Baruch College
Zicklin School of Business

Pitney Bowes Data Challenge Competition

Data Challenge Poster Team 14

ABSTRACT

We were tasked with creating a model that predicts when a Pitney Bowes machine will fail. We decided the best way to go about this was to minimize the false negatives and work on a model with the best recall. We were given a training data set 45000 rows and 53 columns, and a test data set where we were expected to predict when the machine would fail. We started off by cleaning the data and decided to leave the null values as zeros in the cleaned data. We ran exploratory tests on the data to determine which variables had the highest impact on the failure, with feature importance. Finally, we moved our data to the training and testing portion.

We decided our performance measure would be recall as we wanted to be sure there were fewer false negatives. The model that had the highest recall was logistic regression with a recall of 0.80, accuracy of 0.61, and AUC of 0.68. There were other models that had performed better with accuracy, but to prevent the loss of a client due to machine failure recall was the more important feature.

The most important features in the data provided are the avg_time_discharging_lag7 and discharging_rate_lag7, followed by average_time_charging_lag7. These features were most indicative of meter failure within seven days.

Team 14

David Lado
Syed Hossain
Victoria Vayner
Daniel Rubens
Immanuel Ryan Augustine

Introduction

Pitney Bowes is a global leader in the parcel postage industry and has remained so for much of the century since its inception in 1920. Businesses the world over, both large and small, rely on Pitney Bowes to streamline their parcel processing and reduce costs in several areas essential to their operations. They offer a wide range of products tailored to the needs of businesses in all sectors, of all sizes. Beyond the amazing products Pitney Bowes (PB) provides it is the assurance that their machines will have the greatest uptime possible which allows them to retain such a loyal userbase. It is with this goal in mind that we have conducted our analysis.

Using predictive modeling to parse out potential machine failures allows PB to stay ahead of procurement, inventory, and escalated customer service matters. Depending on the size of the client PB may decide to preemptively replace a fleet of machines presenting a high likelihood of failure. Data analysis of this nature also plays a crucial role in the engineering and R&D departments of PB as well as aiding in quality control of current and future machines. A crucial intermediary of Pitney Bowes' business operations, the data analysis team is able to provide invaluable information to key stakeholders and clients alike, reducing cost and friction, while allowing them to maintain dominance in a dynamic and ever-changing industry.

Cleaning and Parsing Data

While cleaning and parsing the data there were a few characteristics of note. The original data contained null values with a unique linearity. Particularly on the lag data, there appeared to be null values increasing over time. We assumed this was some sort of progressive failure rate and attempted to incorporate that in our modeling. We changed the two date columns to datetime elements and used them to create an additional column "DaysInService" which served as a running tally of the time machines were in use.

