Data Mining Project

Team Leader: Cole Polychronis
Team Members: Jasmine Boonyakiti, Kelsey Henrichsen, Merritt Ruthrauff

Project Details

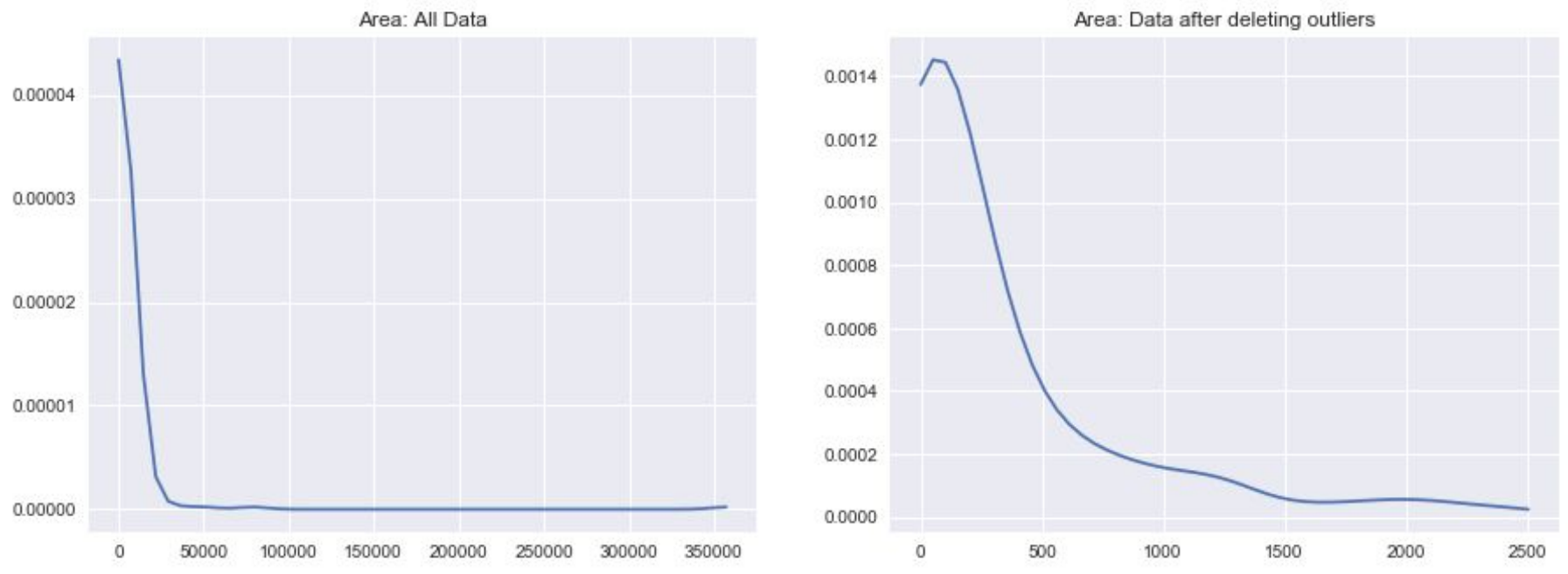
- Using a flag data set from UCI's website
  - 194 Instances
  - 30 attributes
  - Mostly categorical data, with 3 numerical variables
- Goal: predict the religion of a country based upon its flag
- To test our model, we generated 13 additional instances of countries formed after 1990.

