National College of Ireland

Project Submission Sheet – 2020/2021

Student Name: Omkar Ratnoji Tawade

Student ID: 19232136

Programme: MSCDAD_A **Year:** 2020-21

Module: Data Mining and Machine Learning

Lecturer: Michael Bradford

Submission Due Date: 10-01-2021

Project Title: Predicting House Prices using Machine Learning Techniques

Word Count: 7346 words

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the references section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature: Omkar Ratnoji Tawade

Date: 10-01-2020

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. Projects should be submitted to your Programme Coordinator.
3. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
4. You must ensure that all projects are submitted to your Programme Coordinator on or before the required submission date. **Late submissions will incur penalties.**
5. All projects must be submitted and passed in order to successfully complete the year. **Any project/assignment not submitted will be marked as a fail.**

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Predicting House Prices using Machine Learning Techniques

Omkar Taawade
Data Analytics
National College of Ireland
Dublin, Ireland
x19232136@student.ncirl.ie

***Abstract*— Predicting house price is crucial for both families (who is investing their earnings on the new home) and also for the construction company to select a range of price which would be ideal to them and also for their customers. In case of predicting wrong prices, the family would end up giving some extra money and had to comprise with the house for the rest of their life. Whereas, predicting wrong house prices by construction companies will lead to a loss in their business. Now I have experience of living in different countries (India and Ireland) I can relate the approach for predicting house prices in different countries are different, so a person who is moving to another country will require such a model to buy a house at a reasonable price. This model will bring transparency between customers and construction companies or brokers. I will be building a model to predict the house prices in "King County, Washington State, USA", "Melbourne, Australia", "California, USA". The motive of this study is to build regression models using these respective datasets. The price will be used as the dependent variable and Rooms, Area, Locality will be the features in this model. Regression techniques such as Multi-Linear regression, Lasso regression, Ridge regression, Decision Tree, and RandomForest will be used to produce regression models to predict the price of the house.**

Keywords—Regression, Multi-Linear regression, Lasso regression, Ridge regression, Decision Tree and RandomForest, Prediction, House Price.

I. Introduction (*Heading 1*)

The main motivation of this project is to produce models which can predict house price of particular locality based on a few features. I am trying to produce models such that these models can predict house price accurately, I will be also able to compare which regression method serves best for a particular dataset. I will be working on the three datasets which are "California Housing Market", "House Sale in King County" and "Melbourne Housing Market". These datasets contain previous market trends and prices. I will utilize these features in the analysis and will build a model based on these features. The housing market mostly has an upward trend and features which influence price will be always the same, so we can predict house price in the future with help of the same models. The housing market is always in great demand so it is necessary to build such models which can predict house price at a very low error rate. These models will be helpful for real estate agents while pitching prices for their customers. Also, it will be helpful for customers by introducing transparency between brokers and customers. Real Estate is a different kind of market, people invest their life savings in real estate to buy a house because real estate is the only market that will keep on increasing. So, there is a need for models like this which will help to predict the current value of the house as well as forecast the future value of the house so that people can invest in real estate with much more confidence. Real estate helps the economic growth of the country, so the importance of such models become more in such cases. House prices always depend on the factors such as the number of rooms, bedroom, bathroom, area of living, distance from the city center, locality, parking, etc. and most of these factors are continuous. Regression methods are well suited for the continuous variables. Following are the research questions of this project.

Research Questions:

1. Is the multilinear regression method suited best for producing models that can able to forecast house prices?
2. Among Lasso and Ridge regression, which method performs better than other for KC Housing Data?
3. Among DecisionTree and RandomForest, which method performs better than the other for Melbourne Housing Data?

Housing market data consists of numerous variables. Few of these variables are relevant while forecasting the price and while other variables did not even impact the price to any extent. So, it is very critical to figure out the independent variables because if we do not choose the correct variables it will add noise, and the error rate for predicting the price will increase. Each city has a unique house market so using the same variables for predicting the house prices will be incorrect. Nowadays, house prices somewhat depend on various socio-economic factors but today also it strongly depends on the area, number of rooms, etc. In all three datasets, we will work on all have these strongly independent variables present. California Housing Market dataset is an old dataset. It is derived from 1990 census data. In this dataset, median house value will be the dependent variables, and variables such as house age, total bedrooms, households will be more helpful for predicting median house value. I will be using multilinear regression to produce a model for predicting the house price of California. Sometimes these variables which strongly influence the value of the house are strongly correlated with each other. In this case, we will need to drop anyone variable. Rooms, Bedrooms, and Bathrooms are such variables which can be highly correlated because as the number of rooms increases, other tow variables which are bedroom and bathrooms also increases. So, in such a situation it is better to drop the bedroom and bathroom. In the Melbourne Housing dataset and as well as KC housing dataset we have both room and bedroom variables in the dataset. We will see if they are correlated or not in our analysis. We will use metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), R- Squared value, and Root Mean Squared Error (RMSE) to identify the performance of our models. Also visualizing the model like observing the scatter plot of residual error is a must because only observing the metrics will not give an idea of overfitting. Overfitting can be easily visualized in the scatterplot.

II. Related Work

In their research, L. Gattini, & P. Hiebert [1] presented the platform for forecasting and analyzing house prices in the euro area using the vector error correction model. They had considered features such as real housing investment, real income per capita, and real interests. They used VECM because all these features are related to time series. VECM is used to predict the short- and long-term effects of time series data. Aaron NG presented a mobile application that will forecast the price of houses in London [2]. He created an application such that all computation work will execute on the server-side and all visualization will be displayed on the client-side. For producing the model, he had used gaussian process regression. Also, he had used the linear prior mean function in the model to forecast future house prices which will be more accurate. Yashraj Guard et al. [3] discussed hedonic pricing. Hedonic pricing suggests that the price of an object depends on both internal and external factors. For house marketing, factors such as neighborhood, distance from the city center, schools, etc. are considered while estimating the price of houses. They used Random Forest and Lasso regression to generate models to predict house prices. P. Durganjali and M. Vani Pujitha, proposed a house resale price prediction using classification algorithms [4]. In this paper, the resale price prediction of the house is done using different classification algorithms like Logistic regression and Naive Bayes is used and I will be using the Decision Tree, Random Forest, KNN, Lasso, and Linear Regression techniques. Models generated using the Logistic and Naive Bayes method had an efficiency of 81.5% and 86.5% respectively. Various parameters are considered while predicting house prices like the size of the room, neighborhood, and economic variables. We will consider the RSME value which is a performance metric to determine the best model we can use to predict house prices for a particular dataset. A lot of researches have been conducted at the international level to predict the price of real estate. According to the reference [5] study, they analyzed the long-run integration relationship between equity and real estate prices in 30 developed and emerging economies divided into four subpanels according to income levels and financial market structure. The study presented in reference [6] presented an innovative method called “historical market price”, which used mathematics, statistical database founded algorithms for valuation. They use input data from the databases that gather, analyze, and evaluates data connected with real estate market development. They not only considered the past transactions of rent or flats but also considered the structural data such as the condition of the houses, locality, and various other factors for building their application. According to the article [7], they studied approaches to and options for identifying disequilibrium asset price movements which also means that asset value shifting away from its fundamental-based value. Rohan Bafna et al. [8] discussed the prediction of housing price as an art. They almost discussed all statistical learning which can help to predict the house price. They discussed how prices of houses increase due to increasing purchasing power which led to an increase in demand for the property. They also discussed how machine learning methods such as multilinear regression, AHP, ARIMA, ANN, and Fuzzy logic techniques are used for the prediction of houses. Hujia Yu, Jiafu Wu produced a model to predict prices of residential property using lasso regression [9]. Results of the model suggested that living area square feet, the material of the roof, and neighborhood had the greatest statistical significance in predicting a house’s sale price. Ceyhun Ozgur et al. [10] presented the model using multilinear regression to predict house price. Their dataset consisted of the following variables price, size, Region_h, Type_h, yards, bedrooms, garage_h, floors, basement_h, age, and hoa. Their analysis concluded that only the hoa (House owner association) variable was very significant. They included only the hoa variable in their equation. They also concluded that even though the hoa variable had great influence over price, but it would be wrong to neglect other variables in the equation since they have also some extent of influence over the price of the house. So, it is necessary to include variables such as rooms, bedrooms, etc even though they have less significance in our model. R Manjual et al. [11] presented the model to predict the house price of King County. They used a multilinear regression technique to produce the model. They produced multiple models to check which combination of variables are more significant to predict the price of the house. They have also used simple linear regression in which they used square feet as the independent variable. But model produced using simple linear regression had a very high error rate. In one model they used square feet, bedroom, and bathroom as independent variables. This model was the best fit among all other models. They concluded that for the better model they will need to use a combination of these multilinear regression models, but this will bring high bias and if the complexity of the model increases then it increases variances, so they have suggested using lasso and ridge regression to reduce overfitting of the model. We will be using lasso and ridge regression for the King County dataset so it will be interesting to watch how our model performs. Danh Phan implement models to predict Melbourne City House prices using SVM, Polynomial Regression, and Neural Network [12]. He used the stepwise method to select variables that influence price more. According to the result of the stepwise method, Rooms, Distance, Latitude, Longitude, and Type of house are the five variables which influence the price more. He had used the SVM method with stepwise and also with PCA. According to PCA results, Rooms, Bathroom, Car, Distance, Landsize, and property count are the six variables which accounted for 80% of all predictors variance. For polynomial regression, he concluded that degree three was an optimal degree for polynomial regression. The model produced by the neural network has only two hidden layers since this model obtained the least mean squared error. He compared all models based on mean squared error and concluded that PCA and tuned SVM and stepwise and SVM models showed very less mean squared error which suggests that it was highly accurate models among all. But there was a mode produced using PCA and tuned SVM was overfitted. Model produced using polynomial regression performed worst among all models. The neural network model didn’t perform effectively. Ayush Varma et al. [13] presented the model of estimating the price of houses located in Mumbai. They used features such as square feet area, Bedrooms, Bathrooms, Type of Flooring, Lift availability, Parking availability, and Furnishing condition. To include features about locality they had used Google maps API to identify public places to that particular property which can influence the price of the house. They had used Linear regression and Forest regression to build models. These models were fed into neural networks along with boosting regression which helped to achieve accurate results.

