**Steckel Mill Downtime Prediction**

The main task of this challenge is to identify the downtime of the Asset - HAGC – Hydraulic System Supply (Asset 544-630).In order to build the model,I have started doing data analysis of the datasets given.The initial data analysis will not only help in understanding the data but also in picking up the right model.I have performed all the data analysis in python.We have three main datasets which consists of work orders,Indicator Reading and Downtime.

Work order consists of Planned and Unplanned maintenance information.Each entry in this table is identified by the work order.I am assuming the field “Requested Completion Date” as the date when downtime has occurred and a request is raised for unplanned maintenance .Below are the following analysis based on this dataset.

The total number of planned and unplanned maintenance are as below

|  |  |
| --- | --- |
| **Work Classification** | **Counts** |
| Planned | 542 |
| Unplanned | 155 |

The Mean downtime cost/time for planned and unplanned maintenance-

|  |  |
| --- | --- |
| **Work Classification** | **Mean Downtime Cost** |
| Planned | 712.766125$ |
| Unplanned | 1972.706065$ |

This will be useful while performing **cost benefit analysis** of the Models

Next looking the time frame of Work orders data we get the below results:

|  |  |
| --- | --- |
| **2012** | 377 |
| **2011** | 291 |
| **2013** | 44 |
| **2010** | 3 |

We can clearly see that most of the work-order data we have is of the year 2012.We have very few data from the year 2010 and 2013.And below is the start and end of this work orders time line

first 2010-07-07 06:08:00  
last 2013-02-04 17:21:00

The Indicator Reading dataset consists of all the asset reading data on daily basis.The most important column in this dataset is Indicator name and Indicator reading,These two should be combined together to get the features to fit the prediction model.Below are the list of unique Indicator names

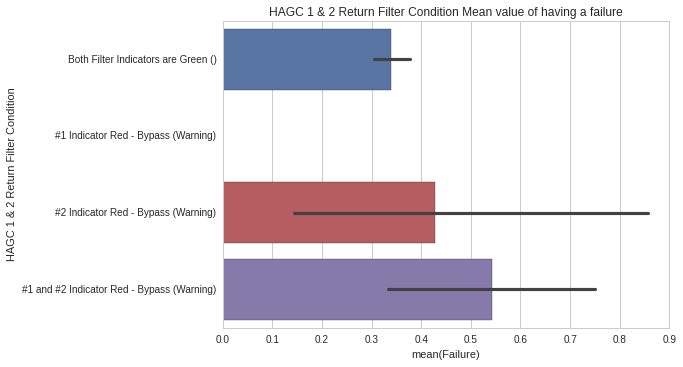
1. HAGC Hydraulic Fluid', 'HAGC Total Weekly Fluid Loss
2. HAGC 3 Day Average Fluid Loss
3. HAGC Fluid LOSS Calculation (E Bin, W Bin, Refill Bins)
4. HAGC Hydraulic Tank Level
5. HAGC Pump #8 Filter Condition
6. HAGC Pump #7 Filter Condition
7. HAGC Pump #6 Filter Condition
8. HAGC Pump #5 Filter Condition
9. HAGC Side Stream Filter Condition
10. HAGC Hydraulic tank CCJensen
11. HAGC Pump #1 Filter Condition
12. HAGC Pump #2 Filter Condition
13. HAGC Pump #3 Filter Condition
14. HAGC Hydraulic Fluid Temperature
15. HAGC 7 & 8 Return Filter Condition
16. HAGC 3 &4 Return Filter Condition
17. HAGC 1 & 2 Return Filter Condition
18. HAGC Hydraulic UR Bin CCJensen
19. HAGC Pump #4 Filter Condition
20. HAGC Total Monthly Fluid Loss

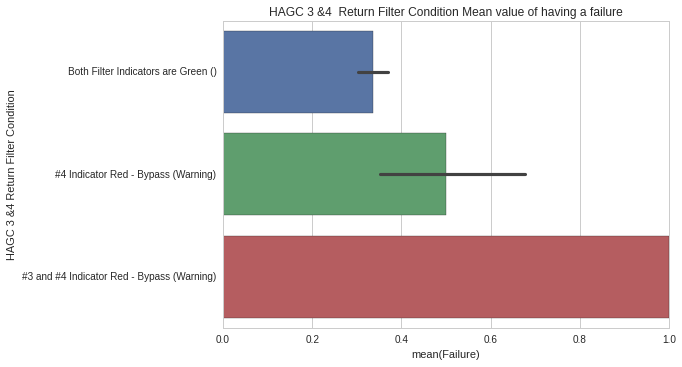
All the filter conditions are categorical variables and rest of the Indicators like fluid loss,temperature and tank level are ordinal.The “Date and Time Collected” column is used to merge with work orders data to fetch on which day downtimes has occurred.

