



Statistical Assessment of Operational Risks for
Induced Seismicity from Hydraulic Fracturing in the
Montney, Northeast BC

Geoscience BC Report 2020-12
(Geoscience BC Project 2019-008)

Final Report

by Amy Fox, Scott McKean, Neil Watson

Enlighten Geoscience Ltd.

October 2020

Disclaimer and Document Authentication

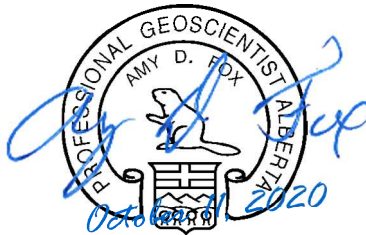
All work performed for this project and information provided in this report is of the highest technical quality possible given data limitations and uncertainties at the time the work was completed. Enlighten Geoscience Ltd. cannot guarantee the accuracy of the material included in this report (and any corresponding files or presentations) and bears no responsibility for the use of the material.

This report was written as a contribution to understanding of induced seismicity related to hydraulic fracturing. The authors of, and contributors to, this report make no guarantees to the predictability of, economic valuation of associated assets, risks to, or consequences of, induced seismicity.

The statistical methods employed in this report, including exploratory data analysis, feature selection, and machine learning use correlation-based methods that generally assume independence between observations. Correlation does not equate with causation, and the results presented in this report are one statistical view of numerous probable outcomes. No rigorous attempt has been made to control for a wide range of biases possible when doing statistical analysis including confounding and sampling bias.

Enlighten Geoscience Ltd.

APEGA Permit P13351



Executive Summary

This report describes the results of a study aimed at better understanding the relationship between hydraulic fracturing and induced seismicity in the Montney formation of northeast British Columbia. The study applies multivariate statistical analysis in an attempt to identify which hydraulic fracturing parameters show the strongest correlation to induced seismic events. The analysis also looked at a limited set of geological parameters. Using machine learning techniques, it addresses both induced seismicity likelihood and magnitude without relying on the supposition of a causal mechanism. One of the additional goals of the study was to provide an open-source, reproducible analysis for public and/or expert scrutiny, providing a framework for evaluating other formations in which induced seismicity is believed to be caused by hydraulic fracturing.

Data for the study came from several seismic event catalogues, published and proprietary maps and two collections of hydraulic fracturing data. Extensive data preparation and filtering were applied prior to modeling. All analyses were performed using R, an open-source programming language. Of the roughly 2,000 wells in the data after filtering, between 21% and 27% were classified as seismogenic, depending on the data set used. For the regression workflow, performed only on the seismogenic wells, just under 600 wells were included after filtering and correcting for missing features. Several exploratory data analysis techniques were used to determine the optimal features to include in both the classification and the regression workflow. During the feature selection process, geological features consistently ranked higher than completions features, prompting a re-run of the feature ranking process with completions features only.

Several different models, covering a range of model complexity, were run for both the seismogenic classification and the magnitude regression analyses. Feature importance and interaction were analyzed for each model, and partial dependence and individual conditional expectation were examined. In addition, interpretability techniques were applied to one specific well, 100/09-35-081-18W6/00, which was associated with a magnitude 2.7 induced seismic event, in order to examine how the models related input features to the seismic outcome and magnitude prediction.

This study concludes that:

1. Both the classification of wells as seismogenic and the prediction of induced event magnitude are highly dependent on several factors, including the data set used, the specific model used, and the subset of well, completions and geological features selected for inclusion in the final models. As a result, the analysis does not single out one or more clearly causal features that are responsible for induced seismicity from hydraulic fracturing in the Montney in NEBC. Notably, deciding which features to carry forward into the final modeling could not be achieved by the machine learning workflow alone but required a considerable amount of human intervention.
2. In the seismogenic classification, feature interaction tends to mimic feature importance, suggesting that feature interactions contribute to a higher model variance, especially for the more complex models. The simplest model places a relatively high negative importance on minimum horizontal stress and high positive importance on geothermal gradient, distance between wells and mean proppant per stage, while the more complex models place a relatively

high importance on Paleozoic structure and distance to faults. In the magnitude regression, most models show a relatively high importance for top of Montney structure and distance to faults (normal and thrust). Interactions tend to be higher for completions parameters than for geological parameters. It is difficult to know, however, whether some of these parameters may be serving as a proxy for other, more difficult to measure, features, either of the completion or of the reservoir.

The following appendices are available separately:

Appendix A: Feature Definitions

Appendix B: Methodology Details

Appendix C: Full-Size Figures

Appendix D: Database

Appendix E: Github Repository

Contents

Disclaimer and Document Authentication.....	ii
Executive Summary.....	iii
I. Introduction	7
Methods.....	7
Anticipated Outcomes	7
Study Area.....	8
Geological Background	9
II. Data	9
Earthquake Catalogues	9
Well and Hydraulic Fracture Completion Data	12
Geological Data	12
Note on Fault Terminology	12
Data Preparation.....	13
III. Analysis	14
Analysis Overview	14
Seismogenic Classification	16
Magnitude Regression	16
Model Generation.....	16
Model Summary.....	17
Data QA/QC and Outlier Removal.....	17
Exploratory Data Analysis	18
Feature Ranking and Selection	24
Model Tuning & Training	32
IV. Model Interpretation	37
Feature Importance and Interaction	38
Partial Dependence and Conditional Expectation	42
Local Model Interpretability	48
V. Summary, Discussion and Recommendations.....	52
VI. Acknowledgements	53

VII. Appendices	53
Appendix A: Feature Definitions	53
Appendix B: Methodology Details	53
Appendix C: Full-Size Figures	54
Appendix D: Database.....	54
Appendix E: Github Repository	54
VIII. References	54

I. Introduction

Induced seismicity is a concern in northeast British Columbia (NEBC) and many other parts of the world. Numerous studies (e.g., BC OGC, 2012; Kao et al., 2018), have linked anomalous induced seismicity in NEBC to the practice of hydraulic fracturing, which involves the large-scale stimulation of oil and gas reservoirs by injecting large volumes of fluid to break the rock in the subsurface. In February, 2019 the British Columbia (BC) Minister of Energy, Mines and Petroleum Resources issued its *Report of the Scientific Hydraulic Fracturing Review Panel*, which included input from a large number of stakeholders and experts regarding the ability of BC's regulatory framework to manage the risks of hydraulic fracturing including, but not limited to, induced seismicity. The report emphasized the important role that peer-reviewed science could have in informing best-practice regulations and increasing the public's confidence in the regulator.

Following the publication of the panel's report, both the British Columbia Oil and Gas Commission (BC OGC) and Geoscience BC initiated research projects to increase understanding of hydraulic fracturing induced seismicity specifically in the Montney formation in NEBC. After a preliminary geomechanical investigation of the Kiskatinaw Seismic Monitoring and Mitigation Area, Fox and Watson (2019) recommended that statistical methods be applied to look for correlations between hydraulic fracture operational parameters, such as injected volumes and rates, and induced seismic events in NEBC. This followed similar recommendations by others including the BC OGC (2012).

Methods

This study applies multivariate statistical analysis in an attempt to identify which hydraulic fracturing parameters show the strongest correlation to induced seismic events in the Montney in NEBC. Multivariate statistical approaches have been used successfully to link induced seismicity to geological risk factors in the Duvernay formation in Alberta (Pawley et al., 2018; Schultz et al., 2018) and to link well completion factors to both higher hydrocarbon production and lower completion costs in the Montney of the Swan-Elmworth area in Alberta and BC (Lenko and Foster, 2016). This study employs a similar workflow and also applies some additional data science techniques aimed at feature selection and model interpretation.

This study centers around the use of machine learning to model induced seismicity likelihood (using a classification model) and magnitude (using a regression model). One of the benefits of a machine learning approach is it does not rely on the supposition of a causal mechanism for the phenomenon of interest. Machine learning is a broad and complex topic, the definition of which is outside of the scope of this report. Many resources are available to derive a more detailed understanding of machine learning and statistical methods in general (e.g., Molnar, 2019 and James et al., 2017).

Anticipated Outcomes

Upon initiation of the study, it was anticipated that the results would include:

- A report summarizing the study findings and recommendations for formation-specific mitigation measures to reduce the influence of high-importance features on induced seismicity

- A database of completion parameters and limited database of geological risk factors for Montney wells in the study area provided in a peer-reviewed format for public review and use in the study
- A Montney formation-specific predictive model for induced seismicity likelihood and magnitude
- An open-source, reproducible analysis for public and/or expert scrutiny, providing a framework for evaluating other formations in which induced seismicity is believed to be caused by hydraulic fracturing

Study Area

The NEBC Montney formation has been divided by the BC OGC into two primary regions for the understanding of the causes of and the mitigation of induced seismicity. These regions, illustrated in Figure 1, are the Kiskatinaw Seismic Monitoring and Mitigation Area (KSMMA) and the North Peace Ground Motion Monitoring Area (NPGMMA). Data from both areas were included in this analysis.

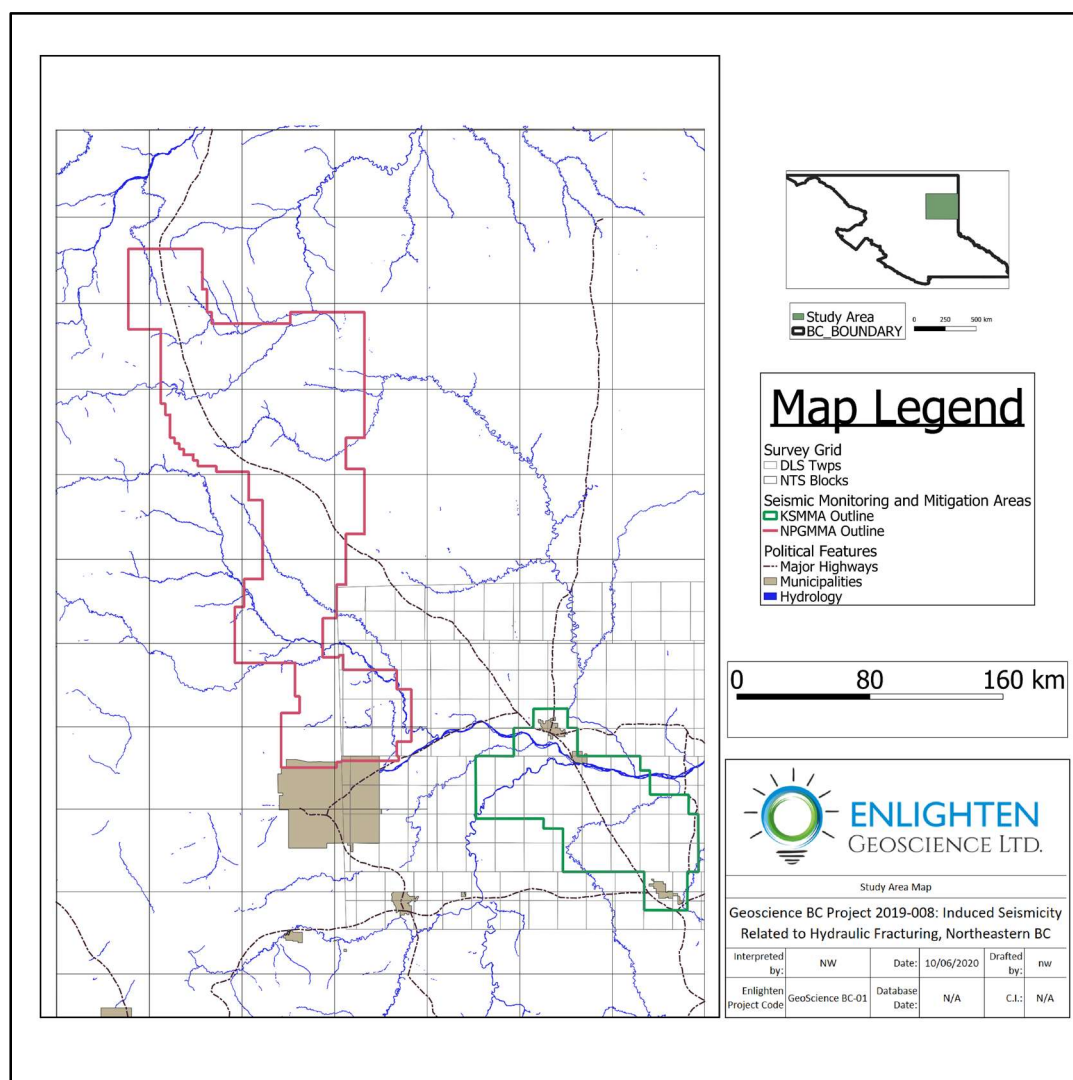


Figure 1. Study area map illustrating extents of the KSMMA and the NPGMMA.

Geological Background

The Montney formation of NEBC and west-central Alberta represents a Lower Triassic (Griesbachian to Spathian) unconformity bounded wedge of mixed clastic and carbonate sediments (Euzen et al., 2018). This formation records a progradation from its eastern subcrop to a thickness of over 300 metres and a westwards transition from shoreline sediments to a system dominated by distal shelf siltstones punctuated by mass wasting deposits (i.e. turbidites) (Davies et al., 2018).

The extremely fine-grained nature of the Montney siltstones and compaction due to the maximum burial depth of the Montney has resulted in extremely low permeability over much of NEBC. The transition to these lower permeability sediments corresponds to the onset of an over-pressured (i.e. >10 kPa/m) deep basin (pervasive hydrocarbon saturations dominated by natural gas and liquids). This massive hydrocarbon trap covers an area of over 26,000 km² (Hayes, 2012). The ability to access this resource through fractured horizontal wells in a region with significant production infrastructure has made the NEBC Montney a highly attractive target for drilling and completion, specifically hydraulic fracture stimulation, activity.

The structural setting of the NEBC Montney is particularly relevant to the subject of induced seismicity. The KSMMA overlies the Dawson Creek Graben Complex (Barclay et al., 1990), a feature that resulted from the collapse of the Peace River Arch at the end of the Devonian through the early Mississippian by way of a series of normal faults. These structures were reactivated by compression during the Columbia and Laramide Orogenies at the end of the Jurassic and Cretaceous respectively to create a transpressional structural setting.

The NPGMMA can be characterized as having a somewhat less complex structural history. The dominant structural style is predominated by right-lateral strike-slip movement along the Hay River Fault Zone. Dip-slip faults, including some inter-formational thrust faults contribute to the secondary structural component.

II. Data

The study required the collection of completions and geological features from a variety of seismic event catalogues and well, hydraulic fracture stimulation and geological data sources. Each is discussed separately in this section. Definitions of all of the features used in the modeling are provided in Appendix A.

Earthquake Catalogues

Three earthquake catalogues were combined in this study: the Composite Alberta Seismicity Catalog (Fereidoni and Cui, 2015), a comprehensive Geological Survey of Canada (GSC) catalogue (Visser et al., 2017), and a catalogue from the BC OGC and GSC (Babaie Mahani et al., 2020). Within the Montney extents, the Composite Alberta catalogue contains 1,196 events dating from May 16, 1969 to April 16, 2019. Similarly, the GSC catalogue contains 3,011 events from January 1, 2014 to December 31, 2016. The BC OGC/GSC catalogue contains 10,694 events from January 1, 2017 to December 31, 2018. The catalogues were individually processed and combined. Each catalogue presented the moment

magnitude (M_w) when available, with some older events being recorded in local magnitudes (M_L). Seismic events before January 1, 2000 were also removed.

In order to remove duplicate seismic events after combining the catalogues, a one-way search was conducted on the combined catalogue. Events occurring within 10 seconds and within 10 km of each other were flagged as a duplicate, resulting in the removal of 2,298 seismic events. A magnitude of completeness analysis was performed using the maximum magnitude method (Wiemer and Wyss, 2000) with a Rice Rule binning (Scott, 1979). This reported a maximum magnitude of completeness of 0.25 for the KSMMA region, providing a balance between other common binning methods; for example, the Sturges (1926) binning method yields a magnitude of completeness of 1 while the Freedman and Diaconis (1981) binning method yields a magnitude of completeness of 0.05. Seismic events below the KSMMA magnitude of completeness (0.25) were removed to retain granularity in that region of the study. This is non-conservative for the NPGMMA, which is reported with a magnitude of completeness around 1.4. This yielded a final catalogue with 9,843 events. A b-value analysis using the maximum likelihood (Aki, 1965; Bender, 1983), maximum curvature, and goodness of fit methods (Wiemer and Wyss, 2000) shows that the earthquake catalogues from the two regions display slightly different b-values (Figure 2) near 1, however this may be a result of censoring. All plot functions for the b-value calculations can be found in the project data repository (Appendix E).

The catalog provides two targets for our analysis. The first is the seismogenic target, which indicates whether a well is associated with seismicity or not. The second is the maximum magnitude of all the seismic events associated with each well. For the regression analysis, wells without associated seismicity are excluded and the maximum magnitude is encoded as zero.

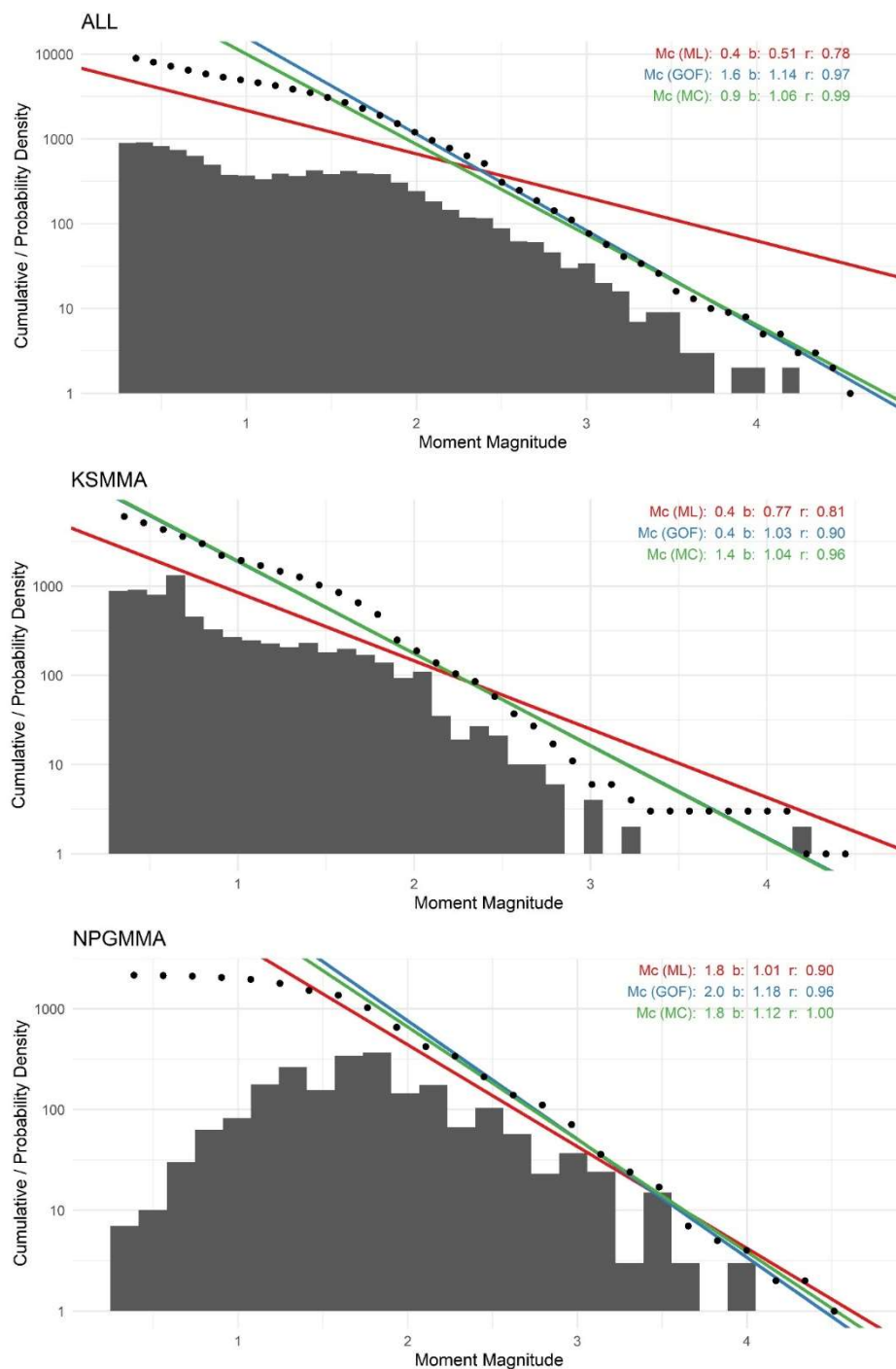


Figure 2. Magnitude distribution plot of the seismic event data from the entire catalogue (top), KSMMA (middle), and NPGMMA (bottom). The histogram of events is shown in grey, and the cumulative probability mass function is plotted by black dots. The b -value (b) and magnitude of completeness (Mc) estimated by the maximum likelihood method (red), the goodness-of-fit method (blue) and the maximum curvature method (green) are shown along with the coefficient of determination (r).

