Machine Learning Approaches to Traffic Accident Analysis

Cem Baykal, Justin Indla, Denizhan Kilic, Jiangyuan Yuan

May 2, 2024

Abstract

Traffic collisions are an immensely import area of societal concern. In 2020 and in the U.S. alone, there were 35,766 fatal car crashes and another 1,593,390 crashes resulted in injury (Bieber, 2023) [2]. Our project utilized collision data for the city of Seattle that is updated weekly by the Seattle Department of Transportation to employ machine learning methods [3]. We utilized logistic regression, support vector machine (SVM), and decision tree models that were trained using information regarding the weather condition, relation to intersections, road conditions, light conditions, and if the driver was under the influence variables to predict whether or not the collision resulted in an injury. Our findings found that there was a strong correlation between these variables and whether or not an injury was sustained. Consequently, our work has significant implications for traffic and road planning as well as policies regarding alcohol use and driving.

1 Introduction

Traffic collisions are an ongoing and critical challenge for societies around the world. It has a high cost in terms of human lives lost, economic impact and injuries sustained. Understanding the factors that contribute to these collisions and developing models to forecast their occurrence are important jobs for policymakers, public safety officials, and transportation planners.

The primary objective of the paper is to conduct a comparative analysis of various machine learning models, including decision trees, linear regression, and Support Vector Machines (SVM), with the aim of identifying the model that achieves the highest accuracy for predicting whether or not there is an injury by utilizing the conditions present at the time of the crash. Our models seek to identify variables such as weather conditions, road conditions, proximity to intersections, driver impairment, and light conditions as discussed in the abstract. Our paper utilizes data from the city of Seattle provided by the Seattle Department of Transportation [3].

Our findings are significant beyond academic research. They have far-reaching implications for practical interventions in traffic and road planning, alongside the enforcement of alcohol-related policies. As we identify the most influential variables for whether or not an injury occurs, the data from our findings can help to inform targeted interventions aimed at reducing the number of fatalities and injuries on our roads.

In this paper, we describe the methodology that has been used, discuss the results that were obtained, and mention the implications of our findings. Through our findings, our goal is to contribute to traffic safety and protect road users' well-being.

2 Related Works

The utilization of traffic data for predicting and gathering information about the severity of traffic accidents is an active research field. A team from the University of Évora in Portugal implemented a DBSCAN method to map and locate traffic accident clusters in Lisbon to find hotspots (Santos et al., 2021) [5]. Furthermore, work by the same team implemented logistic regression, decision tree, and random forest models to predict fatal/non-fatal traffic accidents by utilizing information about the accident and its relation to clusters (Santos et al. 2021) [5]. They found that a random forest model was the most accurate at 73% and found the accident's relation to a cluster was the most important feature (Santos et al., 2021) [5]. Another team based out of the Jordan University of Science and Technology implemented similar methods of mapping accidents in Zarqa, Jordan to find hotspots of traffic accidents and analyzed the severity of accidents (Al-Mistarehi, 2022) [1]. Their work found that the highway, vehicle, and environment variables were the most important in determining the severity of the crash and that the time of day was the single most important variable (Al-Mistarehi, 2022) [1]. A recent study published by a team based out of the European University Cyprus studied the effectiveness of machine learning techniques in the

prediction of accident severity (an injury or not) and found the highest success using logistic regression and random forest models achieving 87% accuracy while the engine capacity and age of the vehicle were the most important features (Obasi & Benson, 2023) [4]. Thus, this echoes the findings of the team at the University of Évora in regards to the most effective models.

3 Data Processing and Results

3.1 Data Processing

Due to the real-world nature of the collision data set, a series of data processing procedures needed to be deployed. Firstly, we only kept the attributes in the data set that could tell us information about whether or not there was an injury. For example, there were multiple attributes that served the process of uniquely identifying the collision and while important for data collection, these did not have any relation with the severity of the collision. Data entries with missing column values were removed from the data set as many of these fields were categorical and could not be averaged. Moreover, each string data value was mapped to an integer value that represented its perceived impact. For example, we encoded the most common 5-6 values which represented over 90% of the dataset using a range of integers that represented the scale of the value. Since a "SEVERITYCODE" of '2', 2b', and '3' represent a sort of injury while '0' or '1' represent no injury, we created a binary injury "INJURY_BINARY" target variable that mapped whether or not an injury was sustained. The data set was checked for outliers.

3.2 Logistic Regression

For predicting injury severity from car collisions, we employed logistic regression, leveraging features such as weather, road and light conditions, junction and intersections, and whether the driver was under the influence. We partitioned the data into a 60% training set, 20% validation set, and 20% testing set, a distribution that ensures a robust tuning and testing process.

The regularization parameter C, critical for controlling model complexity and ensuring generalization, was varied systematically during hyperparameter tuning. The best performance on the validation set was observed at $C = 0.01$, a value that was then applied to the model for the testing set. This configuration yielded an accuracy of 65.87% on the testing set, indicating a model that generalizes well to unseen data. The corresponding confusion matrix obtained from the testing set is presented in Figure 1.

