Μηχανική Μάθηση

Τελική Εργασία

Ονοματεπώνυμο: Ευάγγελος Αποστόλου

ΑΕΜ: 245

1. Περίληψη Εργασίας
   Σκοπός της παρούσας εργασίας είναι η ανάπτυξη και αξιολόγηση μοντέλων Μηχανικής και Βαθιάς Μάθησης για την πρόβλεψη του επιπέδου φτώχειας νοικοκυριών, αξιοποιώντας πραγματικά δεδομένα από την Παγκόσμια Τράπεζα. Το πρόβλημα προσεγγίστηκε ως πρόβλημα παλινδρόμησης με στόχο την πρόβλεψη της συνεχούς μεταβλητής κατανάλωσης ανά άτομο. Η επίλυση τέτοιων προβλημάτων είναι κρίσιμης σημασίας για την ορθή κατανομή πόρων και τη χάραξη κοινωνικής πολιτικής σε αναπτυσσόμενες χώρες.

2. Data Processing
   Η διαδικασία προετοιμασίας των δεδομένων ξεκίνησε με την ενοποίηση των χαρακτηριστικών και των ετικετών βάσει του μοναδικού αναγνωριστικού νοικοκυριού. Για τη διαχείριση των ελλειπουσών τιμών εφαρμόστηκε η μέθοδος της συμπλήρωσης, χρησιμοποιώντας τη διάμεσο για τα αριθμητικά χαρακτηριστικά και τη συχνότερη τιμή για τα κατηγορικά. Ακολούθως, πραγματοποιήθηκε κανονικοποίηση των αριθμητικών δεδομένων και κωδικοποίηση των κατηγορικών μεταβλητών μέσω της τεχνικής One-Hot Encoding. Τέλος, εφαρμόστηκε λογαριθμικός μετασχηματισμός στη μεταβλητή στόχο για την εξομάλυνση της κατανομής της και τη βελτίωση της απόδοσης των μοντέλων.

3. Data Analysis
   Στο στάδιο της διερευνητικής ανάλυσης εξετάστηκαν οι στατιστικές ιδιότητες των χαρακτηριστικών και οι μεταξύ τους συσχετίσεις. Η οπτικοποίηση των δεδομένων ανέδειξε τη σημαντική ασυμμετρία στην κατανομή της κατανάλωσης, δικαιολογώντας την επιλογή του λογαριθμικού μετασχηματισμού. Παράλληλα, εντοπίστηκαν σημαντικά χαρακτηριστικά που σχετίζονται με την υποδομή του νοικοκυριού και την εκπαίδευση, τα οποία φάνηκε να διαδραματίζουν καθοριστικό ρόλο στην πρόβλεψη του επιπέδου ευημερίας.

4. Models

   Για την επίλυση του προβλήματος υλοποιήθηκαν τέσσερις διαφορετικοί αλγόριθμοι ώστε να καλυφθεί ένα ευρύ φάσμα προσεγγίσεων. Συγκεκριμένα, χρησιμοποιήθηκε ο LinearSVR για τη γραμμική χαρτογράφηση σε υψηλότερες διαστάσεις και ο Random Forest Regressor ως μέθοδος συνόλου για τη μείωση της διακύμανσης. Επιπλέον, εφαρμόστηκε ο αλγόριθμος XGBoost για την εκμετάλλευση της τεχνικής ενίσχυσης κλίσης. Τέλος, αναπτύχθηκε ένα μοντέλο Βαθιάς Μάθησης (Deep Learning MLP) με πολλαπλά κρυφά επίπεδα και τεχνική Dropout, προκειμένου να εντοπιστούν πολύπλοκα μη γραμμικά πρότυπα στα δεδομένα.

