# CAUSES OF ROAD ACCIDENTS IN CHICAGO

Predictive Model for Accident Causes

GROUP 5

# OVERVIEW

This project aims to conduct a thorough analysis of traffic crashes in the city of Chicago, utilizing multiple modeling techniques. The primary goal is to uncover insights into the factors influencing traffic accidents in Chicago, and suggest proactive measures to reduce traffic accidents.

# STAKEHOLDERS

1. **Chicago City planners and traffic engineers:** interested in reducing traffic accidents.

2. **Chicago City Vehicle Safety Board**: interested in becoming aware of any interesting patterns.

# OBJECTIVES

1. Identify the most significant factors contributing to road accidents.

2. Provide insights to help stakeholders implement targeted interventions.

3. Develop a predictive model for road accidents.

# DATASETS

**1. Crashes dataset:**
This dataset contains information about each traffic crash within the City of Chicago limits and under the jurisdiction of Chicago Police Department (CPD).

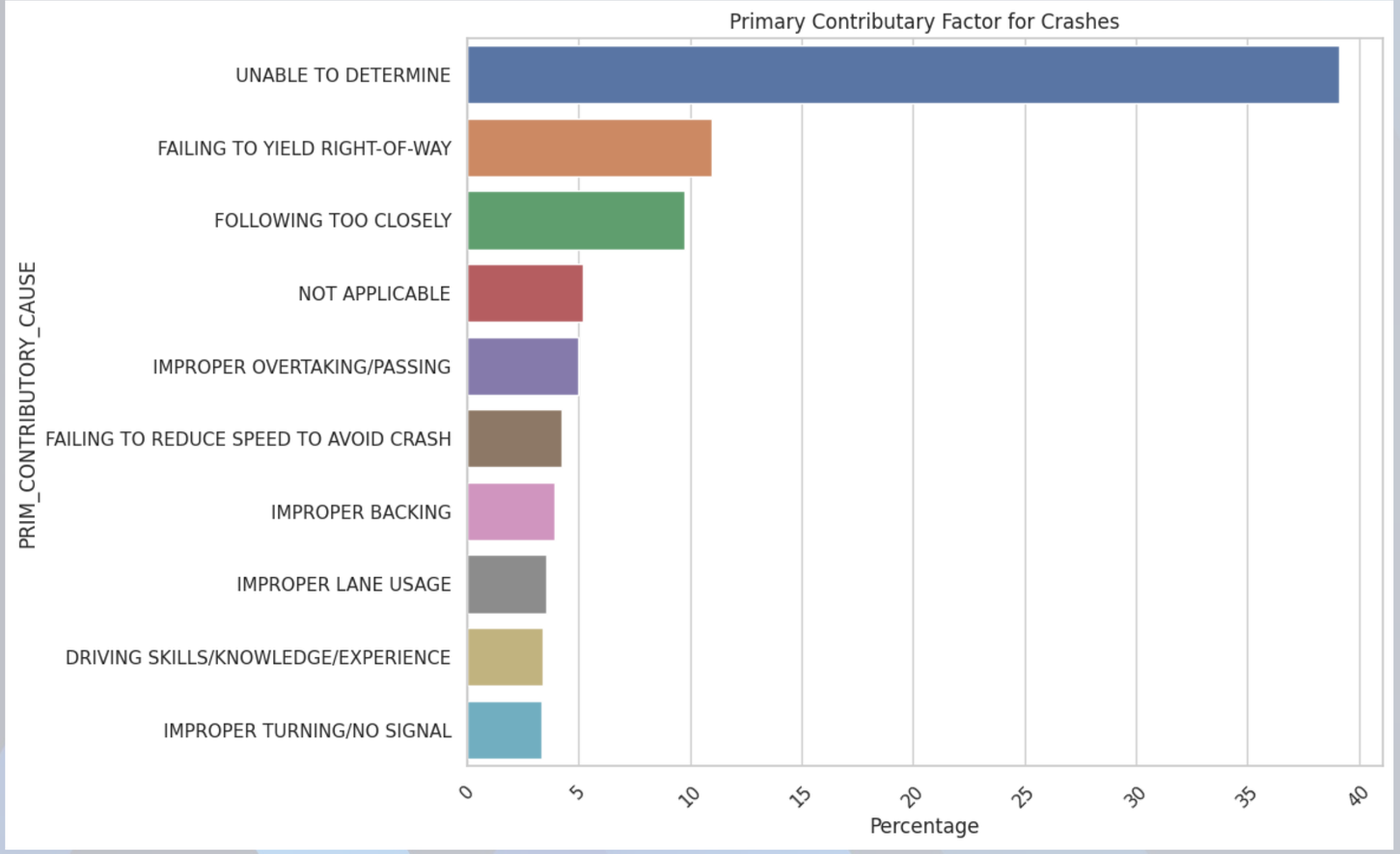
**2. Vehicles dataset:**
This dataset contains information about vehicles (or units as they are identified in crash reports) involved in a traffic crash.

**3. People dataset:**
This dataset contains information about people involved in a crash and if any injuries were sustained. Each record corresponds to an occupant in a vehicle listed in the Crash dataset.

# EXPLORATORY DATA ANALYSIS

The Chicago Traffic Crashes dataset consists of around 880,000 rows and 48 columns. This dataset captures various details about road accidents, such as the date of the crash, weather and lighting conditions, types of vehicles, and details about injuries sustained by people involved.

From the initial data analysis, we identified the most significant contributing factors to road accidents in Chicago by exploring the primary causes as well as the secondary causes.




