Project Report
on
Housing Price Prediction

BY:

MAYANK SHUKLA

Background & Introduction

We are about to deploy an ML model for Housing price prediction and analysis. This kind of system becomes handy for many people.

Imagine a situation where you have an house and want to sell it. You may of course approach an agent for this and find the market price, but later may have to pay a lot of money for his repairing work in selling your house. But what if you can know your house selling price without the intervention of an agent. Or if you are an agent, definitely this will make your work easier. Yes, this system has already learned about previous selling prices over years of various Houses.

So, to be clear, this deployed web application will provide you will the approximate selling price for your house based on the plot area, number of rooms and balconies, parking space, garden area, number of floors and etc are the foucs points are there.

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. 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

Approach to Data Cleaning

* First we check the information of the given dataset because it tells that how many rows and columns are present in our dataset and data type of the columns whether they are object, integer or float.

- Drop duplicates rows if present in dataset.+Then we 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 tells about the statistics of our dataset i.e. mean, median, max value ,min values and also it tell whether outliers are present in our dataset or not.

Approach to Data Cleaning (Cntd…)

* We also check the correlation of our dataset to check the correlation of the columns with each other. If columns are highly correlated with each other let's say 90% or above then remove those columns to avoid multi coli-nearity problem.

* We extract data from date column and make new columns like day, month and year to see the outcomes with our target column that is label.

* We delete the pcircle column because it has only one unique value that tells that collected data is only for one circle.

* We cannot remove outliers because more than 20% of our data are removed.

Visualization.

* We plot correlation matrix via heatmap to see the correlation of the columns with other columns.

* We also visualize the correlation of columns with target column via bar graph to see which column is highly correlated with target column.

* We see the number of defaulter and non defaulter customers with the help of count plot.

* We plot histogram to displays the shape and spread of continuous sample data.

* We also see the customers labels i.e defaluter/Non-defaulter according to date and month with count plot.

* We also see the distribution of the data with the help of distribution plot whether it is left skewed or right skewed.

Modelling part

* We know that this is classification problem so we use accuracy score, classification report and confusion matrix as our evaluation matrix. We also see the AUC score and also plot the AUC_ROC curve for our final model.

* As we know this dataset is imbalance so we don’t too much focus on accuracy score . We see the precision and recall value along with f1_score.

* First we see the result without doing any sampling technique and for that I use Ridge classification model

* At the end we use Lasso Regressionm which is a popular type of regularized linear regression that includes an L1 penalty

Conclusion

- So here 'Lasso Model' is the best model out of all model tested above and Number of predictors selected by double the optimal alpha for lasso are:129