Data Science Predictive Application for Country Instability (DSAPCI)

Courtney Beckham and Lancie Affonso
College of Charleston Department of Computer Science

Abstract

Abstract: This project explores the follow-up model of the Analyzing Complex Threats and Operations Readiness (ACTOR) model employed by the United States government. The ACTOR model aggregates data on 12 factors from 159 countries over the period 1975-1999 to predict the level of instability a country will experience. The new modeling system is known as the Integrated Crisis Early Warning System (ICEWS) and incorporates the same background as the previous model, but with a few updates and a public database. This project used that database to explore the importance of data visualization in predictive modeling. Building off of data from the ICEWS database, an interactive website was created to fill the gap between non-technical and technical people alike. By allowing the data visualizations to be customized, users are allowed to explore the data and build a report out of their findings.

Quick Summary: The building of a predictive application in the form of a website to predict country instability and cater to the need for data visuals to tell the story of numbers.

Objectives

This project was created as a follow-up to a previous project on the ACTOR (Analyzing Complex Threats for Operations and Readiness) model built by the United States government. Main objectives include:

- Interactivity for the customer
- Ability to select many options from which to build a data visual or model
- The ability to download generated graphics
- The ability to create and download pdfs of selected visuals
- Polished front end that gives the feel of a professional application

Methods

Software: After experimenting with several different backends, I ended up settling on Django. Django offers a well-documented community, high versatility, and an extremely extensive package library. For the data visuals, I used Dash, which offers polished graphics with embedded interactive features.

Data: Data was used directly from the Integrated Crisis Early Warning System (ICEWS) Database from Harvard University. The specific dataset used aggregated provincial data on 9 countries from the years 2005 to 2013. Event codes included are listed in the chart below. Events were rated with intensities based on the Goldstein scale which ranges from -10 to 10 with "negative numbers representing hostile actions and positive numbers representing cooperative actions." (Lautenschlager)

Model: To predict future values, trends and forecasting were used from the Tableau analytics page. Below are examples of each.

Figure 1. The image on the left displays the forecasting visualization for each event type. These are broken down by country and generated using exponential smoothing. The image on the right displays the trend lines for each event type. They are broken down by country and generated using a polynomial trend line.

Website

Website layout: The website is made up of three main content pages: Visualizations, Reports, and Predictions. The Visualizations page allows the user to generate customizable graphs and download them as PNG images. It consists of 5 subtabs, as pictured below.

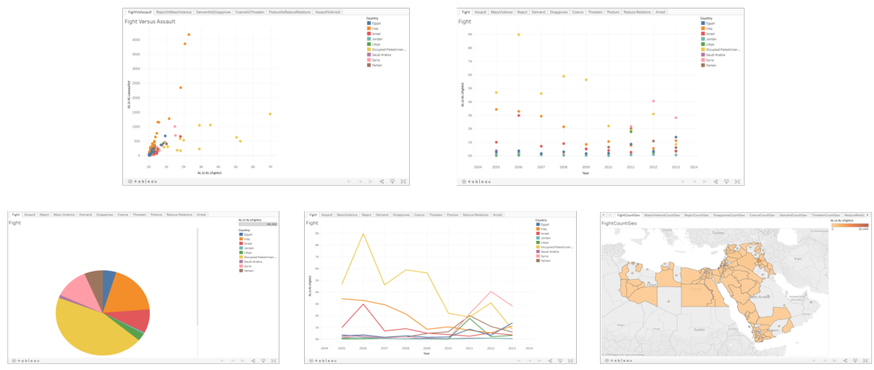

Figure 2. These images represent each subtab of the Visualizations page (from left to right, top to bottom): Scatter Comparisons, Scatter Plots, Pie Graphs, Line Graphs, and Geo Graphs.

The graphs above can be downloaded by the user in a myriad of data types, leaving customization in the hands of the user. The legends give the ability to select a subset of the data to be displayed, which can give a better profile of each individual country. Similarly, the Reports page allows the user to download visualizations by 3 separate groupings: ScatterVsLine, LineVsPie, and ScatterVsGeo. Each subpage is shown below.

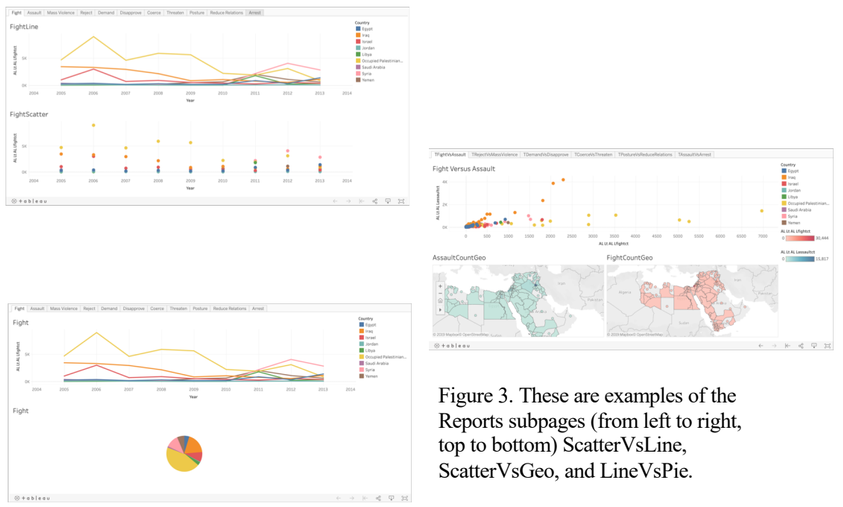

Figure 3. These are examples of the Reports subpages (from left to right, top to bottom) ScatterVsLine, ScatterVsGeo, and LineVsPie.

