
# HOUSE PRICE PREDICTION

Submitted by:

ARPITA MISHRA

# Introduction

Housing and real estate market is one of the major contributors in the world's economy. With the help of Data Science it has become easier to help the companies increase their overall revenue, profits, improving their marketing strategies and focusing on changing trends in house sales and purchases. The present project is associated with a US-based housing company named Surprise Housing which is planning 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.

We are trying 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. Thus, with the present model we will predict 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.

# DataSet Information

Two datasets were given to us as train and test data in csv format. Train data set was used to train the model while test data set was used for house price prediction with the model built. Train data was found to have 81 columns and 1168 rows while test data set had 80 columns and 292 rows. The dataset had both categorical as well as numerical columns.

## Data Preprocessing

Both the train and test data were analysed for null values. There were five features (PoolQC, MiscFeature, Alley, Fence, FireplaceQu) with more than 45-50 percent null values. These features were dropped from both the data set. Features with lesser null values were treated accordingly. Null values in the categorical columns were replaced by mode of the respective column. Null values in features with continuous values were replaced by median of that column.

# Data Visualization

Data distribution of all the numerical features were observed by distplot method of seaborn library. Data in most of the numerical features were found to be skewed. Boxplot method was used to check for the outliers. Extreme outliers were present in most of the columns. These outliers were removed by setting an upper limit manually and dropping the values greater than that limit. Log transformation method was used to treat skewness in all the numerical features. Correlation of all numerical feature with the target ie house price were studied by heat map method. Graphical analysis of the numerical data showed the following relationship

- House price was found to be positively correlated to the Overall Quality,Total area of house in square feet,
- Number of bathrooms, Garage Area, Ground floor living area, Total rooms above ground and woodDeck area also affected price.

Categorical columns were observed by count plot method and how these features affect the house price were analysed by plotting bar graph with groupby method. It was observed that-

- Houses with attached garage had higher price as compared to detached one.
- 1Family houses had more price as compared to duplex
- Houses which had the concrete foundation had the highest price while the one with wooden foundation had the least price.
- Houses with Gable type of roof were more expensive.
- Vinyl sided exteior has higher price.

# Data preprocessing

Categorical features were transformed to numerical features. Since there were features with ordinal as well as nominal data. Features like 'ExterQual', 'ExterCond','BsmtQual' etc which had ordinal data were transformed by ordinal encoding while that of nominal data like neighbourhood, lot shape etc were transformed by One hot encoding. All the features were standardised by Standard Scaler.

One hot encoding led to an increase in features from 80 to 210 in train dataset and to 186 in test data set.

Higher number of features and difference in the number of features which needed to be addressed before model building.

Therefore, PCA was used for reducing the number of features in train data set. PCA led to significant decrease in features from 210 to 115 features. Now these features were aligned with the test data set with 'inner join' method which led to the retention of only those features that were present in train dataset. Thus, we finally had only those features in train and test dataset which were common.

# Algorithms used

The preprocessed train data was split into train and test data. While the actual test data was kept separately for prediction by the model built.

Since it was a regression problem, we tried to build models on Linear Regression. We also used Random Forest Regressor(RFR) and tried to find a better model. The parameter chosen was R2 score.

R2 score of Linear Regressor was found to be 85 percent and that of RFR was around 86 percent.

Crossvalidation was also performed and the score was 76 percent and 86 percent respectively.

The difference in r2 score and CV score was lesser in RFR so we proceeded with hyperparameter tuning of RFR. Randomized searchCV was used for hyperparameter tuning. Final model was built with best parameter. Model was saved by using joblib.

Now the final model was used to predict house price of the test data.

# Summary and Conclusion

- The given data were first preprocessed/cleaned. Null values were found and replaced with suitable data.
- Graphical representation of all the features were done by Seaborn library.
- Categorical features were transformed to numerical ones.
- Standardisation of data was done before model building
- PCA was performed to reduce the number of features
- Linear Regressor and Random Forest Regressor was tried and compared for maximum R2 score
- RFR was finally used as final model after hyperparameter tuning. R2 score was found to be 87 percent.
- Final model was used to predict the house price of test data set.