Αριστοτέλειο Πανεπίστημιο Θεσσαλονίκης

Τμήμα Πληροφορικής

---

Μηχανική Μάθηση - Τελική Εργασία

---

Γεώργιος Θανασούλης, Αριθμός Μητρώου: 205

Ιανουάριος 2026, Θεσσαλονίκη

# 1 Επεξεργασία Δεδομένων & Ανάλυση Δεδομένων

Η προεπεξεργασία των δεδομένων σχεδιάστηκε με βασικό στόχο την αποφυγή διαρροής πληροφορίας (target leakage) και τη διασφάλιση της γενίκευσης του μοντέλου στα μη παρατηρημένα survey panels. Από το αρχικό σύνολο περίπου 88 χαρακτηριστικών ανά νοικοκυριό, αφαιρέθηκαν όλες οι μεταβλητές που σχετίζονται άμεσα ή έμμεσα με την κατανάλωση, όπως η πραγματική ημερήσια κατά κεφαλήν κατανάλωση (cons_ppp17), οι δαπάνες (utl_exp_ppp17) και όλα τα poverty-line χαρακτηριστικά (_pline), τα οποία προκύπτουν από την κατανομή της κατανάλωσης και θα οδηγούσαν σε τεχνητά χαμηλό σφάλμα. Επιπλέον, αφαιρέθηκαν μεταβλητές που σχετίζονται με τη δειγματοληψία, όπως τα βάρη (weight) και τα strata, καθώς αυτές δεν αποτελούν πληροφορία διαθέσιμη στο στάδιο της πρόβλεψης και χρησιμοποιούνται αποκλειστικά στο aggregation για τον υπολογισμό των ποσοστών φτώχειας.

Το τελικό σύνολο εισόδου αποτελείται από 26 αριθμητικά χαρακτηριστικά, τα οποία περιγράφουν τη δημογραφική σύνθεση του νοικοκυριού (μέγεθος, ηλικία, αριθμό παιδιών και ηλικιωμένων), την εκπαίδευση και την απασχόληση των μελών, τις συνθήκες στέγασης και πρόσβασης σε βασικές υποδομές (νερό, ηλεκτρισμός, αποχέτευση), καθώς και γεωγραφικούς δείκτες περιοχής και αστικότητας. Οι κατηγορικές μεταβλητές διατηρήθηκαν σε αριθμητική κωδικοποίηση χωρίς one-hot encoding, ώστε να περιοριστεί η διάσταση του προβλήματος και να αποφευχθεί υπερπροσαρμογή, αξιοποιώντας τη μη γραμμική ικανότητα του MLP. Η μεταβλητή στόχος, δηλαδή η κατανάλωση ανά άτομο, μετασχηματίστηκε λογαριθμικά με τον τύπο log(1 + cons_ppp17), καθώς η

κατανομή της είναι έντονα δεξιά ασύμμετρη και τα poverty metrics είναι ιδιαίτερα ευαίσθητα στις χαμηλές τιμές κατανάλωσης.

Η ίδια ακριβώς προεπεξεργασία εφαρμόστηκε τόσο στο training όσο και στο test σύνολο, χωρίς χρήση καμίας πληροφορίας από τα test surveys κατά την εκπαίδευση, εξασφαλίζοντας αυστηρή απομόνωση των συνόλων. Τα βάρη πληθυσμού διατηρήθηκαν εκτός μοντέλου και χρησιμοποιούνται μόνο στο στάδιο της αξιολόγησης, για τον υπολογισμό των ποσοστών φτώχειας ανά survey μέσω σταθμισμένων αθροισμάτων. Η προεπεξεργασία αυτή επέτρεψε τη σταθερή εκπαίδευση βαθιών MLP μοντέλων, την εφαρμογή ensemble στρατηγικών και quantile mapping, και τελικά τη βελτιστοποίηση του βασικού metric του διαγωνισμού (weighted poverty wMAPE), το οποίο εξαρτάται κυρίως από τη σωστή αναπαράσταση της κατανομής της κατανάλωσης και όχι μόνο από το μέσο σφάλμα πρόβλεψης.

