ISyE 6740 - Spring 2023
Project Report

Team Member Names: Thomas Swarbrick

Project Title: Flow Instability detection for oil wells using Machine Learning methods

Problem Statement Oil production wells are complex installations which provide a conduit for reservoir fluids to travel from deep underground to surface. Oil production wells are typically monitored with high frequency data measurements of pressure and temperature at different points along the well, i.e. at the tubing head and at a gauge located close to the bottom of the well. Problems with the well can be extremely costly for operators of oil wells, as any downtime will result in a loss of production and revenue. In addition, wells must be kept within safe operating conditions to ensure that no loss of oil occurs into the surrounding environment.

Whilst it can be possible to diagnose problems with wells from the high frequency data, it can be very time consuming for field technicians and engineers to constantly review the data. Therefore, machine learning methods can be used to help these professionals by classifying faults or problematic events for them. Using machine learning techniques for time-series event classification has applications to many industries and can provide predictions with reduced human bias and often faster and more efficiently than humans.

The objective of this project is to build an accurate classifier using machine learning techniques to identify whether an oil well is in normal operating conditions or experiencing flow instability based on multi-variate time series data. Flow instability occurs when the fluid traveling up the wellbore does not have enough energy to continuously reach the surface. Flow instability is usually characterised by alternating slugs of gas and liquid flowing out of the well. This phenomena can result in stress to the equipment, leading to damage, as well as problems at the processing facility downstream. Being able to detect these events automatically will reduce the time taken to take corrective action, leading to less well downtime and safer operations.

Data Source The dataset used in this analysis will be the 3W dataset, which is a open source collection of high frequency measurements publshed by Petrobras [2]. The data set contains multi-variate time series data for a number of wells. Each well has readings corresponding to

- Pressure at the Permanent Downhole Gauge (P-PDG)
- Pressure at the Tubing Head (P-TPT)
- Temperature at the Tubing Head (T-TPT)
- Pressure upstream of the surface control valve (P-PCK)
- Temperature downstream of the surface control valve (T-PCK)
- Gas Lift Injection Rate (QGL)

Pressure measurements are recorded in Pascals, temperature in degrees Celcius and gas flow rate in meters-per-second. The data set also contains labels for the observations, indicating whether the time-series was taken from a normal operating event or if it was from an undesirable events, such as hydrate formation, scaling and flow instability. An example of these events is shown in Figure 1.

The 3W data contains 597 csv files for time periods containing *normal* conditions and 344 time periods of *flow instability*. Each csv file has data recorded at one-second intervals for a 2-hour window, resulting in 7200 observations in each file.

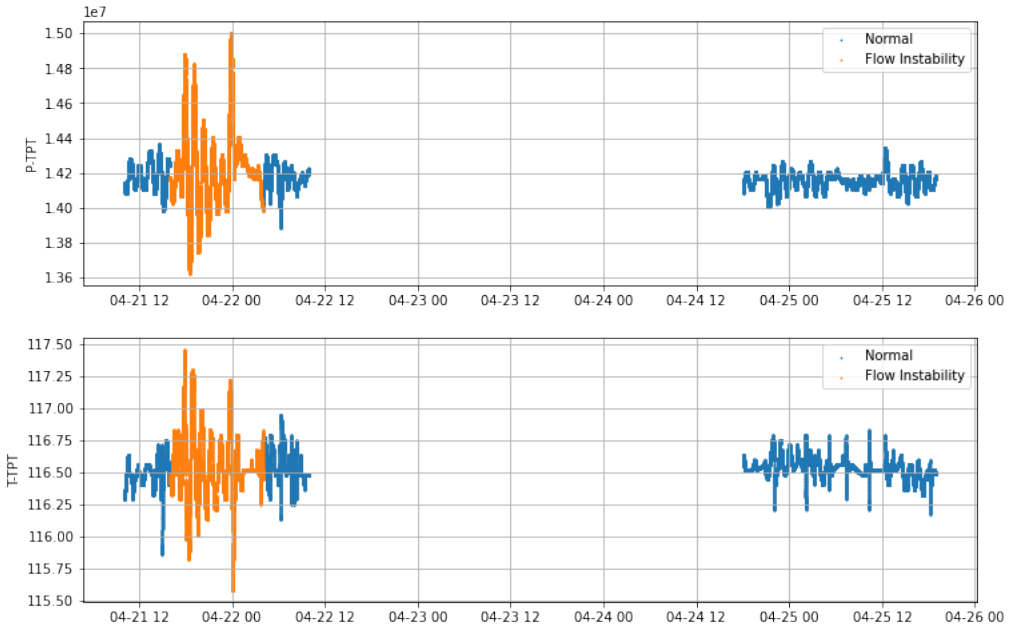

Figure 1: Tubing Pressure (P-TPT) and Tubing Temperature (T-TPT) from Well 00001 experiencing normal conditions (blue) and flow instability (orange)