III. Data Mining Methodology

The definition for CRISP-DM is a data mining process or methodology that helps you or provides you a blueprint to conduct a data mining project. The expansion of CRISP is the Cross-Industry Standard Process. It provides a roadmap, structures for better and faster results of using data mining. This is non-proprietary, documented, and freely available so that everyone can use it.

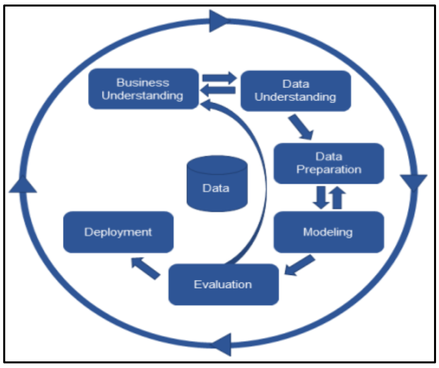
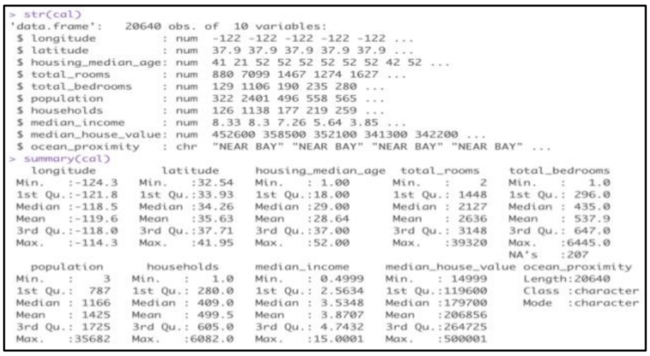


Figure 1. CRISP-DM Process

Business Understanding is where you convert a business objective, or you understand the project from the business perspective and then you convert it to data mining subtasks. So, you convert a business objective into a data mining objective or data mining task where you can technologies for modeling. There are four major tasks that we have to focus on business understanding. It starts with determining your business objective where you focus and understand the true goal of the project and what are some of the important factors that we need to know about the business and then the second is assessing the situation where you list out what are the assumptions that we need to make, a cost-benefit analysis that we need to do and the third is to determine the data mining goals where you set objectives for the business and then the fourth is where you provide a project plan where you set specific outlines and also propose a timeline and you see these are all the tools and techniques that we are going to implement in our project. The next process is data understanding, it starts with the initial collection of the data and where you increase the familiarity with the data, and also you have to create a hypothesis based on the data quality. So, you can provide an initial hypothesis with the hidden information that you have collected. Data understanding has four major tasks. It starts with collecting the data, describing the data, exploring the data, and data quality. So, these tasks are pretty much self-explanatory. In describing the data, you examine the surface of the data, and if you see any problem that you have during acquiring the data. We also have an option to see what formats we can set and how much quantity and quality you have. We just need to check whether the data that is acquired is satisfying our business understanding or not. So, the third task is the exploration of the data, in this task, you create a data exploration report and then list out your first findings on your initial hypothesis that you have. The fourth task is the data quality which is a significant task here. In this step, we find the missing attributes and then we check if there are any blank fields. We just make sure the quality of data must be good at this step. The third step of CRISP is Data preparation. Now we have the data and also verified the quality of the data. In this step, we structured the data as a final dataset. We will be using this dataset for modeling which is the next phase. So, in data preparation, we collect the data, set the data, and fed this set data into the modeling tools. Data preparation also has five subtasks which are collect the data, clean the data, construct the data, integrate the data, and format the data. This process is the most time consuming and it is highly critical. Next process is Modeling, in this step, you propose a various modeling technique and select and apply them and see if you can apply that. We have four major tasks here. We select the model and decide this is the model we are going to build for example it could be a linear or multilinear regression model or whatever techniques that we have. After this, we are going to test the model, we are going to generate a test, and see the test quality and we will see if there is an empirical test. The third task is to create a model and the fourth task is where you assess the model. In assessing the model one needs to work with a business analyst and give a business output. The next process is Evaluation, in this step we work with our business objective and then we come up with the business evaluation and then we come up with process reviewing and then will see if there anything to determine our next steps. Here we summarized our whole result and then we give it as business criteria. The sixth and final phase of CRISP-DM is deployment. Deploying is where we actually present the report or decide to carry the project to the next level, or we take and carry forward to the business steps. There are four major tasks in the deployment process. The first is to plan the deployment, the second task is to monitor, if it is a day-to-day activity, we need to monitor it or maintenance on it. The third task is the final report that we are going to submit to the actual business and then we review the project. Now if we observe Figure 1, we can see that process can be altered at the evaluation phase and it can go back to business understating process which suggests that we may require multiple iterations, we may evaluate and go back to business understanding process and can set the objective again and go through all the process one more time. We will now see how we apply CRISP-DM methodology in our project in detail. We will discuss this methodology for each of our datasets separately.

