Analyzing Chicago Car Crashes:

Unveiling Insights and Enhancing Road Safety

Introduction

Car crashes are a pressing concern in any city, impacting the lives of individuals, families, and communities. Understanding the factors that contribute to these incidents is crucial in developing effective strategies to improve road safety.

In this presentation, we will dive deep into the rich dataset of car crashes in Chicago, exploring the various dimensions of the problem, identifying key patterns and trends, and uncovering valuable insights that can guide policy-makers, law enforcement agencies, and other stakeholders in their efforts to reduce accidents and enhance traffic safety.

Project Overview

The project aims to analyze and gain insights from the extensive dataset of car crashes in Chicago. By delving into this valuable data, our objective is to uncover patterns, identify contributing factors, and provide actionable recommendations to enhance road safety and reduce the frequency and severity of accidents in the city.

The Data

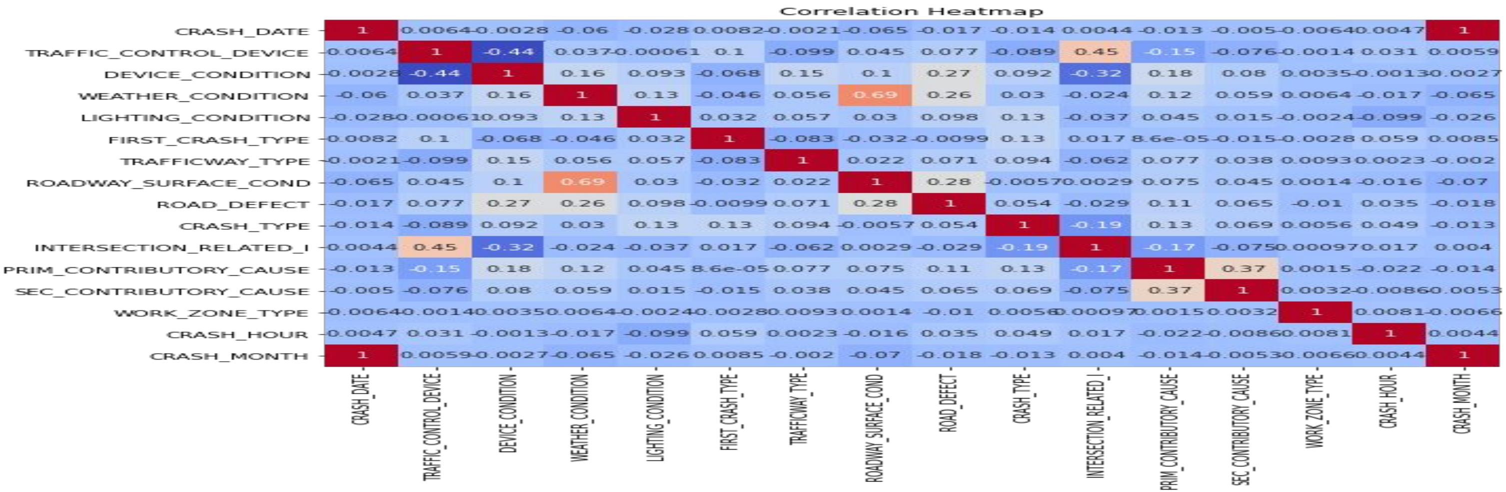
The dataset contains information about car crashes in Chicago, including various attributes such as crash date and time, traffic control devices, weather conditions, road defects, contributing causes, and more. It consists of several columns providing details about each crash incident.

The dataset consists of 722,467 instances/rows and 18 columns/attributes.

Correlation was done to check for the column that were highly correlated to the target variable(primary contributory cause of car crashes).

Correlation Heatmap

The features that are highly correlated with the target feature are: device condition, weather condition, trafficway type, secondary contributory cause and road defect.



Modeling

Three models were built for this project.

A Logistic Regression model was built as the baseline model. The logistic regression model yielded low accuracy and F1 score of 40.4% and 27.3% respectively, suggesting that the model's performance in predicting the primary contributory cause of car accidents is not satisfactory.

The second model was a Random Forest Classifier that slightly increased the performance of the model to an accuracy of 45.7% and F1 score of 34.9%.

The final model was a classification model using Extreme Gradient Boosting with tuned hyperparameters. This model yielded better performance overly as shown:

Accuracy: 45.8% , F1 Score: 36.3% , Precision: 39.0% , Recall: 45.8%

Results

Overall, predicting the primary contributory cause of car accidents based on the available features is a challenging task. It requires a deeper understanding of the underlying factors contributing to accidents, as well as the availability of more comprehensive and accurate data. Further model refinement and exploration of advanced techniques may be necessary to achieve a higher level of predictive accuracy and reliability.

Conclusion and Recommendations

The model would be useful in the following situations:

Investigating Contributing Factors: The model predictions can assist stakeholders in understanding the contributing factors behind accidents.

Evaluating Policy Interventions: If the stakeholders implement new policies or interventions to improve road safety, the model can be used to assess the effectiveness of those measures.

Situations where the model would not be useful:

Real-Time Accident Prediction: The model's predictions are based on historical data and existing variables. It may not be suitable for real-time accident prediction or immediate response scenarios where data on certain variables may not be available or rapidly changing.

Individual Driver Behavior: The model focuses on broader patterns and factors contributing to accidents at a population level. It may not capture individual driver behavior or specific instances of reckless driving, distracted driving, or other driver-related factors.