Washington Property Sale Price Prediction Final Report

Richa Sharma Rishi Bamb Sonali Gupta Tirth patel

***Abstract* — Due to economic cycles, the real estate market in Washington has been up and down for numerous decades. While economic variables have influenced Washington house prices, the fundamental attributes of a home may also influence the price of a home. It would be helpful to analyze what the driving factors of a costly property in the U. S. are for house traders and purchasers to make sensible purchasing decisions. In this project, we are going to predict the price of houses in Washington based on multiple factors affecting the price of the houses. It has been observed that the pandemic has also affected the prices of houses in the United States. To predict the sale price of the properties we will use the machine learning models and to improve the accuracy of the prediction we will use the tuning parameters, analytics helps transform leads into prospects by closing at good deals and making the business more profitable.**

I. INTRODUCTION

In this project we have used the historic data of houses collected from the source Kaggle. The original dataset contains 21 features and a total of 21613 observations to perform various analyses to extract meaningful insights which helps in making quick and efficient decisions. We have performed analytics to predict the target feature trend that is the price of the property along with taking into consideration other features like square feet living, number of bedrooms and bathrooms, date renovated, waterfront, number of floors etc. We are trying to design a price prediction model by using ML algorithms for the users to help them find houses under their budget on the basis of important features which contribute more towards the price of the property. Our goal is to help customers make more efficient decisions by using our model for buying or selling the property in Washington state.

II. MOTIVATION

The project's inspiration came from academic coursework and a significant paper study , both of which were Real-estate related. We decided to use the kaggle dataset for our project which would help in analyzing the properties price based on multiple factors. Our industry case study is based on real estate and the databases, and signifies how the databases play a major role in the real estate business to generate profit for the stakeholders. As the property offers valuations of the houses using machine learning techniques. The motivation of our project is to predict the price of a house in the USA based on the characteristics of the house. As people prefer buying property based on their budget and the requirements hence to solve this problem we are using the machine learning techniques, which helps in predicting the price of the houses in future based on the current market. Also customers can make investments based on the predicted prices in order to make profits in the future. We have used multiple models to predict the price of houses, in our study we have focused on Ordinary Least Squares (OLS) Regression, XGBoost, Gradient Boosting Machines, LightGBM ,CatBoost and apply bagging and boosting techniques to predict more accurately. Using and understanding such techniques helps us to understand the machine learning models and predictions working.

III. LITERATURE SURVEY

A. Visualizing Real Estate Property Information on the Web

Theodor Hong, author of Imperial College London, has developed a system called ReV, or Real Estate Visualizer, that describes the classification and exploration of real estate. High-dimensional domain data is difficult to graph, so use graph and color-based map visualizations to convey data to consumers and stakeholders for business growth. Using this to show the effect of real estate components in our project dataset.

B. Applied research on Real estate price prediction by neural network

Authors Hu Xiaolong and Zhong Ming wrote this work, highlighting the ways of home price forecasting and valuation in the face of strong demand and the continuous flow of real estate data that is difficult to analyze in function. As a result, neural networks are used to generate results based on human behavior, decision making, and various other factors. We can use the pricing and control components to analyze and evaluate property categories that match users in our project.

