Μοντελοποίηση και Πρόβλεψη Επιπέδων Φτώχειας

1. Ποιότητα και Σημαντικότητα Χαρακτηριστικών

Κατά την ανάλυση των δεδομένων (EDA) μέσω του ProfileReport, εντοπίστηκαν τα εξής:

- **Ποιότητα:** Τα δεδομένα περιλαμβάνουν 88 χαρακτηριστικά. Εντοπίστηκαν υψηλές συσχετίσεις (High Correlation) μεταξύ μεταβλητών, όπως η μεταβλητή hsize (μέγεθος νοικοκυριού) με την ύπαρξη αποχέτευσης.

- **Σημαντικότητα:** Τα μοντέλα βασίζονται σε χαρακτηριστικά που αντικατοπτρίζουν τις συνθήκες διαβίωσης. Το μέγεθος του νοικοκυριού και οι υποδομές αποδείχθηκαν κρίσιμα, καθώς παρουσιάζουν ισχυρή στατιστική συσχέτιση με το επίπεδο κατανάλωσης και φτώχειας.

2. Προεπεξεργασία Δεδομένων (Preprocessing)

Για την προετοιμασία των δεδομένων εφαρμόστηκε ένας σύνθετος μετασχηματισμος (ColumnTransformer) που περιλάμβανε:

- **Διαχείριση Ελλιπών Τιμών:** Χρήση SimpleImputer για τη συμπλήρωση κενών τιμών.

- **Αριθμητικά Χαρακτηριστικά:** Εφαρμογή StandardScaler και RobustScaler για την κλιμάκωση των τιμών, ώστε να μην επηρεάζονται τα μοντέλα από διαφορετικές κλίμακες μεγεθών.

- **Κατηγορικά Χαρακτηριστικά:** Χρήση OneHotEncoder, OrdinalEncoder και LabelEncoder για τη μετατροπή κατηγορικών δεδομένων σε μορφή αναγνωρίσιμη από τους αλγόριθμους.

- **Target Transformation:** Εφαρμόστηκε λογαριθμικός μετασχηματιστής (np.log1p) στη μεταβλητή-στόχο για να εξομαλυνθεί η κατανομή της και να βελτιωθεί η σύγκλιση των μοντέλων.

3. Μοντέλα Μηχανικής και Βαθιάς Μάθησης

Για την επίλυση του προβλήματος επιλέχθηκαν τρεις αλγόριθμοι Μηχανικής Μάθησης (Machine Learning) και ένας Βαθιάς Μάθησης (Deep Learning), ώστε να καλυφθεί ένα ευρύ φάσμα μεθοδολογιών.

Ανάλυση Επιλογής Μοντέλων & Περιορισμοί

Random Forest Regressor (Baseline):

- **Γιατί επιλέχθηκε:** Είναι ένα "ensemble" μοντέλο που βασίζεται σε πολλαπλά δέντρα απόφασης. Είναι εξαιρετικά ανθεκτικό σε ακραίες τιμές (outliers) και δεν απαιτεί αυστηρές υποθέσεις για την κατανομή των δεδομένων.

- **Πού δουλεύει καλά:** Σε περιπτώσεις όπου υπάρχουν σύνθετες αλληλεπιδράσεις μεταξύ των χαρακτηριστικών.

- **Περιορισμοί:** Μπορεί να γίνει πολύ αργό στην πρόβλεψη αν τα δέντρα είναι πάρα πολλά και τείνει να κάνει "overfit" αν δεν περιοριστεί το βάθος των δέντρων. Δυσκολεύεται να προβλέψει τιμές εκτός του εύρους των δεδομένων εκπαίδευσης.

LightGBM (Light Gradient Boosting Machine):

- **Γιατί επιλέχθηκε:** Χρησιμοποιεί την τεχνική "leaf-wise" ανάπτυξης, η οποία επιτρέπει την ταχύτερη σύγκλιση και υψηλότερη ακρίβεια σε μεγάλα σύνολα δεδομένων.

- **Πού δουλεύει καλά:** Όταν η ταχύτητα εκπαίδευσης και η χαμηλή χρήση μνήμης είναι προτεραιότητα.
- **Περιορισμοί:** Είναι ευαίσθητο στο overfitting όταν τα δεδομένα είναι λίγα. Απαιτεί προσεκτική ρύθμιση των υπερπαραμέτρων (hyperparameter tuning).

CatBoost Regressor:

- **Γιατί επιλέχθηκε:** Είναι κορυφαίο στη διαχείριση **κατηγορικών δεδομένων** (categorical features) χωρίς την ανάγκη για εκτεταμένο encoding (όπως One-Hot), μειώνοντας τον κίνδυνο απώλειας πληροφορίας.
- **Πού δουλεύει καλά:** Σε δεδομένα με πολλές μεταβλητές (όπως οι απαντήσεις σε ερωτηματολόγια ερευνών).
- **Περιορισμοί:** Η εκπαίδευσή του μπορεί να είναι χρονοβόρα σε GPUs αν οι κατηγορικές μεταβλητές έχουν πάρα πολλά distinct values (high cardinality).

