CITY OF CHICAGO TRAFFIC CRASHES

STUDENT NAMES; Aggrey Timbwa, Richard Macharia, Pamela Jepkorir Chebii

GROUP; Group 4

STUDENT PACE; Part time

SCHEDULE PROJECT REVIEW DATE; phase four(14/10/2024)

INSTRUCTOR NAME; Winnie Anyoso

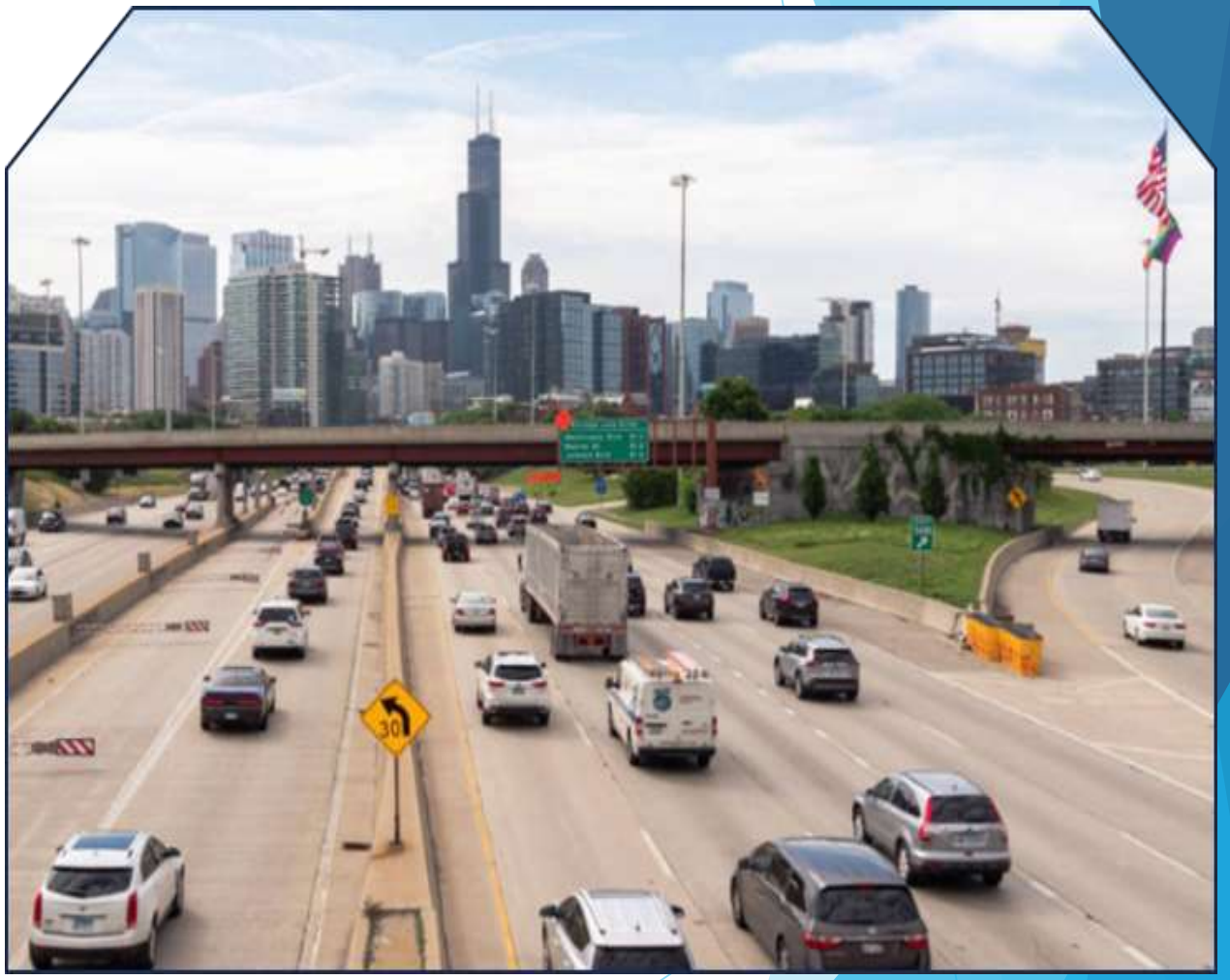
PROJECT INTRODUCTION

Chicago’s crash dataset details traffic incidents within city limits under the Chicago Police Department (CPD) jurisdiction, compiled from the electronic E-Crash reporting system.

