PREDICTING CONTRIBUTORY CAUSES OF CAR ACCIDENTS IN CHICAGO

PRESENTATION BY : MAUREEN WANJERI

Overview

Objective: Develop a predictive model to identify the primary contributory cause of car accidents in Chicago.

Stakeholders:

- Vehicle Safety Board: Interested in identifying key factors contributing to accidents.
- City of Chicago: Seeks data-driven strategies to reduce traffic accidents and improve public safety.

Business Understanding

- Challenge: High rates of traffic accidents in Chicago leading to injuries, fatalities, and financial losses.
- Goal: Identify patterns in accident data that can inform targeted interventions to reduce accidents.
- Key Questions:
  - What are the most common causes of traffic accidents?
  - How can we predict the primary cause of an accident based on available data?

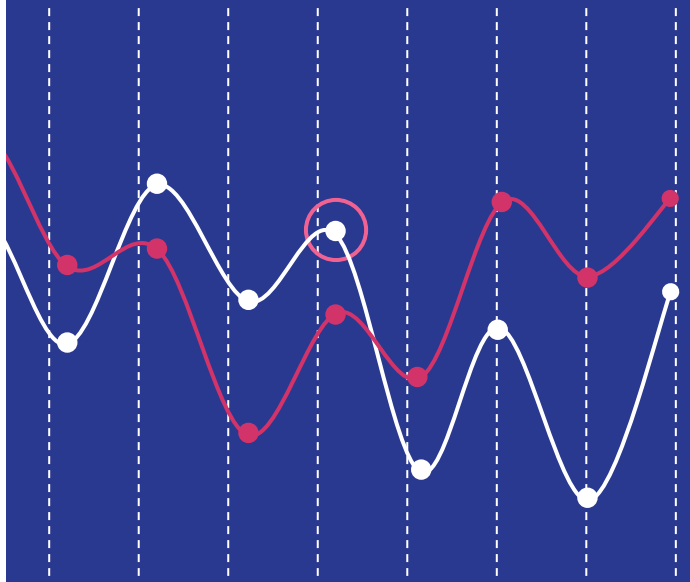
Data Understanding

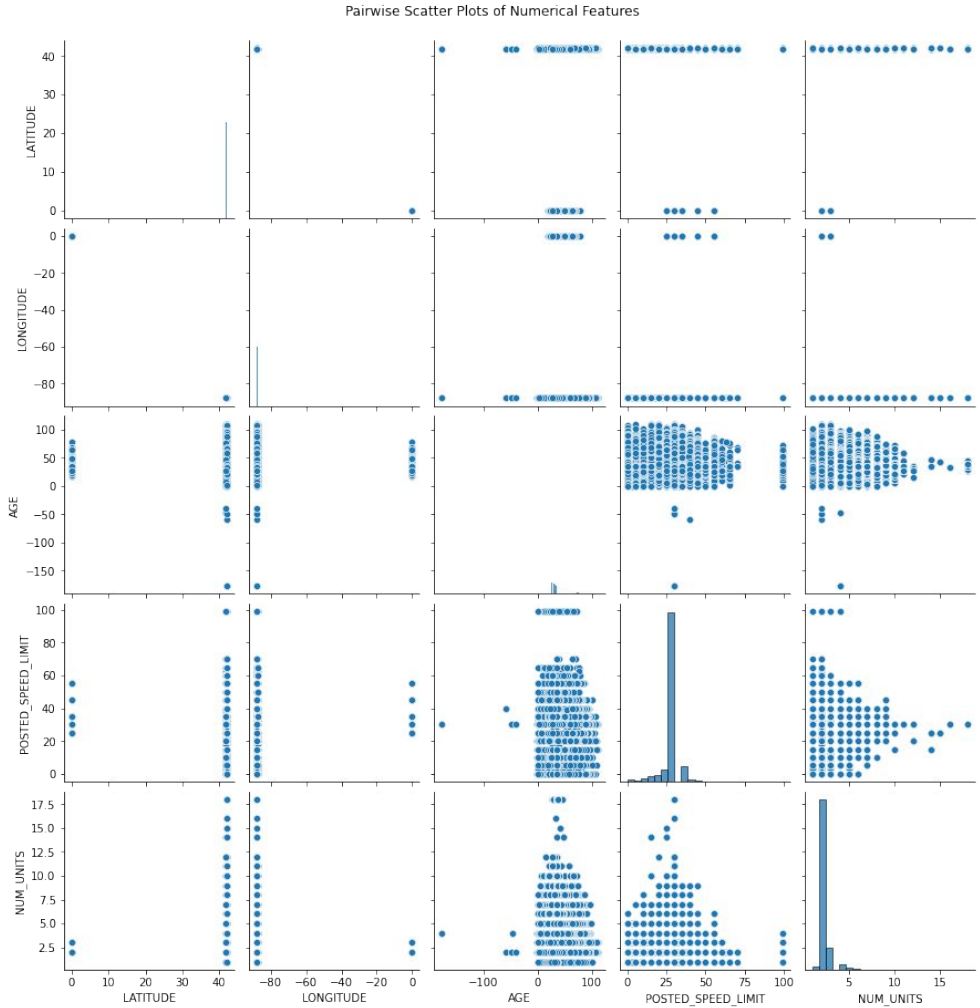
Data Sources:

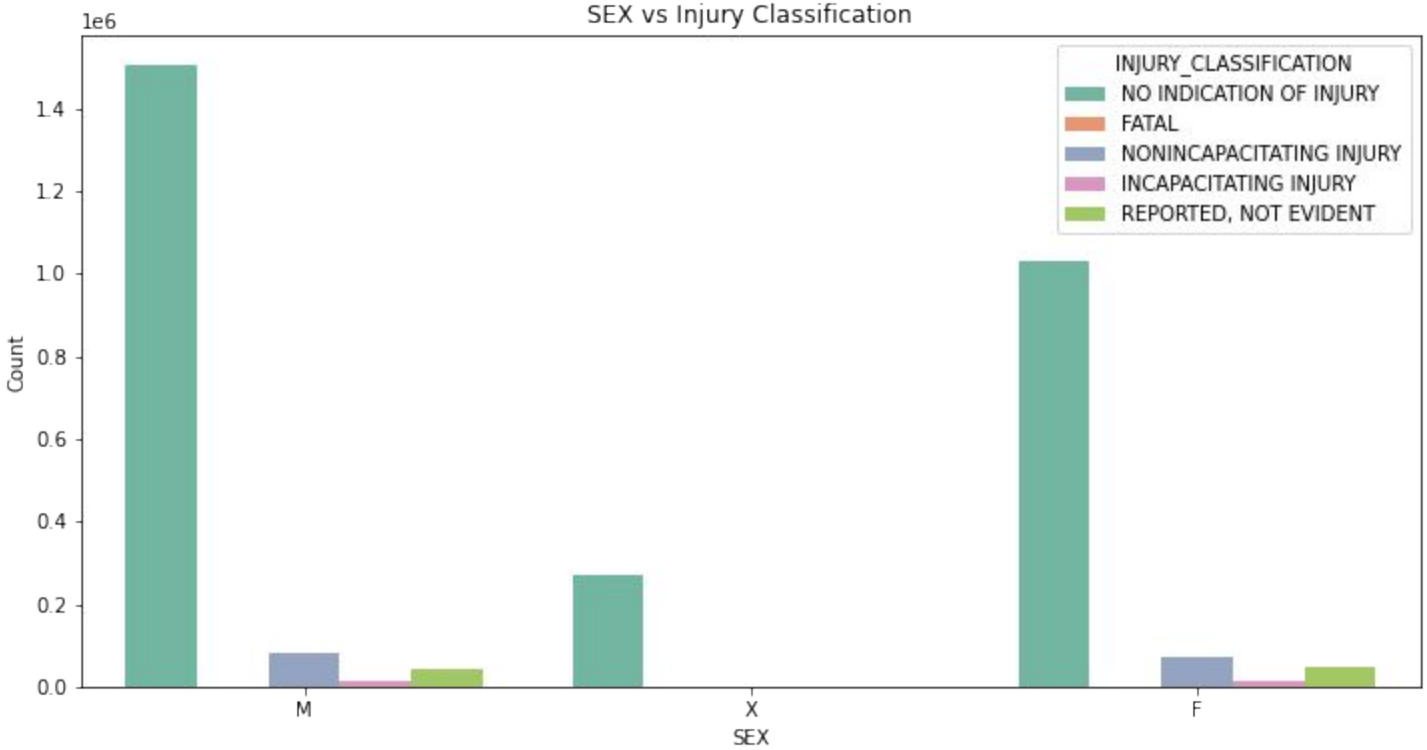
- Traffic Crashes - Crashes
- Traffic Crashes - People
- Traffic Crashes - Vehicles