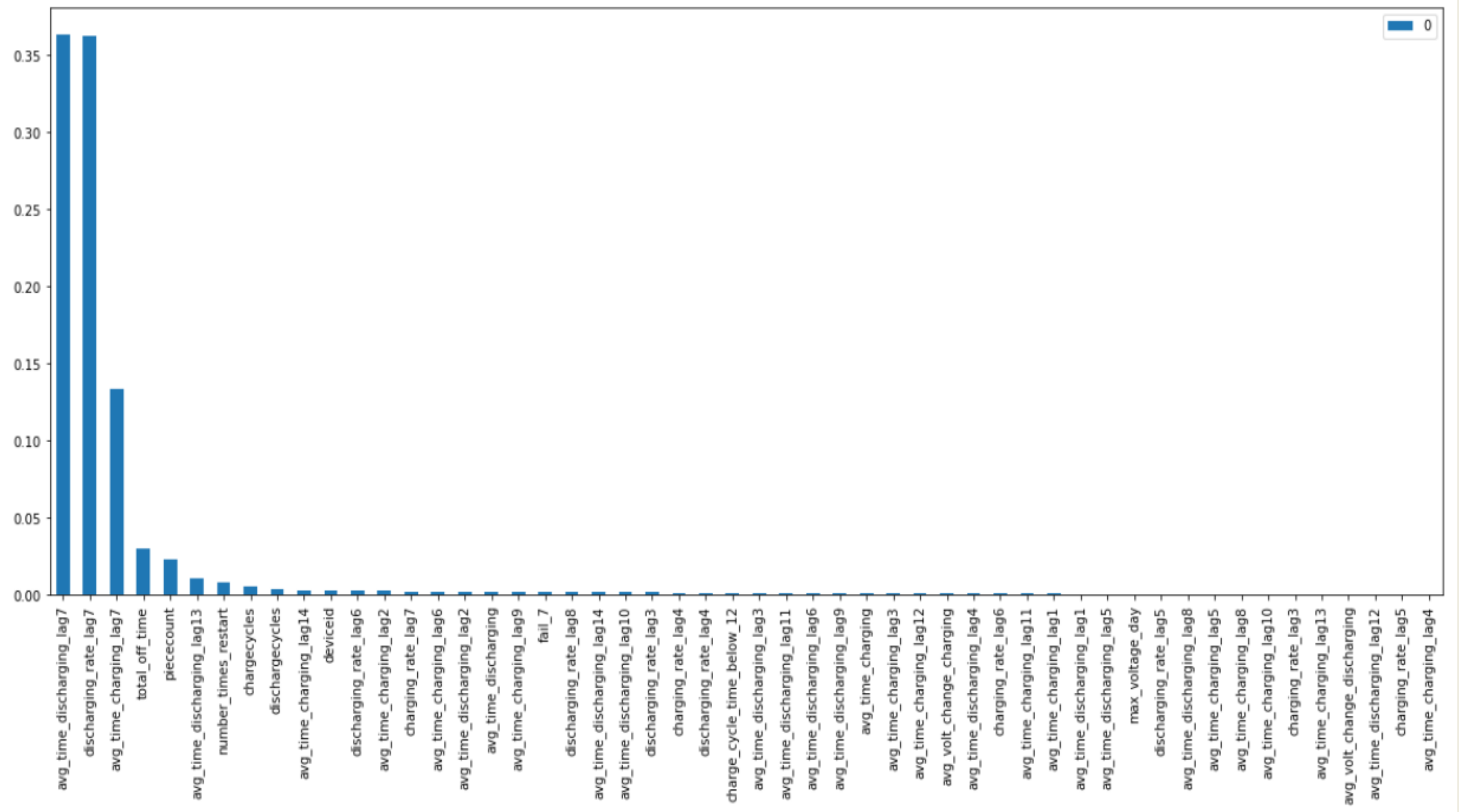
Outliers were dealt with by quantile. Both the nulls contained in the original dataset as well as null values created in the outlier code were converted to 0 to retain their presence in the data. Dataframes were created throughout the cleaning and parsing process and plotted using Matplotlib. The final dataframe used in our models ("tr1") had all original null values zeroed, outliers zeroed, "DaysInService" added, and the two date columns, as well as the "deviceid" column, removed as they were not considered material to our analysis.

Interpretation of Performance Measures

There are four main performance metrics that our models test for: Accuracy, Precision, Recall, and F1-Score To align with the business objectives of Pitney Bowes the recall will be more heavily weighed for consideration of best model. The reason being because recall calculates the number of actual positives our model finds. For this reason, it is the best model for costs that could be associated with a False Negative and a lost customer due to a faulty machine.

These metrics will be tested for different models. The models tested are Logistics Regression, Naïve Bayes, Decision Tree, Random Forest, Gradient Boosted Tree, and Neural Network. Each will be evaluated against the performance metrics, an AUC, and an ROC Curve analysis.

