Sprint 0

The Problem Area

My area of interest is within the manufacturing domain. More specifically, I am interested in leveraging production line data to help improve overall quality and production output by predicting the occurrence of internal failures. Improvements in this area of manufacturing can occur when optimizing maintenance strategies to minimize downtime, automating decision making, and utilizing data to predict and make informed decisions. A common area of concern when manufacturing any product in mass, is being able to consistently catch defective products that would fail quality checks. An alternative perspective on this issue is - is there is an effective way to use data to predict when a product will fail? This discovery will help facilities minimize waste, improve quality standards, and increase production output.

The User

In this moment in time, all leading manufacturers are looking for cutting edge technology to find the smallest of improvements that would make their manufacturing process more efficient. On the other hand, based off of research and my experience within the industry, smaller manufacturing facilities do not have access or the skillset to use their own data to optimize their manufacturing process. Small - mid sized manufacturing companies will benefit from my project because it will highlight how rich their data is and allow them to improve revenue from a human capital perspective, production output, and their ability to monitor and control quality.

The Big Idea

The most common machine learning technique that can be used to detect internal failures within the manufacturing process is numerical prediction. However, next point forecast may have potential use cases. For example, using next point forecasting can allow us to predict when the next failure will occur as opposed to predicting if the product has failed or not. A model that uses next point forecast for manufacturing would be best to develop optimal predictive maintenance strategies, not necessarily within the scope of this project.
Numerical prediction on the other hand, will give us the tools needed to classify whether or not a failure will occur. Given that the scope of this project is predicting when internal failures will occur, we can leverage classification tools like, XGBoost and logistic regression to categories failures.

The Impact

With this project I am looking to learn what it takes to implement machine learning algorithms in an environment where I can make mistakes and be mentored. In reality, I am looking to help small - medium sized manufacturing facilities utilize their data to optimize their overall production. The societal impact is simply breaking down the barriers to entry for smaller facilities and allowing them to better utilize their rich data. The business value can come from improving

production output and minimizing waste. For example, when I worked with Air Molded Plastics from 2022 - 2023, a manufacturing facility that manufactured large tractor toys for kids. I noticed that they had a reactive approach to failures given that they were treated as problems that need to be solved after the product came out visibly defective. This kind of project can help me integrate proactive solutions for these smaller manufacturing facilities, improving their overall quality and production output.

The Data

Data Set 1:
https://www.kaggle.com/c/bosch-production-line-performance/overview

Bosch Production Line Data:
The dataset includes 6 different csv files broken down into 2 different categories. A training set and a test set. At a high level:

The three sets of data are:

(1) Numeric Data
(2) Categorical Data
(3) Date Data
The dataset contains fields that are structured in the following format:
L0_S00_F0000

Where:

L is the production line
S is the Station on the line
F is the feature number - (feature being measured)

Using the test training set we must train a model to predict when a failure will occur using the test set.

Data Set 2:
https://www.kaggle.com/competitions/mercedes-benz-greener-manufacturing/overview

Mercedez Benz greener manufacturing data:
We are looking to assess the different features that will be tested and understand which testing sequence will lead to faster testing period, leading to lower carbon dioxide emissions from the reduced time stress testing.

The files provided for this dataset is a training set with an anonymized set of variables, that represent a feature in a Mercedes car. There are numeric, binary, and categorical variables in the dataset.

The Alternative

I hope it doesn't come down to this, but I am a huge fan of the UFC (ultimate fighting championship). It is a physical fighting sport where 2 fighters, meet in an octagon and fight for 3 or 5 rounds. There are 3 outcomes in a fight:
Outcome 1: ko/tko/doctor stoppage
Outcome 2: Submission
Outcome 3: Judges Decision
In order to make the most money out of betting, predicting a ko/tko/doctor stoppage or a submission is the most profitable. Using fight data like significant strikes, number of takedowns/ number of takedowns stopped, is there a way that we can predict whether a fight will end in outcome 1 or outcome 2 and where should we bet our money.