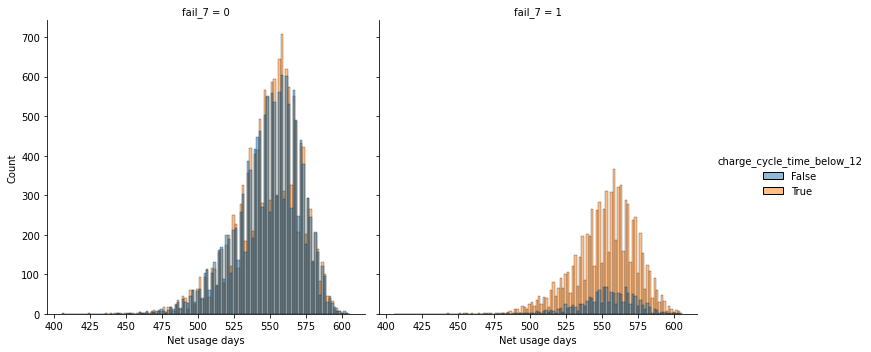
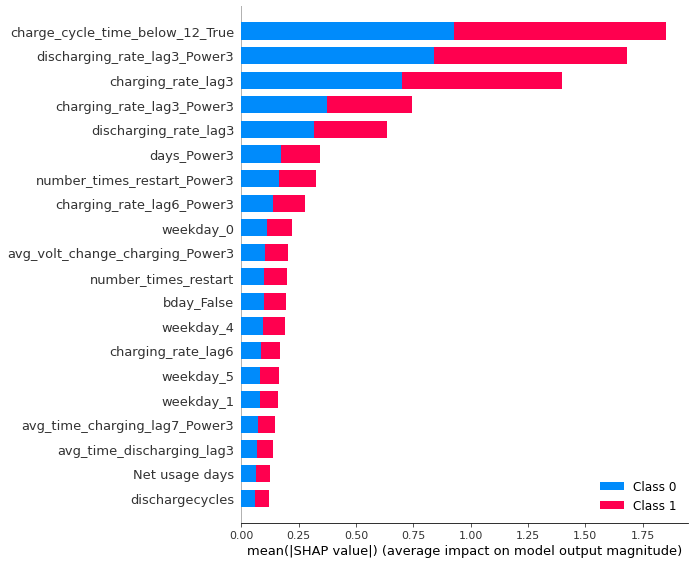
Pitney Bowes Data Challenge 2021 Team 22

**Poster**

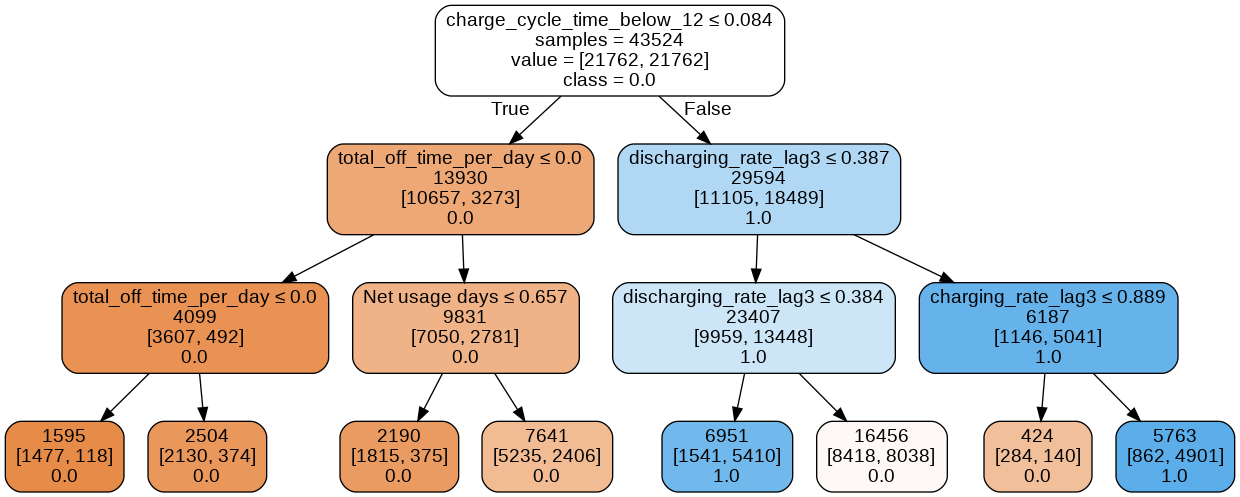
1. **Summary:** 
   1. The best model/model selected is LGBM with an accuracy of 81.09% and F1 score of 51%. Decision Tree is underperforming (accuracy: 79.17%, f1 score: 50.50%). However, it's easier to explain than LGBM. Features: discharging rate, charging rate and charged cycle time below 12 should be paid attention to predict fail meter..
2. **Business Understanding**
   1. Battery Life Time Problem: we must identify the features that affect the battery life of meters so we can learn to predict their behavior which leads to meter failure.
   2. **To assess our model selection, we will focus on model performance metrics: accuracy and F1 score.**
3. **Data Preprocessing**
   1. **Missing Value: we implemented:Mean,Median,Linear regression,KNN and Imputation by previous value on deployed date**

**Final decision/method selected: Linear regression imputation.**

* 1. Outliers: removed and kept outliers, noticed no impact in model performance. 

Final decision/method selected: keep outliers given that there is no domain knowledge regarding them.

* 1. Data Formatting and Z Score Normalization : Yeo-johnson Transformation was applied given that some features are not Gaussian Distribution and reduce the skewness of data.
  2. **SMOTE on unbalanced data: we applied 5 different versions of SMOTE, similar output was received; did not provide significant improvement in model performance.**
  3. PCA: applied Linear PCA and Kernel PCA to reduced dimension in data; did not improve model performance.
  4. Correlation Matrix: strong correlation between avg time charging and avg time discharging from lag 4-14 and variable cycle time.

1. **Feature Engineering**
   1. Technical Methods:
      1. T-test: failed null hypothesis, indicates that pattern in dataset is not strong by chance.
      2. **Polynomial Features: best fit among tested ones was polynomial feature 3**
   2. **Business Methods:**
      1. **Transform DateDeployed and LastRecord into multiple features,such as *usage days, weekday, is businessday*.**
2. **Modelling**
   1. Model comparison with cross validation: 14 models were tested from 3 different model platforms.
   2. **Accuracy and Explainability:**
      1. **The best model/model selected is LGBM with an accuracy of 81.09% and F1 score of 51%**
      2. **Decision Tree is underperforming (accuracy: 79.17%, f1 score: 50.50%). However, is easier to explain than LGBM**
      3. **SHAP explains the most important features that can predict the meter’s failure.**
3. **Business Conclusion & Recommendation:**
   1. Pitney & Bowes should monitor meters during the latest three days of their usage, especially on features: discharging rate, charging rate and charged cycle time below 12. These are the features that strongly determine which meter will possibly fail within the next 7 days.
   2. Pitney & Bowes should also consider tracking temperature from meters, given that it is a critical feature to the battery life of meters and eventually their failure. And customize the models based on business requirement with threshold adjustment. For example, decrease false positive but increase false negative.