HOUSING: PRICE PREDICTION

Submitted by:

Rutuja Patil

ACKNOWLEDGMENT

The real estate sector is an important industry with many stakeholders ranging from regulatory bodies to private companies and investors. Among these stakeholders, there is a high demand for a better understanding of the industry operational mechanism and driving factors.

Today there is a large amount of data available on relevant statistics as well as on additional contextual factors, and it is natural to try to make use of these in order to improve our understanding of the industry.In some cases, non-traditional variables have proved to be useful predictors of real estate trends. For example, in it is observed that Seattle apartments close to specialty food stores such as Whole Foods experienced a higher increase in value than average.

This project can be considered as a further step towards more evidence-based decision making for the benefit of these stakeholders. The project focused on assessment value for residential properties in Calgary between 2017-2020 based on data from. The aim of our project was to build a predictive model for change in house prices in the year 2021 based on certain time and geography dependent variables.

INTRODUCTION

- ## Business Problem Framing

  To implement model we 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.

- ## Conceptual Background of the Domain Problem

  - Create an effective price prediction model.
  - Validate the model's prediction accuracy.
  - Identify the important home price attributes which feed the model's predictive power.

- ## Review of Literature

House price prediction can be done by using a multiple prediction models (Machine Learning Model) such as support vector regression, artificial neural network, and more. There are many benefits that home buyers, property investors, and house builders can reap from the house-price model. This model will provide a lot of information and knowledge to home buyers, property investors and house builders, such as the valuation of house prices in the present market, which will help them determine house prices. Meanwhile, this model can help potential buyers decide the characteristics of a house they want according to their budget Previous studies focused on analyzing the attributes that affect house price and predicting house price based on the model of machine learning separately. However, this article combines such a both predicting house price and attributes together.

Analytical Problem Framing

- Data Preprocessing Done

  1.Data Exploration

  2.Data cleaning

  3.Data Visualization

  4.Feature Selection

  5.Build model

- Hardware and Software Requirements and Tools Used

  a. Jupyter Notebook

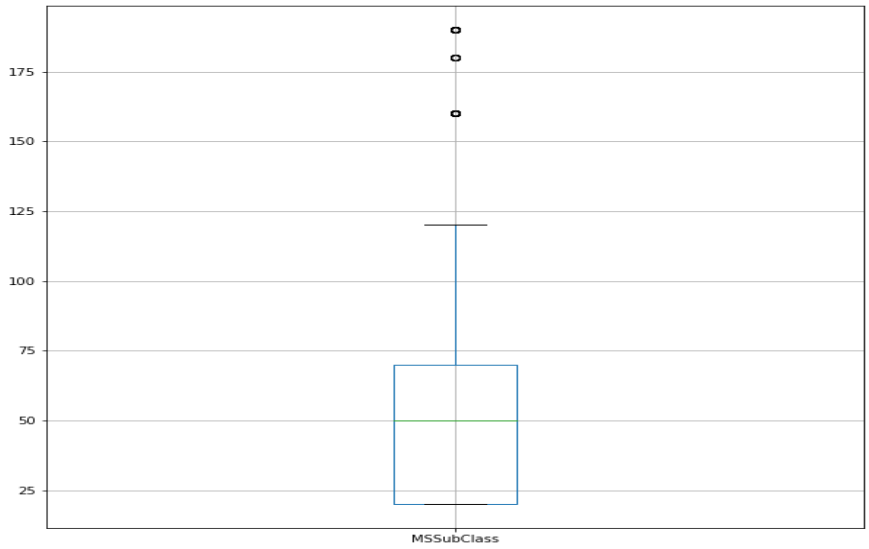
  b. Libraries –

```
from sklearn.model_selection import train_test_split
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier, AdaBoostClassifier,
ExtraTreesClassifier, GradientBoostingClassifier
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis,
QuadraticDiscriminantAnalysis
from sklearn.metrics import accuracy_score, log_loss
from sklearn.neighbors import KNeighborsClassifier
from sklearn.linear_model import LogisticRegression
from sklearn.decomposition import PCA
from sklearn.preprocessing import StandardScaler, RobustScaler
from sklearn.naive_bayes import GaussianNB
import statsmodels.api as sm
from scipy.stats import probplot
from sklearn import metrics
from sklearn import preprocessing
from sklearn import utils
from sklearn.metrics import mean_absolute_error
from statistics import mean, stdev
from sklearn.metrics import f1_score
from sklearn.metrics import classification_report
from sklearn.metrics import accuracy_score, roc_curve, auc, confusion_matrix,
plot_confusion_matrix
from sklearn.model_selection import KFold, cross_val_score, cross_validate,
StratifiedKFold
from sklearn.svm import SVC
from sklearn.feature_selection import chi2
from sklearn.feature_selection import SelectKBest
```