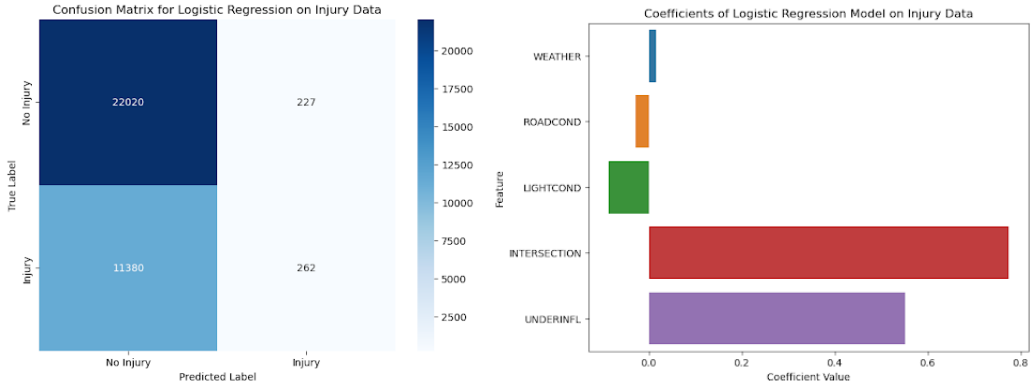

Figure 1: Confusion Matrix (left) and Corresponding Coefficients (right) of Logistic Regression Model

Further analysis of feature significance, depicted in Figure 1 on the right, indicated that "UNDERINFL", and "INTERSECTION", denoting whether the driver was under the influence and whether the incident occurred at an intersection had a pronounced positive effect on predicting injury severity. Contrary to initial assumptions, "LIGHTCOND" showed that darker conditions correlate with lower injuries, potentially due to heightened caution exercised by drivers.

3.3 Support Vector Machines

For our 2nd way to model the data, we chose to explore Support Vector Machines, or SVMs. SVMs are used especially for classification problems that have a large amount of features. SVMs accomplish this by modifying a

hyperplane to separate the data by its classes optimally. Since SVMs can establish this hyperplane, SVMs are less prone to outliers affecting their parameters, which is one drawback to logistic regression.

When we used SVMs to classify an accident as injurious or not, we used a 60-20-20 train-validate-test split for the data. Then, we trained 6 different SVM models, each with C values of 0.001, 0.01, 0.1, 1, 10, and 100. We experimented with this hyperparameter because of how it affects the balance between classifying all the points correctly and how complex the model is. Higher values of C give preference to classifying as many points as possible correctly, whereas lower values of C yield a model that is able to generalize better on new data, but may lead to underfitting.

We found that the C value that yielded the highest accuracy was $C = 1$, with an accuracy of 65.77% as found in Figure 2 on the right. We also found that the model had an f1 score of 0.79 for classifying an accident as non-injurious and an f1 score of 0.04 for classifying an accident as injurious, indicating that the model did a good job of classifying an accident as non-injurious but it did a poor job of classifying it as injurious. This is also evident in Figure 2 on the left as actual injuries were rarely classified correctly by our SVM model but actual non-injuries were classified with great accuracy by our SVM model.

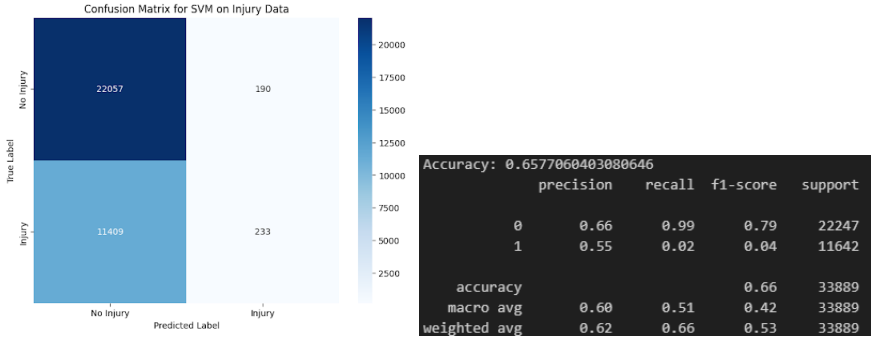

Figure 2: Confusion Matrix (left) and Accuracy Metrics (right) of SVM Model

Unlike logistic regression and decision trees, we cannot necessarily find which variables have more of an impact on the result compared to others. This is because, with this SVM, we utilize the Radial Basis Function, which translates the data provided and plots it into a higher dimensional space in order for the SVM to establish a boundary between the two sets of data. As a result, the original attributes of each data point do not necessarily correlate to one specific "axis".

3.4 Decision Trees

As part of a comprehensive approach to seeing the usefulness of machine learning techniques in predicting injuries, we also employed a decision tree that used the same attributes used in our logistic regression and SVM models. For our decision tree, we implemented 60% training set, 20% validation set, and 20% testing set split for the data. The maximum depth of the decision tree can have a profound impact on the accuracy of the model. A very large maximum depth can lead to overfitting where, due to the extremely detailed nature of the tree, the model is influenced by noise in the data and is not an accurate representation of the data as a whole. On the other hand, having a very shallow tree depth can lead to the issue of underfitting where the tree is not complex enough to precisely see the patterns within the data.

