PROCEEDINGS, 47th Workshop on Geothermal Reservoir Engineering
Stanford University, Stanford, California, February 7-9, 2022
SGP-TR-223

Predicting Geothermal Favorability in the Western United States by Using Machine Learning: Addressing Challenges and Developing Solutions

Stanley P. MORDENSKY[1], John J. LIPOR[2], Jacob DEANGELO[1], Erick R. BURNS[1], and Cary R. LINDSEY[1]

[1]U.S. Geological Survey, 2130 SW 5th Ave, Portland OR 97201, USA

[2]Portland State University, 19000 SW 4th Ave, Suite 160, Portland OR 97201, USA

smordensky@usgs.gov

Keywords: geothermal resource assessment, machine learning, logistic regression, support-vector machine, XGBoost, class imbalance

ABSTRACT

Previous moderate- and high-temperature geothermal resource assessments of the western United States utilized weight-of-evidence and logistic regression methods to estimate resource favorability, but these analyses relied upon some expert decisions. While expert decisions can add confidence to aspects of the modeling process by ensuring only reasonable models are employed, expert decisions also introduce human bias into assessments. This bias presents a source of error that may affect the performance of the models and resulting resource estimates. Our study aims to reduce expert input through robust data-driven analyses and better-suited data science techniques, with the goals of saving time, reducing bias, and improving predictive ability. We present six favorability maps for geothermal resources in the western United States created using two strategies applied to three modern machine learning algorithms (logistic regression, support-vector machines, and XGBoost). To provide a direct comparison to previous assessments, we use the same input data as the 2008 U.S. Geological Survey (USGS) conventional moderate- to high-temperature geothermal resource assessment. The six new favorability maps required far less expert decision-making, but broadly agree with the previous assessment. Despite the fact that the 2008 assessment results employed linear methods, the non-linear machine learning algorithms (*i.e.,* support-vector machines and XGBoost) produced greater agreement with the previous assessment than the linear machine learning algorithm (*i.e.,* logistic regression). It is not surprising that geothermal systems depend on non-linear combinations of features, and we postulate that the expert decisions during the 2008 assessment accounted for system non-linearities. Substantial challenges to applying machine learning algorithms to predict geothermal resource favorability include severe class imbalance (*i.e.,* there are very few known geothermal systems compared to the large area considered), and while there are known geothermal systems (*i.e.,* positive labels), all other sites have an unknown status (*i.e.,* they are unlabeled), instead of receiving a negative label (*i.e.,* the known/proven absence of a geothermal resource). We address both challenges through a custom undersampling strategy that can be used with any algorithm and then evaluated using F1 scores.

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 was completed by Williams and DeAngelo (2008), Williams et al. (2008), and Williams et al. (2009). This assessment produced 28 models to identify locations of high geothermal favorability in the western United States (examples shown in Fig. 1) using two modeling methods (*i.e.,* weight-of-evidence and logistic regression) that varied combinations of 9 geological input feature sets (see Williams and DeAngelo (2008) for complete reference information). These 9 feature sets are divided into 5 input feature types, and each model uses no more than one feature set from each type:

- Quaternary faulting
  - Distance to Quaternary faults from the USGS Quaternary fault and fold database (Machette et al., 2003)
- Magmatic activity from Donnelly-Nolan (1988), MacLeod et al. (1995), Walker et al. (2006), and Hildreth (2007)
  - Distance to all magma bodies
  - Distance to felsic magma bodies
  - Distance to mafic magma bodies
- Heat flow
  - Heat flow interpolated from unpublished data compiled for Williams et al. (2007)
  - Heat flow interpolated from Blackwell and Richards (2004)
- Seismic Activity
  - Earthquake density within 4 km from the *ANSS Comprehensive Earthquake Catalog*
  - Log of the sum of seismic moments of earthquakes within 10 km from the *ANSS Comprehensive Earthquake Catalog*
- Stress
  - Maximum horizontal stress interpolated from Reinecker et al. (2005)

Although these assessment models 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, some feature sets had to be binned (*i.e.,* bucketed or categorized). Other feature sets used buffer distance to create a binary classification (*e.g.,* within 4

km of a feature or beyond 4 km from a feature). Thus, parameters like bin sizes, number of bins, and threshold values had to be selected. While these expert decisions potentially add value, they also introduce a potential source of bias.

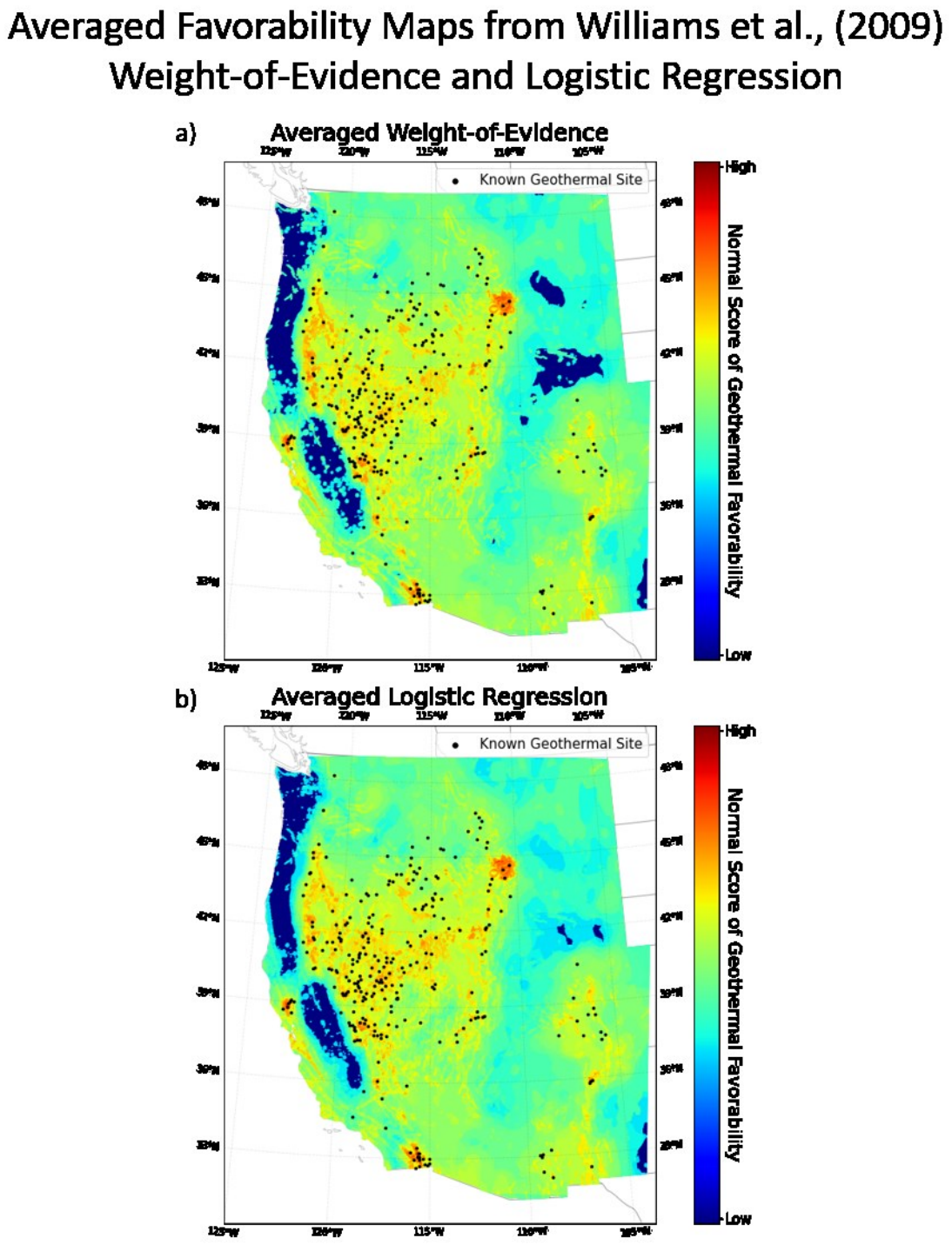

Figure 1: Geothermal favorability maps of the western United States averaged from 12 models as presented in Williams et al. (2009) using the: a) weight-of-evidence; and b) logistic regression methods from Williams and DeAngelo (2008). The individual models used for averaging are differentiated by their unique input feature combinations. Favorability is the normal score transform of averaged probability.

