# C 139 SF401 Machine Learning Summary

## Methodology/Contents



## Data Exploration

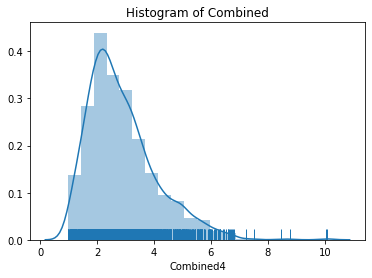
Nature of the feature datasets:

|  |  |
| --- | --- |
| Scope | C139 Shorncliffe 401 line |
| Source datasets | * TRC 20170131 * TRC 20171017 * TRC 20170704 * TRC 20180131 * GPR 2018 SF401 * Drainage locations C139 |
| Approximate number of original records | 50,000 |
| Number of features | 46 |
| Number of instances used | 1,650 |

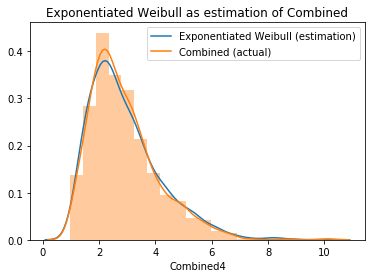
Response variable:

|  |  |
| --- | --- |
| **Response Structure** | |
| Min | 0.961385 |
| Max | 10.094321 |
| Mean | 2.874873 |
| Standard Deviation | 1.206252 |

|  |  |
| --- | --- |
| Distribution of the response variable:  There does not appear to be an obvious structure associated with the response variable e.g. high values of the Combined metric are often preceded by low values in adjacent meterage.  The distribution of the response variable is skewed towards lower values of Combined i.e. there are proportionally fewer samples that indicate higher degradation: | C:\Users\Marcus\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\FF54665D.tmp |



The distribution resembles an exponentiated Weibull distribution and was found to be closely estimated by the following probability density function:



|  |  |
| --- | --- |
| The cumulative density function is as follows.  Based on current practice, a threshold for Combined value was set at 4.9 i.e. values above this are of high interest in maintenance decision making: | C:\Users\Marcus\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\123EBD2B.tmp |

The probability of a sample exceeding this threshold was 7.64% i.e. the dataset is highly biased to samples that are lower than the threshold of interest.

A trivial classifier that only predicts {Combined < threshold} would achieve 92.36% accuracy. It would be more informative to predict the value of the target variable for a future quarter (rather than class). As such, regression was employed.

## Pre-processing

Key pre-processing actions include:

|  |  |
| --- | --- |
| 1. Alignment of TRC datasets | by minimising the standard deviations in differences between Gauge and Super across the 4 datasets |
| 1. Alignment of GPR to TRC | using meterage measures |
| 1. Calculate standard deviations and the Combined metric | across 20 metre sections of TRC measures |
| 1. Alignment of drainage points | using meterage measures |
| 1. Train/test split | the dataset was split into training (75%) and test (25%) sets |
| 1. Standardisation | feature data was standardised to mean 0, standard deviation of 1 (as required by models such as K-NN and SVR) |

## Prediction Baseline

Using the most trivial regression model: , achieved 63.86% test accuracy.

|  |  |
| --- | --- |
| The baseline prediction could be improved by projecting the most recent quarter Combined value i.e. . The baseline prediction achieved 80.92% test accuracy.  Note, the gradient of line of best fit between predicted and actual points was only 0.77 (as opposed to the ideal of 1.0), meaning the baseline significantly overestimates the actual Combined value (possibly due to maintenance work undertaken in the quarter). | C:\Users\Marcus\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\A13F319B.tmp |

|  |  |
| --- | --- |
| The relatively high test accuracy indicates the TRC alignment process appears to be successful.  The baseline prediction of “high priority” datapoints (i.e. those where prediction and/or actual exceed the threshold), showed a low level of correlation with the actual target values indicating a poor fit. | C:\Users\Marcus\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\EF8CD961.tmp |

Baseline metrics summary:

|  |  |
| --- | --- |
| Test accuracy | 80.92% |
| Gradient of best-fit for predicted and actual points | 0.77 |
| Test accuracy “High Priority” data points | 73.69% |
| Correlation between “high priority” predicted and actual points | 0.06 |

As the Combined metric is a linear combination of the standard deviations of each of Top Left, Top Right and Twist 3, the baseline prediction was also assessed for these measures. The test accuracy for the 3 sub-measures was ~68% i.e. substantially lower than the Combined prediction accuracy.

