Name: OLOKEDE Oluwatobiloba Stephen

Project: DSN_AI Bootcamp Qualification 2023

INTRODUCTION

The aim of the project is to help Wazobia Real Estate Limited, a leading real estate company in Nigeria tackle one of its crucial challenges by helping them develop a powerful and accurate predictive model that can estimate house prices in Nigeria and enable them to make informed pricing decisions, enhance its competitiveness in the market, and deliver enhanced value to its customers.

DATASET AND APPROACH

The datasets provided on Zindi was used in the project. the train dataset (Housing_dataset_train.csv) contained 14000entries and the test dataset (Housing_dataset_test.csv) contained 6000 entries. The train dataset contained the following columns:

1. ID
2. loc
3. title
4. bedroom
5. bathroom
6. parking_space
7. price

DATA CLEANING

I decided to remove null entries from the dataset in columns 'loc' and 'ID' and filled in the null values of the other numerical columns with the median value for the group they were subset under using both columns 'loc' and 'title'. At the end of the data cleaning process, the dataset was reduced to 10531 entries.

EDA

I created several plots to analyse the relationship between columns in the dataset especially how they related to the 'title' column and to the target column, 'price'. I was able to come to the conclusion that the 'ID', 'loc', 'title', and 'bedroom' columns were very vital to the model while the 'parking_space' and 'bathroom' columns were not as vital. So I proceeded to creating more relevant columns in the next step

DATA PROCESSING, FEATURE SELECTION AND FEATURE CONVERTING

I decieded to create more columns to further enhance the model and these columns include one containing a ratio between the bedroom and bathroom columns, another columne containing the total

number of tome (i.e sum of bathroom and bedroom), in another column, I proceeded to map values of the ‘loc’ column (states) into a superset of zones and then one_hot encode the column. Furthermore, I mapped the values of the ‘title’ column into a ranking from smallest to biggest based on size, the ranking is based off answers from bard. I then proceeded to normalize the values of the loc column based on their value counts

Observation: the 0ne-hot encoding of the zones really made the model much better(result not reflected in zindi submission as that was closed yesterday and I still readjusted my code cause of the confirmed extension)

MODELING

I tested 2 models on the dataset, CatBoostRegressor and LGBMRegressor. I used 75% of the data as training data, and 20% of the data as testing data

After running both models, I calculated the RMSE to tell how off the predicted prices are from the true prices. Having similar results, I then used an8-fold cross validation on a shuffled version of the data to run the LGBMRegressor model

SUBMISSION

I ran the dataset to be tested through the model and created a new column in the submission sample file then saved.