Herein, we detail a means to minimize expert bias while producing models that reliably predict geothermal favorability in the western United States using fundamental machine learning algorithms (*i.e.,* logistic regression, support-vector machines, and eXtreme Gradient Boosting) with two strategies and the data from the 2008 USGS geothermal resource assessment.

1.1 What Are Machine Learning Algorithms?

Machine learning algorithms, like logistic regression, support-vector machines, and eXtreme Gradient Boosting, provide a data-driven means to produce models. Machine learning algorithms aim to learn the statistics of a dataset (*i.e.,* the training data) in order to create optimal decision functions (*i.e.,* models) using minimal human input. Implicit to the name, data-driven decisions are choices algorithmically made based upon the data during the fitting of a model that optimizes a performance metric (*e.g.,* accuracy, precision, recall, and F1 score as defined in Equations 1 – 4).

$$Accuracy = \frac{True\ Positives + True\ Negatives}{True\ Positives + True\ Negatives + False\ Positives + False\ Negatives} \tag{1}$$

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives} \tag{2}$$

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives} \tag{3}$$

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

Performance metric optimization is primarily achieved through the selection of an algorithm's internal variables that balances the tradeoff between underfitting (*i.e.,* when a model fails to predict well with the training data; that is, when the model is too simple) and overfitting (*i.e.,* when a model predicts well with the training data, but fails to predict well with data not used for training; that is, when the model is too complex). The adjustable internal variables are called hyperparameters, and the selection of hyperparameter values that optimize the chosen performance metric is called hyperparameter optimization. Fundamentally, data-driven decisions mean the nature of the data determines hyperparameter values. Therefore, hyperparameters are optimized when the hyperparameters produce a model that is most representative of the phenomenon as is possible by the chosen algorithm. Depending on the quality of the data, this determination (*i.e.,* what qualifies as most representative) may require human discretion; however, given sufficient quality, the data make this decision.

Hyperparameter optimization helps algorithms handle the unique qualities of datasets. One such quality is the relative frequency of the occurrence of classification labels. Machine learning algorithms operate most effectively when the occurrences of classification labels in the data are nearly equal in frequency (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 (*i.e.,* downsampling) removes data of the majority class (*i.e.,* the class with the more frequent occurrence). Undersampling presents the risk of removing valuable data. 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 (*e.g.,* logistic regression versus eXtreme Gradient Boosting; see generally Branco et al. (2015)).

1.2 Challenges of the Data from the 2008 Geothermal Resource Assessment

The data in the 2008 USGS geothermal resource assessment had severe class imbalance. The 2008 USGS geothermal resource assessment gridded the western United States into 725,442 2-km-by-2-km cells, 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 725,164 cells are unlabeled, though it is deemed likely that most cells are negative, and for the 2008 assessment, all the unlabeled cells were assumed to be negative. One immediate difficulty evident from this severe class imbalance (*i.e.,* a < 1:2600 positive:negative ratio) is that a simple model that predicts every cell as negative has an accuracy of > 99.96%, even though that model predicts no geothermal systems. In other words, this highly accurate model provides no insight into where sparse geothermal systems exist. Furthermore, under the assumption that all unlabeled cells are negative, undiscovered geothermal systems are incorrectly labeled as negatives, and a challenge of geothermal resource analysis is to properly allow and account for these incorrectly labeled cells.

2. METHODS

With consideration for the advances in machine learning over the last decade, we seek to develop and implement a philosophy of unbiased (or minimally biased) data analysis in order to model the favorability of conditions for geothermal resources. We use the data of Williams and DeAngelo (2008) to facilitate the comparison between the past assessment and the machine learning approaches developed herein. The F1 score (Equation 4) is selected as the metric of performance. Below, the algorithms are summarized, and the two strategies of addressing the severe class imbalance are described.

We describe model predictions in terms of *favorability*. We define *favorability* as a measure of the presence of geologic conditions believed to be associated with the presence of a geothermal system. In order to permit the comparison of geothermal favorability between each algorithm, the native output from each algorithm is normal score transformed (see generally Pyrcz and Deutsch, 2018), thereby accounting for differences in measurement units (*e.g.,* probability or distance measures).

2.1 The Data

To provide a direct comparison of data-driven machine learning algorithm performance with that of the 2008 methods, we use five of the feature sets from Williams and DeAngelo (2008). Specifically, we select one raw feature set from each feature type: heat flow; distance to a magma body; distance to a fault; density of epicenters for seismic events ≥ M3 within a 4-km radius; and maximum horizontal stress. These feature sets are used to create new geothermal favorability maps using the methods of Williams and DeAngelo (2008) and the data-driven, machine learning approaches we introduce herein. As is common practice in data-driven methods, we standardize and normalize the data prior to application of each machine learning algorithm (see generally Burkov, 2019).

2.2 Performance Metric Selection

Per the methods review by Bekker and Davis (2020), the F1 score is the most appropriate for binary positive-unlabeled classifications like those found in the geothermal data used herein. Hence, we select the F1 score (Equation 4) as the performance metric for all simulations.

The F1 score gets a maximum value of 1 when all positive locations are identified as positive and all unlabeled locations are identified as negative. In this way, a lower F1 score reflects a model with poorer performance. We note that with positive-unlabeled data, we cannot be certain of false positives, which are a consideration with the F1 score; however, other performance metrics (*e.g.,* accuracy, precision, recall) have been found to be even less adequate for positive-unlabeled data (Bekker and Davis, 2020). By nature of the F1 score taking into account true positives, false negatives, and false positives, the F1 score provides the best assessment of a model's performance with positive-unlabeled data.

2.3 The Algorithms Considered

We compare the three data-driven algorithms used for analysis – logistic regression, support-vector machines (*i.e.,* SVMs), and eXtreme Gradient Boosting (commonly referred to as XGBoost) – to the expert decision-dependent application of weight-of-evidence and logistic regression methods in the 2008 USGS geothermal resource assessment. We choose these three data-driven algorithms for several reasons. We select logistic regression because Williams and DeAngelo (2008) also used this algorithm, albeit with expert decisions, allowing for the comparison with all other conditions being the same between the machine learning and expert decision-dependent models. We also include two non-linear methods: SVMs and XGBoost. SVMs and XGBoost are general-purpose classifiers (see generally Fernández-Delgado et al., 2014), but when compared to each other, they rely on fundamentally different approaches to produce decisions, providing a contrast between common non-linear methods. Hence, selecting these three data-driven algorithms expands our perspective in the behavior of machine learning with the geothermal data. More details and reference information are provided for each algorithm in the subsequent three subsections.

2.3.1 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 Equation 5 (Fig. 2):

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

in which the coefficients, $\beta_0, \beta_1, \beta_2, \ldots, \beta_n$, are empirically fit, and $x_1, x_2, \ldots, x_n$ are the input features (see Berkson (1944) for complete details). We normal score transform *Probability* to produce a plot of favorability.

A decision threshold (often probability = 0.5) defines classification labels (*e.g.,* 1 or 0, Yes or No, Geothermally Favorable or Not Geothermally Favorable) above and below that decision threshold. The computational requirements of logistic regression scale linearly with additional training data.

Using a custom *k*-folds cross-validation (see section 2.4 for more details), we use the common 0.5 decision threshold with logistic regression and optimize two hyperparameters, the inverse regularization strength and the class weight. The inverse regularization strength hyperparameter inversely 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 necessary 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 sites over non-positive sites. Misclassification of the majority class occurs more frequently as the minority class receives greater class weighting (an example for 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).