Feature Importance



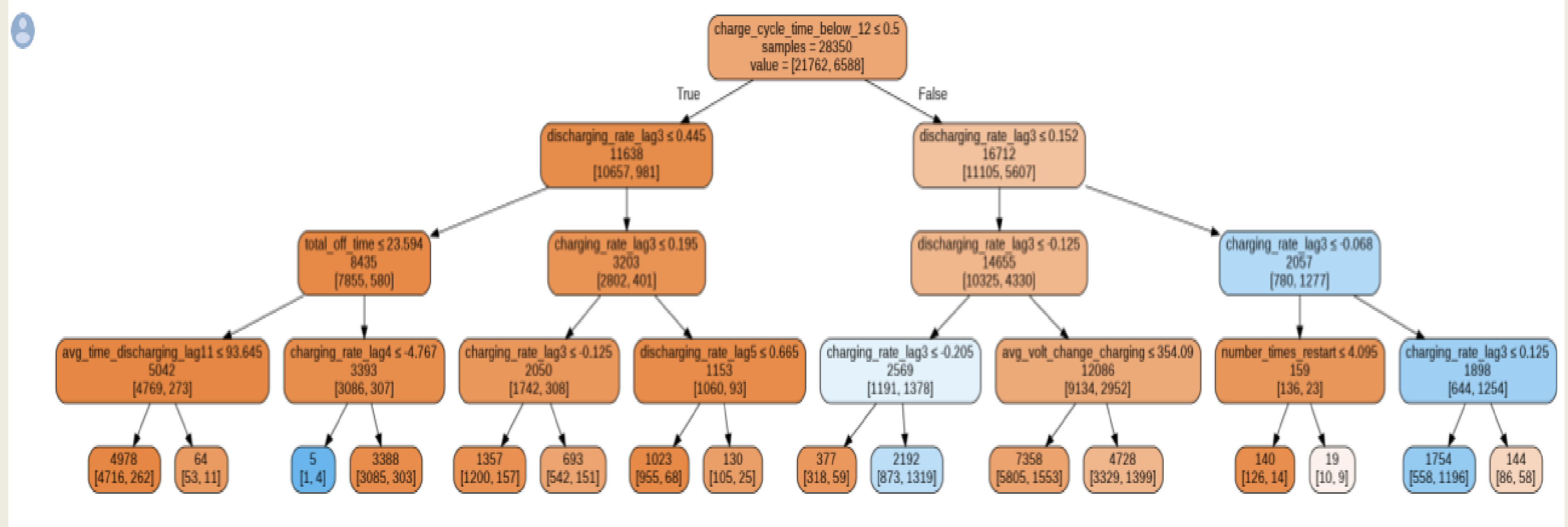


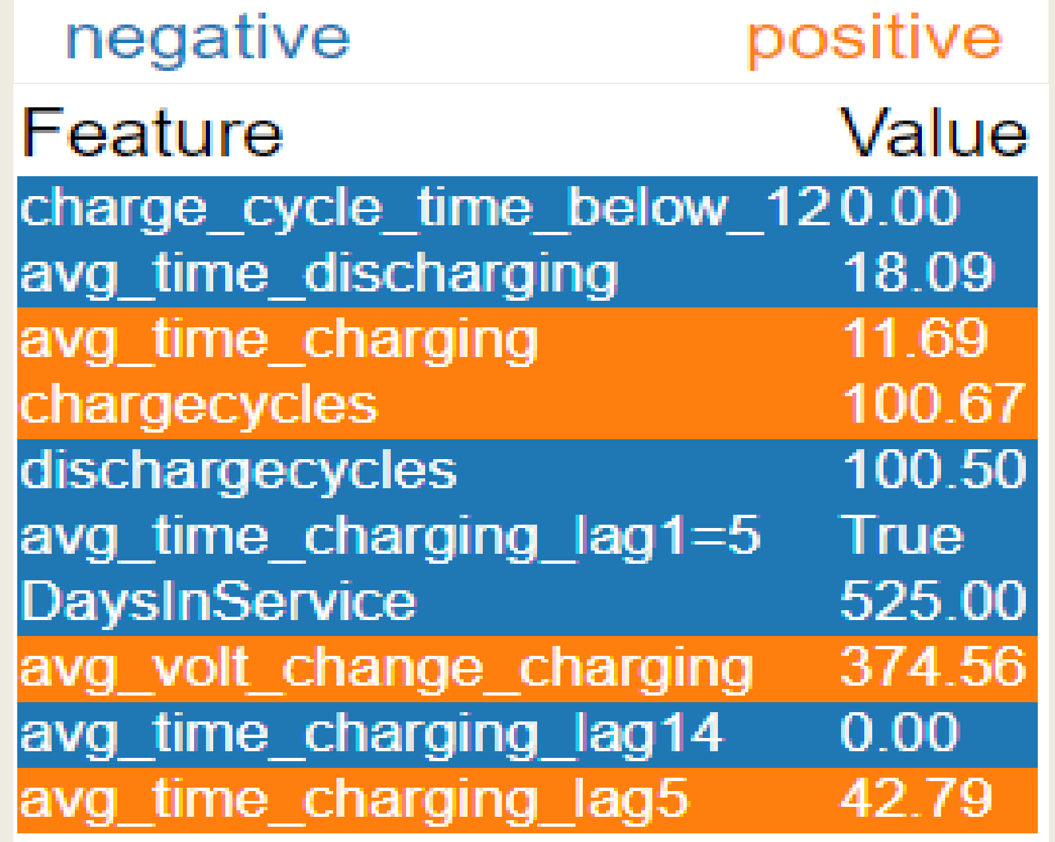
Figure 01: Decision Tree

Baseline AUC

LogisticRegression Train	0.679651024
LogisticRegression Valid	0.674552401
DecisionTree Train	0.658279547
DecisionTree Valid	0.657571692
NaiveBayes Train	0.653838997
NaiveBayes Valid	0.65224597
RandomForest Train	1
RandomForest Valid	0.59574629
GradientBoostedTree Train	0.664418049
GradientBoostedTree Valid	0.657622768
NeuralNetworks Train	0.727816889
NeuralNetworkss Valid	0.711501447

