Energy 313 (2024) 133904

Contents lists available at ScienceDirect

Energy

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

Forecasting energy poverty using different machine learning techniques for Missouri

Sarah Balkissoon [a,*], Neil Fox [a], Anthony Lupo [a], Sue Ellen Haupt [c], Stephen G. Penny [d,b], Steve J. Miller [e], Margaret Beetstra [f], Michael Sykuta [g], Adrienne Ohler [g]

[a] Atmospheric Science Program, School of Natural Resources, University of Missouri, USA
[b] Cooperative Institute for Research in Environmental Sciences, Boulder, CO, USA
[c] Research Applications Lab, NSF National Center for Atmospheric Research, Boulder, CO, USA
[d] Sofar Ocean, San Francisco, CA, USA
[e] Department of Environmental Studies, University of Colorado, Boulder, USA
[f] The Nurture Nature Center, Easton, PA, USA
[g] Financial Research Institute, University of Missouri, USA

ARTICLE INFO

Keywords:
Energy poverty
Decision trees
Random forest
Extreme gradient boosting (XGB)
Support vector machine (SVM)

ABSTRACT

Energy poverty in Missouri was analyzed using the four quadrant approach using both state and county level data sets for two separate definitions of the grid. Predictions used machine learning techniques including decision trees, random forest, extreme gradient boosting and support vector machines. It was determined that the extreme gradient boosting performed the best when compared to all the other models after the hyperparameters were tuned. The F1 scores for the county level data sets were higher than for the state levels thus indicating greater predictability for the National Oceanic and Atmospheric Administration (NOAA) climatological regional runs. For the county level data, the F1 score was the highest for region 1, which coincided with one of the highest expenditure risk values whilst regions 2–4 were the lowest scoring area. In grid 2, the largest class distribution changed from grid 1's expenditure risk to the no risk category. This grid had more variability in terms of the double risk class when compared to grid 1 and, as such, its predictability in terms of its F1 scores was reduced. There were similarities in the ranking of the prediction scores for the regions for both grids as regions 1 and 6 incurred the largest F1 values. Thus energy poverty can be classified and predicted for Missouri, which in turn may aid policy makers via quantitative regional risk analysis. This data-driven informed policy making can lead to the development and implementation of laws and social programs to help ameliorate energy poverty.

1. Introduction

1.1. Energy poverty and its definition

There is no universal definition of energy poverty [1]. However, energy poverty can be defined, in developed countries, as the inability of households to pay their energy bills [2]. There are objective and subjective indicators of energy poverty. The objective indicators are based on a measure of energy poverty that includes household income and energy expenditure, whilst the subjective indicators are based on perception of energy cost and maintaining the basic standard of living [3]. The objective indicators include the Ten Percent Rule, Low Income High Cost, Minimum Income Standard and Compound Energy Poverty Indicator. The Ten Percent Rule applies when a household spends 10% or more of its income on energy which results in cases when households with high income are considered energy poor [3]. The Low Income High Cost definition rectifies this problem by considering both the income and energy expenditure. The objective indicator methodology of Low Income High Cost is used in our study with two defined grids and their corresponding energy poverty classifications compared. Two grids were used as there is an arbitrariness in the selection of income and expenditure thresholds, as it is difficult to define energy poverty objectively [4].

1.2. Policy implications

The aim of this project is to aid policy makers in preventing energy poverty by quantitative analysis of the risk areas within Missouri. This data driven informed policy making via state legislature can aid in the development and implementation of laws, social programs, funding opportunities and other initiatives to ameliorate their economic

* Corresponding author.
E-mail address: sarahsharlenebalkissoon@mail.missouri.edu (S. Balkissoon).

https://doi.org/10.1016/j.energy.2024.133904
Received 18 March 2024; Received in revised form 17 October 2024; Accepted 19 November 2024
Available online 27 November 2024

hardship and energy poverty. The social and economic roots that lead to a household experiencing energy poverty, in the first instance, can be addressed. The reduction of energy poverty can lead to the decrease in aliments of physical and mental health as well as gender, education and standard of living disparities as well as socioeconomic deterioration [3,5].

Prediction is also relevant particularly to businesses, such as electric utilities, that provide necessary services, but do not want to collect and access sensitive data such as income. However, affordability of these services is an important factor to utility regulators, and better prediction methods of energy poverty could help utilities and utility regulators implement programs towards households most likely to need assistance without having to collect these sensitive information.

1.3. Machine learning (ML) techniques in classification

There is a lack of conventional big data in the analysis of energy poverty [1]. Despite this limitation, socioeconomic data was sourced for Missouri and machine learning techniques were used in this study. Traditional regression methods were not utilized because of their difficulty in handling large data and their assumptions that there exists correlations amongst the features [4,5]. ML methods are capable of analyzing big data by automatically detecting patterns in the data without assuming a priori, correlations amongst the variables. Also, nonlinearity of the complex systems can be investigated using ML techniques unlike conventional regression methods, which are purposed for linear problems [3,4]. The application of eXplainable Artificial Intelligence (XAI) can lead to the knowledge about the input–output relationships as well as the predictors of energy poverty [2]. According to [5], there is no one socioeconomic driver that determines household energy poverty. Empirical evidence from [5] also shows that the socioeconomic aspects of a household are important indicators of susceptibility to energy poverty. Hence the determinants as well as their relationship with energy poverty is of importance.

2. Data

There is a scarcity of socioeconomic data for Missouri that include household habitat data [1]. Despite the lack of data, energy poverty for the state and counties in Missouri for households was obtained from [6] for the year of 2018. The area median income data from the Low Income Energy Affordability Data (LEAD) was utilized in this study. There are, for the state of MO, 776,060 data points in this data set. For this state the variables and their categorical values considered in the analyses are given in Table 2. HINCP, ELEP, GASP, FULP, defined in Table 2, included null values, which were inputted after pre-processing of the data with their respective median inputs. The predictand, Colors, was derived from the four quadrant approach. Colors are defined as the risk type categories. The yearly income threshold was based on the minimum wage of MO in 2018. It was determined to be $14,082.90. This was in keeping with [4], where the income threshold was defined as the 2015 minimum income level stipulated by the Dutch government. Yearly expenditures was given as the sum of average household electricity, gas and other fuel. This was also done as in [4]. For the first grid, yearly expenditures exceeding 10% were also determined from the ratio of FUEL_EXPENDITURE to the max FUEL_EXPENDITURE; this was labeled 'PERCENT_EnergyExp'. These two numeric values are the lines of demarcation for the 4-quadrant grid. The quadrants were labeled as No risk or Color 0 when HINCP > $14,082.90 and PERCENT_EnergyExp < 10. Expenditure risk or Color 1, occurs when the households income was greater than the annual accumulation of minimum wages but the percent expenditure is greater than 10%. Income risk, Color 2 is where the annual income is less than or equal to the defined income line of demarcation and energy expenditure is less than or equal to the 10% line. Double risk, Color 3, is where households spends more than 10% but earns less than the yearly minimum wage accumulation for 2018.

Table 1
Statistics for the variables used in the state wide data set.

Statistics	HINCP	FUEL_EXPENDITURE
count	77.760600e+05	77.760600e+05
mean	4.115009e+04	1.921706e+03
std	2.587607e+04	8.286539e+02
min	−9.11537e+03	1.168294e−08
max	8.8142383e+05	1.787769e+04

The 10% demarcation line of energy expenditure was chosen in keeping with the standard classification procedures used in literature [4].

The statistics for the variables used in the grids construction of the state wide data set is shown Table 1.

Three data sets were considered as in [4]. The first included all the variables in the table except for fuel expenditure, as including this feature will incur 100% prediction accuracy of methods. The second, had all variables except expenditure and income. Thirdly, the last permutation only considered the income column. These were labeled respectively as df_MO_A, df_MO_B and df_MO_C.

To prevent overfitting and to find the optimal parameters in the training process, cross validation was used. The model was verified using 3 repeated 10-fold cross validations. This means that the original data were split randomly into 10 new training and validation sets. These evaluation data sets were repeated 3 times to conduct a performance test in the validation stage. The data are shuffled and then split into the k (in our case, 10) unique groups. After the model was fitted and evaluated, the model is discarded whilst the score is retained before the process is repeated. The stratification aspect ensures that each fold has the same portion of observations belonging to a particular category [3].

