Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης

Σχολή Θετικών Επιστημών

Τμήμα Πληροφορικής

Πρόγραμμα Μεταπτυχιακών Σπουδών

Τεχνητή Νοημοσύνη

Μηχανική Μάθηση

Τελική Εργασία

Παναγιώτης Καραμητόπουλος

Α.Ε.Μ. 213

Email: karamitopp@ece.auth.gr

Διπλ. Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Α.Π.Θ.

Θεσσαλονίκη, 2026

Περιεχόμενα

Κατάλογος Πινάκων

Κατάλογος Σχημάτων

Εισαγωγή

Στα πλαίσια της παρούσας εργασίας αξιοποιήθηκαν, επεξεργάστηκαν και αναλύθηκαν τα δεδομένα από την πλατφόρμα Driven Data. Πιο συγκεκριμένα, σκοπός της εργασίας είναι η εφαρμογή τεχνικών μηχανικής και βαθιάς μάθησης για την επίλυση ενός προβλήματος παλινδρόμησης που αφορά την πρόβλεψη φτώχιας.

1 – Περιγραφή Δεδομένων

Αρχικά, φορτώνονται τα αρχεία train_hh_features.csv, train_hh_gt.csv και test_hh_features.csv, μεγέθους (104234, 88), (104234, 3) και (103023, 88) αντίστοιχα. Πιο συγκεκριμένα, τα 88 χαρακτηριστικά (στήλες) των αρχείων train_hh_features.csv και test_hh_features.csv είναι τα εξής:

1) hhid: Αναγνωριστικός κωδικός (id) ενός σπιτιού/οικογένειας.

2) com: Αναγνωριστικό (id) μέλους της οικογένειας.

3) weight: Δειγματοληπτικό βάρος της οικογένειας.

4) strata: Μεταβλητή διαστρωμάτωσης.

5) utl_exp_ppp17: Δαπάνες για υπηρεσίες κοινής ωφέλειας.

6) male: Ο αρχηγός της οικογένειας είναι άνδρας;

7) hsize: Αριθμός μελών της οικογένειας.

8) num_children5: Αριθμός παιδιών κάτω των 5 ετών εντός της οικογένειας.

9) num_children10: Αριθμός παιδιών ηλικίας 5 – 10 ετών εντός της οικογένειας.

10) num_children18: Αριθμός παιδιών ηλικίας 10 – 18 ετών εντός της οικογένειας.

11) age: Ηλικία του αρχηγού της οικογένειας.

12) owner: Ιδιοκτήτης κατοικίας;

13) water: Η κατοικία έχει πρόσβαση στο σύστημα ύδρευσης;

14) toilet: Υπάρχουν τουαλέτες στην κατοικία;

15) sewer: Οι τουαλέτες συνδέονται με το αποχετευτικό σύστημα;

16) elect: Η κατοικία έχει πρόσβαση στο σύστημα ηλεκτρικής ενέργειας;

17) water_source: Πηγές πόσιμου νερού.

18) sanitation_source: Κύρια πηγή αποχέτευσης.

19) dweltyp: Τύπος κατοικίας.

20) num_adult_female: Αριθμός ενηλίκων γυναικών ηλικίας 18 – 69 ετών εντός της οικογένειας.

21) num_adult_male: Αριθμός ενηλίκων ανδρών ηλικίας 18 – 69 ετών εντός της οικογένειας.

22) num_elderly: Αριθμός ηλικιωμένων ηλικίας 70+ ετών εντός της οικογένειας.

23) employed: Ο αρχηγός της οικογένειας είναι εργαζόμενος;

24) sworkershh: Ποσοστό ενηλίκων που εργάζονται στην οικογένεια.

25) share_secondary: Ποσοστό ενηλίκων αποφοίτων δευτεροβάθμιας εκπαίδευσης στην οικογένεια.

26) educ_max: Υψηλότερο επίπεδο εκπαίδευσης στην οικογένεια.

27) sfworkershh: Ποσοστό των ενηλίκων εργαζομένων σε επίσημη απασχόληση.

28) any_nonagric: Κάθε μέλος της οικογένειας που εργάζεται σε μη γεωργικό τομέα.

29) sector1d: Τομέας απασχόλησης του αρχηγού της οικογένειας.

30) region1: Γεωγραφικό χαρακτηριστικό.

31 – 35) ... Γεωγραφικά χαρακτηριστικά.

36) region7: Γεωγραφικό χαρακτηριστικό.

37) urban: Αστικός δείκτης - Γεωγραφικό χαρακτηριστικό.

38) consumed100: Ψωμί - Χαρακτηριστικό σχετικό με την κατανάλωση τροφίμων.

39 – 86) ... Χαρακτηριστικά σχετικά με την κατανάλωση τροφίμων.

87) consumed5000: Άλλα τρόφιμα - Χαρακτηριστικό σχετικό με την κατανάλωση τροφίμων.

88) survey_id: Αναγνωριστικός κωδικός (id) της έρευνας.

Παράλληλα, οι στήλες του αρχείου train_hh_gt.csv είναι οι εξής:

1) hhid: Αναγνωριστικός κωδικός (id) ενός σπιτιού/οικογένειας.

2) survey_id: Αναγνωριστικός κωδικός (id) της έρευνας.

3) cons_ppp17: Καθημερινές δαπάνες ανά μέλος της οικογένειας.

