Prelim Examination - Evaluation Report

Course Code: CPE 019

Course Title: Emerging Technologies 2

Section: CPE32S3

Members:
Cuevas, Christian Jay L
Jimenez, Jerviz MIco

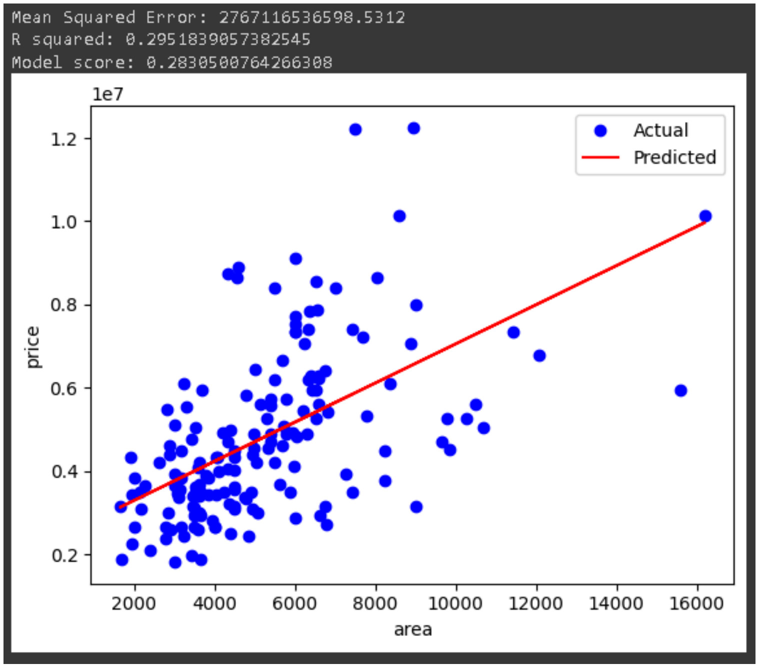
Instructions:

- ➔ Choose any dataset applicable for classification and/or prediction analysis problems
- ➔ Show the application of the following algorithms
    - ◆ Linear Regression
        - Singular LR
        - Multiple LR
        - Polynomial LR
    - ◆ Logistic Regression
    - ◆ Decision Tree
    - ◆ Random Forest
- ➔ Provide Evaluation Reports for all models

Linear Regression

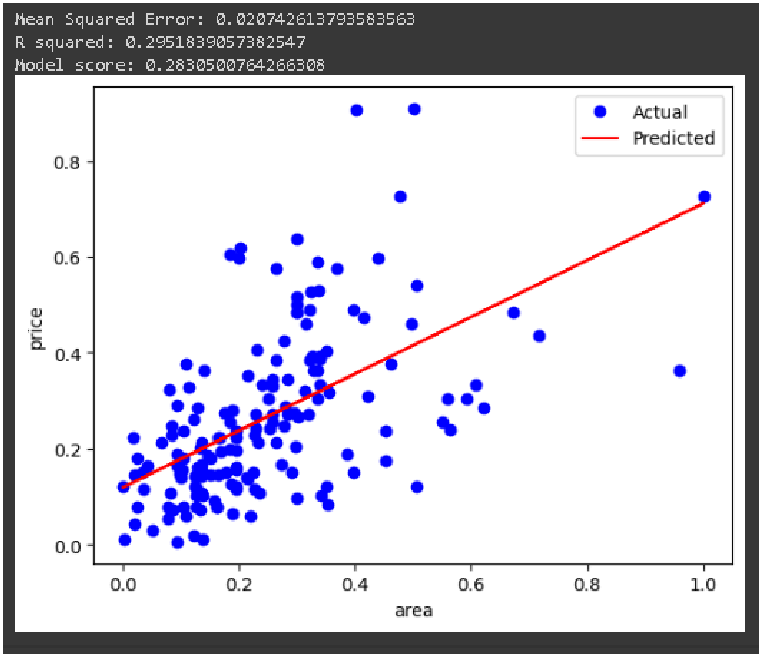
- Singular Linear Regression

(Not Normalized Data)



```
Mean Squared Error: 2767116536598.5312
R squared: 0.2951839057382545
Model score: 0.2830500764266308
```

(Normalized Data)



