Prediction of Traffic Accidents with Logistic Regression, Decision Tree, and Random Forest Models

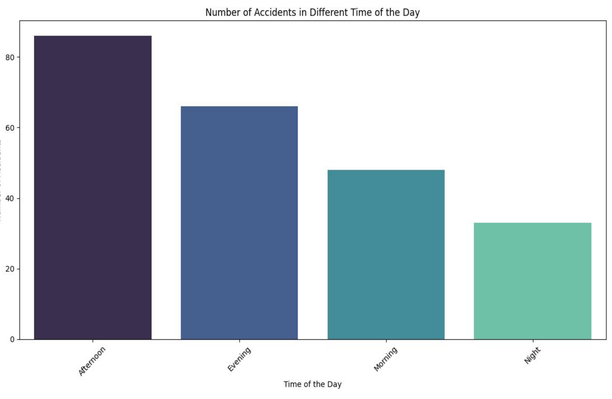
Kyi Lei Aye | NorthEast Big Data Innovation Hub | National Student Data Corps | Transportation Data Science Project Accelerator

Traffic accidents pose significant threats to public safety, resulting in loss of life, economic costs, and societal impacts. This research aims to develop a predictive framework for identifying potential traffic accidents using machine learning models based on various contributing factors, such as weather conditions, road types, time of day, traffic density, speed limit, number of vehicles, alcohol consumption of drivers, experience of drivers, and road conditions.

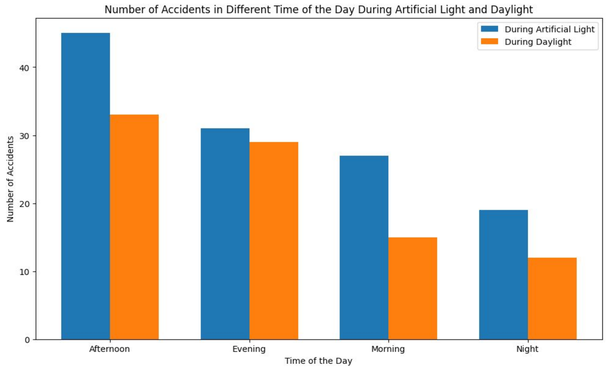
Models such as Logistic Regression, Decision Tree, and Random Forest are employed with an optimization to enhance the accuracy and performance. Key performance metrics including precision, recall, F1-score, are used to evaluate the models' effectiveness in predicting traffic accidents. The findings demonstrate that the Random Forest Model outperforms other classifiers in terms of predictive accuracy and generalization. This research contributes to proactive traffic management systems and improve road safety by enabling real-time accident prediction.

Exploration and Visualization

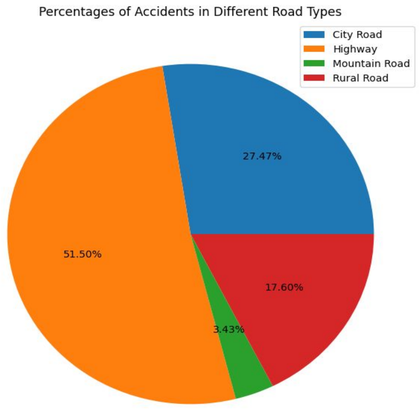
- Most accidents happened in the afternoon followed by the evening; by contrast, lowest rate at night.



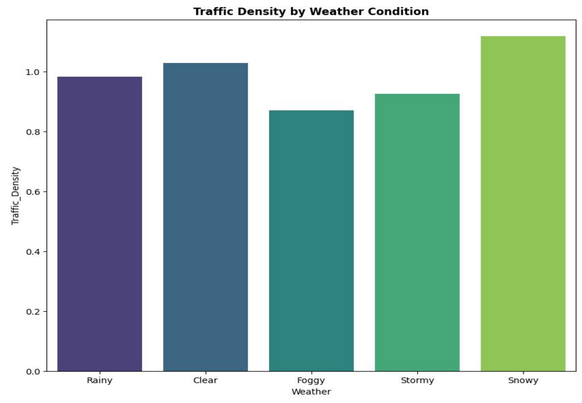
- Accidents were less likely to occur in daylight than during artificial light.



Exploration and Visualization Cont.



- More than half of the accidents (51.5%) happened on highways.
- Accidents least likely to happen on mountain roads (3.43%)



- Traffic Density was the highest when the weather was snowy or clear. At those weathers, accidents were likely to happen more.

