Report

Εισαγωγή

Σκοπός της παρούσας εργασίας είναι η ανάπτυξη και αξιολόγηση μοντέλων Μηχανικής Μάθησης για την πρόβλεψη της κατά κεφαλήν κατανάλωσης νοικοκυριών και την εκτίμηση δεικτών φτώχειας. Η προσέγγιση περιλαμβάνει προεπεξεργασία δεδομένων, εφαρμογή και σύγκριση πολλαπλών αλγορίθμων, fine-tuning υπερπαραμέτρων και τελική επιλογή μοντέλου για παραγωγή προβλέψεων.

1. Επεξεργασία Δεδομένων

Η προεπεξεργασία περιλάμβανε:

- **Φόρτωση των αρχείων με τα δεδομένα:** Φορτώθηκαν τα αρχεία που περιείχαν τα σύνολα χαρακτηριστικών, τα labels, τα poverty rates, καθώς και τα σύνολα χαρακτηριστικών του test set.
- **Δημιουργία κοινού train set:** Δημιουργήθηκε το train set με ένωση labels και χαρακτηριστικών.
- **Αφαίρεση στηλών:** Αφαιρέθηκαν οι στήλες ID που δεν κρίθηκαν χρήσιμες για την εκπαίδευση των μοντέλων.
- **Διαχείριση ελλιπών τιμών**: Χρήση SimpleImputer με median για αριθμητικά και most frequent για κατηγορικά χαρακτηριστικά.
- **Κωδικοποίηση κατηγορικών μεταβλητών**: Εφαρμογή one-hot encoding ώστε τα δεδομένα να είναι συμβατά με τα μοντέλα.
- **Κλιμάκωση χαρακτηριστικών**: Τα χαρακτηριστικά προσαρμόστηκαν ώστε να είναι όλα στην ίδια κλίμακα, ιδιαίτερα σημαντικό βήμα για το νευρωνικό δίκτυο (MLP).

2. Περιγραφή Δεδομένων

Η περιγραφή των δεδομένων περιλάμβανε:

- **Ερμηνεία των πληροφοριών που δίνουν τα χαρακτηριστικά:** Ανάλυση κάθε ομάδας χαρακτηριστικών.
- **Περιγραφή των μεταβλητών και της ποιότητάς τους:** Χρήση της describe() και γραφημάτων.
- **Εύρεση σημαντικότητας χαρακτηριστικών:** Υπολογισμός correlation κάθε χαρακτηριστικού με την μεταβλητή στόχο και αναπαράσταση του.
- **Εύρεση συσχετίσεων μεταξύ των χαρακτηριστικών:** Υπολογισμός correlation κάθε χαρακτηριστικού με τα υπόλοιπα χαρακτηριστικά και αναπαράστασή του.

3. Εφαρμογή Αλγορίθμων Μηχανικής Μάθησης

Χωρίστηκε το σύνολο δεδομένων σε train & validation sets και εφαρμόστηκαν τρεις αλγόριθμοι Μηχανικής Μάθησης και ένας αλγόριθμος Βαθιάς Μάθησης:

- **Γραμμική Παλινδρόμηση**: Χρησιμοποιήθηκε ως baseline μοντέλο λόγω απλότητας και ερμηνευσιμότητας.
- **Random Forest Regressor**: Επιλέχθηκε για την ικανότητά του να μοντελοποιεί μη γραμμικές σχέσεις και αλληλεπιδράσεις χαρακτηριστικών.
- **Gradient Boosting Regressor**: Χρησιμοποιήθηκε για την υψηλή του ακρίβεια σε προβλήματα παλινδρόμησης και τη σταδιακή βελτίωση των προβλέψεων.
- **Neural Network (MLP)**: Εφαρμόστηκε ως μοντέλο βαθιάς μάθησης, ικανό να εκμεταλλευτεί πολύπλοκα μοτίβα στα δεδομένα.

Για κάθε μοντέλο πραγματοποιήθηκε **fine-tuning** βασικών υπερπαραμέτρων, με στόχο τη βελτιστοποίηση της απόδοσης στο validation set.

