**Part I: Research Question**

A.  Describe the purpose of this data mining report by doing the following:

1.  Propose **one** question relevant to a real-world organizational situation that you will answer using **one** of the following prediction methods:

•  decision trees

•  random forests

•  advanced regression (i.e., lasso or ridge regression)

Can we determine what features are important to predicting monthly charges using decision trees?

2.  Define **one** goal of the data analysis. Ensure that your goal is reasonable within the scope of the scenario and is represented in the available data.

One goal of the data analysis is to determine what variables/features are most important to the monthly charges.

**Part II: Method Justification**

B.  Explain the reasons for your chosen prediction method from part A1 by doing the following:

1.  Explain how the prediction method you chose analyzes the selected data set. Include expected outcomes.

Decision trees are useful with classification and regression techniques, which is useful for answering question in “A1” since we will be using a regression model for a continuous variable monthly charge. Decision tree regressor uses branches and nodes that will contain true or false for decision making of the root “MonthlyCharge” and determine the results of the prediction in a tree-like diagram.

2.  Summarize **one** assumption of the chosen prediction method.

Assumption: During the beginning, all variables in the training set are the root.

3.  List the packages or libraries you have chosen for Python or R, and justify how each item on the list supports the analysis.

Packages and libraries used for Python:

1. Numpy
2. Pandas
3. Seaborn
4. Matplotlib
5. Sklearn

Numpy, Pandas, Matplotlib are the most basic imports used in data science projects and provide methods that allow for visualizations, reading, and saves for the dataset. Also, Seaborn was used for more enhanced visualization in graphs, plots, and matrices. Finally, Sklearn helped split, train, predict, handle different algorithms “DecisionTreeRegressor”, reports, and metrics for the data.

**Part III: Data Preparation**

C.  Perform data preparation for the chosen data set by doing the following:

1.  Describe **one** data preprocessing goal relevant to the prediction method from part A1.

The data contains many features/variables that are categorical, so will have to transform those from categorical to numerical. For example, Churn is binary “Yes/No” and will be transformed to “1,0”.

2.  Identify the initial data set variables that you will use to perform the analysis for the prediction question from part A1, and group each variable as continuous or categorical.

In the analysis I will be using a combination of numerical and categorical variables “which are listed below”. Also, I will be choosing the continuous variable “MonthlyCharge” as my target variable and all other variables for all possible outcomes based on certain conditions within the decision tree model.

Continuous Variables:

* Children
* Income
* Outage\_sec\_perweek
* Email
* Contacts
* Yearly\_equip\_failure
* Tenure
* MonthlyCharge
* Bandwidth\_GB\_Year
* Age

Categorical Variables:

* Contract
* Techie
* Phone
* Tablet
* Port\_modem
* InternetService
* Multiple
* OnlineSecurity
* OnlineBackup
* DeviceProtection
* TechSupport
* StreamingTV
* StreamingMovies
* PaperlessBilling
* PaymentMethod
* Marital
* Churn

Ordinal Variables:

* Timely response
* Timely fixes
* Reliability
* Options
* Respectful response
* Courteous exchange
* Evidence of active listening

3.  Explain the steps used to prepare the data for the analysis. Identify the code segment for each step.

1. Import packages and libraries.
2. Read dataset into a data frame.
3. View columns, drop meaningless columns, and update survey questions to appropriate names.
4. Check if any null or empty columns.
5. Check for any misspellings in the data.
6. Check if all previous modifications to the data took effect ex: survey questions have proper names.
7. Get statistics on the data.
8. Get unique values for columns.
9. Prepare Visualizations.
10. Compare model of Decision Tree to Multiple Linear Regression
11. Save a clean copy of the data set.
12. Backup to google drive.

