HOUSING: PRICE PREDICTION

Submitted by:

Himani Uniyal

ACKNOWLEDGMENT

I am really grateful for this project opportunity and would sincerely thank to “Flip Robo” for trusting me with this project.

Besides this, I would like to express my gratitude towards Srishti Mann Ma’am (SME) to provide us quick revert on the queries raised and continuous support throughout the project.

Also, I have utilized a few external resources that helped me to complete the project. All the external resources that were used in creating this project are listed below:

https://www.google.com/

https://www.youtube.com/

https://scikit-learn.org/stable/user_guide.html

https://github.com/

https://www.kaggle.com/

https://medium.com/

https://towardsdatascience.com/

https://www.analyticsvidhya.com/

INTRODUCTION

- ## Business Problem Framing

  Housing is one of the basic needs of every person in the world and so the housing market is one of the major contributors to the global economy. It is a very large market and there are various companies operating in the domain.

  Data science comes as a very important tool for problem-solving in the domain to help companies increase their total revenue, profits, improve their marketing strategies and focus on changing trends in real estate sales. Predictable modelling, market-based modelling, recommendation systems and other machine learning strategies used to achieve the business objectives of real estate companies. Our problem is related to one such housing company.

  We need to model housing prices with available independent variables. This model will then be used by managers to understand how prices vary widely. They can by controlling the firm's strategy and focusing on areas that will bring the highest profits. In addition, the model will be a good way for managers to understand price fluctuations in the new market.

- ## Conceptual Background of the Domain Problem

  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. The data is provided in the CSV file below.

  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:

1. Which variables are important to predict the price of a variable?
2. How do these variables describe the price of the house?

- ## Review of Literature

  Based on the sample data provided to us from our client database where we have understood that the company is looking at prospective properties to buy houses to enter the market. The data set explains it is a regression problem as we need 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.

- ## Motivation for the Problem Undertaken

  Our main objective of this project is to build a real estate forecasting model with the help of other supporting features. We will predict using machine learning algorithms.
  Sample data is provided to us from our customers' website. In order to improve customer selection, the client seeks specific predictions that can help them invest more and improve customer selection.
  The House Price Index is often used to measure changes in the price of a house. Since house prices are closely linked to other factors such as location, location, population, it requires more information than HPI to predict the value of individual homes.
  There have been a large number of papers that use traditional machine learning methods to accurately predict house prices, but they rarely worry about the performance of individual models and ignore the lesser-known but more complex models.
  As a result, to explore the various effects of the factors on predictive methods, this paper will use standard and advanced machine learning methods to investigate the differences between a few advanced models.

Analytical Problem Framing

- Mathematical/ Analytical Modeling of the Problem

  We created a model in Machine Learning to predict the actual number of future buildings and decide whether to invest in them or not. Therefore, this model will help us determine what changes are necessary to predict the price of a variable and how these variables define the price of the house. This will help determine the price of available flexible private housing. They can by controlling the company's strategy and focusing on areas that will bring the highest profits.

- Data Sources and their formats

  Data set provided by Flip Robo was in the format of CSV (Comma Separated Values). The dimension of data is 1168 rows and 81 columns. There are 2 data sets that are given. One is training data and one is testing data.
  1) Train file will be used for training the model, i.e., the model will learn from this file. It contains all the independent variables and the target variable. Size of training set: 1168 records.
  2) Test file contains all the independent variables, but not the target variable. We will apply the model to predict the target variable for the test data. Size of test set: 292 records.

  .

- Data Preprocessing Done

  Data pre-processing in Machine Learning refers to the technique of preparing (cleaning and organizing) the raw data to make it suitable for a building and training Machine Learning models. I have used some following pre-processing steps:

  a. Loading the training dataset as a dataframe
  b. Used pandas to set display I ensuring we do not see any truncated information

c. Checked the number of rows and columns present in our training dataset
d. Checked for missing data and the number of rows with null values
e. Dropped all the unwanted columns and duplicate data present in our data frame
f. Separated categorical column names and numeric column names in separate list variables for ease in visualization
g. Checked the unique values information in each column to get a gist for categorical data
h. Performed imputation to fill missing data using mean on numeric data and mode for categorical data columns
i. Used Pandas Profiling during the visualization phase along with count plot, scatter plot and the others
j. With the help of ordinal encoding technique converted all object data type columns to numeric data type
k. Separate feature and label data to ensure feature scaling is performed avoiding any kind of biasness
l. Checked for the best random state to be used on our Regression Machine Learning model pertaining to the feature importance details
m. Finally created a regression model function along with evaluation metrics to pass through various model formats

- Data Inputs- Logic- Output Relationships

  When loading the training database, we had to go through various data processing steps to understand what we were given and what we were expected to predict about the project. When it comes to the logical part of domain knowledge to understand how real estate works and how we should provide it to customers is even helpful to train the model with modified installation data.

