Risk Prediction of Gas Wells Based on the Existing Inactive Wells in Alberta

by Siavash Fard

June 2020

Contents

Abstract

Alberta has an estimated 95,000 inactive and 2,963 orphan wells as of May 1, 2020. With growing number of inactive wells in Alberta, a study on risk classes and abandonment complications of these wells would be informative for oil and gas contractors and service providers. These inactive wells are grouped into three classes of low, medium, and high risks. A well risk type is mostly determined by its well type (e.g. fluid type disposal, injection, etc.), oil and gas production history, H2S content. Both suspension and abandonment requirements vary based on AER risk assessment. If the probability of risk category of a well at certain coordinates can be predicted within their first few months of production (or before well completion stage), operators can proactively design a remediation program which addresses the concerns, and therefore, reduce the costs later at the suspension and abandonment stage.

In this report, public data from AER and geoSCOUT have been used with six different algorithms to predict the risk category of gas wells based on the existing data. The predictive models used in this report performed better than random guessing. Out of these six predictors, Random Forest performed better than other models with the Accuracy and F1 Score of 0.91. For Random Forest model, TVD has the highest importance followed by gas production data.

It is highly recommended to use these models with drilling data for prediction of risk classes. However, only production data and some limited well properties were available.

Introduction

Background

Alberta is called Canada's energy province. The province produces over 80 percent of Canada's crude oil and contributes over 25 percent to the province's GDP. Since 2015, the province oil sector has been suffering a great deal mostly due to crude oil prices resulting in bankruptcies for a significant number of companies, and therefore, inability to properly decommission and abandon oil wells across the province, and leaving the clean up at costs of falling to taxpayers. More importantly, these inactive wells pose a major threat to public health and safety. Alberta has an estimated 95,000 inactive and 2,963 orphan wells as of May 1, 2020.

Alberta Energy Regulator (AER), the regulating body of the province, has several directives for well abandonment which operators must follow. AER assesses inactive wells into three risk categories (i.e. high, medium, and low risk). A well risk type is mostly determined by its completion type, gas production history, H2S content. Both suspension and abandonment requirements vary based on AER risk assessment. Associated costs and safety, therefore, are greatly impacted based on the risk category.

Unfortunately, to the best of author's knowledge, there has not been any study on risk evaluation of oil and gas wells during their lifetime. Montague et. al tried predicting gas migration from existing wells using machine learning algorithms and found Random Forest to be the best classifier.

Motivation

Millions of dollars are spent on remediation and reclamation of inactive wells with medium and high risks. If the probability of risk category of a well at certain coordinates can be predicted within their first few months of production, operators can proactively design a remediation program which addresses the concerns, and therefore, reduce the costs later at the suspension and abandonment stage.

In this report, six machine learning models will be used to predict the risk category of gas wells based on the existing data. The author believes that this model can be used by operators to better evaluate their drilling and completions program for future abandonment programs.

Methodology

Data

The data for this report is collected from multiple sources. AER frequently published a list of all inactive wells (http://www1.aer.ca/ProductCatalogue/360.html) and their risk categories with their legal subdivisions (LSD). LSD is based on Alberta Township Survey System (ATS) which needs to be

converted to coordinates using data downloaded from Altalis Ltd. (https://www.altalis.com/), geospatial data distributor based in Calgary, Alberta.

Next, the data for each well is extracted from geoSCOUT, a tool which has a library of oil and gas database. From the geoSCOUT public database for inactive wells, a set of parameters (e.g. production data) will be imported, normalized, and modelled using different classification approaches.

Data Wrangling and Exploratory Data Analysis

AER Inactive Dataset

AER inactive well list contains information for more than 95,000 wells (as of May 1, 2020). The csv file includes information such as UWI, Company, License Number, LSD, Field Area, Final Drill Date, Inspection Date and AER Risk Class. As per AER Directive 13, inactive (suspended) wells are categorized into Low, Medium and High risk levels (see Glossary for definitions). Figure 1 shows the risk classes for all inactive wells in Alberta. Majority of the inactive wells are either low or medium and mostly are compliant with AER regulations. However, the number of 'Medium' risk wells is misleading because low risk class wells are categorized as 'Medium' if suspended more than 10 years (type 6 medium). There are 28,858 Type 6 Medium wells among these suspended wells. As a result, all 'Type 6' and 'Not Reported' wells have been removed from the dataset. If type 6 wells are removed from the dataset, there are only 5,338 wells with 'Medium' risk class. This may cause imbalance in our dataset later when predicting the risk class type.

