Statistical Missions: Determining a Machine Learning Model of Gospel Access

Amelia, Lilly, Grace

3/27/2023

Abstract

Using a dataset from the Joshua Project, we evaluate five different machine learning models to determine which is best at predicting a people group's adoption of Christianity. Predictions were made based of a people groups indigenous status, bible access, world region, location, population, and amount of countries the group was present in. It was determined that a k-nearest neighbors model was the most effective at optimizing the AUC metric for this dataset.

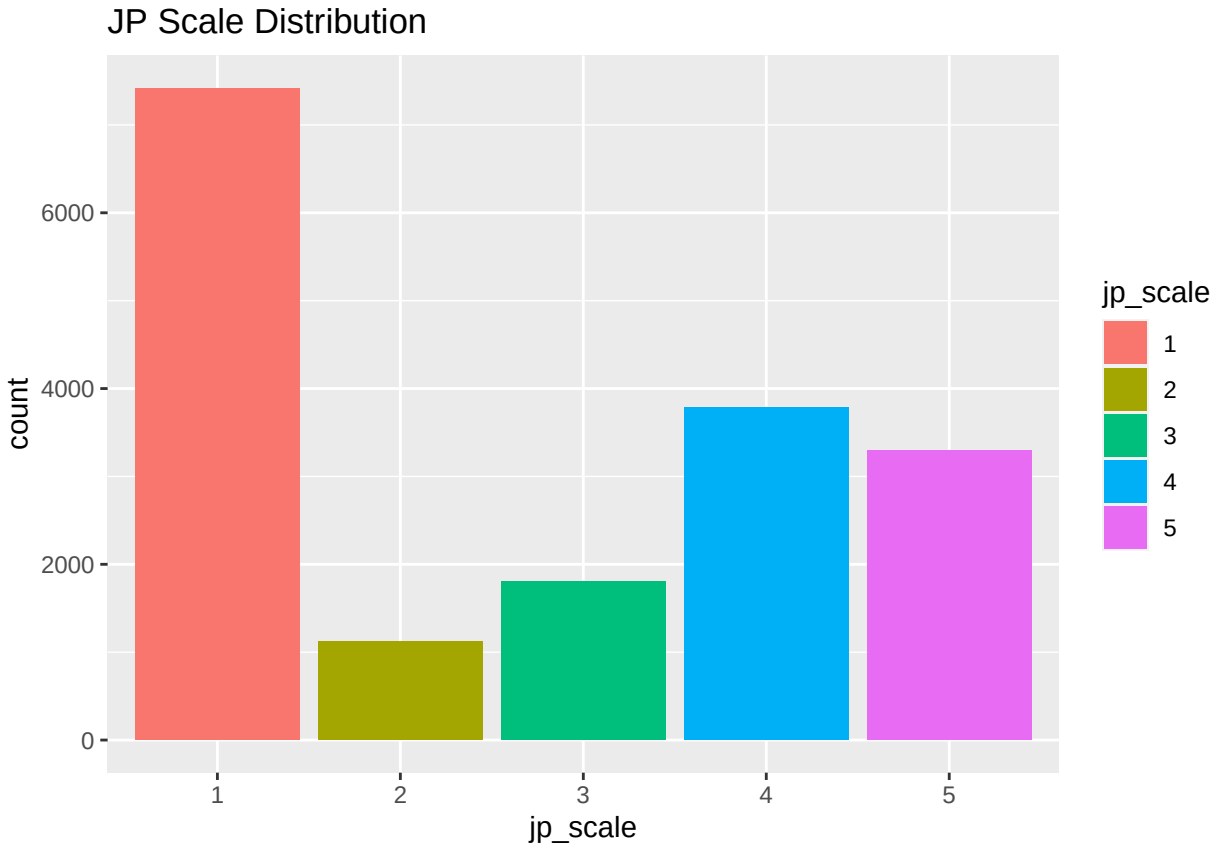
Introduction

As of 2023, 42.5% of the world population belongs to a people group of which less than 5% of the members are Christian adherentsˆ [The Joshua Project, joshuaproject.net]. These people groups are referred to as unreached peoples, and many of these groups face lack of access to scripture and fear of persecution if they choose to convert to Christianity. Joshua Project is a research initiative that seeks to collect information on the activity of gospel movements and availability of scripture of all world people groups, especially those that are categorized as unreached. Their current dataset was assembled by