Στη συνέχεια, τα δεδομένα από τα αρχεία train_hh_features.csv και train_hh_gt.csv συγχωνεύονται με βάση το hhid, δηλαδή το id της οικογένειας, καθώς και το survey_id, δηλαδή το id της έρευνας. Αυτό γίνεται καθώς η συγχώνευση με βάση μόνο το hhid ενδεχομένως να οδηγούσε σε λανθασμένες αντιστοιχίσεις, αν μια οικογένεια έχει συμμετάσχει σε δύο διαφορετικές έρευνες (π.χ. σε διαφορετικά έτη). Επιπλέον, τέθηκε how='inner', ώστε να διατηρούνται μόνο οι εγγραφές που υπάρχουν και στα δύο αρχεία (π.χ. αν μια οικογένεια υπάρχει μόνο στο $1^{ο}$ αρχείο, αλλά όχι στο $2^{ο}$, τότε αυτήν η οικογένεια δεν θα διατηρηθεί στο συγχωνευμένο αρχείο, ισχύει και το αντίστροφο). Ως αποτέλεσμα, το νέο συγχωνευμένο σύνολο δεδομένων, περιλαμβάνει 104234 εγγραφές (οικογένειες) με 89 στήλες.

2 – Ανάλυση Δεδομένων

Σε αυτό το βήμα επεξεργάζονται τα δεδομένα του προβλήματος. Αρχικά, απαλείφονται τα περιττά χαρακτηριστικά και αναλύεται η μεταβλητή στόχος. Στη συνέχεια, μελετώνται οι συσχετίσεις των χαρακτηριστικών με τη μεταβλητή στόχο και συμπληρώνονται οι ελλιπείς τιμές. Τέλος, ακολουθεί ο μετασχηματισμός των δεδομένων.

2.1 Ανεξάρτητες Μεταβλητές

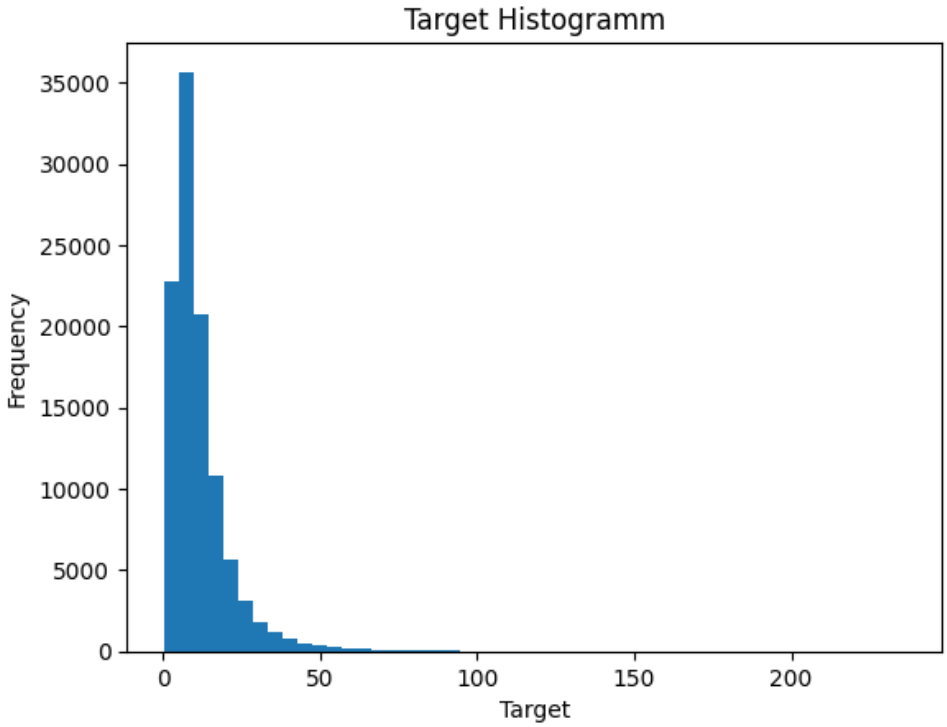
Από τις 89 στήλες του συγχωνευμένου συνόλου δεδομένων, ως ανεξάρτητες μεταβλητές θα χρησιμοποιηθούν όλες εκτός από τα ids (hhid, survey_id και com) αφού αυτά είναι περιττά και δεν έχουν κάποια συσχέτιση με τον στόχο, καθώς και τη μεταβλητή στόχο, cons_ppp17.

2.2 Μεταβλητή Στόχος

Όπως έχει ήδη ειπωθεί η μεταβλητή στόχος είναι η cons_ppp17. Τα στατιστικά στοιχεία καθώς και το ιστόγραμμα της μεταβλητής αυτής παρουσιάζονται στον Πίνακα 1 και στο Σχήμα 1 αντίστοιχα. Επομένως, προκύπτει ότι η κατανομή που ακολουθεί η μεταβλητή στόχος μοιάζει με την εκθετική κατανομή, καθώς παρουσιάζει έντονη ασυμμετρία.

	cons_ppp17
Count	104234
Mean	11.555229
Std	10.034225
Min	0.359563
25%	5.439294
50%	8.774002
75%	14.216931
Max	236.115680

Πίνακας 1: Στατιστικά στοιχεία μεταβλητής στόχου



Σχήμα 1: Ιστόγραμμα μεταβλητής στόχου

2.3 Συσχετίσεις

Στους Πίνακες 2 και 3 παρουσιάζονται τα 5 αριθμητικά χαρακτηριστικά με την υψηλότερη θετική και αρνητική συσχέτιση με τη μεταβλητή στόχο αντίστοιχα. Από τους πίνακες αυτούς, παρατηρείται ότι οι μεταβλητές utl_exp_ppp17 (δαπάνες για υπηρεσίες κοινής ωφέλειας) και sfworkershh (ποσοστό των ενηλίκων εργαζομένων σε επίσημη απασχόληση) έχουν τη μεγαλύτερη θετική συσχέτιση με τον στόχο. Αυτό σημαίνει ότι όσο αυξάνονται οι εν λόγω μεταβλητές, αυξάνονται και οι καθημερινές δαπάνες ανά μέλος της οικογένειας (στόχος). Από την άλλη πλευρά, οι μεταβλητές strata (χαρακτηριστικό διαστρωμάτωσης) και region5 (γεωγραφικό χαρακτηριστικό) παρουσιάζουν την υψηλότερη αρνητική συσχέτιση με τον στόχο.

