Don’t Let Negatives Hold You Back: Accounting for Underlying Physics and Natural Distributions of Hydrothermal Systems When Selecting Negative Training Sites Leads to Better Machine Learning Predictions

Pascal D. Caraccioli[1,2], Stanley P. Mordensky[1], Cary R. Lindsey[1,4], Jacob DeAngelo[3], Erick R. Burns[1], John J. Lipor[2]

[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

[4]Great Basin Center for Geothermal Energy, Nevada Bureau of Mines and Geology, University of Nevada, Reno NV 89557, USA

Keywords

machine learning, geothermal, Nevada, class imbalance, positive-unlabeled classification, XGBoost, geothermal favorability, play fairway analysis

ABSTRACT

Selecting negative training sites is an important challenge to resolve when utilizing machine learning (ML) for predicting hydrothermal resource favorability because ideal models discriminate between hydrothermal systems (positives) and all types of locations without hydrothermal systems (negatives). The Nevada Machine Learning project (NVML) fit an artificial neural network to identify areas favorable for hydrothermal systems by selecting 62 negative sites where the research team had confidence that no hydrothermal resources exist. We compare the implications of the expert selection of negatives (i.e., the NVML strategy) against a random sample strategy, where it is assumed that areas outside the favorable structural ellipses defined by NVML are negative. Because hydrothermal systems are sparse, it is highly probable that, in the absence of a favorable geological structure, hydrothermal favorability is low. We compare three training strategies using: 1) the positive and negative labeled examples from NVML; 2) the positive examples from NVML with randomly selected negatives in equal frequency as NVML; and 3) the positive examples from NVML with randomly selected negatives reflecting the expected natural distribution of hydrothermal systems relative to the total area. We apply these training strategies to the NVML feature data (input data) using two ML algorithms (XGBoost and logistic regression) to create six favorability maps for hydrothermal resources. When accounting for the expected natural

distribution of hydrothermal systems, we find that XGBoost performs better than the NVML neural network and its negatives. Model validation was less reliable using F1 scores, a common performance metric, than comparing probability estimates at known positives, likely because of the extreme natural class imbalance and the lack of negatively labeled sites. This work demonstrates that expert selection of negatives for training in NVML likely imparted modeling bias. Accounting for the sparsity of hydrothermal systems and all the types of locations without hydrothermal systems allows us to create better models for predicting hydrothermal resource favorability.

1. Introduction

The mean electric output capacity for conventional hydrothermal resources (identified and undiscovered) for Nevada's portion of the Great Basin region is estimated to be ~5.7 GWe (Williams et al., 2008). Challenges associated with identifying viable geothermal resources (Richards and Blackwell, 2002) are evidenced in that only about 10% of resources have been developed (Ayling, 2020) and that many of the remaining undiscovered resources are blind (i.e., no easily identifiable surface expression; Coolbaugh et al., 2007). Research seeking to identify geothermal resources in the Great Basin has advanced from methods that relied dominantly upon expert decisions for the Nevada Play Fairway Analyses (Nevada PFA; Faulds et al., 2015a; Faulds et al., 2015b; Faulds et al., 2016; Faulds et al., 2017; Faulds et al., 2021b) to more recent data-driven approaches, like the Nevada Machine Learning Project (NVML; Brown et al., 2020; Faulds et al., 2020; Faulds et al., 2021a; Smith, 2021; Smith et al., 2023), though fundamental challenges (e.g., low number of training data for the known presence or absence of a hydrothermal system) are still being addressed by infusion of expert knowledge to aid in the selection and engineering of input features and during the careful evaluation of model results.

The NVML research team applied machine learning (ML) techniques to the dataset compiled under the Nevada PFA using an artificial neural network (ANN) to identify favorable areas for geothermal exploration (Fig. 1). In doing so, the NVML team labeled 83 known hydrothermal systems as positive and recognized that they needed negative sites for the ANN to develop a model that could separate the two classes. Hence, the NVML team selected 62 negative sites where the NVML research team had confidence that no hydrothermal resources existed.

Mordensky et al. (2023) completed a similar analysis for hydrothermal favorability of the entire western United States using the input features of Williams et al. (2008). Similarly, Mordensky et al. (2023) had known positives (278) but no known negatives. Instead of selecting known negatives, it was assumed that due to the sparseness of hydrothermal systems (i.e., having few positives with most of the area as negative), most locations other than known hydrothermal systems are negative. These other locations were sampled randomly and assigned as negative, then this process was repeated many times to make sure an unlucky random sample did not bias the final model. Mordensky et al. (2023) also accounted for class imbalance (i.e., the unequal distribution) of positive to negative hydrothermal systems. Both the random sampling of negatives and class imbalance problems were addressed by using a custom training strategy tailored for geothermal systems. The results from Mordensky et al. (2023) demonstrated that ML algorithms could be used to remove or minimize the need for expert decisions employed previously in Williams et al. (2008).

In the context of supervised ML classification, a lack of examples can potentially lead to overfitting (see generally Mordensky et al. 2022 for additional information regarding supervised ML classification with geoscience data). Overfitting occurs when the algorithm 'memorizes' the training data and performs poorly when faced with new, unseen testing data. Hence, we ask how the differences between the NVML and Mordensky et al. (2023) training strategies influence model predictions when using the PFA data, and we seek to identify the best strategy for detecting hydrothermal resources using ML techniques.

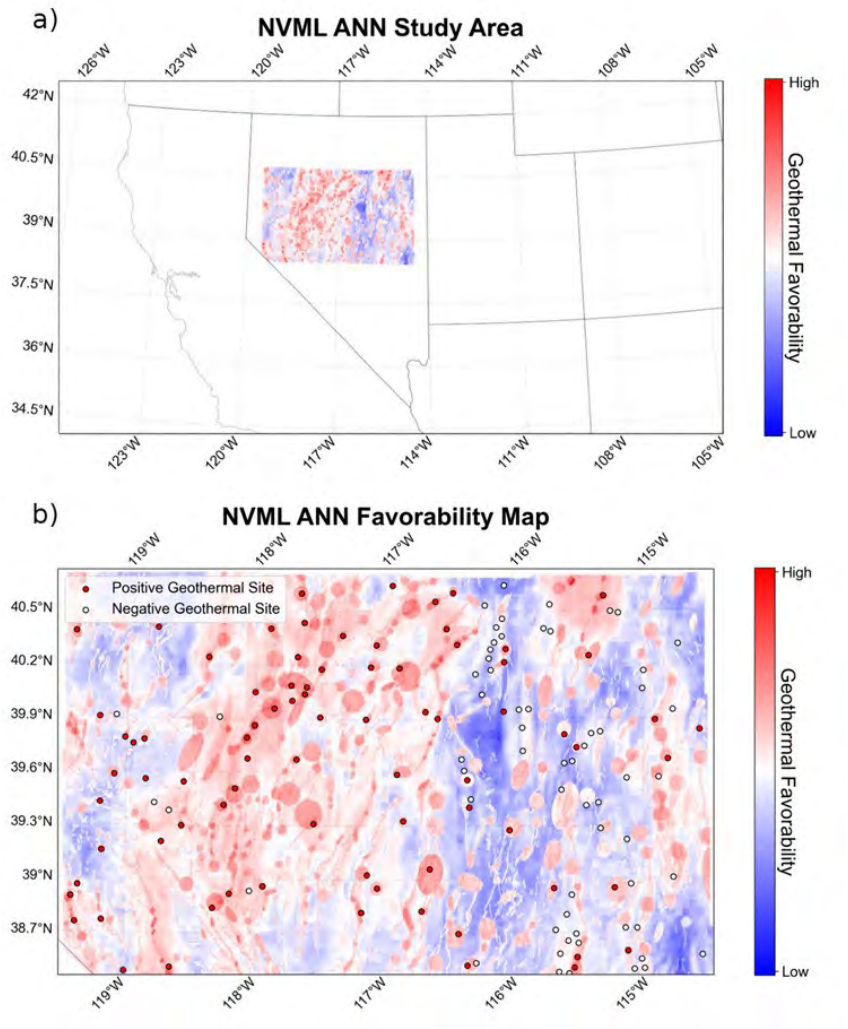

Figure 1: Nevada Machine Learning project (NVML) a) study area and b) favorability map. White points are negative training/testing examples defined through expert decisions. Red points are the positive examples (i.e., known hydrothermal systems). Favorability scores from the artificial neural network (ANN) have been normal score transformed for easy comparison with the results below from our study. The base map has been made using data from Natural Earth.

