PHASE 4 PROJECT – CHICAGO CAR CRASHES

JESSE GITOBU

BUSINESS UNDERSTANDING

- **Context:** This project aims to analyze car accident data to identify the primary contributing causes. Understanding these causes is crucial for developing effective strategies to reduce traffic accidents, improve road safety, and ultimately save lives.
- **Stakeholders:**
  - **Vehicle Safety Board:** Interested in understanding vehicle-related factors and informing vehicle safety standards.
  - **City of Chicago (or other municipalities):** Interested in optimizing traffic management, infrastructure planning, and public safety initiatives.

PROBLEM STATEMENT

- he Challenge: Traffic accidents are a significant public health concern, resulting in injuries, fatalities, and economic losses. Understanding the underlying causes of these accidents is essential for implementing effective prevention strategies.
- Specific Questions:
  - What are the most frequent primary contributing causes of car accidents?
  - Which factors (e.g., driver behavior, road conditions, weather, vehicle-related issues) are most strongly associated with different accident causes?
  - Can we accurately predict the primary contributing cause of an accident based on available data?
  - Are there specific patterns or trends in accident causes based on location, time of day, weather conditions, or driver demographics?

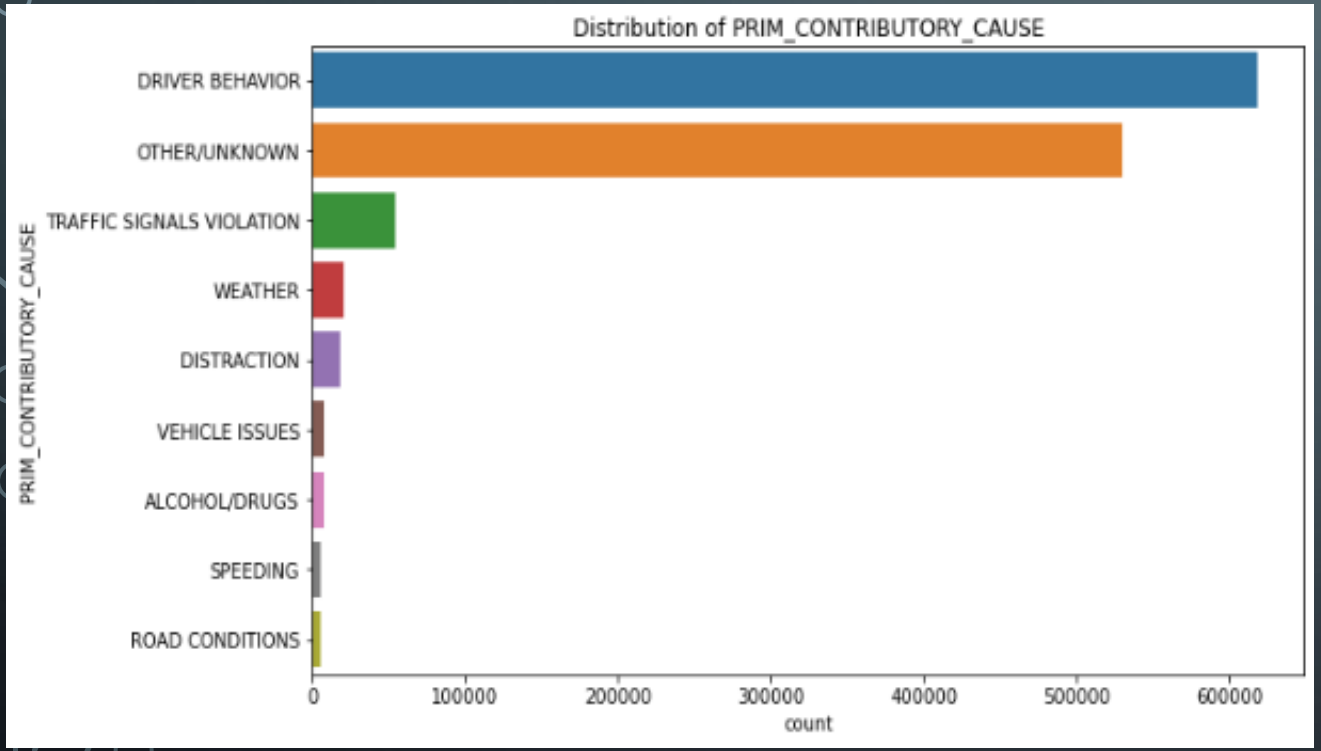