C. Analysis and design of the real estate property right registration information system for the whole life cycle.

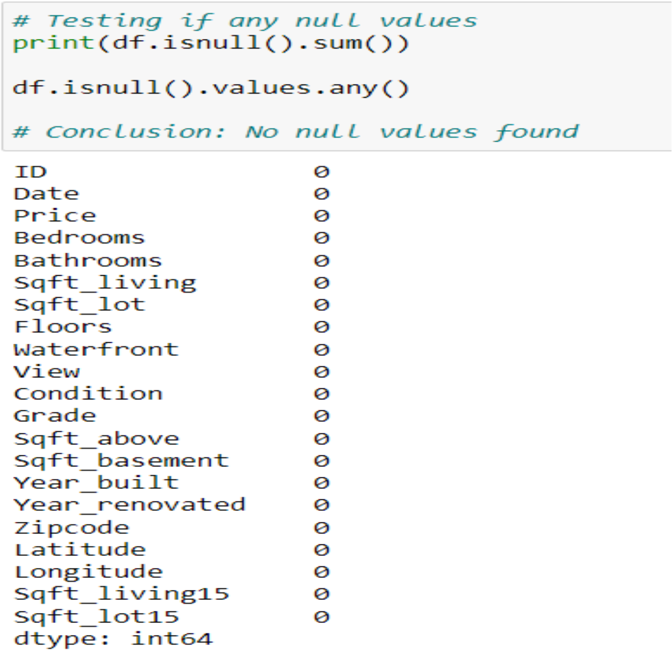
Author Jing Liu, Faculty of Business Administration, Xian Northwestern Institute of Technology, Songzheng Zhao, Faculty of Business Administration, Xian Northwestern Institute of Technology. This paper presents their finding that the concept of real estate property rights and the core application phase are based on registration types. Some visual representations and diagrams, such as requirements models, logic models, and dynamic models are created using the OOA tool Unified Modeling Language (UML). An information system is then implemented to process all types of registrations and receive information about the real estate management cycle. From this, we aim to understand the development of real estate. This helps us to get more efficient results for making business-oriented decisions.

IV. PROPOSED WORK

In this project, we have obtained the dataset from the source kaggle. We have considered several steps for the execution of this project. The collected data was cleaned and prepared using python and its libraries. To find a pattern, descriptive analysis has been performed between the target feature with the rest of the features. Inferential analysis has been performed to get the correlation. And finally, model building followed by evaluating and validating results to predict the target feature, price.

1. Data Cleaning and Preparation

Handling raw data and getting insights is difficult to understand hence data preparation is a crucial step to conduct further analysis. In our project we have used python by testing null values and conducted data engineering. In Figure 1 we can see that no null values were found. Data engineering steps can be seen in figure 2, dropping features like Id, date as they were having no significant impact on the target feature, price. The column Year built was converted to the variable 'Age of the house', which makes more business sense and easier to analyze the dataset. Lastly, convert the data types of all the features to their relevant types to get accurate results.



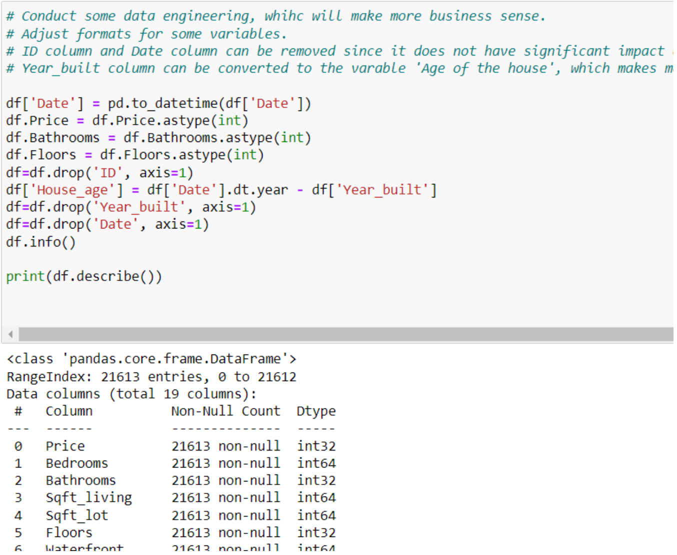
```
# Testing if any null values
print(df.isnull().sum())

df.isnull().values.any()

# Conclusion: No null values found
```

```
ID                 0
Date               0
Price              0
Bedrooms           0
Bathrooms          0
Sqft_living        0
Sqft_lot           0
Floors             0
Waterfront         0
View               0
Condition          0
Grade              0
Sqft_above         0
Sqft_basement      0
Year_built         0
Year_renovated     0
Zipcode            0
Latitude           0
Longitude          0
Sqft_living15      0
Sqft_lot15         0
dtype: int64
```