Machine Learning Models & Optimizations For Prediction of Traffic Accidents

Logistic Regression Model

$$g(z) = \frac{1}{1 + e^{-z}}$$

- Built a logistic binary classification regression model to predict the probability of vehicle accidents. [0 - No, 1 - Yes]
- Model's accuracy rate was around **51%** with 23 false negatives and 58 false positives.

Boosting Logistic Regression Model's Performance with Gradient Descent & Feature Scaling

$$J(\vec{w}, b) = \frac{1}{m} \sum_{i=1}^{m} y^{(i)} log(f_{\vec{w},b}(\vec{X}^{(i)})) + (1 - y^{(i)}) log(1 - f_{\vec{w},b}(\vec{X}^{(i)}))$$

Cost function of the logistic regression

Machine Learning Models & Optimizations Cont.

- Compute the loss and repeatedly find the best parameters w and b to promote model accuracy and performance

$$w_j = w_j - \alpha \frac{\partial}{\partial w_j} J(\vec{w}, b)$$

$$b = b - \alpha \frac{\partial}{\partial b} J(\vec{w}, b)$$

- After optimizing with gradient descent and feature scaling, the model's accuracy was increased to about **68%.**

```
Accuracy: 0.679
              precision    recall  f1-score   support

           0       0.75      0.83      0.79       117
           1       0.43      0.31      0.36        48

    accuracy                           0.68       165
   macro avg       0.59      0.57      0.57       165
weighted avg       0.65      0.68      0.66       165

Confusion Matrix:
[[97 20]
 [33 15]]
```

Decision Tree Model

$$IG(D_p, f) = I(D_p) - \frac{N_{left}}{N} I(D_{left}) - \frac{N_{right}}{N} I(D_{right})$$

f: feature split on
D_p: dataset of the parent node
D_{left}: dataset of the left child node
D_{right}: dataset of the right child node
I: impurity criterion (Gini Index or Entropy)
N: total number of samples
N_{left}: number of samples at left child node
N_{right}: number of samples at right child node

- Using the Tree module from Scikit-Learn package, a Decision Tree Classifier Model was built to predict the possibility of Traffic Accidents.
- The accuracy was around **57%** with 13 true positives and 81 false negatives.

Optimizing Decision Tree Classifier Model's Performance with Hyperparameter Tuning

- By tuning hyperparameters (such as entropy and gini), Decision Tree Classifier was improved to the accuracy of around **62%.**
- However, true positives values was decreased to 6, which can be the result of class imbalance in dataset.

```
Accuracy: 0.618
Classification Report:
              precision    recall  f1-score   support

           0       0.70      0.82      0.75       117
           1       0.22      0.12      0.16        48

    accuracy                           0.62       165
   macro avg       0.46      0.47      0.46       165
weighted avg       0.56      0.62      0.58       165

Confusion Matrix:
 [[96 21]
 [42  6]]
```

Machine Learning Models & Optimizations Cont.

Random Forest Model

- Random Forest Classifier Model, which is based on bootstrap (random sampling replacement) technique and random feature selection, was constructed along with balancing classes in dataset to predict the probability of accidents.
- The model was found to be accurate at a rate of approximately **71%**.

Enhancing Random Forest Model with Hyperparameter Tuning

- After adjusting the parameters in the model, the model's performance remains the same at about **71%.**

```
Accuracy: 0.709
Classification Report:
              precision    recall  f1-score   support

           0       0.71      0.98      0.83       117
           1       0.50      0.04      0.08        48

    accuracy                           0.71       165
   macro avg       0.61      0.51      0.45       165
weighted avg       0.65      0.71      0.61       165

Confusion Matrix:
 [[115   2]
 [ 46   2]]
```

Conclusion, Suggestions, & Future Directions

- Random Forest Classifier outperforms Decision Tree and Logistic Regression Classifiers in predicting traffic accidents.
- Yet, models are not significantly accurate at this point and required to improve more to better predict accidents.
- According to data analysis and observations, highway light systems should be developed and speed limit should be adjusted based on traffic density and weather conditions.

References

- https://nebigdatahub.org/dsrr/
- https://www.geeksforgeeks.org/how-to-tune-a-decision-tree-in-hyperparameter-tuning/
- https://www.kaggle.com/datasets/denkuznetz/traffic-accident-prediction
- https://www.geeksforgeeks.org/random-forest-hyperparameter-tuning-in-python/

Acknowledge

I really thank you to NorthEast Big Data Innovation Hub & National Student Data Corps, which host Transportation Data Science Project in collaboration with the U.S. Department of Transportation Federal Highway Administration.

Contact

ayekyilei2003@gmail.com
LinkedIn

RESEARCH POSTER PRESENTATION DESIGN © 2015
www.PosterPresentations.com