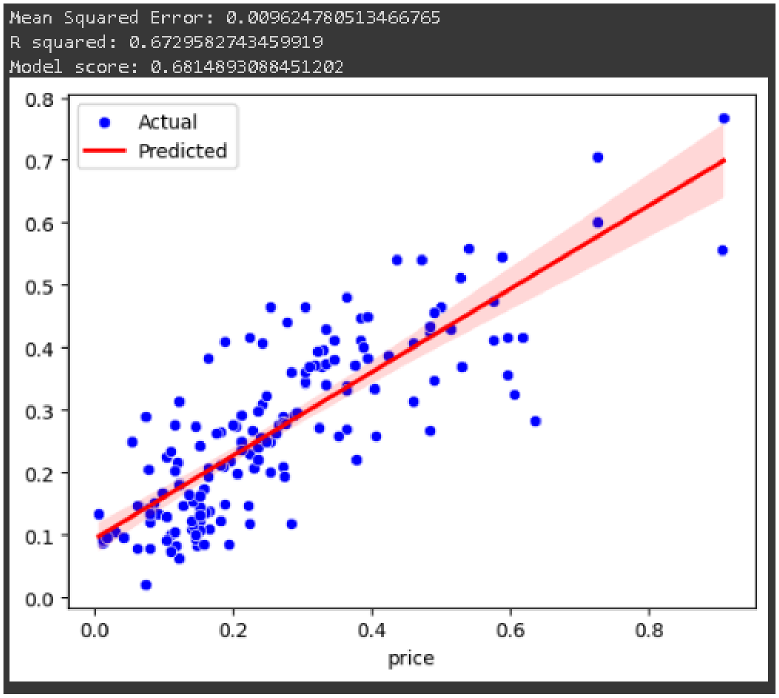
```
Mean Squared Error: 0.020742613793583563
R squared: 0.2951839057382547
Model score: 0.2830500764266308
```

- In this model, we used singular linear regression to predict the value of the price using the predictor variable “area”. The predictor “area” has the highest correlation with the target variable “price” with the value of correlation equal to 0.5359 that can be interpreted as “Moderate Positive Correlation”.
- The model is evaluated using the MSE, R^2, and Model Score. MSE or Mean Squared Error measures the mean of the error but it is squared. The figure below is the formula for the MSE.

$$\sum_{i=1}^{D}(x_i - y_i)^2$$

- This is one of the reasons why MSE can be biased since if your data points range is so high, the value of the error is also high. The value range of our “price” column is between 1,000,000 to 13,000,000 while the “area” column is between 2000 to 16000, this can result in a high value of MSE. This is the reason why we are normalizing the data using MinMaxScaler() so that the data will be plotted in a range from 0 to 1.
- The original value of MSE is around 2 trillion, but by using the MinMaxScaler() to scale all the columns with continuous values, the dataset was normalized and the current value of the MSE is 0.02.
- The coefficient of determination or the R^2 is the measure of the variability and how well your model explains the variability or the outlier in the data. The closer the value of R^2 to 1, the better your model explains the variability in your data.
- In the case of our model, we have low MSE and low R^2. This can be interpreted as having high accuracy but it can’t explain the variability or the outlier in the dataset.
- Overall, this model is not very well-fitted with our dataset, mainly because one variable can’t explain the trend of the target variable “price”.

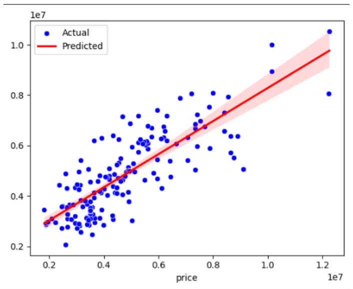
- ## Multiple Linear Regression



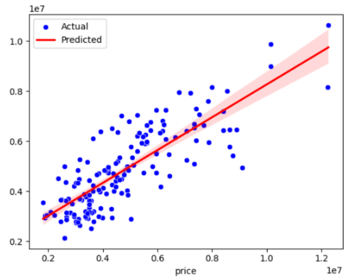
```
Mean Squared Error: 0.009624780513466765
R squared: 0.6729582743459919
Model score: 0.6814893088451202
```

