Assignment 3

Group 2: Guanhao Hu, Qing Yin, Junyi Hua, GianCarlo Samayoa

2024-03-18

Introduction

In this project our group built and compared four predictive models (Logistic Regression, KNN, LASSO, and Random Forest) in order to most accurately predict whether labor action in the Labor Action Tracker (LAT) data set is a “strike” [1] or “non-strike” [0]. All models were considered on the basis of accuracy and not area under the ROC curve because of the priority placed on correct predictions over few false positives/negatives and because of the balanced predictor variable classes.

Our logistical regression provides 55.4% accuracy in predicting labor action for unseen data. With regard to the LASSO model that has the lowest root-mean-square error (RMSE) when setting the penalty level at 0, the accuracy of its prediction is 55.8%. After tuning number of trees, number of randomly sampled predictors and minimum number of data points in a node, the Random Forest model reports 54.2% accuracy in predicting with a best performing parameter–4 randomly sampled predictors, 10 decision trees and minimum 8 data points in each node.

The K-nearest neighbors (KNN) model number 5 selecting 43 neighbors was chosen as the final model to predict labor actions. The goal of this project is to predict labor actions as accurately as possible. Seeing that the KNN model was the most accurate of all the models tested, at a 58% accuracy rate, it was chosen over all other models. *(see table 1)*

Table 1: Accuracy of Each Model (% of correct predictions)

Logistic	KNN	Lasso	Random Forest
0.554	0.58	0.558	0.542

Furthermore, as mentioned earlier, this decision was made because when predicting labor actions, it's crucial to prioritize accuracy above all other evaluation metrics. To explain, the cost of failing to predict most labor actions correctly is likely higher than the cost of having more false positives or false negatives. In other words, the KNN model that selects 43 neighbors was chosen because it had the highest recorded accuracy in predicting labor actions based on the testing data set, making it the most cost-minimizing option for any real-world applications.

Data Manipulation

In cleaning and preparing the LAT dataset there were four major changes made: Data Transformation into Tidy Format, Parsing Coordinate Variable, Merging New Variables into the Data Set, and Creation of New Variables

1. Each row was manipulated so that it represented one observation instead of multiple. This was done by duplicating rows based on whether the values in the "number of locations" column was greater than one.
2. The "Latitude, Longitude" column was split into separate respective variables ("lat", "lon").
3. The "GEOID" column in the countries data set from the tigris library package was used as a key to merge the LAT data set with an imported ACS county-level dataset. In general, our group intends to examine worker's financial situations, costs to work and basic demographic features as strong indicators in predicting a possible strike, and thus, from the ACS county data set ten variables were imported and merged into the LAT dataset:

    - Median income, total household earners, public assistance for past 12 months, total population below poverty population, mean travel time to work, median gross rent, total population over 25, total population over 25 with BA degree, population over 16 years employed.

4. From the new variables merged into the LAT data set, variables for poverty rate and college degree rate were made. Poverty rate was created by dividing the total population below poverty by the total population, and the college (BA) degree rate was created by dividing the college degree population over 25 by the total population.

With regards to missing data, there were only two instances in which it was encountered. The first was one longitude data missing for action in "2023-06-30 13:47:25" from the LAT training data set during the data manipulating process. This was resolved by replacing it with the corresponding longitude value from the LAT training data set. Second, when creating models for prediction, all predictor variable missing values were replaced with their corresponding means and then normalized.