A. *California Housing Market:*

In this study, our objective is to predict the Median House Value. We will be following the CRISP-DM approach for this study. So, to predict median house value is our business objective. According to the CRISP-DM approach, the next process is to understand the data. I have downloaded this dataset from Kaggle, so we are not dealing with the data collection process for this study. Now we will start with the data exploration.



```
> str(cal)
'data.frame':	20640 obs. of  10 variables:
 $ longitude         : num  -122 -122 -122 -122 -122 ...
 $ latitude          : num  37.9 37.9 37.9 37.9 37.9 ...
 $ housing_median_age: num  41 21 52 52 52 52 52 52 42 52 ...
 $ total_rooms       : num  880 7099 1467 1274 1627 ...
 $ total_bedrooms    : num  129 1106 190 235 280 ...
 $ population        : num  322 2401 496 558 565 ...
 $ households        : num  126 1138 177 219 259 ...
 $ median_income     : num  8.33 8.3 7.26 5.64 3.85 ...
 $ median_house_value: num  452600 358500 352100 341300 342200 ...
 $ ocean_proximity   : chr  "NEAR BAY" "NEAR BAY" "NEAR BAY" "NEAR BAY" ...
> summary(cal)
   longitude         latitude     housing_median_age  total_rooms    total_bedrooms  
 Min.   :-124.3   Min.   :32.54   Min.   : 1.00      Min.   :    2   Min.   :   1.0  
 1st Qu.:-121.8   1st Qu.:33.93   1st Qu.:18.00      1st Qu.: 1448   1st Qu.: 296.0  
 Median :-118.5   Median :34.26   Median :29.00      Median : 2127   Median : 435.0  
 Mean   :-119.6   Mean   :35.63   Mean   :28.64      Mean   : 2636   Mean   : 537.9  
 3rd Qu.:-118.0   3rd Qu.:37.71   3rd Qu.:37.00      3rd Qu.: 3148   3rd Qu.: 647.0  
 Max.   :-114.3   Max.   :41.95   Max.   :52.00      Max.   :39320   Max.   :6445.0  
                                                                     NA's   :207     
   population      households     median_income     median_house_value ocean_proximity   
 Min.   :    3   Min.   :   1.0   Min.   : 0.4999   Min.   : 14999     Length:20640      
 1st Qu.:  787   1st Qu.: 280.0   1st Qu.: 2.5634   1st Qu.:119600     Class :character  
 Median : 1166   Median : 409.0   Median : 3.5348   Median :179700     Mode  :character  
 Mean   : 1425   Mean   : 499.5   Mean   : 3.8707   Mean   :206856                       
 3rd Qu.: 1725   3rd Qu.: 605.0   3rd Qu.: 4.7432   3rd Qu.:264725                       
 Max.   :35682   Max.   :6082.0   Max.   :15.0001   Max.   :500001                       
```

Figure 2. Summary of California Housing Market Dataset

Figure 2 gives an idea about the dataset. We can have up to nine independent variables, but we need to check many things before this. Now we will check the format of data and also check if there are null values present in the dataset. According to Figure 3, ocean_proximimty is the only variable whose data type is a string or character. We will convert this variable type to factor first and the integer. We can also observe that 'total_bedrooms' have 207 NA values. We have a various method to fill NA values like taking the mean or median of total_bedrooms and fill this value in place of NA's. But for this particular dataset, we can use another approach. Ideally, total_rooms and total_bedrooms should be linear. We will check if such a linear relationship exists in our dataset or not.

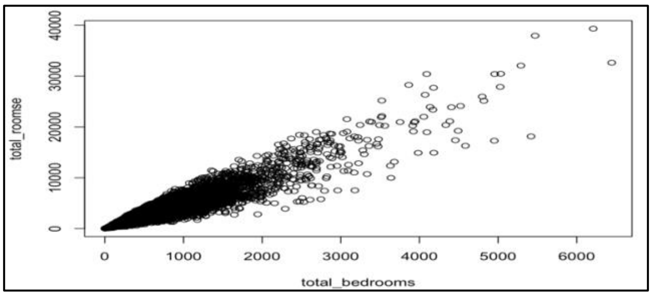


Figure 3. Linear Relation between total_rooms and total_bedrooms

Figure 4 shows that there is a linear relation between total_rooms and total_bedrooms. We will produce a linear regression model using these two variables and will then predict the total_bedrooms values with the help of this model. We will replace NA's in the total_bedrooms with these predicted values. From Figure 1 we can see that data is structured area-wise or sector-wise, like in the first row we get the information about the number of rooms and bedrooms of 126 households. Instead of total rooms and bedrooms, we can have an average number of rooms and bedrooms available in that particular area. So, we will transform total_rooms and total_bedrooms to avg_rooms and avg_bedrooms based on households. Now we will explore data visually. We will observe the frequency distribution of all variables.

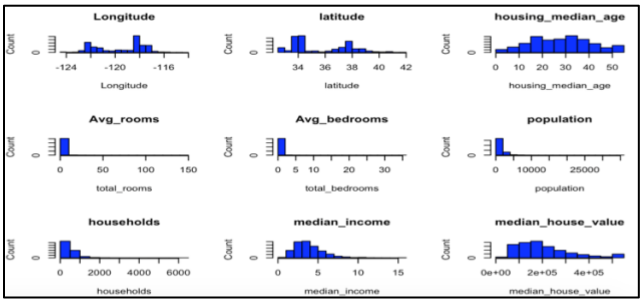


Figure 4. Frequency plots of all variables.

From Figure 5, we can conclude that avg_rooms, avg_bedrooms, population, households have more skewness. Also, median_income and median_house_value can have a uniform spread. So, we will apply log transformation on these respective variables and will see if how they are distributed.

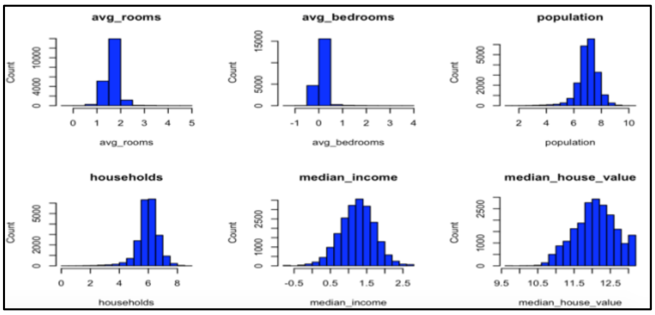


Figure 5. Log Transformation

After applying log transformation all variables now seem to be normally distributed. Now we will check if any outliers are present in the data. Figure 6 shows the boxplot of all independent variables.

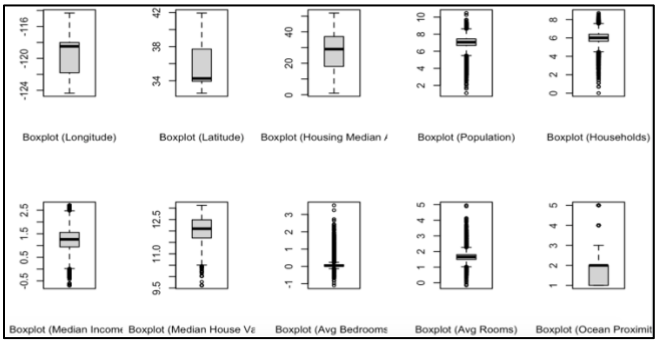


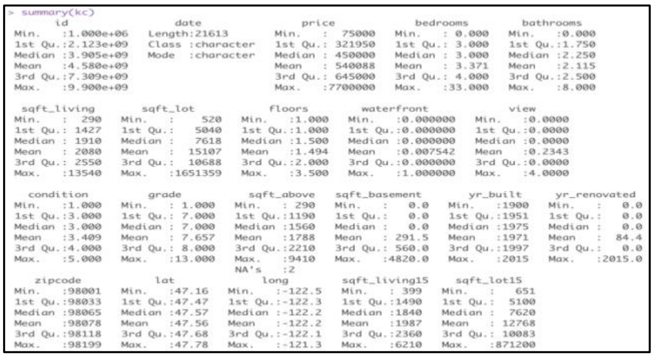
Figure 6. Boxplots of independent variables.

There are too many outliers in the dataset. We will try to remove the outliers, but we will also need to take care of the dataset size and it should not change significantly as it can also lead to overfitting. Our original data size was 20640 observations and after cleaning the data, there were 16197 observations. I believe we can build the model on this clean dataset. The cleaned dataset contains only a few outliers.

