Cursed? Why One Does Not Simply Add New Data Sets to Supervised Geothermal Machine Learning Models

Stanley P. Mordensky[1], Erick R. Burns[1], John J. Lipor[2], Jacob DeAngelo[3]

[1]U.S. Geological Survey, Portland OR 97201, USA

[2]Portland State University, Portland OR 97201, USA

[3]U.S. Geological Survey, Moffett Field CA 94035, USA

Keywords

geothermal, features, supervised machine learning, play fairway analysis, PFA, dimensionality

ABSTRACT

Recent advances in machine learning (ML) identifying areas favorable to hydrothermal systems indicate that the resolution of feature data remains a subject of necessary improvement before ML can reliably produce better models. Herein, we consider the value of adding new features or replacing other, low-value features with new input features in existing ML pipelines. Our previous work identified stress and seismicity as having less value than the other feature types (i.e., heat flow, distance to faults, and distance to magmatic activity) for the 2008 USGS Geothermal Resource Assessment; hence, a fundamental question regards if the addition of new but partially correlated features will improve resulting models of hydrothermal favorability. Therefore, we add new maps for shear strain rate and dilation strain rate to fit logistic regression and XGBoost models, resulting in new 7-feature models that are compared to the old 5-feature models. Because these new features share a degree of correlation with the two least informative features from the 2008 assessment (i.e., stress and seismicity), we also consider replacement of the two lower-value features with the two new features, creating new 5-feature models.

Adding the new features improves the predictive skill of the new 7-feature model over that of the old 5-feature model, but that improvement is not statistically significant because the new features are correlated with the old features and, consequently, the new features do not present considerable new information. However, the new 5-feature XGBoost model has greater predictive skill with known positives than the old 5-feature and new 7-feature models. The improved performance of the new 5-feature model over the new 7-feature model is due to the lower-dimensional feature space of the new 5-feature model than that of the new 7-feature model. In higher-dimensional feature space, relationships between features and the presence or absence of hydrothermal systems are harder to discern (i.e., the 7-feature model likely suffers from the “curse of dimensionality”). Hence, the production of the best ML models for geothermal resource assessments is dependent upon thoughtful feature engineering and selection, consideration to which ML algorithms perform well with relatively few examples, and, perhaps, dimensionality reduction.

1. Introduction

The U.S. Geological Survey (USGS) produces periodic national geothermal resource assessments (White and Williams, 1975; Muffler, 1979; Reed, 1983; Williams and DeAngelo, 2008; Williams et al., 2008; Williams et al., 2009). The most recent moderate- to high-temperature conventional geothermal energy assessment of naturally occurring hydrothermal systems was completed in 2008 and relied upon expert decisions in an otherwise data-driven modeling process (Williams and DeAngelo, 2008; Williams et al., 2008; Williams et al., 2009). More recently, the USGS has been preparing for the next round of geothermal resource assessments by developing machine learning (ML) tools that use purely data-driven approaches and minimize the dependency on expert decisions by comparing the assessment results from 2008 to new results from ML approaches that have used the same input data (Mordensky et al., 2022, 2023). Herein, we evaluate if the addition of new data into those data-driven approaches gives improved predictions.

In the 2008 assessment, the western United States was gridded in to 725,442 2-km-by-2-km cells in which 278 cells (< 0.04% of the domain) were labeled as positive, because they contain identified hydrothermal systems (see Mordensky and DeAngelo [2023] for additional details). The remainder of the cells were classified as unlabeled. Cells that were known to not contain a hydrothermal system were classified as negative, but in practice, these labels were rarely assigned. Because conventional hydrothermal systems are sparse, classification of all cells would result in most unlabeled cells being negative (i.e., hydrothermal systems are few and far between). Each cell had five corresponding features used as input to predict hydrothermal favorability (i.e., heat flow, distance to nearest Quaternary fault, distance to nearest Quaternary magmatic activity, maximum horizontal stress, and seismic event density).

To compare the results from the 2008 assessment with results from modern data-driven ML strategies, Mordensky et al. (2023) used the same data as the 2008 assessment with seven ML approaches of increasing complexity to fit seven ML models of hydrothermal favorability. These approaches addressed challenges related to using positive-unlabeled examples (i.e., no negatives) and extreme class imbalance (i.e., an extremely high ratio of areas without hydrothermal systems to sites with hydrothermal systems; roughly 2,600:1). Because ML algorithms produce output that is not always directly comparable, a quantile-to-quantile transform was used for comparison between methods. In particular, the normal score transform was used to produce maps of hydrothermal favorability (e.g., Fig. 1). Mordensky et al. (2023) concluded that the best performing approach used XGBoost with a training strategy that downsampled testing and validation data (but not the training data) to the expected natural occurrence of hydrothermal systems (i.e., Single XGBoost in Mordensky et al. [2023]).

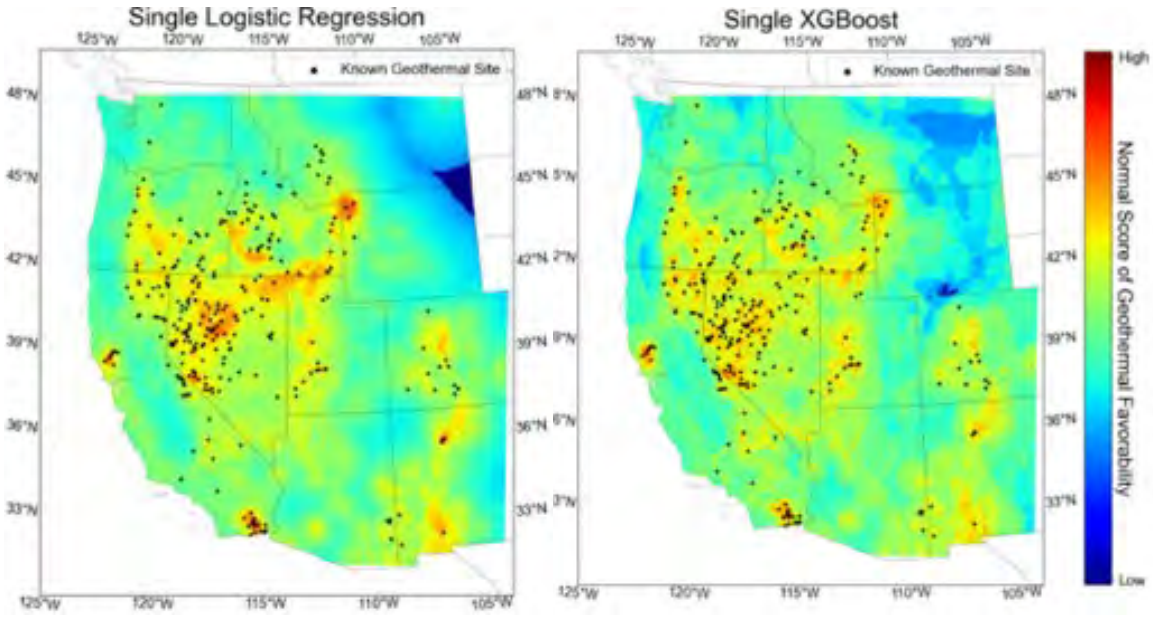


Figure 1. Hydrothermal favorability maps (Mordensky et al., 2023) from logistic regression and XGBoost using a training strategy that downsampled the testing and validation data (but not the training data) to the expected natural occurrence of hydrothermal systems to address class imbalance. The base map has been made using data from Natural Earth.

Mordensky et al. (2023) compared seven measures of feature importance to evaluate which input features were most important for making reliable hydrothermal favorability maps. Although some features ranked as comparatively unimportant across the different approaches (e.g., maximum horizontal stress and seismic event density), a key conclusion suggested that the resolution of even the most important features detrimentally influenced model performance (e.g., the most important feature, heat flow, was a smooth interpolated surface that did not represent the finer-scale natural geological conditions that control hydrothermal circulation). Mordensky et al. (2023) hypothesized that including additional and improved features would improve model performance.

We seek to understand how the selection of features affects the performance of ML models. In this study, we test the hypothesis from Mordensky et al. (2023) that supplementing the existing feature set with new features (i.e., shear strain rate and dilation strain rate) from Zeng (2022) will improve model performance. We then investigate how replacing seismic event density and stress, the least important features identified in Mordensky et al. (2023), with shear strain rate and dilation strain rate impacts model performance.

2. Methods

In order to evaluate how changes to predictive skill resulting from the addition or replacement of input features varies between different structures of supervised ML, we select two ML approaches. Both approaches use the training strategy that addresses the positive-unlabeled classifications and extreme class imbalance through downsampling the testing and validation data but not the training data to the expected natural frequency of hydrothermal systems (see Mordensky et al., 2023). We select XGBoost (Chen and Guestrin, 2016) with downsampling because this approach was the best

performing model-strategy pair in Mordensky et al. (2023). We also select logistic regression (Berkson, 1944, 1951) for comparative purposes, because logistic regression was the simplest algorithm used in Mordensky et al. (2023).

In the remainder of this section, we describe the different data set combinations used to fit the models, briefly review hyperparameter optimization, discuss the normal score transformation needed to evaluate and compare model predictions, and outline the measures of feature importance used in this study.

2.1 Data Sets

For both logistic regression and XGBoost, we train models using three different data sets (Table 1). The first, hereafter referred to as the old 5-feature model, uses the same data as Mordensky et al. (2023). These original 5 data sets are made available in Mordensky and DeAngelo (2023). The second, termed the new 7-feature model, augments the data used in Mordensky et al. (2023) with post-2008 maps of shear strain rate and dilation strain rate from Zeng (2022) with a resolution of 0.1 degrees and re-sampled to the same grid spacing as Mordensky and DeAngelo (2023). The third model, termed the new 5-feature model, replaces the least important features identified by Mordensky et al. (2023) (i.e., maximum horizontal stress and seismic event density) with the two features (i.e., shear strain rate and dilation strain rate) from Zeng (2022).