OBJECTIVES

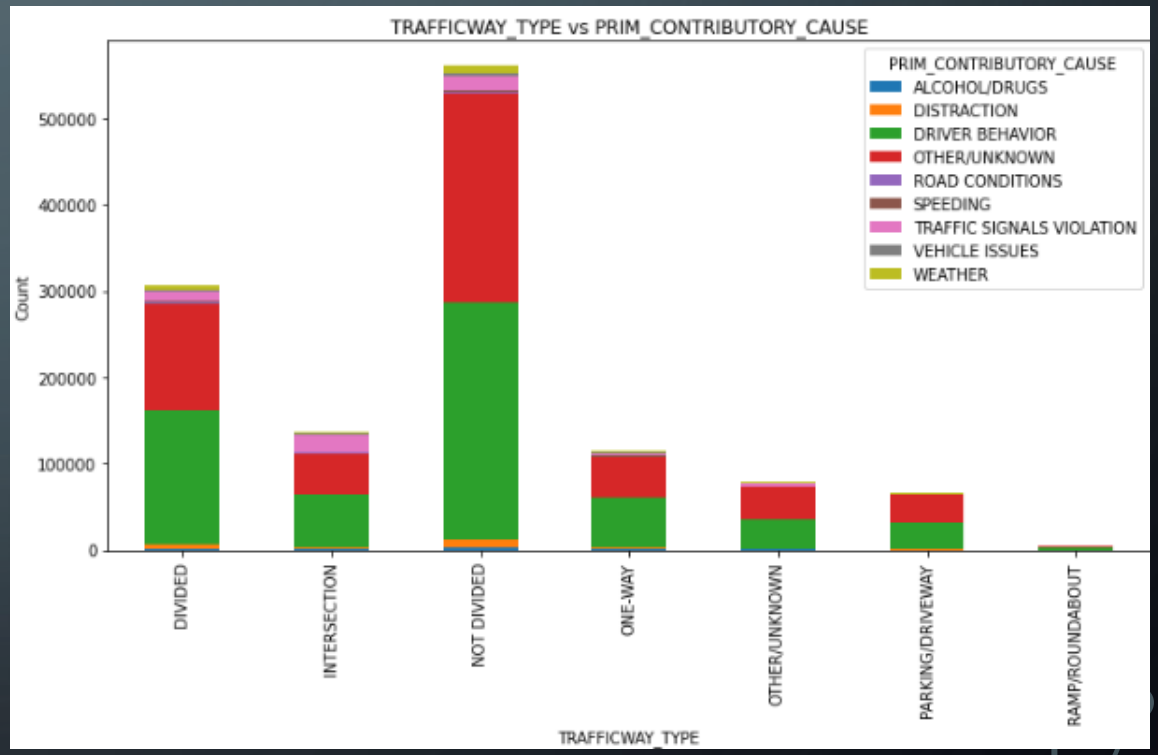
- The main Objective of this dataset is to:
- a) Develop a machine learning model that predicts the primary contributory cause of vehicle accidents with reasonable accuracy.
- b)Identify the most significant factors influencing crashes, such as driver behavior, road conditions, weather, and vehicle attributes.
- c) Determine which variables contribute most to accidents, such as reckless driving, speeding, distracted driving, or poor road conditions

DATA UNDERSTANDING

- Dataset Source: The data used in this project comes from [Specify the data source, e.g., the City of Chicago Data Portal, a specific government agency, etc.]. Provide a link if available.
- Data Description: The dataset contains information about car accidents, including:
  - Target Variable: PRIM_CONTRIBUTORY_CAUSE (Primary Contributing Cause of the Accident)
- Features:
  - Vehicle-related factors: VEHICLE_TYPE, VEHICLE_DEFECT, VEHICLE_USE, VEHICLE_AGE_CATEGORY
  - Driver-related factors: DRIVER_ACTION, DRIVER_VISION, DRIVER_CATEGORY, SEX, OCCUPANT_CNT
  - Environmental factors: WEATHER_CONDITION, LIGHTING_CONDITION, ROADWAY_SURFACE_COND, ROAD_DEFECT
  - Location and Time: POSTED_SPEED_LIMIT, TRAFFICWAY_TYPE, ALIGNMENT, CRASH_YEAR, CRASH_WEEKDAY, TIME_OF_DAY, CRASH_SEASON
  - Crash characteristics: FIRST_CRASH_TYPE, TRAFFIC_CONTROL_DEVICE, DEVICE_CONDITION
  - Other: TRAVEL_DIRECTION

