**Prediction of Normal Butane Volume Percentage**

* Supervised Machine Learning:
* With our project we wanted to explain:
  + How a machine learning algorithm is used in data analytics.
  + Create training and test groups from Normal Butane Volumen Percentage dataset.
  + Implement the logistic regression, decision tree, random forest, and Support Vector Machine algorithms.
  + Interpret the results of the logistic regression, decision tree, random forest, and support vector machine algorithms.
  + Compare the advantages and disadvantages of each supervised learning algorithm.
  + Determine which Supervised Learning algorithm is best used for Normal-Butane-Volume-Percentage data set.
  + Use ensemble and resampling techniques to improve model performance.

We were able to demonstrate our machine learning abilities. When resampling models to Predict of Normal Butane Volume Percentage, there was an accuracy score and confusion matrix for all three algorithms. And, we were able to simultaneously generate a classification resort for all three respective algorithms, we did a great job in regards to predicting Normal Butane Volume Percentage for the SMOTEENN algorithm. There is a confusion matrix and classification report. We used ensemble classifier to predict Normal Butane Volume Percentage. Proper implementation of accuracy score and confusion matrix for one algorithms. We were able to create a classification report and for one algorithms. We were able to sorting our list features, as well. The balanced accuracy score and the precision and recall scores for all our algorithms are described.

Feature Importances: We also went ahead and calculated some feature importances, feature coefficients, and correlations to understand the relationships of the features among each other and with different model types. The feature importance functions such as model.coef\_ and model.feature\_importances\_, along with naitive process knowlege of team group members, allowed for a systematic ranking of features. Though 54 of the original 56 features were left in the model, the feature importance ranking is a valuable output and will be used to further tune the model outside of the class project end date.

* ETL:

We were able to create a function to extract, transform and load our Normal Butane Volumen Percentage dataset. Our dataset were extracted into DataFrames and We displayed the DataFrames with columns. We cleaned Normal Butane Volumen Percentage data. We filtered out the Normal Butane Volumen Percentage, we used try-except to drop duplicates. We list comprehension was used to drop columns with null values. Our dataset were cleaned properly. Unnecessary columns were dropped from this DataFrame, columns were renamed and missing Kaggle data was filled using Normal Butane Volumen Percentage data. Normal Butane Volumen Percentage data was cleaned and merged with the Normal Butane Volumen Percentage DataFrame. The empty rows were filled with zero. The DataFrames are displayed and have the correct columns.

* Web with HTML/CSS:

All of prediction-of-Normal-Butane-Volume-Percentage images, tables, code were correctly displayed, stored, and retrieved. The data is properly displayed using HTML and aesthetically enhanced through CSS and additional bootstrap 3 components.

* Pandas:

We were able to cohesively present a written analysis that summarizes the data displayed in the dataframe and multiple-line chart. We also explained the implication of both analyses. We created our DataFrames in our notebook. It cont ained all the prediction-of-Normal-Butane-Volume-Percentage. The notebook also generated a summary DataFrame with all the requirements. We were effectively able to demonstrate the graph with annotations, x-axis is by dates and the correct graph style is used.

* Matplotlib:

We were able to cohesively present a written analysis that summarizes the data displayed in the dataframe and multiple-line chart. We created our DataFrames our notebook. It contained all the information and averages. Our notebook also generated a summary DataFrame with all the requirements. We were effectively able to demonstrate the graph with annotations, x-axis is by months and the correct graph style were used.

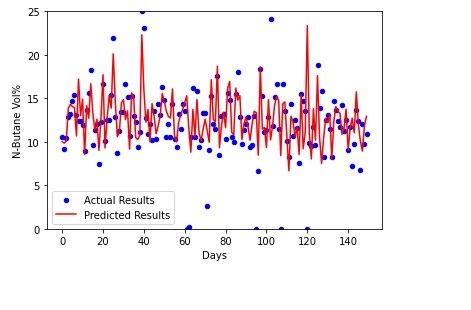
* SQL:

We more than delivered on our query establishment that enabled us to create a table with unique Normal-Butane-Volume-Percentage tables, all of which were successfully exported as CSV files. In addition, the format of our findings were arranged accordingly.