Χαρακτηριστικό	Συσχέτιση με τον στόχο
utl_exp_ppp17	0.445324
sfworkershh	0.369261
region7	0.300099
region1	0.129630
region2	0.064078

Πίνακας 2: Αριθμητικά χαρακτηριστικά με την υψηλότερη θετική συσχέτιση με τη μεταβλητή στόχο

Χαρακτηριστικό	Συσχέτιση με τον στόχο
strata	-0.445995
region5	-0.309193
hsize	-0.272773
num_children18	-0.221311
num_children10	-0.203503

Πίνακας 3: Αριθμητικά χαρακτηριστικά με την υψηλότερη αρνητική συσχέτιση με τη μεταβλητή στόχο

2.4 Προεπεξεργασία Δεδομένων

Παρατηρείται ότι υπάρχουν 56 χαρακτηριστικά με ελλιπείς τιμές, ωστόσο το ποσοστό ελλιπών τιμών για τα 54 από τα 56 χαρακτηριστικά είναι μικρότερο του 0.1%. Τα δύο χαρακτηριστικά με το μεγαλύτερο ποσοστό ελλιπών τιμών είναι το sector1d (τομέας απασχόλησης του αρχηγού της οικογένειας) και το dweltyp (τύπος κατοικίας) με ποσοστά 13.56% και 1.16% αντίστοιχα. Οι ελλιπείς τιμές των αριθμητικών χαρακτηριστικών συμπληρώθηκαν με τη διάμεσο γιατί είναι πιο ανθεκτική στις ακραίες τιμές, ενώ οι ελλιπείς τιμές των κατηγορικών χαρακτηριστικών συμπληρώθηκαν με την πιο συχνή τιμή, καθώς αποτελεί το πιο πιθανό αντιπροσωπευτικό δείγμα για τις τιμές που λείπουν.

2.5 Μετασχηματισμός Δεδομένων

Επιπλέον, τα κατηγορικά χαρακτηριστικά κωδικοποιήθηκαν ώστε να μετατραπούν σε κατάλληλη μορφή για τους αλγορίθμους μάθησης. Πιο συγκεκριμένα, εφαρμόσθηκε κωδικοποίηση onehot, ούτως ώστε να αποφευχθεί η εισαγωγή λανθασμένης ιεραρχίας και να αντιμετωπίζεται ισότιμα κάθε κατηγορία.

Έπειτα, το σύνολο δεδομένων διαχωρίστηκε στα υποσύνολα x_train, x_test, y_train και y_test με αναλογία 80% για εκπαίδευση και 20% για έλεγχο. Μετά τον διαχωρισμό, εφαρμόσθηκε κανονικοποίηση standard scaling, η οποία μετατρέπει όλα τα χαρακτηριστικά έτσι ώστε να έχουν μέση τιμή 0 και τυπική απόκλιση 1. Αξίζει να σημειωθεί, ότι οι παράμετροι της κανονικοποίησης υπολογίστηκαν αποκλειστικά στο υποσύνολο εκπαίδευσης και στη συνέχεια εφαρμόστηκαν στο υποσύνολο ελέγχου για την αποφυγή διαρροής δεδομένων. Με αυτή τη μέθοδο κανονικοποίησης, εμποδίζεται η κυριαρχία των μεταβλητών με μεγάλες τιμές και

επιτυγχάνεται ταχύτερη σύγκλιση των αλγορίθμων ιδίως σε εφαρμογές βαθιάς μάθησης.

3 – Εφαρμογή Αλγορίθμων Μηχανικής Μάθησης

Σε αυτό το στάδιο της εργασίας εφαρμόζονται τρεις αλγόριθμοι μηχανικής μάθησης, καθώς και ένας αλγόριθμος βαθιάς μάθησης. Ως μετρική αξιολόγησης των μοντέλων επιλέχθηκε η Τετραγωνική Ρίζα του Μέσου Τετραγωνικού Σφάλματος (Root Mean Squared Error (RMSE)), καθώς τιμωρεί σε μεγαλύτερο βαθμό τα μεγάλα σφάλματα πρόβλεψης. Με αυτόν τον τρόπο το μοντέλο αποφεύγει ακραίες εσφαλμένες προβλέψεις. Επιπλέον, η χρήση του RMSE καθιστά το μέγεθος του σφάλματος περισσότερο κατανοητό, καθώς εκφράζεται στις ίδιες μονάδες μέτρησης με τη μεταβλητή στόχο.

3.1 Μοντέλο Ridge Regression

Αρχικά, αποφασίστηκε να εκπαιδευτεί ένα απλό γραμμικό μοντέλο ως μοντέλο αναφοράς. Πιο συγκεκριμένα, επιλέχθηκε το μοντέλο Ridge (L2 norm) regression, καθώς διαχειρίζεται καλύτερα τα πολλά χαρακτηριστικά που δημιουργούνται μετά την κωδικοποίηση onehot. Το μοντέλο αυτό διατηρεί όλα τα χαρακτηριστικά μειώνοντας απλά την επίδρασή (βάρη) τους, σε αντίθεση με το μοντέλο Lasso (L1 norm) που μηδενίζει κάποια βάρη. Με αυτόν τον τρόπο συντελούν στην πρόβλεψη της μεταβλητής στόχου ακόμα και χαρακτηριστικά που φέρουν λιγότερο χρήσιμη πληροφορία.