Model Assessment

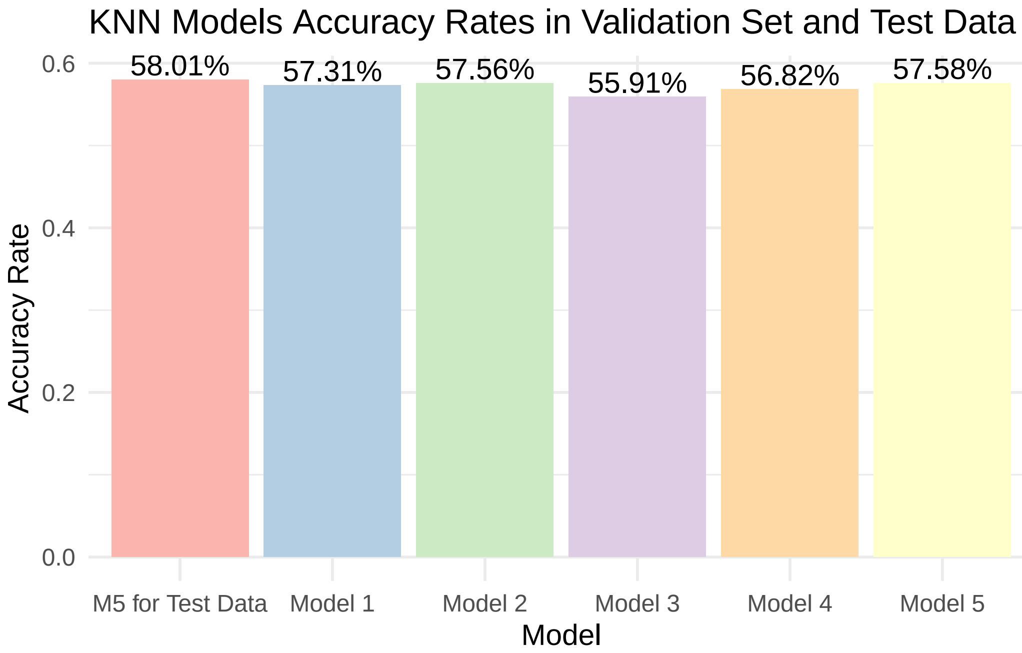
This part will show the detailed steps of how the model has been trained and how well the final model performs in the validation sets and test data.

First, it is important to note that not all of the variables present in the "new" LAT data set were used, only the following variables were used: median income, household earners, mean travel time to work, median rent, public assistance for past 12 months, total employed population, poverty rate, and college degree rate. All other variables were excluded due to being used to create the two new variables college degree rate and poverty rate, and if included could bias the results.

The prediction model had the predicted variable *strike_f* (categorical variable for strike or non-strike as [1] or [0]) on the left hand side and the aforementioned eight variables on the right hand side. All predictor variables are normalized. Then, five different KNN models are made using the following values for k (nearest neighbors): square root of $n = 59$, square root of $n/2 = 29$, square root of $2 * n = 119$, halfway between the the square root of n and square root of $2 * n = 89$, and halfway between square root of n and square root of $n/2 = 43$, with n as the number of observations. Choosing such diverse k values, and points in between, for KNN model development can help identify the optimal balance between sensitivity to local data structures and generalization ability, aiming to enhance model accuracy by exploring a range of model complexities. This approach can help find a k that minimizes bias and variance, which is critical for achieving the best predictive performance.

Then, workflows were set up and 5 KNN models (with unique k values) were fitted to LAT training data. 10-fold cross validation was used to resample each of the 5 KNN models, and model number 5 ($k = 43$) had the highest mean accuracy rate.

KNN model number 5 in the validation data set performed almost equally as well as it did with the test set, having an accuracy rate of 57.9% and 58.01% respectively. The little change in accuracy from the cross validation to the test data set indicates a high probability of it being unlikely that the current model is overfit to the training data set, strengthening the validity of the models prediction capabilities. Additionally, it is important to note that KNN models do not offer any indication of which predictor variables were most important in determining the prediction accuracy.



Conclusions

One of the major weaknesses of KNN model number 5 is that every time the model is run, despite setting a seed to keep random variation consistent, the accuracy of the model changes by one or two percentage points. However, of the five models, KNN model number 5 consistently remains the most accurate even after cross-validation. Another major weakness of the model is that the optimum number of nearest neighbors has not been identified; this can be a point of improvement.

Based on this analysis it cannot be said that KNN model number 5 is the best prediction model for predicting labor actions, due to not knowing the best number(s) of nearest neighbors and its sensitivity to changes in it. Additionally, this model is likely not suited for long term use due to more information being added over time, due to its sensitivity to noisy data.

The usefulness of KNN model number 5 is somewhat limited by its variability and the lack of optimization for the number of nearest neighbors. While it may not be the most reliable for long-term predictions, it seems to have some immediate applicability for short-term predictions about whether a labor action would be a strike. This could be useful for organizations to prepare for immediate labor disputes, although the implications must be carefully considered, given the model's sensitivity to data changes.

In terms of policy implications, the model's current state suggests caution should be taken when using it as the basis for decision-making. If the model is to be used in policy decisions, it would be important to:

1. Establish clear protocols for model updates, including regular recalibration with new data.
2. Identify the optimal number of neighbors to improve prediction stability.
3. Consider combining the KNN model with other models or approaches to mitigate its weaknesses and enhance overall predictive accuracy.

Policymakers might use this model to prepare for imminent labor actions, but they should be aware of its limitations and ensure that decisions are supported by multiple data points and analyses, not solely on the model's predictions. Additionally, they should stay vigilant to the evolving nature of the data and be ready to adapt the model as needed to maintain its relevance and accuracy.