

djdrjeff /  
phase3project\_KelvinNjenga

<> Code

Issues

Pull requests

Actions

Projects

Wiki

Security

Insights

Set

☆ 0 stars

🍴 0 forks

👁 1 watching

🔗 Branches

🏠 Activity

🏷 Tags

🌐 Public repository

1 Branch

0 Tags

Go to file

t









Go to file

+



Add file ▾

Code

⋮

 djdrjeff	PDF Update	fd93e6d · now	
 Dataset link.txt	Create Dataset link.txt	22 minutes ago	
 Presentation1.pdf	PDF Update	now	
 Presentation1.pptx	PDF Update	now	
 README.md	Update README.md	3 minutes ago	
 Traffic_Crashes_KelvinNjen...	Update Traffic_Crashes_KelvinNje...	4 days ago	
 ~\$Presentation1.pptx	Power point and ipynb update	8 hours ago	

📖 README

# phase3project\_KelvinNjenga

## Project Overview:

This project focuses on analyzing traffic crash data from the City of Chicago to predict the primary factors contributing to accidents. The objective is to support organizations like Vehicle Safety Boards and city administrations in identifying trends and implementing strategies to reduce traffic accidents.

## Dataset

Source: The dataset is obtained from the City of Chicago's Traffic Crashes dataset.

## NB: To access the Dataset, the steps below need to be followed

1. On the git repository there is a .txt document which has a ink to the dataset used for this project.

2. Click on the link it will take you to a google drive folder where the data is residing

3. The above was done since the data was too big to fit in the git repository.

## Modeling Process

---

### Baseline Model:

---

Logistic Regression was used as the initial baseline model.

### Feature Selection:

---

Features were chosen based on their correlation with the target variable and domain expertise.

### Model Comparison:

---

Various models, including Logistic Regression, Ridge Classifier, Random Forest, and Gradient Boosting, were trained and evaluated.

### Hyperparameter Tuning:

---

Hyperparameters were optimized using RandomizedSearchCV to enhance model performance.

### Final Model:

---

Gradient Boosting emerged as the top-performing model, selected based on its accuracy and other evaluation metrics.

## Results

---

Best Model: Gradient Boosting

Accuracy: 1.00%

Key Features: The most important features contributing to the model's predictions were Weather\_Condition\_Mode, Speed\_Weather\_Interaction, and Is\_Weekend.

## Insights and Recommendations

---

### Traffic Management:

---

The analysis indicates that weather conditions and speed limits play a critical role in contributing to crashes. Enforcing stricter speed regulations during adverse weather conditions could help reduce accident rates.

### Policy Changes:

---

Prioritize enhancing road infrastructure and improving traffic control systems in areas identified with high accident frequencies.

### Further Analysis:

---

Conduct additional studies on the influence of time of day and specific locations (latitude and longitude) on crash rates to enable more targeted and effective interventions.

## Modeling Approach & Evaluation Metrics

1. Accuracy: The proportion of correct predictions made by the model.
2. Precision: The ability of the model to correctly identify positive instances.
3. Recall: The ability of the model to capture all positive instances.
4. F1-Score: The harmonic mean of precision and recall, balancing the two.
5. Confusion Matrix: A detailed breakdown of the model's performance across different classes.

## Model Comparison

The Model combines the strengths the flexibility of Decision Trees, allowing it to model complex non-linear relationships, handle interactions between features and generalize well on unseen data. Its performance were further enhanced through Hyperparameter Tuning, which likely led to it outperforming the other models tested.

## Data Modeling

Several machine learning models will be trained and evaluated, including: Logistic Regression Regularized Logistic Regression Decision Trees Random Forest Gradient Boosting

## Best Model Establishment

Best model in our case as indicated on the chart is Gradient Boosting Gradient Boosting emerged as the top-performing model for this problem primarily due to the large size of the dataset. It leverages the strengths of ensemble learning and the flexibility of Decision Trees, enabling it to capture complex non-linear relationships, handle feature interactions effectively, and generalize well to unseen data. Additionally, its performance was further optimized through hyperparameter tuning, which likely contributed to its superior results compared to the other models.

## Conclusion

1. The analysis of the traffic crash data has provided valuable insights into the primary contributory causes of accidents and the critical factors influencing road safety. By identifying the top features and causes associated with accidents, we can make informed recommendations to enhance road safety in the City of Chicago
2. The insights gained from this analysis provide a clear pathway to making the city's roads safer for everyone by reducing traffic accidents, enhancing road safety, and ultimately save lives.

## Nest Steps

Putting into effect the recommendations made so as to enhance the road safety and reduce incidences and



### Releases

No releases published

[Create a new release](#)

---

## Packages

No packages published

[Publish your first package](#)

---

## Languages

● Jupyter Notebook 100.0%