# 2 Εφαρμογή Αλγορίθμων Μηχανικής Μάθησης

Η αρχιτεκτονική που χρησιμοποιήθηκε βασίζεται σε ένα Mixture of Experts (MoE) σχήμα, όπου κάθε expert είναι ένα βαθύ MLP με residual συνδέσεις. Αντί να εκπαιδευτεί ένα μόνο μοντέλο, εκπαιδεύτηκαν πολλαπλά ανεξάρτητα MLPs με διαφορετικά random seeds, ίδια αρχιτεκτονική και ίδια δεδομένα. Κάθε expert μαθαίνει ελαφρώς διαφορετική προσέγγιση της συνάρτησης κατανάλωσης λόγω της διαφορετικής αρχικοποίησης και της στοχαστικότητας του optimizer. Κατά το inference, οι προβλέψεις των experts συνδυάζονται (μέσος όρος ή quantile-aware aggregation), επιτρέποντας στο τελικό σύστημα να μειώσει τη διακύμανση, να εξομαλύνει ακραίες προβλέψεις και να προσεγγίσει καλύτερα ολόκληρη την κατανομή της κατανάλωσης κάτι κρίσιμο για την ακρίβεια των poverty metrics.

Κάθε expert υλοποιείται ως βαθύ Residual MLP. Η αρχιτεκτονική ξεκινά με ένα κρύφα επίπεδα που προβάλλουν τα 26 χαρακτηριστικά εισόδου σε έναν υψηλότερης διάστασης χώρο. Ακολουθεί ένα stack πλήρως συνδεδεμένων επιπέδων με Batch Normalization, ReLU ενεργοποιήσεις και Dropout, που επιτρέπουν σταθερή και βαθιά εκπαίδευση χωρίς υπερπροσαρμογή. Στο τελευταίο στάδιο προστίθεται ένα ResidualMLPBlock, το οποίο μαθαίνει μια διορθωτική συνάρτηση πάνω στο ήδη εξαγόμενο embedding. Η residual σύνδεση (skip connection) βοηθά στη διατήρηση της πληροφορίας και στη σταθερότητα των gradients, επιτρέποντας στο μοντέλο να εστιάσει σε λεπτές μη γραμμικές διορθώσεις της πρόβλεψης. Το τελικό head είναι ένας γραμμικός προβολέας που επιστρέφει μία scalar τιμή: το log-transformed cons_ppp17.

Το συνολικό pipeline υλοποιήθηκε σε διακριτές φάσεις με καθαρό διαχωρισμό ευθυνών. Αρχικά, δημιουργήθηκαν custom PyTorch Dataset classes που φορτώνουν τα προεπεξεργασμένα CSV και επιστρέφουν tensors μαζί με τα

αναγνωριστικά survey_id και hhid. Στη συνέχεια, μέσω builder functions (build_loaders) κατασκευάστηκαν DataLoaders για train, validation και test, με διαχωρισμό των survey panels (π.χ. το survey 300000 χρησιμοποιείται αποκλειστικά για validation). Η εκπαίδευση γίνεται σε epochs με κοινό loss (L1 πάνω στο log-target) για train και validation, ενώ σε κάθε epoch υπολογίζεται επιπλέον το weighted poverty wMAPE πάνω στο validation survey, χρησιμοποιώντας population weights και poverty thresholds. Το pipeline ολοκληρώνεται με early stopping, αποθήκευση checkpoints ανά seed και τελικό ensemble inference, το οποίο τροφοδοτεί απευθείας το submission script για την παραγωγή των δύο απαιτούμενων αρχείων της διοργάνωσης.

## 2.1 Quantile Mapping

Μετά την αρχική εκπαίδευση του μοντέλου, παρατηρήθηκε ότι παρότι το household-level loss (MAE / L1) ήταν ικανοποιητικό, οι κατανομές της προβλεπόμενης κατανάλωσης δεν ταίριαζαν καλά με τις πραγματικές κατανομές των surveys. Αυτό είναι κρίσιμο, γιατί το βασικό metric του διαγωνισμού βασίζεται όχι σε μεμονωμένες προβλέψεις, αλλά στη σωστή εκτίμηση των poverty rates, οι οποίες εξαρτώνται άμεσα από τη μορφή της συνολικής κατανομής.

Για τον λόγο αυτό εφαρμόστηκε quantile mapping, μια τεχνική κατανομικής βαθμονόμησης. Συγκεκριμένα, οι προβλέψεις του μοντέλου ευθυγραμμίστηκαν ποσοστημοριακά (quantile-wise) με την κατανομή της πραγματικής κατανάλωσης στο validation survey (survey 300000). Έτσι, διατηρείται η σχετική κατάταξη των νοικοκυριών (ranking), αλλά διορθώνεται η κλίμακα και η κατανομή ώστε τα ποσοστά κάτω από κάθε poverty threshold να προσεγγίζουν καλύτερα τα πραγματικά.

