Traffic accident prediction with Machine Learning

This is the first part of our three-part final project for DSE 511

Russ Limber

The University of Tennessee, rlimber@vols.utk.edu

Sanjeev Singh

The University of Tennessee, ssingh42@vols.utk.edu

EonYeon Jo

The University of Tennessee, ejo1@vols.utk.edu

Additional Keywords and Phrases: Logistic Regression, Naïve Bayes, Machine Learning, Dept. of Transportation, Neural Network, Random Forest, Vehicle Accident Data, Road Conditions, XGBoost, Ensemble

1 INTRODUCTION

The aim of this project is to model vehicle accident severity based on weather and road conditions. Once a final model is selected, we plan on performing exploratory factor analysis (or similar methodology) to identify which variables contribute the most to the severity of accidents. Additionally, we are exploring whether or not accident severity varies significantly by major city within the United States. This is a topic of great significance as vehicular accidents make up approximately 38,000 deaths in the United States each year and cause about 4.4 million hospitalizations. The specific hypotheses that we will be exploring are:

1. Can we produce a classification model with a strong F1 Score, as well as high recall and precision values, that determines the severity of a road accident based on a 1-4 scale?
2. Do certain factors influence the chance of having a more severe accident more than others, and if so, which factors?
3. Is there a statistically significant difference between accident severity by city based on ANOVA?

The dataset for this project was collected by researchers at The Ohio State University. The data collection process was a collaboration between the research team as well as the US and state departments of transportation, law enforcement agencies, traffic cameras, and traffic sensors within the road networks. The research team went on to publish: "Accident Risk Prediction based on Heterogeneous Sparse Data: New Dataset and Insights." The main result of this work was the creation of DAP: A Deep neural-network-based Accident Prediction model. The authors went on to explain how the F1 Score for accidents was the most significant criteria for testing their neural network against their baseline methods (logistic regression, deep neural network, and gradient boosting classifiers) and that DAP outperformed all of the baseline methods.

2 DATA

We have selected the dataset titled: “US Accidents (updated) A Countrywide Traffic Accident Dataset (2016 - 2020).” The admin for this dataset is Sobhan Moosavi, and it was posted to Kaggle.com two months ago. The dataset can be found here:
https://www.kaggle.com/sobhanmoosavi/us-accidents

The dataset consists of 1.5 million observations where each sample represents an accident that occurred in the United States between 2016 and 2020. To make the data size more manageable, we plan to down sample the data by restricting our analysis to major US cities w.r.t. the population size like New York City, Los Angeles, Chicago, Houston, Phoenix, and Philadelphia.
Regarding the features, there are forty-seven variables. The target variable (as of right now) is accident severity which is a ranking from 1 - 4 where 1 is not very severe, and 4 is most severe. The features can be broadly categorized into four parts:

1. **Event Log:** It carries the information related to the time and duration of the accident, how many miles of traffic was impacted by accident, along with a brief description of the event.
2. **Location:** It has the geographic coordinates and other information like Street, State, City, County, Zip Code, and additional info that locates the event.
3. **Weather Conditions:** Temperature, Humidity, Wind Chill, Pressure, Visibility, Day, Night, etcetera.
4. **Road Infrastructure:** These features specify what road infrastructure was present at the accident, like Bump, Crossing, Roundabout, Traffic Signal, and others.

This dataset can be utilized as a supervised learning problem where the target variable is severity. The dataset is fully inclusive of all of the information we need to answer our questions. We don’t anticipate needing any other dataset.

3 METHODS

This project will include both in-depth data analysis and machine learning algorithms to solve a classification problem. We will start with exploratory data analysis to develop an understanding of the underlying correlations within the dataset. Also, our factor analysis as well as ANOVA fall under the category of data analysis since they only draw a conclusion relating to the structure of the dataset. The classification models we look to produce will be examples of machine learning algorithms since they will be capable of outputting a prediction based on the data.

Our baseline model of choice is going to be a logistic regression model. This is a standard classification method that can use both continuous as well as categorical data types and is therefore a reasonable selection for our baseline. The machine learning methods that we will be implementing for classification will be: logistic regression, a multinomial naïve Bayes classifier, support vector machine and we will apply the ensemble module using random forest, XGBoost and adaboost. For evaluation our main scoring criteria for classification will be the F1 score but we will also report and consider recall and precision values. For the ANOVA our level will be equal to 0.05. We are going to measure success by the F1-Score; the higher, the better. However, along with the F1-Score, we'll also see how well our methods help us decide the critical factors that determine the accident's severity.

4 CONCLUSION

Ideally, we hope to produce a classification model that can predict the severity of a vehicle accident based on the dataset. Even if that model should underperform our expectations, we would still expect to have identified which factors contribute the most to accident severity. Under the worst-case scenario, where neither our model nor our factor analysis seems to draw any substantial conclusions, we will still be able to answer the question: "Does accident severity seem to vary by major US city?"

If this study were to be successful, it would help save lives by implementing predictive systems that'll provide real-time warnings for potential road accidents based on changing road conditions. For example, if certain weather conditions combined with specific road features lead to severe accidents, it would help authorities send alerts to on-road drivers informing them about safe driving practices, which would help to avoid severe road accidents.

REFERENCES

[1] Association for Safe International Road Travel. 2021. Road Safety Facts: Association for Safe International Road Travel. [online] Available at: https://www.asirt.org/safe-travel/road-safety-facts.

[2] Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, and Rajiv Ramnath. "A Countrywide Traffic Accident Dataset.", 2019.https://arxiv.org/pdf/1906.05409.pdf

[3] Moosavi et al. "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. https://arxiv.org/pdf/1909.09638.pdf