# Predicting the Primary Cause of Car Crashes.

Phase 4 Project by Alexandra Yakovleva.

# Business Problem.

Stakeholder: City of Chicago.

The business goal: develop a data-driven Driver Awareness Campaign to reduce accidents in business and residential areas.

# Proposed Solution.

Use publicly-available Chicago Crashes Dataset.

Develop a predictive model using ML tools.

Analyze the key features contributing to the accidents.

# Car Crashes Datasets.

CAR CRASHES – CRASHES.

CAR CRASHES – VEHICLES.

CAR CRASHES – PEOPLE.

# Traffic Crashes - People

- Dataset size: 1877321 rows x 29 columns.
- Age, sex, physical condition.
- Passengers, Pedestrians, Cyclists.
- vehicle_id, crash_record_id.

CHICAGO
DATA PORTAL

# Traffic Crashes - Vehicles

- Dataset size: 1743922 rows x 71 columns.
- Vehicle Type, Make, Model.
- Vehicle Use.
- vehicle_id, crash_record_id.

CHICAGO
DATA PORTAL

# Traffic Crashes - Crashes

- Dataset Size: 854910 rows x 48 columns.
- Crash scene related information: weather, roadway, signals, etc..
- Primary and Secondary Crash Causes.
- crash_record_id.

CHICAGO DATA PORTAL

# Data Transformation Strategy.

Identify features, relevant to Primary Contributory Cause.

Explore and clean subset with the features.

Merge Datasets.

Perform EDA.

# Dataset Transformation Overwiew.

- People:1877321 rows,  29 columns.
- Vehicles: 1743922 rows, 71 columns.
- Crashes: 854910 rows, 48 columns.
- Combined Dataset: 1541625 rows, 60 columns.

# Dataset Key Limitations.

Inconsistent Data Logging.

Imbalanced

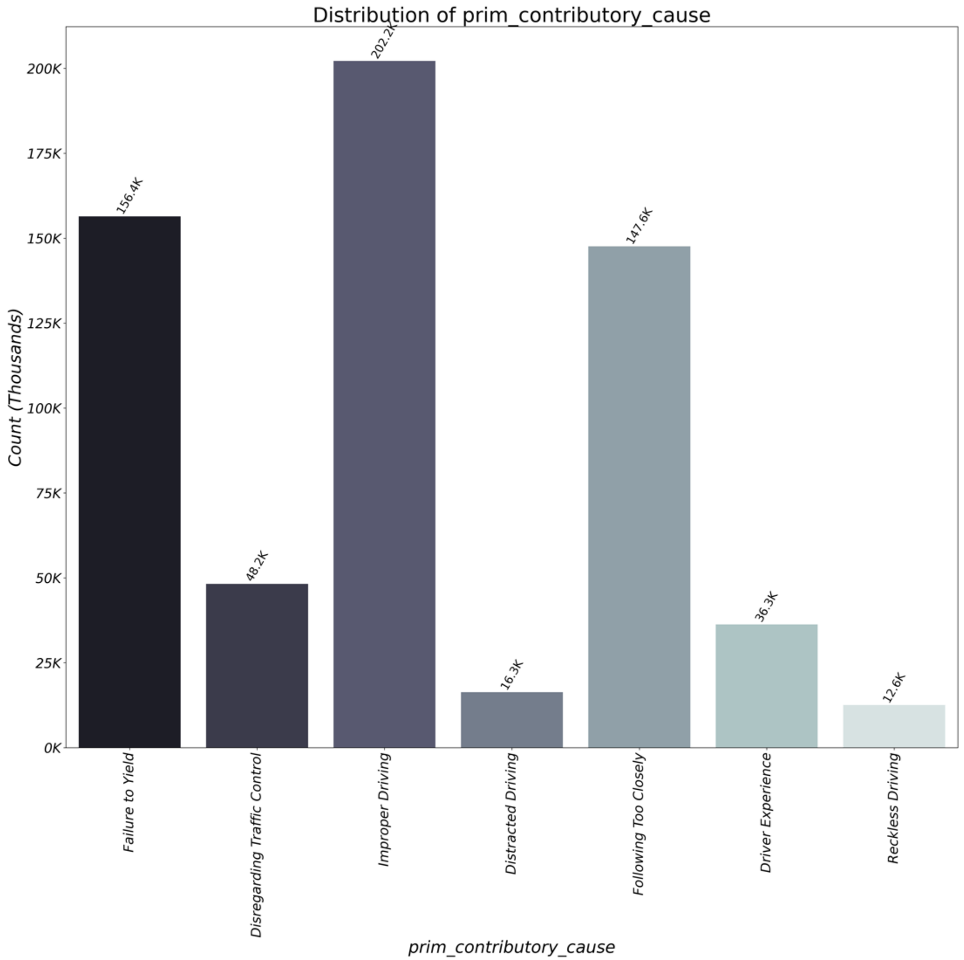
Data Representation.

