Επεξεργασία Δεδομένων

- Για την αποτελεσματική διαχείριση των δεδομένων και την αποφυγή ασυμβατότητας χαρακτηριστικών μεταξύ των συνόλων εκπαίδευσης (Train) και δοκιμής (Test), συνενώθηκαν αρχικά τα δύο αρχεία σε ένα κοινό DataFrame, εφαρμόζοντας ενιαίους κανόνες προ-επεξεργασίας: συμπλήρωση των ελλειπουσών τιμών (με τη διάμεσο για τα αριθμητικά και την ένδειξη 'missing' για τα κατηγορικά) και μετατροπή των κατηγοριών σε αριθμητική μορφή μέσω One-Hot Encoding. Με την μέθοδο αυτή διασφαλίστηκε ότι και τα δύο σύνολα δεδομένων απέκτησαν ακριβώς την ίδια δομή στηλών, εξαλείφοντας τον κίνδυνο τεχνικών σφαλμάτων λόγω διαφορετικών κατηγοριών. Μετά την ολοκλήρωση της κωδικοποίησης, διαχωρίστηκαν εκ νέου τα δεδομένα, προετοιμάζοντας το μεν σύνολο εκπαίδευσης με την ενσωμάτωση των ετικετών στόχων (targets), το δε σύνολο δοκιμής για την παραγωγή των τελικών προβλέψεων του μοντέλου.
- **Ένωση (Merging):** Τα δεδομένα ήταν χωρισμένα σε χαρακτηριστικά (features) και ετικέτες (gt). Η ένωση έγινε με βάση το μοναδικό κλειδί το οποίο αποτελείται από 2 στήλες, τις hhid και survey_id ώστε να αντιστοιχηθεί κάθε νοικοκυριό με την κατανάλωση.
- **Missing Values:** Εντοπίστηκαν ελλείψεις σε στήλες όπως η sector1d και η dweltyp οπότε χρησιμοποιήθηκαν η τεχνική της διάμεσης τιμής (median) για αριθμητικές τιμές και η δημιουργία ειδικής κατηγορίας "missing" για τις κατηγορικές, ώστε να μη χαθεί πληροφορία.
- **One-Hot Encoding:** Επειδή οι αλγόριθμοι δεν καταλαβαίνουν κείμενο (π.χ. "Urban", "Rural"), όλες τις κατηγορικές μεταβλητές μετατράπηκαν σε δυαδικά χαρακτηριστικά/στήλες (0/1).

Ανάλυση Δεδομένων

Συσχετίσεις: Από τον πίνακα συσχέτισης (Correlation Matrix) βρέθηκε ότι:

- **Θετική Συσχέτιση:** Η μεταβλητή utl_exp_ppp17 (έξοδα για κοινοχρηστα/Expenditure on utilities) έχει την ισχυρότερη θετική σχέση με την κατανάλωση. Όσο περισσότερα ξοδεύει κάποιος για ρεύμα/νερό, τόσο πιθανότερο είναι να έχει υψηλότερη κατανάλωση γενικά.
- **Αρνητική Συσχέτιση:** Το hsize (μέγεθος νοικοκυριού) έχει αρνητική συσχέτιση. Μεγάλες οικογένειες τείνουν να έχουν μικρότερη κατά κεφαλήν κατανάλωση.

Ανάλυση των 4 Αλγορίθμων

Χρησιμοποιήθηκαν 4 διαφορετικοί αλγόριθμοι:

1. Linear Regression (Γραμμική Παλινδρόμηση)

Ο πιο απλός αλγόριθμος ο οποίος προσπαθεί να τραβήξει μια ευθεία γραμμή που να περνάει όσο το δυνατόν πιο κοντά από όλα τα σημεία των δεδομένων. Η εξίσωση

είναι της μορφής y = w1*x1 + w2*x2 + ... + b. Επειδή η φτώχεια είναι πολύπλοκο φαινόμενο και δεν ακολουθεί απλούς γραμμικούς κανόνες, λογικά έχει χαμηλή απόδοση.

2. Random Forest Regressor (Τυχαίο Δάσος)

Ανήκει στην κατηγορία **Ensemble (Bagging)**. Αντί να φτιάξουμε ένα δέντρο αποφάσεων, φτιάχνουμε πολλά (π.χ. 50 δέντρα - n_estimators=50). Κάθε δέντρο εκπαιδεύεται σε ένα τυχαίο υποσύνολο των δεδομένων και κάνει τη δική του πρόβλεψη. Στο τέλος, βγάζουμε τον **μέσο όρο** όλων των δέντρων. Ο αλγόριθμος πετυχαίνει συνήθως πολύ καλή ακρίβεια, δεν παθαίνει εύκολα overfitting και μπορεί να βρει μη γραμμικές σχέσεις στα δεδομένα (π.χ. αγροτική περιοχή ΚΑΙ 5 παιδιά -> φτώχεια).

3. Gradient Boosting Regressor

Επίσης κατηγορία **Ensemble (Boosting)**, αλλά με διαφορετική λογική από το Random Forest.