Table 1. Feature sets in each model type (old 5-feature, new 7-feature, new 5-feature).

Old 5-Feature	New 7-Feature	New 5-Feature
Heat flow	Heat flow	Heat flow
Distance to nearest fault	Distance to nearest fault	Distance to nearest fault
Distance to nearest magmatic activity	Distance to nearest magmatic activity	Distance to nearest magmatic activity
Maximum horizontal stress	Maximum horizontal stress	
Seismic event density	Seismic event density	
	Shear strain rate	Shear strain rate
	Dilation strain rate	Dilation strain rate

Due to an incomplete overlap of the data from Zeng (2022) and the 725,442 cells used in Mordensky et al. (2023), we discard 15,642 unlabeled cells (i.e., 2.2% of the study area in Mordensky et al., 2023) at the most southern extent of the western United States. Hence, the models in this study are fit from 709,800 cells, but all the original 278 known geothermal systems (e.g. positive cells) are retained.

We focus description of the feature data on the differences in the distributions of feature values between the positive and unlabeled cells using a two-sample t-test and the Pearson and the Spearman correlations. Before inspecting the statistical difference between labeled distributions

or correlation coefficients, the input data are standardized (i.e., mean = 0, standard deviation = 1) to convert feature values to the same unitless scale of similar magnitude.

2.2 Hyperparameter Optimization and Evaluating Performance

Evaluating the performance of models trained from positive-unlabeled data with extreme class imbalance has no best practice and remains an ongoing field of research. Requiring a metric by which to optimize hyperparameters (i.e., to tune model parameters that improve the predictive skill of a model), we follow the same hyperparameter optimization routine summarized in Mordensky et al. (2023). Hence, the hyperparameters (positive class weight, max depth, number of estimators, and learning rate; see Chen and Guestrin, 2016) are optimized to maximize the F1 score (Equation 1) by treating unlabeled examples as having negative labels in 120 train-test splits, in which 80% of the data are used for training and 20% of the data are used for testing, on the USGS supercomputer referred to as DENALI (Falgout et al., 2021).

$$F1\ Score = \frac{True\ Positives}{True\ Positives + \frac{1}{2}(False\ Positives + False\ Negatives)} \quad (1)$$

The F1 scores and confusion matrices for positive-unlabeled data (see Table 2) reported herein are from these 120 train-test splits. However, we acknowledge that the F1 score remains poorly suited to evaluate the performance of models trained from positive-unlabeled data because the F1 score penalizes unlabeled positives predicted as positive; hence, there are no reliable false positives. Therefore, in order to surmount challenges in evaluating models fit from positive-unlabeled data with extreme class imbalance, we follow the same practices for the evaluation of positive-unlabeled models presented in Mordensky et al. (2023) by comparing the distributions of predictions for known positives. In this method, the predictions of a model are normal score transformed before separating the predictions for the positive examples from the predictions for the unlabeled examples. The separation of the distributions allows for an examination of how different the predictions for positive labels are from the predictions for the larger set of unlabeled cells, which retain a nearly standard normal distribution (i.e., a distribution with mean = 0 and a variance = 1). Assuming the models have predictive skill, the positive labeled cells should have a distribution of predictions spanning a higher range of values than that of the unlabeled cells; hence, higher normal score transformed predictions for the positive labels result in a greater distinction between the positive labeled and unlabeled examples and indicate better model performance. We examine the distributions of predictions using cumulative distribution functions.

Table 2: Layout for positive-unlabeled confusion matrices. Rows correspond to labeled classifications. Columns correspond to predictions.

	Predicted Negatives	Predicted Positives
Unlabeled	Negative Predictions from Unlabeled Examples	Positive Predictions from Unlabeled Examples
Positives	Negative Predictions from Positive Labels	Positive Predictions from Positive Labels

2.3 Feature Importance

We evaluate the relative importance of each of the input features in making predictions by using model-agnostic (i.e., not approach-specific) measures (i.e., F1 score sensitivity, Receiver Operating Characteristic/Area Under the Curve [ROCAUC] sensitivity, and SHapely Additive exPlanation [SHAP] values) and model-gnostic (i.e., approach-specific) measures (i.e., coefficients for logistic regression and weight, gain, cover, and F score for XGBoost). Additional detail on these measures of feature importance are available in Mordensky et al. (2023).

3. Results

In this section, correlation between the input features is analyzed and predictions for hydrothermal favorability from the different models are presented through difference maps. Then, model performance is evaluated in terms of F1 scores, confusion matrices, and cumulative distribution functions of hydrothermal favorability for the known positives. Lastly, feature importance is summarized. Optimal hyperparameters are provided in Appendix A.

3.1 Data Sets

The distribution of feature values generally share two distinct qualities: 1) all features have distinct distributions of values between unlabeled and positive cells; and 2) each feature correlates moderately with at least one other feature.

The mean values between the unlabeled and positive examples are significantly different at $p < 0.001$ for every feature by measure of a two-sample t-test. Differences between the distributions of hydrothermal favorability for known positive and unlabeled examples are strongest where peaks in the distribution are distinct (Fig. 2). Where the distributions are most distinct, it can be inferred that the corresponding feature has value for separating known positive examples from unlabeled examples. By this measure, heat flow has the greatest difference between positive and unlabeled cells, followed by, in order, distance to nearest fault, shear strain rate, distance to nearest magmatic activity, maximum horizontal stress, seismic event density, and dilation strain rate.

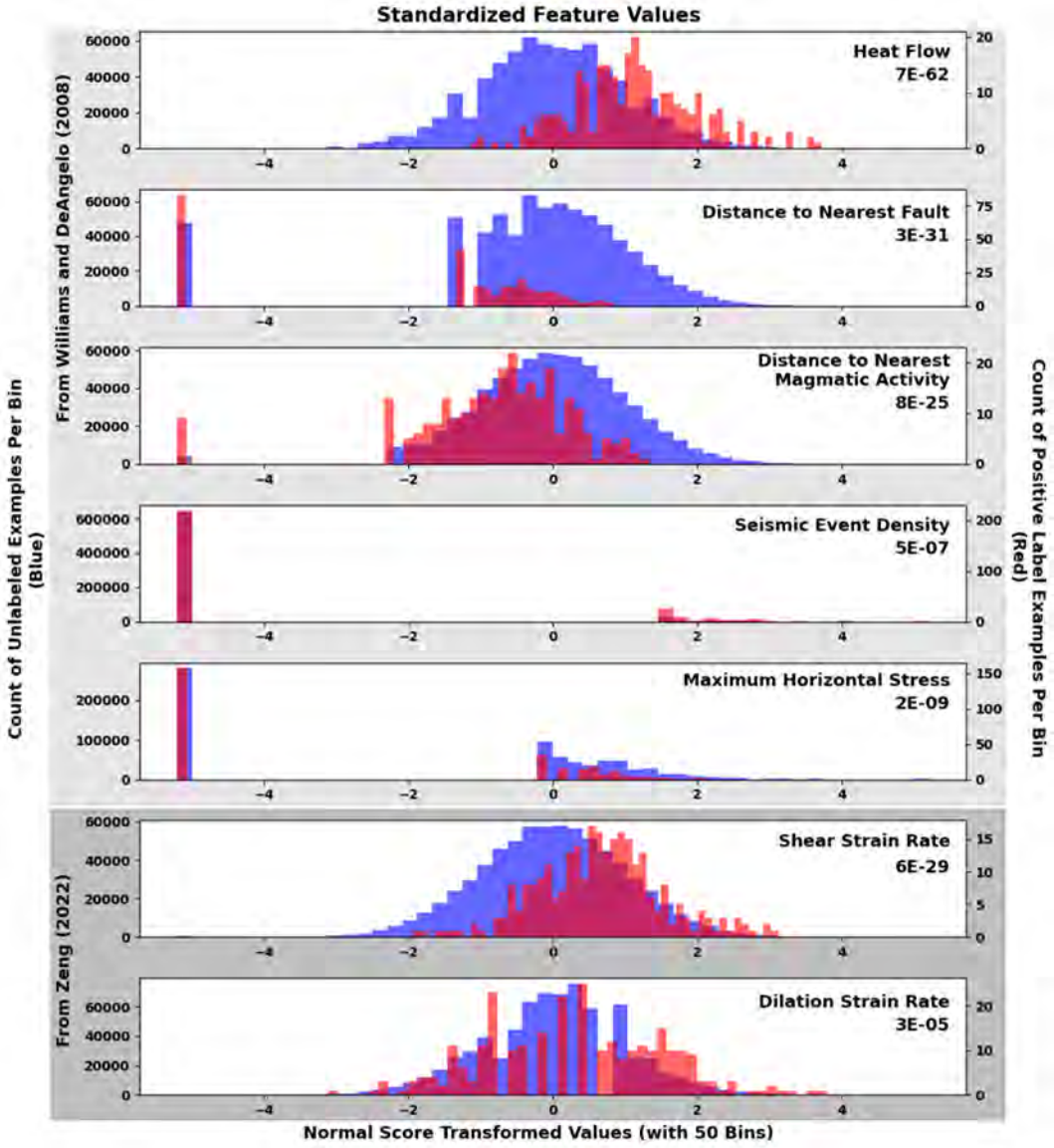


Figure 2. Histograms of standardized features. Red (right axis) represents the distribution of examples with positive labels. The numbers in the upper right of each feature pane provide the *p*-value from a two-sample t-test with a null hypothesis assuming that distributions for the known positive and unlabeled examples are indifferent. Blue represents the distribution of unlabeled examples (left axis). Purple appears when the distributions for the two classes overlap. The upper 5 features are the features used in Mordensky et al. (2023) and used to train the old 5-feature model. The bottom two features are data from Zeng (2022) and contribute toward the new 7-feature and 5-feature models.

