Flags Data Mining Project

Team Leader: Cole Polychronis

Jasmine Boonyakiti

Kelsey Henrichsen

Merritt Ruthrauff

Westminster College

CMPT 300B Data Mining

Final Data Mining Project Report

Table of Contents

Abstract

Our dataset consisted of 194 instances, with 30 attributes, mostly made of categorical data. The goal of this project was to predict the religion of a country based upon its flag. The biggest problem we faced as a team was needing to normalize our data in order to improve the accuracy and the quality of our data and diminish overfitting. Random Forest was the best classifier we chose for our dataset because of the uneven distribution of our data and how large our data set was; correctly predicting the religion of a country with an accuracy of 72.68 percent. The accuracy being so low can be explained by the sudden increase of Muslim countries in current times versus the historical data of what countries were considered Muslim. This prediction is supported by our finding that on new data, our model only has approximately 25 percent prediction accuracy, but when we test it on an untouched portion of the data from the original dataset, our accuracy jumps to over 90 percent. In future projects, some aspects we should work on are: feature creation, normalizing data, and creating equally distributed dated; these help make data more accurate and improves data visualization.

Keywords: dataset, religion, normalization, muslim

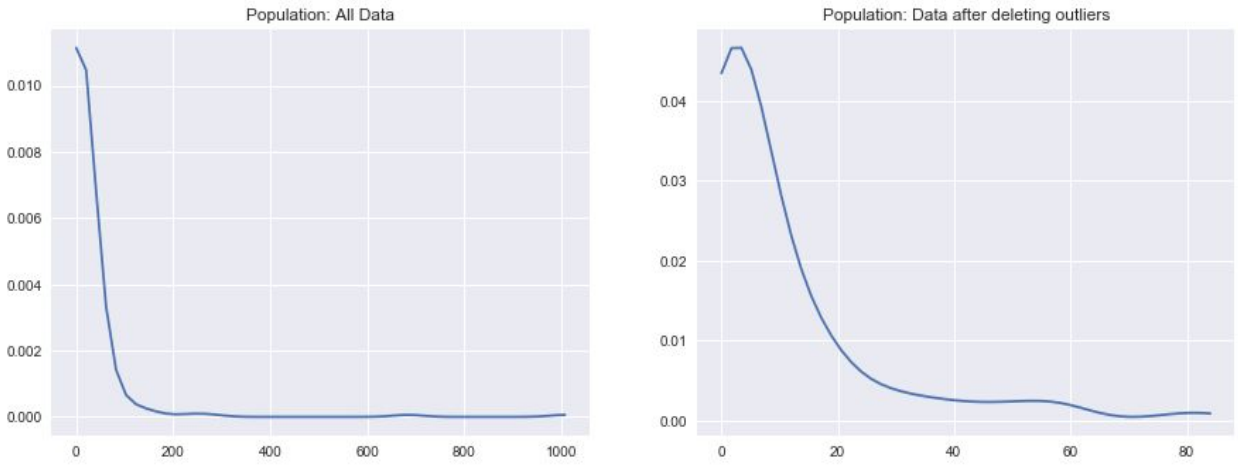
Introduction to the Dataset

For our final project, we chose to analyze a dataset from the University of California, Irvine on flags from around the country. This dataset consisted of 194 instances for different countries formed prior to 1990, where each instance had 30 attributes. We wanted to explore whether or not the dominant religion of a country had any bearing on the design of that country's flag. Specifically, we wanted to see if we could predict the religion of a country based on various attributes of its flag, such as dominant color and symbols on the flag. While there is not necessarily a big need for a model that can predict religion from attributes of a country's' flag in the real world, we felt that given the complexity of this task, it would provide us with an interesting way of exploring various data mining and analytic techniques.