2. Methods

We conduct our comparison of the NVML and Mordensky et al. (2023) training strategies using two supervised ML algorithms: logistic regression and eXtreme Gradient Boosting (XGBoost). Logistic regression is a simple, linear classification algorithm that produces a probability value for every example and then classifies the example as positive or negative using a decision threshold (commonly chosen to be 0.5; Berkson, 1944; Berkson, 1951). XGBoost is a non-linear, boosted decision-tree approach, in which a series of decision trees (i.e., estimators) produce a probability for each example and then classify the example using a decision threshold (again, commonly chosen to be 0.5; Chen and Guestrin, 2016). Because logistic regression, XGBoost, and ANNs all create probabilities differently and in ways that may not be directly comparable, all favorability maps shown herein are the normal score transform (see generally Pyrcz and Deutsch, 2018) of the model-specific predicted probabilities. The post-transformed prediction values are shown as geothermal favorability. Because a normal score transform is a quantile-to-quantile transform, most and least favorable parts of a map occupy the same total area, allowing for an easy comparison where methods agree about the most favorable locations.

For the remainder of this section, we briefly describe the data processing, exploratory data analysis, training strategies, and hyperparameter optimization.

2.1 Data Processing and Exploratory Data Analysis

NVML gridded 96,000 km^2 of the Great Basin into 1,728,000 250-m-by-250-m cells, in which 83 cells contain a known hydrothermal system and were labeled as positive (Smith et al., 2021); 62 cells were labeled as negative by expert examination. The remaining cells were unlabeled. The labeled and unlabeled cells were populated with data from 11 features (i.e., independent input datasets). Ten of the 11 features were interpolated maps of the following geologic properties: strain rate, elevation, distance to nearest Quaternary fault, fault recency, horizontal gravity gradient, seismic density, heat flow, magnetic field gradient, slip rate, slip and dilation tendency (Fig. 2). The eleventh feature was crafted by defining ellipses that contain geologic structures deemed by experts to be favorable for the occurrence of hydrothermal systems (i.e., known favorable structural setting; Fig. 3; Faulds et al., 2016; Brown et al., 2020).

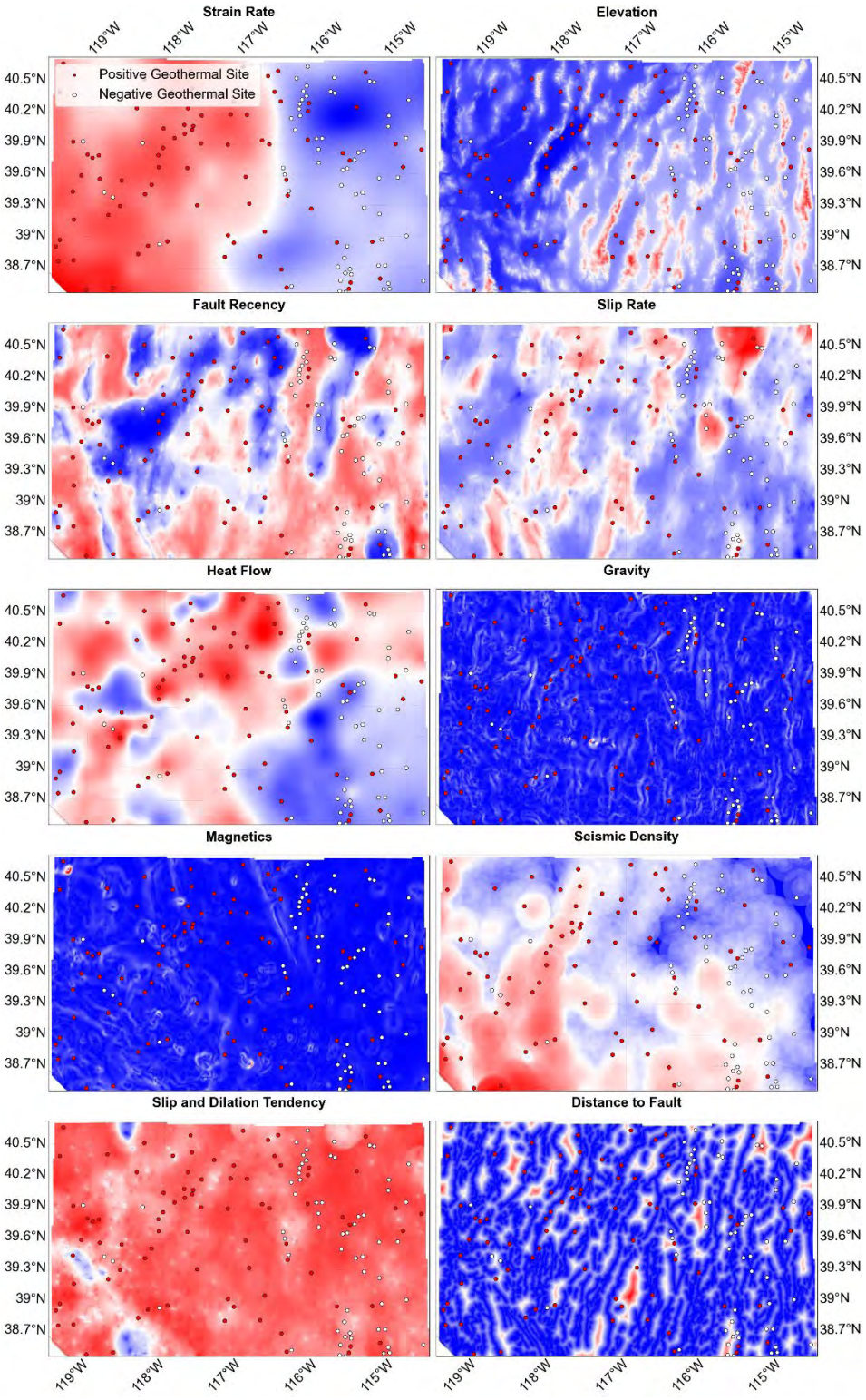

Figure 2: Maps of the ten Nevada Machine Learning project (NVML) input features used for analysis herein. Red represents high values. Blue represents low values. Positives (red dots) and negatives (white dots) are training data used by NVML.

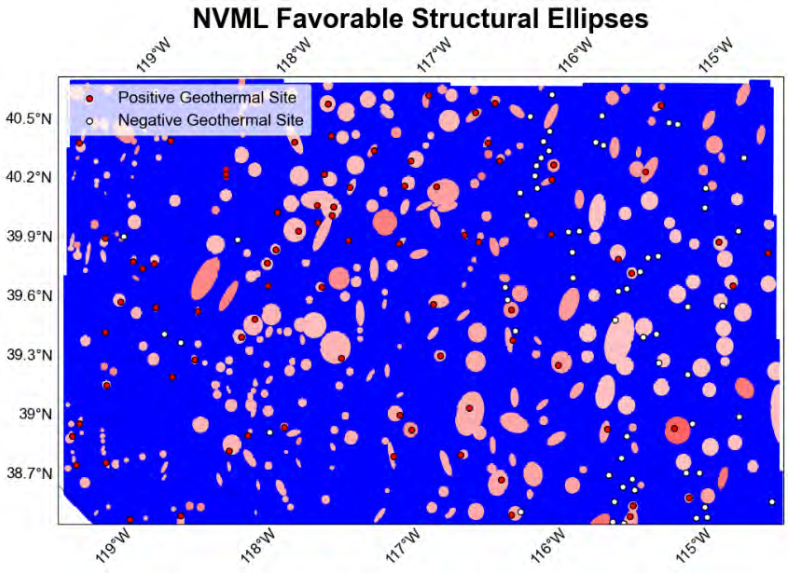

Figure 3: Map of the eleventh Nevada Machine Learning project (NVML) feature (favorable structural setting ellipses [red areas]) used for the artificial neural network (ANN) model (Fig. 1), but not used as an input feature for analyses herein. For random negative sampling strategies, negatives were selected from areas outside of ellipses (i.e., blue areas). Positive (red dots) and negative (white dots) show training data used by NVML.