All features share moderate correlation (i.e., an absolute correlation coefficient ≥ 0.29) with at least one other feature (Fig. 3); yet, none of the features have a strong correlation with the presence of a hydrothermal system. That is, the absolute Pearson and Spearman correlation coefficients between the labels and any feature remain ≤ 0.02 in the first columns of the correlation tables in Fig. 3. Heat flow shares an absolute correlation coefficient ≥ 0.29 with the greatest number of other

features (i.e., distance to fault, distance to magmatic activity, maximum horizontal stress, and dilation strain rate). Distance to fault and distance to magmatic activity share the greatest absolute correlation (i.e., a Pearson correlation coefficient = 0.63). Seismic event density has an absolute correlation coefficient ≥ 0.29 with the fewest number of other features (i.e., only with shear strain rate).

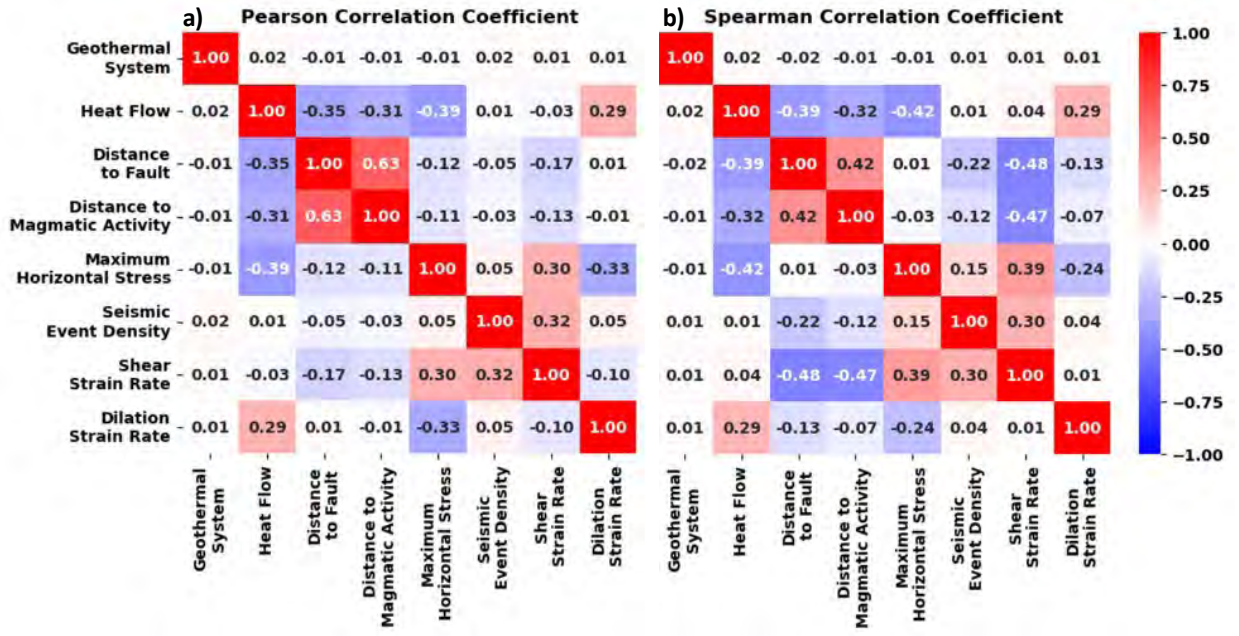


Figure 3. a) Pearson and b) Spearman correlation coefficients for label and feature pairs. Stronger red and blue colors represent stronger positive and negative correlations, respectively.

3.2 Model Predictions

A comparison of the resulting favorability maps reveals that the XGBoost predictions are more strongly affected by the addition or substitution of the new features from Zeng (2022) than the logistic regression predictions (Fig. 4). The greatest changes between the XGBoost approaches are in the Great Basin and Rio Grande Rift, whereas the greatest changes between the approaches with logistic regression are focused along the San Andreas Fault. The changes in predictions between the XGBoost approaches appear more granular than changes in predictions between the approaches using logistic regression.

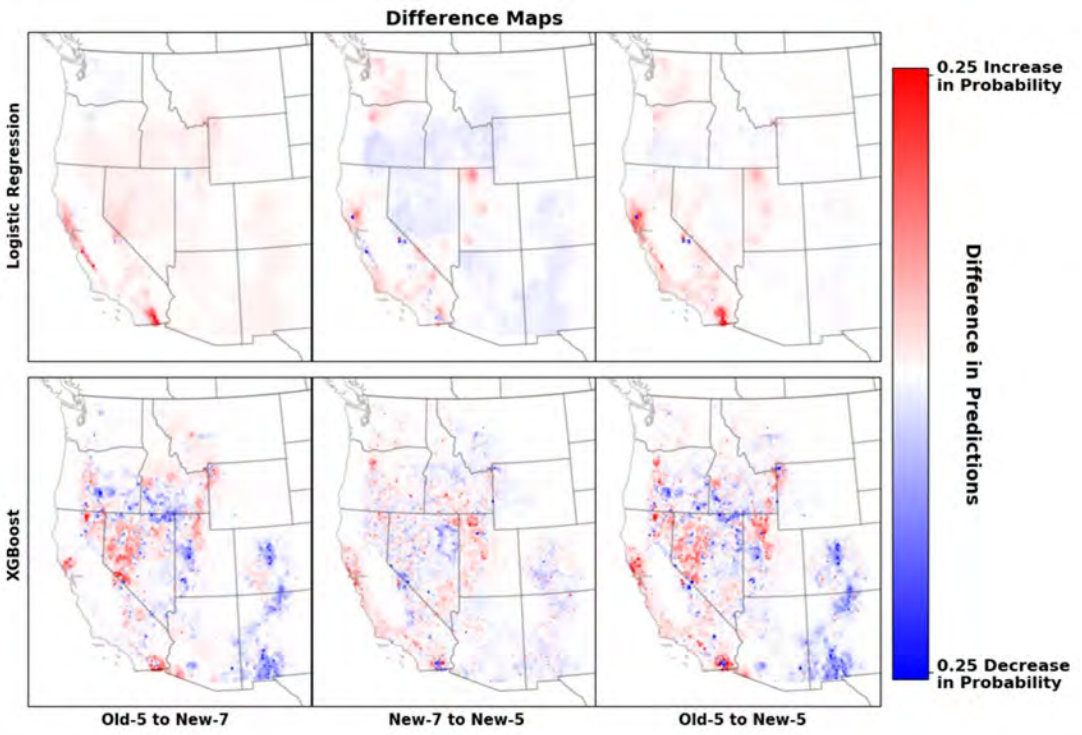


Figure 4. Difference maps comparing hydrothermal favorability from the models fit by different feature sets (see Table 1). Red shows areas with increased favorability for raw model predictions (given on a 0 to 1 scale). Blue shows areas with decreased favorability for raw model predictions. White means no difference between the model predictions. All difference maps use the same scale (-0.25 to 0.25 on the scale of raw model predictions) to allow for easier comparison between plots. The base map has been made using data from Natural Earth.

3.3 Model Performance

The relative performance of the models is similar when evaluating the confusion matrices (Table 3) and F1 scores (e.g., median F1 scores < 0.04 with a possible range of zero to one; Fig. 5); yet the predictions from new 5-feature XGBoost and old 5-feature XGBoost have a statistically significant difference in distributions at $p = 0.06$, whereas the differences in predictions from the different logistic regression models are not statistically significant at a $p < 0.20$ (Table 4). Likewise, the predictions for known positives from the new XGBoost models have higher mean and median values than the old-5-feature XGBoost model (Table 5) with the new 5-feature model having the highest predictions for known positives relative to predictions for the unlabeled cells (Fig. 6).

Table 3. Confusion matrices of training and testing data across 120 train-test splits. The bold values represent mean values. Unbolded values represent one standard deviation. See Table 2 for the layout of the positive-unlabeled confusion matrices. The different feature sets are defined in Table 1.

Logistic Regression

	Old 5-Feature		New 7-Feature		New 5-Feature	
Training	**669,175** ± 382	**2,141** ± 382	**667,969** ± 384	**3,347** ± 384	**667,866** ± 409	**3,450** ± 409
	207 ± 1	**14** ± 2	**202** ± 2	**19** ± 2	**204** ± 2	**17** ± 2
Testing	**38,081** ± 24	**123** ± 24	**38,011** ± 27	**193** ± 27	**38,086** ± 27	**198** ± 27
	52 ± 1	**3** ± 1	**51** ± 1	**4** ± 1	**51** ± 1	**4** ± 1

XGBoost

	Old 5-Feature		New 7-Feature		New 5-Feature	
Training	**669,721** ± 359	**1,595** ± 359	**668,830** ± 586	**2,486** ± 586	**668,579** ± 642	**2,737** ± 642
	202 ± 3	**19** ± 3	**198** ± 4	**23** ± 4	**190** ± 5	**31** ± 5
Testing	**38,112** ± 23	**92** ± 23	**38,0862** ± 36	**142** ± 36	**38,046** ± 40	**158** ± 40
	53 ± 1	**2** ± 1	**53** ± 1	**2** ± 1	**53** ± 1	**2** ± 1