  - Multiple Linear Regression uses multiple predictor variables to predict the values of the target variable.
  - In the Multiple LR graph above, we are using all the available predictor values, to predict the target variable “price”. You can see that the graph of Multiple LR has a certain range of values along the line unlike the Singular LR where it only has a straight line.
  - This model is evaluated using MSE, R^2 and model score. Fundamentally, R^2 and model score are the same, but their application differs in terms of their function.
  - From the value of MSE, we can infer that this model is very accurate since the value of MSE is very close to 0.
  - From the R^2 and the model score, with their value of around 0.68, we can interpret this as a realistic model but it is a poor model since it did not reach the industry standard for 70% - 90%.
  - Overall, this model is well-fitted to our data which can be further improved by data preprocessing, so that it can reach the industry standard. This model is very suitable for our dataset unlike the earlier Singular LR, this is because this model can explain the outlier data of our dataset.

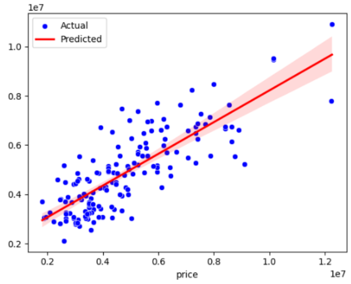
Feature Selection using Backward Elimination In Multiple LR



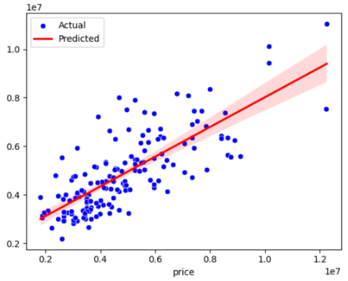
Removed "unfurnished"
Accuracy = 67%



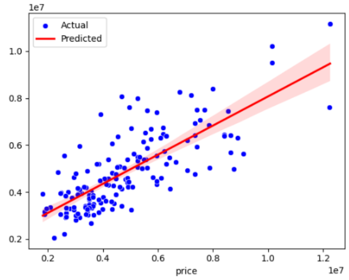
Removed "semi-furnished"
Accuracy = 67%



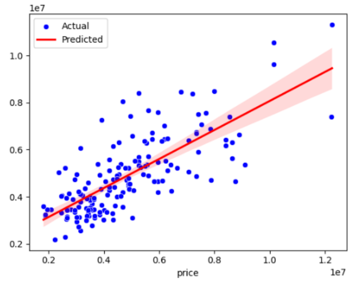
Removed "prefarea"
Accuracy = 65%



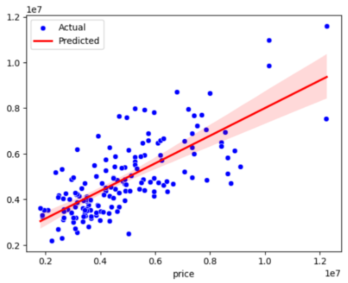
Removed "airconditioning"
Accuracy = 61%



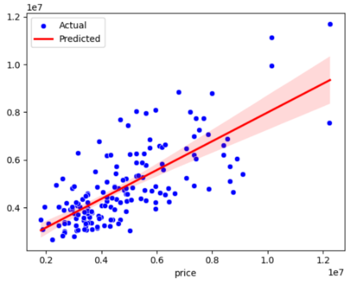
Removed "hotwaterheating"
Accuracy = 61%