Well and Hydraulic Fracture Completion Data

The workflow was run on two sets of data – one published by the BC OGC and one sourced from geoLOGIC Systems Ltd. (geoLOGIC) through an agreement between geoLOGIC and Geoscience BC. This data included basic well information, deviation surveys and production data. It also includes detailed completion parameters such as stage-by-stage injected fluid and proppant, stage durations, the start and end point of each stage, and instantaneous shut-in pressure (ISIP) where recorded by the operator. The same features were extracted from both data sets for consistency, but the actual data varied between the two; for example, the wells in each set, the data completeness, and absolute values of some of the features were not the same. In general, the geoLOGIC data is slightly more granular and has more complete observations. We present results from both data sets, although the analysis can only be reproduced for the BC OGC data set due to licensing restrictions that apply to the geoLOGIC data.

Geological Data

We included only a limited number of geological factors in the analysis due to the high level of uncertainty in the current understanding of critical geological factors such as the stress state in the subsurface. Some geological parameters were included as proxies. For example, isotherm is easily calculated and can provide a proxy for depth, vertical stress and connection to basement. An example of a similar use of geological analogues is given by Pawley et al. (2018), who employed lithium concentrations in their study of the geological susceptibility of the Duvernay to induced seismicity.

The maps utilized in the study included:

- First order residual on the Paleozoic structure
- Third order residual on the Paleozoic structure
- True vertical depth (TVD) to the top of the Montney
- Top of Montney structure
- Geothermal gradient
- Montney isotherm
- Minimum horizontal stress (S_{hmin} , from Grasby et al., 2012)
- Vertical stress (S_v , from Grasby et al., 2012)
- Significant faults (see note below)
- Ratio of pore pressure vs. TVD

These maps were either published in the literature, other public domain technical reports referenced above or proprietary Enlighten mapping generated/provided for the use in this report.

Note on Fault Terminology

In this study, specific faults within the KSMMA and the NPGMMA are divided into types as classified by Berger et al. (2008) and Davies et al. (2018). A more general discussion of fault definitions may be found in Ragan (1973) and USGS (2019). Simplified definitions of the faults used in this analysis are:

- Normal: steeply inclined faults generally related to tectonic extension in which the hanging wall (rock above the fault) has moved in a downward sense relative to the foot wall (rock beneath the fault)
- Listric: faults that show relative displacement as in a normal fault but in which the dip of the fault (fault inclination) decreases with depth
- Thrust: a type of reverse fault (inclined faults created during compression in which the foot wall has moved upwards relative to the hanging wall) in which the dip of the fault decreases with depth
- Strike-slip: individual, sub-vertical faults with primarily horizontal movement between fault blocks resulting from shear deformation
- Divergent: strike-slip faults that occur in a set and show divergence as a result of transpression

Data Preparation

Well completion data, deviation surveys, geological data, the combined earthquake catalogue and production data were loaded and processed in R, an open source programming language, to standardize Unique Well Identifiers (UWIs) and geospatial projections. The stage-by-stage completion data for each vertical, deviated and horizontal well was aggregated, since the analysis is conducted on a single well level. Wells were analyzed individually, meaning that wells on the same pad were not aggregated but were treated as separate analysis units. This also meant calculating summary statistics for parameters such as injected fluid and proppant and injection rates. On-production dates and production metrics-to-date were summarized for each well. Geospatial layers and deviation surveys were projected to NAD83 UTM Zone 10N prior to distance calculations. A 300 m buffer was drawn around each well and mean geological parameters extracted. The wells were then seismogenically associated with the combined earthquake catalogue and the data combined into a single data set. Information on the data sets after data processing and filtering out null values is provided in Table 1. Additional details on the data processing can be found in the notebooks included with the source code for this project (Appendix E), which can also be used to reproduce the preparation for the BC OGC data set.

	Total Observations	Original Features	Final Non-Null Observations	Final Features	Seismogenic Percentage
Classification (Induced Seismicity Likelihood)					
BC OGC	2,904	67	1,640	12	27.0%
geoLOGIC	4,376	79	1,841	12	20.7%
Regression (Magnitude)					
BC OGC	2,904	67	583	12	100%
geoLOGIC	4,376	79	518	12	100%

Table 1. Summary of data set information after processing and filtering out null values.

The seismic catalogue events were associated to hydraulically fractured wells using the Pawley et al. (2018) spatiotemporal filter. This filter associated seismic events with completions if the events exceeded a magnitude of 1.1, were within a cluster with at least two other seismic events, occurred within 1 day before and 30 days after the end of the completion, and occurred within 5 km of the well or

the maximum expected uncertainty for the majority of seismic events. In cases where multiple seismic events are associated with the same well, the maximum magnitude of the associated events was taken for the regression analysis. A HDBSCAN algorithm (Ester et al., 1996; Hahsler et al., 2017) was used for the clustering. The algorithm has three main sensitivities: the cluster size (minpts), the distance from the well to events, and the magnitude of completeness used. The effect of these parameters on the number of seismogenic events and mean magnitude is illustrated in Figure 3. Overall, the parameters chosen for a clustering model provide a balance between overclassification of seismogenic events and reducing the magnitude of seismic events available for the statistical analysis.

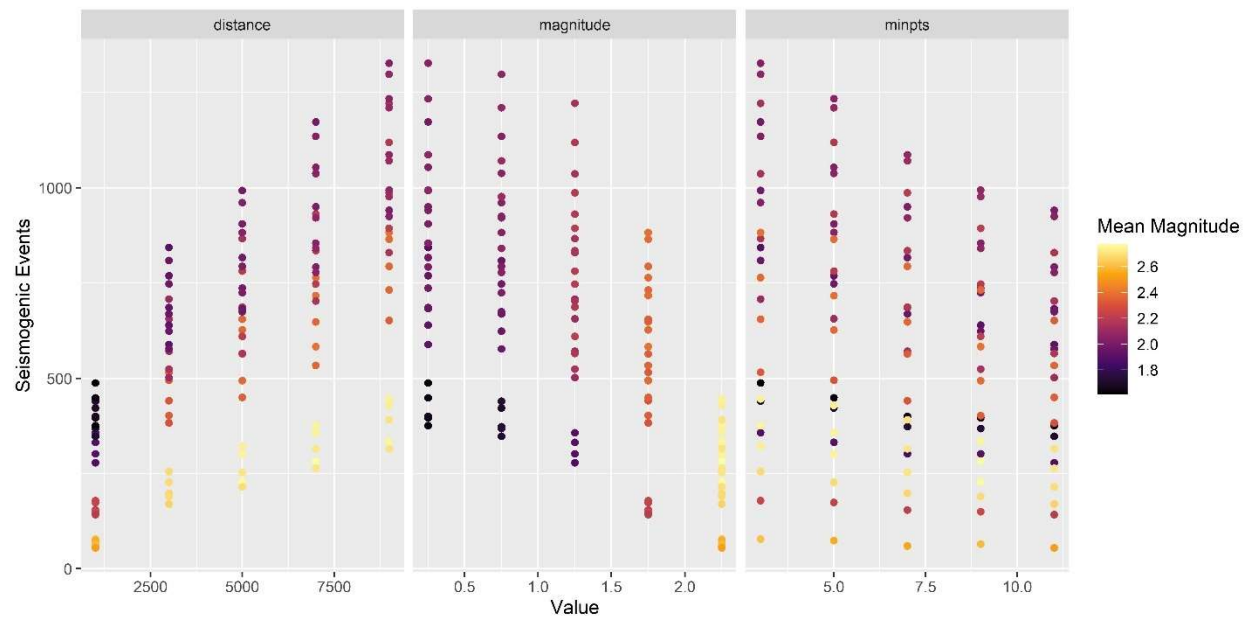


Figure 3. Sensitivity study results for the HDBSCAN algorithm. The effect of distance (in m), magnitude, and the minimum cluster size (minpts) is illustrated against the number of seismogenic events identified (y-axis) and the mean magnitude of those seismogenic events (colour scale). The mean magnitude increases as the number of seismogenic events decrease because more events of lower magnitude are seismogenically associated with a particular well.

III. Analysis

Analysis Overview

We used a statistical approach for the analysis, specifically a supervised machine learning workflow with feature selection and interpretation techniques. This section provides a discussion of each component of the analysis workflow (Figure 4), but additional technical details may be found in Appendix B and the notebooks included with the source code for this project (Appendix E).

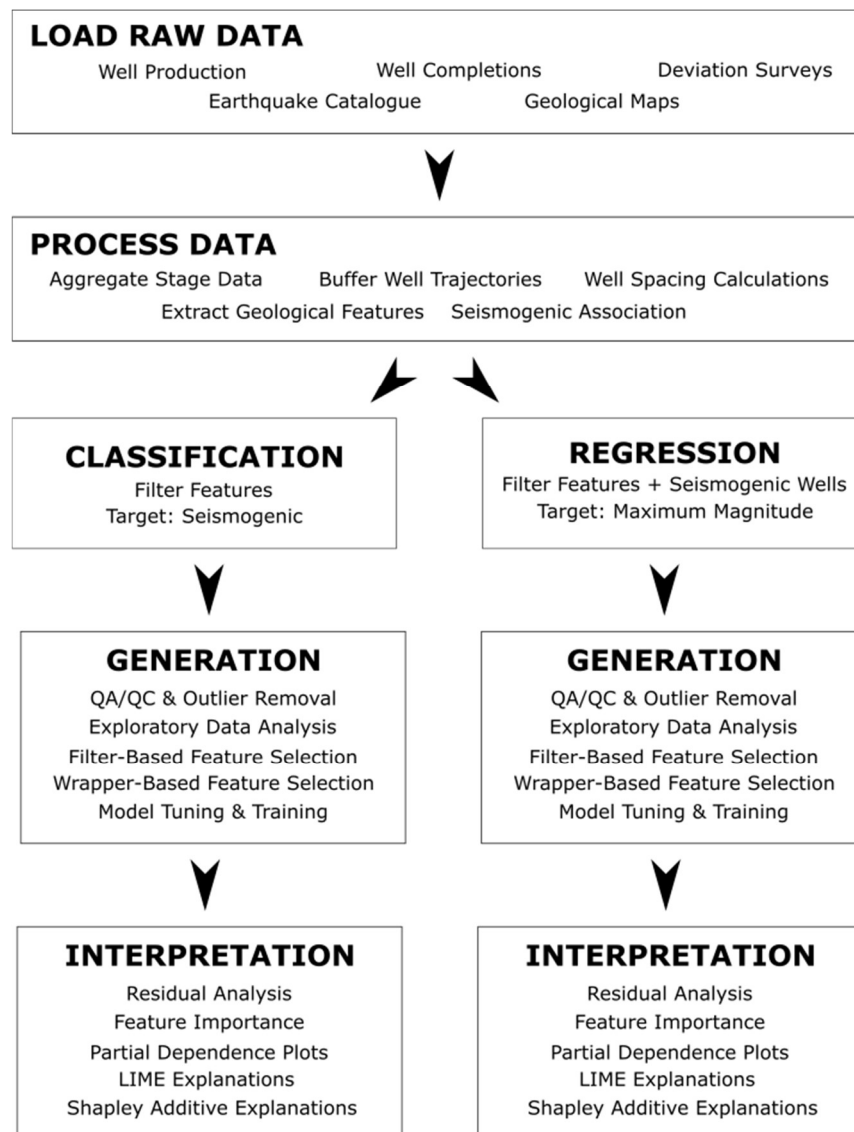


Figure 4. An illustration of the statistical workflow, including the main steps of data loading and processing, followed by the two models – seismogenic classification and magnitude regression.

Supervised machine learning creates statistical models that are trained on observed data without assumptions on the data generating process (Breiman, 2001). It finds a relationship between a *target*, or a variable of interest (e.g., the magnitude of an induced seismic event), and model features, or variables in our data (e.g., injection volume) that explain the variability of the target. The resulting model, which is statistical (or empirical) in nature, can be used to explain what is driving the variability of the target, or predict the target outcome based on new features.

In this study, we fit models for two targets – the presence or absence of an associated induced seismicity event, and the observed magnitude of an associated induced seismicity event. These two models can be treated as nearly independent analyses: a classification model that predicts the presence

or absence of induced seismicity, and a regression model that predicts the magnitude of associated induced seismicity events.

The initial data was the same for both workflows (Table 1). Unfortunately, missing values for important features reduced the number of wells for analysis.

Seismogenic Classification

The classification model provides a likelihood of a well being seismogenic. The log loss is used to train the classification model and tune hyperparameters. This metric measures the classification accuracy (predicted vs. actual class), quantifying the balance between false positives and false negatives across a decision threshold. The log loss score is a good metric for the evaluation of problems with uneven class distributions, such as this study where only 25% of the observations present a positive class.

Magnitude Regression

The regression model predicts the magnitude of the maximum earthquake for a seismogenic well. The average of Root Mean Square Error (RMSE) and Maximum Absolute Error (MAE) was used as a metric for evaluating the model and hyperparameter tuning. The regression workflow can only be conducted on wells associated with induced seismicity, a criterion that reduced the regression data set below 600 wells.

Model Generation

The MLR (Machine Learning in R, Bischl et al., 2016) framework in R is used for model generation, along with several external packages detailed in the R markdown notebooks (Appendix E). The model generation approach is shown in Figure 5. It begins with data quality assurance and quality control (QA/QC) and outlier removal. Some missing secondary feature values are also imputed to increase the number of observations available for model training. Exploratory data analysis and feature selection are then used to focus on important features. This is especially important in the current study because of the large number of potential features. Filter-based and wrapper-based feature selection techniques are then used, followed by model tuning and training. These steps are described below, with a more detailed explanation provided in Appendix B.

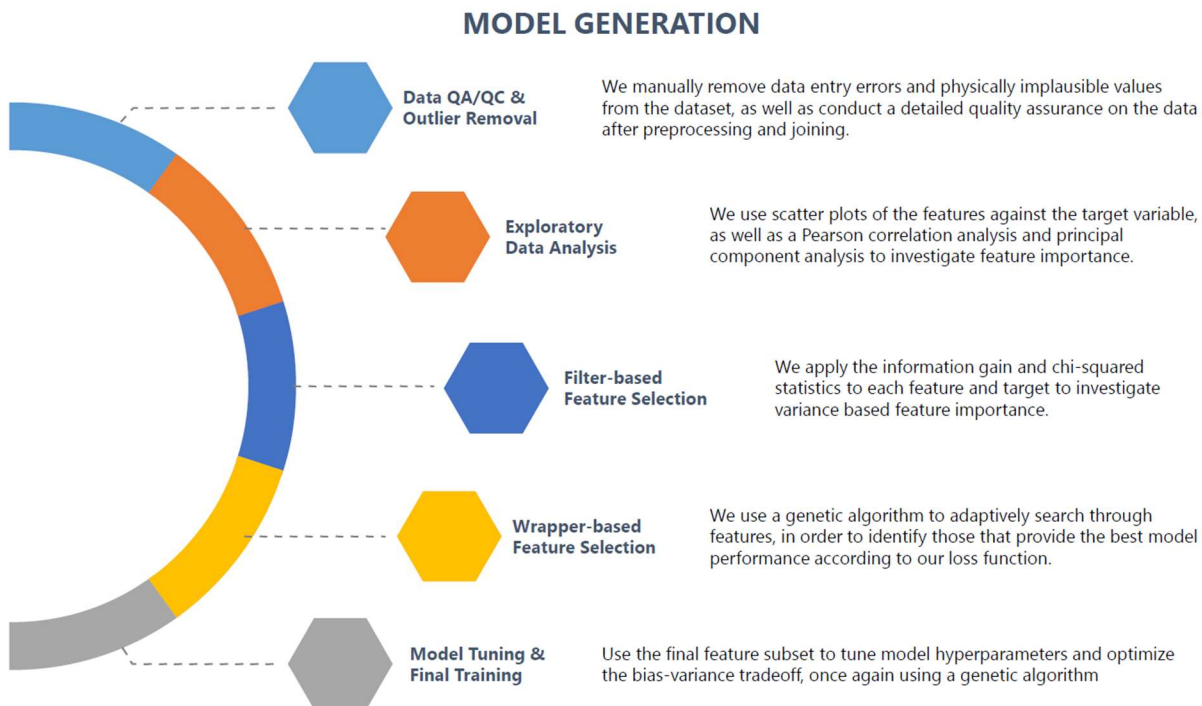


Figure 5. Steps in the generation of the classification and regression models.

Model Summary

Four different models were used in the model generation process. Because of the small size of the data sets, the models vary in complexity and are compared in order to evaluate the bias-variance trade-off in the machine learning workflow. Additional details are provided in the model training section below and Appendix B. A summary of the models including a brief description of each is provided in Table 2.

Model	Abbreviation	Description
Generalized Linear Model	GLM	Relatively simple model that can encode interdependence of features. Classification and regression.
Classification Decision Tree	CART	Simplest tree-based modelling approach, using a single pruned decision tree. Classification only.
Multivariate Adaptive Regressive Splines	MARS	Uses multiple linear slopes for each feature to increase complexity relative to a GLM. Regression only.
Extreme Gradient Boosting	XGBoost	High complexity ensemble model utilizing a forest of decision trees. Classification and regression.

Table 2. Model summary for the seismic classification and magnitude regression workflows.

Data QA/QC and Outlier Removal

An observation is defined as a set of features and target for a single well. Removing observations with any missing features (that is with any null value across any of the features) resulted in the removal of over half of the observations in each data set. There are also features with outlying values that represent data entry errors and non-physical measurements. These outliers were first removed, and

observations missing critical features (proppant and fluid volumes for example) were also filtered out. Features considered non-critical were imputed with the mean of the data in order to assist with feature ranking and selection. These features included `sd_treating_mpa`, `mean_treating_mpa`, `sd_treating_mpa`, `mean_intervals_per_stage`, and `sd_breakdown_mpa` for the geoLOGIC data and `mean_rate_m3_min` for the BC OGC data. The maximum number of missing values was 48 out of 1640 for `mean_rate_m3`. The imputation limited the number of wells that are removed from the data set prior to exploratory data analysis and feature selection and does not affect feature importance.

Exploratory Data Analysis

Box plots, scatter plots, correlation plots and principal component analysis were used to visualize the relationship between features and the target for the seismogenic classification and magnitude regression workflows.

Box plots were used to visualize the relationship between features and the seismogenic classification. Features with different distributions between seismogenic and non-seismogenic wells are more likely to be important. An example box plot is shown in Figure 6, where the number of completion stages is shown for seismogenic wells in red and for non-seismogenic wells in black. All of the box plots are provided in Appendix C. Based on the box plots, potentially important features include horizontal wells within 1 km, number of stages, average stage spacing, the mean number of intervals per stage, the total fluid injected in each well, the mean injection rate, the fluid intensity, the Paleozoic structure, and the top of Montney structure.

Scatter plots are used to visualize the relationship between features and the maximum earthquake magnitude associated with a seismogenic well. An example scatter plot for maximum stage duration is shown in Figure 7. All of the scatter plots are provided in Appendix C. The scatter plot shows why a relatively complex model is required for predicting the magnitude of induced seismicity, since none of the features shows a clear relationship with the target variable. Based on the scatter plots, features that may be important in regression include the number of stages, average stage spacing, mean fluid per stage, total fluid injected in each well, fluid intensity, mean ISIP, maximum ISIP, and geothermal gradient.

Of the potentially important features identified with the scatter plots, only the number of stages, mean fluid per stage and fluid intensity overlap with the features identified with the box plots. It is important to note that the features are not actually expected to overlap, since the target and data in each workflow differ. The causal variables for whether a well is seismogenic or not may not relate directly to whether the induced earthquake is of a significant magnitude.