For the models, the metrics used were accuracy and F1 score. Accuracy is defined as the ratio of correctly predicted observations to the total number of observations. However, this metric is more suited to cases where the data sets are balanced in terms of class distribution. F1 score corrects this by determining class-wise performance, which is written in terms of precision and recall as seen in Eq. (1). In the multiclass calculation of this metric, this score is determined individually for each class, as in our study. The determination of the precision and the recall parameters for say, class 0 is calculated from Eq. (2). The implementation of this metric is done using Python sci-kit learn library. The net F1 score is computed by utilizing various averaging techniques such as macro averaging, micro averaging and weighted-averaging. The weighted average, used in this study, is best for imbalanced class distribution as in our case. The weighted F1 score for N classes is defined in Eq. (3). Please see Fig. 1 for the data set for MO, df_MO_A where there is a disproportionate distribution for Color 0 or no risk.

$$F1score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{1}$$

$$Precision(class = 0) = \frac{TruePositive(class = 0)}{TruePositive(class = 0) + FalsePositive(class = 0)}$$
$$Recall(class = 0) = \frac{TruePositive(class = 0)}{TruePositive(class = 0) + FalseNegative(class = 0)} \tag{2}$$

$$weightedF1score = \sum_{i=1}^{N} w_i \times F1score_i$$
$$w_i = \frac{No.OfSamplesInClassi}{TotalNumberOfSamples} \tag{3}$$

3. Methodology

Machine learning algorithms are utilized because of their ability to explore nonlinear relationships between the response variables and the predictor variables [3]. According to [1], the most commonly used algorithms for AI energy poverty prediction are decision trees and artificial neural networks. Decision trees are one method implemented in this study with the methodology described below.

Table 2
Description of variables input into analyses.

Variables	Description	Categories
TEN_CATEGORY	Type of tenants	0 -Owner 1-Renter
YBL6_CATEGORY	Year the building was first constructed	0- 2010 and later 1 -2000 to 2009 2- 1980 to 1999 3- 1960 to 1979 4-1940 to 1959 5- before 1940
BLD_CATEGORY	Number units in building/ type of dwelling	0 - single family detached home 1 - single family attached home 2 - 2 unit multifamily home/ apartment 3 - 3–4 unit multifamily home/apartment 4 - 5–9 unit multifamily home/apartment 5 - 10–19 unit multifamily home/ apartment 6 - 20–49 unit multifamily home/apartment 7 - 50 + unit multifamily home/ apartment 8 - boat, recreational vehicle or van 9 - mobile or trailer home
HFL_CATEGORY	Primary space heating fuel type	0 -utility/natural gas 1- bottled/propane or liquefied petroleum gas 2- electricity 3-fuel Oil 4-coal 5- wood 6-solar 7- other 8- none
AMI68_CATEGORY	Area Median Income	0 - < 30% 1- 30%–60% 2- 60–80% 3-80–100% 4- 100+ %
UNITS	Number of occupied housing units or households	–
HINCP	Average annual household income	–
HINCP_UNITS	Total HINCP x UNITS	–
FUEL_EXPENDITURE	Average household electricity (ELEP) + gas (GASP) + other fuel (FULP) expenditures	–

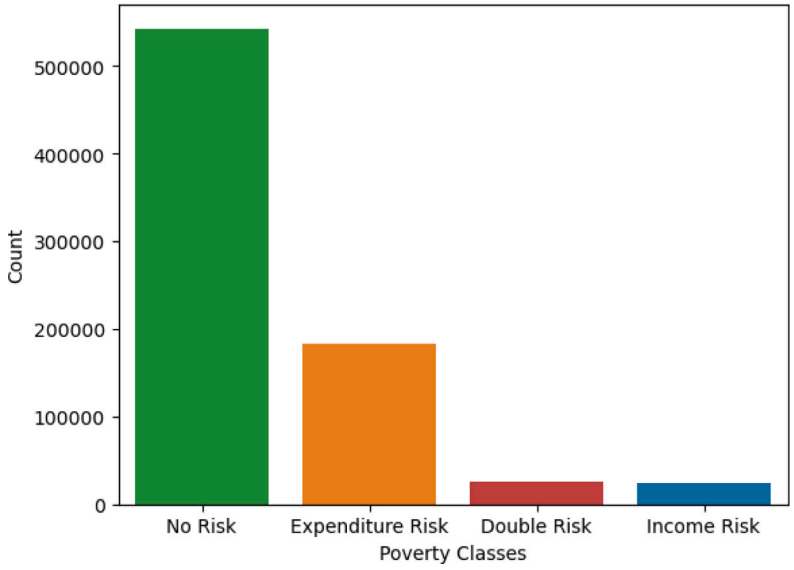

Fig. 1. Class distribution of households within MO or count.

3.1. Decision trees

Decision trees are tree-like structure algorithms with connections or nodes to learn patterns for the purpose of classification and prediction [3]. This method uses the decision tree to split the data by asking questions based on the columns or features [7]. This process of branching is continued until the algorithm reaches a desired accuracy [7]. According to [3], there are four steps in this method. The growing step considers suitable split and stopping rules that determines the time for expanding the nodes. The pruning step is where branches, which cause overfitting, are removed. The validation step is where the decision trees are evaluated using cross-validation with data. In the

prediction and the interpretation phases of the process, the tree model is established and prediction is completed. The split criterion is based on the impurity index called the Gini Index [3]. This error method is used by the decision tree to determine how the split should be made. The lower the Gini index, the lower the error. The Gini index is define mathematically below.

$$gini = 1 - \sum_{i=1}^{c} (p_i)^2 \tag{4}$$

where p_i is the probability of the split resulting in the correct value or the fraction of the set with the correct output label and c is the total number of classes.

The description of the hyperparameters to be tuned are as follows [7].

1. max_depth — this is the depth of the tree or the number of times splits are made. This is utilized to reduce overfitting.
2. min_samples_leaf — this determines the minimum number of samples a leaf may have. This is also used to minimize overfitting.
3. max_leaf_nodes — this hyperparameter specifies the maximum number of leaves in the tree.
4. max_features — this considers a subset of the features for making a split.
5. min_samples_split — this gives the minimum number of samples required before a split is made.
6. splitter — this determines how the model would select the features to split each branch. The 'best' splitter selects features which would give the greatest gain of information whilst the 'random' option splits randomly.
7. criterion — this is the scoring method for the splits.
8. min-impurity_decrease — this specifies when to split for a particular impurity.
9. min_weight_fraction_leaf — of the total weights, this is the minimum weighted fraction necessary to be a leaf.

3.2. Random forest (RF)

Random forests are ensemble models that combine various versions of the same model, decision trees, through bagging and bootstrapping [3,7]. To reduce prediction errors, bootstrapping (sampling with replacement) is used to randomly select samples and features. It was estimated mathematically that for each decision tree, two thirds of the samples are unique whilst the remainder are duplicates [7]. This creates independent decision trees that collectively reduce overfitting issues [3]. This algorithm utilizes Out-of-bag (OOB) infrastructure where the test data are the data samples not extracted from the bootstrapping method [3]. The random forest uses majority rule for classifiers, where after each decision tree makes its prediction, the final result is aggregated using majority rules [7].

There are additional hyperparameters for the random forest [7]. They are:

1. oob_score — When this is true, it is not necessary to split the data into train and test. The samples not chosen in the bagging procedure will be used as the test data set.
2. n_estimators — This parameter determines the number of trees in this algorithm. As this parameter increases, the score plateaus.
3. warm_start — This hyperparameter is useful when determining the optimal number of trees. When true, the model does not start from the beginning when adding more trees; it restarts from where the previous model ended.
4. bootstrap — This is sampling with replacement, if this is listed as false, then oob_score cannot be set as true.
5. verbose — This is used to display more information of the model.

The other hyperparameters which are the same as for the decision tree algorithm, but they are not as important for this model.

3.3. Extreme gradient boosting (XGB)

This method, unlike the random forest, is based on boosting, an algorithm that learns the mistakes of the trees and adjusts the new trees based on errors derived from previous trees [2,7]. Unlike bagging, these trees are constructed and operated in isolation [7]. Since boosting does not focus on developing a strong baseline model, the algorithm pays attention to the transformation of weak to strong learners via iterations [3,7]. To determine the score of the model, the sum of the residuals of the individual trees is calculated.