Επιπλέον, Το χαρακτηριστικό κατά κεφαλήν κατανάλωση (cons_ppp17) παρουσιάζει έντονη δεξιά ασυμμετρία (heavy tail): λίγα νοικοκυριά έχουν πολύ υψηλή κατανάλωση, ενώ η πλειονότητα βρίσκεται σε χαμηλές τιμές. Η απευθείας εκπαίδευση σε γραμμική κλίμακα οδηγεί το μοντέλο να δίνει δυσανάλογη έμφαση στα μεγάλα outliers, κάτι που βλάπτει τόσο το household MAE όσο και την εκτίμηση των poverty rates, για τον λόγο αυτό, το target μετασχηματίστηκε σε λογαριθμική κλίμακα, και για τον τελικό υπολογισμό των αποτελεσμάτων επέστρεφε στην αρχική κλίμακα.

## 2.2 Επεξήγηση Αποτελεσμάτων & Περιορισμοί Μοντέλων

Παρά τη βελτιωμένη απόδοση, τα μοντέλα που χρησιμοποιήθηκαν (MLP με ensemble και quantile calibration) έχουν συγκεκριμένους περιορισμούς. Πρώτον, πρόκειται για καθαρά επιβλεπόμενα μοντέλα παλινδρόμησης, τα οποία βασίζονται αποκλειστικά στα διαθέσιμα survey χαρακτηριστικά και δεν ενσωματώνουν ρητά

οικονομικούς ή κοινωνικούς μηχανισμούς που επηρεάζουν την κατανάλωση (π.χ. τοπικές τιμές αγαθών, πληθωρισμό, άτυπη οικονομία). Δεύτερον, το μοντέλο υποθέτει ότι η σχέση μεταξύ χαρακτηριστικών και κατανάλωσης είναι σταθερή εντός κάθε survey, κάτι που δεν ισχύει πάντα στην πράξη. Τέλος, η ανάγκη για quantile mapping δείχνει ότι το μοντέλο, από μόνο του, δεν αναπαράγει τέλεια τη συνολική κατανομή της κατανάλωσης, αλλά απαιτεί μεταγενέστερη βαθμονόμηση.

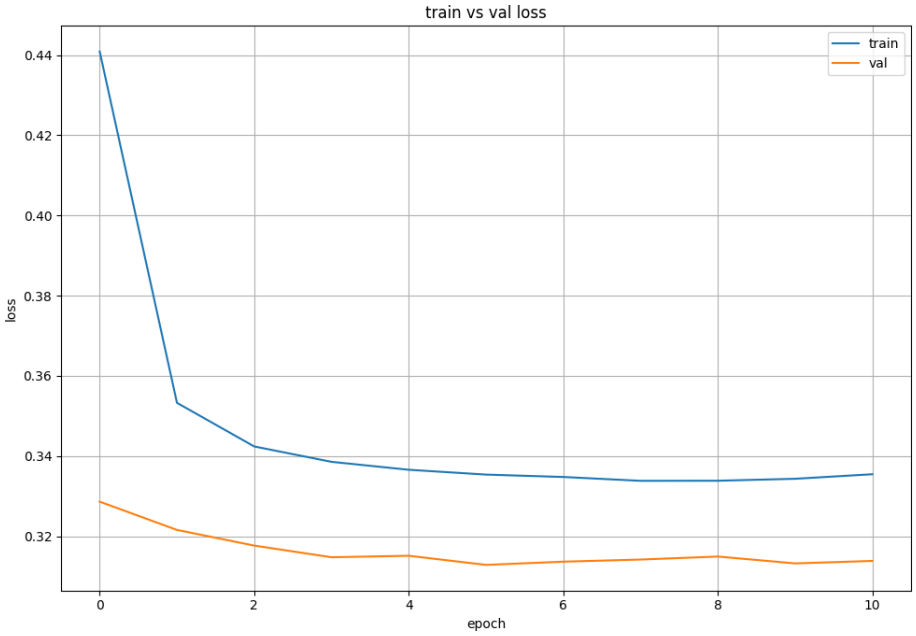
Τα μοντέλα αποδίδουν καλύτερα σε μεσαία επίπεδα κατανάλωσης, όπου υπάρχει υψηλή πυκνότητα δειγμάτων και επαρκής πληροφορία στα χαρακτηριστικά. Εκεί επιτυγχάνεται καλή πρόβλεψη τόσο σε επίπεδο νοικοκυριού όσο και σε επίπεδο poverty rates. Αντίθετα, παρουσιάζουν μεγαλύτερη αβεβαιότητα στα άκρα της κατανομής: πολύ φτωχά ή πολύ εύπορα νοικοκυριά. Στις χαμηλές τιμές, μικρά απόλυτα σφάλματα οδηγούν σε μεγάλα σχετικα σφάλματα (MAPE), ενώ στις υψηλές τιμές η σπανιότητα δεδομένων δυσκολεύει τη γενίκευση. Επιπλέον, η απόδοση είναι καλύτερη όταν το test survey είναι κατανομικά κοντά στα training surveys, σε περιπτωσεις απόκλισης (distribution shift) η ανάγκη για calibration γίνεται εντονότερη.

