EVALUATION OF AUTOMATED TUBING LEAK DETECTION USING BOTH ANALYTICAL AND MACHINE LEARNING METHODS

Applied Analytics
Practicum

By

Thomas Swarbrick

In Partial Fulfillment
of the Requirements for the Degree
Master of Science in Analytics
School of Industrial and Systems Engineering

Georgia Institute of Technology

November 2023

ACKNOWLEDGMENTS

I would like to thank Artificial Lift Performance Ltd for their support in this project. It would not have been possible to complete this kind of project without their support and invaluable access to data. I would also like to extend my appreciation to my supervisor, Sandy Williams, for his guidance and mentorship throughout the project.

Any views and conclusions contained herein are those of the author, and do not necessarily represent the official positions, express or implied, of the funders.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1
INTRODUCTION

In the era of the fourth industrial revolution, many industries are applying big data, connectivity and advanced analytics to increase efficiency and equipment reliability. Supervisory Control and Data Acquisition (SCADA) systems have become indispensable in supporting this effort, however, the expanding volume of data available presents many new challenges.

The use of SCADA systems to monitor equipment has been a part of the Oil & Gas industry for many years. These systems enable skilled technicians to monitor equipment remotely and look for significant deviations from standard operating conditions. In fact, many regulators require wells to have real-time monitoring devices installed [1].

Modern SCADA systems will typically provide a means to alert users when a measurement has exceeded a predefined threshold. These alarms can be very useful when used properly and help mitigate process upsets and troubleshoot problems. As the number of equipment and measurements grow, these systems often fall short of accurately detecting specific critical events at the required speed and precision. To overcome these challenges, data-driven analytical techniques such as Machine Learning are becoming increasingly prevalent in order to automate reviewing large amounts of data and extract valuable insights.

This dissertation addresses this pressing issue by exploring the application of both analytical and machine learning techniques to detect problematic events in oil wells using data commonly used for monitoring. This should bridge existing gaps in event detection methodologies and contribute to the advancement of intelligent systems in industrial settings.

1.1 Motivation

An area of active development at Artificial Lift Performance Ltd (ALP) is to automatically detect undesirable events, such as a tubing leak, using the data available within our application, Pump Checker. The ability to automatically detect a tubing leak using multiple streams of available data will enable our clients to take prompt action, reducing the time taken to return the well to normal operating conditions and thereby increasing production and revenue for their organization.

In order to find the best solution, both an analytical and machine learning approach will be evaluated. Whilst a machine learning approach may generalize better and adapt to changing data patterns, they can be less interpretable (i.e. a black-box) and may require large computational resources to train and test. An analytical solution would provide a more transparent and interpretable solution however, may not generalize as well and may become too rigid based on the rules used.

This project contains sensitive information and so data will be referred to in general, anonymized terms.

CHAPTER 2

BACKGROUND INFORMATION

2.1 Oil and Gas Production Systems

The production of oil and gas from underground reservoirs is a complicated process which involves many different disciplines of science and engineering. Once hydrocarbon fluids are discovered in commercial volumes, they are transported to surface through a production well. As the hydrocarbons reach the surface, some of the molecules will evolve from the liquid phase and into the gaseous phase as the pressure and temperature drop below the bubble point. In addition, water from within the reservoir may also be produced to surface. Therefore, we typically expect three fluid phases to be produced; oil, water and gas.

After drilling the borehole from surface into the reservoir, wells are typically installed with a section of pipe called casing which is cemented to the bore wall. This section of pipe provides structural stability for the wellbore and prevents fluids from above the reservoir interval entering the well. After this, a smaller diameter section of pipe, known as tubing, is installed and is the primary conduit for reservoir fluid to travel to surface. A diagram of a typical production system is shown in Figure 2.1

Over time, it may become necessary for production wells to require additional energy to continue to lift volumes of fluid to surface as natural energy within the reservoir declines. This is referred to as *artificial lift*, as it involves providing energy artificially to lift fluids to surface. There are many forms of artificial lift, including installing an electrical submersible pump (ESP) deep in the well which help pump the fluids to surface. Oil wells themselves are costly investments, usually costing millions of dollars, and can lead to catastrophic events if not continuously monitored.

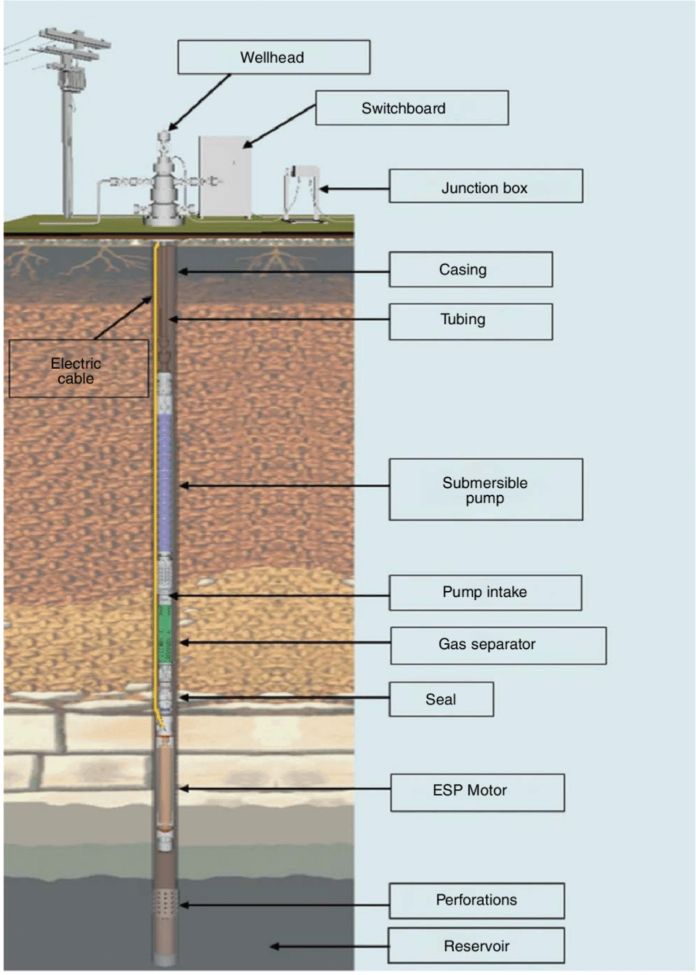

Figure 2.1: Typical oil production well with ESP installed [2].

2.1.1 Monitoring Oil and Gas Wells

Oil wells are typically installed with sensors to measure parameters such as pressure and temperature at different points on the well. For wells that are installed with an ESP, measurements of pressure and temperature are often measured at the intake of the pump along with electrical parameters like the motor current, motor temperature and pump frequency. The data will be collected and transmitted using a SCADA system back to a human-machine interface (HMI) for a skilled technician to review and monitor. Monitoring the wells ensures their performance can be optimized remotely and that any problems can be identified before they become more serious. These measurements are usually provided at high-frequency, with values recorded every few minutes and are referred to as *real-time* parameters.

The produced rates of oil, water and gas are very important for monitoring wells but are not as easily monitored as real-time parameters. As these three phases are produced together from the well, they will need to be separated into distinct streams in order to measure the volumetric rate of each one. This typically involves routing one well away from the main production system and into a dedicated separator vessel which will split the fluid into distinct phases. The duration of the well test is dictated by the time taken for the flow from the well to stabilize and can range between 1 to 48 hours [3].

As a result, the three phase rates over this time period are averaged and together provide an indication of well production, also known as a *well test*. Due to the duration, well tests are usually provided once every few days. It is also possible to install multiphase flow meters to provide inline phase measurements in real-time but these devices are much more expensive than a separator vessel and become prohibitively costly for operators of hundreds of wells.

2.1.2 Tubing Leaks

Like most industrial equipment, oil wells may operate outside of their intended operating conditions and experience undesirable events. One undesirable event which can occur is a leak in the tubing, known as a *Tubing Leak* or a *Hole-In-Tubing*. This will result in fluid flowing from the tubing and into the space between the casing and tubing, known as the annulus, instead of reaching the surface. These leaks occur for many reasons such as corrosion, wear from intervention tools and erosion due to local turbulence in fluid flow. Tubing leaks are problematic and result in deferred production & revenue from oil and gas not reaching the surface, well integrity issues and breaches of environmental regulations which could result in fines. It is for these reasons that operators continuously monitor their wells to detect these events early and remediate them before becoming more serious.

2.1.3 Detecting Tubing Leaks

Tubing leaks can be detected from reviewing both real-time and well test measurements and typically exhibit themselves from the following trends:

1. An increase in pump intake pressure, due to fluid accumulating at the bottom of the well
2. A decrease in tubing head pressure, due to less fluid reaching the surface
3. A relatively constant pump frequency, as changing pump frequency will change hydraulic conditions around the well
4. A severe reduction in total liquid production, as liquid would be circulating inside the well rather than reaching the surface.