Hight Dimensionality.

# Preparing Dataset for ML.

- Reducing both features and primary contributory causes: 619627 samples x 22 features, with target containing 7 classes.
- Use 15% of the data as a holdout set.
- Use 45% of the data for training and testing (70%-30%).

# Using F1 Score as a Model Performance Metric.




*F1 score balances precision and recall, making it ideal for our imbalanced dataset.*

# Modelling Results.

| Model | Accuracy | F1 Score |
|---|---|---|
| KNN | 0.63 | 0.61 |
| Random Forest | 0.71 | 0.68 |
| Extreme Gradient Boosting | 0.71 | 0.7 |
| Deep Neural Network | 0.71 | 0.69 |

*Accuracy and F1 scores computed on test set.*

# Optimized XGB Model.

- Accuracy score 0.72.
- F1 Score 0.7.
- The optimized XGB model achieves a balanced trade-off between precision and recall, effectively handling class imbalances.
- The model captures most but not all of the patterns.

*Accuracy and F1 scores computed on holdout set.*

# Optimized XGB Model: Confusion Matrix.

# Optimized XGB Model: False Discovery Ratio.

| Primary Contributory Cause | FDR |
|---|---|
| Disregarding Traffic Control | 0.34 |
| Distracted Driving | 0.47 |
| Driver Experience | 0.51 |
| Failure to Yield | 0.32 |
| Following Too Closely | 0.16 |
| Improper Driving | 0.30 |
| Reckless Driving | 0.56 |

*A lower FDR indicates fewer incorrect predictions for the cause.*

# Optimized XGB Model: Interpretability.

| Feature | Value | Importance (relative units) |
|---|---|---|
| First Crash Type | Rear End | 280.237396 |
| Driver Action | Improper Maneuver | 89.850182 |
| First Crash Type | Angle | 73.313103 |
| Driver Action | Distance | 62.447044 |
| Driver Action | Disregarding Controls/Signs | 56.923775 |
| Driver Action | Distraction | 41.656368 |
| First Contact Point | Rear | 22.867151 |
| First Crash Type | Turning | 21.865906 |
| First Crash Type | Sideswipe Same | 20.575909 |
| Driver Action | Other | 17.656883 |

# Recommendations.

- **Target Rear-End collisions.**

  Focus on campaigns that promote safe following distances, especially in high-traffic areas.

- **Address Improper Maneuvers.**

  Educate drivers about common improper maneuvers, such as unsafe lane changes, abrupt stops, and risky turns.

- **Mitigate Distracted Driving.**

  Stress the dangers of texting, phone use, and other distractions while driving.

# Future Work (Improving Performance).

- The combined dataset serves as a robust foundation for further analysis and modeling.
- Consider using Deep Neural Network (similar performance, better model interpretability).
- Reduce features based on DNN’s weights.
- Re-evaluate using DNN or XGB.