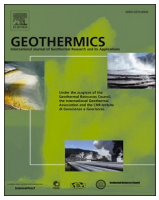
Geothermics 110 (2023) 102662

Contents lists available at ScienceDirect

Geothermics

journal homepage: www.elsevier.com/locate/geothermics



When less is more: How increasing the complexity of machine learning strategies for geothermal energy assessments may not lead toward better estimates

Stanley P. Mordensky [a,*], John J. Lipor [b], Jacob DeAngelo [c], Erick R. Burns [a], Cary R. Lindsey [a]

[a] U.S. Geological Survey, Portland OR 97201, United States
[b] Portland State University, Portland, OR 97201, United States
[c] U.S. Geological Survey, Moffett Field, CA 94035, United States

ARTICLE INFO

Keywords:
Geothermal resource assessment
Machine learning
Class imbalance
Logistic regression
Support-vector machines
XGBoost
Artificial neural network
Feature importance
Positive-unlabeled data
Positive-unlabeled learning

ABSTRACT

Previous moderate- and high-temperature geothermal resource assessments of the western United States utilized data-driven methods and expert decisions to estimate resource favorability. Although expert decisions can add confidence to the modeling process by ensuring reasonable models are employed, expert decisions also introduce human and, thereby, model bias. This bias can present a source of error that reduces the predictive performance of the models and confidence in the resulting resource estimates.

Our study aims to develop robust data-driven methods with the goals of reducing bias and improving predictive ability. We present and compare nine favorability maps for geothermal resources in the western United States using data from the U.S. Geological Survey's 2008 geothermal resource assessment. Two favorability maps are created using the expert decision-dependent methods from the 2008 assessment (*i.e.,* weight-of-evidence and logistic regression). With the same data, we then create six different favorability maps using logistic regression (without underlying expert decisions), XGBoost, and support-vector machines paired with two training strategies. The training strategies are customized to address the inherent challenges of applying machine learning to the geothermal training data, which have no negative examples and severe class imbalance. We also create another favorability map using an artificial neural network.

We demonstrate that modern machine learning approaches can improve upon systems built with expert decisions. We also find that XGBoost, a non-linear algorithm, produces greater agreement with the 2008 results than linear logistic regression without expert decisions, because the expert decisions in the 2008 assessment rendered the otherwise linear approaches non-linear despite the fact that the 2008 assessment used only linear methods. The F1 scores for all approaches appear low (F1 score < 0.10), do not improve with increasing model complexity, and, therefore, indicate the fundamental limitations of the input features (*i.e.,* training data). Until improved feature data are incorporated into the assessment process, simple non-linear algorithms (*e.g.,* XGBoost) perform equally well or better than more complex methods (*e.g.,* artificial neural networks) and remain easier to interpret.

1. Introduction

The U.S. Geological Survey (USGS) has produced 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 (Williams and DeAngelo, 2008; Williams et al., 2008; Williams et al., 2009). This assessment produced 28 models to identify locations of high geothermal favorability (*i.e.,* the likelihood of conditions favoring the presence of a geothermal system) in the western United States (examples shown in Fig. 1) using two modeling methods (*i.e.,* weight-of-evidence and logistic regression). The 2008 geothermal resource assessment varied combinations of nine geological input feature sets (see Williams and DeAngelo [2008] for complete reference information). These nine feature sets were divided

* Corresponding author.
E-mail address: smordensky@usgs.gov (S.P. Mordensky).

https://doi.org/10.1016/j.geothermics.2023.102662
Received 21 October 2022; Received in revised form 28 December 2022; Accepted 24 January 2023
Available online 7 February 2023

into five input feature types, and each model used no more than one feature set from each type:

- Heat flow
  - Heat flow interpolated from unpublished data compiled for Williams et al. (2007)
  - Heat flow interpolated from Blackwell and Richards (2004)
- Mapped Quaternary faulting
  - Distance to nearest Quaternary fault from the USGS Quaternary fault and fold database (Machette et al., 2003)
- Mapped Quaternary magmatic activity (*i.e.,* intrusives, flows, and vents), inert and extant, from Donnelly-Nolan (1988), MacLeod et al. (1995), Walker et al. (2006), and Hildreth (2007)
  - Distance to nearest felsic magmatic activity
  - Distance to nearest mafic magmatic activity
  - Distance to nearest magmatic activity irrespective of composition
- Summarized seismic activity data from the Advanced National Seismic System Comprehensive (ANSS) Earthquake Catalog (2022)
  - Earthquake density within 4 km
  - Log of the sum of seismic moments of earthquakes within 10 km
- Stress
  - Maximum horizontal stress interpolated from Reinecker et al. (2005)

Although the assessment models of Williams and colleagues used data-driven fitting methods to assign measured correlations between input features and geothermal sites, data selection and pre-processing occurred at several stages of the analyses based upon expert judgment. For example, all of the feature sets of otherwise continuous data were binned, thereby requiring expert decisions to be made for parameters like bin sizes, number of bins, and threshold values. While these expert decisions potentially add value by incorporating patterns supported by professional judgement, they impose binning methods onto a problem that does not require binning and, therefore, may reduce the predictive skill of the models. Hence, expert decisions may introduce a potential source of human bias and model bias. The mixture of data- and expert-driven decision making is not unique to USGS assessments of geothermal resources, but is also found in many modern assessments including geothermal play fairway analysis in the U.S. (*e.g.,* Aleutian Arc [Hinz et al., 2015], the Cascades [Shevenell et al., 2015], the Great Basin [Faulds et al., 2017], Hawaii [Ito et al., 2017; Lautze et al., 2017; Lautze et al., 2020], the Modoc Plateau [Siler et al., 2017], the Snake River Plain [Nielson et al., 2015; Shervais et al., 2020; Shervais et al., 2021], across parts of Washington State [Forson et al., 2017]) and outside the U.S. (*e.g.,* Argentina [Lindsey et al., 2021], Brazil [Lacasse et al., 2022], China [Meng et al., 2021], Egypt [Abuzied et al., 2020], Taiwan [Meng et al., 2021]).

Machine learning presents an opportunity to remove the expert decisions used in the 2008 geothermal resource assessment by instead relying on data-driven decisions (see generally Musumeci et al., 2019). The use of machine learning to perform these tasks reduces the potential for human bias and error and allows the researcher to focus on other topics (see generally Boutaba et al., 2018). Mordensky et al. (2022) completed the initial steps to adapt machine learning approaches for use with data for geothermal assessments; however, the integration of machine learning into a geothermal resource assessment raises the question about the reliability of the predictions from the machine learning approaches compared to those dependent upon expert decisions.

Herein, we detail machine learning approaches to predict favorability for conventional hydrothermal resources in the western United States with equal or improved performance compared to the 2008 USGS geothermal resource assessment using the same data from the assessment. We emphasize that this study primarily serves as a means of inquiry into the capabilities of machine learning for performing resource assessments, including geothermal play fairway analysis. Consequently, this work is not a comprehensive or complete geothermal resource assessment in itself. In pursuit of developing a better understanding of the relationship between geologic data and resource predictions, seven different machine learning approaches are employed: logistic regression (without expert binning), eXtreme Gradient Boosting (commonly referred to as XGBoost), and support-vector machines (SVMs), each using two training strategies, and one multilayer perceptron artificial neural network (ANN) using one training strategy.

1.1. Machine Learning Algorithms

Machine learning algorithms provide a data-driven means to generate models that can predict conditions favorable for the presence or absence of geothermal systems capable of producing electricity. Machine learning algorithms operate by learning directly from data in order to create optimal decision functions, more commonly called models in geoscience. Implicit to the name, data-driven decisions are choices algorithmically determined to optimize model performance by maximizing performance metrics like accuracy (Eq. 1), precision (Eq. 2), recall (Eq. 3), and F1 score (Eq. 4), which rely upon knowing the number of true positives (*i.e.,* positive training sites predicted as positive), false positives (*i.e.,* negative training sites predicted as positive), true negatives (*i.e.,* negative training sites predicted as negative), and false negatives (*i.e.*, positive training sites predicted as negative).

$$Accuracy = \frac{True\ Positives + True\ Negatives}{True\ Positives + True\ Negatives + False\ Positives + False\ Negatives} \quad (1)$$

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives} \quad (2)$$

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives} \quad (3)$$

$$F1\ Score = 2\frac{Precision \times Recall}{Precision + Recall} = \frac{True\ Positives}{True\ Positives + \frac{1}{2}(False\ Positives + False\ Negatives)} \quad (4)$$

Performance metric optimization is primarily achieved through the selection of an algorithm's internal variables that balance the tradeoff

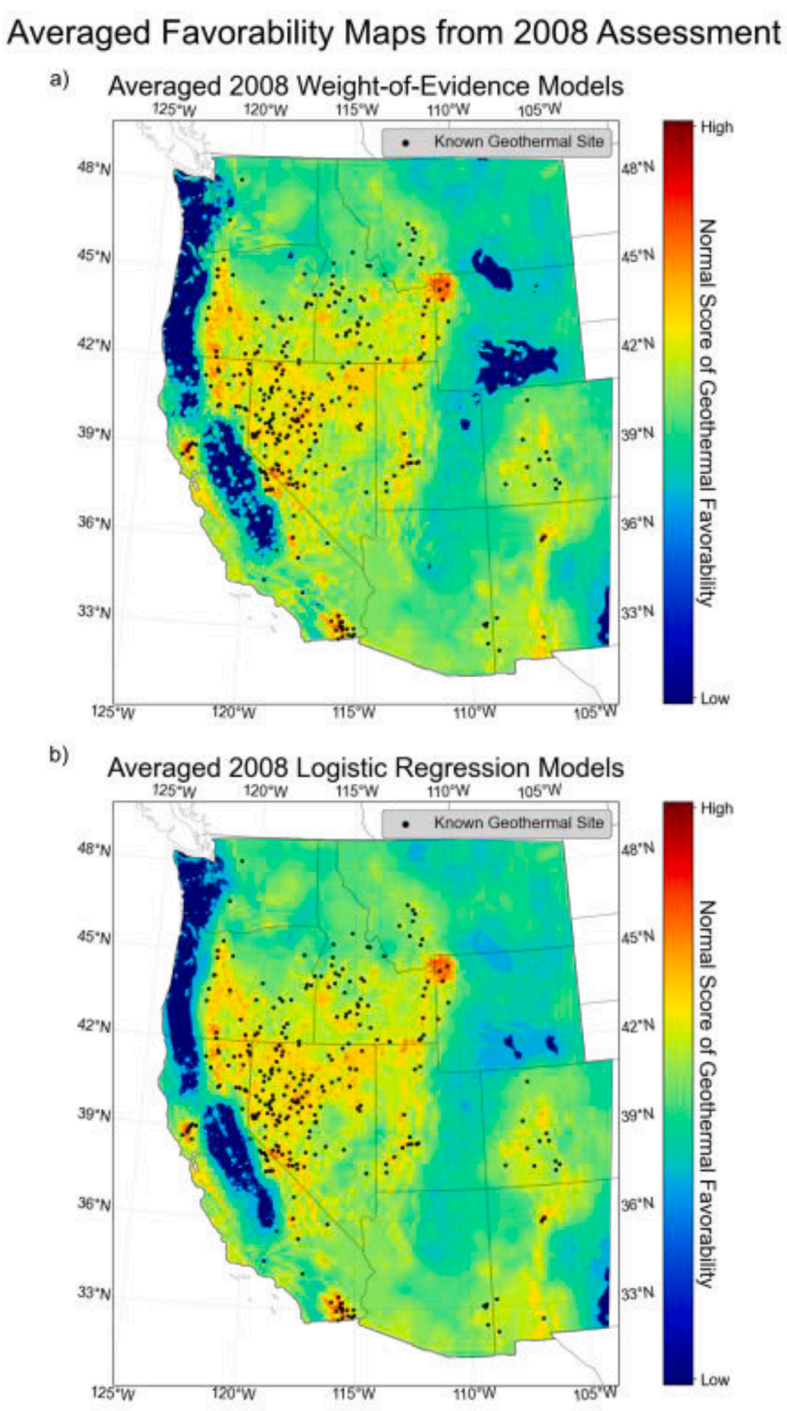


Fig. 1. Geothermal favorability maps of the averaged probability of occurrence predicted from 12 different models (as presented in Williams et al. [2009]) for the western United States using the: a) weight-of-evidence; and b) logistic regression methods from Williams and DeAngelo (2008). The 12 individual models used for averaging are differentiated by their unique input feature combinations. For comparison purposes in this manuscript, favorability is plotted using the normal score transform of the native output from each model.

between underfitting and overfitting. The adjustable internal variables are called hyperparameters, and the selection of hyperparameter values that optimize the chosen performance metric is called hyperparameter optimization. Hyperparameter optimization is completed by experimentally finding the combination of hyperparameter values that correspond to a best performing model (*e.g.,* a model with the greatest predictive skill or lowest error; see generally Burkov, 2019a).

Hyperparameter optimization helps algorithms handle the unique qualities of datasets. One such quality is the relative frequency of the occurrence of each classification label (*e.g.,* positive or negative; presence or absence of a geothermal system). Most machine learning algorithms perform best when there are approximately equal numbers of each type of data (see generally Fernández et al., 2018). Substantial deviation from a similar occurrence of labels is termed class imbalance and impairs the ability of data-driven algorithms to learn from the data (see generally Branco et al., 2015). Class imbalance can range from slight (*e.g.,* 1:10) to severe (e.g., 1:> 100; see generally Krawczyk, 2016). There are several means to address modest class imbalance. Three of the most common are oversampling, undersampling, and penalization (see generally Fernández et al., 2018). Oversampling duplicates existing data of the minority class (*i.e.,* the class with the less frequent occurrence) and increases the risk of overfitting the data because the new data are derived from the smaller, pre-existing dataset. Undersampling removes data of the majority class (*i.e.,* the class with the more frequent occurrence). Undersampling presents the risk of removing valuable data in the larger class. Penalization (*e.g.,* class

weighting) weights label types to place greater emphasis on predicting minority class labels over majority class labels during training. Other options to address class imbalance include using different performance metrics (*e.g.,* accuracy versus F1 score) and algorithms (see generally Branco et al., 2015).

1.2. Challenges Using the Data from the 2008 Geothermal Resource Assessment

Two fundamental challenges exist when applying modern machine learning approaches to geothermal data: (1) although many geothermal systems are known (*i.e.,* labeled as positive), the remainder of the landscape is unlabeled; and (2) geothermal systems are sparse and thus they present severe class imbalance.

The first challenge for a modern, data-driven geothermal resource analysis regards understanding how to account for unlabeled cells. The 2008 USGS geothermal resource assessment gridded the western United States into 725,442 2-km-by-2-km cells (see Williams and DeAngelo, 2008; see Williams et al., 2008), of which 278 contained known conventional hydrothermal systems (Fig. 1). If a cell contained a known geothermal system, the cell was given a positive label. One geothermal system could not span two cells. The remaining cells were assumed to be negative for the 2008 assessment. However, the cells labeled as negative are more correctly identified as unlabeled, since some of these cells may contain geothermal systems. Classic machine learning algorithms and performance metrics are structured to work with positive-negative data and not with positive-unlabeled data.

The second challenge for geothermal resource analysis with machine learning is the severe class imbalance. Only 278 of the 725,442 cells were labeled as positive, resulting in severe class imbalance (*i.e.,* a $<$ 1:2,600 positive:unlabeled ratio). If it were assumed that only 10% of geothermal systems have been identified, and that adding these undiscovered systems to the analysis results in 2,780 positive cells, the class imbalance would still be severe ($<$1:260). The problem with severe class imbalance can be illustrated by considering a simple model that predicts every cell as negative (assuming most unlabeled cells are negative) has an accuracy (Eq. 1) of $>$ 99%, even though that model predicts that no geothermal systems exist. In other words, this highly accurate model provides no insight into where geothermal systems exist.