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

The assumption part for me was relying strictly on the data provided to me and taking into consideration that the separate training and testing datasets were obtained from real people surveyed for their preferences and how reasonable a price for a house with various features inclining to them were .

- Hardware and Software Requirements and Tools Used

Hardware Used:

i. RAM: 8 GB
ii. Wi-Fi router

Software Used:

i. Programming language: Python
ii. Distribution: Anaconda Navigator
iii. Browser based language shell: Jupyter Notebook

Libraries/Packages Used:

Pandas, NumPy, matplotlib, seaborn, scikit-learn.

Model/s Development and Evaluation

- Identification of possible problem-solving approaches (methods)

  We need to predict the sale price of houses, means our target column is continuous so this is a regression problem. I have used various regression algorithms and tested for the prediction. By doing various evaluations I have selected Extra Trees Regressor as best suitable algorithm for our final model as it is giving good r2-score and least difference in r2-score and CV-score among all the algorithms used. Other regression algorithms are also giving me good accuracy but some are over-fitting and some are with under-fitting the results which may be because of fewer amounts of data. In order to get good performance as well as accuracy and to check my model from over-fitting and under-fitting I have made use of the K-Fold cross validation and then hyper parameter tuned the final model.

  Once I was able to get my desired final model I ensured to save that model before I loaded the testing data and started performing the data pre-processing as the training dataset and obtaining the predicted sale price values out of the Regression Machine Learning Model.

- Testing of Identified Approaches (Algorithms)

  The algorithms used on training and test data are as follows:

  A. Linear Regression Model
  B. Random Forest Regression Model
  C. Gradient Boosting Regression Model

- Run and Evaluate selected models

```
from sklearn.model_selection import cross_val_score
ml_models=[LinearRegression(),SVR(),RandomForestRegressor(),GradientBoostingRegressor()]
for m in ml_models:
    m.fit(x_train,y_train)
    predm=m.predict(x_test)
    mse=mean_squared_error(y_test,predm)
    mae=mean_absolute_error(y_test,predm)
    r2=r2_score(y_test,predm)
    print(f'metrics of {m}:')
    print(f' mean_absolute_error: {mae}\n mean_squared_error: {mse}\n r2_score: {r2} ')
    score=cross_val_score(m,x_scaled,y, cv=5)
    print(' mean cv score:',score.mean())
    print('\n\n')
```

```
metrics of LinearRegression():
 mean_absolute_error: 0.10219821153348653
 mean_squared_error: 0.02092735993565864
 r2_score: 0.8646198624524845
 mean cv score: 0.8533261845767184


metrics of SVR():
 mean_absolute_error: 0.1087455139273754
 mean_squared_error: 0.030182121731819538
 r2_score: 0.8047503457630483
 mean cv score: 0.8175661098506278


metrics of RandomForestRegressor():
 mean_absolute_error: 0.10768067821532128
 mean_squared_error: 0.025588619994928036
 r2_score: 0.8344659381204724
 mean cv score: 0.8516339503679953


metrics of GradientBoostingRegressor():
 mean_absolute_error: 0.10038632720302777
 mean_squared_error: 0.022731717449039304
 r2_score: 0.8529473835016091
 mean cv score: 0.86811135778508
```

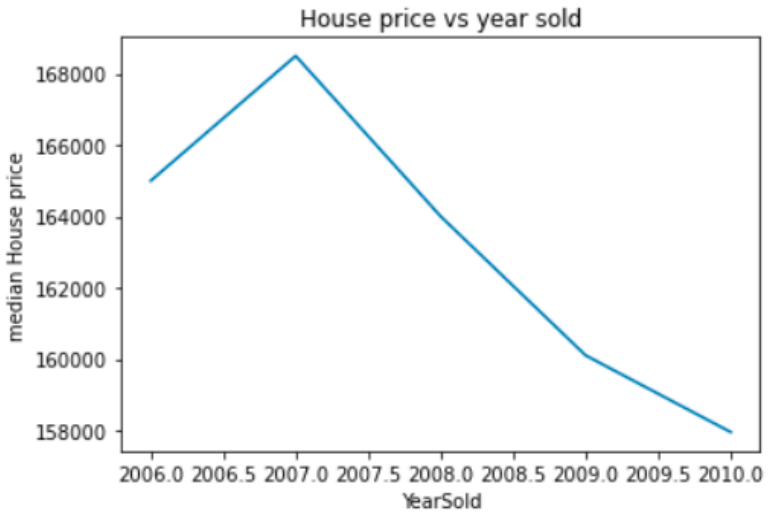
- Key Metrics for success in solving problem under consideration