Figure 1



```
# Conduct some data engineering, whihc will make more business sense.
# Adjust formats for some variables.
# ID column and Date column can be removed since it does not have significant impact
# Year_built column can be converted to the varable 'Age of the house', which makes m

df['Date'] = pd.to_datetime(df['Date'])
df.Price = df.Price.astype(int)
df.Bathrooms = df.Bathrooms.astype(int)
df.Floors = df.Floors.astype(int)
df=df.drop('ID', axis=1)
df['House_age'] = df['Date'].dt.year - df['Year_built']
df=df.drop('Year_built', axis=1)
df=df.drop('Date', axis=1)
df.info()

print(df.describe())
```

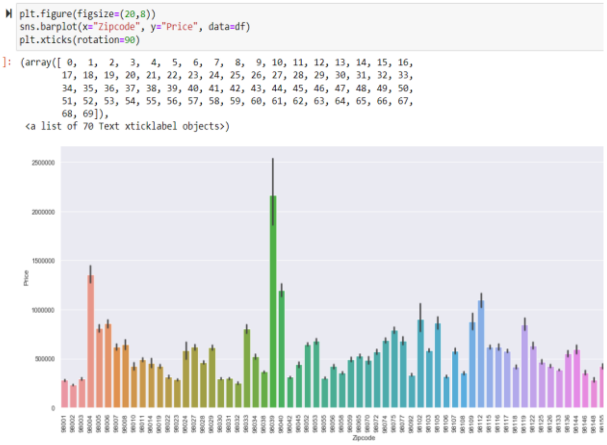
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 21613 entries, 0 to 21612
Data columns (total 19 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   Price           21613 non-null  int32
 1   Bedrooms        21613 non-null  int64
 2   Bathrooms       21613 non-null  int32
 3   Sqft_living     21613 non-null  int64
 4   Sqft_lot        21613 non-null  int64
 5   Floors          21613 non-null  int32
 6   Waterfront      21613 non-null  int64
```

Figure 2

2. Descriptive Analysis

As we are dealing with several features, descriptive statistics played an important part to get basic information about them in the dataset. Also, it helped to highlight the potential relationships in between the features and with the target feature, price to gain relevant insights for building prediction models with good accuracy. There are several graphical and visual methods that enhance the researcher's understanding of individual variables and the relationships between them. The graphical and photographic methods provide a visual representation of the data. Some of these methods are: histogram, scatter plots etc.

From figure 3 we can conclude that Zipcode 96039 has the average highest price in King County. Zipcode 96004, 96040 and 96112 are also leading the house price in King County. Based on the price, zipcode, longitude and latitude, we can build three sales teams that focus on different markets. Team 1 should focus on wealthy clients sell the houses located in the zipcode 96004, 96040 96112. Team 2 should focus in new home buyers who are not able to afford the high house price. The targeted house for them locates in the zipcode 96001, 96002, 96003, 96030, 96031.96032,96148 and 96168. Team 3 focus on other clients and the target houses could locates in any area in King County.



```
plt.figure(figsize=(20,8))
sns.barplot(x="Zipcode", y="Price", data=df)
plt.xticks(rotation=90)
```