collecting information manually. The purpose of this paper is to determine if there is a machine-learning model that could predict the "Joshua Project status" of a people group (which ranges from 1-unreached to 5-significantly reach) based on a limited amount of information. If successful, this model could be useful to the Joshua Project by allowing them to determine which people groups should be prioritized in missional movements without spending long periods of time collecting data.

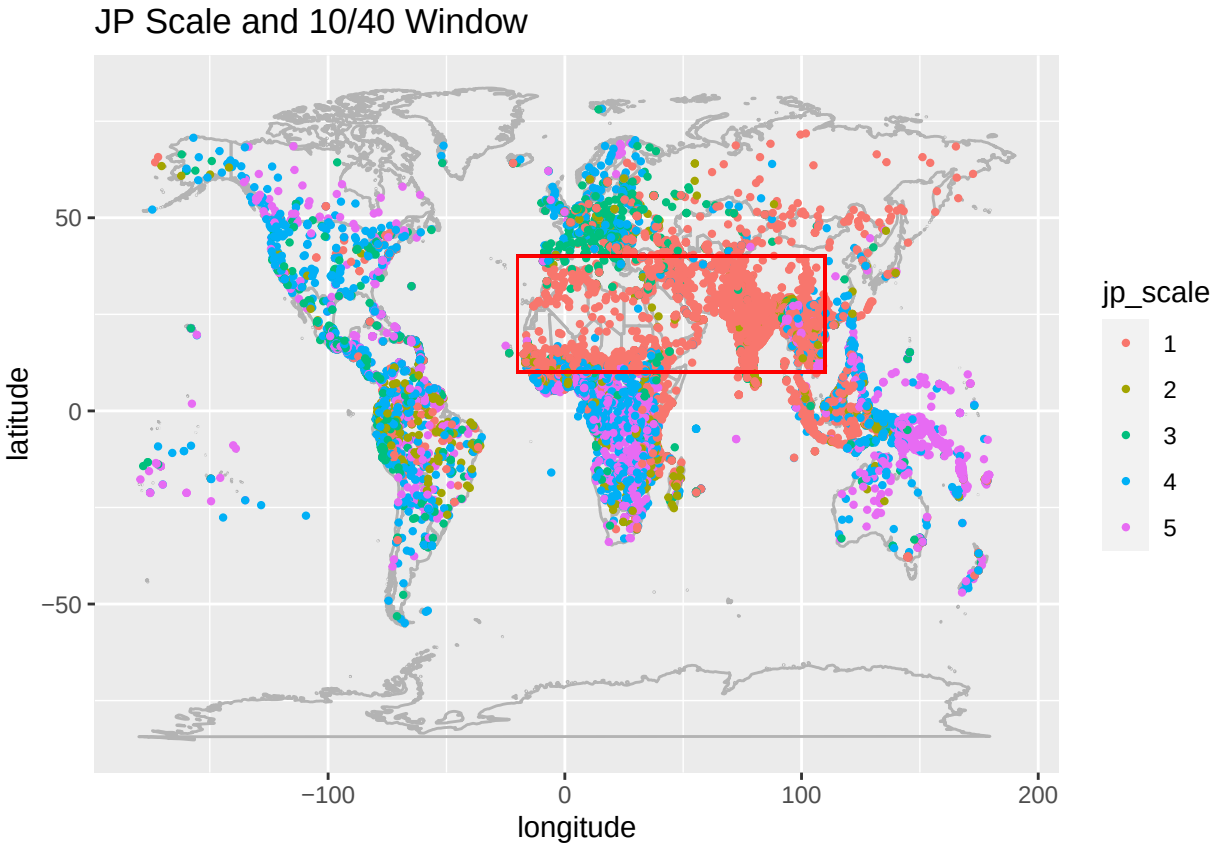
Our initial dataset includes a variety of variables including the name of the people group, their primary language, country of residence, their affinity bloc (a group based on language, culture, religion, politics, of which there are 16 worldwide), world region name, count of countries that people group is present in, bible access status (from 1-5), population, longitude, and latitude. The dataset also includes a dummy variable for whether that people group is indigenous to the country of residence, and whether that group falls within the 10-40 window (a geographic location that is believed to be home to a majority of unreached people groups). Our outcome variable of interest is the Joshua Project Scale variable.