Text

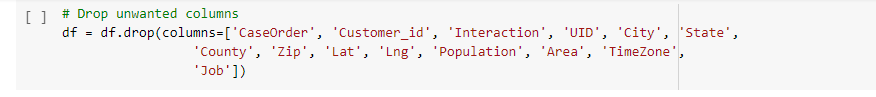
Description automatically generatedGraphical user interface, text

Description automatically generatedTable

Description automatically generatedTable

Description automatically generated with medium confidenceGraphical user interface, application

Description automatically generatedTable

Description automatically generatedA picture containing table

Description automatically generatedA screenshot of a computer

Description automatically generated with medium confidenceTable

Description automatically generatedText

Description automatically generatedA picture containing diagram

Description automatically generatedDiagram

Description automatically generatedChart

Description automatically generatedA picture containing table

Description automatically generatedChart, bar chart

Description automatically generatedGraphical user interface

Description automatically generated with medium confidenceChart, bar chart

Description automatically generatedChart, bar chart

Description automatically generatedChart, bar chart

Description automatically generatedChart, bar chart

Description automatically generatedChart, bar chart

Description automatically generatedChart, bar chart

Description automatically generatedChart, bar chart

Description automatically generatedText

Description automatically generatedGraphical user interface, text, application

Description automatically generatedGraphical user interface, application

Description automatically generatedGraphical user interface, application

Description automatically generatedGraphical user interface, application

Description automatically generated

4.  Provide a copy of the cleaned data set.

Graphical user interface, text

Description automatically generated with medium confidence

**Part IV: Analysis**

D.  Perform the data analysis and report on the results by doing the following:

1.  Split the data into training and test data sets and provide the file(s).

A picture containing text

Description automatically generated

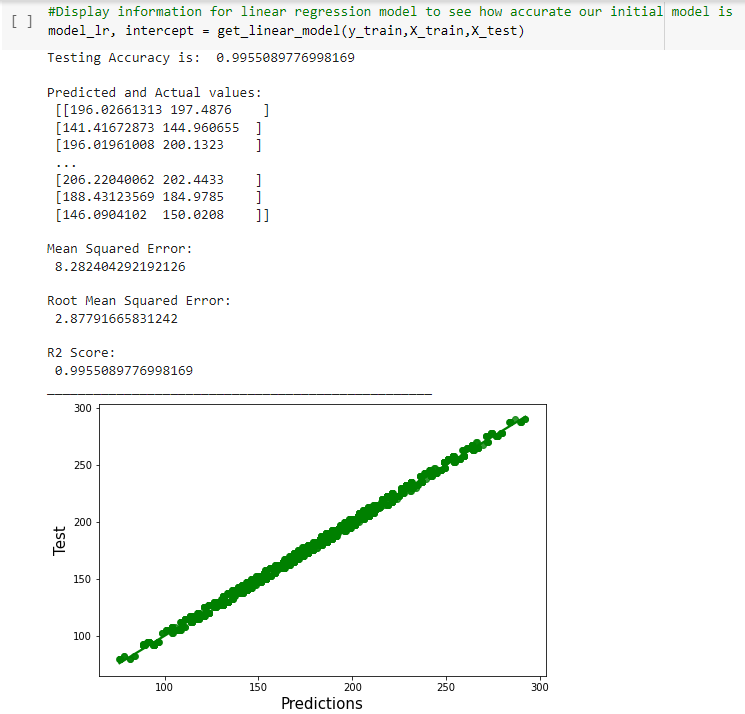
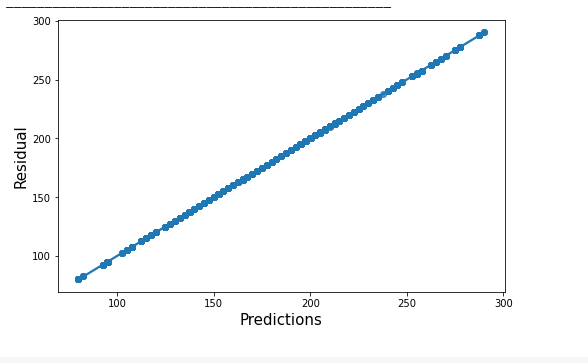
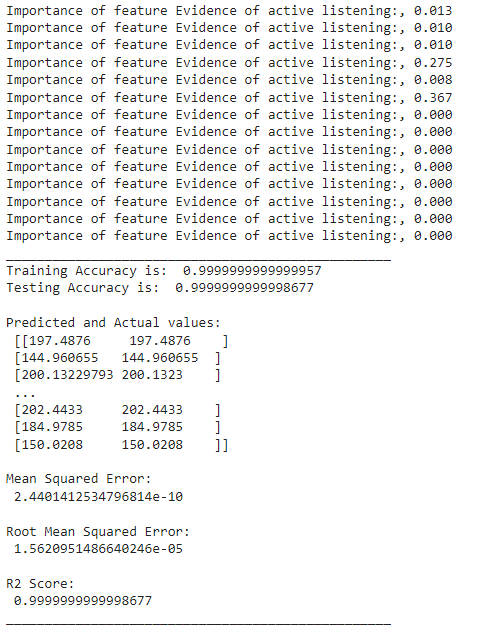
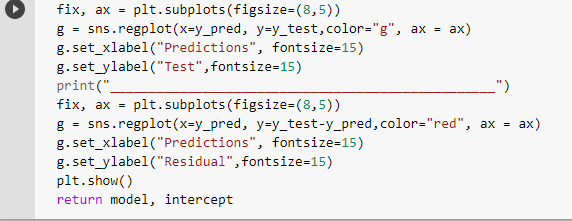
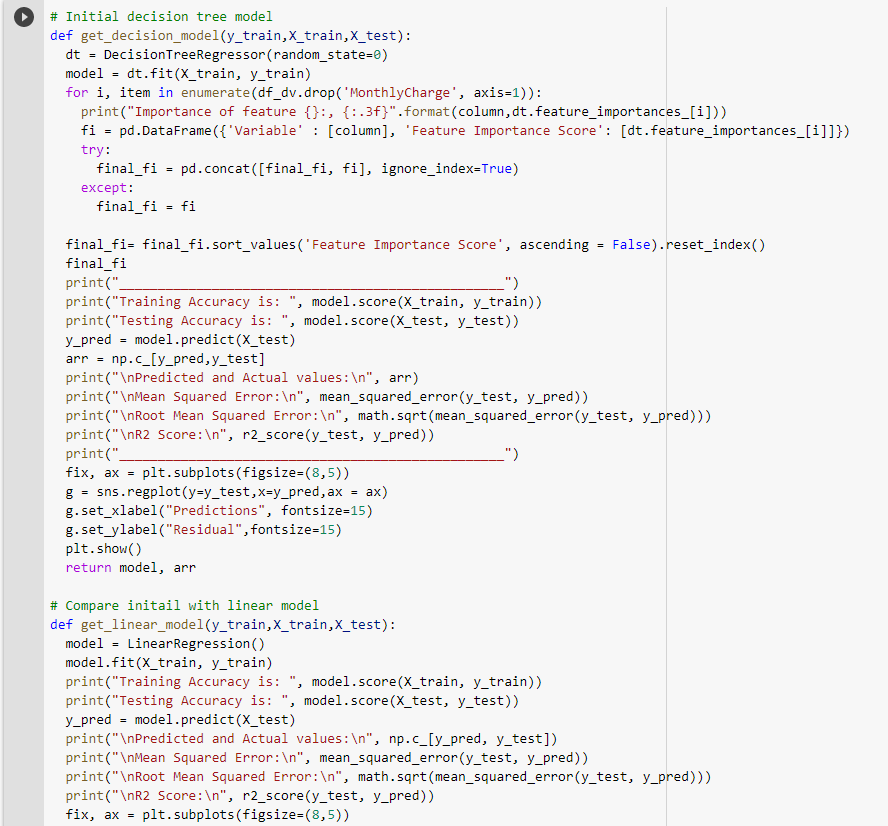
2.  Describe the analysis technique you used to appropriately analyze the data. Include screenshots of the intermediate calculations you performed.

