Car Crash Severity Project

Soya Han
2020

Introduction

With advancements in technology and the constant need to be connected, more and more drivers are driving distracted. In 2018, the total number of fatal crashes was 33,564. Of those 33,564 crashes, 2.628 were distraction-affected fatal crashes was and the total number of crashes were the cellphone was in use was 349 (Insurance Information Institute). As the age of cellphone use gets younger and younger and more people starting to rely on texting instead of speaking on the phone, distracted-driving due to cellphone use has become a major problem. However, with the growth of technology and Artificial Intelligence (AI) things such as autonomous driving has come to light.

Although the goal of autonomous driving is to have cars drive themselves completely without any human engagement, there are safety factors that are out of the control of the human driving and the car itself. These safety factors include road conditions, weather, location of accidents and human error, such as driver intoxication. In order to get a comprehensive idea, different combinations of safety factors need to be looked at not only separately, but together as well.

The target of this observation are municipal emergency services, such as the police department, fire departments, paramedics. We will also target car insurance companies and car companies that work on autonomous vehicles.

Data

The data that will help predict the severity of an accident were obtained from https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv. This data set gives us the large data that we need to create a model and predict the severity of the car crashes. The data consists of 38 independent variables which can be used as potential features and a collection of 194,673 recorded accidents.

Cleaning Data

With the large data set, the data needs to first be preprocessed. Objects need to be converted into values and need to remove NaN and empty values. Since NaN does not help with classification, it is removed from the target and features. NaN values were replaced with 0 and object values such as ‘Y’ and ‘N’ were converted to 0 and 1, respectively.

Building Models

For this project, we decided to use Decision Tree. Originally, the plan was to use Decision Tree, KNN, and Logistic Regression, but when trying to install imblearn so we can use SMOTE to unsample training, an error kept showing up saying the version is not compatible.

The data was split into a train set and a test set. We trained our model with the train set and the evaluation was done with the test set to keep from having a biased prediction.

```
Train set: (155738, 9) (155738,)
Test set: (38935, 9) (38935,)
```

Results

After running decision tree the F1-score for the test dataset came out to **0.8349**, while th e F1-score for the training set turned out to be **0.8516**. Although we were not able to run kNN and Logistic Regression to compare the tree models to see which one would be th e best model, our results show that decision tree is a good model to use to predict accid ent severity.

Conclusion

Among 40 features that can potentially have a value in predicting the severity of a car cr ash, human intoxication is one of the most important factors. It is one of the factors that the humans/the driver can control. Things such a weather and road conditions cannot be controlled and just have to be taken into consideration when driving. The f1 score obtained by this model can be used to predict accident severity in certain areas when on their trip. The target audience of this observation, police department, fire department,

as well as paramedics and car insurance companies can also use this information to help with their calls relating to accidents and their policies for their companies.