Arie Kfiri

Data Science Project – Plan & Goal

Summary

Our client is a firm that products metal parts for machines in manufacturer and industrial fields. The production process is complicated, but precisely for this reason, it can potentially generate data that will help streamline it itself. Heat sensors, noise sensors and other sensors were built upon the production machines, and supply us the main data that we will use in our project.

Data

This data from the sensors, are organized by samples taken every 1/25 of a second, in a table named “Session_data”. Most of the columns in this table are the various indications that we get from the sensors.

One column in this table that doesn’t fit this definition is the “SESSION_INFO_ID” that connect us to the metadata table, named “Session_info”. In this table we have the information of each machine, including its tools, battery charged level, and more.

With the data from the mentioned two tables, we want to have the ability to reach conclusions and insights that will help our client. the main content that interests our client is embodied in table named “Labels”. In this table we discover for each sample from “Session_data” weather it reflects anomaly or not, and weather it includes touch between some of the metal tools. This information will connect us to our goal in this project.

Goal

Our goal is predicting the tool touch and anomaly labels, for each sample.

Predicting anomalies can be very important to our client, since anomalies can reflect potential to damage in their very expansive machines. Early detection of anomaly enables taking care of it before real damage occurs, a bit like early detection of a disease.

Predicting tool touch is also important, because it’s important factor to other calculations, relevant to the production process. It helps in the calculation of wear, informs operational metrics like "utilization" of the machine etc.

Model Evaluation metrics

The two labels for prediction are very different one from the other, in many ways. Two differences are very significant, while choosing evaluation method:

1. Regarding the anomaly label, the data is very imbalanced data. We can assume that samples that will be considered as anomal, will be very rare, certainly much less than one percent. On the other hand, the touch label in balanced; bases on the data in our hands, the distribution between its values is about 53% and 47%.
2. Regarding the anomaly label, the potential damage of “false negative” mistake is much larger than the potential damage of “false positive”. In the case of the touch label, we can’t point out a big difference in the importance of each kind of mistake.

These two differences obligate us to choose different evaluation metric for each label. Therefore, for the anomaly label we will choose recall/sensitivity metric, which focus finding as much as possible of the anomal samples. The touch label, on the other hand, requires more general metric, which evaluate mistakes from all kinds equally. Therefore, I will choose the F1 score metric for this label. In both cases, we will aspire very high accuracy label, when in the case of the anomaly label our inspirations should be even higher. At this point, we will set as a goal accuracy of 85% for the touch label, and 90% for the anomaly label.

Work plan

- Further exploration of the data: understanding the variables, correlations, differences, impact on the labels etc.
- Fixing problems in the data - handling outliers and nulls as needed.
- In the case of the anomaly label – choosing method to handle the imbalanced data (oversampling, undersampling, SMOTE, etc.)
- As needed – adding features and feature selection
- Checking models and choose the best predicting model for each label
- Production: assuming that not all our client staff are developers, and anyway all of them will prefer the work to be as comfortable as possible, we would like the end product to be inference model at a hosted endpoint. Also, because of the potential urgency of discovering anomal samples, I would like to create system which run the anomaly model consistently and alert (by email?) every time anomaly discovered.