TabNet (Deep Learning):

- **Γιατί επιλέχθηκε:** Συνδυάζει τη δύναμη των Νευρωνικών Δικτύων με τη δομή των δέντρων απόφασης. Χρησιμοποιεί "sequential attention" για να επιλέγει ποια χαρακτηριστικά θα προσέξει σε κάθε βήμα.
- **Πού δουλεύει καλά:** Σε πολύπλοκα tabular δεδομένα όπου υπάρχουν κρυφά μοτίβα που τα κλασικά δέντρα μπορεί να χάσουν.
- **Περιορισμοί:** Απαιτεί πολύ περισσότερα δεδομένα για να αποδώσει καλά σε σχέση με τα XGBoost/LightGBM και είναι "μαύρο κουτί" (δύσκολη ερμηνεία του αποτελέσματος).

4. Σημαντικότητα Χαρακτηριστικών (Feature Importance)

Τα μοντέλα έδωσαν ιδιαίτερη έμφαση στα παρακάτω χαρακτηριστικά:

- **Σύνθεση Νοικοκυριού:** Ο αριθμός των μελών είναι ο καθοριστικότερος παράγοντας, καθώς επηρεάζει άμεσα το κατά κεφαλήν εισόδημα και τις ανάγκες κατανάλωσης.
- **Γεωγραφία & Υποδομές:** Η περιοχή διαμονής (αστική/αγροτική) και η ποιότητα της κατοικίας λειτουργούν ως έμμεσοι δείκτες (proxies) για το επίπεδο φτώχειας.
- **Επίπεδο Εκπαίδευσης:** Η εκπαίδευση του νοικοκυριού αποδείχθηκε ισχυρός προγνωστικός παράγοντας για την οικονομική σταθερότητα.

Ποιότητα Προβλέψεων: Η έμφαση σε αυτά τα χαρακτηριστικά είναι ζωτικής σημασίας. Αν το μοντέλο αγνοούσε το μέγεθος του νοικοκυριού, οι προβλέψεις θα ήταν ανακριβείς, καθώς ένα υψηλό συνολικό εισόδημα σε ένα 10μελές νοικοκυριό μπορεί να υποκρύπτει φτώχεια.

5. Τρόποι Βελτίωσης των Δεδομένων

Για να βελτιωθεί η ακρίβεια των προβλέψεων στο μέλλον, προτείνονται τα εξής:

Πρόσθετη Πληροφορία:

- Δεδομένα Απασχόλησης: Πληροφορίες για το αν τα μέλη εργάζονται στον επίσημο ή ανεπίσημο τομέα.

- Πρόσβαση σε Χρηματοπιστωτικές Υπηρεσίες: Αν το νοικοκυριό έχει τραπεζικό λογαριασμό ή δάνεια.
- Τιμαριθμικά Δεδομένα: Το κόστος ζωής ανά περιοχή θα βοηθούσε το μοντέλο να "σταθμίσει" καλύτερα την κατανάλωση.

Feature Engineering:

- Δημιουργία ratios, όπως "αριθμός δωματίων ανά άτομο" ή "ποσοστό εξαρτώμενων μελών (παιδιά/ηλικιωμένοι) προς εργαζόμενους".

Αντιμετώπιση Θορύβου:

- Περαιτέρω καθαρισμός των δεδομένων σε μεταβλητές με πολλά "μηδενικά" ή κενά, τα οποία μπορεί να οφείλονται σε άρνηση απάντηση.

6. Επικύρωση Μοντέλων (Validation Strategy)

Η διαδικασία αξιολόγησης σχεδιάστηκε με γνώμονα την ιδιαιτερότητα των δεδομένων ερευνών νοικοκυριών, όπου κάθε γραμμή (παρατήρηση) δεν έχει την ίδια βαρύτητα και η τυχαία δειγματοληψία μπορεί να οδηγήσει σε λανθασμένα συμπεράσματα.

Α. Διαχωρισμός Δεδομένων (Hold-out Split)

Αρχικά, το συνολικό σύνολο δεδομένων χωρίστηκε σε δύο μέρη:

- **Training Set (Σύνολο Εκπαίδευσης):** Χρησιμοποιήθηκε για την εκπαίδευση των αλγορίθμων και τη ρύθμιση των παραμέτρων τους.
- **Test Set (Σύνολο Δοκιμής):** Ένα ποσοστό των δεδομένων κρατήθηκε "κρυφό" από το μοντέλο. Η τελική αξιολόγηση έγινε σε αυτό το σύνολο, προσομοιώνοντας την απόδοση του μοντέλου σε πραγματικές συνθήκες με άγνωστα δεδομένα.

