APS Failure on Scania trucks

Building a predictive tool

Author: Juan Carlos

3rd November 2021

Agenda

- Overview of data
- How data has been manipulated
- Modelling process
- Selecting an algorithm
- Model understanding
- Recommendations

Overview of data

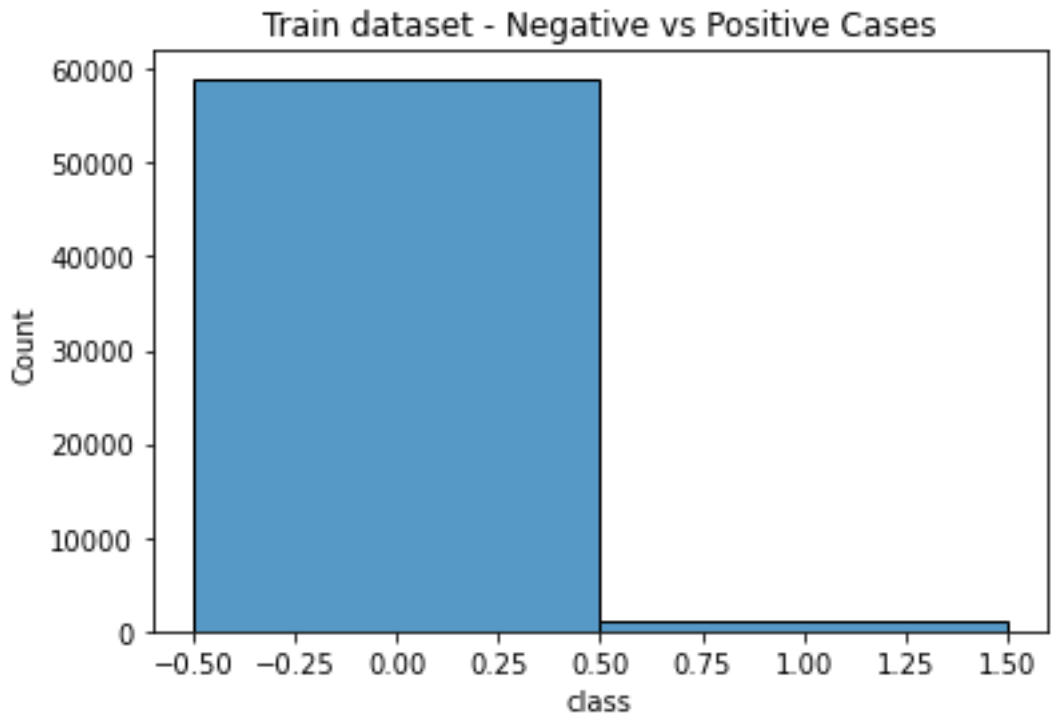
- All the data points are composed of data coming from sensors on Scania trucks and other sources.
- Records are related to broken trucks, all of them.
- It is a binary dataset; there are two classes:

1. Positive class means failures related to the Air Pressure System (APS).
2. Negative class also represents broken trucks, but these failures are not related to the APS.

- Correct classification is essential, since cost of late mechanical check on APS failure is $500, while unnecessary checks cost $10 only.

Overview of data

- The dataset contains 76,000 records in total.
- Just around 1% of total records belong to the positive class.
- There are 171 variables (the class we want to predict + 170 other attributes).



Data Glacier
Your Deep Learning Partner

Data Manipulation

These are the steps followed in order to get a useful dataset which is ready to be analyzed and model the data.

1. Exploring the data files (train and test sets).
2. Manipulating the target variable: assigning 0s and 1s to negative and positive classes respectively; counting; converting to integer data type.
3. Handling missing values: imputing the mean to null values.

Data Manipulation

4. Detecting possible outliers using the z-score: values 3 standard deviations or further from the mean, were considered outliers.
5. Outliers presence is one of the reasons why I decided to standardized the dataset.
6. The final dataset is composed of standardized values for 171 attributes (the target + 170 other variables).

Modelling the data

- As mentioned at the beginning, only around 1% of the target is positive class; we clearly have an imbalanced dataset.
- I decided to use SMOTE algorithm to resample and get a balanced dataset.
- Finally, I got half of negative class values and half the positive class ones.

