**Data423 Assignment 3**

**Yu Xia (62380486)**

Data description

The data consists of 999 observations of 21 variables. The target variable is “Response”.

There are missing values in a portion of the numeric variables. The missing values visually appear random. There are no excessively missing variables or observations.

The numeric predictor data has some uni-variable outliers but these disappear when the IQR multiplier reaches 2.6 so these are not of great significance.

There are several nominal variables. One which, “BloodType” has a cardinality of 4. The numeric data has a variety of scales. The means range from 1.99 to 3034 across the variables. There is a wide disparity in the variances (1.35 to 874,337).

The predictor correlation shows that a block of numeric variables are highly correlated. These all have similar variable names i.e. “Reagent\_\*”.

Several date formatted character variables exist. The format is unambiguously YYYY-mm-dd. These have been converted to date variables.

The pairs plots show some non-linear relationships between the predictors and the target.

Strategies

Missing data

There are no excessively missing variables or observations to discard.

Because we have a raw obs/variables ratio of only 48 we do not in a strong position to choose partial deletion. There is NO evidence of missing values being “Missing Completely at Random” so partial deletion might generate bias.

The methods that implicitly handle missing values can be tried on the raw data.

For the other methods we shall employ knn (neighbours=5) imputation as a standard approach.

Once selected, the best model was improved slightly by using bag imputation.

Since there are no missing target values we do not need to discard these observations.

Outliers

The uni-variable outliers are of little concern. The observations are all to be retained. Some robust

methods were tried just in case they offer some advantage. The critical issue was the significance of

residual outliers.

The impact of outliers should be assessed through model residuals rather than relying solely on uni-variable analysis,and we will leave this issue on desk and investigate it with our final chosen best four models