# Primary Causes

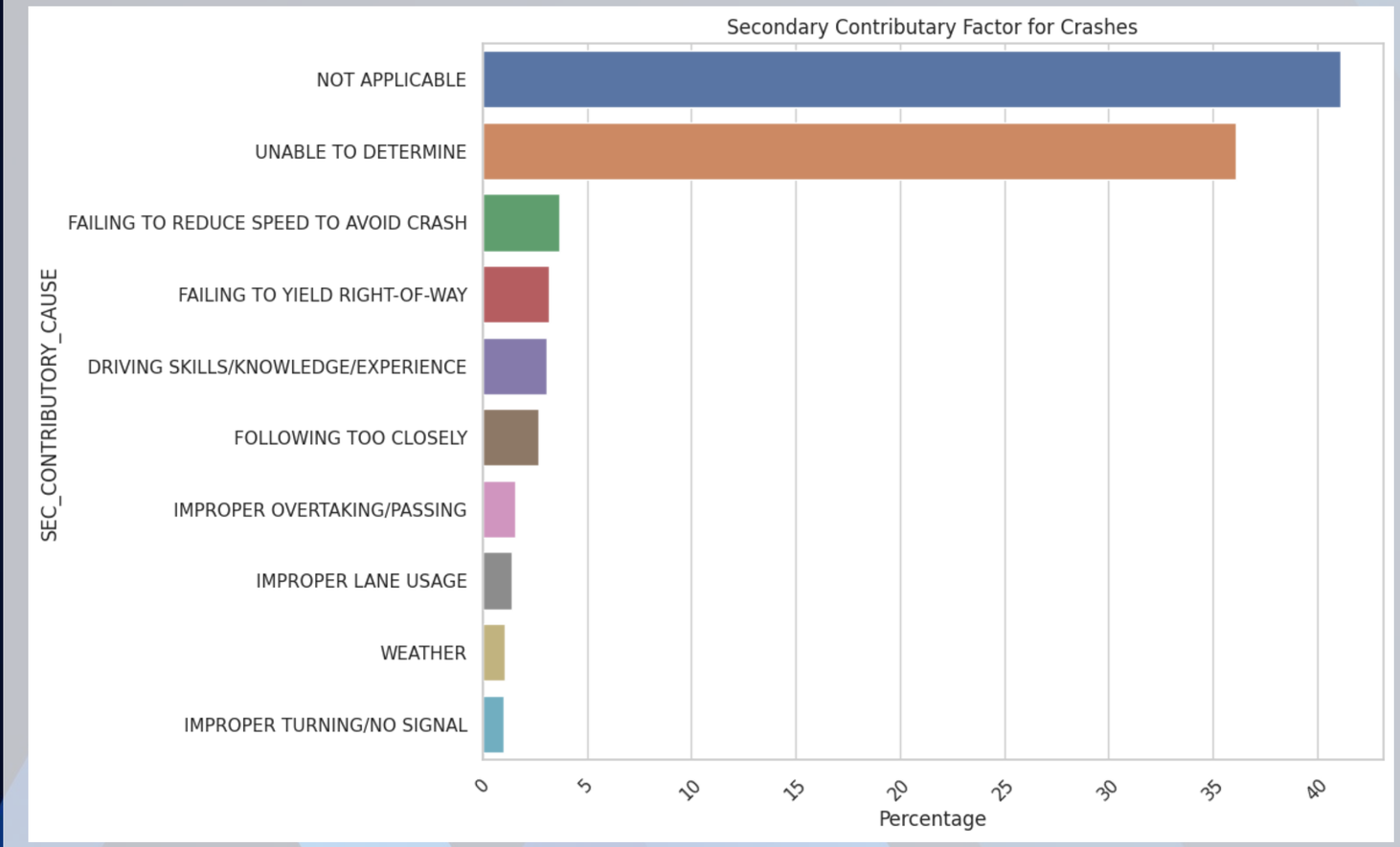
The horizontal barplot shows the top 10 primary causes of traffic crashes as reported by the officer:

39% of the traffic crashes reported, the reporting officer was unable to determine cause and 5% of the traffic crashes primary contributary factors were recorded as NOT APPLICABLE.

- Other factors, ranked by decreasing percentage, include:
  1. Failing to Yield Right-of-Way
  2. Following Too Closely
  3. Not Applicable
  4. Improper Overtaking/Passing
  5. Failing to Reduce Speed to Avoid Crash
  6. Improper Backing
  7. Improper Lane Usage
  8. Driving Skills/Knowledge/Experience
  9. Improper Turning/No Signal

* Factors like "Failing to Yield Right-of-Way" and "Following Too Closely" are significant contributors to crashes.

# Secondary Causes




The horizontal barplot shows the top 10 secondary causes of traffic crashes as reported by the officer:

36% of the traffic crashes reported, the reporting officer was unable to determine cause and 41% of the traffic crashes primary contributary factors were recorded as `NOT APPLICABLE`.
Factors like DRIVING SKILLS/KNOWLEDGE/EXPERIENCE, FAILING TO YIELD RIGHT-OF-WAY and FAILING TO REDUCE SPEED TO AVOID CRASH, are the highest secondary contributors to crashes as reported by the officer, though they're significantly low.