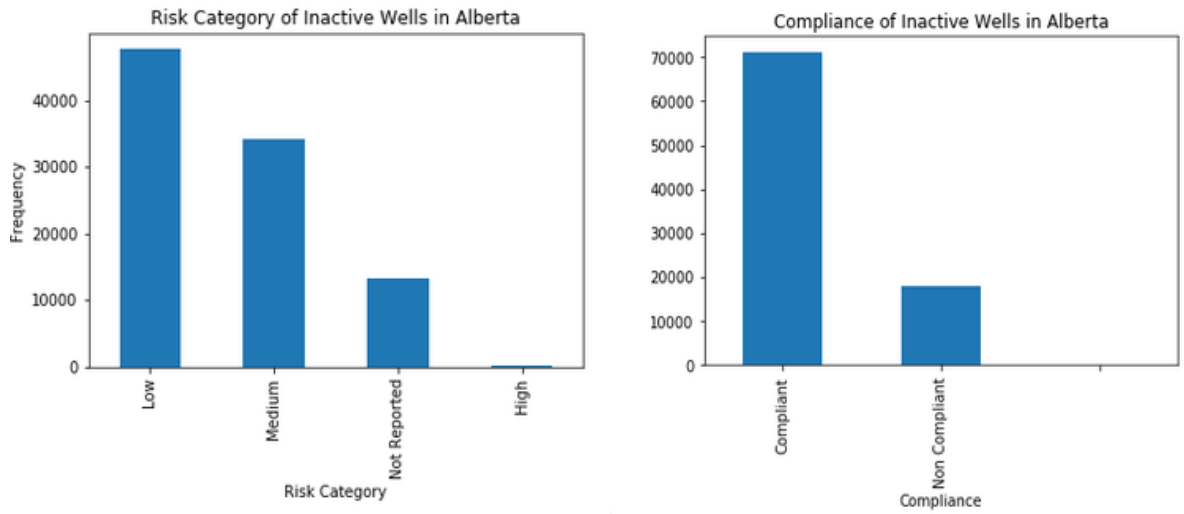


Figure 1 Risk category and compliance of Alberta inactive wells

The categorical column of risk class needs to be converted to numerical values in order to be included in the classification modelling. As a result, 'Low', 'Medium' and 'High' risk classes have been replaced by 0, 1 and 2 respectively.

Last, AER inactive list contains well type information (e.g. H_2S level, injection well, disposal well, etc.). The risk class of an inactive well, therefore, is determined by its well type. For example, wells with more

than 50 mole/kmole H_2S level are ‘Medium’ risk or Class 4 (steam or portable water) injection wells are considered ‘Low’ risk (see Section 2 of Directive 51). As a result, risk classes of most wells are determined even before drilling starts. Therefore, only well types that still can be used in prediction are those with gas production. Table 1 shows the first five rows of the final dataset after removing eight (8) columns.

Table 1 First five rows of AER inactive dataset after removing eight (8) columns

Company	Abbreviated Well ID	LSD	Well Type	AER Risk Class
Long Run Exploration Ltd.	00/11-21-057-21W4/2	11-21-57-21W4	Gas Wells < 28000 m3/Day that are low risk as ...	Low
Crescent Point Energy Corp.	00/13-31-056-06W4/0	13-31-56-6W4	Gas Wells < 28000 m3/Day that are low risk as ...	Low
Long Run Exploration Ltd.	00/02-36-031-14W4/2	2-36-31-14W4	Gas Wells < 28000 m3/Day that are low risk as ...	Low
NEP Canada ULC	00/07-35-050-26W4/3	7-35-50-26W4	Gas Wells < 28000 m3/Day that are low risk as ...	Low
NEP Canada ULC	00/02-28-051-26W4/2	2-28-51-26W4	Gas Wells that can be Medium Risk as...	Medium

