JIMWAMAE / Chicago_car_crashes_project Public

☆ 0 stars 0 forks

☆ Star Watch

Code Issues Pull requests Actions Projects Wiki Security Insights Se

main ▾ ...

JIMWAMAE ... 6 minutes ago

View code

README.md

Chicago_car_crashes_project

Author- Jimcollins wamae
LinkedIn - Jim Wamae

Overview

Our objective was to develop inferential classification models for the Vehicle Safety Board of Chicago using cleaned and formatted data on crashes, vehicles, and individuals involved in accidents from 2016 to 2020. Our goal was to categorize the primary contributory causes of car accidents into two groups: avoidable and unavoidable.

Business Understanding

The Vehicle Safety Board of Chicago (our stakeholder) has entrusted us with the task of conducting a comprehensive analysis of car crash data to enhance their understanding of the causes and factors influencing avoidable and unavoidable accidents in the region. Our project aims to delve deep into the available data, spanning from 2016 to 2020, to identify the primary contributory causes of crashes and uncover valuable insights that can drive actionable recommendations.

By leveraging advanced data analysis techniques and employing various classification models, our objective is to decipher the underlying patterns, correlations, and trends within the crash data. This analysis will enable us to differentiate between avoidable and unavoidable accidents, shedding light on the key factors that contribute to each category.

Through this project, we seek to address the following crucial business questions: What are the primary causes of avoidable accidents in Chicago? Which age groups are more susceptible to being involved in avoidable crashes? Are there specific road conditions or locations that pose a higher risk for avoidable accidents?

By gaining a deeper understanding of the factors influencing avoidable accidents, we aim to provide strategic recommendations that will assist the Vehicle Safety Board in formulating effective interventions and policies. These recommendations may include targeted driver education campaigns, infrastructure improvements in high-risk areas, or other measures aimed at reducing the occurrence of avoidable accidents and promoting overall road safety.

Ultimately, our project endeavors to contribute to the mission of the Vehicle Safety Board by providing data-driven insights and actionable recommendations. By equipping the Board with a comprehensive understanding of the causes and factors associated with avoidable accidents, we can collectively work towards creating a safer driving environment for the residents of Chicago and reducing the human and economic toll of car crashes in the region.

Data Understanding

The Vehicle Safety Board would like to better understand the causes of crashes in the Chicago area. That way, they can focus their campaign on potentially preventing some of those crashes in the future. We used data from the City of Chicago Data Portal, which contains information about Chicago Car Crashes from January 2016 to December 2020. The target variable was “primary contributory cause” as recorded by the police officer at the scene of the crash. Some of our inferential variables “defect road”, ”bad road conditions”, and “obscured vision” will help us in our analysis to see if a crash was preventable.

Data Preparation

We combined 3 different datasets that we found on the City of Chicago website. Those datasets were crashes, people, and vehicles.

We conducted thorough data preparation on the provided datasets from the City of Chicago, which included vehicles, people, and crashes data. Our goal was to clean, transform, and integrate the data to ensure its quality and suitability for analysis. By addressing missing values, inconsistencies, we enhanced the reliability and accuracy of the data. Through feature engineering and data integration, we created new variables and established meaningful relationships between the datasets. The data preparation process forms the foundation for our subsequent analysis and modeling, enabling us to derive valuable insights and make informed recommendations to the Vehicle Safety Board of Chicago.

Modeling

Our target had a distribution of .75 to .25. The 2 categories in the target are 1 for preventable and 0 for not preventable. Because our class balance was 2:1 , we did not SMOTE the minority class.

We modeled the data through iterative modeling. The following are the models that were used and their order.

1. Decision Tree: Decision trees are often chosen as the initial model due to their simplicity and interpretability. They provide a baseline performance and can help in understanding the underlying patterns and relationships in the data. Decision trees are relatively easy to implement and provide a good starting point for more complex models.

2. Gradient Boosting: Gradient boosting is a powerful ensemble method that combines multiple weak models (decision trees in this case) to create a stronger predictive model. It improves upon the shortcomings of individual decision trees by reducing bias and variance. Gradient boosting often produces better predictive performance compared to a single decision tree and can capture more complex patterns in the data.

