Pitney Bowes Baruch Data Challenge – Team 18

Anna Bae, Zafirah Baksh, Guoyi Chen, Deepa Rajareddy, Ridhi Likhi, Janani Ravichandran

Lean Canvas Role Definition

KEY PARTNERS

Servicing team for meters (in-house, client or third party)

KEY ACTIVITIES

Analytics on device monitoring data, service work order assignment and tracking, servicing meters

KEY RESOURCES

Data pipelines and governance, computers and cloud service subscriptions (if needed), spare meter parts

VALUE PROPOSITION

Seamless business operations for mailing meter customers achieved through preventative maintenance activities

CUSTOMER RELATIONSHIPS

Dedicated assistance through lease support contract

CHANNELS

Direct to existing leasing customers, Indirect as advertised service

CUSTOMER SEGMENTS

Leasing customers with medium to high quantity of meters, customers that serve businesses

COST STRUCTURE

Labor from Data Analytics team, Maintenance/Asset Management team, customer support teams

REVENUE STREAMS

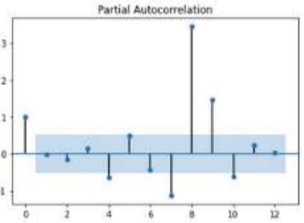
Subscription add-on to leases, automatically included for contracts with high quantity of leased meters and factored into lease price

Interactions Recommended

- Identify customers to assess lowest tolerability for business disruption.
- Discuss meter servicing workload with maintenance and IT teams at PB and with customers to understand capacity to take on additional tasks. If additional support is needed, identify third parties who can support.

Data Understanding

- Generally, positive trend (0.9) between corresponding charging and discharging variables, which means only one needs to be used.
- Negative trend (-0.8) between charge cycles and cycle time, as well as charging time.
- For each variable, we compared the values of the two groups in the training set, fail (red) and not fail (blue), and there was not much of a difference between the two.
- To identify important lag variables, we generated the partial autocorrelation plot for a device for the average time charging variables (lag 1 through 14) and found that lags 7, 8 and 9 were important for predicting future values. We did not generalize this to the entire dataset.



- Large amounts of values were outside of the normal distribution +/- 3 std. deviations which we perceived as outliers but decided not to remove or modify them.

Data Preparation

- Thousands of null values for a few of the lag variables which we replaced with zero (avg_time_charging_lag11-14, avg_time_discharging_lag11-14).
- We changed the data type of dates (Date deployed, Date recorded)

Model Development

We used split validation- trained various models on the training set and tested their performance on the validation set. The accuracy for all models was low. We performed feature engineering (used attributes ‘last_record and ‘date_deployed’ to calculate life of meter in days.) This improved the accuracy of all our models.

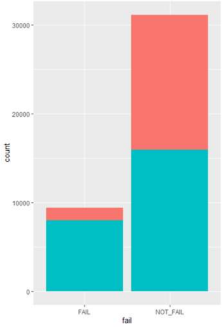
Model Evaluation

There is a trade-off between proactively servicing machines that will fail versus servicing machines unnecessarily (false positives). We would like to only service machines that are in danger of failing soon (within a week) as there is labor and cost required for this task. We aim to identify the minimum number of machines that would fail in order to preserve resources. But if there are too many false negatives then we will not service enough meters and cost our clients money plus our reputation. The F-score best reflects this scenario by finding a balance between the two and addressing the uneven distribution between failing and not failing (roughly 20% versus 80%).

Performance Results

Model	Precision	Recall	F- score	Accuracy
Decision Tree Classifier	0.66	0.38	0.48	0.81
Logistic Regression	0.45	0.025	0.049	0.77
Naive Bayes	0.34	0.74	0.47	0.62
Random Forest	0.71	0.25	0.37	0.81
Gradient Boost	0.66	0.35	0.46	0.81
Adabooster	0.68	0.34	0.46	0.81

We predicted that 588 out of 4500 meters will fail in the next 7 days, using the decision tree model as it has the highest accuracy score and F-score of all the models. An honorable mention is the Naive Bayes model which has the highest recall score of 74%, more than double all the other models. For the decision tree, the most important feature is charge cycle time below 12 - where there is a difference in the values between the fail and not failing groups. For not failing groups the distribution between values true (blue) and false (red) was about 1 to 1 ratio, but for the failing group it was 6:1.


count
FAIL
NOT_FAIL
fail