Unemployment Rates Predicted from Tree Equity Data

Adina Kugler Hrishi Shah Shiva Yeshlur Aparna Kumar Jinze Wu

Columbia University
Applied Machine Learning - Group 11

1 Introduction

Our objective is to investigate the feasibility of utilizing socio-demographic and environmental attributes to predict unemployment. Given the annual rise in societal unemployment rates, this analysis holds the potential to contribute significantly to social progress by comprehending and ameliorating these rates. Identifying the root causes of unemployment and predicting fluctuations in unemployment rates can inform policies and interventions, ultimately alleviating homelessness and enhancing overall social well-being.

2 Methodology

We explored various machine learning models during our analysis. The considered models encompassed ElasticNet (Linear Regression), XGboost, random forest, decision tree, and deep learning. Through experimentation, we aimed to identify the most effective model for our specific requirements.

2.1 ElasticNet

ElasticNet is a regularization technique that combines both L1 (Lasso) and L2 (Ridge) penalties, making it effective for feature selection and handling multicollinearity in the dataset. It strikes a balance between Lasso and Ridge regression, providing the advantages of both.This is valuable for identifying the most influential socio-demographic and environmental attributes in predicting unemployment.

2.2 Decision Tree

Decision trees are simple yet powerful models that can capture non-linear relationships in the data and are easy to interpret. Decision trees inherently perform feature selection by selecting the most informative features for splitting nodes. This transparency can be valuable in understanding the decision-making process and the hierarchy of features contributing to unemployment predictions.

2.3 XGBoost

XGBoost is an ensemble learning algorithm known for its high performance and ability to handle complex relationships in data. It is particularly suitable for regression tasks and is robust against overfitting. XGBoost employs methodical theory by combining weak learners (individual decision trees) to form a strong predictive model. Its boosting approach sequentially corrects errors, emphasizing instances that are more challenging to predict. This can reveal intricate patterns in the data, offering a nuanced understanding of the factors influencing unemployment. For XGBoost, randomized search was used for hyperparameter tuning, and the parameters used were min_child_weigh, max_depth, gamma, and eta. In addition, two versions were tested.

2.4 Random Forest

Random Forest is an ensemble learning method that builds multiple decision trees and merges their predictions. It is robust, handles non-linearity well, and provides a measure of feature importance. The feature selection mechanism in Random Forest is inherent in its construction. By considering different subsets of features and averaging their predictions, Random Forest automatically assesses feature importance. This allows you to identify key socio-demographic and environmental attributes contributing to unemployment.

2.5 Deep Learning Model

Deep learning models, such as neural networks, are known for their ability to capture complex patterns and relationships in data. Deep learning models can automatically learn hierarchical representations of features. The depth of the network allows it to discover intricate patterns in the data, potentially revealing nuanced relationships between socio-demographic and environmental attributes and unemployment.

3 Experiment

3.1 Data

3.1.1 Data source

In our analysis, we employed the Tree Equity Score data set, a tool developed by the American Forest Foundation to gauge the impact of tree development projects on low-income, marginalized communities. Each unique location is identified using a GEOID, a census identification based on locality. Our target variable is the normalized unemployment rate, and the feature variables encompass tree canopy information (land area, biome, data source, goal, percentage, gap), demographic measures (priority index, percentage of people of color, percentage of people in poverty, dependency ratio, percentage of households where no one speaks English, percentage of children, percentage of seniors), environmental measures (tree equity score, composite tree equity score of locality, home owner loan grade), and climate and health measures (health burden index, heat extremity, temperature difference).

3.1.2 Data preprocesssing

To maintain data integrity, we initially conducted an Exploratory Data Analysis (EDA), followed by cleaning steps for highly correlated data and missing data. We prioritized normalized data over raw data, ensuring consistency and accuracy in our analysis. Location-specific information was excluded to maintain generalizability. Columns with high collinearity, such as the total population of entire block groups, were removed to prevent redundancy. The dataset was then divided into development and testing sets, allocating 80% for development and 20% for testing. For categorical variables, biome and tree canopy goal, we implemented target encoding and ordinal encoding respectively, ensuring appropriate handling of these data types.

