Αναφορά Τελικής Εργασίας Μηχανική Μάθηση

Κωνσταντίνα Μαρίνα Μπλέτσα, ΑΕΜ: 243

Ανάλυση Δεδομένων

Το αρχείο train_hh_features.csv περιλαμβάνει 104234 νοικοκυριά και πολλά χαρακτηριστικά, τα οποία περιγράφουν την μορφή, τις συνθήκες διαβίωσης και τις καταναλωτικές συνήθειες κάθε νοικοκυριού. Κάθε εγγραφή αντιστοιχεί σε ένα μοναδικό νοικοκυριό, το οποίο έχει το αναγνωριστικό hhid.

Οι κύριες κατηγορίες που χωρίζονται τα χαρακτηριστικά είναι, τα δημογραφικά χαρακτηριστικά, όπως το φύλο του αρχηγού του νοικοκυριού (male), το συνολικό μέγεθος του νοικοκυριού (hsize) και ο αριθμός των παιδιών σε διαφορετικές ηλικιακές ομάδες. Οι μεταβλητές αυτές προσφέρουν πληροφορίες για τη δομή του νοικοκυριού και επηρεάζουν τις ανάγκες κατανάλωσης και τον βαθμό οικονομικής ευαλωτότητας. Επιπλέον, άλλη κατηγορία είναι οι οικονομικές μεταβλητές, με κυριότερη την κατανάλωση ανά άτομο σε ισοτιμία αγοραστικής δύναμης (utl_exp_ppp17), λειτουργούν ως βασικός δείκτης ευημερίας. Η συγκεκριμένη μεταβλητή αποτελεί τον πυρήνα της εκτίμησης φτώχειας, καθώς αποτυπώνει με άμεσο τρόπο το πραγματικό επίπεδο διαβίωσης των νοικοκυριών. Συμπληρωματικά, σημαντικό ρόλο παίζουν τα στατιστικά βάρη (weight) και οι μεταβλητές δειγματοληψίας, οι οποίες δεν επηρεάζουν άμεσα την πρόβλεψη σε ατομικό επίπεδο, αλλά είναι απαραίτητες για τη σωστή ερμηνεία και γενίκευση των αποτελεσμάτων επίπεδο πληθυσμού. Τέλος, το πιο μεγάλο μέρος του συνόλου δεδομένων είναι δυαδικές μεταβλητές κατανάλωσης τροφίμων, οι οποίες δηλώνουν αν το νοικοκυριό κατανάλωσε συγκεκριμένα αγαθά, και χρησιμοποιούνται ως δείκτες οικονομικής κατάστασης και ποιότητας ζωής. Νοικοκυριά με υψηλότερο επίπεδο κατανάλωσης τείνουν να παρουσιάζουν μεγαλύτερη ποικιλία διατροφικών αγαθών.

Εφαρμογή Αλγορίθμων Μηχανικής και Βαθιάς Μάθησης

Επιλέχθηκαν τρεις διαφορετικοί αλγόριθμοι Μηχανικής Μάθησης, καθώς και ενός αλγόριθμος Βαθιάς Μάθησης. Η προεπεξεργασία των δεδομένων είναι παρόμοια και στα 4 μοντέλα. Αρχικά έκανα καθαρισμό των δεδομένων, χειρισμό ελλιπών τιμών και κατάλληλη κωδικοποίηση κατηγορικών χαρακτηριστικών, ώστε τα δεδομένα να είναι συμβατά με τους αλγορίθμους που θα εφαρμοζόταν σε κάθε μοντέλο και στη συνέχεια τα δεδομένα διαχωρίζονται σε σύνολα εκπαίδευσης και αξιολόγησης.