It is expected for total liquid production to decline over time, along with tubing head pressure, as energy in the reservoir declines. A tubing leak, however, typically shows a much steeper production decline than expected.

An example of a tubing leak is shown Figure 2.2, where the pump frequency, pump intake pressure, tubing pressure and motor current are plotted over time. The duration at which the tubing leak has occurred is highlighted in red for demonstration, this would not normally be highlighted on the HMI.

Once detected, a tubing leak will require remediation action to occur on the well which will contribute to non-productive time. Whilst it may be possible for an engineer to routinely review the data for a small number of wells, this task soon becomes tedious for hundreds or thousands of wells. To overcome this problem, advanced analytical techniques can be used to review the data automatically across many wells.

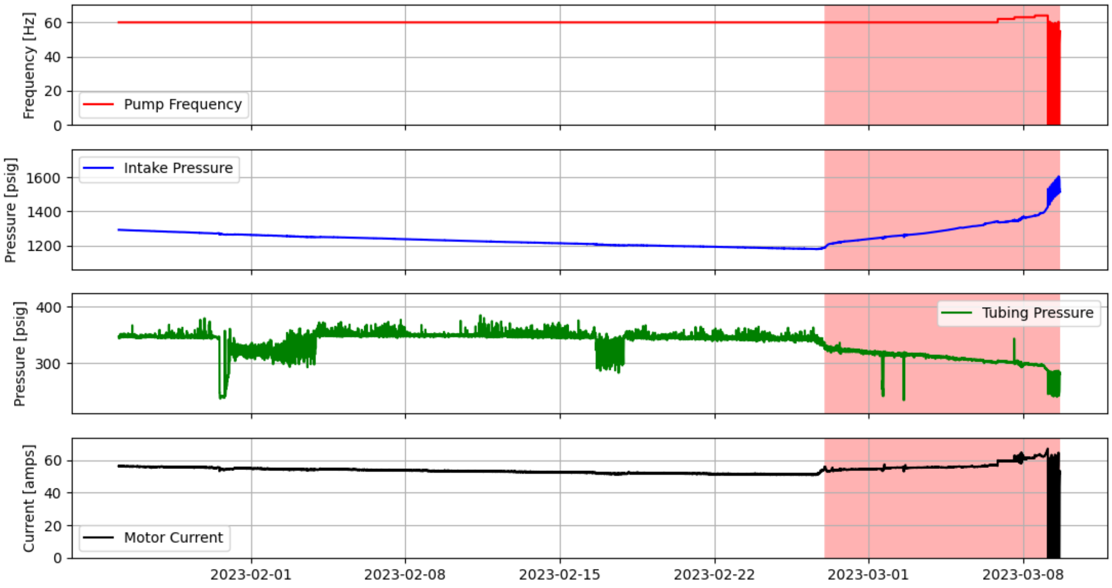

Figure 2.2: Example of typical measured parameters on a well where a tubing leak has occurred, as highlighted in red

2.2 Machine Learning

Machine Learning (ML) is a broad subsection of Artificial Intelligence (AI) involving techniques to build models based only on empirical data [4]. These models do not require explicit programming and learn the relationship between variables purely based on the data provided. As a result, ML models usually require large amounts of data to train themselves on which can also be computationally expensive.

2.2.1 Models

The ML models evaluated for supervised classification as part of this project are described below.

Random Forest

A single decision tree model is made up of a sequence of binary decisions (i.e. branches) which are each relatively simple. A random forest (RF) model consists of averaging the results of many decision tree models, as shown in Figure 2.3. This step of averaging many noisy, unbiased learners is known as *bagging* and is a key feature of the random forest model. Each tree is trained on a random selection of input variables which helps avoid overfitting [5].

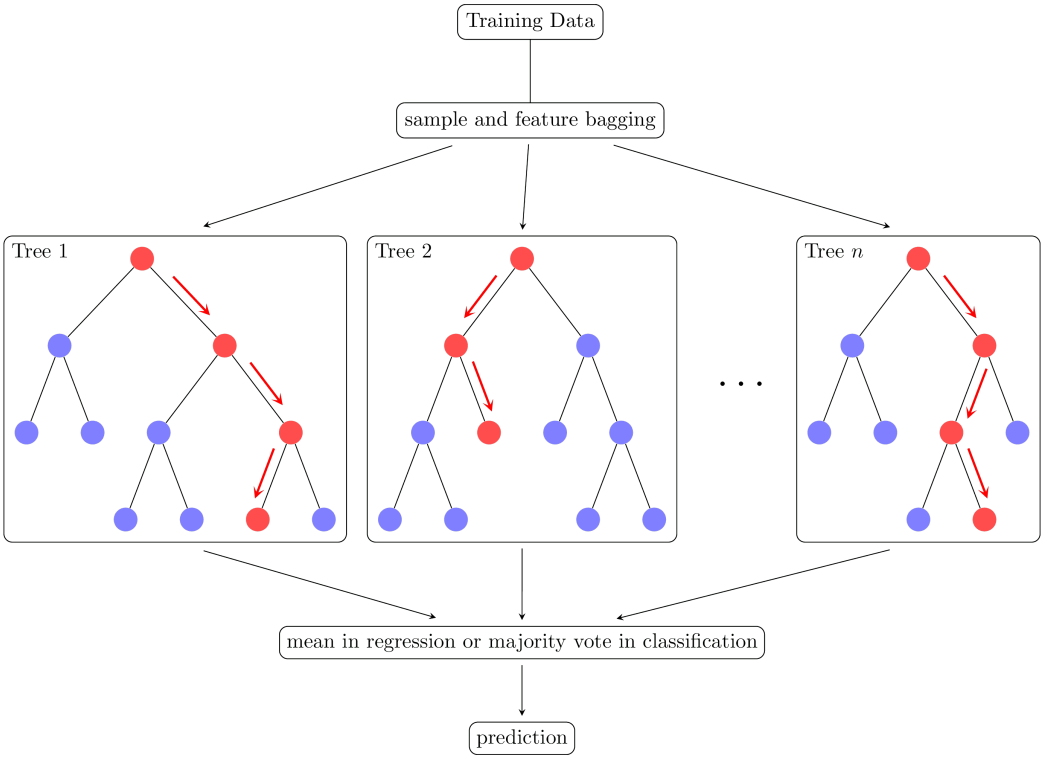
Training Data
sample and feature bagging
Tree 1
Tree 2
. . .
Tree n
mean in regression or majority vote in classification
prediction

Figure 2.3: Diagram showing how a RF model makes a classification [6]

Gradient Boosted Decision Tree Classifier

Boosting is another method for ensemble learners where weak learners are iteratively trained on weighted datasets where misclassified observations are given a higher weight than those correctly classified. Each successive classifier is therefore forced to concentrate on areas where it performed poorly, resulting in better performance. Algorithms like gradient descent are used at each iteration to reduce the difference between the actual and predicted values.

Support Vector Machine

The support vector machine (SVM) model constructs an optimal hyperplane which separates the dataset into two distinct classes. The hyperplane chosen is one which has the maximum possible distance from all points. These models are mainly used for classification problems and it can be difficult to interpret the models, particularly if the feature space was transformed.

k-Nearest Neighbor

The k-nearest neighbor (kNN) method uses the k closest data points to determine its classification. These models do not require any training but require the entire dataset in order to make a prediction.

2.2.2 Evaluating Model Performance

It is important that a model is able to generalize well to data outside of what it has been trained upon. When a model fits its training data too closely and fails to generalize to independent data, this is known as *overfitting*. To avoid selecting a model that is overfit, the total dataset is split into two distinct groups where one group is used to train the model (*training* data) and the other to provide an unbiased evaluation of the model's performance (*testing* data).

Each learning algorithm may have several *hyperparameters*, or tuning parameters, which are not directly learnt and need to be selected. Selecting these parameters is done by a similar process of evaluating the model's performance with them on data unseen during the training phase. For these reasons, it is common to split the training dataset into a training dataset and a *validation* dataset, which is used to evaluate the hyperparameters. This process of selecting hyperparameters on the training & validation set is usually done through a process called Cross-Validation (CV).

k-Fold Cross-Validation is a technique where the dataset is partitioned into k distinct groups. For each k partition, the model will train itself on $k - 1$ groups and evaluate its performance on the remaining group [7]. The performance metrics are averaged across the k partitions to provide an estimate of model performance with lower variance than with just a single partition. This process is shown in Figure 2.4.