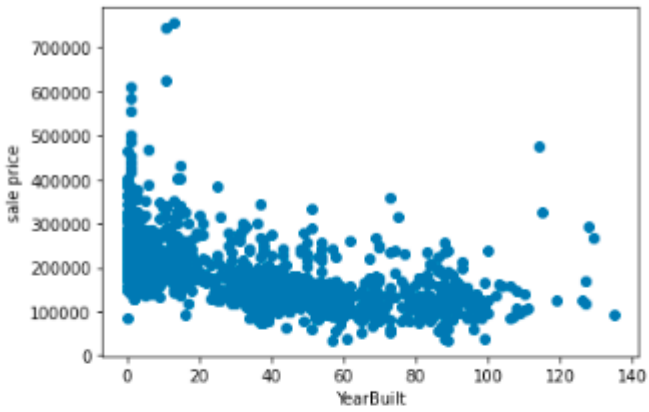
  The key metrics used here were r2_score, cross_val_score, MAE, MSE and RMSE. We tried to find out the best parameters and also

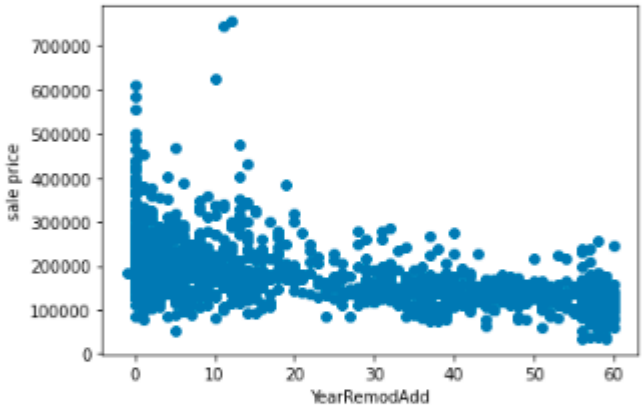
to increase our scores by using Hyperparameter Tuning and we will be using GridSearchCV method

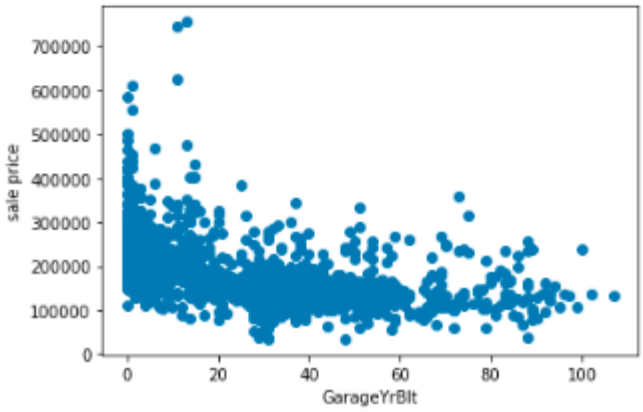
- Visualizations

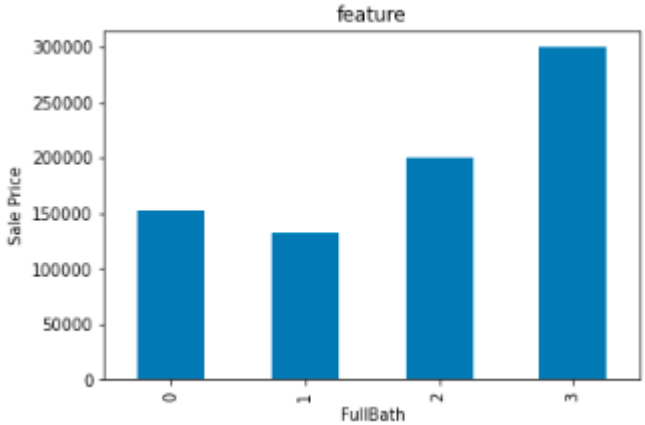
  The dataset was splitted into object type, numeric type and temporal type. Also the numeric data was further divided into discrete and continuous. And the data set was visualized accordingly.