```
]: (array([ 0,  1,  2,  3,  4,  5,  6,  7,  8,  9, 10, 11, 12, 13, 14, 15, 16,
        17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33,
        34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50,
        51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67,
        68, 69]),
 <a list of 70 Text xticklabel objects>)
```

Figure 3

Figure 4 and figure 5 concludes that Based on the histogram of Price, we know the mean price is 540,182, and the median price is 450,000. Some properties have been sold for a price far higher than typical for King County. Based on the boxplot, Price has a lot of outliers. Since the mean of the price is higher than the median, it means there are half of the houses inventory in the country whose prices are lower than $450,000. It is a good selling point for our sales team to look for investors and first home buyer who do not have high budget.

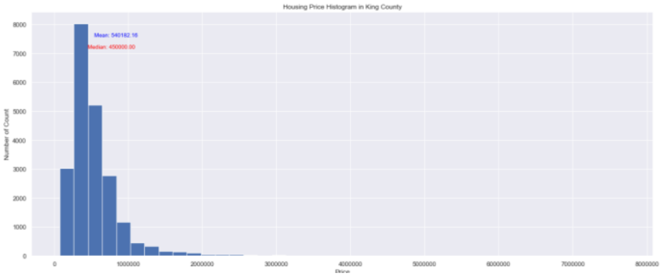


Figure 4

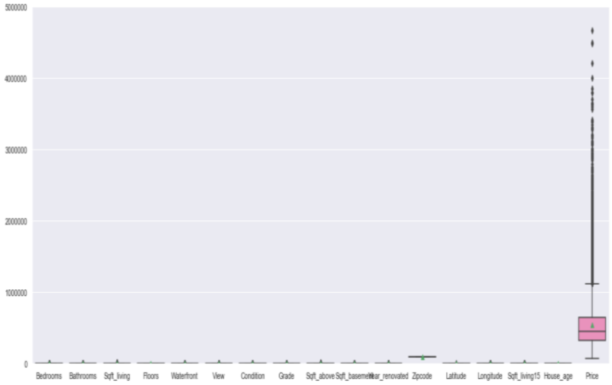


Figure 5

From figure 6 and figure 7 we have concluded that There are 14953 entries for the outliers compared to the 21613 total dataset entries. If we remove the outliers with higher percentage of the dataset, the mean of Price will be lower and the shape of the graph will change a lot. It is good to not remove outliers for this dataset to visualize the data and predict the price. Since the percentage of outliers is high. it may not be a good idea to create a regression model without outliers.

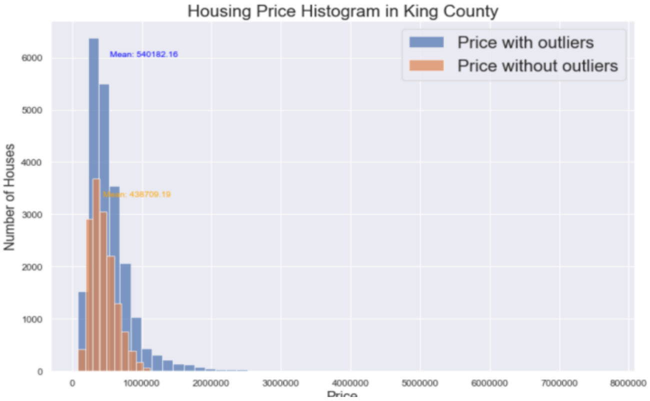


Figure 6

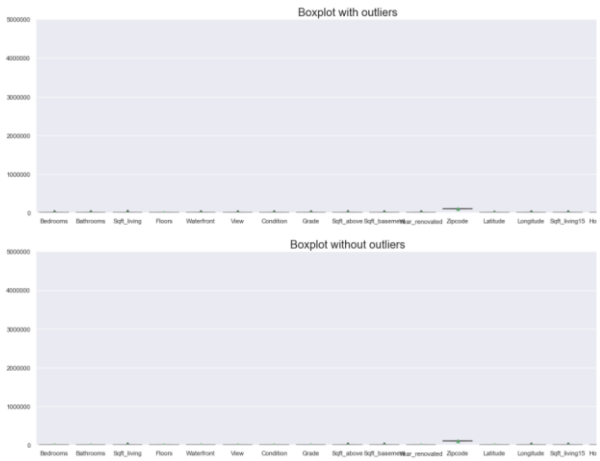


Figure 7