2. Methods

With consideration for the challenges detailed in Section 1.2, we seek to develop an approach for the minimally biased modeling of geothermal resource favorability. Although additional data have been collected since the 2008 assessment, we choose to use the data of Williams and DeAngelo (2008) to allow for a direct comparison between the past assessment methods and the machine learning approaches developed herein. In the remainder of this section, we describe the data sets selected as features, briefly detail the selected machine learning algorithms, and define the two training strategies for addressing the positive-unlabeled labels and severe class imbalance of the data. Then, we describe why the F1 score is selected as the performance metric, 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. Feature Selection

We select only one feature from any type (Section 1 lists possible features) because Williams and DeAngelo (2008) only selected one feature of any type in their modeling. Selecting only one feature of each type also serves to reduce correlation between the selected feature sets. Under these criteria, the features for distance to faults and stress were self-selected because they were the only features of their respective types. For heat flow and seismic activity, we chose the features for which we had the clearest understanding of their development. Respectively, we select the heat flow map of Williams et al. (2007) and density of epicenters for seismic events $\geq$ M3 within a 4-km radius. For distance to magmatic activity, we select the most general feature of the type, distance to all Quaternary magmatic activity regardless of composition. These five datasets are georeferenced and published in the data release that accompanies this manuscript (Mordensky and DeAngelo, 2023). As is common practice in data-driven methods, we standardize the data (*i.e.,* from each data point, subtract the sample mean and divide by the sample standard deviation) prior to the application of each machine learning algorithm (see generally Burkov, 2019a).

We note that none of the 28 models produced in the 2008 geothermal resource assessment used more than four features for any one model, whereas we select five features by including every feature type; therefore, we reproduce the methods from the 2008 assessment to create five-feature models for weight-of-evidence and expert decision-dependent logistic regression to provide a direct comparison between the strategies from 2008 and the new data-driven machine learning strategies.

2.2. The Algorithms Considered

Here, we describe the four machine learning algorithms to be used (*i.e.,* logistic regression, XGBoost, SVMs, and an ANN) for comparison with the methods from the 2008 USGS geothermal resource assessment. We choose these four data-driven, machine learning algorithms for several reasons. We select logistic regression because Williams and DeAngelo (2008) also used this algorithm, albeit with expert decisions (*e.g.,* binning the data), thereby providing insight into how this machine learning algorithm can change its output when biased by expert decisions. Additionally, selecting logistic regression provides a linear algorithm for comparison with weight-of-evidence, the other linear method from the 2008 geothermal resource assessment. We also select three general-purpose classifiers (*i.e.,* XGBoost, SVMs, and ANNs; see generally Fernández-Delgado et al., 2014; Chollet, 2021). We choose these three non-linear machine learning algorithms because, when compared to each other, they rely on fundamentally different frameworks to produce decisions and thereby provide a contrast between common non-linear algorithms representing a range of complexity. Hence, selecting four machine learning algorithms for the current analysis expands our perspective on the behavior of machine learning when used with the geothermal data. More details and reference information are provided for each algorithm in the subsequent four subsections.

2.2.1. Logistic Regression

With its initial introduction in Berkson (1944) and subsequent developments in the years that followed (*e.g.,* Berkson, 1951), logistic regression remains one of the older and simpler algorithms in machine learning. At its core, logistic regression fits the input feature set(s) linearly to the logit of *Probability*, which is then transformed to *Probability* with the logit function as summarized in Eq. 5 (Fig. 2):

$$Probability = \frac{e^{(\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_n x_n)}}{1+\ e^{(\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_n x_n)}} \tag{5}$$

in which the coefficients, β_0, β_1, β_2, …, β_n, are empirically fit, and x_1, x_2, …, x_n are the input features (see Berkson (1944) for complete details). Consequently, the computational requirements of logistic regression scale linearly with additional training data.

A decision threshold (often probability = 0.5) defines classification predictions (*e.g.,* 1 or 0, Yes or No, Geothermally Favorable or Not Geothermally Favorable) above and below that decision threshold (see generally Fernández et al., 2018). Herein, we use the common 0.5 decision threshold with logistic regression and optimize the F1 score using two hyperparameters, the inverse regularization strength and the class weight. The inverse regularization strength hyperparameter inversely

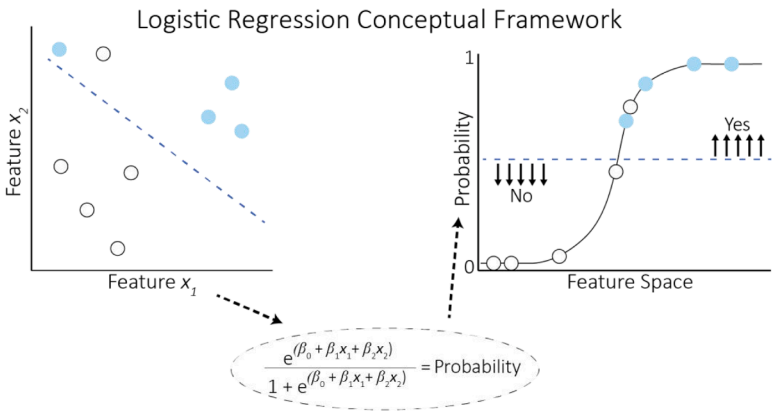


Fig. 2. Conceptual framework for logistic regression (schematic shows only two features for illustrative purposes, but the concept easily extends to *n* features through Eq. 5). The dashed blue line represents a 0.5 probability threshold (a common choice in the machine learning community). The solid, blue circles are examples of a positive label. The hollow, black circles are examples of a negative label, so the hollow circle above the threshold would be a false positive. Solid arrows indicate the classification prediction dictated by the chosen threshold. Probability values range between 0 and 1, and a normal score transform of these values is used in this manuscript for plots of favorability (see Section 2.5).

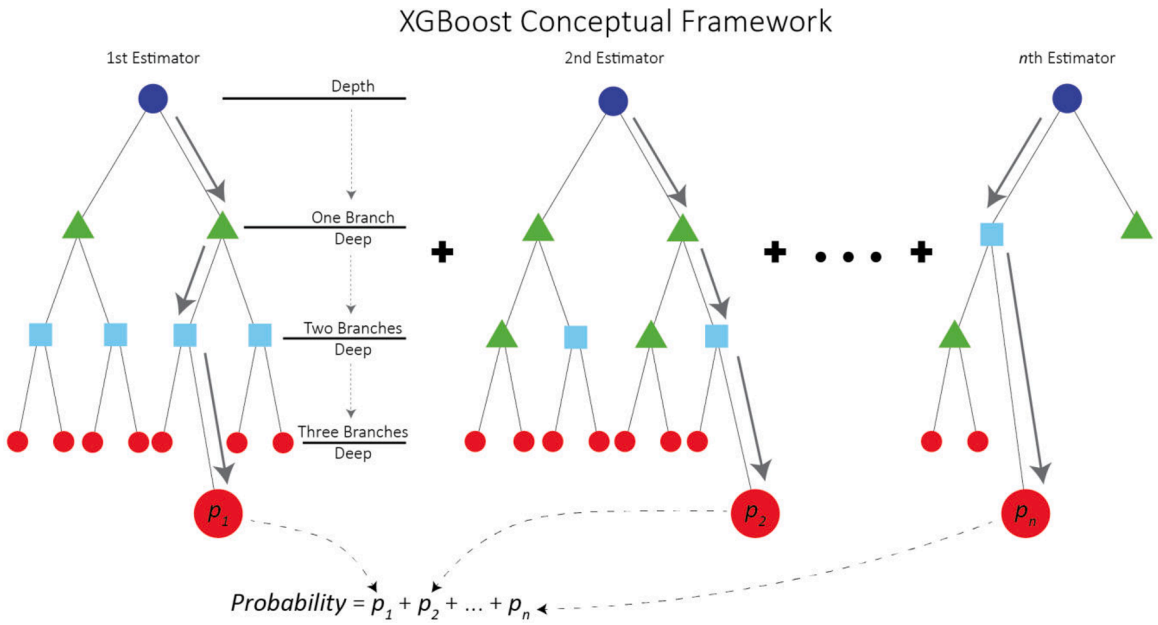


Fig. 3. Conceptual framework for XGBoost. This figure depicts three of *n* estimators (*i.e.,* trees) in an XGBoost classifier. For each cell in a map, a probability value is computed for each estimator, given by its own path (*e.g.,* solid grey arrows) from the root node (purple circle) through the branch nodes (green triangle or blue square), each with a condition dependent upon a feature value, differing between branches and estimators. Ultimately, a cell arrives at an end node (red circle). Each end node has an assigned probability (*e.g.,* p_1, p_2,…, p_n) for a particular class of predictions (*i.e.,* positive sites in this study) found during fitting. The final classification at each map location is predicted by the summation of the probability values across all of the estimators (*e.g.,* dashed black arrows). The final probability values are normal score transformed to produce favorability maps for comparison between approaches (see Section 2.5).

correlates with the propensity of the algorithm to overfit without regularization. The lower the optimal inverse regularization strength hyperparameter, the greater the degree of regularization used to prevent overfitting. The class weight hyperparameter is a means to correct for class imbalance. The greater the optimal class weight, the greater the emphasis the model imparts on correctly identifying positive labels over non-positive labels. Misclassification of the majority class occurs more frequently as the minority class receives greater class weighting (an example of which is provided in Fig. 2). We leave the other parameters of logistic regression at the default values found in the Python's Scikit-Learn module, as they pertain to the specifics of the optimization routine and have only a modest impact on performance (Pedregosa et al., 2011; Kuhn and Johnson, 2013).

2.2.2. XGBoost

XGBoost, first introduced in Chen and Guestrin (2016) uses a process called boosting, that creates a series of decision trees, which are aggregated to produce a single model (Fig. 3). That is, XGBoost produces a series of simple decision trees (*i.e.,* estimators) with each subsequent tree based upon the residuals of the preceding tree. Each subsequent estimator is evaluated and improved from the previous estimator. The amount of information communicated from a previous estimator to a new estimator is called the learning rate. The number of estimators used in the final classifier is determined when additional estimators begin to overfit the training data. Similarly, the depth of the estimators (*i.e.,* the number of branch levels in the trees) is also optimized so as to not overfit the training data. The final node (*i.e.,* the node at the end of a terminal branch) in every estimator has an associated probability value. A sample's prediction value is determined from the sum of the probability values across all of the estimators (see summation of probability values from each estimator in Fig. 3). The computational requirements of XGBoost grow at greater than a linear rate (*i.e.,* greater than that of logistic regression) but less than a quadratic rate (*i.e.,* less than that of SVMs) with each additional sample in the training data.

We optimize four hyperparameters for XGBoost: the class weight, learning rate, number of estimators, and maximum depth of estimators. Class weight in XGBoost differs in exact implementation compared with logistic regression, but this hyperparameter serves much the same purpose: a greater class weight places greater emphasis on accurately predicting positive labels as positives than non-positive labels as negatives. The other parameters are used to maximize prediction performance while also avoiding overfitting (Chen and Guestrin, 2016). We leave the other parameters of XGBoost at the default values found in Python's XGBoost module as they pertain to the specifics of the optimization routine and have only a modest impact on performance (Chen and Guestrin, 2016).

2.2.3. Support-Vector Machines (SVMs)

SVMs provide a more modern machine learning algorithm and an increase in complexity with respect to logistic regression (Cortes and Vapnik, 1995). SVMs classify predictions by finding a hyperplane in an n-dimensional space with n defined by the number of input features (in our case, five input features define a five-dimensional space). The hyperplane serves as a decision boundary (*i.e.,* a maximum margin classifier) that maximizes the n-dimensional distance between data with different predictions (Fig. 4 shows a linear 2-dimensional example). Although finding a hyperplane is a linear process, non-linearities are accommodated through the so-called *kernel trick* (Shalev-Shwartz and Ben-David, 2014b), which uses a non-linear transform to map the data to a new space where a linear decision boundary is found. Given their framework, SVMs do not provide a probability like logistic regression, but instead directly provide a classification prediction and the distance from that observation (*i.e.,* data point) to the decision boundary. SVMs work well for smaller datasets (*i.e.,* thousands of samples or less), because the computational requirements grow quadratically with each additional sample in the training data (Chapelle, 2007); hence, SVMs are less efficient for large datasets.

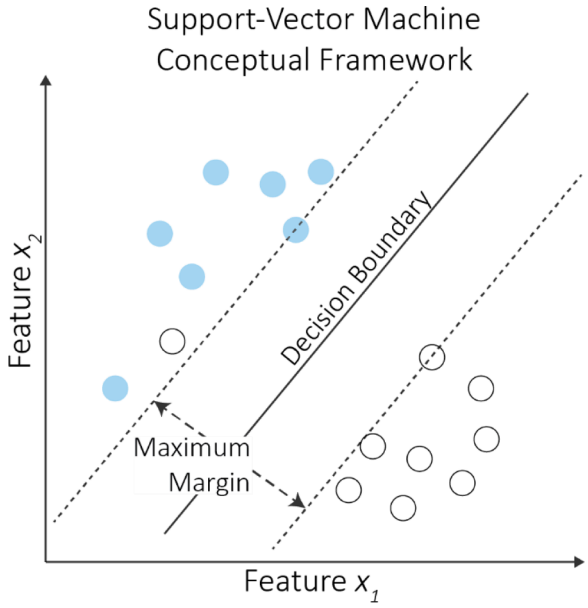


Fig. 4. Conceptual framework of an SVM showing a simple two-feature (x_1 and x_2) example, which mathematically generalizes to higher dimensions using hyperplanes. The solid, blue circles are examples of one label. The hollow, black circles are examples of another label. The decision boundary (*i.e.,* the maximum margin classifier), which maximizes the distances to the nearest examples (*i.e.,* data points or support vectors) of differently classified predictions, is a solid black line. Distance from the decision boundary indicates the confidence associated with a prediction. Note that this example SVM misclassifies one hollow, black sample as that of the solid blue sample. Distance between the dashed black lines is the maximum margin. We normal score transform the distance and direction between each example (*i.e.,* each cell) in the n-dimensional space and decision boundary (positive distance on the positive side of the boundary, and negative distance on the negative side) to produce favorability maps for comparison between approaches (see Section 2.5).

We utilize an SVM with the radial basis function (RBF) kernel. Like with logistic regression, with SVMs, we optimize the inverse regularization strength and class weight. We also add the kernel parameter gamma as a third hyperparameter to optimize. Although not implemented identically between the algorithms, the influence of inverse regularization strength and class weight on the behavior of SVMs is similar to that of logistic regression (Section 2.2.1). The kernel parameter gamma controls how the *kernel trick* is applied; hence, gamma controls the non-linear complexity of the decision boundary hyperplane. The higher the gamma, the more complex the decision boundary, and, therefore, a greater likelihood of overfitting. We leave the other parameters of SVMs at the default values found in the Python's Scikit-Learn module, as they either do not apply to the specific form of SVM used (*e.g.,* apply only to other kernel choices) or have minimal impact on performance (Pedregosa et al., 2011; Kuhn and Johnson, 2013).

2.2.4. Multilayer Perceptron Artificial Neural Network

ANNs operate by passing feature data through a series of activation functions in nodes of varying interconnectivity to transform the input feature data into a prediction (*i.e.,* data can be combined and recombined multiple times in several layers). That is, an ANN is a function containing several layers with each layer containing nodes (Fig. 5). The initial layer (*i.e.,* the input layer) has one node for each feature. The output layer commonly has only one or two nodes but can consist of additional nodes depending on how many types of predictions (*e.g.,* number of classes) are being made with a model. With the data from the 2008 geothermal resource assessment, only one node in the output layer is used for probability. There are an N_h number of hidden layers between the input and output layers with N_h being a hyperparameter. Each hidden layer contains M_h nodes with M_h also serving as a hyperparameter. Each node contains an activation function in which the input feature value is transformed and passed forward to the nodes of the next layer (see generally Burkov, 2019b).