A Pearson correlation plot shows the similarity between numerical features. Groups of parameters with high correlation are generally assumed to represent the same physical process, at least statistically. This suggests that a single feature can statistically represent an entire group of correlated features. Highly correlated features can also confound statistical analyses and pose a risk to model interpretation (Pearl,

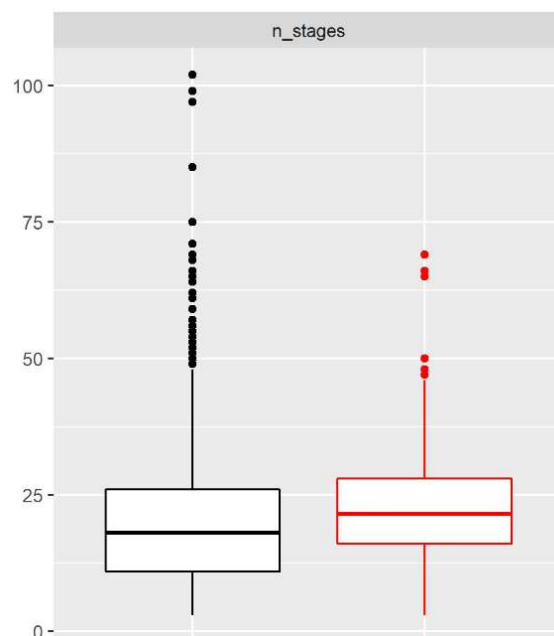


Figure 6. Box plot showing the distribution of the number of completion stages in seismogenic wells (red) and non-seismogenic wells (black). The box represents the interquartile range (IQR) of the data, with the horizontal line representing the mean. The whiskers represent 1.5xIQR above and below the box, with points representing outliers.

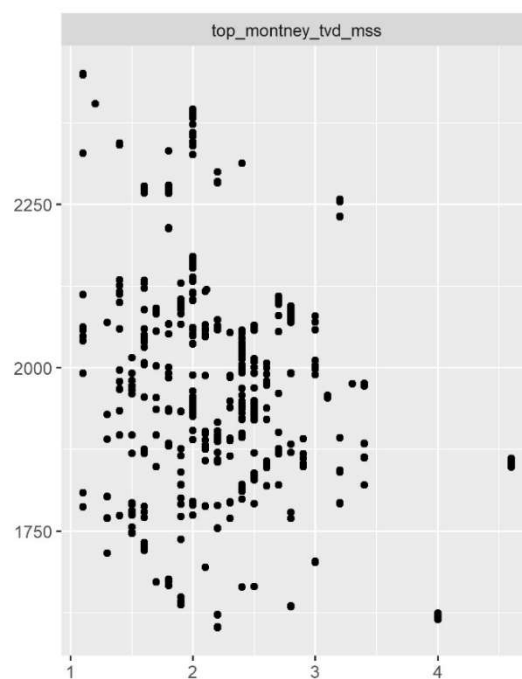


Figure 7. Scatter plot of TVD to the top of the Montney in meters versus maximum earthquake magnitude.

2009). The correlation plot for the seismogenic classification features is shown in Figure 8. The correlation plot for the magnitude regression features is shown in Figure 9. Full-size versions of both figures are provided in Appendix C.

The strength of the correlation is denoted both by the shape of the ellipse and colour, with thin ellipses and dark colours representing high correlations. Blue is a positive correlation, and red is a negative correlation. In general, most of the parameters show a low correlation and the two figures appear very similar. The correlation plots show that there are numerous groups of input features with high correlation, suggesting that a single parameter from each group can be selected for the final model. The groups include:

- fluid intensity, proppant intensity, total fluid, and total proppant
- horizontal wells within 1, 5, 10, and 25 km
- mean fluid per stage, mean proppant per stage, mean stage duration, mean well TVD and mean breakdown pressure (moderate correlation)
- number of stages, average stage length, and average stage spacing

Principal component analysis (PCA) is a dimensionality reduction technique that can also identify important features in the data set (Jolliffe, 2002; Peres-Neto et al., 2005). It assumes that many of the features are correlated with each other and that a linear combination of features can explain the data's variance. By recombining the features into principal components and investigating the parameters in the first two principal components, one can see a) which parameters are important; b) which parameters group together; and c) how many parameters might be needed in a model. A unit circle plot (Figure 10) is used to investigate a) and b) by plotting the influence of each original feature in the primary and secondary principal components. A scree plot (Figure 11) is used to explain the variance that can be explained by each principal component, helping to answer how many components might be needed in a reasonable model (c). The data is standardized using a z-score transformation prior to PCA analysis. Both plots are available as full-size versions in Appendix C.

The unit circle in Figure 10 shows again that many features are correlated with each other, and that the data is distributed relatively evenly across features in the first two principle components. This means that numerous features will be needed to predict either seismogenic association or magnitude and that a complex model will likely be required. The scree plot in Figure 10 shows that a large amount of the total variance of the data set can be represented by around 7 or 8 principal components (the elbow in the scree curve), providing qualitative guidance on the number of features to consider in a model.

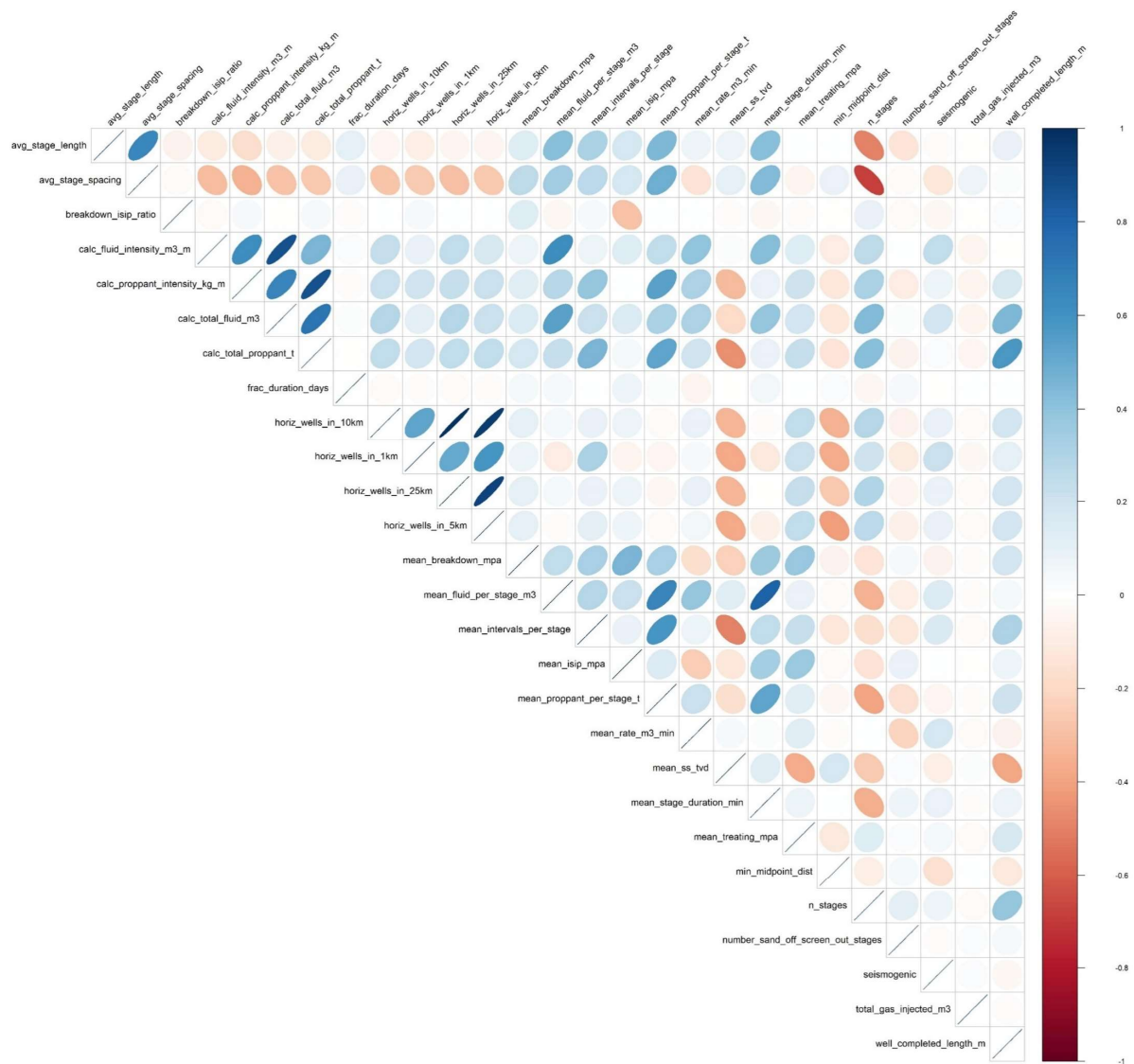


Figure 8. Correlation plot of all model features and the seismogenic classification. Thin ellipses and dark colours represent higher correlations. Blue is a positive correlation, and red is a negative correlation.

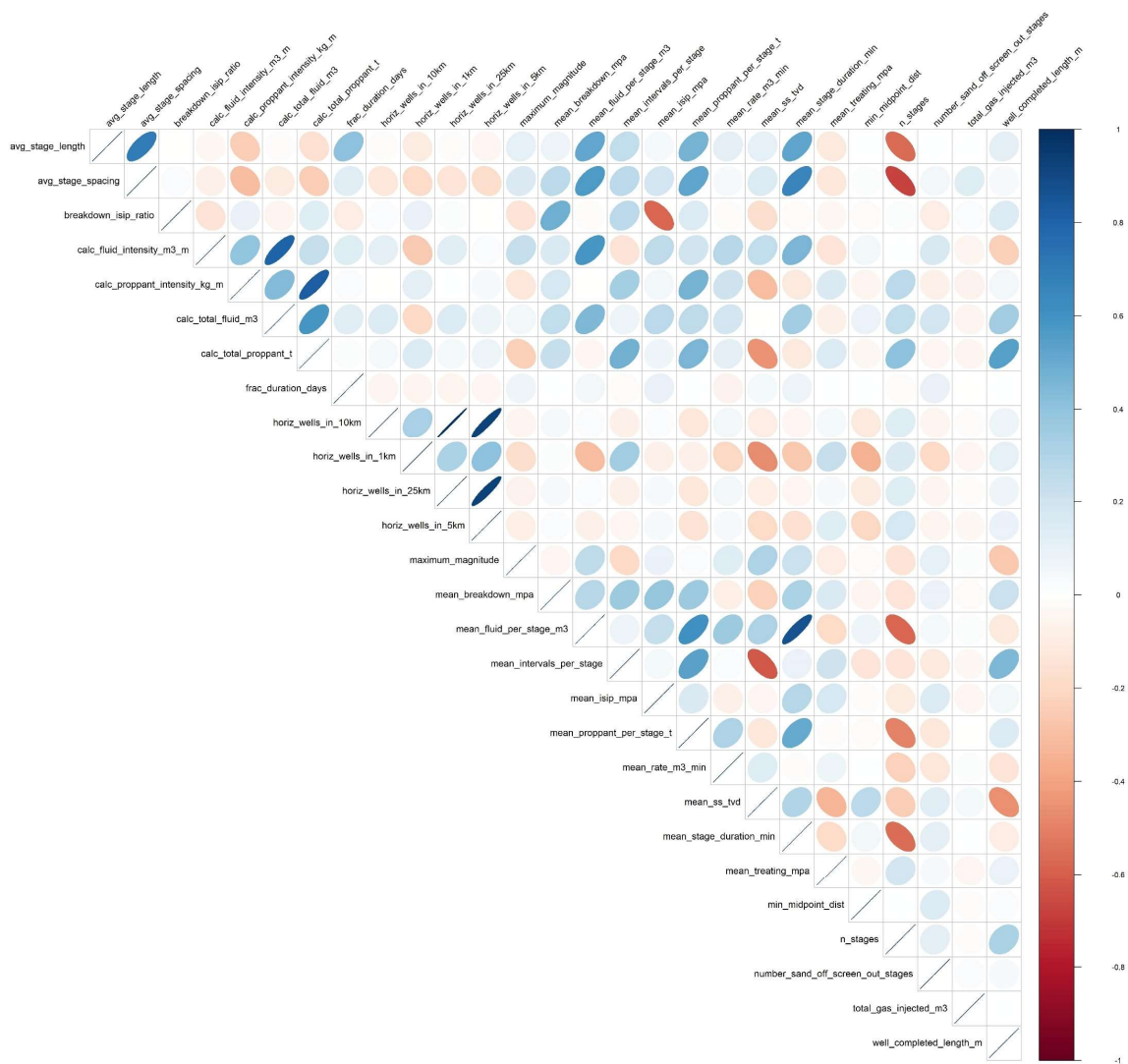


Figure 9. Correlation plot of all model features and earthquake magnitude. Thin ellipses and dark colours represent higher correlations. Blue is a positive correlation, and red is a negative correlation.

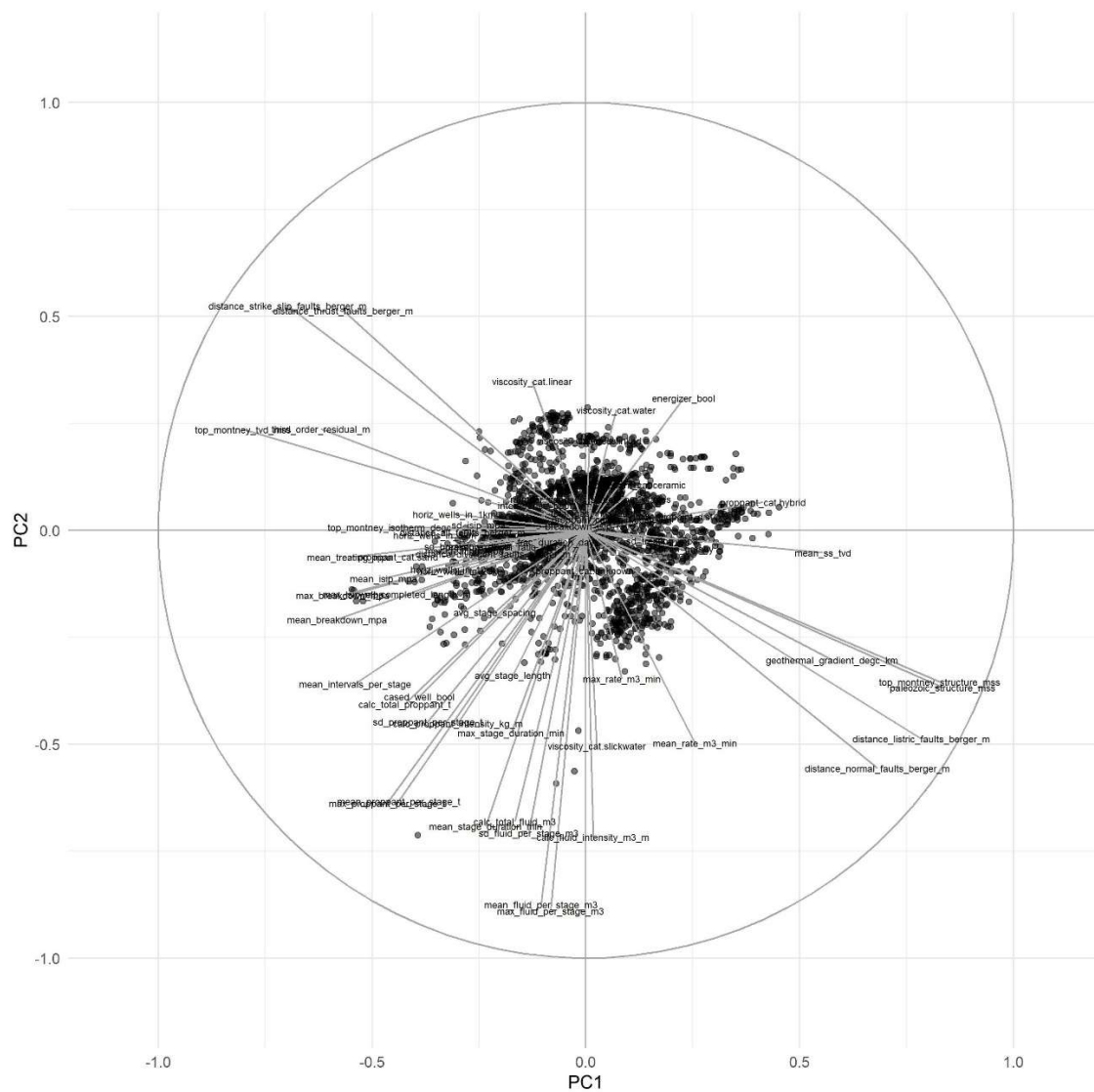


Figure 10. A unit circle plot showing the proportion and weight of each feature in the first two principal components. Parameters closer to the edge of the unit circle explain more variance. Parameters that are grouped closer together are correlated with each other in terms of explaining the first two principal components.

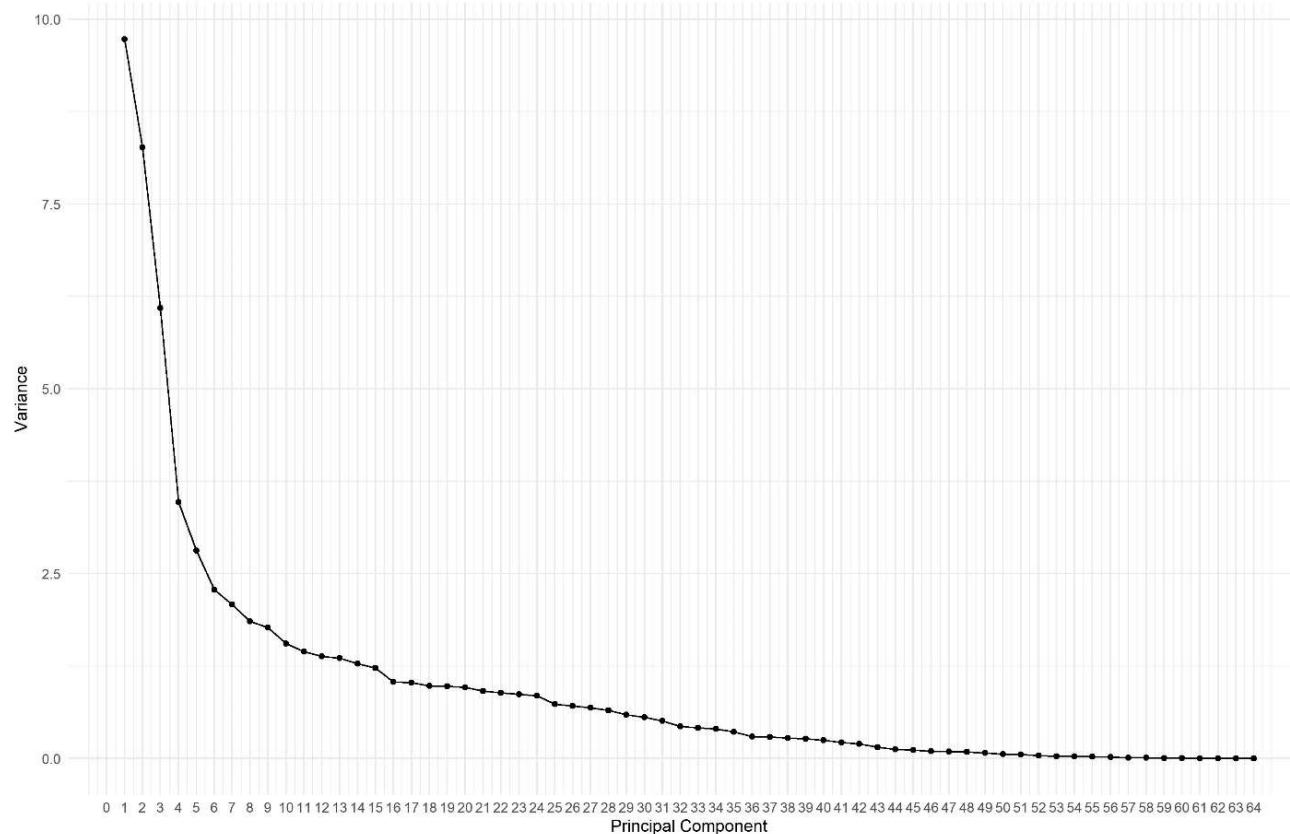


Figure 11. A principal component analysis scree plot showing the variance explained by each principal component. The ‘elbow’ or peak inflection in the curve is generally construed to approximate the optimal number of parameters in a parsimonious model.