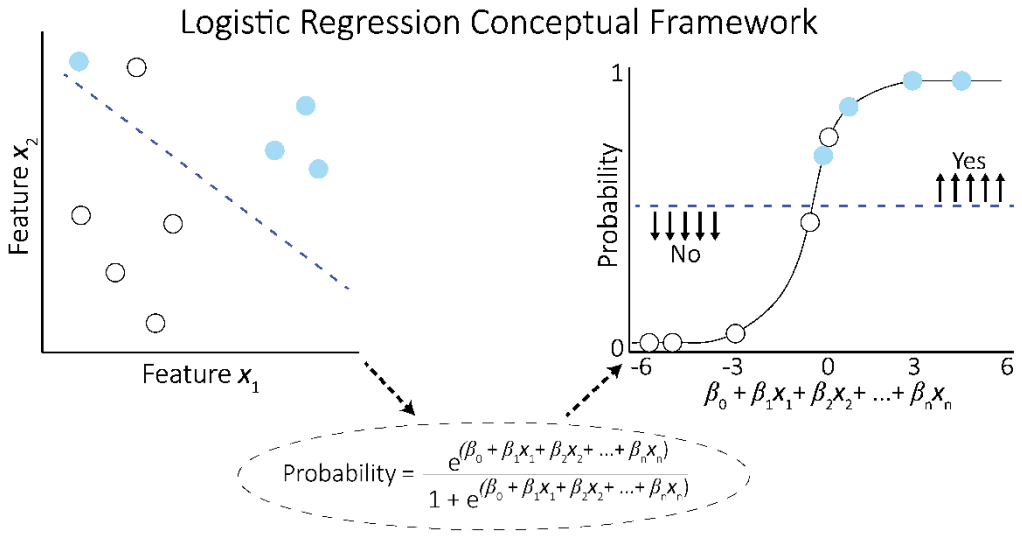

Figure 2: Conceptual framework for logistic regression (schematic shows only two features for illustrative purposes, but the concept easily extends to *n* features through Equation 5). The dashed blue line represents a 0.5 probability threshold (a common choice in the machine learning community). The 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 label dictated by the chosen threshold. Probability values range between 0 and 1, and a normal score transform of these values are used in this manuscript for plots of favorability.

2.3.2 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 labels by finding a hyperplane in an *n*-dimensional space with *n* defined by the number of input features (in our case, 5 input features define a 5-dimensional space). The hyperplane serves as a decision boundary (*i.e.,* maximum margin classifier) that maximizes the *n*-dimensional distance between data with different labels (Fig. 3 shows a linear 2-dimensional example). While finding a hyperplane is a linear process, non-linearities are accommodated through the so-called *kernel trick* (Shalev-Shwartz and Ben-David, 2014), which uses a non-linear transform to map the data to a new space where a linear decision boundary is found. 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.

Given their framework, SVMs do not provide a probability like logistic regression, but instead directly provide a label and the distance from that label to the decision boundary. We normal score transform the distance between the label and decision boundary to produce favorability plots with SVMs.

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 with a custom *k*-folds cross-validation (see section 2.4). We also add the kernel parameter gamma as a third hyperparameter to optimize. Although not implemented identically between the two algorithms, the influence of inverse regularization strength and class weight on the behavior of SVMs is similar to that of logistic regression (section 2.3.1). 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).

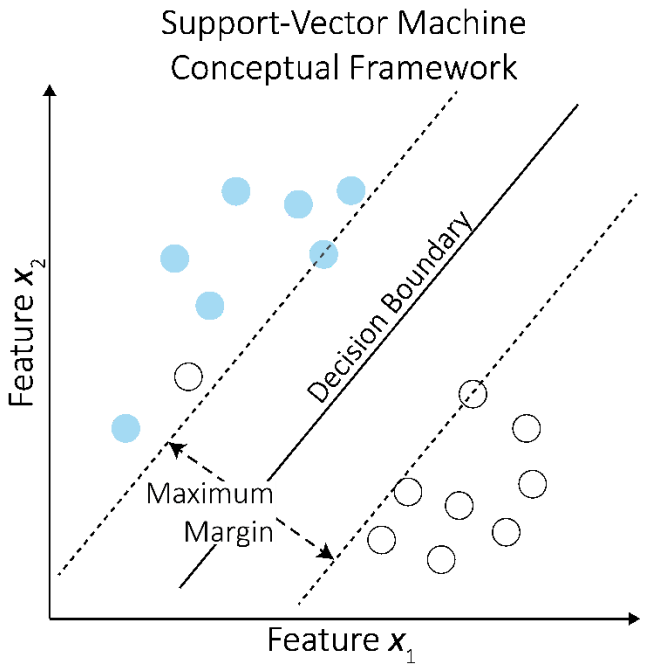

Figure 3: Conceptual framework of an SVM showing a simple two-feature (x_1 and x_2) example, which mathematically generalizes to higher dimensions using hyperplanes. The solid, blue circles are examples of one label. The hollow, black circles are examples of another label. The decision boundary (*i.e.*, the maximum margin classifier), which maximizes the distances to the nearest examples of each predicted label, is a solid black line. 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.

2.3.3 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. 4). XGBoost produces a series of simple decision trees (*i.e.,* estimators). 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 classification label is determined from the sum of the probability values across all of the estimators (see summation of probability values from each estimator in Fig. 4). We normal score transform the sum of the probabilities to produce favorability maps. 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.

With the custom *k*-folds cross-validation (see section 2.4), we optimize four hyperparameters for XGBoost: class weight, learning rate, number of estimators, and maximum depth of estimators. Class weight in XGBoost differs in exact implementation compared with logistic regression and SVMs, but this hyperparameter serves much the same purpose: a greater class weight places greater emphasis on accurately predicting positive labels (*i.e.,* known geothermal systems) than non-positive labels (*i.e.,* unknown resource potential). 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).

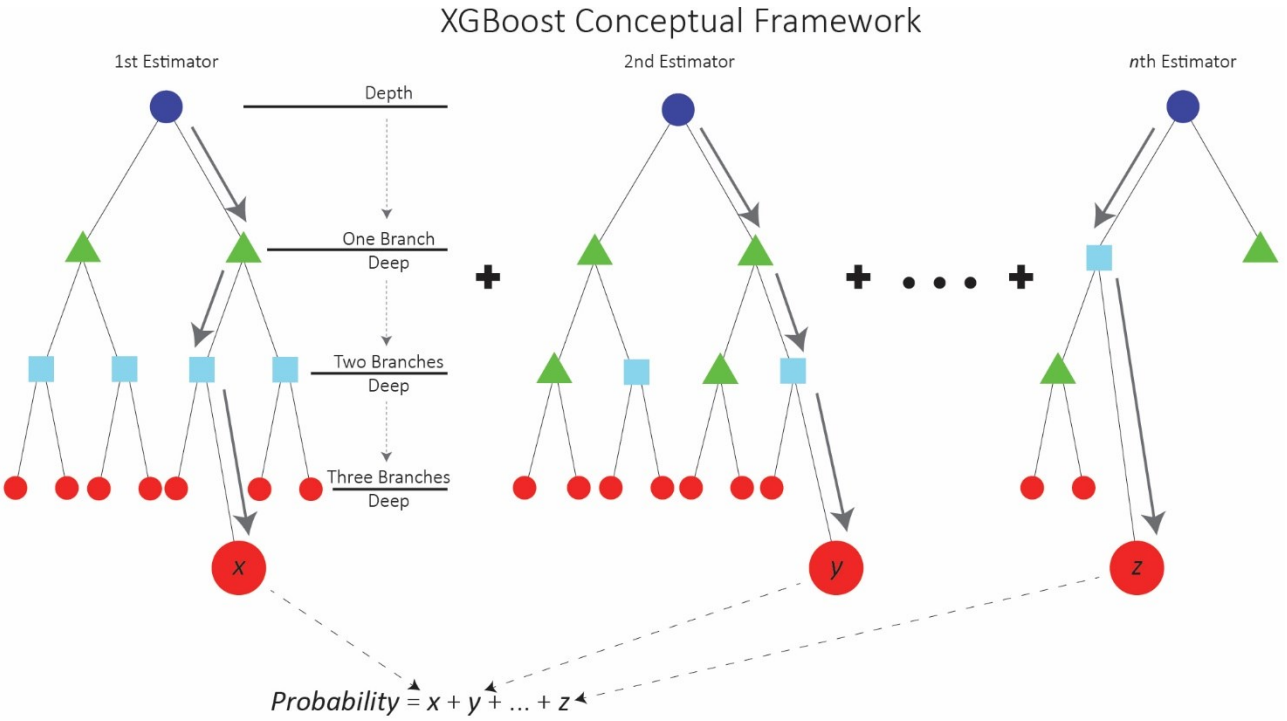

Figure 4: Conceptual framework for XGBoost. This figure depicts three of *n* estimators (*i.e.,* trees) in an XGBoost classifier. For each cell in a map, a probability value is computed for each estimator, given by its own path (*e.g.* solid black arrows) from the root node (purple circle) through the 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., x, y,…, z*) 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.