The question we seek to answer with this paper is what statistical model is most effective at predicting the Joshua Project Scale variable. This is a multiclass classification problem, with five possible classifications. We examine decision trees, linear discriminant analysis, k-nearest neighbor, and support vector machine models. We seek to optimize the AUC score of the model, which indicates how much better our model performs than chance. Because many of the variables in the dataset are collinear, in all of our models we predict Joshua Project scale based on count of countries, population, affinity bloc, region name, bible status, and the dummy variables for indigenous and 10-40 window.

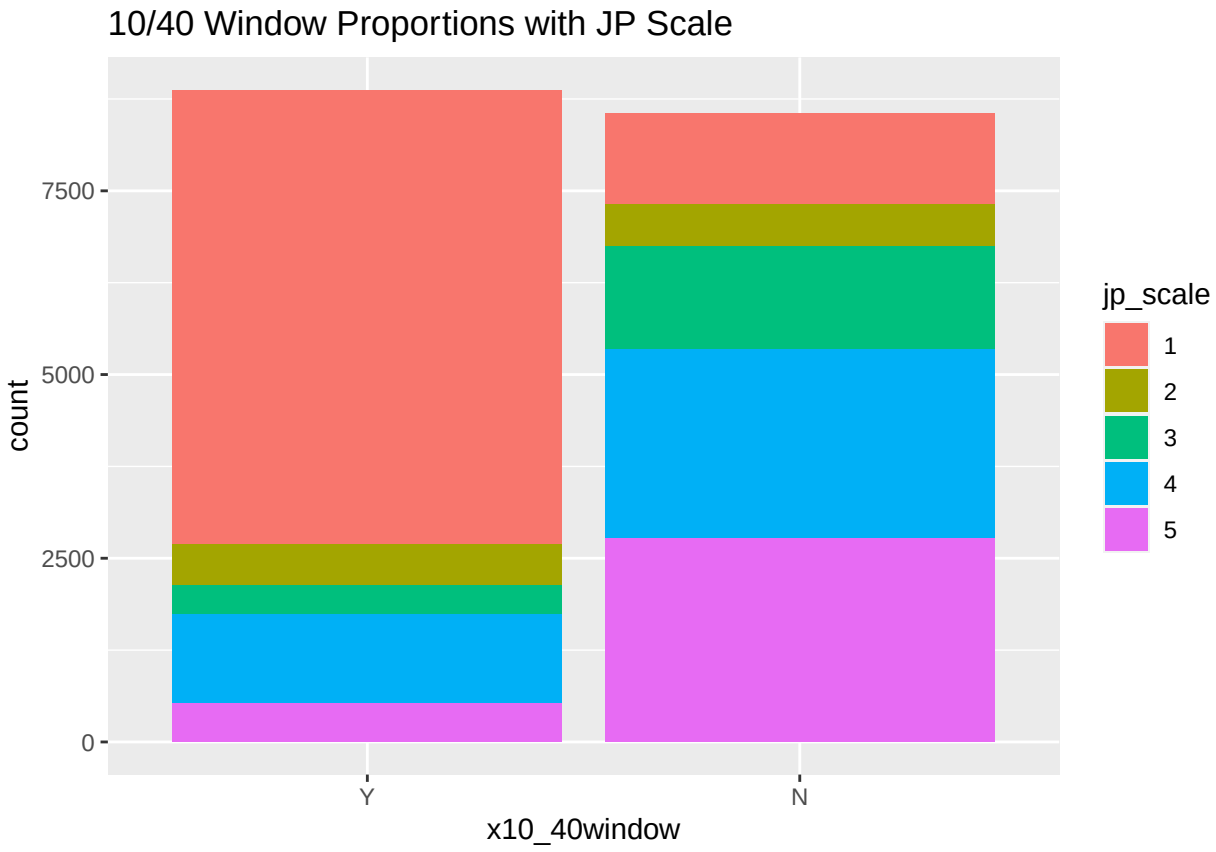
EDA Our data from the Joshua Project gave us 17,423 observations to work with. The following chart shows how this data is distributed among the outcome variable. A plurality of the data is in the 1 category, meaning they have limited access to Christianity.



