Early Warning Signal for Pitney Bowes Meters
Identifying Failed Meters to Reduce Down-time Risk

Chi Nguyen, Huong Nguyen, Aizaz Ur Rahman, Linh Bui

Baruch College

phamhaichi.nguyen1@baruchmail.cuny.edu, huongthao.nguyen@baruchmail.cuny.edu, aizazurrahmantaha.mohammed@baruchmail.cuny.edu, linh.bui@baruchmail.cuny.edu

Business Understanding

Pitney Bowes is one of the leading solution providers in Postage Meters, which allows businesses to simplify their shipping and mailing process.

KEY PARTNERS
- Pitney Bowes
- Its customers (small businesses, enterprises, retails, government clients)
- USPS

KEY ACTIVITIES
Provide postage meters

KEY RESOURCES
Postage meters

VALUE PROPOSITION
Help customer:
- Save time (skip the Post Office and easily buy and print USPS® postage)
- Save money (access to discounted postage)

CUSTOMER SEGMENT
Businesses who ship large quantities of mail through the USPS

COST STRUCTURE
- Production and maintenance cost of postage meters
- Labor cost
- Selling cost
- Marketing cost
- Research, Development & Intellectual Property cost

REVENUE STREAM
- Direct sales (selling postage meter equipment)
- Subscription plan (month posting printing fee)

The goal of this data mining challenge is to predict which meters will fail within the next 7 days in order to reduce down-time risks of meters deployed at Pitney Bowes' customers to avoid any sort of disruption.

Data Understanding

Data Description:

The data used in this challenge includes two datasets: train dataset and test dataset.

- train.csv (15.3 MB) – Sample data for meters with a flag whether they fail within the next 7 days, including 40,500 records with 55 attributes.
- test.csv (1.69 MB) – Sample data without fail flag, which is the objective of this data mining used for prediction, including 4,500 records with 54 attributes.

The datasets provide charge and discharge information in terms of average time, cycle, rate, volve; total time off, number of restart times, max voltage … Lag data from 1 to 12 for some average time and rate are also provided.

Data Exploration:

- ***Charging_rate_lag3 and charging_rate_lag7*** have minimum of negative values. We assume this is technical problem in data collection and will transform the negative values to null values.
- ***Total_off_time*** range from 0 to 2013. The variable is worth considering for abnormal values.
  - ***Life_of_device (new variable)*** is calculated as: **LastRecord - Date Deployed**
- ***Total_off_life (new variable)*** from ***total_off_time and life_of_device.***
  - ***Total_off_time*** is proportional with the life of the devices from the date of being deployed to the date of last record.
  - ***Time_off_life*** variable has outliers. Thus, we replace the outliers with values from percentile 5 to 95.

Data Preparation

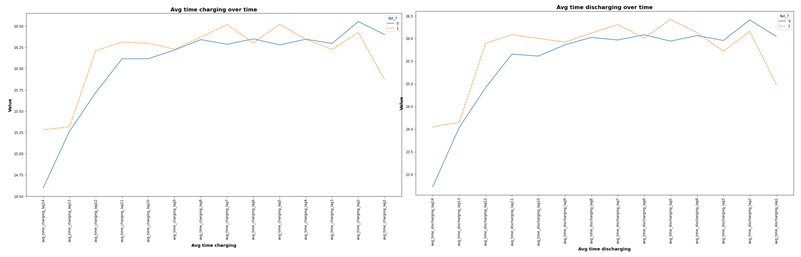
Dealing with missing values

- ***charging_rate_lag7 and charging_rate_lag3:*** 50% of the values are missing after replacing the negative values in with null values. Therefore, **we drop these 2 features.**
- ***avg_time_charging_lag14*** **and** ***avg_time_discharging_lag14:*** missing values are about 16.5% each, which are the largest missing proportion, but are still within margin of error. **Therefore, we decided to keep these variables.**
- ***avg_time_charging_lag14, avg_time_discharging_lag14, avg_time_charging_lag13, avg_time_discharging_lag13, avg_time_charging_lag12, avg_time_discharging_lag12:*** variables with the most missing values.
  - These variables have the strongest correlation with ***cycle_time.***
  - The missing values of variables are usually in the machines whose ***cycle_time*** are between 30 and 60.Therefore, we will fill the missing values with median values of machines with ***cycle_time*** from 30-60.
- Other variables whose missing value proportions are below 6%
  - We fill values normally with the corresponding median values calculated on the available data.

	Missing Values	% of Total Values
charging_rate_lag7	20321.0	50.2
charging_rate_lag3	20039.0	49.5
avg_time_discharging_lag14	6668.0	16.5
avg_time_charging_lag14	6668.0	16.5
avg_time_discharging_lag13	4382.0	10.8
avg_time_charging_lag13	4382.0	10.8

Examining the association between variables

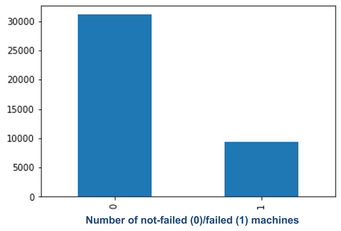
- **Removing either discharging variables or charging variables:**
  - There is a strong correlation between **discharging variables and charging variables**.
  - The trend over time of **discharging variables and charging variables** for **failed** and **not failed machines** shows similar patterns**.** Therefore, we can remove either discharging variables or charging variables.
  - In this case, we will **remove *avg_time_charging variables*.**
- ***charging_rate*** and ***discharging_rate*** show different patterns on the outcome (fail vs not fail).
  - We assume the two variables have different effects on the machines.