Step One

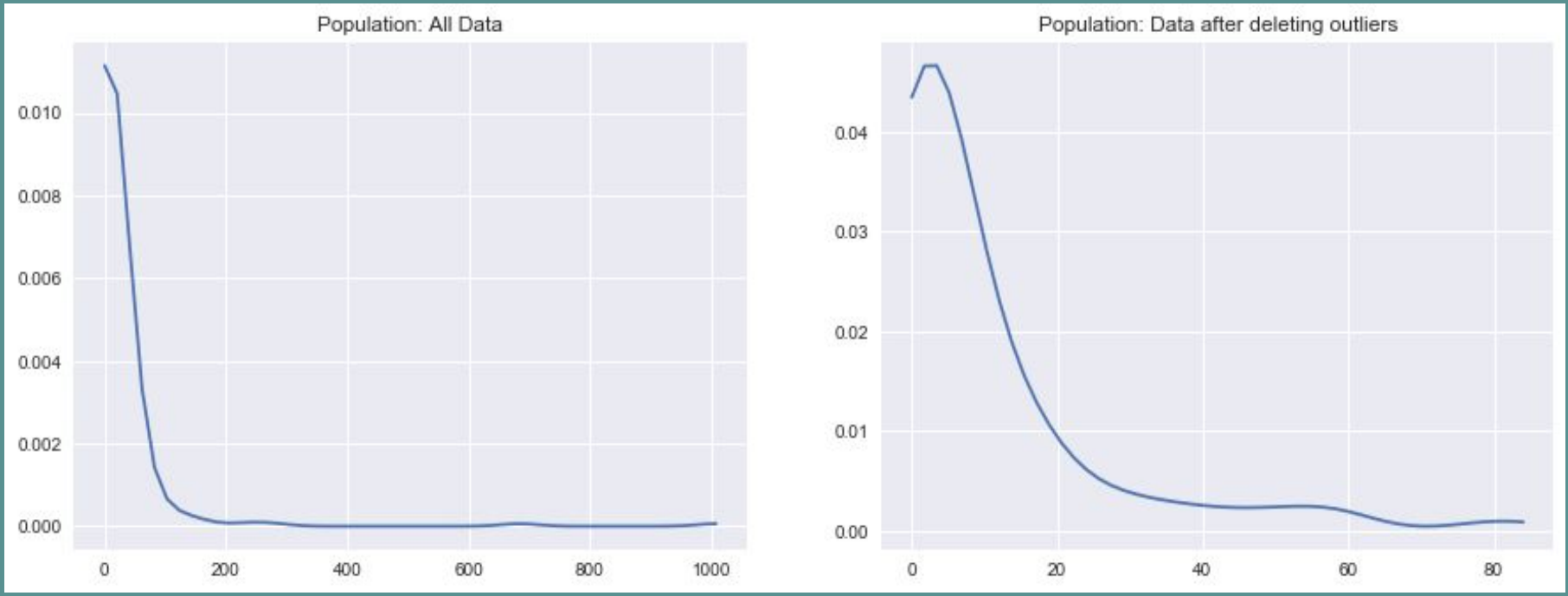
Tasks:

1. Collect data
2. Clean data(noise, outlier, n/a value)
3. Apply one algorithm to run the data explain the results

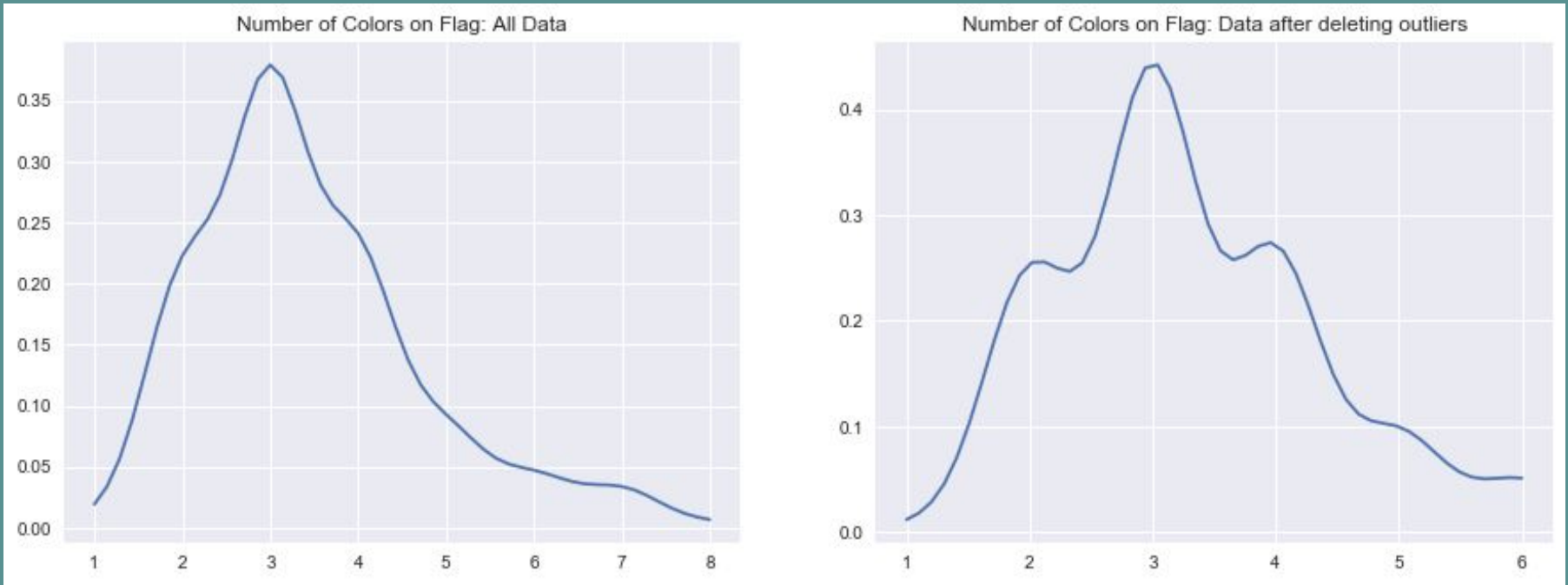
The Importance of Normalization



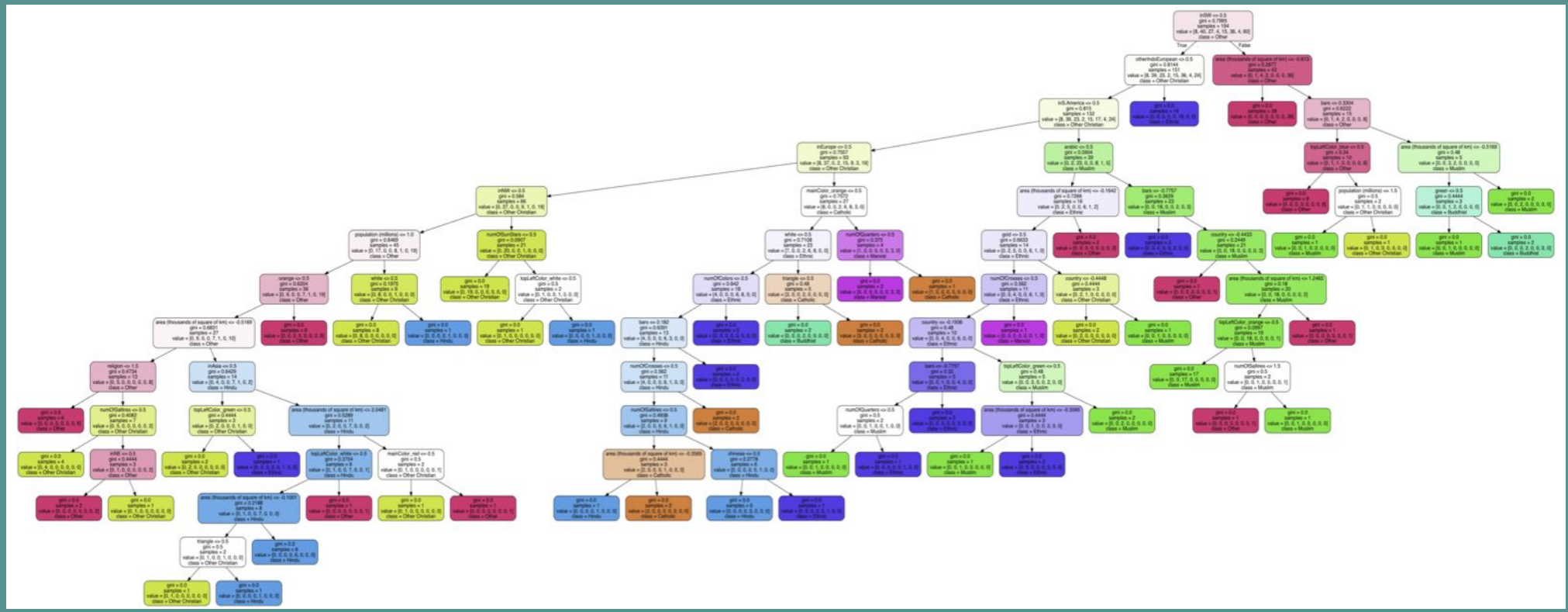
The Importance of Normalization



The Importance of Normalization



Applying the Decision Tree Classifier to Our Data



Problems Faced in Step One:

- While we didn't realize it until Step Two, failing to normalize the data made our model completely overfit the data
- Creating a web scraper to get dataset off UCI's website
- Changing our categorical data to "dummy variables" in order to be able to run them through the algorithms offered by sklearn

Step Two

Tasks:

1. Choose at least five different algorithms to run the data.
2. Use grid search and cross-validation to choose the best parameters and model.
3. Use k-fold, ROC, or other methods to evaluate these models.
4. Draw at least FOUR visualized graphs of the results, using Matplotlib, seaborn, Bokeh, and plotly.

Justification for Algorithms Chosen:

1. **Decision Tree Classifier**
   - First classifier we learned that was specifically meant for predicting class of an instance based on its attributes
2. **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
3. **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
4. **Support Vector Classifier**
   - SVC is one of the most popular machine learning techniques in use
5. **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

Accuracy of Chosen Algorithms:

1. Decision Tree Classifier : 58.25%
2. Random Forest Classifier : 72.68%
3. Gradient Boosting Classifier : 60.82%
4. Support Vector Classifier : 56.19%
5. Neural Network Classifier : 64.43%

Why the Random Forest might be Best:

- Since our data is unevenly distributed among different religions, taking random subsets might eliminate some of the overestimation of Muslim
- Since our data has a large set of features, taking random subsets of the features could help reduce over dependence on a small set of features that a model may overfit the data with

Problems Faced in Step Two:

- Our estimated fits were almost, or actually, perfect. Most likely due to overfitting.
- Before normalization, our only important feature was population, which contributed ~85% of the importance
- After normalizing the data, our models performed significantly worse, which means that we had to spend more time determining why our best model only had ~70% accuracy