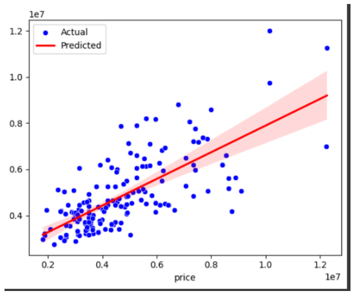
Removed "basement"
Accuracy = 60%



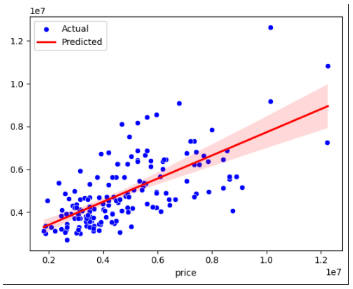
Removed "guestroom"
Accuracy = 58%



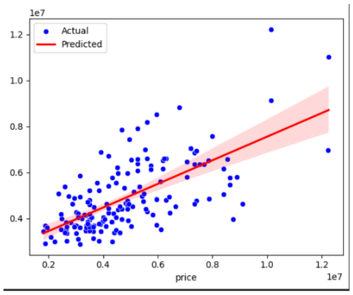
Remove "mainroad"
Accuracy = 56%



Remove "parking"
Accuracy = 55%



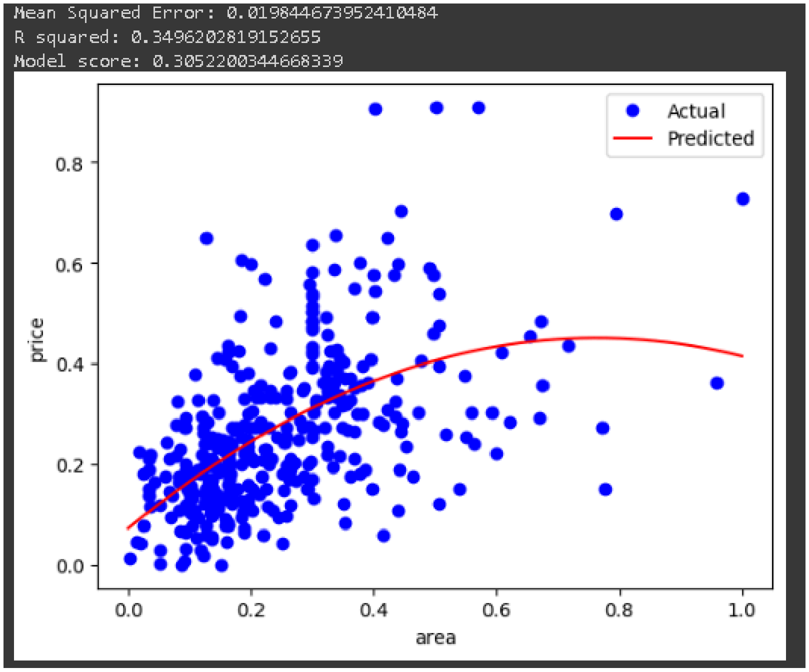
Remove "stories"
Accuracy = 50%



Remove "bedrooms"
Accuracy = 47%

- The figures above are the Multiple LR graph that resulted from the backward elimination that we employed.
- This is done for further evaluation and to understand the dataset more deeply and to observe the weight of the data points to the overall accuracy of the model.
- From what we can observe above, the most significant change in accuracy when it was removed was “airconditioning” and “stories”. Although not all the combinations of columns are tested, we can infer from this that “airconditioning” and “stories” columns have a significant weight towards the prediction of our target variable “price”.
- Another insight that we can gather from this is that some variables, even if removed, has little to no impact to the overall performance of the model. This is important since by removing some predictor variables, we can simplify the model and easily interpret it.
- Overall, the Multiple LR can be a very powerful tool if it is in the right hands. It is very versatile with the predictor variables and it can explain outlier data which can’t be explained by Singular LR or Simple LR. This model is well-fitted with our dataset and with just a little tweaks, it can be considered as a good model.

- Polynomial Linear Regression



```
Mean Squared Error: 0.019844673952410484
R squared: 0.3496202819152655
Model score: 0.3052200344668339
```

- The Polynomial linear regression is just an extension of the Singular LR but it can better accommodate real-life dataset because of its capability to plot curvature or non-linear relationships.
- The Polynomial LR uses a single predictor variable to predict the target variable which is actually the same as the Singular LR. Their difference lies with the fact that we can observe the plotted values in a much more detailed way than Singular LR.
- In the graph above, we can see that the red line goes up until 0.4 and then it

