσω **inverse transform και expm1**, ώστε οι τελικές τιμές

να βρίσκονται στην πραγματική τους κλίμακα. Τέλος, οι πραγματικές και οι προβλεπόμενες τιμές συγκεντρώνονται στο DataFrame **predictions_df**.



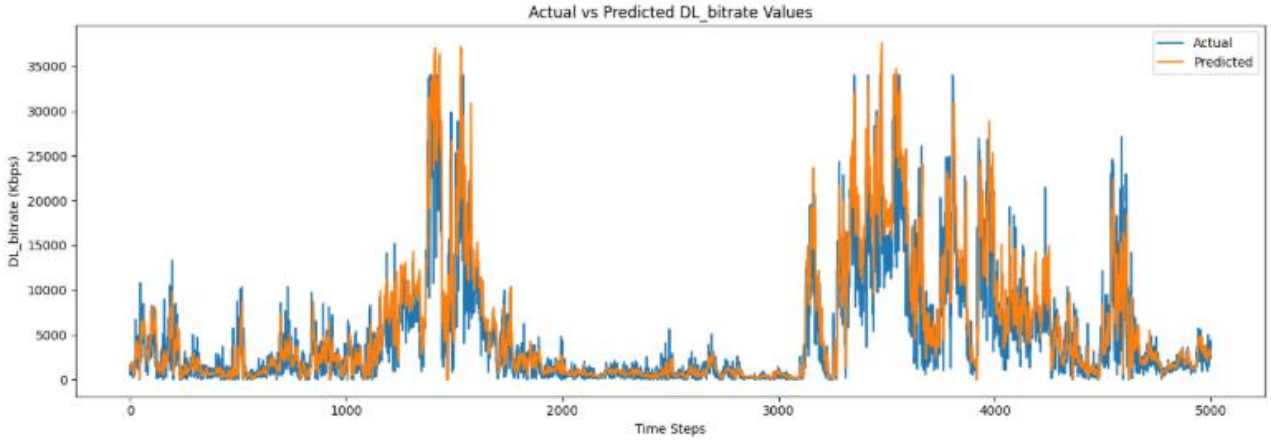

Από τα learning curves παρατηρείται ότι η εκπαίδευση δεν είναι σταθερή και εμφανίζεται overfitting. Συγκεκριμένα, ενώ το training loss μειώνεται σταθερά, το validation loss παρουσιάζει αυξομειώσεις, γεγονός που δείχνει ότι το μοντέλο μαθαίνει υπερβολικά καλά το training set εις βάρος της γενίκευσης. Αυτό υποδηλώνει πως, παρότι τα 96 time steps βελτιώνουν αριθμητικά το MAE, η γενική συμπεριφορά του μοντέλου δεν είναι ιδανική.

**Μετρικές σφάλματος (kbps)**

```
Train RMSE: 3941.90
Train MAE : 2273.09
----------------------------------------
Validation RMSE: 9273.09
Validation MAE : 6439.79
----------------------------------------
Test RMSE: 13667.70
Test MAE : 8336.65

Mean actual DL_bitrate in test set: 12251.24 Kbps
Relative MAE: 68.05%
Relative RMSE: 111.56%
```

Τέλος, για την οπτικοποίηση των αποτελεσμάτων του μοντέλου, απεικονίζονται οι **πρώτες 5000 τιμές** του συνόλου δοκιμής (**test set**), συγκρίνοντας τις **πραγματικές** με τις **προβλεπόμενες** τιμές του **DL_bitrate**.