Model Metrics

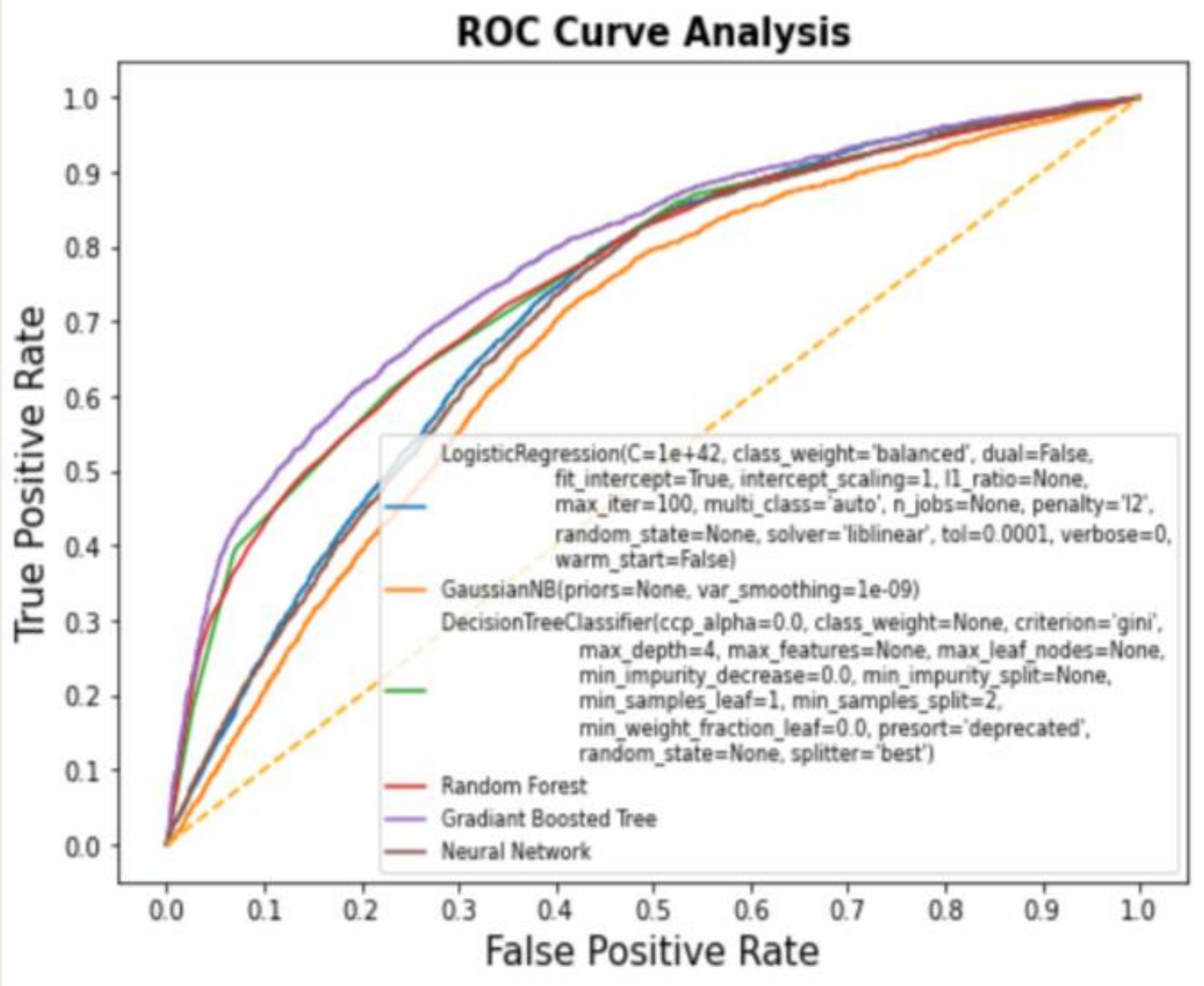
Name	Train Accuracy	Test Accuracy	Train Precision	Test Precision	Train Recall	Test Recall	Train F1 score	Test F1 score
Logistic Regression	0.6140	0.6070	0.3540	0.3482	0.8023	0.8002	0.4913	0.4853
Decision Tree	0.8059	0.8052	0.6375	0.6310	0.3823	0.3825	0.4780	0.4763
Naive Bayes	0.62	0.6149	0.3465	0.3426	0.7170	0.7216	0.4672	0.4646
Random Forest	1	0.7978	1	0.7035	1	0.2193	1	0.3344
Gradient Boosted Trees	0.81856	0.8137	0.7053	0.6816	0.3764	0.3668	0.4908	0.4770
Neural Network	0.77	0.7649	0.5332	0.4529	0.0803	0.0735	0.1395	0.1266


negative
positive

Feature	Value
charge_cycle_time_below_12	0.00
avg_time_discharging	18.09
avg_time_charging	11.69
chargecycles	100.67
dischargecycles	100.50
avg_time_charging_lag1=5	True
DaysInService	525.00
avg_volt_change_charging	374.56
avg_time_charging_lag14	0.00
avg_time_charging_lag5	42.79

Figure 02: Model Interpretability

ROC Curve Analysis



Conclusion

Our best model was the Logistic Regression model with 0.80 recall, 0.61 accuracy and AUC of 0.68. The Naïve Bayes model came in a close second with a recall of 0.72, an accuracy of 0.62, and AUC of 0.65. We had a few other models with better accuracy, but for the purposes of this analysis, we went with the model that had the best recall. Random Forest was dismissed due to overfitting issues. Moving forward we would suggest normalizing the data due to skewedness in the dataset for some attributes.

The most important features to help determine meter failure are the charge_cycle_time_below_12 and avg_time_discharging. It seems evident that to better predict the failure of a machine within a week, looking at these features will give the best prediction of whether a machine will require maintenance in a week.

Acknowledgements

A special thanks to the Pitney Bowes team and Baruch for collaborating to make this competition possible for Baruch College students and for the guidance, insights, and encouragement given to us along the way.