2.4 Preventing Bias Due To Class-Imbalance

In an effort to address the severe class imbalance, we experiment with two strategies using the data-driven algorithms: 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.

Both strategies require an estimate of how many geothermal systems exist in the study area (*i.e.,* identified systems + undiscovered systems) so that the expected natural positive-negative ratio guides class imbalance weighting and the number of samples selected during undersampling. To estimate the number of undiscovered systems, we estimate the mean power generation of the identified systems in Equation 6. Then, assuming the same average will hold true for undiscovered systems, the number of undiscovered systems can be computed from the estimated undiscovered resources by Equation 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 from 95% probability with 3,675 MWe to 5% probability with 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 from 7,797 MWe at 95% probability to 73,286 MWe at 5% probability. The total number of geothermal systems is then found by summing the number of identified systems and the number of undiscovered systems (Equation 8).

$$Average\ Power\ Generation\ of\ a\ System = \frac{Power\ Generation\ of\ Identified\ Systems}{Number\ of\ Identified\ Systems} \quad (6)$$

$$Number\ of\ Undiscovered\ Systems = \frac{Total\ Undiscovered\ Power\ Potential}{Average\ Power\ Generation\ of\ a\ System} \quad (7)$$

$$Total\ Number\ of\ Geothermal\ Systems = Identified\ Systems + Undiscovered\ Systems \quad (8)$$

Considering the power production estimates at 95% and 5% probability in Williams et al. (2008), we estimate a range of 760 – 1314 conventional hydrothermal systems exist in the western United States. Herein, we use the mean estimate of 1040 systems across the 725,442 2-km-by-2-km cells, thereby estimating a natural class imbalance of 1:700.

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. 5). This split is random (*i.e.,* the training and testing data are randomly sampled

from the data), and to prevent an unfortunate, unlucky split that results in a poor model, this procedure is repeated 100 times. The optimal hyperparameters are then averaged to train a model from a single train-test split and predict geothermal favorability using all of the available data.

Within each iteration of the 100 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, 2019). 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 5 folds as is common in machine learning practice to avoid overfitting or underfitting a model (see generally Burkov, 2019).

In both class imbalance strategies we used, the testing data and the data in the fold used during validation are randomly downsampled from the class imbalance of the data set (< 1:2600) to the estimated natural class imbalance (1:700). The two 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, all of the data in the remaining folds are used for fitting 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 such that 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 Equation 9.

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

Hence, with the data in this study, the ensemble strategy creates 4 subsets of data per fold. A model is then fit to each of these subsets and the performance of these models is evaluated and validated in aggregate.

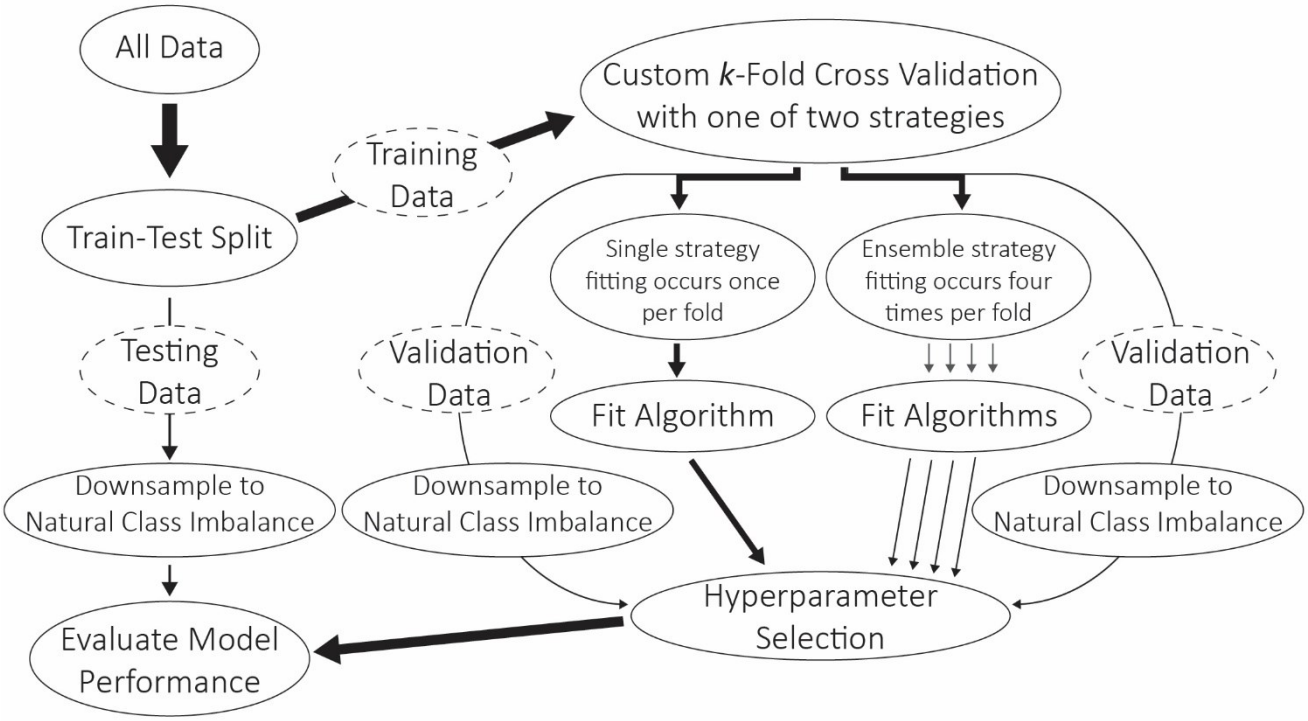

Figure 5: Workflow for all data-driven algorithms, including class-imbalance correction. The "Custom k-Fold Cross Validation" uses one of two strategies during stratified k-fold cross validation. The *single* strategy fits a single model with the remaining four fifths of the folds. The *ensemble* strategy splits the unlabeled data within the remaining four fifths of the folds to create 4 subsets of the data so that each subset approximately has the estimated 1:700 positive:negative estimated natural class imbalance for a 2-km-by-2km 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.

3. RESULTS & DISCUSSION

3.1 Comparing Favorability Maps

The geothermal favorability maps constructed using the methods from the 2008 geothermal resource assessment (Fig. 6) and the machine learning algorithms (*i.e.,* logistic regression [Fig. 7], SVMs [Fig. 8], and XGBoost [Fig. 9]) generally show a broad agreement with each other, particularly for areas of high geothermal favorability. Hence, we demonstrate that the machine learning algorithms can reproduce

and, perhaps, even improve geothermal favorability prediction without the human bias implicit to expert decision-dependent methods. However, despite this broad agreement, there are distinctions between the results of the different approaches.

The data-driven logistic regression favorability maps have a smooth geospatial distribution of favorability (Fig. 7) relative to the favorability maps from the more expert decision-dependent approaches (*i.e.,* weight-of-evidence and expert decision-dependent logistic regression; Fig. 6) and the non-linear data-driven approaches (*i.e.,* SVMs [Fig. 8] and XGBoost [Fig. 9]). The smooth distribution is the result of the linear fit of smoothly varying continuous input feature sets (see Equation 5).

In contrast with data-driven logistic regression, the apparent similarity in granularity between the 2008 results, which used linear models, and the non-linear models in this study (*i.e.,* SVMs and XGBoost) indicates that one effect of selecting expert-informed bins and thresholds is the inherent addition of non-linear features to the favorability maps. This unanticipated occurrence is the result of the processing of feature sets required for the 2008 methods. The binning and buffering effectively transformed the linear methods to become non-linear. Hence, we find that expert decision making can have as much influence on the favorability models of geothermal resource assessments as the methods selected to create those models. The degree of contrast and granularity between the methods from the 2008 geothermal resource assessment and the machine learning approaches suggests that non-linear algorithms are more appropriate than linear algorithms to reproduce the models in the 2008 geothermal resource assessment. Additionally, these non-linear machine learning approaches present the potential to introduce methods to produce more accurate and precise models for predicting geothermal resources (*e.g.,* through ensembling several unique algorithms).