Selecting an algorithm

I decided to apply four learning algorithms over the data:

1. Support Vector Classifier
2. K Nearest Neighbors
3. Logistic Regression
4. Random Forest

Selecting an algorithm

Here, we are interested in getting a higher **recall** (this is to reduce false negatives, as the cost is x50 higher than predicting a positive case when it is not).

The **Hyperparameter Tuning** technique that was used showed that **KNN** was one of the best options, based on recall obtained and speed of training, using 5 neighbors as parameter.

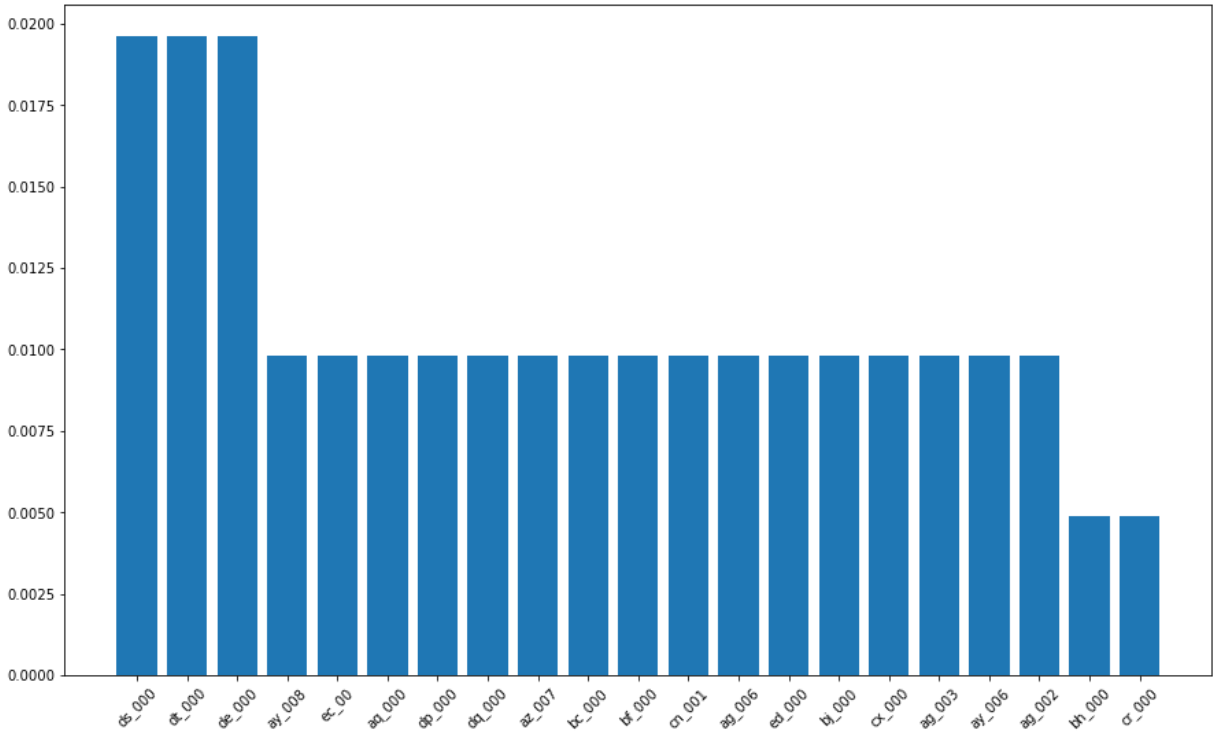
Model insights

I performed **Permutation Importance** to have an approximation to predictive power of variables and this is the result:

21 features have been found to have some influence on the target, while measuring recall, which is the metric we are most interested in this time.

Model insights

Variables that seem to be more important to the predicted class



Data Glacier
Your Deep Learning Partner

General recommendations

- Variables that seem to be more important to the target deserve our attention: technicians could mainly focus on them as they may weigh heavier on the final failure cause (either the APS or other system).
- Model can be improved in the future with feature engineering to get a higher accuracy. The cost of making unnecessary checks is big (x50 times).

Thank You!