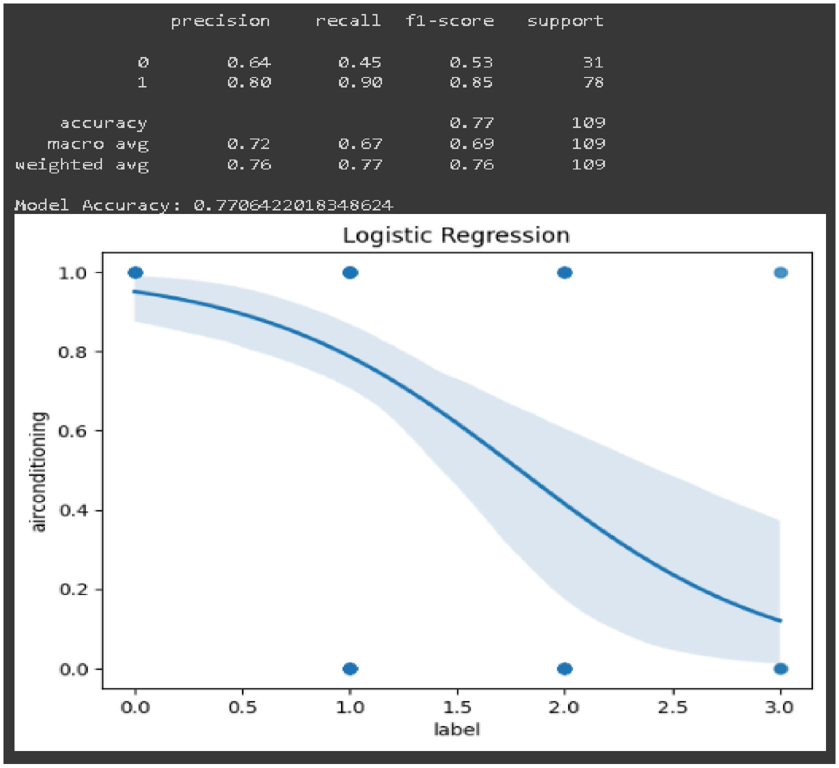
starts to curve a little bit. This is contrary to the Singular LR which goes straight up in a diagonal direction.

- The Polynomial Regression gives us more insights with regards to the other data points of the dataset. We can infer from the graph that although the price goes up as the area goes up, at some point it became stagnant and even went downwards even for just a little bit. This means that area is not the only factor which can affect the target variable price. The price of the house can be affected by many factors such as the different predictor variables present in our dataset.
- This model has a low MSE which means that its predictions are fairly accurate.
- This model has R^2 of 0.34 which means that it is below the standard of accepted machine learning models accuracy. Low R^2 also means that it can’t accurately explain the variance of other data points.

*** Note***

- The following algorithm has a modified dataset from the original. The variables having values such as Yes and No have been changed to 0 and 1 respectively. We also classified the prices into 4 types. The price ranges are as follows:
    1. 1,740,000 - 3,000,000 as Affordable (Label = 0)
    2. 3,000,001 - 6,000,000 as Average (Label = 1)
    3. 6,000,001 - 9,000,000 as Expensive (Label = 2)
    4. 9,000,001 and above as Luxury (Label = 3)

- *Logistic Regression*

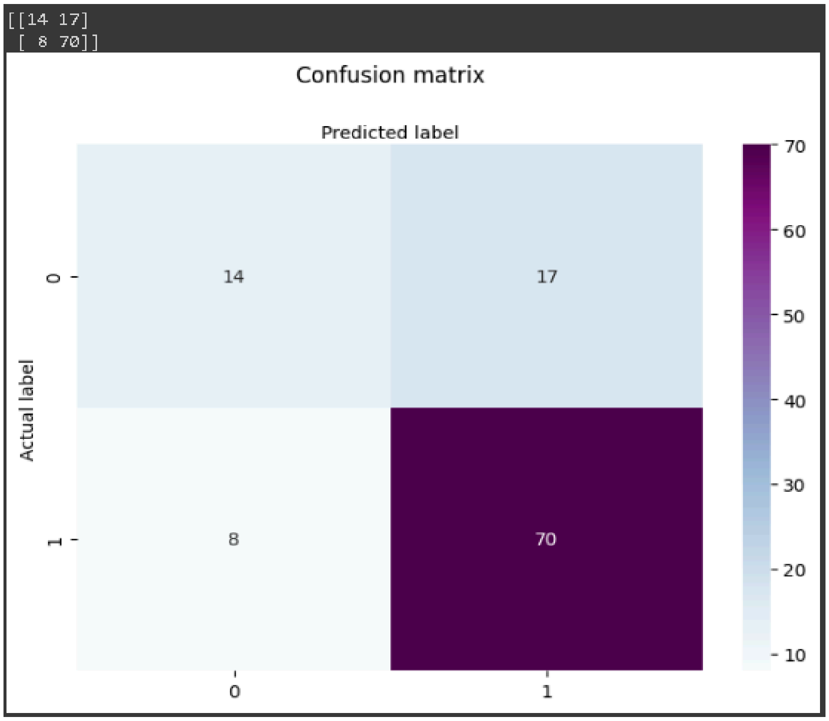


```
              precision    recall  f1-score   support

           0       0.64      0.45      0.53        31
           1       0.80      0.90      0.85        78

    accuracy                           0.77       109
   macro avg       0.72      0.67      0.69       109
weighted avg       0.76      0.77      0.76       109

Model Accuracy: 0.7706422018348624
```