Models Development and Evaluation

- Basic Stats

`describe()` shows a summary of numerial features, which can be visualized using boxplots and histograms. `value_counts()` can be used to generate a summary of categorical features.



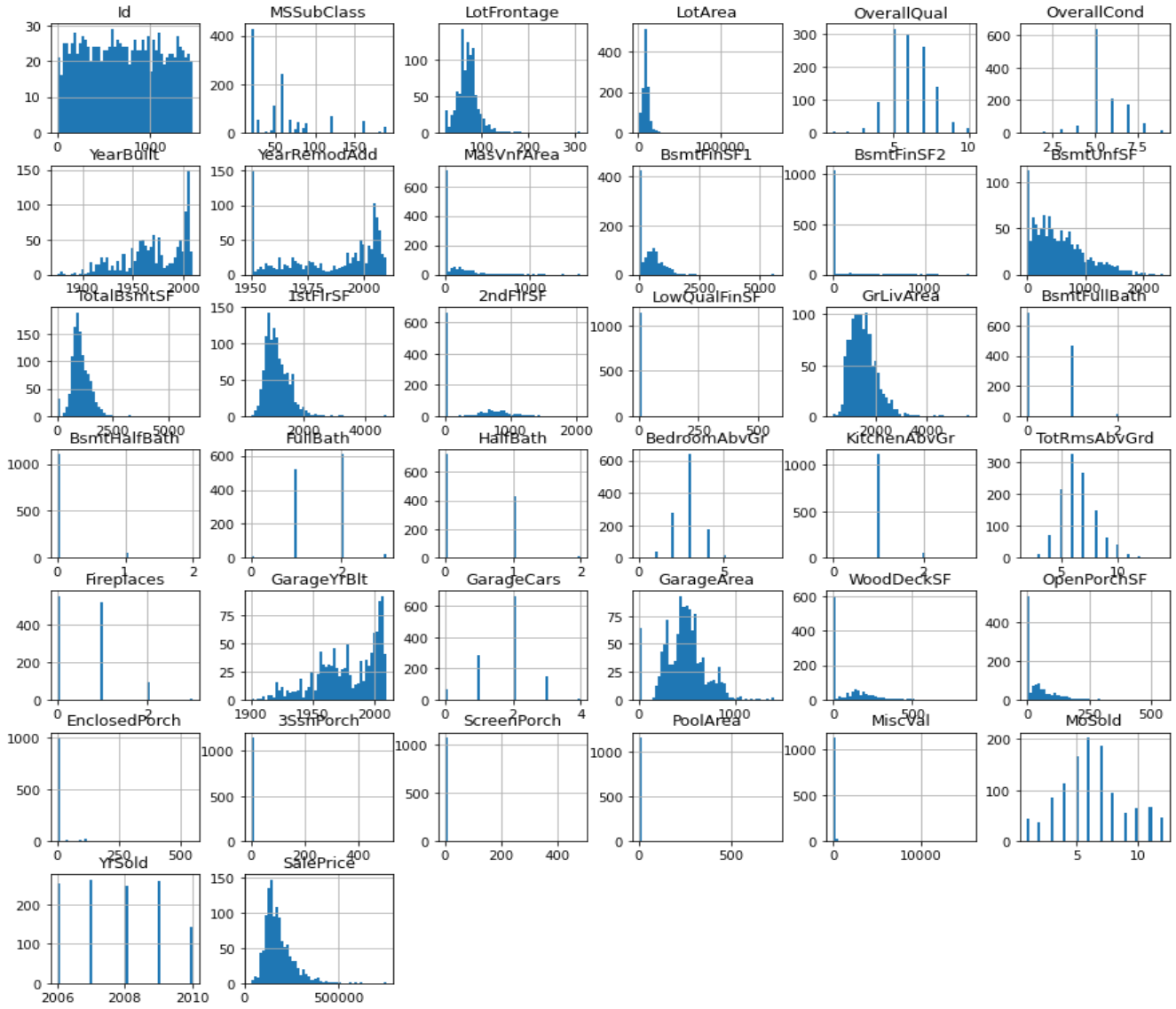
1. Do the data make sence? scan each column and see whether the data make sense at a high level.
    - longitude, latitude and house median age seem to be OK.
    - total rooms and total bedrooms are in hundreds and thousands - this does make sense given each row is a district
    - population seems to be OK but you want to know what's the unit for the number, e.g., thousands? millions? households are numbers of people living together, which is OK. Households mean is 499 and population mean is 1425, so you can tell that population is just the total number of people in each district not in thousands or millions.
    - median income is apparently problematic - how can mean income be 3.87? Actually, this is because median income has been scaled and capped between 15.0001 (max) and 0.4999 (min). Preprocessed attributes like this are common. Knowning how the data is calculated is critical.
    - median house value data is OK and this is our target variable, i.e., we want to build a model to predict this value.
2. Feature scaling: you have noticed that the features have very different scales, which we need to handle later
3. Distribution: from the histograms, we can tell many of them are skewed, i.e., having a long tail on one side or the other. In many cases, we need to transform the data so that they have more bell-shaped distributions.

