

HOUSING PRICE PREDICTION PROJECT

Submitted by:

SANJAY SIVAKUMAR

**ACKNOWLEDGMENT**

In this project I have done only the analysis which I learned by doing hands-on during my learning period of data analytics. So, In this acknowledgement I would like to thank my professors of Data Trained Education.

**INTRODUCTION**

* Business Problem Framing

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.

* Conceptual Background of the Domain 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.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.

* Review of Literature

The company is looking at prospective properties to buy houses to enter the market. You are required 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. For this company wants to know:

• Which variables are important to predict the price of variable?

• How do these variables describe the price of the house?

* Motivation for the Problem Undertaken

Real estate 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.As a data enthusiasist there is nothing much more motivation needed than this.

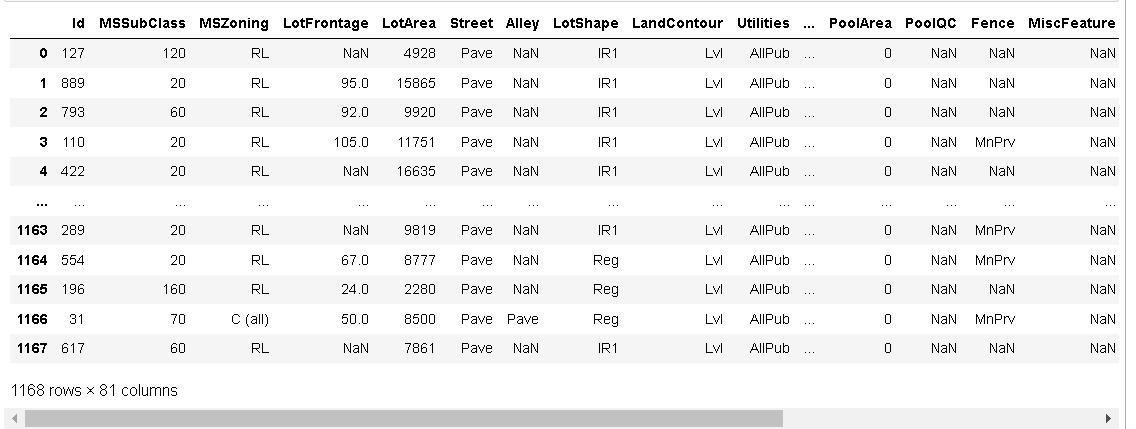
**Analytical Problem Framing**

* Mathematical/ Analytical Modelling of the Problem

You are required to model 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.

* Data Sources and their formats

The dataset consists of 81 columns out of which 80 input and one target variable.



* Data Pre-processing Done

1.Firstly,our dataset requires much tuning process to be ready for the analysis stage.

2.Then we can check for any null values present in the dataset.

3.Then Encoding the objective data.

* Data Inputs- Logic- Output Relationships

Currently we have is as much as 81 columns in our dataset out of which the column “SalePrice” is our target/output column.Thus we have 80 columns as input to predict our target column.

* Hardware and Software Requirements and Tools Used

For this project I have used python as programming language and Jupyter notebook as its tool.The libraries and packages I have used in this project are,

Text

Description automatically generated

1.pandas- for dataframing the dataset

2.numpy- for any analytical functions

3.seaborn and matplotlib-for visualization process

4.train\_test\_split- for splitting data into training and testing

5.And I have imported regression algorithms such as Random Forest Regressor,KNeighbors Regressor etc… for model building.

6.And finally imported metrics for regression problem for proper model selection.

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

Methods used during this project are,

1.Z-Score method- for removing skewness/outliers

2.Box-Cox transformation- for removing skewness/outliers

3.Correlation method- for analyzing the correlation of input columns with our target column

4.VIF method- to check for any multicollinearity present in our input columns

5.Standard Scaler method- to standardize our input column data.

* Testing of Identified Approaches (Algorithms)