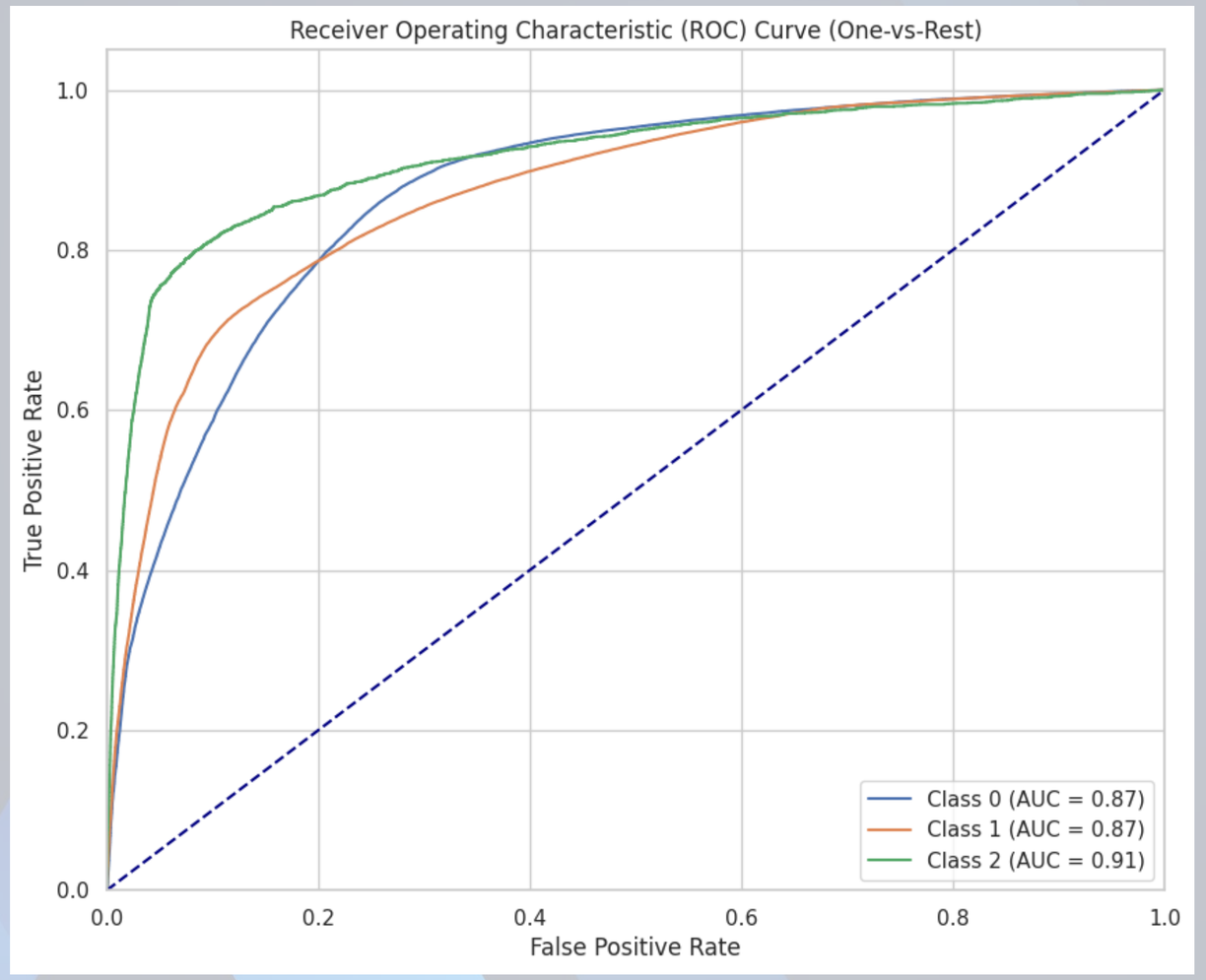
# MODELLING

A number of machine learning models were explored to predict the primary contributory causes of road accidents based on the features available in the dataset.
The goal was to create a model that could help identify patterns and risk factors, ultimately guiding interventions to reduce accidents.

# Logistic Regression

- The logistic regression model performed reasonably well, achieving an accuracy of **76.2%**. This indicates good generalization to unseen data without significant overfitting.
- **Precision (80.1%)**: The model was correct 80.1% of the time when predicting a specific accident cause, demonstrating effective minimization of false positives.
- **Recall (76.2%)**: The model successfully identified 76.2% of all actual cases, indicating competence in capturing relevant accident causes while leaving room for improvement in identifying true positives.
- **F1-Score (77.7%)**: With an F1-score of 77.7%, the model balances precision and recall, reflecting overall effectiveness but suggesting that some non-linear relationships may not be fully captured.

# ROC CURVE


Receiver Operating Characteristic (ROC) Curve (One-vs-Rest)
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Class 0 (AUC = 0.87)
Class 1 (AUC = 0.87)
Class 2 (AUC = 0.91)

From the ROC Curve,

- **Class 0 - Driver Error (AUC = 0.87):**

> The ROC curve for Class 0 is depicted in blue. (AUC) of 0.87 indicates that the classifier is fairly good at distinguishing Class 0 from the other classes. An AUC of 0.87 suggests that there's an 87% chance that the model will correctly rank a random positive instance from Class 0.

- **Class 1 - Environmental Factors / Other (AUC = 0.87):**

> The ROC curve for Class 1 is shown in orange. AUC of 0.87, indicates a performance similar to Class 0. This suggests the classifier is performing equally well in distinguishing Class 1 from the rest of the classes as it does for Class 0.