Feature Ranking and Selection

This study uses a quantitative framework to rank features, followed by a manual selection process based on causal groupings of features. There are generally two robust approaches to quantitatively ranking feature importance in machine learning workflows to assist with feature selection - filter and wrapper methods (Kohavi and John, 1997). Filter methods apply external algorithms to measure the influence of each feature on the variance of the data. Wrapper methods iteratively assess model performance using random selection of features in order to determine the features that optimize model performance against the training data. A brief summary of each method is provided below, with a more technical explanation in Appendix B. Multiple models were used for the classification and regression workflows with many sequential and random runs per model. This generates a large number of results (see Appendix E) that are used to rank features and ultimately select the ones to carry forward into the final modeling. In seismogenic classification, features with higher rankings are more influential for differentiating seismogenic and non-seismogenic wells. In magnitude regression, features with higher rankings carry higher weights for prediction of magnitude, but the weight has no bearing on the model coefficients. That is, a higher ranking means that a feature carries a lot of weight and for example, has a large absolute univariate slope, but that ranking carries no information regarding the sign of the univariate slope.

Filter-Based Feature Ranking

Two statistical tests were used for filter-based feature ranking for both the seismogenic classification and magnitude regression workflows: information gain and chi-squared testing. Information gain tests the degree to which features share information with the target. Chi-squared testing quantifies the dependency of the target on a feature. Filter-based feature ranking methods are model agnostic and therefore consistent across all models used. Features with high information gain or chi-squared statistics are ranked higher in importance.

The filter-based results for seismogenic classification are shown in Figure 12 for all features including geological features, well information and completion features from the BC OGC data set using the CART model. The results between the two methods are in close agreement. Amongst the highest ranked geological features in both sets of results are Paleozoic structure, top Montney structure and distance to normal faults. Amongst the highest ranked well or completion features are mean TVD of the horizontal well section and the number of wells within 1 km. It is clear from both sets of feature rankings that geological features dominate over the well and completions features.

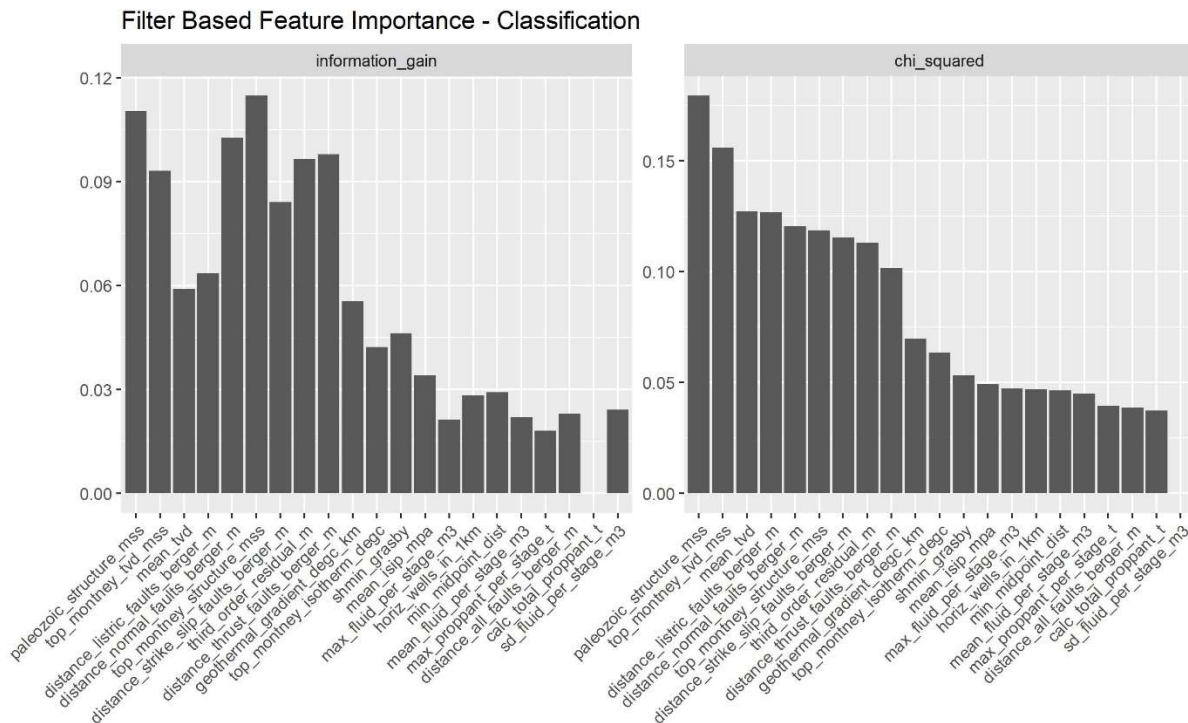


Figure 12. Filter-based feature importance for the classification problem. The left panel shows the information gain scores. The right panel shows the chi squared scores. Features are shown in the same order on both plots.

The filter-based results for magnitude regression are shown in Figure 13 for all features including geological features, well information and completion features from the BC OGC data set using the MARS model. The results both rank some of the same geological features highly including, for example, S_{hmin} gradient, distance to thrust faults and geothermal gradient. The results do not agree as closely for

completion parameters. As with the seismogenic classification feature ranking, geological features dominate over well and completions features.

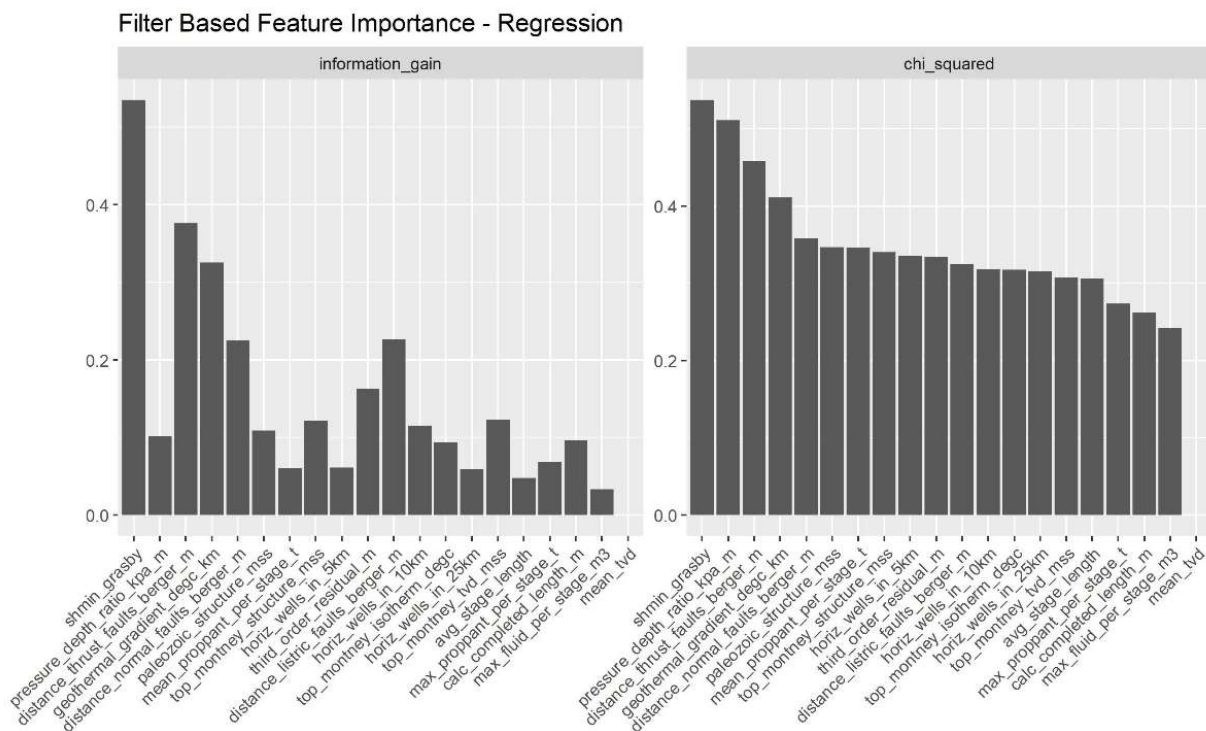


Figure 13. Filter-based feature importance for the regression (maximum magnitude) problem. The left panel shows the information gain scores. The right panel shows the chi squared scores. Features are shown in the same order on both plots.

Wrapper-Based Feature Ranking

Two methods were used for wrapper-based feature ranking. The first was a sequential floating backwards search where features are removed and randomly added back in until a selection of models with the highest performance is achieved (approximately 1,000 of 50,000 model runs). The second is a random search where features are ranked in the best performing models out of a large random set (approximately 1,000 of 100,000 model runs).

An example of the sequential and random wrapper-based results for seismogenic classification is shown in Figure 14 for all features including geological features, well information and completions features from BC OGC data set using the CART model. While the specific results differ between the two methods, in both cases the top ranked features are a combination of geological and well or completions features. In general, the sequential results provide a stronger differentiation of features than the random results for seismogenic classification.

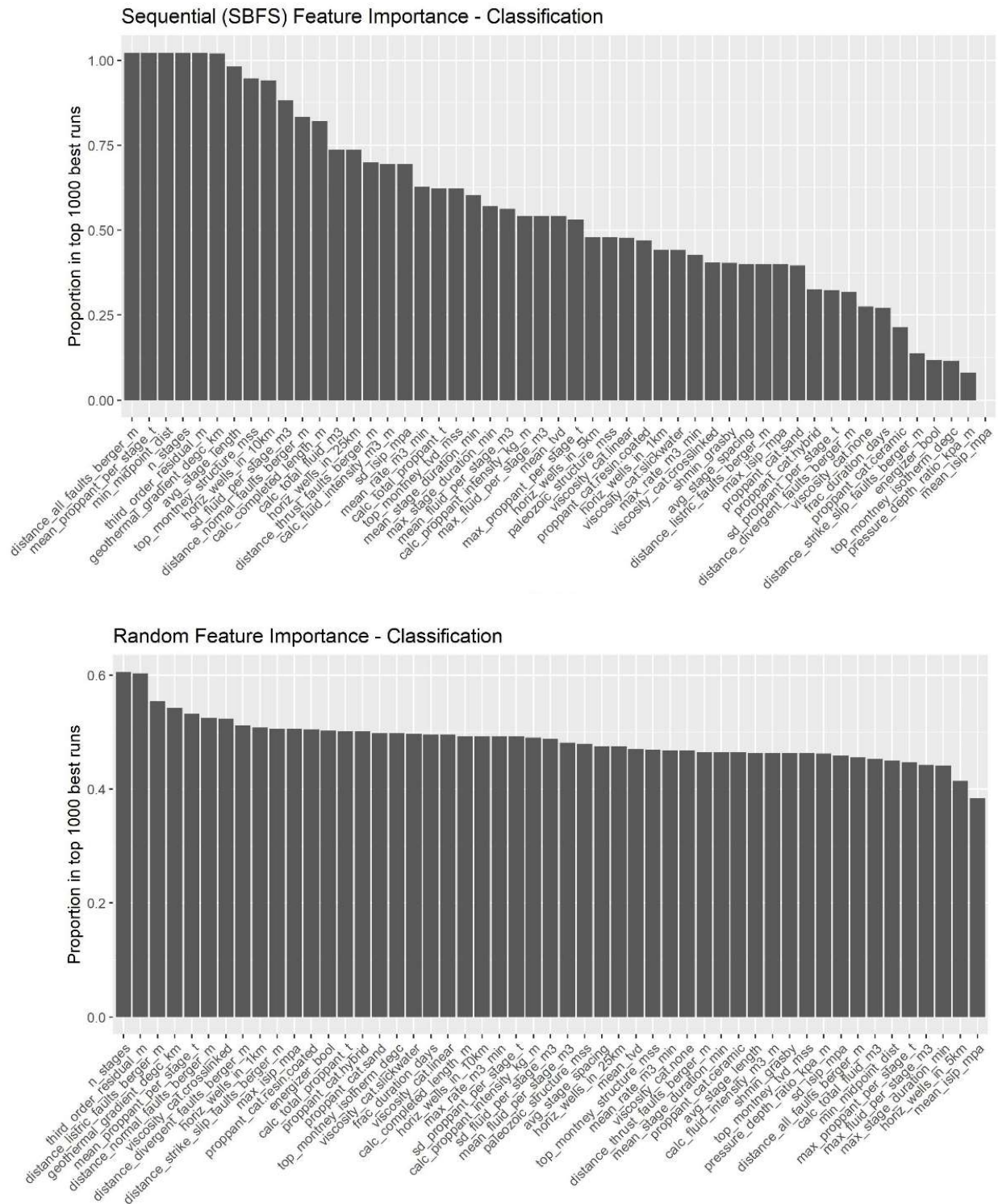


Figure 14. Sequential (top) and random (bottom) wrapper-based feature rankings for the classification analysis using a CART model. In each plot, features are shown in order of importance in the top 1,000 runs, based on the log loss score.

An example of the sequential and random wrapper-based results for magnitude regression is shown in Figure 15 for all features including geological features, well information and completions features from BC OGC data set using the MARS model. The first observation is that there is low differentiation in the sequential feature ranking, which may be an indicator of high interdependence of features in the regression analysis and/or poor predictive power of the features. The random feature importance may provide a more reliable differentiation in features, at least for the example presented. In the random results it is again seen that geological features dominate over well or completions features in the rankings.

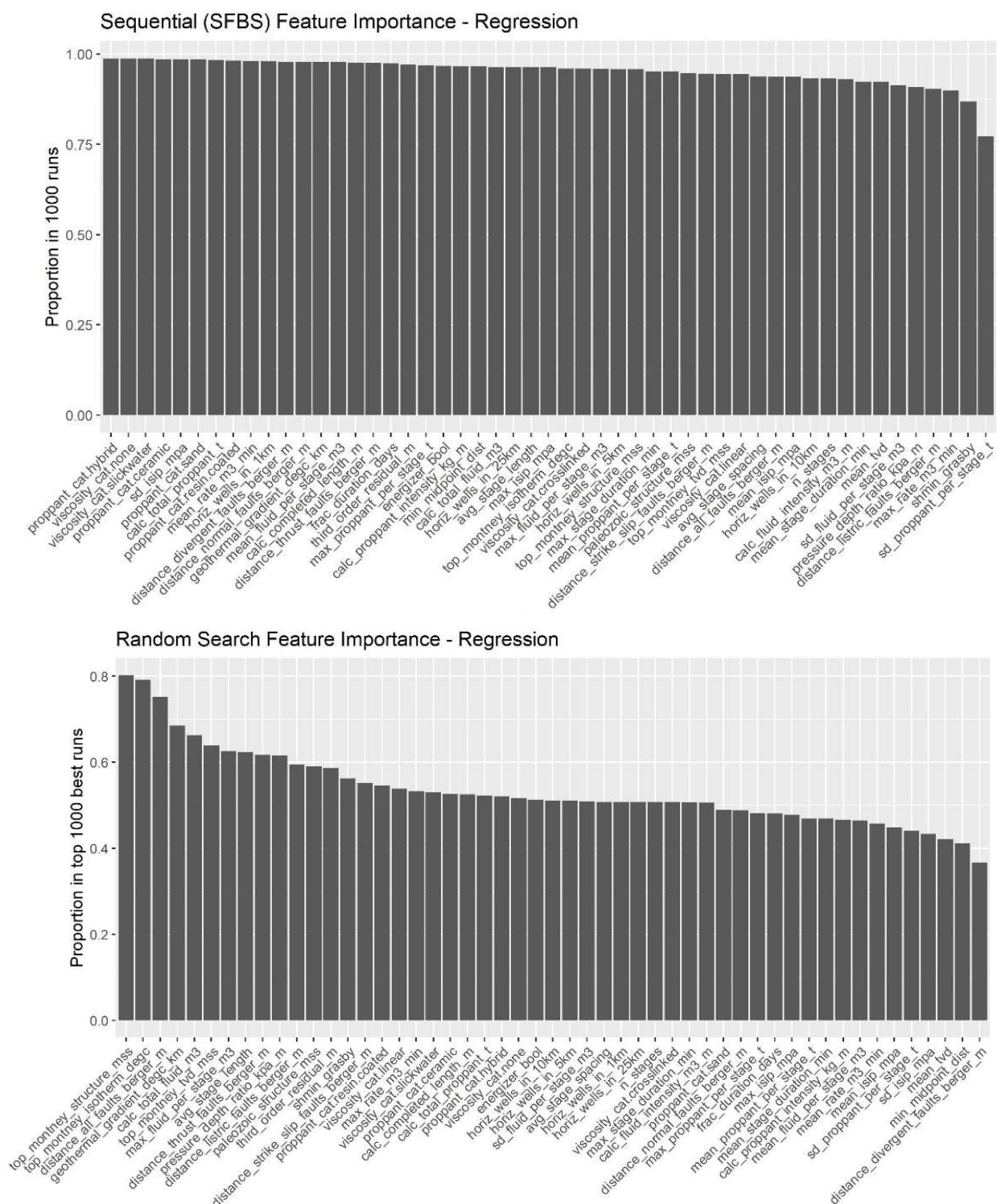


Figure 15. Sequential (top) and random (bottom) wrapper-based feature importance for the regression analysis using a MARS model. In each plot, features are shown in order of importance in the top 1,000 runs, based on the mean of maximum absolute error and root mean squared error.

Discussion and Final Feature Selection

The filter-based, sequential wrapper-based, and random wrapper-based results were compiled across data sets and models for both the classification and regression analyses. The classification analysis, for example, had three models (GLM, CART, and XGBoost), each wrapped using the BC OGC and geoLOGIC completions feature sets. This resulted in two rankings for filter-based methods (one for each data set since filter-based methods are model agnostic), six rankings for sequential wrapper-based methods and six rankings for random wrapper-based methods. For a final numerical ranking, a mean rank was calculated for each of the three groups, and then a final ranking was calculated by averaging the means. The individual model rankings, group rankings and final rankings are all provided in the repository accompanying this report (Appendix E).

The primary goal of this analysis was to try to associate completions parameters with induced seismicity risk, either by being related to induced seismic events or related to induced event magnitude. The results of the feature rankings across all models were in agreement in one respect, and that was that geological features consistently ranked higher than completions features. In order to better highlight the relative importance of the completions features only, the entire feature ranking process was run again on a set of features that did not include geological features (referred to as “completions features only”). The individual and final rankings for the completions features only are also provided in the repository accompanying this report (Appendix E).

For both feature selection runs – all features and completions features only – a considerable amount of time was spent manually comparing the ranking results from the different models and data sources (BC OGC vs. geoLOGIC completions data) as well as the final numerical rankings. Because of the disagreement between individual model rankings, the final rankings did not always reflect how often a feature ranked highly in the models themselves. A likely reason for this is interaction between features. Amongst the geological features, for example, depth is likely related to geothermal gradient or minimum horizontal stress. Amongst the completions parameters, the number of stages is related to completed length and stage spacing. To help examine potential feature interactions, the features were categorized into groups as listed below (see Appendix A for individual feature definitions). When several features from a single group ranked highly, only one or two were eventually selected for the final modeling, and it was attempted to include at least one feature from each group. The final feature selections are presented in Table 3.

- Completions fluids
 - calc_fluid_intensity_m3_m
 - calc_total_fluid_m3
 - energizer_bool
 - max_fluid_per_stage_m3
 - max_rate_m3_min
 - mean_fluid_per_stage_m3
 - mean_rate_m3_min
 - sd_fluid_per_stage_m3
 - viscosity_cat.crosslinked
 - viscosity_cat.linear

- viscosity_cat.none
 - viscosity_cat.slickwater
- Proppants
 - calc_proppant_intensity_kg_m
 - calc_total_proppant_t
 - max_proppant_per_stage_t
 - mean_proppant_per_stage_t
 - proppant_cat.ceramic
 - proppant_cat.hybrid
 - proppant_cat.resin.coated
 - proppant_cat.sand
 - sd_proppant_per_stage_t
- Time
 - frac_duration_days
 - max_stage_duration_min
 - mean_stage_duration_min
- Well design
 - avg_stage_length
 - avg_stage_spacing
 - calc_completed_length_m
 - n_stages
- Montney development density
 - horiz_wells_in_1km
 - horiz_wells_in_5km
 - horiz_wells_in_10km
 - horiz_wells_in_25km
 - min_midpoint_dist
- Distance to faults
 - distance_all_faults_berger_m
 - distance_divergent_faults_berger_m
 - distance_listric_faults_berger_m
 - distance_normal_faults_berger_m
 - distance_strike_slip_faults_berger_m
 - distance_thrust_faults_berger_m
- Structure
 - paleozoic_structure_mss
 - top_montney_structure_mss
 - third_order_residual_m
- Temperature
 - geothermal_gradient_degc_km
 - top_montney_isotherm_degc
- Stress and pressure
 - max_isip_mpa

- mean_isip_mpa
- pressure_depth_ratio_kpa_m
- sd_isip_mpa
- shmin_grasby
- Other
 - mean_TVD
 - max_isip_mpa
 - mean_isip_mpa
 - sd_isip_mpa
 - top_montney_tvd_mss

Completions	Geological
calc_total_fluid_m3 mean_rate_m3_min mean_proppant_per_stage_t proppant_cat.hybrid horiz_wells_in_5km min_midpoint_dist calc_completed_length_m	paleozoic_structure_mss geothermal_gradient_degc_km shmin_grasby distance_listric_faults_berger_m distance_normal_faults_berger_m

Table 3a. Features selected for the final model runs for the seismogenic classification.