The dataset excludes personal identifiers and is updated as reports are finalized or amended.

Approximately half of the reports are self-reported by drivers at police stations, while the remainder is recorded by officers on-site.

We aim to predict the primary causes of these accidents



OVERVIEW

The goal of this project is to build a model that predicts the primary contributory cause of a car accident based on factors such as road conditions, vehicle characteristics, and the people involved

BUSINESS PROBLEM

Traffic accidents are a major concern in large cities like Chicago, leading to property damage, injuries, and fatalities.
Understanding the primary causes of these accidents can enable city planners, traffic safety boards, and policymakers to implement proactive measures to reduce incidents and enhance road safety.
This project utilizes a dataset provided by the City of Chicago, which contains comprehensive information on accidents, vehicles, and individuals involved.
This data offers valuable insights into the root causes of traffic crashes.

OBJECTIVES

1. Main Objective:
Build a model to predict the **PRIM_CONTRIBUTORY_CAUSE** of car accidents.

2. Data Quality:
Ensure the dataset is of high quality by maintaining completeness and accuracy, especially in critical variables like road conditions, weather, and vehicle information. Reliable data will enable more accurate predictions and ensure robust models

.

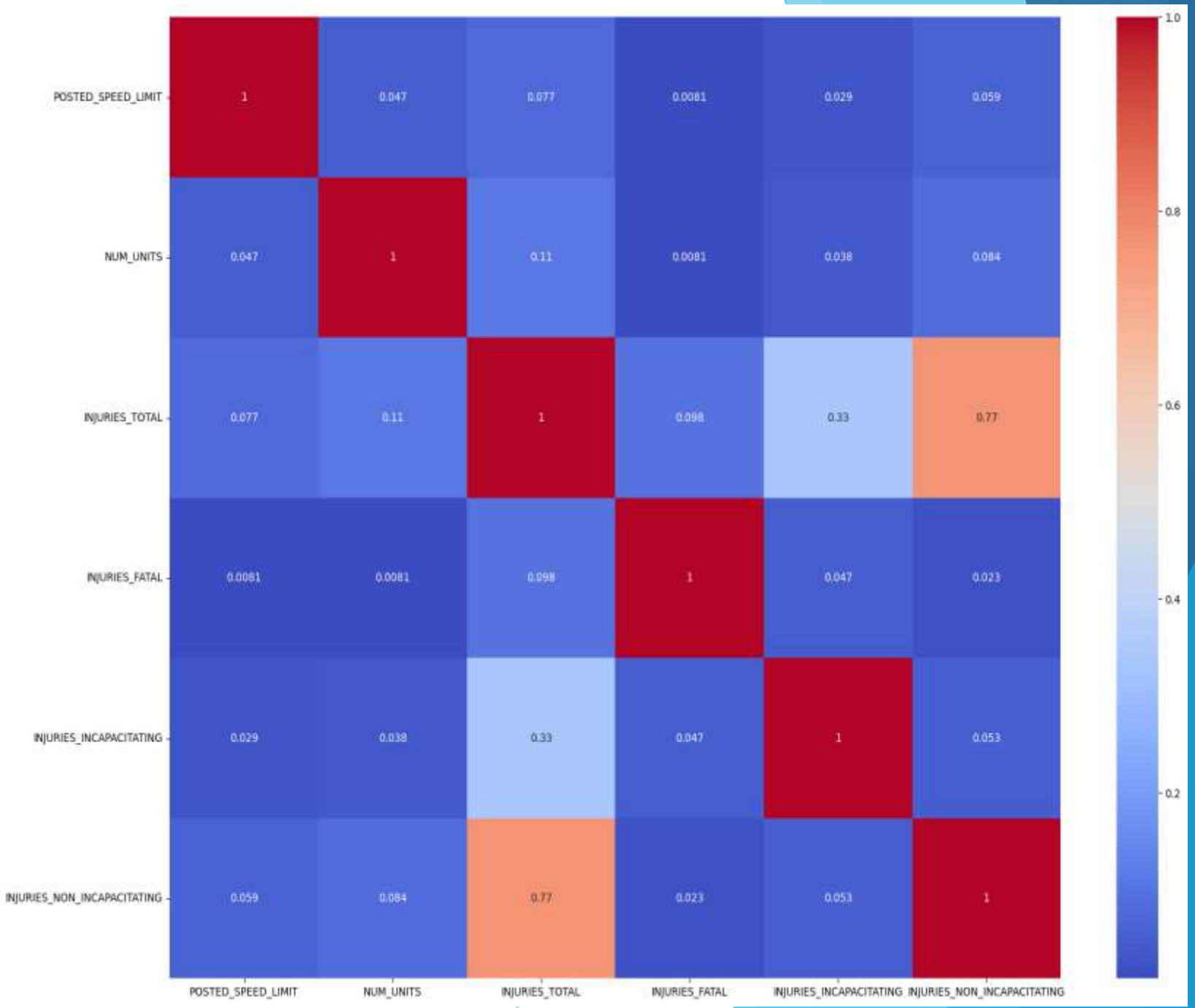
3. Data Imbalance:
Address the severe imbalance in the target variable (PRIM_CONTRIBUTORY_CAUSE) by applying techniques such as **SMOTE**, **class weighting**, or **ensemble methods**.