There are hyperparameters for this model as well. The two most important parameters are learning_rate and n_estimators as described below. Generally, as the n_estimators increase, the learning_rate should be decreased [7].

1. n_estimators — this as mentioned previously, determines the number of trees in the ensemble.
2. learning_rate or eta — this determines the contribution of individual trees to the model. This value is contained in the closed interval [0, 1]. It defaults to 0.1, which implies that the trees in the model have an influence of 10%.
3. max_depth — this determines the depth of the tree
4. Lagrange multiplier — nodes must exceed this value before other splits are made in accordance with the loss function- it restricts when splits are made, and thus causes shallower trees.
5. min_child_weight — This is the minimum sum of the weights required for a split.
6. subsample — this determine the number of rows for each boosting.
7. colsample_bytree — this randomly selects columns or features for model run.
8. booster or base learner — this is the initial decision tree of the ensemble. It defaults to the common usage of gbtree or gradient boosted tree.

3.4. Support vector machine (SVM)

SVM is a generalization of the maximal margin classifier. First we define what is a hyperplane. In p-dimensional space, it is a subspace that is flat, affine and of $p-1$ dimensions. Thus for 2 or 3 dimensions, a hyperplane is a line or a plane, respectively. Mathematically, for 2 dimensions a hyperplane is defined by Eq. (5) [8]

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 = 0 \tag{5}$$

where β_0, β_1 and β_2 are parameters. Any $X = (X_1, X_2)^T$ which Eq. (5) holds, is a point on the hyperplane. Generally, in p-dimensions, a hyperplane can be written as

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p = 0 \tag{6}$$

where any point $X = (X_1, X_2, \ldots, X_p)^T$ in p-dimensional space that satisfies Eq. (6), lies on the hyperplane. If Eq. (6) is greater than or less than 0, then X lies on opposite sides on the hyperplane. The concept of a separating hyperplane is used as a classifier for the data. For the classification of test observation x^*, the sign of the equation $f(x^*) = \beta_0 + \beta_1 x_1^* + \cdots + \beta_p x_p^*$ determines the class observation assignment. Also, if the magnitude of the $f(x^*)$ is furthest away from 0, then we are more certain or confident about the assignment of the class [8].

The optimal way of separating the hyperplane, termed maximal margin hyperplane, is the hyperplane that has the largest margin or has the largest minimal distance from the observations to the hyperplane. The maximal margin classifier is based on the classification of the test observations' relative positioning with respect to the maximal margin hyperplane. The support vectors are the observations that are contained

within the width or on the wrong side of the margin in which the maximal margin hyperplane is dependent upon [8].

However, there may exist cases where observations that belong to the classes may not be separated by the hyperplane. The support vector classifier allows for the misclassification of points from the training set. The number of violations of the margin is determined by the non-negative hyperparameter C, where C is defined as a bound of the sum of the ϵ_is, where ϵ_i are the slack variables, which allow the observations to be on the wrong side of the margin or the hyperplane. For a small value of C, margins are narrow, thus implying a classifier that is highly fit to the data. For larger C values, more violations or support vectors are allowed [8].

The support vector machine (SVM) is an extension of the support vector classifier where the feature space is enlarged using kernels or functions to accommodate non-linear class boundaries between classes. The linear support classifier can be mathematically represented as Eq. (7).

$$f(x) = \beta_0 + \sum_{i=1}^{n} \alpha_i \langle x, x_i \rangle \tag{7}$$

where α_i are n parameters each for a training observation. To estimate these parameters we need the inner products of all pairs of the training set, $\langle x_i, x_i' \rangle$. This can be further simplified as below.

$$f(x) = \beta_0 + \sum_{i \epsilon S} \alpha_i \langle x, x_i \rangle \tag{8}$$

where S is the set of indices where the training observations are support vectors and thus α_i are non-zero. We can replace the inner products in Eq. (8) by a generalization, that is K; a function which determines the similarity between two observations, referred to as a kernel. The linear polynomial of degree d and radial kernels can be written respectively as Eqs. (9)–(11)

$$K(x_i, x_{i'}) = \sum_{j=1}^{p} x_{ij} x_{i'j} \tag{9}$$

$$K(x_i, x_{i'}) = \left(1 + \sum_{j=1}^{p} x_{ij} x_{i'j}\right)^d \tag{10}$$

$$K(x_i, x_{i'}) = exp\left(-\gamma \sum_{j=1}^{p} \left(x_{ij} - x_{i'j}\right)^2\right) \tag{11}$$

where d is a positive integer representing the degree the polynomial and $d > 1$ implies a more flexible boundary. Also, γ is a positive constant [8]. Our problem is not a binary classification as there are $K > 2$ classes. For this we need to further extend our approach to one-versus-one classification or one-versus-all classification. The former method, $\binom{K}{2}$ SVMs compares pairs of classes. The largest total number of times a class was assigned to an observation is deemed the final classification. The latter fits K SVMs where each class is compared to the rest of the $K - 1$ classes. From this fit, the parameters are $\beta_{0k}, \beta_{1k}, \ldots, \beta_{pk}$. The observation, x^*, is assigned to the class whose value of $\beta_{0k} + \beta_{1k} x_1^* + \cdots + \beta_{pk} x_p^*$ is the largest [8].

4. Results

Fig. 2 shows the Energy-Poverty grid where each quadrant represents a class described in the data section. The classes are no risk, expenditure risk, income risk and double risk.

When the three MO data sets were used, df_MO_A (all variables except expenditure), df_MO_B (all variables except expenditure and income) and df_MO_C (only income), it was determined that the data set with all the variables except expenditure incurred the least errors and had the most accuracy for the decision tree model. The first to the third data sets had a percentage accuracy of 85.3%, 79.9% and 78.7%, respectively. The confusion matrices for these models are shown in Figs. 3 to 5. The confusion matrices show that along the diagonals, the predicted labels, which are equal to the true labels are highest for

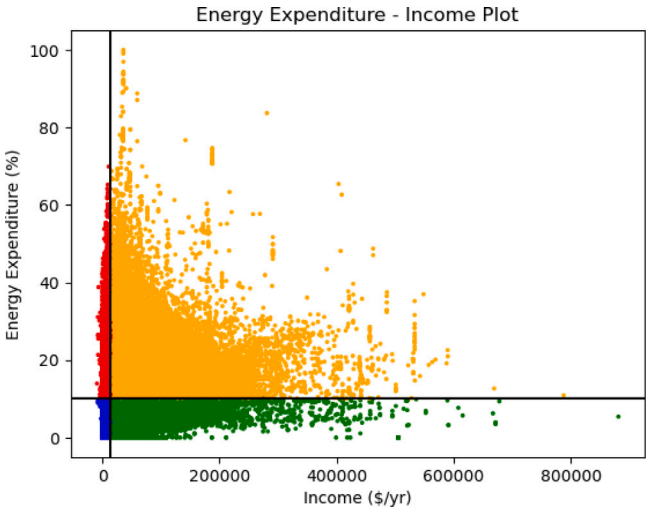

Fig. 2. Energy poverty grid for MO where green, orange, blue and red quadrants represents no risk, expenditure risk, income risk and double risk categories, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

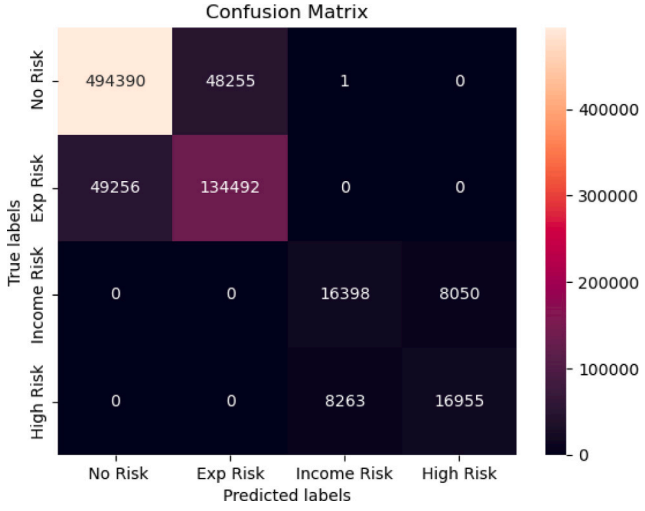

Fig. 3. Confusion matrix for the first data set using decision trees. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

the first data set. Consequently, throughout our analyses henceforth, we investigated only this first data set for MO using its specified features.

