Visualizing and Predicting Conflict in Africa

This repository provides the files for the capstone project for five students in the George Washington University data analytics bootcamp. It represents the culmination of the skills developed across the six-month class from March to September 2020.

For the project, we developed a full-stack web application that created interactive ways for users to explore data related to internal strife in African countries. The files in this repository are designed to interact with each other to call data, arrange it, visualize on maps and charts, and calculate predictions. The files include:

* Flask to establish a connection between our database and our web application;
* SQL to transform a variety of Excel spreadsheet into a database that is easy to pull into the app;
* Python to organize data for loading to a Amazon Web Services database and for data analysis and machine learning to predict future conflict;
* JavaScript with D3.js to design maps, charts, animations, tables, and interactive tools, using libraries such as Plotly and Mapbox;
* HTML, CSS, and Bootstrap to format, design, and make the application responsive.

The project was voted by our classmates as the best of six presented on the final day of class.

**Who We Are [everyone writes a one or two sentence bio plus how you contributed to the project; we also could include links to our LinkedIn profiles]**

As in any group coding project, each of us took the lead on sections and relied on the group to troubleshoot code and finetune the app. We had extended Zoom sessions where we worked independently and collaboratively over the course of two weeks, while balancing our professional responsibilities. At key moments, we sought help from our instructor, Dartanion Williams, and tutors.

The team behind this app is:

**Carly Kelly**

**Chris Hanafin**

**Cole Fingerut**

**Yasir Omar**

**David Hoff** is a communications professional who enrolled in the GW bootcamp to add data skills to enrich his storytelling. For this project, he collaborated with Cole on the statistical analysis for machine learning, designed maps and charts in Tableau, and drafted the readme.

At the end of the project, we presented the app to our final class session.

**About the Data**

From the beginning, we wanted our project to stand out as unique and built on our interests and expertise. To encompass all that we learned in the bootcamp, we wanted to integrate machine learning that either predicted or classified future events, meaning that data needed to diverse and extensive.

We explored several options, but the Africa Conflict stood out. COLE, COULD YOU BUILD THIS OUT?

Based on our review of academic literature identifying factors related to conflict, we searched for data covering the ethnicity and population, economy, governance, and foreign direct investment (FDI). To ensure we would be able to conduct sophisticated machine learning and provide time series visualizations, we collected data from 1998 to 2018. The specific indicators included GDP and GNI; a variety of indices related to ethnicity; government effectiveness, accountability, and corruption; population; FDI inflows; and mortality rate.

rate, and ethnic diversity – all of which were the independent variables for analysis. The dependent variable was the number of conflict events in each year. The events ranged from peaceful protests to riots to civil wars. We also collected the number of fatalities due to the conflicts as a potential dependent variable. But it didn’t test well in our machine learning approach.

**About the Application**

**About Machine Learning**

We tested a variety of statistical approaches using both linear and logistic regressions. Before our analysis, we surveyed the conflict data – both the number of events and number of fatalities. The numbers varied widely from country to country and even within countries. From that, we hypothesized that it would be difficult to build a model that precisely predicted either events or fatalities. We predicted that the best approach would be a logistic regression that classified countries based on the potential for conflict.

Still, we started our analysis by running several multivariate regressions. Our goal was to see how the interacted and identify pieces that contributed to large r-squared scores. In our initial runs, we discovered we correlations among out datasets; the highest r-squared scores were around 0.5. We sought advice from social science researchers and statisticians. When we described that our data represented different sectors including from economic, governance, and population. In addition, many data columns were indices. They told us that, given those types of data, we couldn’t expect to get r-squared much higher than 0.5. In fact, social science research often draws conclusions based on that level of correlation. With that support, we moved forward to identify the factors that led to the highest r-squared. They included a combination of economic and governance data. In general, population data didn’t contribute to the correlations.

Our next step was to identify the classification model. Using the Support Vector Classification is SKLearn, we entered data from our selected datapoints to identify the best of the following models: Logistic Regression, Decision Tree Classifier, Random Forest Classifier, KNeighbors Classifier, Linear Discriminant Analysis, and Gaussian NB. The Random Forest was clearly was the best model fit with the highest the mean cross value score of 0.949 and the lowest standard deviation 0.03.

by finThat led to our hypothesis from the start that a classification model would be the best fit for our data. The conflict events in our data were in a wide range of categories. Moreover, they were not well-defined statistical indicators, such as economic or stock market data. We found it useful to put the data through multivariate regressions to test how different factors could predict conflict. At the beginning, we were surprised that our r-squared results were less than 0.5. We consulted with social science researchers and our instructor about our results. They informed us that researchers who analyze data that are indices representing diverse datasets often publish based on r-squared in that area. That gave us the confidence to move forward and reinforced our approach of developing a classification model that would identify the economic, social, and governmental factors that could lead to conflict.

Based on that advice, we decided moved forward with logistic regressions using two methods: the random forest and neural networks. Using data that produced the best r squared in the linear regressions, we created regressions with accuracy rates of better than 97 percent. The independent variables aligned with factors identified in the academic literature: ethnic diversity, economic success, government corruption and effectiveness.

The regressions provided that we could use our data to create two interfaces for users:

* A dashboard that identified countries that were at risk of conflict in the future based on their past data;
* A tool that would calculate the potential for conflict based on new data for a country.

In the random forest, we tested several combinations of data to identify the best data for the regression. We decided on 10 data points: ethnic score, rule of low, government effectiveness, mortality rates, total GDP, total GNI, FDI inflow as GDP, and FDI inflows total.