HOUSING PRICE PREDICTION

A project report submitted in

partial fulfilment of the requirement for the award of internship provided by

FlipRobo Technologies, Bangalore

Submitted by:

ELTON GRIVITH D SOUZA

engardaelton.work@gmail.com

OCTOBER 2021

ACKNOWLEDGMENT

I would like to offer my sincere thanks to **FlipRobo Technologies** for providing the opportunity to intern under their company.

I would like to thank **Ms Khusboo Garg** for coordinating and providing valuable suggestions during the process of this project and internship.

Finally, I would like to express my sincere gratitude to my family and friends who have been the embodiment of love and support which helped us carry out this project in a smooth and successful way.

INTRODUCTION

Problem Statement

Houses are one of the necessary needs of each and every person around the globe and therefore housing and real estate market is one of the markets which is one of the major contributors in the world's economy. This domain has a very large market and there are various companies working in the domain.

Data science comes as a very important tool to solve problems in the domain to help the companies increase their overall revenue, profits, improving their marketing strategies and focusing on changing trends in house sales and purchases.

Predictive modelling, market mix modelling, recommendation systems are some of the machine learning techniques used for achieving the business goals for housing companies. Our problem is related to one such housing company.

A US-based housing company named **Surprise Housing** has decided to enter the Australian market. The company uses data analytics to purchase houses at a price below their actual values and flip them at a higher price.

The company is looking at prospective properties to buy houses to enter the market. This projects requirement is to build a model using Machine Learning in order to predict the actual value of the prospective properties and decide whether to invest in them or not.

Business Goal

The requirement of this project was to model the price of houses with the available independent variables.

This model would then be used by the management to understand how exactly the prices vary with the variables.

They would then be able to manipulate their strategy to concentrate on areas that would lead to high returns. Furthermore, the model will be a good way for the management to understand the pricing dynamics of a new market.

Analytical Problem Framing

Data Sources and their formats

The data source was provided by **FlipRobo Technologies** stating the source of the data is from a US-based housing company named **Surprise Housing**. The format as provided by the company was a CSV file containing more than 1600 rows with 81 features

Data Pre-processing

The provided data consisted of a lot of missing values. This was treated by imputing max occurring value for categorical features and mean value for continuous data.

The given data consisted of noise where the scraped values consisted of html methods. These tuples were detected and dropped accordingly

Outliers were detected and dropped using the Z – Score method.

Skew was treated by using scikits' power transform method. yeo-johnson algorithm used for the treatment. This process was followed by normalisation to make the distribution more Gaussian.

Data Relationships

The relationships analysed from the data was done through exploring the correlation, distribution of feature values, distribution of feature sets, distribution of feature values in regard to the target feature etc.

The resulting relationships explored the features that have high correlation with the target label. This was done by using single, dual and multi feature exploration.

Hardware and Software Requirements and Tools Used

The development of this project required the following tools:

- Python 3.8
- CUDA 10
- Scikit-learn 1.0
- XGBoost
- LightGBM

Model Development and Evaluation

Model Development

The model was trained on a 25% split of the training data which was fed to a pipeline of different models. These models are listed below:

- ElasticNet
- Logistic Regression
- Decision Tree Regressor
- Random Forest Regressor
- AdaBoost Regressor
- Support Vector Machine
- LGBM Regressor

The model was scored using root mean squared error (RMSE). And was filtered based on its CV score that took neg-RMSE as its scorer.

Hyper Parameter Tuning *(HPT)*

Parameter tuning was done using GridSearchCV on Logistic Regression. This model was chosen based on its CV scores. The grid parameters that were used are given below:

- **penalty**: l1, l2, elasticnet, none
- **solver**: newton-cg, lbfgs, liblinear, sag, saga
- **max_iter**: 12 randomised numbers between 100 and 1200
- **multi_class**: auto, ovr, multinomial

The best parameters obtained from this process was then used to provide the submission model.

CONCLUSION

The submitted model provides consistent results that can be validated under other circumstances. The result obtained from the provided data can be assumed competitive and can provide better insights to companies looking forward to the real estate business. The provided solution satisfies all the requirements stated by the company and hence this project can be considered successful.

Future Work could include cleaner localised data with more feature parameters to provide better estimations and analytical relationships.