Chicago Car Accidents Analysis

Justin Lee

Overview

Analyst: Justin Lee

Stakeholder: Vehicle Safety Board of Chicago

Motive: Aims to understand the primary contributory causes of a car accident in order to reduce traffic accidents.

Data Understanding

Dataset features:

- 29 columns
- 2,005,876 rows
- Target variable column: DRIVER_ACTION


```
NONE                            568041
UNKNOWN                         406729
FAILED TO YIELD                 145118
OTHER                           143764
FOLLOWED TOO CLOSELY             93147
IMPROPER BACKING                 46726
IMPROPER TURN                    42036
IMPROPER LANE CHANGE             41055
IMPROPER PASSING                 35904
DISREGARDED CONTROL DEVICES      28288
TOO FAST FOR CONDITIONS          23312
WRONG WAY/SIDE                    6449
IMPROPER PARKING                 5856
OVERCORRECTED                    3230
EVADING POLICE VEHICLE           2473
CELL PHONE USE OTHER THAN TEXTING 2313
EMERGENCY VEHICLE ON CALL        1493
TEXTING                           626
STOPPED SCHOOL BUS                193
LICENSE RESTRICTIONS               69
Name: DRIVER_ACTION, dtype: int64
```

Baseline Model Evaluation

1) High overall accuracy (78%) but biased toward majority class
   a) 97% recall for majority class
2) Poor detection of minority class
   a) Class -2 has only 5% recall
   b) Class -1 has only 2% recall
3) Confusion matrix shows severe misclassification
   a) Class -2 misclassified as Class 0 in 33,462 cases
   b) Class -1 misclassified as Class 0 in 42,688 cases

Balanced Model Evaluation

1) Improved recall for minority classes -2 and -1
   a) Class -2 improved from 5% to 28%
   b) Class -1 improved from 2% to 61%
2) Lower accuracy trade-off
   a) Accuracy dropped from 78% to 49%
3) Overall more balanced predictions
   a) Class 0 recall dropped from 97% to 50% - fewer cases of the majority class are being correctly predicted

Recommendations

Balanced Model

In order to reduce traffic accidents it is best to prioritize recall and fairness of all traffic cases rather than accuracy of a model. A balanced performance across all causes should be prioritized over pure accuracy.

Next Steps

Random Forest Model

A Random Forest Model will help handle imbalanced data better than a single Decision Tree.

Add Environmental Data

We could consider adding external factors to our data. Weather data, traffic density, and road conditions would all be great factors to add to our data to help improve model accuracy.

Policy Recommendations

We could create policy recommendations from our model results. We could enforce stricter rules and regulations based on leading causes of traffic accidents.

Contact Information

Justin Lee

Email: jlee8150@gmail.com

LinkedIn:
https://www.linkedin.com/in/justinlee2023/