Για την εύρεση της καλύτερης παραμέτρου ομαλοποίησης (alpha), χρησιμοποιήθηκε η μέθοδος GridSearchCV με 5-fold cross-validation. Επιπλέον, πριν από κάθε δοκιμή τα δεδομένα ανακατευόταν τυχαία (shuffle=True), ώστε το μοντέλο να μην επηρεάζεται από τη σειρά των δεδομένων. Παράλληλα, τέθηκε το random state ίσο με 42, ώστε το ανακάτεμα των δεδομένων να είναι το ίδιο σε κάθε εκτέλεση του κώδικα και τα αποτελέσματα να παραμένουν σταθερά. Στη συνέχεια, αφού βρέθηκε η βέλτιστη τιμή της παραμέτρου alpha (0.3), το βέλτιστο μοντέλο αξιολογήθηκε στο σύνολο ελέγχου.

Από τα αποτελέσματα που παρουσιάζονται στον Πίνακα 4, προκύπτει ότι το RMSE στο σύνολο ελέγχου (6.8533) είναι ελαφρώς καλύτερο από το RMSE που υπολογίστηκε μέσω cross-validation (6.8607), συνεπώς το μοντέλο έχει πολύ καλή γενίκευση και δεν υπάρχουν ενδείξεις υπερπροσαρμογής.

	Βέλτιστο Μοντέλο Ridge
alpha	0.3
Cross-Validation RMSE	6.8607
Test Set RMSE	6.8533

Πίνακας 4: Βέλτιστο μοντέλο Ridge

3.2 Μοντέλο Random Forest Regression

Ως δεύτερο μοντέλο, επιλέχθηκε το Random Forest καθώς πρόκειται για ένα μοντέλο το οποίο βασίζεται σε πολλά δέντρα απόφασης και μπορεί να εντοπίσει σύνθετες μη γραμμικές σχέσεις στα δεδομένα. Για την εύρεση των καλύτερων τιμών των παραμέτρων, χρησιμοποιήθηκε και πάλι η μέθοδος GridSearchCV με 5-fold cross-validation, shuffle=True και random_state=42. Λόγω των περιορισμένων πόρων του google colab, καθώς και της χρονοβόρας διαδικασίας εκπαίδευσης ο αριθμός των δέντρων (n_estimators) τέθηκε 200, ενώ οι παράμετροι που εξετάστηκαν είναι το μέγιστο βάθος τους (max_depth), καθώς και ο ελάχιστος αριθμός δειγμάτων που απαιτούνται για τον διαχωρισμό ενός κόμβου (min_samples_split). Στη συνέχεια, αφού βρέθηκαν οι βέλτιστες τιμές των παραμέτρων, το βέλτιστο μοντέλο αξιολογήθηκε στο σύνολο ελέγχου.

Από τα αποτελέσματα που παρουσιάζονται στον Πίνακα 5, προκύπτει ότι το RMSE στο σύνολο ελέγχου (6.1013) είναι ελαφρώς καλύτερο από το RMSE που υπολογίστηκε μέσω cross-validation (6.1909), συνεπώς το μοντέλο έχει πολύ καλή γενίκευση και δεν υπάρχουν ενδείξεις υπερπροσαρμογής. Αξίζει να σημειωθεί, ότι το RMSE βελτιώθηκε κατά 11% σε σύγκριση με το γραμμικό μοντέλο Ridge. Συνεπώς, υπάρχουν σύνθετες μη γραμμικές σχέσεις στα δεδομένα που δεν μπορούσαν να εντοπιστούν από το γραμμικό μοντέλο.

Βέλτιστο Μοντέλο	Random Forest
n_estimators	200
max_depth	None
min_samples_split	10
Cross-Validation RMSE	6.1909
Test Set RMSE	6.1013

Πίνακας 5: Βέλτιστο μοντέλο Random Forest

3.3 Μοντέλο XGBoost

Ως τρίτο μοντέλο, επιλέχθηκε το Extreme Gradient Boost (XGBoost), καθώς είναι υπολογιστικά πιο αποδοτικό, διαχειρίζεται καλύτερα τη μνήμη και συνήθως επιτυγχάνει υψηλότερη ακρίβεια σε δομημένα δεδομένα. Σε αντίθεση με το Random Forest, το οποίο δημιουργεί ανεξάρτητα δέντρα, το XGBoost υλοποιεί διαδοχικά δέντρα, όπου το κάθε νέο δέντρο προσπαθεί να διορθώσει τα σφάλματα των προηγούμενων. Για την εύρεση των καλύτερων τιμών των παραμέτρων, χρησιμοποιήθηκε και πάλι η μέθοδος GridSearchCV με 5-fold cross-validation, shuffle=True και random_state=42. Πιο συγκεκριμένα, οι παράμετροι που εξετάστηκαν είναι ο αριθμός των δέντρων (n_estimators), το μέγιστο βάθος τους (max_depth), ο ρυθμός μάθησης (learning_rate) και το ποσοστό δεδομένων ανά δέντρο (subsample). Στη συνέχεια, αφού βρέθηκαν οι βέλτιστες τιμές των παραμέτρων, το βέλτιστο μοντέλο αξιολογήθηκε στο σύνολο ελέγχου.