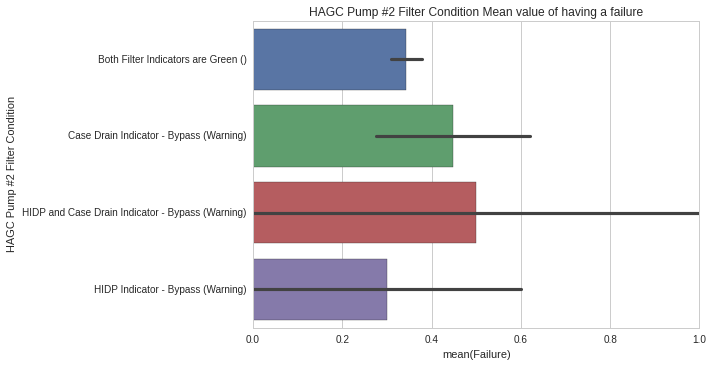
Below are few of the visualizations which helps us in understanding these Variables.

The below bar plots show what is the probability of having a failure if the Indicator Reading are any of the shown categorical variables.For example for HAGC 3 & 4 Filter Condition the probability of having a failure is equal to 1,if the reading says 3&4 Indicator Red bypass warning.And for HAGC 1 & 2 Filter condition the probability of having a failure is zero if its in Indicator red bypass warning.

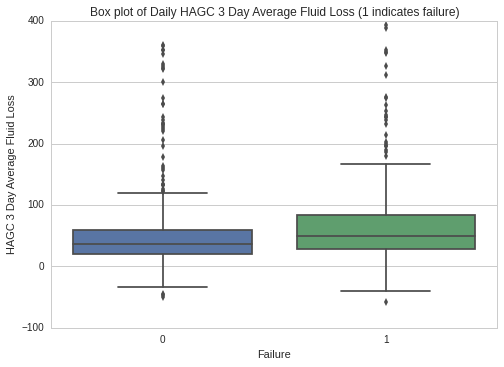
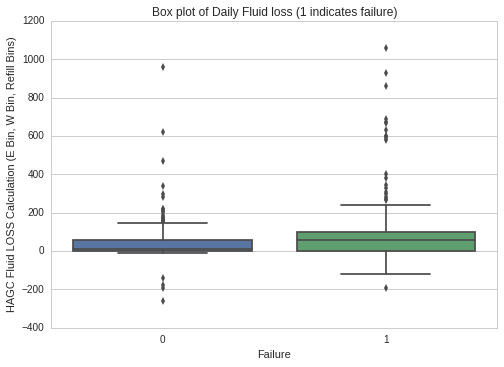
I have visualized these kind of plots for all the filter conditions,By looking at these plots we get a fare estimate of the predicting power of these variables.





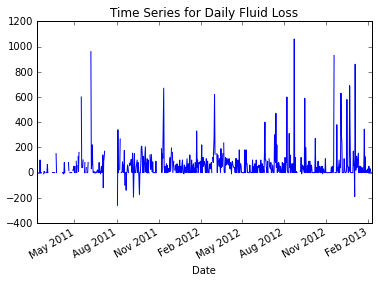


For variables with ordinal data,I am visualizing using the side by side boxplots to know their predicting power.

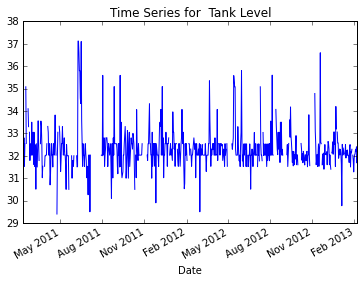
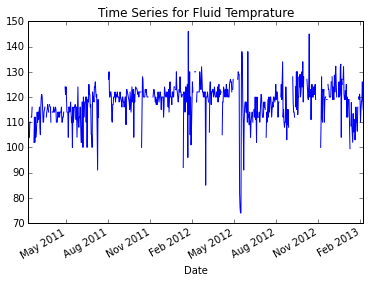


From these box plots we can infer that the median of fluid loss is certainly higher in case of failure,and also it shows more variance in case of failure.

In order to fill the null values of the fluid loss,temperature and tank level,We need to visualize how the data is distributed.Below are time series plots for these ordinal variables



We can see for the Fluid loss the values are reverting down to mean,hence I am filling the nulls with mean value.