For all three datasets, the decision tree models were rerun using synthetic minority oversampling technique (SMOTE). This method is done to balance class distribution by replicating and randomly increasing the minority class examples. It is the most commonly used oversampling technique. The best performing dataset with specific features was the same as without SMOTE, df_MO_A. For these datasets, the percentage accuracy using SMOTE, when compared to runs without SMOTE, were almost the same for df_MO_A (85.3%) and df_MO_C (79.4%). The dataset with all variables except income and expenditure, df_MO_B, saw a significant decrease in accuracy by almost 10%; its accuracy score was 70.6%. However, this accuracy decrease is not an implication of the poor performance of SMOTE. SMOTE, used as in our case because of class imbalance of the original dataset, rectifies the model's propensity to predict the majority class well whilst performing poorly in predicting

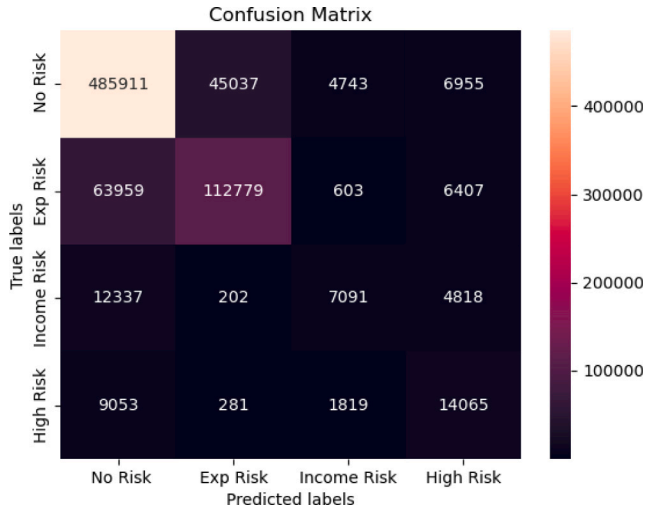

Fig. 4. Confusion matrix for the second data set using decision trees. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

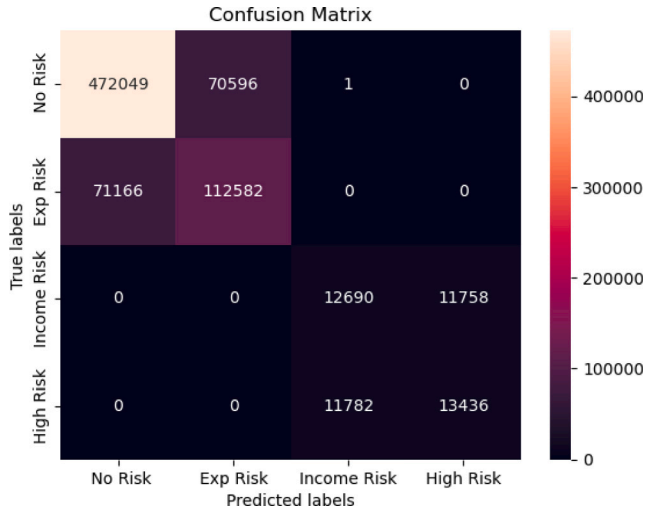

Fig. 5. Confusion matrix for the third data set using decision trees. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

the minority classes. This allows the model to predict all the classes evenly which may result in a decrease in the overall accuracy of the model as its performance in terms of accuracy of the majority class may decrease.

For our MO data set under consideration, RandomizedSearchCV was used to tune the hyperparameters. This method, unlike GridSearchCV, tries a random number of hyperparameter combinations. For the decision tree model, after inputting a range of values for the hyperparameters described in the methodology, from the randomized search, the max_features, max_leaf_nodes, min_impurity_decrease, min_samples_split and min_weight_fraction_leaf were determined to be 0.85, 30, 0.005, 10, 0.0075, respectively. The accuracy of this tuned model was computed to be 86.6%.

SMOTE was used in this tuned model. The accuracy was almost similar in magnitude to the tuned model without SMOTE. Its accuracy value was 87.0%. The various scores of the metrics for each of the classes are tabulated in Table 3 for this model run of df_MO_A.

The most important features evaluated from this model are depicted in Fig. 6. From this plot, the average annual household income is the

Table 3
Metric scores for each of the classes using the SMOTE in the tuned decision tree model with data, df_MO_A.

Class	Precision	Recall	F1 score
No risk	0.93	0.91	0.92
Expenditure risk	0.75	0.81	0.78
Income risk	0.67	0.68	0.67
Double risk	0.68	0.67	0.67
Accuracy			0.87
Macro average	0.76	0.77	0.76
Weighted average	0.88	0.87	0.87

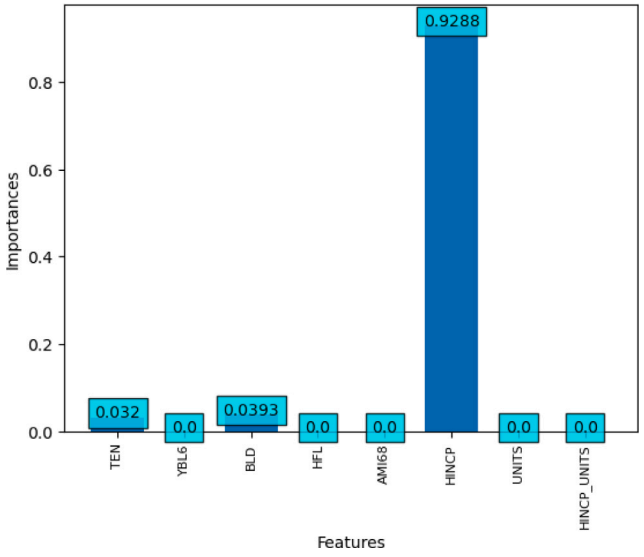

Fig. 6. Important Features determined from the decision tree model for grid 1.

most important feature followed by the type of dwelling and type of tenants. Thus we can predict energy poverty using this model with 86.6% accuracy using the three most important features of HINCP, BLD and TEN.

From [1–5], income or wealth is the most important feature. In our study, this is also the most important driver. We expected primary space heating fuel type to have more contribution in terms of feature importance because an important factor of energy poverty is access to modern forms of energy instead of the low efficiency, high pollution given by firewood, coal and kerosene amongst others [9]. However, after further analysis of the data, we discovered that for this column the data were repeating each category five times, hence it added little or no information to the model. The second and third most important features are type of dwelling and the type of tenants, respectively. Studies have shown that these two factors are indicators to those susceptible to energy poverty [3]. Detached houses are more vulnerable to energy poverty than apartments. This is also true for rented households where the houses may have lower energy efficiency [3].

Next we applied the random forest technique. From the hyperparameters mentioned in the methodology for the random forest model, the oob_score was set to true and the n_estimators were determined for a warm_start. A plot of various scores for increments of 50 trees showed that the optimal n_estimators is 350. This can be seen in Fig. 7, where for this value, there is a peak in the score before it plateaus. The accuracy using this model after the hyperparameters were tuned and defined increased to 87.5%.

SMOTE was also used in the random forest model using the same dataset used without SMOTE. It was determined that the accuracy score further increased to 88.2%. The classification report is given in Table 4.

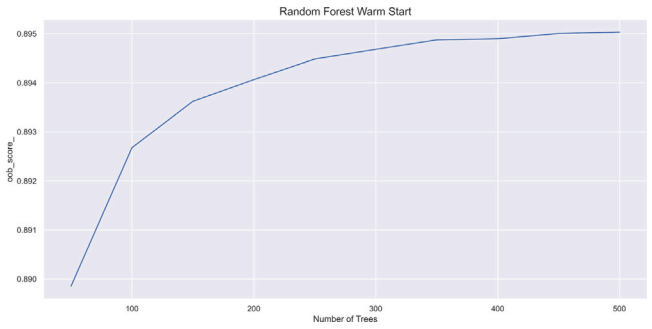

Fig. 7. Scores for various n_estimators using Random forest warm start model.

