FLIP ROBO

HOUSING PROJECT

Submitted by:

RAJ VAGHASIA

INTRODUCTION

- ## Business Problem

  Houses are one of the necessary need 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. It is 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

- ## Conceptual Background of the Domain Problem

  An intermediate level of real estate combined with civil engineering expertise would be beneficial to understand the data more precisely. Basically the cost of material in the local market would enhance a data analyst to understand and design the predictive model accordingly.

- ## Motivation for the Problem Undertaken

  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 company is looking at prospective properties to buy houses to enter the market who is looking to build a Machine Learning in order to predict the actual values of prospective properties and decide whether to invest 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?

Analytical Problem Framing

- Mathematical/ Analytical Modeling of the Problem

For the statistical purpose, the chi-squared test can be performed to decide on the hypothesis which can be established to check whether there is dependencies between identical attributes. For that matter, correlation matrix is plotted and the features resulting in higher correlation can be taken into account hypothesis testing whether there is significance to affect the prediction by model.

After completing statistical part, analysis says that there is high amount of correlation between the following feature:

1. GarageArea & GarageCars. Since the garage cars does not have any deciding factor to affect the Saleprice of the house, we can remove the feature.
2. Among the high correlation were the attributes like Exterior1st – Exterior2nd, ExterQual – MasVnrType, BsmtQual – BsmtExposure which had similar distribution of data and hence we can normalize them using log transformation to verify if there's any normalization happening.
3. After analysing the existing data and relation of various feature with target variable, we can fill the null values sighting other similar feature data and fill nan values accordingly.

- Data Preprocessing

What were the steps followed for the cleaning of the data? What were the assumptions done and what were the next actions steps over that?

The raw data included 81 feature and 1460 records. Therefore, the deliberate processing of data is carried to achieve maximum model performance.

The steps for preprocessing included:

(Not in particular order)

1. Separating numerical and categorical feature
2. Getting the idea of null values in each feature
3. Dropping feature who has more than 70% missing values.
4. Checking the distribution of the data and if the bias is high towards particular values, dropping that particular feature since model will be not generalized.
5. Analysing the outliers through boxplot
6. Filling the missing values.

- State the set of assumptions (if any) related to the problem under consideration

The assumption taken are only legit and considered taking care that it should not affect the model performance largely.

Model/s Development and Evaluation

- Testing of Identified Approaches (Algorithms)

Listing down all the algorithms used for the training and testing.

Since the dependant variable is continuous feature, we have used Regression algorithms:

1. Base model: Linear Regression
2. Lasso and Ridge Regressor with hyperparameter tuning
3. Decision Tree Regressor
4. Random Forest Regressor
5. XGBoost Regressor

- Key Metrics for success in solving problem under consideration

  The metrics used for performance evaluation is R^2 score and RMSE score along with 5-fold cross validation. To achieve better performance, R^2 score should be higher and RMSE score should be least since it indicates the error in data points and distance to best fit line.

CONCLUSION

From the model evaluation, we can set XGBoost Regressor as the base model to predict the unseen data. But due to limited knowledge of the domain, more progress can be achieved. Further, comprehensive preprocessing can be done like standard scaling, filling null values taking the reference of other feature inputs.

Hence, there is always a scope of improvement in terms of what Data Scientist can do.