- Logistic regression compares the data values of two variables. It is used to predict the value of the variable based on the other.
- In this visualization of logistic regression, the variables to be compared are if

there is air conditioning on the different price types of a house.

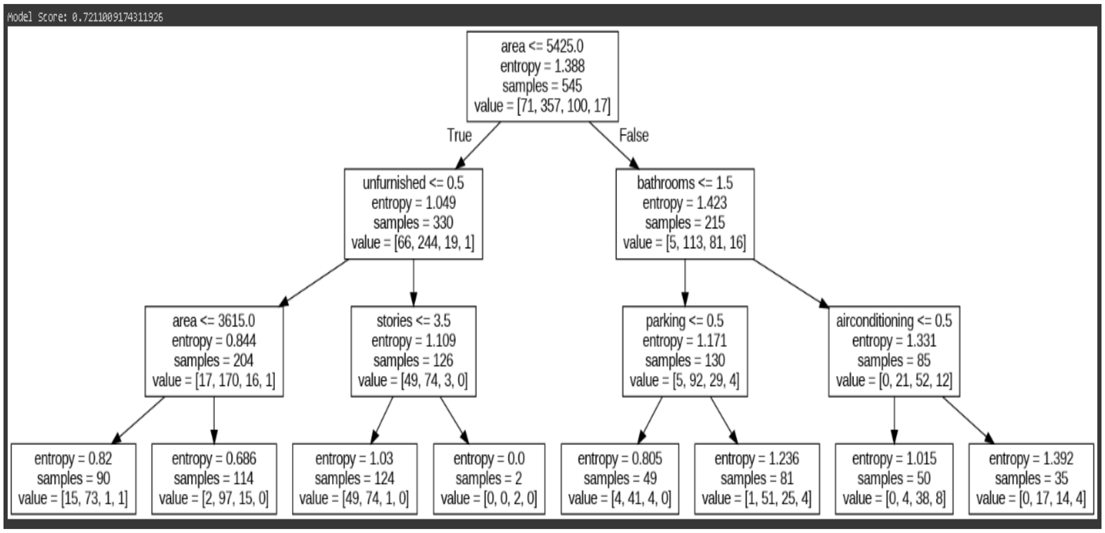
- Let us look into the houses marked as expensive or labeled as 2.0. Based on the data in the dataframe, there are houses which have air conditioning and some do not. Based on the flow of the graph, the line starts to curve towards 0.4. This means that 40% of the houses may have air conditioning.



```
[[14 17]
 [ 8 70]]
```

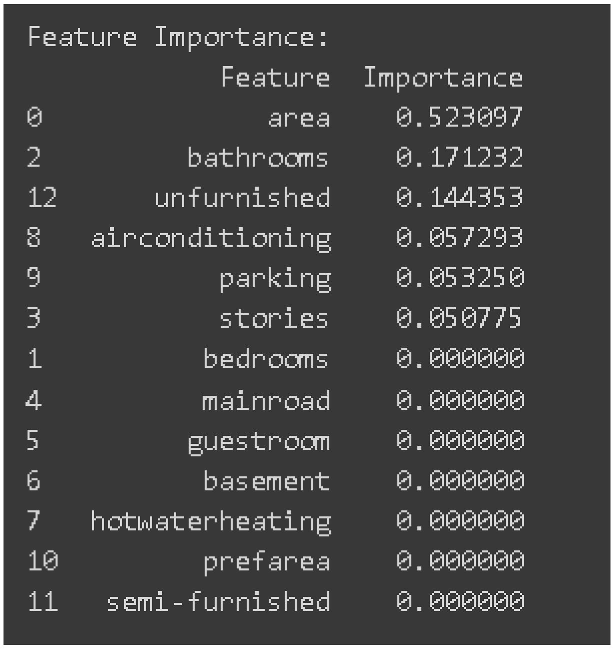
- The confusion matrix determines if the predicted label is the same as the actual label. This shows the values of the amount of correct predictions.
- In this confusion matrix, It shows that the top left and lower right part of the matrix are equal for the predicted and actual label. This means that there are 70 plus 14 correct predictions which is higher than the wrong predictions