In addition to the N_h number of hidden layers and the M_h number of nodes in a hidden layer, we optimize four other hyperparameters in the ANN: class weight, learning rate, epochs, and batch size. Class weight in an ANN differs in exact implementation compared with logistic regression, XGBoost, and SVMs, but this hyperparameter serves much the same purpose: a greater class weight places greater emphasis on accurately

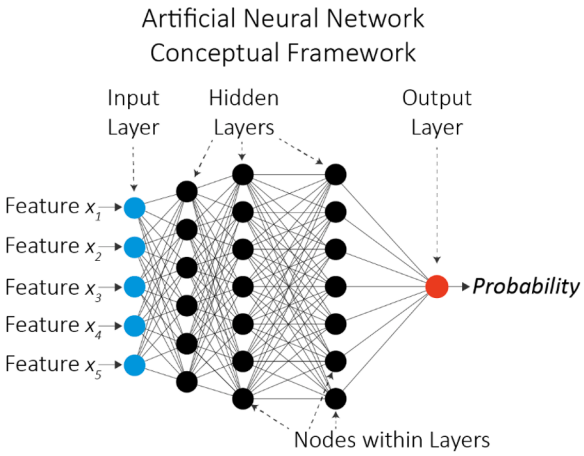


Fig. 5. Conceptual Framework for an ANN with five input features requiring five nodes at the input layer (blue), three hidden layers (black), and one node at the output layer (red). The final probability values are normal score transformed to produce favorability maps for comparison between approaches (see Section 2.5).

predicting positive labels (*i.e.,* known geothermal systems) than non-positive labels (*i.e.,* unknown resource classification). Batch size refers to the number of examples used to fit the ANN at any one time during fitting. Epochs refer to the number of propagations of a batch through the ANN during fitting. The learning rate is the rate at which the model fits to new training data (see generally Geron, 2017). The output layer uses a sigmoidal activation function to transform the values of the final hidden layer into a probability value ranging from zero to one with a threshold probability of 0.5 to separate positive and negative predictions. We set other parameters of the ANN to their commonly accepted default values (*i.e.,* a rectified linear unit [*i.e.,* ReLU] serves as the activation function for the input and hidden layers; binary cross-entropy serves as the loss function with an accuracy metric; Adam serves as the optimizer; early stopping is employed if the model fails to improve after 100 epochs; see generally Chollet (2015)). Whereas logistic regression, SVMs, and XGBoost are forms of shallow learning, an ANN is a form of deep learning for which *depth* refers to the additional interconnected hidden layers in the model. The computational requirements for an ANN are dependent upon the depth and node count of the ANN.

2.3. Addressing the Class-Imbalance and Unlabeled Examples

Because geothermal systems are sparse, most unlabeled locations are negative. In an effort to address the severe class imbalance and positive-unlabeled classifications in this dataset, we experiment with two undersampling training strategies that treat the non-positive class as negative during the training and then as unlabeled during testing and performance evaluation. These two training strategies are: 1) the *single strategy*, in which algorithms are fit with all the available training data, and; 2) the *ensemble strategy*, in which the majority class (*i.e.,* that of the unlabeled cells) is subdivided into four datasets for training and the sub-models fit from those data subsets are averaged into one model. We forego modeling with the ensemble training strategy with the ANN because of the anticipated limited expected gain at substantial computational cost (see Section 4.1.2).

In order to properly undersample, both strategies require an estimate of how many identified and undiscovered geothermal systems exist in the study area so that the undersampling adheres to the expected underlying natural distribution. To estimate the number of undiscovered systems, we use the estimate of undiscovered power potential from Williams et al. (2008). Mean power generation of the identified systems is estimated using Eq. 6 below. Assuming the same average will hold true for undiscovered systems, the number of undiscovered systems can be computed from the estimated undiscovered power potential by Eq. 7. Williams et al. (2008) estimated the mean power potential from identified geothermal resources as 9,057 MWe, but also provided a range of estimates with a 95% probability that these have at least 3,675 MWe to a 5% probability that these have at least 16,457 MWe. Similarly, Williams et al. (2008) estimated the mean power potential from undiscovered geothermal resources as 30,033 MWe and provided a range of estimates with a 95% probability that these undiscovered resources have at least 7,797 MWe to a 5% probability that they have at least 73,286 MWe. The total number of geothermal systems is then found by summing the number of identified systems and the number of undiscovered systems (Eq. 8).

$$\textit{Average Power Generation of a System} = \frac{\textit{Power Generation of Identified Systems}}{\textit{Number of Identified Systems}} \tag{6}$$

$$\textit{Number of Undiscovered Systems} = \frac{\textit{Total Undiscovered Power Potential}}{\textit{Average Power Generation of a System}} \tag{7}$$

$$\textit{Total Number of Geothermal Systems} = \textit{Identified Systems} + \textit{Undiscovered Systems} \tag{8}$$

Considering the power production estimates at 95% and 5% probability in Williams et al. (2008), we estimate a range of 760 – 1,314 conventional geothermal systems exist in the western United States. Herein, we use the mean estimate of 1,040 systems to estimate a natural class imbalance of 1:700 to compare algorithms and training strategies; however, in addition to the mean estimated class imbalance, we also train models for logistic regression and XGBoost using both training strategies with the estimated natural class imbalance derived from the 95% and 5% probability estimates in Williams et al. (2008) (*i.e.,* evaluated the class imbalances 1:955 and 1:550, respectively). Evaluating the range of class imbalance estimates allows us to gauge how model performance responds to changes in this estimate of positive-negative natural class imbalance, the results of which are presented in Appendix C.

Each machine learning algorithm employs a train-test split, in which 80% of the data are used for training and 20% are used for testing, to evaluate the performance of the training model (Fig. 6). This split is random (*i.e.,* the training and testing data are randomly sampled from the data), and to prevent an unfortunate split that results in a poor model, this procedure is repeated 120 times using the USGS supercomputers referred to as YETI, DENALI, and TALLGRASS (Falgout and Gordon, 2021; Falgout et al., 2021a; Falgout et al., 2021b). The optimal hyperparameters are then averaged to train the final resulting model from a single train-test split and predict geothermal favorability for all available data.

Within each iteration of the 120 train-test splits, the training data are further split into smaller partitions (*i.e.,* folds) for custom stratified *k*-fold cross validation. In *k*-fold cross validation, one of the folds is set aside and the remaining folds are used to train a model, and the performance of that model is then evaluated with the initial fold that was set aside (see generally Burkov, 2019a). This process is repeated *k* times until every fold has evaluated the model fit by the other folds. Then, the performance of all the folds is averaged. The *stratified* in stratified *k*-fold cross validation means that the positive labels are evenly distributed amongst the folds. In this study, we use five folds as is common in machine learning practice to avoid overfitting and underfitting a model (see

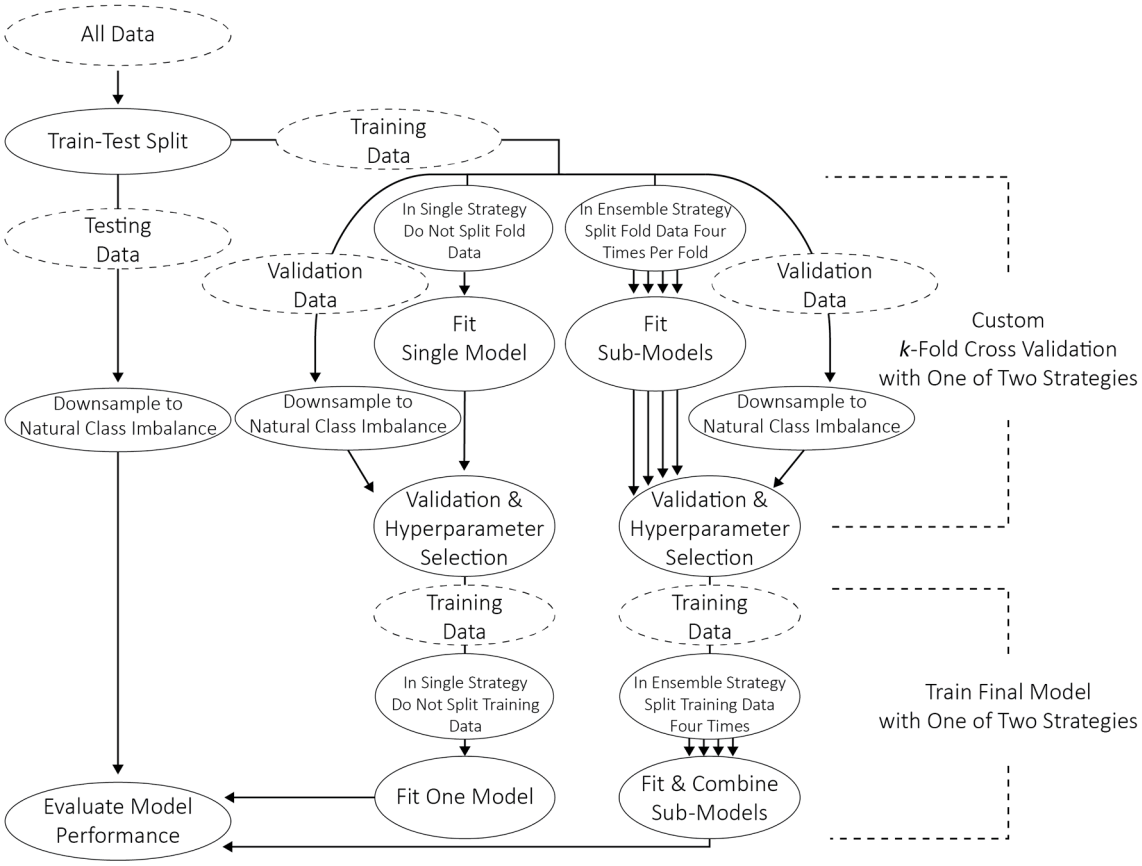


Fig. 6. Training strategies to address class imbalance. During *k*-fold cross validation and the final training of a model, one of two training strategies (*i.e.,* single or ensemble) is pursued. The *single* strategy fits a single model with the remaining four fifths of the folds. The *ensemble* strategy splits the unlabeled data within the remaining four fifths of the folds to create four subsets of the data so that each subset approximately has the estimated 1:700 positive:negative natural class imbalance for a 2-km-by-2-km grid of the western United States. A sub-model is then fit to each of these subsets of data and the sub-models are evaluated in aggregate. Solid circles identify steps in the pipeline. Dash circles identify groups of data.

generally Burkov, 2019a).

In both training strategies, the testing data and the data in the fold used during validation are randomly undersampled from the class imbalance of the data set (known positive samples to unlabeled samples is < 1:2,600) to the estimated natural class imbalance (1:700). The two training strategies differ in how the data in the remaining folds (*i.e.,* the folds not set aside for validation) are used for training a model. With the single strategy, the samples in the majority class in the remaining folds are used to train a single model. With the ensemble strategy, the data from the majority class (*i.e.,* the unlabeled cells) from the remaining folds are randomly distributed into smaller subsets. Each subset has approximately the expected natural class imbalance with each subset receiving all the known positives from the training folds; therefore, the number of subsets created is found by Eq. 9.

$$\textit{Number of subsets of data created in the ensemble strategy} = \frac{\textit{Class Imbalance of Dataset}}{\textit{Natural Class Imbalance}} \tag{9}$$

Hence, with the data in this study, the ensemble strategy creates four subsets of data per fold. A model is then fit to each of these subsets and the final predictor is the average of these models, resulting in an ensemble predictor.

To evaluate model performance, a final model is trained using all training data. At this point, the algorithms using the single training strategy produce a single model, whereas the algorithms using the ensemble training strategy again partition the training data into subsets to achieve the natural class imbalance before training separate sub-models, which are then averaged into one model to predict geothermal favorability.

2.4. Performance Metric Selection

Bekker and Davis (2020) recommend the F1 score as the most appropriate performance metric to use for binary positive-unlabeled classifications like that found in the geothermal data used in this study. Additionally, the F1 score accommodates for class imbalance better than other performance metrics (*e.g.,* accuracy; Guo et al., 2008). Hence, we select the F1 score (Eq. 4) as the performance metric for hyperparameter optimization and model evaluation.

2.5. Comparing Model Results

The F1 score penalizes false positives and false negatives; hence, the F1 score obtains a maximum value of 1 when all positive locations are identified as positive and all unlabeled locations are identified as negative. Ideally, lower F1 scores reflect a model with poorer performance. However, we note that with positive-unlabeled data, we cannot

Table 1
Measures of feature importance by approach. Model-agnostic measures are bolded font. Model-gnostic measures are in normal font. Abbreviations: SHAP: SHapely Additive explanation, ROCAUC: Receiver Operating Characteristic/Area Under the Curve.

Weight-of-Evidence	Logistic Regression	XGBoost	Support-Vector Machines	Artificial Neural Network
Information Value	**F1 Score Sensitivity** **ROCAUC Sensitivity** **SHAP** Coefficients	**F1 Score Sensitivity** **ROCAUC Sensitivity** **SHAP** F Score Weight Cover	**F1 Score Sensitivity** **ROCAUC Sensitivity** **SHAP**	**F1 Score Sensitivity** **ROCAUC Sensitivity** **SHAP**

be certain that an unlabeled location is not a true positive (*i.e.,* a true positive at an unlabeled location contributes to F1 as a false positive), biasing the F1 score. Other performance metrics (*e.g.,* accuracy, precision, recall) have been found to be even less adequate for positive-unlabeled data (Bekker and Davis, 2020).

In addition to the obscurity of F1 scores imparted by the positive-unlabeled data, F1 scores were not used for construction of the 2008 geothermal resource assessment models. Instead, the 2008 assessment predicted relative geothermal favorability for every cell relative to other cells in the 2-km-by-2-km grid without providing a decision threshold for classification. Because F1 scores are not available for the 2008 models, we require a different means to compare all nine models.

Another impediment for comparing the nine approaches is presented by their different units of prediction. Specifically, the expert decision-dependent logistic regression and weight-of-evidence methods predict relative probability on a custom scale (*i.e.,* the posterior probability over the prior probability so that the predicted value represents the number of times greater than simple random chance that a geothermal system is likely to be present), while the logistic regression without expert decisions, XGBoost, and the ANN algorithms predict probability values on a different (i.e., zero-to-one) scale, and SVMs do not supply a probability value but instead produce a distance to a decision boundary as a measure of relative certainty for every prediction.

In order to surmount the challenges in comparing these different model approaches, we propose new mechanisms of comparing the predictions for the different approaches presented in this study. First, the predictions of an approach are normal score transformed (see generally Pyrcz and Deutsch, 2018) so that the different types of model predictions are converted to a common scale since the units from the predictions vary between the different approaches. In particular, the normal score transform provides a quantile-to-quantile transform to a standard normal distribution with a mean of zero and a variance of one. We define these normal score transformed predictions as comparative favorability (*i.e.,* a relative measure of the predicted presence of geologic conditions believed to be associated with the presence of a geothermal system). This quantile-to-quantile transform allows easy comparison of where methods agree on most and least favorable locations using cross-plots and allows favorability maps to be plotted in the same color range for ease of comparison. Because all predicted points are normal scored, plotting the histogram of unlabeled and positive samples allows an examination of how different the known positives are from the larger set of unlabeled data, which have a distribution that is nearly normal with mean zero and variance one.

2.6. Measures of Feature Importance

For every modeling approach, we evaluate the relative importance of each input feature in making predictions (*i.e.,* heat flow, seismic event density, distance to magmatic activity, distance to a fault, and maximum horizontal stress). Unfortunately, being different mathematical constructs, every machine learning approach has a different way of estimating feature importance. However, we are still able to compare the relative measures of feature importance using model-gnostic (*i.e.,* approach-specific) and model-agnostic (*i.e.,* not approach-specific) measures of feature importance (Table 1). When possible, we apply several measures of feature importance to explore the variability between measures and develop a more general understanding of feature importance than any single measure would offer. To allow comparison between the different measures, each measure is min-max normalized to a zero-to-one scale.

2.6.1. Feature Importance with Weight-Of-Evidence

Feature importance for the 2008 weight-of-evidence method can be gauged using information value (IV). Information value provides a relative measure for features by measuring the contribution of features' bins and their associated weights with consideration for their association with events (*e.g.,* the presence of a geothermal system) or non-events (*e.g.,* the absence of a geothermal system) as provided in Eq. 10 (see generally Zdravevski et al., 2011):

$$\textit{Information Value of a Feature} = \sum_{i=1}^{h} (WoE_i\ (PoE - PoNE)) \tag{10}$$

in which *WoE* is a weight of a bin for a given feature, *PoE* is the percent of events associated with that bin, *PoNE* is the percent of non-events associated with that bin, where *h* is the total number of bins for a given feature.