For Tank level and temperature the values looks to be randomly fluctuating,hence to fill the nulls for these to variables am using time-series interpolation method.And from the above plots we can also see the start and end dates of the indicator reading available in the given dataset.

The Downtime dataset consists all the information related to asset downtime,below are the analysis of why downtime occurs and their mean costs and durations.

|  |  |  |
| --- | --- | --- |
| **Downtime Reason** | **Downtime Costs** | **Production Downtime Duration(Mins)** |
| **Automation** | 11334.392222 | 50.277778 |
| **Electrical** | 24967.816667 | 118.666667 |
| **Hydraulic** | 48010.380000 | 174.900000 |
| **Mechanical** | 33245.607500 | 150.625000 |
| **Operational** | 1219.980000 | 6.000000 |
| **Production** | 15379.100000 | 41.000000 |

We can see from above analysis Hydraulic downtime costs and duration is much higher than any other.When I looked for the time frame of the downtime dataset,we get below results

|  |  |
| --- | --- |
| **2012** | 59 |
| **2011** | 3 |
| **2013** | 3 |

The Downtime data belongs to the year 2012.I tried matching the work order number in Downtime dataset and Work order data set,But there are only 3 work order numbers which match in both datasets.

**Data Modeling & Transformation**

* The main task of transformation is to merge the downtime dates with its indicator reading during that day.
* Since we are trying to predict the failure of the asset 48hrs prior to the actual failure,I am pulling the date two days before failure and predicting the failure based on the indicator reading on that day.
* I am choosing the downtime dataset to get the failure dates as it seems more accurate to me.As downtime dataset has records mostly of year 2012,I would build my model only for 2012 data.

Below are step by step tasks involved for data transformation-

1. **Filtering on Asset ID-** After loading the Indicator reading dataset we need to first filter it on HAGC Asset Number which is 540-630.
2. **Pivoting the Data-** All the Indicator names are stored in a single columns ,we need to pivot the data such that each indicator name is an individual column and its readings for each date as rows

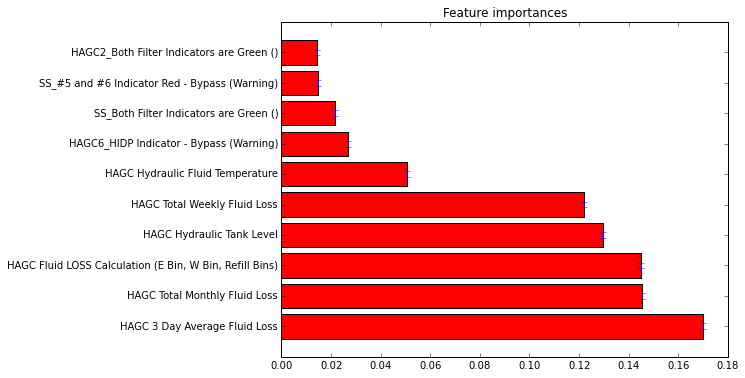
|  |  |  |  |
| --- | --- | --- | --- |
| **Date** | **Indicator Name-1** | **Indicator Name-2** | **Indicator Name-3** |
| Date-1 | Readings | Readings | Readings |

1. **Type Conversions-**  The Numeric readings are converted into integer format.
2. **Filling the Null Values-** While filling the null values I have used interpolation method for numeric values which have some kind of auto-correlation.And for other numeric values I have used the mean value.While filling nulls for categorical values I used a backfill.
3. **Calculating Moving Average and Sum-**Calculated moving average and sum to fill in nulls values for fluid loss readings.
4. **Filtering the Year-** Filtered the indicator data to year 2012 as the downtimes are only available for this period.So we have a dataset of 365 days.Am filtering the year after filling the nulls because the moving averages gets filled from previous years data.
5. **Merging with Downtime dates-** We fetch the downtime dates 2 days before to the downtime and mark as failure.
6. **Creating Dummies for filter variables-**In Python we convert the categorical variables to binary indicators using dummies function.
7. **Dropping the records of actual Downtime-** Dropping the actual downtime dates and its previous dates indicator readings as it can disturb the model.

