Project Overview

Predicting casualties in a terror attack is a complex and challenging task that involves the analysis of various factors, including the type of attack, target location, and the nature of the weapon used. Machine learning models and statistical analyses can be employed to assess historical data, patterns, and trends to identify potential risk factors and develop predictive models. However, due to the dynamic and unpredictable nature of terrorism, achieving precise predictions is inherently difficult. Ultimately, while predictive tools can contribute to risk assessment, comprehensive counter-terrorism strategies should encompass a multi-faceted approach involving intelligence, international cooperation, and community engagement to enhance overall preparedness and response capabilities.

Our focus in the project is to predict if there will be any casualties in terror attacks in Europe. Global Terrorism Database (GTD) was used in the project which downloaded from Kaggle. The Global Terrorism Database (GTD) is an open-source database including information on terrorist attacks around the world from 1970 through 2017. The GTD includes systematic data on domestic as well as international terrorist incidents that have occurred during this time period and now includes more than 180,000 attacks. The database is maintained by researchers at the National Consortium for the Study of Terrorism and Responses to Terrorism (START), headquartered at the University of Maryland.

Problem Statement

The problem at hand revolves around predicting the likelihood of a terror attack leading to casualties. In the realm of counterterrorism and security, it is crucial to develop a predictive model that can assess the potential impact of a given incident. The objective is to leverage historical data on terrorist attacks, taking into account various factors such as the type of attack, location, target, and time, to create a predictive framework. By analyzing patterns and trends within the data, the aim is to develop a reliable model capable of predicting whether a particular terrorist incident is likely to result in casualties. This predictive capability holds significant importance for preemptive security measures, resource allocation, and emergency response planning, ultimately contributing to the enhancement of public safety and national security.

Metrics

To evaluate the performances of the models we will use following metrics

1. **Accuracy Score**: It is the ratio of correctly predicted observations to the total observations. It is a common metric for classification problems. The formula is as follows

$$Accuracy\ Score = \frac{correctly\ predicted}{total\ observations}$$

2. **Confusion Matrix**: As the name suggests a confusion matrix is a matrix/table that gives simple yet detailed view of model’s performance and summarizes it in a form of a matrix. Following is an example of a confusion matrix

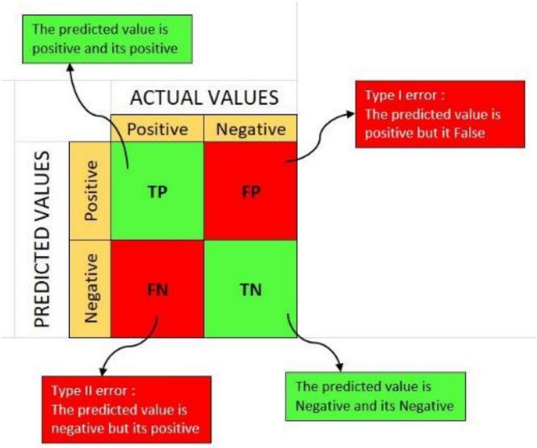
The predicted value is positive and its positive
ACTUAL VALUES
Positive
Negative
Type I error : The predicted value is positive but it False
PREDICTED VALUES
Positive
Negative
TP
FP
FN
TN
Type II error : The predicted value is negative but its positive
The predicted value is Negative and its Negative

Figure 1 Confusion Matrix Example

3. **Precision**: It is the ratio of correctly predicted positive observations to the total predicted positives. It measures the accuracy of the positive predictions.

$$Precision = \frac{True\ Positives}{True\ Positive + False\ Positives}$$

4. **Recall**: It is the ratio of correctly predicted positive observations to the all observations in actual class. It measures the ability of the model to capture all the positive instances.

$$Recall = \frac{True\ Positives}{True\ Positive + False\ Negatives}$$

5. **F1-Score**: It is the harmonic mean of precision and recall. It provides a balance between precision and recall.

$$f1 = 2\ \times \frac{Precision \times Recall}{Precision + Recall}$$

6. **ROC-AUC:** ROC curve is the plot between Recall (Sensitivity) and the FP rate for various threshold values. The area under curve (AUC) is the area under this ROC curve; it is used to measure the quality of a classification model. The larger the area, the better the performance.

