Data Challenge

APS Failure at Scania Trucks

Goal

Minimize maintenance costs of the air pressure system (APS) of Scania trucks

Input data

- Training set : 60’000 × 171
  - col 1 : target feature
    class = neg || pos :
    ’neg’ - a truck with failures for components not related to the APS
    ’pos’ - component failures for a specific component of the APS system
  - cols 2-171: 170 numeric features, 70 of which belong to 7 histograms with 10 bins each
- Test set : 16’000 × 171

Challenge metric

$$f = 10 \times Err_{type_I} + 500 \times Err_{type_{II}}$$

Methodology

- Missing values treatment : 8.33%
  - Remove the variables which contain more than 20% of NaNs
  - Impute the rest of missing values with median
  - Reduced DS : 60’000 × 147
- Sump up the variables that represent each histogram
  - Reduced DS : 60’000 × 84
- Explore variation within variables, outlier detection
  - Split the data into two DS, one of which has only negative observations and another - only positive
  - Compute the whiskers, i.e. 1.5 × IQR above and below 3rd and 1st quartiles
  - Replace ’positive’ extreme values with the median of posititves
  - Remove ’negative’ extreme values
  - Reduced DS : 18’075 × 84
  - Remove the variables, variance of which is equal to 0
- Explore correlation between variables, feature significance and selection
  - PCA, Variable and Individual factor maps
  - Linear Regression Model
  - Random Forest Model
  - Reduced DS : 18’075 × 52

Methodology

- Imbalanced data
  - Oversampling
  - Undersampling
- Prediction model
  - Divide DS into training and validation sets, 75% and 25%
  - Tune the cost, c, parameter in **SVM** using **CV** technique
  - Repeat each experiment n times, build confusion matrices and compute the average of evaluation metrics (Error, Precision, Recall, F_1)
- Predict the test samples
  - Transform DS to the same form as training DS
  - New test DS : 16'000 × 52
  - Feed built model with transformed test DS