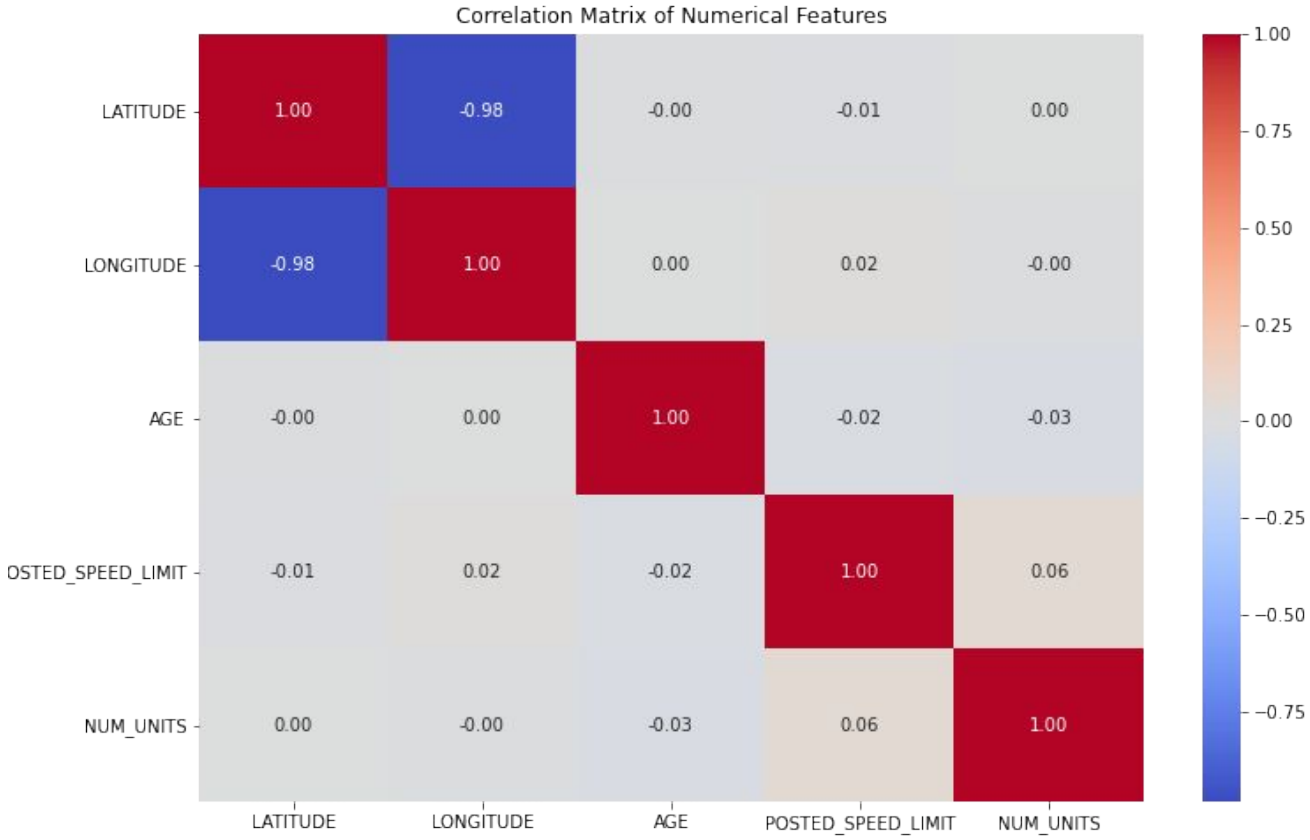
Data Description

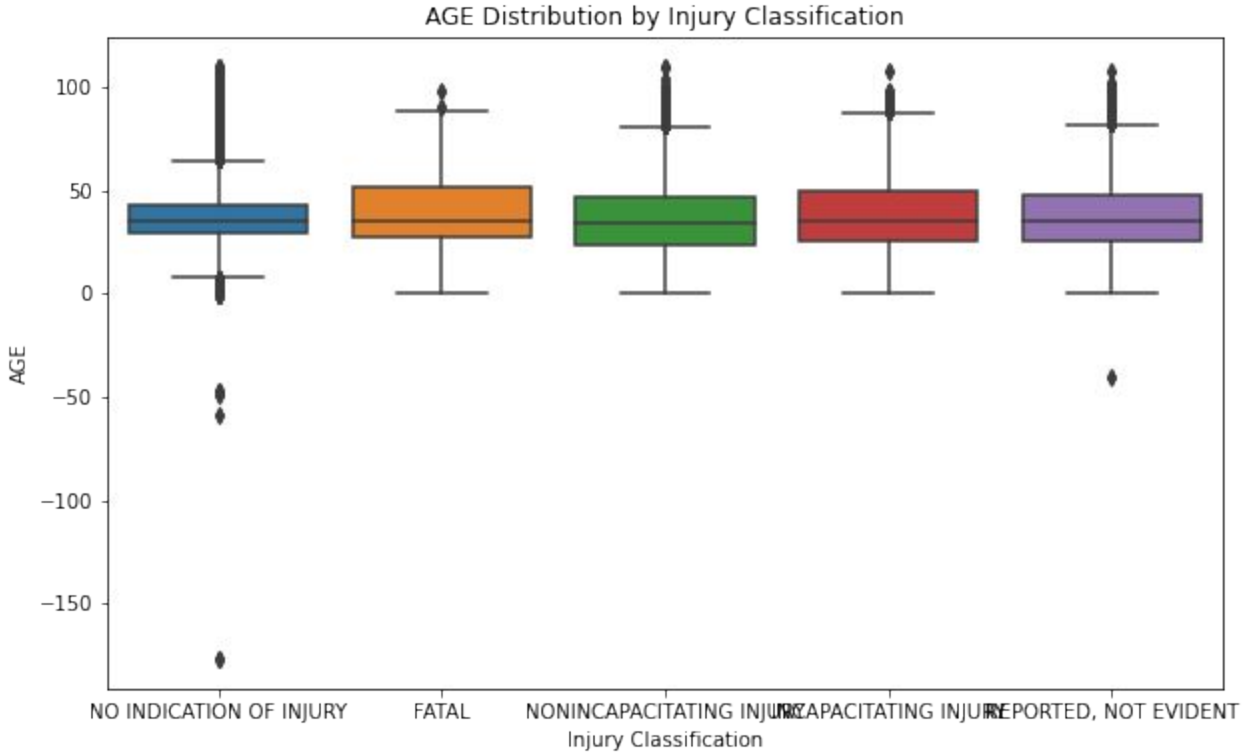
- Key Variables:
  - Target Variable: Injury Classification (e.g., Incapacitating Injury, Non-Incapacitating Injury)
  - Features: Vehicle Type, Weather Conditions, Traffic Control Device, etc.
- Preprocessing Steps:
  - Handling missing values
  - Merging datasets




Pairwise Scatter Plots of Numerical Features
LATITUDE
LONGITUDE
AGE
POSTED_SPEED_LIMIT
NUM_UNITS