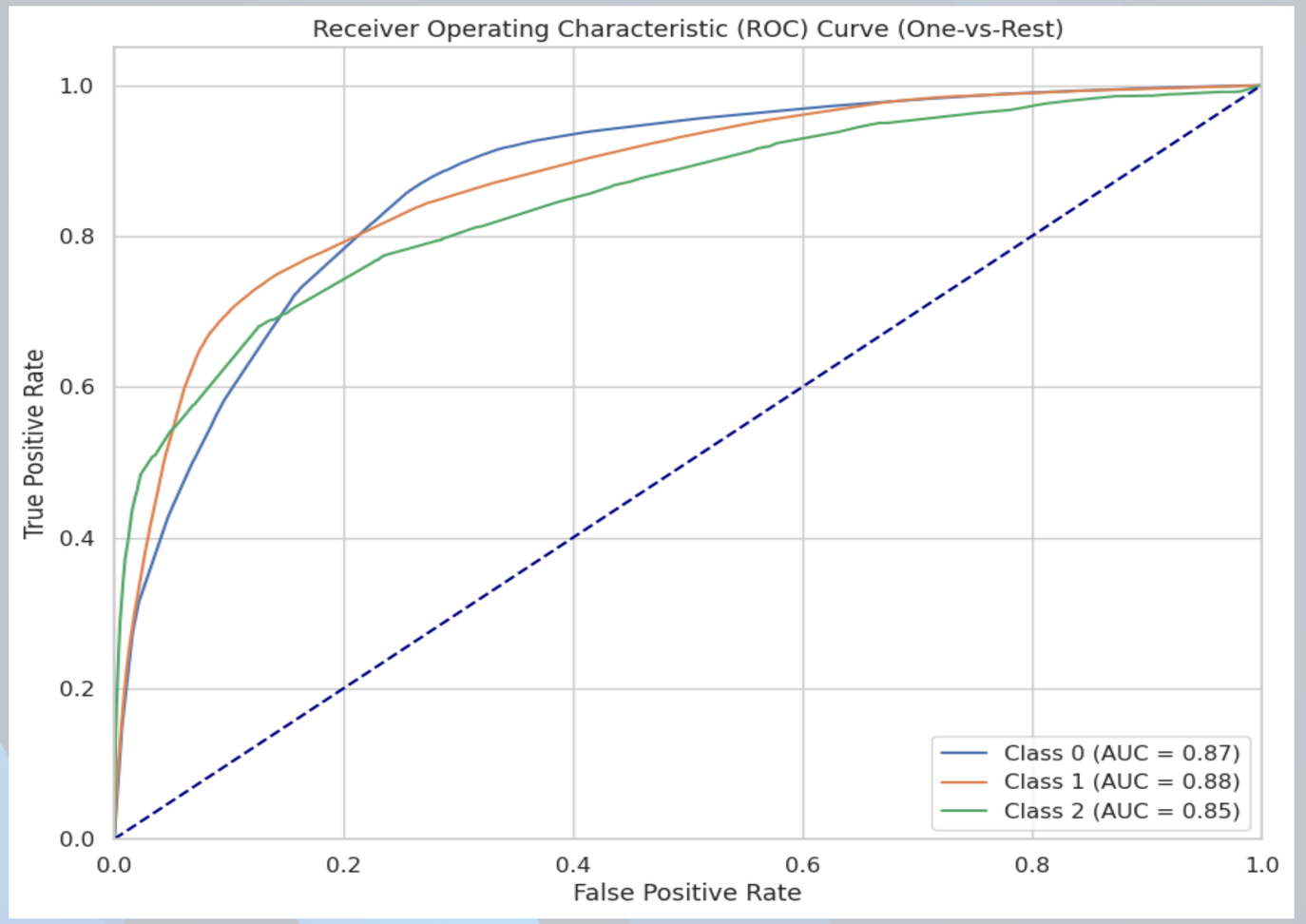
- **Class 2 - Vehicle Conditions (AUC = 0.91):**

> The ROC curve for Class 2 is displayed in green. AUC is 0.91 is slightly higher than the other two classes. This means that the classifier performs the best for Class 2, with a 91% chance of correctly distinguishing instances of Class 2 from other classes.

# Decision Tree Classifier

- The decision tree classifier exhibited a notable improvement post-tuning, achieving an accuracy of **80.0%**. This indicates effective classification of accident causes.
- **Precision (80.0%)**: The model was correct 80.0% of the time when predicting a cause, showing strong performance in minimizing false positives..
- **Recall (80.0%)**: The model captured 80.0% of all actual cases, indicating proficiency in identifying true instances while suggesting some potential for missed classifications.
- **F1-Score (79.0%)**: An F1-score of 79.0% reflects a good balance between precision and recall, highlighting the model's effectiveness but indicating room for further optimization.

# ROC CURVE


Receiver Operating Characteristic (ROC) Curve (One-vs-Rest)
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Class 0 (AUC = 0.87)
Class 1 (AUC = 0.88)
Class 2 (AUC = 0.85)

From the ROC Curve,

- **Class 0 - Driver Error (AUC = 0.87):**

> The ROC curve for Class 0 is depicted in blue. (AUC) increased from 0.71 to 0.87 with hyperparameter tuning, whish indicates a good performance in distinguishing Class 0 from the other classes. This means that the model is 87% likely to rank a randomly chosen positive instance of Class 0 higher than a negative one.

- **Class 1 - Environmental Factors / Other (AUC = 0.88):**

> The ROC curve for Class 1 is shown in orange. AUC increased from 0.71 to 0.88 with hyperparameter tuning, which shows that the model performs very well in identifying Class 1. An AUC of 0.88 signifies a strong classifier for this class.

