Predicting Severity of Traffic Accidents in New York City

1st Aaron Valdes
Department of Electrical and Computer Engineering
University of Miami
Coral Gables, Florida
axv725@miami.edu

2nd Kyle Riley
Department of Computer Science
University of Miami
Coral Gables, Florida
kmr145@miami.edu

***Abstract*—**
***Index Terms*—Collision, Statistical Learning, Random Forest**

I. Introduction

Cars have been one of the biggest accomplishments in technology innovation since the late 1800s. Today, cars have evolved so much from just a simple motorized vehicle that couldn't travel faster than 10 miles per hour or carry more than 2 people at a time to an electric vehicle that can drive itself. Despite all of the great efforts made to improve vehicles over time, it has come at a great price. Car accidents have become a big problem in the United States and the world as a whole. According to the Association for Safe and International Road Travel, more than 38,000 people in the U.S. and 1.35 million people in the world die every year due to car accidents [1]. In addition to this, another 20-50 million people worldwide suffer from non-fatal injuries that leave them with long-term disabilities. Even though many car manufacturers go to extremes in an effort to prevent this, the numbers are still high. Part of this is due to factors that the manufacturers themselves can't control. These uncontrollable factors are the reason why it is relevant for us to take a deeper look into the accidents themselves and the manner in which they occur in an effort to try and predict the severity of car accidents. One main reason to research this issue is because it can be a major step in preventative safety. Car manufacturers can use this research to make better safety systems for the vehicles they produce in the future. Another aspect of safety that this research can contribute to is management safety. If the prediction of car crash severity proves to be successful and accurate, it can provide vital information to medical professionals that would allow efficient and prompt medical treatment in the event of an accident.

According to recent studies, there have been a few successful attempts in predicting traffic accident severity. In 2021,a similar study by using machine learning to analyze traffic accidents in the district of Setúbal, Portugal from 2016-2019 [2]. In their research, they used methods such as decision trees, random forests, logistic regression, and naive Bayes to develop models to classify the severity of the accidents that occured. With their findings, they were also able to identify hotspots of where accidents are more likely to happen. Another study predicted multiple types of accident severity across China [3]. In their research, they were about to create a multi-task learning model and the first known deep neural network based model for predicting traffic accidents. The model is able to predict different levels of injury, death, and property loss. Moreover, the model can also identify key factors that cause the three types of severity through layer-wise relevance propagation. This generates explanations based on the structure and weights of deep neural networks. Lastly, another research team performed a study to accurately predict the severity of traffic accidents that occured in Victoria, Australia from 2014-2019 by creating a hybrid model [4]. The model consists of using principal component analysis with multiplayer perception neural networks and support vector machines. With this model, accurate severity predictions were

obtained once the principal components were applied to the models.

In order to predict the severity of accidents, this project's goals will consist of predicting the amount of individuals killed or injured that were involved. These individuals include drivers, passengers, and any other person that may have been harmed as a result of the accident. The focal area of the project will include data points from New York City. As a very popular and modern area, New York City provides a lot of useful information to work with.

II. Methods/Approach

A. The Dataset

For this project we used two major dataset which constitute the weather and crashes for the city of New York from 2012 to 2022. For the collision dataset, we used two dataset,the collision information and the person-involved in the accident information provided by the New York City Open Data [5]. Both person and crash information have on average 2.5 million observations. For the weather data we used OpenWeatherMap weather history API to find the weather per day and hour from January 1, 2012 to March 20, 2022 [6].

After gathering the dataset, we followed the simple procedure of removing all the missing values for each of the columns. This was followed by the removal of columns that either didn't provide a necessary value to our research, were redundant, or varied too much. Moreover, columns such as vehicle type which could provide us with value information about a collision had data in dozens of formats from year to year which made it difficult to add to our dataset. Furthermore, the number one contributing factor column could have been really useful however given the fact that there were more than 32 factors in this column which impeded us to run the decision tree model made us discard this column. Next, we added some columns such as the day of the week and if the date was a holiday. As for the people-involved dataset, we extracted only two necessary columns which is the age and sex of the individuals involved.

Column Name	Description
Crash Date	Date of the traffic Accident
Crash Time	Time of the traffic Accident
Person Age	Age of the person involved in the collision
Person Sex	Sex of one of the person involved in the collision
Killed/Injured	Number of killed and injured combined
Weekday	Day of the week
Holiday	Is the date a holiday

TABLE I
FINAL COLUMNS FROM THE COLLISION DATASET

Most of the work, that it was done in the weather dataset was the removal of unnecessary columns that will have provided us with irrelevant information such as the dew point or the sea level.

Column Name	Description
Temperature	Temperature in celsius
Visibility	Visibility
Pressure	Air pressure
Humidity	Humidity
Wind speed	Speed in m/s
Weather Main	One word description of the weather
Weather Description	Brief Description of the weather conditions

TABLE II
FINAL COLUMNS FROM THE WEATHER DATASET

The first step in the merging of the dataset, was to first merge the collision information from the person's dataset and the information out of the collision by using the vehicle id as the key. Next, we merged the weather and the collision information by using the date and time of the collision.

1) Downsampling: After thorough investigation of our dataset, it was discovered that our dataset was highly imbalance which meant that it needed to be fixed. Our biggest issue, was class 0 which occurs when no one either dies or becomes injured during a traffic accident which occurs the majority of the time. This class constituted to at least 74 percent of the entire dataset which meant that we needed to reduced it by at least half. By making class 0 be twice the size of the second biggest class, we reduced class 0 to be 50 percent of the entire

dataset. We performed test on both the regular dataset and the downsampled dataset to see the effect of it.

