Bosch Production Line Performance
Capstone 1 – Data Wrangling

Tom Widdows

The data source for this project was a past Kaggle competition. The data represents anonymized measurements of a part as it moves through production line(s). Each feature heading is coded to include the production line, the station on that line and a feature number. For example, feature column header 'L1_S17_F4548' would represent line 1, station 17 and feature 4548. The raw data has been separated into three comma delimited files representing numerical, date and categorical information.

The dataset has over 4,000 anonymized features. Because of the size and number of data elements, the dataset required reshaping for cleaning and future Exploratory Data Analysis (EDA). I split each CSV file (numeric, date and categorical) into 10 equally sized files (30 files total) so I had the memory and processing power to read and manipulate the data. Utilizing Pandas melt, I reshaped the data in each file by transforming the column headers into rows ultimately making the tables significantly narrower and longer. During this process, I added three columns named Line, Station and Feature and populated the columns with data from the column header. For example, a column header of 'L1_S17_F4548' would result in the Line, Station and Feature columns being populated with 1, 17 and 4548 respectively. Finally, I concatenated the reshaped data and stored the result as a new CSV files (numeric, date and categorical) .

The numeric and date data consists of approximately 80% missing values and the categorical data has greater than 99% missing values. Because of the modified data structure, I was able to easily drop records with null values enabling me to only work with populated data. Since the values of the measurements are normalized, I ultimately plan to fill any missing values with zero. For categorical data, I plan on utilizing One-Hot Encoding.

Because of having anonymized features and the volume of data, I did not detect any outliers. In addition, because of the raw volume of data, it is unlikely any outliers (unless extreme) would impact the ultimate model performance.