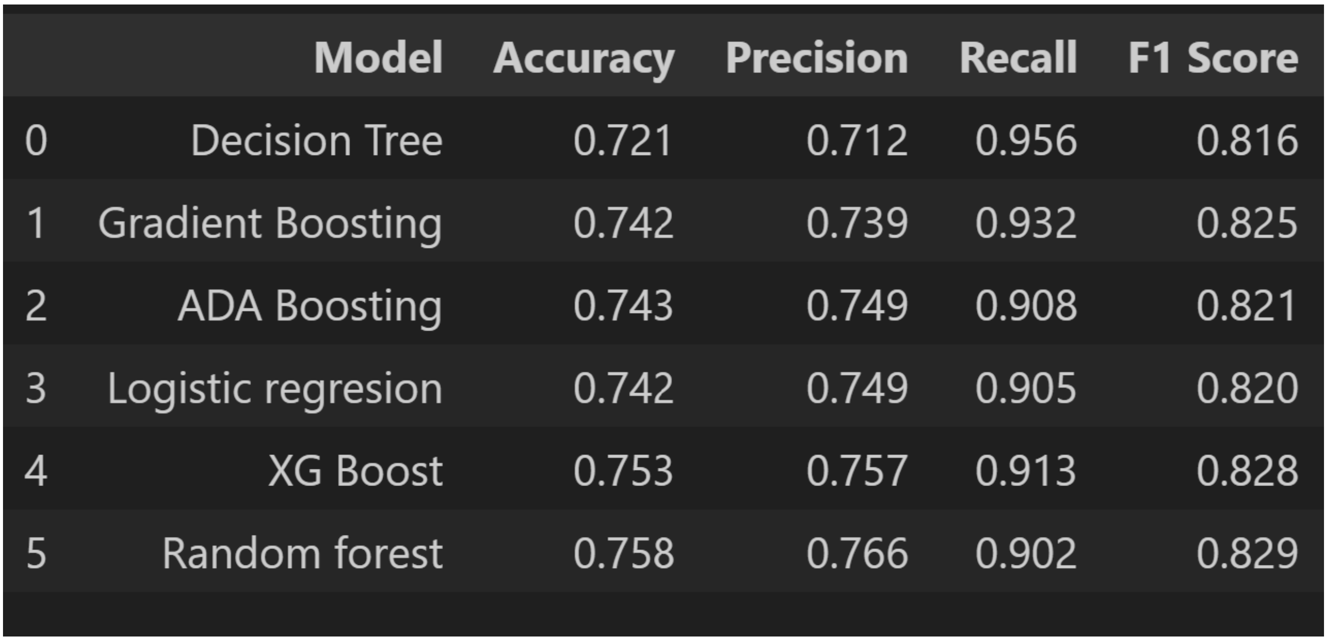
3. ADA Boosting: ADA boosting is another boosting algorithm that iteratively adjusts the weights of misclassified instances to build a strong model. It focuses on correcting the mistakes made by the previous models, thereby improving overall performance. ADA boosting can be effective in handling imbalanced datasets and can further enhance the predictive accuracy of the model.

4. Logistic Regression: Logistic regression is a popular model for binary classification problems. It is used when the relationship between the predictors and the target variable is expected to be linear. Logistic regression provides interpretable coefficients and can help in understanding the impact of each predictor on the outcome. It is often included in the model selection process to compare the performance of linear models with more complex ones.

5. XG Boost: XG Boost is an optimized implementation of gradient boosting that offers faster computation and better handling of large datasets. It is known for its efficiency and scalability, making it suitable for handling complex problems with a large number of features. XG Boost often provides competitive performance and can be used as an alternative or complement to other boosting algorithms.

6. Random Forest: Random forest is an ensemble model that combines multiple decision trees and aggregates their predictions. It helps in reducing overfitting and improving generalization. Random forest models are robust and can handle a wide range of data types and feature interactions. They are often used as a benchmark or final model to compare against other algorithms and evaluate overall performance.

From the iterative model the models performed as such from the table below.



	Model	Accuracy	Precision	Recall	F1 Score
0	Decision Tree	0.721	0.712	0.956	0.816
1	Gradient Boosting	0.742	0.739	0.932	0.825
2	ADA Boosting	0.743	0.749	0.908	0.821
3	Logistic regresion	0.742	0.749	0.905	0.820
4	XG Boost	0.753	0.757	0.913	0.828
5	Random forest	0.758	0.766	0.902	0.829

The target variable was categorized into two classes: avoidable and unavoidable accidents.The best model overall was our final model with an accuracy score of 75% with the least accuracy score in our baseline model the decision tree classifier.