Completions	Geological
calc_total_fluid_m3 mean_rate_m3_min calc_total_proppant_t calc_completed_length_m n_stages min_midpoint_dist	pressure_depth_ratio_kpa_m top_montney_structure_mss third_order_residual_m geothermal_gradient_degc_km distance_listric_faults_berger_m distance_normal_faults_berger_m

Table 3b. Features selected for the final model runs for the magnitude regression.

The feature selections for the seismogenic classification and seismic event magnitude regression are not, and do not have to be, the same, because the two problems are different. That stated, there is considerable agreement between the two, both for the completions features and the geological features. Common completions features include total fluid volumes and distance between wells, while common geological features include geothermal gradient and distance to both listric and normal faults.

Model Tuning & Training

Model tuning and training involved using the final, selected features and both data sets (BC OGC and geoLOGIC completions parameters) to create tuned models for interpretation and prediction. The models used in the feature selection process were run with default hyperparameters, which are the settings used to control complexity and model fitting. The default hyperparameters create models with complexity that is not well tuned to the data and doesn't optimally balance bias and variance. Due to the small size of the data sets (especially in the case of magnitude regression), the risk of overfitting a model is quite high. Overfitting occurs when the model fits the training data very accurately (i.e. has low

bias) yet fails to generalize to unseen or new data well (i.e. has high variance). The purpose of model training and tuning is to create a model that balances bias and variance appropriately and quantify how it performs across various sets of data. In this study, the following strategies were used to assess and reduce overfitting:

1. Models are trained using five-fold cross validation. Cross validation creates multiple ‘folds’ of training data. The hyperparameters are trained to optimize model fit across all the folds, sequentially training the model on four of the five folds while evaluating its performance against the holdout fold. This is considered to be one of the best ways to avoid overfitting in a model, but as is observed in the XGBoost results, can still lead to overfitting when the model complexity overwhelms the data set size.
2. Use bootstrap resampling to quantify a tuned model’s bias and variance. While cross-validation helps avoid overfitting, it only provides a single performance metric (the out of sample test metric on each fold) for the model. By randomly permuting the data set and train / test split while holding the hyperparameters constant, an estimate of the in-sample bias and variance of each model is obtained.
3. Keep a small holdout set to evaluate out-of-sample performance. The irreducible error of models can be evaluated in this way; however, the performance is sensitive to which data is selected for the holdout. This measure is used as an estimate of how the model will generalize to new data, and thus its out of sample variance and irreducible error. For example, the XGBoost model has near perfect performance on the cross-validated data, but still performs relatively poorly on the holdout set.

Each model type is discussed briefly below. Additional details on each model along with a description of its hyperparameters can be found in Appendix B.

- Generalized linear models (GLMs) are extensions of multivariate linear regression. In this study, GLMs were applied for both seismogenic classification and magnitude regression. For classification, a logistic link function was used to transform the continuous output of the GLM to a binomial classification (e.g. seismogenic or not seismogenic).
- Classification and regression trees (CART) are decision trees used for seismogenic classification. The tree uses an impurity metric, or the degree of efficiency in separating classes using a feature, to create a decision tree. This tree is pruned using a complexity penalty.
- Multivariate Adaptive Regression Splines (MARS) models are an extension to multivariate linear regression, where each surface is allowed to ‘bend’ around knots. These knots provide a form of non-linearity where the model can represent more complex surfaces by changing slope through each variable. A complexity-penalized optimization is used to select the number of knots. In this study, a MARS model was applied to the magnitude regression.
- An extreme gradient boosted machine (XGBoost) was used for both the seismogenic classification and magnitude regression (Friedman, 2001). A gradient boosted machine is a tree-based ensemble model, similar to the widely used random forest style of models, but with

added regularization terms which provides a balance between model complexity (accuracy) and overfitting (bias).

The final features from the feature selection workflow described above were used to subset the data into final training data sets for both the BC OGC and geoLOGIC data sets. A total of 12 models were then trained, six for seismogenic classification and six for magnitude regression. As detailed above, a Bayesian optimization algorithm with five-fold cross validation was used to tune hyperparameters for each model. The tuned model was then evaluated using bootstrap resampling and against the small holdout data set. The performance of each model can therefore be described using the following metrics:

- The performance of the tuned model against the training and holdout data sets: This is illustrated using a performance versus threshold plot (Figure 16) and receiver operating characteristic (ROC) curve (Figure 17) for seismogenic classification and a residual plot (Figure 18) for magnitude regression. Plots for each model are provided in the project repository (Appendix E). Single point performance metrics are also provided for easier comparison between models and data sets. These metrics (MMCE, F1, Log Loss, RMSE, MAE) are summarized in Tables 4 and 5.
- The performance metrics, bias, and variance determined using bootstrap resampling: By evaluating the tuned model against a resampled training set many times, it is possible to estimate how the model will perform against another data set that is similar in distribution to the training set. The bias and variance for the seismogenic classification and magnitude regression problems are illustrated in Figure 19. The test bias and variance are always higher than the train bias and variance, and the discrepancy between the training and test data increases as the data set size decreases, illustrated by the difference between the classification and regression results. It should be noted that the calculation for bias and variance is substantially different for classification and regression problems, which also creates differences when interpreting the plot. The performance of the models is summarized in Table 4 for seismogenic classification and Table 5 for magnitude regression.

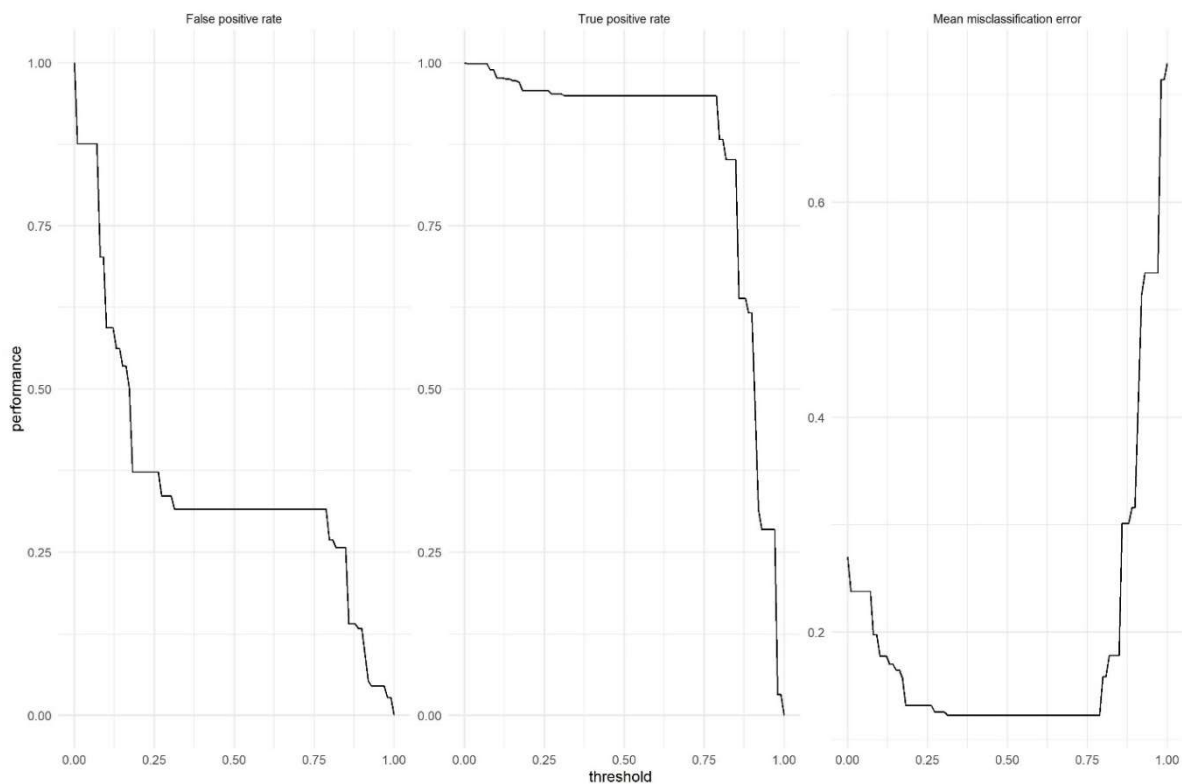


Figure 16. An example of a performance versus threshold plot from the CART model applied to the BC OGC data set for seismogenic classification. The false positive rate, true positive rate, and mean misclassification error of the seismogenic classification model.

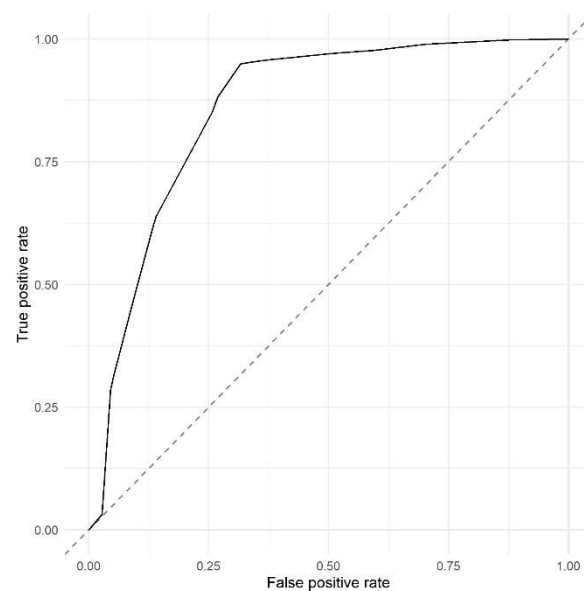


Figure 17. An example of a receiver operating characteristic curve from the CART model applied to the BC OGC data set for seismogenic classification. The true positive rate is plot against the false positive rate as the threshold changes.

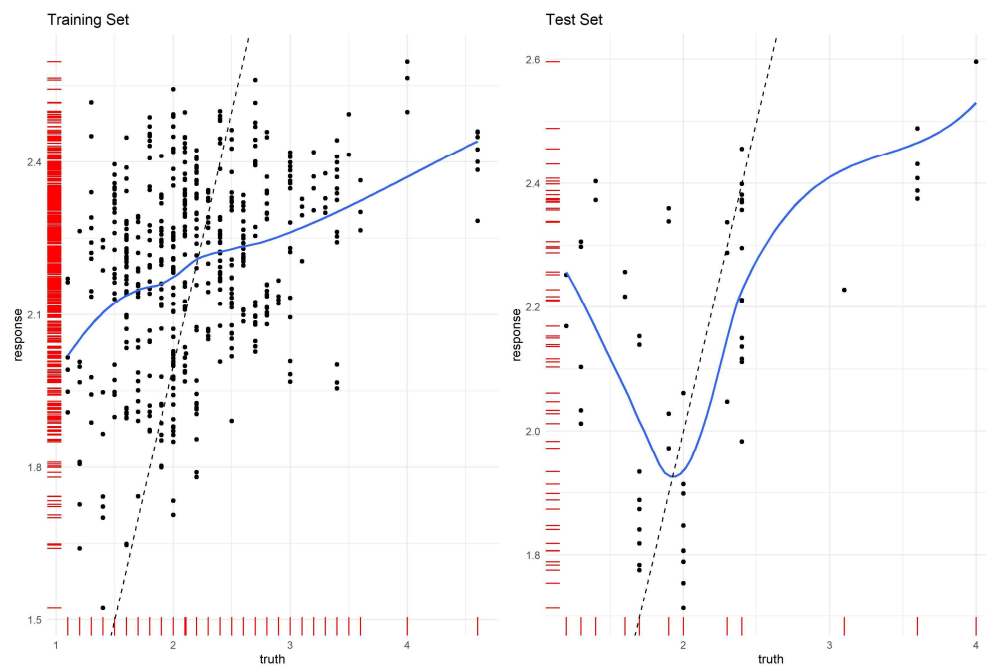


Figure 18. An example of a residual plot from the GLM model applied to the BC OGC data set for magnitude regression. Red lines show the data observations (bottom) and predicted values (left). Note the binning bias in the model, shown by the distinct truth values. The blue line shows a LOESS average of the fit for a visual aid only.



Figure 19. Summary of the bias and variance in the seismogenic classification (top) and magnitude regression (bottom).

Model	GLM	GLM	CART	CART	XGBoost	XGBoost
Data set	geoLOGIC	BC OGC	geoLOGIC	BC OGC	geoLOGIC	BCOGC
Train Bias	0.19	0.26	0.11	0.12	0.00	0.00
Train Variance	0.19	0.25	0.11	0.12	0.04	0.03
Holdout Bias	0.19	0.26	0.16	0.17	0.08	0.08
Holdout Variance	0.19	0.25	0.11	0.12	0.07	0.07
Train Log Loss	0.39	0.50	0.25	0.31	0.04	0.03
Train F1	0.90	0.85	0.95	0.93	1.00	1.00
Train MMCE	0.18	0.25	0.09	0.10	0.00	0.00
Holdout Log Loss	0.75	0.62	0.85	3.49	0.57	0.69
Holdout F1	0.78	0.82	0.77	0.81	0.83	0.82
Holdout MMCE	0.32	0.29	0.31	0.30	0.23	0.27

Table 4. Summary of model performance for the seismogenic classification analysis. Each model and data set are presented with the training and holdout bias and variance. Three performance metrics are presented for the training and holdout sets: the log loss score, the F1 score, and the mean misclassification error (MMCE).

Model	GLM	GLM	MARS	MARS	XGBoost	XGBoost
Data set	geoLOGIC	BC OGC	geoLOGIC	BC OGC	geoLOGIC	BCOGC
Train Bias	0.341	0.340	0.207	0.198	0.012	0.017
Train Variance	0.002	0.001	0.044	0.038	0.002	0.004
Holdout Bias	0.363	0.352	-	0.523	0.160	0.150
Holdout Variance	0.002	0.001	-	0.304	0.022	0.024
Train MAE	0.444	0.454	0.329	0.348	0.078	0.093
Train RMSE	0.581	0.584	0.443	0.467	0.104	0.127
Holdout MAE	0.466	0.436	0.438	0.346	0.399	0.397
Holdout RMSE	0.640	0.599	0.607	0.444	0.545	0.487

Table 5. Summary of model performance for the magnitude regression analysis. Each model and data set are presented with the training and holdout bias and variance. Two performance metrics are presented for the training and holdout sets: the maximum absolute error (MAE) and the root mean squared error (RMSE).

IV. Model Interpretation

The objective of model interpretation is to evaluate the global and local behaviour of a single model relative to the data used to train that model. The use of multiple datasets and models makes model interpretation difficult since the fit of each model to a particular dataset is highly variable when employing an empirical (i.e. statistical) framework. This problem is especially apparent when model complexity increases, or tree-based methods are used. This section attempts to examine the similarities, or lack thereof, between the various models using model interpretation techniques implemented in the MLR and IML libraries in R. The techniques include feature importance and feature interaction (Breiman, 2001), partial dependence and individual conditional expectation plots (Friedman, 2001; Goldstein et al.,

2015), local interpretable model-agnostic explanations (LIME, Ribeiro et al., 2016) and SHapley Additive exPlanations (SHAP, Shapley, 1953; Lundberg and Lee, 2017). A general overview of a model interpretation workflow is illustrated in Figure 20.

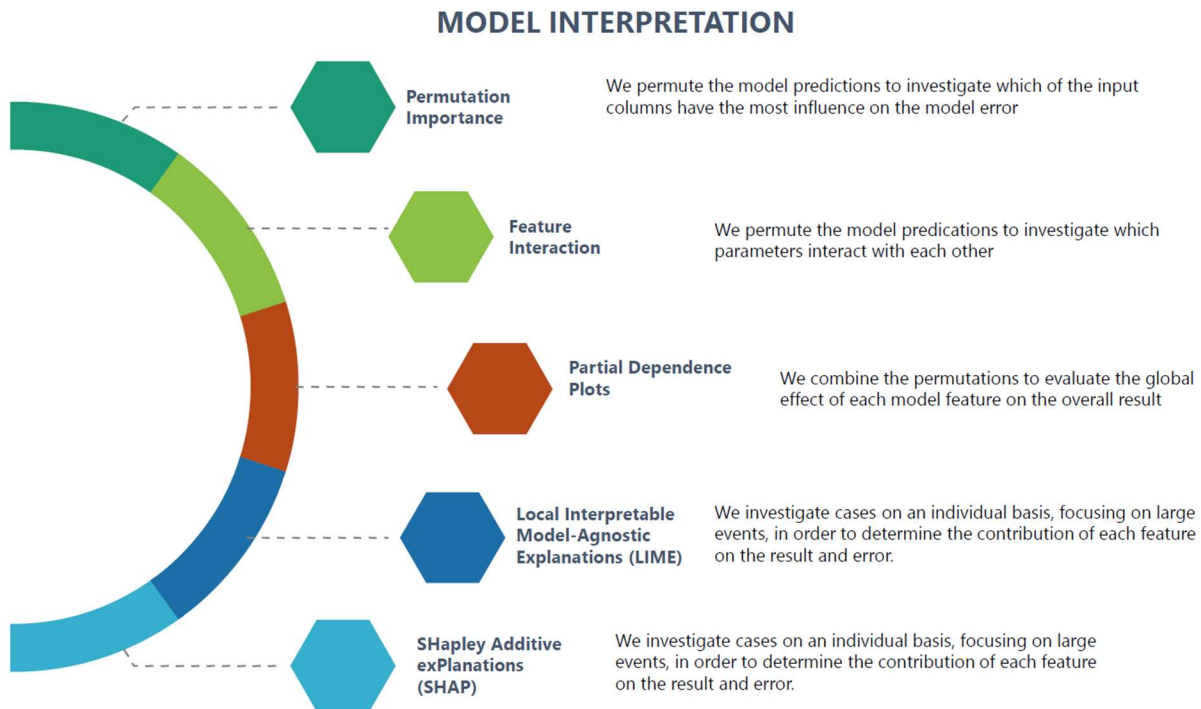


Figure 20. Illustration of the model interpretation workflow.

Feature Importance and Interaction

Permutation-based feature importance was used to interpret the global importance of each input feature on the target. Features with higher importance have a higher effect on the model prediction. Due to the varying nature of the seismogenic classification models used, it was necessary to use different methods to quantify feature importance for each, which makes it difficult to directly compare the results between models. The following importance measurements were used for each model:

- **GLM:** The coefficients are reported for each feature. In linear models, features with larger absolute coefficients have a larger effect on model predictions. It is important to note the impact of the input variable scale, since the outcome of each variable is equal to the input value times the coefficient. Features with negative coefficients decrease the probability of seismogenic classification whereas positive coefficients increase it.
- **CART:** The sum of the decrease in impurity for each of the features at each node in the tree is reported. Features that cause the largest aggregated decrease in impurity are considered the most important. This importance is absolute, that is does not indicate how this feature affects the seismogenic classification.

- XGBoost: The sum of the relative contribution of each of the features at each node is reported. Features that have the largest relative contribution are considered the most important. This metric, while scaled between 0 and 1, is expected to be somewhat similar to the impurity decrease reported in the CART model.

For the magnitude regression models, a consistent permutation importance framework was used. The maximum absolute error (MAE) of the model is measured before and after shuffling values of individual features. The increase in MAE is used as a proxy for the feature's importance. In the results presented below, a value of 1 represents the global model error and larger values indicate more generated error and therefore a larger importance. The values provide no indication of how a parameter effects the model prediction in absolute terms (i.e. positively or negatively).

The interaction between features is also quantified, since strongly interacting features can create issues with statistical models (Hall, 1999). The interaction measure varies between 0 and 1, with 0 representing no interaction and 1 meaning that 100% of the model variance is explained by the interaction between two features. Features with a high interaction value indicate that features might be combining to explain a latent (non-quantified) parameter, or that variables may be confounding each other. Ideally a model would reduce feature interactions, but due to the high degree of correlation between geological and completion parameters, this was found to be difficult in the current study.