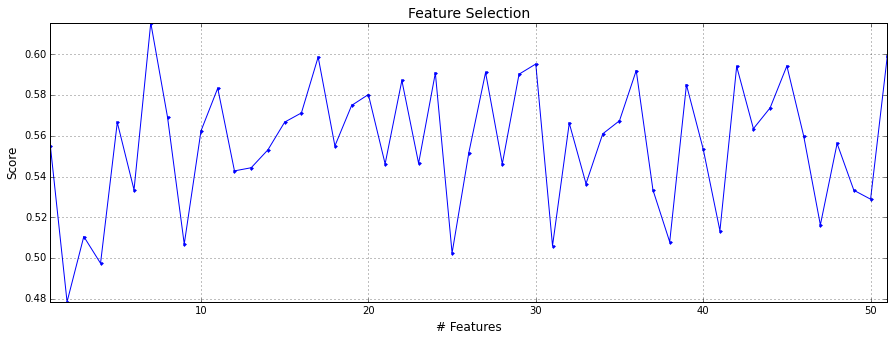
**Feature Selection**

In order to find the best estimators in the given data,I have performed Random Forest Classification on the data.Below are the list of best 10 Indicator with their importance scores

**Feature ranking:**



We can see from the above result that 3 Day Avg Fluid loss is the most strong predictor.And when I plot the accuracy scores with number of features selected I get the below graph.From this below graph I don't see any constant decreased value,hence am not selecting any subset of features for modeling my prediction.



Another method which I used to know about the strength of features is by knowing its correlation with the target label.For this I have used two methods

* Spearman-Rank correlation for categorical Vs categorical data
* Point-Biserial correlation for categorical Vs continuous data

Below are the results are the features which have the highest correlation with the failure.We can clearly see that The Fluid loss has the highest predicting power out of all other features.

|  |  |  |
| --- | --- | --- |
| **Features** | **correlation** | **abs\_corr** |
| **HAGC Fluid LOSS Calculation (E Bin, W Bin, Refill Bins)** | 0.253247 | 0.253247 |
| **HAGC Total Weekly Fluid Loss** | 0.175300 | 0.175300 |
| **HAGC Total Monthly Fluid Loss** | 0.163177 | 0.163177 |
| **HAGC 3 Day Average Fluid Loss** | 0.137314 | 0.137314 |
| **HAGC6\_HIDP Indicator - Bypass (Warning)** | 0.095569 | -0.095569 |
| **HAGC12\_#1 and #2 Indicator Red - Bypass (Warning)** | 0.086134 | 0.086134 |
| **URCC\_Normal ()** | 0.081752 | 0.081752 |
| **HAGC1\_HIDP Indicator - Bypass (Warning)** | 0.081752 | -0.081752 |
| **HAGC1\_HIDP and Case Drain Indicator - Bypass (Warning)** | 0.072806 | 0.072806 |

**Applying Machine Learning Models**

The Given problem is Supervised learning problem having two classes(Having failure in 48hrs/or Not having failure in 48hrs).Since the data is a combination of categorical and numerical data with two classes which are not equally distributed,I have used the below 4 machine learning Algorithms

1. **Logistic Regression**- It uses regularized regression using the liblinear solver sklearn which is preferred for smaller dataset.I used l2 penalty using scikit learn
2. **Gaussian Naive Bayes**- GaussianNB implements the Gaussian Naive Bayes algorithm for classification. The likelihood of the features is assumed to be Gaussian
3. **Random Fores**t-A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and use averaging to improve the predictive accuracy and control over-fitting.
4. **Gradient Boosting Classifier**- Gradient Boosting builds an additive model in a forward stage-wise fashion; it allows for the optimization of arbitrary differentiable loss functions.

Apart from the above 4 algorithms,I tried using SVM and KNN classifier algorithms,But their performance scores(recall) are too low.

I have used Grids search in python to search for best parameters to run the model.I have performed grid search along with cross validation in order to avoid over fitting of the data.

**Model Performance**

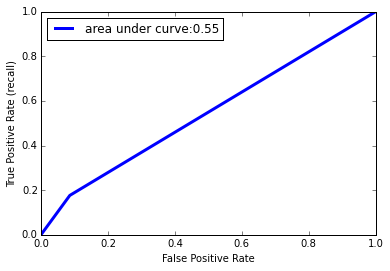
To evaluate the performance of the model,I have drawn the Confusion Matrix for each algorithm.Since we are working on a Binary classification-the confusion matrix looks as below

|  |  |  |
| --- | --- | --- |
| **Actual\Predicted** | **0** | **1** |
| **0** | True Negatives | False Positives |
| **1** | False Negatives | True Positives |

I am considering the Class-1(positive class) as failure.To Evaluate the model we can either use Accuracy,Precesion,Recall or Area under curve. For this problem Recall is a good evaluation metric which tells us what percentage of downtimes we predicted.Also the area under curve reduces the cost of false positives and false negatives.Below are the results for each of the Algorithm