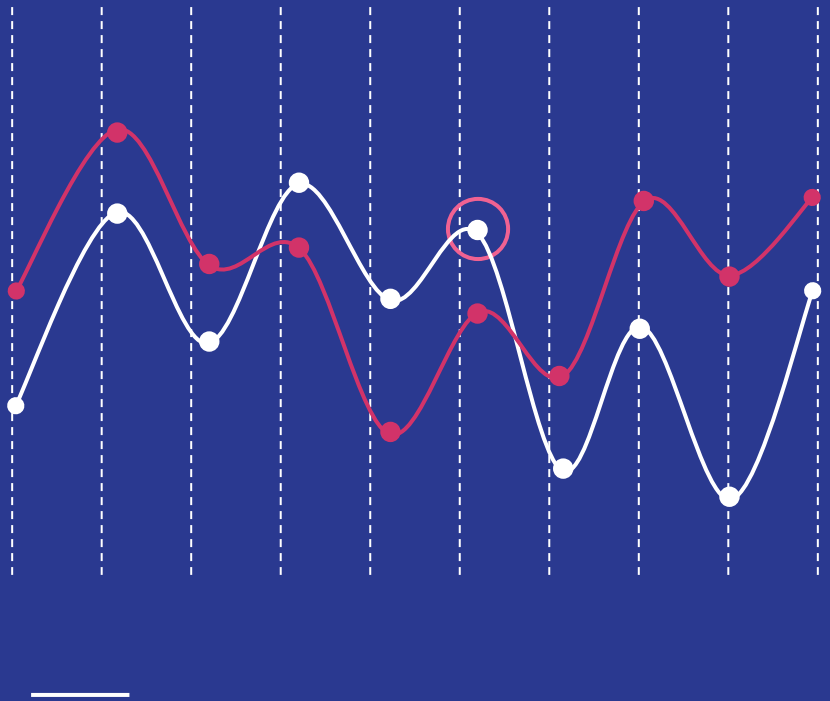

SEX vs Injury Classification
1e6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
Count
M
X
F
SEX
INJURY_CLASSIFICATION
NO INDICATION OF INJURY
FATAL
NONINCAPACITATING INJURY
INCAPACITATING INJURY
REPORTED, NOT EVIDENT