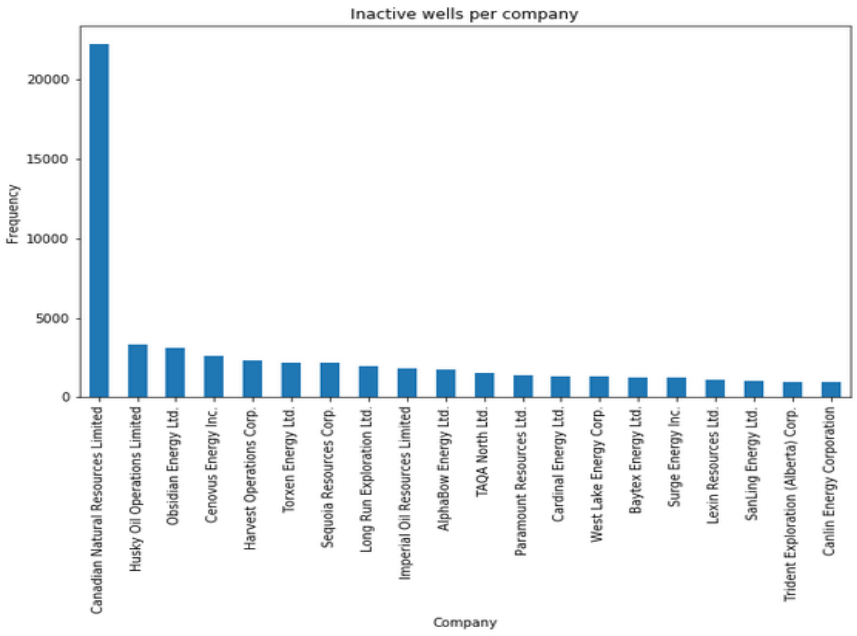


Figure 2 Companies with highest number of inactive wells in Alberta (top 20). CNRL has 22,227 inactive wells in Alberta as of May 1, 2020.

goeSCOUT dataset

The data on well properties were extracted from geoSCOUT public data. The data included construction information (e.g. spud dates, drilling days, etc.), physical properties (e.g. total depth, casing size and diameter, casing weight) and most importantly the production data from first 12 months, cumulative and

last 12 months. The dataset comprises of 94,968 wells but because of missing values, the size significantly reduced after cleaning.

Irrelevant information from the geoSCOUT file (for the list of inactive wells) were removed. Number of columns reduced from 103 to 54 after removing the irrelevant information. Unfortunately, only data accessible from geoSCOUT was production and some other limited information such as bottom hole temperature, casing sizes and depths. Only available drilling related data was 'Drilling Problems' which is a summary of pressure related problems during drilling. 'Drilling Problems' were converted to numerical values to include in the modelling.

Table 2 Numerical values for Drilling Problems of the inactive wells

Drilling Problem	Total Number of Wells
Blow	4
Water Flow	3
Kick	2
High Pressure	1
No Problem	0
Lost Circulation	-1

Because we did not have access to drilling and completions data, the first 12 months of production (of gas and water) as well as drilling problems and total depth were used to predict the risk class of the gas wells. Note that as per Directive 13, gas wells with more than 28,000 cu m/ day production are categorized as 'Medium' risk. Therefore, it may be interesting to know how the first 12 months production can affect the risk class.

After removing null values from the production data and merging two datasets (geoSCOUT and AER inactive wells list), 6,880 wells remained for modelling. Out of these 6,880 wells, there are 6,386 wells with 'Low' risk class leaving only 493 wells with 'Medium' risk and 1 with 'High' risk. The only 'High' risk well was removed from the dataset in order to have a binary classification of the output.

Since the size of 'Medium' risk class is small, we should use oversampling approach to balance the dataset. After implementing SMOTETomek on the dataset, our dataset size increased to11,818 for both 'Low' and 'Medium'.

Machine Learning Algorithms

Six different methods were used to predict the risk classification of these 11,818 wells. Same training and test datasets were used for each model for consistency and comparing the results. These predictive models include KNN, Decision Tree, SVM, Logistic Regression, AdaBoost and Random Forest.