B. King County Housing:

We will follow the CRISP-DM approach for this study also. Since we are using the dataset from Kaggle we can directly start the process of understanding the data. Our business objective is to predict the price of a house in King

County. We will proceed now with the exploration of the data.



```
> summary(kc)
       id                  date               price            bedrooms        bathrooms
 Min.   :1.000e+06   Length:21613       Min.   :  75000   Min.   : 0.000   Min.   :0.000
 1st Qu.:2.123e+09   Class :character   1st Qu.: 321950   1st Qu.: 3.000   1st Qu.:1.750
 Median :3.905e+09   Mode  :character   Median : 450000   Median : 3.000   Median :2.250
 Mean   :4.580e+09                      Mean   : 540088   Mean   : 3.371   Mean   :2.115
 3rd Qu.:7.309e+09                      3rd Qu.: 645000   3rd Qu.: 4.000   3rd Qu.:2.500
 Max.   :9.900e+09                      Max.   :7700000   Max.   :33.000   Max.   :8.000

  sqft_living       sqft_lot           floors        waterfront            view
 Min.   :  290   Min.   :    520   Min.   :1.000   Min.   :0.000000   Min.   :0.0000
 1st Qu.: 1427   1st Qu.:   5040   1st Qu.:1.000   1st Qu.:0.000000   1st Qu.:0.0000
 Median : 1910   Median :   7618   Median :1.500   Median :0.000000   Median :0.0000
 Mean   : 2080   Mean   :  15107   Mean   :1.494   Mean   :0.007542   Mean   :0.2343
 3rd Qu.: 2550   3rd Qu.:  10688   3rd Qu.:2.000   3rd Qu.:0.000000   3rd Qu.:0.0000
 Max.   :13540   Max.   :1651359   Max.   :3.500   Max.   :1.000000   Max.   :4.0000

   condition         grade          sqft_above   sqft_basement       yr_built     yr_renovated
 Min.   :1.000   Min.   : 1.000   Min.   : 290   Min.   :   0.0   Min.   :1900   Min.   :   0.0
 1st Qu.:3.000   1st Qu.: 7.000   1st Qu.:1190   1st Qu.:   0.0   1st Qu.:1951   1st Qu.:   0.0
 Median :3.000   Median : 7.000   Median :1560   Median :   0.0   Median :1975   Median :   0.0
 Mean   :3.409   Mean   : 7.657   Mean   :1788   Mean   : 291.5   Mean   :1971   Mean   :  84.4
 3rd Qu.:4.000   3rd Qu.: 8.000   3rd Qu.:2210   3rd Qu.: 560.0   3rd Qu.:1997   3rd Qu.:   0.0
 Max.   :5.000   Max.   :13.000   Max.   :9410   Max.   :4820.0   Max.   :2015   Max.   :2015.0
                                  NA's   :2
    zipcode           lat             long         sqft_living15    sqft_lot15
 Min.   :98001   Min.   :47.16   Min.   :-122.5   Min.   : 399   Min.   :   651
 1st Qu.:98033   1st Qu.:47.47   1st Qu.:-122.3   1st Qu.:1490   1st Qu.:  5100
 Median :98065   Median :47.57   Median :-122.2   Median :1840   Median :  7620
 Mean   :98078   Mean   :47.56   Mean   :-122.2   Mean   :1987   Mean   : 12768
 3rd Qu.:98118   3rd Qu.:47.68   3rd Qu.:-122.1   3rd Qu.:2360   3rd Qu.: 10083
 Max.   :98199   Max.   :47.78   Max.   :-121.3   Max.   :6210   Max.   :871200
```

Figure 7. Summary of King County Housing Dataset.

In this study "price" is our dependent variable and all other variables can be independent variables. From Figure 7 we can get an overview of the dataset and we can conclude that there is a date column that has a character data type. We will need to clean the date column because dates in the date column are appended by an unwanted string. Now our date column is cleaned, but we cannot keep dates in our data as it would not help in our analysis. We can keep only months instead of dates as it would suggest that if price changes based on months or not. So, we will clean the data accordingly and then rename the date column as a month in the dataset. We can also observe there are two NA's in sqft_above. We will simply remove these NA's from our dataset. Now we will try to explore the data visually.

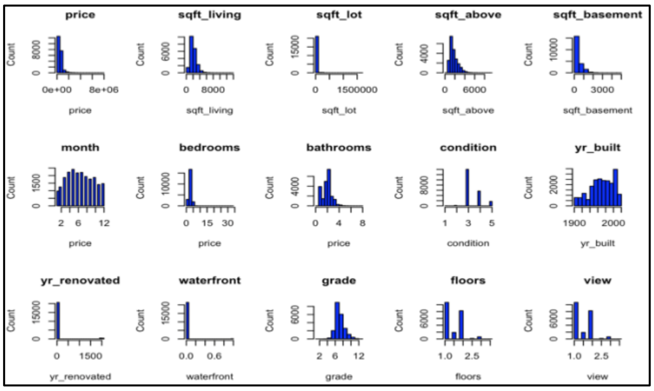


Figure 8. Frequency plots of all variables.

From Figure 8 we can see that price, sqft_living, sqft_lot, sqft_above, and sqft_basement are skewed. We can apply log transformation to reduce this skewness. Now we will check the association of each variable with the price and then we can decide which variables can be our independent variables.

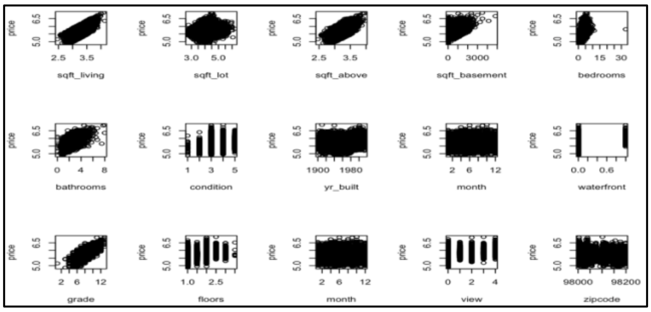


Figure 9. Plots of all variables against price.

Based on Figure 9, we can say that sqft_living, sqft_lot, sqft_above, bedrooms, and grade have a good linear relationship with the price. sqft_basement and bathrooms also have a somewhat linear relationship with the price. condition plot suggests that houses with good condition scores have high prices. Similarly, the waterfront plot suggests that houses with waterfront have higher price comparatively. The price of the house also increases when the number of floors is more and the view score of houses is high. I have also checked the correlation of each variable with the other. From the correlation result, I have decided to remove a few variables from the dataset. We will need to remove sqft_aobve since it is highly correlated with sqft_living. Also, we will need to remove sqft_living15 and sqft_lot15 because these variables are correlated with sqft_living and sqft_lot. Along with these variables, we will remove id since it is irrelevant in our analysis. yr_renovated does not have a great association with the price so we will drop this variable. Since we have zipcode data we will remove latitude and longitude data. Now we can observe the boxplots of our updated dataset.

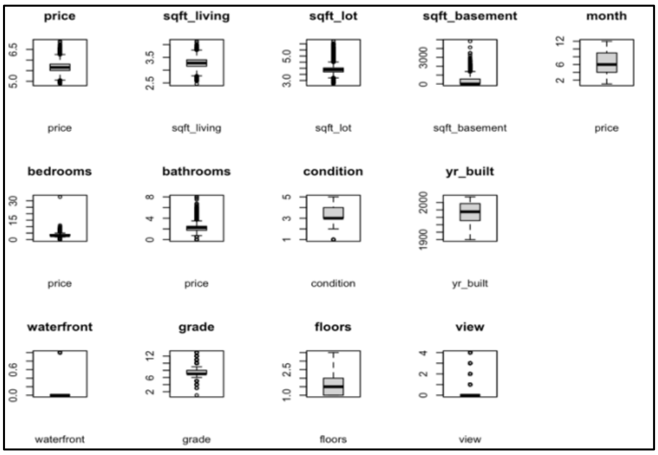
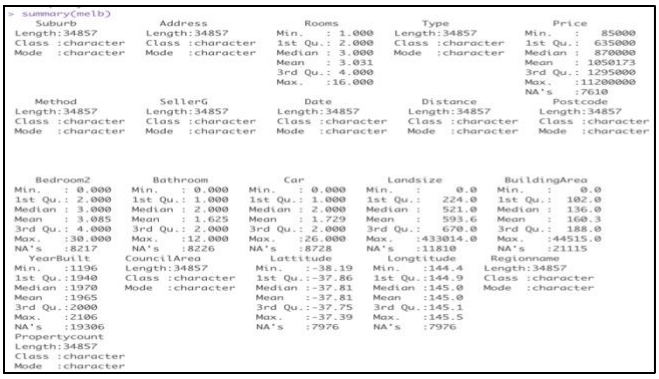