The data included latitude and longitude coordinates, as well as a dummy variable indicating whether a country was in the 10/40 window. Below is a map plot showing the primary location of each observation; the color indicates the JP Scale status of the people. The red box approximately indicates the 10/40 window; many of the peoples in the 10/40 window are in the most unreached category. Intuitively, it seems that geospatial location has a notable relationship with the JP Scale status; using the 10/40 window dummy variable was much easier than using longitude and latitude in our models.



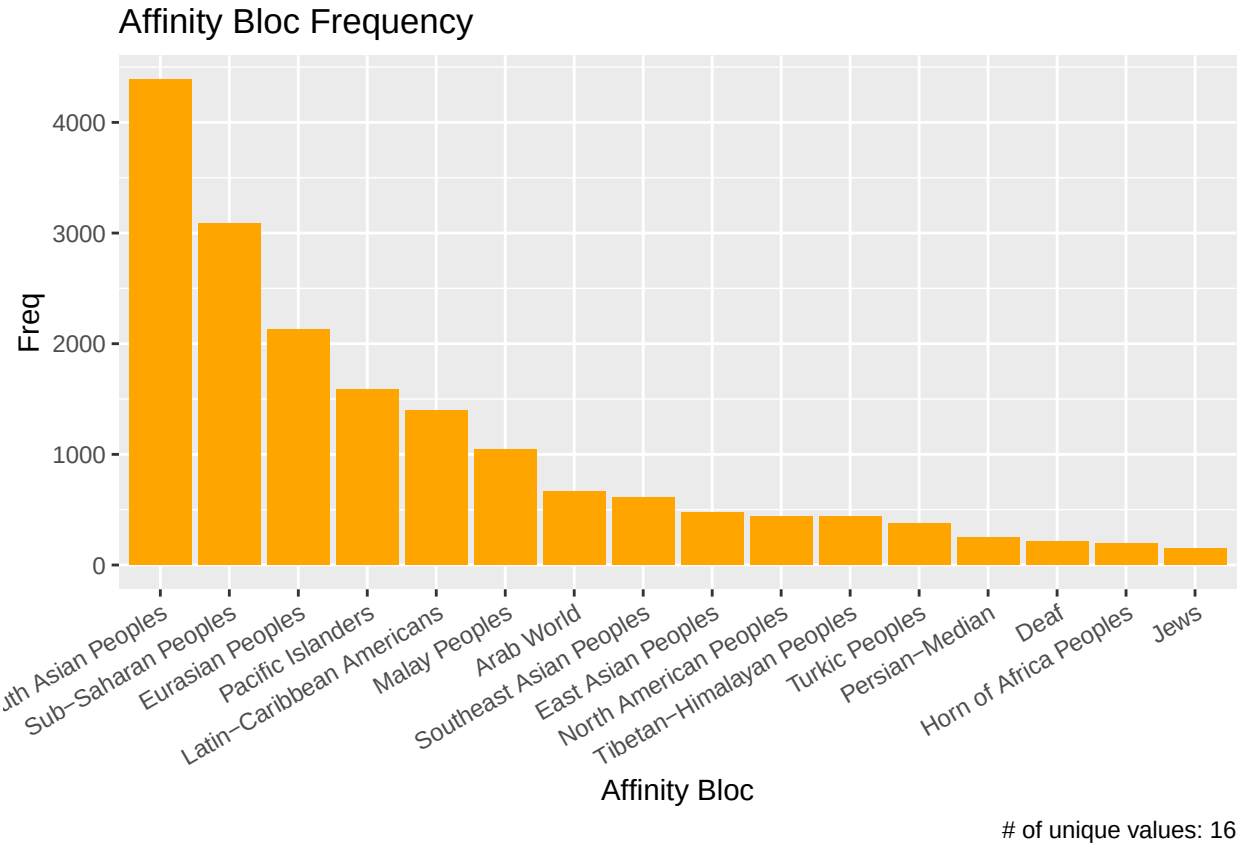
The chart below further demonstrates the relationship between primary location in the 10/40 window and JP