Accuracy is a straightforward metric, ROC-AUC and F1-Score are preferred over accuracy because they provide more nuanced insights, particularly when dealing with imbalanced dataset such as our case (as shown in the pie chart there is class imbalance in the dataset 30.39% of the data belong to class 1 and the rest belong to class 0). So, our emphasis is on constructing models that optimizes (maximizes) the ROC-AUC.

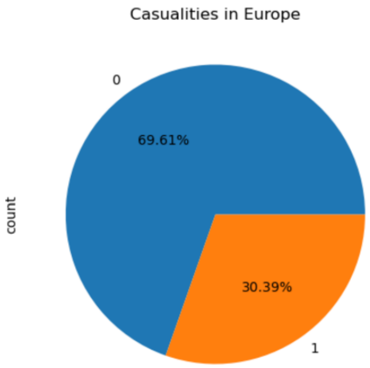

Figure 2 Class Imbalance

Analysis

Data Exploration

The dataset encompasses 135 features with global data. However, our current focus is specifically on Europe. Consequently, we have filtered the dataset to include only the rows corresponding to the European region. Within the 135 features, redundancies exist, such as having two features for each country—one for its name and the other for its code. Additionally, certain information, such as the resolution presented after the attack, is deemed unnecessary for our present analysis. Furthermore, some features exhibit a substantial number of null values, such as the count of US citizens kidnapped. As a result, we are disregarding these irrelevant or incomplete features and directing our attention to key aspects, including the type of attack, hostages involved, location, target, and time. The sole irregularity observed after selecting this subset of the is the absence of data in numerous rows regarding the number of hostages. We infer from this absence that no hostages were involved in those instances and proceed to substitute the null values with zeros.

Data Visualizations

Numerous visualizations have been created. The following is a summary of these visualizations, though due to their abundance, it is impractical to include all of them here (please refer to Project.ipynb for the complete set) here are some visualizations as well.

- Europe experienced its highest number of terror attacks during the years 2014 and 2015. While in 1979, the occurrence of terror attacks reached the third-highest number on record.
- Poland has a relatively low number of terror attacks, with the highest recorded incidence being in 1990, totaling 7 incidents.
- The United Kingdom holds the unfortunate distinction of experiencing the highest number of terror attacks, with Spain ranking second. The disparity in the count is significant, with Spain's incidents representing nearly 50% less than those recorded in the United Kingdom.
- The years 2014 and 2015 witnessed the highest number of terror attacks, with Ukraine being the country most significantly affected during this period.
- Bombing and explosions constitute the most prevalent type of terror attacks.
- The majority of terror attacks, whether occurring in 2014, 2015, or within Poland, lack information about the responsible organization.
- Explosives emerge as the most frequently employed weapon type in terror attacks, whether in 2014, 2015, or within the context of Poland.
- A substantial portion of terror attacks in both Europe and Poland target businesses and private properties.
- In 2014 and 2015, the primary targets of terror attacks were military installations, a consequence of the Russia-Ukraine conflict.
- In general, 85% of terror attacks are reported as successful, and within Poland, a slightly higher rate of 88% is observed in terms of attack success.
- Across the dataset, only 0.6% of these terror attacks involved a suicide perpetrator, and notably, there were no instances of suicide perpetrators within Poland.
- In Europe, 30% of attacks resulted in casualties, while in Poland, a slightly higher percentage of 36%, led to casualties.

- There is generally a lack of correlation among the variables, except for the evident relationship between casualties and the number of kills. Additionally, a moderate correlation is observed between the number of hostages and the number of kills. These insights suggest that, in many cases, the occurrence of casualties is closely related to the number of fatalities, while the number of hostages may also be associated with the severity of the attack.
- It appears that in the majority of cases, there are no casualties associated with various factors such as the type of attack, weapon used, and the target type. However, following are the types where casualties occur more frequently than in other instances.
  - Attack type of Armed Assault and Assassination. This implies that assassination attempts were successful most of the time.
  - Weapon of Firearms
  - When target is police
- The occurrence of successful terror attacks does not always result in casualties. There are instances where attacks deemed successful may not lead to any significant harm or casualties. The success of an attack, in some cases, may be attributed to achieving the intended objectives without causing physical harm or casualties.