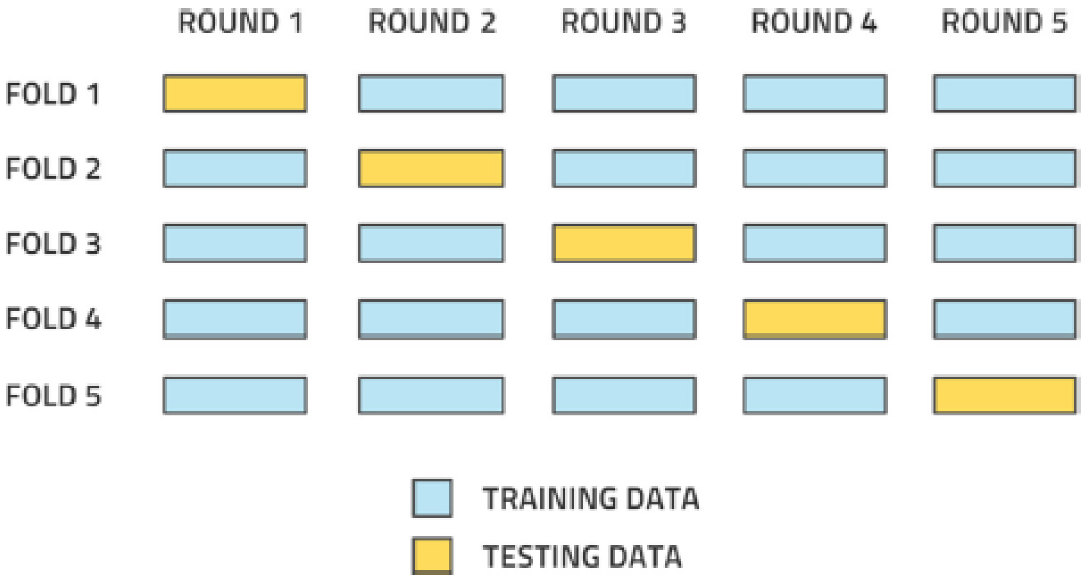

Figure 2.4: Diagram showing 5 fold cross-validation to evaluate performance [4]

CHAPTER 3
METHODOLOGY

3.1 Problem Formulation

As described in chapter 1, the objective of this project was to evaluate using machine learning and analytical methods to detect tubing leaks in oil wells. Having this functionality would be very valuable for clients of Artificial Lift Performance Ltd, as they could take prompt action to remedy wells with tubing leaks and increase their production & revenue.

A key consideration was the adherence of any model to the underlying physical trends described in subsection 2.1.3. Regardless of any solution's accuracy, users would need to feel confident in the results and comfortable with how the model came upon the classification made. It is for these reasons that both methods were evaluated and the interpretability of any ML model will also be considered.

The overall methodology for this project can be split into two broad sections, Data Analysis and Modelling. This process is highlighted in Figure 3.1 and shows the flow between the data set and the final model.

3.2 Data Analysis

3.2.1 Dataset

The dataset for this project was taken from 15 oil production wells fitted with ESPs where a tubing leak was detected. The tubing leak was confirmed with Root Cause Failure Analysis (RCFA) report after the well was worked-over to ensure that the correct physical phenomena was occurring. The data set is comprised of available measurements from the wells, including both real-time and well test parameters. The full list of available measurements is shown in Table 3.1.

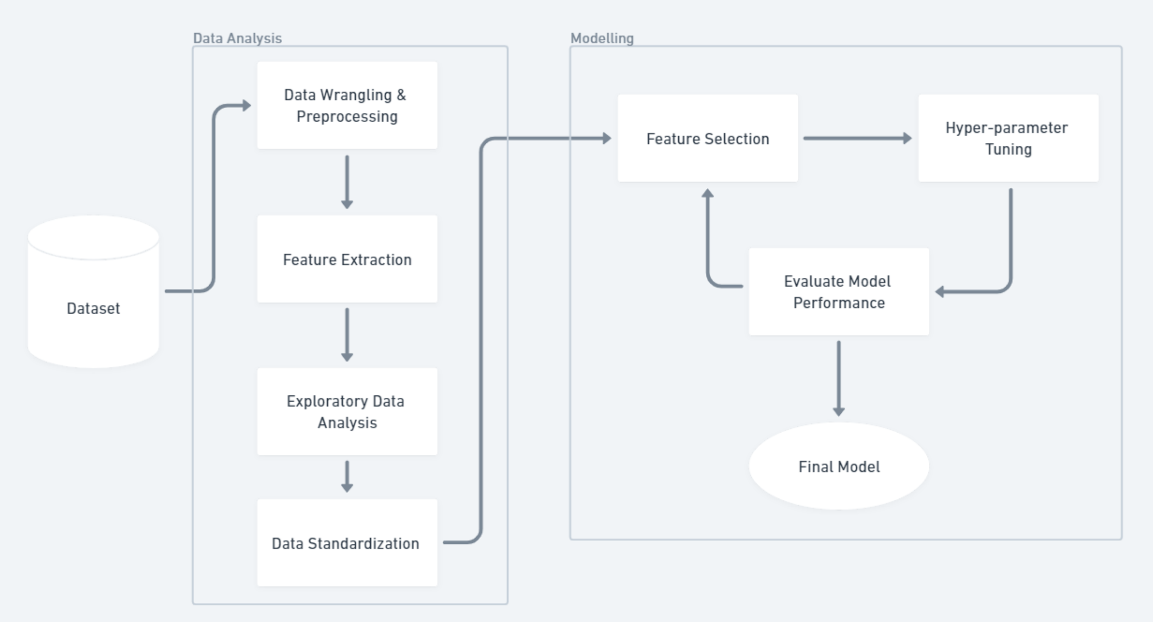

Figure 3.1: Modelling Workflow. Image credit: created by author.

Table 3.1: Measurements available from wells, including both real-time and daily well test parameters

Measurement	unit	Data source
Tubing head Pressure	psig	real-time
Casing head Pressure	psig	real-time
Drive Output Current	amps	real-time
Drive Output Voltage	amps	real-time
Pump Frequency	hertz	real-time
Pump Intake Pressure	psig	real-time
Pump Intake Temperature	degrees F	real-time
Motor Current	amps	real-time
Motor Temperature	degrees F	real-time
Oil flow rate	stb/d	well test
Gas flow rate	mscf/d	well test
Water flow rate	stb/d	well test

3.2.2 Data Wrangling & Preprocessing

The real-time and well test data was available in separate database tables. In order to align the real-time and well test measurements (i.e. every 5 minutes vs daily), an initial step was to represent the real-time measurements with an average taken over the day. This would allow the two data sets to be merged into a single table with a row for each day and a column for each feature with well datasets stacked vertically on each other. This process would be expanded to other statistical measures like standard deviation, range and skew if a suitable model was not able to be developed. The data was then filtered to remove any unphysical values, for example negative pressures or flow rates, and any missing values.

3.2.3 Feature Engineering

A number of additional features was created based on the measured data including Liquid Rate, Water Cut, Gas-Liquid Ratio (GLR) and Gas-Oil Ratio (GOR). These measures are commonly used to evaluate well performance and their formulation is provided below.

$$\text{Liquid flow rate (stb/d)} = \text{Water flow rate (stb/d)} + \text{Oil flow rate (stb/d)} \tag{3.1}$$

$$\text{Water Cut (\%)} = \frac{\text{Water flow rate (stb/d)}}{\text{Water flow rate (stb/d)} + \text{Oil flow rate (stb/d)}} \times 100 \tag{3.2}$$

$$\text{Gas-Oil Ratio (scf/stb)} = \frac{\text{Gas flow rate (scf/d)}}{\text{Oil flow rate (stb/d)}} \tag{3.3}$$

$$\text{Gas-Liquid Ratio (scf/stb)} = \frac{\text{Gas flow rate (scf/d)}}{\text{Water flow rate (stb/d)} + \text{Oil flow rate (stb/d)}} \tag{3.4}$$

In order to capture the trend in the measured parameters, a feature was calculated which reflected the change in percent between days. This was calculated for each of the parameters, both real-time and well test. The percentage change was decided over an absolute change so that the model would generalize to other operating ranges.

The point at which the tubing leak occurred was also manually picked from the real-time trends. This is consistent with the physical understanding of tubing leaks described previously in subsection 2.1.3. The days after the tubing leak occurred were denoted with a new feature called ‘HITFlag’, where ‘HIT’ refers to ‘Hole-In-Tubing’. This feature would be “1” on days after the tubing leak occurred and “0” before.

In order to keep the dataset as balanced as possible, an equal period of time was taken before the tubing leak occurred. For example, if a well showed 5 days of operating with a tubing leak, 5 days prior to the leak were also included for the negative case.