Figure 10. Boxplots

Figure 10 suggests that there is a need for data cleaning. We will try to remove as many outliers as possible. Our dataset had 21,613 observations before cleaning and after cleaning we have 17867 observations. I have tried to remove as many outliers as possible and if I had tried to remove all outliers from the dataset then the dataset would have contained only 6700 observations. There would be a risk of overfitting of the model if we would have removed all the outliers. Now our data is ready to be fed into the model

C. Melbourne Housing Dataset:

Our business objective for this study is to predict the price of the houses in Melbourne city. As per the CRISP-DM approach, our next step is to explore the data. In the last two studies, we used a summary to explore the data first, and then we used to observe data visually. Based on our observations we cleaned out the dataset and made them ready to feed in the model. In this study due to more variables, I have divided variables into three types. There are Location-based, home-based, and seller-based variables in the dataset. We will observe and clean the data according to these features.



```
> summary(melb)
    Suburb             Address              Rooms              Type               Price         
 Length:34857       Length:34857       Min.   : 1.000   Length:34857       Min.   :   85000  
 Class :character   Class :character   1st Qu.: 2.000   Class :character   1st Qu.:  635000  
 Mode  :character   Mode  :character   Median : 3.000   Mode  :character   Median :  870000  
                                       Mean   : 3.031                      Mean   : 1050173  
                                       3rd Qu.: 4.000                      3rd Qu.: 1295000  
                                       Max.   :16.000                      Max.   :11200000  
                                                                           NA's   :7610      
    Method             SellerG              Date             Distance           Postcode        
 Length:34857       Length:34857       Length:34857       Length:34857       Length:34857      
 Class :character   Class :character   Class :character   Class :character   Class :character  
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character  



    Bedroom2         Bathroom           Car            Landsize         BuildingArea    
 Min.   : 0.000   Min.   : 0.000   Min.   : 0.000   Min.   :     0.0   Min.   :    0.0  
 1st Qu.: 2.000   1st Qu.: 1.000   1st Qu.: 1.000   1st Qu.:   224.0   1st Qu.:  102.0  
 Median : 3.000   Median : 2.000   Median : 2.000   Median :   521.0   Median :  136.0  
 Mean   : 3.085   Mean   : 1.625   Mean   : 1.729   Mean   :   593.6   Mean   :  160.3  
 3rd Qu.: 4.000   3rd Qu.: 2.000   3rd Qu.: 2.000   3rd Qu.:   670.0   3rd Qu.:  188.0  
 Max.   :30.000   Max.   :12.000   Max.   :26.000   Max.   :433014.0   Max.   :44515.0  
 NA's   :8217     NA's   :8226     NA's   :8728     NA's   :11810      NA's   :21115    
   YearBuilt     CouncilArea          Lattitude        Longtitude      Regionname       
 Min.   :1196   Length:34857       Min.   :-38.19   Min.   :144.4   Length:34857      
 1st Qu.:1940   Class :character   1st Qu.:-37.86   1st Qu.:144.9   Class :character  
 Median :1970   Mode  :character   Median :-37.81   Median :145.0   Mode  :character  
 Mean   :1965                      Mean   :-37.81   Mean   :145.0                     
 3rd Qu.:2000                      3rd Qu.:-37.75   3rd Qu.:145.1                     
 Max.   :2106                      Max.   :-37.39   Max.   :145.5                     
 NA's   :19306                     NA's   :7976     NA's   :7976                      
 Propertycount     
 Length:34857      
 Class :character  
 Mode  :character  
```

Figure 11. Summary of Melbourne Housing Dataset

Based on initial observation of the dataset, we can say there are too many NA's in the dataset and also there are few variables whose data type is character. We will need to modify our data accordingly. I have changed the data type of Distance, Regionname, CouncilArea, Method to numeric. Also, I have changed the datatype of Postcode, PropertyCount, and Type to numeric. I have also changed the datatype of SellerG to factor. Our dependent variable Price has NA values so we will clear those NA's from our dataset. Now, we will start exploring the data feature-wise. So, we will explore the variables which come under the category of location. Suburbs, Address, PostCode, CouncilArea, Latitude, Longitude Regionname, Propertycount, and Type are location-based variables. Address is unique to each row and Suburbs have large unique values. So, I have decided to remove Address, Suburbs, Latitude, and Longitude because instead of these columns we can use Regionname. We will explore the location-based variables visually now.

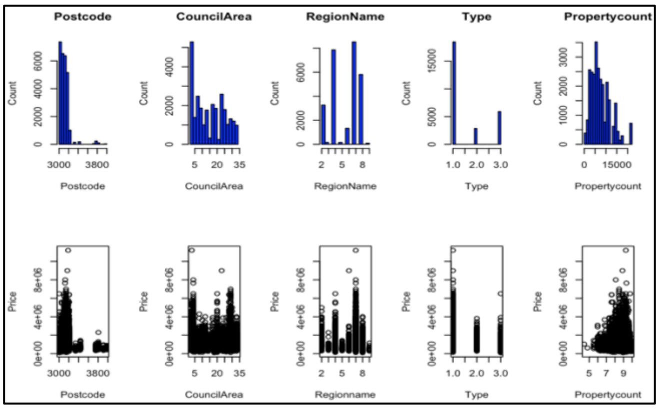


Figure 12. Frequency plot of location variables and plot of location variables against price variable.

Based on Figure 12 observation, we can apply log transformation on the Propertycount variable to get the normal curve. Now we will explore the home-based variables. Rooms, Bedroom2, Bathroom, Landsize, Distance, Car, BuildingArea, and YearBuilt are the variables that come under the home category. From Figure 11 we can observe that BuildingArea and YearBuilt have a larger number of NA values, so I have decided to drop these variables from our dataset. The Bathroom also has a significant number of NA's but first, we will check its relationship Rooms. After plotting the graph between Bathroom and Rooms we came to know that there is a linear relationship among. Them. So, we will make a linear regression model using Bathroom and Rooms and then we will replace the NA's of Bathroom with the predicted values of Bathroom. Similarly, for Car, there was a linear relationship with Rooms. I have used the predicted values of the car to replace NA's of Car. For NA's of the Landsize variable, we will use the median value and substitute median value in place of NA's. Bedrooms2 variable is highly correlated with Rooms variable so I have decided to drop the Bedroom2 variable. Now we will explore the House variables visually.

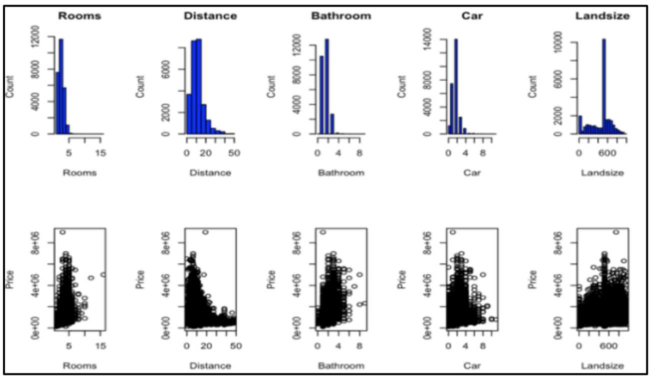


Figure 13. Frequency plot of house variables and plot of house variables against price variable.

Based on Figure 13 we can say that price of the house increases when the number of Rooms, Bathroom, and Car increases. When distance increases the price of the house is decreased. Now we will explore the Seller related variables. SellerG, Method, and Data are the variables that come under the category of Seller. SellerG has 346 unique values so I have decided to drop this variable from the dataset. We will need to modify the Date columns slightly. We will extract only month data from the Date column. Also, we have converted the Method variable to numeric. Now we visualize the seller-based variables.