Methodology Previous work has been completed on building a multi-class random forest classifier to classify events in the 3W dataset [1]. Rather than building a multi-class classifier, this project will focus on building a binary classifier for the two categories *normal* and *flow instability*. The following steps will be required:

- The ranges of values present in the dataset will first need to be reviewed. In particular, an understanding of missing values will be required as well as any unphysical measurements such as negative pressure.
- Feature extraction will be required to translate the data from raw time-series observations into meaningful features. This will consist of taking descriptive statistics from a time window such as mean, standard deviation, minimum, maximum, skewness and kurtosis.
- After feature extraction, a review of the features will be performed and correlation between features will be evaluated. Features with high correlation will be rationalised in order to avoid multicollinearity.

A number of different classifiers will be evaluated on the dataset including a Logistic Regression model, Naive Bayes classifier and Random Forest model. Each model will be trained on measurements from one set of wells with parameter tuning evaluated using 10-fold cross validation. The model's performance will then be evaluated on a different set of wells. This will help to build confidence that the learnings from one well can be generalized to many other wells.

A range of performance metrics will be utilised, including Accuracy, Precision and Recall to provide a more balanced view of the performance. The most important metric for this work will be precision, the ratio between the number of true positives (TP) and total number of true and false positives ($TP + FP$), shown in Equation 1.

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

One problem with automated event detection systems is that false positives will result in many false alarms for the user, providing more of a nuisance and leading to users distrusting the model. Therefore, having a model with high precision is very important.

Evaluation and Final Results

Data Pre-Processing After combining the data from multiple .csv files, the number of missing values was reviewed, as shown in Table 1. Most parameters have a large number of missing values, except for P-PDG, P-PTP and T-TPT. Rather than imputing missing values or keeping all features but removing rows with missing values, the features P-MON-CKP, T-JUS-CKP, P-JUS-CKP, T-JUS-CKP and QGL were not considered for further feature engineering.

Parameter	Missing values
timestamp	0
P-PDG	4,316
P-TPT	4,625
T-TPT	4,627
P-MON-CKP	1,009,747
T-JUS-CKP	1,473,281
P-JUS-CKP	3,795,210
T-JUS-CKP	12,418,867
QGL	3,116,880

Table 1: Missing values in 3W dataset

Another consideration was the ranges of values present. The data set was filtered such that only values greater than 0 was used for pressure to ensure physical consistency and temperature ranges were also reviewed. One concern was that more than 75% of the values for the feature P-PDG were 0, which may be due to the downhole pressure gauge not working. As the feature contained identical, and non-physical, values for most of the observations, this feature was also not considered.

Feature Engineering The two measurements, P-TPT and T-TPT, were then used to generate a number of features based on windows of the raw time-series measurements. Whilst many of the other measurements were dropped, it is believed that the phenomena of flow instability can be observed using these two, as shown in Figure 1. A fixed window size of 30 minutes was used, as it is small enough to capture the changes in behaviour for flow instability compared to normal operating conditions. Whilst a larger window size could be used, as observed in Figure 1, this will help generate more observations of the two classes. For each of the windows, mean, standard deviation, minimum, maximum, skewness and kurtosis were extracted.

Figure 2 shows the correlation between features after extracting from the time-series data. As there is strong positive correlation between minimum, maximum, average and median for both P-TPT and T-TPT, only the average was retained the rest dropped. This resulted in using average, standard deviation, skew and kurtosis for the P-TPT and T-TPT measurements, or 8 features from the time-series data to train the classifier. As the ranges of the features was very different, the features were mean centered and standardized before further analysis.