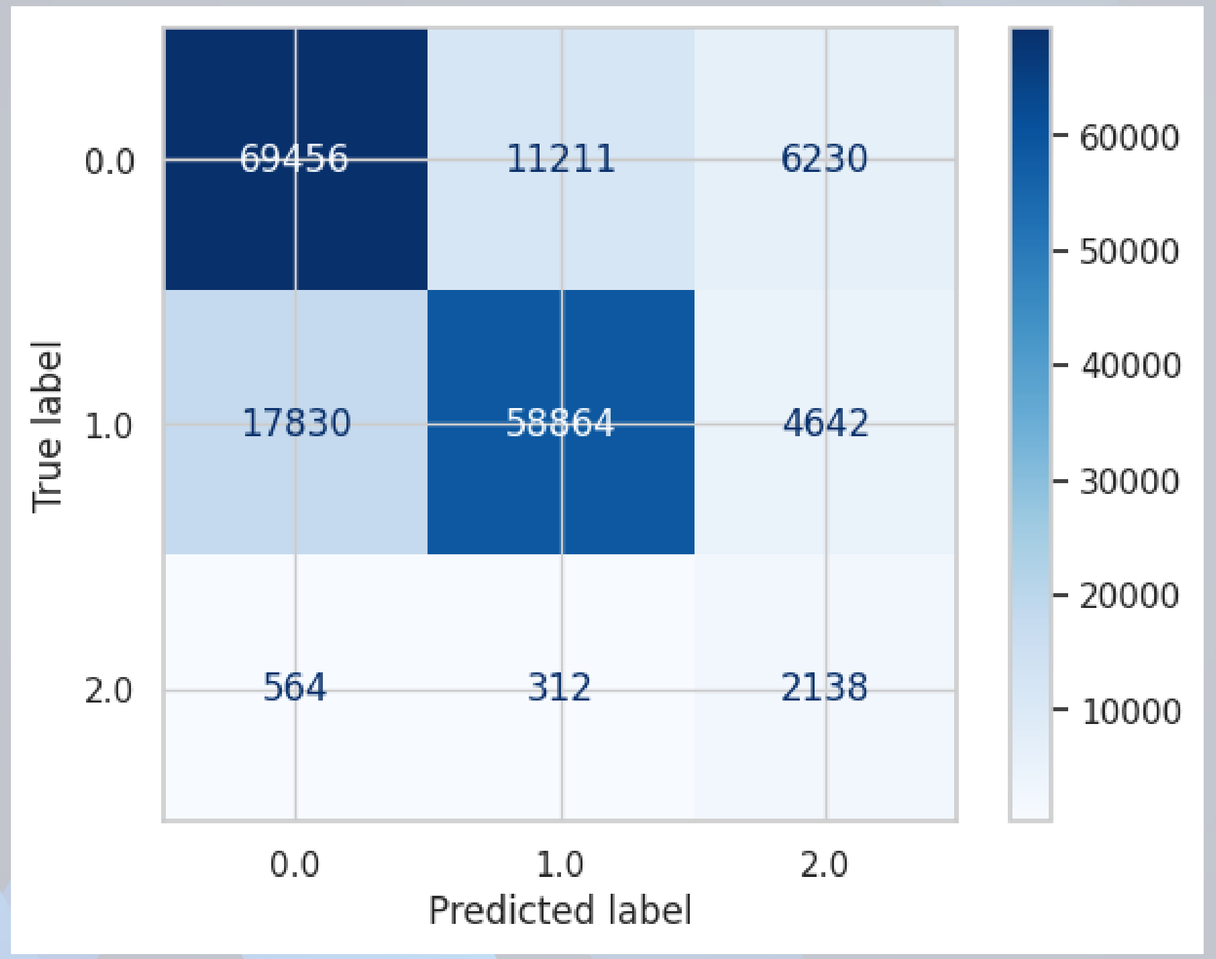
- **Class 2 - Vehicle Conditions (AUC = 0.85):**

The ROC curve for Class 2 is displayed in green. AUC increased from is 0.63 to 0.85 indicates that the classifier is also performing well for Class 2, though slightly worse than for Class 0 and Class 1.

# Random Forest Classifier

- The random forest classifier achieved an overall accuracy of **76.2%**, showcasing a solid performance while effectively managing overfitting through its ensemble learning method.
- **Precision (80.1%)**: The model's precision of 80.1% indicates that it was correct when predicting a specific cause 80.1% of the time, effectively minimizing false positives.
- **Recall (76.2%)**: With a recall of 76.2%, the model successfully identified a significant portion of actual cases but also suggests some potential for missed instances.
- **F1-Score (77.7%)**: The F1-score of 77.7% indicates a well-balanced performance, reflecting overall effectiveness in identifying accident causes while leaving room for improvement.

# CONFUSION MATRIX



From the Confusion Matrix,

- **Class 0.0:**

> True Positives (Correctly Predicted as 0.0): 69,456

> False Positives (Predicted as 0.0 but True Label was 1.0 or 2.0): Not directly shown but can be inferred.

> False Negatives (True Label is 0.0 but Predicted as 1.0 or 2.0): 11,211 + 6,230 = 17,441

- **Class 1.0:**

> True Positives: 58,864

> False Negatives: 17,830 (True Label is 1.0 but Predicted as 0.0) + 4,642 (Predicted as 2.0) = 22,472

> False Positives: 11,211 (Predicted as 1.0 but True Label was 0.0) + 312 (True Label was 2.0) = 11,523