The feature importance and interaction plots for the seismogenic classification using all three models on both data sets (BC OGC and geoLOGIC completions features) are shown in Figure 21. The first observation is that each model generally shows the same results regardless of which completions data set was used. The second is that feature interaction tends to mimic feature importance and that when models assign a certain feature importance, they may be doing it at the cost of another feature that interacts with that feature. This suggests that more complex models (CART and XGBoost) will likely experience a higher variance not only due to overfitting, but also due to feature interaction. That said, both the CART and XGBoost models place a relatively high importance on Paleozoic structure and distance to faults. The GLM model places a high negative importance on minimum horizontal stress and high positive importance on geothermal gradient, distance between wells and mean proppant per stage.

The feature importance and interaction plots for the magnitude regression using all three models on both data sets are shown in Figure 22. Most models show a relatively high importance for top of Montney structure and distance to faults (normal and thrust). Interactions tend to be higher for completions parameters than for geological parameters. As with the classification models, interaction tends to mimic importance. It is also worth noting how the XGBoost model balances feature importance better than the MARS and GLM models, which tends to focus importance on two or three features.

Statistical Assessment of Operational Risks for Induced Seismicity from Hydraulic Fracturing in the Montney, Northeast BC (Geoscience BC Project 2019-008) – Final Report

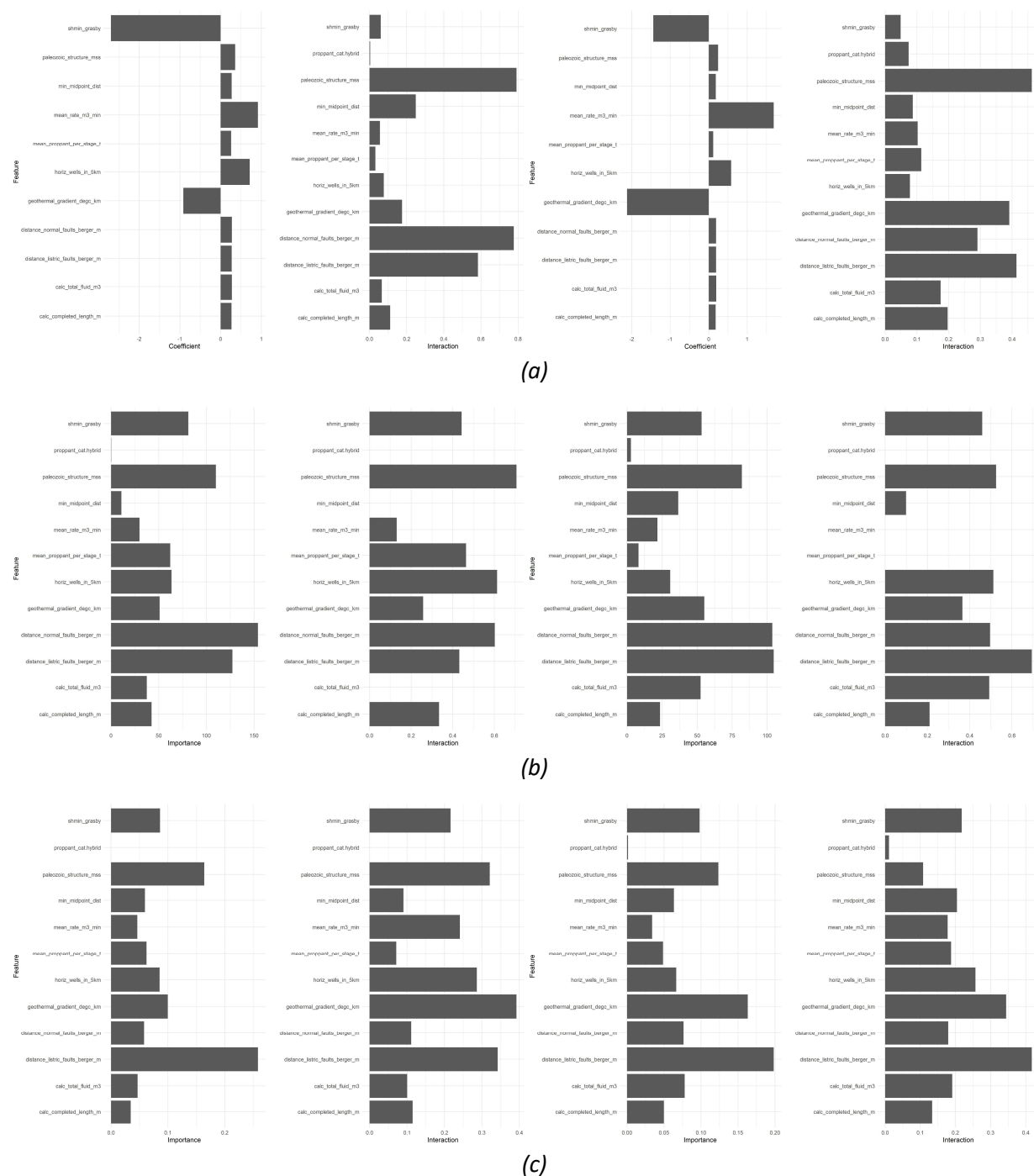


Figure 21. Feature importance and interaction for the classification models run using, BC OGC (left) and geoLOGIC (right) completions data: (a) GLM models, (b) CART models, (c) XGBoost models. Different feature importance measures were used: the model coefficients are directly shown for the GLM model while the sum of impurity decrease is shown for the CART model, and the sum of feature contribution is shown for the XGBoost model. Full-size versions of each plot are provided in Appendix C.

Statistical Assessment of Operational Risks for Induced Seismicity from Hydraulic Fracturing in the Montney, Northeast BC (Geoscience BC Project 2019-008) – Final Report

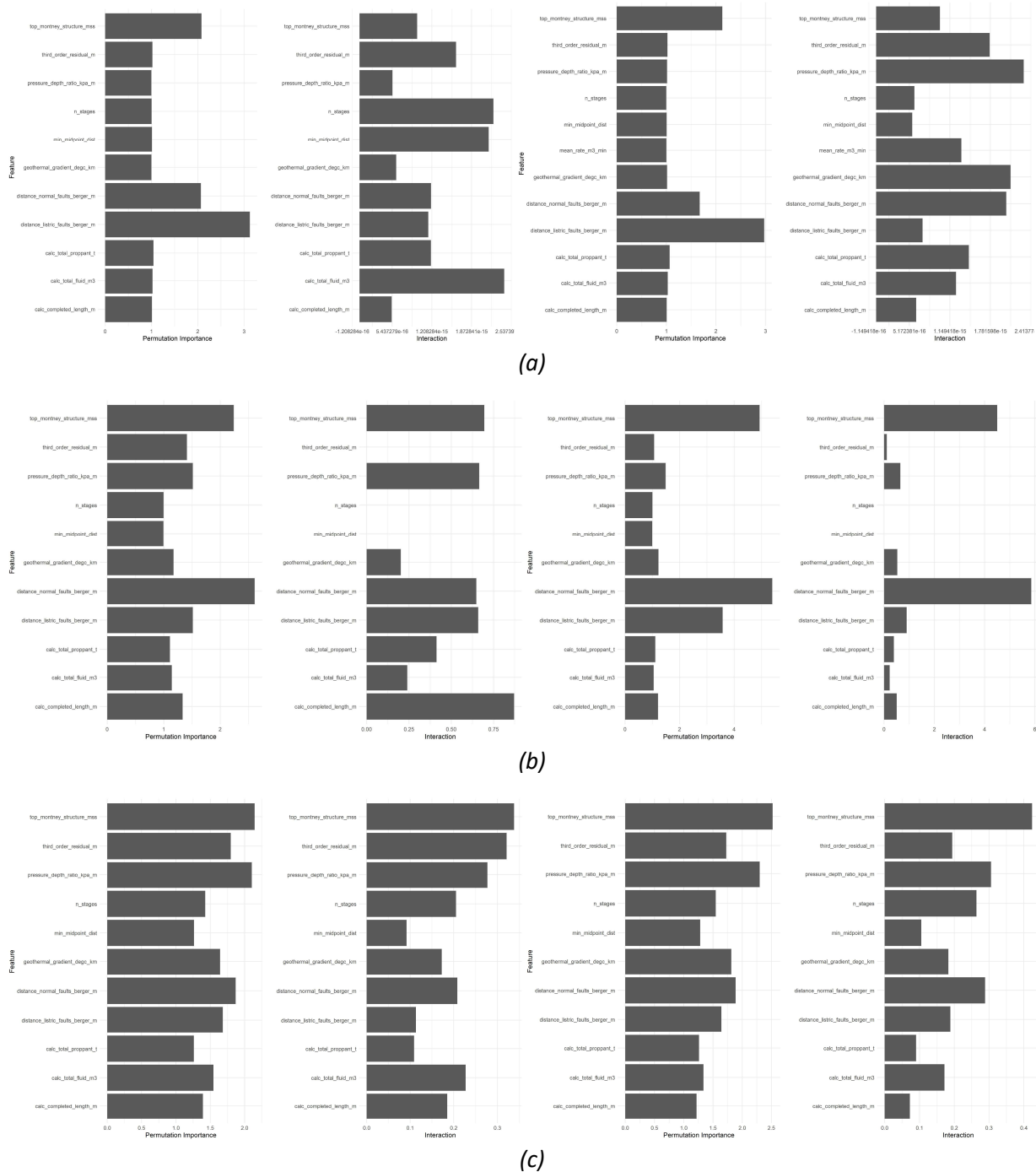


Figure 22. Feature importance and interaction for the regression models run using both sets of completions features, BC OGC (left) and geoLOGIC (right): (a) GLM models, (b) MARS models, (c) XGBoost models. Permutation feature importance is used for all models. Full-size versions of each plot are provided in Appendix C.

Partial Dependence and Conditional Expectation

Partial Dependence Plots (PDPs) compare the change in the average predicted value as a particular feature (or features) varies over its marginal distribution. This is done by holding all variables constant for each observation in the training data set and then applying the unique values of one feature for each observation. Individual Conditional Expectation (ICE) curves are an extension of PDPs. Instead of plotting the average marginal effect on the response variable, they show the change in the predicted response variable for each observation as each predictor variable is varied. Since ICEs plot combinations of variables that fall outside of the observed data, they tend to appear chaotic and show high variability across the features. They are useful, however, for providing a measure of model variance and potential bias on unseen data.

The PDP and ICE curves for the seismogenic classification problem using the GLM model and the BC OGC completions features are provided in Figure 23. The plot illustrates how a seismogenic classification (X1) or non-seismogenic classification (X0) changes as each feature changes. Note that because the classification is binary, the X1 plots and X0 plots are exact opposites of each other. With the exception of hybrid proppant (a categorical variable), each feature increases or decreases the probability of a well being seismogenic. Large, non-linear responses are observed with geological features (e.g., Paleozoic structure and distance to listric faults). These are compared with relatively muted, linear responses from completion parameters (e.g., total fluid volume and horizontal wells within 5 km). More specifically, for example, a larger geothermal gradient decreases the probability of a well being seismogenic, whereas a larger total injected fluid volume increases the probability. It is also worth noting the rug plots at the bottom of each PDP, which shows the distribution of the variables. Extrapolation of a PDP past the data (as with the minimum horizontal stress feature) is fraught with interpretation error, as statistical models perform poorly at extrapolating past the bounds of observed data.

Similar plots for the seismogenic classification using the CART and XGBoost models are shown in Figures 24 and 25, respectively. The CART model shows negligible variability for some parameters, which likely indicates that there aren't any nodes in the tree for that parameter, or that the PDP analysis couldn't permute the feature outside of one or two nodes where it was used. For other parameters it shows discrete decision nodes. For example, it appears that the model encodes distance to normal faults using two nodes, placing a higher seismogenic potential for wells within 500 and 2500 m from a normal fault. The XGBoost model PDPs shows the much higher variability encoded by the ensemble model. This variability unfortunately mutes some of the ICE trends, but the PDPs suggest that seismogenic potential correlates with an increase in distance from listric faults, decrease in geothermal gradient, increase in development density, and increase in total fluid injected. The XGBoost PDPs also demonstrate the much more complex response surface of the ensemble model, which can encode many decision nodes for each feature. This is one of the reasons ensemble tree models are prone to overfitting.

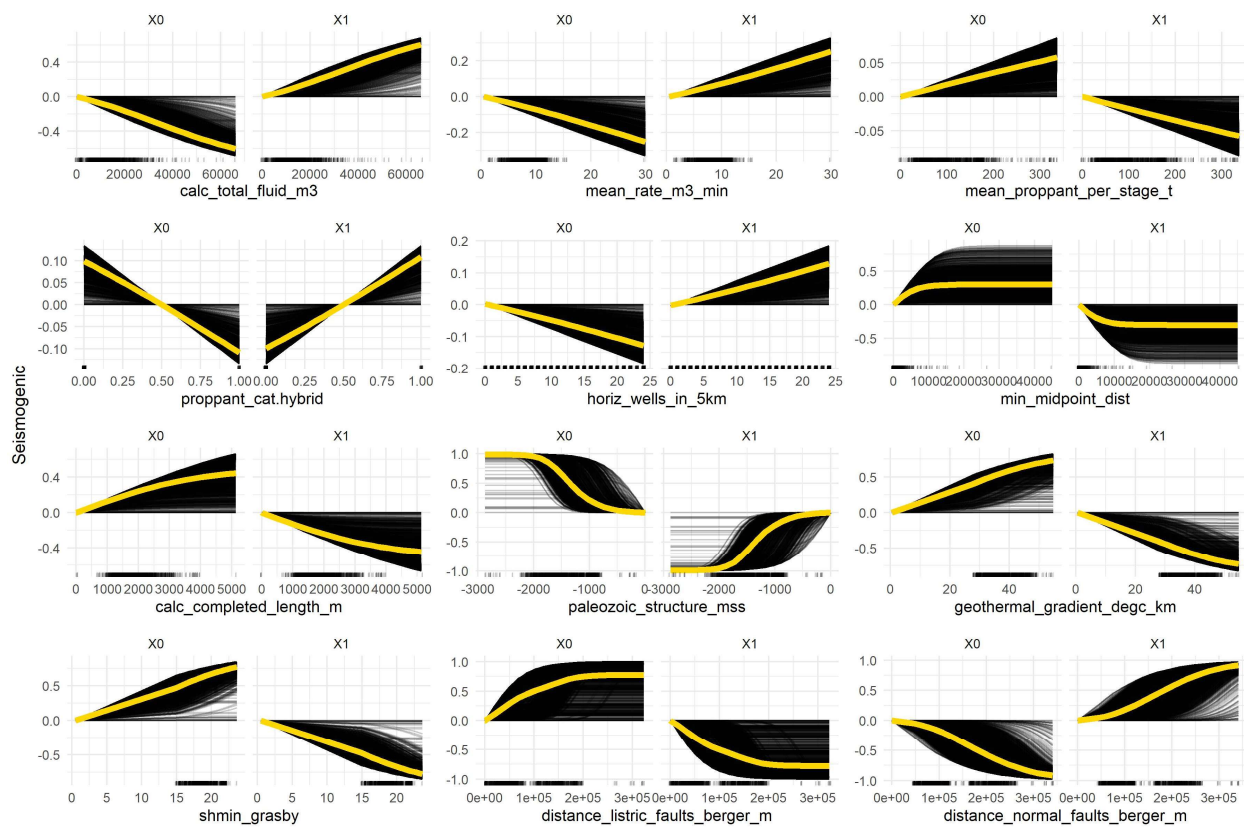


Figure 23. Partial dependence plot (yellow) and individual conditional expectation plots (black) for the seismogenic classification using the GLM model and BC OGC completions features. Variables are standardized and centered between the two classes. The rug plot at the bottom of each PDP plot shows the data distribution of each feature.

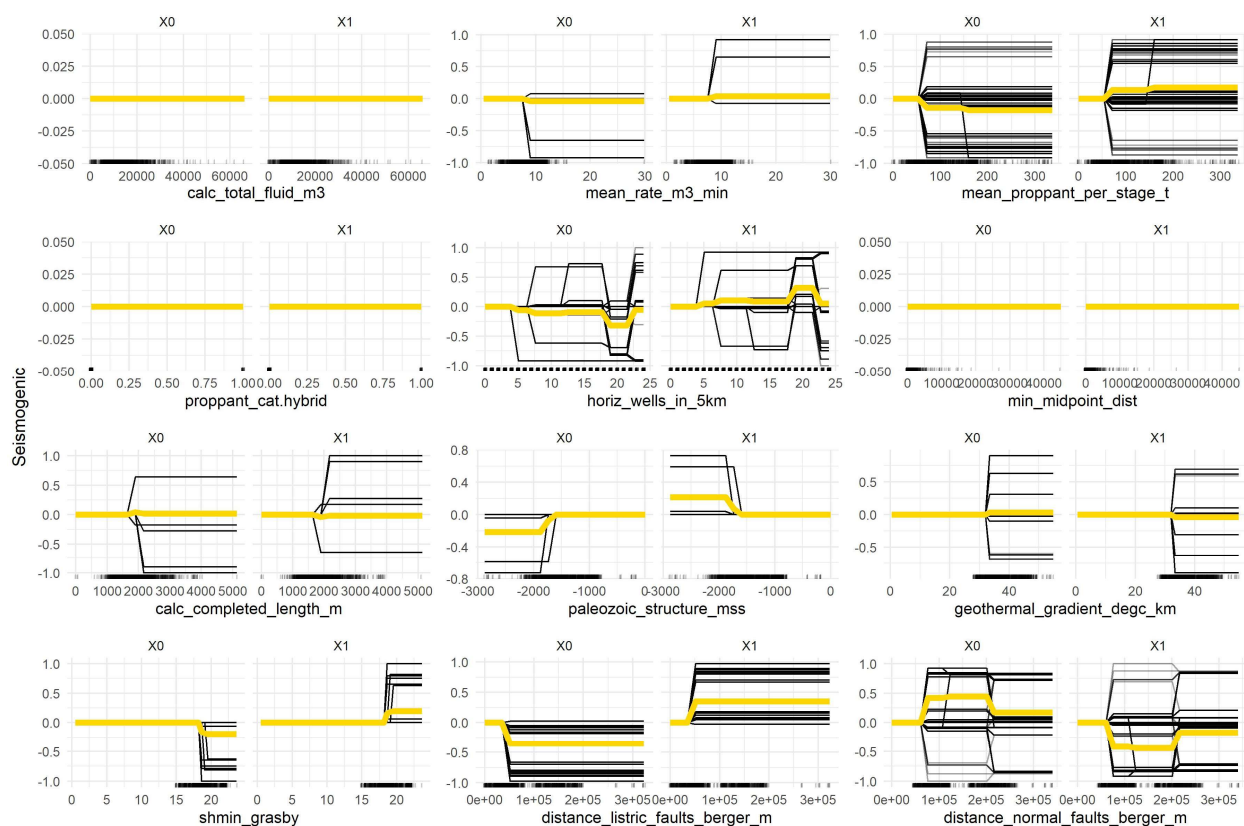


Figure 24. Partial dependence plot (yellow) and individual conditional expectation plots (black) for the seismogenic classification using the CART model and BC OGC completions features. Variables are standardized and centered between the two classes. The rug plot at the bottom of each PDP plot shows the data distribution of each feature.