2.6.2. Feature Importance with Logistic Regression

Machine learning logistic regression relies upon feature coefficients (see Eq. 5) which provide a means to assess feature importance from the absolute values of the coefficients when training and predicting from standardized data (Berkson, 1944, 1951). The logistic regression method from the 2008 assessment (*i.e.,* the logistic regression method that used binned values) does not permit a direct comparison between the features' coefficients (see Eq. 5) because these features were categorically separated into bins that do not contain standardized data; however, it can be shown that comparing the standard deviation of the bin coefficients for each feature is an analogous measure of feature importance (Appendix A), allowing an evaluation of feature importance from the 2008 logistic regression analysis.

2.6.3. Feature Importance with XGBoost

XGBoost hosts several unique model-gnostic measures of feature importance; these are weight, cover, gain, and F score (not to be confused with the F1 score). Weight refers to the number of instances that feature was used to split the data. Cover refers to the number of observations affected by a split with that feature. Gain refers to how much each feature contributes toward better predictions for a model with consideration for all the splits using that feature. F score considers both the number of splits and the number of correctly classified samples resulting from those splits with that feature (Chen and Guestrin, 2016).

2.6.4. Feature Importance with Model-Agnostic Measures

Model-agnostic measures can be applied to nearly any existing machine learning model. We use three model-agnostic measures in this study (*i.e.,* sensitivity analysis using an F1 score, sensitivity analysis using the area under the receiver operating characteristic curve (*i.e.,* the ROCAUC), and SHapely Additive exPlanation (*i.e.,* SHAP) values).

Sensitivity analysis (also termed permutation importance) provides a model-agnostic measure of feature importance by shuffling the values of

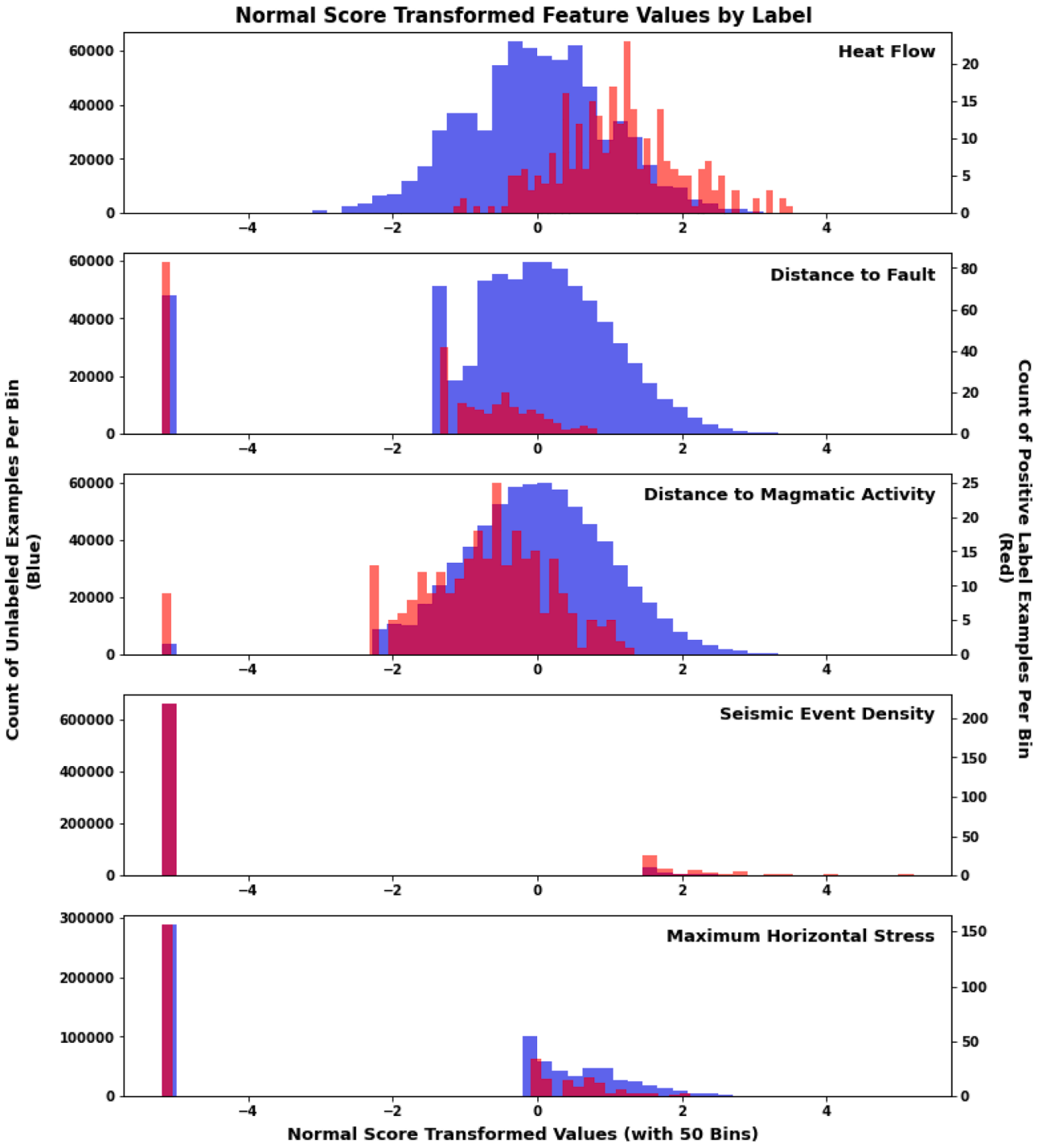


Fig. 7. Histograms of the normal score transformed features. Red represents the distribution of examples with positive labels. Blue represents the distribution of unlabeled examples. Purple appears when the distributions for the two classes overlap.

a single feature while the other features remain unshuffled, using the model to make new predictions, and then comparing these new predictions with the predictions from the originally unshuffled data. By sequentially completing this process through all the features, sensitivity analysis is able to gauge the magnitude of the contribution of each feature toward a prediction. Sensitivity analysis compares the predictions from the shuffled data and the unshuffled data using the same performance metrics used to evaluate data-driven models (*e.g.,* accuracy, precision, recall, F1 score; see generally Pedregosa et al., 2011).

When inspecting sensitivity, we use the F1 score and the ROCAUC to provide two measures of sensitivity. The receiver operating curve (*i.e.,* the 'ROC' of ROCAUC) plots the tradeoff between the true positive rate (Eq. 11) and the false positive rate (Eq. 12) over changing classification thresholds (see generally Murphy, 2012). The ROCAUC is the area bounded by the receiver operating curve and a false positive rate of 0.

$$True\ Positive\ Rate = \frac{True\ Positives}{True\ Positives + False\ Negatives} \tag{11}$$

Table 2
Machine learning model performance. Median F1 score values are in bolded font. 95th-percentile values are provided in italicized, bolded font. Mean optimal hyperparameter values are in normal font. One standard deviation for optimal hyperparameter values is provided in italicized font. Abbreviations: F1: F1 Score, Inverse Reg. St.: Inverse Regularization Strength, LR: Logistic Regression, ANN: Single Artificial Neural Network, 95th – 95th percentile value, SD: Standard Deviation.

Strategy & Algorithm	F1	Class Weight	Inverse Reg. Str.			
Single Logistic Regression	**0.036**	258	3			
Single LR 95th / SD	***0.076***	*47*	*11*			
Ensemble Logistic Regression	**0.036**	112	0.0013			
Ensemble LR 95th / SD	***0.064***	*22*	*0.0091*			
Strategy & Algorithm	F1	Class Weight	Learning Rate	*n* of Estimators	Max Depth	
Single XGBoost	**0.023**	206	0.22	61	3	
Single XGBoost 95th / SD	***0.056***	*21*	*0.14*	*19*	*1*	
Ensemble XGBoost	**0.025**	59	0.32	5	4	
Ensemble XGBoost 95th / SD	***0.045***	*15*	*0.32*	*2*	*2*	
Strategy & Algorithm	F1	Class Weight	Inverse Reg. Str.	Gamma		
Single SVM	**0.016**	575	2	0.032		
Single SVM 95th / SD	***0.038***	*97*	-	*0.042*		
Ensemble SVM	**0.011**	156	10	0.010		
Ensemble SVM 95th / SD	***0.035***	*50*	-	*0.023*		
Algorithm	F1	Class Weight	Learning Rate	Hidden Layers	Node Count	Epoch	Batch Size
Single ANN	**0.0177**	285	0.00086	2	35	300	256
Single ANN 95th / SD	***0.0559***	*77*	*0.00247*	*0.50*	*8*	-	-

$$False\ Positive\ Rate = \frac{False\ Positives}{False\ Positives + True\ Negatives} \quad (12)$$

SHAP values provide another model-agnostic measure of feature importance. SHAP values operate similarly to sensitivity analysis at a conceptual level but with some fundamental differences. The SHAP function varies values for every possible combination of feature sets, whereas sensitivity analysis sequentially shuffles only one feature at a time. Also, SHAP measures the differences between predictions and does not rely on a specific performance metric. More specifically, every sample for every feature with consideration for every combination of feature sets is assigned a SHAP value that is the difference between the original and permutated predictions, and the sample SHAP values are then averaged by feature to provide the average feature SHAP values (Lundberg and Lee, 2017).

3. Results

In this section, the input feature data are briefly described, model performance is presented, favorability predictions are provided, and feature importance is given. The input feature and model prediction data are also available in the accompanying data release (Mordensky and DeAngelo, 2023).

3.1. Exploratory Data Analysis

We focus the exploratory data analysis on the differences in the distributions of feature values between the positive and unlabeled cells. A normal score transformed distribution for each is shown as histograms in Fig. 7 and the zero-to-one, min-max normalized distributions are provided in Appendix B (Fig. B1). The normal score transformation is a quantile-to-quantile transform designed to transform data to resemble a standard normal distribution (*e.g.,* unlabeled data distribution for heat flow in Fig. 7), but the abundance of zero values (*i.e.,* there are a large number of minimum values = 0) in the pre-transformed features for distance to nearest fault, distance to nearest magmatic activity, seismic event density, and maximum horizontal stress result in an otherwise abnormally high occurrence of lowest values in the normal score transformed spaces for these features. Differences between the distributions of positive and unlabeled data are strongest where peaks in the distribution are distinct. When different, it can be inferred that the corresponding data type has value for separating positives from unlabeled data when used as a predictor. By this measure, heat flow has the greatest difference between positive and unlabeled cells with distance to the nearest fault and distance to the nearest magmatic activity having the second and third, respectively, greatest differences between unlabeled and positive cells in normal score transformed space. With seismic event density and maximum horizontal stress, zero is the most common pre-transformed feature value.

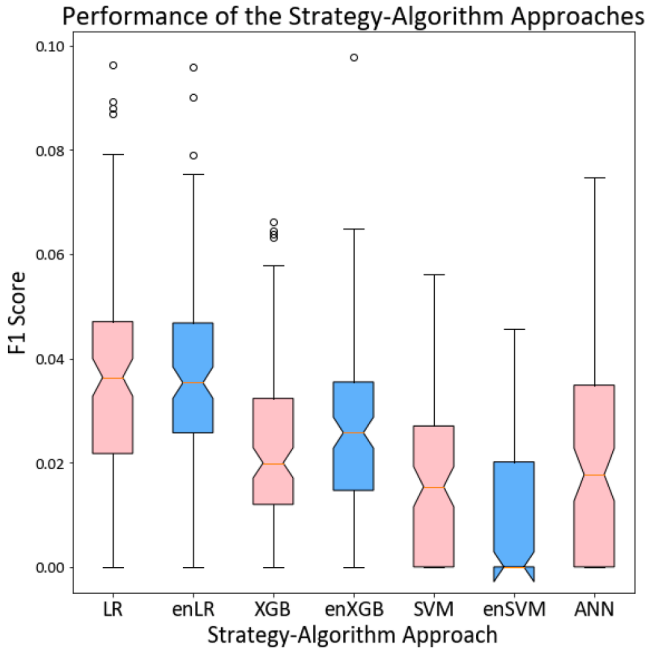


Fig. 8. Box-and-whisker plots of F1 scores for test data for each machine learning approach from the 120 train-test splits. The single strategy approaches are in red, and the ensemble strategies are in blue. Boxes extend from the first quartile (Q1) to the third quartile (Q3) with a notch and line at 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. Abbreviations: LR: Single Logistic Regression, enLR: Ensemble Logistic Regression, XGB: Single XGBoost, enXGB: Ensemble XGBoost, SVM: Single Support-Vector Machine, enSVM: Ensemble Support-Vector Machine, ANN: Single Artificial Neural Network.

3.2. Optimal Hyperparameter Values and Model Performance

Optimal hyperparameter values for the machine learning approaches are provided in Table 2, and the performance of the corresponding models across the 120 train-test splits have considerable overlap (Fig. 8) with generally decreasing performance as the complexity of the strategy-algorithm approaches increases, with the exception that the single ANN performs similarly to logistic regression. Although the median F1 scores of the seven machine learning models are similarly low (< 0.04), two important distinctions can be made. First, the simplest algorithm (*i.e.,* logistic regression) has the highest median F1 score compared to that of other algorithms when either strategy is considered. Second, the SVMs and the single ANN have a first-quartile (*i.e.,* 25th-percentile) F1 score of zero; hence, these models are more likely to misclassify known positives than any of the other machine learning approaches. Lastly, we note that varying the expected positive-negative class imbalance from 1:700 to 1:550 and 1:955, respectively corresponding with the 5% and 95% potential resource estimates from Williams et al. (2008), did not substantially change model performance (Figs. C1, C2 in Appendix C), indicating the models resulting from both training strategies are relatively robust to the different estimates of potential resources in Williams et al. (2008).

3.3. Model Predictions

The geothermal favorability maps constructed using the methods from the 2008 geothermal resource assessment (Fig. 9) and the machine learning algorithms (*i.e.,* logistic regression [Fig. 10], XGBoost [Fig. 11],

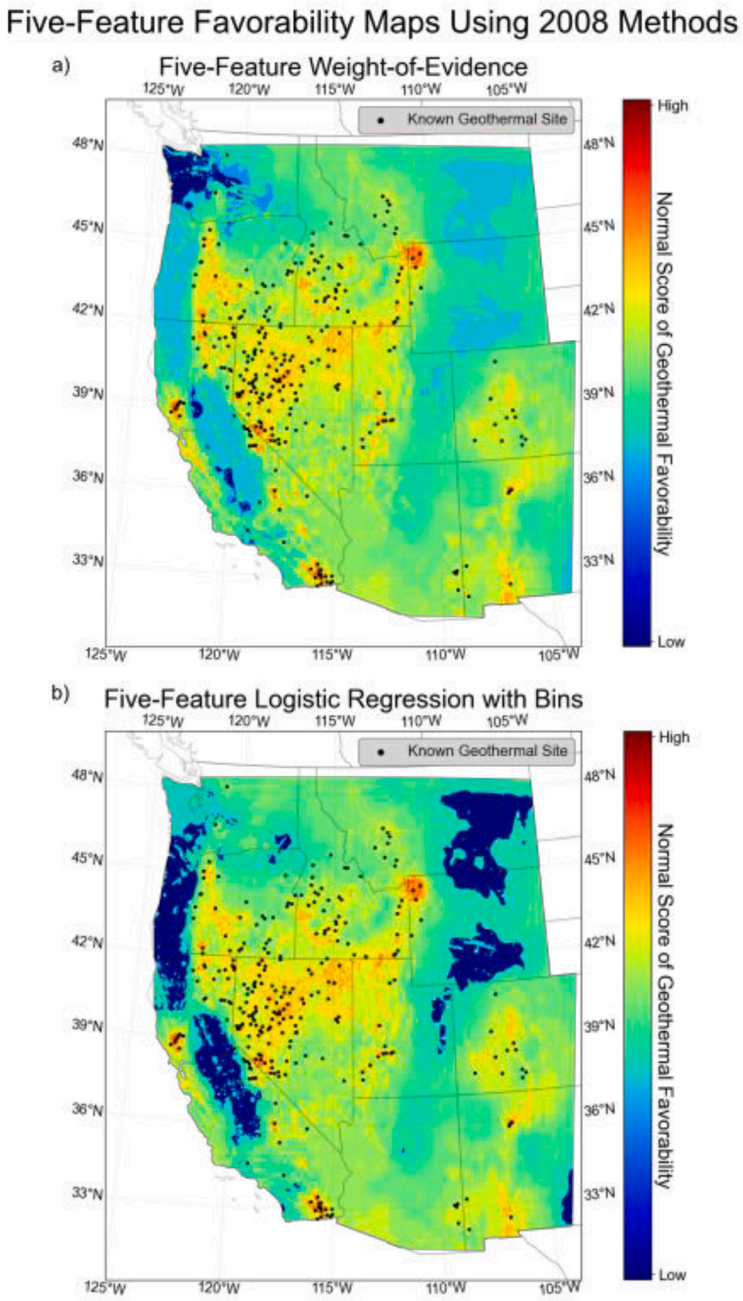


Fig. 9. Geothermal favorability maps of the western United States using the five features selected for training machine learning models in this study and the methods in Williams and DeAngelo (2008): a) weight-of-evidence and b) logistic regression with underlying expert decisions. Favorability is the normal score transform of the predicted probability of occurrence.