- **Class 2.0:**

> True Positives: 2,138

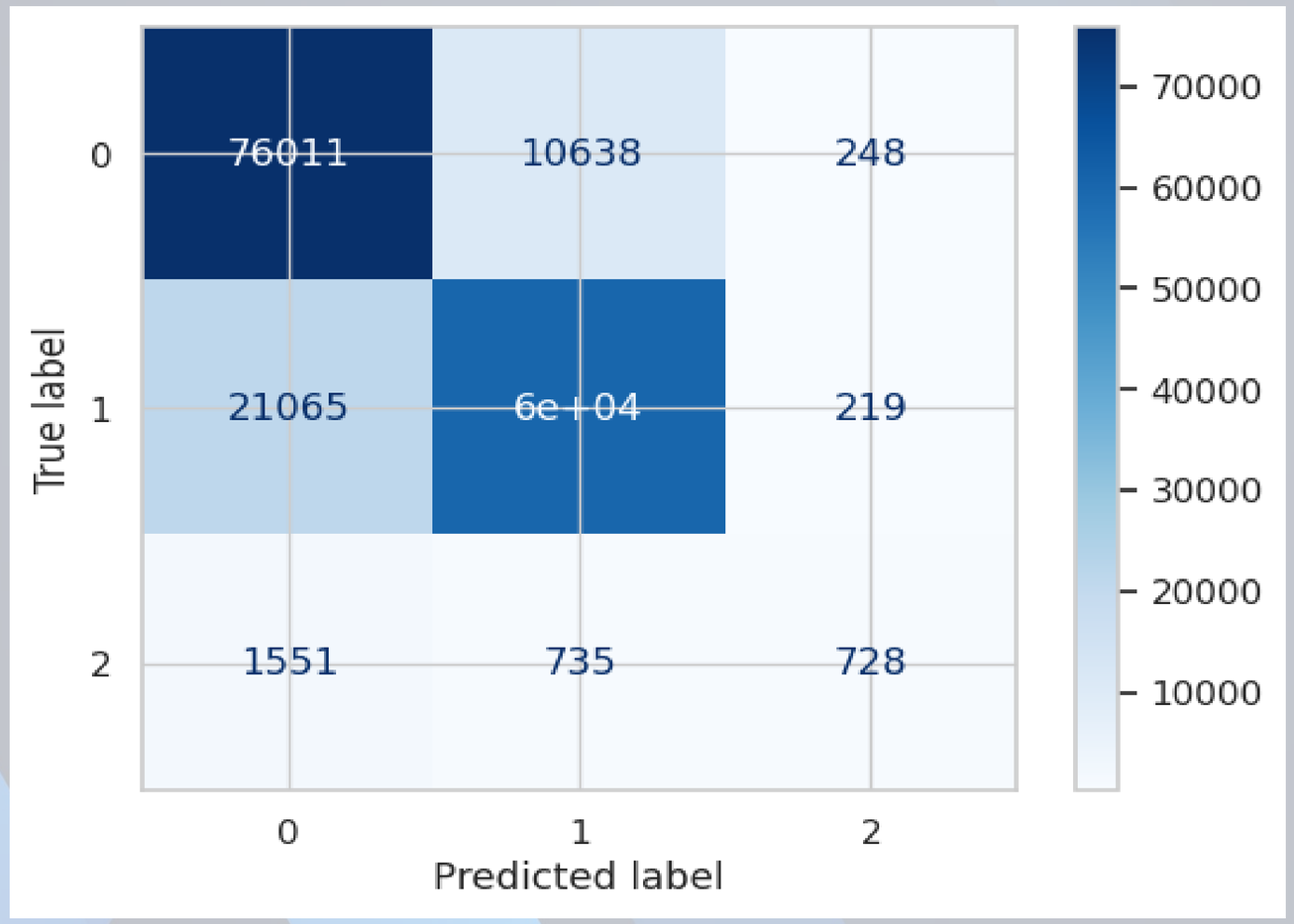
> False Negatives: 564 + 312 = 876

> False Positives: 6,230 (True Label was 0.0) + 4,642 (True Label was 1.0) = 10,872

# XGBoost Classifier

- The XGBoost classifier demonstrated strong performance with an overall accuracy of **79.9%**, indicating its effectiveness in generalizing to new data through advanced boosting techniques.
- **Precision (80.1%)**: With a precision of 80.1%, the model was correct when predicting a specific cause 80.1% of the time, effectively minimizing false positives..
- **Recall (79.9%)**: The recall score of 79.9% indicates that the model successfully captured a substantial number of true instances but still has room for improvement in reducing false negatives.
- **F1-Score (79.5%)**: The F1-score of 79.5% reflects a balanced performance, signifying that the model effectively identifies true positives while maintaining a low rate of misclassifications.

# CONFUSION MATRIX



From the Confusion Matrix,

- **Class 0:**

  > True Positives: 76,011 (correctly predicted as class 0)

  > False Negatives: 2,1065 (incorrectly predicted as class 1)

  > False Positives: 10,638 (incorrectly predicted as class 0)

- **Class 1:**

  > True Positives: 60,000 (correctly predicted as class 1)

  > False Negatives: 1,551 (incorrectly predicted as class 2)

  > False Positives: 735 (incorrectly predicted as class 1)

- **Class 2:**

  > True Positives: 728 (correctly predicted as class 2)

  > False Negatives: 248 (incorrectly predicted as class 0)

  > False Positives: 1,551 (incorrectly predicted as class 2)