Από τα αποτελέσματα που παρουσιάζονται στον Πίνακα 6, προκύπτει ότι το RMSE στο σύνολο ελέγχου (5.8066) είναι ελαφρώς καλύτερο από το RMSE που υπολογίστηκε μέσω cross-validation (5.8929), συνεπώς το μοντέλο έχει πολύ καλή γενίκευση και δεν υπάρχουν ενδείξεις υπερπροσαρμογής. Αξίζει να σημειωθεί, ότι το RMSE βελτιώθηκε περεταίρω σε σύγκριση με το μοντέλο Random Forest. Συνεπώς, το XGBoost αποτελεί το καλύτερο μοντέλο μηχανικής μάθησης για αυτό το σύνολο δεδομένων.

Βέλτιστο Μοντέλο	XGBoost
n_estimators	300
max_depth	6
learning_rate	0.05
subsample	0.8
Cross-Validation RMSE	5.8929
Test Set RMSE	5.8066

Πίνακας 6: Βέλτιστο μοντέλο XGBoost

3.4 Μοντέλο Βαθιάς Μάθησης

Ως μοντέλο βαθιάς μάθησης, επιλέχθηκε ένα τεχνητό νευρωνικό δίκτυο τύπου Multi – Layer Perceptron (MLP), έτσι ώστε να διερευνηθεί αν οι μη γραμμικοί μετασχηματισμοί που πραγματοποιούνται στα κρυφά επίπεδα (Hidden Layers) μπορούν να εντοπίσουν ακόμα πιο πολύπλοκες συσχετίσεις μεταξύ των χαρακτηριστικών.

Τα δεδομένα εκπαίδευσης διαχωρίζονται σε υποσύνολα εκπαίδευσης και επικύρωσης με αναλογία 90% - 10%, ώστε το μοντέλο να προσαρμόζει το ρυθμό μάθησης και να σταματά έγκαιρα αν ανιχνευθεί υπερπροσαρμογή. Για την υλοποίηση του νευρωνικού δικτύου επιλέχθηκε μια αρχιτεκτονική τεσσάρων κρυφών επιπέδων (Dense Layers). Πιο συγκεκριμένα:

- ✓ **Input Layer:** Δέχεται ως είσοδο τα κανονικοποιημένα δεδομένα, συνεπώς ο αριθμός των νευρώνων ισούται με τον αριθμό των χαρακτηριστικών.

- ✓ **1st Hidden Layer:** Είναι το μεγαλύτερο στρώμα, καθώς αποτελείται από 512 νευρώνες. Βασικός σκοπός του είναι να δημιουργήσει έναν πολύ μεγάλο αριθμό πιθανών σχέσεων μεταξύ των χαρακτηριστικών. Παράλληλα, στα δύο πρώτα στρώματα χρησιμοποιείται BatchNormalization, ώστε οι τιμές των νευρώνων να παραμένουν σε μια σταθερή κλίμακα. Αξίζει να σημειωθεί, ότι ως συνάρτηση ενεργοποίησης χρησιμοποιείται η LeakyRelu στα δύο πρώτα στρώματα, διασφαλίζοντας ότι οι νευρώνες παραμένουν ενεργοί ακόμα και όταν οι είσοδοί τους πάρουν αρνητικές τιμές. Επιπλέον, η χρήση του Dropout (0.3) απενεργοποιεί τυχαία το 30% των νευρώνων σε κάθε κύκλο εκπαίδευσης. Με αυτόν τον τρόπο, το μοντέλο αναπτύσσει πολλαπλά μονοπάτια και δεν εξαρτάται από συγκεκριμένους νευρώνες, με αποτέλεσμα να μειώνεται ο κίνδυνος υπερπροσαρμογής.

- ✓ **2nd Hidden Layer:** Αποτελείται από 256 νευρώνες, καθώς το δίκτυο ξεκινά να φιλτράρει τις σχέσεις από το πρώτο στρώμα κρατώντας τις πιο σημαντικές. Επιπλέον, η χρήση του Dropout (0.2) απενεργοποιεί τυχαία το 20% των νευρώνων, καθώς ο αριθμός τους είναι μικρότερος. Πρακτικά, αποτελεί μια επιπρόσθετη δικλείδα ασφαλείας, εμποδίζοντας το μοντέλο να αποστηθίσει συγκεκριμένα μοτίβα των δεδομένων εκπαίδευσης.

- ✓ **3rd – 4th Hidden Layers:** Αποτελούνται από 128 και 64 νευρώνες αντίστοιχα, ώστε οι σύνθετες αφηρημένες σχέσεις των προηγούμενων επιπέδων να μετατραπούν σε συγκεκριμένες αριθμητικές αναπαραστάσεις που προσεγγίζουν τη μεταβλητή στόχο. Ως συνάρτηση ενεργοποίησης σε αυτά τα δύο επίπεδα χρησιμοποιείται η Relu, η οποία λειτουργεί ως φίλτρο που μηδενίζει τις αρνητικές αποκρίσεις. Με αυτόν τον τρόπο, το δίκτυο εστιάζει στις πιο χρήσιμες πληροφορίες.

- ✓ **Output Layer:** Αποτελείται από έναν νευρώνα, ο οποίος συνδυάζει τα βάρη των 64 νευρώνων του προηγούμενου επιπέδου για την πρόβλεψη της μεταβλητής στόχου.

