Predviđanje cena kuća

Stefan Dragičević, E9-4-2022

1. Uvod

U okviru ovog projekta, postavljen je cilj da se predvide cene kuća na osnovu podataka iz dostupne baze. Baza podataka sastoji se od dva CSV fajla: Train.csv, sa dimenzijama 29451x12, i Test.csv, sa dimenzijama 68720x11. Podaci su već podeljeni na trening i test skupove. U bazi se nalazi ukupno 12 obeležja, od kojih su 3 kategorička, dok su preostalih 9 obeležja numerička. Važno je napomenuti da u bazi ne postoje nedostajuće vrednosti.

2. Inženjerstvo karakteristika

Podaci su prolazili kroz niz koraka kako bi se pripremili za dalju analizu i modeliranje. Detaljan pregled obrade podataka prikazan je u nastavku.

Prvo kategoričko obeležje ‘POSTED_BY’ je pretvoreno u numeričko tako što je ‘Owner’ zamenjen sa 0, ‘Dealer’ sa 0.5, a ‘Builder’ sa 1.

Drugo kategoričko obeležje ‘BHK_OR_RK’ ima vrednosti ‘BHK’ i ‘RK’ koje su zamenjene sa 0 i 1.

Treće kategoričko obeležje označava adresu na kojoj se objekat nalazi i ono je izbačeno jer postoje druga dva obeležja koja predstavljaju geografske koordinate i pružaju dovoljno informacija za analizu.

Obeležja 'TARGET(PRICE_IN_LACS)' i 'SQUARE_FT' imaju autlajere i oni su uklonjeni.

Nakon obrade podataka, primećeno je da obeležje 'BHK_OR_RK' ima sve vrednosti jednake 0. Zbog toga je ovo obeležje izbačeno iz dalje analize, jer ne donosi korisne informacije.

Nakon svih prethodnih koraka urađena je normalizacija podataka. Dimenzije skupa za treniranje nakon obrade podataka iznose 29434x10, što znači da je izbačeno 17 uzoraka i 2 obeležja, dok su dimenzije skupa za testiranje ostale iste.

3. Modeli

Za rešavanje ovog problema korišćeno je ukupno 8 modela i to Random Forest Regressor, XGB Regressor, Support Vector Regressor, Linear Regression model, Ridge Regression Model, KNN Regressor, Random Forest Regressor za koji je korišćeno podešavanje hiperparametara i Multilayer Perceptron Regressor čija je arhitektura određena uz pomoć AutoML-a.

Za Random Forest Regressor, hiperparametri koji su podešavani su n_estimators, max_features i max_depth. Nakon eksperimentisanja sa različitim kombinacijama ovih parametara, najbolji rezultati su postignuti sa vrednostima max_depth=20, max_features='sqrt' i n_estimators=150.

Najbolja arhitektura za MLP Regressor je ona koja se sastoji iz 3 skrivena sloja sa po 98, 80, 77 neurona.

Podaci koji se nalaze u Train.csv su podeljeni na trening i test skup tako što je 20% podataka rezervisano za testiranje modela.

4. Rezultati

Evaluacija modela je izvršena na osnovu R2 Scora, Mean Squared Error (MSE) i Mean Absolute Error (MAE). Dobijeni rezultati su prikazani u tabeli 1, gde se može videti da Random Forest Regressor ima najbolji rezultat sa MAE vrednošću od 30.42, a odmah iza njega se nalazi XGB Regressor sa MAE vrednošću 32.63

	R2 Score	Mean Squared Error	Mean Absolute Error
Random Forest Regression	0.949042	17868.455451	30.421287
RFR_hyperparam	0.950599	17322.485174	30.767437
XGBoost Regression	0.949915	17562.361601	32.639315
KNN Regression	0.916883	29144.908936	39.314861
MLPRegressor	0.907326	32496.138356	61.766929
Linear Regression	0.751762	87044.186462	77.9829
Ridge Regression	0.750093	87629.463966	78.478471
Support Vector Regression	-0.004686	352291.899573	95.552694

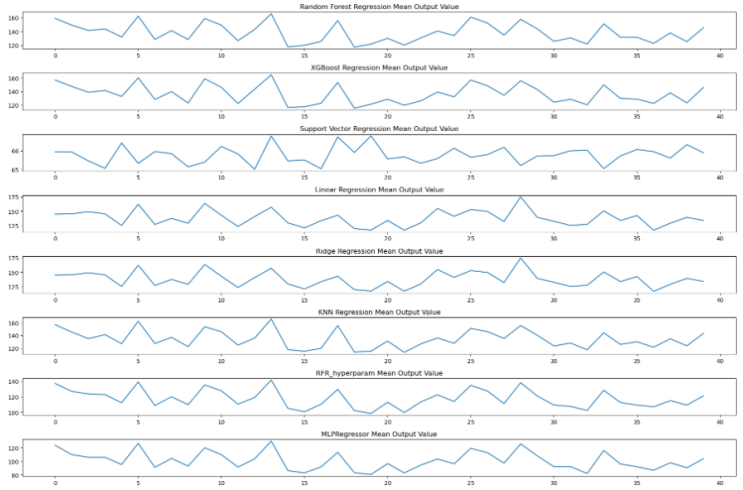
Tabela 1 - Dobijeni rezultati

5. Monitoring

Praćenje performansi AI modela omogućava identifikaciju promena u kvalitetu predviđanja ili odluka tokom vremena. Ako se performansa modela pogorša, to može ukazivati na probleme u podacima ili potrebu za podešavanjem modela. Monitoring pomaže da se identifikuju ove promene i preduzmu odgovarajuće mere za poboljšanje performansi.

Zbog nedostatka stvarnih vrednosti za y_test umesto njih je korišćena srednja vrednost svih predviđenih vrednosti y_pred_mean kako bi se izračunala srednja kvadratna greška, srednja apsolutna greška i R2 skor za svaki model.

Da bi se simulirao rad modela u određenom vremenskom periodu, y_pred je podeljen na 40 podskupova, pri čemu svaki podskup sadrži 1718 vrednosti. Nakon toga, za svaki podskup je pronađena srednja vrednost, a dobijene vrednosti su prikazane na slici 1.



Slika 1 Promena srednje vrednosti izlaza tokom vremena za svaki model

6. Zaključak

Ovaj projekat se bavio predviđanjem cena kuća na osnovu dostupnih podataka. Kroz proces inženjerstva karakteristika, podaci su bili obrađeni i pripremljeni za dalju analizu i modeliranje. Nakon obrade podataka i normalizacije, primenjeno je osam različitih modela za rešavanje problema predviđanja cena kuća. Dalje istraživanje i eksperimentisanje sa dodatnim modelima i tehnikama moglo bi unaprediti rezultate i preciznost predviđanja.