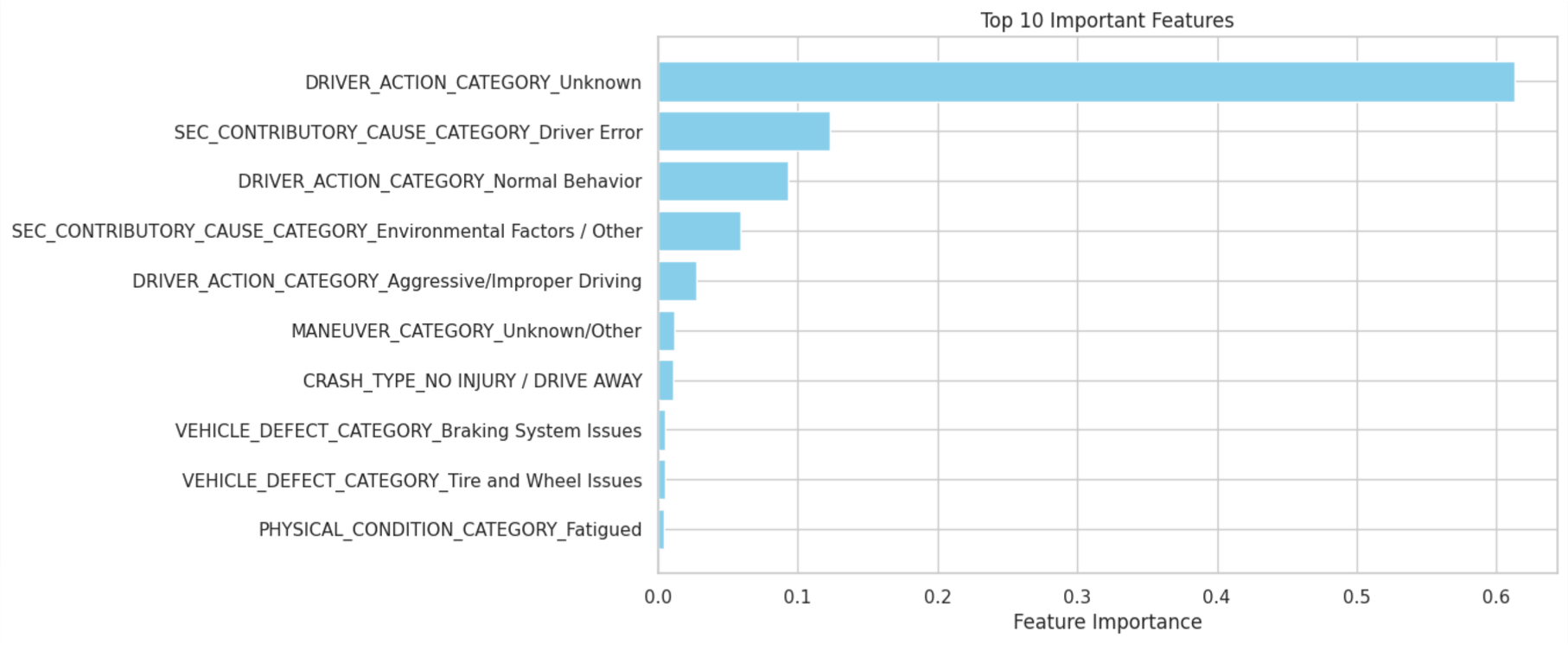
# KEY FINDINGS

- **XGBoost** exhibited strong performance with a test accuracy of 79.9%, precision of 80.1%, and recall of 79.9%. It effectively balanced precision and recall, demonstrating its ability to generalize well to unseen data, making it a robust choice for predicting the primary causes of accidents.
- **Random Forest** also performed admirably with an accuracy of 76.2%, precision of 80.1%, and recall of 76.2%. Its ability to avoid overfitting while accurately identifying important features underscores its reliability as a predictive model.

# KEY FINDINGS

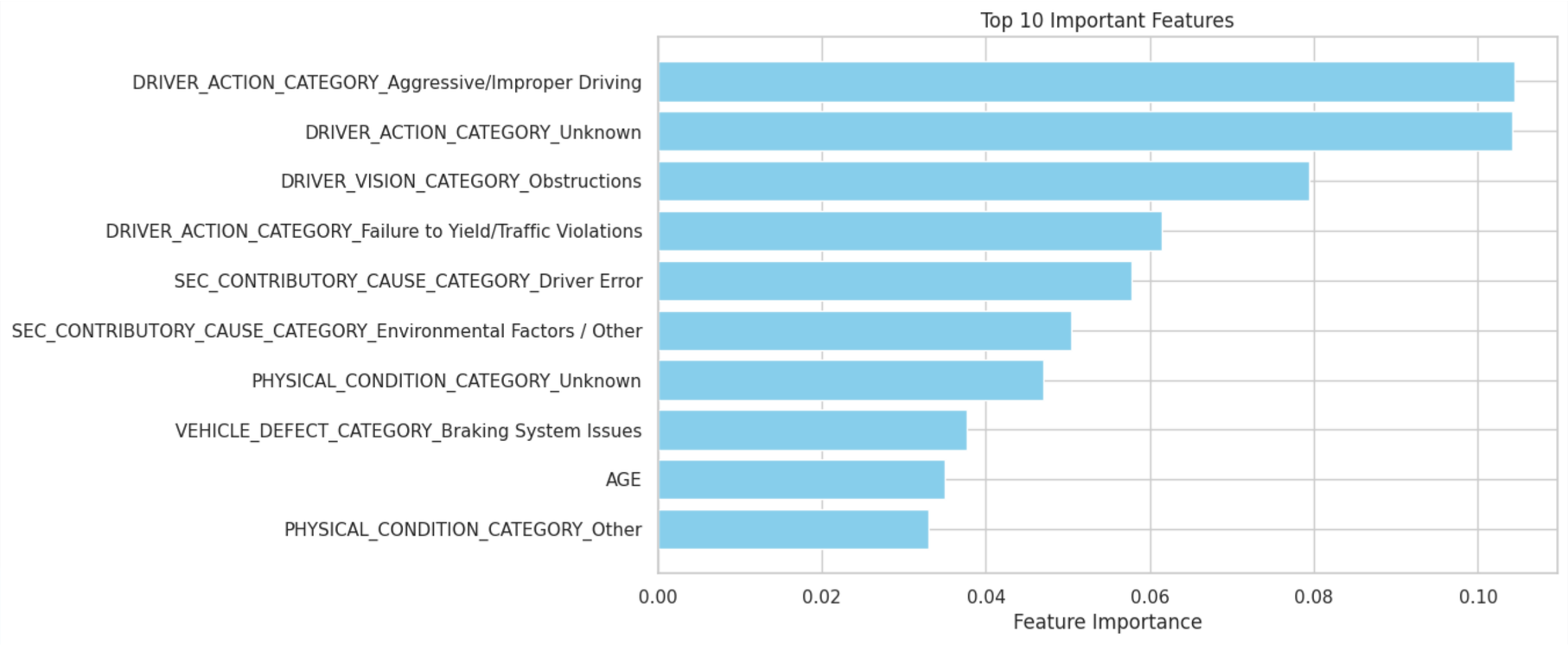
- **Decision Tree** demonstrated moderate performance with an accuracy of 80.0%. Although it showed improved classification post-tuning, the presence of false negatives suggests that it may miss some accident causes, limiting its overall effectiveness in this predictive task.
- **Logistic Regression** showed the weakest performance overall, with an accuracy of 76.2%, and lower precision and recall metrics. Its linear nature made it less effective in capturing complex relationships within the dataset, resulting in limited effectiveness compared to more sophisticated models.