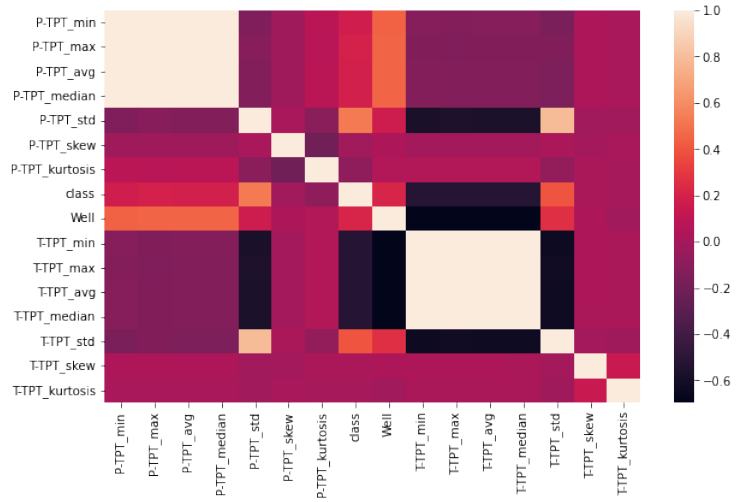

Figure 2: Correlation of features after extracting from time-series

The data was split into distinct sets for training and testing the models performance. In order to evaluate if the model could generalize to different wells, the two sets were created from distinct wells from the dataset. The breakdown of wells used for testing and training is shown in Table 2, with 5644 observations used to train the model and 1034 used to test.

Well	Set	Class	Observations
0001	Train	Normal	674
		Flow Instability	146
0002	Train	Normal	2040
		Flow Instability	452
0003	Train	Normal	260
0004	Train	Normal	48
		Flow Instability	172
0005	Train	Normal	550
		Flow Instability	152
0006	Train	Normal	1150
0007	Test	Normal	8
		Flow Instability	40
0008	Test	Normal	570
0010	Test	Flow Instability	332
0014	Test	Flow Instability	84

Table 2: Distribution of classes and wells between training and test sets

Hyper-Parameter Tuning A range of different number of trees were evaluated between 30 and 300 for the random forest model. The results are shown in Figure 3 and show the best performing model with 270 trees. Having said this, the precision results appear insensitive to the number of trees, all performing similarly. In addition, the standard deviation for all the models is large, indicating that the results are very dependent on the cross-validation split.

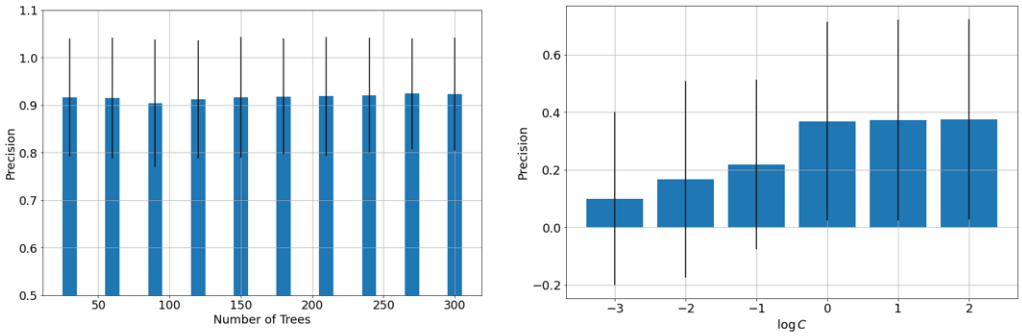

Figure 3: Hyper-parameter tuning for Random Forest (left) and Logistic Regression (right)

For the logistic regression model, a range of different values for C were evaluated on the training set. The C parameter controls the level of regularization applied, where smaller values of C result in stronger regularization. As shown in Figure 3, a model with $C = 10$ resulted in the highest precision, however, both $C = 1$ and $C = 100$ returned almost identical results from cross-validation.

Results & Findings The results of the three models evaluated is shown in Table 3, with results from both the test set and cross-valdiation of the training set. The Random Forest model has the highest precision on both the test and training set, with a precision of 1.0 on the test set. The other two models provide very similar precision scores of 0.98 and 0.99 respectively on the test set.