To conduct a true comparison between training strategies while minimizing expert decisions, we use the same 10 features interpolated from maps for all models developed herein (Fig. 2). As is common practice in ML, standardized features are used as input for each ML model (i.e., subtract the mean and divide by the standard deviation of each input feature).

As part of exploratory data analysis, we inspect the linear correlation of the standardized features using the Pearson correlation coefficient (see generally Lee Rodgers and Nicewander, 1988). The Pearson correlation is a statistical measure that quantifies the linear relationship between two variables by providing values between -1 and 1, where -1 indicates a strong negative correlation, 1 indicates a strong positive correlation, and 0 indicates no correlation.

We compare the distribution of NVML labeled data to the full range of the input features using cumulative distribution functions (CDFs). A CDF represents the distribution of data, so comparing CDFs for positive, negative, and unlabeled data allows a visual evaluation of whether most positive and negative data plot in discrete intervals within the larger range of the input feature.

2.2 Training Strategies

We compare three training strategies: 1) the NVML training strategy; 2) an intermediary strategy that implements random sampling with an equal number of positives and negatives to the NVML strategy (the Random Sampling training strategy); and 3) the training strategy from Mordensky et al. (2023) that implements random sampling and accounts for class imbalance (the Natural Class Imbalance training strategy). The NVML strategy uses the same negative and positive examples from NVML (i.e., 62 negative sites and 83 positive sites). The Random Sampling and Natural Class Imbalance strategies use the same known positives as the NVML strategy, but randomly

select negative sites from outside the favorable structural ellipses (blue area on Fig. 3). The Random Sampling strategy samples 62 negatives, and the Natural Class Imbalance strategy selects the number of negative sites corresponding to the expected positive-negative natural class imbalance (i.e., a roughly 1:5,100 positive-to-negative ratio). Following Mordensky et al. (2023), we estimate that one in four hydrothermal systems have already been found in the study area. Although this estimate is approximate, Mordensky et al. (2023) demonstrated that the corresponding model predictions are insensitive within the expected range of uncertainty (see Mordensky et al. 2023 for complete details).

2.3 Hyperparameter Optimization

We optimize hyperparameters using 60 80:20 train-test splits with five-fold cross-validation to prevent possible bias from any singular train-test split. For the Random Sampling and Natural Class Imbalance strategies, new sets of random negatives are selected with each train-test split.

We optimize each model by maximizing the F1 score given in Equation 1 as:

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

Final hyperparameters are selected by finding the set of hyperparameters that give the largest F1 score from all of the models considered. We fit a final model using all the labeled data, the median optimal hyperparameters from the 60 train-test splits, and one last random sample of negative sites for the Random Sampling and Natural Class Imbalance strategies.

To test and prevent overfitting with XGBoost, we impose loss-of-generalization early stopping (see generally Prechelt, 2002). With generalization loss, the fitting of new estimators stops immediately after the loss (a metric that is penalized by decreased model performance) increases. We train a final model for each XGBoost approach using the median estimator at which early stopping is engaged from the 60 train-test splits.

2.4 Measures of Feature Importance

For every modeling approach, we evaluate the relative importance of each input feature (Fig. 2) in making predictions. We use three measurements of feature importance in this study to explore the variability between measures: 1) sensitivity analysis using an F1 score; 2) sensitivity analysis using the area under the receiver operating characteristic curve (i.e., ROCAUC), and 3) Shapely Additive exPlanation (i.e., SHAP) values (see Mordensky et al., 2023 for a more detailed summary). To allow comparison between the different measures, each measure is min-max normalized to a zero-to-one scale.

3. Results

In this section, we briefly describe the input feature data, present model performance, provide favorability predictions, and supply feature importance.

3.1 Exploratory Data Analysis

The distribution of NVML values for NVML-labeled positive and negative sites generally serve as upper and lower bounds for the distribution of unlabeled data for five features (e.g., strain rate, seismic density, heat flow, fault distance, and slip rate; Fig. 4). However, the values for NVML positives and negatives do not bound the distributions of unlabeled values for five features (i.e., elevation, magnetics, fault recency, gravity, and slip and dilation tendency.

In terms of the Pearson correlation coefficient, strain rate is the feature most correlated with the greatest number of other features (e.g., seismic event density and elevation; -0.59 and 0.33, respectively; Fig. 5). Fault distance and gravity are least correlated with other features (e.g., $\leq |-0.15|$ and $|-0.20|$, respectively).

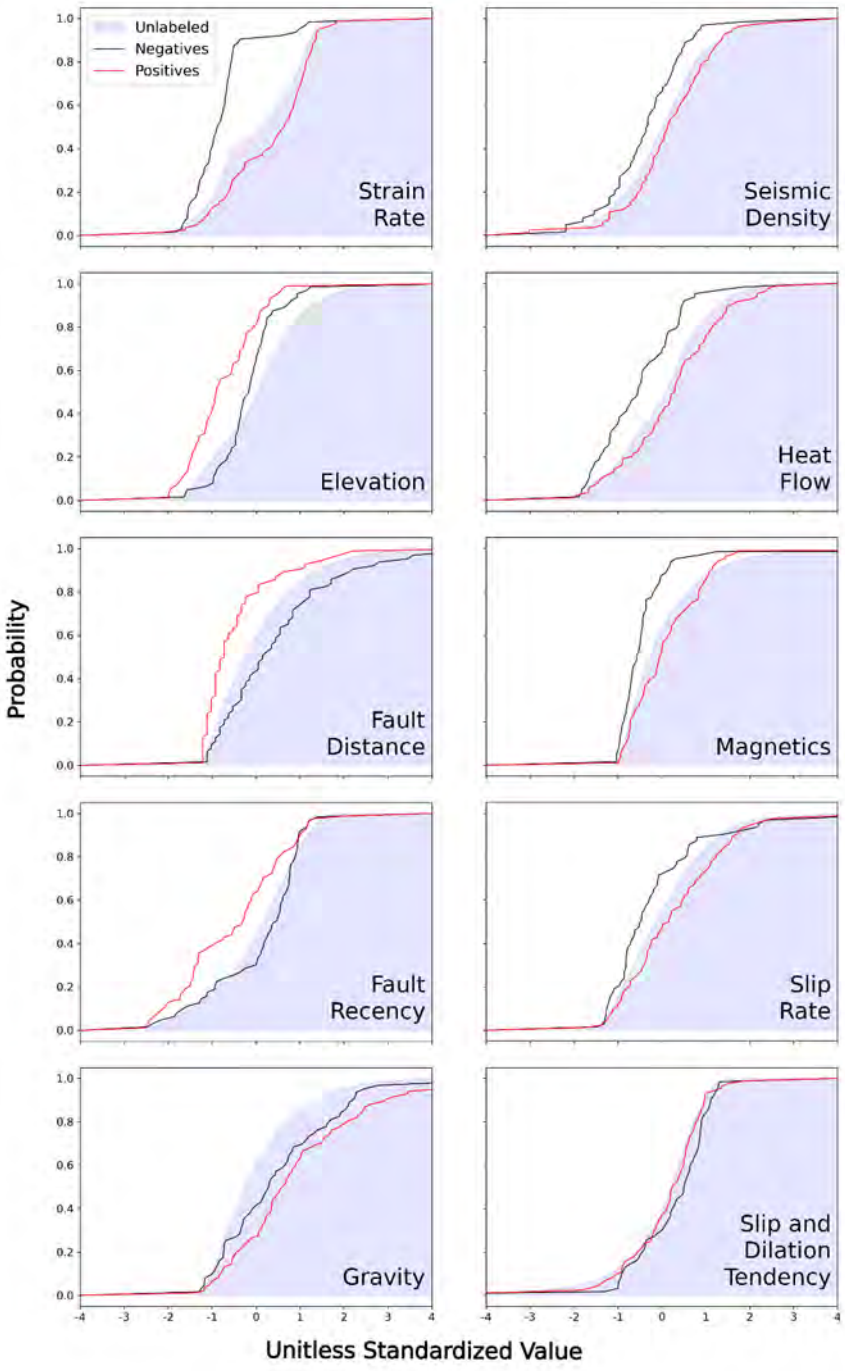

Figure 4: Cumulative distributions of the standardized features. The shaded blue area corresponds to the unlabeled examples. The red line corresponds to positive-labeled examples. The black line corresponds to the negative examples selected by the Nevada Machine Learning project (NVML). Separation of cumulative distribution functions show that distributions are markedly different, so a difference between the positive and negative lines implies the feature may be useful for discriminating between the positive and negative training sites.