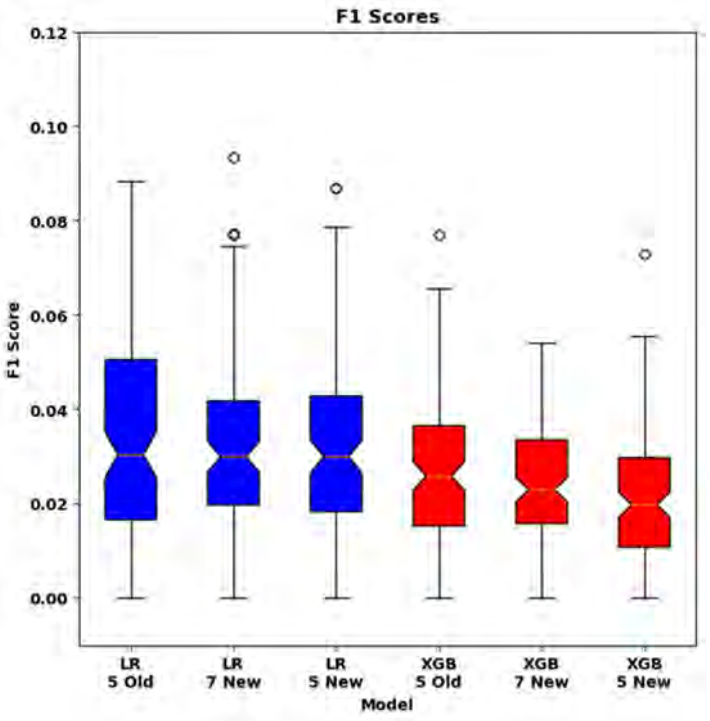


Figure 5. Box-and-whisker plots of F1 scores for test data for each ML approach from the 120 train-test splits. The logistic regression F1 scores are blue, and the XGBoost F1 scores are red. Boxes extend from the first quartile (Q1) to the third quartile (Q3) with a notch and line at the median. The whiskers extend 1.5 times the inter-quartile range (*i.e.*, 1.5 × [Q3 – Q1] while F1 score > 0). Flier points are individual points with values beyond the whiskers.

Table 4. Probability that the distributions of predictions for known positives from the models fit by different feature sets (see Table 1) do not have different means when using a two-sample t-test.

Logistic Regression	p-value	XGBoost	p-value
Old 5 vs New 7	0.22	Old 5 vs New 7	0.38
New 5 vs New 7	0.53	New 5 vs New 7	0.33
Old 5 vs New 5	0.54	Old 5 vs New 5	0.06

Table 5. Mean and median values of the distributions of normal score transformed predictions for known positive examples (shown in Fig. 6). The mean value of transformed predictions for all examples is zero. The different feature sets are defined in Table 1.

Logistic Regression	Mean	Median	XGBoost	Mean	Median
Old 5	1.32	1.25	Old 5	1.64	1.57
New 7	1.33	1.23	New 7	1.68	1.64
New 5	1.30	1.27	New 5	1.72	1.74

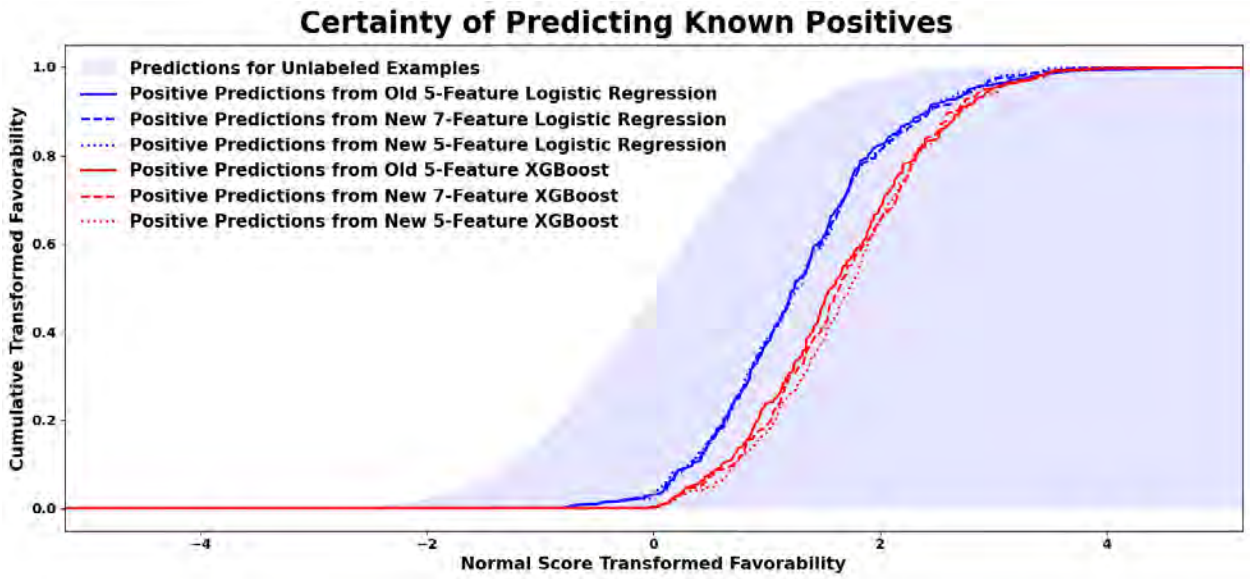


Figure 6. Comparing normal score transformed predictions for unlabeled (light blue) and known positive (blue for logistic regression and red for XGBoost) examples from the different models fit by using different feature sets (see Table 1).

3.4 Feature Importance

Several measures of feature importance are used to identify the value of each feature. In all cases (i.e., the old-5, new-7, and new-5 feature sets), heat flow and distance to nearest fault are the most important features (Fig. 7). One of the two new features (i.e., shear strain rate) has an intermediate magnitude of importance like that of distance to nearest magmatic activity, whereas the other new feature (i.e., dilation strain rate) has importance arguably similar to that of the original two least important features from the 2008 assessment (i.e., maximum horizontal stress and seismic event density).

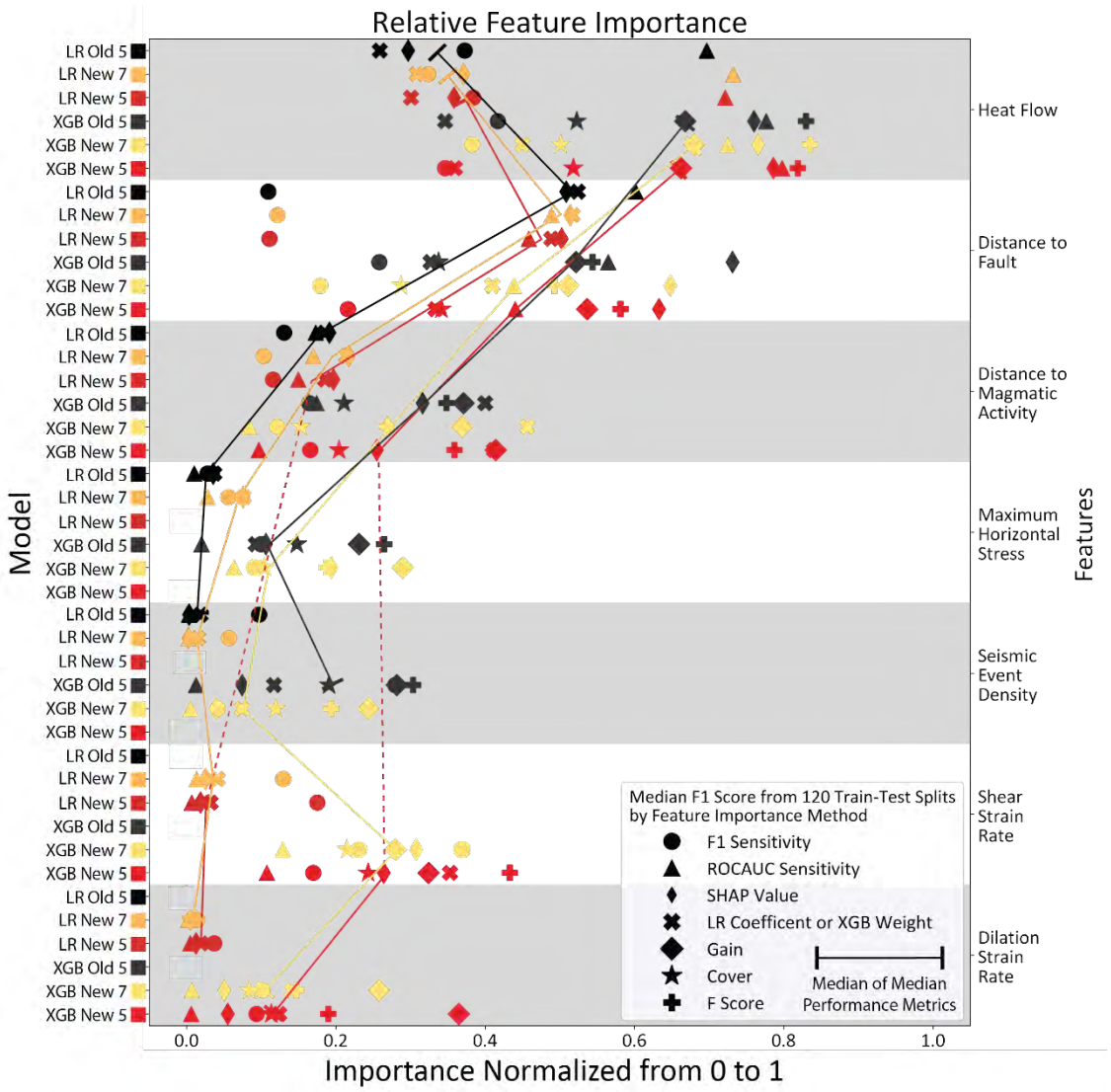


Figure 7. Feature importance from the 120 train-test splits using the different approaches. The measures of feature importance were 0-to-1 min-max normalized using the data within the interquartile range multiplied by 1.5 (i.e., by removing the flier points from the box-and-whisker plots in Appendix B before normalization) because some measures of feature importance would appear as having nearly no importance due to extreme outliers if the flier points had been retained. Solid lines connect the median value of the median measures of feature importance per approach and feature. Dashed lines span features not used in that approach. Colors correspond to the approach they depict as defined along the left y-axis. Abbreviations: LR: Logistic regression, XGB: XGBoost, ROCAUC: Area Under the Receiver Operating Characteristic Curve, SHAP: SHapely Additive explanation.

