FLIP ROBO

Housing Price Prediction

Submitted by:

ASHISH KUMAR DEWANGAN

INTRODUCTION

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. 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. 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.

Objective

The company is looking at prospective properties to buy houses to enter the market. You are 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?

There is a need to model the price of houses with the available independent variables. This model will then be used by the management to understand how exactly the prices vary with the variables. They can accordingly manipulate the strategy of the firm and concentrate on areas that will yield high returns. Further, the model will be a good way for the management to understand the pricing dynamics of a new market.

Exploratory Data Analysis

- Data Pre-processing and Visualizations

The training dataset has 1168 entries and test set has 292 where each entry has 80 features and 1 label. 3 features have float values, 35 features have integer values and 43 features are categorical in nature. The feature `'Id'` is serial number.

The inference from the analysis of missing values are as follows:

1. Lot Frontage: It 214 missing values. We will use Iterative Imputation technique to treat this feature.

2. Alley-1091: More than 93% missing values. Hence, it is safe to drop.
3. BsmtQual (30 NaN), BsmtCond (30 NaN), BsmtExposure (31 NaN), BsmtFinType1 (30 NaN), BsmtFinType2 (31 NaN): We will replace the missing values with one of the category 'NA
4. FireplaceQu: It has 551 missing entries. We will fill the missing values with no fireplace.
5. GarageType (64 NaN), GarageYrBlt (64 NaN), GarageFinish (64 NaN), GarageQual (64 NaN), GarageCond (64 NaN): We will use Iterative Imputation technique to treat these features.
6. PoolQC: It has 1161 missing values. We will fill the missing values with no pool.
7. Fence: It has 931 missing entries. We will fill the missing values with no fence.
8. MiscFeature: It has 1124 missing values. We will fill the missing values with no feature.

Univariate Analysis

The following conclusions can be derived from the strip plot of categorical features:

1. The decreasing order of price based on various zone is:
   1-Residential Low Density
   2-Floating Village Residential
   3-Residential High Density
   4-Residential Medium Density
   5-Commercial
2. Only 4 out of 1168 properties have paved roads.
3. The minimum price of 'Moderately Irregular' and 'Irregular' shaped properties is higher than 'Regular' and 'Slightly Irregular' shaped.
4. The minimum price of 'Hillside - Significant slope from side to side' properties is higher than other land contours.
5. The minimum price of 'Inside Lot' is least among all Lot Configuration.
6. The minimum price of properties which are in close proximity to positive offsite feature like park, green belt, etc. or Within 200' of East-West Railroad is very high as compared to other properties.
7. The minimum price of properties whose dwelling type is Single-family Detached is the least.
8. The minimum price of properties whose dwelling style is '2.5 Story' and 'Split Level' is very high as compared to '1 Story', '1.5 Story', '2 Story' or 'Split Foyer'.
9. The minimum price of properties based on roof type in decreasing order is:
   1-Shed
   2-Mansard
   3-Flat
   4-Hip
   5-Gable
   6-Gabrel (Barn)

10. The highest price of Flat, Shed, Gabrel(Barn) and Mansard is very less as compared to Gable and Hip.
11. Majority of the properties are made up of Standard (Composite) Shingle.
12. The minimum price of the properties whose roof are made up of Wood Shakes is highest.
13. The minimum price of properties are in the lower range when the exterior covering is made up of Wood Siding, Vinyl Siding, Brick Face or Asbestos Shingles.
14. The minimum price is very high if the masonry veneer is of Stone.
15. The minimum price is high if the foundation is made up of Wood, Stone or Poured Concrete.
16. Majority of the properties have Gas forced warm air furnace.

Data Pre-Processing

The boxplot of all the features showed that there are lot of outliers in the dataset. To eradicate outliers the method of Z score was used. After outlier removal we were left with 1077 entries in the dataset.

The outlier treatment handled the skewness to some extent which was verified by visualization the distribution of cleaned dataset.

There are 80 features for every entry hence, there are higher chances that the feature will be correlated to each other. Hence, feature selection can prove to be a deciding step to develop an accurate model. We can visualize the correlation matrix with the help of heat map to identify the dependent features.

The features which have high degree of correlation are - 'Exterior2nd', 'GarageCars', and 'PoolArea'. Hence, we can drop these features. After this step we have 78 features which play a significant role in identifying the defaulters.

We have the information about construction year and renovation year. From these two features we are introducing two mew features

1. 'Age' – It is the age of the property.
2. 'Effective Age' – It is the age of property after renovation.

Now we need encode the data before fitting into model. The features which have ordered values are encoded with the help of Ordinal Encoder while others were encoded using Lable Encoder.

After data pre-processing labels and features were separated, the features were split into train and test dataset in the ratio of 3:1.

The model which were tested in this data are-

1. Logistic Regression
2. Lasso
3. Ridge

4. KNN Regressor
5. Decision Tree Regressor
6. Random Forest Regressor
7. Ada Boost Regressor
8. Gradient Boost Regressor
9. XG Boost Regressor

The parameters of the model were tuned were ever applicable to improve model accuracy and precision. The parameters precision, recall and F1 score were also compared.

The model performance is tabulated below for reference:

S. No.	Model Name	MSE
1.	Linear Regression	23975
2.	Lasso	23971
3.	Ridge	23944
4.	KNN	37154
5.	Decision Tree	35772
6.	Random Forest	25153
7.	Ada Boost	30270
8.	Gradient Boost	21713
9.	XG Boost	23723

The accuracy scores of model after hyper parameter tuning are:

S. No.	Model Name	Score
1.	Decision Tree	0.76
2.	XG Boost	0.89
3.	Random Forest	0.86
4.	Gradient Boost	0.88

The performance of XG Boost Regressor is better than any other model. Hence it is selected as the final model for prediction of property price.

The prediction was performed on the test data using the final model.

Conclusion

We cleaned the data, performed EDA and successfully trained a model and tuned the hyper parameters to predict the price of properties based on the features and dataset available with us. The XG Boost Regressor proved to be the best model in prediction due to its high

- **Score – 0.89**
- **MAE** – 14390

- **RMSE** – 21176
- **Hyper Parameters – {max_depth=5, learning_rate=0.1, min_child_weight=7, gamma=0.4, colsample_bytree=0.3}**

The feature which are key factors in deciding the price are:

1. OverallQual
2. MiscVal
3. ExterQual
4. GrLivArea
5. GarageArea
6. KitchenQual
7. GarageFinish
8. FullBath
9. BsmtQual
10. Age
11. TotalBsmtSF
12. GarageYrBlt
13. 1stFlrSF
14. EffectiveAge
15. MSZoning
16. Heating
17. Street
18. BsmtFinSF1
19. MasVnrArea
20. GarageType
21. Foundation
22. Fireplaces
23. CentralAir
24. TotRmsAbvGrd
25. OpenPorchSF