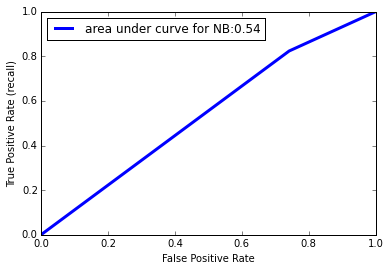
**Logistic Regression**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Confusion Matrix**   |  |  |  | | --- | --- | --- | | **Actual\Predicted** | **0** | **1** | | **0** | 53 | 5 | | **1** | 28 | 6 | | **Classification report**  precision recall f1-score support   0 0.65 0.91 0.76 58  1 0.55 0.18 0.27 34  avg / total 0.61 0.64 0.58 92 |



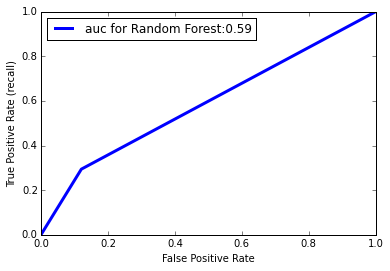
**Gaussian Naive Bayes**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Confusion Matrix**   |  |  |  | | --- | --- | --- | | **Actual\Predicted** | **0** | **1** | | **0** | 15 | 43 | | **1** | 6 | 28 | | **Classification report**  precision recall f1-score support   0 0.71 0.26 0.38 58  1 0.39 0.82 0.53 34  avg / total 0.60 0.47 0.44 92 |



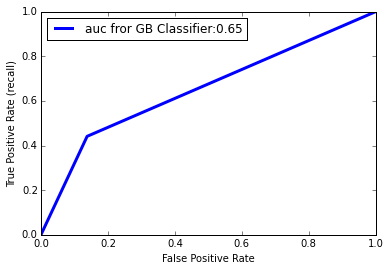
**Random Forest**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Confusion Matrix**   |  |  |  | | --- | --- | --- | | **Actual\Predicted** | **0** | **1** | | **0** | 51 | 7 | | **1** | 24 | 10 | | **Classification report**  precision recall f1-score support   0 0.68 0.88 0.77 58  1 0.59 0.29 0.39 34  avg / total 0.65 0.66 0.63 92 |



**Gradient Boosting Classifier**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Confusion Matrix**   |  |  |  | | --- | --- | --- | | **Actual\Predicted** | **0** | **1** | | **0** | 50 | 8 | | **1** | 19 | 15 | | **Classification report**  precision recall f1-score support   0 0.68 0.88 0.77 58  1 0.59 0.29 0.39 34  avg / total 0.65 0.66 0.63 92 |



**Comparing the models**

|  |  |  |  |
| --- | --- | --- | --- |
| **Models** | **Recall** | **ROC\_AUC\_Score** | **Total Cost Saved/Cost-Benefit Analysis** |
| **Logistic Regression** | .18 | .55 | (6\*1260)-(5\*712) =4000$ |
| **Naive Bayes** | .82 | .54 | (28\*1260)-(43\*712)=4664$ |
| **Random Forest** | .29 | .59 | (10\*1260)-(7\*712)=7616$ |
| **Gradient Boosting** | .29 | .65 | (15\*1260)-(8\*712)=**13204$** |

The Total cost saved can be calculated by knowing the mean unplanned and planned maintenance costs,which I calculated in the beginning during data analysis

**Formula**=[(True Positives)\*(Planned cost -Unplanned cost)] - [(False Positives)\*(Planned Cost)]

Mean Planned cost=**712$**

Mean Unplanned Cost=**1972$**

Difference=**1260$**

By predicting the downtime,we are converting unplanned maintenance to planned,so the difference is multiplied with total number of correct downtime predictions and this value is subtracted from false positives as it costs unwanted planned maintenance

**Selecting the Best Model**

From the above analysis we can clearly say that **Gradient Boosting Classifier** is the best model,As it Total cost saved and area under curve is the higher than any other model.Though Naive Bayes has best recall score its total cost saving is lower due to high false positive rate.Gradient Boosting and random forest having higher AUC score shows that algorithms with decision trees fit the data better.

**Tool Used**

Programming - Python

Platform - Ipython/Jupyter

Packages- Sklearn,Scipy,Numpy,Pandas

Visualizations- Matplotlib,Seaborn

**Other Model**

I tried to predict the downtime based on the unplanned maintenance dates from orders dataset.This method has more data as Work order dataset has more than 2 years of data.Though the performance on this was good,I dropped pursuing this approach as we are trying to predict the downtime but the not the unplanned maintenance.