Για τη διαδικασία της εκπαίδευσης χρησιμοποιήθηκε ο βελτιστοποιητής (optimizer) Adam με αρχικό ρυθμό μάθησης 0.001. Επιπλέον, χρησιμοποιήθηκαν οι τεχνικές ReduceLROnPlateau και EarlyStopping για τους εξής λόγους:

- ✓ **ReduceLROnPlateau:** Μειώνει τον ρυθμό μάθησης όταν το σφάλμα στα δεδομένα επικύρωσης σταθεροποιείται, με αυτόν τον τρόπο επιτυγχάνεται μεγαλύτερη ακρίβεια του μοντέλου.

- ✓ **EarlyStopping:** Τερματίζει την εκπαίδευση του μοντέλου όταν δεν επιτυγχάνεται περαιτέρω βελτίωση. Μέσω της επιλογής restore_best_weights, εξασφαλίζεται ότι οι παράμετροι του τελικού μοντέλου θα είναι αυτοί που πέτυχαν τη βέλτιστη γενίκευση.

Η εκπαίδευση πραγματοποιήθηκε για μέγιστο αριθμό 200 εποχών, ενώ το μοντέλο επεξεργάστηκε τα δεδομένα σε υποσύνολα των 128 παρατηρήσεων πριν την ενημέρωση των βαρών του, ώστε να επιτυγχάνεται ταχύτερη εκπαίδευση μέσω παράλληλης επεξεργασίας, αλλά και μια πιο σταθερή και ομαλή σύγκλιση προς το ελάχιστο σφάλμα. Τέλος, το μοντέλο αξιολογήθηκε στο σύνολο ελέγχου. Τα αποτελέσματα παρουσιάζονται στον Πίνακα 7.

Μοντέλο	Βαθιάς Μάθησης (MLP)
Test Set RMSE	5.9757

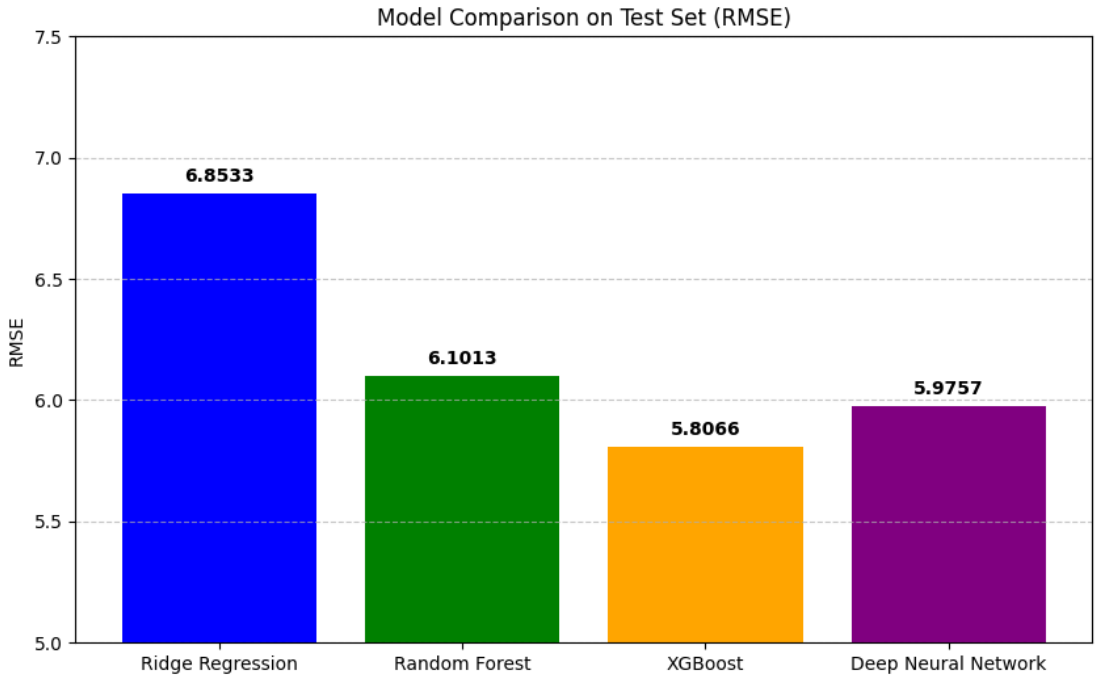
Πίνακας 7: Μοντέλο MLP

3.5 Σύγκριση Μοντέλων Μάθησης

Στον Πίνακα 8 και στο Σχήμα 2 παρουσιάζονται τα αποτελέσματα από τα τρία μοντέλα μηχανικής μάθησης και από το μοντέλο βαθιάς μάθησης. Το γραμμικό μοντέλο Ridge παρουσιάζει το μεγαλύτερο RMSE (6.8533). Στο μοντέλο Random Forest το RMSE (6.1013) βελτιώθηκε κατά 11% γεγονός που αποδεικνύει ότι υπάρχουν σύνθετες μη γραμμικές σχέσεις μεταξύ των χαρακτηριστικών και της μεταβλητής στόχου. Οι καλύτερες επιδόσεις επιτεύχθηκαν από το μοντέλο MLP (5.9757) και το μοντέλο XGBoost όπου το RMSE (5.8066) βελτιώθηκε κατά 15% περίπου σε σχέση με το γραμμικό μοντέλο. Το αποτέλεσμα αυτό επιβεβαιώνει τη βιβλιογραφία, σύμφωνα με την οποία οι αλγόριθμοι Gradient Boosting συνήθως υπερτερούν σε δομημένα δεδομένα, καθώς είναι πιο ανθεκτικοί στον θόρυβο και αξιοποιούν αποτελεσματικότερα τα αραιά χαρακτηριστικά που προκύπτουν από την κωδικοποίηση onehot των κατηγορικών χαρακτηριστικών.