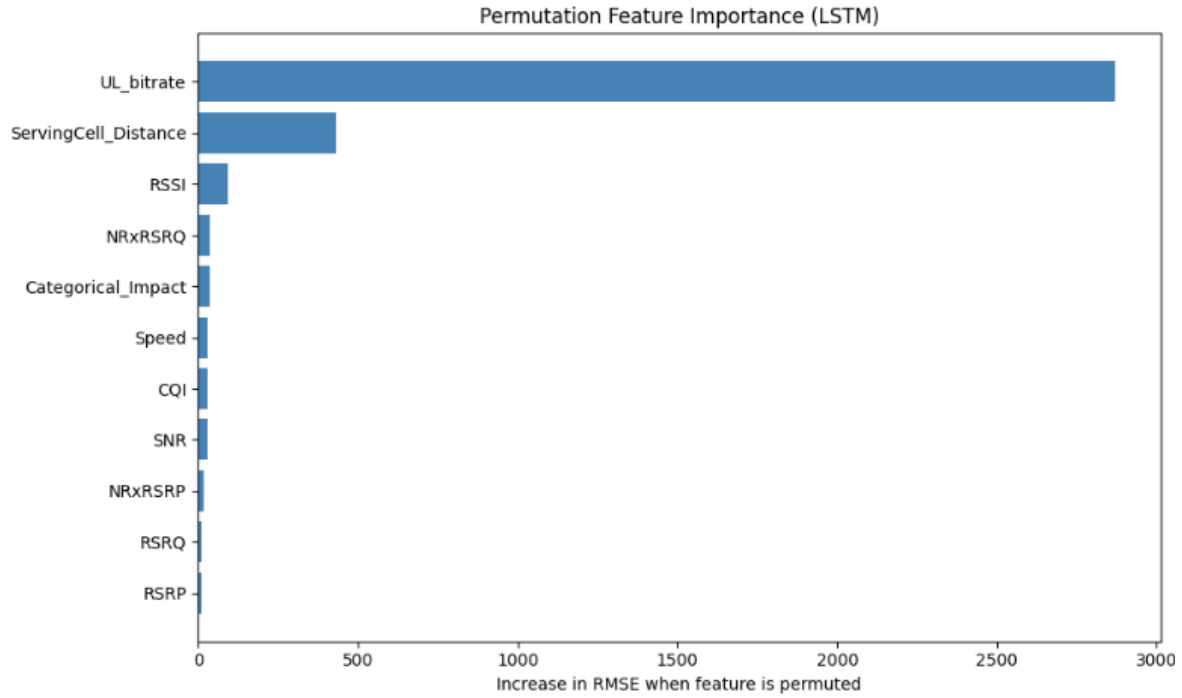


# 7 LSTM EXPLAINABILITY

## 7.1 Permutation Feature Importance

Εφαρμόζεται η τεχνική **Permutation Feature Importance** στο μοντέλο LSTM, με σκοπό να μετρήσει τη συμβολή κάθε χαρακτηριστικού στην ακρίβεια πρόβλεψης. Αρχικά, εξάγονται οι πραγματικές (**y_test_true_full**) και προβλεπόμενες (**y_test_pred_full**) τιμές του **DL_bitrate** από το **predictions_df** και υπολογίζεται το αρχικό σφάλμα RMSE, το οποίο θα χρησιμεύσει ως σημείο αναφοράς (baseline). Στη συνέχεια, η είσοδος **X_test_scaled** μετασχηματίζεται σε τρισδιάστατη μορφή (samples × time_steps × features), ώστε να ευθυγραμμίζεται με την απαίτηση του LSTM μοντέλου για σειριακή είσοδο. Για κάθε χαρακτηριστικό ξεχωριστά, γίνεται αναδιάταξη των τιμών του κατά μήκος όλων των χρονικών βημάτων, ώστε να διακοπεί η πληροφορία που μεταφέρει. Το νέο σύνολο εισόδου χρησιμοποιείται για πρόβλεψη από το μοντέλο, και υπολογίζεται εκ νέου το RMSE. Η διαφορά μεταξύ του νέου και του baseline RMSE εκφράζει τη σημασία του χαρακτηριστικού: όσο μεγαλύτερη η αύξηση του σφάλματος, τόσο πιο κρίσιμο θεωρείται το χαρακτηριστικό για την πρόβλεψη. Τέλος, οι τιμές αποθηκεύονται σε DataFrame, ταξινομούνται σε φθίνουσα σειρά και απεικονίζονται με bar plot.

```
Feature: Speed, RMSE Δ: 30.6410
Feature: RSRP, RMSE Δ: 11.3198
Feature: RSRQ, RMSE Δ: 12.0713
Feature: SNR, RMSE Δ: 29.7276
Feature: CQI, RMSE Δ: 30.4161
Feature: RSSI, RMSE Δ: 92.7021
Feature: UL_bitrate, RMSE Δ: 2869.3361
Feature: NRxRSRP, RMSE Δ: 18.2986
Feature: NRxRSRQ, RMSE Δ: 36.6098
Feature: ServingCell_Distance, RMSE Δ: 432.1035
Feature: Categorical_Impact, RMSE Δ: 35.2499
```




Παρατηρείται ότι το **UL_bitrate** επηρεάζει περισσότερο το μοντέλο, καθώς η τυχαία αναδιάταξή του προκαλεί δραματική αύξηση του RMSE (πάνω από 2800 μονάδες), υποδεικνύοντας ισχυρή συσχέτιση με την τιμή του DL_bitrate. Το **ServingCell_Distance** εμφανίζει επίσης σημαντική συνεισφορά στο μοντέλο, καθώς η τυχαία αναδιάταξή του οδηγεί σε αύξηση του σφάλματος (RMSE). Αυτό υποδηλώνει ότι η απόσταση από τον

εξυπηρετητή σταθμό βάσης (cell tower) επηρεάζει την απόδοση του (**DL bitrate**). Τέλος, τα χαρακτηριστικά όπως τα **RSRP, RSRQ, SNR, CQI** έχουν μικρότερη επίδραση.

