DS301 Project Proposal

Jaeyeon Won

Terrorism is Happening: What Are the Consequences?

Terrorism is "the threatened or actual use of illegal force and violence by a non-state actor to attain a political, economic, religious, or social goal through fear, coercion, or intimidation" (GTD).

Project Description

This project is aimed to find the model that best predicts the casualties of terrorism. Specifically, the casualty of terrorist attacks would be defined by the total number of hostages, kidnapping victims, fatalities, and injured from the victims as well as the perpetrators.

As predicting the number and magnitude of the casualties is a complicated task, it is important to develop a model that accurately estimates these parameters. This machine learning project can especially be helpful for governments in terror-prone regions to understand the behavior of terrorist activities and the factors behind it. The model I would develop can also be used for the governments to effectively design strategies and be more prepared to deal with the consequences of terrorist incidents ahead of time.

Overview of Data Set

The dataset can be assessed at https://start.umd.edu/gtd/.

The codebook can be assessed at https://start.umd.edu/gtd/downloads/Codebook.pdf.

The dataset ***globalterrorismdb_0919dist*** comes from the **Global Terrorism Database (GTD**), an open-source database that provides systematic information on terrorist activities occurred around the world since 1970.

The ***globalterrorismdb_0919dist*** dataset contains 135 variables for 191,465 global terrorism cases that happened between 1970 and 2018. With these features, a wide range of exploration is available, including detailed information on the incident, attacks, weapons, targets, perpetrators, and casualties. Out of these original 135 variables, 9 variables are selected as a y variable, and 27 variables related to the y variable are selected as x variables. The 9 variables will be aggregated into 1 final dependent variable. As many of the qualitative variables are encoded as numbers, in the data cleaning process, the categories will be decoded and renamed with appropriate names based on the coding rules described in the codebook, as needed.

Types of Machine Learning Systems

As the purpose of this project is to predict the casualties based on some relating factors, it will follow the supervised learning systems. The model would be developed by the multiple regression process because the numeric results are targeted to be estimated with more than one feature. Specifically, multiple variables, including the country, region, attack type, target type, weapon type, nationality, number of perpetrators involved, etc., will be used to produce a single prediction of the number of casualties for each terrorist activity.

Performance Measures

For this regression machine learning project, the root mean squared error (RMSE), mean absolute error (MAE), and R-squared (R^2) together will be used as metrics to diagnose the model performance. Moreover, to see the dispersion of the values in the models, the standard deviation will also be measured.

Even though all performance measures are equally important in distinguishing the model, the RMSE would be the main metric since it is expected for the error in the model to be non-linear. However, to avoid catastrophic failure, more questions on which performance measures would be most important for model selection will be examined in the data exploration step.

Assumptions

The data are assumed to follow a polynomial regression model. That is, the relationship between the dependent variable and the features is expected to be more complex than a simple linear line. I also assume that the features are independent of each other, and the errors are independent and normally distributed with a mean of 0.

Works Cited

"Global Terrorism Database." GTD, The National Consortium for the Study of Terrorism and Responses to Terrorism, start.umd.edu/gtd/.