4. Feature Importance:
Investigate the relationships between key features, such as **road conditions**, **vehicle types**, and **driver behavior**.

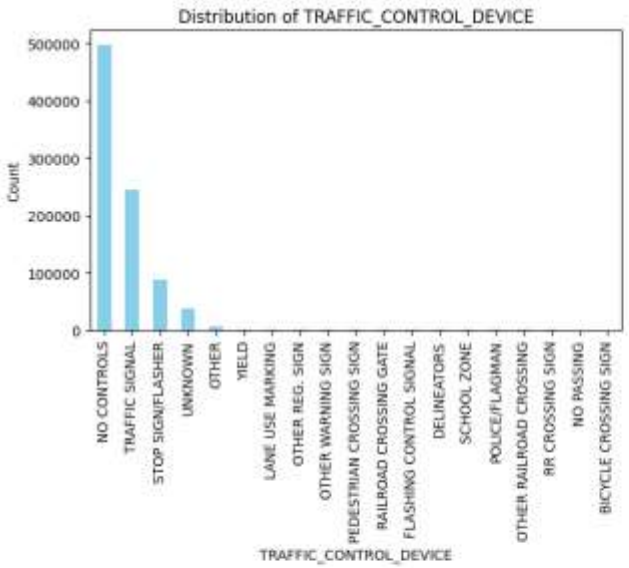
VISUALIZATIONS

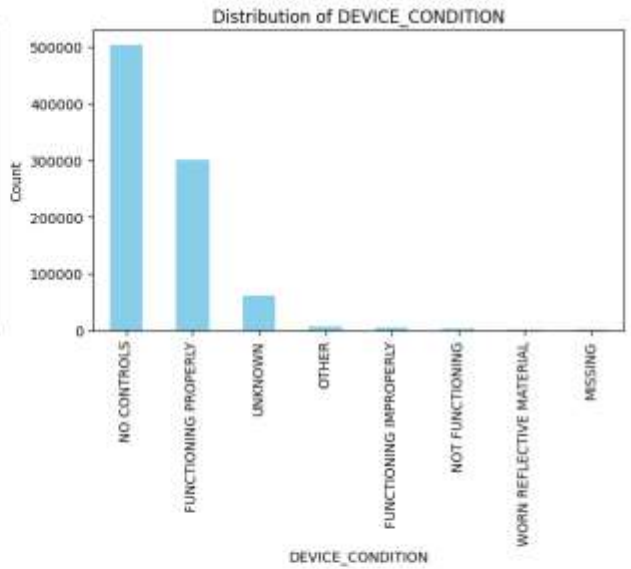
CORRELATIONS MATRIX

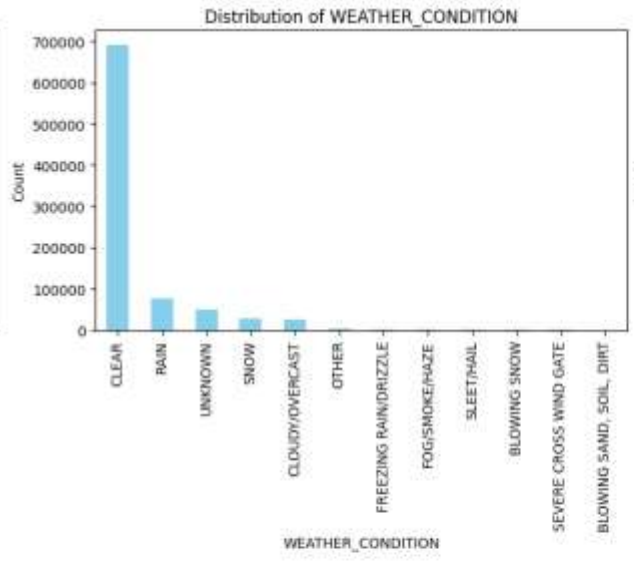
Our correlation analysis only includes numerical features with only INJURIES_TOTAL and INJURY_NON_INCAPACITATING having high correlation (77%) which is to be expected