Table 4
Metric scores for each of the classes using the SMOTE in the random forest model with data, df_MO_A.

Class	Precision	Recall	F1 score
No risk	0.96	0.90	0.93
Expenditure risk	0.75	0.88	0.81
Income risk	0.72	0.71	0.71
Double risk	0.72	0.72	0.72
Accuracy			0.88
Macro average	0.79	0.80	0.79
Weighted average	0.89	0.88	0.89

Fig. 8. Accuracy scores of the number of estimators or trees for the Extreme Gradient Boosting model.

For the larger models of XGB and SVM, Repeated Stratified KFold Cross Validation were utilized. This is where stratified sampling is used instead of random sampling.

For the XGB model, the hyperparameters were tuned and the results were plotted using box and whisker in Figs. 8 through 11. Thereafter, for each of these hyperparameters, tables showing the values of the mean accuracy scores are shown (see Tables 5–9). These plots show from the top: the max, upper quartile, median (horizontal line in box), mean (triangular value), lower quartile, min and outliers values, respectively. The number of estimators, shown in Fig. 8, was increased from the default of 100 to 1000 as the accuracy score plateaus from that value. Similarly for the tree depth, there is an increasing trend of model performance with this hyperparameter as seen in Fig. 9. The tree depth was defined to be 10, an increase from the default of 6. The largest accuracy score was given from a learning rate of 1.0. However, as seen later, this parameter value was too large and gave poor results. From Fig. 12, the subsample was determined to be 0.6, after this value, the mean accuracy score plateaus. The max of all the mean values of the number of features parameter was chosen to be the optimal value as the accuracy score kept increasing (see Fig. 11). This indicates that

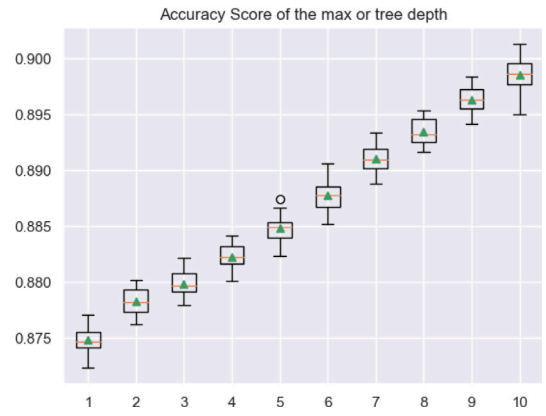

Fig. 9. Accuracy scores of the max or trees depth for the Extreme Gradient Boosting model.

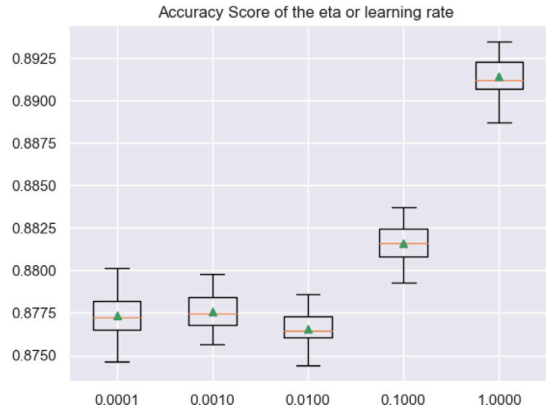

Fig. 10. Accuracy scores of eta or learning rate for the Extreme Gradient Boosting model.

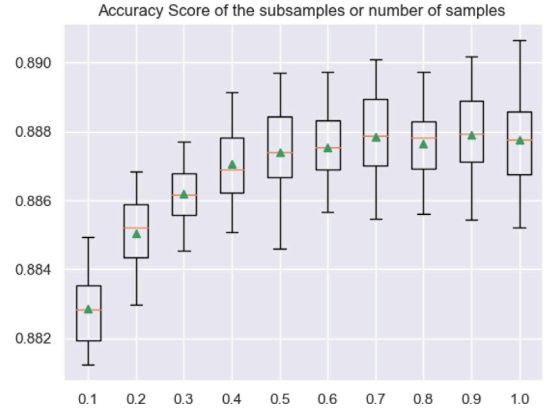

Fig. 11. Accuracy scores of the colsample_bytree or number of features for the Extreme Gradient Boosting model.

all eight features were utilized in these analyses. The model was run with these hyperparameters, however as alluded before, the accuracy score was poor compared to the random forest, it was determined to be 0.644 or 64.4%. However, after experimentation with the learning rate, or eta, the largest accuracy thus far was given with this parameter value of 0.1. The accuracy increased to 90.5%.

The final model we tuned is the Support Vector Machine (SVM). The radial basis function was chosen after tuning, as the kernel function.

Table 5
Mean accuracy scores for the number of estimators or trees.

Number of trees	Mean accuracy scores
10	0.879
50	0.883
100	0.888
500	0.898
1000	0.901
5000	0.901

Table 6
Mean accuracy scores for the max or tree depth.

Tree depth	Mean accuracy scores
1	0.875
2	0.878
3	0.880
4	0.882
5	0.885
6	0.888
7	0.891
8	0.893
9	0.896
10	0.899

Table 7
Mean accuracy scores for the eta or learning rate.

Learning rate	Mean accuracy scores
0.0001	0.877
0.0010	0.878
0.0100	0.877
0.1000	0.882
1.0000	0.891

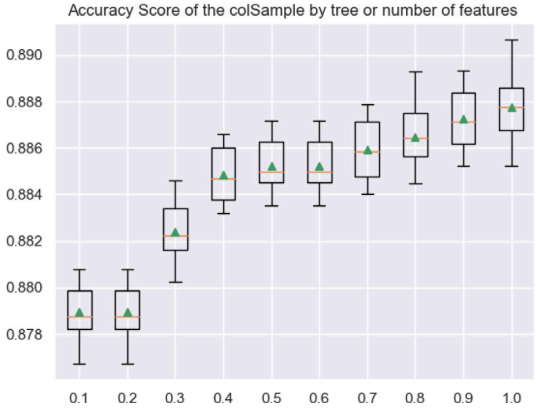

Fig. 12. Accuracy scores of the subsamples or number of samples for the Extreme Gradient Boosting model.

Table 8
Mean accuracy scores for the subsamples or number of samples.

Subsamples	Mean accuracy scores
0.1	0.883
0.2	0.885
0.3	0.886
0.4	0.887
0.5	0.887
0.6	0.888
0.7	0.888
0.8	0.888
0.9	0.888
1.0	0.888

Table 9
Mean accuracy scores for the colSample by tree or number of features.

Number of features	Mean accuracy scores
0.1	0.879
0.2	0.879
0.3	0.882
0.4	0.885
0.5	0.885
0.6	0.885
0.7	0.886
0.8	0.886
0.9	0.887
1.0	0.888

Table 10
Accuracy and F1 Scores for the models for Grid 1.

Model	Accuracy (%)	F1 Score (%)
Decision tree	87.0	87.0
Random forest	88.2	89.0
Extreme gradient boosting	90.5	90.6
Support vector machine	82.5	85.4

Table 11
Grid 2 metric scores for each of the classes using the SMOTE in the decision tree model with data, df_MO_A.

Class	Precision	Recall	F1 score
No risk	0.98	0.95	0.96
Expenditure risk	0.44	0.67	0.53
Income risk	0.43	0.53	0.47
Double risk	0.91	0.87	0.88
Accuracy			0.92
Macro average	0.69	0.75	0.71
Weighted average	0.94	0.92	0.93

This was possibly selected because of the non-linearity of the data. The associated *gamma* parameter, which determines the curvature in the decision boundary, was determined to be 10. The higher this parameter the larger the curvature. The C parameter, which controls the error in the SVM model, was also determined to be 10. The lower the C value, the lower the error, however a lower C value as model input does not imply a better model. The accuracy of this model evaluation was determined to be 82.5%.

The Table 10 below summarizes the average accuracy and F1 scores for each of the models. These results are for the models using SMOTE and Repeated Stratified KFold. For the Repeated Stratified KFold, the average of all the repeated splits were taken, in our case to be 30. From these results, the Extreme gradient boosting does the best followed by random forest, decision tree and SVM models. This is the reverse for the features and data set of [3] in which the random forest had a F1 score of 0.9794 compared to the extreme gradient boosting score of 0.9698. The SVM however, with regards to performance evaluation, does worst compared to these other three models for both studies.

