Pitney Bowes Data Challenge

Baruch College

Team 8

About us

Chi Nguyen
MSc Statistics

Huong Nguyen
MSc Statistics

Linh Bui
MS Business Analytics

Aizazurrahmantah
MS Business Analytics

Table of contents

Business Understanding

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

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

REVENUE STREAM

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

COST STRUCTURE

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

GOAL

Predict which meters will fail within the next 7 days to reduce down-time risks of meters deployed at Pitney Bowes' customers to avoid any sort of disruption

ML TASK

- Categorization
- Input: 54 attributes providing information regarding charge/discharge, restart times, time-off...
- Output: fail_7

OFFLINE EVALUATION

Apply:
- PCA method to select significant variables
- ML methods for classification (random forest, KNN, Naive Bayes...) to build prediction model

DATA SOURCE

Train set & test set provided by Pitney Bowes

Data Understanding

Train Set

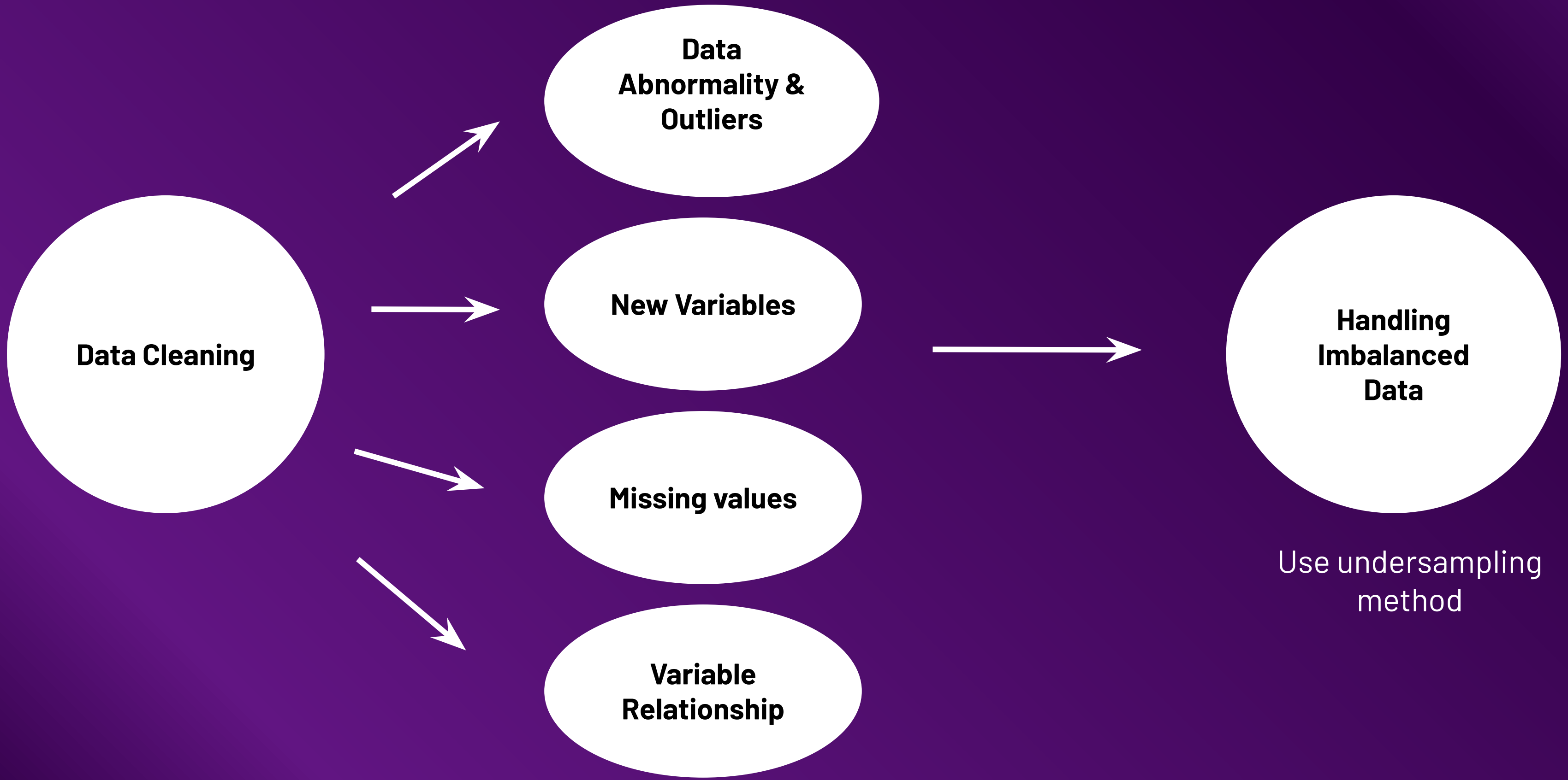
Test Set

40,500 records and 55 attributes

4,500 records and 54 attributes

New added variables: Life_of_device and Total_off_life

Data Preparation



Modelling Process

01 — Prepare Training Data after Feature Engineering

02 — Fitting Classification Algorithms

03 — Apply PCA with data set and fit model

04 — Select model with highest accuracy score

Feature Engineering

Get dummy

charge_cycle_time_below_12 is transformed into a dummy variable.