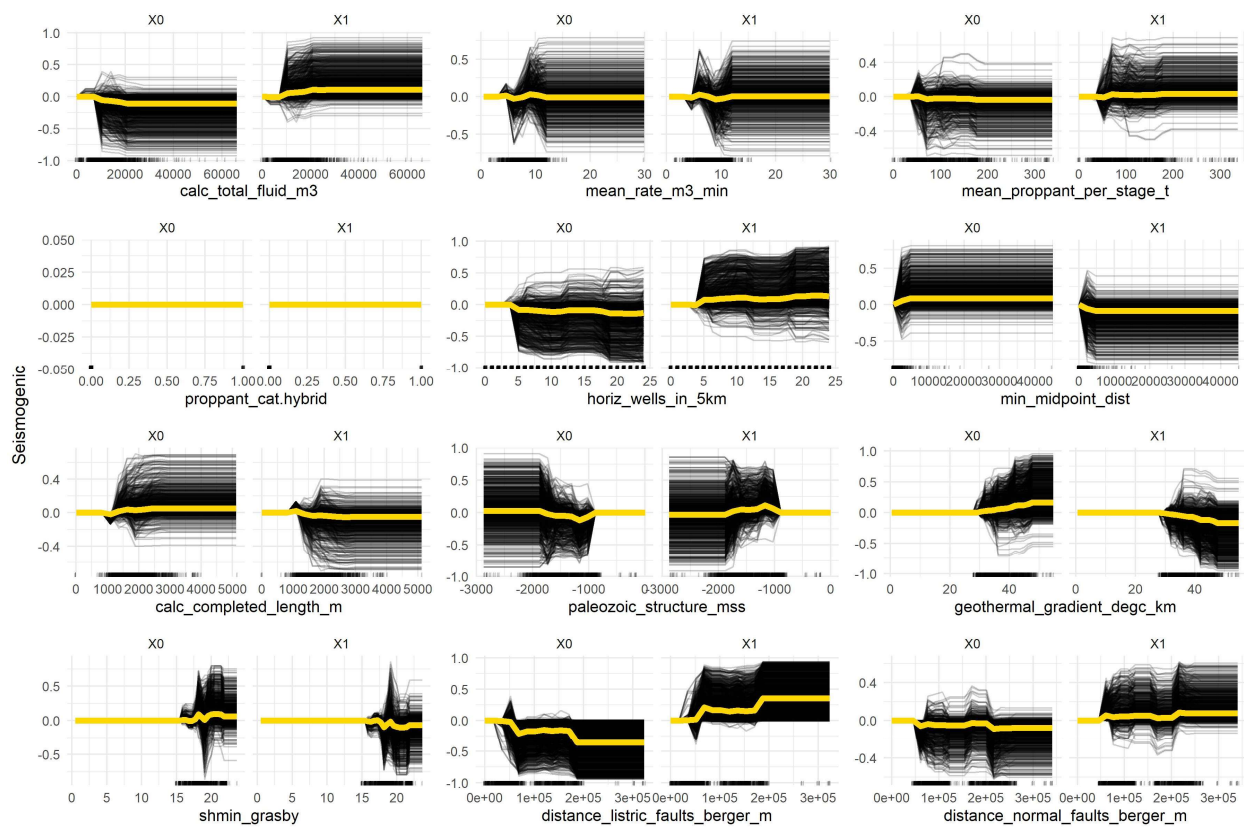


Figure 25. Partial dependence plot (yellow) and individual conditional expectation plots (black) for the seismogenic classification using the XGBoost model and BC OGC completions features. Variables are standardized and centered between the two classes. The rug plot at the bottom of each PDP plot shows the data distribution of each feature.

The PDPs and ICE curves for the magnitude regression using the BC OGC completions features and the GLM, MARS and XGBoost models are provided in Figures 26 to 28. As opposed to the classification PDPs, which map a model to a binomial response (0 or 1), the PDPs map a model to a continuous response.

The response for the GLM model (Figure 26) shows the increase or decrease in magnitude as each univariate parameter increases. In a GLM, the sum of these changes provides the final prediction, thus the magnitude of each slope provides the importance of each feature. Since the model only encodes a single coefficient, there is no variability in the model's response as a feature is permuted (in other words, there are no PDPs for a GLM of a continuous variable). This isn't the case with the classification model shown above because of the logistic link function, which encodes non-linearity in the model through a probability-based prediction with a tuned threshold. For the GLM model, total fluid volume, total proppant, number of stages, distance between wells third order residual appear to be the most importance parameters. Interestingly, the majority of the geological parameters show little influence on the model prediction.

The response for the MARS model is shown in Figure 27. The first observation for this PDP is the large variance in the PDP response, which is likely due to a combination of parameters that is well outside of observed data. For example, at the extremes of the PDPs calculated completed length predicts a

magnitude of -20, and geothermal gradient predicts a magnitude of 7.5. This is an effect of the PDP quantification process and does not reflect the model's predictive performance, but still serves to show some of the variability inherent in the MARS model. The PDP/ICE plot does indicate the large influence of top of Montney structure and indicates an increasing predicted event magnitude as total fluid and total proppant increase, while calculated completed length and third order residual seem to decrease the magnitude.

The response for the XGBoost model in Figure 28 is prone to much less extrapolation error than the MARS model, but still shows a relatively high variance. The response surface of this model provides several interesting observations. First, the magnitude appears to increase with top of Montney structure and decreases with third order residual. Second, it appears that once wells exceed approximately 1.5 km in length, the influence of length on magnitude attenuates. The effects of increasing total fluid and number of stages are also apparent, and there appears to be an initial increase in magnitude with pressure/depth ratio but then the opposite influence occurs.

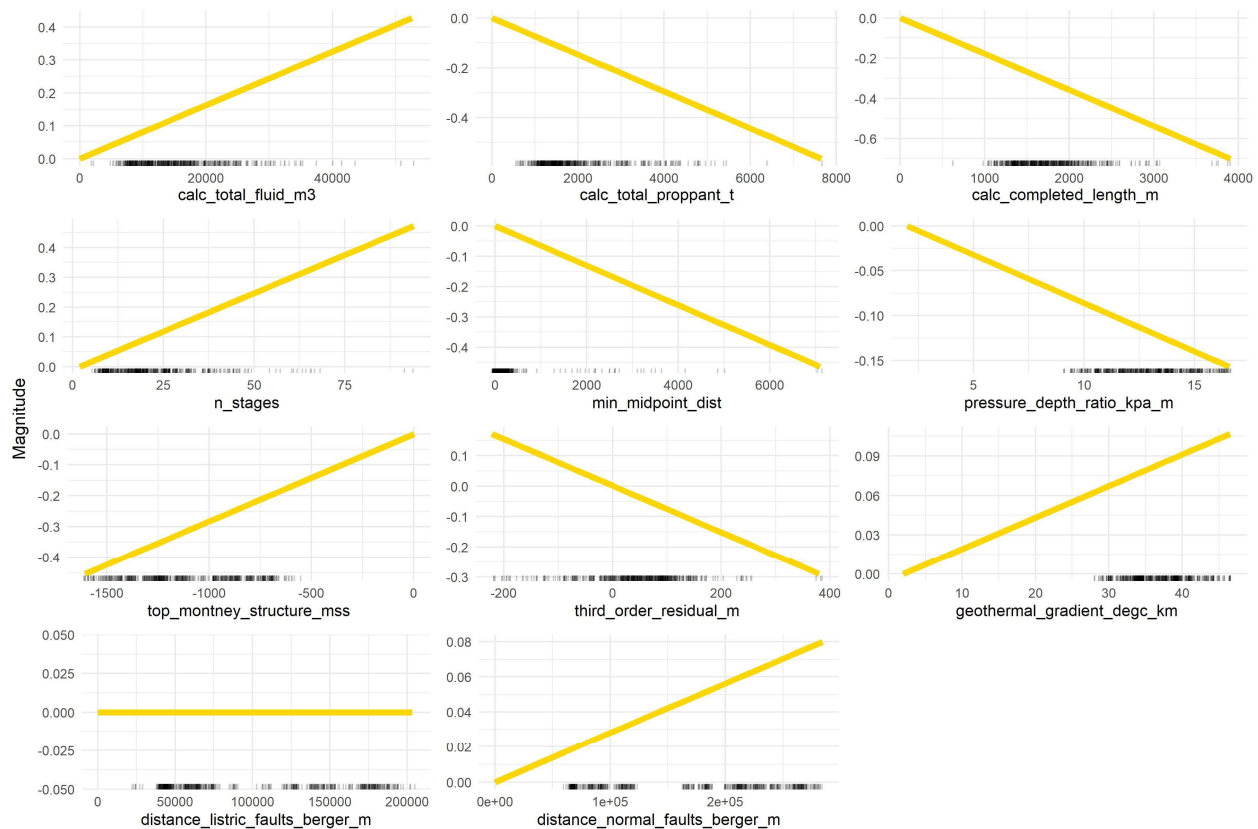


Figure 26. Partial dependence plot (yellow) for the magnitude regression using the GLM model and BC OGC completions features. The rug plot at the bottom of each PDP plot shows the data distribution of each feature.

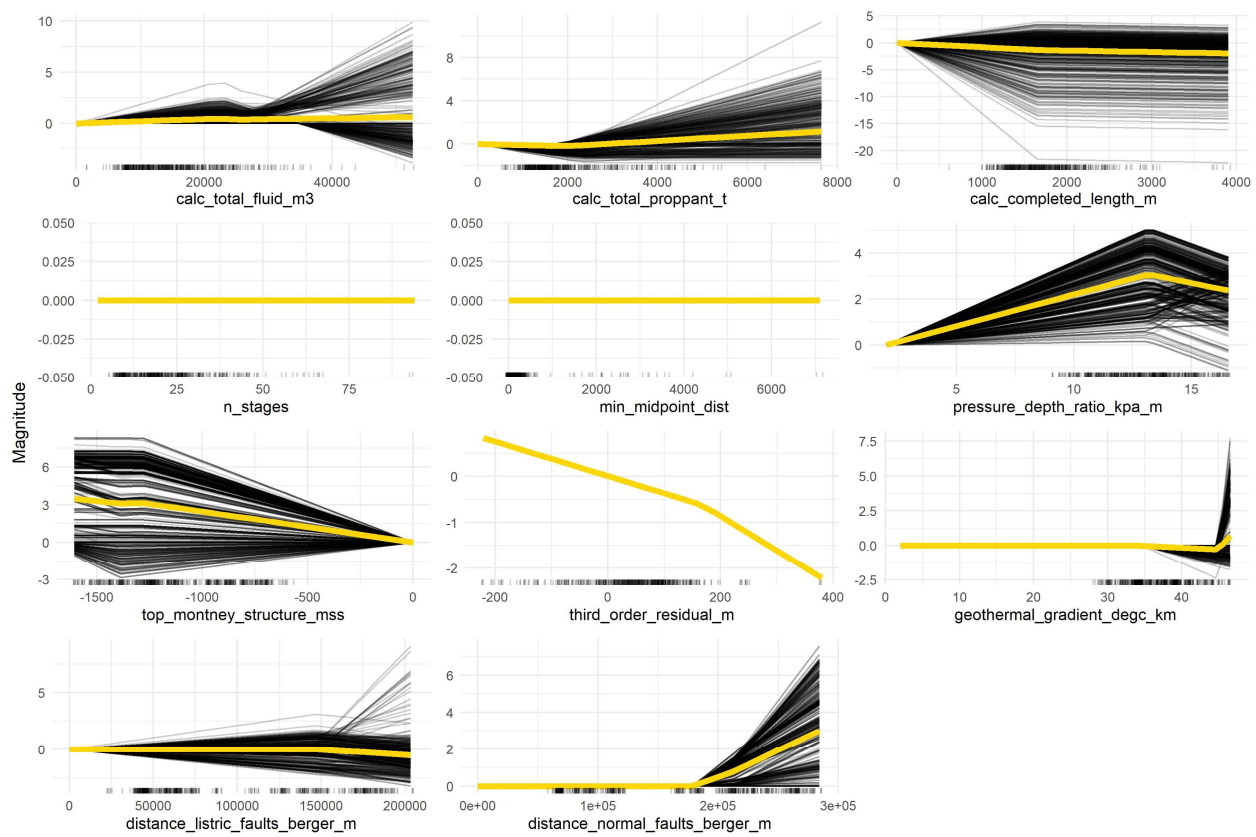


Figure 27. Partial dependence plot (yellow) and individual conditional expectation plots (black) for the magnitude regression using the MARS model and BC OGC completions features. The rug plot at the bottom of each PDP plot shows the data distribution of each feature.

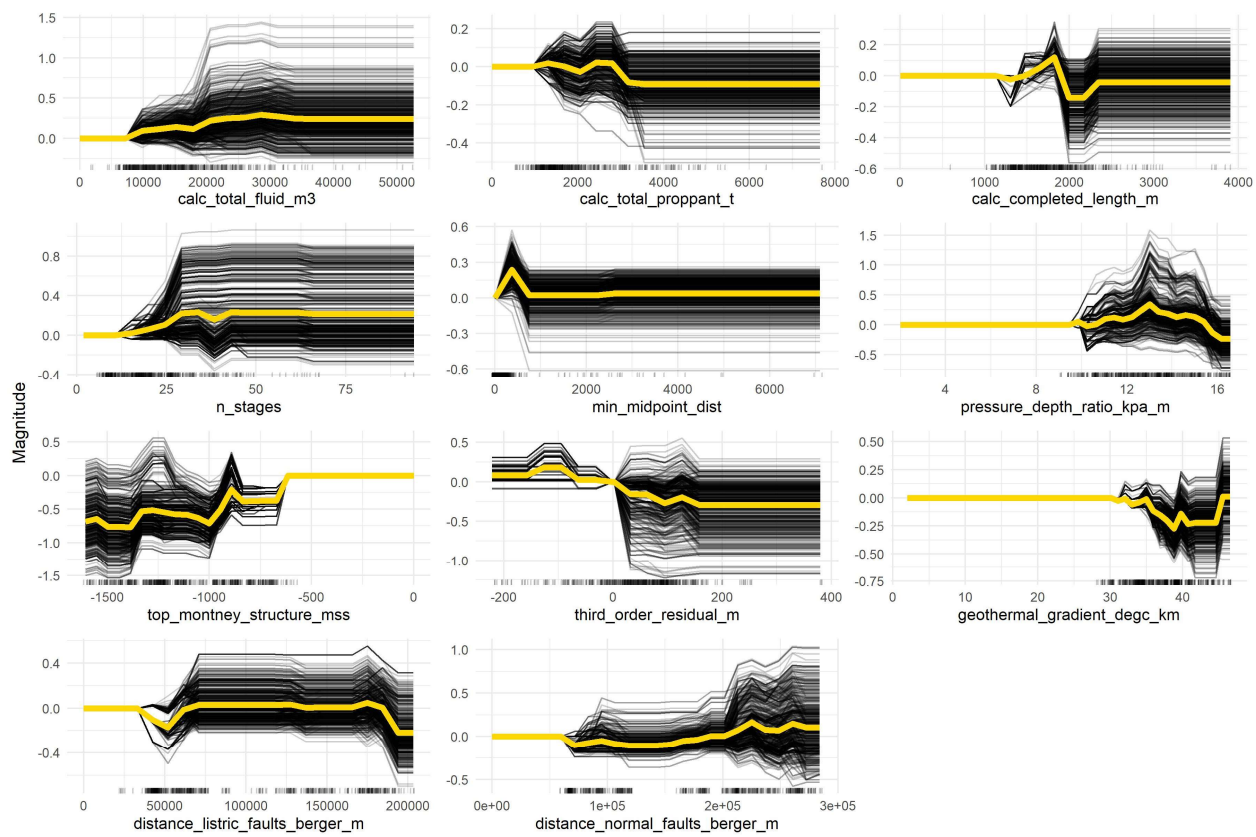


Figure 28. Partial dependence plot (yellow) and individual conditional expectation plots (black) for the magnitude regression using the XGBoost model and BC OGC completions features. The rug plot at the bottom of each PDP plot shows the data distribution of each feature.

Local Model Interpretability

Specific wells can be investigated using local model interpretability techniques. Local interpretable model-agnostic explanations (LIME) use a local surrogate model to explain individual predictions (Ribeiro et al., 2016). An interpretable model is used so that the influence of each parameter can be quantified relative to the model response. The visualization for LIME shows the effect of each feature, which is the weight of the feature times the feature value when a linear regression (i.e. GLM) is employed. This is directly analogous to the coefficient in a linear model and is subject to interpretation bias when the features are not normalized to a standard scale, which is the case for this study.

SHapley Additive exPlanations (SHAP, Shapley, 1953; Lundberg and Lee, 2017) computes feature contributions for single predictions with the Shapley value, an approach from cooperative game theory. The ‘fair value’ for each feature is compared to the difference from the well’s prediction and the average of the entire data set. The possible combinations of features are analyzed, with a weighted average used to determine the final Shapley value. The SHAP plot visualizes the contribution of each feature to the model result (ϕ), with all contributions equalling the difference between the prediction and the average of the data set.

A comparison of LIME and SHAP results for the magnitude regression in well 100/09-35-081-18W6/00 (WA 29429) is provided in Figure 29 for a MARS model. This well was classified as seismogenic and was associated with a magnitude 2.7 seismic event by Babaie Mahani et al. (2020). The figure illustrates the large discrepancy between the two interpretation methods. LIME effects often show an inverse relationship to the SHAP phi values, as is the case with the distance to faults in Figure 29. The SHAP plot shows that distance to normal faults decreases the observed magnitude, whereas the LIME Effect shows a positive coefficient. This is interpreted as indicating that the further you get from normal faults, the higher magnitude event you can expect, all else being equal. It also shows the difficulty in comparing the two interpretability approaches. LIME Effects are scale sensitive and therefore can be misleading. Also, due to the local surrogate model limitations, the LIME Effects can be highly variable and are generally considered less reliable than SHAP values. For this reason, we investigate additional example results using SHAP alone, although LIME plots are provided in Appendix C for the cases shown.

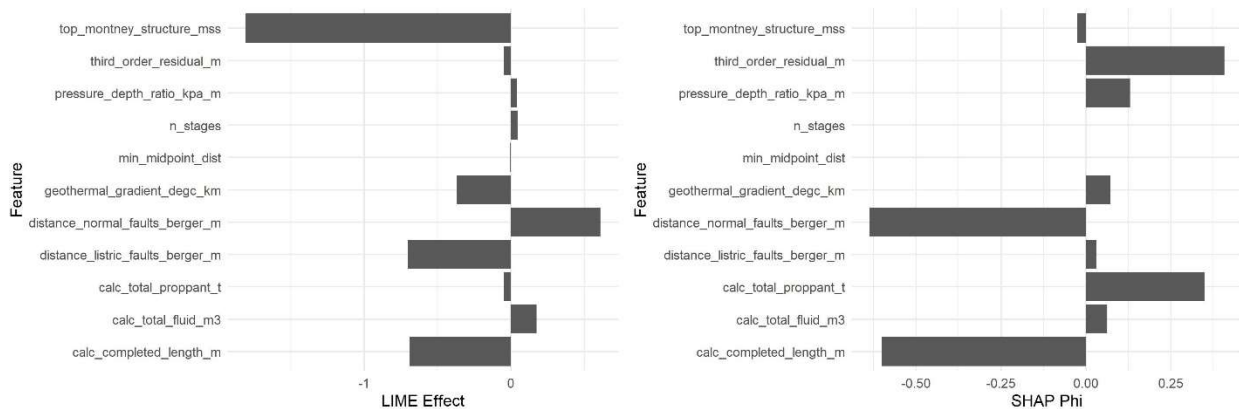


Figure 29. A comparison of LIME and SHAP plots for the magnitude regression of well 100/09-35-081-18W6/00 (WA 29429) using a MARS model with the BC OGC dataset.

The full suite of SHAP local interpretation results for well 100/09-35-081-18W6/00 is shown in Figures 30 and 31 for the seismogenic classification and magnitude regression, respectively, using both the BC OGC and geoLOGIC completions data. The results illustrate how each model encodes different features and uses multiple components to explain each individual observation. This is an indicator of feature interaction and model variance. It also shows that each individual case is driven by numerous factors and that no single factor drives either seismogenic classification or magnitude regression. Importantly, it also illustrates the different results obtained when using different data sets. The classification SHAP plots (Figure 30) suggest that the development density, the distance to listric faults, the total fluid injected, and the geothermal gradient or Paleozoic structure are seemingly most important for the positive seismogenic classification. The results for the magnitude regression (Figure 31) suggest that the number of stages, and a variety of structural geology proxies (e.g., Montney structure, third order residual and distance to faults) are the largest contributors to an increased magnitude, and several factors including, e.g., total proppant and completed length actually contribute to a decrease in predicted magnitude.