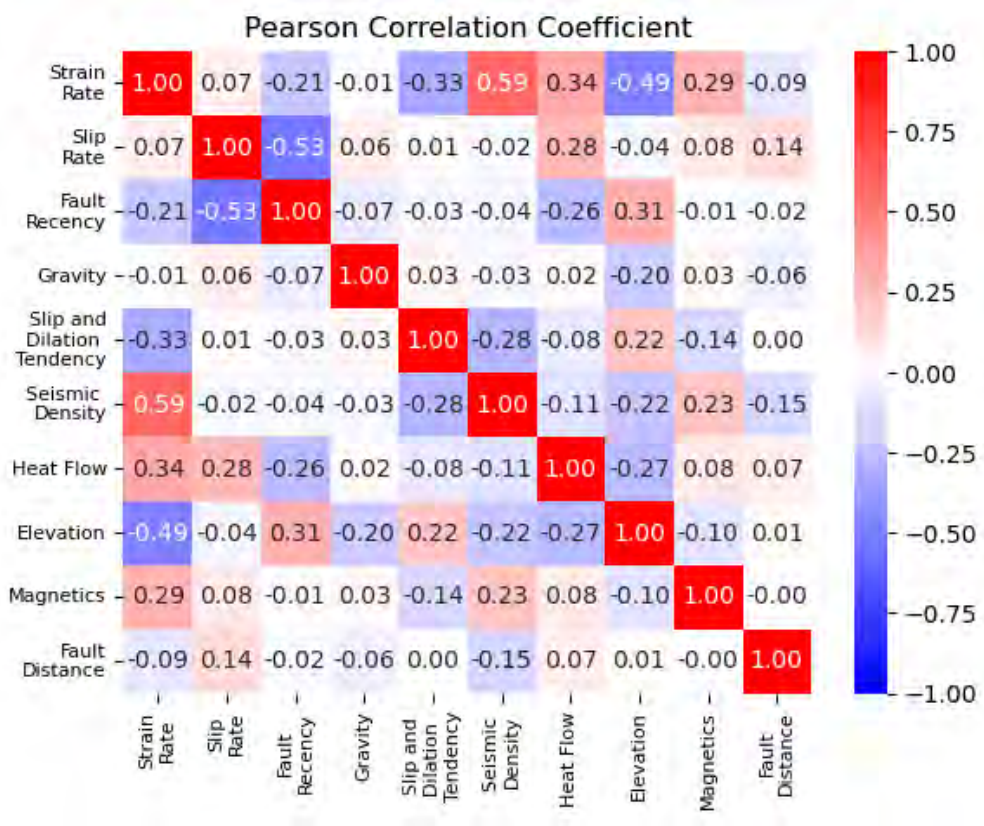

Figure 5: Pearson correlation matrix of the different features for all examples. Brighter colors indicate a higher correlation between pairs. Blue indicates a negative correlation, and red indicates a positive correlation.

3.2 Optimal Hyperparameter Values and Model Performance

We provide the median optimal hyperparameter values and loss-vs-estimator relationships of the 60 train-test splits for early stopping in Appendix A.

The NVML training strategy has the best performance in terms of F1 scores (Fig. 6). The Natural Class Imbalance strategy produces the lowest F1 scores and only correctly predicts a small number of known positives (Table 1). By measure of the F1 score, XGBoost outperforms logistic regression on the NVML and Natural Class Imbalance strategies (Fig. 6), although the differences in the performance in F1 score is statistically significant only for the NVML strategy (p-value < 0.002). A more detailed view of the models' classification performance (Table 1) reveals that training strategies with randomized selection of negatives sites (i.e., the Random Sampling and Natural Class Imbalance strategies) results in a higher predicted number of false positives resulting in lower F1 scores (see Equation 1).

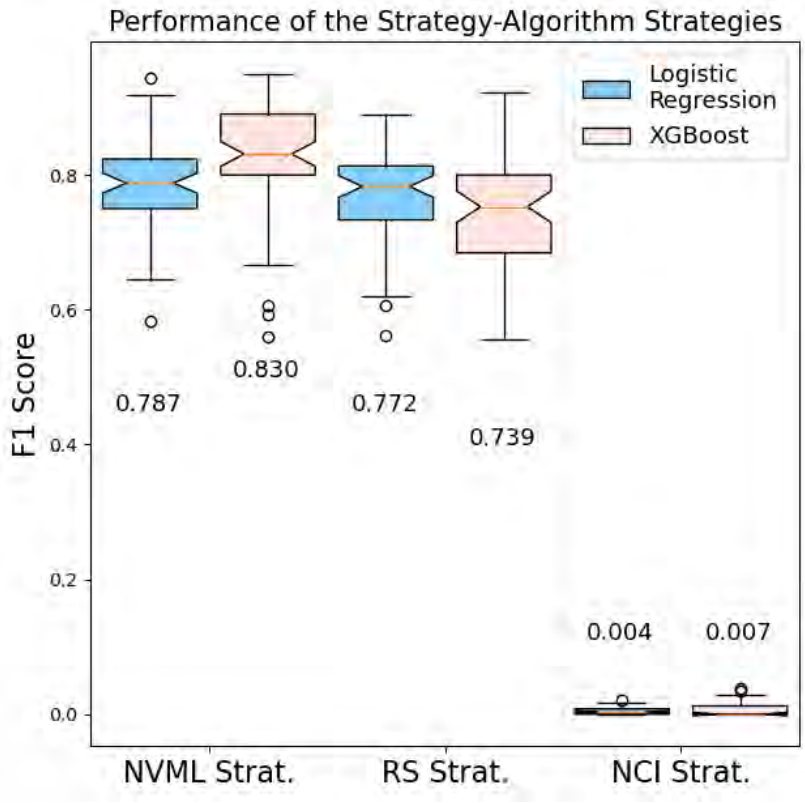

Figure 6: Box-and-whisker plots of F1 score for test data for each machine learning strategy with their corresponding algorithm across 60 train-test splits grouped by training strategy. Logistic regression is blue, and XGBoost is red. Boxes extend from the first quartile (Q1) to the third quartile (Q3) with a red 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. Mean F1 scores are provided in the same column as their respective box-and-whisker plot. Abbreviations: NVML Strat: Nevada Machine Learning Strategy, RS Strat: Random Sampling Strategy, NCI Strat: Natural Class Imbalance Strategy.

Table 1: Confusion matrices of training and testing for logistic regression and XGBoost for the three strategies plus/minus one standard deviation. True Positive corresponds to a positive-labeled cell predicted correctly by the model. True Negative corresponds to a negative-labeled cell predicted correctly by the model. False Negative corresponds to positive labels wrongly predicted as negatives. False Positive corresponds to a negative-labeled cell wrongly predicted as positive. We note that there are no True Negatives or False Positives with the Random Sampling and Natural Class Imbalance strategies because the labeled negatives are truly unlabeled examples that may contain unlabeled positives, but we refer to positives predicted from unlabeled cells as False Positives and negatives predicted from unlabeled cells as True Negatives. Abbreviations: TN: True Negative, TP: True Positive, FP: False Positive, FN: False Negative, NVML Strat: Nevada Machine Learning Strategy, RS Strat: Random Sampling Strategy, NCI Strat: Natural Class Imbalance Strategy.