Step Three

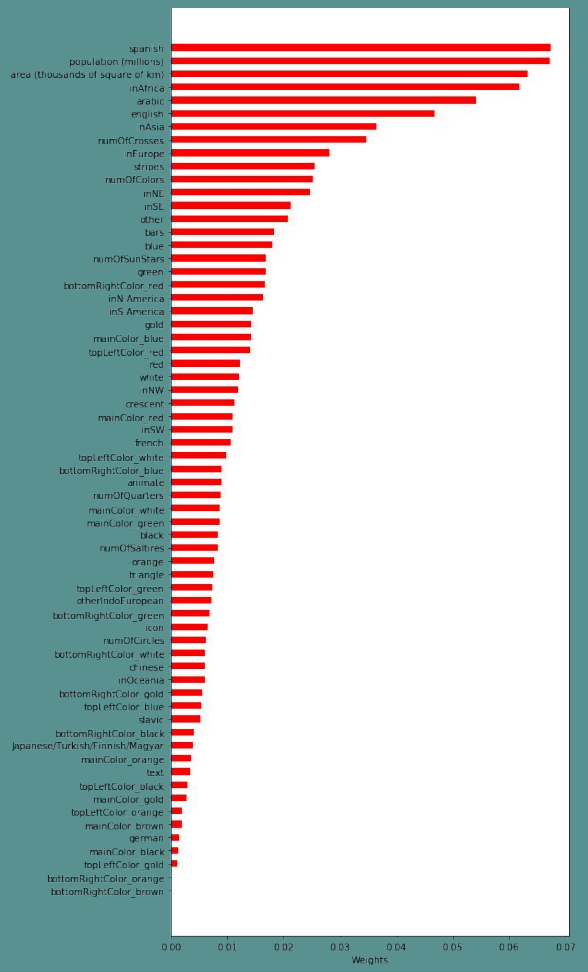
Tasks:

1. Generate your own data for prediction.
2. Explain your visualizations and results.
3. Answer each of the questions.

OLD

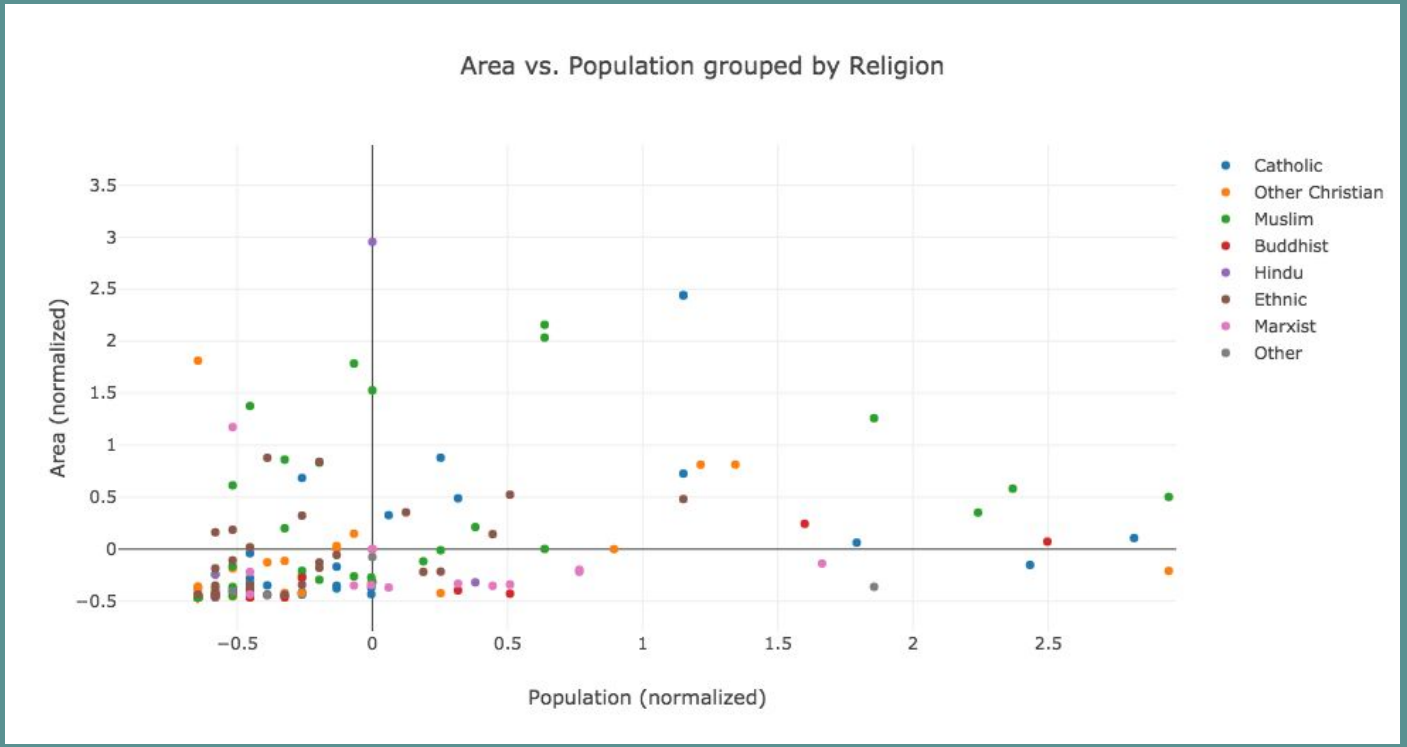

Feature Importances
population (millions)
inS America
inAfrica
bottomRightColor_orange
bottomRightColor_green
bottomRightColor_gold
bottomRightColor_brown
bottomRightColor_blue
bottomRightColor_black
topLeftColor_white
topLeftColor_red
topLeftColor_orange
topLeftColor_green
topLeftColor_gold
topLeftColor_blue
topLeftColor_black
mainColor_white
mainColor_red
mainColor_orange
mainColor_green
mainColor_gold
mainColor_brown
mainColor_blue
mainColor_black
other
Japanese/Turkish/Finnish/Magyar
arabic
chinese
otherIndoEuropean
slavic
german
french
spanish
english
inNW
inSW
inSE
inNE
inOceania
inAsia
inEurope
inN America
text
animate
icon
triangle
crescent
numOfSunStars
numOfQuarters
numOfSaltires
numOfCrosses
numOfCircles
orange
black
white
gold
blue
green
red
numOfColors
stripes
bars
religion
area (thousands of square o' km)
country
features
0.0