	Train	Test		
Model	Precision	Precision	Accuracy	Recall
Random Forest	0.92	1.0	0.88	0.73
Niave Bayes Classifier	0.25	0.98	0.99	1.0
Logistic Regression	0.38	0.99	0.89	0.76

Table 3: Model scores on both training and test sets

For the random forest model, the importance of each feature was evaluated using the average of the reduction in impurity at each node. This score metric, mean decrease in impurity (MDI), is shown for each feature in Figure 4. This shows that both the average and standard deviation of the two measurements, P-TPT and T-TPT, are much more important than the skew or kurtosis. This aligns with the observations in Figure 1, where the flow instability period is characterised by a wider range of values compared to the normal period.

Both the Logistic Regression and Naive Bayes models perform well on the test set, however, the performance on the test set is much better than on the training set. In fact, the Naive Bayes model has a much higher accuracy compared to the other two models on the test set, able to almost perfectly seperate the two classes. This is likely not a true representation of the models predictive ability given the poor precision from the training set and is likely due to the specific observations in the test set, also known as high variance.

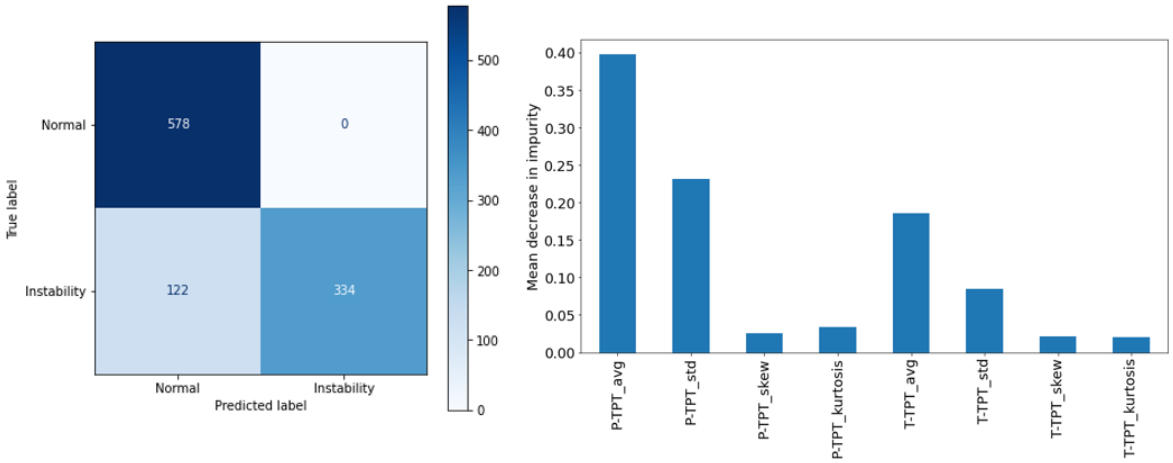

Figure 4: Confusion Matrix for test set (left) and Feature Importance in Random Forest Model (right)

Conclusions and Further Work The analysis completed in this report demonstrated the predictive ability of machine learning methods to classify problematic events in oil wells using high frequency data. A number of different methods were evaluated on the data from the 3W dataset and found the Random Forest model provided the highest precision on both the test and training set. The analysis also utilized only two of the available data sources, P-TPT and T-TPT, out of the 8 provided. This demonstrates that a robust classifier could be used without having to obtain lots of measurements from different areas for classifying flow instability.

Having said this, further work could be done on integrating the sparse data provided from different sources and evaluating whether the predictive ability was better with more data. In addition, further work could be done on evaluating different window sizes and other classification methods such as Neural Networks.

References

[1] M. A. Marins, B. D. Barros, I. H. Santos, D. C. Barrionuevo, R. E. Vargas, T. de M. Prego, A. A. de Lima, M. L. de Campos, E. A. da Silva, and S. L. Netto. Fault detection and classification in oil wells and production/service lines using random forest. *Journal of Petroleum Science and Engineering*, 197:107879, 2021.

[2] R. E. V. Vargas, C. J. Munaro, P. M. Ciarelli, A. G. Medeiros, B. G. do Amaral, D. C. Barrionuevo, J. C. D. de Araújo, J. L. Ribeiro, and L. P. Magalhães. A realistic and public dataset with rare undesirable real events in oil wells. *Journal of Petroleum Science and Engineering*, 181:106223, 2019.