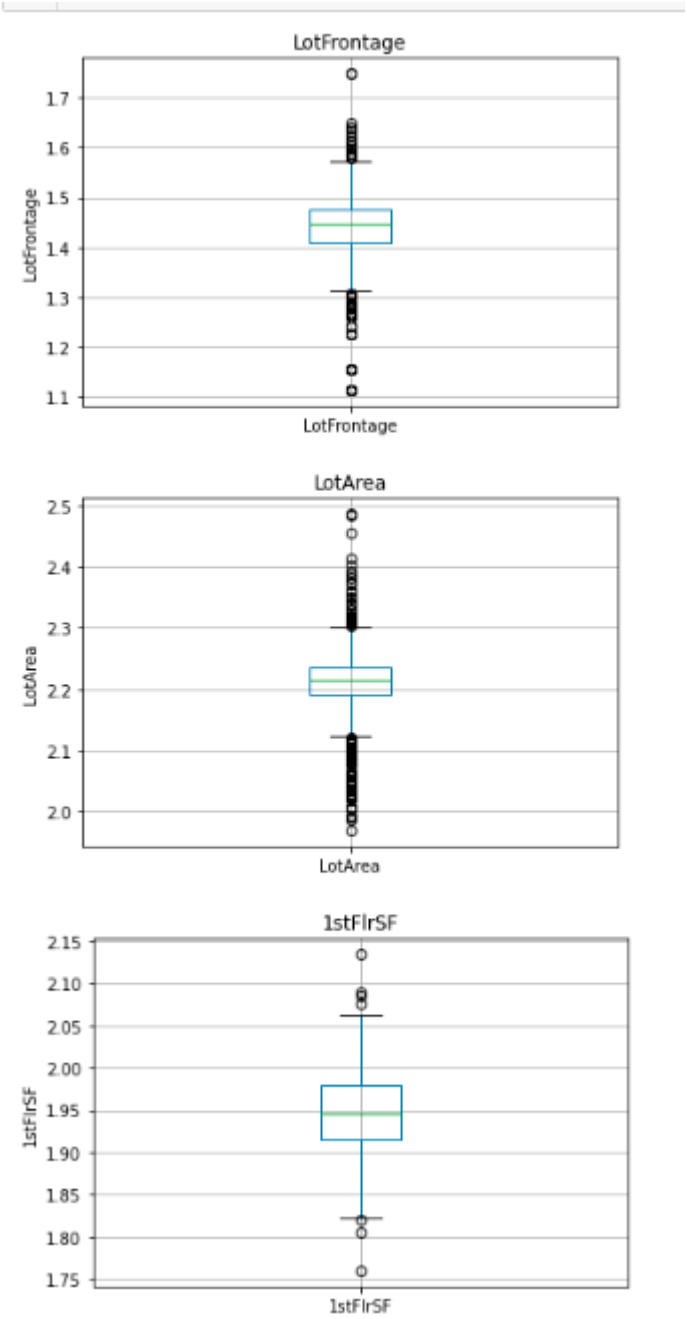



sale price
YearBuilt


sale price
YearRemodAdd


sale price
GarageYrBlt


feature
Sale Price
FullBath



- Interpretation of the Results

  Visualizations: It helped me to understand the correlation between independent and dependent features. Also, helped me with feature importance and to check for multi collinearity issues. Detected outliers/skewness with the help of boxplot and distribution plot.

  Further we used feature selection method to get the best feature and started making ML model.

CONCLUSION

- Key Findings and Conclusions of the Study

  Post model building and choosing the appropriate model I went ahead and loaded the testing dataset. After applying all the data pre-processing steps as the training dataset I was then able to get the predicted sale price results.

- Learning Outcomes of the Study in respect of Data Science

  The above study helps one to understand the business of real estate. How the price is changing across the properties. With the Study we can tell how multiple real estate amenities like swimming pool, garage, pavement and lawn size of Lot Area, and type of Building raise decides the cost. With the help of the above analysis, one can sketch the needs of a property buyer and according to need we can project the price of the property

- Limitations of this work and Scope for Future Work

  There is large number of missing values presents in this data set, so we have to fill those missing values in correct manner. We can still improve our model accuracy with some feature engineering and by doing some extensive hyper parameter tuning on it