The below figures, figure 8 and figure 9 depicts the target feature relationship with each of the other features. The conclusion drawn from those graphs is that the price is positive related to square feet of living room, square feet above ground and average size of the closest 15 houses in square feet, condition and grade and the level of view, bedroom and bathroom. The houses with waterfront have higher average prices than the ones without waterfront. But the gap is not too much. Price is not obviously related to size of the lot in square feet, square feet below ground and average size of the closest 15 houses' lots, in square feet, house age, year renovated. The house with 2 floors have average higher price than the houses with 1 or 3 floors

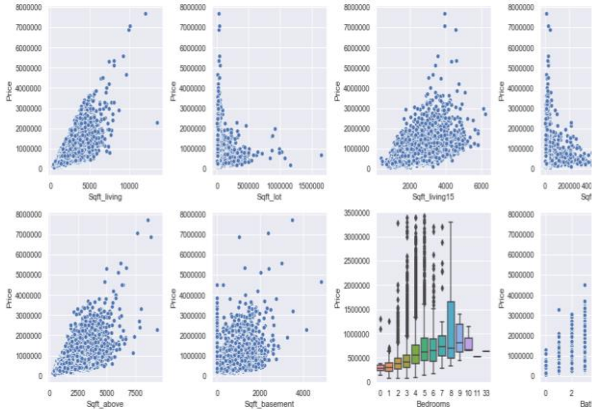


Figure 8

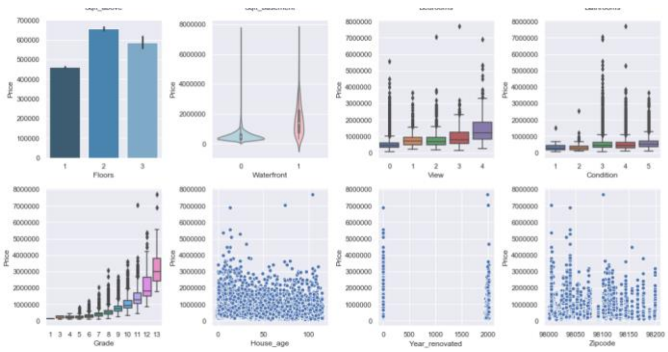


Figure 9

3. Inferential Analysis\

To understand the impact of features on target variables inferential analysis plays an important role.It is crucial and important to know correlation between variables. Inferential statistics helps us with exploring the trends and phenomenon or we can rephrase that it tells about situations.

Correlation Matrix

We have tried to examine the correlation between features with the target feature which is Price.

From the heat map figure 10, we can see how strongly features are correlated to our target feature which is PRICE. These variables which have high correlation with price includes Bathroom, Sqft_living, Grade, Sqft_above, Sqft_living15.

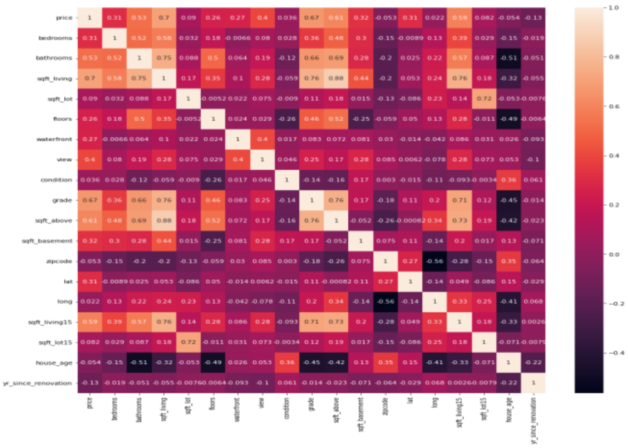


Figure 10

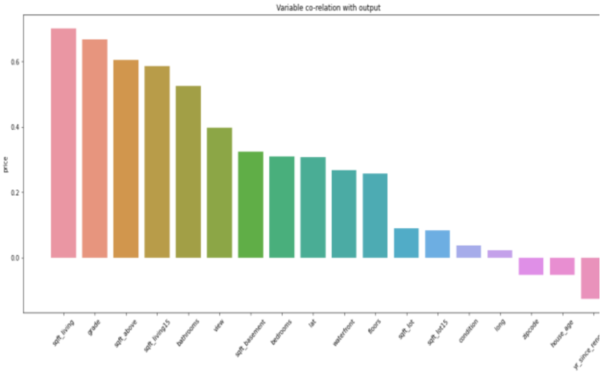


Figure 11

We have generated some of the plots using Python to show the number and range of values for each attribute as well as the covariance among the all attributes.

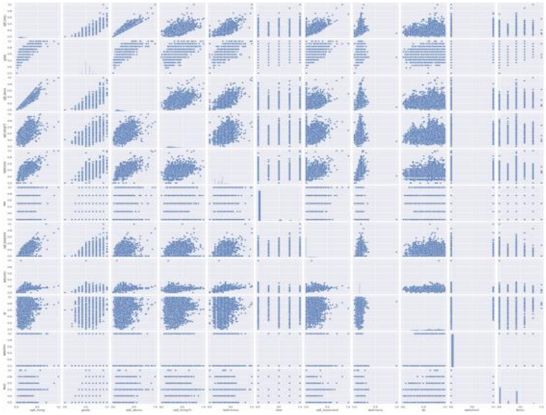


Figure 12

Model Building

We have implemented 6 machine learning models on our dataset for designing price prediction of the properties. These models are Ordinary Least Squares (OLS) Regressor, XGBoost Regressor, MLP Regressor, Catboost Regressor, LightGBM Regressor, and Random Forest Regressor.

