



HOUSING PRICE PROJECT

Submitted by:
PUROBI DEBNATH

ACKNOWLEDGMENT

I like to thank Flip Robo Technologies for giving projects like this and thanks to Sajid Choudhary who thought that we will do this project. This project helped me understanding more about data analysis and got to improve my skills. I have taken lot of efforts to complete this project, but without help of internet and our team leader could not have been able to complete this project.

INTRODUCTION

- Business Problem Framing

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.

- Conceptual Background of the Domain Problem

Earlier advancement of data science real estate companies has huge risk of getting into real estate market without proper knowledge. Now with the help of previous data which provides to data scientist I can perform and can give proper suggestion to client and it reduce the risk of loss in the real estate market.

- Review of Literature

The data consists 1168 rows and 81 columns. Housing are one of the necessary need of every person around the globe and therefore housing and real estate market is the major contributor in the world economy. The company uses data analysis to purchase house below their actual values and flip them at a higher price.

- Motivation for the Problem Undertaken

Every projects begins with idea that are further developed and inspired variety of situation and circumstances. The client want some analyzation on data to improve the real estate housing price. So to help them we have made this project. My motivation behind this project is to upgrade my skills and to learn new things from internet by doing this project. So after doing all the necessary data analysis and programming to give a perfect solution to the client.

Analytical Problem Framing

- Mathematical/ Analytical Modeling of the Problem

After seeing the null values from the data given, I have seen the percentage of the null data and I found that PoolQC, MiscFeature, Alley, Fence has no variance .so maximum percent of data is missing .hence I dropped these columns. After the remaining data I filled with NA, mode and median. By applying this the data is clean and ready for analysing.

- Data Sources and their formats

The sample data of this project which provided to us from our client database. The dataset is in csv file.

- Data Inputs- Logic- Output Relationships

Heating', 'SaleCondition', 'CentralAir', 'LowQualFinSF', 'KitchenAbvGr', 'Functional', 'SaleType', 'MoSold', 'MiscVal', 'PoolArea, ScreenPorch', '3SsnPorch', 'PavedDrive', 'EnclosedPorch', 'BsmtFinSF2', 'Condition2', 'BldgType', 'RoofStyle', 'RoofMatl, ExterCond', 'Utilities', 'Electrical', 'BsmtCond', 'BsmtFinType2', 'GarageQual', 'GarageCond', contains maximum percentage of variance so I have dropped the columns which were effecting our model and output.

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

Since the data were expensive I cannot lose more data that is why I filled the nan values with mode, mean ,median and NA. The data set has more categorical values so I created dummy values and merged it with our data set.

- Hardware and Software Requirements and Tools Used

Processor-Intel Core i3 -5005U CPU @ 2.00ghz

Installed Ram-8Gb

System type-64 bit OS

Software-Windows 10 Home Single Language

Model/s Development and Evaluation

- Testing of Identified Approaches (Algorithms)

The algorithms and libraries that I have used are:

1]Pandas

2]Numpy

3]Seaborn

4]Matplotlib

5]Scalar

6]Train test split

7]Linear regression

8]Lasso

9]Ridge

10]Grid search cv.

- Run and Evaluate selected models

Since it was supervised machine learning ,after evaluating performance of the model I performed with regularization algorithms and grid search cv. The model was performing well in both lasso and ridge regression. I found that ridge was performing better and I train and test the model in ridge regression.

- Interpretation of the Results

1]BsmtQual_Gd

2]KitchenQual_Gd

3]MSSubClass

4]KitchenQual_TA

5]ExterQual_TA.