HOUSE PRICE PREDICTION

Submitted by:

Shweta Rajani

ACKNOWLEDGMENT

I would like to thank all my mentors of Data Trained, who taught me the concepts of Data Analysis, building a machine learning model, and tuning the parameters for best outcomes.

For this particular task, I referred the following websites and articles when stuck:

- https://towardsdatascience.com/a-common-mistake-to-avoidwhen-encoding-ordinal-features-79e402796ab4

- https://stackoverflow.com/questions/43590489/gridsearchcvrandom-forest-regressor-tuning-best-params

- https://www.codegrepper.com/codeexamples/delphi/scikit+pca+preserve+column+names+pca+pipeline

- https://stackoverflow.com/questions/22984335/recoveringfeatures-names-of-explained-variance-ratio-in-pca-with-sklearn

I would also like to thank my mentor in Fliprobo, Sapna Verma, for providing me the dataset and problem statement for performing this wonderful task.

INTRODUCTION

Business Problem Framing

The objective of the model is to determine the price of houses with the available independent variables. This model can be used by the management to understand the prices variation with the available features. They can accordingly manipulate the strategy of the firm and concentrate on areas that will yield high returns. Further, the model help the management to understand the pricing dynamics of a new market.

Conceptual Background of the Domain Problem

Houses are very necessary for each and every person around the globe and therefore housing and real estate market is one of the markets which has the major contributors in the world's economy. It is a very large market and various companies are working in this domain. Data science appear as one of the important tools 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 which are used for achieving the business goals for housing companies. Our problem statement 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. For the same purpose, the company has collected a data set from the sale

of houses in Australia. The data is provided in the CSV file below. The company is looking at prospective properties to buy houses to enter the market. I was required 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. For this company wants to know:

- Which variables are important to predict the price of variable?
- How do these variables describe the price of the house?

Technical Requirements:

- Data has 1460 entries each having 81 variables.
- Data has Null values which is needed to be treated using the domain knowledge and own understanding.
- Extensive EDA has to be done to gain relationships of important variable and price.
- Data has both numerical as well as categorical variable which is handled accordingly.
- Finally, we have to build Machine Learning models, apply regularization and determine the optimal values of Hyper Parameters.
- Need to find important input variables which affect the price positively or negatively.

Data Sources and their formats

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. For the same purpose, the company has collected a data set from the sale of houses in Australia. The data is provided in the CSV file.

The last Feature: Sale Price is the target variable. The above Snapshots show all the features and the top 10 rows. As mentioned earlier, there are 1460 rows and 81 columns.

CONCLUSION

- Key Findings and Conclusions of the Study:

- MS Sub Class seems to have the biggest impact on House Prices, followed by Basement Full Bath and Basement Half Bath

- Other than the Basement related features, Condition 2, Exterior Quality and Lot Area are some of the other important features.

• Learning Outcomes of the Study in respect of Data Science

- Got to understand about the concept of Data Leakage. All transformation must be done after splitting the data to test and train, otherwise the parameters are affected.
- Used RFE for the first time. It is a great technique for Feature Selection.
- Learned about the usage of Lasso and Ridge Regression.

• Limitations of this work and Scope for Future Work

The `biggest limitation I observed was that not all categories of a particular feature were available in the training data. So, if there is a new category in the test data/new data, the model would not be able to identify the new categories.

Example: All 8 categories in MS Zoning are:

```
MSZoning: Identifies the general zoning classification of the sale.

       A      Agriculture
       C      Commercial
       FV     Floating Village Residential
       I      Industrial
       RH     Residential High Density
       RL     Residential Low Density
       RP     Residential Low Density Park
       RM     Residential Medium Density
```

However, in the dataset, MS Zoning only has 5 categories available. So, if the other 3 categories are present in the test set, it would become difficult for the machine to identify