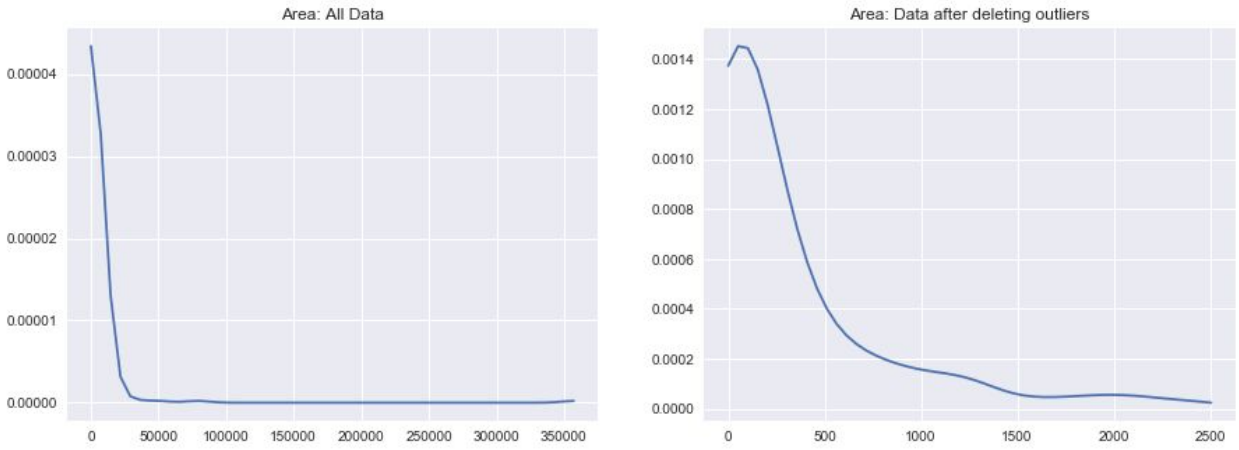
In regards to formulating this task into a data mining problem, there are eight potential religions that we can identify a country as being predominantly comprised of: Catholic, "Other Christian", Muslim, Buddhist, Hindu, Ethnic, Marxist, and "Other" in our dataset. Considering that each one of these is a distinct identifier, then our task is clearly an exercise in classification. Specifically, given a set of attributes on a country, such as population, area, and a host of attributes describing that country's flag, we want to classify the religion of that country as one of the eight religious classes described previously.

Step One

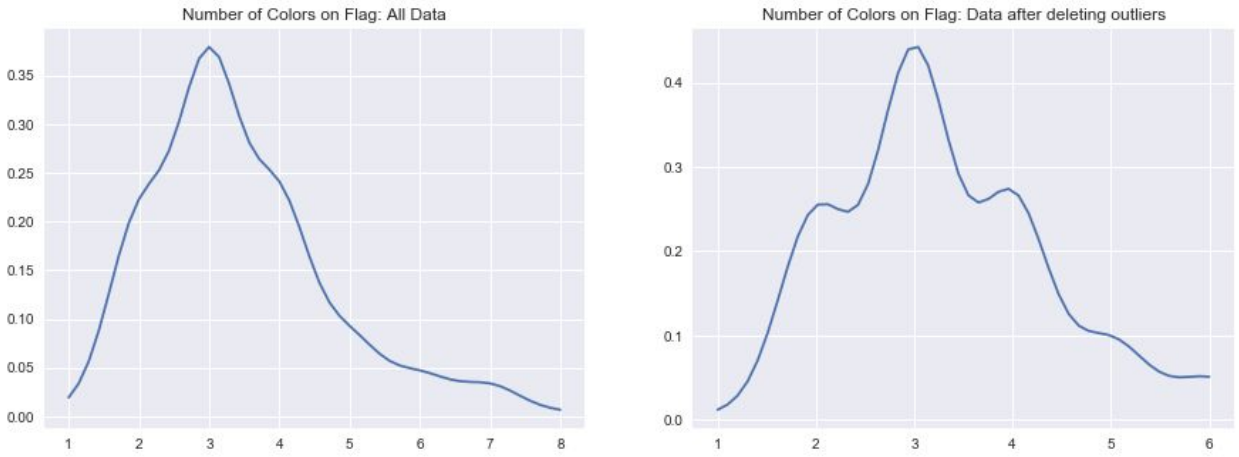
In step one we were asked to collect and clean data, then apply one of the algorithms we learned during the semester and then run our chosen data through the algorithm (decision tree) and explain the results. Before we were able to truly even begin 'cleaning' our data we first faced a couple small inconveniences. In order to get our dataset off of the UCI website, we had to create a web scraper that took the information from UCI and transferred it to our model. Also, because our dataset consisted of mostly categorical data we had to create dummy variables in order to run them through the algorithms offered by sklearn. The biggest mistake that our team made was we did not believe originally that we needed to really normalize our data: we had no missing values and so all we thought we had to do was smooth out the noise and do basic cleaning, it did not seem like we needed to normalize. However, without normalizing our data the quality of our model would have been completely diminished. In order to normalize we removed the top ten percent of our data which remove roughly 10 instances from the dataset, and lead to the distributions of the numerical categories for our dataset shown below:



Graphs showing 'Population' before and after deleting outliers



Graphs showing 'Area' before and after deleting outliers



Graphs showing 'Number of Colors on a Flag' before and after deleting outliers

Step Two

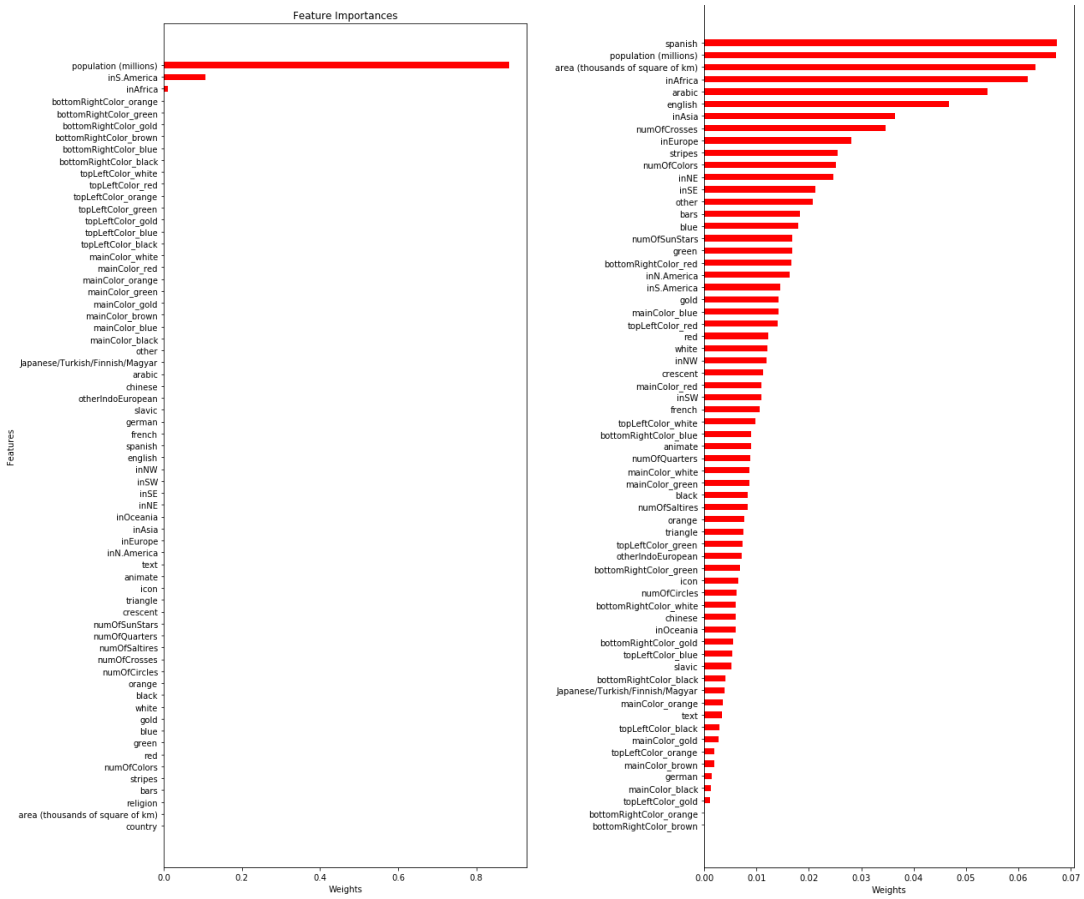
For step two we had to choose at least five different algorithms to run our data. The classifiers we chose to use was the Decision Tree Classifier, Random Forest Classifier, Gradient Boosting Classifier, Support Vector Classifier, and Neural Network Classifier. Below, we give the rationale behind choosing to test each one of these classifiers on our data:

- Decision Tree Classifier: this was the first classifier we learned that was specifically meant for predicting class of an instance based on its attributes. Random Forest Classifier: a large collection of decision trees that randomly sample subsets of attributes and instances in the dataset may decrease our model's sensitivity to unimportant attributes.
- Gradient Boosting Classifier: since each of our attributes only contribute a maximum of ~8% importance, an ensemble of weak predictor may be best suited for our task.
- Support Vector Classifier (SVC): is one of the most popular machine learning techniques in use.
- Neural Network Classifier: neural networks are one of the newest machine learning classifiers and have been used to solve complex problems concerning, among other things, human behavior and culture.

The most accurate model that we were able to produce was the Random Forest Classifier, which yielded an accuracy of 72.68 percent. We imagine that the reason that this classifier performed so well was that it functions by taking random subsets of features and instances. This is beneficial because our data is not evenly distributed, with a large number of countries being Muslim or Catholic. By taking random subsets of these instances, we could avoid some of the possible overfitting that could arise from our data being predominantly Muslim or Catholic. Additionally, as we discovered later on in Step 3, many of our feature have less than 1 percent importance to the classification model, which means that by taking random subsets of the attributes, we avoid assigning too much weight to any of these inconsequential attributes by using a Random Forest.

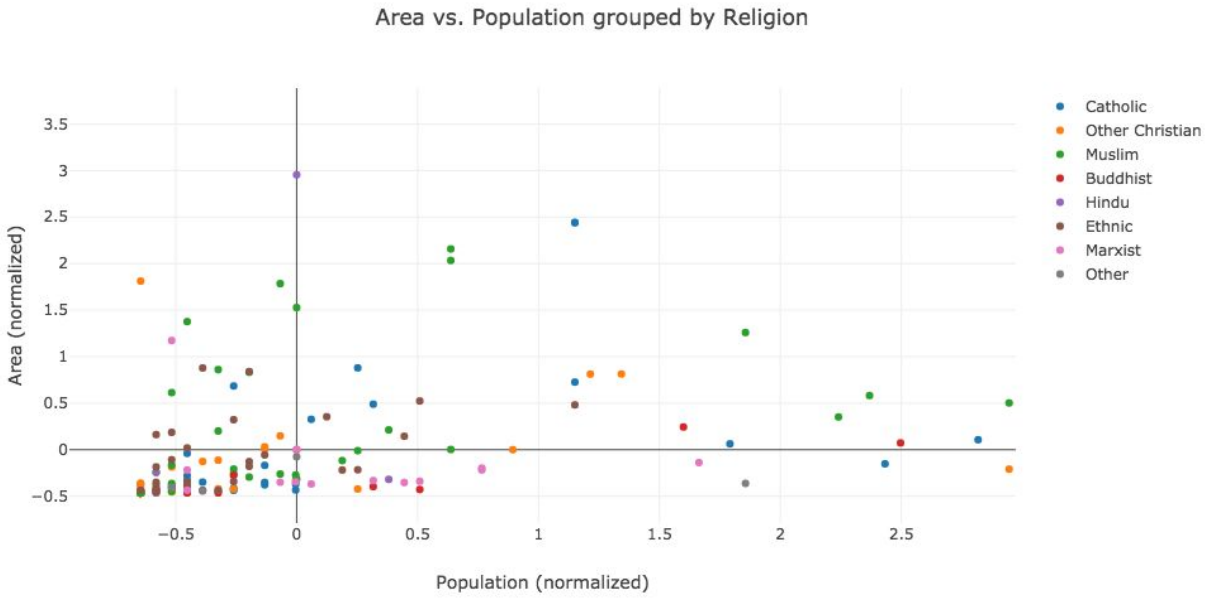
Step Three

In step three we generated 13 new instances of our own in order to test our model. Specifically, since our dataset was last updated in 1990, we added several countries that were founded/formed after 1990. A change in flags had the potential to drastically change what religion we predicted. We also made improvements to our model by normalizing our data. This changed our important features to rely less heavily on the population of a given country and moreso on the attributes of the flag, such as dominant color, and position of symbols.

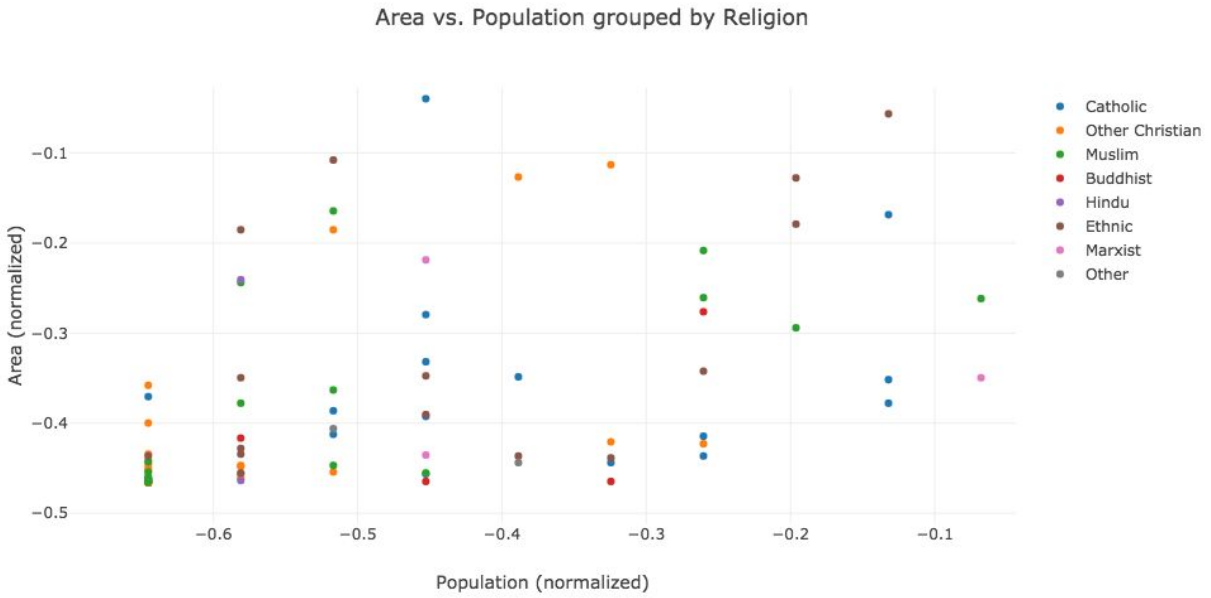


On the left, we have a bar chart of our feature importances from the Random Forest before we normmalized the data. On the right, we have feature importances after normalization.

The above figure show our feature importances before and after normalizing our date. This shows just how big of a difference normalization made. Normalizing our data also made us notice some trends about religion. If the population and area of a country is large, then we tend to predict that the country is Muslim. If only the population is large, then we tend to predict that the country is Catholic.



Above, *we have a graph showing the tendency of countries to be Muslim when population and area are large, whereas countries tend to be Catholic when the population alone is large.*



Above, we have a graph that shows the tendency of countries to be Ethnic when area and population are small. However, we also notice a slight oddity, in that the absolute smallest countries in terms of area and population tend to be Muslim as well.