- Εδώ τα δέντρα χτίζονται σειριακά (το ένα μετά το άλλο).
- Το Δέντρο 1 κάνει μια πρόβλεψη.
- Το Δέντρο 2 βλέπει τα λάθη του Δέντρου 1 και προσπαθεί να διορθώσει μόνο αυτά.
- Το Δέντρο 3 διορθώνει τα λάθη του Δέντρου 2, κ.ο.κ.

Επίσης συνήθως πετυχαίνει πολύ υψηλή ακρίβεια.

4. Deep Learning (MLP - Multi-Layer Perceptron)

Ένα Τεχνητό Νευρωνικό Δίκτυο που μιμείται τη λειτουργία του εγκεφάλου.

- **Input Layer:** Τα χαρακτηριστικά εισόδου (μετά το scaling).
- **Hidden Layers (Dense):** Ενδιάμεσα επίπεδα όπου οι "νευρώνες" συνδυάζουν τα χαρακτηριστικά. Η συνάρτηση **ReLU** (activation='relu') επιτρέπει στο δίκτυο να μαθαίνει πολύπλοκα, μη γραμμικά μοτίβα (κόβει τις αρνητικές τιμές και κρατάει τις θετικές).
- **Dropout (0.2):** Μια τεχνική όπου σε κάθε βήμα εκπαίδευσης, απενεργοποιείται τυχαία το 20% των νευρώνων. Αυτό αναγκάζει το δίκτυο να μην εξαρτάται από συγκεκριμένα μονοπάτια και αποτρέπει το Overfitting.
- **Output Layer:** Ένας μόνο νευρώνας στο τέλος, που βγάζει την τελική τιμή (κατανάλωση).

Σύγκριση

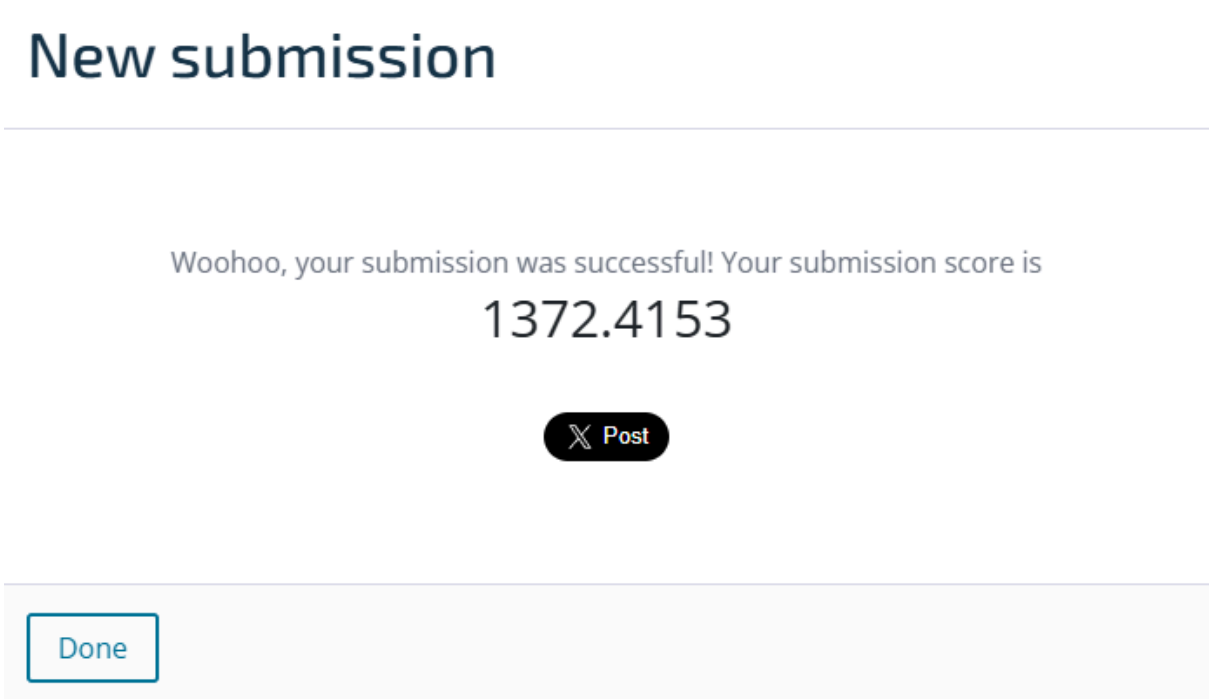
1. **Linear Regression:** Ως **Baseline**. Πέτυχε την μικρότερη ακρίβεια στην μετρική R2(0.5299), γεγονός που δείχνει ότι η σχέση των χαρακτηριστικών με τη φτώχεια δεν είναι γραμμική.
2. **Random Forest:** Ένας ισχυρός αλγόριθμος που χειρίζεται μη γραμμικές σχέσεις. Πέτυχε ικανοποιητικό R2 (0.6064).