3.2 Model Training

3.2.1 ElasticNet

In our experiment with the Elastic Net regression model, we observed mixed outcomes. The model achieved a Root Mean Square Error (RMSE) of 0.16, suggesting a moderate level of prediction error. However, the R^2 value, a measure of the proportion of variance explained by the model, was 0.09. This indicates that the model explains only a small portion of the variance in the data, reflecting limited predictive power in this context.

3.2.2 Decision Tree & Random Forest

In our decision tree analysis, the initial model scored an R2 of -0.624, significantly underperforming. Feature importance analysis revealed an over-reliance on the "tes" (tree equity score) feature. Through Random and Grid Search cross-validation, focusing on "max_depth," "min_samples_leaf," and "max_features," we improved the model. The optimized model highlighted "ua_pop," "pctpovnorm,"

"tes," and "health_nor" as key features. Using these for our final model resulted in a much-improved R2 score of 0.179. Additionally, the feature-selected model was more time-efficient, training in just 0.391 seconds compared to the original model's 5.645 seconds, and also slightly outperformed the grid-searched model in R2 score by about 0.056. This indicates that feature selection notably enhanced the model's efficiency and generalizability. The Random Forest model utilized the same model parameters and achieved better performance.

3.2.3 XGBoost

The XGBoost model produced an R2 coefficient score of 0.3877 on a model without any hyperparameter tuning. However, there was a large amount of overfitting and therefore this model produced an R2 of 0.224 on the test set. To combat overfitting, randomized cross validation was used to determine the best hyperparameter values (*min_child_wieght=1, max_depth=7, gamma=0.005, eta=0.08*).

A subsequent model trained with these hyperparameters achieved R2 scores of 0.3343 and 0.2293 on the training and test sets, respectively. Further refinement using the top four features (ua_pop, pctpovnorm, tes, health_nor) reduced the R2 to 0.2546 on the training set and 0.2038 on the test set, indicating some reduction in overfitting.

3.2.4 Deep Learning Model (Regression Net)

The architecture is centered around a custom neural network class, *RegressionNet*, which incorporates residual blocks for efficient learning. Each residual block contains fully connected layers, batch normalization, and ReLU activations, with a skip connection for facilitating the training of deeper networks by preventing the vanishing gradient problem. The network's input size is dynamically set based on the training data's features. For optimization, we employ the Adam optimizer with a learning rate of 0.0005, and the model's performance is evaluated using Mean Squared Error loss. The training process includes a loop over 20 epochs, with both training and validation phases to monitor the model's learning progress and generalization capability, using loss and R-squared as metrics with 0.0230 and 0.221 respectively. This approach exemplifies a modern machine learning practice, combining advanced neural network architectures with robust training and evaluation techniques.

3.2.5 Model Evaluation

For our project, we utilize Mean Squared Error (MSE) and R-squared (R^2) as evaluation metrics, chosen for their effectiveness in regression analysis.

Mean Squared Error (MSE) is a measure of accuracy, indicating the average squared difference between the estimated values and actual values. Its sensitivity to large errors makes it suitable for our project, as it helps in identifying models that significantly deviate from observed data.

R-squared (R^2) gauges the proportion of the variance in the dependent variable explained by the independent variables. A higher R^2 value in our project signifies a model's strong explanatory power, an essential aspect for our predictive analysis.

Table 1: Model Performance Comparison

Model	MSE	R-square
ElasticNet	0.0268	0.090
Random Forest	0.0235	0.160
Decision Tree	0.0241	0.179
XGBoost	0.0227	0.229
Regression Net	0.0230	0.221

3.3 Conclusion

Our analysis showed that the Regression Net and XGBoost models performed relatively well, especially in predicting peak values. However, there is still room for improvement, possibly due to limitations in our dataset or a weak correlation between features and unemployment rates. Overall, this project was a valuable exploration of different models, providing insights into predictive modeling and highlighting the need for more robust data for improved accuracy.