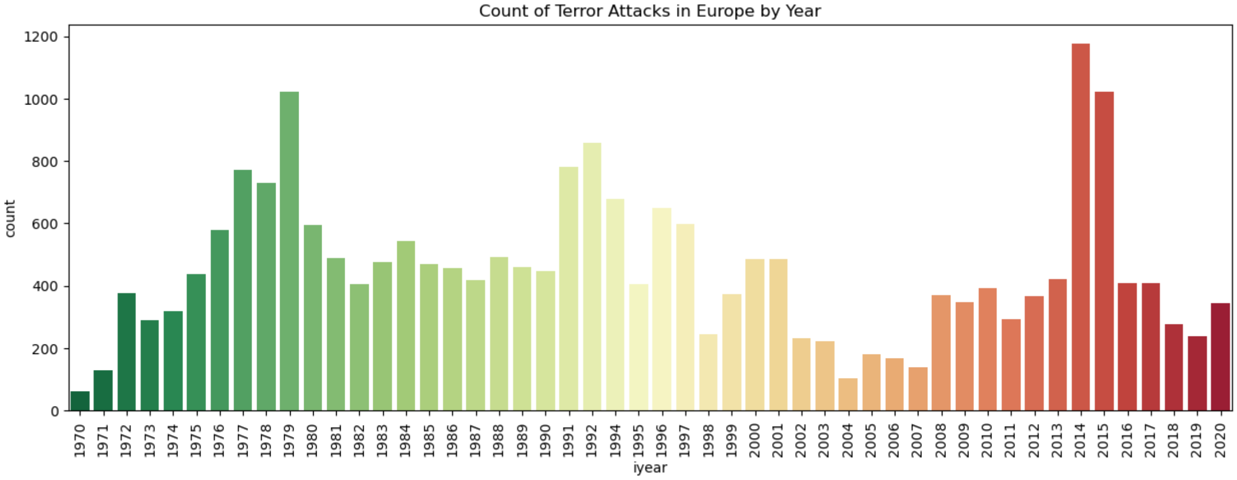

Figure 3 Terror Attacks in Europe over the years

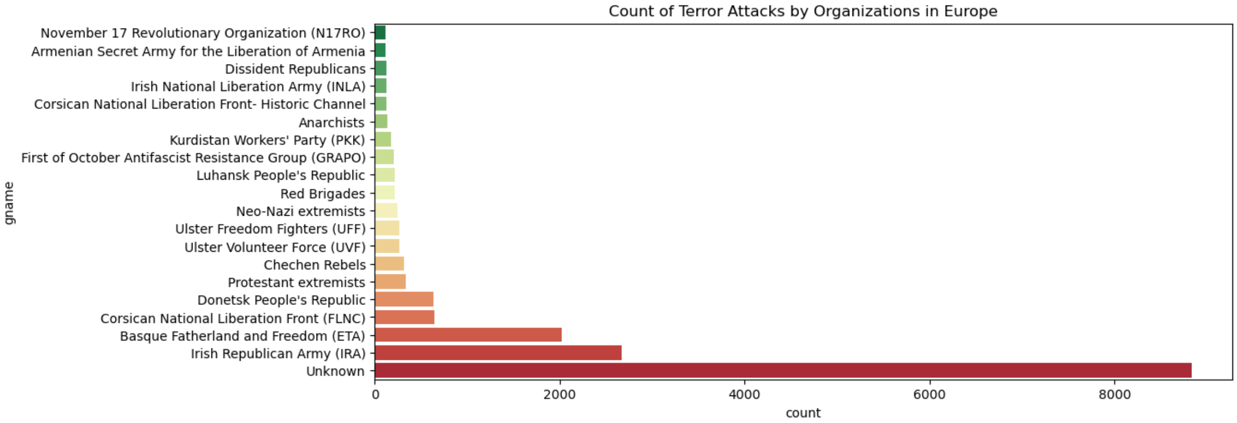

Figure 4 Organizations responsible for terror attacks in Europe

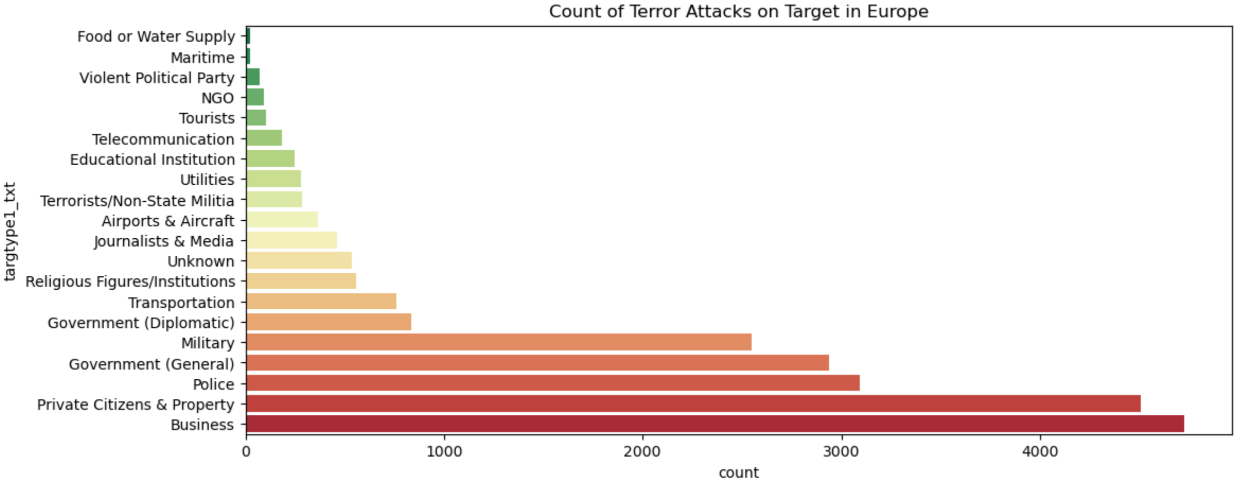

Figure 5 Targets of the terror attacks in Europe

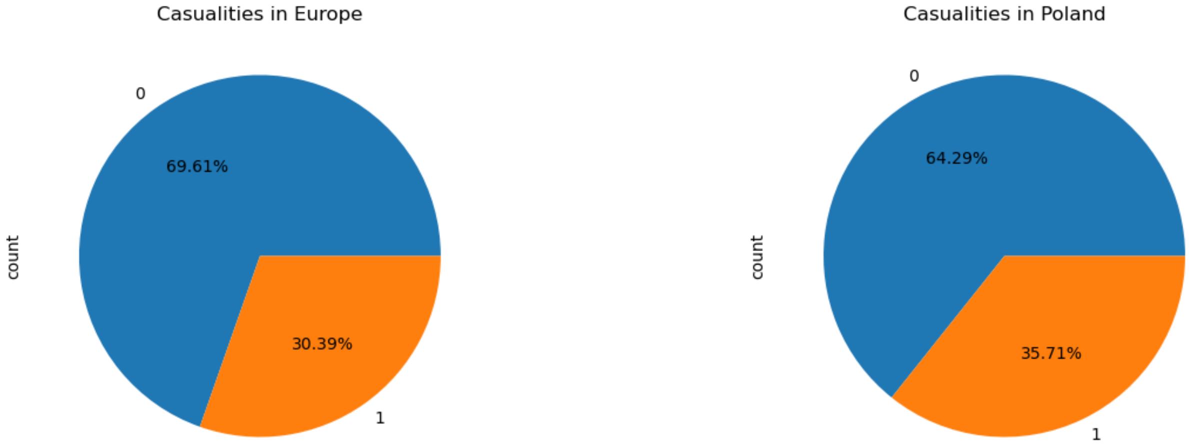

Figure 6 Casualties in Europe and Poland

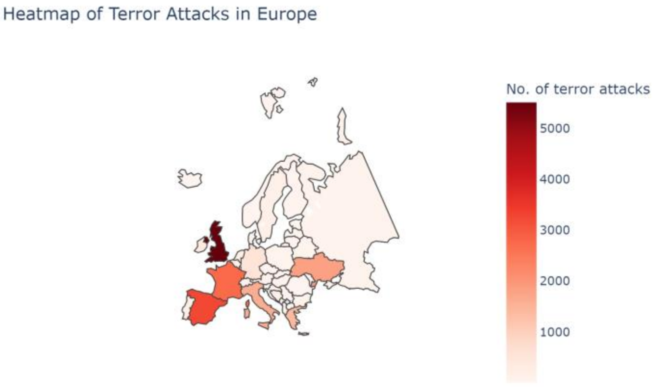

Figure 7 Heatmap of number of attacks in Europe by Country

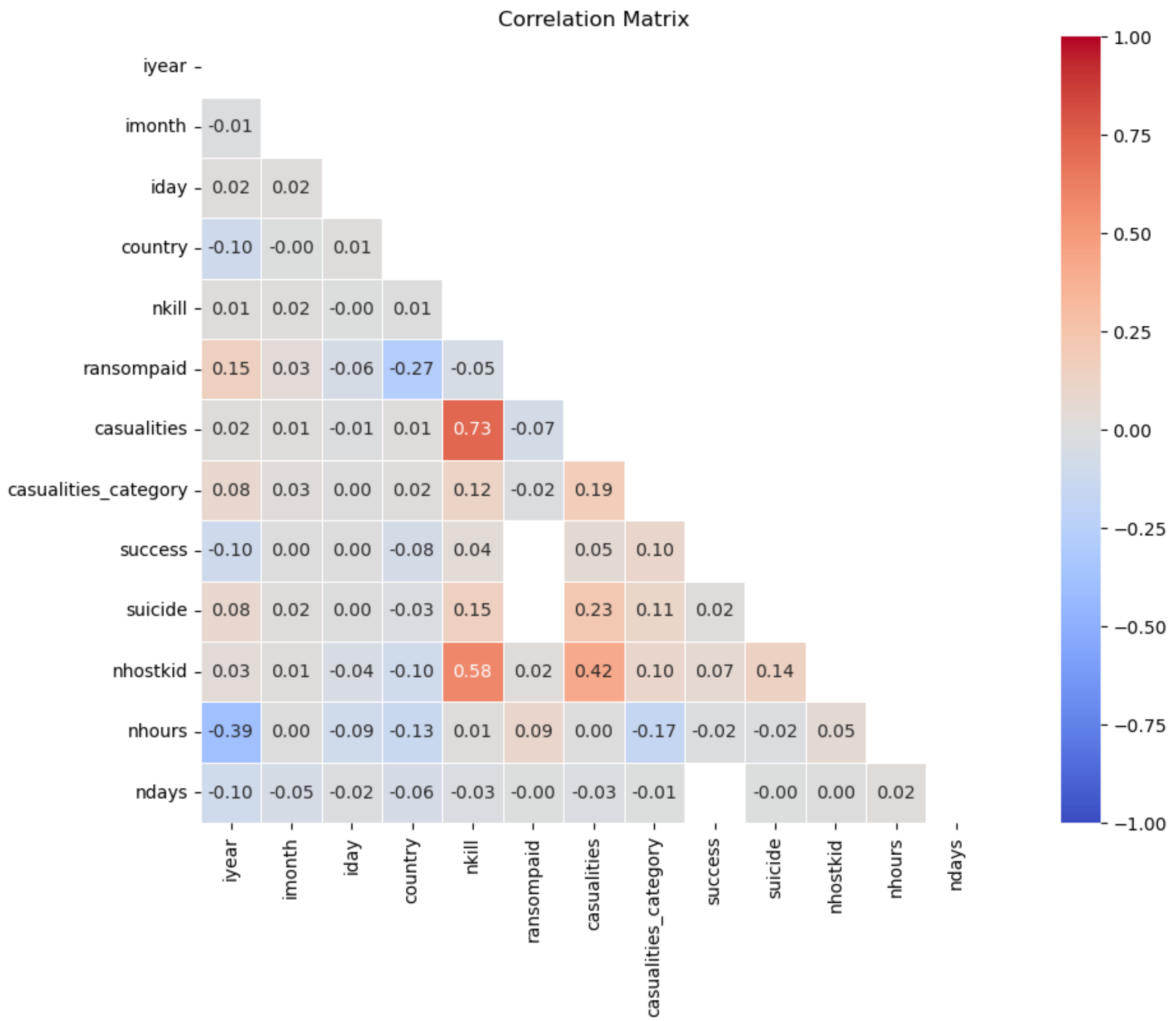

Figure 8 Correlation Matrix

Methodology

Data Preprocessing: The methodology employed for predicting whether a terrorist attack results in casualties involves several key steps. Initially, a comprehensive dataset comprising relevant features such as the type of attack, location, target, and time is gathered. The dataset is then preprocessed to handle missing values, eliminate irrelevant features, and ensure uniformity in representation. Min-Max scaling is utilized to normalize the range of numeric features, ensuring that they fall within a specified interval, typically [0, 1], preserving the relative relationships between values while preventing dominance by variables with larger scales. Subsequently, the dataset is split into training and testing sets to facilitate model evaluation.