3.2.4 Exploratory Data Analysis

After pre-processing and feature engineering, the data set contained 380 rows of data with 32 features. The dataset contained 207 observations before the tubing leak occurred and 173 observations after. Histograms were generated to review the range, mean and skew of the features. An example is shown in Figure 3.2 which shows the distribution of phase rates from the well tests. The histogram is colored based on the 'HITFlag', or whether or not the test was taken before or after the tubing leak had occurred. Whilst there does not appear to be any significant difference between the two samples, the Liquid Rate and Gas Rate show some separation between the two. We can see in both of these that the measured liquid and gas rates after the tubing leak has occurred is generally lower than before.

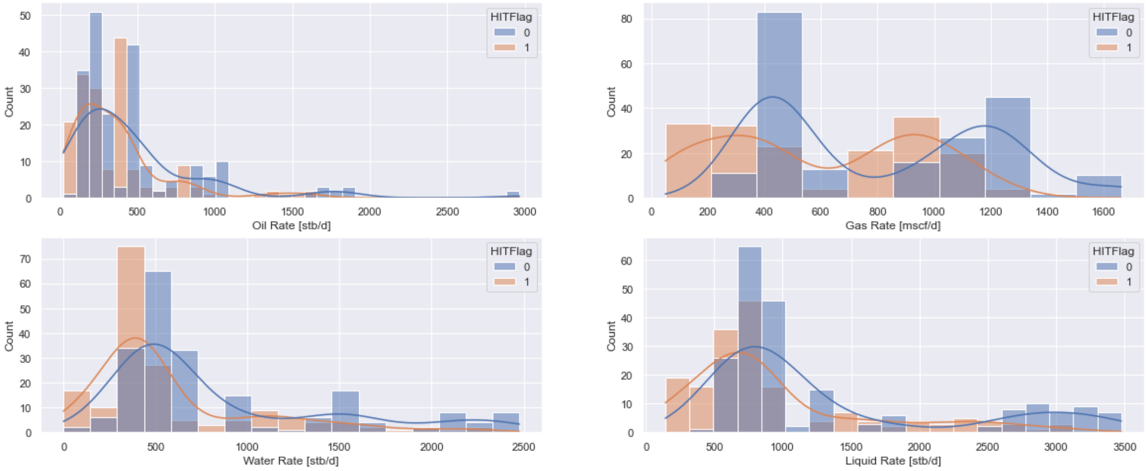

Figure 3.2: Distribution of oil, water and gas rates in dataset

A similar process was done for the change between days, as shown in Figure 3.3, where the change in liquid rate and change in pump intake pressure is displayed. It can be observed that the two groups are more distinct, with the change in pump intake pressure being generally negative between days before a tubing leak and generally positive after a tubing leak. Similar behavior can be observed for the change in liquid rate between days, where the distribution is more left-tailed after a tubing leak than before. This is again consistent with our physical understanding of the phenomena occurring during a tubing leak.

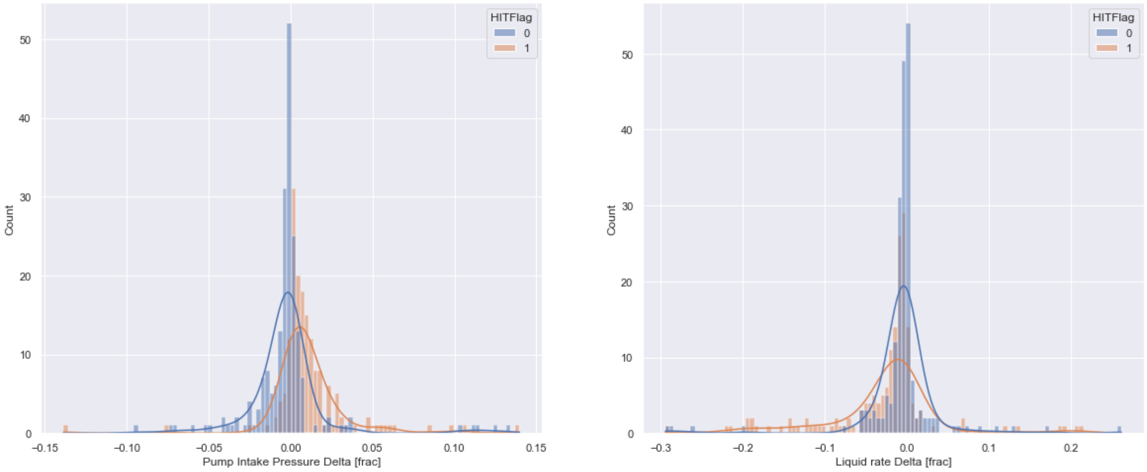

Figure 3.3: Distribution of change in real-time parameters in dataset

The data was further reviewed by reviewing correlation between features. Where features were highly correlated, this provided an opportunity to develop our understanding of their relationship and remove redundant variables. Ideally, as few features should be used in the model as possible in order to ensure the model is learning the correct relationships and the interpretability remains high.

For example, the correlation matrix shown in Figure 3.4 demonstrates high correlation between oil rate, water rate and liquid rate. As the liquid rate is the sum of the water and oil, it is intuitive that they would be correlated. The water and oil rate features were not taken into the model building as they were captured implicitly by the liquid rate feature. Similarly, motor current correlated highly with liquid rate and so too was not taken into the modelling.

Based on the physical understanding of what parameters should change during a tubing leak and the correlation between them, 8 parameters were selected to take into the model building phase and are shown in Table 3.2. If a suitably predictive and interpretable model could not be found, the feature engineering and selection would be revisited.

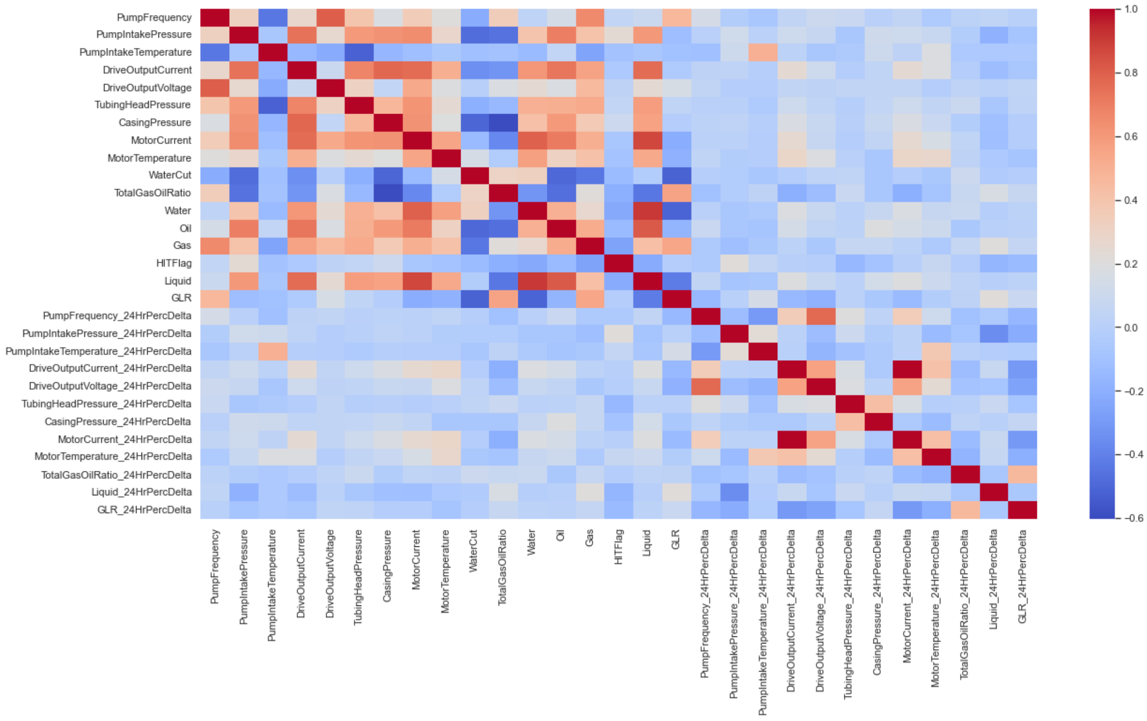

Figure 3.4: Correlation Matrix of features

3.2.5 Data Standardization

The data was then standardized so that the mean of each feature μ_j was zero and the standard deviation σ_j was 1. This is so that the features have a similar range and have approximately gaussian distributions. This is a requirement for many learning algorithms as features with a much larger scale than others may dominate the objective function. This process was done using the z-score transformation, shown in the equation below:

$$x_{ij}^{z} = \frac{x_{ij} - \mu_j}{\sigma_j} \tag{3.5}$$

Table 3.2: Features used for model building

Feature	unit	Data source
Tubing head pressure	psig	real-time
Pump frequency	hertz	real-time
Pump intake pressure	psig	real-time
Liquid flow rate	stb/d	well test
Change in tubing head pressure	%	real-time
Change in pump frequency	%	real-time
Change in pump intake pressure	%	real-time
Change in liquid flow rate	%	well test

3.3 Machine Learning Model

3.3.1 Model Formulation

The task of identifying tubing leaks from time series data can be described as a *supervised* ML problem, as the algorithm used will have access to both the labelled input dataset and the desired output. This problem is also known as a *classification* problem, as the desired output has a discrete form, either "0" or "1".