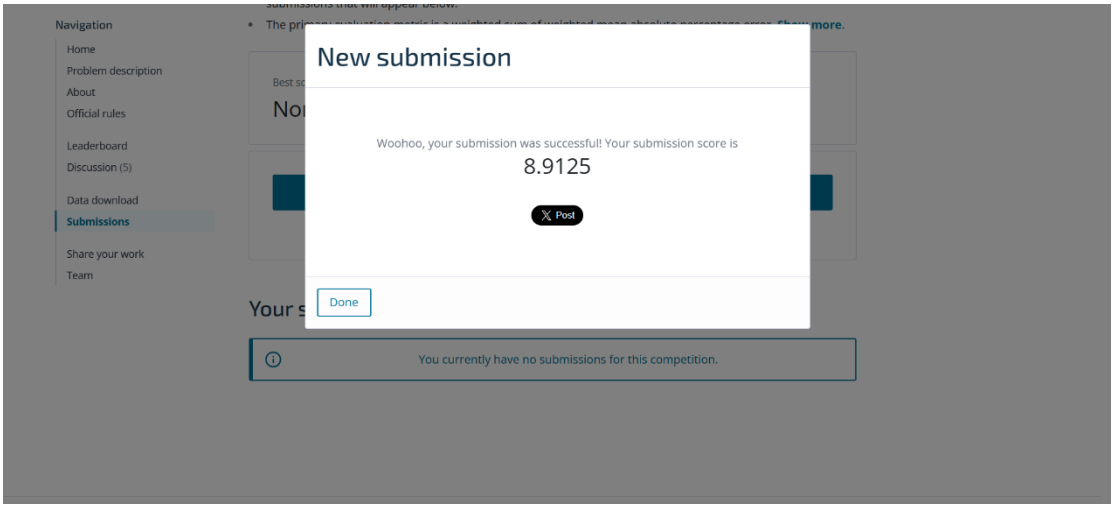
5. Validation

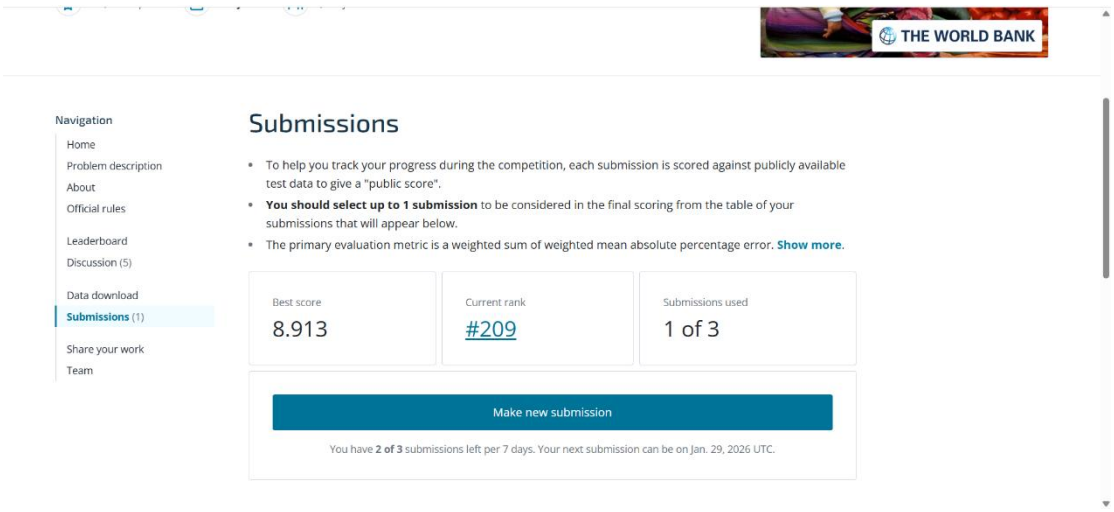
   Η αξιολόγηση των μοντέλων πραγματοποιήθηκε με χρήση ενός ανεξάρτητου συνόλου επικύρωσης, βάσει των μετρικών RMSE και R2. Από τη συγκριτική ανάλυση προέκυψε ότι ο αλγόριθμος XGBoost πέτυχε τη βέλτιστη επίδοση, παρουσιάζοντας το χαμηλότερο σφάλμα πρόβλεψης και την υψηλότερη προσαρμοστικότητα στα δεδομένα. Το μοντέλο Βαθιάς Μάθησης και το Random Forest κινήθηκαν σε παρόμοια επίπεδα ακρίβειας, ενώ ο LinearSVR αποτέλεσε τη βάση αναφοράς με ελαφρώς χαμηλότερη απόδοση.

6. Συμπεράσματα

   Η εργασία κατέδειξε ότι οι μέθοδοι ενίσχυσης κλίσης, και συγκεκριμένα το XGBoost, υπερτερούν σε προβλήματα δομημένων δεδομένων τέτοιου τύπου. Το τελικό μοντέλο που επιλέχθηκε χρησιμοποιήθηκε για την παραγωγή προβλέψεων στο σύνολο δοκιμής, οι οποίες υποβλήθηκαν επιτυχώς στην πλατφόρμα DrivenData. Η διαδικασία ανέδειξε τη σημασία της σωστής προεπεξεργασίας και της παραμετροποίησης των αλγορίθμων για την επίτευξη υψηλής ακρίβειας σε πραγματικά προβλήματα πρόβλεψης φτώχειας.

7. Competition Results


Navigation
Home
Problem description
About
Official rules
Leaderboard
Discussion (5)
Data download
Submissions
Share your work
Team
New submission
Woohoo, your submission was successful! Your submission score is
8.9125
Post
Done
You currently have no submissions for this competition.


THE WORLD BANK
Navigation
Home
Problem description
About
Official rules
Leaderboard
Discussion (5)
Data download
Submissions (1)
Share your work
Team
Submissions
To help you track your progress during the competition, each submission is scored against publicly available test data to give a "public score".
You should select up to 1 submission to be considered in the final scoring from the table of your submissions that will appear below.
The primary evaluation metric is a weighted sum of weighted mean absolute percentage error. Show more.
Best score
8.913
Current rank
#209
Submissions used
1 of 3
Make new submission
You have 2 of 3 submissions left per 7 days. Your next submission can be on Jan. 29, 2026 UTC.