Μοντέλο	Test Set RMSE
Ridge Regression	6.8533
Random Forest	6.1013
XGBoost	5.8066
Deep Neural Network (MLP)	5.9757

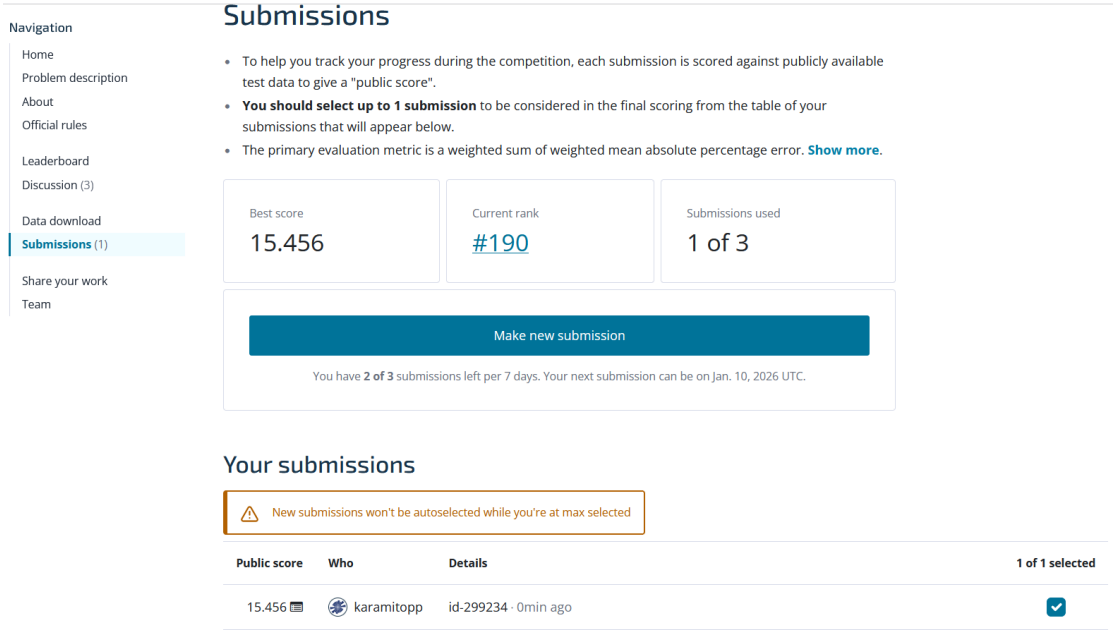
Πίνακας 8: Σύγκριση επίδοσης μοντέλων



Σχήμα 2: Σύγκριση επίδοσης μοντέλων

4 – Εφαρμογή του Βέλτιστου Μοντέλου

Για την ολοκλήρωση της εργασίας το βέλτιστο μοντέλο XGBoost χρησιμοποιήθηκε για την πρόβλεψη της μεταβλητής στόχου στο αρχείο ελέγχου test_hh_features.csv. Για τον σκοπό αυτόν, τα δεδομένα του συγκεκριμένου αρχείου αρχικά υπόκεινται στην ίδια επεξεργασία με τα αρχικά δεδομένα. Στη συνέχεια, το βέλτιστο μοντέλο προβλέπει τη μεταβλητή στόχο. Τα τελικά αποτελέσματα των προβλέψεων αποθηκεύονται στο ζητούμενο αρχείο predicted_household_consumption.csv. Παράλληλα, δημιουργείται ακόμα ένα αρχείο (predicted_poverty_distribution.csv) όπου αποθηκεύεται το ποσοστό των οικογενειών στις οποίες η μεταβλητή στόχος είναι μικρότερη από κάποια συγκεκριμένα όρια σύμφωνα με τις οδηγίες του διαγωνισμού. Τέλος, στο Σχήμα 3 παρουσιάζεται το αποδεικτικό της συμμετοχής στον διαγωνισμό, σύμφωνα με το οποίο η κατάταξη της παρούσας εργασίας είναι 190.



Σχήμα 3: Αποδεικτικό συμμετοχής στον διαγωνισμό

5 – Επεξήγηση Αποτελεσμάτων

Σε αυτήν την ενότητα αρχικά παρουσιάζονται τα πιο σημαντικά χαρακτηριστικά του βέλτιστου μοντέλου. Στη συνέχεια, αναφέρονται τυχόν περιορισμοί αυτού του μοντέλου και τρόποι βελτίωσης.

5.1 Σημαντικότητα Χαρακτηριστικών

Στον Πίνακα 9 και στο Σχήμα 4 παρουσιάζεται η σημαντικότητα των χαρακτηριστικών (feature importance) του βέλτιστου μοντέλου XGBoost. Ειδικότερα, τα 4 σημαντικότερα χαρακτηριστικά είναι τα εξής:

- ✓ **strata:** Η διαστρωμάτωση αποτελεί τον ισχυρότερο προγνωστικό παράγοντα, καθώς καθορίζει σε μεγάλο βαθμό το επίπεδο διαβίωσης.

