Introduction

For this project, I was tasked with building a model that can forecast the probability of unrest one month in advance given unrest.csv, a dataset containing monthly indicators for 50 regions over a 10-year period. These monthly indicators included economic, environmental, and sociopolitical data, along with whether any significant unrest occurred that month.

Methods

I first performed exploratory analysis of the features to gain an insight into which indicators had a strong correlation with unrest occurring the next month. After gaining an understanding of the distribution and importance of individual features, I moved on to building a set of baseline models. These simple models served as a reference to assess whether tuning the model or feature engineering offers meaningful improvements in predictive performance. I evaluated the performance of the four models, Random Forest, Naive Bayes, support vector machine (SVM), and Logistic Regression.

In the next iteration of model development, I used Synthetic Minority Oversampling Technique (SMOTE) to address class imbalance and assessed the performance of the same four models.

I finally did feature engineering based off of previous research on the topic, and used time-based splitting into test and train sets to preserve temporal integrity. This iteration of model development yielded the best results, with some models having nearly perfect accuracy.

Results

The table below depicts accuracy, precision, and recall recorded for each of the four models assessed using the three approaches mentioned above.

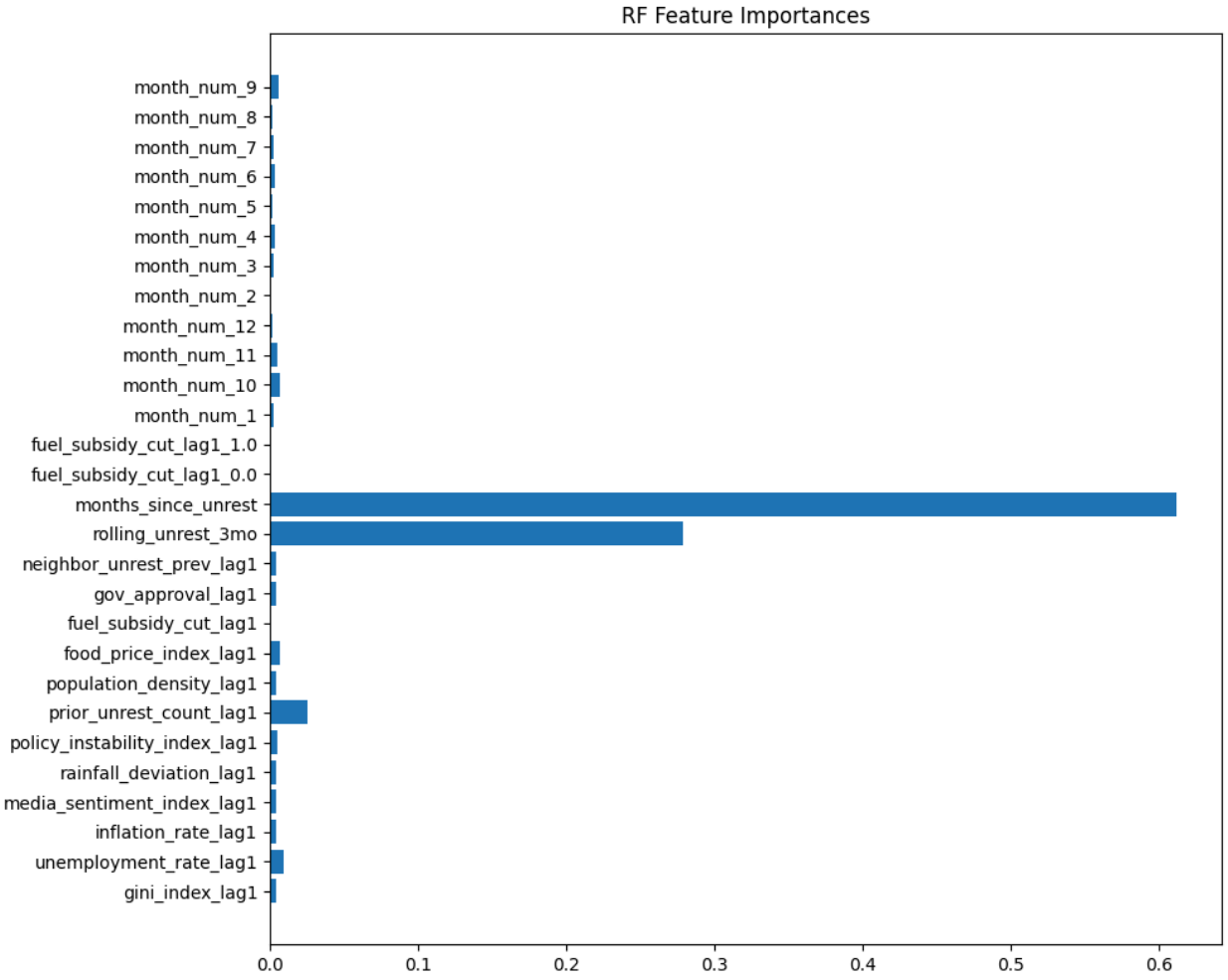
Model Name	**Approach 1: Baseline**			**Approach 2: Using SMOTE**			**Approach 3: Feature Engineering**		
	Accuracy	**Precision**	**Recall**	**Accuracy**	**Precision**	**Recall**	**Accuracy**	**Precision**	**Recall**
Random Forest	0.75	0.56	0.57	0.69	0.55	0.58	1	1	1
Naive Bayes	0.8	0.61	0.6	0.79	0.61	0.6	1	1	1
SVM	0.5	0.53	0.55	0.39	0.56	0.59	0.94	0.87	0.89
Logistic Regression	0.75	0.59	0.62	0.77	0.6	0.62	1	1	1

Recommendations

Based on the factors used to assess the performance of the models, I found that the models performed best after the addition of two derived features: *rolling_unrest_3mo* and *months_since_unrest*, and using SMOTE to balance the training data.

These two features capture patterns in the occurrence of unrest. Recent unrest events emerged as highly predictive indicators of future unrest, as shown in the figure below

.



Rules for Policy Analysts: Policy analysts should be aware of the fact that the presence of unrest in the preceding months is strongly associated with an elevated risk of unrest events in later months, suggesting that unrest tends to cluster. Therefore, should the ‘months_since_unrest’ variable be zero, there will likely be unrest the following month.