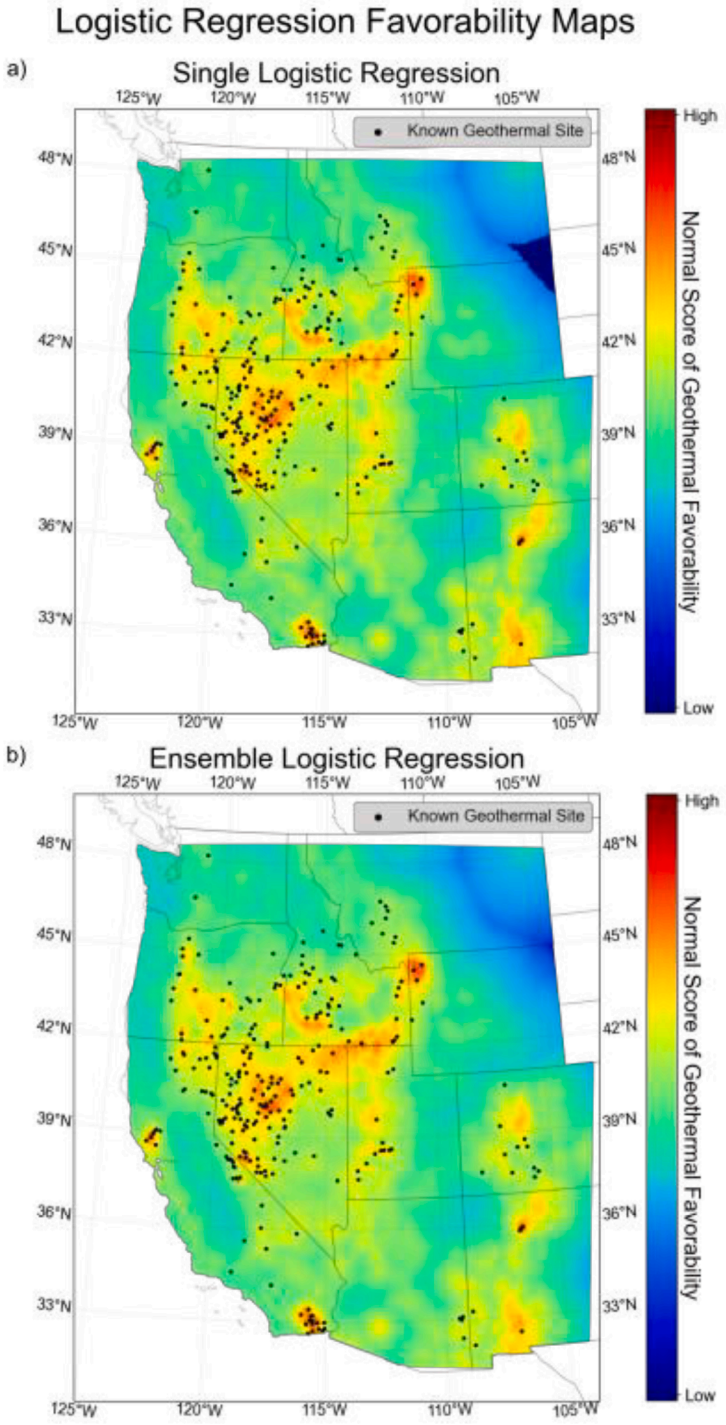


Fig. 10. Favorability maps for modern, machine learning (*i.e.,* without underlying expert decisions) a) single logistic regression and b) ensemble logistic regression. Favorability is the normal score transform of probability.

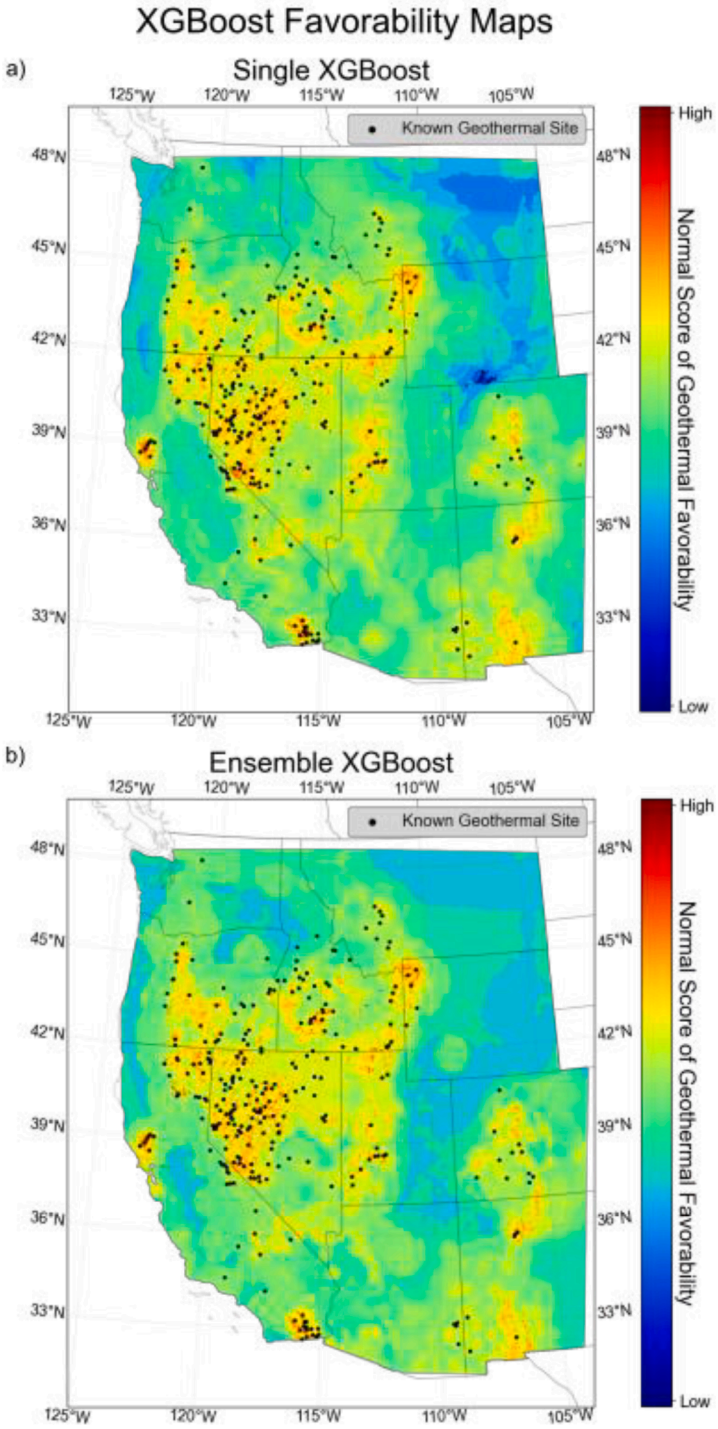


Fig. 11. Favorability maps for a) single XGBoost and b) ensemble XGBoost. Favorability is the normal score transform of probability.

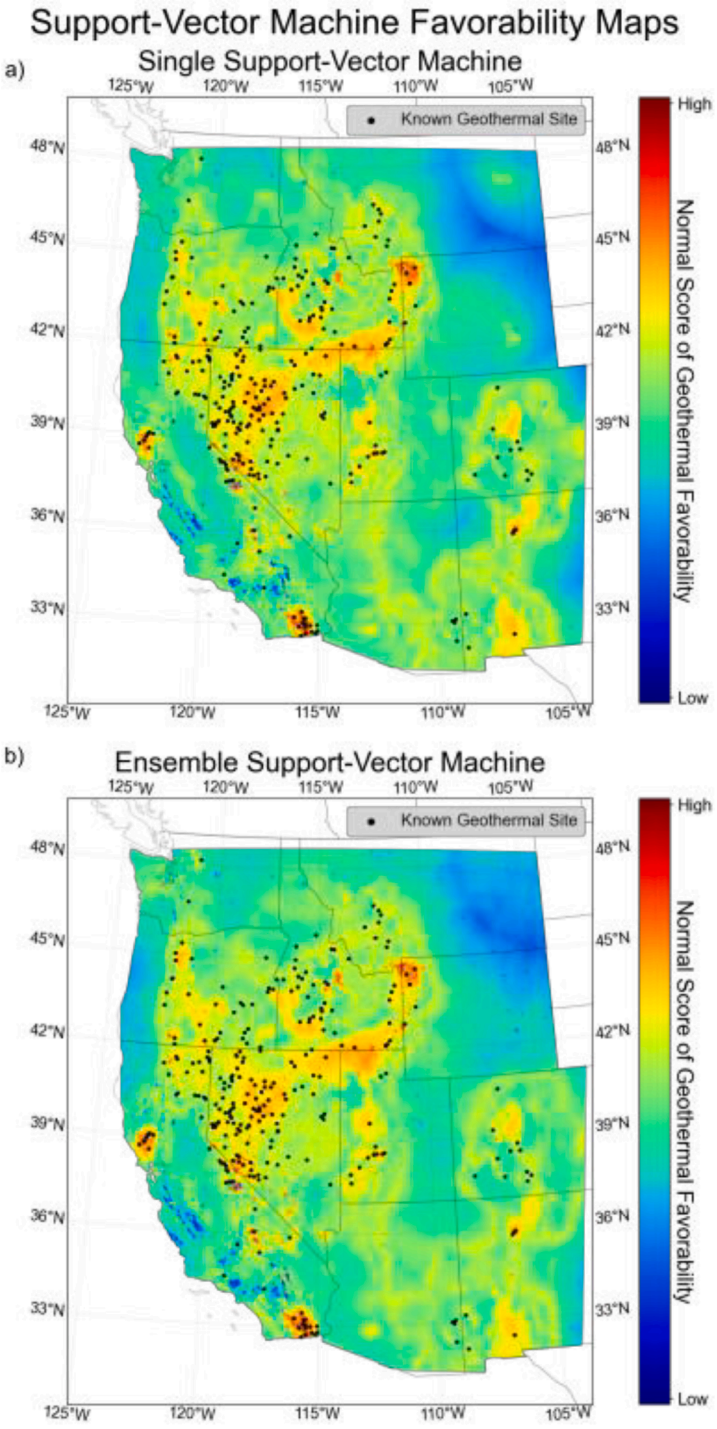


Fig. 12. Favorability maps for a) single SVM and b) ensemble SVM. Favorability is the normal score transform of the *n*-dimensional distance of a cell to the decision boundary in the space defined by the kernel trick. Distance is positive on the positive side of the boundary and negative on the negative side of the boundary.

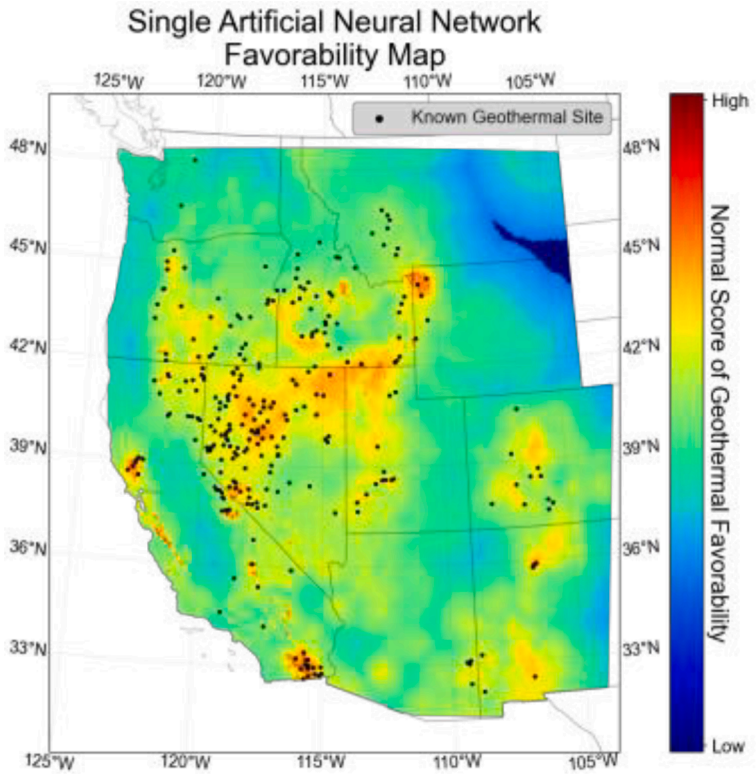


Fig. 13. Favorability map for the single ANN. Favorability is the normal score transform of probability.

SVMs [Fig. 12], the ANN [Fig. 13]; Table C1) generally show a broad agreement in terms of geospatial patterns of regional geothermal favorability, particularly for areas of high geothermal favorability (*e.g.,* as seen by predictions in the Yellowstone, Geysers, and Great Basin regions) with the greatest disagreement between models corresponding to predictions of low geothermal favorability (*e.g.,* eastern Montana, the Central Valley of California, northwest Washington).

3.4. Feature Importance

The relative ranking of different measures of feature importance is mostly consistent across the different approaches. Fig. 14, which summarizes feature importance as the median value from the 120 train-test splits (distributions for each measure shown in Appendix D), depicts the general distribution of importance across the different approaches. In general, heat flow and distance to faults are, respectively, the most and second most important features (*e.g.,* 2008 logistic regression, 2008 weight-of-evidence, ensemble logistic regression, single and ensemble XGBoost). Reciprocally, the feature importance of seismic event density and stress are, respectively, the second least and least important features (*e.g.,* 2008 weight-of-evidence, ensemble XGBoost). Across these different approaches and measures of feature importance, there is one notable exception from these generally observed tendencies; the ensemble SVM ranks seismic event density as the most important feature (Fig. 14, D6) with F1 score sensitivity analysis and SHAP values.

4. Discussion

In this section, we demonstrate that the machine learning algorithms can produce geothermal favorability maps that are generally consistent with those from the 2008 geothermal resource assessment (Williams and DeAngelo, 2008) but do not have the bias implicit to the expert decisions from that assessment (Figs. 9 versus Figs. 10, 11, 12, 13). However, despite this broad agreement, there are distinctions between the results of the different approaches. These distinctions are a product of the differing frameworks between the approaches and their resulting variation in complexity.

4.1. Model Performance

From the perspective of the F1 score, the machine learning approaches appear to perform poorly (F1 score < 0.10). The poor performance can be attributed to two considerations: 1) the quality of the data and 2) the suitability of the F1 for positive-unlabeled data.

The data used in this machine learning study can be considered to have limited quality from the perspective of how representative the data are to true geological conditions. For example, the geospatial data aggregated by Williams and DeAngelo (2008) informed regions of the western United States with varying density, and for heat flow and stress, a radial basis function interpolation populated the cells between known values (*i.e.,* maps of properties varied smoothly between measurement locations). Anomalously high heat flow values (> 120 mW/m^2) were set to equal 120 mW/m^2 to reduce the effect of the convective wells and sampling bias. Today, modern geostatistical approaches offer more robust methods to account for sampling bias (*e.g.,* declustering; see Lindsey et al., 2022). It is also worth consideration that geological conditions do not vary smoothly between measurement locations. Instead, geologic features, like faults, unconformities, and paleotopography, are expected to result in abrupt changes, reducing the accuracy of interpolated values with respect to true conditions.