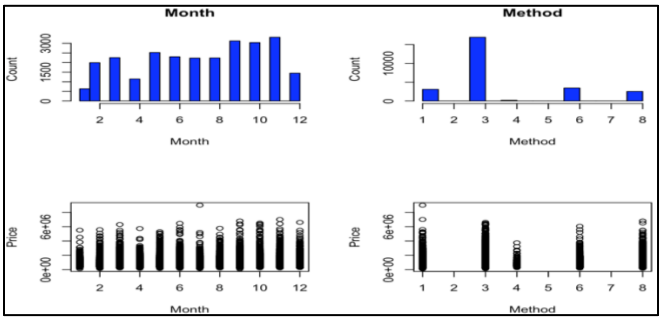


Figure 14. Frequency plot of seller variables and plot of house variables against price variable.

We will now check the outliers in our dataset using boxplots.

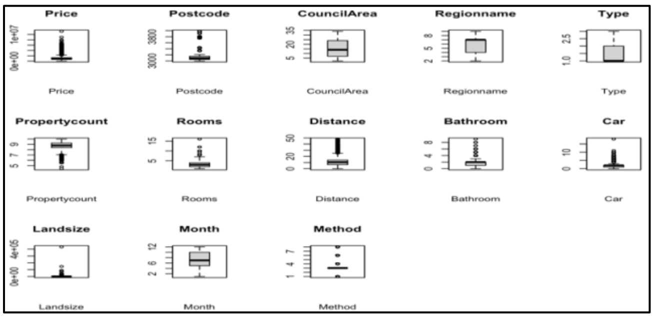


Figure 15. Boxplots of all variables.

Based on Figure 15, we can say that Price, PropertyCount, Rooms, Distance, Bathroom, Car, and Landsize have outliers and we will remove these outliers from our data. After removing these outliers, I believe our dataset is ready to build the model.

IV. EVALUATION

A. *California Housing (Multilinear Regression):*

Multiple linear regression analysis follows a few assumptions. The first assumption is multivariate normality which suggests that residuals are normally distributed. The second assumption is no multicollinearity which suggests that all independent variables must not be highly correlated with each other. This can be tested using the correlation matrix and also by VIF. The third assumption is Homoscedasticity which suggests that variance of error terms should be equal across each independent variable. We have cleaned our data and now we can separate data into training and testing. I have used 70 percent of data for training and 30 percent for testing. Before this separation of data, we can check the correlation of variables with each other.

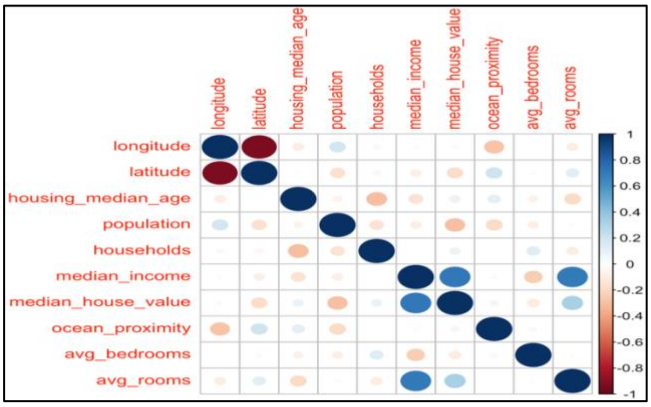
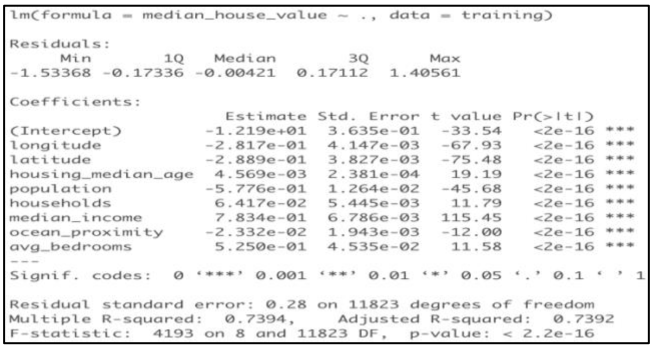


Figure 16. Correlation plot of California Housing Dataset.

Based on Figure 16, we can conclude that avg_rooms and median_income are highly correlated, so I have decided to drop avg_rooms because we have avg_bedrooms in our dataset. Now there is no multicollinearity between independent variables. Now we can feed training data to linear regression function to produce the model. We will now interpret the result of our model.



```
lm(formula = median_house_value ~ ., data = training)

Residuals:
     Min       1Q   Median       3Q      Max
-1.53368 -0.17336 -0.00421  0.17112  1.40561

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)       -1.219e+01  3.635e-01  -33.54   <2e-16 ***
longitude         -2.817e-01  4.147e-03  -67.93   <2e-16 ***
latitude          -2.889e-01  3.827e-03  -75.48   <2e-16 ***
housing_median_age 4.569e-03  2.381e-04   19.19   <2e-16 ***
population        -5.776e-01  1.264e-02  -45.68   <2e-16 ***
households         6.417e-02  5.445e-03   11.79   <2e-16 ***
median_income      7.834e-01  6.786e-03  115.45   <2e-16 ***
ocean_proximity   -2.332e-02  1.943e-03  -12.00   <2e-16 ***
avg_bedrooms       5.250e-01  4.535e-02   11.58   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.28 on 11823 degrees of freedom
Multiple R-squared:  0.7394,    Adjusted R-squared:  0.7392
F-statistic:  4193 on 8 and 11823 DF,  p-value: < 2.2e-16
```

Figure 17. Summary of the model using multilinear regression.

The adjusted R-squared value is 0.7392 which suggests that 73.92% of the variation in the output variables are explained by the input variables. All independent variables coefficients are significant, and we can use the coefficients of all independent variables in the linear equation. We will now explore the model visually.

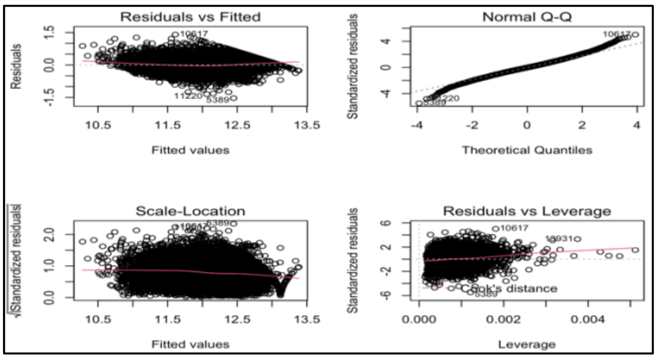


Figure 18. Plots of the model

Based on Figure 18, we can conclude that variation of residuals looks constant. Residuals vs Fitted values graph is used to detect non-linearity, unequal error variances, and outliers. Since the red line in this graph does not deviate too much from the dotted ideal line so we can say model that the model fits well. A normal Q-Q plot is used to check if residuals are normally distributed. In our model, residuals are almost normally distributed. Scale Location graph is used to check if residuals are spread equally along with the range of predictors. It also checks the homoscedasticity. Residuals vs Leverage are used to check the outliers. Since we removed the outliers during the cleaning of the data, so we are now unable to see the dotted line in this graph which indicates cook distance. Data outside the cook distance plotted lines are the outlier. So, by analyzing all graphs and summary we can say that model produced is a good fit model. Our model's prediction error rate is 2.31% which suggests that the model is performing better.