When considering favorability in cross-plot format (Fig. 10), the different approaches generally share similar predictions for cells with high geothermal favorability, as evidenced by a tightening of the data cloud at high values. Conversely, each approach predicts different regions with lowest geothermal favorability (see Figs. 6, 7, 8, 9). Single logistic regression and ensemble logistic regression have the greatest similarity in predictive behavior, indicating an insensitivity to the strategies of handling class imbalance; whereas, XGBoost has the poorest agreement between single and ensemble strategies for class imbalance. The predictions from the ensemble XGBoost approaches nearly always resulted in the greatest root mean squared error (RMSE) when considered with the predictions of any other approach (with ensemble SVM versus 2008 logistic regression as the outlying exception). With every comparison of approaches, the high RMSE values (*e.g.,* > 0.90) are commonly heavily influenced by differing low geothermal favorability predictions. If the goal is to find which methods agree strongly on favorable sites, the RMSE of entire models may not be the best measure. In fact, qualitatively, single and ensemble SVMs appear to differ from the other approaches the most substantially when predicting high geothermal favorability (that is, more diffuse scatter plots at higher geothermal favorability when paired with the other approaches) but still have a moderate RMSE value (*e.g.,* < 0.90), so perhaps the normal score RMSE of data where both predictors produce a normal score of geothermal favorability > 0 is a better measure of agreement of geothermal favorability between predictors for the purposes of resource estimation. Additionally, disagreement between the single SVM approach and the ensemble SVM approach is apparent in Fig. 8 as increased granularity in the ensemble simulations compared with the relatively smooth single simulation approach. Hence, the non-linear algorithms produce the greatest variance between approaches when predicting moderate to high geothermal favorability. Ensembling these and additional new, non-linear approaches that introduce novel conceptual frameworks may provide a means to produce a more accurate and precise machine learning algorithm in future geothermal resource assessments and aid in providing a different perspective on what constitutes geothermally favorable conditions.

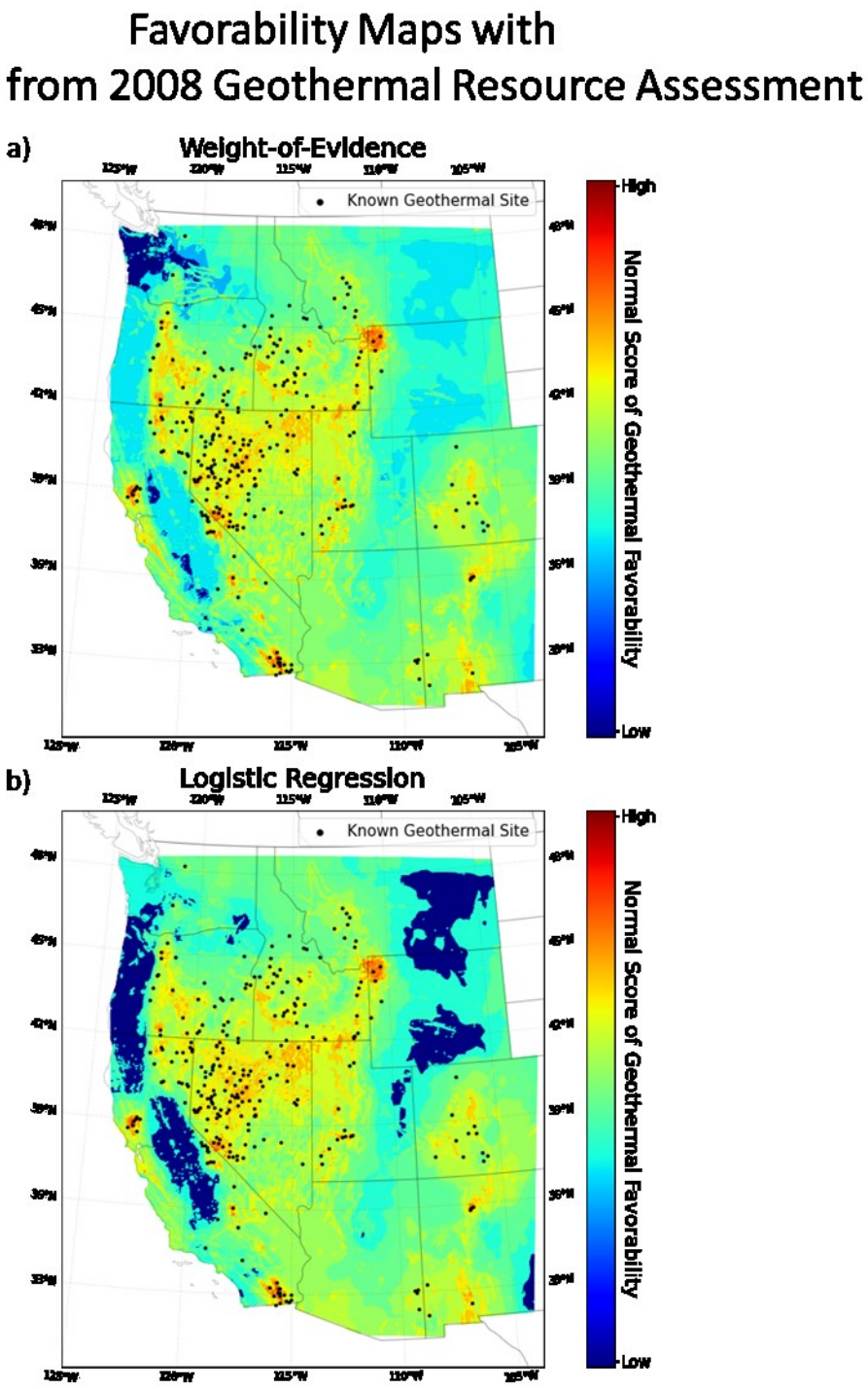

Figure 6: Geothermal favorability maps of the western United States using the methods of Williams and DeAngelo (2008): a) weight-of-evidence and b) logistic regression. Favorability is the normal score transform of probability.