Ο πρώτος αλγόριθμος που εφαρμόστηκε είναι ο Random Forest, επιλέχθηκε το συγκεκριμένο μοντέλο επειδή λειτουργεί δημιουργώντας πολλά διαφορετικά δέντρα αποφάσεων και συνδυάζοντας τις προβλέψεις τους σαν μέθοδος Bagging. Σκοπός ήταν να μειώσει τον κίνδυνο υπερπροσαρμογής και να δώσει πιο σταθερά και αξιόπιστα αποτελέσματα σε σχέση με ένα μόνο δέντρο και παρουσιάζει καλή απόδοση σε δεδομένα με μεγάλο αριθμό χαρακτηριστικών. Στον αντίστοιχο κώδικα γίνεται ρύθμιση βασικών υπερπαραμέτρων, όπως ο αριθμός των δέντρων και το μέγιστο βάθος, με στόχο την αποφυγή του overfitting και τη βελτίωση της

γενίκευσης του μοντέλου (η εκπαίδευση έγινε με αριθμό επαναλήψεων που να επιτρέπει το Colab). Επιπλέον, βασικό θετικό του Random Forest είναι ότι δείχνει τους δείκτες σημαντικότητας χαρακτηριστικών και ερμηνεύεται πιο εύκολα.

Ο δεύτερος αλγόριθμος που εξετάστηκε είναι ο Gradient Boosting είναι ένας αλγόριθμος μηχανικής μάθησης που ανήκει στην κατηγορία των ensemble methods, δηλαδή συνδυάζει πολλά απλά μοντέλα για να δημιουργήσει ένα καλύτερο. Συγκεκριμένα, βασίζεται σε δέντρα απόφασης, τα οποία εκπαιδεύονται διαδοχικά κάθε νέο δέντρο προσπαθεί να διορθώσει τα λάθη που έκαναν τα προηγούμενα. Με αυτόν τον τρόπο, το μοντέλο βελτιώνεται σταδιακά και μπορεί να μάθει πολύπλοκες, μη γραμμικές σχέσεις στα δεδομένα. Ο Gradient Boosting επιλέχθηκε για αυτό το πρόβλημα γιατί τα δεδομένα είναι δομημένα, περιέχουν πολλές κατηγορικές και αριθμητικές μεταβλητές και οι σχέσεις μεταξύ τους δεν είναι απλές. Τέτοιοι αλγόριθμοι είναι ιδιαίτερα αποδοτικοί σε προβλήματα πρόβλεψης κοινωνικοοικονομικών δεικτών, όπως η κατανάλωση νοικοκυριών και τα ποσοστά φτώχειας. Στον κώδικα, το βασικό μοντέλο εκπαιδεύεται πάνω στα δεδομένα εκπαίδευσης και χρησιμοποιεί παραμέτρους όπως ο ρυθμός μάθησης (learning rate), το βάθος των δέντρων (depth) και ο αριθμός των iterations, που καθορίζουν πόσο γρήγορα και πόσο πολύπλοκα μαθαίνει το μοντέλο. Στη συνέχεια εφαρμόζεται cross-validation ανά survey, ώστε να ελεγχθεί αν το μοντέλο γενικεύει καλά σε διαφορετικές έρευνες και όχι μόνο στα δεδομένα που είδε κατά την εκπαίδευση. Η συγκεκριμένη προσέγγιση αποδείχθηκε ιδιαίτερα αποτελεσματική για το πρόβλημα, πετυχαίνει την υψηλότερη βαθμολογία σε σχέση με τα υπόλοιπα μοντέλα στον διαγωνισμό.

Ο τρίτος αλγόριθμος είναι ο Support Vector Regression (SVR), επειδή μπορεί να προβλέψει αριθμητικές τιμές όταν οι σχέσεις μεταξύ των δεδομένων δεν είναι γραμμικές. Τα δεδομένα που αφορούν τα νοικοκυριά περιέχουν πολλές και διαφορετικές πληροφορίες, καθώς και θόρυβο, και το SVR είναι καλό για τέτοιες περιπτώσεις επειδή γενικεύει καλά και δεν επηρεάζεται εύκολα από ακραίες τιμές. Επιπλέον, μέσω των παραμέτρων του, επιτρέπει τον έλεγχο της ακρίβειας του μοντέλου και της υπερπροσαρμογής. Ένας βασικός περιορισμός του είναι ότι απαιτεί αυξημένο υπολογιστικό κόστος (σε χρόνο) σε μεγάλα σύνολα δεδομένων, γι αυτό έγινε η εκπαίδευση του μοντέλου και η ρύθμιση των παραμέτρων έγινε σε μικρό τυχαίο υποσύνολο του training set, ενώ το τελικό μοντέλο εκπαιδεύτηκε στο πλήρες σύνολο δεδομένων.