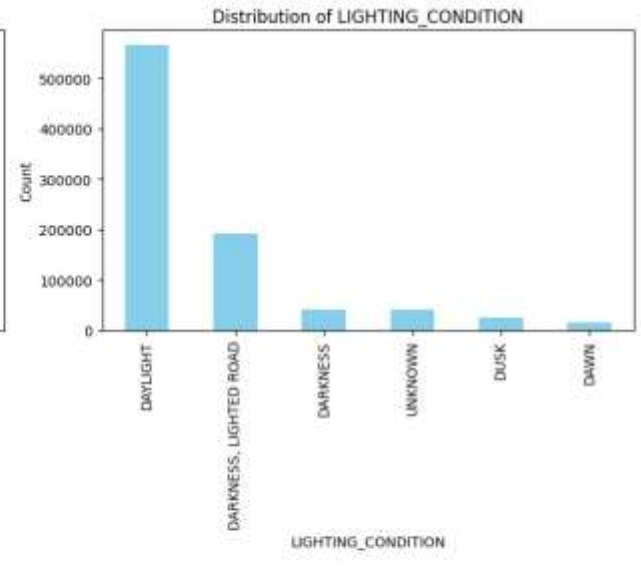


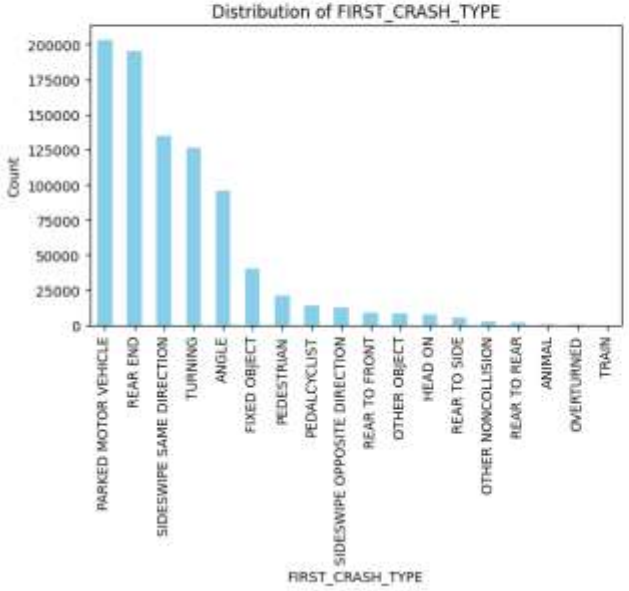
BAR PLOTS

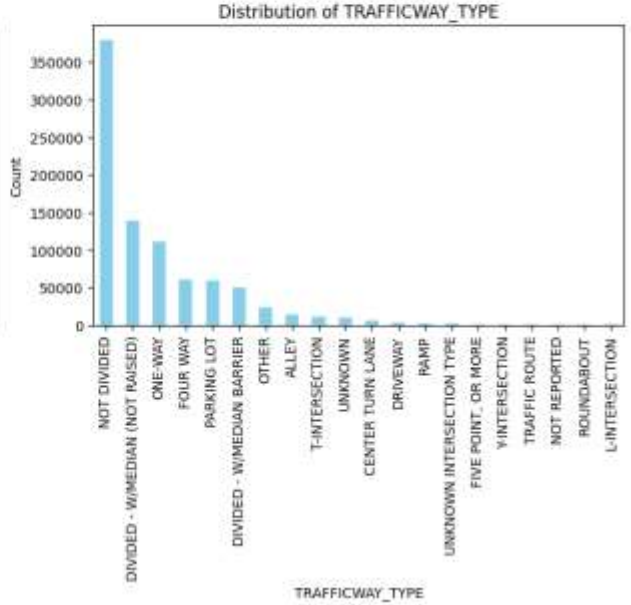

Distribution of TRAFFIC_CONTROL_DEVICE
Count
NO CONTROLS
TRAFFIC SIGNAL
STOP SIGN/FLASHER
UNKNOWN
OTHER
YIELD
LANE USE MARKING
OTHER REG. SIGN
OTHER WARNING SIGN
PEDESTRIAN CROSSING SIGN
RAILROAD CROSSING GATE
FLASHING CONTROL SIGNAL
DELINEATORS
SCHOOL ZONE
POLICE/FLAGMAN
OTHER RAILROAD CROSSING
RR CROSSING SIGN
NO PASSING
BICYCLE CROSSING SIGN
TRAFFIC_CONTROL_DEVICE

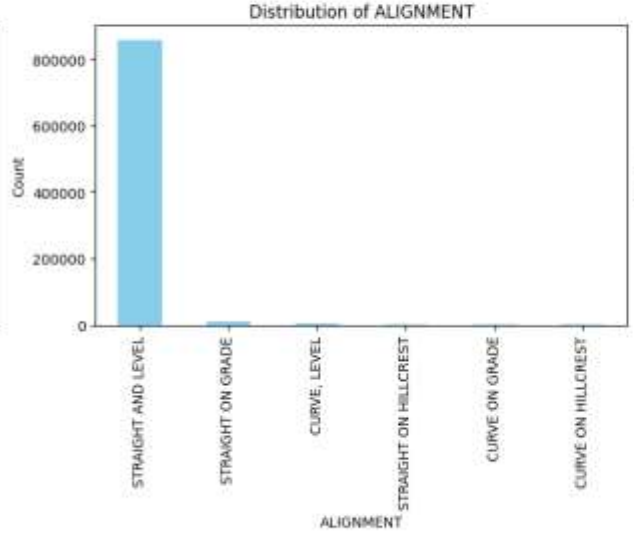

Distribution of DEVICE_CONDITION
Count
NO CONTROLS
FUNCTIONING PROPERLY
UNKNOWN
OTHER
FUNCTIONING IMPROPERLY
NOT FUNCTIONING
WORN REFLECTIVE MATERIAL