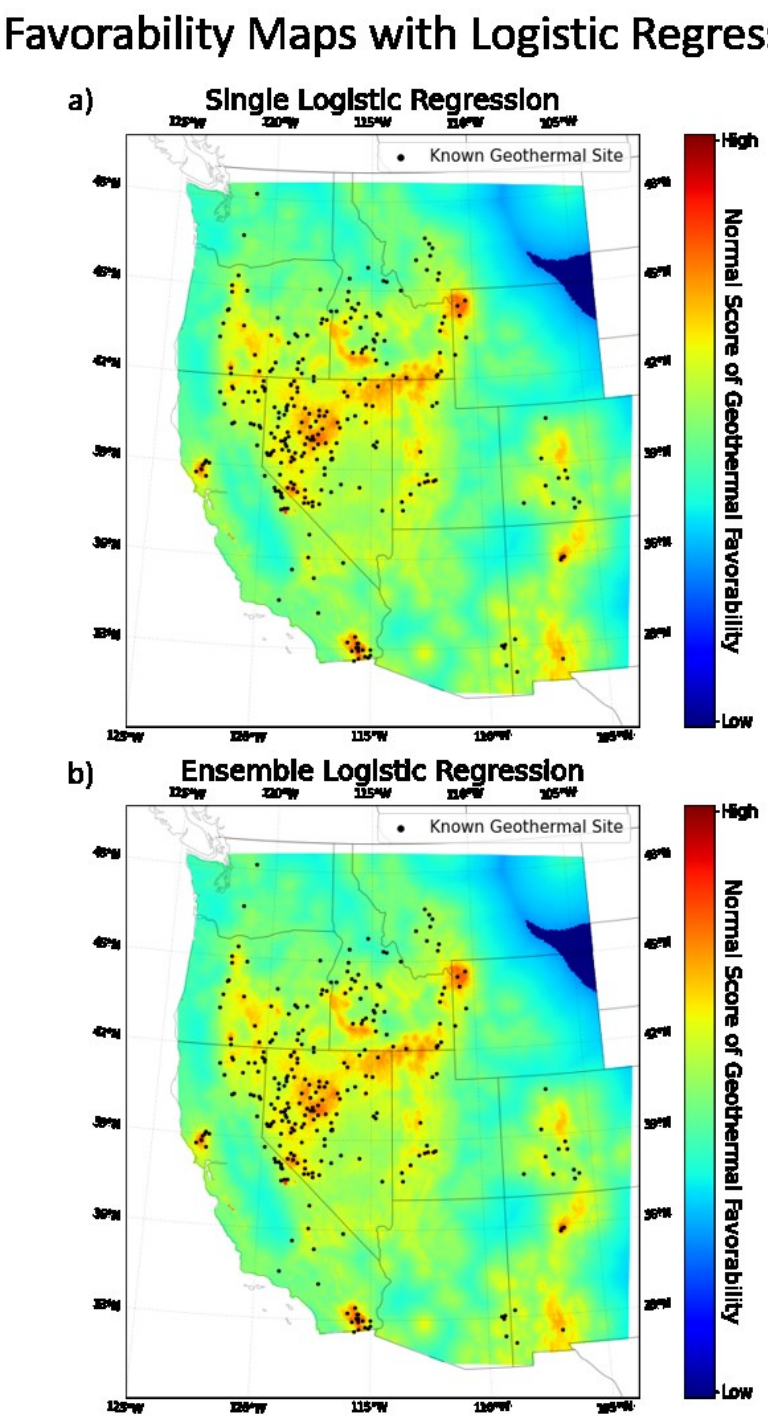

Figure 7. Favorability maps for a) single logistic regression and b) ensemble logistic regression. Favorability is the normal score transform of probability.

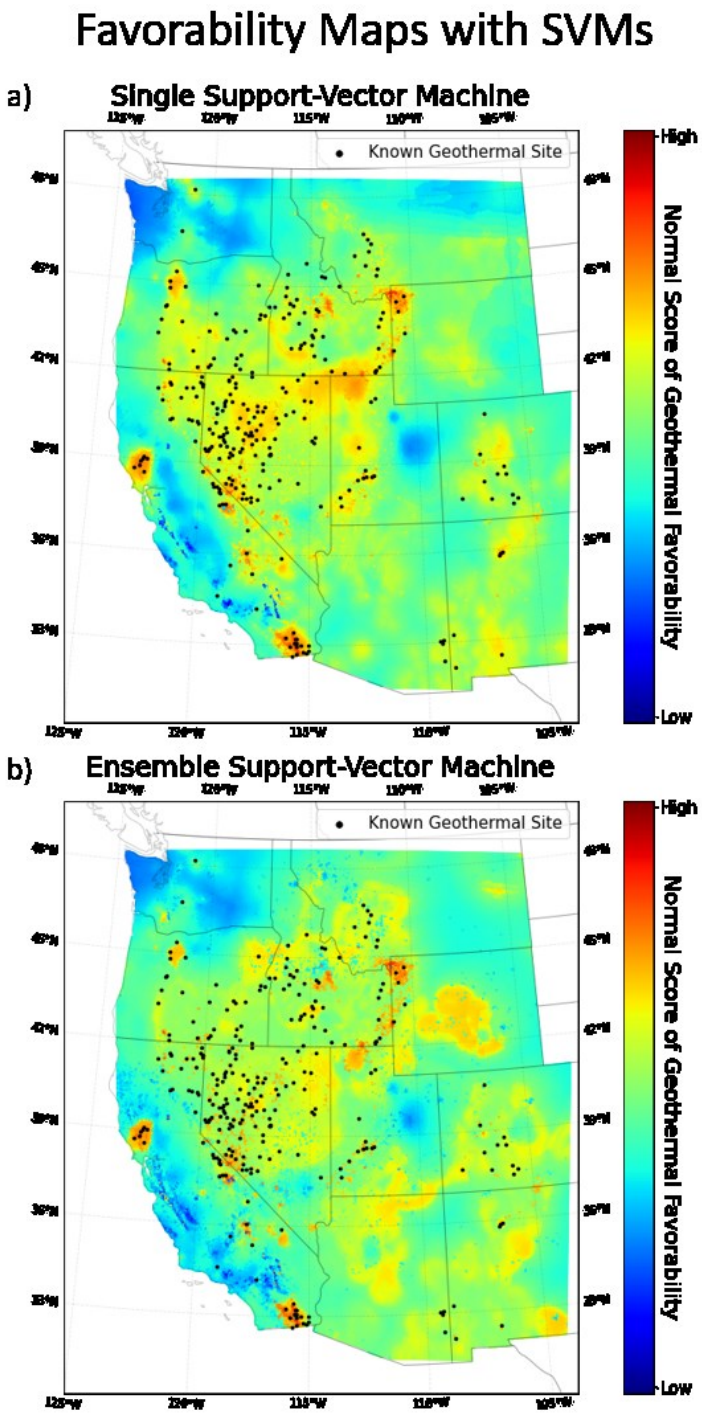

Figure 8. Favorability maps for a) single SVM and b) ensemble SVM. Favorability is the normal score transform of the *n*-dimensional distance of a label to the decision boundary in the space defined by the kernel-trick. Distance is positive on the positive side of the boundary and negative on the negative side of the boundary.

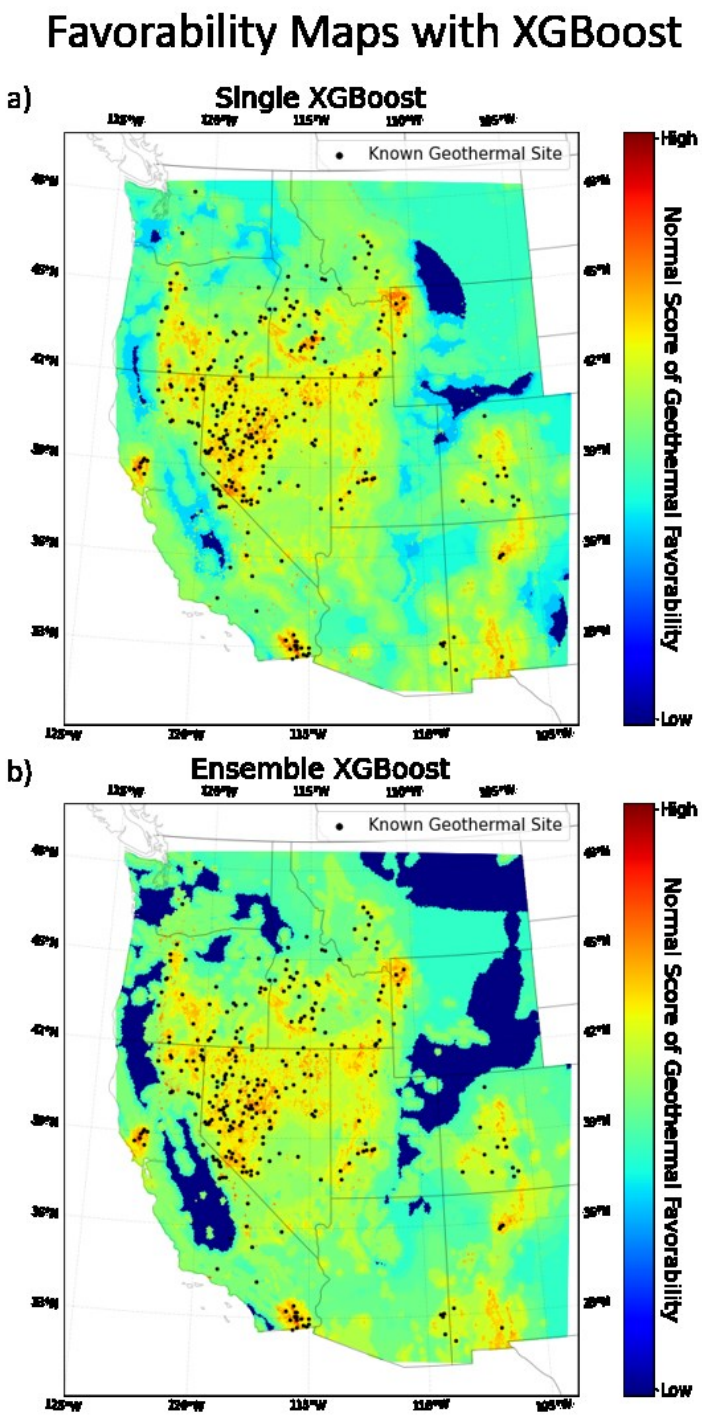

Figure 9. Favorability maps for a) single XGBoost and b) ensemble XGBoost. Favorability is the normal score transform of probability.

