Executive Summary

Business Forecasting Final Exam – Fall 2021

Jeremy James, Praveen Kumar, Sam Wagner, Zach Mazgaj, and Lakshmi Sravani Garimella

# Introduction

Terrorism is one of the most serious challenges to modern civilization. It disrupts a society's need for law and order and reduces quality of life for its members, who end up physically and emotionally oppressed. The goal of this project is to identify the Tehreek-e-Taliban Pakistan (TTP) armed group attacks. The TTP’s stated objectives are the expulsion of Islamabad’s influence in the Federally Administered Tribal Areas and neighboring Khyber Pakhtunkhwa Province in Pakistan; the implementation of a strict interpretation of sharia throughout Pakistan; and the establishment of an independent Pashtun state. Within this summary, we will outline our data, models, and forecast.

# Data

Our data comes from the dataset called “pakistanClean.csv”; a file containing information of attacks within Pakistan. Terrorist organizations sometimes claim their attacks and sometimes do not. There is a lot of information for each attack from 2007 to 2018, the only column we used was the TTP which represents whether the attack was responsible by Tehreek-e-Taliban Pakistan (TTP) or not.

# Models

Random Forest

Random Forest is one of the most widely used tree-based algorithms and is one of many ensemble methods. It can produce good predictions that can be understood easily.

### XGBoost

[XGBoost](https://xgboost.ai/)is a decision-tree-based ensemble Machine Learning algorithm that uses a [gradient boosting](https://en.wikipedia.org/wiki/Gradient_boosting) framework. In prediction problems involving unstructured data (images, text, etc.) artificial neural networks tend to outperform all other algorithms or frameworks. However, when it comes to small-to-medium structured/tabular data, decision tree-based algorithms are considered best-in-class right now.

## Comparison

Of the models used for forecasting, we found Random Forest had better accuracy. Table 1 shows the accuracy of both models. We have also used visual analysis and found its predictions generally closer to the history of the actual attacks than XGBoost. Although more testing may be needed to verify that Random Forest is the better model, there is another reason to choose Random Forest as well.

We chose Random Forest because of its ease of use and interpretability. Random Forest is designed to capture the trends and seasonality in a time series without much effort required in finding the correct parameters. Once it finds the trends and patterns in the time series, it is necessary to understand what they are, and how they combine to produce the final forecast.

Table:1 Models with accuracy

|  |  |
| --- | --- |
| Model | Accuracy |
| Random Forest | 82% |
| XGBoost | 80% |

# Forecast

As mentioned previously, we forecast that TTP is responsible for 58% of unclaimed attacks in Pakistan. Below are plots of our final forecast for unclaimed attacks by TTP.

For a better understanding of the model’s logic, we can not only verify it is correct but also work on improving the model by focusing only on the important variables. Figure 1 shows the importance of each variable features using [XGBoost](https://xgboost.ai/).

Figure 1: Comparison for feature importance using XGBoost

Table

Description automatically generated with low confidence

Figure 2 shows the importance of the variable features using Random Forest.

Figure 2: Comparison for feature importanceA picture containing chart

Description automatically generated

Shapley values can help us understand what and how features contributed to our final prediction. The goal is to understand how a model got to the prediction it did. In Figure 3, we see that the number of terrorists killed, along with a couple of location features, pushed our TTP probability over 50%. While there were other features that would lead us to believe it was not a TTP attack, they were not enough to bring us below 50%.

Figure 3: TTP attack label justification

Chart

Description automatically generated

After the model was built, we used partial dependence plots to show the relationships the model had captured. We found that Khyber Pakhtunkhwa had more attacks by TTP since it is on the Afghanistan–Pakistan border where the TTP is more active, while Baluchistan had fewer attacks since it is a region with less TTP activity. Figure 4 shows the relationship for Baluchistan.

Figure 4: Baluchistan TTP attack relationship

Chart, bar chart

Description automatically generated