0.2
0.4
0.6
0.8
Weights

NEW


spanish
population (millions)
area (thousands of square o' km)
inAfrica
arabic
english
inAsia
numOfCrosses
inEurope
stripes
numOfColors
inNE
inSE
other
bars
blue
numOfSunStars
green
bottomRightColor_red
inN America
inS America
gold
mainColor_blue
topLeftColor_red
red
white
inNW
crescent
mainColor_red
inSW
french
topLeftColor_white
bottomRightColor_blue
animate
numOfQuarters
mainColor_white
mainColor_green
black
numOfSaltires
orange
triangle
topLeftColor_green
otherIndoEuropean
bottomRightColor_green
icon
numOfCircles
bottomRightColor_white
chinese
inOceania
bottomRightColor_gold
topLeftColor_blue
slavic
bottomRightColor_black
Japanese/Turkish/Finnish/Magyar
mainColor_orange
text
topLeftColor_black
mainColor_gold
topLeftColor_orange
mainColor_brown
german
mainColor_black
topLeftColor_gold
bottomRightColor_orange
bottomRightColor_brown
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
Weights

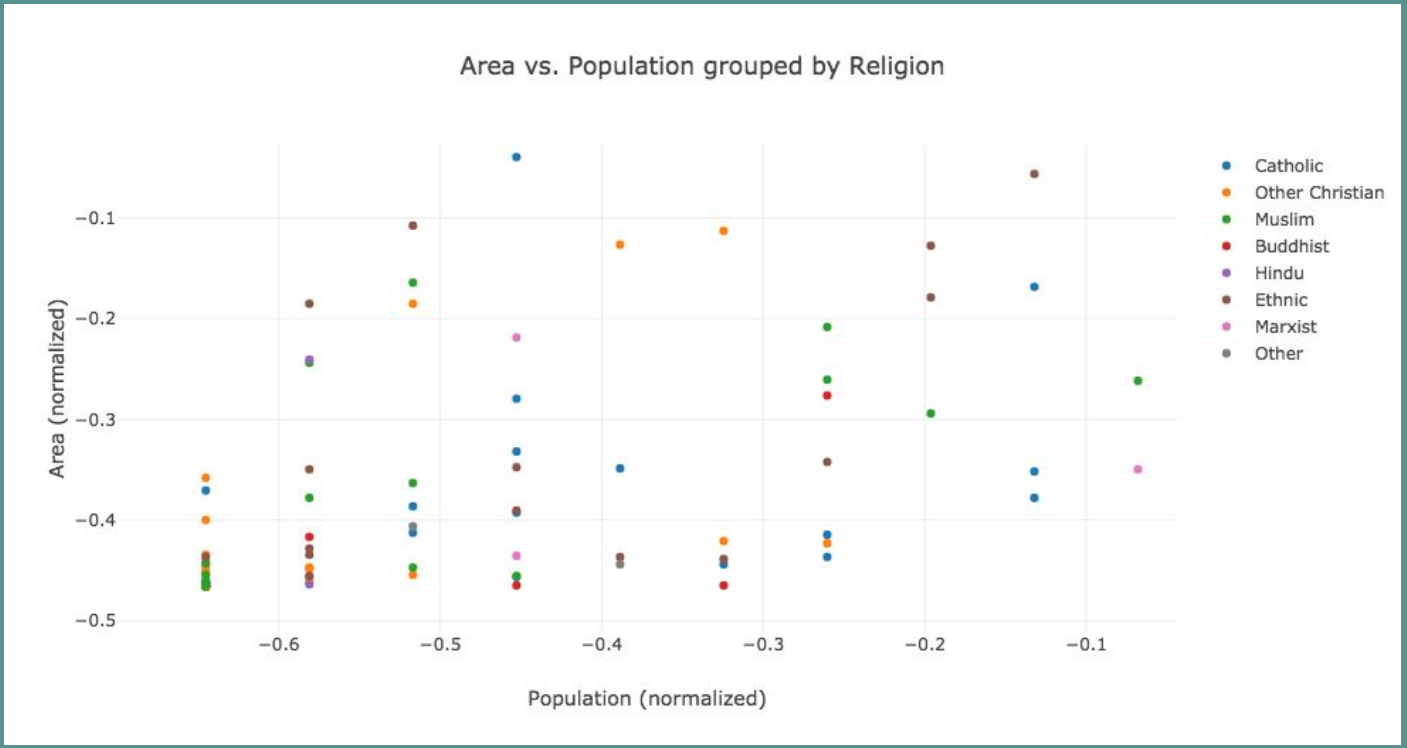
Exploring Important Features (Plotly)