B. *King County Housing (Ridge Regression):*

We have cleaned our dataset and now we can proceed with the data partition step. We will use 70-30 combination for train and test data. We have 15,125 observations in train data and 6,486 observations in test data. We will use the train control function within the caret package to create our custom control parameters. In this function, we will use a method called repeated cv which is known as repeated cross-validation. The second parameter specifies the number of cross-validations. I have used 10-fold cross-validation which means that training data is broken into 10 parts and then the model is made from 9 parts and 1 part is used for error estimation. This is repeated 10 times with different parts used for error estimation. We have also passed VerboseIter equals to true to see what is happening in the background on the terminal. Ridge regression tries to shrink the coefficients, but it keeps all variables in the model. We have called our model a ridge and we will be using a method known as glmnet. This package allows us to fit ridge lasso and elastic regression models. Now we will add a tune grid, in this we will specify alpha=0 which suggests that we will be using ridge regression. For lambda, we can create a sequence. We have started with 0.0001and will go up to 0.1 and length=5 suggest that we have 5 value between this range. Now we can add our custom control in this function. Now we can run our model. Once it completes all the runs it finds that the best value of lambda is 0.00004. This lambda is a hyperparameter and is estimated using cross-validation that we have specified. It is the strength of the penalty on the coefficients. So, as we increase the lambda, we are increasing the penalty and coefficients will shrink. Now we can plot our model.

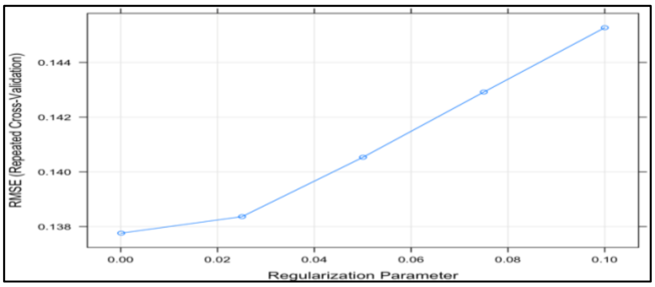


Figure 19. plot of the model

Based on Figure 19, we can say that for higher values for lambda, error increases. Now we will plot log lambda vs coefficients.

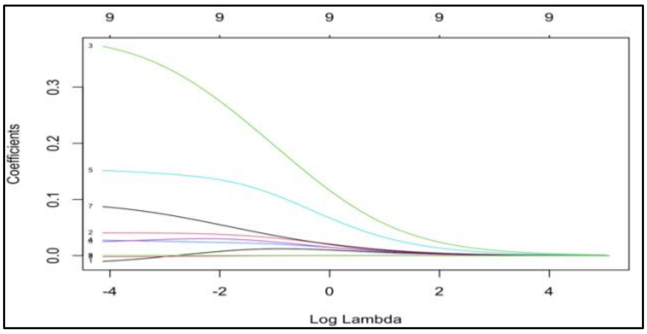


Figure 20. Plot of log lambda vs coefficients.

We can observe that when lambda is increased, coefficients get shrink. At log lambda=4 all coefficients become 0 and we can see at the top of the plot that the number of predictors didn't change. So as log lambda is increased it doesn't make the coefficient of the variables zero which is not contributing that much. Now we will plot fraction deviation against coefficients. This graph explains the variance.

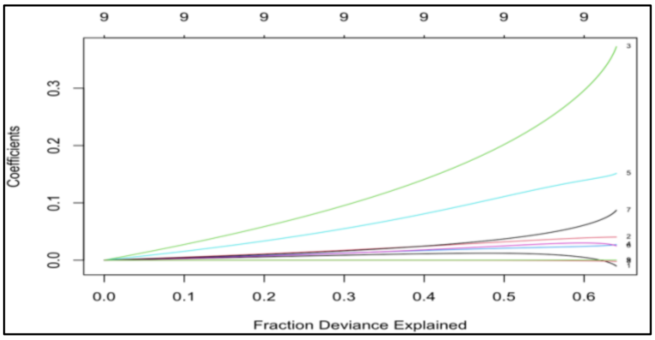


Figure 21. Plot of Fraction Deviance vs Coefficients.

At 20 percent of deviance, we can see a slight growth of the coefficients. At 60 percent deviance coefficients become highly inflated and at this point, we are likely doing overfitting. We have also plotted the variable importance graph which gives an idea about which predictor variable is more important in our model.

C. *King County (Lasso Regression):*

Lasso regression shrinks the coefficients and also performs feature selection. If there is a highly correlated variable that is causing multicollinearity then lasso regression select one variable among them. We have called our model as lasso and parameters in the function will be the same as of ridge expect in this case our alpha will now become 1 which suggests that we are using lasso regression. Also, we have changed the sequence of lambda, for lasso it will start from 0.0001 and ended at 0.01. We have got the same lambda for lasso regression which is 0.00001. Also, the RMSE plot of the lasso is the same as that of ridge regression. To see which variable is performing better, we can analyze the plot of log lambda vs coefficients.

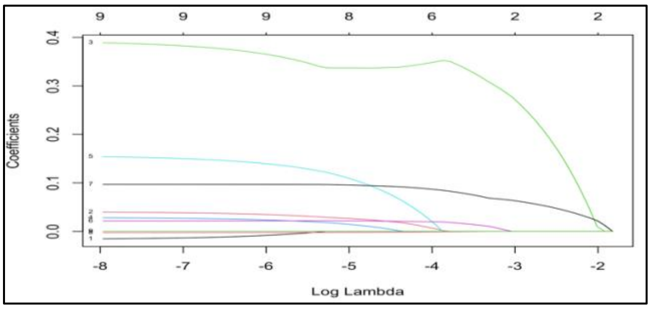


Figure 22. Plot of log lambda vs coefficients.

Based on Figure 22, we can say that bedroom's coefficient is performing much better than any other variable. Also, you can observe that as lambda increases the number of predictors got decreased. Now we will analyze the fraction deviation plot.

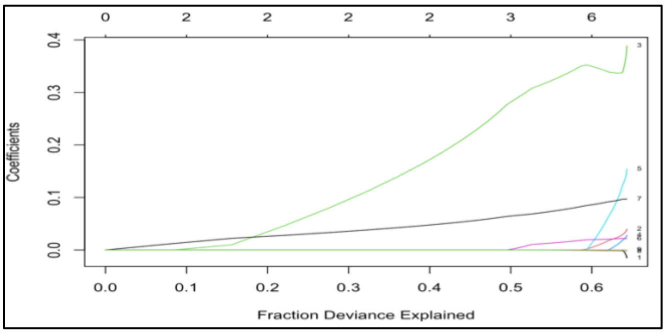


Figure 23. Plot of Fraction Deviance vs Coefficients.

Based on Figure 23, we can say that 60 percent of the variability of the model is explained by 6 variables. Due to lasso regression, we can be able to explain 60 percent of variability with help of 6 variables. We have also plotted the variable importance graph which gives an idea about which predictor variable is more important in our lasso model. We have compared both models and the R-squared value of each model was almost equal, lasso regression has a slightly higher R-square value than the ridge. Also, RMSE values using train and test of both models were almost equal.

D. *Melbourne Housing (Decision Tree):*

We have cleaned the dataset and now we can proceed with data partitioning. We will use 70 percent of data as test data and the remaining 30 percent as test data. I have used rpart package for decision tree regression. Now we will pass the train data in the rpart function and run the model. We will now observe the tree using rpart.plot package. We have passed cp=0.0001. cp is known as the complexity parameter. Validation of the tree is done by using the complexity parameter and cross validated error. We got a very complex tree after running this function. We will need to prune the tree and for that, we can use the printcp function which provides optimal pruning based on the cp value. Plotcp() plots a graph between cp and cross-validation relative error. The cp values are plotted against the geometric mean to depict the deviation until the minimum value is reached.