NVML Strat. Logistic Regression				**NVML Strat. XGBoost**			
Training				***Training***			
TN	38 ± 7	**FP**	13 ± 7	**TN**	38 ± 5	**FP**	13 ± 5
FN	9 ± 5	**TP**	58 ± 6	**FN**	2 ± 1	**TP**	65 ± 3
Testing				***Testing***			
TN	9 ± 2	**FP**	5 ± 3	**TN**	8 ± 2	**FP**	5 ± 2
FN	4 ± 2	**TP**	14 ± 2	**FN**	2 ± 2	**TP**	16 ± 3
RS Strat. Logistic Regression				**RS Strat. XGBoost**			
Training				***Training***			
TN	27 ± 7	**FP**	23 ± 6	**TN**	29 ± 7	**FP**	22 ± 7
FN	6 ± 3	**TP**	61 ± 4	**FN**	2 ± 2	**TP**	65 ± 3
Testing				***Testing***			
TN	5 ± 2	**FP**	7 ± 2	**TN**	5 ± 2	**FP**	8 ± 3
FN	3 ± 1	**TP**	15 ± 2	**FN**	3 ± 2	**TP**	15 ± 3
NCI Strat. Logistic Regression				**NCI Strat. XGBoost**			
Training				***Training***			
TN	1,362,217 ± 1897	**FP**	5,680 ± 1897	**TN**	1,366,237 ± 971	**FP**	1659 ± 971
FN	61 ± 2	**TP**	6 ± 2	**FN**	31 ± 26	**TP**	36 ± 26
Testing				***Testing***			
TN	86,433 ± 126	**FP**	362 ± 126	**TN**	86,688 ± 62	**FP**	107 ± 62
FN	17 ± 1	**TP**	1 ± 1	**FN**	17 ± 1	**TP**	1 ± 1

3.3 Model Predictions

The maps created using the NVML training strategy with logistic regression and XGBoost exhibit a west-east trend in predicted geothermal favorability (Fig. 7). Both approaches show a general inclination to predict the west side of the study area as being more favorable to hydrothermal activity than the east side of the study area. When using the Random Sampling training strategy high predictions instead correspond to locally low elevation (i.e., basins) and low favorability corresponds to locally high elevation (i.e., ranges; Fig. 8). The prediction maps resulting from the Natural Class Imbalance strategy appear much like those from the Random Sampling strategy (Fig. 9).

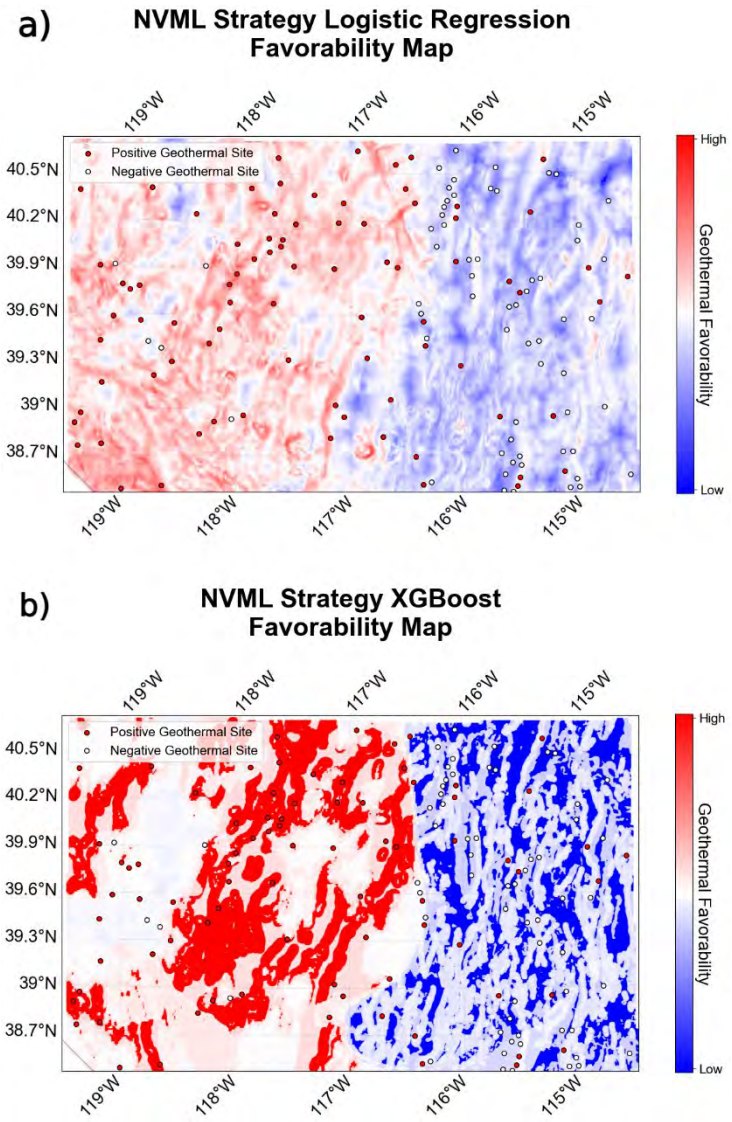

Figure 7: Geothermal favorability maps from using the Nevada Machine Learning training strategy (NVML Strategy) with a) logistic regression; and b) XGBoost with early stopping (at 13 estimators). Geothermal favorability is the normal score transform of probability.

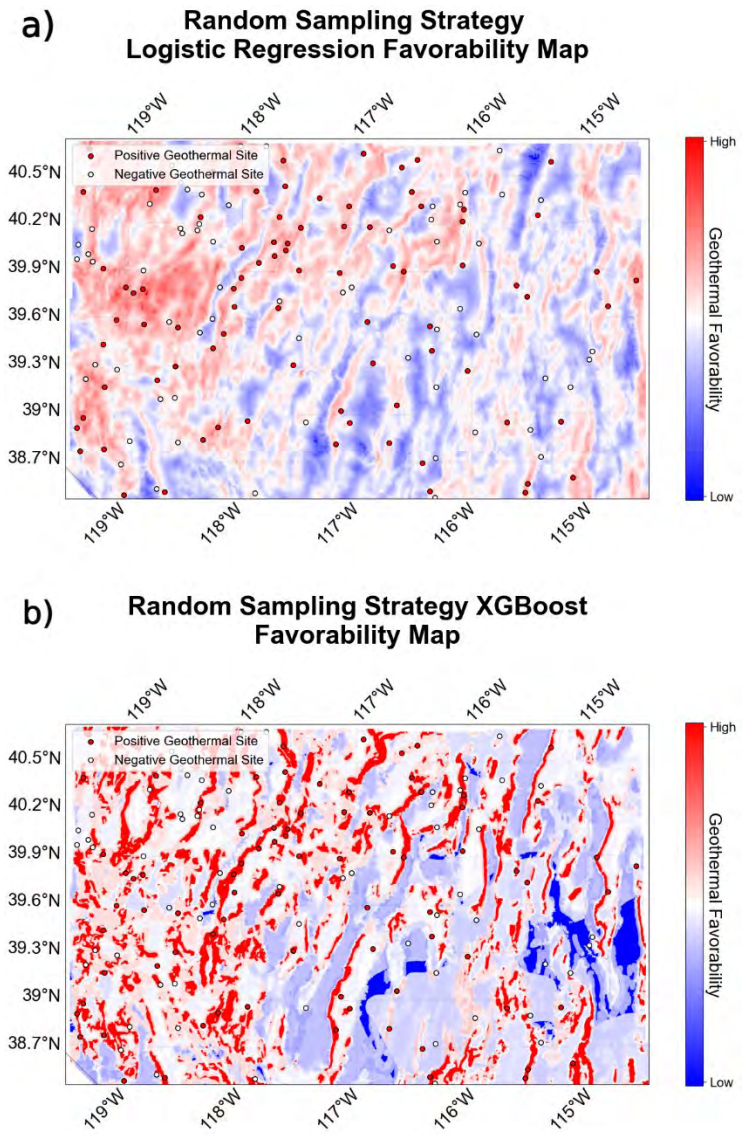

Figure 8: Geothermal favorability maps from using the Random Sampling training strategy with a) logistic regression; and b) XGBoost with early stopping (at 5 estimators). Geothermal favorability is the normal score transform of probability.