* Statistics:

We were able to demonstrate our statistics abilities with this project, we got a very marketable skill during this Bootcamp. The linear regression to predict Normal Butane Volume Percentage was great. We imported the csv file and read into a DataFrame. We wrote a clean R script for a linear regression model to perform on all our variables. An R script was also written well to create a statistical summary of the regression and it addresses all the requerements for this project as well .The total summary has all the metrics. An R script was written for a t-test that compares Prediction of Normal Butane Volume Percentage Vs. Actual Real Time, an R script was also written for all three t-tests. We were able to address the results across all Normal Butane Volume Percentage dataset. We have an awesome statistical design. We had a metric to be tested, a null or alternate hypothesis was described, our statistical test and the data for the statistical test is well described. This was an amazing project, we really got to showcase our R and statistical modeling knowledge in an amazing comprehensive manner. We wish to keep up the good work and this amazing flow we have been able to establish for the next project.

Figure 1. Linear Regression

[](https://user-images.githubusercontent.com/101227930/185302134-7e35ff5c-ce25-4316-9b91-4122aa869cc7.png)

* Python: we proficiently implement python code and the correct pathing.
* Our Webpage: we were able to demonstrate our Webpage creation abilities. We created titles and paragraphs in the containers and we played with colors and shapes. We were able to explore head, body and footer. We used bootstraps and differents features. As soon as the web application is loaded up, all the information in behaves exactly as it should. We correctly displayed the dashboard at our webpage and our Team contact information. All elements were created and when the webpage loads, everything were working good with links and buttons. We were able to created a new environment that fit for this project and with the necessary requirements in order to deploy Heroku for external views. Our webpage is fully responsive and clean when the app loads. The webpage has all three customizations and loads and updated all the required tables and graphics without any errors.
* Big Data:

Our notebook file set a Normal Butane Volume Percentage dataset and it is extracted as a DataFrame , then our extracted dataset is transformed into a DataFrames with the correct columns and all our DataFrames are loaded into their respective tables in pgAdmin. The data is filtered to create a DataFrame or tables where there are 20 dataset. The data is filtered to create a DataFrame or table where the Prediction of Normal Butane Volume Percentage is equal to or greater than 20% .

* Unsupervised Machine Learning:

We were able to preprocess the data for Prediction of Normal Butane Volume Percentage. We were able to perform all the following requirements on the Normal Butane Volume Percentage DataFrame. All the information that we did not need were removed and the ones that did not have a defined algorithm were removed, all the ones with at least one null value were dropped. Also we created a new DataFrame that stores Prediction of Normal Butane Volume Percentage. We have an optimal method in place to create variables for all the text features, and we were also able to store in a DataFrame. The features from the DataFrame have been standardized with the proper function. We were able to reduce data dimensions using PCA. Our PCA algorithm reduces dimensions of the X dataFrame to three principle components. We were able to create the Normal Butane Volume Percentage DataFrame with the columns. Our DataFrame uses the index from our Prediction of Normal Butane Volume Percentage DataFrame. When it came to clustering our Prediction of Normal Butane Volume Percentageyour using K-means, we were able to establish an elbow curve using hvPlot to find the best value for K. Optimal predictions were made. A new DataFrame were created with the same index as our Prediction of Normal Butane Volume Percentage DataFrame and had the necessary columns. Finally, the clusters were well plotted using a 3-D scatter plot and each data point showed the respective information and algorithm associated with it upon hovering. We did great job visualizing and materializing our results. We were really excited with our first project. Money Saver Team is the best. Also we created a table with our Prediction of Normal Butane Volume Percentages using a function, which is then printed. Our DataFrame was created containing all clustered respective data. This was a tough project and our team did an absolute stellar job.