4. Discussion

In this section, we demonstrate that, while adding new features can be beneficial to model predictions, models may perform better with fewer features. The impact of adding new features on model performance is dependent upon several factors (e.g., the number of features already used to fit the model, the correlation between the new and existing features, the correlation between the new features and labels, the number of examples available for training).

Although the XGBoost models do not perform better with respect to the F1 scores (Fig. 5), the XGBoost models consistently outperform the logistic regression models when the distributions of predictions for the unlabeled and known positive examples are examined (Fig. 6); these results are consistent with Mordensky et al. (2023), who identified XGBoost as a better performing supervised ML algorithm when working with geothermal data sets that have positive-unlabeled data with few positives and extreme class imbalance. With consideration for the limitations of the F1 score when working with extreme class imbalance in positive-unlabeled data, in which all the unlabeled cells (and thereby, some potentially unlabeled positive cells) are treated as negatives to the effective detriment of the F1 score (see Equation 1), we emphasize the performance of a model by the confidence in its prediction of known positives (e.g., relatively higher predicted values for known positives than unlabeled examples; Fig. 6).

4.1 Effects Adding Replacing Features

XGBoost is better at adapting its fit to the new data than logistic regression (Figs. 4, 6; Table 4) because of its greater algorithmic architectural complexity (i.e., being a non-linear, boosted, tree-based algorithm; Chen and Guestrin, 2016). Likewise, the relative insensitivity of logistic regression to the changing feature sets stems from the simplicity of the linear algorithm (Berkson, 1944, 1951). More specifically, not all the information presented in the new features is novel as evidenced by the degree of correlation between the new and pre-existing features (Fig. 3). With the novel information being a nuanced component of the newly added data, XGBoost is better suited, to adapt its fit to this new nuanced information because of its non-linear, boosted, tree-based architecture than logistic regression. Consequently, logistic regression is unable to adapt its fit of the new models as ably as XGBoost.

Both new XGBoost models outperform the old XGBoost model, but, contrary to the idea that more input data gives better models, the new 5-feature model outperforms the new 7-feature model even though the 7-feature model is given all of the 5-feature input features and more. The improvement in predictive ability of the new 7-feature model over the old 5-feature model (Fig. 6) demonstrates that the new features contained valuable information for predicting hydrothermal resources. This conclusion is not unexpected, but the increase in performance was not statistically significant (Table 4). Yet, by removing the two least important features of the old 5-feature model (i.e., maximum horizontal stress and seismic event density; Fig. 7) from the new 7-feature model, the new 5-feature model develops a statistically significant improvement at $p = 0.06$ when predicting known positives relative to the old 5-feature model (Fig. 6; Table 4).

4.2 The Curse of Dimensionality

Increasing the predictive skill by removing less informative features suggests more features do not necessarily result in better models. Instead, these results demonstrate that models produced from geothermal data like that used in Williams and DeAngelo (2008) and Mordensky et al. (2023) are extremely susceptible to a phenomenon known as the curse of dimensionality, a term first introduced by Richard E. Bellman (Banks and Fienberg, 2003). In short, the curse of dimensionality is when the sequential addition of new features leads to a model that does a worse job at making predictions due to the intrinsic properties of ML algorithms and strategies. Technically, the curse of dimensionality means that by increasing the complexity of the feature space in which models fit decision boundaries, Euclidean distances become decreasingly effective

for properly fitting decision boundaries between classes. Practically, the curse of dimensionality means that, for a fixed number of labeled examples, the predictive skill of any model first increases as new features are added, but after a certain number of features are added, the predictive skill of the resulting model begins to deteriorate (Fig. 8; Hughes, 1968). Hence, even if new features bear novel, otherwise useful information, the resulting model may not have better performance despite the novel information contained therein. Instead, the increased dimensionality of the feature space may produce a poorer performing model.

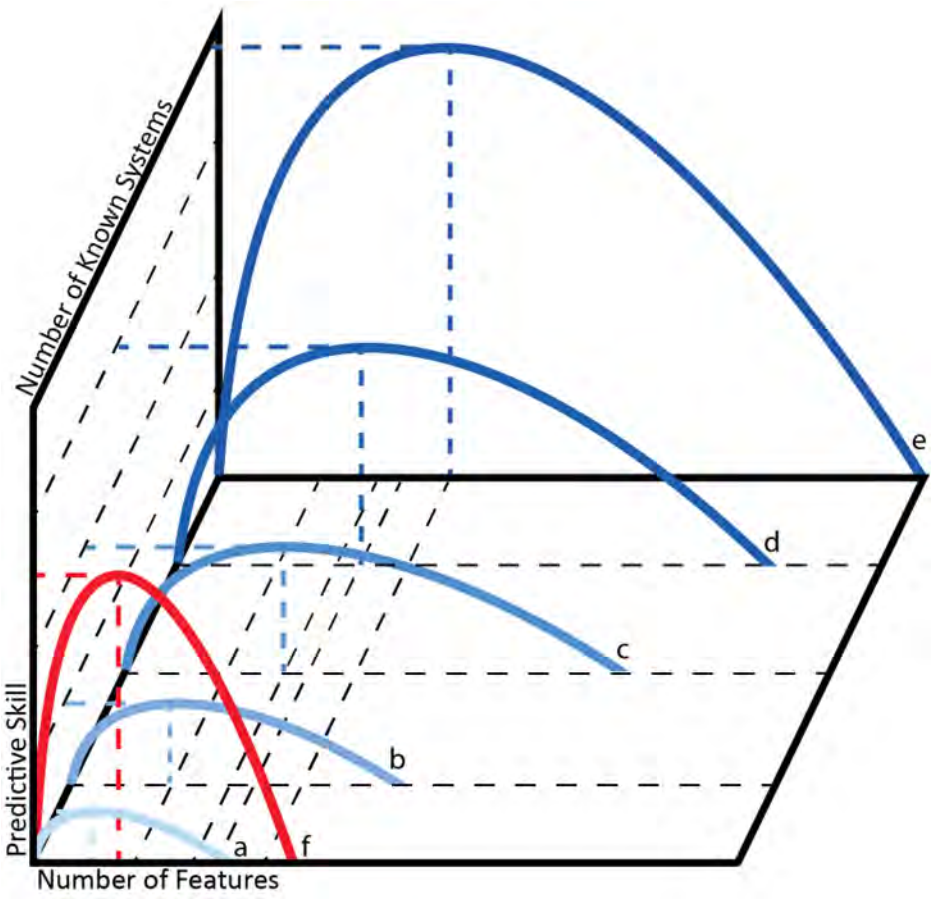


Figure 8. Number of features versus model performance. As the number of features is increased, model performance increases up to a specific number of features and, thereafter, performance decreases as new features are added per a given number of labels (modified from Figure 3 in Hughes, 1968). Additional labels allow models to fit using additional features without detrimental impact to model performance (curve series *a* through *e* [light blue to dark blue], which assume features of equal informative value). More informative features also allow models to fit using additional features without detrimental impact to model performance (e.g., this figure assumes curve *f* [red] has more informative features than curve *a* [blue], thereby allowing curve *f* [red] to fit using more features with higher performance than curve *a* [blue] per equal number of labels).

The increased performance of XGBoost resulting from removing features (e.g., maximum horizontal stress and seismic event density) is a bit unanticipated, because the curse of dimensionality is usually identified when fitting with hundreds or even thousands of features (e.g., features engineered from spectral data; see generally Verleysen and François [2005]). Additionally, tree-based algorithms, particularly boosted tree-based algorithms, like XGBoost, are generally more resilient to the curse of dimensionality than other forms of supervised learning (e.g., support-vector machines and artificial neural networks) because tree-based algorithms do not rely on a distance metric, and therefore Euclidian distances, to generalize (see generally Kuhn and Johnson, 2020).

One possible explanation for the seemingly premature commencement of the curse of dimensionality resides with how few labeled examples (i.e., 278) there are for these data. On the scale of most ML applications, having 278 labeled examples to fit a model would be considered relatively insufficient. Indeed, Mordensky et al. (2023) noted evidence of too few labeled examples with these data from observed decreasing model performance with increasing algorithmic complexity (i.e., with support-vector machines and artificial neural networks not performing as well as XGBoost). This principal observation suggested that the algorithmic complexity of support-vector machines and artificial neural networks was too great to produce models with so few labeled examples without resulting in performance-limiting bias.

The consistently moderate absolute correlation between the features (Fig. 3) is likely another factor contributing to the seemingly premature commencement of the curse of dimensionality. The moderate correlation between features means no one feature adds solely unique information. Hence, the addition of a new correlated feature increases the dimensionality of the feature space without maximizing the potential additive value of that feature.

4.3 What Can Be Done to Minimize the Curse of Dimensionality?

The improved performance of the new 5-feature model over the new 7-feature model (Fig. 6; Table 4) emphasizes the detrimental impact additional features, even if somewhat informative, might impart on model performance when working with data that have as few labeled examples as these geothermal data. In these circumstances, more features do not lead toward better predictions without careful forethought and analysis, because each new feature is inherently costly to model performance.