The F1 score may be considered the most appropriate performance metric for positive-unlabeled data, but as implemented here, the F1 score is still not ideally suited to the role; the F1 score penalizes positive predictions at unlabeled cells as false positives while these cells may indeed be positive. Bekker and Davis (2020) cover several novel performance metrics adapted to positive-unlabeled data (*e.g.,* Lee and Liu,

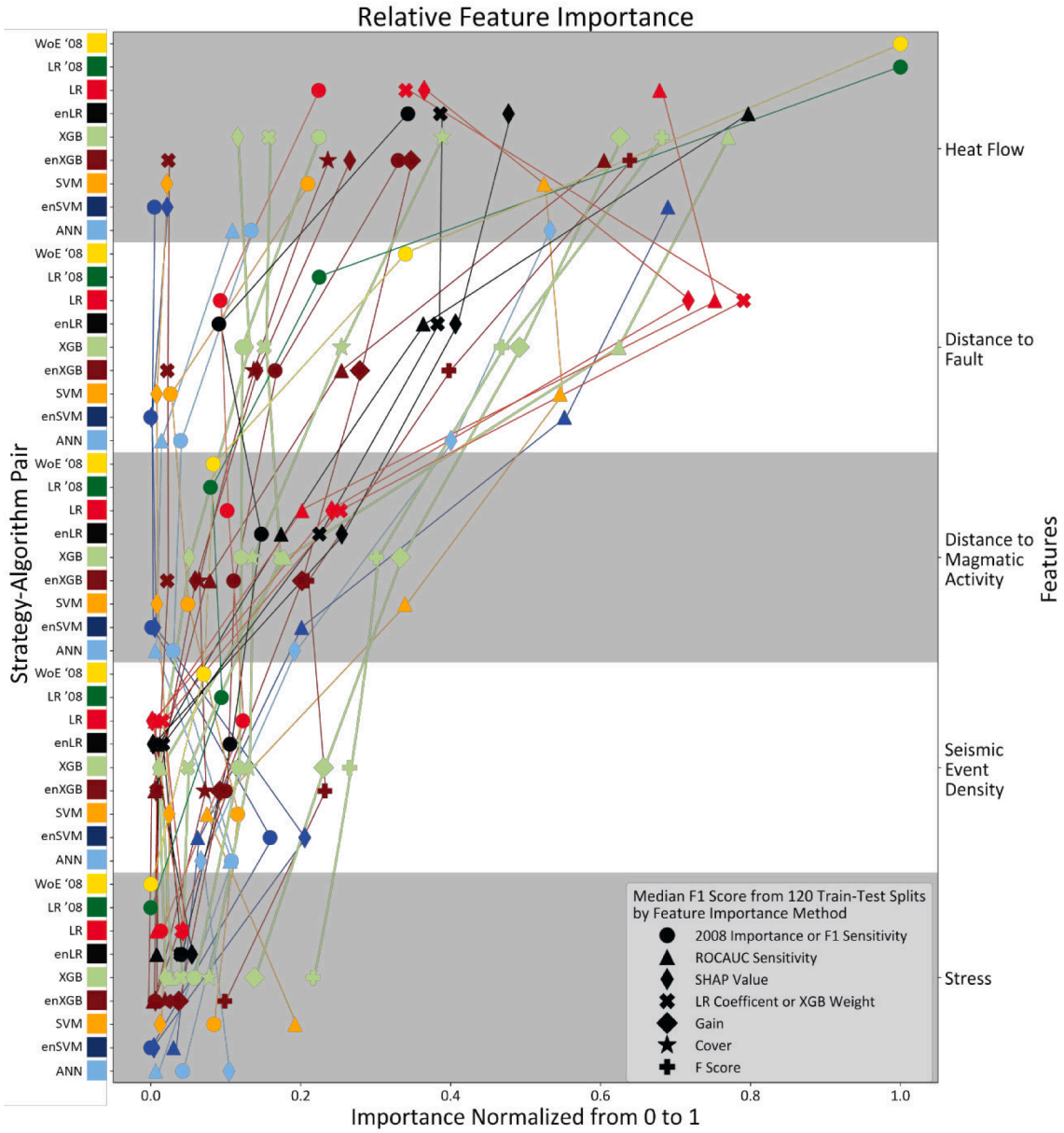


Fig. 14. Median normalized feature importance values from the 120 train-test splits using the different strategy-algorithm approaches. Abbreviations: WoE '08 (yellow): Weight-of-Evidence from the 2008 geothermal resource assessment, LR '08 (green): Logistic Regression from the 2008 geothermal resource assessment, LR (red): Single Logistic Regression, enLR (black): Ensemble Logistic Regression, XGB (purple): Single XGBoost, enXGB (brown): Ensemble XGBoost, SVM (orange): Single Support-Vector Machine, enSVM (dark blue): Ensemble Support-Vector Machine, ANN (light blue): Single Artificial Neural Network, ROCAUC: Area Under the Receiver Operating Characteristic Curve, SHAP: SHapely Additive explanation.

2003), but these novel metrics are not simultaneously designed to also handle the severe class imbalance of geothermal data. An ideal performance metric for positive-unlabeled data with severe class imbalance might reward the prediction of true positives and a limited number of positive predictions from the unlabeled cells (*i.e.,* most unlabeled cells are negative, but not all are) while penalizing the predictions of false negatives and too many positive predictions from unlabeled cells (*i.e.,* most unlabeled cells are negative). To that end, we identify the need for a new performance metric intended to accommodate positive-unlabeled data and severe class imbalance for use in machine learning applications involving the exploration of geothermal systems and other natural resources.

4.1.1. *Relative Model Performance*

Since the methods from the 2008 geothermal resource assessment did not explicitly predict where geothermal systems are favorable, no traditional performance metric can be used to compare the methods from the 2008 assessment to the seven machine learning approaches presented here. Therefore, we present an alternative means to compare model performance; however, we emphasize that the manner of this comparison is not a new performance metric unto itself but only a means to evaluate the performance of the nine different approaches when used with the same data.

To compare the models, we perform a normal score transform on the predictions from each approach. After the transformation, the unlabeled

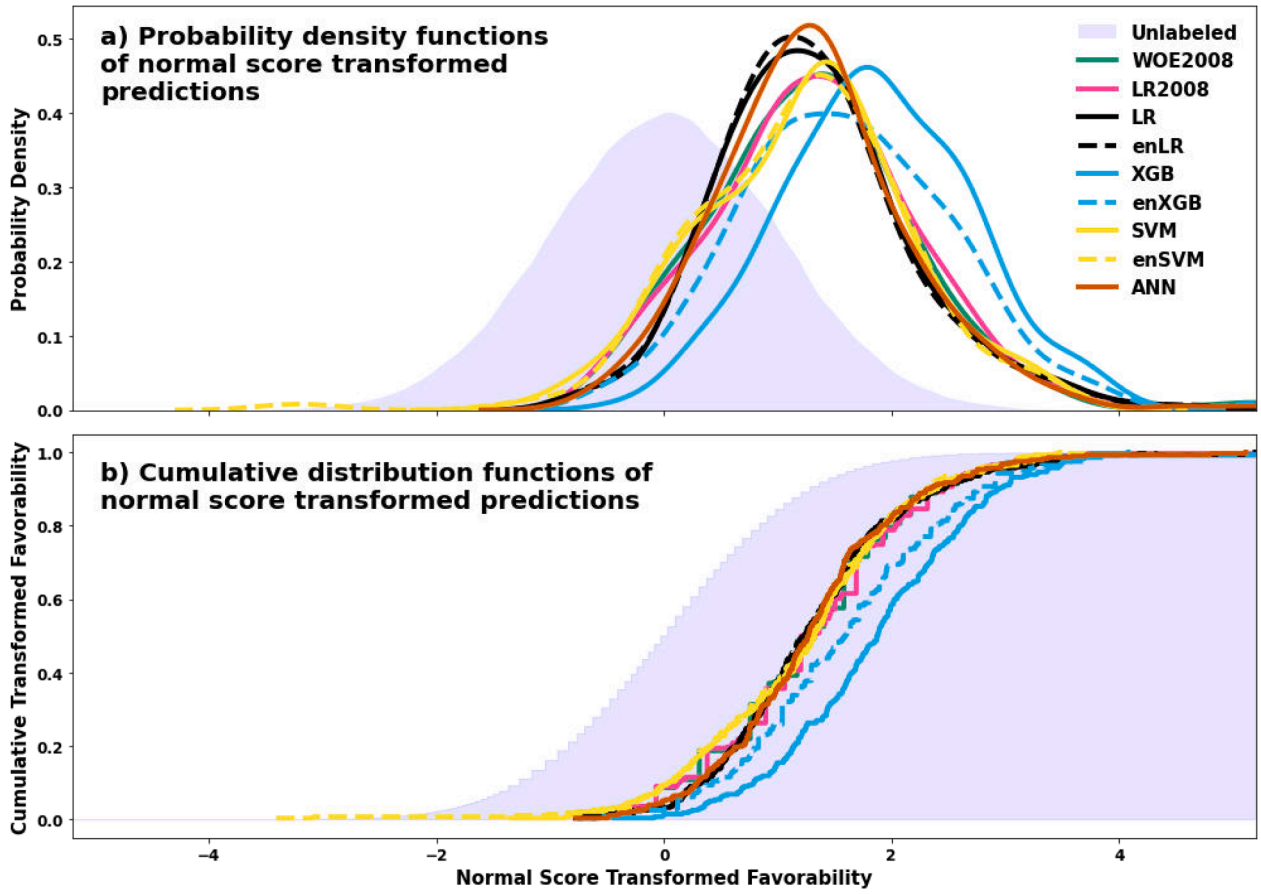


Fig. 15. Comparing normal score transformed predictions between the different approaches using a) probability density functions of normal score transformed predictions and b) cumulative distribution functions of normal score transformed predictions. The shaded blue provides predictions for the unlabeled cells. The lines depict the relative distribution of predictions of positive labels from different approaches. Abbreviations: WoE '08 (yellow): Weight-of-Evidence from the 2008 geothermal resource assessment, LR '08 (green): Logistic Regression from the 2008 geothermal resource assessment, LR (red): Single Logistic Regression, enLR (black): Ensemble Logistic Regression, XGB (purple): Single XGBoost, enXGB (brown): Ensemble XGBoost, SVM (orange): Single Support-Vector Machine, enSVM (dark blue): Ensemble Support-Vector Machine, ANN (light blue): Single Artificial Neural Network.

Table 3
Mean, median, peak values, and variance of normal score transformed predictions as depicted in Fig. 15. Abbreviations: WoE '08: Weight-of-Evidence from the 2008 geothermal resource assessment, LR '08: Logistic Regression from the 2008 geothermal resource assessment, LR: Single Logistic Regression, enLR: Ensemble Logistic Regression, XGB: Single XGBoost, enXGB: Ensemble XGBoost, SVM: Single Support-Vector Machine, enSVM: Ensemble Support-Vector Machine, ANN: Single Artificial Neural Network.

Predictions	Mean	Median	Peak	Variance
Unlabeled	0.00	-0.01	0.06	1.00
WoE '08 Positives	1.31	1.29	1.42	0.84
LR '08 Positives	1.33	1.32	1.36	0.80
LR Positives	1.32	1.25	1.19	0.75
enLR Positives	1.30	1.22	1.12	0.71
XGB Positives	1.88	1.86	1.81	0.72
enXGB Positives	1.61	1.56	1.45	0.88
SVM Positives	1.24	1.27	1.42	0.82
enSVM Positives	1.21	1.24	1.37	0.94
ANN Positives	1.30	1.25	1.28	0.72

cells have nearly a standard normal distribution (Fig. 15). Assuming the models have predictive skill, the positive-labeled cells should have a distribution of predictions with a mean greater than that of the unlabeled cells; hence, higher normal score transformed predictions for the positive cells result in a greater distinction between the positive-labeled and unlabeled cells and better model performance.

Applying the transformation to the predictions reveals that several of the approaches perform similarly (*e.g.,* weight-of-evidence, expert decision-dependent logistic regression, single logistic regression, ensemble logistic regression, single SVMs, ensemble SVMs, and the ANN; Fig. 15). Meanwhile, single XGBoost produces the greatest distinction between known positive and unlabeled samples (Table 3), and ensemble XGBoost predicts the second greatest distinction between known positive and unlabeled samples. Hence, the comparison of the normal score transformed predictions suggests that the simplest non-linear algorithm had the best performance from the different approaches considered, and the other approaches performed similarly to one another despite their varying complexity. The superlative performance of XGBoost with both training strategies suggests that the inherent shape of a decision boundary from XGBoost (*i.e.,* a step function-like boundary due to the decision-tree structure of XGBoost) is more like the true decision boundary for a perfect predictor than the decision boundaries from the other approaches in this study.

4.1.2. *Model Complexity Does Not Improve Performance*

The general agreement of predictions from the different approaches (Figs. 9, 10, 11, 12, 13, 16) is consistent with the agreement of the relative feature importance in the models (Fig. 14) with heat flow and distance to faults as the most important features and maximum horizontal stress and seismic event density as the least important features.

We note that measures that produce feature importance values where ranges do not strongly overlap (*e.g.,* ROCAUC sensitivity in Figs. D1, D2, D3, D4, D5, D6; F score for XGBoost in Figs. D3, D4 in Appendix

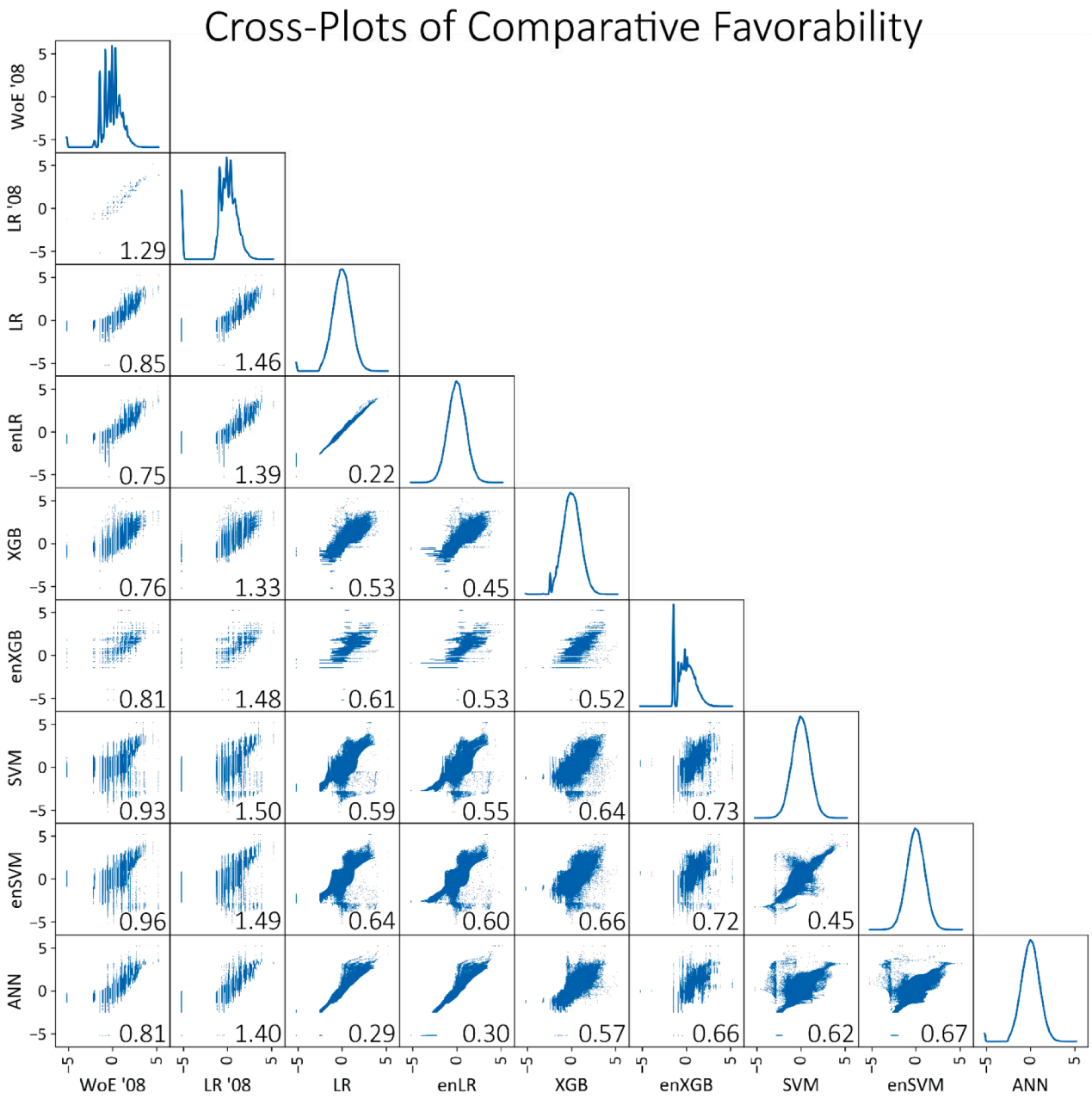


Fig. 16. Cross-plots of predicted normal score transformed favorability at every location for the different approaches (*i.e.,* favorability predictions from Figs. 9, 10, 11, 12, 13). The number in each plot is the root mean square error (sum of square differences at all cells), so low values indicate better cell-by-cell agreement in the favorability maps. The main diagonal shows the histogram of data on each map, which should be a normal distribution of mean = 0 and variance = 1. Because the histograms are a quantile-to-quantile transform, the spikes are a high count of the same value as a result of binning, which also produces regular gaps in favorability values in the cross-plots. Abbreviations: WoE '08: Weight-of-Evidence from the 2008 geothermal resource assessment, LR '08: Logistic Regression from the 2008 geothermal resource assessment, LR: Single Logistic Regression, enLR: Ensemble Logistic Regression, XGB: Single XGBoost, enXGB: Ensemble XGBoost, SVM: Single Support-Vector Machine, enSVM: Ensemble Support-Vector Machine, ANN: Single Artificial Neural Network.