For the analysis I created two functions one for initial regression algorithm “Decision Tree Regressor” and one for comparability of the initial model using “Multiple Linear Regression”. Both functions will provide accuracy scores “Training and Testing”, prediction values “Predicted and Actual”, mean squared error, root mean squared error, and R2 scores. Furthermore, the initial model includes the feature importance of all variables in the dataset, which will be printed to the screen.

Text, letter

Description automatically generated

3.  Provide the code used to perform the prediction analysis from part D2.



**Part V: Data Summary and Implications**

E.  Summarize your data analysis by doing the following:

1.  Explain the accuracy and the mean squared error (MSE) of your prediction model.

From the initial model the accuracy of the training data: 0.999999999999999957, and testing data: 0.9999999999998677, which shows that our model is very accurate at predicting unseen data. Also, the MSE is 2.4401412534796814e-10 for the initial model shows that the model is good at predicting data accurately because when MSE is converted to a decimal “0.0244….”, which is very close to 0 a perfect model, gives a hint of the lower the score the better the model is a prediction accuracy.

.

2.  Discuss the results and implications of your prediction analysis.

From the initial analysis results of accuracy of training data, testing data, MSE, and RMSE results in an extremely good model, so to test this we used another regression model to compare the results to see if our model is to good to be true.

|  |  |  |
| --- | --- | --- |
| **VALUES** | **DECISION TREE REGRESSOR** | **LINEAR REGRESSION** |
| Training | 0.999999999999999957 | 0.9955048609952436 |
| Testing | 0.9999999999998677 | 0.9955089776998169 |
| MSE | 2.4401412534796814 | 8.282404292192126 |
| RMSE | 1.5620951486640246 | 2.87791665831242 |

We can see that the accuracies from the above table show that both are very similar except the MSE is way higher in the Linear model than the decision tree. With that information of MSE it is possibly better to choose the decision tree regressor for the analysis because of the lower MSE score.

3.  Discuss **one** limitation of your data analysis.

The limitation on this data analysis is overfitting, since the decision tree continues creating hypotheses that can reduce the training set error while increasing the test set error. Furthermore, since overfitting can happen, we will have to resolve this issue by pruning and adding constraints to the model parameters.

4.  Recommend a course of action for the real-world organizational situation from part A1 based on your results and implications discussed in part E2.

With good scores from the decision tree model the stakeholders will have a high confidence in predicting the monthly charge rates. The stakeholders should look at the highest features importance’s and see if their rates are comparable to competitors and adjust in price where there is a need. Also, marketing team can design a strategy for promotions for new customers and customers already with the company, to help increase customer tenure, and retain customers.

**Part VI: Demonstration**

F.  Provide a Panopto video recording that includes a demonstration of the functionality of the code used for the analysis and a summary of the programming environment.

Note: The audiovisual recording should feature you visibly presenting the material (i.e., not in voiceover or embedded video) and should simultaneously capture both you and your multimedia presentation.

Note: For instructions on how to access and use Panopto, use the "Panopto How-To Videos" web link provided below. To access Panopto's website, navigate to the web link titled "Panopto Access," and then choose to log in using the “WGU” option. If prompted, log in using your WGU student portal credentials, and then it will forward you to Panopto’s website.

To submit your recording, upload it to the Panopto drop box titled “Data Mining I – NVM2.” Once the recording has been uploaded and processed in Panopto's system, retrieve the URL of the recording from Panopto and copy and paste it into the Links option. Upload the remaining task requirements using the Attachments option.

G.  Record the web sources used to acquire data or segments of third-party code to support the analysis. Ensure the web sources are reliable.

None were used

H.  Acknowledge sources, using in-text citations and references, for content that is quoted, paraphrased, or summarized.

None were used

I.  Demonstrate professional communication in the content and presentation of your submission.