B. Techniques

This problem was approached initially as a regression problem due to the fact that making killed/injured a set of classes would have let to many classes, however as we later learned is that the best approach to solve this problem is by using classification strategies. For our regression approach, we used two model for the testing which was linear regression and boosting. As for our classification approach, we used logistic regression, decision trees, KNN, LDA, Random Forest,and Support Vector Machine.

C. Evaluation

Due to the discovery of class imbalance in the dataset, we decided that we needed to see how it affected our dataset and how it was affected by changes in the size of our dataset so we performed multiple test. For all of our models, we ran them with varied size for two reasons,1)we wanted to see how our accuracy changes with increasing the size and 2) some of our models would have taken us weeks to run with our entire dataset. To evaluate the our results we follow two simple metrics which is accuracy and RMSE. Cross-validation was performed only for linear and logistic regression, furthermore the reason it was not used on the other models was due to the small difference between the results of using cross validation and not using it. This small difference in results was probably due to the fact that our dataset was extremely large and had low levels of variance. For models such as randomforest, logistic regression, and KNN we optimize the results by performing grid optimization.For random forest we optimized the number of trees, the number of variables randomly sampled as candidates at each split, and the node size. For KNN, we tested different values of k from 5 to 13 where it was seen that the higher the k value the better the accuracy.

III. Results

	Simple		**Downsampling**	
Type	**Size**	**Accuracy**	**Size**	**Accuracy**
Linear Regression	2500000	17	2500000	30.41
Logistic Regression	2500000	17	2500000	29.75
Tree	2500000	74.45	2500000	39.98
LDA	2500000	74.3	2500000	52.55
KNN	10000	73.96	50000	48.34
Random Forest	**1000000**	**87**	**1000000**	**57.63**
Support Vector Machine	50000	76.16	2500000	48.36

TABLE III
SUMMARY OF OUR RESULTS WITH THE STANDARD DATASET AND THE DOWNSAMPLED DATASET

One of the most clear observations from Table 3 is that there is tremendous difference between the standard dataset and the downsample dataset. Due to this difference in the results we can say that the class imbalance has a clear importance on the accuracy. Furthermore, there seems to be a correlation between the size of class 0 and the accuracy of our module since when we have 74 percent of our values in class 0 we obtained an accuracy of close to 74 percent, while when we have 50 percent we obtained an accuracy of around 50 percent. However, we have to point out the fact that accidents where people are injured or killed are always less likely to occur.

Model	RMSE	
	Simple	Downsampled
Linear Regression	1.08746498927265	1.19693080784508
Boosting	0.8949589	1.188997

TABLE IV
RESULTS FOR THE REGRESSION MODELS

It can be observed that regression methods are not the best for predicting the severity as seen from the low values for the accuracy and RMSE. However, boosting provided us with better results than linear regression when it came to regression. Among the classification methods we used, the most powerful was random forest after we optimized it for both dataset. Overall, our results indicate one clear issue which is that we need to perform a better job at pre-processing since information such as contributing factor, zipcode, or vehicle type can provided us with crucial information.

IV. Conclusion

Traffic accident severity is extremely important and essential information that can be used to help contain and prevent traffic

accidents. In this study, we attempted to predict traffic accident severity by trying to predict the amount of individuals that were killed or injured in connection to a traffic accident. The dataset used consisted of traffic accident data from New York City from 2012 to 2022. This study analyzed the performance of linear regression, logistic regression, trees, LDA, KNN, random forest, and SVM models on the dataset. When analyzing the results, it was discovered that random forest served as the best model for predicting traffic accident severity. Even though it was the most accurate model, it didn't prove to be very useful for the project as a whole due to the class imbalance. For next steps,we need to improve the class imbalance of our dataset, perform 10-fold cross validation an all models, perform a more detail job during pre-processing, optimize all the models, and use more models to see how well they perform. In the future, we would like to explore the possibility of trying to predict the hotspots of where the most severe traffic accidents occur or predict the contributing factor for each of the accident.

ACKNOWLEDGMENT

Thank you Dr. Aguiar for all the advise you gave us throughout this project and for the information you provided us during class.

REFERENCES

[1] "Road Safety Facts." [Online]. Available: https://www.asirt.org/safe-travel/road-safety-facts/

[2] D. Santos, J. Saias, P. Quaresma, and V. B. Nogueira, "Machine Learning Approaches to Traffic Accident Analysis and Hotspot Prediction," *Computers*, vol. 10, no. 12, p. 157, Nov. 2021. [Online]. Available: https://www.mdpi.com/2073-431X/10/12/157

[3] Z. Yang, W. Zhang, and J. Feng, "Predicting multiple types of traffic accident severity with explanations: A multi-task deep learning framework," *Safety Science*, vol. 146, p. 105522, Feb. 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925753521003659

[4] K. Assi, "Traffic Crash Severity Prediction—A Synergy by Hybrid Principal Component Analysis and Machine Learning Models," *International Journal of Environmental Research and Public Health*, vol. 17, no. 20, p. 7598, Oct. 2020. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7589286/

[5] C. of New York, "Nyc open data." [Online]. Available: http://nycod-wpengine.com/

[6] "Openweathermap." [Online]. Available: https://openweathermap.org/