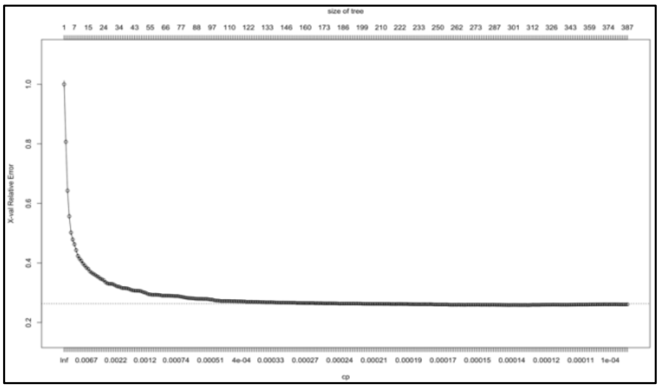


Figure 24. plotcp(melb.tree)

Based on Figure 24, we can say that as the cp value decreases the size of the tree increases, and relative error decreases. But if we keep on decreasing the cp value then our model will get over-fit. By observing Figure 24, we can say that cross-validation relative error becomes constant after cp=0.0004. So, we will use this cp value to prune our tree. Our tree produced is very large, but it will not make the model overfit. We can evaluate the model by observing the R-square value. We got the R-square value of 0.74 which suggests that the model produced is good.

E. Melbourne Housing (Random Forest):

Random forest is nothing but a combination of multiple decision trees. We will use the RandomForest package in R. It has the ability to deal with a large number of features and also helps in feature selection. Now, will pass our train data in the RandomForest function. This function has two parameters, the first parameter is ntree which suggests a number of trees, by default number of tress is 500. The second parameter of this function is mtry and its value for regression is mostly p/3 where p is the independent variable. In our case there are 12 independent variables, so four variables were tried at each split. RandomForest function is very simple to use. We will now run our model and observe the graph of our model.

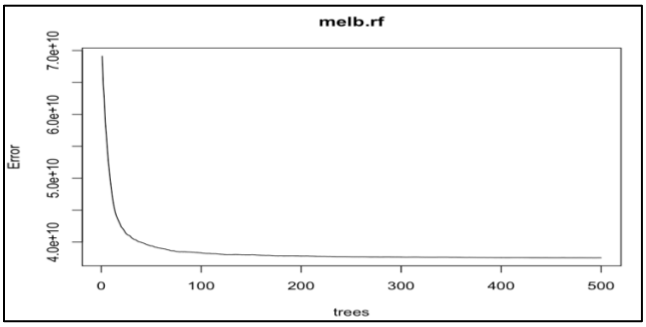


Figure 25. Plot of random forest model

We can see that the number of trees increases, error decreases but an increasing number of trees will make our model overfit. Also, when compared to decision tree output, we can see the R-square value of the random forest model is higher than the decision tree model. We can find out which variables were highly important while building the model.

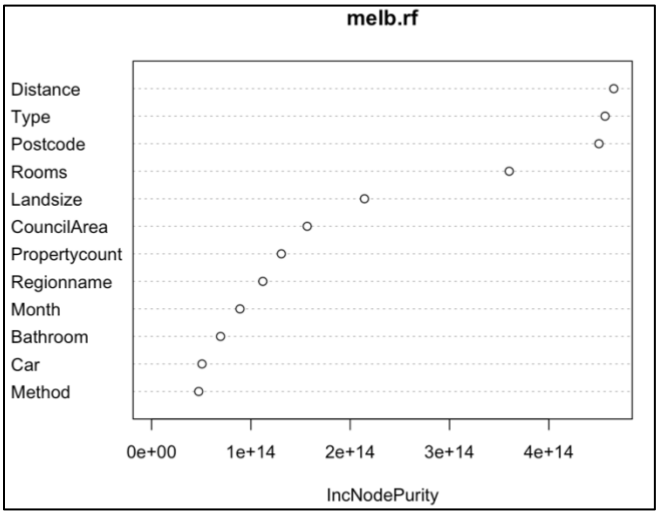


Figure 26. Importance of variables

Based on Figure 26, we can say that Distance, Type, Postcode, and Rooms are the important variables in the model.

V. Conclusion

I have performed a multilinear regression on the California housing dataset. Model performance was good, and the error rate was quite low. I have got this efficiency just because of cleaning the data properly. There were too many outliers in the data. Also, I have checked all the assumptions of multiple linear regression. There was multicollinearity between the avg_rooms and median_income, so I have neglected avg_rooms since we have avg_bedroom data. I have used log transformation to have a normal distribution for a few variables. For the King County housing dataset, there were too many outliers. I have used the ridge and lasso function to produce the models. I have performed various steps to clean the data. I have also checked the lambda values and the result suggested that when the value of lambda was low then the error was low. The best lambda value for both ridge and lasso regression was the same. Ridge regression successfully shrinks all the coefficients at log lambda=4. Lasso regressions shrink all coefficients at log lambda=-2 and also used only 2 variables to determine 60 percent of the variability of the model. The Lasso model performed better than the Ridge model. For the Melbourne Housing dataset, I have used Decision Tree and Random Forest techniques. This dataset required more cleaning effort compared to other datasets. I have removed most of the outliers in the data. While performing the Decision Tree, the complexity parameter is very critical to determine. Using the appropriate cp value, we got a better model that has a better R-square value. Random Forest was the easiest and efficient technique to produce the model. Also, it has a greater R-square value compared to the rest of the other models.

VI. Future work

California Housing model produced was a good model, but we can still further increase the efficiency of the model by using Random Forest. Random forest help for feature selection. KC model produced using lasso regression performed well. Since the housing datasets are large and there are many features in the dataset so here also random forest will be useful because the random forest can perform with a large number of features. Melbourne Housing model performed best among all the models.

References

[1] L. Gattini, & P. Hiebert, Forecasting and assessing euro area house prices through the lens of key fundamentals‖, Working Paper Series, No.124/ October, 2010, 2010.

[2] Aaron NG, Machine Learning for a London House Price Prediction Mobile Application, 2015

[3] YashrajGarud, HemanshuVispute, NayanBisai, Housing Price Prediction using Machine Learning Techniques, IRJET, Volume: 07 Issue: 05, May 2020

[4] P. Durganjali, M. Vani Pujitha, ”House Resale Price Prediction Using Classification Algorithms”, 2019 International Conference on Smart Structure and Systems(ICSSS), Chennai, India, 2019, pp.1-4, DOI:10.1109/ICSSS.2019.8882842.

[5] ČEH ČASNI, A; VIZEK, M. Interactions between real estate and equity markets: An investigation of linkages in developed and emerging countries. In: Finance a Uver - Czech Journal of Economics and Finance. Prague. Volume 64, Issue 2, 2014, p. 100-119. ISSN: 0015-1920.

[6] ARDIELLI, J.; JANASOVA, E. Creation of real property database for determination of capitalization rate of real estate. In: 12th International Multidisciplinary Scientific GeoConference and EXPO, SGEM 2012; Varna; Bulgaria. Volume 4, 2012, p. 877-882.

[7] KOMÁREK, L.; KUBICOVÁ, I. Methods of identification asset price bubbles in the Czech economy. In: Politicka Ekonomie. Prague. Volume 59, Issue 2, 2011, p. 164-183. ISSN: 0032-3233.

[8] Rohan Bafna, Anirudh Dhole, Ankit Jagtap, Asif Kazi, Arbaz Kazi Prediction of Residential Property Prices – A State of the Art, IARJSET, Vol. 5, Issue 3, March 2018.

[9] Hujia Yu and Jiafu Wu (2016), “Real Estate Price Prediction with Regression and Classification”, CS 229 Autumn 2016 Project Final Report.

[10] Ceyhun Ozgur, Ph.D., CPIM , Zachariah Hughes , Grace Rogers , Sufia Parveen Multiple Linear Regression Application in Real Estate Pricing, IJMSI, Vol. 4, Issue 8, October 2016

[11] Manjula, R & Jain, Shubham & Srivastava, Sharad & Kher, Pranav. (2017). Real estate value prediction using multivariate regression models. IOP Conference Series: Materials Science and Engineering. 263. 042098. 10.1088/1757-899X/263/4/042098.

[12] The Danh Phan, "Housing Price Prediction using Machine Learning Algorithms: The Case of Melbourne City Australia", *2018 International Conference on Machine Learning and Data Engineering (iCMLDE)*, December 2018.

[13] Ayush Varma, Abhijit Sarma, Sagar Doshi, and Rohini Nair, "House Price Prediction Using Machine Learning and Neural Networks", *Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*, April 2018.