This is because the Combined prediction benefits from the averaging of the 3 sub-measures. As such, prediction of the Combined metric remained the focus.

## Modelling Objectives

Based on analysis of the baseline predictions, the objectives were defined to be:

|  |  |  |
| --- | --- | --- |
| **Scope** | **Metric #1** | **Metric #2** |
| All test data points | Prediction accuracy on the (unseen) test dataset | Gradient of best-fit between actual and predicted points |
| “High Priority” data points | Prediction accuracy on “high priority” test dataset | Strength of correlation between predicted and actual for “high priority” datapoints |

Prediction horizons considered:

|  |  |
| --- | --- |
| **Prediction Horizon** | **Scope** |
| 1 quarter into the future | * Baseline prediction accuracy * All ML models using all available datasets |
| 2 quarters into the future | * Baseline prediction accuracy * Random Forest Regression using only reduced feature dataset |

## Feature Selection

The feature selection results are summarised as:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Feature Selection Method** | **Best Test Accuracy** | **Gradient of best-fit for Best Model** | **Best Test Accuracy “High Priority”** | **Best Correlation “High Priority”** |
| Linear regression (i.e. investigation of coefficient p-values) | 83.05% | 0.98 | 80.52% | 0.18 |
| LASSO | 81.78% | 1.17 | 80.27% | 0.22 |
| Elastic Net1 | 83.21% | 1.01 | 80.74% | 0.16 |

Note 1: not technically a feature selection method as includes coefficient shrinkage

The features selected by the models are as follows:

|  |  |  |
| --- | --- | --- |
| **Linear Regression (“OLS”)** | **LASSO alpha=0.01 (“20”)** | **LASSO alpha=0.1**  **(“9”)** |
| |  | | --- | | BDMCentre | | BDMRight | | BVMCentreCategory | | BVMCentreVolume | | BVMLeftCategory | | BVMRightCategory | | Drainage | | LRICentre | | LRILeft | | SDTopLeft1 | | SDTopLeft2 | | SDTopLeft3 | | SDTopRight2 | | SDTopRight3 | | SDTwist103 | | SDTwist33 | | SDVersL3 | | SDVersR3 | | TDILeft | | |  | | --- | | BTILeft | | BVMLeftCategory | | BVMLeftVolume | | BVMRightCategory | | LRICentre | | LRILeft | | PVCCentre | | PVCLeft | | PVCRight | | SDTopLeft1 | | SDTopLeft2 | | SDTopLeft3 | | SDTopRight1 | | SDTopRight3 | | SDTwist101 | | SDTwist103 | | SDTwist33 | | SDVersL1 | | SDVersL3 | | TDILeft | | |  | | --- | | PVCCentre | | PVCLeft | | PVCRight | | SDTopLeft3 | | SDTopRight3 | | SDTwist101 | | SDTwist103 | | SDTwist33 | | SDVersL1 | |

Legend: common to all 3, common between LASSO 20 and OLS, common between LASSOs

Note the LASSO (alpha = 0.1) coefficients highlight the importance of Top Left, Top Right, Twist 3 (all consistent with the Baseline predictor) and Twist 10:

|  |  |
| --- | --- |
| **OLS Feature** | **Coefficient** |
| PVCCentre | -0.000034 |
| PVCLeft | 0.000195 |
| PVCRight | -0.000156 |
| **SDTopLeft3** | **0.139517** |
| **SDTopRight3** | **0.445248** |
| SDTwist101 | 0.005967 |
| **SDTwist103** | **0.226816** |
| **SDTwist33** | **0.44012** |
| SDVersL1 | 0.00097 |

Highlighted are LASSO variables with coefficient > 0.1

Feature selection summary performance:

|  |  |
| --- | --- |
| **Objective** | **Performance** |
| Test accuracy | ~2% improvement |
| Gradient line-of-best-fit | High improvement |
| Test accuracy “High Priority” points | ~6% improvement |
| “High Priority” prediction correlation | Moderate improvement |

## Machine Learning Models

Several machine learning algorithms were implemented with the results summarised below.

