U.S. Car Accidents (2016-2023) - Group 9 - Milestone 2

By: Clay Shubert, Nagendra Upadhyay, Robert Grady Williams, Zach Williams

Background

Our team plans to use machine learning techniques to determine the features most important in predicting the severity of a car accident as well as create a model that will accurately predict the severity of the accident based on those features. We also are wanting to better understand contributing environmental factors and/or road factors that can have an impact on the traffic caused by the accident also reflected in severity. We are also looking to find some correlations between certain features within the car accident dataset and determine any possible reasons for these correlations. The analysis of car accidents is a critical area and it is important to learn the reasons that cause them in order to mitigate driving risks.

Motivation

The motivation behind investigating this topic and specifically this dataset stems from the huge impact that car accidents have on public safety and the economy. Car accidents are the leading cause of death in most countries around the world. This is challenging to analyze because there are a wide range of factors that could influence the beginning or likelihood of an accident making it hard to determine the root cause or most significant factors involved. The impact of this project is to attempt to predict the severity of a car accident using several machine learning techniques. This will give insight to community leaders on ways to reduce the likelihood of a car accident and if there are any factors they can influence to prevent one.

Dataset

We decided to use a US accidents dataset from the year 2016-2023. The dataset is quite huge with the size being 3.06 GB. The dataset covers 49 states and contains 7.7 million accident records. The dataset has 46 columns and over 7.7 million rows of data points. Some of the potential challenges that we might face could be with handling some data regarding time, negative time durations as we might have to drop rows with negative values. Basically the most

challenges we think are going to be with preprocessing the dataset as there is so much data and cleaning it up to be able to study it efficiently is going to be very important. One other thing which is going to be essential is to decide what features we would like to use and which we would like to drop as there is so much data that boiling it down to things that will help us come up with proper results will be very important.

State-of-the-Art

The state-of-the-art performance that we discovered for this dataset was by Ronghui Zhou, PH.D. on Kaggle here. Their notebook had the highest number of upvotes and test set prediction accuracy for a prediction notebook on Kaggle among the top notebooks. They used a subset of the dataset pertaining to their personal location in order to predict accident severity based on the various features of the dataset. They first removed any erroneous data, then they hand picked features of importance to use in the severity prediction. This is something that we plan to expand upon in the feature extraction step, and then even further by using feature reduction. Once they had preprocessed the data, they then used a variety of prediction models including Random Forest, Decision Trees, Logistic Regression, and kNN with various k values. From running these prediction models, they achieved a maximum accuracy of 97.4% with Random Forest which is a very high accuracy statistic. We hope to achieve a similar accuracy, but on a larger sample of the dataset. We also plan to use multiple prediction models, similar to their notebook, but will also add classifier fusion and more performance evaluation steps to add even further. Overall, this source is a high target for us to try and achieve similar results while also adding steps to improve the process as a whole.

References

Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, Radu Teodorescu, and Rajiv Ramnath. “Accident Risk Prediction based on Heterogeneous Sparse Data: New Dataset and Insights.” In proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM, 2019.

Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, and Rajiv Ramnath. “A Countrywide Traffic Accident Dataset.”, arXiv preprint arXiv:1906.05409 (2019).