MISSING
DEVICE_CONDITION

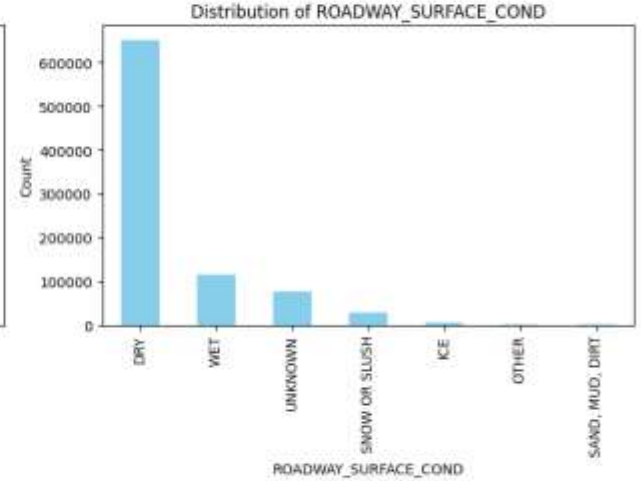

Distribution of WEATHER_CONDITION
Count
CLEAR
RAIN
UNKNOWN
SNOW
CLOUDY/OVERCAST
OTHER
FREEZING RAIN/DRIZZLE
FOG/SMOKE/HAZE
SLEET/HAIL
BLOWING SNOW
SEVERE CROSS WIND GATE
BLOWING SAND, SOIL, DIRT
WEATHER_CONDITION


Distribution of LIGHTING_CONDITION
Count
DAYLIGHT
DARKNESS, LIGHTED ROAD
DARKNESS
UNKNOWN
DUSK
DAWN
LIGHTING_CONDITION


Distribution of FIRST_CRASH_TYPE
Count
PARKED MOTOR VEHICLE
REAR END
SIDESWIPE SAME DIRECTION
TURNING
ANGLE
FIXED OBJECT
PEDESTRIAN
PEDALCYCLIST
SIDESWIPE OPPOSITE DIRECTION
REAR TO FRONT
OTHER OBJECT
HEAD ON
REAR TO SIDE
OTHER NONCOLLISION
REAR TO REAR
ANIMAL
OVERTURNED
TRAIN
FIRST_CRASH_TYPE