In summary, removed 14 following variables: *avg_time_charging_lag14, avg_time_charging_lag13, avg_time_charging_lag12, avg_time_charging_lag11, avg_time_charging_lag10, avg_time_charging_lag9, avg_time_charging_lag8, avg_time_charging_lag7, avg_time_charging_lag6, avg_time_charging_lag5, avg_time_charging_lag4, avg_time_charging_lag3, avg_time_charging_lag2, avg_time_charging_lag1*

Dealing with imbalanced data to prevent misclassification:

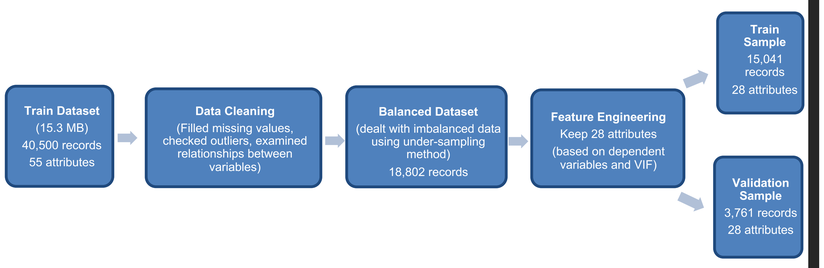
- The number of failed machines is much smaller than the number of not-failed machines. The method used is **under-sampling –** the method downsizes the majority class to balance with the minority class.



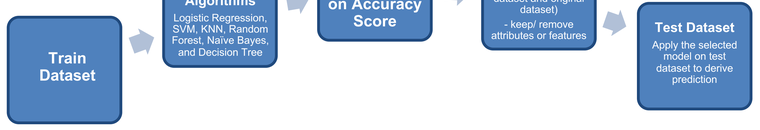
Feature Engineering

- ***charge_cycle_time_below_12*** is not numerical. We transform the variable into a dummy one.
- **Scaling and normalizing** selected features
- **Calculating VIF to check multicollinearity on the selected features.**
  - Variables with VIF > 10 will be excluded from the model.
  - As VIF of ***chargecycles, dischargecycles, avg_volt_change_charging, avg_volt_change_discharging, avg_time_discharging, max_voltage_day, cycle_time, device_life*** are greater than 10, meaning there are highly collinear relationships of these variables to other variables. **Therefore, we considered to drop these features.**

Summary of Data Preparation Steps



Modelling Process



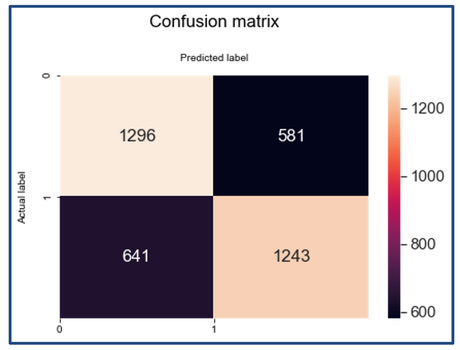
Models & Results

Model	Accuracy	Precision	Recall	F1 Score
Random Forest	0.685988	0.698795	0.687629	0.693167
Naive Bayes	0.669503	0.645999	0.794845	0.712734
Logistic Regression	0.679075	0.642109	0.853608	0.732906
SVM (Linear)	0.678543	0.641502	0.854124	0.732700
Kernel SVM	0.671896	0.641426	0.825258	0.721821
Decision Tree	0.595055	0.613377	0.581443	0.596983
KNeighbors	0.565541	0.569231	0.648454	0.606265

- Basic algorithms (Random Forest, Kernel SVM, Logistic Regression, Naïve Bayes, SVM (Linear), Decision Tree, KNeighbors) were fitted on the given dataset.
- **Random Forest is the algorithms with the highest accuracy score (0.6750)**

Cross Validation:

- Continue with best model **(Random Forest),** from our k-fold Cross Validation results indicate that we would have an accuracy anywhere between **65% to 69%** while running this model on any test set. **This is not the best model as the accuracy score is still low.**



Conclusions

Conclusion

- Out of 4,500 test records, **2,173 meters**, equivalent to **48%,** are **predicted to fail within the next 7 days** when apply our selected model (**Random Forest**) with **accuracy of about 67%**.

Constraints

- Due to limit of time, we cannot try different methods in dealing with imbalanced data. After under-sampling, the size of data has been reduced a lot and this might make the model less predictive.
- There are still a lot of features in the model, and we could not condense those into few of more meaningful ones. We tried PCA to reduce the dimension of the data, but the performance is not as good as the original dataset.

Future improvements

- Implementing different method to deal with imbalanced data.
- Improving feature engineering to select meaningful attributes to the models.
- Running detailed time-series analysis on charging time and discharging time over the span of 14 days.

Acknowledgement

The poster is partially funded Baruch Data Science Challenge 2022 We really appreciate their support and the opportunity that they provided.