

Housing Price Prediction Analysis

By Using

Machine Learning.

Submitted by:

Akchay Pradhan

**ACKNOWLEDGMENT**

I would like to sincerely thank our SME Mr. Shubham Yadav for providing the Dataset and helping us to solve the problem by giving help as to when was needed.

**Introduction:**

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. This is the Regression problem so we have to build a model which predict the price of the house.

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. 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.

**Analytical Problem Framing:**

**Dataset:**

The dataset consists of **1460** individual’s data. There are **81** columns in the dataset, however the first ID is not so important to us and the Dataset Also have a lot of Null Values.

**Problem Definition:**

We have to build a Regression model which can be used to predict in terms of a Price of the houses based on different features, here we are going to find out the prices of the Houses and basis of different features. We need to drop duplicates rows if present in dataset.

We have to check for the null values present in our dataset. If null values are present then fill it via mean, median or mode. Or also you can remove that rows but kindly check it properly.

After that we check the summary statistics of our dataset. This part talks about the statistics of our dataset i.e., mean, median, max value, min values and also it tells whether outliers are present in our dataset or not.

**Data Analysis:**

The First Step is to check the shape and size of the dataset, second step is to analysis the numerical and categorical columns.

After completing the first step next step is to check the null values in the dataset and normalizing them, it’s very important to remove the null values from Dataset .in our case their lot of Null value present in Dataset.

Third step is of feature Engineering during this step we have to remove all duplicate Records, also check for categorical and numerical columns.

**EDA Analysis:**

EDA Analysis is categorized into Univariate Analysis, Bivariate Analysis, and Multivariate Analysis, here we check the relationship between different independent variable and dependent Variable.

* There’s lot of null values in Dataset
* We drop the columns having more than 90 percent of missing values.
* We also check for corelation and Multicollinearity
* We cannot remove the outliers because more than 50 percent of information is lost
* We remove the skewness from features.
* We have also drawn Q-Q Plot to check the distribution of the datapoints

**Correlation Analysis:**

Here we check how the different independent features are corelated with each other and their strength of Relationship, weather they are positively correlated or negatively corelated. The value of Correlation ranges from -1 to +1. -1 indicate the negative correlation with increase in one independent variable the dependent variable decreases +1 indicates the positive correlation means with increase in independent variable dependent variable also increase. In our case some of the independent feature are highly corelated with each other with a person corelation strength more than 0.9 hence we have to dropped all the features which are highly multicollinear to avoid multicollinearity.

**Outliers Analysis:**

Outliers are the data points that differ significantly from other observations, any data points greater than +3 and -3 standard deviations are called as outliers. Z-score is the Automated method used for handling outliers, it’s important to remove outliers as it impacts on the Accuracy of the Model. In our case each and every information is important to us we can’t afford to lose 7% of information but during the outlier’s analysis we found that we are losing 23% of Information hence we have decided not to remove outliers.

**Checking for Skewness:**

Checking for skewness, skewed data are not normally distributed either they are positive skewed or negative skewed, if the data is skewed its seriously impact on accuracy of the model so it’s very important to remove the skewness for right and left skewed data boxcox and logarithm transformation will work fine. for graphically visualization we use QQ Plot to check the distribution of data points. Any value greater than 0.55 is considered to be skewed data,

In our case most of the data are skewed hence we have to remove the skewness during Scaling because if we remove the skewness by log or boxcox method it will induce nan values.

**Standardization:**

Standard scaler is used to bring the data points to standard Normal Distribution having mean = 0 and SD +- 1 to enhance accuracy of the model, In our Case its most important due to presence of high skewed data and Outliers.

**Splitting:**

First, data is divided into two parts using component splitting. In this experiment, data is split based on a ratio of 80:20 for the training set and the prediction set. The training set data is used in the logistic regression component for model training, while the prediction set data is used in the prediction component.

**Regression Models:**

The following regression models are used - Linear Regression, Random Forest regression, Lasso regularization and Ridge regularization.

**Cross Validation:**

It’s very important to perform cross validation because we pick up the training and testing data randomly there's no structure of it, so each and every time we refresh the random state, we can see that the Accuracy of the model changes. so, it’s very important to check our scores by using cross validations, for better and optimal result. There are two types of cross validation. Cross-validation also gives us an idea whether a model is suffering from underfitting and overfitting conditions,

In our Case we are using a Classification Model hence we will go for Stratified K fold cross Validation to make sure that no of instances of each output-class is equally distributed in training and testing phase.

1.Kfold Cross Validation

2.Startified K fold cross Validation.

**Hyperparameter Tuning:**

It’s a process in Machine Learning through which we obtain the best and optimal value for our dataset, hyperparameter is a process of searching an ideal model Architecture. It’s very important to perform Hyperparameter tuning for obtaining optimal result.

1.Grid Search CV used in all Regression and Classification Problem.

2.Randomized Search CV for Random Forest

In our Case we have used Both Grid Search and Random Search.

**Predictions:**

* We know that this is Regression problem so we use R\_2 score, MAE, MSE and RMSE as our evaluation matrix. We also see the cross validated score.
* There’s lot of Categorical features present in Dataset so we have to convert them to numerical features
* We also applied Min Max Scaler for standardization.
* We also used hyperparameter tuning to find out best value.

**Evaluations:**

In our Case our Dataset is have lot of Categorical features we need to convert them to numerical before implementation.

* Linear Regression gave us the score R2 Score of 87% mean R2 Score 87%, Standard deviation of 0.0317, MAE Score of 0.107, MSE Score of 0.023, RMSE Score of 0.153.
* Random Forest Regression gave us the score R2 Score of 84% mean R2 Score 86%, Standard deviation of 0.023 after performing hyperparameter tuning.
* Lasso Regularisation gave us the score mean R2 Score 87%, Standard deviation of 0.029, MAE Score of 0.107, MSE Score of 0.023, RMSE Score of 0.153 after hyperparameter tuning.
* RMSE: Root Mean Square Error
* MSE: Mean Squared Error
* MAE: Mean Squared Absolute Error

**Software and Library used:**

* Python
* Pandas
* NumPy
* Matplotlib
* Seaborn
* Skit-Learn

**Conclusions:**

* In this project the sample data is provided to us from our client database.
* We are required to 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
* We make a machine learning model in order to improve the house price prediction according to the given features.

.