Logistic Regression is the simplest model that takes in a linear combination of input parameters and predicts a binary response. All five different solvers were used to find the parameter weights and minimize the cost function.

An SVM is a form of neural network that uses training data to form a discriminating hyperplane that separates data with the maximum possible margin between data classifications. Random forests are a

collection of decision trees, each trained on a small, random subset of features. All four kernel types were tried to find out the highest accuracy.

KNN (K-Nearest Neighbor) categorizes the datapoints by calculating the distance between each two points. The algorithm is a simple, easy-to-implement supervised machine learning algorithm that can be used to solve both classification and regression problems. The corresponding K value for the highest accuracy (for values from 1 to 15) was chosen for the final predictive model.

An AdaBoost classifier is a meta-estimator that begins by fitting a classifier on the original dataset and then fits additional copies of the classifier on the same dataset but where the weights of incorrectly classified instances are adjusted such that subsequent classifiers focus more on difficult cases.

The decision tree classifier (Pang-Ning et al., 2006) creates the classification model by building a decision tree. Each node in the tree specifies a test on an attribute, each branch descending from that node corresponds to one of the possible values for that attribute.

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.

Model Performance

Several metrices were used to evaluate the performance of the predictive models. Accuracy, F1 Score and Log Loss (if applicable) were used as three metrices for model performance.

Accuracy is the most typical metric used in machine learning problems. It is simply the ratio of correct predictions to the total number of input samples. Since we have already solved the problem of imbalance dataset, it can be used as one of our metrices.

Logarithmic Loss is the second metric used in the model evaluation If there are N samples belonging to M classes, then the Log Loss is calculated as below:

$$LogarithmicLoss = \frac{-1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij} \log\left(p_{ij}\right)$$

where,

y_{ij}, indicates whether sample i belongs to class j or not

p_{ij}, indicates the probability of sample i belonging to class j

Log Loss has no upper bound and it exists on the range [0, ∞). Log Loss nearer to 0 indicates higher accuracy, whereas if the Log Loss is away from 0 then it indicates lower accuracy.

In general, minimizing Log Loss gives greater accuracy for the classifier.

Finally, F1 Score is used to measure test accuracy. F1 Score is the mean between precision (a measure of result relevancy) and recall (measure of how many truly relevant results are returned). High precision but lower recall, gives an extremely accurate, but it then misses a large number of instances that are difficult to classify. The greater the F1 Score, the better is the performance of our model. Mathematically, it can be expressed as:

$$F1 = \frac{2}{\frac{1}{precision} + \frac{1}{recall}}$$

where,

$$Precision = \frac{TruePositive}{TruePositives + FalsePositives}$$

and,

$$Recall = \frac{TruePositive}{TruePositives + FalseNegatives}$$

Results and Discussion

Companies with Lowest and Highest Risk Factors

In order to identify the companies with wells with lowest risk classes, we need to convert the risk classes to risk factors. 'Low', 'Medium' and 'High' risk classes have been replaced by 0, 1 and 2 respectively and averaged to give the companies with lowest risk classes. Table 1 and 2 show the average risk class for five companies with the lowest and highest risk levels (including all type 6 wells).