To circumvent these issues, the model was tested on maximum depths between 1 and 12. This range was determined by testing values to see at what point accuracy began to increase and decrease. Through this process, it was determined that the optimal maximum depth was 2 at which point we attained an accuracy of 65.54%. The confusion matrix attained from this model is seen in Figure 3. This high accuracy conveyed that weather, road and light conditions, intersection type, and whether the driver was under the influence had a strong connection with whether or not there were injuries during the collision. Analysis of the importance of features revealed that "INTERSECTION" followed far behind by "UNDERINFL" had the largest positive impact on predicting whether or not there was an injury as seen in Figure 3 on the right. Consequently, this supported the similar findings found by employing a logistic regression model.

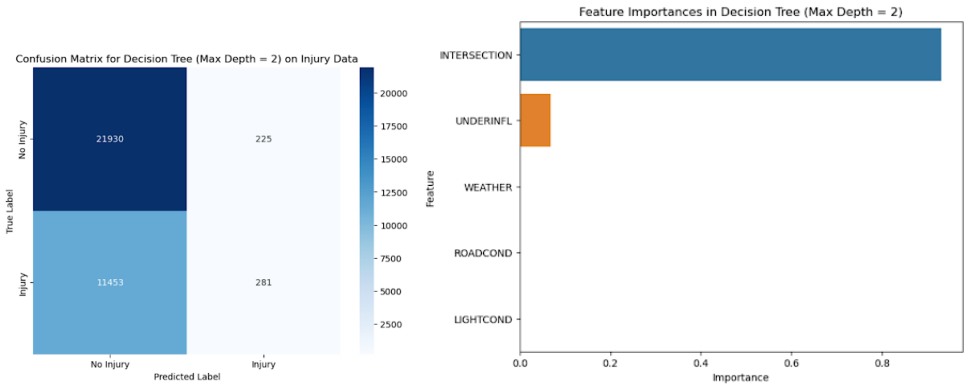

Figure 3: Confusion Matrix (left) and Feature Importances (right) of Decision Tree Model

4 Discussion

4.1 Interpretation of Results

Our analysis revealed that logistic regression and decision trees identified the variables "UNDERINFL" (whether the driver was under the influence) and "INTERSECTION" (proximity to intersections) as significant predictors of whether or not an injury occurred in a traffic collision. However, the SVM, employing the Radial Basis Function kernel, was unable to provide explicit feature importances, which complicates the interpretation and application of its outcomes in policy-making.

4.2 Comparison with Prior Work

The effectiveness of logistic regression in our study is consistent with findings from Obasi & Benson [4], who also reported high accuracy with logistic models in traffic accident prediction. Additionally, the success of other teams' findings with random forests and cluster models such as those of Santos et al. suggests that this is an avenue for further research [5]. However, unlike Al-Mistarehi et al. [1], who found environmental factors to be the most influential in the outcome of a collision, our study highlights variables such as whether a driver was under the influence or where the accident occurred to be more crucial. This disparity might be due to differences in traffic patterns and enforcement policies between Seattle and Zarqa, Jordan, suggesting that this model may only apply to the city of Seattle or similar cities.

4.3 Limitations

One limitation that we found in training our models is the imbalance in the dataset, where instances of non-injuries significantly outnumbered the number of instances of injury cases, which may bias the models towards predicting "No Injury". This issue was visible in the performance disparities across all three models, where they performed well in predicting non-injuries but performed poorly when classifying injury cases. Another limitation we encountered during data processing was our procedure of assigning integers during encoding of the most common text values. As no prior literature sets a precedence on how to encode these values, we manually assigned values depending on the expected impact of the value.

5 Conclusion

Our study demonstrates how machine learning models can be utilized to predict traffic collision injuries with a reasonable degree of accuracy. The findings advocate for a data-driven approach to traffic safety analysis and they emphasize the roles that certain predictors, especially intersection relation and driver sobriety, have in shaping traffic safety outcomes.

References

[1] Bara' W. Al-Mistarehi, Ahmad H. Alomari, Rana Imam, and Mohammad Mashaqba. Using machine learning models to forecast severity level of traffic crashes by r studio and arcgis. https://www.frontiersin.org/articles/10.3389/fbuil.2022.860805/full, Mar 2022.

[2] Christy Bieber. Car accident statistics for 2023. https://www.forbes.com/advisor/legal/car-accident-statistics/, Feb 2024.

[3] Jonathan Leon. Seattle sdot collisions data. https://www.kaggle.com/datasets/jonleon/seattle-sdot-collisions-data/data.

[4] Izuchukwu Chukwuma Obasi and Chizubem Benson. Evaluating the effectiveness of machine learning techniques in forecasting the severity of traffic accidents. https://doi.org/10.1016/j.heliyon.2023.e18812, Aug 2023.

[5] Daniel Santos, José Saias, Paulo Quaresma, and Vítor Beires Nogueira. Machine learning approaches to traffic accident analysis and hotspot prediction. https://doi.org/10.3390/computers10120157, Nov 2021.