Also, grid 2 was constructed similarly to grid 1 but the energy expenditure percentage was given as the ratio of energy expenditure to the corresponding household income. This grid construction utilizes a more economically intuitive definition. The accuracy and F1 scores for the tuned decision tree and random forest models without SMOTE were determined to be 95.0%, 93.6% and 95.1%, 94.8% respectively. For these models, using SMOTE, the classification reports are tabulated in Tables 11 and 12.

The accuracy and the F1 scores for all of the models were determined, as well, for this construction using SMOTE and Repeated Stratified KFold. This is shown in Table 13. The data set used in this analysis consisted of all the features except expenditure as this combination of variables also gave the highest accuracy using the

Table 12
Grid 2 metric scores for each of the classes using the SMOTE in the random forest model with data, df_MO_A.

Class	Precision	Recall	F1 score
No risk	0.98	0.95	0.97
Expenditure risk	0.47	0.72	0.57
Income risk	0.51	0.53	0.52
Double risk	0.91	0.90	0.91
Accuracy			0.93
Macro average	0.72	0.78	0.74
Weighted average	0.95	0.93	0.94

Table 13
Accuracy and F1 Scores for the models using Grid 2.

Model	Accuracy (%)	F1 Score (%)
Decision tree	92.4	93.0
Random forest	93.1	94.0
Extreme gradient boosting	95.9	95.7
Support vector machine	94.7	93.9

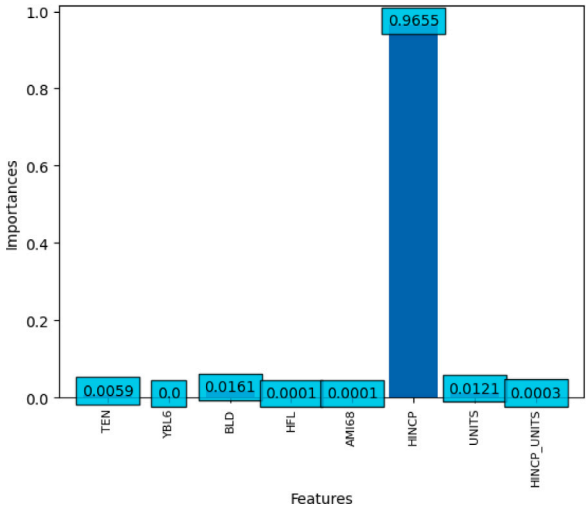

Fig. 13. Important Features determined from the decision tree model for grid2.

Table 14
Statistics for the variables used in the county wide data set.

Statistics	HINCP	FUEL_EXPENDITURE
count	1.043520e+06	1.043520e+06
mean	3.738124e+04	2.074139e+03
std	2.139671e+04	7.928294e+02
min	−7.169553e+03	1.208800e−08
max	8.814238e+05	1.787769e+04

decision tree model. The most important features given by this decision tree model is depicted in Fig. 13. We note that the most important feature for this grid is also income.

The gradient boosting is our best performing model for both grids. This is the motivation for its utilization in the subsequent analysis using the county level data set. It should be noted though that future works will include the computations of each model run for the six climatological regions described below in order to establish the best performing algorithm for each data subset. The statistics for this county level data is shown in Table 14 . As above, this model's hyperparameters were tuned to obtain optimal performance.

The analyses were rerun for the NOAA six climate divisions of Missouri as shown in [10]. These regions are labeled as $1, 2, \ldots, 6$ in Fig. 14. The Energy Expenditure-Income Plots for the various regions for grid 1 are depicted in Fig. 15.

Fig. 14. Six climate regions of missouri defined by the national oceanic and atmospheric administration (NOAA).

The percentage of contribution of each class for the various regions to the total count of each sector for grid 1 is shown in Table 15. From this table, it can be noted that the county level data used in aggregation to produce the climatological regional data sets has more entries classified as expenditure poor compared to the Missouri data set which saw the highest classification of no risk. The county-level data set has more data points than the state wide-data set. There were 1,043,520 households in this data set.

The F1 scores for the prediction of the classes of each of the test sets of the climatological regions ranged from 95.5–98.6%. This can be seen in the Map of Fig. 16. There is an increase from the 90.6% score when using the state level data set. This can be attributed to the largest class distribution for the state-level data being no risk with approximately 70% of the total count. In contrast, the largest class distribution for the county-level data set was an estimated 88% for the expenditure risk category. It can be noted that high model predictability came from region 1, which had the largest number of data points and represented the area in Missouri where the expenditure risk was the second largest in Table 15. The regions 2, 3 and 4, all color coded in the lowest part of the scale, had comparatively the most variability of class distribution or less class imbalance as these were the areas that had the largest no risk percentages.

To determine the most at-risk of the population, the county data is thus used for the analysis using the redefined grid. Using grid 2, the energy poverty classes are computed for the six climate regions. As mentioned previously, grid 2 was constructed as in grid 1 except that the energy expenditure percentage was derived from the ratio of energy expenditure to the corresponding household income. This is in keeping with the official definition of energy poverty referenced in [11].

The energy poverty categories from this redefined grid can be seen in Fig. 17. Two assumptions were made to clean the data of spurious entries. All negative values of income and expenditure were removed along with all rows where fuel expenditure exceeded income. As such, the revised count for each of the regions are shown in the Table 16.

It is evident from Table 16 that there is a shift where the dominating class in the grid comes from the no risk category. There is a lower percentage from the income risk compared to grid 1 but there is a larger spread of values among the classes. For example, for grid 1, the range of values for the percentage double risk is 3.4 to 4.4 whilst for grid 2 it is 4.4 to 6.6.

It is noted that the NOAA regions do not map perfectly onto the Missouri electric service areas, but for some zones, they come relatively close. In particular, region 6 is almost a perfect overlay for Ameren Missouri's boothheel footprint. For this grid, from Table 16, this region