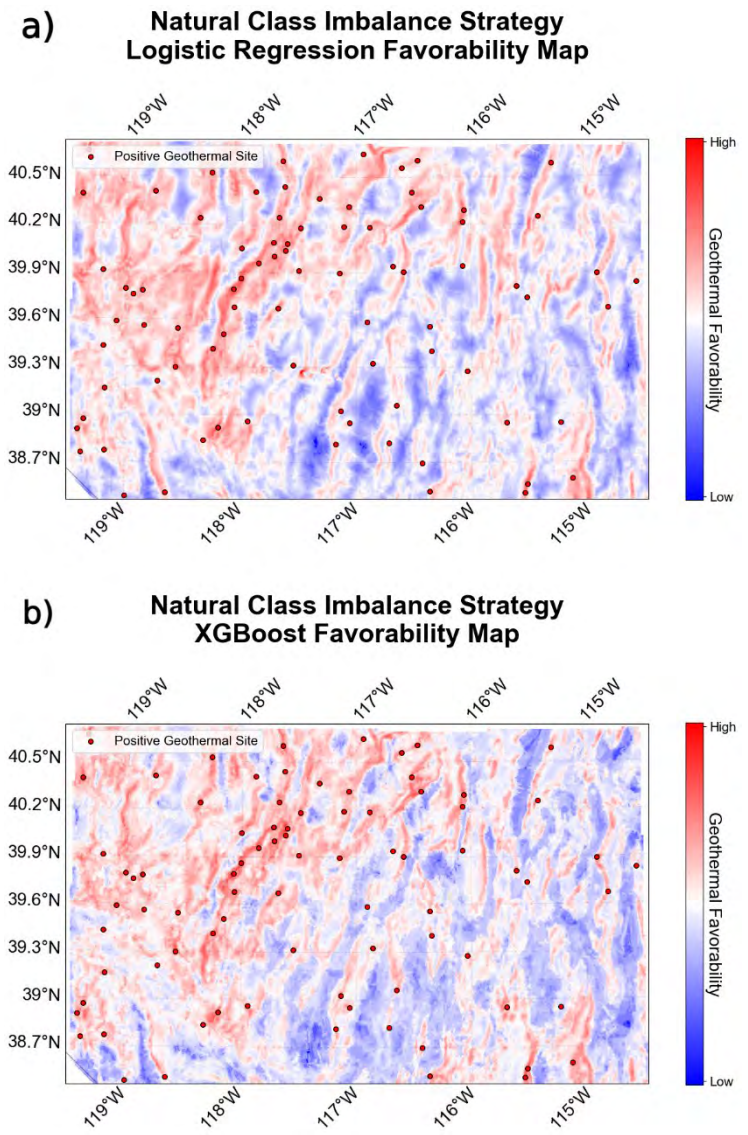

Figure 9: Geothermal favorability maps from using the Natural Class Imbalance training strategy with a) logistic regression; and b) XGBoost with early stopping (at 50 estimators). Geothermal favorability is the normal score transform of probability.

3.4 Feature Importance

The relative ranking of feature importance by different measures (Fig. 10) depicts a strong inclination of all models for one of two features (elevation or strain rate). For the NVML strategy, strain rate is the dominant feature, whereas elevation is the dominant feature for the Random Sampling and Natural Class Imbalance strategies.

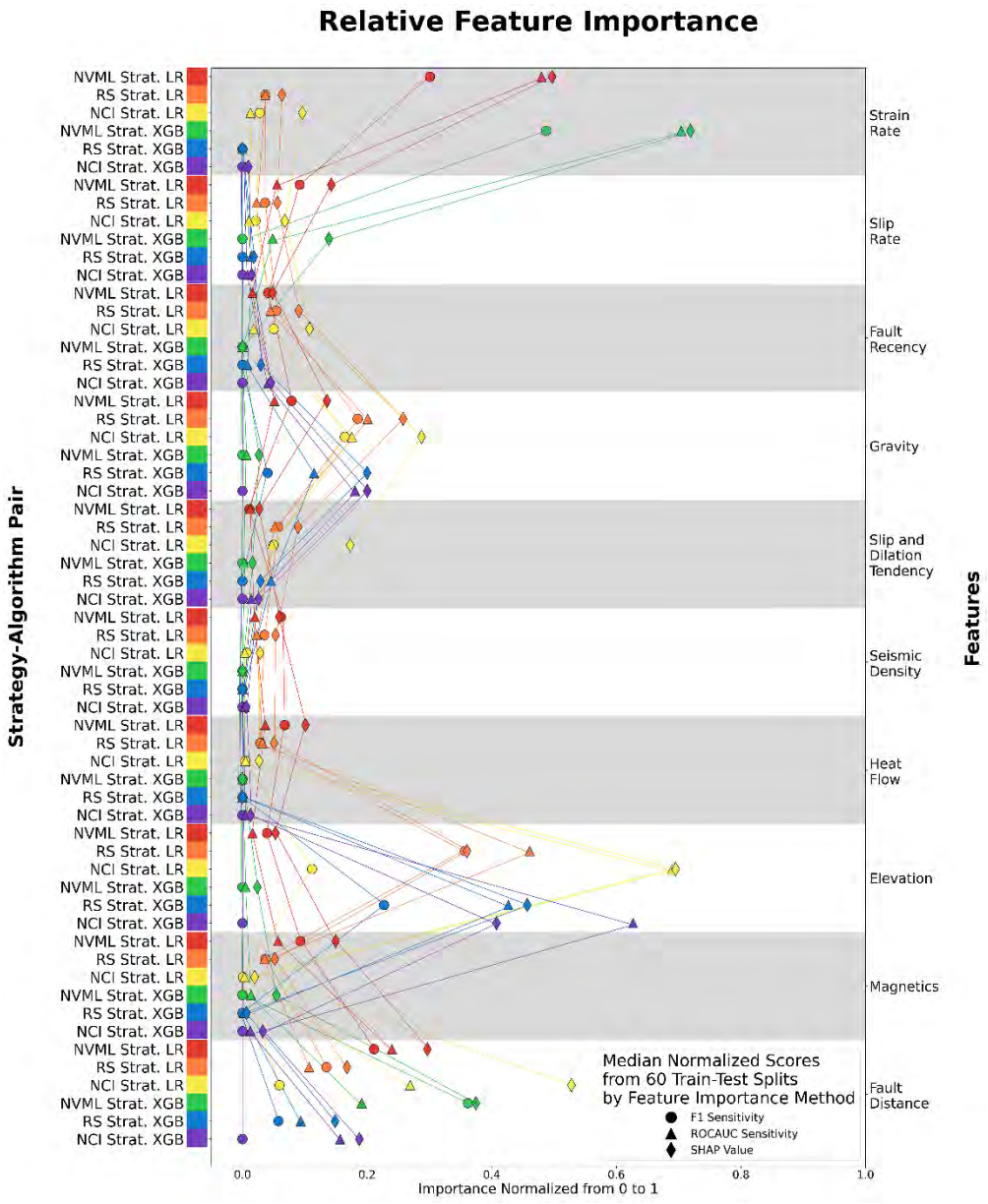

Figure 10: Median normalized feature importance values from the 60 train-test splits using the different strategy-algorithm approaches. Abbreviation: NVML Strat. LR (red) – Nevada Machine Learning strategy Logistic Regression; RS Strat. LR (orange) – Random Sampling strategy Logistic Regression; NCI Strat. LR (yellow) – Natural Class Imbalance strategy Logistic Regression; NVML Strat. XGB (green) – Nevada Machine Learning strategy XGBoost; RS Strat. XGB (blue) – Random Sampling strategy XGBoost; NCI Strat. XGB (purple) – Natural Class Imbalance strategy XGBoost; ROCAUC (triangle) – Area Under the Receiver Operating Characteristic Curve; SHAP (diamond) – Shapely Additive explanation.

4. Discussion

Although F1 scores might suggest the poorest performer is the Natural Class Imbalance training strategy (Fig. 6), the F1 scores may be partially misleading because the “false positives” upon which the metric relies (see Equation 1) may not truly be false positives in the Random Sampling

and Natural Class Imbalance training strategies because these “labeled negatives” may contain unlabeled positives. Therefore, we consider another metric of success - the map that gives the majority of the known hydrothermal systems the highest favorability scores implicitly associates known hydrothermal systems with the smallest geographical area. If the goal is to identify discrete areas for exploration, then this map reduces the effort to only this smaller area. This metric, first proposed in Faulds et al. (2016), compares the relative predictions of known positives with the greater distribution of the unlabeled examples (see Mordensky et al. 2023 for additional details). Curves for known positive and unlabeled cells with greater overlap indicate more alike distributions of predicted values, whereas curves for known positive cells that overlap less with the curve for unlabeled cells indicate better prediction of the known positive cells. Under this metric the Natural Class Imbalance strategy with XGBoost outperforms all other strategies (Fig. 11) and has the narrowest zone for high favorability compared to all maps (Fig. 9) demonstrating that accounting for the expected natural class imbalance in randomized sites in hydrothermal systems leads to better predictions and results in better models for predicting geothermal favorability.

