Summary

This project was done as part of 2 frameworks:

1. As a final project in a training program for data scientists and analysts at Bar Ilan University.
2. As part of an internship at the consulting company “Modernize Solutions”, as part of work done for a client of the company.

This folder is the public version of the work, used for the final project, and a number of secretly related points have been omitted from it. Along the way I navigated the project and the decisions during it according to the requirements of the final project, but when different needs arose from the staff at Modernize Solutions and / or the client, I deviated from the original guidelines.

The customer for whom the work was done is a company that manufactures chips for factories that manufacture metal spare parts for ships, planes and more. The same chips include sensors that transmit information about the production process in the various machines over time, including data about the heat, noise and frequency of movement in the various machines and tools.

The Data

All the relevant data, including the data from the sensors, are concentrated in the SnowFlake cloud service, in 13 tables. Details of the tables are in the "ERD.png" file in this folder. However, for the current project, it was necessary to use 3 of the tables:

1. Session_data - where the data from the sensors are concentrated: the heat, the noise and the frequency of movement in the various machines and tools, One sample per millisecond.
2. Session_info - where the information about each of the machines in the factory is concentrated, including information about the tool type of each machine, its battery charge level, and more. Because each of the samples is related to a specific machine, the table is linked to the previous table.
3. Labels - In this table we find for each sample from Session_data, whether it has touch between a tool and metal parts, and whether an anomaly has been detected in it. As we will soon see, the subject of the touch between the tool and the metal parts is the point we were asked to predict. In addition - there are several columns in this table that have been tagged after the fact, and are related to each of the samples. The data from these columns can not be used for prediction, as it is information collected retrospectively, and will not be available in real time for predication in the future. In some cases, these are data that are an outgrowth of the predication, meaning that the use of the data would have constituted the requested assumption.

   Wider detail of the various variables of the three tables can be found in the "Data_Dictionary.xlsx" file in this folder.

Goal

Our goal is predicting the tool touch label, for each sample. A priori, there was a desire to make a prediction also for the question of anomalies, but at this point this goal was dropped

due to lack of data. Predicting tool touch is important to our client, 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.

The structure of the files and folders in the project

As mentioned, in this folder (the main folder of the project), besides this file there are two files related to understanding the data:

1. ERD.png lists the tables, what features there are in each table, and the relationship between the tables.
2. Data_Dictionary.xlsx - Dictionary of Features; Explains the features of the 3 tables relevant to the project.

Two internal folders where the project's code and data are stored:

1. eda, preprocession, baseline model - This folder, as its name implies, contains code notebooks, data and reports related to the topics included in its name:
   - eda
   - preprocession
   - baseline model
2. training and select models - This folder deals with model training and the selection of the final model for the project.

In these 2 folders, some of the data and files saved while working for backup purposes, were not uploaded to GitHub. This is due to a GitHub error message related to the size of the files. However, the process I went through with these files can be seen in the code notebooks.

Code notebooks and work steps

A priori, we referred to the EDA phase, along with preparing the data and running a baseline model as one phase. The notebook "EDA and Baseline Model.ipynb" (along with the small notebook "mechkar.ipynb") together constitute my first attempt to perform this step. mechkar.ipynb contributes by producing an automated report of the "Mechkar" package in R.

After submitting these files, I received feedback with additional instructions, related to the preparation of the data. Their execution is in the notebook "preprocession and baseline model.ipynb", alongside with details about the changes I was asked to make.

At this point the data was ready and What was left was to do was training and selecting models. This is done in the relevant folder, in the "model selection.ipynb" file.

Next Steps

The deadline for submitting the project has arrived, and I have a nice project in my hands with complex analysis and excellent results. However, there is still things to improve, and a number of steps are expected as part of the internship, and those are hopefully to be done and updated as well on Github soon:

- Run LSTM model for data. Since this is a successful model for Time Series, there is a thought that it might be able to fit more than classic machine learning models. This requires separate examination and study, and will hopefully enrich the project in retrospect.
- Build a predictive model also for the anomalies label, when there will be enough data for that.