### Random Forest

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Machine Learning Model** | **Best Test Accuracy** | **Gradient of best-fit for Best Model** | **Best Test Accuracy “High Priority”** | **Best Correlation “High Priority”** |
| * All features | 85.42% | 1.02 | 84.75% | 0.3 |
| * “OLS” features | 85.06% | 1.01 | 84.39% | 0.37 |
| * LASSO “20” | 85.09% | 1.01 | 85.51% | 0.33 |
| * **LASSO “9”** | **84.35%** | **0.99** | **86.2%** | **0.48** |

Results are for unscaled data. Random Forest fitted for 1,000 trees using a random sample of (number of features)1/2.

|  |  |
| --- | --- |
| **C:\Users\Marcus\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\66BC317.tmp** | **C:\Users\Marcus\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\B561743D.tmp** |
| **Random Forest: 9 features** | |

### Support Vector Regression

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Machine Learning Model** | **Best Test Accuracy** | **Gradient of best-fit for Best Model** | **Best Test Accuracy “High Priority”** | **Best Correlation “High Priority”** |
| * All features | 84.8% | 0.92 | 84.70% | 0.03 |
| * “OLS” features | 83.29% | 0.89 | 84.72% | 0.22 |
| * **LASSO “20”** | **85.14%** | **0.9** | **85.25%** | **0.2** |
| * LASSO “9” | 82.9% | 0.9 | 84.44 | 0.17 |

Results are for scaled data and radial basis function (RBF) kernel which outperformed sigmoid and polynomial kernels. Optimal regularisation parameter (C) = 10 i.e. the model traded a relatively small margin for higher training accuracy.

|  |  |
| --- | --- |
| **C:\Users\Marcus\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\E9A7CF4F.tmp** | **C:\Users\Marcus\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\A9A535.tmp** |
| **Support Vector Regression: 20 features** | |

### K-NN Regression

The lack of “elbow” in the Test RMSE versus K plot using all features indicated the target variable, Combined, is not consistently correlated with a similar set of features i.e. the combination of features and response are relatively unique. On this basis, it was not expected that KNN using all features would perform well on the test dataset. This is contrasted with a clear optimal K (15 neighbours) when using only 9 features.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| C:\Users\Marcus\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\4E0E820B.tmp | | C:\Users\Marcus\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\2DD13351.tmp | | | |
|  | |  | | | |
| **Machine Learning Model** | **Best Test Accuracy** | | **Gradient of best-fit for Best Model** | **Best Test Accuracy “High Priority”** | **Best Correlation “High Priority”** |
| * All features | 81.68% | | 1.04 | 80.08% | 0.13 |
| * “OLS” features | 83.59% | | 0.99 | 83.19% | 0.08 |
| * LASSO “20” | 83.33% | | 1.0 | 80.93% | 0.08 |
| * **LASSO “9”** | **82.21%** | | **1.02** | **81.64%** | **0.23** |

Results are for scaled data using a ball tree algorithm and Manhattan distance.

### Artificial Neural Networks (ANN)

|  |  |
| --- | --- |
| ANNs were developed using the KerasRegressor, Sequential (from the Keras library in Python) and MLPRegressor (from sklearn neural\_network in Python).  Results are shown for the Sequential model implemented with early stopping. | C:\Users\Marcus\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\99AFF52D.tmp |

The importance of early stopping is evident where the test error (orange) starts to rise after relatively few epochs despite training loss continuing to fall.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Machine Learning Model** | **Best Test Accuracy** | **Gradient of best-fit for Best Model** | **Best Test Accuracy “High Priority”** | **Best Correlation “High Priority”** |
| * All features | 85.53% | 0.88 | 83.09% | 0.05 |
| * “OLS” features | 85.33% | 0.90 | 85.08% | 0.32 |
| * **LASSO “20”** | **85.88%** | **0.94** | **85.54%** | **0.34** |
| * LASSO “9” | 84.16% | 0.91 | 80.69% | 0.31 |

Results are for scaled data and network architecture comprising 6 hidden layers using the Adam optimiser.