Distribution of TRAFFICWAY_TYPE
Count
NOT DIVIDED
DIVIDED - W/MEDIAN (NOT RAISED)
ONE-WAY
FOUR WAY
PARKING LOT
DIVIDED - W/MEDIAN BARRIER
OTHER
ALLEY
T-INTERSECTION
UNKNOWN
CENTER TURN LANE
DRIVEWAY
RAMP
UNKNOWN INTERSECTION TYPE
FIVE POINT, OR MORE
Y-INTERSECTION
TRAFFIC ROUTE
NOT REPORTED
ROUNDABOUT
L-INTERSECTION
TRAFFICWAY_TYPE


Distribution of ALIGNMENT
Count
STRAIGHT AND LEVEL
STRAIGHT ON GRADE
CURVE, LEVEL
STRAIGHT ON HILLCREST
CURVE ON GRADE
CURVE ON HILLCREST
ALIGNMENT


Distribution of ROADWAY_SURFACE_COND
Count
DRY
WET
UNKNOWN
SNOW OR SLUSH
ICE
OTHER
SAND, MUD, DIRT
ROADWAY_SURFACE_COND

BAR PLOTS

The data from the bar plots above suggests that many accidents occur under clear weather, daylight, and dry conditions, indicating that environmental factors like adverse weather or poor lighting are not the primary causes of most accidents.

Instead, driver-related factors (like inattentiveness, tailgating, and speeding) and road design (such as undivided roads and intersections without controls) appear to play a more significant role in accident occurrences.

This insight could guide traffic safety measures focusing on driver education, enforcement at high-risk intersections, and safety improvements on undivided roads.

MODELLING TECHNIQUES

Baseline Model: Dummy Classifier

Baseline Model Accuracy: The accuracy of the baseline model is **0.5953**, indicating that the model correctly predicts approximately 59.53% of the instances.

Findings

The model exhibits a strong **imbalance in predictions**, as it only predicts the class **"Pedestrian/Cyclist Errors"** effectively, achieving a recall of **1.00**. This indicates that all instances of this class were correctly identified.

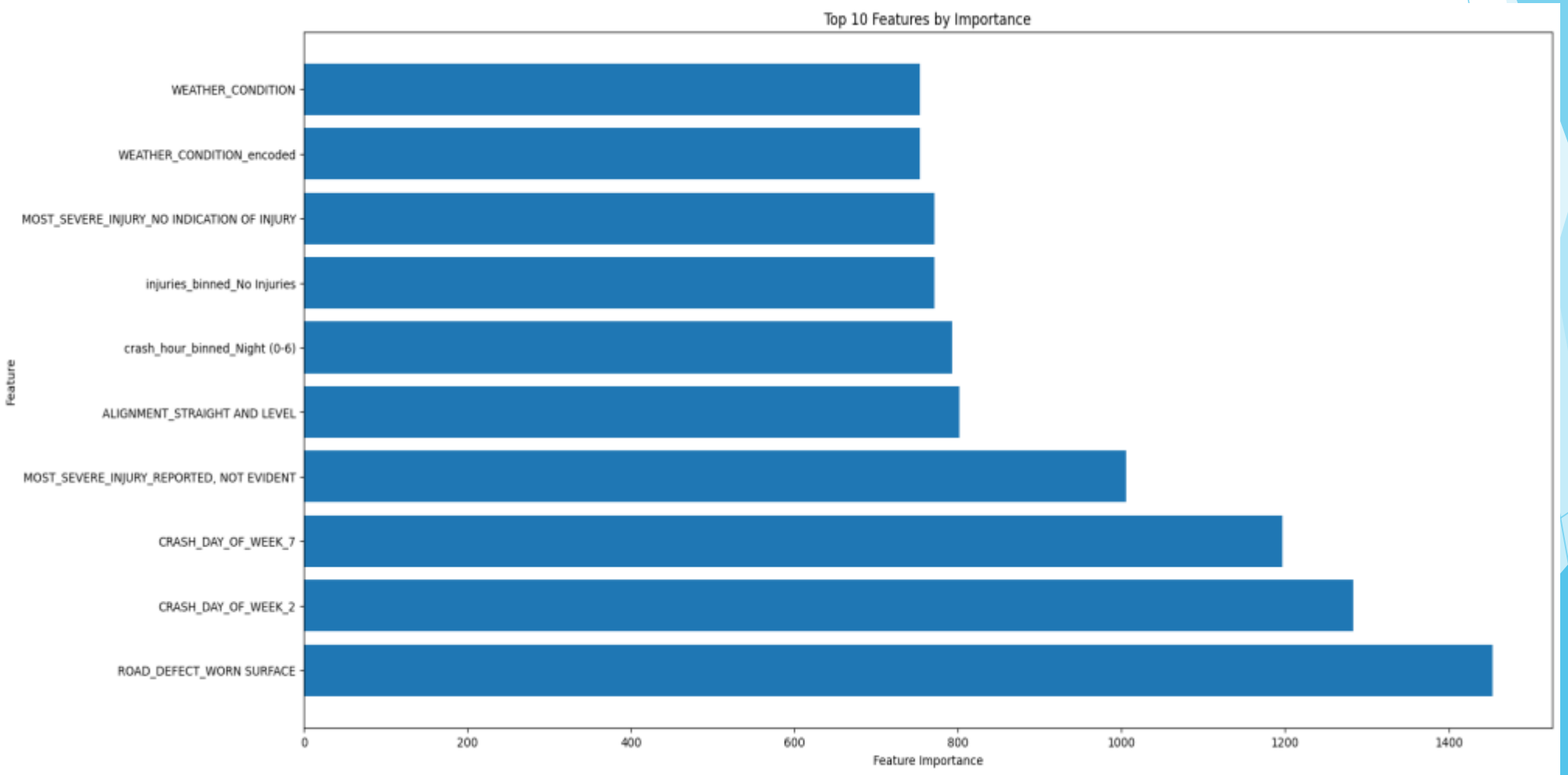
However, the model fails to predict any instances of the other classes, resulting in **zero precision and recall** for them. This highlights a significant shortfall in the model's ability to generalize across all categories.