3. **Gradient Boosting:** Βελτιώνει τα λάθη των προηγούμενων δέντρων. Πέτυχε καλύτερο αποτέλεσμα από το Random Forest (R2 = 0.6381).
4. **Deep Learning (MLP):** Το νευρωνικό δίκτυο πέτυχε το καλύτερο αποτέλεσμα (R2 = 0.6397) και το χαμηλότερο σφάλμα (MSE = 36.1981), αποδεικνύοντας ότι μπορεί να μάθει πολύπλοκα μοτίβα στα δεδομένα αυτά.

Η μετρική R2 διότι προσφέρει μια εικόνα της απόδοσης, ανεξάρτητη από την κλίμακα των οικονομικών μεγεθών. Το R2=0.6397 του βέλτιστου μοντέλου (MLP) δείχνει ότι καταφέραμε να μοντελοποιήσουμε επιτυχώς σχεδόν το 65% της διακύμανσης της κατανάλωσης των νοικοκυριών ενώ η μετρική MSE = 36.1981$ που δείχνει μια απόκλιση κοντά στα 36 δολάρια δεν μας βοηθά να καταλάβουμε αν μια απόκλιση 36 δολαρίων είναι μεγάλη ή μικρή.

Δημοσίευση των Αποτελεσμάτων

- Αλλαγή ονομάτων των στηλών (το hhid γίνεται household_id).
- Υπολογισμός των στατιστικών:
    - Παίρνουμε όλα τα σπίτια μιας έρευνας (π.χ. της έρευνας 400000).
    - Ελέγχουμε π.χ. πόσα από αυτά έχουν κατανάλωση κάτω από 3.17$ -> Έστω το 10%.
    - Ελέγχουμε πόσα κάτω από 3.94$ -> Έστω το 15%.
    - Επανάληψη για όλα τα όρια μέχρι να φτιαχτεί ο πίνακας κατανομής.
- Zip: Μπαίνουν όλα σε ένα φάκελο submission.zip γιατί έτσι το ζητάει η πλατφόρμα DrivenData.



#132	kgl-ctf 2w 3d ago · 1 submission	100.000	90.000	
#133	yooon 1d 22h ago · 1 submission	100.000	90.000	
#134	jrramos 1w 4d ago · 3 submissions	331.545	301.982	
#135	AyushAv05 4d 20h ago · 3 submissions	339.357	69.647	
#136	vpapoglou 1min ago · 1 submission	1372.415	89.833	Share your work?
#137	ramtelpp 2w 1d ago · 1 submission	7768.946	7765.769	

Επεξήγηση Αποτελεσμάτων

Σημαντικότητα Χαρακτηριστικών

Όπως έδειξε η ανάλυση συσχέτισης αλλά και η συμπεριφορά των δενδρικών μοντέλων (Random Forest), τα μοντέλα δίνουν τη μεγαλύτερη έμφαση σε τρεις κατηγορίες χαρακτηριστικών:

1. **Οικονομικοί Δείκτες:** Τα έξοδα για υπηρεσίες κοινής ωφέλειας (utl_exp_ppp17) αποτελούν τον ισχυρότερο δείκτη για την κατανάλωση.
2. **Γεωγραφία:** Η περιοχή κατοικίας (region, urban/rural) καθορίζει σε μεγάλο βαθμό το βιοτικό επίπεδο, λόγω των διαφορών στο κόστος ζωής και τις ευκαιρίες απασχόλησης.
3. **Δημογραφικά Στοιχεία:** Το μέγεθος του νοικοκυριού (hsize) και η αναλογία εξαρτώμενων μελών (παιδιά) παίζουν καθοριστικό αρνητικό ρόλο στην κατά κεφαλήν κατανάλωση.

Αδυναμία μοντέλων

Η ακρίβεια των μοντέλων ίσως να μειώνεται σε αγροτικές περιοχές ή σε νοικοκυριά με άτυπες πηγές εισοδήματος (π.χ. αυτοπαραγωγή τροφίμων) που δεν καταγράφονται εύκολα στα χαρακτηριστικά που έχουμε. Επίσης όταν υπάρχει έλλειψη ιστορικότητας για τα δεδομένα τότε τα μοντέλα δεν θα κάνουν σωστές προβλέψεις. Π.χ. στις αγροτικές οικογένειες η κατανάλωση και το εισόδημα έχουν περισσότερο εποχικά χαρακτηριστικά οπότε η στιγμή της καταμέτρησης των αποτελεσμάτων μπορεί να δώσει εσφαλμένη εικόνα.

Προτάσεις Βελτίωσης Δεδομένων

Για την περαιτέρω βελτίωση των προβλέψεων, θα ήταν χρήσιμη η προσθήκη επιπλέον πληροφορίας, όπως Περιουσιακά Στοιχεία (Assets). Πληροφορίες για την κατοχή συγκεκριμένων αντικειμένων (π.χ. αυτοκίνητο, ψυγείο, smartphone) θα λειτουργούσαν ως ισχυροί δείκτες πλούτου, ειδικά στις περιπτώσεις που τα έξοδα utilities είναι χαμηλά.