1. Algorithms Selected with Justification

Three models were chosen during the modeling process.

- **Logistic Regression:** Because of its simple interpretability and speed, logistic regression was selected as the baseline model. Its performance creates a reliable, and understandable standard that more complex models must show they can surpass.
- **Random Forest:** Because we wanted to test a model that was non-linear and strong, we chose Random Forest. It can capture complex feature interactions that a linear model might overlook since it can create various decision trees on random subsets of data.
- **XGBoost:** Extreme Gradient Boosting, or XGBoost is a gradient boosting technique that is well known for producing excellent results on datasets. It's a strong contender for our best performing model since it creates trees sequentially, fixing the mistakes of the ones that came before it.

2. Data Restructuring and Validation Strategy

To guarantee that the model was appropriately trained and assessed, an extra process was made.

- **Data Restructuring to Prevent Target Leakage:** Our initial models showed unrealistically high performance due to a problem called **target leakage.** This happened because the model was **"cheating"** by using features from an incident to predict the outcome of that very same incident. To build a genuinely predictive model, we needed to use the past information to predict a future event. We solved this by restructuring the data for each victim, pairing the features of a prior crime with the outcome of their next crime. This shifted outcome was stored in our new target column, **"Next_Incident_Is_Severe"**, creating a "past -> future" structure for the model to learn from.
- **Train-Test Split:** The final, restructured dataset was split into a **training set (70%)** and a **test set (30%)**. Stratified sampling (stratify=y) was used to make sure the number of severe assaults was identical in both the training and testing sets, which is a good practice for imbalanced data.

3. Baseline Comparisons (After Leakage Solved)

- **Logistic Regression (Baseline)**
  - Precision: 0.454
  - Recall: 0.710
  - F1-Score 0.554

Interpretation: The baseline model is strong for its recall. It successfully identifies **71%** of all actual severe assaults. However, to achieve this, it has a high rate of false alarms; only **45%** of

its positive predictions are correct. This model is good at finding potential cases but is not very efficient.

- **Random Forest**
    - Precision: 0.451
    - Recall 0.507
    - F1-Score: 0.477

Interpretation: The Random Forest model performed the worst of the three. It struggled to find a good balance, resulting in lower recall than the other models. This suggests that its more complex, non-linear approach may have had difficulty finding a clear signal in the restructured data without further tuning.

- **XGBoost (The Winner So Far)**
    - Precision: 0.465
    - Recall: 0.706
    - F1-Score: 0.561

Interpretation: XGBoost is the best-performing baseline model. It achieved the highest **F1-Score (0.561)**, indicating the best overall balance between precision and recall. Like the logistic regression model, it is optimized for high recall, successfully identifying over **70%** of severe assaults, but it does so with slightly higher precision, making it more efficient and reliable. This result confirms that a victim's prior history contains a meaningful and predictive signal, and XGBoost is the most effective algorithm at capturing it.

4. Hyperparameter Tuning Approach

To optimize the performance of the best-performing model (XG-Boost), a tuning process was used.

- **Method: RandomizedSearchCV** was used to efficiently search through a wide range of potential hyperparameter combinations. This was performed on a 20% sample of the training data to balance thoroughness with computational speed.
- **Validation:** Within the search, a **3-fold StratifiedKFold** cross validation was used. This provided a reliable estimate of each parameter combination's performance by training and testing it on three different subsets of the data.
- **Metric:** The search was created to optimize the **F1-score**, which is a balanced measure of precision and recall and is more suitable for this project's imbalanced classes than accuracy alone.

5. Model Performance Comparison

The table below tracks the performance of the models at each stage of our analysis, focusing on the metrics for predicting whether the next is severe assault (Class 1).

Model Stage	Model Algorithm	Precision	Recall	F1-Score	Notes
1. Initial Test	XGBoost	0.980	0.990	0.980	**Results invalid** due to Target Leakage
2. Leakage-Free Baseline	Logistic Regression	0.454	0.710	0.554	Leakage fixed, realistic baseline
3. Leakage-Free Baseline	Random Forest	0.451	0.507	0.477	Leakage fixed, underperformed
4. Leakage-Free Baseline	**XGBoost**	0.465	0.706	0.561	Leakage fixed, **best baseline model**
5. Final Tuned Model	**Tuned XGBoost**	**0.472**	**0.730**	**0.573**	**Final, Optimized Model**

The Modeling Process

1. **The Impact of Fixing Target Leakage:** The first and most important finding was the performance drop between stage 1 and 2 in the table above.The initial F1-score was a near-perfect 0.980 due to target leakage. After restructuring the data to correctly use past incidents to predict future ones, the scores dropped to a more realistic F1-score of about **0.56**. This difference illustrates the importance of identifying and resolving data leakage to build a truly predictive model.
2. **Finding the Best Model:** After fixing the leakage, a direct comparison of the baseline models (Stage 2, 3, and 4) showed that **XGBoost was clearly the best**, outperforming both Logistic Regression and Random Forest with a balanced F1-Score of 0.561.
3. **Optimization:** The final step, hyperparameter tuning, successfully improved the best model (Stage 5). The final, tuned XGBoost model achieved an **F1-Score of 0.573.** The most significant improvement was in **recall,** which increased from 70.6% to **73.0%**, meaning the final model is better at correctly identifying victims at risk of a future severe assault.