Four different machine learning models were evaluated for classifying tubing leaks based on the available measured data. These models were Random Forest classifier, Gradient Boosted Classifier, Support Vector Machine and k-Nearest Neighbor model.

The overall process for training and evaluating each of the machine learning models is shown in Figure 3.5.

3.3.2 Splitting the dataset

The dataset was split into distinct testing and training datasets. This is so that the performance evaluation of the learning algorithms is not based on data it has trained itself upon and to avoid perfectly fitting the dataset provided. A split of 67% for training and 33% for testing was carried out in a randomized fashion whilst trying to maintain the same proportion of observations before and after a tubing leak. The splits are shown in Table 3.3.

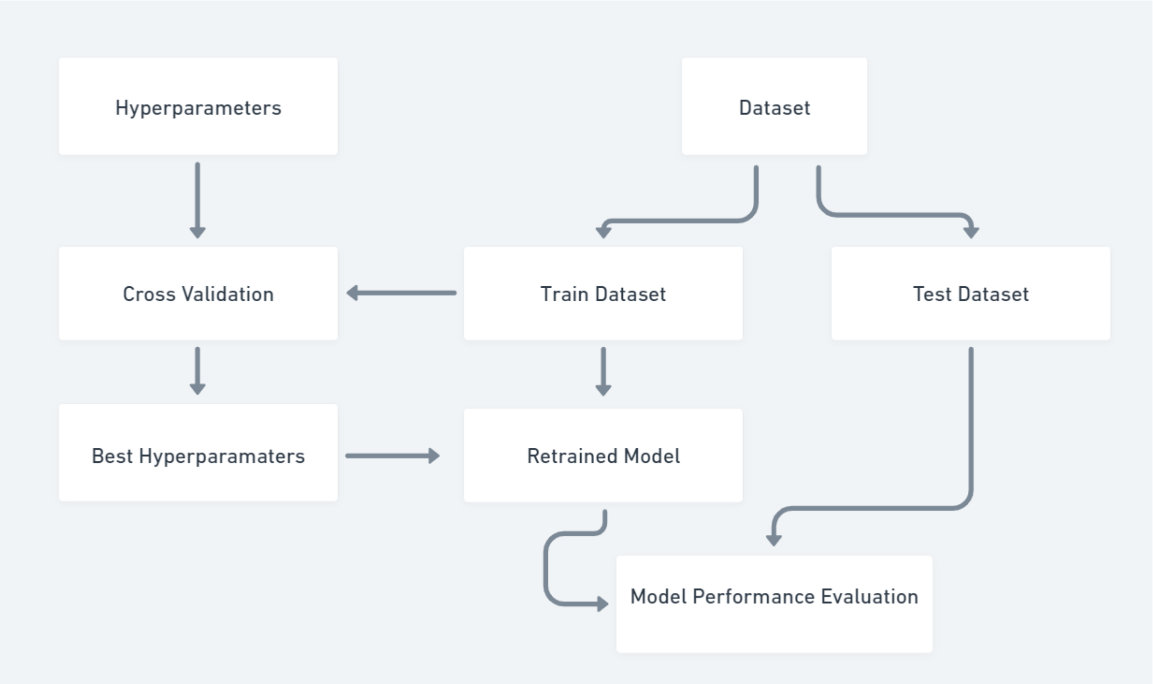

Figure 3.5: Workflow for training and evaluating models

Table 3.3: Split of data between Train and Test sets along with class proportions

	Total Observations	HITFlag = 0	HITFlag = 1
Training Dataset	254	136	118
Test Dataset	126	71	55

3.3.3 Hyperparameter Selection

Each model had a number of hyperparameters which would be evaluated using 10 fold cross-validation of the training dataset. The hyperparameters will be evaluated using an exhaustive grid-search method. The hyperparameters for each model are shown in Table 3.4. After the best hyperparameters were determined, the model with these hyperparameters will be retrained on the entire training dataset and the best model determined by its performance on the test dataset.

Table 3.4: Hyperparameters for each machine learning model evaluated with cross-validation

Model	**Hyperparameter**	**Values**
RF	Number of trees	10, 20, 30, 40, 50, 60, 70, 80, 90
	Maximum Features	5, 6, 7, 8
XGBoost	Number of trees	10, 20, 30, 40, 50, 60, 70, 80, 90
	Maximum Features	5, 6, 7, 8
SVM	C	0.1, 1, 5, 10, 100, 1000
	Gamma	0.0001, 0.001, 0.01, 0.1
kNN	Number of Neighbours (k)	3, 5, 7, 9, 11, 13

3.4 Analytical Model

3.4.1 Model Formulation

The analytical model will be explicitly programmed so it adheres to detecting the expected change in measured parameters for a tubing leak. These are:

1. Increasing Pump Intake Pressure
2. Decreasing Tubing Head Pressure
3. Decreasing Liquid Rate at surface
4. Constant Pump Frequency

When all four criteria is met, the analytical model will classify the observation as a tubing leak. In order to align the measurement frequency, similar to the machine learning model formulation, an average of the real-time measurements over one day will be used. A percentage change between days will also be used by the model to denote changing measurements. These are the change in pump intake pressure (ΔPIP), change in tubing head pressure (ΔTHP), change in pump frequency (ΔFreq), and change in liquid flow rate (ΔQ_L).

In addition, rather than only capturing changes in the expected direction, a threshold parameter will be used for each measurement to denote what is considered a change. This will capture observations where perhaps all three measurements change in the expected direction except one which changes in the opposite direction only by a negligible amount. These thresholds are denoted by θ with the accompanying subscript, for example, the threshold change in tubing head pressure θ_{THP}. As such, the formulation of the analytical model is shown in algorithm 1

```
**Data:** $\Delta$PIP, $\Delta$THP, $\Delta$Freq, $\Delta Q_L$
**Result:** Classification of Tubing Leak, $J \in \{0, 1\}$
**if** $(\Delta PIP > \theta_{PIP}) \wedge (\Delta THP < \theta_{THP}) \wedge (\Delta Q_L < \theta_{Q_L}) \wedge (|\Delta Freq| < \theta_{Freq})$ **then**
    Tubing Leak Detected, $J = 1$;
**else**
    No tubing leak detected, $J = 0$;
**end**
```

Algorithm 1: Analytical method for determining if tubing leak has occurred

3.4.2 Hyperparameter tuning

The selection of the hyperparameters θ will be done by evaluating the performance using the same training dataset. The hyperparameters will be evaluated using an exhaustive grid-search method. As the model does not need to learn parameters, no cross validation will be done. The hyperparameters evaluated for the analytical model are shown in Table 3.5.

Table 3.5: Hyperparameters evaluated for the analytical model

Hyperparameter	**Values**
θ_{PIP} [%]	-5, -4, -3, -2, -1, 0, 1, 2, 3, 4, 5, 6, 7, 8, 9
θ_{THP} [%]	-5, -4, -3, -2, -1, 0, 1, 2, 3, 4, 5, 6, 7, 8, 9
θ_{Q_L} [%]	-5, -4, -3, -2, -1, 0, 1, 2, 3, 4, 5, 6, 7, 8, 9
θ_{Freq} [%]	0, 1, 2, 3, 4, 5, 6, 7, 8, 9

3.5 Evaluating Model Performance

For both the ML and analytical model, a number of different performance metrics will be calculated for selecting the best hyperparameters and determining the most effective model for detecting tubing leaks. These performance metrics are shown below:

$$\text{Accuracy } (\%) = \frac{TP + TN}{TP + FP + TN + FN} \times 100 \tag{3.6}$$

$$\text{Precision}, p\ (\%) = \frac{TP}{TP + FP} \times 100 \tag{3.7}$$

$$\text{Recall}, r\ (\%) = \frac{TP}{TP + FN} \times 100 \tag{3.8}$$

$$\text{F-score } (\%) = \frac{2(p \times r)}{p + r} \times 100 \tag{3.9}$$

Where TP represents the number of true positives, TN is the number of true negatives, FP is the number of false positives and FN is the number of false negatives. The F-score is the harmonic mean of the precision and recall and so implicitly captures both in a single metric. For all model evaluations, the most important metric will be the F-score.

In addition to the quantitative metrics described above, it is important that both models are easily interpretable. Whilst the analytical model has been explicitly programmed and its behavior should therefore be intuitive, the same cannot be said for the ML model. To overcome this issue, techniques such as reviewing the importance of features and Shapley scores will be used to build up intuition about the model's behavior.

