**Summary:**

Data from 20 total sensors was collected for 100 hot finishing mill (hfm) runs that ended in failure. The sensors were broken up into two groups, one containing 8 sensors and one containing 12 sensors. Initial data exploration revealed 6 sensors that provided no information and these were omitted for model training and analysis. Three separate datasets were created, one for the 10 column dataset, one for the 14 column dataset, and one that combined the two. Three models were then trained on these three datasets independently creating 9 unique models in total. The three models were a ridge regression, K nearest neighbors model, and a gradient boosted regression tree. Of the tree types of models the ridge regression generally underperformed compared to the KNN and gradient boosted models, but would be the easiest to operationalize and deploy. Of all 9 models trained the gradient boosted tree trained on the large dataset that combined all sensors had the lowest error and highest coefficient of determination.

**Assumptions:**

* no\_of\_days means days up or online, not how many days remain until an observed failure happened. I will assume that a failure happened the day after the highest no\_of\_days value for each given run. For instance day 192 was the last no\_of\_days for the first hfm run I will assume a failure happened on day 193 before the reading could be taken.
* Each observation is stand alone and does not show autocorrelation with previous days readings. In short each observation is a snapshot in time
* The 10 column and 14 column datasets were collected at the same time. The hfm run numbers and duration in days align between the two datasets so we can assume that the two datasets can be combined into one large dataset if necessary.

**Data Exploration:**

During initial data exploration it was noticed that the hfm runs between the 10 column dataset and 14 column dataset aligned perfectly in both labeling and duration in days. This was the driver for assumption 3 above. Looking at sensor data showed a large range in nominal values for the sensors. For example, sensor 13 trended around 8150 while sensor 14 trended around 8. This large discrepancy in range drove the decision to normalize training data — more on that later.

Viewing plots showed that many sensors sere trending in one direction or the other prior to failure and some sensors also showed no change in value through all data collection, even across hfm runs. The two plots below highlight one sensor, sensor 13, with interesting characteristics and another sensor, sensor 17 that showed no change.

A graph showing different colored lines

Description automatically generated

A screen shot of a graph

Description automatically generated

**Data Cleaning and Transformation:**

The previous visual observation that some sensors never changed value was numerically verified with a min and max across all data for those sensors. Each of the suspect sensor showed 0 or 0.01 units change across all hfm runs. Since this would provide no information to models during training these sensors were removed from the datasets. The sensors removed were 17, 4, 5, 9, 15, and 18.

Based on the observation for the large discrepancy in sensor nominal readings each sensor was normalized. This prevents sensors that have high nominal values from dominating model fitting.

Since we assume that the two datasets can be aligned based on hfm run and no\_of\_days a third dataset was created that combined all valid sensors from the 10 column dataset and 14 column dataset into one large dataset.

There was also a new variable created that was days\_until\_failure because that is the target we are actually striving to predict. This variable is simply the no\_of\_days reversed for each hfm run since we are assuming that all hfm runs ended in failure the next day.

Once all three datasets were created they were split 80/20 into a training and testing sets.

**Baseline Model:**

**Improvements Using a More “State Aware” Model:**

**Testing a More Complex Tree Based Model:**

**Deployment:**