Machine Learning Models Architecture

Below diagram shows the architecture of the models. Most of the models divide the dataset into two parts: training and test and then evaluate the results and finally validate the results for giving prediction accuracy.

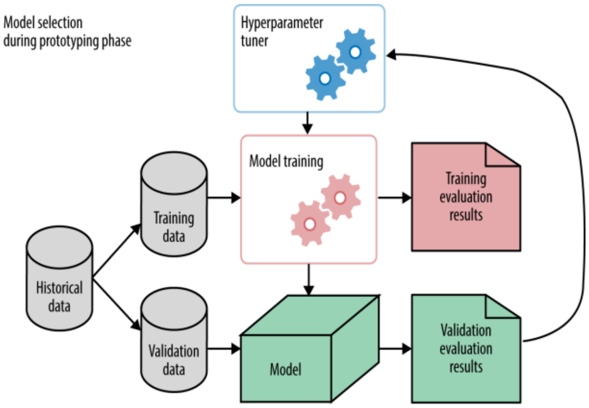


Figure 13

CatBoost Model

CatBoost algorithm is completely based on gradient boosting technique. It works like other boosting algorithms or XGBoost but it supports better with categorical features.The best part about CatBoost is that it does not require extensive data training like other ML models, and can work on a variety of data formats, not undermining how robust it can be. And it is one of the best models which shooted our dataset with highest accuracy.

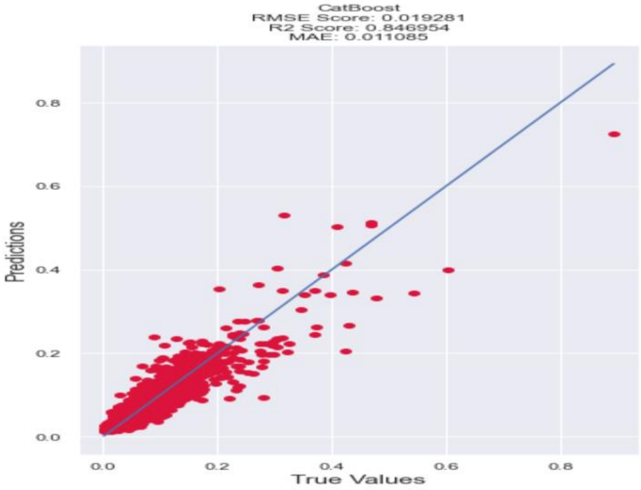


Figure 14

Random Forest Model

Random forest is an ensemble machine learning model which is entirely based on classification or regression trees. Generally, this random forest creates many decision trees and averages their predictions to make the final prediction for the given sample data. For creating each decision tree this random forest algorithm alway uses a subset of all the attributes to avoid overfitting issue.

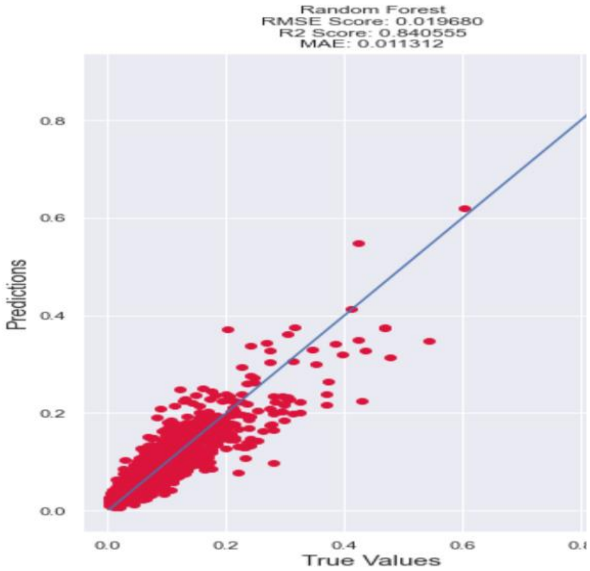


Figure 15

LightGBM

LightGBM algorithm is also based on a gradient boosting algorithm just like CatBoost. It is faster and supports decision tree technique. As it supports a decision tree algorithm, it divides or you can rephrase it splits into leaf wise not level wise like what other models do. Leaf wise distribution reduces the loss significantly which leads to end up with getting far better results or accuracy of the model. And this model is the second best suited model for our price prediction.