CHAPTER 4

RESULTS

4.1 Machine Learning Model

4.1.1 Hyperparameter selection

The best hyperparameters and their associated F-scores for each of the four machine learning models are shown in Table 4.1. This table shows the mean F-score based on 10-fold cross-validation and the standard deviation. All models provided high scores on the training dataset based on the cross-validation results.

Table 4.1: Selected hyperparameters and their performance for each machine learning model evaluated with 10-fold cross-validation

Model	**Hyperparameter**	**Best Value**	**F-score**
RF	Number of trees	30	93.65 ± 3.65
	Maximum Features	5	
XGBoost	Number of trees	40	94.84 ± 4.02
	Maximum Features	7	
SVM	C	100	90.39 ± 5.45
	Gamma	0.1	
kNN	Number of Neighbors (k)	3	87.94 ± 4.46

4.1.2 Evaluation on test set

The models with the best hyperparameters were then retrained on all of the training dataset and their performance evaluated on the test set. The results are shown in Table 4.2 and Figure 4.1.

The results show the XGBoost model has the highest F-score compared to the other ML models based on the test set. Whilst the RF model provides close scores, the kNN and SVM model provide good performance albeit lower scores on the test set.

Table 4.2: Results of ML models on test set

Model	Train Accuracy	Test Accuracy	Test F-score	Test Precision	Test Recall
RF	100	92.1	91.3	88.1	94.5
XGBoost	99.6	94.4	94.4	92.9	94.5
SVM	98.0	88.9	88.8	86.0	89.1
kNN	94.5	90.5	90.3	90.6	87.3

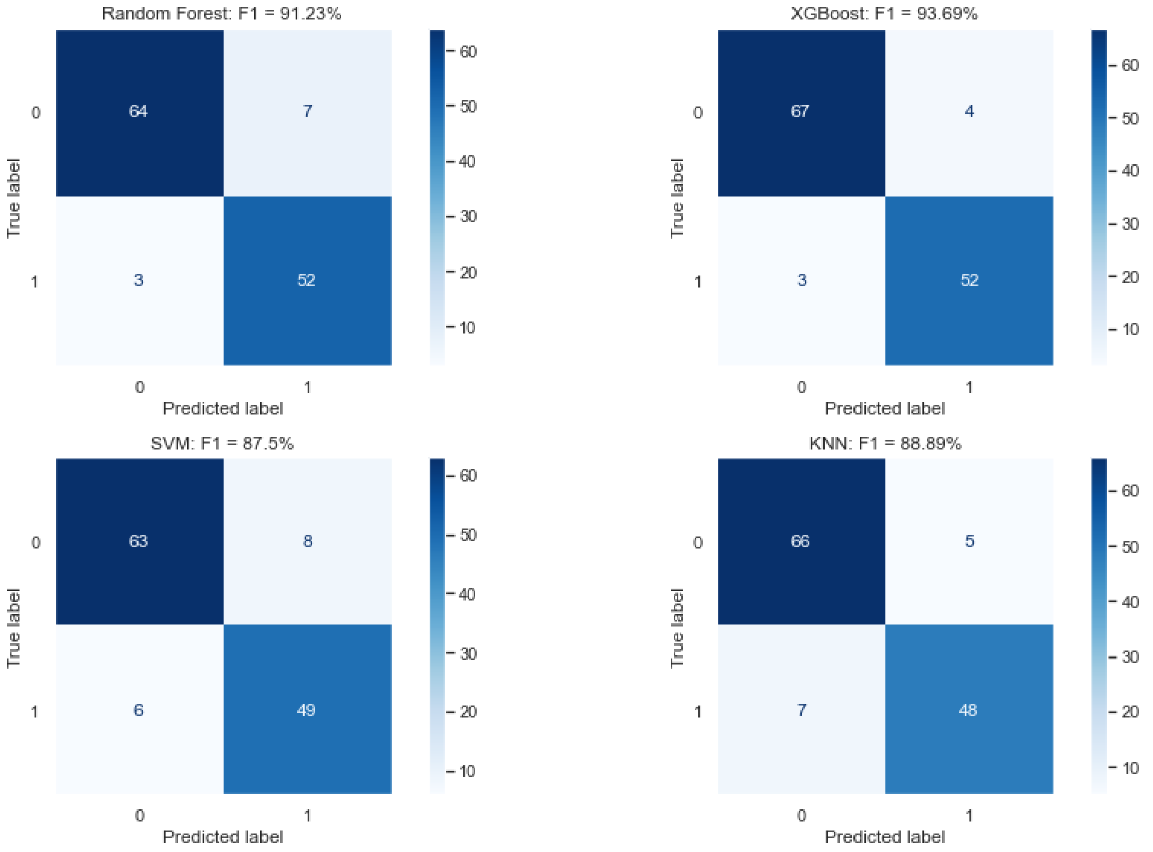

Figure 4.1: Confusion matrix for each ML model based on test set

4.2 Analytical Model

4.2.1 Hyperparameter Selection

The F-score of the analytical model using different hyperparameters is shown in Figure 4.2 with the red dot indicating the selection of hyperparameters. The values which provided the best F-score is shown in Table 4.3, aligning with the red mark on Figure 4.2.

Therefore, the best performing analytical model would detect a tubing leak when

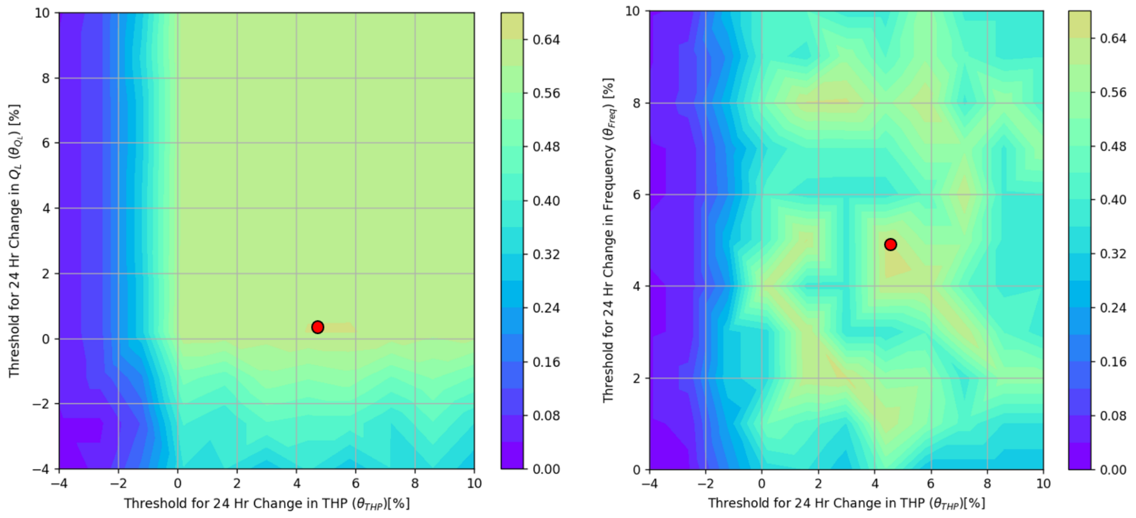

Figure 4.2: Analytical model hyperparameter tuning

Table 4.3: Selected hyperparameters for analytical model

Hyperparameter	**Best value**
θ_{PIP} [%]	0
θ_{THP} [%]	5
θ_{Q_L} [%]	2
θ_{Freq} [%]	5

- The change in pump intake pressure was $> 0\%$
- The change in tubing head pressure was $< +5\%$
- The change in liquid flow rate was $< +2\%$
- The absolute change in pump frequency was $< +5\%$

4.2.2 Evaluation on test set

Using this model, its performance was evaluated on the same test set as the machine learning model. The results for various metrics are shown in Table 4.4.

Table 4.4: Performance metrics of analytical model on test set

Metric	**Train data set**	**Test data set**
F-score	75.3%	80.7 %
Accuracy	75.2 %	82.5%
Precision	70.1%	77.9%
Recall	81.3%	83.6 %

CHAPTER 5
DISCUSSION

5.1 Interpretation of machine learning model results

As the machine learning models are not explicitly programmed, it is important to further understand how the models are making their decisions and the importance of each feature, rather than just review the quantitative performance metrics. A model which is accurate on the test set but has learnt unphysical relationships will not be useful or generalize well.

5.1.1 Mean decrease in F-score

The importance of each feature on all four ML models was reviewed using a mean decrease in F-scores, which are shown in Figure 5.1. These results are provided by calculating the decrease in F-score when a single variable is randomly shuffled. This procedure is repeated a number of times to reduce the variance in the results. The process of randomly shuffling values will eliminate the relationship between the feature and prediction, thereby indicating how much the model depends on it to make a prediction [8].

As observed in Figure 5.1, liquid rate, pump intake pressure and change in pump intake pressure result in the largest decrease in F-score for both the RF and XGBoost model. This is consistent with the physical understanding of what should happen as a result of a tubing leak.