Implementation: Four distinct machine learning models—logistic regression, random forest, decision tree, and gradient boosting—are selected for the predictive task. Logistic regression is chosen for its simplicity and interpretability, while random forest and decision tree models are included for their ability to capture complex relationships within the data. Gradient boosting is incorporated to harness the power of ensemble learning and improve predictive performance.

The models undergo a training phase using the training dataset, during which they learn patterns and relationships within the data. Following training, the models are evaluated on the testing dataset to assess their predictive accuracy and generalization performance. The evaluation metrics employed include ROC-AUC, which gauges the models' ability to distinguish between positive and negative instances, and the F1-score, which balances precision and recall.

Refinement: The overall methodology is iterative, involving fine-tuning of hyperparameters to enhance model performance. Grid Search with cross-validation technique is employed for this purpose. Grid Search is a hyperparameter tuning technique that involves systematically searching a predefined set of hyperparameter values to find the optimal parameters for a machine learning model. Hyperparameters are external configurations for a model that are not learned from the data but are set prior to the training process. In a Grid Search, you define a grid of hyperparameter values that you want to explore. The algorithm then trains and evaluates the model for each combination of hyperparameter values on a predefined evaluation metric. The goal is to find the set of hyperparameters that produces the best performance on the validation set or during cross-validation.

Results

As described earlier, 4 models are used and each one in fine-tuned using their various hyper-parameters.

Model	Hyper-Parameters
Logistic Regression	C, penalty: The regularization parameters C and penalty in logistic regression are crucial for controlling overfitting (or underfitting) by balancing the influence of individual features and preventing excessive complexity in the model.
Decision Tree	Max-Depth, Criterion: The max-depth parameter in a decision tree is important as it determines the maximum depth of the tree, controlling its complexity and preventing overfitting, while the criterion parameter defines the measure used for splitting nodes, influencing the tree's decision-making process.
Random Forest	Max-Depth, No. of Estimators: The max-depth parameter controls the maximum depth of individual decision trees within the ensemble, influencing the model's complexity and potential for overfitting. The number of estimators determines the total number of trees in the Random Forest, impacting the model's overall predictive power and generalization performance.
Gradient Boosting	No. of estimators, Subsample, Max-Features: Subsample controls the fraction of samples used for each boosting iteration, influencing variance and preventing overfitting. Max-features regulates the number of features considered for each split, enhancing model robustness, while the number of estimators determines the total number of boosting stages, impacting the overall predictive performance.