Website Design: The frontend of the DSAPCI website was designed using Bootstrap, a seamless open-source CSS framework that streamlines design across an entire website. Colors were chosen to be attractive to all ages, and fonts were chosen for their readability. The inclusion of an easy-to-use navigation bar was an important feature, as it allows the user to effortlessly navigate between pages. Below is an example of the header and navigation bar for the website.

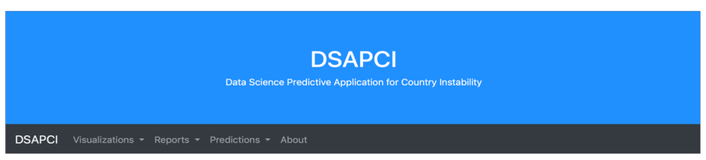

Figure 4. The navigation bar and header for the DSAPCI website.

Modeling Technique

Trends: In Tableau, the analytic package offers trend lines that interpret the data and predict the future trends of the data. For this dashboard, polynomial trend lines were used of order 3. Polynomial modeling was preferred over linear or logarithmic due to the randomness of the data and the presence of several peaks and valleys amongst the average line plot. Tableau's trend lines technology is typically pretty good, but can gain more accuracy with fine-tuned paramaters, like knowing when to use polynomial over logarithmic. By extending the year axis into the future, we have nice visuals that not only seek to describe our current data, but makes a prediction about the trends of data in the future. When hovering over a trend line, a tooltip pop-up contains three items: event type formula, r-squared, and p-value. The event type formula shows the formula calculated and used to generate the trend line. The two statistical values, r-squared and p-value, allow us to know how well our trend line fits the data. R-squared is a goodness-of-fit measure and is always a value between 0 and 1. Values that are closer to 1 tend to express a better model with higher confidence in future predictions. However, before interpreting on the R-squared alone, the p-value must be taken into consideration. A p-value of less than .05 is means that the trend line model Tableau generated may be significant. Combining these measures allows the users to see which trend lines are good predictors for future events, and which ones are not. Due to the random fluctuation of some data, trend lines may not be able to effectively fit certain countries in a certain event type.

Forecasting: In addition to trend lines, the analytic package that Tableau offers also has a feature entitled forecasting. Forecasting "uses a technique known as exponential smoothing." Exponential smoothing models "forecast future values of a regular time series of values from weighted averages of past values of the series." [2] These algorithms try to find patterns in the data that can be projected into the future. During this process, Tableau uses eight different models and picks the one that generates the best quality forecast. This means that Tableau optimizes the modeling process for you, instead of having to sift through which models are best for your data. However, there are some ways to improve forecasts through fine-tuning certain parameters. By allowing periods to be automatic, we allow Tableau to find which patterns are most evident in the data apart from "seasons". Also, through creating a custom model with seasons set to "additive" we are able to get a better idea of the pattern going into the future. An additive model is one in which the model components are summed, and just says that the component that effects the trend is present in the data. The descriptions of each model are also present in three dashboard tabs.

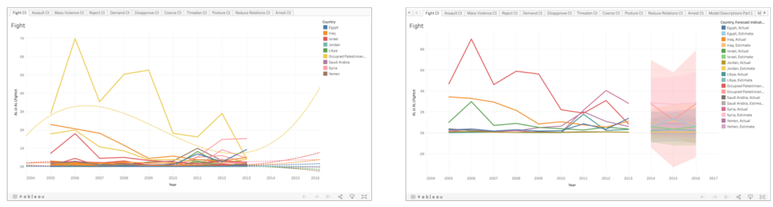

Figure 5. This figure illustrates the scope of each model, with the decision tree model taking a holistic approach to modeling the ICEWS data, all the way up to multinomial logistic regression providing a very specific look at individual data points.

Conclusion

Through this project, every aspect of web development was explored. From the need for data visualization and how to craft a pleasing website to writing backend code, I was able to learn about what it takes to build a website from scratch. While this can be a process with its ups and downs, ultimately the product can be a beautiful website that seeks to help others in their data exploration. I hope that the DSAPCI website can be one of many examples for years to come on how we seek to convey data truths to the average user.

References

[1] Lautenschlager, Jennifer; Shellman, Steve; Ward, Michael, 2015, "ICEWS Event Aggregations", https://doi.org/10.7910/DVN/28117, Harvard Dataverse, V3.

[2]"How forecasting works in tableau," https://onlinehelp.tableau.com/current/pro/ desktop/en-us/forecast how it works.htm.

Acknowledgements: I would like to express my gratitude to Professor Lancie Affonso for his continued guidance on this project. I would also like to thank SMART. The SMART Scholarships is funded by: USD/R&E (The Under Secretary of Defense-Research and Engineering), National Defense Education Program (NDEP) / BA-1, Basic Research. I would like to thank the Honors College as well as the Computer Science and Math departments at the College of Charleston for their continued teaching and support.