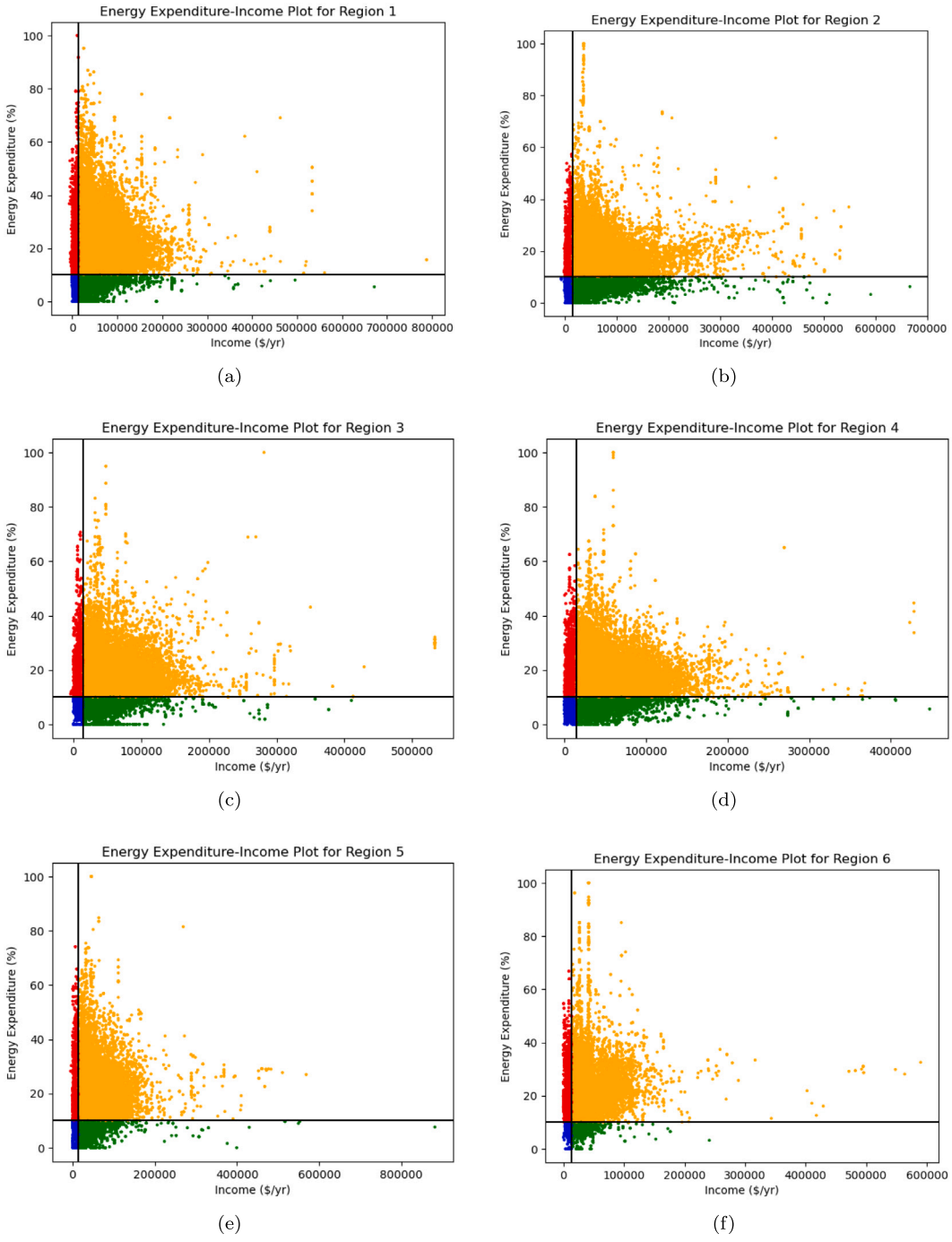

Fig. 15. Energy Poverty using grid 1 definition for the 6 climate divisions of MO where green, orange, blue and red quadrants represents no risk, expenditure risk, income risk and double risk categories respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 15
Percentages of the classes for the various regions in Missouri for grid 1 definition.

Regions	% No risk	% Expenditure risk	% Income risk	% Double risk	Count
1	5.1	89.2	1.3	4.4	278,925
2	9.0	85.8	1.8	3.4	261,490
3	6.8	87.6	1.6	4.0	138,329
4	7.6	86.7	2.0	3.8	187,195
5	4.5	89.6	1.4	4.4	120,281
6	6.4	88.4	1.4	3.8	57,300

has the highest expenditure and double risk percentages in the state. Region 1 is mostly served by either Ameren (in a portion of 10 counties) or Evergy (in all but three counties). However, these regions represent a small share of households served by those two utility companies and

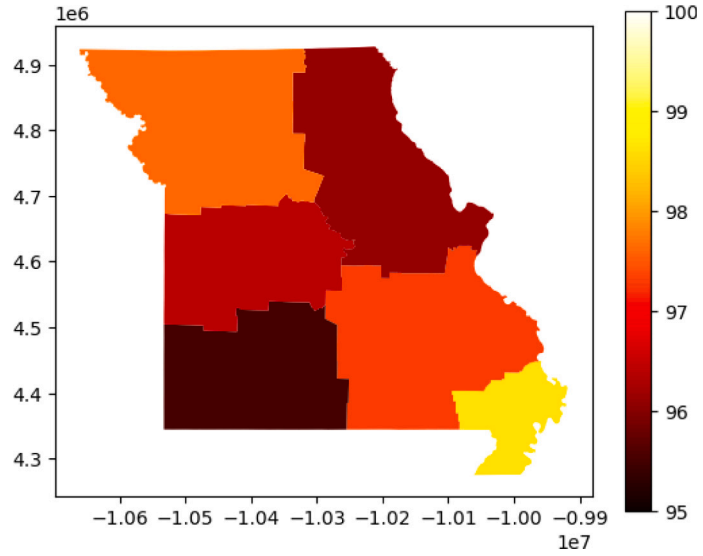

Fig. 16. Map of F1 Scores of test set prediction for the climatological regions in Missouri using Grid 1. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 16
Percentages of the classes for the various regions in Missouri using grid 2 definition.

Regions	% No risk	% Expenditure risk	% Income risk	% Double risk	Count
1	88.8	5.5	0.83	4.6	278,194
2	88.5	6.3	0.53	4.4	260,842
3	88.3	6.1	0.67	4.6	137,909
4	87.8	6.5	0.84	4.7	186,726
5	88.7	5.4	0.86	4.8	120,016
6	83.6	8.6	0.84	6.6	57,111

are a relatively small portion of the geographic composition of the state. Both companies also cover the urban areas in zones 2 and 3. The assumption is that these companies' energy poverty problems may be the most severe in the urban areas. However, as mentioned previously, the results suggest otherwise.

The range of the F1 scores for the predicted climatological regions for grid 2 was close to that of grid 1. Grid 2 had a range of 96.0 to 97.4%. This is a bit lower than grid 1 since grid 1 had less variability in terms of the double risk classes. That is there was more of a class imbalance. There were also similarities in the predictability of regions. For example, regions 1 and 6 had the largest F1 scores for both grids (see Fig. 18).

5. Future work and conclusion

From the analysis of energy poverty in Missouri for grid 1, the data set included all of the variables except expenditure as it incurred the least errors. From the decision trees, the most important feature is the average annual household income. After tuning the hyperparameters, the extreme gradient boosting was the best-performing model. Thus, this method was used for the analysis using the county-level data set. Using this algorithm, the F1 scores increased from 90.6% of to a range of 95.5–98.6%. From the NOAA climatological regional runs, the F1 score was the highest for region 1, which coincided with one of the highest expenditure risk values. For grid 2, where the energy expenditure percentage was derived from the ratio of energy expenditure to the corresponding household income, the highest class distribution shifted from expenditure risk to no risk. There was higher class variability in terms of the double risk category for grid 2, thus less of a class imbalance. The F1 scores for grid 2 was a bit lower than grid 1. There was also similar predictability in the regions for both grids; regions 1 and 6 had the largest F1 scores.

Limitations of this study include data quality of some features of the county and state-wide datasets. There was a pattern of repeating values of rows for the primary space heating fuel type. Another limitation is the lack of data concerning those households under a certain income threshold, who receive subsidies or waivers of their energy bill. This could have been factored in this study.

Additional analysis and further investigation will be done using alternative selection of the features. That is, all of the features including percentage energy expenditure excluding the household annual income. The motivation comes from this dataset having a relatively high percentage accuracy for both descriptions of the grids using the decision tree model run. The accuracies were 96.0% and 97.0% for grid 1 and grid 2 respectively. In other analyses, we also intend to investigate similar techniques in the methodology for the dataset, df_MO_B, which excludes both income and fuel expenditure features as this also poses an interesting problem. Another point of future research is the determination of how energy poverty and particularly their counts for each classification, changes with more recent minimum wages and thus various income thresholds.

Future work also includes using environmental and geographical remote sensing data sets as features in the analysis of energy poverty as done in [9,12]. Such variables include precipitation, temperature, soil moisture, vegetation (from the Normalized Difference Vegetation Index ($NDVI$), fine particulate matter ($PM_{2.5}$), travel time to nearest city and night time lights [9]. Of all the data, the most important