Although an obvious solution to the curse of dimensionality is to identify more labeled examples, there are only a limited number of known hydrothermal systems, with the process of labeling more examples being expensive if not impractical; instead, it may be more feasible to remove less informative features. Alternatively, geothermal ML studies could benefit from dimensionality reduction (e.g., principal component analysis with truncation through removing less informative components) as a solution to reduce the number of features while retaining much of the informational value of the original data given the physical underpinnings that contribute to the correlation between many of the geothermal features (e.g., heat flow and distance to nearest magmatic activity; Fig. 3).

Another practical means to offset the curse of dimensionality is to engineer highly informative features (e.g., as depicted by the red curve in Fig. 8). Heat flow and distance to faults remain the

most important features across the three XGBoost models (Fig. 7); this suggests that the new features derived from Zeng (2022) might not be the best data to differentiate between locations favorable to hydrothermal systems. It is possible that some aspects of the Zeng (2022) strain data (i.e., the broad strain values that change slowly with respect to distance) are incompatible with predicting hydrothermal systems, which are small and discrete. Producing better geothermal predictions is dependent upon engineering features with specific consideration toward the modeling framework (e.g., with consideration for grid resolution).

The USGS geothermal assessment team is working to engineer new, more informative features for the purpose of better predicting geothermal energy favorability and estimating resource potential. For example, where past heat flow maps combine conductive and convective components, DeAngelo et al. (2023) separated these signals by constructing a map of conductive heat flow and estimating the magnitude of hydrothermal convection at individual thermal gradient measurement wells. A second example of where input features can be improved is related to how fault data are used. Even though geothermal scientists have postulated fault intersection complexity controls hydrothermal upflow (e.g., Curewitz and Karson, 1997; Faulds and Hinz, 2015; Jolie et al., 2021), fault features that are typically used are distance to faults regardless of structural complexity (e.g., Williams and DeAngelo, 2008).

Perhaps the greatest opportunity for engineering highly informative features is with seismic data. Domain knowledge establishes a link between hydrothermal systems with some form of seismic expression (i.e., the presence of a hydrothermal system should influence seismic behavior in some form; e.g., Mordensky et al., 2019). The low absolute correlative relationships between seismicity and the other features (Fig. 3) suggests that this link would substantiate novel information to a model; yet, the low feature importance of seismic event density (Fig. 7) indicates that the feature is uninformative. That is, quantifying seismic events greater than M3 within 4 km is not an informative use of seismic data for predicting hydrothermal systems with logistic regression and XGBoost. Hence, we should ask: 1) how we can use seismic data differently so that seismic data may be more informative; and 2) is there another supervised ML algorithm we should consider that would make better use of this feature?

4.4 A Bias Toward More Easily Identified Systems

The label and feature data bias the models to predict more commonly identified types of hydrothermal systems more favorably than less commonly identified types of hydrothermal systems. Herein, we detail these biases.

The use of binary classification biases the predictions for hydrothermal systems toward more easily identified hydrothermal systems. In reality, every hydrothermal system is a unique combination of geological conditions, and every hydrothermal system differs by size, chemistry, and geologic characteristics. No two hydrothermal systems are identical. Hence, some systems are more obvious to identify than others; however, traditional binary ML classification treats every known hydrothermal system as being equal (i.e., a positive). By labeling all known hydrothermal systems as the same value, a model is trained to place greater emphasis on the more commonly known types of hydrothermal systems (i.e., the easier systems to identify).

The feature data also risk biasing positive predictions toward more easily identified hydrothermal systems. An underlying challenge to using ML to predict hydrothermal favorability is that every input feature is also a model based on limited input data. Any sampling biases from the collection of this limited input data propagate through the ML methods. For example, a heat flow map is an interpolated surface based on limited data, and heat flow measurements are likely preferentially made in areas where hydrothermal systems are suspected. Consequently, areas with known hydrothermal systems may be rich with data complexity, whereas areas without known hydrothermal systems may have sparse data and interpolation algorithms may oversimplify heat flow across the relatively undersampled areas.

5. Conclusion

In this study, we show that the addition of two new and arguably improved features (i.e., shear strain rate and dilation strain rate) to the machine learning (ML) approaches from Mordensky et al. (2023) result in only small improvements for models of hydrothermal favorability, because the new features are less important than several of the features used for the previous models. We demonstrate that the simple addition of new features may be a less effective strategy than replacement of relatively uninformative features with improved features due to the curse of dimensionality. We also demonstrate the choice of ML algorithm (e.g., logistic regression versus XGBoost) plays a deciding role regarding how and to what extent new data sets inhibit or augment the performance of new models. To substantially improve hydrothermal favorability predictions, we need to identify and engineer features that provide new information to our models (e.g., information that is uncorrelated with existing data, but is correlated with hydrothermal convection), choose ML algorithms that are resistant to the curse of dimensionality, and consider modeling frameworks that harness the informational value of high-resolution data (e.g., by using finer-resolution grids or abandon grids entirely).

Acknowledgements

This work was supported by the U.S. Department of Energy's Office of Energy Efficiency and Renewable Energy (EERE), Geothermal Technologies Office (GTO) under Contract No. DEAC02-05CH11231 with Lawrence Berkeley National Laboratory, Conformed Federal Order No. 7520443 between Lawrence Berkeley National Laboratory and the U.S. Geological Survey (Award Number DE-EE0008105), and Standard Research Subcontract No. 7572843 between Lawrence Berkeley National Laboratory and Portland State University. Support for Cary Lindsey was provided by the U.S. Department of Energy's Office of Energy Efficiency and Renewable Energy (EERE) under the Geothermal Technologies Office, under Award Number DE-EE0008762. Additional support for John Lipor was provided by the National Science Foundation awards NSF CRII CIF-1850404 and NSF CAREER CIF-2046175. Support for Jake DeAngelo and Erick Burns was provided by the U.S. Geological Survey Energy Resources Program. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government. We thank the USGS Advanced Research Computing (ARC) group for their assistance in using YETI and DENALI. In particular, we thank Nathan (Leon) Foks for dedicating several hours of his time to help parallelize our serial scripts. We extend our thanks to Jeff Pepin and Paul Stackelberg for their review of this proceedings paper.

Appendix A – Hyperparameters

The selected hyperparameters for all three dataset combinations (Table A1) are within one standard deviation of their corresponding equivalents for the same algorithm.

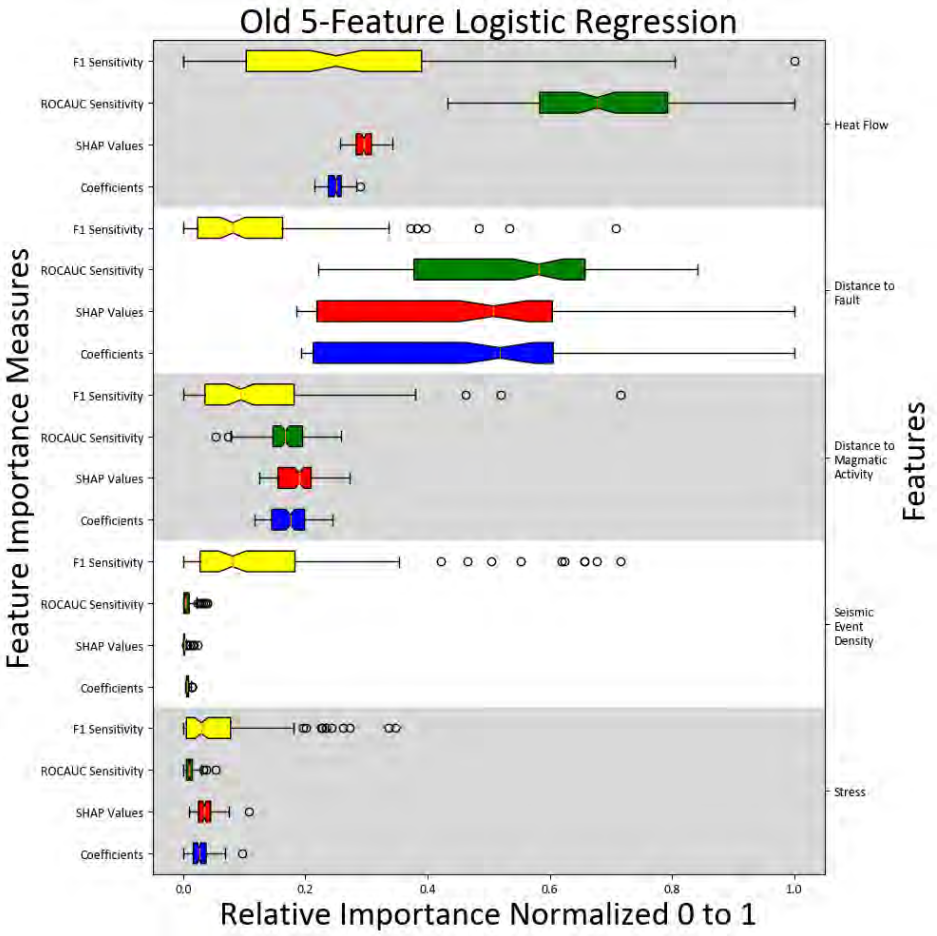
Table A1. Average hyperparameters (± 1 standard deviation) from 120 train-test splits using different feature sets. Weight corresponds to positive class weight. C corresponds to inverse regularization strength. n Estimators corresponds to the number of estimators (i.e., trees). Max depth refers to the maximum depth of the estimators.