Scale.

Furthermore, This dataset splits the world into 12 regions, and we also use this to account for the effect of location on JP Scale.

Joshua Project also groups all people in 16 broad affinity blocs based on language, culture, religion, or politics; the graph below shows how often each affinity bloc appears in the data. This is not necessarily representative of population, but rather how many peoples are grouped together.



Methods

In order to find the best performing model, we test four models that predict jp_scale based on the selected predictors. Those four models are linear discriminant analysis, K nearest neighbors, a decision tree, and a support vector machine.

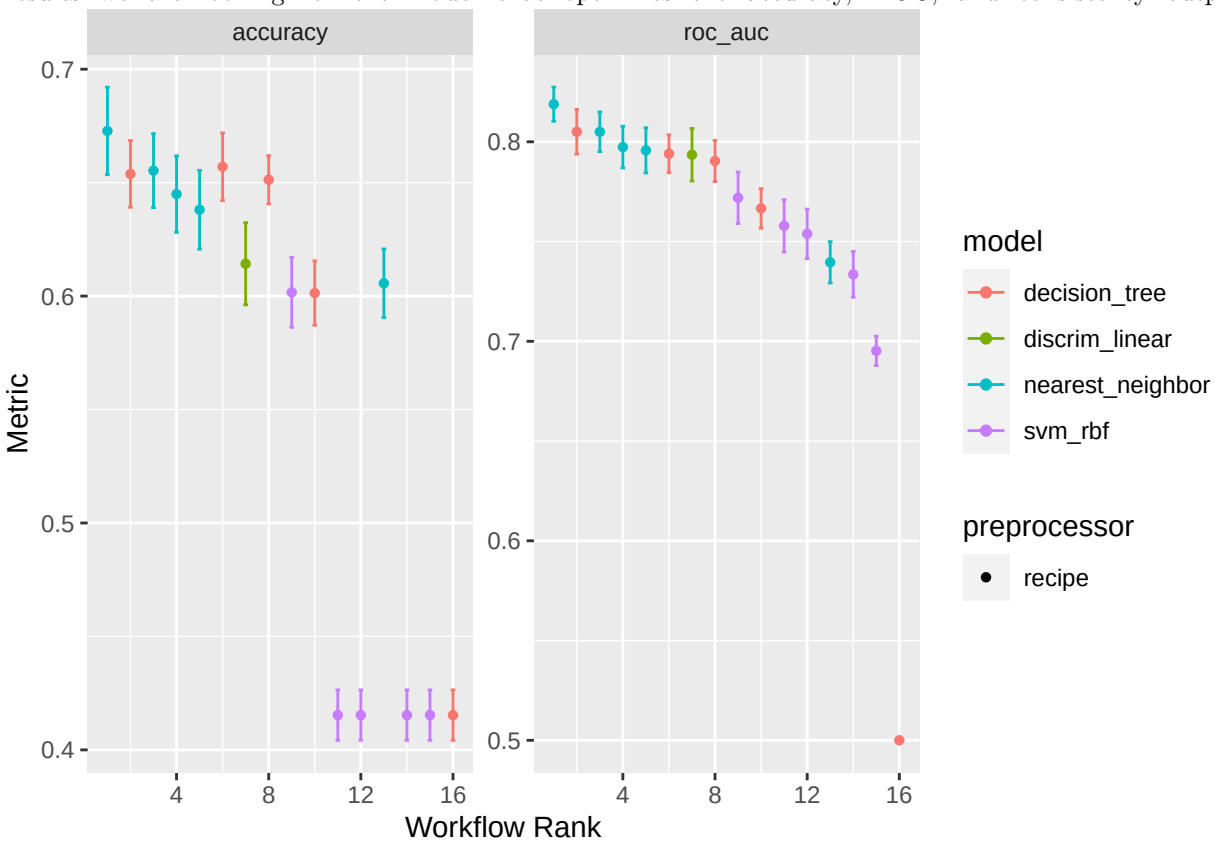
For preprocessing, we create a base recipe for all of the models to begin with. This recipe indicates that our models should predict jp_scale based on the predictors bible status, indigenous_code, count_of_countries, population, affinity block, 10_40_window, and region name. Our first step was to take the log of the population variable, since it varies from quite small values to very large ones. Next, we normalize all numeric predictors to reduce the scale of the variables. Finally, we use step_dummy on our binary character variables (indigenous_code, frontier, least_reached) to convert them into numeric binary predictors. We needed to filter out groups that did not have population values provided, as it caused calculation errors. There were only around 20 groups for which this was the case, and all were small enough that population data could not be found, so we felt comfortable excluding them from the model.