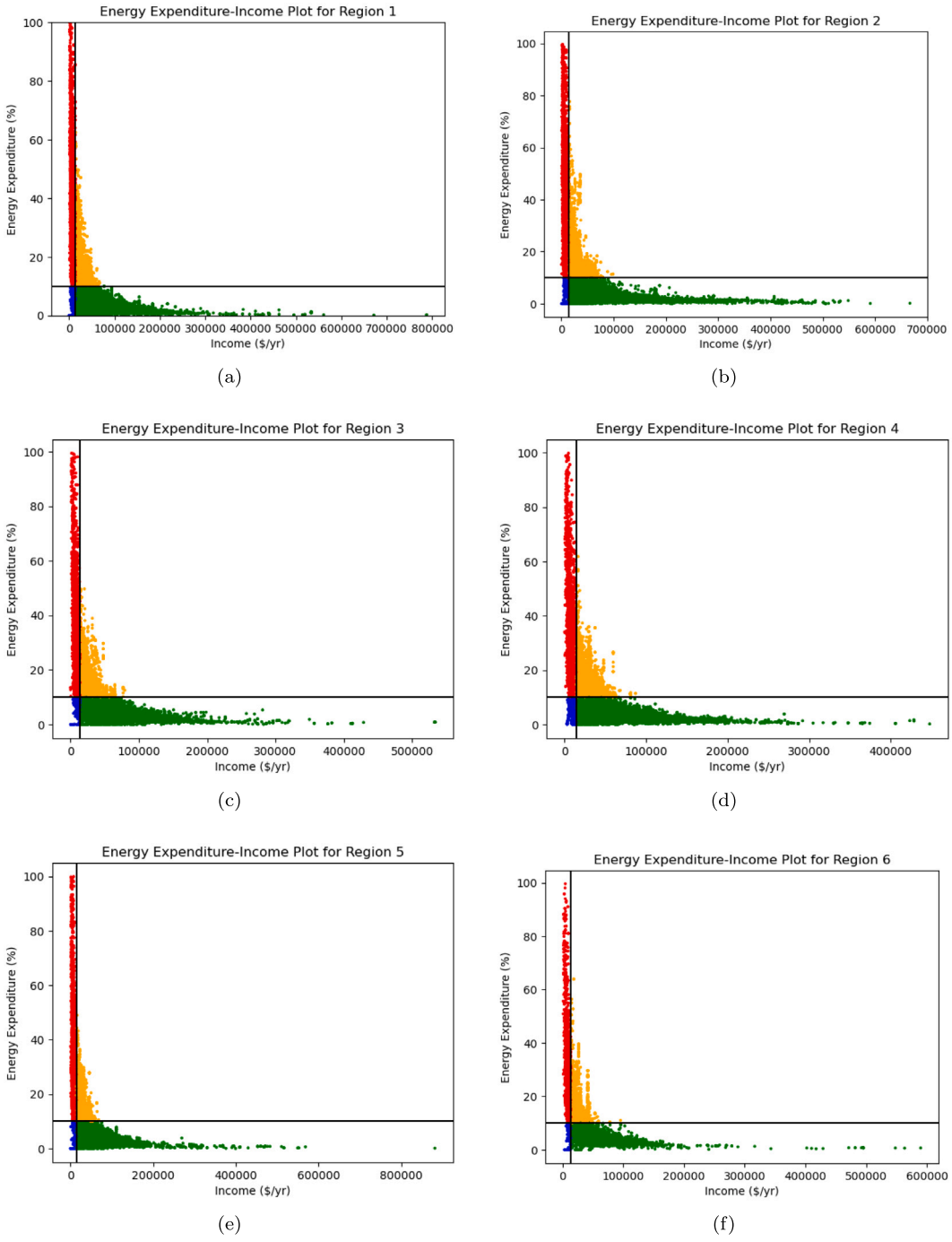

Fig. 17. Energy Poverty using grid 2 definition for the 6 climate divisions of MO where green, orange, blue and red quadrants represents no risk, expenditure risk, income risk and double risk categories respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

contributions given by [9] to model predictability came from precipitation and $PM_{2.5}$. Those households receiving less precipitation are experiencing more energy poverty, as there is a positive relationship between precipitation and per capita GDP in [9]'s study for India. Such a study could be incorporated for Missouri, which is also heavily dependent on agricultural income.

The work conducted in this research paper can be utilized by policy implementers to formulate social programs based on data driven modeling to ameliorate energy poverty in Missouri. This tool may be informative for legislative initiatives. In broadly addressing energy poverty, state legislature can utilize this model to prioritize other types of state low-income assistance programs especially since household income is the primary driver of energy poverty using grid 2. It should be noted that grid 2 should be referred to in policy discussions as it is derived from a more economically intuitive definition. Also, affordability is of concern at the utility regulator and investor-owned utility level. Being able to predict energy poverty may aid in targeting energy efficiency and low-income support programs for those companies.

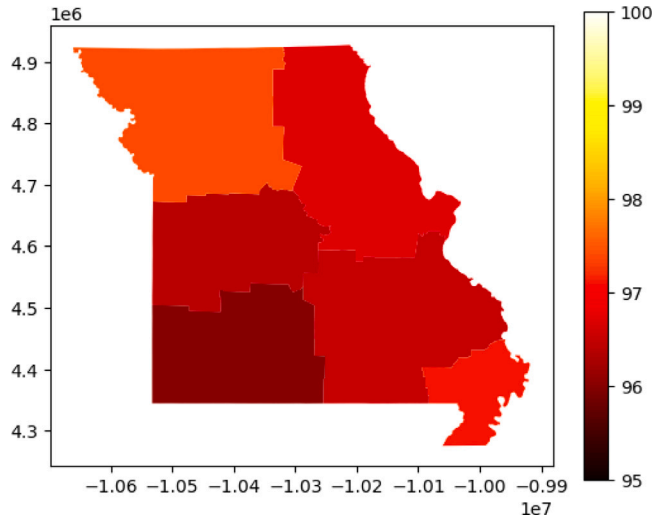

Fig. 18. Map of F1 Scores of test set prediction for the climatological regions in Missouri using Grid 2. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

CRediT authorship contribution statement

Sarah Balkissoon: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Neil Fox:** Visualization, Validation, Supervision, Project administration, Methodology, Investigation, Formal analysis, Conceptualization. **Anthony Lupo:** Visualization, Validation, Supervision, Project administration, Methodology, Investigation, Formal analysis, Conceptualization. **Sue Ellen Haupt:** Writing – review & editing, Visualization, Validation, Supervision, Project administration, Methodology, Investigation, Formal analysis, Conceptualization. **Stephen G. Penny:** Visualization, Validation, Supervision, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization. **Steve J. Miller:** Visualization, Validation, Supervision, Project administration, Methodology, Investigation, Formal analysis, Conceptualization. **Margaret Beetstra:** Writing – review & editing, Visualization, Validation, Supervision, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Michael Sykuta:** Visualization, Validation, Supervision, Project administration, Methodology, Investigation, Formal analysis, Conceptualization. **Adrienne Ohler:** Visualization, Validation, Supervision, Project administration, Methodology, Investigation, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Stephen G. Penny reports financial support was provided by Office of Naval Research (ONR) grants Grant Numbers N00014-19-1-2522 and N00014-20-1-2580. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

Sue Ellen Haupt is with the NSF National Center for Atmospheric Research, which is a major facility sponsored by the U.S. National Science Foundation under Cooperative Agreement No. 1852977.

Data availability

Data is referenced in the paper.

References

[1] López-Vargas A, Ledezma-Espino A, Sanchis-de Miguel A. Methods, data sources and applications of the Artificial Intelligence in the Energy Poverty context: A review. Energy Build 2022;268:112233.
[2] van Hove W, Dalla Longa F, van der Zwaan B. Identifying predictors for energy poverty in europe using machine learning. Energy Build 2022;264:112064.
[3] Hong Z, Park IK. Comparative analysis of energy poverty prediction models using machine learning algorithms'. J Korea Plan Assoc Vol 2021;56(5).
[4] Dalla Longa F, Sweerts B, van der Zwaan B. Exploring the complex origins of energy poverty in The Netherlands with machine learning. Energy Policy 2021;156:112373.
[5] Abbas K, Butt KM, Xu D, Ali M, Baz K, Kharl SH, et al. Measurements and determinants of extreme multidimensional energy poverty using machine learning. Energy 2022;251:123977.
[6] Ma O. Low-income energy affordability data - LEAD tool - 2018 update. U.S. Department of Energy, Office of Energy Efficiency and Renewable Energy; 2020.
[7] Glynn K, Wade C. Hands-on gradient boosting with XGBoost and scikit-learn. Packt Publishing; 2020.
[8] James G, Witten D, Hastie T, Tibshirani R, Taylor J. An introduction to statistical learning with applications in Python. Springer; 2023.
[9] Wang H, Maruejols L, Yu X. Predicting energy poverty with combinations of remote-sensing and socioeconomic survey data in India: evidence from machine learning. Energy Econ 2021;102:105510.
[10] Henson C, Market P, Lupo A, Guinan P. ENSO and PDO-related climate variability impacts on midwestern united states crop yields. Int J Biometeorol 2017;61:857–67.
[11] Rajić MN, Milovanović MB, Antić DS, Maksimović RM, Milosavljević PM, Pavlović DL. Analyzing energy poverty using intelligent approach. Energy Environ 2020;31(8):1448–72.
[12] Putri SR, Wijayanto AW, Sakti AD. Developing relative spatial poverty index using integrated remote sensing and geospatial big data approach: A case study of East Java, Indonesia. ISPRS Int J Geo-Inf 2022;11(5):275.