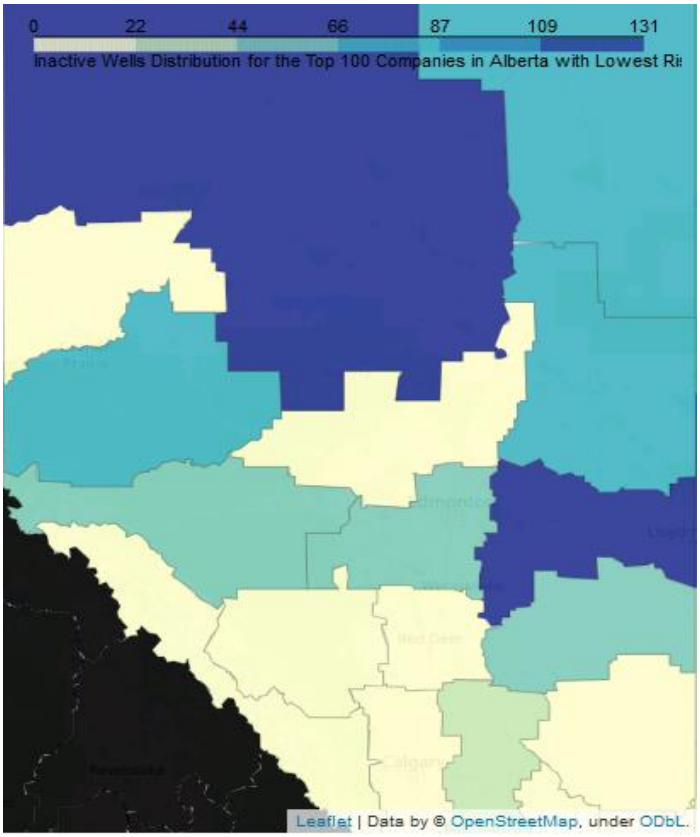
Table 3 Top five companies with lowest risks and more than 500 total number of wells. Closer the average to 0, lower is the risk.

Company	Total Number of Wells	Average AER Risk Class
Gear Energy Ltd.	738	0.144986
Bonavista Energy Corporation	811	0.171393
Karve Energy Inc.	605	0.181818
Torxen Energy Ltd.	2045	0.239120
Crescent Point Energy Corp.	578	0.249135

Table 4 Top five companies with highest risks and more than 500 total number of wells. Closer the average to 1, higher is the risk.

Company	Total Number of Wells	Average AER Risk Class
Domestic Water Well	528	1.000000
Trident Exploration (Alberta) Corp.	617	0.677472
Razor Energy Corp.	616	0.628247
Paramount Resources Ltd.	1176	0.620748
Bonterra Energy Corp.	528	0.590909

Figure 3 Distribution of inactive wells for top 100 companies in Alberta with lowest inactive well risk class



Even though Division 17 has a high number of inactive wells, it is misleading as the division has the largest area in Alberta. The choropleth map was created using the LSDs (need to be converted to coordinates) and shapefile of Canada divisions.

Model Performance

The results for accuracy of K values from 1 to 15 are show in Figure 4. The second highest accuracy belongs to K value of 3 (since K value of 1 is meaningless). The KNN with K value of 3 was used to predict the test dataset.

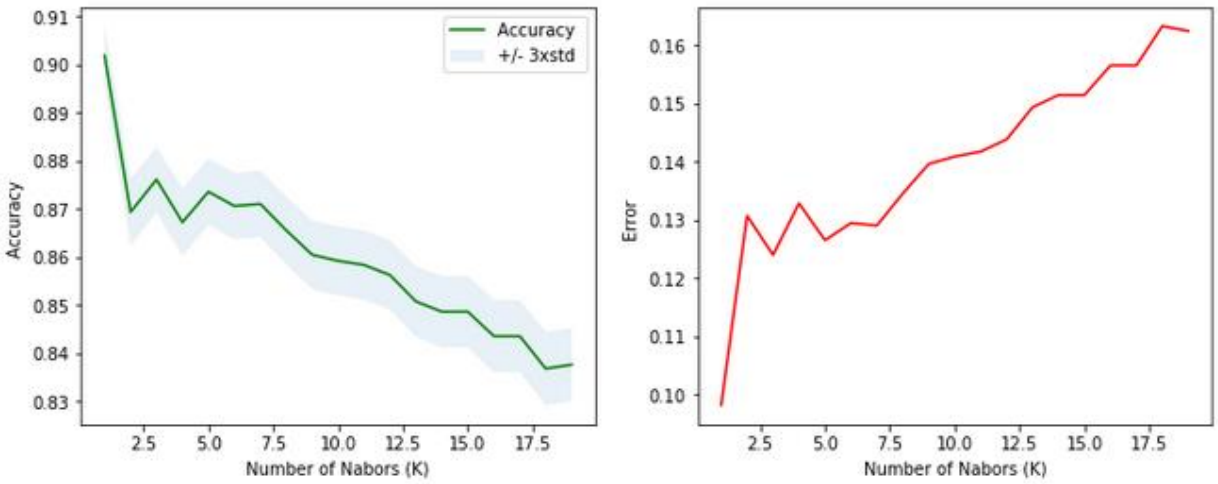


Figure 4 KNN model accuracy (left) and error (right) for K values of 1 to 15. K value of 3 shows the highest accuracy.