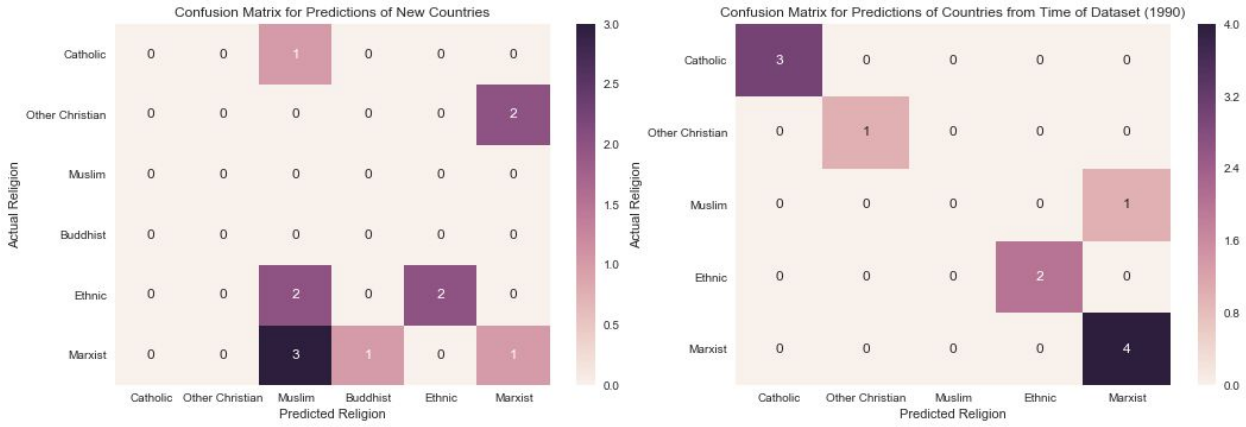
We faced several problems in step three. We tried to improve the accuracy of our model, but didn't end up getting the accuracy as high as we would have liked. We also had a lot of problems getting the visualizations to show up on everyone's computer. We utilized several packages and figuring out which ones everyone had installed and which weren't working properly proved to be difficult at times and cut into the time we had to work on improving accuracy. Another technical issue we had was with Github, we had a lot of trouble with merging conflicts and pulling. This ended up cutting into the time we had to work on our project.

All this being said, we tried several method for improving the accuracy of our model, including trying different cut-offs for normalizing our data (removing the top 15 and 20 percent

of outliers from our data). We also tried providing a larger range of argument values for our Grid Search algorithm to test on the Random Forest Classifier, such as increasing the number of estimators, the maximum depth of the trees, and so on. However, after trying to improve our model in all of these various ways, we only saw about a 2 percent increase in accuracy, which could have been the result of chance. This seems to insinuate that predicting the religion based on attributes of the flag is much more difficult that we originally assumed.