We cohesively present our analysis and draw reasonable conclusion from our Normal-Butane-Volume-Percentage dataset:

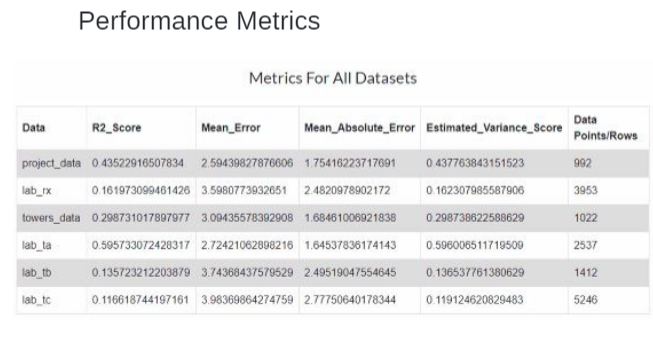
In conclusion, there is certainly limitations to the model. While it does not predict the N-Butane Vol% exactly, it does seem to trend the general direction of the N-Butane Vol%, and generally predicts this figure within 2-5 Vol% points. By only allowing the prediction to occur when all input data streams are within normal operating conditions, performance is likly to improve.

Summary of Project:

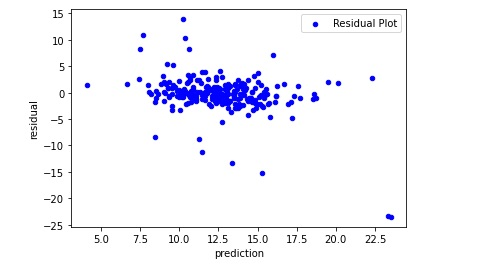
Analysis Output Performance

Performance Metrics The following performace metric results were gathered based on the X\_test predictions.

Figure 2: Metrics

[](https://user-images.githubusercontent.com/101227930/185315649-69ecf9f0-c649-4a31-8407-277a975afd04.png)

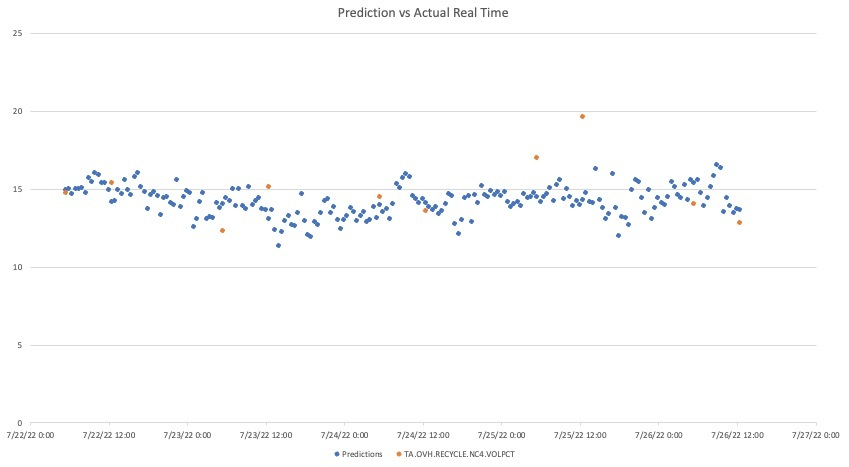
Residual Plot: The plot in Figure 2 show evenly distributed residuals, thus giving reason to believe the model performs well and is not over-fit. Figure 3: Residual Plot

[](https://user-images.githubusercontent.com/101227930/185297754-189facfe-7a53-4692-80d8-6e8cbfc9ea06.png)

The total Mean Absolute Error is 1.92, meaning the average error is only 1.92 Vol%. The total Root Mean Squared Error is 3.51.

Real Time Model Performance Finally, real time data, not included in the original dataset was taken from the past 3 weeks and ran in place of the X\_test data as X\_realtime data. Unfortunately, unit downtime during this time-span allowed only 4 days of continuous runtime data to be input. This real time data was input to the model to simulate an operators view of actual and simulated N-Butane Vol% in TA's recycle line. The results are shown in Figure 4 below.

Figure 4: Real Time Model Performance

[](https://user-images.githubusercontent.com/101227930/185296714-bcfd6502-d606-4bbb-b380-967d5297828d.png)

Conclusion In conclusion, there is certainly limitations to the model. While it does not predict the N-Butane Vol% exactly, it does seem to trend the general direction of the N-Butane Vol%, and generally predicts this figure within 2-5 Vol% points. By only allowing the prediction to occur when all input data streams are within normal operating conditions, performance is likly to improve.