Τέλος, εφαρμόστηκε αλγόριθμος Βαθιάς Μάθησης, για την εκπαίδευση του μοντέλου, χρησιμοποιήθηκε ένα πολυεπίπεδο νευρωνικό δίκτυο (MLP), όπου τα δεδομένα εισόδου περνούν διαδοχικά από κρυφά επίπεδα με μη γραμμικές συναρτήσεις ενεργοποίησης, ώστε το δίκτυο να μάθει σύνθετες σχέσεις μεταξύ των χαρακτηριστικών και της κατανάλωσης. Η εκπαίδευση έγινε επαναληπτικά σε epochs (όσες επιτρέπει το Colab) με χρήση του αλγορίθμου backpropagation και βελτιστοποιητή τύπου AdamW. Αρχικά, το δίκτυο εκπαιδεύτηκε μόνο για την πρόβλεψη της κατανάλωσης, ενώ στη συνέχεια εφαρμόστηκε fine-tuning, όπου η συνάρτηση κόστους τροποποιήθηκε για να λαμβάνει υπόψη και τα ποσοστά φτώχειας ανά survey. Η χρήση του αλγορίθμου βαθιάς μάθησης είχε και κάποιες προκλήσεις, η εκπαίδευση χρειάστηκε περισσότερο χρόνο σε σχέση με άλλους αλγορίθμους, ειδικά όταν χρησιμοποιούνται πολλά κρυφά επίπεδα και επίσης δεν είναι εύκολο να εξηγήσουμε με απλό τρόπο πώς κάθε χαρακτηριστικό επηρεάζει την τελική πρόβλεψη.

Επεξήγηση και Ανάλυση Αποτελεσμάτων

Τα αποτελέσματα δείχνουν ότι κάθε μοντέλο έχει τα δικά του πλεονεκτήματα και δυσκολίες. Το μοντέλο SVR είχε μέτρια απόδοση, με RMSE περίπου 5.48 και R^2 περίπου 0.70. Αυτό σημαίνει ότι μπορεί να προβλέψει τη γενική τάση των δεδομένων, αλλά δυσκολεύεται όταν οι σχέσεις μεταξύ των χαρακτηριστικών είναι πιο σύνθετες. Επίσης, επηρεάζεται αρκετά από την επιλογή των παραμέτρων του. Το Random Forest παρουσίασε καλύτερη και πιο σταθερή απόδοση, με weighted MAPE περίπου 0.15, δείχνοντας ότι κάνει πιο αξιόπιστες προβλέψεις. Επίσης λειτουργεί καλά όταν υπάρχουν πολλές αλληλεπιδράσεις μεταξύ των χαρακτηριστικών, αλλά μπορεί να επηρεαστεί από θόρυβο στα δεδομένα. Το Gradient Boosting έδωσε επίσης από τα καλύτερα αποτελέσματα, καθώς βελτιώνει σταδιακά τα λάθη του και δίνει μεγαλύτερη σημασία στα πιο σημαντικά χαρακτηριστικά, όμως πάλι χρειάζεται προσοχή ώστε να μην υπερπροσαρμοστεί. Το μοντέλο Βαθιάς Μάθησης είχε χαμηλό σφάλμα κατά την εκπαίδευση (loss περίπου 0.31), αλλά η απόδοσή του εξαρτάται πολύ από την ποσότητα των δεδομένων και δεν αποδίδει το ίδιο καλά σε νέα δεδομένα. Γενικά, όλα τα μοντέλα δίνουν μεγαλύτερη έμφαση σε αριθμητικά και ποσοστιαία χαρακτηριστικά, τα οποία παίζουν σημαντικό ρόλο στην ποιότητα των προβλέψεων. Η απόδοση κάθε μοντέλου θα μπορούσε να βελτιωθεί με περισσότερα δεδομένα, καλύτερη προεπεξεργασία, καλύτερες υπερπαραμέτρους για τους αλγορίθμους ή ακόμα και περισσότερες επαναλήψεις στην εκπαίδευση.

3. Υποβολή στον διαγωνισμό

Η καλύτερη υποβολή σε επίπεδο διαγωνισμού ήταν αυτή του αλγορίθμου Gradient Boosting και παρακάτω είναι το screenshot από την κατάταξη στον διαγωνισμό.