**Shooting Fatality Classification: A Machine Learning Approach**

**Project Title: Shooting Fatality Prediction using Gradient Boosting Model**

**1. Introduction**

**1.1 Project Overview**

This project aims to develop a machine learning model to classify shooting incidents as either fatal or non-fatal based on a dataset of features including demographic information, incident details, and location-related data. The deliverables include:

* A trained and validated machine learning model.
* A comprehensive report on the model's development process, performance, and limitations.
* Visualizations and insights to explain factors influencing fatality rates.

**1.2 Problem Statement**

In urban areas, gun violence incidents vary widely in their outcomes. Identifying factors that contribute to the fatality of shootings can help allocate resources more effectively and inform public safety strategies. This project addresses the challenge of predicting whether a given shooting incident will result in a fatality, leveraging various demographic and situational features.

**1.3 Approach Overview**

The approach involves data preprocessing, model selection (Gradient Boosting), hyperparameter tuning, evaluation, and interpretability through feature importance analysis. The final model prioritizes identifying fatal incidents with high recall, ensuring fewer fatal incidents are missed.

**2. Data Understanding and Preprocessing**

**2.1 Dataset Description**

The dataset comprises thousands of shooting incidents with the following features:

* **Demographic Information**: Victim’s and perpetrator’s age, race, sex, etc.
* **Incident Details**: Year of occurrence, location description, and jurisdiction.
* **Outcome**: A binary target variable indicating if the incident was fatal (True) or non-fatal (False).

**2.2 Data Preprocessing**

* **Handling Missing Data**: Missing values were imputed or flagged as "Unknown" to retain useful rows.
* **Encoding Categorical Variables**: Categorical features such as race, age group, and location descriptions were encoded using one-hot encoding.
* **Feature Scaling**: No scaling was needed since tree-based models like Gradient Boosting are unaffected by feature scaling.

**2.3 Feature Selection**

Feature selection focused on keeping the most relevant variables. For instance, **age groups** of both victim and perpetrator were important factors, while redundant or overly granular features were removed.

**3. Model Development**

**3.1 Initial Model Selection**

Various models were considered including:

* Logistic Regression
* Random Forest
* Gradient Boosting

Gradient Boosting was selected for its superior handling of imbalanced datasets and high interpretability, allowing for feature importance analysis.

**3.2 Model Training**

The dataset was split into 80% training and 20% testing. Cross-validation and grid search were used to fine-tune hyperparameters such as learning rate and the number of estimators.

**3.3 Threshold Selection**

The classification threshold was set at **0.35** (lower than the default 0.5), optimizing for **higher recall** on fatal incidents. This decision ensures that more fatal cases are captured, even if it results in a few more false positives.

**3.4 Final Model**

The final model is a **Gradient Boosting Classifier** with tuned hyperparameters:

* **Learning rate**: 0.01
* **Number of estimators**: 500
* **Max depth**: 3
* **Min samples split**: 100

**4. Model Performance and Evaluation**

**4.1 Confusion Matrix**

The confusion matrix for the model (with the threshold at 0.35) shows:

|  | **Predicted non-fatal** | **Predicted Fatal** |
| --- | --- | --- |
| **Actual non-fatal** | 3607 | 3003 |
| **Actual Fatal** | 529 | 1051 |

The model captured 67% of fatal incidents (recall) but at a trade-off with lower precision (26%).

**4.2 Classification Metrics**

| **Metric** | **Value** |
| --- | --- |
| **Precision** | 0.26 |
| **Recall** | 0.67 |
| **F1-Score** | 0.37 |
| **Accuracy** | 57% |
| **ROC AUC** | 0.66 |

**Interpretation**: The model prioritizes **recall** over precision to capture more fatal incidents, accepting more false positives.

**4.3 ROC and Precision-Recall Curves**

* **ROC Curve**: The model’s AUC of 0.66 indicates it moderately distinguishes between fatal and non-fatal incidents.
* **Precision-Recall Curve**: Precision drops significantly as recall increases, justifying the use of a lower classification threshold to maximize recall.

**5. Feature Importance and Interpretability**

**5.1 Feature Importance**

The following chart displays the top 20 features driving the model’s predictions:

**5.2 SHAP Values**

The SHAP summary plot provides an interpretable view of how features impact individual predictions:

**Insights:**

* **Victim and Perpetrator Age Groups (25-44)**: Major contributors to predicting fatal outcomes.
* **Location Description**: Incidents in multi-dwelling public housing or apartments have a higher likelihood of fatality.
* **Race**: The race of both victim and perpetrator also significantly affects the model’s predictions.

**6. Limitations**

**6.1 Model Limitations**

* **Class Imbalance**: The model struggles with precision for fatal incidents due to the smaller number of fatal cases relative to non-fatal ones.
* **Bias Risk**: Features like race may introduce unintended biases, necessitating careful interpretation.
* **Limited Features**: The dataset lacks some potentially significant features like weapon type or response time, limiting the model's performance.

**6.2 Data Limitations**

* **Incomplete Data**: Some fields (e.g., race, age) have missing or unknown values.
* **Geographic Scope**: The model is limited to the dataset's specific geographic region, making it less generalizable to other areas.

**7. Conclusions and Recommendations**

**7.1 Key Findings**

* **Age** and **location** are key drivers in the likelihood of shooting fatalities.
* **Public housing** locations and certain **boroughs** (e.g., Brooklyn) show higher fatality rates.

**7.2 Recommendations**

* **Policy Interventions**: Focus on high-risk age groups and locations for preventive actions.
* **Resource Allocation**: Law enforcement can use this model to allocate resources in areas with higher fatality risks.
* **Future Research**: Expanding the dataset to include other critical factors (e.g., time, weapon) would improve the model.