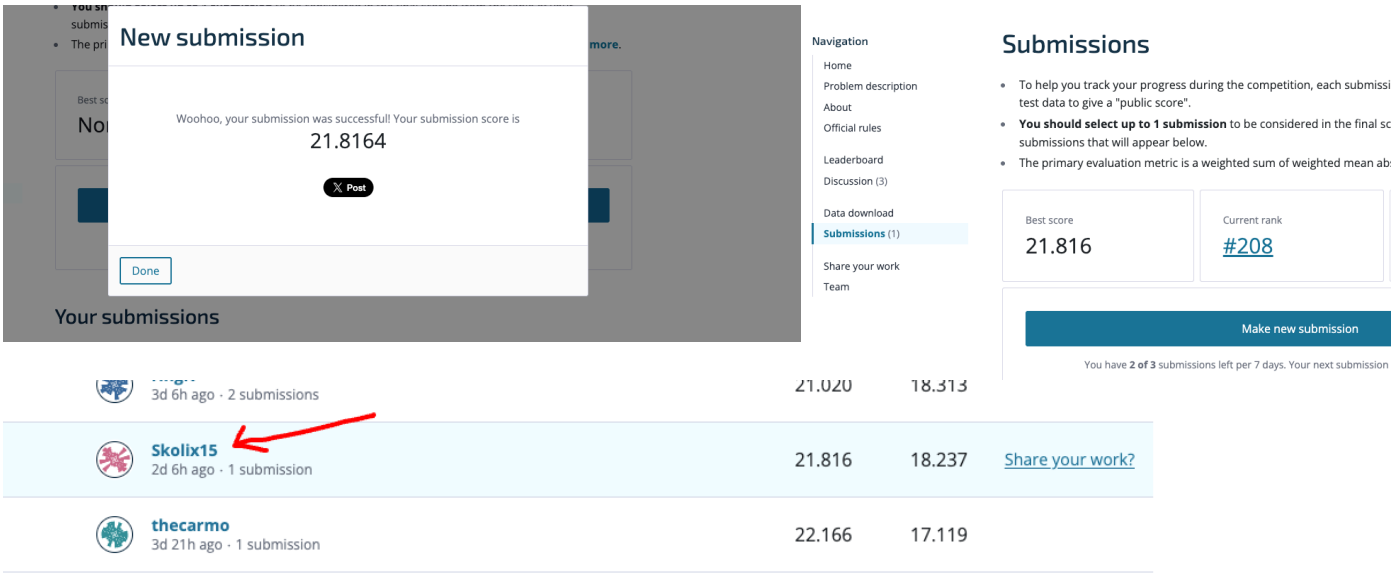
Μηχανική Μάθηση - Τελική Εργασία

Ονοματεπώνυμο: Παπαδόπουλος Ελευθέριος
Email: eleftherios.v.papadopoulos@gmail.com - epapadoax@csd.auth.gr

Αποδεικτικό Συμμετοχής



1. Εισαγωγή & Περιγραφή Προβλήματος

Σκοπός της παρούσας εργασίας ήταν η επίλυση ενός προβλήματος παλινδρόμησης (regression) που αφορά την πρόβλεψη της κατά κεφαλήν ημερήσιας δαπάνης των νοικοκυριών. Χρησιμοποιήθηκαν πραγματικά δεδομένα από την πλατφόρμα DrivenData (World Bank Poverty Prediction). Η ακριβής πρόβλεψη αυτών των δεικτών είναι κρίσιμη για τη χάραξη κοινωνικής πολιτικής σε αναπτυσσόμενες χώρες.

2. Επεξεργασία & Προετοιμασία Δεδομένων

Η προετοιμασία των δεδομένων αποτέλεσε το πιο χρονοβόρο κομμάτι, καθώς τα αρχικά αρχεία απαιτούσαν ενοποίηση και καθαρισμό. Τα βήματα που ακολουθήθηκαν ήταν:

- **Merge:** Έγινε συνένωση των χαρακτηριστικών (*train_hh_features*) με τη μεταβλητή στόχο (*train_hh_gt*) μέσω του κωδικού *hhid*.
- **Feature Selection:** Αφαιρέθηκαν στήλες που θεωρήθηκαν πλεονάζουσες ή που θα μπορούσαν να προκαλέσουν overfitting (όπως τα *hhid, com, survey_id, weight*).