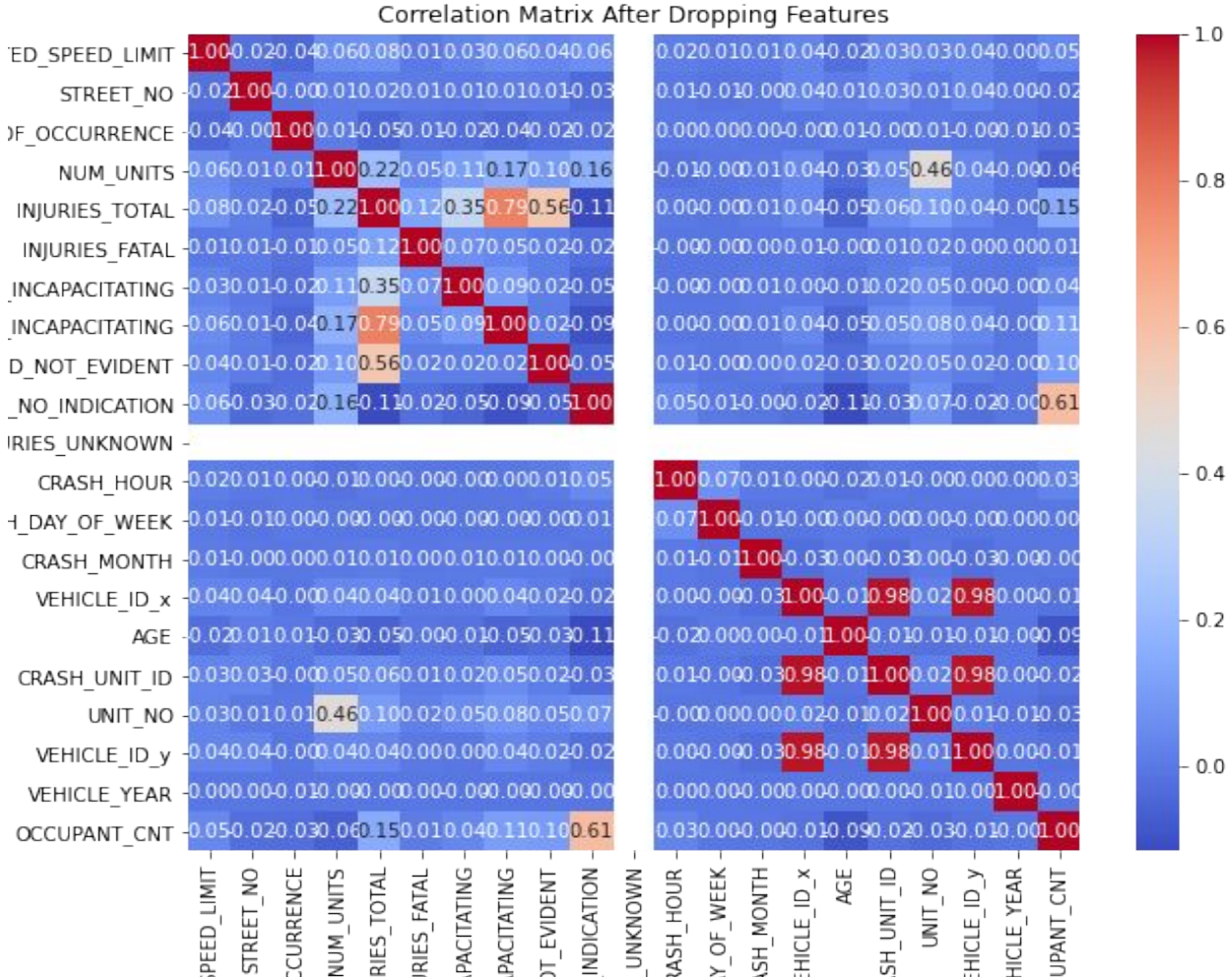

Correlation Matrix of Numerical Features
LATITUDE
LONGITUDE
AGE
OSTED_SPEED_LIMIT
NUM_UNITS
1.00
-0.98
-0.00
-0.01
0.00
-0.98
1.00
0.00
0.02
-0.00
-0.00
0.00
1.00
-0.02
-0.03
-0.01
0.02
-0.02
1.00
0.06
0.00
-0.00
-0.03
0.06
1.00
LATITUDE
LONGITUDE
AGE
POSTED_SPEED_LIMIT
NUM_UNITS
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75


AGE Distribution by Injury Classification
100
50
0
−50
−100
−150
AGE
NO INDICATION OF INJURY
FATAL
NONINCAPACITATING INJURY
INCAPACITATING INJURY
REPORTED, NOT EVIDENT
Injury Classification

Data Preparation

Data Cleaning: Addressed missing values and inconsistencies to ensure data integrity.
Feature Engineering: Created new features to enhance model performance.
Categorical Encoding: Transformed categorical variables into numerical format for accurate model interpretation.
Train-Test Split: To ensure balanced and robust training and testing datasets.