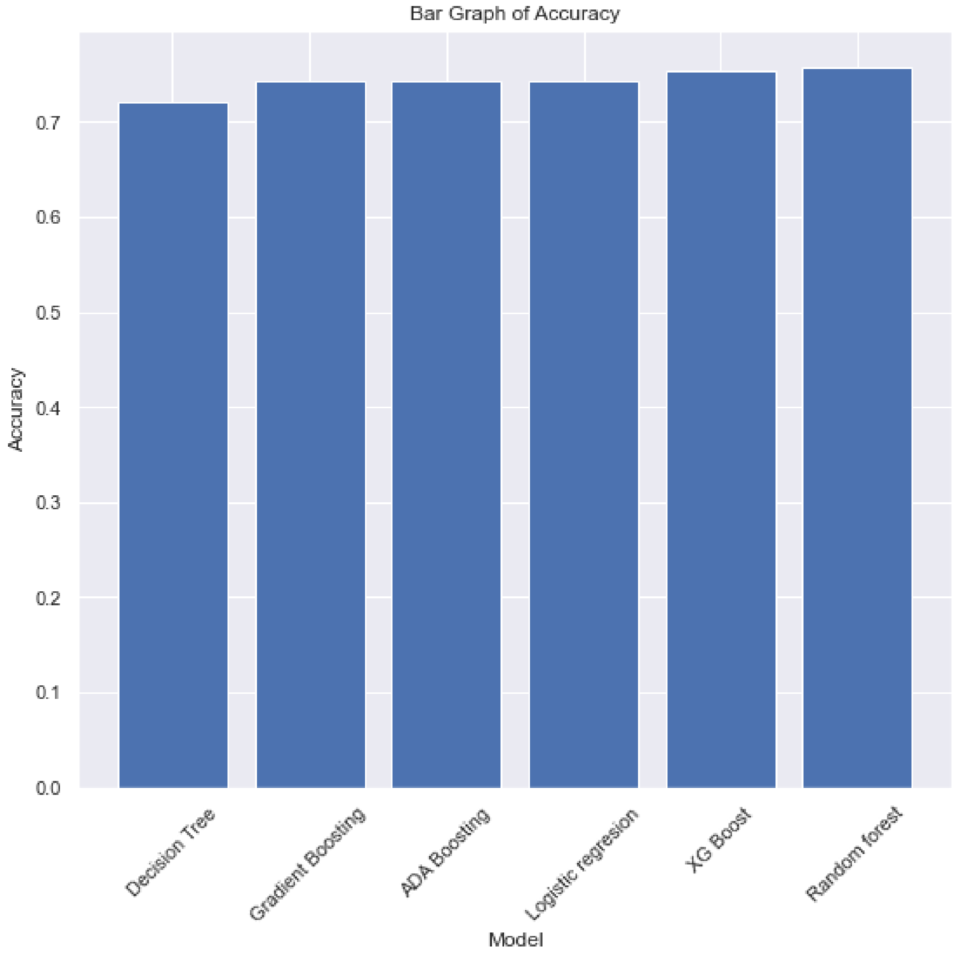
Evaluation

The Random Forest model achieved the following classification metrics:

Precision: 0.7663 Recall: 0.9021 Accuracy: 0.7580 F1 Score: 0.8287

These metrics indicate that the model performed well in predicting both avoidable and unavoidable accidents. The high precision suggests that the model had a low rate of false positives, accurately identifying avoidable accidents. The high recall indicates that the model had a low rate of false negatives, successfully capturing a large portion of actual avoidable accidents. The accuracy score reflects the overall correctness of the model's predictions, while the F1 score balances precision and recall.

Based on accuracy we can view the overal model perfomance compared to previous models.



These results suggest that the random forest model can effectively identify the causes of accidents in Chicago, making it a valuable tool for reducing accidents by targeting specific risk factors. It provides a reliable framework for identifying avoidable and unavoidable accidents, enabling policymakers and relevant stakeholders to allocate resources and implement targeted interventions to mitigate the identified causes.

By leveraging the insights gained from the Random Forest model, it is possible to identify the key factors contributing to accidents in Chicago. This information can be used to implement targeted measures and interventions aimed at reducing the occurrence of accidents and improving road safety in the city.

Conclusions

Based on our analysis, we have several recommendations to Vehicle Safety Board of Chicago to improve road safety and reduce preventable crashes in the Chicago area.

1. Fix damaged or defective roads: We recommend prioritizing the repair of roads in hot spot areas that have a higher frequency of non-preventable crashes. By addressing infrastructure issues, such as potholes or inadequate signage, we can create safer road conditions and minimize the occurrence of such incidents.

2. Invest in an online driver and behavior education campaign: Implementing an online driver education program can have a significant impact on reducing preventable crashes. This approach is not only affordable but also easily accessible to a wide audience. By providing educational resources and promoting safe driving practices, we can enhance driver awareness and decision-making skills.

3. Target the younger audience (age 20-39): Our analysis indicates that drivers within the age range of 20-39 accounted for a significant proportion of preventable crashes. Therefore, we recommend tailoring the driver education campaign to specifically target this demographic. By focusing on this age group, we can effectively address their unique driving behaviors and contribute to a substantial reduction in preventable accidents.

4. Increase traffic policing during the hours for 2 pm - 6pm and moreso on weekends. Most accidence seem to happen during this time period. Targeting peak traffic hours: By focusing on the time period of 2 pm - 6 pm, when traffic congestion is typically high, increased traffic policing can help manage traffic flow more efficiently. This can reduce the likelihood of accidents caused by reckless driving, speeding, or aggressive behavior during peak hours.

By implementing these recommendations, we believe we can make substantial progress in reducing preventable crashes in the Chicago area. It is crucial to allocate resources and collaborate with relevant stakeholders to ensure the successful implementation of these measures. Together, we can create a safer road environment and work towards the goal of minimizing accidents and their associated consequences.

For More Information

See the full analysis in the Jupyter Notebook

Releases

No releases published
Create a new release

Packages

No packages published
Publish your first package

Languages

- ● **Jupyter Notebook** 100.0%