- **Handling Missing Values:** Χρησιμοποιήθηκε ο μέσος όρος για τη συμπλήρωση κενών τιμών σε αριθμητικά δεδομένα.
- **Encoding:** Μετατράπηκαν οι κατηγορικές μεταβλητές (Yes/No, Urban/Rural) σε αριθμητική μορφή (0/1). Για τις μεταβλητές με πολλά επίπεδα (π.χ. πηγή νερού), εφαρμόστηκε η μέθοδος factorize.

3. Ανάλυση Δεδομένων

1. Περιγραφή / Ερμηνεία Χαρακτηριστικών

Τα χαρακτηριστικά του dataset χωρίζονται σε θεματικές ενότητες που περιγράφουν το βιοτικό επίπεδο ενός νοικοκυριού:

- **Δημογραφικά:** Μεταβλητές όπως το hsize (μέγεθος νοικοκυριού) και η ηλικία του αρχηγού (*age*), που δείχνουν τη σύνθεση και τις ανάγκες της οικογένειας.
- **Υποδομές:** Πρόσβαση σε ηλεκτρισμό (*elect),* νερό (*water*) και αποχέτευση (*sewer*), που αποτελούν άμεσους δείκτες διαβίωσης.
- **Οικονομικά/Απασχόληση:** Οι δαπάνες κοινής ωφέλειας (*utl_exp_ppp17*) και ο τομέας εργασίας (*sector1d*), που δείχνουν την οικονομική επιφάνεια.
- **Κατανάλωση:** Δυαδικοί δείκτες (*consumedXXXX*) που καταγράφουν αν το νοικοκυριό αγόρασε συγκεκριμένα τρόφιμα, αποκαλύπτοντας το διατροφικό προφίλ.

2. Περιγραφή Μεταβλητών & Ποιότητα

Τύποι Μεταβλητών: Έχουμε ένα μείγμα αριθμητικών (π.χ. utl_exp_ppp17, age) και κατηγορικών μεταβλητών. Οι κατηγορικές μετατράπηκαν σε αριθμητικές μέσω του clean_data (mapping για binary και factorize για τις υπόλοιπες).

Ποιότητα Δεδομένων: Η στήλη sector1d παρουσίασε το μεγαλύτερο ποσοστό ελλειπουσών τιμών, πιθανώς λόγω ανεργίας, και συμπληρώθηκε με τον μέσο όρο μετά την κωδικοποίηση.

- Οι μεταβλητές κατανάλωσης είναι πολύ "καθαρές" με ελάχιστα κενά.
- Η μεταβλητή στόχος cons_ppp17 έχει εύρος τιμών που απαιτεί προσοχή, καθώς οι πολύ υψηλές τιμές (outliers) μπορεί να επηρεάσουν το *RMSE*.

3. Σημαντικότητα των Χαρακτηριστικών

Βάσει της φύσης του προβλήματος και των στατιστικών των δεδομένων:

1. **utl_exp_ppp17 (Δαπάνες Κοινής Ωφέλειας):** Είναι ο ισχυρότερος "proxy" δείκτης για το συνολικό εισόδημα.

2. **educ_max (Εκπαίδευση):** Υπάρχει άμεση γραμμική σχέση μεταξύ του μορφωτικού επιπέδου και της ικανότητας κατανάλωσης.
3. **hsize (Μέγεθος Νοικοκυριού):** Σε αναπτυσσόμενες οικονομίες, τα πολύ μεγάλα νοικοκυριά συνδέονται στατιστικά με χαμηλότερη κατά κεφαλήν δαπάνη.
4. **consumed2600 / consumed4700:** Συγκεκριμένοι κωδικοί τροφίμων που θεωρούνται "πολυτελείας" ή "ελαστικής ζήτησης" λειτουργούν ως ισχυροί διαχωριστές μεταξύ φτωχών και μη φτωχών νοικοκυριών.

4. Συσχετίσεις μεταξύ Χαρακτηριστικών

- Παρατηρείται ισχυρή συσχέτιση μεταξύ της μεταβλητής urban και της πρόσβασης σε δίκτυα (elect, sewer), καθώς οι υποδομές αυτές είναι πιο ανεπτυγμένες στις πόλεις.
- Υπάρχει αρνητική συσχέτιση μεταξύ του hsize και της κατά κεφαλήν δαπάνης, επιβεβαιώνοντας ότι τα μεγάλα νοικοκυριά είναι πιο επιρρεπή στη φτώχεια.
- Οι δαπάνες κοινής ωφέλειας (utl_exp_ppp17) συσχετίζονται θετικά με το επίπεδο εκπαίδευσης (educ_max), δείχνοντας ότι η μόρφωση οδηγεί σε υψηλότερη κατανάλωση υπηρεσιών.

