**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 Volume 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.

We were able to demonstrate our machine learning abilities. When resampling models to Predict of Normal Butane Volume Percentage, there was a performance metrics analysis. And, we were able to simultaneously generate a classification resort for all three respective algorithms. We used linear regression model to predict Normal Butane Volume Percentage focusing on mean error as our primary performance metric. Are target was a mean error of 2% percentage points, after the model was built and tested, we were able to achieve a mean error of 2.5% percentage points.

**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 native 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 Volume Percentage dataset. Our dataset were extracted into DataFrames and We displayed the DataFrames with columns. We cleaned Normal Butane Volume Percentage data. We filtered out the Normal Butane Volume 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 data was filled using Normal Butane Volume Percentage data. Normal Butane Volume Percentage data was cleaned and merged with the Normal Butane Volume 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 contained 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.

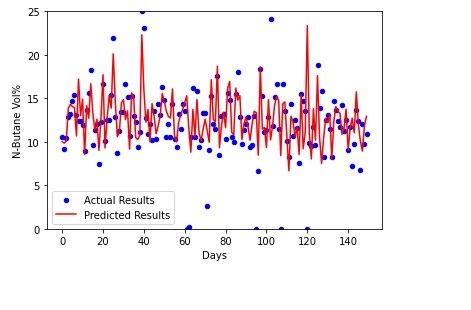
* **SQL:**

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 Python script for a linear regression model to perform on all our variables. An Python script was also written well to create a statistical summary of the regression and it addresses all the requirements for this project as well. The total summary has all the metrics. A Python script was written for a performance summary test that compares Prediction of Normal Butane Volume Percentage Vs. Actual Real Time. We were able to address the results across all Normal Butane Volume Percentage dataset. This was an amazing project, we really got to showcase our Python 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 codes and we used 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, though we were not able to completely deploy the app on Heroku. 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% .

**Our team:**

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

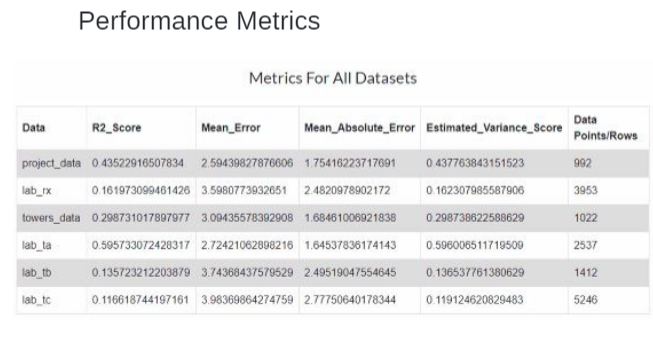
In conclusion, there is certainly limitations to the model. While it does not predict the N-Butane Vol% exactly, it does seem to trend the general direction of the N-Butane Vol%, and generally predicts this figure within 2-5 Vol% points. By only allowing the prediction to occur when all input data streams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

**Summary of Project:**

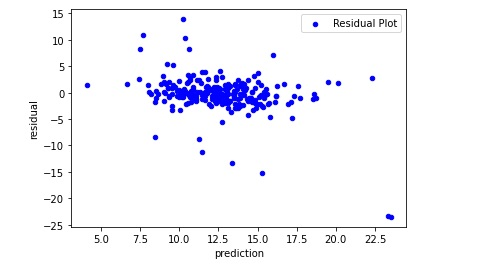
Analysis Output Performance

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)

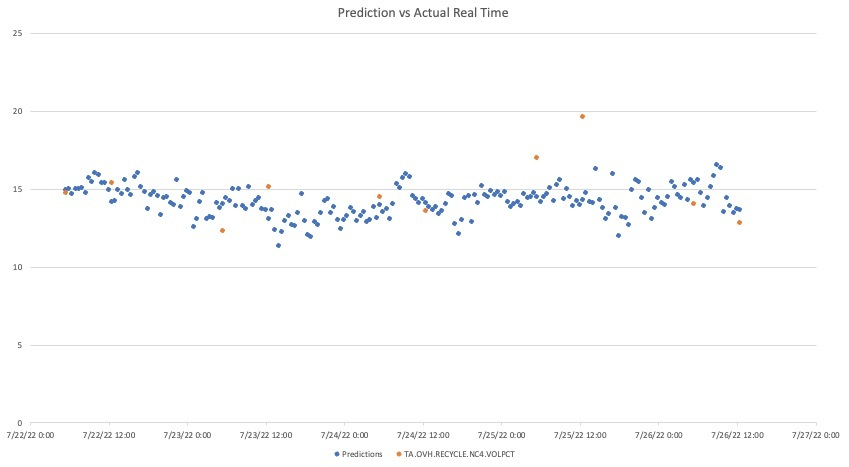
**Figure 2** Residual Plot: The plot is evenly distributed residuals, thus giving reason to believe the model performs well and is not over-fit.

[](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 3: 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.

* Reexamine ETL Process (reduce mergeable data loss)
* Future Models (choose neural network regression model)
* Operational Background (work with operations to determine normal operation conditions)
* Feature Importance and Correlation (further optimize model inputs)