Data Exploration

Data exploration is the first step in data analysis and typically involves summarizing the main characteristics of a data set, including its size, accuracy, initial patterns in the data and other attributes. It is commonly conducted by data analysts using visual analytics tools, but it can also be done in more advanced statistical software, Python. Before it can conduct analysis on data collected by multiple data sources and stored in data warehouses, an organization must know how many cases are in a data set, what variables are included, how many missing values there are and what general hypotheses the data is likely to support. An initial exploration of the data set can help answer these questions by familiarizing analysts with the data with which they are working.

Data Visualization

Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data. In the world of Big Data, data visualization tools and technologies are essential to analyse massive amounts of information and make data-driven decisions.

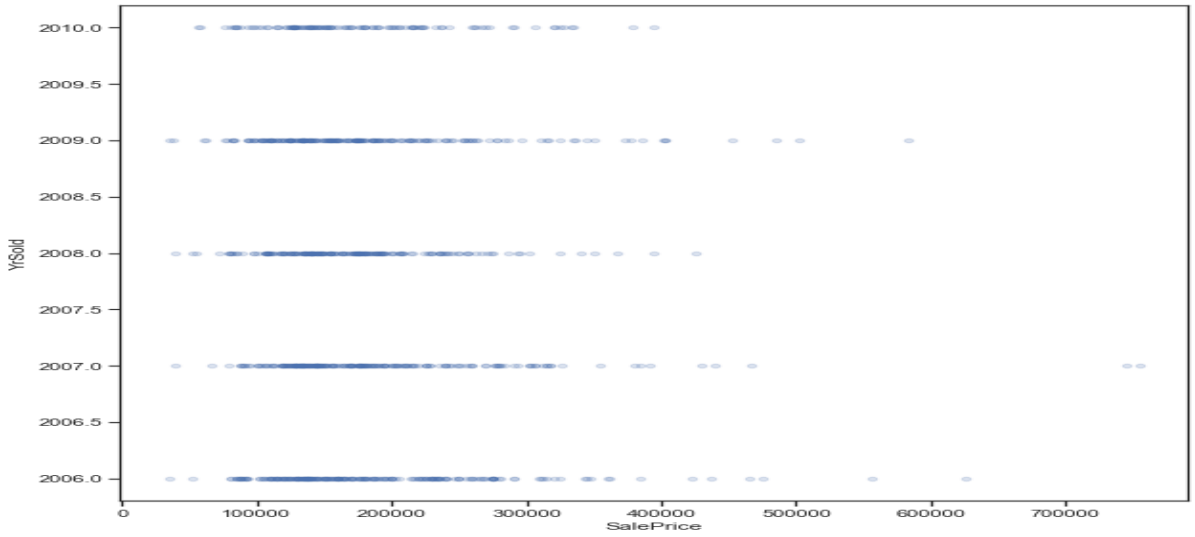


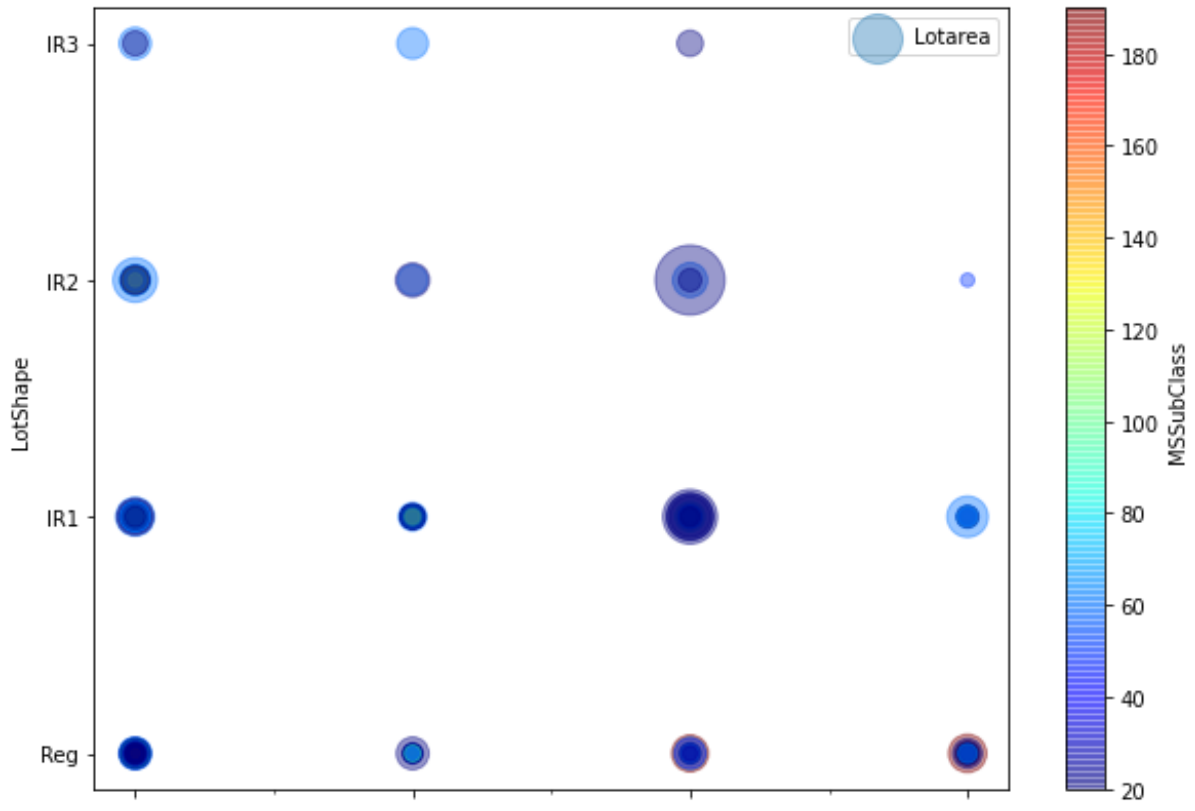
Data Selection

Data selection is defined as the process of determining the appropriate data type and source, as well as suitable instruments to collect data. Data selection precedes the actual practice of data collection. This definition distinguishes data selection from selective data reporting

The process of selecting suitable data for a research project can impact data integrity.

The primary objective of data selection is the determination of appropriate data type, source, and instrument(s) that allow investigators to adequately answer research questions. This determination is often discipline-specific and is primarily driven by the nature of the investigation, existing literature, and accessibility to necessary data sources.





CONCLUSION

The accurate prediction model would allow investors or house buyers to determine the realistic price of a house as well as the house developers to decide the affordable house price. This paper addressed the attributes used by previous researchers to forecast a house price using various prediction models.

These models were developed based on several input attributes and they work significantly positive with house price. In conclusion, the impact of this research was intended to help and assist other researchers in developing a real model which can easily and accurately predict house prices. Further work on a real model needs to be done with the utilization of our findings to confirm them.