The NVML ANN is the second best-performing model we consider, with its known positives (83) and negatives (62) outperforming all but the Natural Class Imbalance strategy with XGBoost approach (Fig. 11). This greater performance could be explained in part because the NVML team included structural information as expert-identified ellipses (Fig. 3) as an input feature to their ANN (Fig. 1). The structural ellipses have two functions: 1) encapsulating the positive sites (i.e., most NVML positive sites are within ellipses); and 2) narrowing the potential geospatial distribution of positives (i.e., positive sites are located within favorable structural settings). These two factors combined allow the ANN to better select areas for the occurrence of potential (i.e., undiscovered) positive sites that are consistent with the structures associated with hydrothermal systems in the study area (Faulds et al., 2006; Faulds et al., 2011; Faulds and Hinz, 2015).

Overall, all resulting maps (Figs. 7, 8, 9) exhibit to some degree a west-east bias in the favorability prediction, regardless of the randomization of the negative sites. The analysis of feature importance (Fig. 10) shows that elevation and strain rate rank as the most important features. However, these features are correlated to each other (Fig. 5), and both have a general east-west trend (Fig. 2). The biased predictions may reflect, in part, the spatial distribution of positive sites being primarily in the west (controlled by basin-bounding faults that govern most geothermal systems in the Great Basin; Faulds et al., 2011; Faulds and Hinz, 2015) that result in the algorithms associating conditions in the west (i.e., high strain rate, low elevation) with positive predictions, and limiting the models’ ability to learn from other local variation in feature values.

Another perspective to consider is the limitation of the data itself. The input data vary smoothly, (Fig. 2) whereas geological conditions do not exhibit this behavior. This smoothness in the input data may lead to inaccuracies in interpolated values, reducing their fidelity in representing true geological conditions.

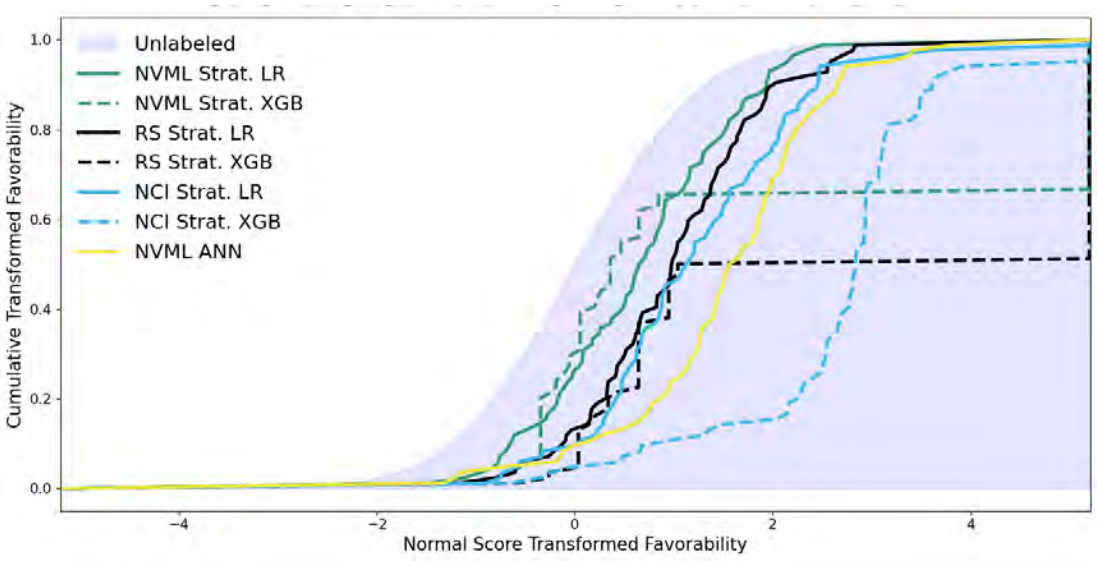

Figure 11: Cumulative distribution function (CDFs) of favorability scores for known geothermal systems relative to most other map locations (i.e., unlabeled cells). CDFs for known hydrothermal systems that plot further to the right represent models that predict known systems as being in a smaller area of the map composed of the highest favorability scores. Shaded blue provides the cumulative distribution of predictions for unlabeled cells. Lines represent the cumulative distributions of predictions for positive-labeled cells from the different approaches. Abbreviations: NVML Strat. LR (solid green line) – Nevada Machine Learning Strategy with Logistic Regression; NVML Strat. XGB (dashed green line) – Nevada Machine Learning Strategy with XGBoost; RS Strat. LR (solid black line) – Random Sampling Strategy with Logistic Regression; RS Strat. XGB (dashed black line) - Random Sampling Strategy with XGBoost; NCI Strat. LR (solid light blue line) – Natural Class Imbalance Strategy with Logistic Regression; NCI Strat. XGB (dashed light blue line) - Natural Class Imbalance Strategy with XGBoost; NVML ANN (solid yellow line) - Nevada Machine Learning Artificial Neural Network.

5. Conclusion

In this study, we compare the predictive performance of different models trained by randomly selecting negative sites for highly sparse hydrothermal systems with the Nevada Machine Learning Project's artificial neural network, which used negative sites where the research team had confidence that no hydrothermal resources exist. In doing so, we demonstrate that the expert selection of negative sites imparted modeling bias to the Nevada Machine Learning Project's artificial neural network. Furthermore, we demonstrate that accounting for the expected natural class imbalance in randomized sites leads to better performance with XGBoost emerging as the top-performing training approach.

Acknowledgment

This work was supported by the Portland State University Louis Stokes Alliance for Minority Participation Program, Portland State University Institute for Sustainable Solutions Program, Portland State McNair Scholars Program, the U.S. Geological Survey Portland State University Partnership, 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 Ryan Cahalan and William Schermerhorn for their valuable feedback and suggestions to improve this manuscript.

Appendix A: Hyperparameters and Early Stopping

The median optimal hyperparameters from the 60 train-test splits (Table A1) are used to train the final models using all the data.

Table A1: Median optimal hyperparameters from the 60 train-test splits. Abbreviations: Inv. Reg. Str: Inverse Regularization Strength, LR: Logistic Regression, XGB: XGBoost, NVML Strat: Nevada Machine Learning Strategy, RS Strat: Random Sampling Strategy, NCI Strat: Natural Class Imbalance Strategy, GL: Early stopping using generalization loss, Est: Estimator.

Strategy & Algorithm	Class Weight	Inverse Reg. Str
NVML Strat. LR	1.5 ± 0.88	10.2 ± 24.4
RS Strat. LR	1.9 ± 0.84	2.1 ± 3.5
NCI Strat. LR	1700 ± 96.31	0.0001 ± 0.04

Strategy & Algorithm	Class Weight	Learning Rate	Max Depth	GL Est.
NVML Strat. XGB	1.6 ± 0.3	0.08 ± 0.06	3 ± 0.6	13 ± 9.2
RS Strat. XGB	2.4 ± 0.7	0.11 ± 0.09	3 ± 0.8	5 ± 4
NCI Strat. XGB	1800 ± 96.4	0.15 ± 0.03	4 ± 0.9	50 ± 27

The median estimator for early stopping in the 60 train-test splits is used as the max depth in the final XGBoost models to prevent overfitting. We investigate one form of early stopping (loss of generalization), which is dependent upon loss-vs-estimator relationships. Training loss and validation loss generally decrease as early estimators are added (Fig. A1), but training loss generally decreases at a faster rate than the validation loss in the NVML and Random Sampling strategies. With the Natural Class Imbalance strategy, the validation and training loss decrease at roughly the same rate, suggesting that the larger number of selected negatives in the Natural Class Imbalance strategy reduces the potential for train-test splits being unrepresentative like some train-test splits when training with the NVML and Random Sampling strategies.

