

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

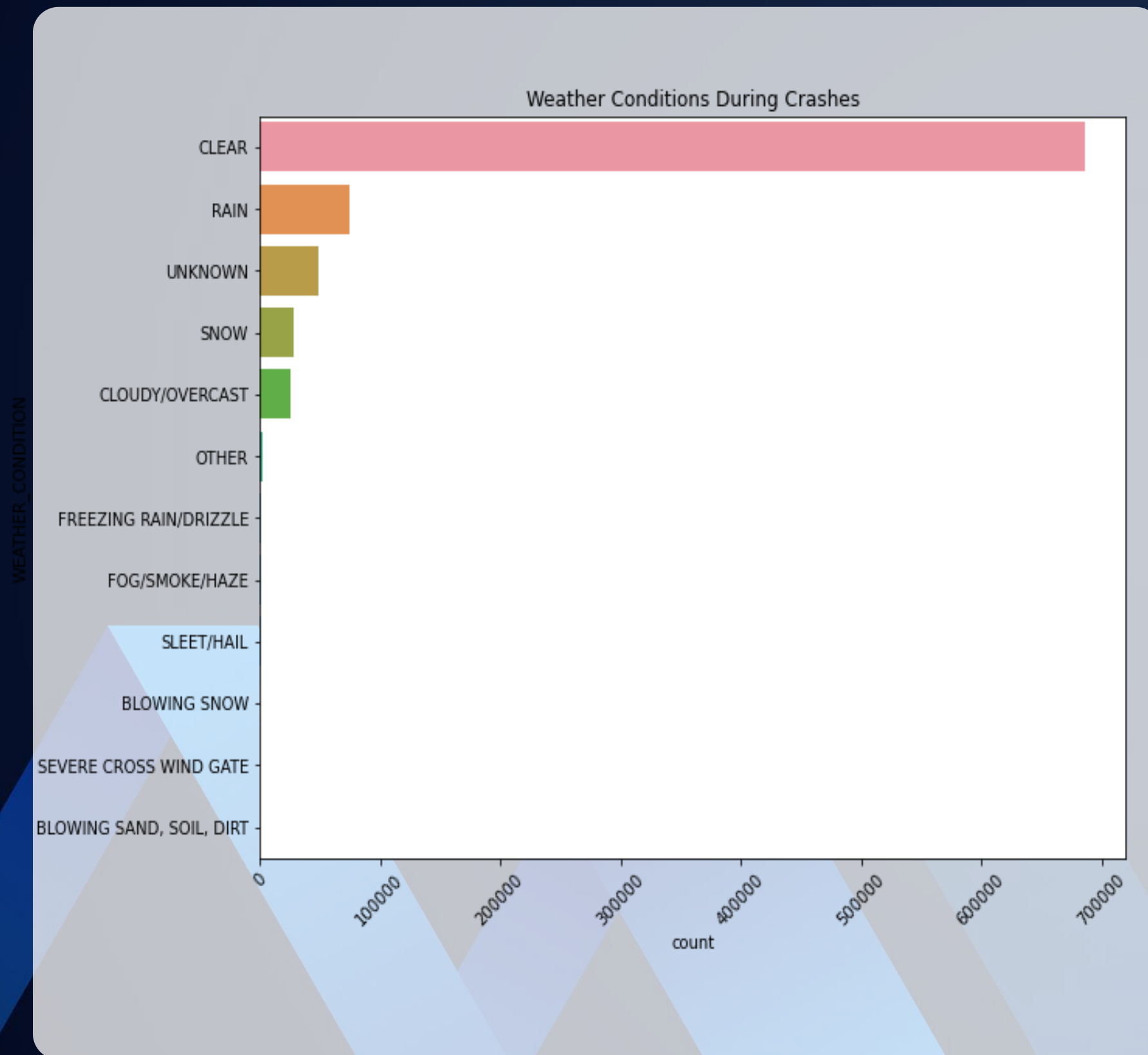
The Chicago Traffic Crashes dataset consists of around 880,000 rows and 48 columns. This dataset captures various details about road accidents, such as the date of the crash, weather and lighting conditions, types of vehicles, and details about injuries sustained by people involved.

The key variables we examined are:

- Weather conditions during the crash
- Lighting conditions (e.g., daylight, nighttime, artificial light)
- Traffic control devices present at the crash site (e.g., traffic signals, stop signs)
- Crash types, such as collisions with fixed objects or other vehicles
- Injury types and their severity

Weather Conditions

Adverse weather conditions (e.g., rain, snow) were found to be a common factor in crashes, but they did not always result in severe injuries. However, dry weather had a higher overall count of accidents due to increased traffic volumes.



Lighting Conditions

From the analysis, a significant number of accidents occurred during daylight, which can be attributed to higher traffic volume. However, accidents that occurred in nighttime or under poor lighting conditions were more likely to result in severe injuries.

Crash Types & Injury Severity

From the data it was clear that:

- Collisions with other vehicles at intersections were the most frequent crash types.
- Crashes involving fixed objects (e.g. poles, barriers) tended to have a higher injury severity compared to vehicle-to-vehicle crashes.
- Pedestrian involvement in crashes was strongly associated with higher injury severity.

Traffic Control Devices

- Accidents were more frequent in areas where no traffic control devices (e.g., signals, signs) were present. However, when crashes occurred at signalized intersections, they often resulted in more serious injuries.
- Improper functioning of traffic signals or driver negligence in observing signals also contributed significantly to intersection crashes.

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.

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.

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.

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.

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 FINDINGS

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. 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.