Four different kernel types (e.g. 'rbf', 'linear', 'poly', 'sigmoid') were used in SVM and the accuracy for each kernel was calculated.

Table 5 Accuracy of different kernels in SVM. Linear kernel showed the highest accuracy.

rbf	linear	poly	sigmoid
0.633427	0.635163	0.612426	0.224645

In Logistic Regression, five different solvers were used to calculate the accuracy. ‘saga’ showed the lowest Log Loss as shown in Table 6.

Table 6 Results for five different kernels for Logistic Regression. 'saga' showed the lowest Log Loss among five.

Solver	Similarity	Confusion	Log Loss
newton-cg	0.6430048242591316, 0.6380153738644304]	[[933, 243], [275, 913]]	0.555843
lbfgs	[0.6430048242591316, 0.6380153738644304]	[[933, 243], [275, 913]]	0.555843
liblinear	[0.63125, 0.6350515463917525]	[[909, 267], [264, 924]]	0.560879
sag	[0.6430048242591316, 0.6380153738644304]	[[933, 243], [275, 913]]	0.55584
saga	[0.6430048242591316, 0.6380153738644304]	[[933, 243], [275, 913]]	0.555842

Finally, Table 7 shows all the results from these six models. Random Forest model seems to be the best classifier out of these five models with 91% accuracy followed by KNN. In addition, Logarithmic Loss for Random Forest algorithm is the smallest among all six predictors. Figure 6 shows the features importance for Random Forest model. TVD has the highest influence on the model response out of all features.

Table 7 Performance metrices for six predictors. Random Forest shows the highest accuracy.

Algorithm	F1-score	LogLoss	Accuracy
KNN	0.875841		0.876058
Decision Tree	0.786316		0.786379
SVM	0.777909		0.777919
Logistic Regression	0.780855	0.555844	0.780880
AdaBoost	0.796003	0.688406	0.796108
Random Forest	0.911998	0.260926	0.912014

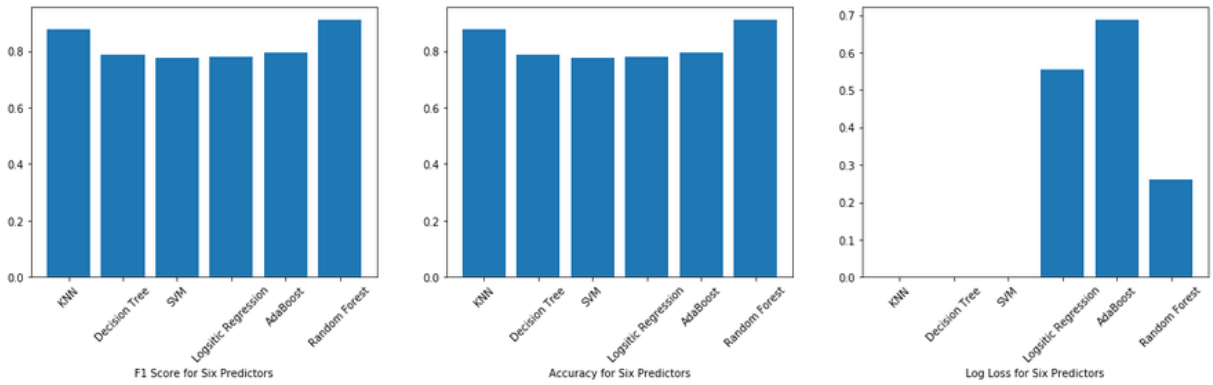


Figure 6 F1 Score (left), Accuracy (middle), and Log Loss (right) for six different Predictors. Random Forest showed the best results out of six models.