# KEY FEATURES



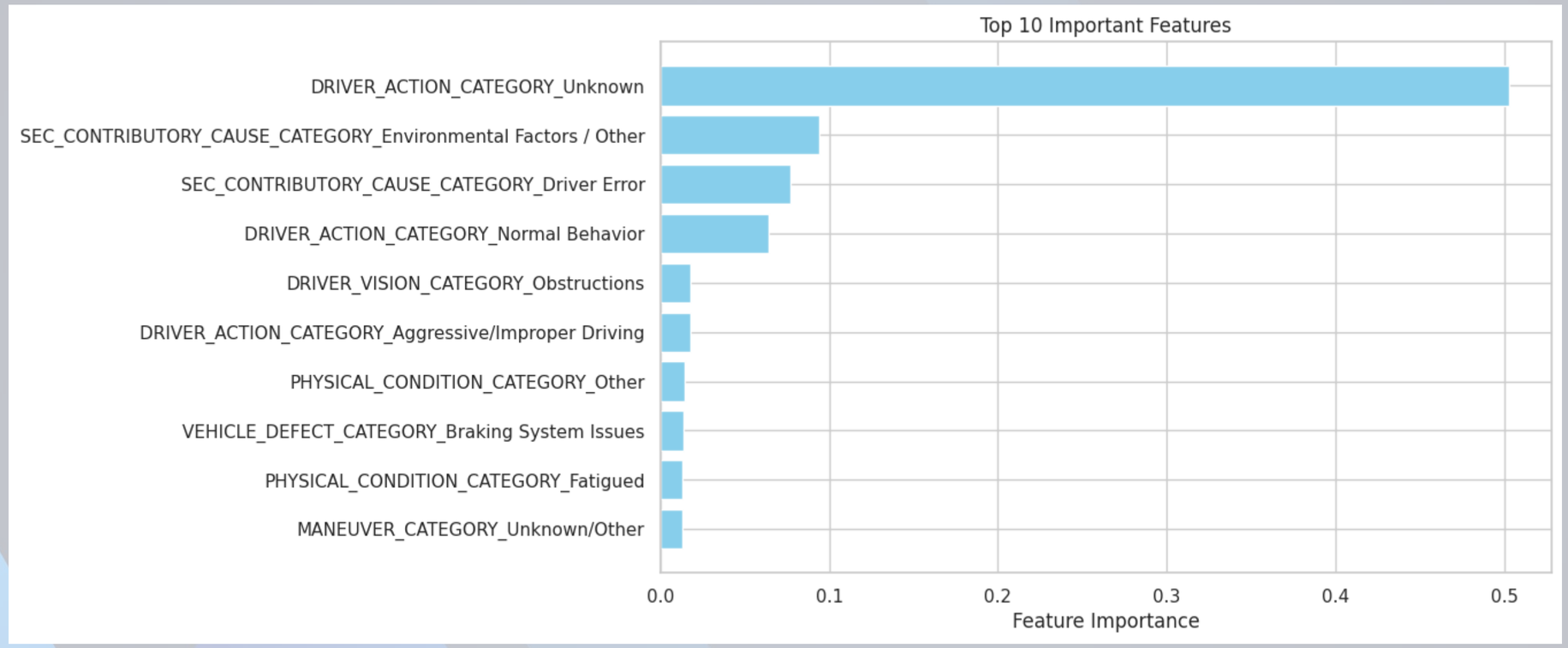

**Feature Importance 1: Decision Tree Model**

# KEY FEATURES




**Feature Importance 2: Random Forest Classifier**

# KEY FEATURES




**Feature Importance 3: XGBoost Classifier**

# KEY FINDINGS

> The key findings in terms of causes of accidents in Chicago City, the four main features of importance across all models, ranked from highest to lowest as seen on the previous slides, are:

**1.Driver Action**
**2.Secondary Contributory Cause: Environmental**
**3.Secondary Contributory Cause: Driver Action**
**4.Driver Vision: Obstruction**

> These findings indicate that while both **XGBoost** and **Random Forest** are strong contenders for this predictive modeling task, **XGBoost** stands out as the most effective model. The identified features suggest that focusing on driver actions and environmental factors is crucial for improving accident prediction models. In contrast, simpler models like Logistic Regression and Decision Tree are less capable of handling the complexity inherent in the dataset.

# RECOMMENDATIONS

Based on the models' findings:

1. **Target Driver Behavior with Education and Enforcement**:
   > Implement public awareness campaigns focused on common driver errors (e.g., speeding, distracted driving).
   > Enhance traffic law enforcement with speed cameras, sobriety checkpoints, and patrols in high-risk areas.

# RECOMMENDATIONS

2. **Upgrade Infrastructure** :
 > Improve road surfaces, drainage systems, and signage, especially in accident-prone zones.
 > Deploy adaptive traffic signals and speed limits that respond to real-time weather and road conditions.

# RECOMMENDATIONS

3. **Strengthen Vehicle Maintenance Programs** :
 > Mandate regular vehicle inspections and integrate health reports into accident databases.
 > Partner with manufacturers and insurers to collect data on common vehicle issues and incentivize proper maintenance.

# RECOMMENDATIONS

4. **Leverage Data Integration and Predictive Tools for Proactive Action:**

> Pool data from insurance companies, automakers, and traffic systems to improve model accuracy.

> Use predictive analytics to allocate resources efficiently, focusing on high-risk areas and accident-prone times.

# CONCLUSION

- The predictive modeling results offer crucial insights into the key contributors to car accidents, empowering data-driven decision-making to improve road safety. By implementing targeted interventions—ranging from driver education and infrastructure upgrades to better vehicle maintenance—stakeholders can significantly reduce traffic accidents. These efforts will promote safer roads and enhance the well-being of all road users.