D) provide more confidence in relative feature importance than those with overlapping ranges (*e.g.,* F1 sensitivity analysis). Where present, the overlap presents ambiguity with respect to the reliability of the analyses. This overlap and consequential ambiguity are not likely a result of the measures themselves, but of two qualities in the data from the 2008 geothermal resource assessment. The first, as already suggested by the low F1 scores (Fig. 8; Table 2), is that the features do not sufficiently reflect geological conditions at a fine enough scale showing properties that change rapidly over short distances. The second pertains to the limited number of cells labeled as known positives (*i.e.,* 278) across the several geologically diverse regions of the western United States. Feature importance likely varies between these different geological regions (*e.g.,* the Great Basin, the Cascades, the Rocky Mountains), especially if there are different types of geothermal systems in each region (*e.g.,* deep circulation within thin crust versus magmatic heat source). The regional geologic variability coupled with the high likelihood of regionally skewed distributions of known positives in the train-test splits accentuate the variance in the measures of feature importance.

Neither of the two above potential explanations can account for the anomalous relative ranking of feature importance by the ensemble SVM (Fig. 14, D6). Close inspection of the SVM favorability plots (Fig. 12) finds seismically active locales display lower favorability than seismically inactive areas, suggesting that SVMs predict seismicity to inversely

correspond to geothermal favorability. This relationship is not observed at a nearly similar magnitude with the other approaches (Fig. 14), which generally rank relative feature importance as consistent with the degree of difference between the positive and unlabeled cells in the feature data (Fig. 7).

The unusual magnitude of the negative relationship of seismicity in the ensemble SVM may be a result of the intrinsic complexity of the algorithm itself. When using the radial basis function kernel, the fewer training data in the ensemble approach coupled with the high complexity of the SVM may be insufficient to properly train a well-performing model. That is, although all the machine learning approaches explored in this study are subject to the bias-variance tradeoff, the potential of SVMs to learn a highly complex boundary imparts low bias (*i.e.,* approximation error). As a result, SVMs require more positive samples to achieve a low estimation error (Shalev-Shwartz and Ben-David, 2014a). When using the data from the 2008 geothermal resource assessment, too few known positives result in some SVM models that seem prone to finding a non-linear transform that accentuates seismicity, giving a very different model than is supported by the other machine learning models. Alternatively, the SVM predictor may be highly sensitive to selecting appropriate support vectors from the data set. This task is especially challenging given the class imbalance of the data used in this study. Similarly, the substantial overlap of feature importance in the F1 and ROCAUC sensitivity analyses of the single ANN (Fig. D7) suggest that this deep learning algorithm was also detrimentally affected by its complexity when faced by the simplicity of the feature data. Hence, the highly complex machine learning models may not be as appropriate for data like that from the 2008 geothermal resource assessment as the less complex algorithms. With consideration for these results and the substantial implicit computational burden, we decided to forego completing 120 train-test splits of the ensemble ANN; however, given that the single and ensemble strategies performed similarly for logistic regression, XGBoost, and SVMs, we do not anticipate vastly different predictive skill between a single and an ensemble ANN.

Likewise, we again note that single XGBoost, the simplest non-linear approach, produced the best performing model evaluated from the perspective of discriminating known positive locations from the background unlabeled locations (Fig. 15), and ensemble XGBoost did not perform as well. XGBoost likely distinguishes itself from the simpler logistic regression algorithm because XGBoost can learn the ranges of heat flow, fault distance, magmatic distance, seismicity, and stress associated with geothermal systems, whereas logistic regression, a linear algorithm, can only identify smoothly varying linear relations. Therefore, choosing the most appropriate approach to predict geothermal systems requires weighing the complexity of the approach with the data available to train and test a model.

4.3. Comparing Predictions

Comparing predictions on a cell-by-cell basis between models (Fig. 16) shows that the expert decisions in the 2008 geothermal resource assessment had as much influence on the models produced as the algorithm selected.

4.3.1. Influence of Expert Decisions in a Cell-by-Cell Comparison

In a cell-by-cell comparison of normal score transformed predictions (Fig. 16), the machine learning models generally agree with each other more than the models from the 2008 geothermal resource assessment (Fig. 16). Predictions from the 2008 logistic regression method have the greatest disagreement with all other models (*i.e.,* largest RMSE), and weight-of-evidence has the second largest disagreement with all other models, while the 2008 logistic regression and weight-of-evidence models do not agree comparatively well with each other (RMSE = 1.29). The higher RMSE between the expert decision-dependent and machine learning models is likely a result of the biases imparted by the expert decisions. Assuming that more models in agreement implies those models are more likely correct, the 2008 Logistic Regression is likely the least reliable estimator of geothermal favorability.

The dichotomy between the expert decision-dependent and the machine learning approaches is most apparent when examining the three approaches for logistic regression. The single logistic regression and ensemble logistic regression have the greatest similarity in predictive behavior when comparing any combination of the approaches investigated in this study (RMSE = 0.22), indicating an insensitivity to the train-test split strategies, whereas the greatest disagreement between pairings with single or ensemble logistic regression and another approach are found with the 2008 geothermal resource assessment. That is, the biases of the expert decisions are explicit when comparing the different forms of logistic regression and demonstrate that the biggest differences between predictions of geothermal favorability are not a result of which strategy or shallow machine learning algorithm is used, but are, instead, an eventual product of the overall philosophy pursued.

4.3.2. Expert Decisions Imposed Non-Linearity in the 2008 Geothermal Resource Assessment

The machine learning logistic regression favorability maps have a smooth geospatial distribution of favorability (Fig. 10) relative to the favorability maps from the more expert decision-dependent approaches (*i.e.,* weight-of-evidence and expert decision-dependent logistic regression; Fig. 9) and two of the non-linear data-driven approaches (*i.e.,* XGBoost [Fig. 11] and SVMs [Fig. 12]). The differences in the continuity of predictions are also apparent in cross-plots, in which non-linear approaches display distinct binning of predicted values (*e.g.,* weight-of-evidence, expert decision-dependent logistic regression, single XGBoost, and ensemble XGBoost in Fig. 16). The smooth distribution in the machine learning logistic regression is a result of the linear fit of continuously valued input features, which contrasts with the weight-of-evidence and expert decision-dependent logistic regression methods, in which input features have binned values.

The apparent similarity in granularity between the results from the expert decision-dependent methods, which use linear models, and two of the non-linear models in this study (*i.e.,* XGBoost and SVMs) indicates that one effect of selecting expert-informed bins and thresholds is the inherent creation of non-linear features through the expert-driven conversion of the continuous values to categorical bins. While this effect was recognized in the work of Williams and DeAngelo (2008), we again find that, like in Section 4.3.1, expert decision can have as much influence on the favorability models of geothermal resource assessments as the approaches selected to create those models.

4.3.3. Effect of Algorithm Complexity on Granularity

It would generally be expected that the ensemble models, being composed of an average of sub-models, would appear smoother (*i.e.,* have less granularity) than their equivalents from the single training strategy. The natively continuous predictions from logistic regression make this determination between the two training strategies practically impossible (Fig. 10). With XGBoost and SVMs, the single variants of these approaches produce similar granularity as the ensemble approaches when predicting higher favorability (Figs. 11, 12, 16); this behavior is a product of each sub-model in the ensemble training strategy receiving the same examples of known positives. Curiously, the ensemble SVM produces greater granularity than single SVM when predicting low geothermal favorability (Fig. 12), indicating the sub-models of ensemble SVM do not express substantial variability despite being trained from different subsets of negative training data.

The low granularity of the ANN (Fig. 13) more closely resembles the smoothly varying favorability predictions produced from the machine learning logistic regression (Fig. 10) and has the lowest RMSE with machine learning logistic regression (RMSE < 0.31) than with the other approaches (Fig. 16). The relative similarity between the ANN and machine learning logistic regression, compared to the ANN and the

other approaches, is likely due to the ANN balancing the complexity resulting from its two hidden layers and 35 nodes per layer (Table 2) and its effort to avoid overfitting (*i.e.,* if a simple linear model is sufficient, or only small non-linearities are required, then the ANN will favor a "nearly" linear model). ANNs have been documented to behave similarly to logistic regression when the ANNs remained simple (*e.g.,* by using fewer layers; see Spackman, 1992; Vach et al., 1996). As a result of its design with a sigmodal activation function in its single-node output layer (see Fig. 2 for a sigmoidal function; see Fig. 5 for the operational significance of the single-node output layer), the ANN produces near-continuous predictions (Fig. 13) and, while these continuous predictions could express greater geospatial granularity reflecting the complexity of the algorithm (*e.g.,* Ayer et al., 2010), the effects of averting overfitting result in a favorability map more like those of machine learning logistic regression than any other approach (Figs. 10, 13, 16).

4.3.4. *A Proposed Modification to RMSE for Future Work*

As previously stated, models differ greatly in areas that are considered to have low favorability. Therefore, RMSE values are commonly heavily influenced by differing low geothermal favorability predictions. This behavior may partly be the result of only using positive and unlabeled data for training, thereby teaching the models a relatively common predilection for identifying positive locations; however, the absence of a geothermal system may result from a range of prohibitive physical conditions (*e.g.,* heat flow may be too low or there may be insufficient permeability), and each of the machine learning strategies may emphasize different conditions of failure, resulting in different map patterns of low favorability.

If the goal is to find which approaches agree strongly on high-favorability sites, the RMSE of entire models may not be the best measure. In fact, qualitatively, ensemble logistic regression and single XGBoost appear to differ from the other shallow learning approaches the most substantially when predicting high geothermal favorability but still have a low RMSE value when comparing the models to each other (*i.e.,* RMSE = 0.45; Fig. 16). Similarly, the ensemble XGB and the ANN appear to share good agreement at high favorability, but have a relatively high RMSE value (*i.e.,* RMSE = 0.57; Fig. 16); hence, the normal score RMSE of data where both predictors produce a normal score of geothermal favorability > 0 (*i.e.,* both models agree that data points are within the highest 50% of data) might be a better measure of agreement between predictors for the purposes of identifying favorable locations.

4.4. *Interpreting Hyperparameter Values*

The differences of hyperparameter values between the single and ensemble strategies reflect the structural differences of the strategies. Foremost, class weights in the models from the ensemble strategy are generally a fraction of the class weights for models from the single strategy (Table 2). This observation is expected given the lower class imbalance in the ensemble strategy than in the single strategy. However, we also note that single logistic regression, single XGBoost, single SVM, and single ANN have positive class weights significantly less than what would be expected by the positive:negative natural class imbalance (*i.e.,* 1:700) as estimated using the results from Williams et al. (2008). The difference between optimized class weights and the estimated natural class imbalance suggests that the estimated number of naturally occurring geothermal systems may be too low. While we use the mean power production as modeled by Williams et al. (2008) to estimate the number of naturally occurring geothermal systems in the western United States (*i.e.,* 1:700), the class weighting for single logistic regression (*i.e.,* 258), XGBoost (*i.e.,* 206), SVM (*i.e.,* 575), and ANN (*i.e.,* 285) may suggest that the true number of naturally occurring geothermal systems is greater than our estimate of 1,040 (Table 2).

The 1:700 positive:negative class imbalance estimate derived from Williams et al. (2008) is a starting point; thereafter, the hyperparameter optimization process tunes class weights, which reflect the class imbalance the algorithms identify as the models are optimized. Using the estimated power potential of geothermal resources in the western United States at 5% probability from Williams et al. (2008), we would anticipate a positive:negative class imbalance of 1:550 (see Eqs. 6–8), which approximately resembles the optimal class weighting found for the single SVM (*i.e.,* 575). Hence, the optimal class weight for the single SVM suggests that the real geothermal power potential in the western United States may be more than twice the mean estimated value of Williams et al. (2008), and perhaps closer to the 5% confidence value of 73,286 MWe; however, we caution that the low F1 scores suggest that strong inferences from class weight on the natural class imbalance may be imprudent based on the findings herein. Nonetheless, if we follow this line of reasoning, we also note that the class weighting of the single ANN (*i.e.,* 285), single logistic regression (*i.e.,* 258) and single XGBoost (*i.e.,* 206), again suggests that the number of naturally occurring geothermal systems in the western United States may exceed the mean probability estimate derived from Williams et al. (2008). As better performing models are developed for predicting the favorability of geothermal resources (*e.g.,* as the future models produce higher F1 scores), perhaps the number of expected geothermal systems can be more accurately constrained.

5. Opportunities to Enhance Geothermal Resource Assessments

The approaches discussed above provide a means to understand past assessments and provide confidence that robust assessments can be developed that rely more fully upon the data-driven decisions with fewer choices by experts. Yet, in addition to positive-unlabeled data and class imbalance, several challenges remain. The USGS geothermal resource assessment team is currently working towards answering the following questions for the next generation of geothermal resource assessments:

- Is the F1 score is an adequate metric for positive-unlabeled data? The F1 score penalizes positive predictions of unlabeled cells (*i.e.,* what would be termed false positives with positive-negative data; see Eq. 4), but these cells may indeed be positive. Instead, geological resources and phenomena need a new performance metric for their unique characteristics.
- Is a decision threshold of 0.5 appropriate for machine learning with geothermal data? Preliminary testing of this assumption (*i.e.,* deviation from a 0.5 threshold) did not find improvement at other decision thresholds, but a full examination is beyond the scope of this paper and a more exhaustive analysis could provide insight.
- How could the distributions of normal score transformed predictions between known positives and unlabeled examples be used for hyperparameter optimization? What other methods could address the positive-unlabeled aspect of the data? Bekker and Davis (2020) suggest several methods for training with positive-unlabeled data, like using semi-supervised approaches with consistency regularization to separate positive and negative classes.
- How can we develop workflows that are not reliant upon gridding a region of study? While there is value in understanding the geothermal favorability of km-sized cells, it may be more useful to understand geothermal favorability directly under foot (or any other arbitrary geographic location). To do so, we would need to break grids to < 100 m in dimension, which would require presently unattainable processing power for regions the size of the western United States or the abandonment of grids in some workflows entirely. This degree of geospatial precision would also be dependent upon engineering more informative features.