EXPLORATORY DATA ANALYSIS





MODELLING

- Random Forest: Best F1 Score = 0.5577
  - Best Parameters: {'classifier__max_depth': 10, 'classifier__min_samples_leaf': 1, 'classifier__min_samples_split': 2, 'classifier__n_estimators': 300}
- KNN: Best F1 Score = 0.5386
  - Best Parameters: {'classifier__n_neighbors': 11, 'classifier__p': 1, 'classifier__weights': 'distance'}
- Gradient Boosting: Best F1 Score = 0.6025
  - Best Parameters: {'classifier__learning_rate': 0.01, 'classifier__max_depth': 5, 'classifier__n_estimators': 200, 'classifier__subsample': 0.8}
-

COEFFICIENT MATRIX

GRADIENT BOOSTING

- Strengths:
  - Highest overall accuracy and weighted F1-score.
  - Reasonable performance on DRIVER BEHAVIOR and OTHER/UNKNOWN, which are the most frequent classes.
- Weaknesses:
  - Extremely poor performance on minority classes (ALCOHOL/DRUGS, ROAD CONDITIONS, SPEEDING, WEATHER, VEHICLE ISSUES, DISTRACTION). The precision and recall values are very low, indicating that the model is not able to identify these causes effectively.
  - The confusion matrix shows that the model often misclassifies these minority classes as DRIVER BEHAVIOR or OTHER/UNKNOWN.
- Interpretation and Recommendations:
  - Gradient Boosting is the best-performing model, but it's still struggling with the class imbalance. Focus on improving its performance on the minority classes.

RANDOM FOREST

- Strengths:
  - High recall for WEATHER and TRAFFIC SIGNALS VIOLATION, meaning it captures a good proportion of these events.
- Weaknesses:
  - Extremely low precision for almost all classes. This means that many of the accidents it predicts as having a certain cause are actually due to something else.
  - The high recall combined with low precision suggests that the model is overgeneralizing and assigning a lot of accidents to these classes, even when they are not the true cause.
  - Lower F1-score, indicating that it's generally a poor model for this dataset.
- Interpretation and Recommendations:
  - Random Forest is not performing well and is likely overfitting or not capturing the underlying patterns.

KNN

- Strengths:
  - None of significance based on the data.
- Weaknesses:
  - Extremely poor performance across the board, especially for minority classes.
  - Very low recall values, indicating that it's not able to identify the causes of accidents effectively.
  - High computation, especially for prediction with no comparable performance to other models.
- Interpretation and Recommendations:
  - KNN is not suitable for this dataset. It's likely being overwhelmed by the high dimensionality and the large number of instances.
  - Model Comparison: Gradient Boosting outperformed Random Forest and KNN on the test set.

RECOMMENDATIONS

- Based on the Model Results, we recommend the following to the City of Chicago and the Vehicle Safety Board:
  - Focus on driver education and awareness campaigns to address improper driving behaviors (targeting DRIVER BEHAVIOR):
  - Investigate and reduce the ambiguity of crash data collection to better understand "Other/Unknown" causes:
  - Implement targeted interventions for areas with high traffic signal violation rates (addressing TRAFFIC SIGNALS VIOLATION):
  - Improve road maintenance during adverse weather conditions (addressing ROAD CONDITIONS and WEATHER):
  - Enhance data collection: Integrate more features.

FUTURE WORK

- **Address Class Imbalance More Aggressively:** Experiment with oversampling and undersampling techniques, as well as cost-sensitive learning.
- **Feature Engineering:** Explore new features based on domain knowledge and interactions between existing features.
- **Explore Other Models:** Consider other machine learning models, such as ensemble methods specifically designed for imbalanced data (e.g., EasyEnsemble, BalancedRandomForest).
- **External Data:** Incorportate data.
- **Geospatial Analysis:** Perform geospatial analysis to identify high-risk locations and patterns.
- **Causal Inference:** Explore causal inference techniques to better understand the causal relationships between factors and accident causes.