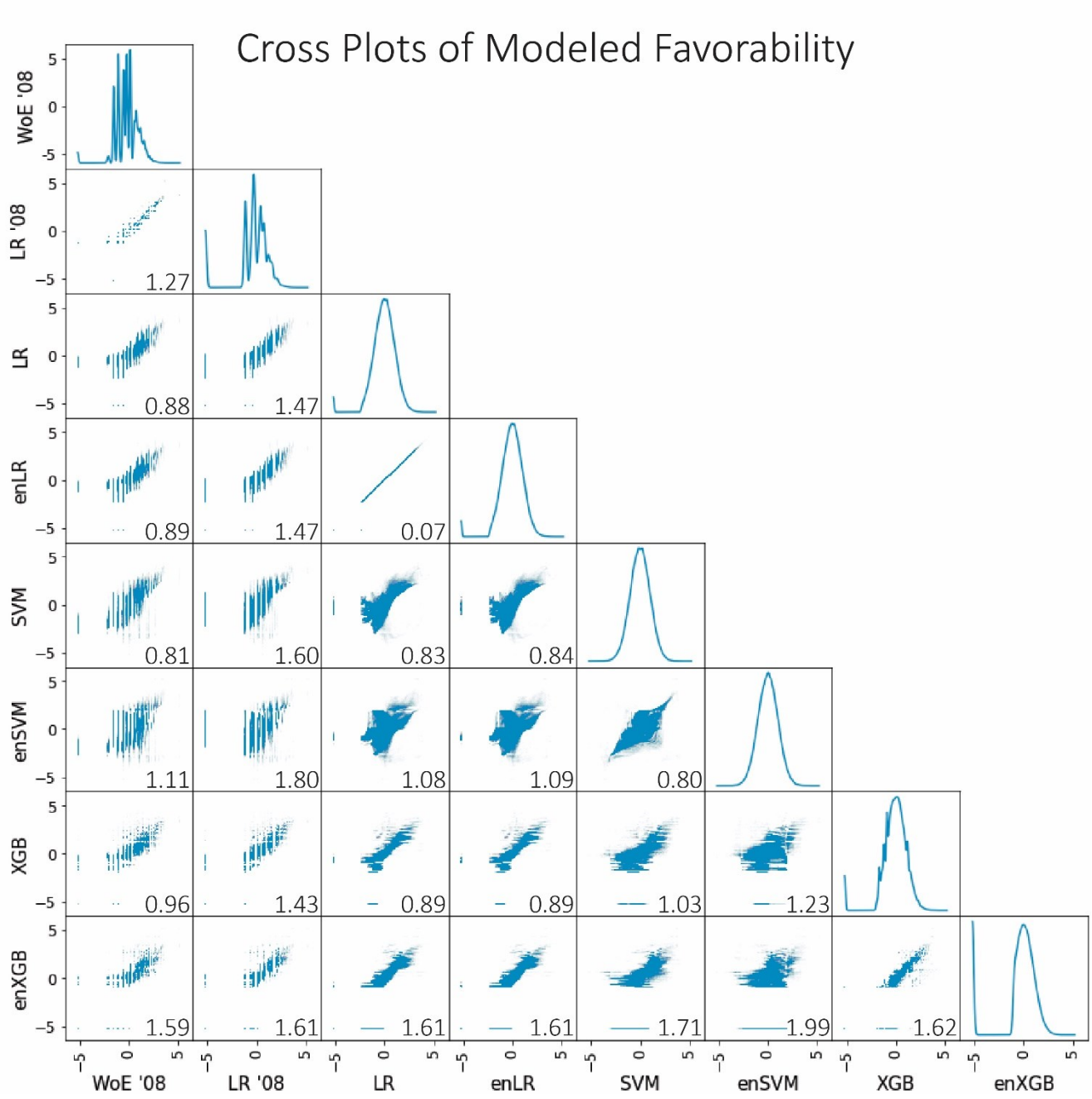

Figure 10. Cross plots of predicted favorability at every location for the different approaches for geothermal resource assessment (*i.e.,* favorability predictions from Figs. 6, 7, 8, 9). The number in each plot is the root mean square error (sum of square differences at all cells), so low values indicate 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 and buffering, 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, SVM – Single Support-Vector Machine, enSVM – Ensemble Support-Vector Machine, XGB – Single XGBoost, enXGB – Ensemble XGBoost.

3.2 Performance: F1 Scores

The strategy-algorithm pairs (*i.e.,* every combination of the two strategies and three algorithms) considerably overlap in their performance with respect to F1 scores (Fig. 11; Table 1). Although the median F1 scores of the 6 algorithm-strategy pairs are similarly low (< 0.04), two important distinctions can be made. First, the simplest algorithm (*i.e.,* logistic regression) appears to have the highest median F1 score compared to that of other algorithms when either strategy is considered. Similarly, logistic regression also has the largest inter-quartile range (Fig. 11). Second, only the ensemble SVM has a first-quartile (*i.e.,* 25th-percentile) F1 score of 0; hence, the ensemble SVM is more

likely to misclassify known positives than any of the other strategy-algorithm pairs. Thus, the ensemble SVM is more conservative than the other strategy-algorithm pairs when predicting positive labels.

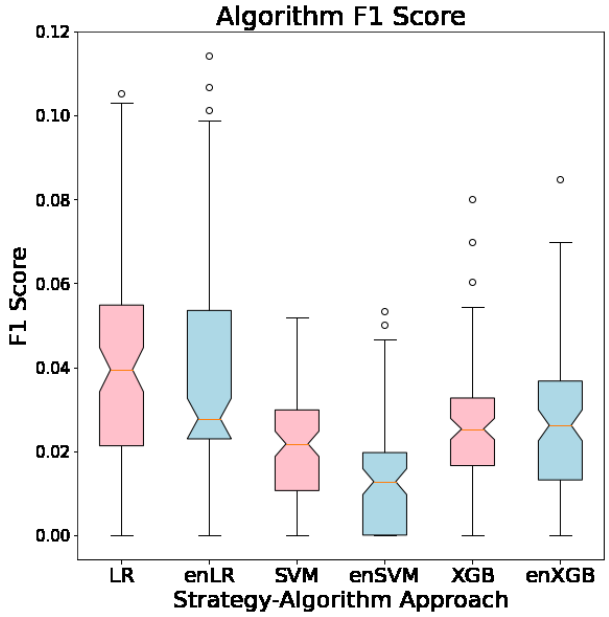

Figure 11: Box and whisker plots of F1 scores for each strategy-algorithm pair (see also Table 1). The single strategy approaches are in red, and the ensemble strategies are in blue. Boxes extend from the first quartile (Q1) to the third quartile (Q3) with a notch and line at the median. The whiskers 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, SVM – Single Support-Vector Machine, enSVM – Ensemble Support-Vector Machine, XGB – Single XGBoost, enXGB – Ensemble XGBoost.

Table 1: Algorithm Performance. Fitting Time provides a relative estimate of the processing time to fit an algorithm using all 725,442 cells from the 2008 USGS geothermal resource assessment to produce a single model on a Windows 10 computer with a 2.4-GHz, 8-core CPU and 64 GB of RAM. 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, 95th – 95th percentile value, SD – Standard Deviation.

