Baruch Data Challenge with Pitney Bowes

4/28/2022

Presented by: Zi Zeng, Khilola Rustamova, Natalia Kurbakova

I. Business understanding and Design Thinking:

Companies use mailing meters to decrease costs of operating a business and increase the pace of their operations. Pitney Bowes mailing meters play a big part in providing small-scale companies the opportunity to grow at low additional investments as well as assisting organizations of a larger scale with maintaining their customers' satisfaction from fast service. Overall, since Pitney Bowes meters are leased hardware, it is important to ensure that the risk of downtime is mitigated and to provide time-sensitive guidelines/help to businesses on how to maintain the meter better if the device is being flagged as at risk.

This project's main stakeholders are 1) Pitney Bowes, who wants to reduce the risk of down-times of their meters to keep their leadership position in the mailing meter space, and 2) their business clients, who need the mailing meters to simplify their operations to serve their customers faster and who would also benefit from an indication to take steps before the device fails.

From a sample of 40500 meters, we will build a predictive model that will help identify when there is an issue with the meter before it fails

II. Data understanding

First, we started with exploring the train dataset by utilizing the Pandas library to create the DataFrame from the dataset and observe samples that helped us notice common patterns for the devices that did fail. Next, we utilized visualization libraries, such as Matplotlib and Seaborn, to confirm previously mentioned patterns and any relationship or correlation between fields - this would help us manipulate the dataset by adding new columns and determine important predictors for our model in the next steps. The following bullet list describes our findings:

1. For most fields of the original train dataset, the mean value is higher than the median value due to the outliers. This can be explained by different types of meters used and/or different operational capacity of a business.

2. Separating the values of the dataset into two dataframes (first containing the lag variables and second - all of the train DataFrame variables without the lag fields) allowed us to study the distribution of predictors and outliers in them more accurately.

3. The devices that did not fail had longer off-time than the devices that did fail.

4. There is a positive relationship between the age of the battery and the occurrence of downtime for the device.

III. Date preparation

In this part of the project, we will be modifying the dataset by handling missing values, organizing the columns, adding new ones, and manipulating the data type.

To handle the missing values, we decided to drop the lag5 - lag14 fields which were the only columns that contained older data about the (dis-)charging time and rate. This way we will train our model on the complete number of rows from the original dataset without dropping any records.

For data formatting, we modified the ‘Date Deployed’ field values to the format yyyy-mm-dd.

For the derived data, we created one DateFrame with ordered lag columns - ‘trainlag’ - and one without the lag columns - ‘train2’. Afterwards, we created 3 new columns to the train2: AVG_charging_rate, AVG_discharging_rate, age_of_battery (months), and added the fields containing lag_1 to lag_4 back.

IV. Modeling

After modifying the original dataset, we have worked with multiple libraries to, first, rank the importance of each field, and, then, build multiple predictions using different models. Finally, we selected the decision tree as our final model. Here are the steps we took to find the solution:

1. We used the Machine Learning package SKLearn to divide the dataset into two parts: train and test; and scale the predictors.
2. Due to the target variable (fail_7) being categorical, we initially put the train2 dataset through a logistic regression which yielded accuracy of more than 60%;
3. Next, we tried a decision tree model using a Datacamp template that lets us evaluate best parameters (for example, the depth of the tree) and best accuracy for the model from a desired range of parameters. The first node of every tree was the boolean column, charge_cycle_below_12
4. In our decision trees, the percentage of false positives outweigh true positives. We believe the tree is overfitting our data, and finding patterns that don't exist. We move to XGBoost, to see if there is improvement if the trees can learn from previous trees. The discovery of feature importance from this model supports our previous findings that lag3 and 4 are significant.
5. We have also tried an unsupervised learning method using the K-Means algorithm.
6. For the prediction model, we decided on the decision tree model due to 1) its simplicity in terms of visualizing the importance of variables and communicating the rule-based partitioning of data, and 2) the ability to capture non-linear relationships.

V. Evaluation

After the model has been finalized, we imported metrics from SKLearn to show the accuracy of the model which is about 80.9%.

To look at different types of errors that occurred when we ran the model we used the confusion matrix, which indicated that the number of well-functioning devices that were labeled at-risk is higher than the number of failed devices that were not flagged.

VI. Business Conclusions

The predictors that the model uses indicate key features that Pitney Bowes should pay more attention to in terms of maintenance, when companies use their meters. From the train2 dataset, some of the fields that has the most significance were whether the device received the 12 normal charge cycles (boolean predictor), the charging and discharging rate gathered three days prior to April 1st 2021, as well as the average voltage change when charging and discharging. Even though the model rated a lot of devices that didn't fail yet as already failed, this only highlights the proactive nature of utilizing different methods to solve the business problem, and those devices might actually need to be taken better care of during its use.