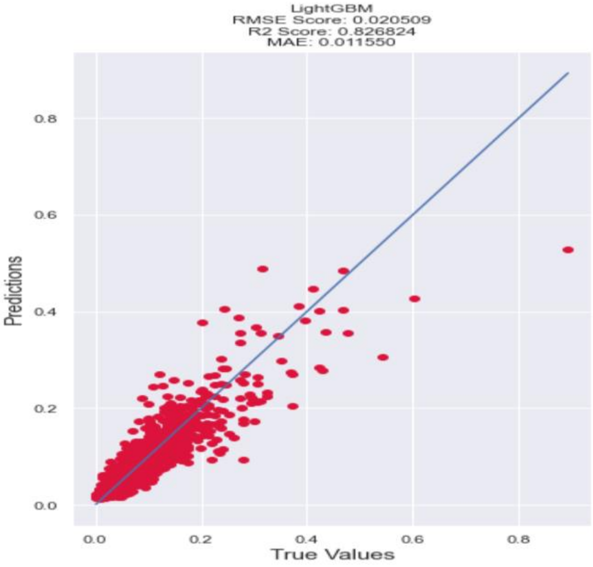


Figure 16

Ordinary Least Squares (OLS) Model

Ordinary Least Squares regression (OLS) is a common technique for estimating coefficients of linear regression equations which describe the relationship between one or more independent quantitative variables and a dependent variable.It works on principle of close relationship between features.Primary goal is to find regression line that fits huge range of the sample data which is provided by calculating intercept and slope of the regression line.

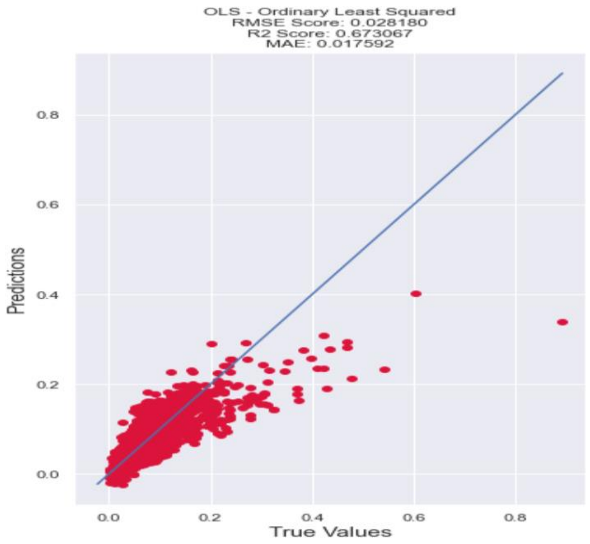


Figure 17

XGBoost Model

XGBools has an immensely high predictive power which makes it the best choice for accuracy. And this XGBoost processes both kinds of models CART ensemble model classifier and regression tree. It makes the algorithm more than 10 times faster than existing gradient booster techniques. It takes output as sum of each leaf node from each tree.Uses iterative computation technique and introduces new function on each iteration (boosting).Good hyperparameter values can be found by trial and error for a given dataset, or systematic experimentation such as using a grid search across a range of values.Randomness is used in the construction of the model. This means that each time the algorithm is run on the same data, it may produce a slightly different model.

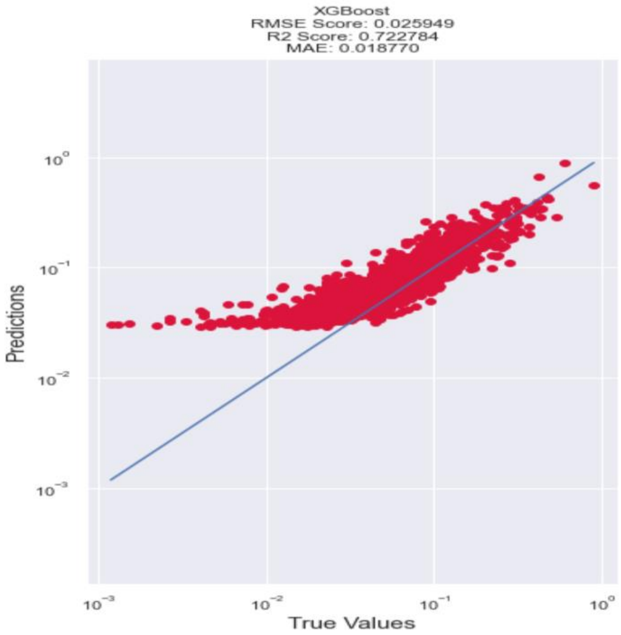
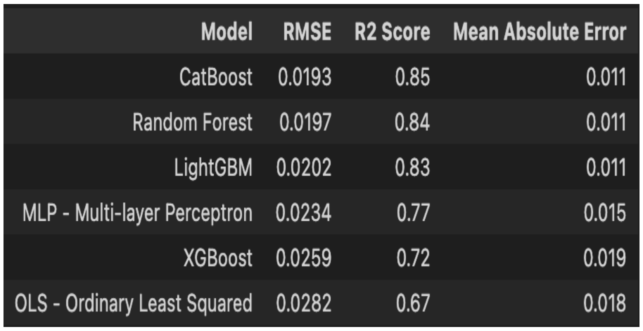


Figure 18

Model Comparison

Below the comparison table shows the results of each prediction model.



Model	RMSE	R2 Score	Mean Absolute Error
CatBoost	0.0193	0.85	0.011
Random Forest	0.0197	0.84	0.011
LightGBM	0.0202	0.83	0.011
MLP - Multi-layer Perceptron	0.0234	0.77	0.015
XGBoost	0.0259	0.72	0.019
OLS - Ordinary Least Squared	0.0282	0.67	0.018

Figure 19