- ***Decision Tree***



Model Score: 0.7211009174311926

- The decision tree is a model that decides between two outcomes. It is based on whether the condition is true or false. Each branch considers different factors that change the value for every iteration.
- The variables included in the decision tree are based on the order of the feature importance

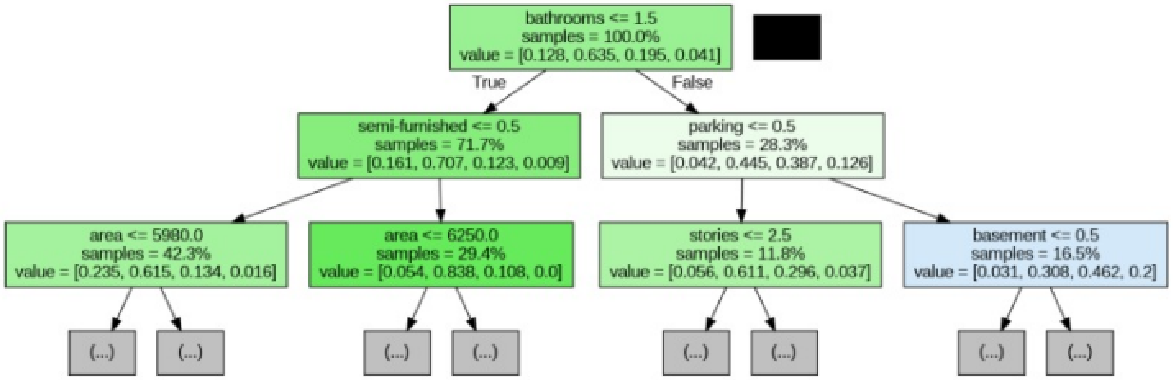


```
Feature Importance:
            Feature  Importance
0              area    0.523097
2         bathrooms    0.171232
12      unfurnished    0.144353
8   airconditioning    0.057293
9           parking    0.053250
3           stories    0.050775
1          bedrooms    0.000000
4          mainroad    0.000000
5         guestroom    0.000000
6          basement    0.000000
7   hotwaterheating    0.000000
10         prefarea    0.000000
11    semi-furnished    0.000000
```

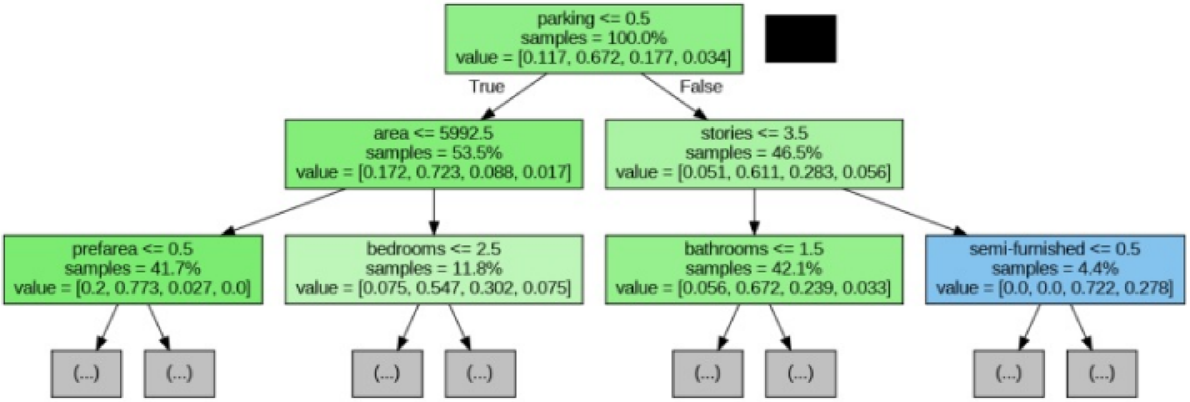
- In this data frame, the best option when looking for affordable houses are the ones with less than 3.5 stories, not furnished, and has an area of less than 5425.
- The best option for average houses are with an area between 3615 and 5425 while being furnished.
- The best option for expensive houses are the one with air conditioning, has more than 1.5 bathrooms, and has an area of greater than 5425.