Looking at trends of large area, large population



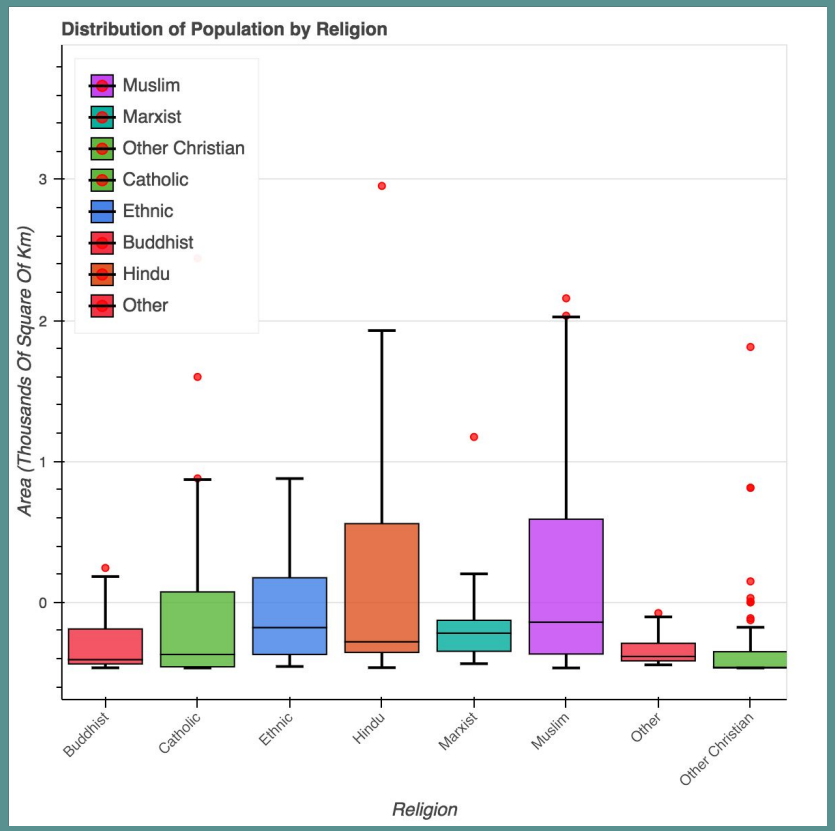
Exploring Important Features (Plotly)

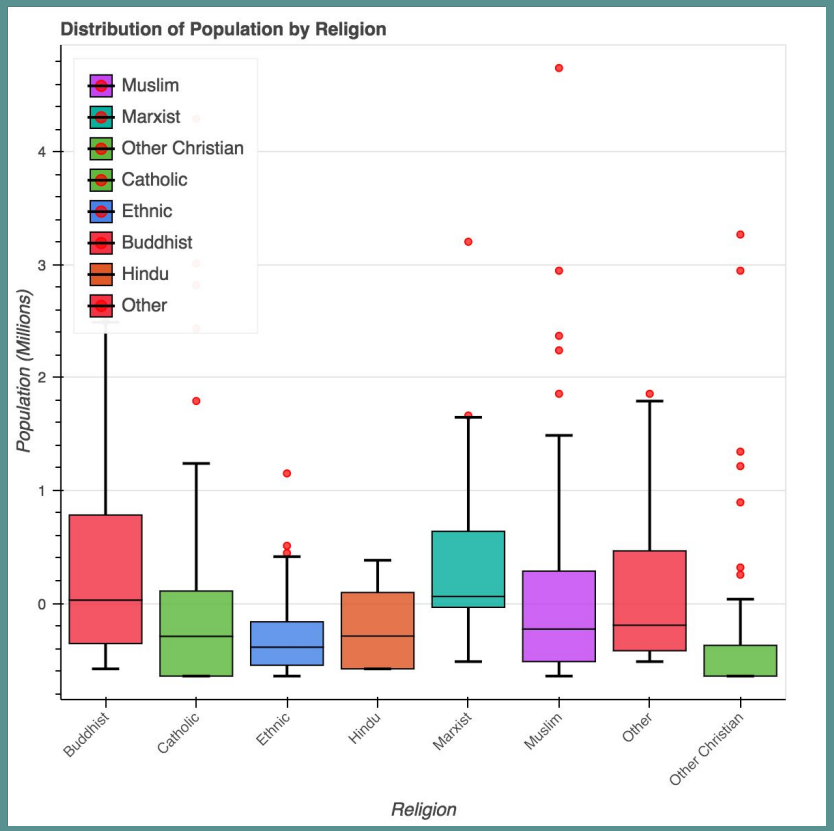
Looking at trends of small area, small population