Strategy & Algorithm	Fitting Time	F1	Class Weight	Inverse Reg. Str.
Single Logistic Regression	< 1 min	**0.039**	226	90.10
Single LR 95th / SD		***0.098***	*14*	*8.18*
Ensemble Logistic Regression	< 2 min	**0.028**	49	0.40
Ensemble LR 95th / SD		***0.085***	*4*	*0.08*

Strategy & Algorithm	Fitting Time	F1	Class Weight	Inverse Reg. Str.	Gamma
Single SVM	10 hours	**0.022**	772	11.60	0.001
Single SVM 95th / SD		***0.038***	*65*	*6.48*	*0.002*
Ensemble SVM	2.5 hours	**0.013**	102	1.00	0.036
Ensemble SVM 95th / SD		***0.041***	*8*	*0.04*	*0.042*

Strategy & Algorithm	Fitting Time	F1	Class Weight	Learning Rate	*n* of Estimators	Max Depth
Single XGBoost	40 min	**0.025**	237	0.37	13	3
Single XGBoost 95th / SD		***0.049***	*54*	*0.22*	*10*	*1*
Ensemble XGBoost	30 min	**0.026**	56	0.05	72	4
Ensemble XGBoost 95th / SD		***0.058***	*6*	*0.01*	*8*	*1*

3.3 Interpreting Hyperparameter Values

The differences of hyperparameter values between the single and ensemble strategies reflect the structural differences of the strategies. Foremost, class weighting is generally an order of magnitude less in the ensemble-algorithm pairs than in the single-algorithm pairs (Table 1). 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 and single XGBoost have a class weight significantly less than 700, which would reflect the estimated 1:700 positive:negative natural class imbalance as estimated using the results from Williams et al. (2008). The difference between optimal class weights and the estimated natural class imbalance suggests that the estimated number of naturally occurring geothermal systems may be inaccurate. While we use the mean power production as modeled by Williams et al. (2009) to estimate the number of naturally occurring geothermal systems in the western United States, the class weighting for single logistic regression and single XGBoost suggest that the true number of naturally occurring geothermal systems may be greater than our estimate of 1040. However, even with the estimated power potential in the western United States at 5% probability from Williams et al. (2009), we would anticipate a positive:negative class imbalance of 1:550 (see Equations 6 – 8), which is still greater than twice that as suggested by the weighting of single logistic regression and single XGBoost. Hence, the class weighting of single logistic regression and single XGBoost suggests that the number of naturally occurring geothermal systems in the western United States exceeds the 5% probability estimate derived from Williams et al. (2009).

The ensemble logistic regression and ensemble SVMs required lower inverse regularization strength (*i.e.,* more regularization) than their single strategy variants (Table 1). These ensemble approaches require greater regularization than their single-algorithm variants because each ensemble-strategy model is fit from a fraction of the cells used to fit a single-strategy model. Fewer cells during fitting make an algorithm more prone to overfitting without some form of regularization (see generally Burkov, 2019). Similarly, accommodation for the increased risk of overfitting with the ensemble strategy is also apparent with XGBoost's hyperparameters. Not having an inverse regularization strength hyperparameter, XGBoost instead relies on tuning the learning rate, number of estimators, and max depth hyperparameters to prevent overfitting. Although ensemble XGBoost has a deeper optimal max depth and a greater number of estimators than single XGBoost, the potential influence of overfitting by these hyperparameter values is offset by ensemble XGBoost's lower learning rate than single XGBoost's learning rate.

3.4 Computational Requirements

For practical consideration, we provide the approximate processing time for each strategy-algorithm pair in Table 1. The ensemble method did not appear to be faster than the single method when using logistic regression, which was already relatively fast (< 1 minute). The negligible change in processing time is primarily related to the processing requirements of logistic regression scaling linearly as training data grow larger. Ensemble logistic regression requires more processing time than single logistic regression because of the resources needed to aggregate the ensemble results. Fitting with the ensemble SVMs and XGBoost is faster than the single variants because the resources required by SVMs and XGBoost increase at greater than a linear rate with each additional sample in the training data. Hence, training from several small datasets is faster than training from a single, large dataset when using SVMs and XGBoost.

4. FUTURE OPPORTUNITIES

The strategies and algorithms 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 with fewer choices by experts. Yet, several challenges remain. The USGS geothermal resource assessment team is currently working towards answering all of the following questions for the next generation of geothermal resource assessments:

- How can the F1 scores be improved? The mean F1 scores from the strategy-algorithm pairs in this study are < 0.04, suggesting a True Positive:[False Positive + False Negative] ratio of < 1:50 (see Equation 4). How can the strategies and algorithms be modified to increase the prediction rate of true positives? Alternatively, Lee and Liu (2003) suggest a new performance metric that does not rely on false positives in its calculation and might serve as a proxy for the F1 score. Would this new performance metric be appropriate for predicting geothermal favorability? Could true negatives be incorporated into assessing model performance?

- Can we identify features that are better predictors of geothermal favorability? Engineering new features that better represent geological conditions as they relate to geothermal favorability would aid in the development of improved geothermal resource assessments. Feature engineering includes producing improved interpolations from point data and engineering features that represent properties not yet included in the 2008 feature sets (*i.e.,* permeability).

- What other methods could address the positive-unlabeled aspect of the data other than using an F1 score as the performance metric? Bekker and Davis (2020) suggest several methods for training with positive-unlabeled data, like using semi-supervised algorithms to identify reliable negatives.

- What other methods can be used to address the class imbalance? Although the ensemble strategy we present in this study serves as a means to reduce class imbalance from 1:2600 to 1:700, class imbalance is still classified as extreme even at 1:100 (see generally Krawczyk, 2016). Reducing the class imbalance will improve the prediction capabilities of any algorithm used.

- 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 would 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 abandon workflows with grids entirely.

- 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 hydrothermal 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. Similarly, 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?

5. CONCLUSIONS

In this study, we compare geothermal favorability models for the western United States created with the data and methods from the 2008 USGS geothermal resource assessment, which relied on expert decisions, to geothermal favorability models created from the same data but with machine learning strategies and algorithms. In so doing, the expert decision-dependent and machine learning approaches show general agreement, demonstrating that the machine learning algorithms present a means to produce the geothermal favorability maps from the 2008 geothermal resource assessment while minimizing the biases of expert decisions. We also demonstrate how the expert decisions from the 2008 geothermal resource assessment (*e.g.,* binning and buffering) of the input feature sets effectively rendered the otherwise linear methods used therein (*i.e.,* logistic regression and weight-of-evidence) to become non-linear. We also find that the non-linear approaches produced the greatest variability in predictions for high geothermal favorability. The models produced by the machine learning approaches performed similarly with ubiquitously low F1 scores, emphasizing the need for additional research to address the challenges inherent to geothermal data (*e.g.,* positive-unlabeled data, extreme class imbalance) and improve the predictive capabilities of machine learning with geothermal resource assessments.

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.

REFERENCES

ANSS 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. *Machine Learning, 109*, 719-760. doi:10.1007/s10994-020-05877-5

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

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

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

Branco, P., Torgo, L., & Ribeiro, R. (2015). A Survey of Predictive Modelling under Imbalanced Distributions. *Computing Research Repository*, 1-48.

Burkov, A. (2019). Chapter 5: Basic Practice. In *The Hundred-Page Machine Learning Book* (pp. 43-60): Burkov, Andriy.

Chapelle, O. (2007). Training a Support Vector Machine in the Primal. *Journal of Machine Learning Research, 19*, 1155-1178. doi:10.1162/neco.2007.19.5.1155

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

Cortes, C., & Vapnik, V. (1995). Support-Vector Networks. *Machine Learning, 20*, 273-297. doi:doi.org/10.1007/BF00994018

Donnelly-Nolan, J. M. (1988). A magmatic model of Medicine Lake volcano, California. *Journal of Geophysical Research, 93*, 4412-4420.

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*. Cham, Switzerland: Springer.

Hildreth, W. (2007). Quaternary magmatism in the Cascades – Geologic Perspectives. *U.S. Geological Survey Professional Paper 1744*, 125 p. doi:10.3133/pp1744

Krawczyk, B. (2016). Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence, 5*, 221-232. doi:10.1007/s13748-016-0094-0

Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling* (1st ed.). New York, USA: Springer.

Lee, W. S., & Liu, B. (2003). *Learning with Positive and Unlabeled Examples Using Weighted Logistic Regression.* Paper presented at the Twentieth International Conference on Machine Learning.

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*, 2 sheets, scale 1:62,500, pamphlet, 23. doi:10.3133/i2455

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

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., . . . others. (2011). Scikit-learn: Machine learning in Python. *Journal of machine learning research, 12*, 2825-2830.

Reed, M. J. (1983). Assessment of low-temperature geothermal resources of the United States- 1982. *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. (2014). Chapter 16: Kernel Methods. In *Understanding Machine Learning: From Theory to Algorithms* (pp. 215-226). New York, New York, USA: Cambridge University Press.

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). *Special Paper of the Geological Society of America, 397*, 61-71. doi: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 Transactions, 32*, 181-188.

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

Williams, C. F., Reed, M. J., Galanis, S. P., & DeAngelo, J. (2007). The USGS national geothermal resource assessment: An update. *GRC Transactions, 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. *U.S. Geological Survey Fact Sheet 2008-3082*, 1-4.