## 7.2 Αφαίρεση features

Ύστερα, ξεκινώντας από το πλήρες σύνολο χαρακτηριστικών (ταξινομημένα βάσει Permutation Feature Importance), αφαιρείται το λιγότερο σημαντικό χαρακτηριστικό και το LSTM επανεκπαιδεύεται με τα εναπομείναντα. Για κάθε επανάληψη δημιουργούνται νέοι πίνακες **train/val/test** μόνο τα επιλεγμένα χαρακτηριστικά, εφαρμόζεται κανονικοποίηση (MinMax), δημιουργούνται time windows (με τη **create_dataset**), και εκπαιδεύεται εκ νέου το μοντέλο LSTM. Στη συνέχεια πραγματοποιείται πρόβλεψη στο test set και υπολογίζεται το σφάλμα RMSE. Τέλος, απεικονίζεται γραφικά η σχέση μεταξύ αριθμού χαρακτηριστικών και απόδοσης του μοντέλου.


Παρατηρείται ότι το μικρότερο RMSE επιτυγχάνεται όταν διατηρούνται όλα τα χαρακτηριστικά (11), υποδεικνύοντας ότι η αφαίρεση έστω και ενός επιφέρει απώλεια πληροφορίας. Ωστόσο, η απόδοση παραμένει συγκρίσιμα καλή και με λιγότερα χαρακτηριστικά (8 ή 9), γεγονός που δείχνει ότι το μοντέλο μπορεί να διατηρήσει ικανοποιητική ακρίβεια ακόμη και με μικρότερο αριθμό εισόδων, εφόσον διατηρηθούν οι πιο σημαντικές.

# 8 ΣΥΜΠΕΡΑΣΜΑΤΑ

### XGBoost Error Metrics

| | Train | Validation | Test | Σφάλματα ως ποσοστό της μέσης τιμής του DL_bitrate στο Testing Set |
|---|---|---|---|---|
| **RMSE** | 1432.33 | 4472.76 | 6810.02 | 56.98% |
| **MAE** | 781.86 | 2815.53 | 2871.10 | **24.02%** |

### LSTM Error Metrics

| | Train | Validation | Test | Σφάλματα ως ποσοστό της μέσης τιμής του DL_bitrate στο Testing Set |
|---|---|---|---|---|
| **RMSE** | 3717.66 | 10084.72 | 14104.63 | 115.13% |
| **MAE** | 2054.22 | 7169.90 | 8386.82 | 68.46% |

Με βάση τις τιμές των μετρικών RMSE και MAE, προκύπτει ξεκάθαρα ότι το μοντέλο XGBoost υπερέχει σημαντικά έναντι του LSTM τόσο σε ακρίβεια όσο και σε γενίκευση. Ειδικότερα, το MAE στο test set για το XGBoost ανέρχεται μόλις στο 24.02% της μέσης τιμής του DL_bitrate, έναντι 68.46% για το LSTM, κάτι που αποδεικνύει ότι το XGBoost κάνει μικρότερες απόλυτες προβλέψεις σφάλματος. Επιπλέον, η διαφορά μεταξύ training και test σφάλματος είναι πολύ μικρότερη στο XGBoost, γεγονός που φανερώνει καλύτερη γενίκευση και απουσία overfitting, ενώ το LSTM εμφανίζει μεγάλη απόκλιση μεταξύ train και validation/test set, κάτι που υποδηλώνει πιθανή υπερεκπαίδευση ή αδυναμία μοντελοποίησης της δυναμικής του σήματος. Τέλος, δεδομένου του υψηλού κόστους εκπαίδευσης των LSTM και της ανάγκης για μεγάλα σε όγκο και ποιότητα datasets, το XGBoost αποδεικνύεται στην πράξη πιο αποδοτικό και πιο σταθερό για προβλήματα με θορυβώδη, μη γραμμικά και διακοπτόμενα πρότυπα όπως αυτά του DL_bitrate.

# 9 ΠΗΓΕΣ


**3 ΠΡΟΕΠΕΞΕΡΓΑΣΙΑ**

**[1]** https://www.analyticsvidhya.com/blog/2021/05/feature-engineering-how-to-detect-and-remove-outliers-with-python-code/#h-iqr-based-filtering

**[2]** https://www.geeksforgeeks.org/how-to-handle-categorical-variables-in-regression/

**4 ΥΛΟΠΟΙΗΣΗ XGBOOST REGRESSOR**

**[3]** https://xgboost.readthedocs.io/en/latest/parameter.html

**5 ΥΛΟΠΟΙΗΣΗ XAI**

**[4]**https://shap.readthedocs.io/en/latest/example_notebooks/tabular_examples/tree_based_models/Front%20page%20example%20%28XGBoost%29.html

**[5]**https://shap.readthedocs.io/en/latest/example_notebooks/overviews/An%20introduction%20to%20explainable%20AI%20with%20Shapley%20values.html

**6 ΥΛΟΠΟΙΗΣΗ LSTM**

**[6]** https://www.geeksforgeeks.org/deep-learning-introduction-to-long-short-term-memory/

**8 ΣΥΜΠΕΡΑΣΜΑΤΑ**

**[7]** https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9880654