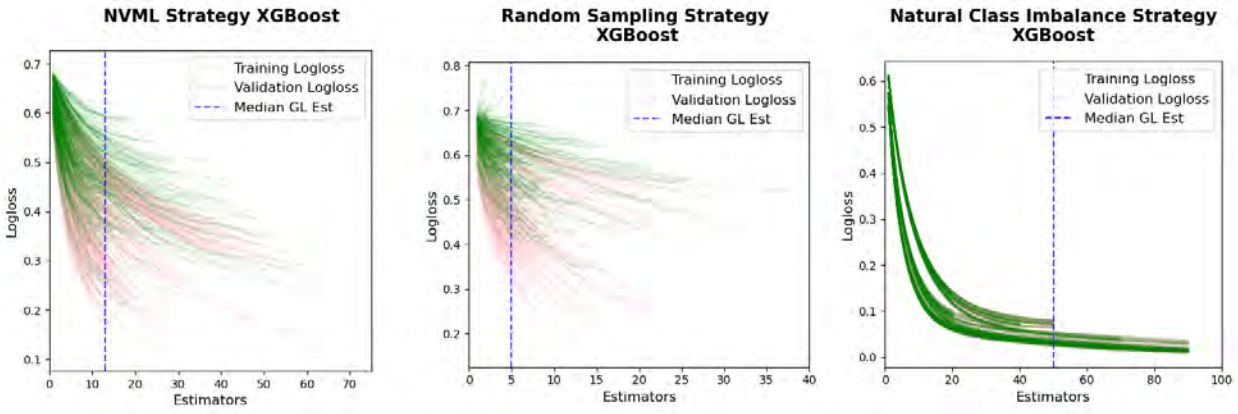

Figure A1: Loss-vs-estimator relationships for XGBoost in the Nevada Machine Learning project (NVML), Random Sampling and Natural Class Imbalance training strategies. Training loss (pink) and validation loss (green) generally decrease as early estimators are added, but training loss decreases at a faster rate than the validation loss in the NVML and Random Sampling training strategies. The median estimator for early stopping using loss in generality (GL Est) is marked by a blue dashed line.

REFERENCES

Ayling, B. F. (2020). *35 Years of Geothermal Power Generation in Nevada, USA: A Review of Field Development, Generation, and Production Histories*. Proceedings of the 45th Workshop on Geothermal Reservoir Engineering, Stanford University, Stanford, California, United States.

Berkson, J. (1944). Application of the logistic function to bio-assay. *Journal of the American statistical association*, *39*(227), 357-365.

Berkson, J. (1951). Why I prefer logits to probits. *Biometrics*, *7*(4), 327-339.

Brown, S. R., Coolbaugh, M. F., DeAngelo, J., Faulds, J. E., Fehler, M., Gu, C., Queen, J., Treitel, S., Smith, C., and Mlawsky, E. (2020). Machine Learning for Natural Resource Assessment: An Application to the Blind Geothermal Systems of Nevada. *Geothermal Resources Council Transactions*, *44*, 920-932.

Chen, T., and Guestrin, C. (2016). *XBGoost: A Scalable Tree Boosting System*. KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, California, United States.

Coolbaugh, M. F., Raines, G. L., and Zehner, R. E. (2007). Assessment of Exploration Bias in Data-Driven Predictive Models and the Estimation of Undiscovered Resources. *Natural Resources Research*, *16*, 199-207.

Faulds, J. E., Brown, S. R., Coolbaugh, M. F., DeAngelo, J., Queen, J. H., Treitel, S., Fehler, M., Mlawsky, E., Glen, J. M., and Lindsey, C. (2020). *Preliminary Report on Applications of Machine Learning Techniques to the Nevada Geothermal Play Fairway Analysis*. Proceedings of the 45th Workshop on Geothermal Reservoir Engineering, Stanford University, Stanford, California, United States.

Faulds, J. E., Brown, S. R., Smith, C. M., Queen, J., and Treitel, S. (2021a). *Machine Learning Model Geotiffs-Applications of Machine Learning Techniques to Geothermal Play Fairway Analysis in the Great Basin Region, Nevada*. https://doi.org/https://doi.org/10.15121/1897036

Faulds, J. E., Coolbaugh, M. F., Vice, G. S., and Edwards, M. L. (2006). Characterizing Structural Controls of Geothermal Fields in the Northwestern Great Basin: A Progress Report. *Geothermal Resources Council Transactions, 30*, 69-76

Faulds, J. E., Hinz, N., Coolbaugh, M., dePolo, C., Siler, D., Shevenell, L., Hammond, W., Kreemer, C., and Queen, J. (2016). *Discovering Geothermal Systems in the Great Basin Region: An Integrated Geologic, Geochemical, and Geophysical Approach for Establishing Geothermal Play Fairways*. Proceedings of the 41st Workshop on Geothermal Reservoir Engineering, Stanford University, Stanford, California, United States.

Faulds, J. E., and 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*. Proceedings of the World Geothermal Congress, Melbourne, Australia.

Faulds, J. E., Hinz, N. H., Coolbaugh, M., Ayling, B., Glen, J., Craig, J. W., McConville, E., Siler, D., Queen, J., and Witter, J. (2021b). *Discovering Blind Geothermal Systems in the Great Basin Region: An Integrated Geologic and Geophysical Approach for Establishing Geothermal Play Fairways: All Phases* (DOE-UNR-06731-01).

Faulds, J. E., Hinz, N. H., Coolbaugh, M. F., Cashman, P. H., Kratt, C., Dering, G., Edwards, J., Mayhew, B., and McLachlan, H. (2011). Assessment of Favorable Structural Settings of Geothermal Systems in the Great Basin, Western USA. *Geothermal Resources Council Transactions, 35*, 777-783.

Faulds, J. E., Hinz, N. H., Coolbaugh, M. F., dePolo, C., Siler, D., Shevenell, L. A., Hammond, W., Kreemer, C., and Queen, J. (2015a). *Discovering Geothermal Systems in the Great Basin region: An Integrated Geologic, Geochemical, and Geophysical Approach for Establishing Geothermal Play Fairways*. Proceedings 41st Workshop on Geothermal Reservoir Engineering,

Faulds, J. E., Hinz, N. H., Coolbaugh, M. F., Sadowski, A., Shevenell, L. A., McConville, E., Craig, J., Sladek, C., and Siler, D. (2017). *Progress Report on the Nevada Play Fairway Project: Integrated Geological, Geochemical, and Geophysical Analyses of Possible New Geothermal Systems in the Great Basin Region*. Proceedings of 42nd Workshop on Geothermal Reservoir Engineering, Stanford University, Stanford, California, United States.

Faulds, J. E., Hinz, N. H., Coolbaugh, M. F., Shevenell, L. A., Siler, D., dePolo, C., Hammond, W., Kreemer, C., Oppliger, G., and Wannamaker, P. (2015b). Integrated Geologic and Geophysical Approach for Establishing Geothermal Play Fairways and Discovering Blind Geothermal Systems in the Great Basin Region, Western USA: A Progress Report. *GRC Transactions, 39*(DOE-UNR-06731-05).

Lee Rodgers, J., and Nicewander, W. A. (1988). Thirteen ways to look at the correlation coefficient. *The American Statistician, 42*(1), 59-66.

Mordensky, S. P., Lipor, J. J., Burns, E. R., and Lindsey, C. R. (2022). *What did they just say? Building a Rosetta stone for geoscience and machine learning*. 2022 Geothermal Rising Conference,

Mordensky, S. P., Lipor, J. J., DeAngelo, J., Burns, E. R., and 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.

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

Prechelt, L. (2002). Early stopping-but when? In *Neural Networks: Tricks of the trade* (pp. 55-69). Springer.

Pyrcz, M. J., and Deutsch, C. V. (2018). Transforming data to a Gaussian distribution. *Geostatistics Lessons*.

Richards, M., and Blackwell, D. (2002). A Difficult Search: Why Basin and Range systems are hard to find. *Geothermal Resources Council Bulletin*, *31*(4), 143-146.

Smith, C. M. (2021). *Machine Learning Techniques Applied to the Nevada Geothermal Play Fairway Analysis* University of Nevada, Reno]. Reno, Nevada. https://scholarworks.unr.edu//handle/11714/7940

Smith, C. M., Faulds, J. E., Brown, S., Coolbaugh, M., DeAngelo, J., Glen, J. M., Burns, E. R., Siler, D. L., Treitel, S., and Mlawsky, E. (2023). Exploratory analysis of machine learning techniques in the Nevada geothermal play fairway analysis. *Geothermics*, *111*, 102693.

Williams, C. F., Reed, M. J., Mariner, R. H., DeAngelo, J., and 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.