The **macro average** metrics indicate overall poor performance, with an F1-score of only **0.12**, emphasizing the need for improvement.

Model Performance Overview

Model	Accuracy	Key Insights
Logistic Regression	0.6303	Struggled with most classes; good recall for **Pedestrian/Cyclist Errors** but poor overall performance.
Logistic Regression (SMOTE)	0.3066	Significant drop in accuracy; failed to effectively address class imbalance.
Random Forest	0.6140	Good recall for **Pedestrian/Cyclist Errors**; poor performance on minority classes.
XGBoost	0.6393	Comparable to the neural network; maintains good performance for certain classes, particularly **Pedestrian/Cyclist Errors**.
Neural Network	0.6429	Highest accuracy; performs well for **Pedestrian/Cyclist Errors** but shows signs of potential overfitting.

Model Comparison: Random Forest vs. XGBoost



CONCLUSION

1. Addressing the Problem with Predictive Models:

Neural Network and XGBoost models effectively identified key factors like road conditions, time of day, and human behavior, providing actionable insights for stakeholders.

The Neural Network achieved the highest accuracy (0.6429) by capturing complex patterns and relationships in the data, but it showed signs of overfitting, indicating a need for further tuning to improve consistency. XGBoost achieved an accuracy of 0.6393, offering strong performance with minimal overfitting, making it a reliable choice for practical applications

2. Insights on Contributory Causes:

Key features identified by the models, such as **road defects** and **day of the week**, align with real-world safety concerns. This demonstrates that the models are not only predictive but also relevant to stakeholder needs.

3. Handling Data Challenges:

Class Imbalance: Despite mitigation efforts like SMOTE, models such as Logistic Regression struggled to accurately predict minority classes, reflecting the complexity of modeling rare accident causes. The Neural Network and XGBoost demonstrated stronger performance across categories, indicating a better ability to manage data imbalance, though further improvement is still warranted.

RECOMMENDATIONS

1. **Hyperparameter Tuning**:

Further refine the **Neural Network** to address overfitting and unlock additional performance gains.

2. Feature Engineering:

Explore new features, such as **weather and traffic congestion interactions**, to capture more nuanced relationships between accident causes.

3. Continuous Learning:

As new data becomes available, retrain models periodically to maintain predictive relevance and adapt to changing traffic patterns.


Thank
you