Housing Sale Price Prediction

Submitted by:

Rohan V Borade

ACKNOWLEDGEMENT

I'm extremely grateful to Flip Robo Company for giving this wonderful opportunity to work on first industry-based project. Also, learned some new skills which are added to my knowledge.

I would also like to thank our SME Shubham Yadav,who has giving us needful instructions on project. By providing dataset, it's detailed explanation in the form of text file and demonstration of the report writing with docx enlisting the subpoints with help we should write our report.

INTRODUCTION

- Business Problem Framing

The Surprise Housing is US based housing company has decided to enter the Australian market. The company's main motive is to buy the house at reasonable and flip them at higher price.

This is relatable to real world because the many aspects we or every one who makes decision for buying new house.

Some factors while buying house are central location for Travelling, School, Supermarkets, colleges which are Important key factors to keep in mind. Above factor are relatable to the real world.

- Conceptual Background of the Domain Problem.

  There 80 independent features in the dataset which are considered in real life for buying a house.

  Every feature column is considered by a person who searching for that type of specific house.

  For example:

If a person wants to buy a DUPLEX - ALL STYLES AND AGES, the zoning of the such type of house will come under Commercial of classification of sale which also considered the major factor to sale price of the house because the properties are in commercial zone which automatically raise the sale price for the company.

Condition is applied like if the property is old basically, it will need renovation. Then the sale price might increase.

- Review of Literature

This is a comprehensive summary of the research done on the topic. The review should enumerate, describe, summarize, evaluate and clarify the research done.

House is one of the important elements in basic human needs. People need a house to stay away from danger, hot weather, rainy day and as well as a place to stay calm. As long as people can fill the comfort of living under a roof then it is called a house.

People who are called the rich and famous can afford a house that is almost called a heaven and some can only live in an ordinary but comfortable house.

We can see that the housing price is increasing as the time goes by. This may be an important area to look upon because more or less it could affect the economic level of a country.

Therefore, a housing price can be defined as the rate of payment that one has to pay in order to purchase a house and for sure there are several factors that lead to housing price determination.

In my own point of view, I believe that the increment of a housing price is due to the price increment in the raw material.

The income and interest rate can be considered has the important factors.

How much an individual can borrow from the bank and financial institutions can also be the one away to determine the price of the

housing. For example: Higher the interest rate, the higher the price of housing.

Important factor affecting the housing prices is the size of the floor area. The second and third most effective factors are the road surface ratio and the floor area, and fourth factor is the distance of housing to the seashore. This means, the bigger the size of floor the more money a purchaser will have to pay.

- Motivation for the Problem Undertaken

Describe your objective behind to make this project, this domain and what is the motivation behind.

My main objective to make this project to learning and develop the skills required to deal with real-world problems.

Also, to apply skills which are base foundation of the machine Learning and dive deep. work on the industry base problems.

Analytical Problem Framing

In statistics, linear regression is a linear approach for modelling the relationship between a scalar response and one or more explanatory variables (also known as dependent and independent variables). The case of one explanatory variable is called simple linear regression.

In linear regression, the relationships are modelled using linear predictor functions whose unknown model parameters are estimated from the data. Such model is called linear model

Linear regression has many practical uses. Most applications fall into one of the following two broad categories

If the goal is prediction, forecasting, or error reduction, linear regression can be used to fit a predictive model to an observed data set of values of the response and explanatory variables.

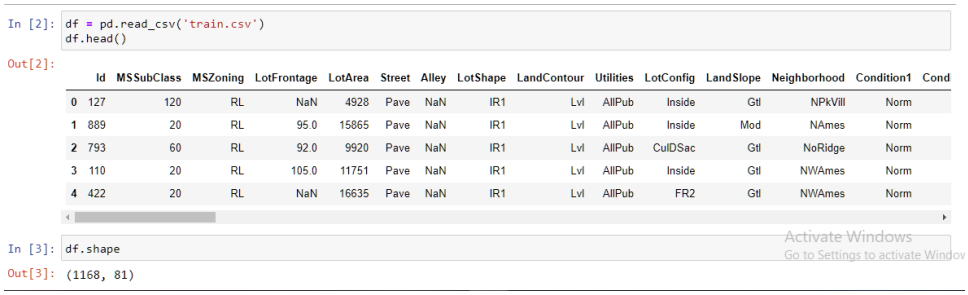
- Data Sources and their formats

  The data sources are from Australia dataset consist of various other independent features which are considered while sales of the house.

  The format of the dataset its CSV (which comma separated value), it also contains numerical as well categorical variable.

  The train data consist of 1168 row and 81 columns,

  The test data consist of 293 row and 80 columns in this data set the dependent column is not present.



```
In [2]: df = pd.read_csv('train.csv')
        df.head()
```

Out[2]:

	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	LotConfig	LandSlope	Neighborhood	Condition1	Cond
0	127	120	RL	NaN	4928	Pave	NaN	IR1	Lvl	AllPub	Inside	Gtl	NPkVill	Norm	
1	889	20	RL	95.0	15865	Pave	NaN	IR1	Lvl	AllPub	Inside	Mod	NAmes	Norm	
2	793	60	RL	92.0	9920	Pave	NaN	IR1	Lvl	AllPub	CulDSac	Gtl	NoRidge	Norm	
3	110	20	RL	105.0	11751	Pave	NaN	IR1	Lvl	AllPub	Inside	Gtl	NWAmes	Norm	
4	422	20	RL	NaN	16635	Pave	NaN	IR1	Lvl	AllPub	FR2	Gtl	NWAmes	Norm	

```
In [3]: df.shape
Out[3]: (1168, 81)
```


- Data Pre-processing Done

  What were the steps followed for the cleaning of the data? What were the assumptions done and what were the next actions steps over that?

  The Pre-processing start with the check for the missing values. Using mean of the feature columns

  Second step to identify the numerical feature column in the data using List comprehension

  Third step to identify the numerical variables are usually of two type continuous and discrete there 17 columns in which we have discrete variable.

  I have created discrete feature and continuous feature variable

Then did some EDA (Exploratory Data Analysis) to get some insight of the data set to make some decision based on the analysis.

Then in next step the data set feature containing Object type are stored in one variable Categorical data. Also identified the unique value in it and printed it of the complier.

Took the copy of the dataset and replace the missing value with the new label

Then with help list comprehension checking for variable that contains missing. we will print the numerical nan variable and percentage of missing values.

Then number feature applying log transformation to the features.

For categorial data, data frame group by on the column of sale price mean of the target column.

Scaling features on all feature excluding Id and Target column using minmax scaler.

- Hardware and Software Requirements and Tools Used

Listing down the hardware and software requirements along with the tools, libraries and packages used. Describe all the software tools used along with a detailed description of tasks done with those tools.

The Minimum Hardware requirement for this project is 4 GB RAM, i3 processor.

The software requirements are Python 3.2 version libraries and packages used are as follows:

Pandas, Numpy, Matplotlib, Seaborn, Warnings, Sklearn Pre-processing, MinMax scaler, Metrics, linear model, tree, Ensemble.model selection, xgboost, Regulration,Lasso, GridSearchCV,Model saving Pickle