SSR Final Presentation Words

Data Exploration and Failure Prediction of SSR-Mining Trucks

Introduction:

SSR relies heavily on hauling trucks to transport material. This can be up to 200,000 tonnes per day making them integral to operations at a mine. Breakdowns of these trucks can be costly for example: last year, 41 % of the Marigold total direct cost went towards maintenance. Breakdowns also jeopardize the efficiency of operations at a mine. Therefore, decreasing the unexpected maintenance events associated with a truck would financially and operationally benefit the company as maintenance before a breakdown would be a lower cost preventative measure.

The Problem:

Predicting unscheduled component failures of haul trucks before they happen based on alarm data streamed from truck sensors.

The Data:

The slide below shows where the data comes from and what information each system collects. The end result was the collection of 50GB of data for the Hitachi EH500 ACII fleet.

Initial Data Exploration:

Because of the large amount of data we began by looking at certain tables to see if anything could be learned before developing a model. We focused on TruckMachineData and WorkOrders tables.

Below begins the summation of failures and alarms associated with all trucks. An alarm is raised by a sensor measuring a parameter of a truck (tire pressure, coolant temperature, etc) associated with some mechanical part. When this alarm is raised, the parameter measurement is recorded along with its alarm level. An alarm has three levels: 1,2,3. In the case, alarm level 1 is critical and is the most dangerous alarm.

Furthermore, If we consider the number of level 2 and level 3 alarms associated with trucks, it starts to make the number of breakdowns associated with each truck make more sense.

Based on the graphs above, it may be indicative that trucks with higher level 2 alarms have a larger number of breakdowns. This may be indicative of a slow build-up in increasing level of warnings that may pre-empt an unscheduled breakdown.

Further consideration was given for the temperature and pressure profiles for all Hitachi truck 478 in the fleet to analyze any trends.

It should also be noted that all analysis done from this point on were on truck 478.

The GPS locations of these alarm events are also shown below in a heat map indicating a larger frequency of events with hotter colours.

There is more work order before November 2017 (around) and very few after that time period. This was found for all trucks within the fleet indicating a turnaround in maintenance issues. This was confirmed by SSR as the trucks started to go for regular maintenance more often.

Despite inevitable holes in the data, an attempt was made to smooth the data of outliers and make it more continuous for truck 478. This was done by averaging all data taken for one parameter in a day. Thus, we get a matrix where we have 365 rows (as we have averaged over each day), and 41 columns (as there are 41 different parameters that are measured). The matrix still had many NaN values, so columns that had more than a certain number where then filtered from the matrix. In this case, columns that had more than 260 NaN values were removed. Then, the correlation matrix was produced below between the remaining parameters.

At this point, we considered predicting unscheduled maintenance events based on linear combinations of the values of the parameters and then taking the sigmoid function of that value. In other words, preforming a logistic regression on the matrix constructed earlier. The labels for each entry of that matrix will be a zero (no unscheduled maintenance event that day) or a one (unscheduled maintenance event). Below, we construct these labels using the work orders issued for a day as a work order means an unscheduled maintenance event (which we have filtered for truck 478 and unscheduled work orders).

Some of the parameters on the X-matrix where removed to achieve a better fit of the data which was found experimentally to be about 13 useful parameters. These also happened to be the most continuous parameters that the systems measuring them were able to give.

Another important note is that the X-matrix was standardized by the mean and variance. When performing the regression, the regular parameters as well as quadratic parameters and interactions. In other words, the tire pressure, the tire pressure squared, and the tire pressured multiplied by some other parameter was also used in the regression.

The code above was ran 20 times, with the y labels and X-matrix shuffled, and with a training size that is 60% of the data and a test set that is 40% of the data for 1 truck. The accuracy of this model was found to be 74.45%.

The failing of the model comes from the fact that for the entire data set for this truck, unscheduled maintenance events only make up 22%. Therefore, If the model has simply guessed all zero labels, it would have a 78% success rate for this data set outperforming the actual regression. This may be due to the incompleteness of the data (many Nan values) or the fact that the model doesn’t really account for the fact that it is a time series. Logistic regression assumes that the data is independently and identically distributed but for a time series this is not necessarily true.

Future areas of interest can consider the potential clustering of these alarms between different parameters and finding out where failure zones are. Another area of consideration is in the lead up to an unscheduled maintenance event in the time series data. The models we considered did not really take these into account. The concepts talked about here are illustrated below.

For future work considerations, based on the fact that there are negative and positive correlations between different parameters as shown by the correlation matrix, principal component analysis could be used as a dimensionality reduction tactic. It can also show the contrasts between which parameters are more important for explaining the variation of the data.