4. Μεθοδολογία & Εφαρμογή Αλγορίθμων

Υλοποιήθηκαν τέσσερα διαφορετικά μοντέλα για τη σύγκριση των αποτελεσμάτων:

A. Μηχανική Μάθηση (Machine Learning)

1. **Ridge Regression:** Χρησιμοποιήθηκε ως baseline μοντέλο. Προσφέρει μια γραμμική προσέγγιση με ρύθμιση L2 για την αποφυγή πολυκορυφικότητας.
2. **Random Forest Regressor:** Ένα ensemble μοντέλο δέντρων απόφασης. Επιλέχθηκε για την ικανότητά του να χειρίζεται outliers και μη-γραμμικές σχέσεις χωρίς ιδιαίτερη ανάγκη για κανονικοποίηση.
3. **XGBoost Regressor:** Μοντέλο Gradient Boosting που εστιάζει στη διόρθωση των σφαλμάτων των προηγούμενων δέντρων. Αποδείχθηκε το πιο ισχυρό μοντέλο, πετυχαίνοντας το χαμηλότερο **RMSE** (Εφαρμόστηκε με 500 estimators και learning rate 0.05).

B. Βαθιά Μάθηση (Deep Learning)

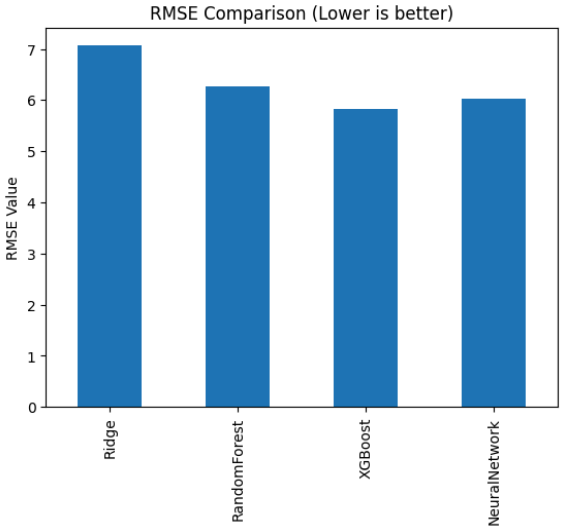
- **Multi-Layer Perceptron (MLP):** Σχεδιάστηκε ένα νευρωνικό δίκτυο με 3 κρυφά επίπεδα (128, 64, 32 νευρώνες) και τεχνικές **Dropout** (0.2) για την αποφυγή υπερεκπαίδευσης. Το μοντέλο αυτό κατάφερε να συλλάβει σύνθετα μοτίβα στα δεδομένα κατανάλωσης.

5. Αποτελέσματα & Επικύρωση

Η αξιολόγηση έγινε στο validation set (20%) με βάση το σφάλμα **RMSE** (Root Mean Squared Error).

Μοντέλο	**RMSE (Lower is better)**	**R² Score**
Ridge	7.065	0.503
Random Forest	6.262	0.610
XGBoost	**5.819**	**0.663**
Neural Network	6.031	0.640

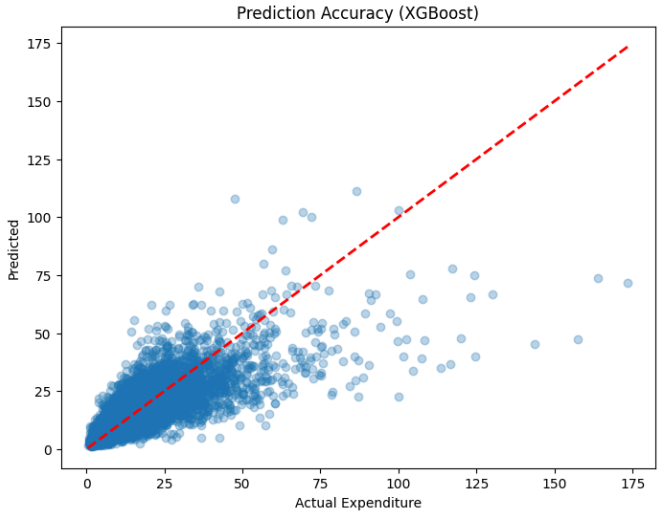
Το **XGBoost** αναδείχθηκε ως το βέλτιστο μοντέλο. Μετά το fine-tuning των υπερπαραμέτρων (learning rate, depth), χρησιμοποιήθηκε για την παραγωγή των τελικών προβλέψεων στο test set.



