**Data Cleaning slide**

The data is cleaned prior to modelling.

Missing values are treated, since it can cause bias and volatility to the model. This is done via dropping features with high missing percentage, dropping observations, and re-grouping feature levels.

The regrouping also serves to simplify features with many levels.

Outliers which could cause bias to model trends are also identified and removed with the Interquartile range rule.

**GBM Slide 1**

We then try using a Boosted Tree model, which is an extension to the random forest via sequential learning. The GBM seems to yield the best results when trained with the repeat balanced dataset.

The softmax objective function is used since this is a multiclass prediction, and a grid search is used for hyperparameter tuning with 10 fold cross validation.

**GBM Slide 2**

With the best set of tuned hyperparameters, the model’s sensitivity on the minor class <30 seems to be the poorest (at 21%). The F1 score also seems to be lower than other competing models.

Compared to the training data which yields very high sensitivity and F1, it is likely that the GBM is overfitting.

**Model Comparison**

Having done 5 different models, the conclusion is that the SVM is the best suited model for our data.

We conclude this by choosing the Recall to be our main determinant metric, since the focus of our problem is on reducing false negatives (ie >30 and NO).

The SVM shows the best Recall metric, as well as F1 score.