## 2.3 Αποτελέσματα Εκπαίδευσης - Επικύρωσης

Η εκπαίδευση του τελικού μοντέλου πραγματοποιήθηκε σε περιβάλλον GPU και παρουσίασε γρήγορη και σταθερή σύγκλιση. Συγκεκριμένα, το μοντέλο εκπαιδεύτηκε για 11 epochs, με early stopping να ενεργοποιείται βάσει του validation loss. Η καλύτερη τιμή validation loss ήταν 0.313 (L1 loss σε λογαριθμική κλίμακα), γεγονός που δείχνει ικανοποιητική ακρίβεια σε επίπεδο πρόβλεψης νοικοκυριών. Παράλληλα, η τελική τιμή του poverty weighted wMAPE στο validation set ανήλθε σε 0.0999, επιβεβαιώνοντας ότι το μοντέλο, σε συνδυασμό με τη λογαριθμική κλίμακα και το calibration, αποδίδει καλύτερα στις κρίσιμες περιοχές της κατανομής που επηρεάζουν τον δείκτη φτώχειας. Ο συνολικός χρόνος εκπαίδευσης ήταν περίπου 51 δευτερόλεπτα, κάτι που καθιστά τη μεθοδολογία αποδοτική και επαναλήψιμη.

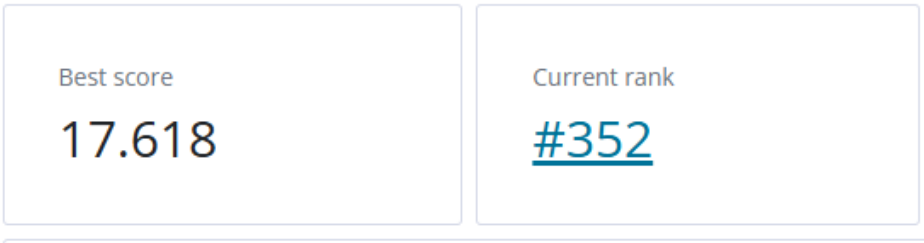
Το διάγραμμα **??** δείχνει ότι το μοντέλο συγκλίνει γρήγορα και σταθερά μέσα στα πρώτα epochs. Παρατηρείται απότομη μείωση του training loss στο αρχικό στάδιο (epoch $0 \rightarrow 1$), γεγονός που υποδηλώνει ότι το μοντέλο μαθαίνει αποτελεσματικά τις βασικές σχέσεις μεταξύ των χαρακτηριστικών και της κατανάλωσης. Στη συνέχεια, τόσο το training όσο και το validation loss μειώνονται πιο ομαλά και σταθεροποιούνται γύρω από τις τιμές 0.33–0.34 και 0.31–0.32 αντίστοιχα. Η μικρή και σταθερή απόσταση μεταξύ των δύο καμπυλών υποδεικνύει καλή γενίκευση και απουσία overfitting. Επιπλέον, η ελαφρά ανοδική τάση του training loss στα τελευταία epochs, χωρίς αντίστοιχη επιδείνωση στο validation loss, δικαιολογεί τη χρήση early stopping και επιβεβαιώνει ότι το μοντέλο

έχει φτάσει σε σημείο κορεσμού, όπου περαιτέρω εκπαίδευση δεν προσφέρει ουσιαστική βελτίωση στην απόδοση.




Σχήμα 1: Καμπύλες απώλειας εκπαίδευσης και επικύρωσης

Τέλος η αναφορά και του αποτελέσματος του από το leaderboard, όταν το έκανα submit ήταν στο 200. Από τότε από όσο φαίνεται μπήκαν παραπάνω χρήστες και πέτυχαν καλύτερα αποτελέσματα στις συγκεκριμένες μετρικές.




Σχήμα 2: Αποτέλεσματα Leaderboard