- The best option for luxury houses is the same with the expensive houses.

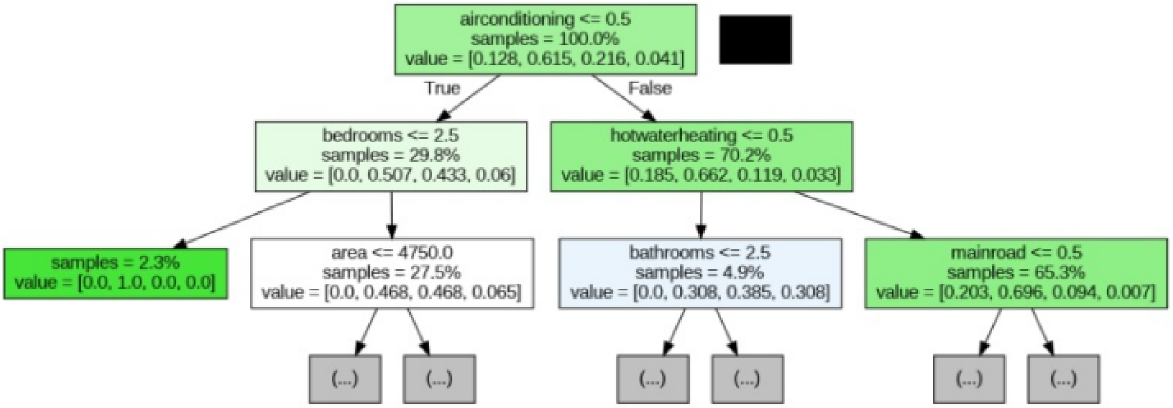
- ***Random Forest***
  - **First Iteration**


[redacted]

  - **Second Iteration**


[redacted]

  - **Third Iteration**


[redacted]

- The random forest works the same as the decision tree when making decisions based on the given variables. Its notable difference is that it is not dependent on the value of importance. The starting decision is randomized making every decision different and can accommodate different combinations of the different variables.

- In the first iteration, we can see that it is able to compare starting from the number of bathrooms while the second and third starts at the number of parkings and the availability of air conditions.
- The values at the end are represented as the percentage of the housing available rather than the exact number of houses.
- The random forest is the best model for this data set. It is able to predict the percentage of housing based on the given conditions. This model allows for certain combinations to be considered. For example if a customer wanted to find a house based on bathrooms, area, and bedrooms, they would need to find an iteration where in it includes those decisions and then find the percentage of available houses in that price range.

Conclusion:

- This prelim exam really gave us so much insight regarding the different Machine Learning models that we use to train and test our data. It also pushed us to learn and study more with regards to this subject so that we can understand it and apply it in future contexts. We also learned that Data Preprocessing is a very important step in data analysis using Machine Learning models, this is because preprocessing the data allows us to manipulate it to fit the model. Preprocessing involves removing null value/changing it to a more meaningful value, changing the formats of categorical variables (Yes/No), creating dummy variables to separate categorical data, removing outliers in the data if necessary, and scaling data so that it can be normalized. This step is very essential because it can affect the overall performance of your model and it also lets you freely manipulate the data to fit a certain model. We also learned how to evaluate a model in different ways like coefficients, accuracy, score, and confusion matrix.

  Overall, this activity allowed us to contrast and compare different Machine Learning algorithms and find the best one that fits our data. In the case of our dataset, the best model is the random forest, this is because of its robust prediction that utilizes many iterations of the decision tree. We really enjoyed this activity and it can introduce students to the world of data science.