Accuracy:

The proportion of true outcomes from the total data set. Accuracy can be explained as:

(True Positive + True Negative) / (True Positive + False Positive + False Negative + True Negative)

Accuracy was used in this study as the problem to be solved is a classification problem. Accuracy works well when data set is balanced.

Precision:

The proportion of predicted positive are true positives. Precision can be explained as:

True Positive / (True Positive + False Positive)

This metric indicates the percentage of your predication which are relevant.

Recall:

The proportion of actual positive are classified correctly. Recall can be explained as:

True Positive / (True Positive + False Negative)

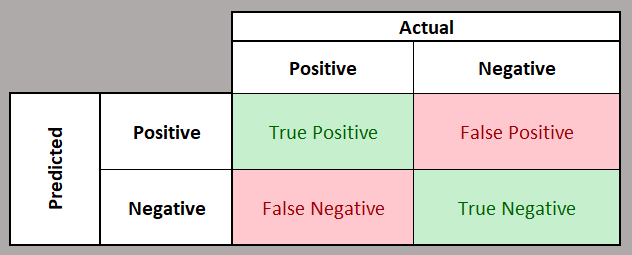
This metric indicates the percentage of total relevant results are correctly classified by the model.

F1 Score:

A score between 0 and 1 which is the harmonic mean of precision and recall. F1 score can be explained as:

2 \* ((precision \* recall) / (precision + recall))

The metric indicated how precise your classifier is, meaning how many times the label was classified correctly.



True Positive: Predicted YES and output was YES

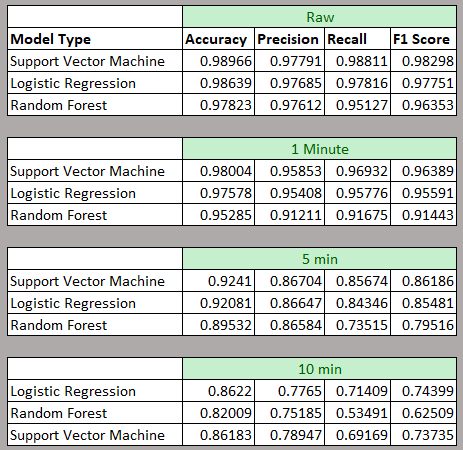
True Negative: Predicted NO and the output was NO

False Positive: Predicted YES and the output was NO

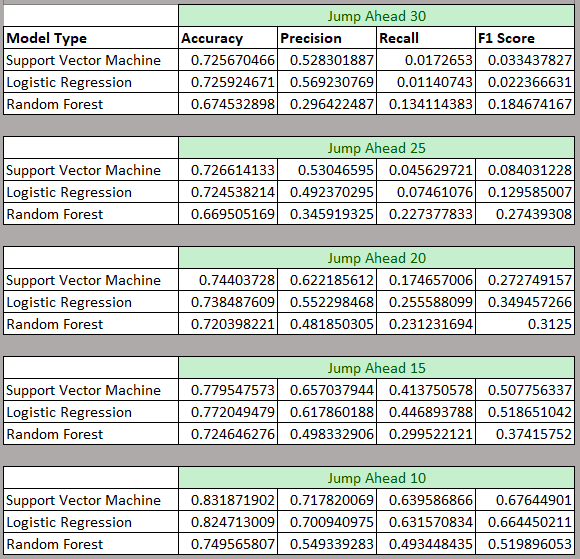
False Negative: Predicted NO and output was YES

Observations:

During the study it was found that resampling the data from ten seconds to every one minute the results did not vary in enough to deem to use the raw data. The variance between the metrics were in the 0.01 to 0.02 for all metrics used in the study. The benefit of using the one-minute data is that data loaded faster, and training/testing of the models were also speed up. The five and ten minute resampling caused the metrics to decrease enough that they were not deemed suitable.



The results when predicting thirty to ten minutes in the future decrease from with time. When predicting ten minutes into the future the logistic regression and SVM model produce results with an accuracy of 82% and 83% while predicting 30 minutes into the future we see accuracy drop to 75%. The Random Forest model results do not perform as well and are deemed not suitable for this study.



Conclusion:

Though the SVM model produce slightly better results (2%) this does not warrant the use of a more complex model for results that are higher. We have found that the logistic regression provides as good of results as that of the SVM model. The logistic regression model does run faster when compared to the SVM model and would be easier to deploy in a business setting.

The model can be used as a starting platform to be deployed in real time drilling operations. The model can be tested as drill operations occur and verified in a real time fashion allowing for practical implementation.