6. Επεξήγηση & Συμπεράσματα

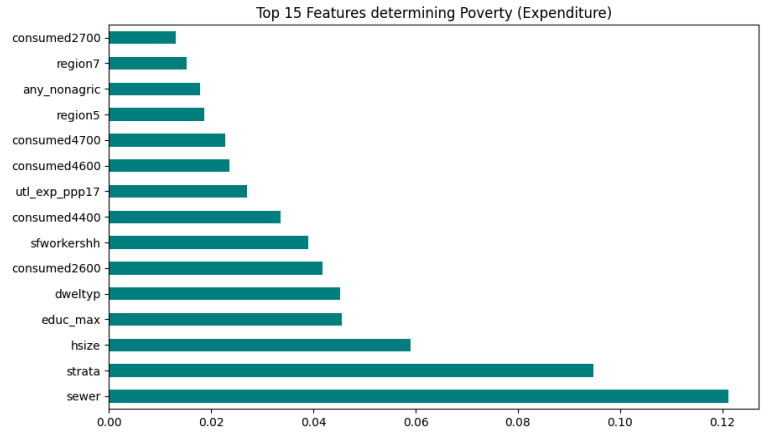
Περιορισμοί Μοντέλων

Τα μοντέλα δυσκολεύονται να προβλέψουν πολύ υψηλές τιμές δαπανών, κάτι που φαίνεται στο γράφημα Actual vs Predicted, όπου οι προβλέψεις "συσσωρεύονται" γύρω από το κέντρο.



Σημαντικότητα Χαρακτηριστικών

Το γράφημα Feature Importance του μοντέλου XGBoost δείχνει ότι την πρόβλεψη καθορίζουν κυρίως:



- **hsize (Μέγεθος Νοικοκυριού):** Περισσότερα μέλη → χαμηλότερη κατά κεφαλήν δαπάνη.
- **utl_exp_ppp17 (Δαπάνες Κοινής Ωφέλειας):** Υψηλότερες δαπάνες συνδέονται με καλύτερο βιοτικό επίπεδο.
- **strata (Στρώμα Δειγματοληψίας):** Αποτυπώνει γεωγραφικές και οικονομικές ανισότητες.
- **sfworkershh (Εργαζόμενα Μέλη):** Περισσότερο σταθερό εισόδημα αυξάνει την αγοραστική δύναμη.
- **urban (Αστικότητα):** Αστικές και αγροτικές περιοχές έχουν διαφορετικά πρότυπα κατανάλωσης.
- **educ_max (Εκπαίδευση):** Υψηλότερη μόρφωση συνδέεται με καλύτερες οικονομικές προοπτικές.
- **dweltyp & sewer (Κατοικία – Υποδομές):** Η ποιότητα στέγασης και οι υποδομές δείχνουν οικονομική ευημερία.
- **sector1d (Τομέας Απασχόλησης):** Ο κλάδος εργασίας επηρεάζει το επίπεδο δαπανών.
- **num_children18 (Παιδιά):** Περισσότερα παιδιά μειώνουν την κατά κεφαλήν διαθεσιμότητα πόρων.
- **consumed2600 / 4700:** Κατανάλωση μη βασικών αγαθών διαχωρίζει τα οικονομικά στρώματα.

Προτάσεις Βελτίωσης

Η προσθήκη δεδομένων σχετικά με την κατοχή συγκεκριμένων περιουσιακών στοιχείων (π.χ. ηλεκτρικές συσκευές, οχήματα) ή πληροφοριών για το τοπικό κόστος ζωής θα μπορούσε να μειώσει το σφάλμα πρόβλεψης.

Επίσης, μια πιθανή βελτίωση θα ήταν η δημιουργία νέων χαρακτηριστικών (feature engineering), όπως η αναλογία εργαζομένων προς το συνολικό μέγεθος του νοικοκυριού.