Β. Μετρική Αξιολόγησης: wMAPE

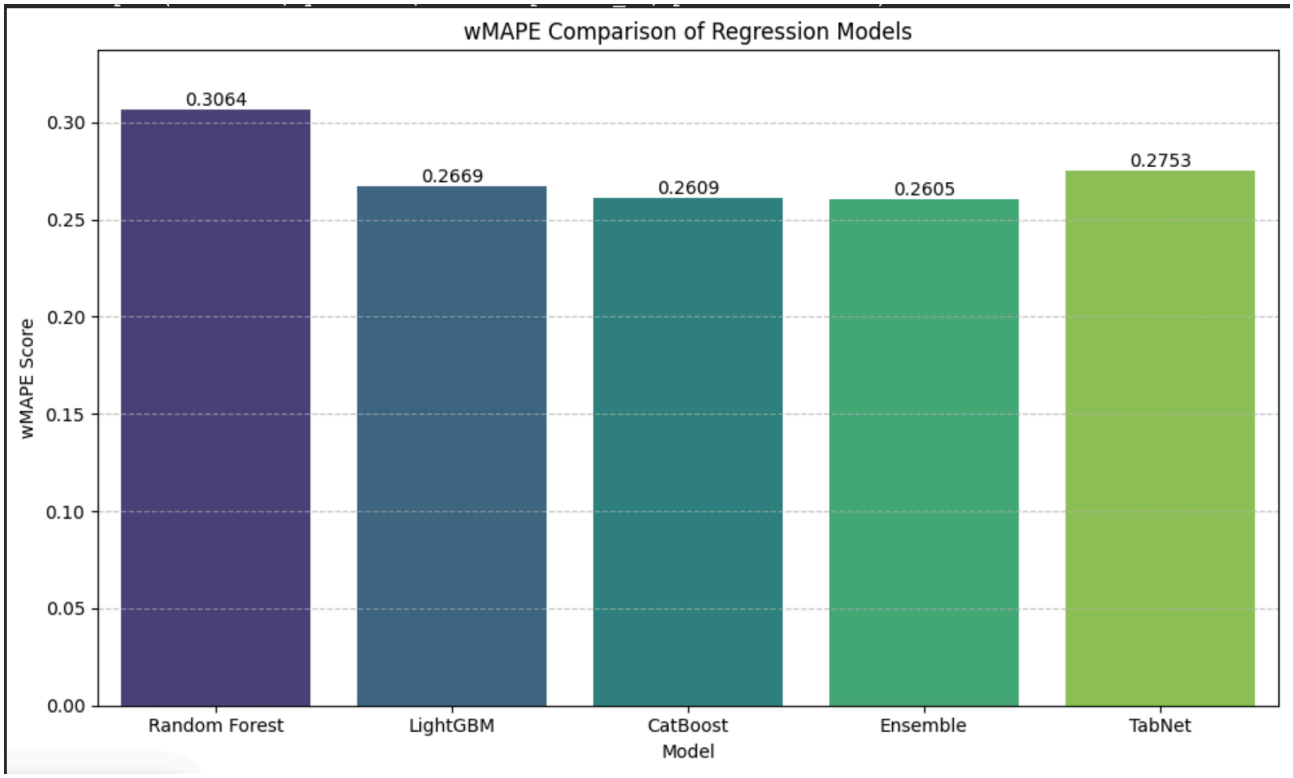
Η επιλογή της μετρικής **wMAPE (Weighted Mean Absolute Percentage Error)** έγινε για τους εξής λόγους:

1. **Ενσωμάτωση Βαρών (Sample Weights):** Στις κοινωνικοοικονομικές έρευνες, κάθε νοικοκυριό αντιπροσωπεύει έναν συγκεκριμένο αριθμό νοικοκυριών στον γενικό πληθυσμό (μέσω της μεταβλητής weight).
2. **Δίκαιη Αξιολόγηση:** Χρησιμοποιώντας τα βάρη, το μοντέλο "τιμωρείται" περισσότερο αν κάνει λάθος σε ένα νοικοκυριό που αντιπροσωπεύει μεγάλο μέρος του πληθυσμού, καθιστώντας την πρόβλεψη στατιστικά πιο έγκυρη για την εθνική στρατηγική κατά της φτώχειας.

Δ. Λογαριθμικός Μετασχηματισμός (Target Transformation)

Κατά την επικύρωση παρατηρήθηκε ότι η μεταβλητή-στόχος είχε έντονη δεξιά κλίση (skewness). Εφαρμόστηκε ο μετασχηματισμός log(1+x) πριν την εκπαίδευση και η αντίστροφη συνάρτηση (exponential) για την αξιολόγηση. Αυτό βοήθησε τα μοντέλα να "μάθουν" καλύτερα τις χαμηλότερες τιμές (που αφορούν τα φτωχότερα στρώματα) χωρίς να επηρεάζονται δυσανάλογα από τις πολύ υψηλές τιμές κατανάλωσης.

7. Σύγκριση Μοντέλων(Γράφημα)



Αποδεικτικό Υποβολής

Best score
12.631

Current rank
#227

Submissions used
2 of 3

Make new submission

You have **1 of 3** submissions left per 7 days. Your next submission can be on Jan. 21, 2026 UTC.