1.Linear regression

2.Decision Tree Regressor

3.Random Forest Regressor

4.KNeighbors Regressor

5.Standard Vector Regressor

* Run and Evaluate selected models

Text, letter

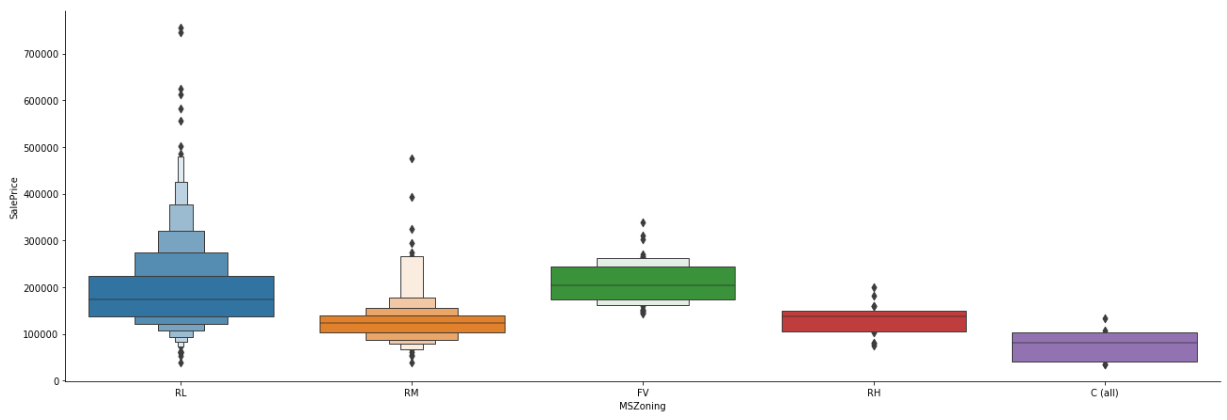
Description automatically generated

* Key Metrics for success in solving problem under consideration

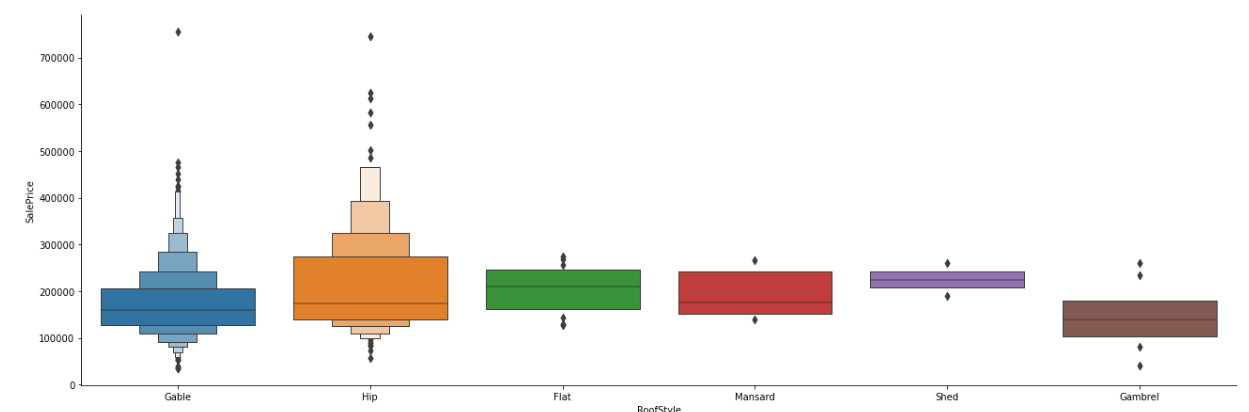
1.r2 score- to know the accuracy of the model

2.mean squared error-error predicting metrics

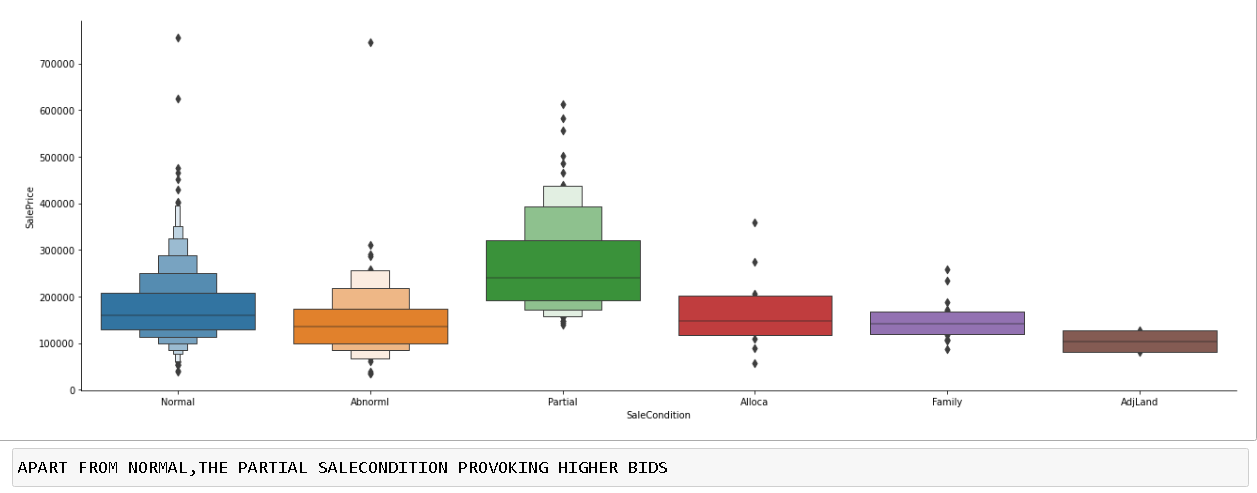
* Visualizations



-RESIDENCY WITH LOW DENSITY ATTRACTS HIGHER PRICE RANGE



-HIP and GABLE ROOFSTYLE HOUSES ATTRACTS HIGHER BID



* Interpretation of the Results

From Data Preprocessing there are many null values which needs to be replaced properly. Then while using Histogram and Box Plot I found that there are more outliers and I tried to reduce it. Then I go for model building where I used various algorithms and found out the best fit algorithm for my model.

**CONCLUSION**

* Key Findings and Conclusions of the Study

Initially I have founded that there is much skewness present in the dataset.So I used various methods to reduce it.Also found that there is no multicolliniearity present in the columns.Then I standardize the dataset for model building and I used as much as 5 algorithms to build my model.Later I found that the algorithm Random Forest Regressor is the best suit for my model by analyzing using the cross-val score and its desired metrics.

* Learning Outcomes of the Study in respect of Data Science

-By this project I have learned about various parameters that affect the Housing Prices.

-I have found that the columns,

"LotShape","ExterQual","BsmtQual","BsmtExposure","HeatingQC""KitchenQual","GarageType","GarageFinish" doesn’t much affect the sale price of the house.

-And the columns,

"OverallQual","YearBuilt","YearRemoteAdd","1stFlrSF",

"GrLivArea","FullBath","TotRmsAbvGrd","Fireplaces",

"GarageCars", "Garage Area","OpenPorchSF" affects more our target column. Thus, these columns are the vital parameters to be considered by the company.

* Limitations of this work and Scope for Future Work

After my final model I got a accuracy which partially good in this scenario and this accuracy can be further increased by adding more columns and using advanced feature Engineering technique to this dataset.