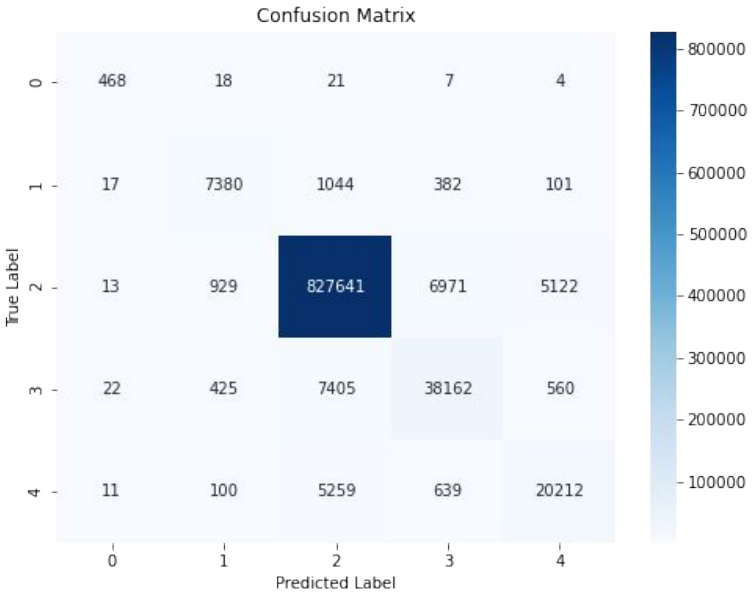


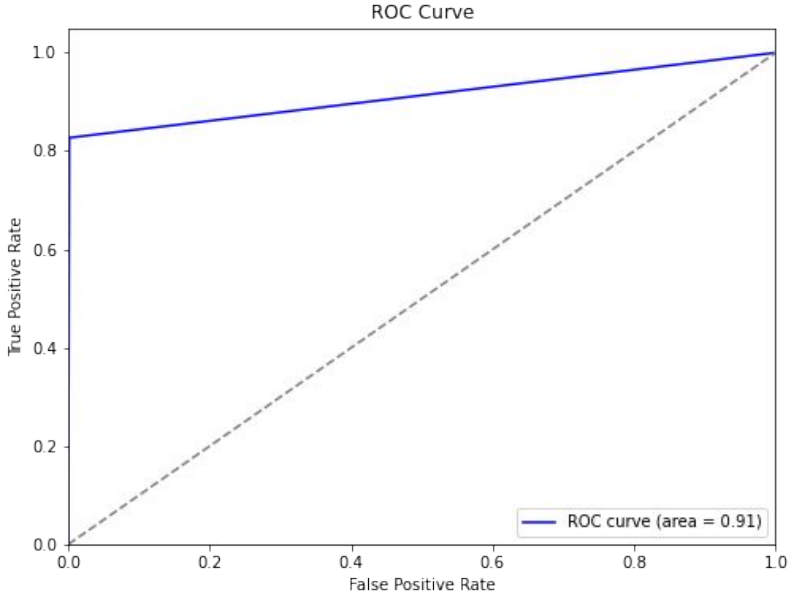

Correlation Matrix After Dropping Features
ED_SPEED_LIMIT
STREET_NO
OF_OCCURRENCE
NUM_UNITS
INJURIES_TOTAL
INJURIES_FATAL
INCAPACITATING
INCAPACITATING
D_NOT_EVIDENT
NO_INDICATION
IRIES_UNKNOWN
CRASH_HOUR
H_DAY_OF_WEEK
CRASH_MONTH
VEHICLE_ID_x
AGE
CRASH_UNIT_ID
UNIT_NO
VEHICLE_ID_y
VEHICLE_YEAR
OCCUPANT_CNT
1.0
0.8
0.6
0.4
0.2
0.0

Modelling

Decision Tree Baseline Model

- Developed a basic model using a Decision Tree to predict the main cause of car accidents in Chicago.
- Model Performance:
  - Prediction Accuracy: The model was able to accurately predict most cases where there was no indication of injury, as seen in the confusion matrix.
- Model Quality: The ROC curve, which measures the model's ability to differentiate between classes, showed a strong performance with a score of 0.91.
- Challenges Encountered
  - Memory Constraints:
    - While trying to improve the model and test other types like Logistic Regression, we faced significant memory limitations.
    - These limitations meant we couldn't run more complex models or tune the existing model further.
- Impact on Project:
  - Due to these technical challenges, we were only able to present results from the baseline Decision Tree model without further refinements or additional models.
- Conclusion
  - Despite the challenges, the baseline Decision Tree model provided valuable insights, but further work is needed with more robust resources to refine and expand the analysis.
  -


Confusion Matrix
True Label
Predicted Label
0
1
2
3
4
468
18
21
7
4
17
7380
1044
382
101
13
929
827641
6971
5122
22
425
7405
38162
560
11
100
5259
639
20212
800000
700000
600000
500000
400000
300000
200000
100000


ROC Curve
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (area = 0.91)

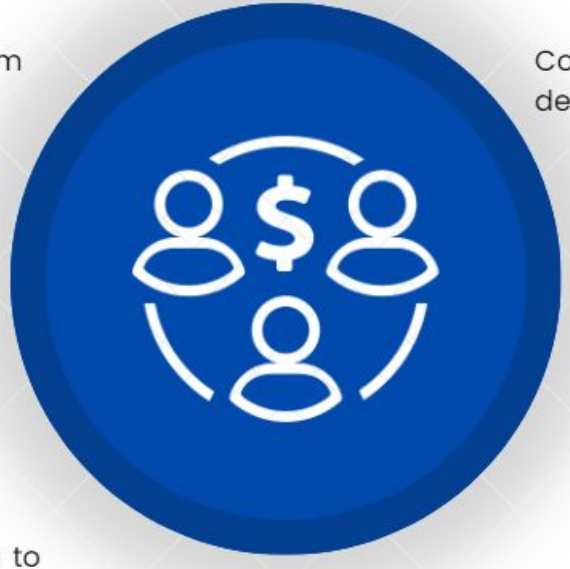
Recommendations

- Focus on improving intersection safety, as many accidents are associated with "TRAFFIC SIGNAL" as a primary cause.
- Increase awareness and enforcement of traffic laws in areas with frequent "NONINCAPACITATING INJURY" accidents.
- Deploy targeted interventions during peak hours and under specific weather conditions identified as high-risk.

Next Steps

Implement the model in a real-time system for predicting accident causes.

Consider additional features like traffic density and real-time traffic data.



Regularly update the model with new data to maintain accuracy.

THANK YOU

Q&A

Maureen Wanjeri Njoroge
Linkedin : https://www.linkedin.com/in/maureen-wanjeri-pmp%C2%AE-pmi-acp-%C2%AE-1x-aws-924364bb/