Exploring Important Features (Bokeh)

Comparing spread and center of Area and Population

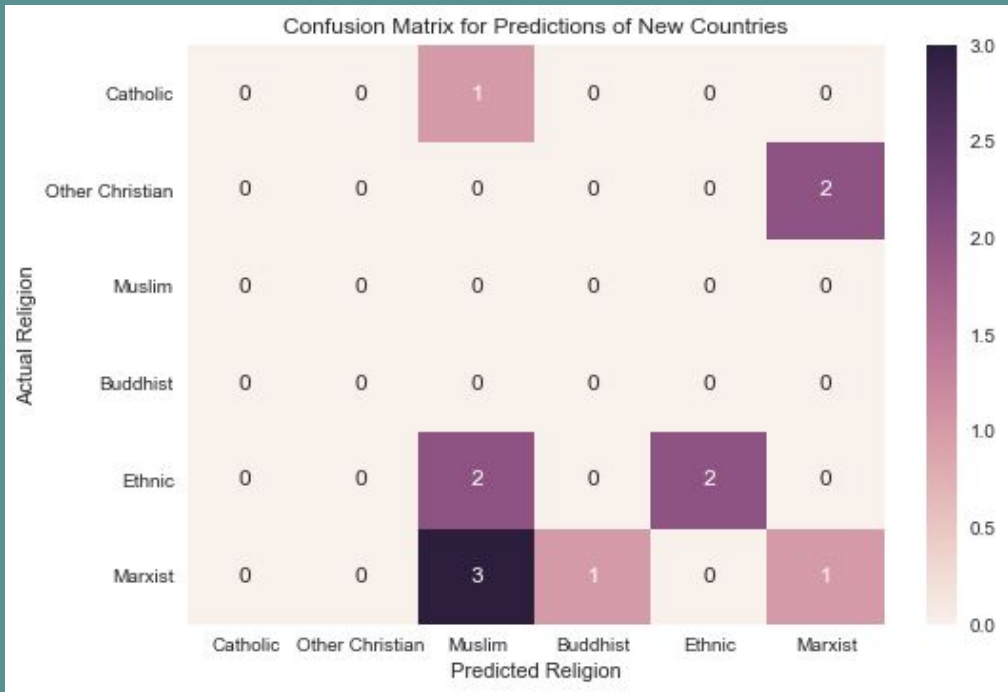


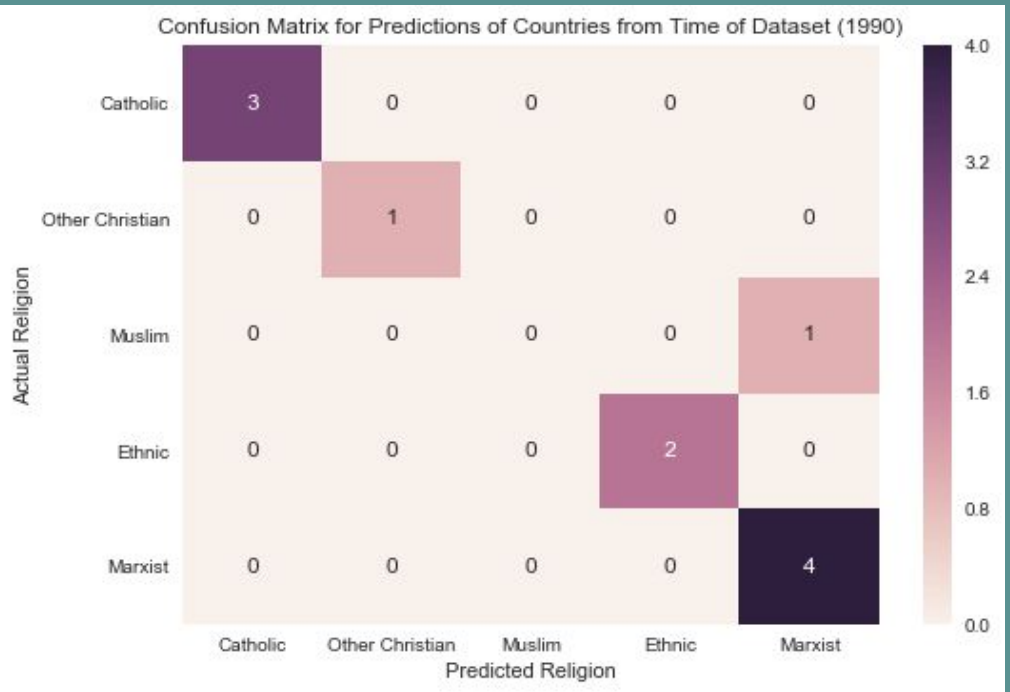


Testing Our Model on New Data

```
Prediction Accuracy for Kernel 1 Data
0.25
```

```
Prediction Accuracy for Kernel 2 Data
0.909090909091
```





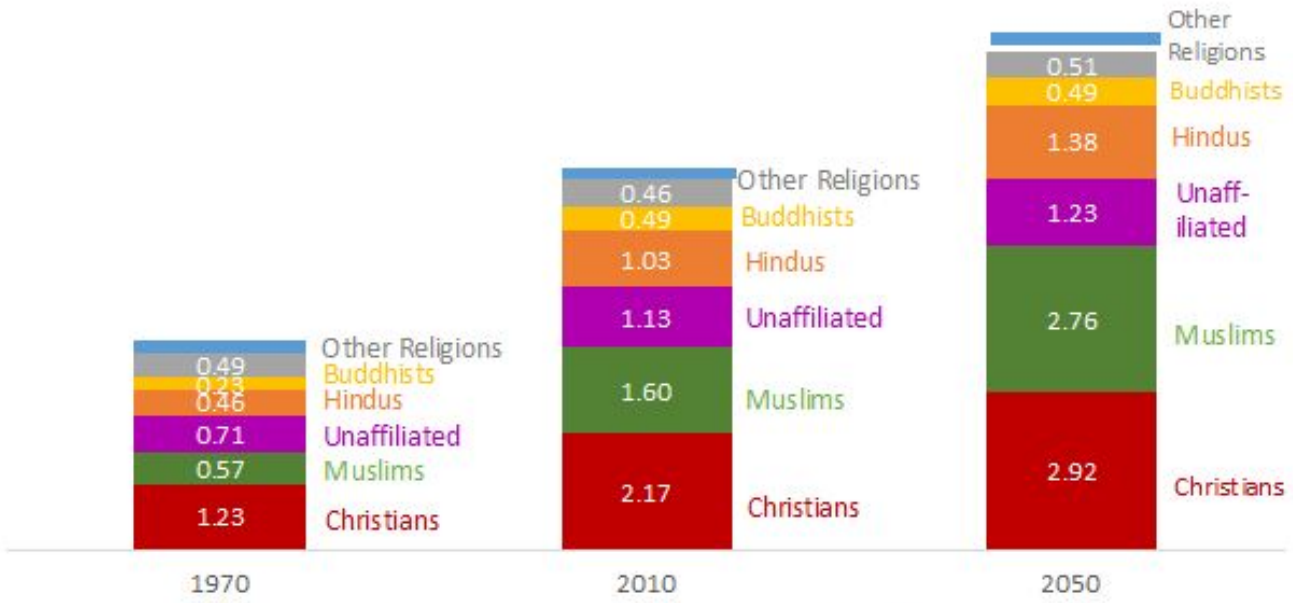
Problems Faced in Step Three

- Trying to improve the accuracy of our chosen algorithms.
- Having the visualizations work on everybody's computers
- GitHub

Wrapping Up

80 Years of Global Religious Change - 1970, 2010 and 2050

Global population by religious group in billions...



Sources: World Religion Database (1970) and Pew Research Center's Future of World Religions (2010 and 2050).
Notes: "Other Religions" include religious traditions not covered elsewhere in this report that do not have sufficient data to have their own category across all country censuses and surveys. Jews numbered about 13 million in 1970, 14 million in 2010 and are expected to be about 16 million in 2050.

Changing religion, changing economies - October 2015 - Religious Freedom & Business Foundation

Why the Model We Chose Worked

- Normalization, Normalization, Normalization
- Also, because Random Forest takes subsets of data it was able to take our unevenly distributed data and decrease our model's sensitivity to the unimportant attributes and focus only on our most important attributes.

Aspects We Can Improve On In Future Work

- Feature Creation
- Normalizing Data
- Creating Equally Distributed Data

All of these “tweakings” of a given data set helps make it be more accurate and overall improves the data making it easier to read and visualize.