For the SVM and kNN models, pump frequency results in a more significant drop of F-score and change in pump intake pressure is not as influential. This behavior is not consistent with our understanding of what should be physically occurring during a tubing leak.

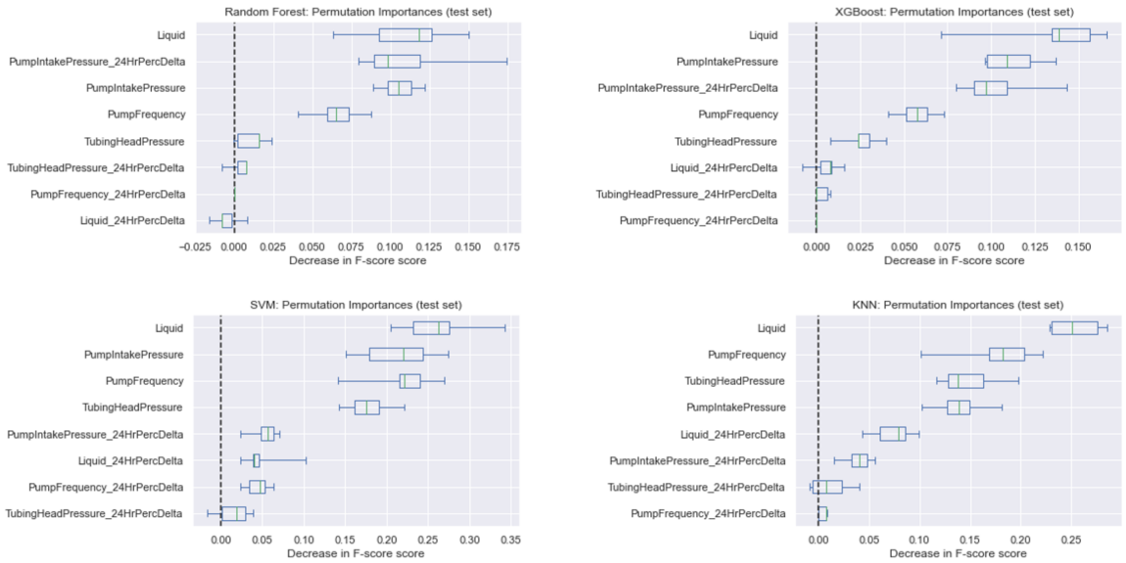

Figure 5.1: Feature importance using mean decrease in f-score for all ML models

5.1.2 Shapley Scores

The understanding of how the ML model is making its decision can be further evaluated using Shapley Scores. Shapley scores reflect the marginal difference in prediction output when adding in a single feature. These are determined by randomly replacing one feature with another observations value and measuring the difference in model output. The value reported is the average contribution for each feature over all permutations. These values can be used to understand which features are important to an ML model and how the model output will respond to the value of a given feature [9].

As the XGBoost is the best performing model, the Shapley scores only for this model were evaluated. The absolute mean Shapley values are shown in Figure 5.2 and demonstrate the change in pump intake pressure, liquid rate and pump intake pressure provide the most significant impact to model output, which is consistent with the physical understanding of how a tubing leak should present itself. The Shapley scores indicate that pump frequency is also more important than changes in liquid rate and tubing head pressure, which is not as intuitive.

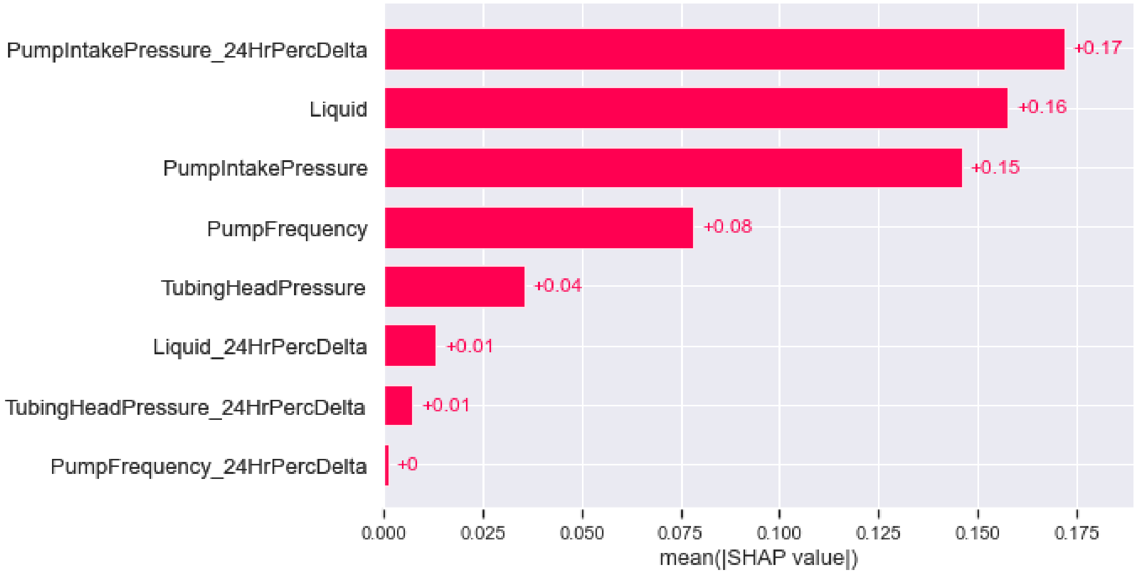

Figure 5.2: Mean shapley scores for XGBoost model

The shapley scores of each observation in the test set is summarized in Figure 5.3. This plot shows the Shapley value of the feature based its position on the x-axis and the color demonstrates its relative value (i.e. “High” or “Low”). These charts help develop an understanding of how the model outputs directionally respond to the features. For example, Figure 5.3 indicates that when the change in pump intake pressure is “High”, this feature generally results in a large positive contribution to the output. This is consistent with our understanding of the inputs, as a rising intake pressure should result in a positive indication of a tubing leak.

5.2 Performance between machine learning and analytical model

From the results shown in subsection 4.1.2, the XGBoost model provides the highest F-score of all the ML models. In addition, the results shown in subsection 4.2.2 highlight that all four ML models provide a higher F-score based on the same test set than the analytical model. There are many reasons for this behavior, including more features used in the ML models and greater model complexity.

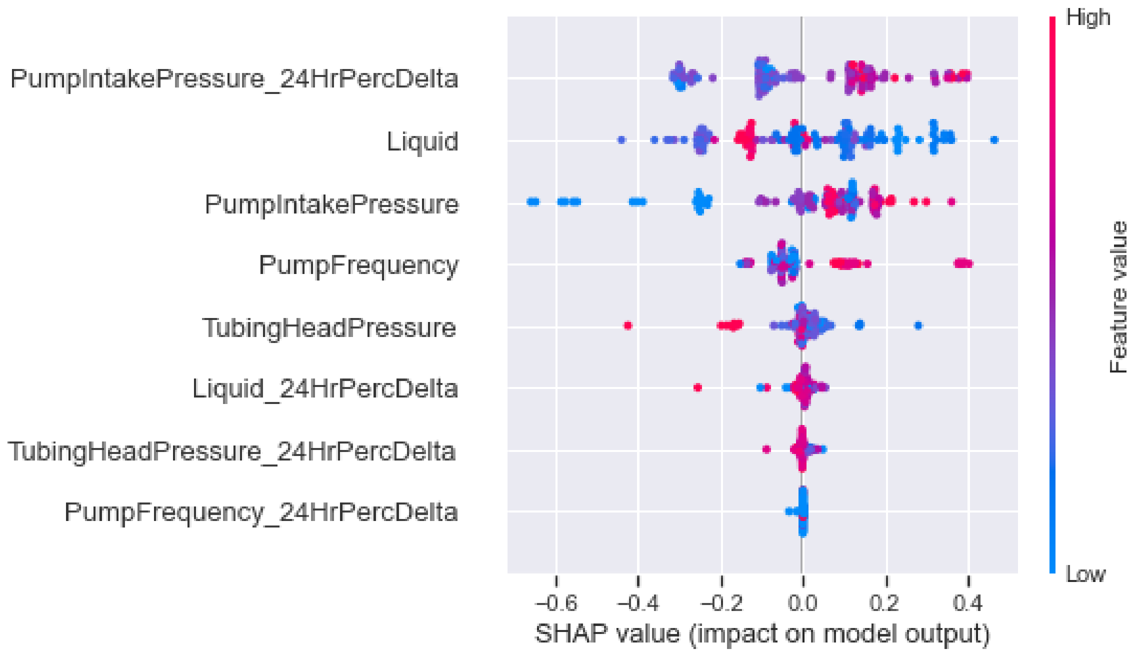

Figure 5.3: Summary of Shapley scores for test set using XGBoost model

An example from one of the wells is shown in Figure 5.4, where the measured parameters are shown along with the prediction from the XGBoost model and the analytical model. In this example, both models provide similarly accurate predictions and are able to correctly indicate the days on which a tubing leak has occurred.