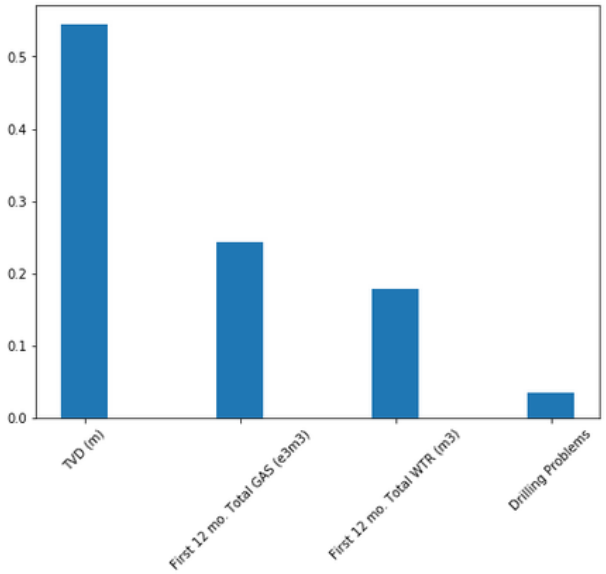


Figure 5 Features importance for Random Forest model. TVD shows the highest importance out of four features.

Conclusion

With available AER data, first 100 companies with lowest risk classes were identified. These wells and their locations may be informative for contractors and service companies in terms of their abandonment complications.

The predictive models used in this report performed better than random guessing. Out of these six predictors, Random Forest performed better than other models with the Accuracy and F1 Score of 0.91. For Random Forest model, TVD has the highest importance followed by gas production.

Because of limited access to geoSCOUT software, only production, TVD and drilling problem (i.e. pressure issues during drilling) were used in the models. However, it is highly recommended to incorporate drilling parameters to predict well performance and risk class before the production stage.

Glossary

Suspended/Inactive well: A well that has not been used for production, injection, or disposal for a specified amount of time (six months for high-risk wells, or 12 months for medium- and low-risk wells). A producer may choose to suspend a well because it is not considered to be economically viable at the time, but it could be in the future.

Abandoned/decommissioned well: Abandonment is the stage before reclamation, where a well that is no longer needed to support oil and gas development is permanently plugged, cut and capped according to Alberta Energy Regulator Directive 20.

Reclamation: The process of returning the abandoned site, as close as possible, to a state that's equivalent to before it was disturbed. Companies are responsible for reclamation liability for 25 years, after which the liability reverts to the Crown. Abandoned sites that have gone through reclamation must go through a certification process before being officially deemed reclaimed.

Orphan: A well, pipeline, or facility that does not have any legally responsible and/or financially able party to conduct abandonment and reclamation responsibilities.

Low risk well: Low risk well is defined as one of the following:

Type 1. Cased-hole wells that are not critical sour and have no perforations (noncompleted)
Type 2. Gas wells less than 28 000 m3/day that are low risk (see appendix 1)
Type 3. Water source wells
Type 4. Class 4 injectors (see *Directive 051*, section 2)
Type 5. Nonflowing1 oil wells with an H2S content less than or equal to 50 moles per kilomole (mol/kmol)

Medium risk well: Medium risk well is defined as one of the following:

Type 1. Gas wells that are medium risk (see the appendix 1)
Type 2. Nonflowing oil wells with an H2S content more than 50 mol/kmol
Type 3. Flowing oil wells

Type 4. Class 2 & 3 injection, carbon dioxide (CO2) injection/disposal wells (see *Directive 051*, section 2)
Type 5. Class 1B waste disposal wells (see *Directive 051*, section 2)
Type 6. Low-risk wells inactive for more than 10 years.

High risk well: High-risk well is defined as one of the following:
Type 1. Critical sour wells perforated or not.
Type 2. Acid gas wells.
Type 3. Class 1A waste disposal wells (see *Directive 051*, section 2).

Type 6: Low-risk wells inactive for more than 10 years

Reference

1. https://www.aer.ca/documents/directives/Directive013.pdf
   https://www.aer.ca/documents/directives/Directive020.pdf
   https://www.aer.ca/documents/directives/Directive051.pdf
2. Montague, James A., George F. Pinder, and Theresa L. Watson. "Predicting gas migration through existing oil and gas wells." *Environmental Geosciences* 25.4 (2018): 121-132.