Final Project: Poverty Prediction

Αστέριος Τερζής, ΑΕΜ: 218

---

1. Εισαγωγή

Σε αυτή την εργασία ασχολήθηκα με την πρόβλεψη της φτώχειας σε επίπεδο νοικοκυριού. Τα δεδομένα είναι από τον διαγωνισμό “World Bank Poverty Prediction Challenge” του DrivenData. Ο στόχος ήταν να προβλέψω την ημερήσια κατανάλωση ανά άτομο (cons_ppp17) για κάθε νοικοκυριό.

Τα δεδομένα περιλαμβάνουν περίπου 14.500 νοικοκυριά για training και 103.000 για test, με 88 χαρακτηριστικά το καθένα (δημογραφικά, στέγαση, απασχόληση, κατανάλωση τροφίμων κλπ).

---

2. Επεξεργασία Δεδομένων

Αρχικά έκανα φόρτωση και ένωση των αρχείων (features με targets) χρησιμοποιώντας το hhid ως κλειδί. Μετά χρειάστηκε να μετατρέψω τις κατηγορικές μεταβλητές σε αριθμητικές:

- **Yes/No → 1/0**
- **Male/Female → 1/0**
- **Access/No access → 1/0**
- **Label Encoding για τις υπόλοιπες**

Για τις ελλείπουσες τιμές χρησιμοποίησα τη διάμεσο (median) γιατί είναι πιο ανθεκτική στα outliers από τον μέσο όρο.

---

3. Ανάλυση Δεδομένων

Η κατανάλωση (cons_ppp17) έχει θετικά ασύμμετρη κατανομή - δηλαδή οι περισσότεροι έχουν χαμηλή κατανάλωση και λίγοι πολύ υψηλή. Γι’ αυτό εφάρμοσα log-transform που βοηθάει τα μοντέλα.

Από τις συσχετίσεις, τα πιο σημαντικά χαρακτηριστικά ήταν:

Χαρακτηριστικό	**Συσχέτιση**	**Ερμηνεία**
utl_exp_ppp17	+0.48	Δαπάνες για ρεύμα/νερό
urban	+0.45	Αστικές περιοχές
sewer	+0.40	Πρόσβαση σε αποχέτευση
sfworkershh	+0.40	Τυπική απασχόληση

Όσοι πληρώνουν περισσότερα για ρεύμα ή νερό και μένουν σε πόλεις με καλές υποδομές, έχουν και μεγαλύτερη κατανάλωση.

---

4. Αλγόριθμοι

Δοκίμασα 4 αλγορίθμους ML και 1 Neural Network:

1. **Ridge Regression:** Γραμμικό μοντέλο με regularization (χρησιμοποιήθηκε ως baseline).

2. **Random Forest:** Ensemble από δέντρα αποφάσεων, κατάλληλο για μη-γραμμικές σχέσεις.

3. **XGBoost:** Gradient Boosting αλγόριθμος, ιδιαίτερα αποτελεσματικός για tabular data.

4. **LightGBM:** Βελτιστοποιημένο Gradient Boosting, ταχύτερο από το XGBoost.

5. **Neural Network:** Νευρωνικό δίκτυο με 3 κρυφά επίπεδα (256→128→64), BatchNorm και Dropout.h

Για την αξιολόγηση χρησιμοποίησα το wMAPE (weighted Mean Absolute Percentage Error) και χώρισα τα δεδομένα 80% training, 20% validation.

---

5. Αποτελέσματα

Τα αποτελέσματα στο validation set:

Αλγόριθμος	**wMAPE (%)**	**MAE**
LightGBM	**27.71**	**2.69**
XGBoost	**27.76**	**2.70**
Neural Network	29.08	2.87
Random Forest	30.24	2.90
Ridge Regression	31.96	3.05

Τα gradient boosting μοντέλα (LightGBM, XGBoost) ήταν τα καλύτερα. Το Neural Network δεν τα κατάφερε καλύτερα, κάτι που είναι αρκετά συνηθισμένο σε tabular data. Για την τελική υποβολή έκανα ensemble των δύο καλύτερων:

Τελική Πρόβλεψη = 0.5 × XGBoost + 0.5 × LightGBM

Αποτελέσματα Leaderboard

Στην πλατφόρμα DrivenData πήρα θέση **251 από 1100** συμμετέχοντες (Top 23% εκείνη την στιγμή).

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	Current rank	Submissions used
11.936	#251	1 of 3

Make new submission

You have **2 of 3** submissions left per 7 days. Your next submission can be on Jan. 28, 2026 UTC.

GitHub :
https://github.com/agterzis/Machine-Learning/blob/main/%CE%9C%CE%B7%CF%87%CE%B1%CE%BD%CE%B9%CE%BA%CE%AE_%CE%9C%CE%AC%CE%B8%CE%B7%CF%83%CE%B7_Final_Project.ipynb