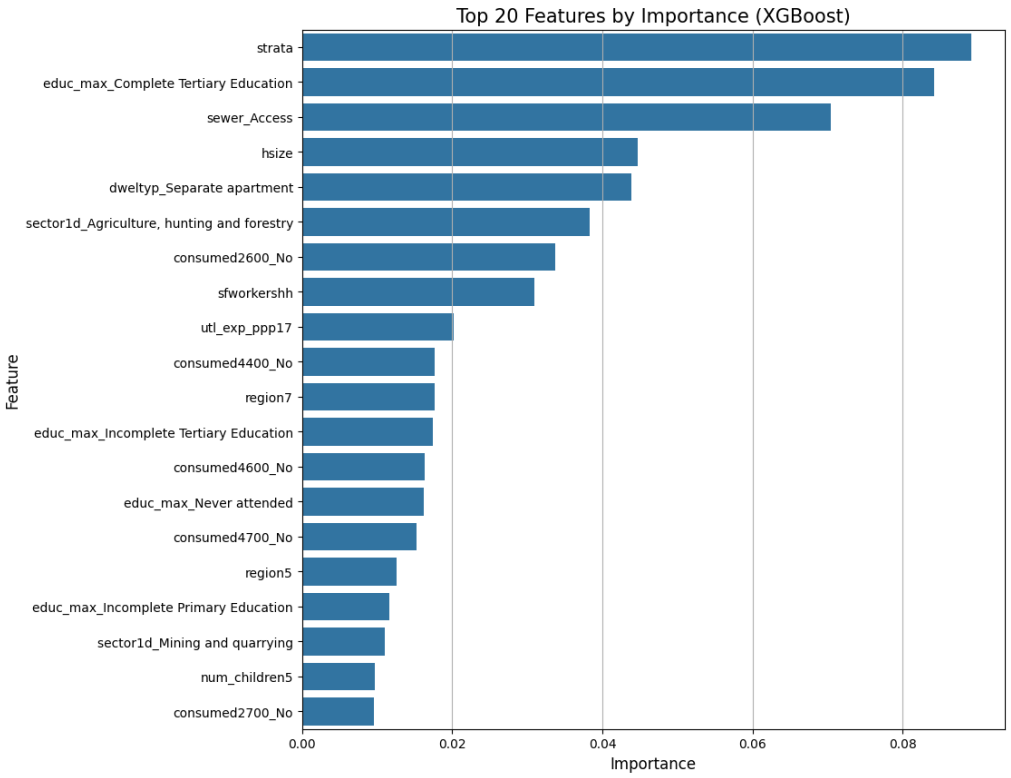
- ✓ **educ_max:** Η ολοκλήρωση της τριτοβάθμιας εκπαίδευσης από τουλάχιστον ένα μέλος της οικογένειας συνδέεται άμεσα με τον στόχο.

- ✓ **sewer:** Η σύνδεση με το αποχετευτικό σύστημα αποτελεί και αυτήν έναν παράγοντα άρρηκτα συνδεδεμένο με την ποιότητα ζωής.

- ✓ **hsize:** Ο αριθμός των μελών της οικογένειας επηρεάζει σημαντικά τις καθημερινές δαπάνες (στόχο), καθώς στις μεγαλύτερες οικογένειες οι δαπάνες ανά άτομο συνήθως είναι χαμηλότερες.

Χαρακτηριστικό	Σημαντικότητα
Strata	0.089086
educ_max	0.084174
Sewer	0.070354
Hsize	0.044678
Dweltyp	0.043837
sector1d	0.038340
consumed2600	0.033660
sfworkershh	0.030923
utl_exp_ppp17	0.020151
consumed4400	0.017719

Πίνακας 9: Σημαντικότητα χαρακτηριστικών



Σχήμα 4: Σημαντικότητα χαρακτηριστικών

5.2 Περιορισμοί του Μοντέλου

Αν και το βέλτιστο μοντέλο XGBoost παρουσιάζει υψηλή ακρίβεια, υπάρχουν ορισμένοι περιορισμοί. Πιο συγκεκριμένα, το μοντέλο αυτό λειτουργεί σωστά σε οικογένειες που ακολουθούν τα κλασσικά πρότυπα (π.χ. στις αστικές οικογένειες υψηλού επιπέδου εκπαίδευσης ή στις αγροτικές οικογένειες με λιγότερες υποδομές). Από την άλλη πλευρά, το μοντέλο ενδέχεται να αποκλίνει συστηματικά σε περιπτώσεις όπου οι οικογένειες σε πλούσιες περιοχές αντιμετωπίζουν μια προσωρινή κρίση ή οικογένειες χαμηλού επιπέδου εκπαίδευσης παρουσιάζουν υψηλά εισοδήματα από πηγές που ενδεχομένως να μην καταγράφονται στην έρευνα. Επιπλέον, η κωδικοποίηση onehot που εφαρμόστηκε οδηγεί σε πολλά χαρακτηριστικά, κάποια από τα οποία μπορεί να είναι ασήμαντα, εισάγοντας ενδεχομένως θόρυβο στο μοντέλο.

5.3 Τρόποι Βελτίωσης – Μελλοντικές Προεκτάσεις

Για τη βελτίωση των προβλέψεων, προτείνονται τα εξής:

- ✓ Η χρήση δυναμικών δεδομένων (χρονοσειρών) θα παρείχε πληροφορίες εποχικότητας καθώς η κατανάλωση (κυρίως στις αγροτικές οικογένειες) συνήθως μεταβάλλεται ανάλογα με την περίοδο του έτους.

- ✓ Δεδομένα για το κόστος ζωής ανά περιοχή ενδεχομένως να συντελούσαν στη βελτίωση των προβλέψεων.

- ✓ Τέλος, θα μπορούσε να εφαρμοσθεί κωδικοποίηση στόχου αντί onehot ώστε να συγκριθούν τα αποτελέσματα των μοντέλων και να επιλεχθεί το καλύτερο.

Βιβλιογραφία

[1] elearning AUTH. “Μηχανική Μάθηση” [Online]. Available: https://elearning.auth.gr/course/view.php?id=7130 (visited on Dec. 20, 2025).