We tuned our models using a workflow map and the tune_grid function with a grid size of 5. For the tree, we tuned the tree_depth and cost_complexity, for the KNN model we tuned the number of neighbors and the dist_power, and for the support vector machine we tune the cost and rbf_sigma value. Rbf_sigma controls the level of non-linearity allowed in the model.

After tuning and fitting the model, we run the autoplot function, which allows us to visualize both the accuracy and AUC performance of the varying models. For a multiclass classification method, the AUC specification is from a 2001 paper by Hand and Till[1]. From these

[1]Hand, D.J., Till, R.J. A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems. Machine Learning 45, 171–186 (2001). https://doi.org/10.1023/A:1010920819831

results we are looking for the model that optimizes the accuracy, AUC, and consistency output.



Results:

Of the four models we tried, the k-nearest neighbors had the highest AUC values for the best tuned models. Some decision tree models had slightly higher rates of accuracy, but priority is given to the AUC value because it penalizes for false positives. Because this is a multiclass classification problem, a false positive is any value of a different class that was assigned to the class in question. Because our dataset has a large number of entries with a jp_scale value of 1, accuracy would be relatively high if the model simply predicted 1 for most values. Prioritizing AUC penalizes the model for doing this, as it is trying to reduce the amount of non-1 values that are categorized as 1. The k-nearest neighbors model with the highest AUC model was tuned to have 15 neighbors, and the distribution power is 0.45. This model had an AUC value of around 82%, and an accuracy metric of around 65%. The LDA model performed the worst, with all other models having a version performing better than it. However, we were not able to tune the model, which may have led to its poor performance. The decision tree models had the widest range of metrics, while the k nearest neighbor model performed consistently.

Conclusion:

In this paper, we used a dataset with information on world people groups to determine the status of Christian movements in that group. After testing 4 different models and tuning them to optimize their parameters, we determined that a k-nearest neighbors model with 15 neighbors was the optimal model. One of the limitations of this model was that we were not able to use very many predictor variables. Most of the variables available in the dataset (such as language spoken, certain geographic variables) were collinear with other important predictor variables, so they had to be excluded from the model. In the future, bringing in outside data on characteristics of people groups such as income levels, societal structure, or means of

employment could allow us to increase the accuracy of the model. With the addition of these variables, further research could investigate what factor (such as location, bible access, or socioeconomic conditions) or combination of factors are best at predicting the Joshua Project scale value.