Logistic Regression	Weight	C
Old 5	303 ± 84	0.16 ± 0.34
New 7	333 ± 86	0.07 ± 0.23
New 5	319 ± 87	0.12 ± 0.31

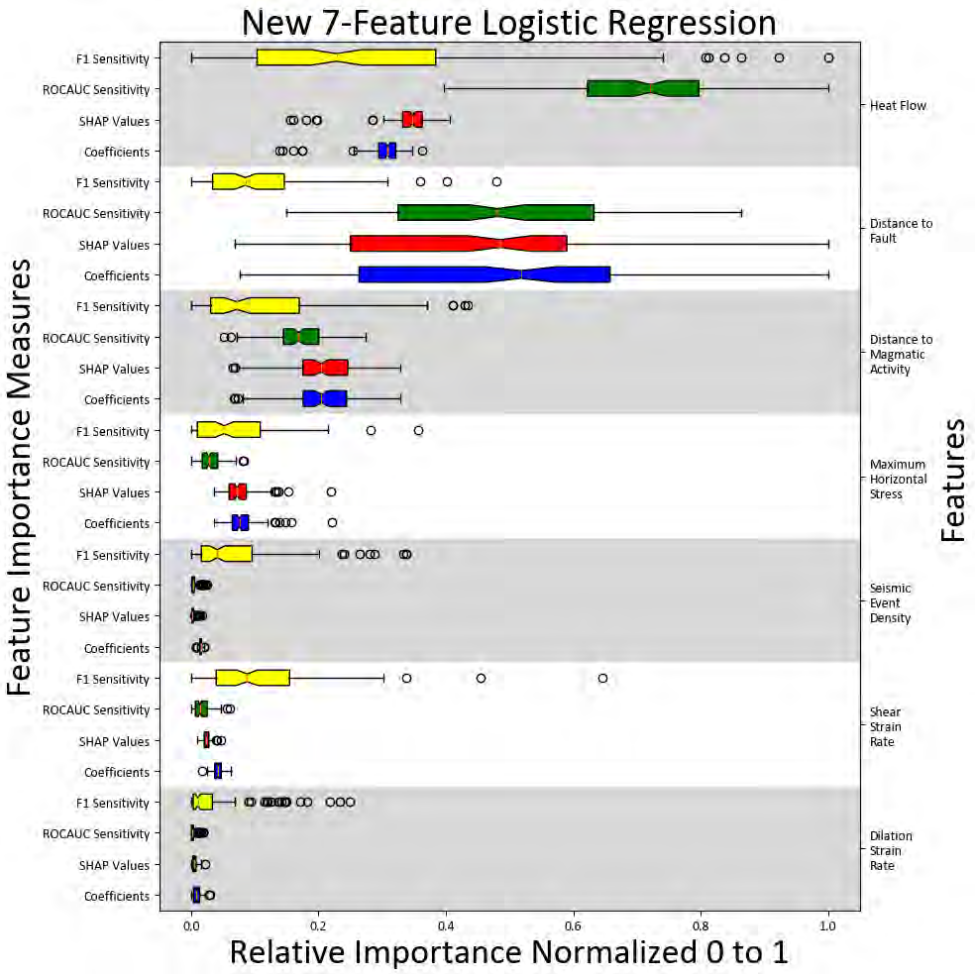
XGBoost	Weight	Max Depth	*n* Estimators	Learning Rate
Old 5	205 ± 18	2 ± 1	63 ± 20	0.22 ± 0.10
New 7	206 ± 20	3 ± 1	63 ± 19	0.21 ± 0.09
New 5	211 ± 19	2 ± 1	68 ± 21	0.25 ± 0.10

Appendix B – Feature Importance

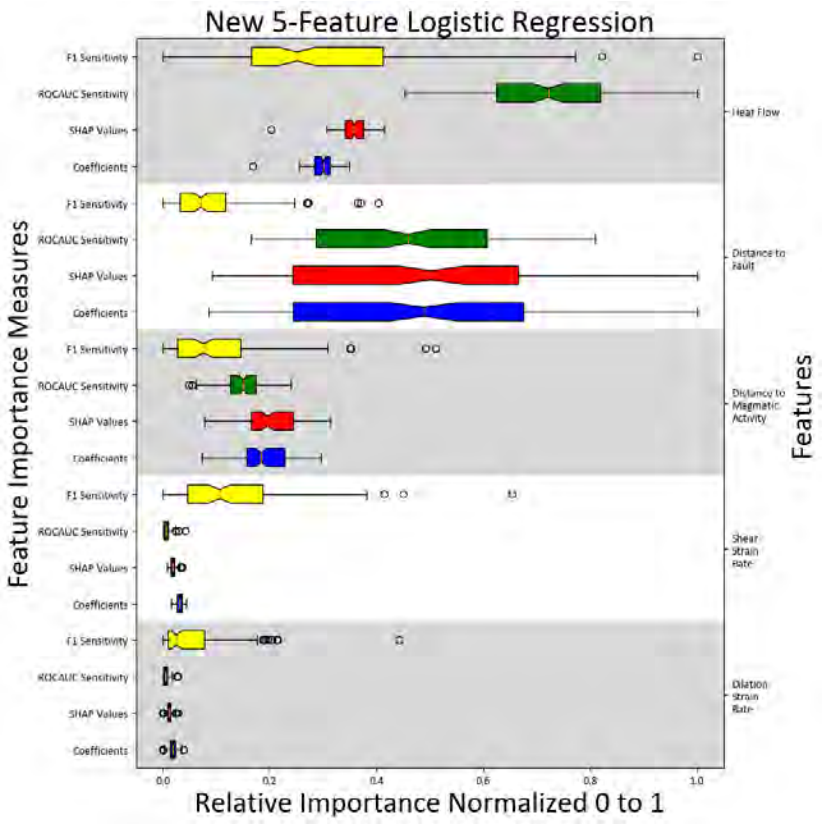
Figures B1 – B6 provide the distribution of relative feature importance from the 120 train-test splits min-max normalized from 0 to 1 for the different training strategy-algorithm approaches.



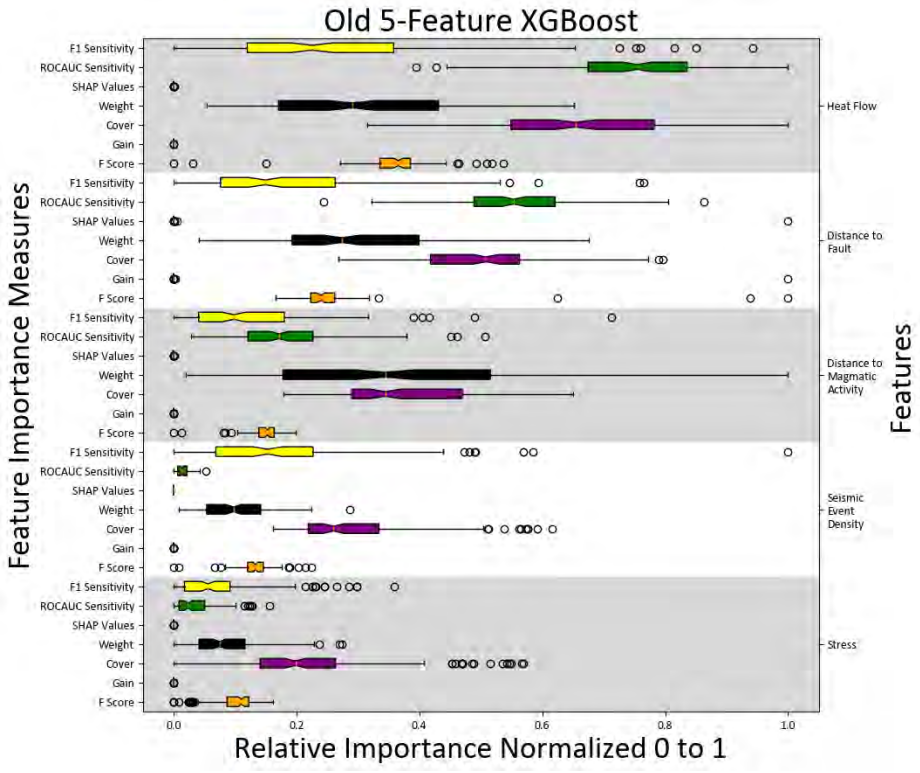
Figures B1. Normalized feature importance values for old 5-feature logistic regression. Abbreviations: SHAP: SHapely Additive exPlanation, ROCAUC: Area Under the Receiver Operating Characteristic Curve.



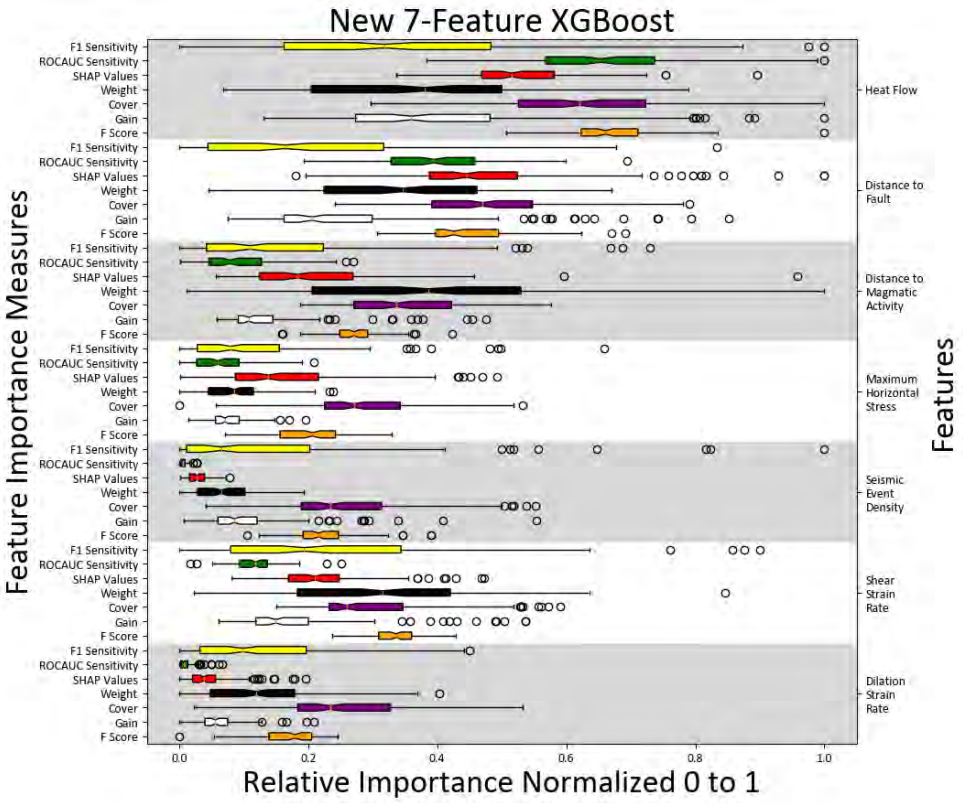
Figures B2. Normalized feature importance values for new 7-feature logistic regression. Abbreviations: SHAP: SHapely Additive exPlanation, ROCAUC: Area Under the Receiver Operating Characteristic Curve.