Statistical Assessment of Operational Risks for Induced Seismicity from Hydraulic Fracturing in the Montney, Northeast BC (Geoscience BC Project 2019-008) – Final Report

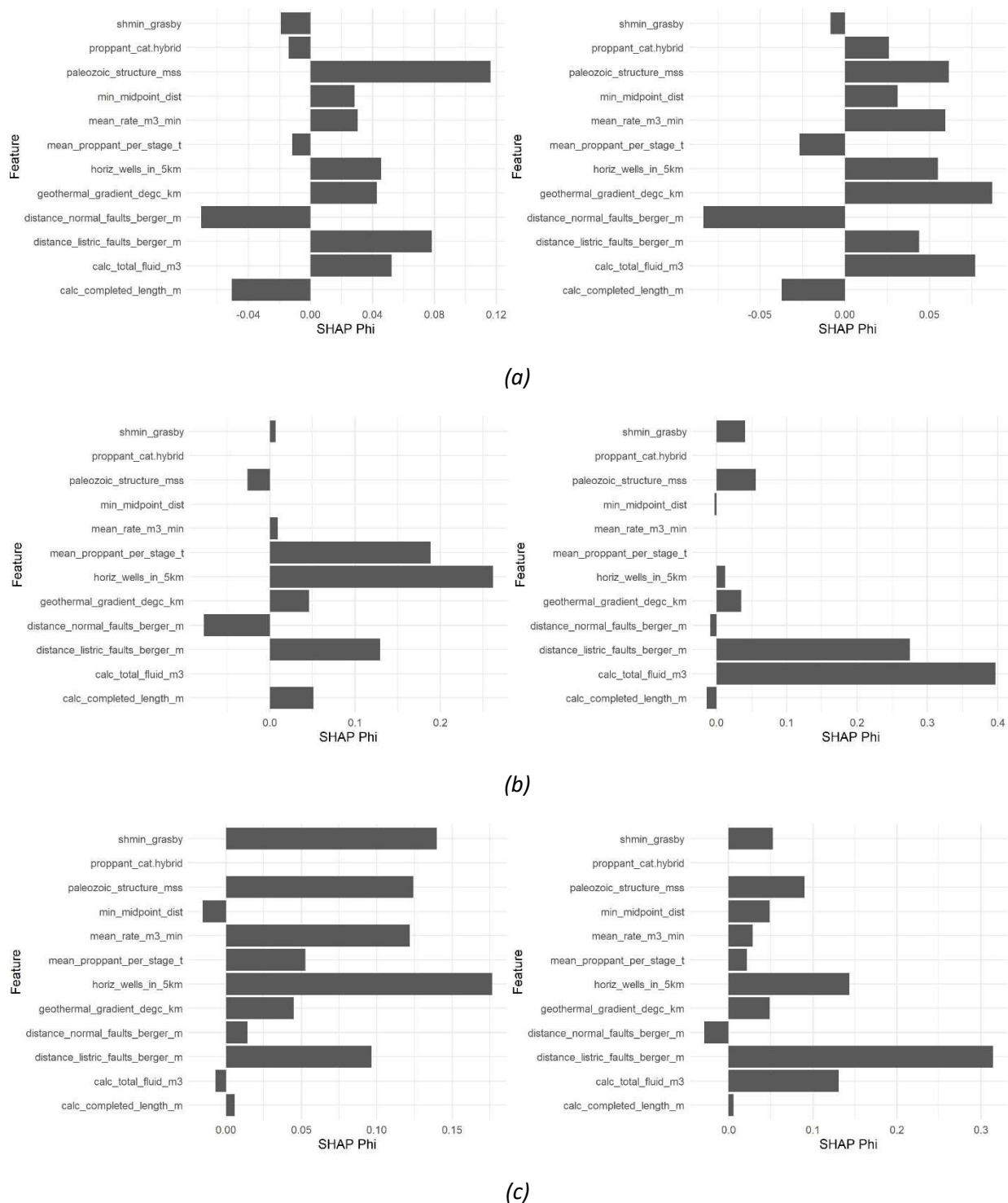


Figure 30. SHAP plots for the seismogenic classification of well 100/09-35-081-18W6/00 using the (a) GLM model, (b) CART model and (c) XGBoost model with the BC OGC (left) and geoLOGIC (right) completions data.

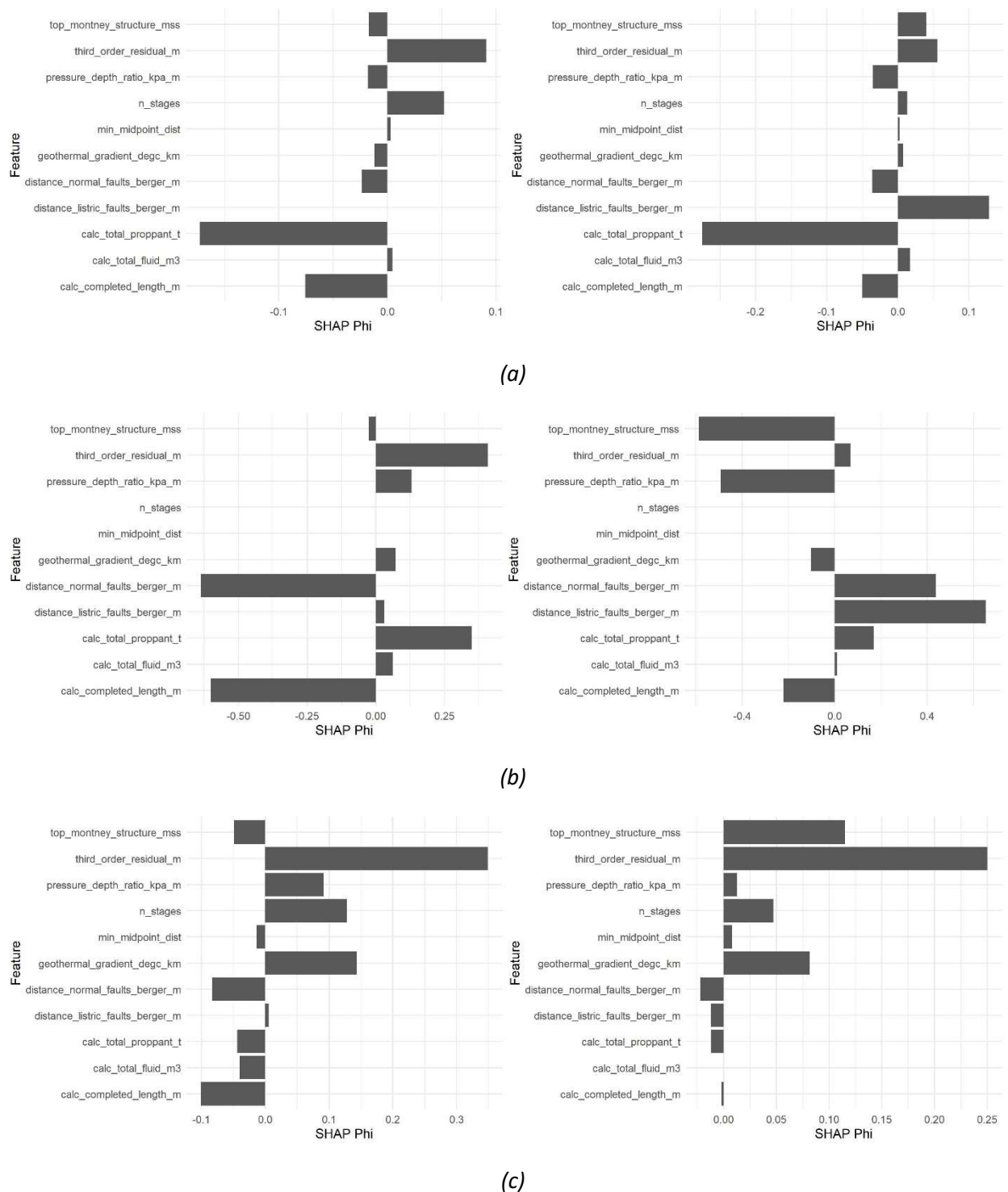


Figure 31. SHAP plots for the magnitude regression in well 100/09-35-081-18W6/00 using the (a) GLM model, (b) MARS model and (c) XGBoost model with the BC OGC (left) and geoLOGIC (right) completions data.

V. Summary, Discussion and Recommendations

In this study, 21 to 27% of the horizontal wells in the Montney are associated with anomalous induced seismicity, depending on the source of the well data (BC OGC versus geoLOGIC). This value may over-represent the percentage of wells that induce seismicity due to the duplicate assignment of a relatively large induced seismic event to more than one well. Generally, an individual well will only be associated to a handful of seismic events (<5), but up to 50 seismic events have been associated with a single well. Conversely, multiple wells (especially those drilled from the same pad) can be associated to one or two seismic events, which increases the percentage of seismogenic association.

The results of both the seismogenic classification and the magnitude regression using multiple models and two, slightly different datasets clearly illustrate that classifying wells as seismogenic and predicting induced event magnitude are highly dependent on several factors, including:

- The specific data set used, due to differences in the features present and missing values
- The model used, due to differences in model complexity and the trade-off between bias and variance
- The subset of well, completions and geological features selected for inclusion in the final models

As a result, the analysis does not single out one or more clearly causal features that are responsible for induced seismicity from hydraulic fracturing in the Montney in NEBC. In the examples shown, numerous features, both completion-related and geological, explain the results of each individual observation.

The data sets used were quite small compared to what may be considered “big data” resulting in a high risk of model overfitting, particularly in the case of the magnitude regression. Several strategies were applied to reduce overfitting. Features that provide a closer link the first principle controls on induced seismicity may potentially improve model fit and reduce confounding bias (Pearl, 2009) by focusing the model on causal/driving features instead of a proxy for those features. For example, minimum well spacing could be serving as a proxy for other, more difficult to measure, parameters such as fluid maturity, well deliverability, and reservoir containment.

In the seismogenic classification, feature interaction tends to mimic feature importance, suggesting that feature interactions contribute to a higher model variance, especially for the more complex models. The simplest model places a relatively high negative importance on minimum horizontal stress and high positive importance on geothermal gradient, distance between wells and mean proppant per stage, while the more complex models place a relatively high importance on Paleozoic structure and distance to faults. In the magnitude regression, most models show a relatively high importance for top of Montney structure and distance to faults (normal and thrust). Interactions tend to be higher for completions parameters than for geological parameters.

An attempt to interpret model results for a specific well (100/09-35-081-18W6/00) highlighted the problems of model variance and feature interaction. The classification models suggest that the development density, the distance to listric faults, the total fluid injected, and the geothermal gradient or Paleozoic structure are seemingly most important for the positive seismogenic classification. The results for the magnitude regression suggest that the number of stages, and a variety of structural

geology proxies (e.g., Montney structure, third order residual and distance to faults) are the largest contributors to an increased magnitude, and several factors including, e.g., total proppant and completed length actually contribute to a decrease in predicted magnitude.

In the feature selection process, geological features repeatedly dominated the feature rankings. In order to more clearly understand the importance of completions features, it was necessary to rerun the feature selection process excluding the geological features. The final model runs included both types of features, and the model results showed significant contributions from both. Deciding which features to carry forward into the final modeling could not be achieved by the machine learning workflow alone but required a considerable amount of human intervention. There remains a significant opportunity for the workflow to be repeated and different final feature sets to be chosen in order to see how the model results are affected.

Previous researchers have also attempted to find factors influencing induced seismicity due to hydraulic fracturing in other formations. Fasola et al. (2019) found that, in general, simultaneous completions had the highest probability of inducing seismic activity in the Eagle Ford of South Texas. Key contributors to the higher probability included effective injection rate, injected volume and number of laterals on a pad. Key geologic contributors were proximity to faults and orientation of faults relative to the stress field. Pawley et al., (2018) focused on geologic factors in the Duvernay formation in Alberta and found that the key factors contributing to the likelihood of inducing a $M > 2.5$ earthquake were Duvernay overpressure, minimum horizontal stress magnitude, proximity to Swan Hills reef margins, lithium concentration (a proxy for distance to basement) and natural seismicity rate. A comparison of these results to those presented here may indicate that predictors are formation-specific.

VI. Acknowledgements

The authors are grateful for the financial and administrative support of Geoscience BC, without which this study would not have been possible. We also want to thank geoLOGIC Systems Ltd. for allowing us access to their well completions database for the duration of the study. Comments on an early draft of this report from three anonymous reviewers greatly improved both the analysis and the final report. Finally, we thank Anton Biryukov for helpful comments on our analysis and reporting.

VII. Appendices

The following appendices are available as separate documents:

[Appendix A: Feature Definitions](#)

The filename is **2019-008 Final Report APPENDIX A – Feature Definitions.PDF**

[Appendix B: Methodology Details](#)

The filename is **2019-008 Final Report APPENDIX B – Methodology Details.PDF**

Appendix C: Full-Size Figures

This is a collection of individual files in a folder. The folder name is **2019-008 Final Report APPENDIX C – Full-size Figures**. The file names are self-explanatory.

Appendix D: Database

This is a folder containing the input data used in the analysis, provided as comma-separated value (.csv) or Microsoft Excel files. The folder name is **2019-008 Final Report APPENDIX D – Database**. The file names are self-explanatory.

Appendix E: Github Repository

The GeoscienceBC_9019-008 Github repository contains all of the data, source code and resulting output for the BC OGC completions data set only (the geoLOGIC completions data set is not distributable). The code is fully documented and includes data preparation, exploratory data analysis, and machine learning model details. The URL to the repository is https://github.com/Enlightengeo/GeoscienceBC_2019-008, and a compressed archive of the repository was delivered with this report. The online repository will be maintained for at least one year after delivery of this report.

VIII. References

Aki, K., Maximum likelihood estimate of b in the formula $\log N = a - bM$ and its confidence limits, Bulletin of the Earthquake Research Institute, Tokyo Univ., 43, 1965.

Babaie Mahani, A., Esfahani, F., Kao, H., Gaucher, M., Hayes, M., Visser, R. and S. Venables, A systematic study of earthquake source mechanism and regional stress field in the Southern Montney unconventional play of northeast British Columbia, Canada, Seismological Research Letters, v. 91(1), doi.org/10.1785/0220190230, 2020.

Barclay, J. E., Krause, F. F., Campbell, R. I. and J. Utting, Dynamic casting of a graben complex: Basin infill and differential subsidence during the Permo-Carboniferous; Peace River Embayment, Western Canada. Bulletin of Canadian Petroleum Geology, v. 38A, 1990.

Bender, B., Maximum likelihood estimation of b values for magnitude grouped data, Bulletin of the Seismological Society of America, v. 73(3), 1983.

Berger, Z., Boast, M. and M. Mushayandebvu, The contribution of integrated HRAM studies to exploration and exploitation of unconventional plays in North America: Part 1. The Peace River Arch, Canadian Society of Petroleum Geologists Reservoir, v. 35, 2008.

Berger, Z., Boast, M. and M. Mushayandebvu, The contribution of integrated HRAM studies to exploration and exploitation of unconventional plays in North America: Part 2. Basement structures control on the development of the Peace River Arch's Montney/Doig resource plays, Canadian Society of Petroleum Geologists Reservoir, v. 36, 2009.

Bischi, B., Lang, M., Kotthoff, L., Schiffner, J., Richter, Studerus, E., Casalicchio, G. and Z. M. Jones, mlr: Machine Learning in R, Journal of Machine Learning Research, v. 17(1), 2016.

Breiman, L., Random Forests, Machine Learning, v. 45(1), 2001.

British Columbia Minister of Energy, Mines and Petroleum Resources, Report of the Scientific Hydraulic Fracturing Review Panel, 2019.

British Columbia Oil and Gas Commission (BC OGC), Investigation of Observed Seismicity in the Horn River Basin, 2012.

Davies, G. R., Watson, N., Moslow, T. and J. A. MacEachern, Regional subdivisions, sequences, correlations and facies relationships of the lower Triassic Montney formation, West-central Alberta to northeastern British Columbia, Canada — With emphasis on role of paleostructure, Bulletin of Canadian Petroleum Geology, v. 66(1), 2018.

Ester, M., Kriegl, H., Sander, J. and X. Xu, A density-based algorithm for discovering clusters in large spatial databases with noise, Association for the Advancement of Artificial Intelligence, 1996.

Euzen, T., Moslow, T.F., Crombez, V. and S. Rohais, Regional stratigraphic architecture of the Spathian deposits in Western Canada – Implications for the Montney resource play, Bulletin of Canadian Petroleum Geology, v. 66(1), 2018.

Fasola, S. L., Brudinski, M. R., Skoumal, R. J., Langenkamp, T., Currie, B. S. and K. J. Smart, Hydraulic fracture injection strategy influences the probability of earthquakes in the Eagle Ford shale play of South Texas, Geophysical Research Letters, v. 26(22), doi.org/10.1029/2019GL085167, 2019.

Fereidoni, A. and L. Cui, Composite Alberta Seismicity Catalog: CASC2014-x, 2015.

Fox, A. D. and N. D. Watson, Induced seismicity study in the Kiskatinaw Seismic Monitoring and Mitigation Area, British Columbia, report to the BC Oil and Gas Commission, 2019.

Freedman, D. and P. Diaconis, On this histogram as a density estimator: L2 theory, Zeit. Wahr. ver. Geb., v. 57, 1981.

Friedman, J. H., Greedy function approximation: A gradient boosting machine, Annals of Statistics, 2001.

Goldstein, A., Kapelner, A., Bleich, J. and E. Pitkin, Peeking inside the black box: visualizing statistical learning with plots of individual conditional expectation, Journal of Computational and Graphical Statistics, v. 24(1), 2015.

Grasby, S.E., Allen, D.M., Bell, S., Chen, Z., Ferguson, G., Jessop, A., Kelman, M., Ko, M., Majorowicz, J., Moore, M., Raymond, J., and R. Therrien, Geothermal energy resource potential of Canada, Geological Survey of Canada, Open File 6914 (revised), doi:10.4095/291488, 2012.

Hahsler, M., Piekenbrock, M., Arya, S. and D. Mount, dbSCAN: Density Based Clustering of Applications with Noise (DBSCAN) and related algorithms, R package version, 2017.

Hall, M. A., Correlation-Based Feature Selection for Machine Learning, 1999.

Hayes, M., Montney formation play atlas NEBC, British Columbia Oil and Gas Commission, www.bcogc.ca/montney-formation-play-atlas-nebc, 2012.

James, G., Witten, D., Hastie, T. and R. Tibshirani, An Introduction to Statistical Learning with Applications in R, Springer, 2017.

Jolliffe, I.T., Principal Component Analysis, 2nd edition, Springer, New York, 2002.

Kao, H., Visser, R., Smith, B. and S. Venables, Performance assessment of the induced seismicity traffic light protocol for northeastern British Columbia and Western Alberta, The Leading Edge, v. 37(2), 2018.

Kohavi, R. and G. H. John, Wrappers for feature subset selection, Artificial Intelligence, 97(1-2), 1997.

Lenko, M. and S. Foster, Multivariate statistical analysis of Montney completions: Taking aim at design improvements, Unconventional Gas Technology Forum, 2016.

Lundberg, S. M. and S. I. Lee, A unified approach to interpreting model predictions, In Advances in Neural Information Processing Systems, 2017.

Molnar, C., Interpretable Machine Learning, christophm.github.io/interpretable-ml-book, 2019.

Pawley, S., Schultz, R., Playter, T., Corlett, H, Shipman, T., Lyster, S. and T. Hauck, The geological susceptibility of induced earthquakes in the Duvernay play, Geophysical Research Letters, v. 45, 2018.

Pearl, J., Causal inference in statistics: An overview, Statistics Surveys 3, 2009.

Peres-Neto, P. R., Jackson, D.A. and K. M. Somers, K.M., How many principal components? Stopping rules for determining the number of non-trivial axes revisited, Computational Statistics & Data Analysis, v. 49(4), 2005.

Ragan, D., Structural Geology: An Introduction to Geometrical Techniques, Cambridge University Press, 1973.

Ribeiro, M. T., Singh, S. and C. Guestrin, Why should I trust you?: Explaining the predictions of any classifier, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016.

Shultz, R., Atkinson, G., Eaton, D. W., Gu Y. J. and H. Kao, Hydraulic fracturing volume is associated with induced earthquake productivity in the Duvernay play, Science, v. 359, 2018.

Scott, D.W., On optimal and data-based histograms, Biometrika, v. 66, 1979.

Shapley, L. S., A value for n-person games. Contributions to the Theory of Games, 2(28), pp.307-317, 1953.

Sturges, H., The choice of a class-interval, Journal of the American Statistical Association, v. 21, 1926.

United States Geological Survey (USGS), What is a fault and what are the different types?, [/www.usgs.gov/faqs/what-a-fault-and-what-are-different-types?qt-news_science_products=0#qt-news_science_products](https://www.usgs.gov/faqs/what-a-fault-and-what-are-different-types?qt-news_science_products=0#qt-news_science_products), 2019.

Visser, R., Smith, B., Kao, H., Hutchinson, J. and J. E. McKay, A comprehensive earthquake catalogue for northeastern British Columbia and Western Alberta, 2014-2016, Geological Survey of Canada, 2017.

Wiemer, S. and M. Wyss, Minimum magnitude of completeness in earthquake catalogues: Examples from Alaska, the Western United States, and Japan, Bulletin of the Seismological Society of America, v. 90(4), 2000.