Check multicollinearity

- Using VIF score to quantify the extent of correlation between one predictor and the other predictors.

- Remove variables with VIF > 10

Scale data

- Standardization scale

Model Fitting

Random Forest
Accuracy: **68.60%**

Logistic Regression
Accuracy: **66.95%**

SVM (Linear)
Accuracy: **67.85%**

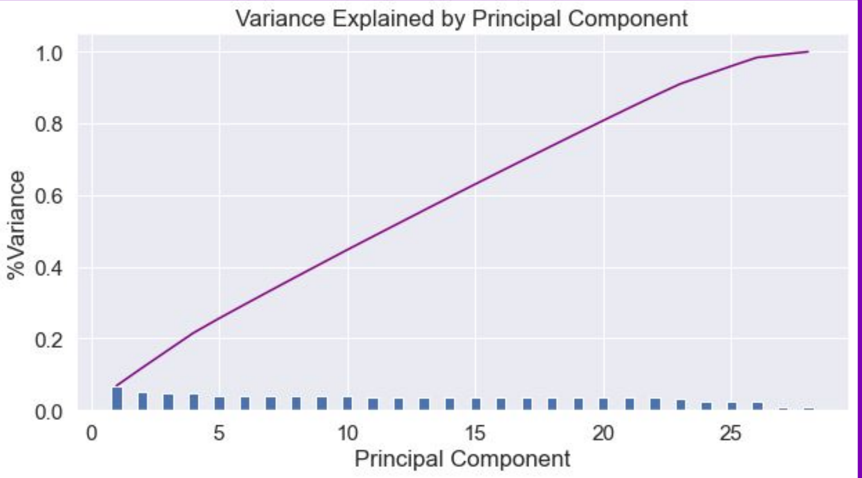
Naive Bayes
Accuracy: **66.96%**

Kernel SVM
Accuracy: **67.19%**

Decision Tree
Accuracy: **59.50%**

Knn
Accuracy: **56.55%**

Principal Component Analysis

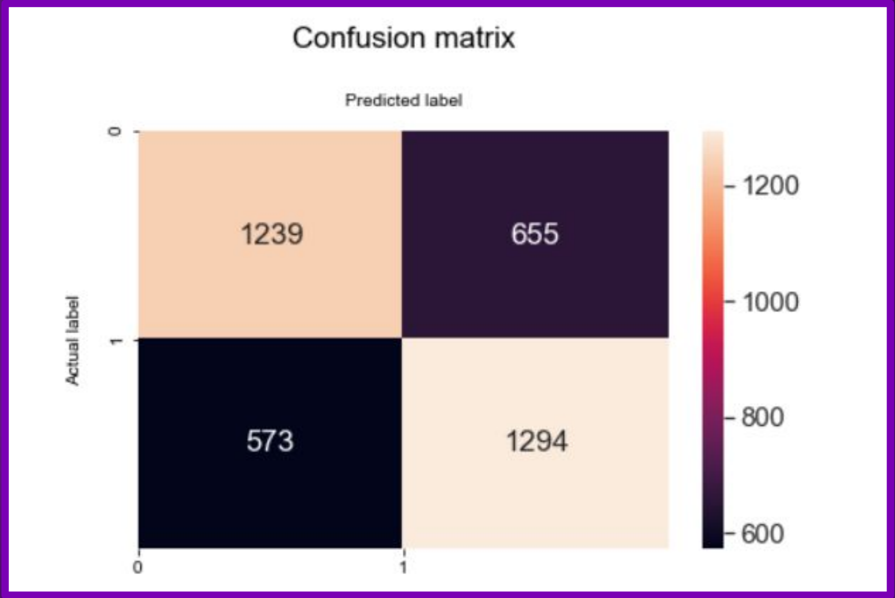


	Model	Accuracy	Precision	Recall	F1 Score
0	Kernel SVM	0.540282	0.531025	0.632566	0.577365
1	Logistic Regression	0.528849	0.524548	0.543653	0.533930
6	Random Forest	0.521138	0.518415	0.497590	0.507789
2	Naive Bayes	0.524860	0.517778	0.623996	0.565946
3	SVM (Linear)	0.521670	0.517764	0.530798	0.524200
4	Decision Tree	0.514225	0.510893	0.502410	0.506616
5	KNeighbors	0.512098	0.507561	0.575254	0.539292

Keep 23 Principal Components as they explain more than 90% variance of the data

The accuracy score is lower than the original data set

Model Selection



Random Forest

k-fold Cross Validation results indicate that we would have an accuracy anywhere between 65% to 69%

Precision: 66.40%
Recall: 69.30%

Improvement

69%

Low Accuracy Score

50%

Data Reduced After Balance

48%

Failed Machine - Unrealistic

Recommendations

TO DO LIST	
Task 1	Dealing with imbalanced data without losing many observations
Task 2	Better feature engineering to have higher accuracy score
Task 3	Running model to specific types of machines to increase accuracy

Thanks!