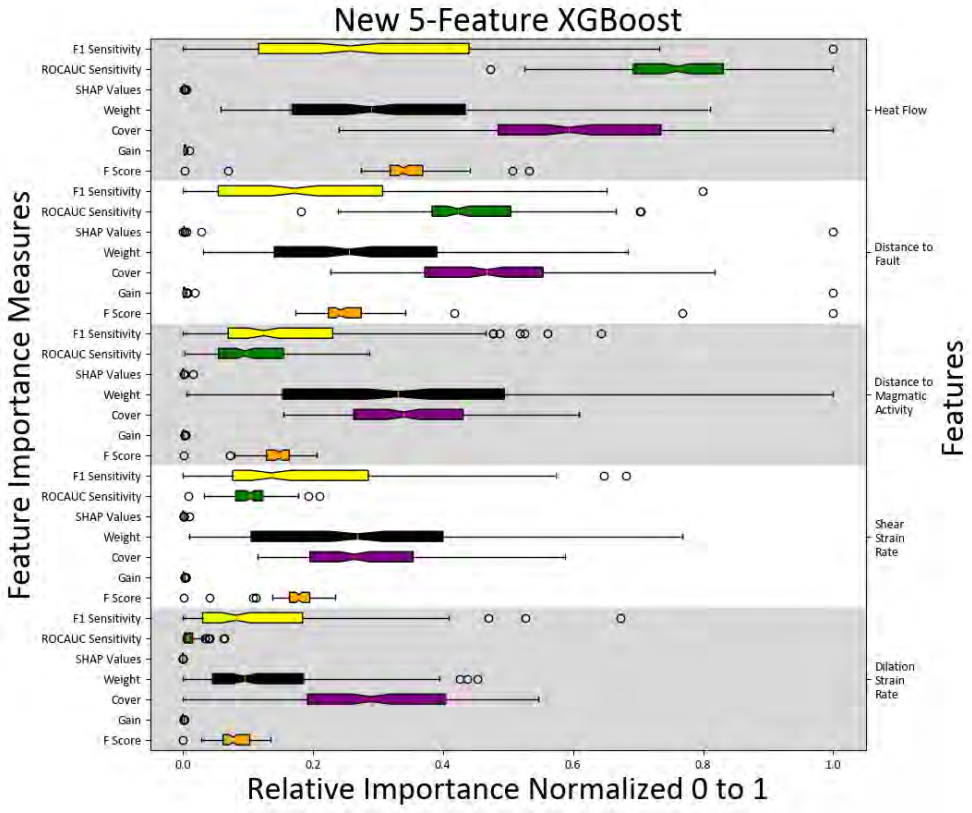
Figures B3. Normalized feature importance values for new 5-feature logistic regression. Abbreviations: SHAP: SHapely Additive exPlanation, ROCAUC: Area Under the Receiver Operating Characteristic Curve.



Figures B4. Normalized feature importance values for old 5-feature XGBoost. Abbreviations: SHAP: SHapely Additive exPlanation, ROCAUC: Area Under the Receiver Operating Characteristic Curve.



Figures B5. Normalized feature importance values for new 7-feature XGBoost. Abbreviations: SHAP: SHapely Additive exPlanation, ROCAUC: Area Under the Receiver Operating Characteristic Curve.



Figures B6. Normalized feature importance values for new 5-feature XGBoost. Abbreviations: SHAP: SHapely Additive exPlanation, ROCAUC: Area Under the Receiver Operating Characteristic Curve.

REFERENCES

Banks, D. L., & Fienberg, S. E. (2003). Data Mining, Statistics. In R. A. Meyers (Ed.), *Encyclopedia of Physical Science and Technology (Third Edition)* (pp. 247-261): Academic Press.

Berkson, J. (1944). Application of the Logistic Function to Bio-Assay. *Journal of the American Statistical Association, 39*, 357-365.

Berkson, J. (1951). Why I Prefer Logits to Probits. *Biometrics, 7*, 327.

Chen, T., & Guestrin, C. (2016). *XGBoost: A Scalable Tree Boosting System.* Paper presented at the Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA.

Curewitz, D., & Karson, J. A. (1997). Structural settings of hydrothermal outflow: Fracture permeability maintained by fault propagation and interaction. *Journal of Volcanology and Geothermal Research, 79*(3-4), 149–168. doi:10.1016/S0377-0273(97)00027-9

DeAngelo, J., Burns, E. R., Gentry, E., Batir, J. F., Lindsey, C. R., & Mordensky, S. P. (2023). *New Maps of Conductive Heat Flow in the Great Basin, USA: Separating Conductive*

and Convective Influences. Paper presented at the 48th Workshop on Geothermal Reservoir Engineering, Stanford, CA, USA.

Falgout, J. T., Gordon, J., Williams, B., & Davis, M. J. (2021). *SGS Advanced Research Computing, USGS Denali Supercomputer*. U.S. Geological Survey.

Faulds, J. E., & Hinz, N. H. (2015). *Favorable Tectonic and Structural Settings of Geothermal Systems in the Great Basin Region , Western USA: Proxies for Discovering Blind Geothermal Systems*. Paper presented at the World Geothermal Congress 2015, Melbourne, Australia.

Hughes, G. (1968). On the mean accuracy of statistical pattern recognizers. *IEEE Transactions on Information Theory, 14*(1), 55-63. doi:10.1109/TIT.1968.1054102

Jolie, E., Scott, S., Faulds, J. E., Chambefort, I., Axelsson, G., Gutiérrez-Negrín, L. C., . . . Zemedkun, M. T. (2021). Geological controls on geothermal resources for power generation. *Nature Reviews Earth and Environment, 2*, 324-339. doi:10.1038/s43017-021-00154-y

Kuhn, M., & Johnson, K. (2020). Effect of Irrelevant Features. In *Feature Engineering and Selection: A Practical Approach for Predictive Model* (1st ed.). Boca Raton, FL, USA: CRC Press.

Mordensky, S. P., & DeAngelo, J. (2023). Geothermal Resource Favorability: Select Features and Predictions for the Western United States Curated for DOI 10.1016/j.geothermics.2023.102662 , U.S. Geological Survey data release, https://doi.org/10.5066/P9V1Q9XM

Mordensky, S. P., Heap, M. J., Kennedy, B. M., Gilg, H. A., Villeneuve, M. C., Farquharson, J. I., & Gravley, D. M. (2019). Influence of alteration on the mechanical behaviour and failure mode of andesite: implications for shallow seismicity and volcano monitoring. *Bulletin of Volcanology, 81*. doi:10.1007/s00445-019-1306-9

Mordensky, S. P., Lipor, J. J., DeAngelo, J., Burns, E. R., & Lindsey, C. R. (2022). *Predicting Geothermal Favorability in the Western United States by Using Machine Learning: Addressing Challenges and Developing Solutions*. Paper presented at the 47th Stanford Geothermal Workshop, Stanford, California (Virtual).

Mordensky, S. P., Lipor, J. J., DeAngelo, J., Burns, E. R., & Lindsey, C. R. (2023). When Less Is More: How Increasing the Complexity of Machine Learning Strategies for Geothermal Energy Assessments May Not Lead toward Better Estimates. *Geothermics, 110*, 102662. doi:10.1016/j.geothermics.2023.102662

Muffler, L. P. J. (1979). Assessment of geothermal resources of the United States-1978. *U.S. Geological Survey Circular 790*, 163. doi:10.3133/cir790

Natural Earth. (2023). August, 24, 2023. www.naturalearthdata.com

Reed, M. J. (1983). Assessment of low-temperature geothermal resources of the United States-1982. *Geological Survey Circular 892*, 73.

Verleysen, M., & François, D. (2005). The Curse of Dimensionality in Data Mining and Time Series Prediction. In *Lecture Notes in Computer Science* (Vol. 3512, pp. 758-770).

White, D. E., & Williams, D. L. (1975). Assessment of geothermal resources of the United States. *U.S. Geological Survey Circular 726*, 155.

Williams, C. F., & DeAngelo, J. (2008). Mapping Geothermal Potential in the Western United States. *GRC Transactions, 32*, 181-188.

Williams, C. F., Reed, M. J., DeAngelo, J., & Galanis, S. P. (2009). Quantifying the undiscovered geothermal resources of the United States. *Transactions, 33*, 882-889.

Williams, C. F., Reed, M. J., Mariner, R. H., DeAngelo, J., & Galanis, S. P. (2008). Assessment of Moderate-and High-Temperature Geothermal Resources of the United States. *U.S. Geological Survey Fact Sheet 2008-3082*, 1-4.

Zeng, Y. (2022). GPS Velocity Field of the Western United States for the 2023 National Seismic Hazard Model Update. *Seismological Research Letters*, *93*(6), 3121-3134. doi:10.1785/0220220180