4. Αξιολόγηση και Σύγκριση Μοντέλων

Η σύγκριση των μοντέλων πραγματοποιήθηκε με βάση τη μετρική **RMSE** στο validation set. Τα αποτελέσματα έδειξαν ότι:

- Η Γραμμική Παλινδρόμηση παρουσίασε τη μεγαλύτερη τιμή σφάλματος.
- Τα ensemble μοντέλα (Random Forest και Gradient Boosting) πέτυχαν σαφώς χαμηλότερο RMSE.
- Το νευρωνικό δίκτυο (MLP) κατέγραψε τη χαμηλότερη τιμή RMSE, επιτυγχάνοντας την καλύτερη συνολική απόδοση.

Τα αποτελέσματα παρουσιάστηκαν μέσω πινάκων και γραφημάτων, επιτρέποντας την άμεση οπτική σύγκριση των μοντέλων.

Με βάση τη σύγκριση, επιλέχθηκε το μοντέλο με τη χαμηλότερη τιμή RMSE ως τελικό μοντέλο. Το μοντέλο αυτό εκπαιδεύτηκε στο πλήρες σύνολο εκπαίδευσης και χρησιμοποιήθηκε για την παραγωγή προβλέψεων στο αρχείο επικύρωσης (test set χωρίς labels). Οι προβλέψεις οργανώθηκαν σε αρχεία CSV σύμφωνα με τις προδιαγραφές της πλατφόρμας υποβολής.

---

5. Επεξήγηση Αποτελεσμάτων

Τα μη γραμμικά μοντέλα αποδείχθηκαν καταλληλότερα για το συγκεκριμένο πρόβλημα, ωστόσο παρουσιάζουν αυξημένο υπολογιστικό κόστος και εξάρτηση από την ποιότητα των δεδομένων. Το MLP, αν και πιο ακριβές, είναι ευαίσθητο σε missing values και στην επιλογή υπερπαραμέτρων. Τα μοντέλα αποδίδουν καλύτερα όταν τα δεδομένα περιγράφουν επαρκώς την οικονομική κατάσταση των νοικοκυριών και χειρότερα όταν λείπουν κρίσιμες πληροφορίες.

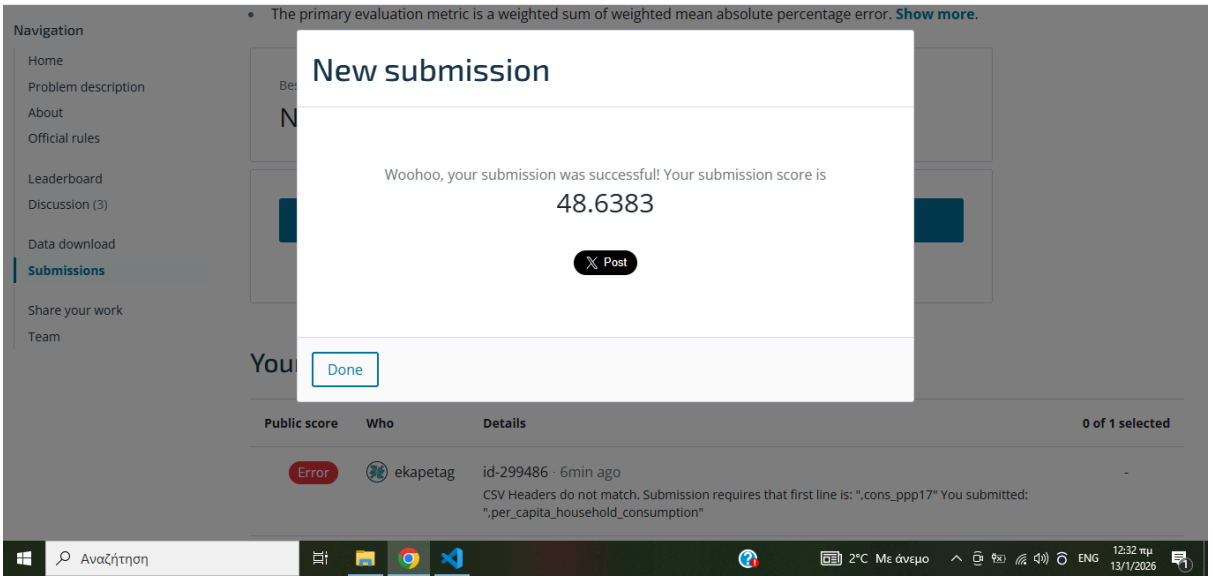
Πιθανές βελτιώσεις περιλαμβάνουν:

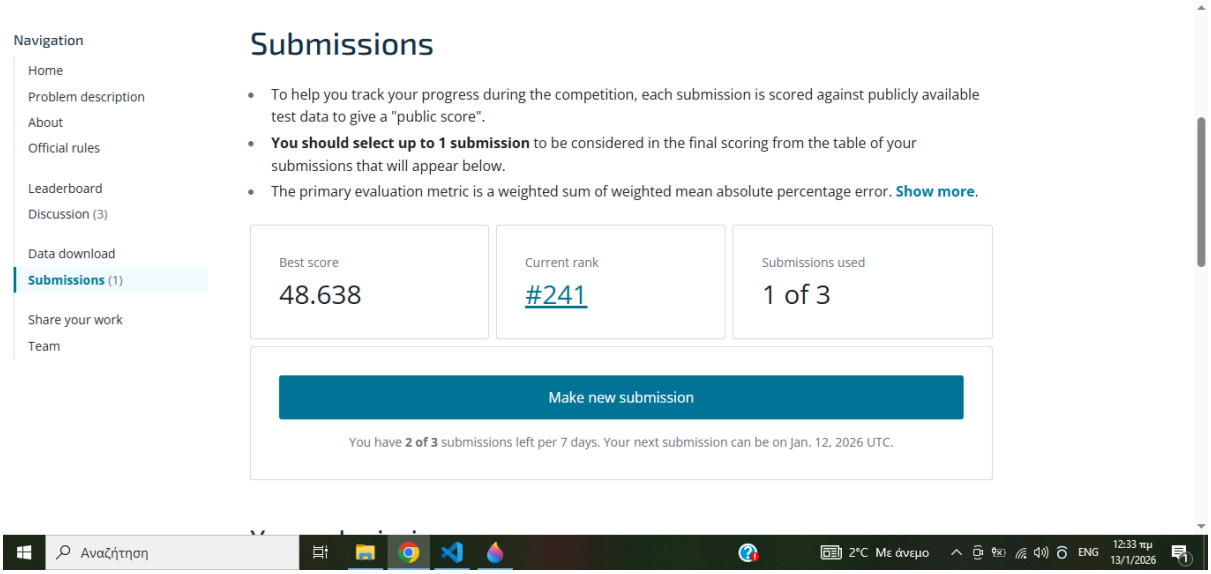
- Ενσωμάτωση επιπλέον χαρακτηριστικών (γεωγραφικά, χρονικά, δείκτες πρόσβασης σε υπηρεσίες).
- Καλύτερη ποιότητα και πληρότητα δεδομένων.
- Περαιτέρω fine-tuning και δοκιμή πιο εξελιγμένων αρχιτεκτονικών βαθιάς μάθησης.

Συμπεράσματα

Η εργασία ανέδειξε τη σημασία της σωστής προεπεξεργασίας και της επιλογής κατάλληλων μοντέλων για προβλήματα πρόβλεψης φτώχειας. Τα αποτελέσματα δείχνουν ότι τα ensemble και τα μοντέλα βαθιάς μάθησης μπορούν να προσφέρουν πιο ακριβείς και αξιόπιστες προβλέψεις, όταν εφαρμόζονται με προσεκτική μεθοδολογία.

Αποδεικτικά Υποβολής και Θέσης στο Leaderboard




Navigation
Home
Problem description
About
Official rules
Leaderboard
Discussion (3)
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
48.638
Current rank
#241
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. 12, 2026 UTC.
Αναζήτηση
2°C Με άνεμο
ENG
12:33 πμ
13/1/2026