Conclusion

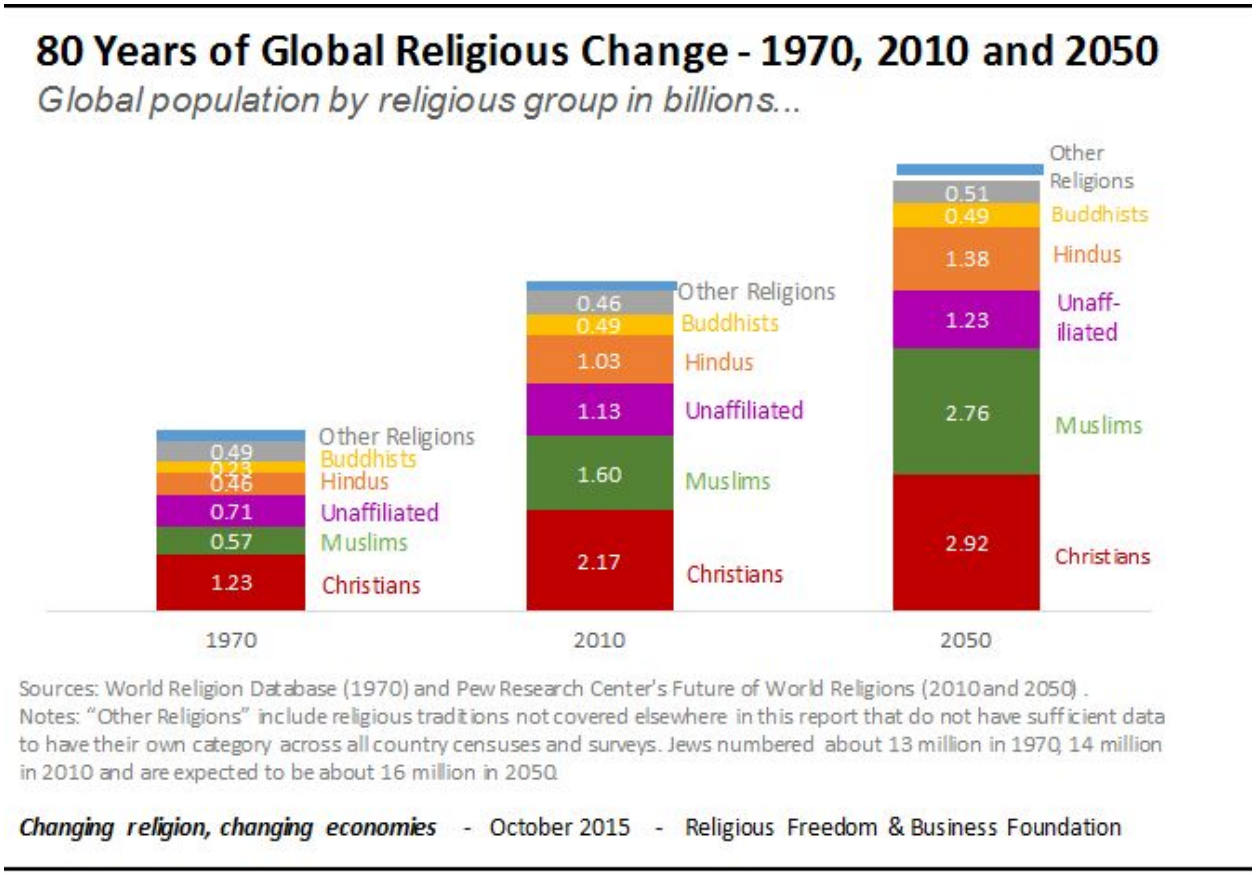
In conclusion, when we tested our model on an untouched subset of the original UCI dataset, and thus, only tested our model on data from countries that existed prior to 1990 the prediction accuracy of our model was approximately 90.9%. In contrast, when we tried to test our model on the contemporary data we generated based on countries that were formed after 1990, the prediciton accuracy of our model dropped to approximately 25%, which is less than a coin flip, which means our model was meaningless for trying to predict on this data.



Left: Confusion matrix using contemporary data (prediction accuracy at 25%)

Right: Confusion matrix using historical data (prediction accuracy at 90.91%)

We believe that this is because in the recent years, there has been a significant increase of countries overall becoming Muslim, which in return creates more outliers which then throws off our data and prediction accuracy.



Graph showing the increase of different religions across recent history (1970-2010; predicting 2050)

Future Work to Improve our Model

One way in which our model could be improved is by experimenting with feature creation. Given the fact that area and population are so important, it is possible that combining these two features together into something like number of people living in a country per square km could also have been an important feature for predicting religion. Another area that has room

for improvement is our exploration of ideal normalization methods. In particular, we were not able to spend much time experimenting with different cutoffs for what was considered an outlier of our data. It is possible that if we repeated our GridSearch and Parameter Searching process multiple times, each of which using a different outlier cutoff (5%, 6%, etc.), we could make our model more accurate. One final aspect that we could have explored in our pursuit of improving our model was curating our dataset to be more evenly distributed, by which we mean that we could have taken greater lengths to ensure that we had roughly equivalent amounts of countries that were Catholic, Muslim, Hindu, etc.