|  |  |
| --- | --- |
| **C:\Users\Marcus\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\10C546C9.tmp** | **C:\Users\Marcus\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\82BD5CBF.tmp** |
| **ANN: 20 features** | |

The machine learning results are summarised as:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Machine Learning Method** | **Best Test Accuracy** | **Gradient of best-fit for Best Model** | **Best Test Accuracy “High Priority”** | **Best Correlation “High Priority”** |
| Random Forest (9 features) | 84.35% | 0.99 | 86.2% | 0.48 |
| SVR (20 features, scaled) | 85.14% | 0.9 | 85.25% | 0.2 |
| KNN (9 features, scaled) | 82.21% | 1.02 | 81.64% | 0.23 |
| ANN (20 features, scaled) | 85.88% | 0.94 | 85.54% | 0.34 |

Feature selection summary performance:

|  |  |
| --- | --- |
| **Objective** | **Performance** |
| Test accuracy | ~3% improvement |
| Gradient line-of-best-fit | High improvement |
| Test accuracy “High Priority” points | ~12% improvement |
| “High Priority” prediction correlation | High improvement |

## Feature Transformation

It was noted the ML models used a feature dataset that sourced data from different time horizons: GPR (12-months old), TRC (quarterly) and drainage points (relatively fixed).

To assess the impact of this, the TRC datasets were replaced with:

1. the most recent TRC (i.e. prior to that to be estimated) was retained to preserve the most currently known information regarding the features, and
2. derived features intended to capture the rate of change over prior TRC runs, calculated as follows:

where *γ* is a decay coefficient reducing the impact of historic rates of change. It was noted the most useful rate of change features used *γ*=0 i.e. only the rate of change between the most recent feature and the corresponding prior TRC feature was used.

In general, the ML models did not demonstrate superior performance using this feature transformation approach.

## Consider Longer Prediction Horizons

Predictions were made for a 2-quarter time horizon. The baseline performance was compared with Random Forest regression. The ML method clearly outperforms the baseline in the longer time horizon. Additionally, Random Forest produced significantly higher test accuracy on the “high priority” predictions.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Prediction Method** | **Best Test Accuracy** | **Gradient of best-fit for Best Model** | **Best Test Accuracy “High Priority”** | **Best Correlation “High Priority”** |
| * Baseline | 51.86% | 0.23 | 28.78% | -0.56 |
| * Random Forest | 74.19% | 1.29 | 70.93% | 0.15 |

Random Forest used 9 feature, unscaled data.

|  |  |
| --- | --- |
| **Baseline prediction: Combinedi(t) = Combinedi(t-1)** | **Random Forest Regression (9 features)** |
| C:\Users\Marcus\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\6D73C941.tmp | C:\Users\Marcus\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\CF384425.tmp |
| C:\Users\Marcus\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\BB149DF7.tmp | C:\Users\Marcus\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\3FD3AE7B.tmp |

It is argued that unlike the baseline method, the Random Forest model produced predictions for 2 future quarters that could be useful.

## Summary of Results

Key findings include:

1. The target variable (i.e. Combined) can be relatively well approximated as an exponentiated Weibull PDF.
2. Feature selection can be used effectively to reduce the number of features to as few as 9 while still enabling reasonable performance from the ML methods.
3. With robust alignment, utilising the most recent value of the target variable delivered ~80% test accuracy when predicting the target variable in the next quarter.
4. Machine Learning methods such as Random Forests, ANNs and SVR delivered ~3% improvement in prediction accuracy for the following quarter.
5. Over the short time horizon, the ML model prediction accuracy was ~12% higher for the “high priority” points.
6. For the longer prediction horizon of 2 quarters, the baseline method delivered poor estimates, achieving only 52% test accuracy. Its prediction accuracy was only 29% on the “high priority” points in the longer horizon.
7. In the longer prediction horizon, Random Forest achieved 74% test accuracy and 71% accuracy on the “high priority” points. Unlike the baseline prediction method, may be useful for generating longer-term horizon predictions.

## Potential Improvements

The following warrant further investigation:

1. Control for maintenance work (i.e. using work order data) in the ML algorithms
2. Continue experiments in longer-term prediction horizons with a view to optimising the ML methods

## Project Team

JiaJun Liu

Melvin Yin

Rahul Chowdhury

Marcus Salouk