Using these results, we found the CatBoost model is the best suited prediction model that works best for our predictive task of price prediction for the properties based on important highly correlated features. The model, with the accuracy of 85%, shows the best result. And other two best models are Random Forest and LightGBM with the accuracy 84% and 83%.

V. CONCLUSION

On the basis of our ML model implementation we observed there are three models which fitted best to our dataset for predicting the price of the houses for the Washington States.These three models are CatBoost followed by Random forest and LightGBM.

VI. FUTUREWORK

Future scope for our project is to gather global real estate data to design a price prediction model for larger customers to help them to find the best property which falls into their budget and interests for buying or selling. This will benefit the real estate businesses to expand their business globally and target the larger customers and grow their businesses.

REFERENCES

[1] Hong, T., 1999, July. Visualizing real estate property information on the web. In *1999 IEEE International Conference on Information Visualization (Cat. No. PR00210)*(pp. 182-187). IEEE.

[2] Xiaolong, H. and Ming, Z., 2010, July. Applied research on real estate price prediction by the neural network. In *2010 The 2nd Conference on Environmental Science and Information Application Technology* (Vol. 2, pp. 384-386). IEEE.

[3] Liu, J. and Zhao, S.Z., 2012, October. Analysis and design of the real estate property right registration information system for the whole life cycle. In *2012 International Conference on Information Management, Innovation Management and Industrial Engineering* (Vol. 2, pp. 397-401). IEEE.

Link for dataset (Gdrive):
https://drive.google.com/file/d/1sFkTITVjQNBUUS3VcVTyep9hTWSU5Cz4/view?usp=sharing