- How can features be better engineered to predict geothermal favorability? Our results suggest that the 2008 data were insufficient to predict the occurrence of geothermal systems well. In addition to collecting refined data, new features might be engineered that better represent geological conditions as they relate to geothermal favorability.
- Should other forms of processing for feature data be used? In this study, the feature sets are standardized, but would a different transformation be more appropriate to target the specific conditions that permit geothermal systems or reduce the effect of outliers in the feature data (*e.g.,* a quantile-to-quantile transform that removes or emphasizes outliers)?
- Is it best to call all known geothermal systems positive, or are there distinct systems that should all have separate labels *(e.g.,* magmatic systems, deep-circulation systems)? Hitherto, we have been discussing geothermal exploration in pursuit of all conventional geothermal systems. Should we expect shallow, magmatically driven geothermal systems to share the same qualities as deep-circulation, fault-driven systems? If not, the geothermal data would benefit from more than one type of positive label. Are there distinctions between "big" or "small" geothermal systems? That is, do "big" systems occur where conditions are more favorable?
- How will the algorithms need to be applied differently to identify conditions favorable to engineered geothermal systems (*i.e.,* EGSs) or blind geothermal systems? How do we approach the data-driven exploration of direct-use geothermal energy?

6. Conclusion

In this study, we demonstrate that, when using the same data, modern machine learning approaches can perform as least as well as, if not better than the methods used in the 2008 U.S. Geological Survey geothermal resource assessment, which relied upon expert decisions, to predict geothermal favorability in the western United States. The models produced by the machine learning approaches perform similarly with ubiquitously low F1 scores (*i.e.,* F1 scores < 0.10), emphasizing the need for improving input feature data and handling intrinsic problems with labeled geothermal data (*e.g.,* positive-unlabeled data, severe class imbalance). The expert decision-dependent and machine learning approaches show general agreement, demonstrating that the machine learning algorithms present a means to produce and even improve the geothermal favorability maps from the 2008 geothermal resource assessment while minimizing the biases of expert decisions. By using several measures of feature importance across the nine approaches, we find that heat flow and distance to a fault are the two features of predominant importance when producing models to predict geothermal favorability from the five input features used. We posit that highly complex algorithms do not perform as well or as consistently with the data from the 2008 geothermal resource assessment as simpler algorithms, and postulate that the differences in performance are a product of the bias-variance tradeoff and/or the inherent shape of the decision boundary native to the algorithms considered. Finally, we demonstrate how the expert decisions from the 2008 geothermal resource assessment (*i.e.,* binning) of the input feature sets effectively rendered the otherwise linear methods used therein (*i.e.,* weight-of-evidence and logistic regression) to become non-linear and that the greatest variability between the predictions from the different models is a result of their degree of dependence on expert decisions.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The input feature data and model predictions are available through a U.S. Geological Survey data release available at https://www.sciencebase.gov/catalog/item/63090a9cd34e3b967a8c19c4 (DOI 10.5066/P9V1Q9XM).

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 Cevat (Özgen) Karacan, Mark Coolbaugh, and one anonymous reviewer for their review of this manuscript.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.geothermics.2023.102662.

References

Abuzied, S.M., Kaiser, M.F., Shendi, E.-A.H., Abdel-Fattah, M.I, 2020. Multi-criteria decision support for geothermal resources exploration based on remote sensing, GIS and geophysical techniques along the Gulf of Suez coastal area, Egypt. Geothermics 88, 101893. https://doi.org/10.1016/j.geothermics.2020.101893.

Advanced National Seismic System Comprehensive Earthquake Catalog. (2022). Retrieved from https://earthquake.usgs.gov/data/comcat/.

Bekker, J., Davis, J., 2020. Learning from positive and unlabeled data: a survey. Int. J. Mach. Learn. Cybern. 109, 719–760. https://doi.org/10.1007/s10994-020-05877-5.

Berkson, J., 1944. Application of the Logistic Function to Bio-Assay. J. Am. Statist. Assoc. 39, 357–365.

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

Blackwell, D.D., Richards, M., 2004. Geothermal Map of North America. AAPG Map, scale 1:6,500,000.

Boutaba, R., Salahuddin, M.A., Limam, N., Ayoubi, S., Shahriar, N., Estrada-Solano, F., Caicedo, O.M., 2018. Acomprehensive survey on machine learning for networking: evolution, applications and research opportunities. J. Internet Servic. Appl. 9 (1), 1–99.

Branco, P., Torgo, L., Ribeiro, R., 2015. A Survey of Predictive Modelling under Imbalanced Distributions. Computing Research Repository, pp. 1–48.

Burkov, A., 2019a. Chapter 5: Basic Practice. The Hundred-Page Machine Learning Book. Burkov, Andriy, pp. 43–60.

Burkov, A., 2019b. Chapter 6: Neural Networks and Deep Learning. The Hundred-Page Machine Learning Book. Burkov, Andriy, pp. 61–76.

Chapelle, O., 2007. Training a Support Vector Machine in the Primal. J. Mach. Learn. Res. 19, 1155–1178. https://doi.org/10.1162/neco.2007.19.5.1155.

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

Chollet, F., 2015. Keras. GitHub Repository. Retrieved from. https://github.com/fchollet/keras.

Chollet, F., 2021. Chapter 1. What is deep learning? Deep Learning with Python, 2nd ed. Simon and Schuster.

Cortes, C., Vapnik, V., 1995. Support-Vector Networks. Int. J. Mach. Learn. Cybern. 20, 273–297, 10.1007/BF00994018.

Donnelly-Nolan, J.M., 1988. A magmatic model of Medicine Lake volcano, California. J. Geophys. Res. 93, 4412–4420.

Falgout, J.T., Gordon, J., 2021. USGS Advanced Research Computing, USGS Yeti Supercomputer. U.S. Geological Survey.

Falgout, J.T., Gordon, J., Davis, M.J., 2021a. USGS Tallgrass Supercomputer. U.S. Geological Survey.
Falgout, J.T., Gordon, J., Williams, B., Davis, M.J., 2021b. SGS Advanced Research Computing, USGS Denali Supercomputer. U.S. Geological Survey.
Faulds, J.E., Craig, J.W., Hinz, N.H., Coolbaugh, M.F., Glen, J.M., Earney, T.E., Siler, D. L., 2017. Discovery of a blind geothermal system in southern Gabbs Valley, western Nevada, through application of the play fairway analysis at multiple scales. Geotherm. Resour. Counc. Trans. (42).
Fernández-Delgado, M., Cernadas, E., Barro, S., Amorim, D., 2014. Do We Need Hundreds of Classifiers to Solve Real World Classification Problems? J. Mach. Learn. Res. 15, 3133–3181.
Fernández, A., García, S., Galar, M., Prati, R., Krawczyk, B., Herrera, F., 2018. Learning from Imbalanced Data Sets. Springer, Cham, Switzerland.
Forson, C., Steely, A., Cladouhos, T., Swyer, M., Davatzes, N., Anderson, M., Stelling, P., 2017. Geothermal Play-Fairway Analysis of Washington State Prospects Phase 2 Technical Report(DOE-WGS-6728-1).
Geron, A., 2017. Chapter 10: Introduction to Artificial Neural Networks with Keras. In: Roumeliotis, R., Tache, N. (Eds.), Hands-On Machine Learning with Scikit-Learn & TensorFlow : concepts, tools, and techniques to build intelligent systems, 2nd ed. O'Reilly, Canada.
Guo, X., Yin, Y., Dong, C., Yang, G., Zhou, G., 2008. On the Class Imbalance Problem. In: Paper presented at the 2008 Fourth International Conference on Natural Computation, Jinan, China.
Hildreth, W., 2007. Quaternary magmatism in the Cascades – Geologic Perspectives, p. 125. https://doi.org/10.3133/pp1744. U.S. Geological Survey Professional Paper 1744.
Hinz, N.H., Coolbaugh, M.F., Shevenell, L., Melosh, G., Cumming, W., Stelling, P., 2015. Preliminary Ranking of Geothermal Potential in the Cascade and Aleutian Volcanic Arcs, Part II: Structural—Tectonic Settings of the Volcanic Center. GRC Trans. 39, 717–726.
Ito, G., Frazer, N., Lautze, N., Thomas, D., Hinz, N., Waller, D., Wallin, E., 2017. Play fairway analysis of geothermal resources across the state of Hawaii: 2. Resource probability mapping. Geothermics 70, 393–405. https://doi.org/10.1016/j.geothermics.2016.11.004.
Krawczyk, B., 2016. Learning from imbalanced data: open challenges and future directions. Progr. Artific. Intellig. 5, 221–232. https://doi.org/10.1007/s13748-016-0094-0.
Kuhn, M., Johnson, K., 2013. Applied Predictive Modeling, 1st ed. Springer, New York, USA.
Lacasse, C.M., Prado, E.M.G., Guimarães, S.N.P., Filho, O.A.d.S., Vieira, F.P., 2022. Integrated assessment and prospectivity mapping of geothermal resources for EGS in Brazil. Geothermics 100 (102321). https://doi.org/10.1016/j.geothermics.2021.102321.
Lautze, N., Ito, G., Thomas, D., Frazer, N., Martel, S.J., Hinz, N., Martin, T., 2020. Play Fairway analysis of geothermal resources across the State of Hawai'i: 4. Updates with new groundwater chemistry, subsurface stress analysis, and focused geophysical surveys. Geothermics 86, 101798. https://doi.org/10.1016/j.geothermics.2019.101798.
Lautze, N., Waller, T.D., Frazer, N., Hinz, N., Apuzen-Ito, G., 2017. Play fairway analysis of geothermal resources across the state of Hawaii: 3. Use of development viability criterion to prioritize future exploration targets. Geothermics 70, 406–413. https://doi.org/10.1016/j.geothermics.2017.07.005.
Lee, W.S., Liu, B., 2003. Learning with Positive and Unlabeled Examples Using Weighted Logistic Regression. In: Paper presented at the Twentieth International Conference on Machine Learning.
Lindsey, C.R., Ayling, B.F., Asato, G., Seggiaro, R., Carrizo, N., Larcher, N., Coolbaugh, M.F., 2021. Play fairway analysis for geothermal exploration in northwestern Argentina. Geothermics 95. https://doi.org/10.1016/j.geothermics.2021.102128.
Lindsey, C.R., Price, A.N., Burns, E.R., 2022. Exploring Declustering Methodology for Addressing Geothermal Exploration Bias. In: Paper presented at the Geothermal Rising Conference, Reno, NV.
Lundberg, S.M., Lee, S.-I., 2017. A Unified Approach to Interpreting Model Predictions. In: Paper presented at the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.
Machette, M.N., Haller, K.M., Dart, R.L., Rhea, S.B., 2003. Quaternary fold and fault database of the United States. United States Geological Survey Open-File Report. https://www.usgs.gov/programs/. earthquake-hazards/faults?qt-science_support_page_related_con=4#qt-science_support_page_related_con.
MacLeod, N.S., Sherrod, D.R., Chitwood, L.A., Jensen, R.A., 1995. Geologic map of Newberry volcano, Deschutes, Klamath and Lake Counties, Oregon. U.S. Geological Survey Miscellaneous Investigations Series Map I-2455, p. 523. https://doi.org/10.3133/i2455, 2 sheets, scale 1:62,500, pamphlet.
Meng, F., Liang, X., Xiao, C., Wang, G., 2021. Geothermal resource potential assessment utilizing GIS-based multi criteria decision analysis method. Geothermics 89, 101969. https://doi.org/10.1016/j.geothermics.2020.101969.
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., 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. In: Paper presented at the 47th Stanford Geothermal Workshop, Stanford, California (Virtual).
Muffler, L.P.J., 1979. Assessment of geothermal resources of the United States-1978. U.S. Geological Survey Circular 790, 163. https://doi.org/10.3133/cir790.
Murphy, K.P., 2012. Machine Learning: A Probabilistic Perspective. The MIT Press, Cambridge, Massachuesetts.
Musumeci, F., Rottondi, C., Nag, A., Macaluso, I., Zibar, D., Ruffini, M., Tornatore, M., 2019. An overview on application of machine learning techniques in optical networks. IEEE Commun. Surv. Tutor. 21 (2), 1383–1408. https://doi.org/10.1109/COMST.2018.2880039.
Nielson, D.L., Shervais, J.W., Evans, J., Liberty, L.M., Garg, S.K., Glen, J.M., Sonnenthal, E.L., 2015. Geothermal Play Fairway Analysis of the Snake River Plane, Idaho. In: Paper presented at the Fourtieth Workshop on Geothermal Reservoir Engineering. Stanford University, Stanford, California.
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., 2011. Scikit-learn: Machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830.
Reed, M.J., 1983. Assessment of low-temperature geothermal resources of the United States- 1982. U.S. Geological Survey Circular 892, 73.
Reinecker, J., Heidbach, O., Tingay, M., Sperner, B., Muller, B., 2005. The release 2005 of the World Stress Map.
Shalev-Shwartz, S., Ben-David, S., 2014a. Chapter 5: The Bias-Complexity Trade-Off. Understanding Machine Learning: From Theory to Algorithms. Cambridge University Press, New York, New York, USA, pp. 215–226.
Shalev-Shwartz, S., Ben-David, S., 2014b. Chapter 16: Kernel Methods. Understanding Machine Learning: From Theory to Algorithms. Cambridge University Press, New York, New York, USA, pp. 215–226.
Shervais, J.W., Evans, J.P., Newell, D.L., Glen, J.M., Siler, D.L., DeAngelo, J., Ritzinger, B., 2021. Play Fairway Analysis of the Snake River Plain, Idaho: Final Report. Retrieved from Geothermal Data Repository.
Shervais, J.W., Glen, J.M., Siler, D.L., Liberty, L.M., Nielson, D., Garg, S., Neupane, G., 2020. Play Fairway Analysis in Geothermal Exploration: The Snake River Plain Volcanic Province. In: Paper presented at the Workshop on Geothermal Reservoir Engineering. Stanford University, Stanford, California.
Shevenell, L., Coolbaugh, M.F., Hinz, N., Stelling, P., Melosh, G., Cumming, W., 2015. In: Geothermal Potential of the Cascade and Aleutian Arcs, with Ranking of Individual Volcanic Centers for their Potential to Host ElectricityGrade Reservoirs. Retrieved from Geothermal Data Repository.
Siler, D.L., Zhang, Y., Spycher, N.F., Dobson, P.F., McChain, J.S., Gasperikova, E., Cantewell, C., 2017. Play-fairway analysis for geothermal resources and exploration risk in the Modoc Plateau region. Geothermics 69, 15–33. https://doi.org/10.1016/j.geothermics.2017.04.003.
Spackman, K.A., 1992. Maximum Likelihood Training of Connectionist Models: Comparison with Least Squares Back-propagation and Logistic Regression. In: Paper presented at the Fifteenth Annual Symposium on Computer Applications in Medical Care, Washingon, DC.
Vach, W., Roßner, R., Schumacher, M., 1996. Neural networks and logistic regression, Part II. Comput. Stat. Data Anal. 21 (6), 683–701. https://doi.org/10.1016/0167-9473(95)00033-X.
Walker, J.D., Bowers, T.D., Black, R.A., Glazner, A.F., Farmer, G.L., Carlson, R.W., 2006. A geochemical database for western North American volcanic and intrusive rocks (NAVDAT). Spec. Pap. Geol. Soc. Am. 397, 61–71. https://doi.org/10.1130/2006.2397(05).
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 Trans. 32, 181–188.
Williams, C.F., Reed, M.J., DeAngelo, J., Galanis, S.P., 2009. Quantifying the undiscovered geothermal resources of the United States. Trans. Natl. Saf. Congr. 33, 882–889.
Williams, C.F., Reed, M.J., Galanis, S.P., DeAngelo, J., 2007. The USGS national geothermal resource assessment: An update. GRC Trans. 31, 99–104.
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, pp. 1–4. U.S. Geological Survey Fact Sheet 2008-3082.
Zdravevski, E., Lameski, P., Kulakov, A., 2011. Weight of evidence as a tool for attribute transformation in the preprocessing stage of supervised learning algorithms. In: Paper presented at the The 2011 International Joint Conference on neural Networks (IJCNN), San Jose, California, USA.