Following is the summary of the results

Model	Accuracy	ROC-AUC	F1-Score
Decision Tree	0.784	0.769	0.726
Logistic Regression Tuned	0.752	0.801	0.681
Logistic Regression	0.753	0.801	0.683
Random Forest	0.772	0.812	0.659
Decision Tree Tuned	0.796	0.821	0.735

Gradient Boosting	0.799	0.835	0.739
Gradient Boosting Tuned	0.803	0.848	0.737
Random Forest Tuned	0.807	0.848	0.737

All the models performed very closely where the difference between highest and lowest performing models' ROC-AUC scores difference is 0.079. Random Forest after fine-tuning demonstrated superior performance, achieving the highest recall and ROC-AUC score of 0.85. Here is the confusion matrix and ROC curve for that model. All models have slightly lower recall scores. The recall scores for all models are slightly lower, indicating that the models may not be highly accurate in identifying instances when casualties occur.

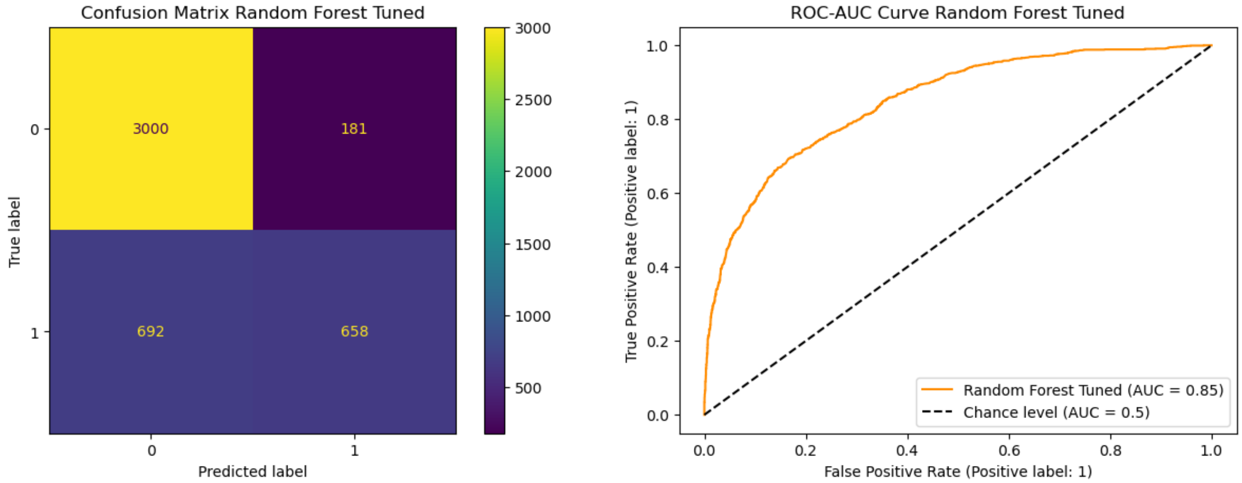

Figure 9 Performance of Fine-Tuned Random Forest

Conclusion

In conclusion, our predictive modeling efforts to determine the likelihood of casualties in terror attacks have yielded insightful results. The comprehensive analysis employing logistic regression, random forest, decision tree, and gradient boosting revealed distinct patterns and predictive strengths across the models. Notably, the exploration of key features such as attack type, location, target, and time proved crucial in understanding the dynamics influencing the outcomes of these incidents. While the selected models exhibited commendable performance but they performed slightly poor on predicting when there are casualties i.e., poor recall score, future improvements could involve refining feature engineering, exploring more advanced algorithms, and integrating data of all the regions for enhanced predictive recall. Additionally, a deeper examination into the geopolitical context and sociodemographic factors could contribute to a more nuanced understanding of the complexities associated with predicting casualties in terror attacks. These findings provide a solid foundation for ongoing research and the continual refinement of models to bolster counterterrorism efforts and public safety measures.