BUSINESS OVERVIEW

The City of Chicago has a vested interest in the safety and well-being of its residents. By analyzing car crash data, the city can gain valuable insights into the factors that contribute to accidents within its jurisdiction. This information can assist city officials and policymakers in making informed decisions regarding traffic infrastructure, road design, traffic regulations, and enforcement strategies. Understanding the prevalent causes of accidents can guide the city in implementing targeted measures to reduce accidents, improve traffic flow, and enhance overall road safety. Additionally, the analysis may reveal patterns or correlations between accidents and specific locations, weather conditions, or other variables, allowing the city to focus resources on areas that require attention and intervention.

PROJECT OBJECTIVES

1. Develop a predictive model: The primary objective of the project is to develop a predictive model that can accurately predict car crash occurrences in Chicago. The model considered various factors such as traffic control devices, weather conditions, road defects, and contributing causes to provide insights into the likelihood of accidents.
2. Improve road safety: The project aims to contribute to improving road safety in Chicago by identifying high-risk areas, understanding contributing factors, and implementing targeted interventions. The model can assist stakeholders in identifying locations or circumstances prone to accidents, allowing them to prioritize resources and interventions effectively.
3. Identify key contributing causes: The project seeks to identify the primary contributory causes behind car crashes in Chicago. By analyzing historical data and examining the importance of different variables, the model can provide insights into the factors that significantly impact road safety. This information will help stakeholders focus their efforts on addressing the most critical causes.

DATA UNDERSTANDING

Dataset Description: 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.

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

Attribute Types: The dataset primarily consists of categorical attributes, such as traffic control device, device condition, weather condition, lighting condition, crash type, intersection-related flag, roadway surface condition, road defect, primary contributory cause, and secondary contributory cause. Additionally, there are a few numerical attributes, including crash hour, crash day of the week, and crash month.

Missing Values: Some columns, such as intersection-related flag, work zone type, and location, have missing values.

Target Variable: The primary contributory cause column appears to be the target variable, indicating the primary cause of each car crash incident.

DATA CLEANING

1. Handling Missing Values: There were missing values in the columns 'INTERSECTION_RELATED_I' and 'WORK_ZONE_TYPE' which were filled.
2. Handling Duplicates: Duplicated values were removed from the dataset.
3. Encoding Categorical Variables: Categorical column variables were convered into numerical representations suitable for modeling. Label encoding was used.
4. Removing Unnecessary Columns: Any columns that are irrelevant or redundant for the modeling task were removed. These columns may not contribute meaningful information or may introduce noise into the model.
5. Correlation was also done to check for the column that were highly correlated to the target variable.

MODELING

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. This indicates that the current model may not be capturing the underlying patterns and relationships effectively.

Alternative models were therefore be built to improve the performance.

The second model was a Random Forest Classifier that slighltly 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 overally as shown:

Accuracy: 0.45832467262523385

F1 Score: 0.3630366615661475

Precision: 0.39000601668021057
Recall: 0.45832467262523385

RESULTS

The logistic regression model trained on a 10% sample of the Chicago car crash dataset achieved an accuracy of 40.4% and an F1 score of 27.3%. Although these metrics indicate some level of predictive performance, they are still relatively low. The model struggled to accurately classify the primary contributory cause of car accidents based on the selected features, such as device condition, weather condition, trafficway type, secondary contributory cause, and road defect. The confusion matrix revealed a diverse range of predicted classes, with some classes having very few correct predictions. This suggests that the model may be biased towards certain classes and lacks the ability to effectively differentiate between multiple contributory causes.

To further improve the model's performance, alternative approaches such as ensemble methods can be explored. One such method is XGBoost, which was also trained on the 10% sample data. The XGBoost model achieved a slightly higher accuracy of 45.01% and an improved F1 score of 35.46%. The precision and recall values were calculated to be 37.21% and 45.01%, respectively. Although the model's performance remains suboptimal, XGBoost demonstrated a slight improvement over logistic regression in capturing the complex relationships between features and the primary contributory cause. Nonetheless, addressing the class imbalance and fine-tuning hyperparameters could potentially enhance the model's predictive capability.

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

Situations where predictions made by the model would be useful:

Investigating Contributing Factors: The model predictions can assist stakeholders in understanding the contributing factors behind accidents. By examining the importance of various variables in the prediction, stakeholders can identify the primary causes or conditions leading to accidents. This knowledge can be used to develop specific strategies, such as improving road infrastructure, enhancing traffic control devices, or implementing educational campaigns to address those factors effectively.

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. By comparing the predicted outcomes before and after implementing the changes, stakeholders can evaluate the impact of their interventions and make data-driven decisions for further improvements.

Situations where the model predictions 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. For real-time accident prediction, stakeholders would require a different approach, such as real-time sensor data, traffic cameras, or predictive models specifically designed for immediate responses.

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. Stakeholders should consider other approaches, such as driver education programs, traffic law enforcement, or telematics-based solutions, to address individual driver behavior.