Another example well is shown in Figure 5.5, where the two models show more different predictions. Before the tubing leak occurs, the analytical model provides many false positive indications of a tubing leak. This is due to small changes in parameters like pump intake pressure where small decreases or increases have a large impact on the prediction. The machine learning model also shows some fluctuations on some days before the tubing leak occurs but as these are below 50%, these are not classified as a tubing leak.

After the tubing leak has occurred, indicated by the red shaded area, the analytical model provides a number of false negative indications due to one parameter moving in a counter-intuitive direction, like the small decrease in pump intake pressure. The machine learning model indicates similar behavior, showing some false negatives after the tubing leak occured, but not as many as the analytical model. This indicates it is perhaps better

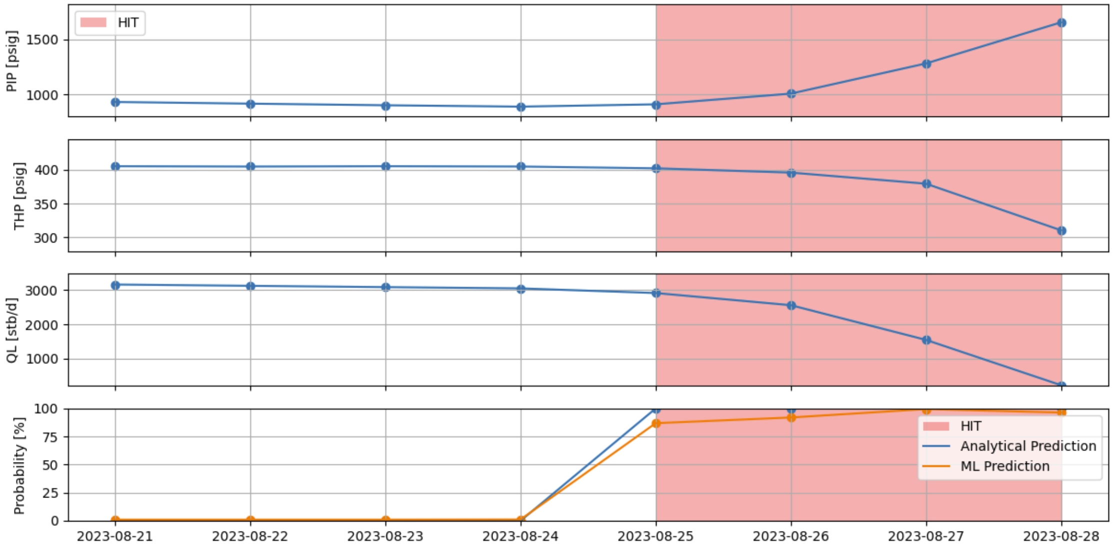

Figure 5.4: Comparison of prediction between ML and analytical model, where both models show good agreement

able to balance multiple readings where not all are in agreement.

5.3 Feature Selection

For both the XGBoost and analytical model, the pump frequency is a required parameter to make a prediction of whether a tubing leak has occured. When reviewing the feature importance plots in Figure 5.1 and Figure 5.2, the impact of the change in pump frequency appears much less influential than the other parameters.

As the pump frequency has such a large impact on the hydraulics of the system, it was decided to keep this feature in the model. The data set used for this project was relatively small and based on our domain knowledge, it was decided to keep this fundamental feature in the data set. In addition, no additional features were added into the model given the high performance on the test set.

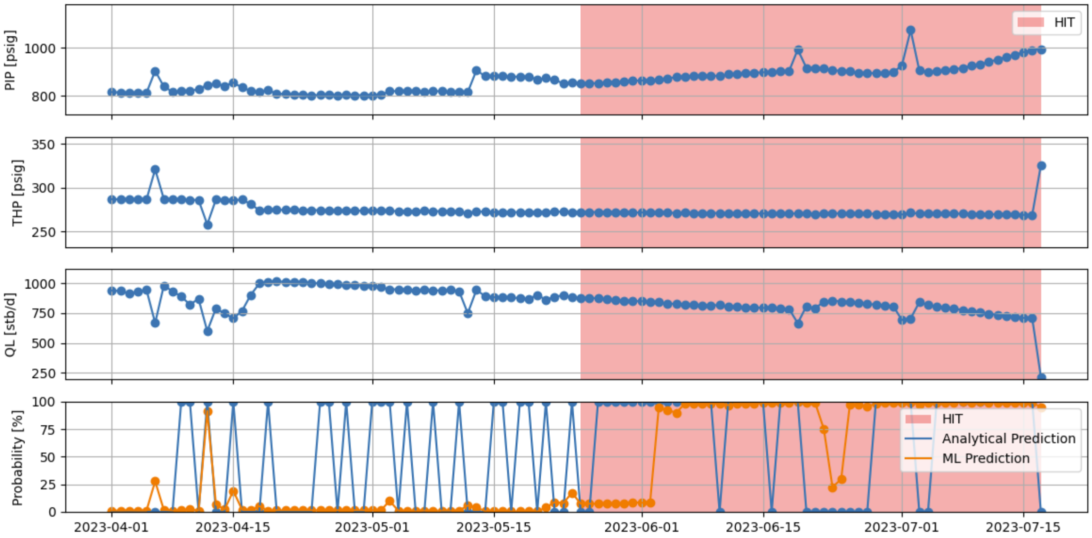

Figure 5.5: Another comparison between ML and analytical models, where analytical model shows many false positives

CHAPTER 6
CONCLUSION

The objective of this project was to evaluate the use of both machine learning and analytical models to detect tubing leaks in oil wells using available measured parameters. A methodology was outlined which involved preprocessing the data, extracting features, exploratory data analysis and data standardization before building the models.

From the dataset of 15 wells, a range of different ML models were evaluated and all provided strong performance on the test set, with the XGBoost model showing the best performance with an F-score of 94.4 %.

An analytical model was also built which used the physical understanding of how a tubing leak would manifest and showed strong performance on the test set with an F-score of 80.7 %. This score, however, was less than any of the ML models evaluated.

The performance of the ML model was evaluated further using a number of metrics such as feature importance with mean decrease in F-score and Shapley scores. Both of these indicated conformance with the physical understanding of how a tubing leak would manifest, which brings confidence to its predictive ability.

6.1 Recommendations for further work

Whilst the work in this project did achieve the objectives outlined, further work could be done to extend the findings, such as:

- One of the assumptions of the work is that the observations were all independent of each other. Apart from the change in parameters between days, each classification was made in isolation of the observations around it. This resulted in large variations in predictions between days. The probability of a tubing leak, however, should increase relatively monotonically with time. More advanced ML techniques like recurrent neural networks could be used to overcome this problem and consider more available data prior to making a prediction.

- More time-based features could be reviewed, such as moving averages with larger windows or momentum between long time periods and their impact on prediction results.

- Given the strong performance of the ML model, no further work on feature engineering or feature selection was performed. Having said this, further work could be done on evaluating how few parameters could be used without degrading performance significantly.

- The dataset used was selected from wells where the tubing leak could be observed in the monitoring data available. This was done so that the model would learn the correct patterns. As a result, the performance of the model reported in this project could be optimistic compared to when deployed into a live environment. Further work could be done on using samples with noisy monitoring signals and ones which are difficult to interpret for a human.

REFERENCES

[1] T. B. of Safety and E. Enforcement, “Best practices for real time monitoring of offshore well construction,” Summary of Best Practices, 2017.

[2] G. Oliva, H. Galvão, D. Santos, A. Maitelli, R. Costa, and C. Maitelli, “Gas effect in esp system stage by stage analysis,” May 2015.

[3] J. Kikani, *Reservoir Surveillance*. Society of Petroleum Engineers, ISBN: 978-1-61399-304-0.

[4] P. Bangert, “Chapter 3 - machine learning,” in *Machine Learning and Data Science in the Oil and Gas Industry*, P. Bangert, Ed., Gulf Professional Publishing, 2021, pp. 37–67, ISBN: 978-0-12-820714-7.

[5] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning* (Springer Series in Statistics). New York, NY, USA: Springer New York Inc., 2001.

[6] J. Golze, S. Zourlidou, and M. Sester, “Traffic regulator detection using gps trajectories,” *KN - Journal of Cartography and Geographic Information*, vol. 16, Jul. 2020.

[7] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin, Heidelberg: Springer-Verlag, 2006, ISBN: 0387310738.

[8] F. Pedregosa *et al.*, “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[9] C. Molnar, *Interpretable Machine Learning, A Guide for Making Black Box Models Explainable*, 2nd ed. 2022.