Road Accidents Analysis

Mohit Gaggar
PES University
Bangalore, India
gaggarmohit@gmail.com

Shreya Prabhu
PES University
Bangalore, India
shreya.pr611@gmail.com

Abhishek Adinarayan
PES University
Bangalore, India
abhishek.adinarayan@gmail.com

***Abstract*—Finding measures to prevent traffic accidents is a critical matter, given the significant impact they have on public health and safety, as well as concerning maintaining public infrastructure and averting property damage. They are a prevalent cause of injuries and deaths. Furthermore, vehicle crashes and accidents also are a predominant cause of traffic congestion and delay. It is thus imperative to study and analyze the different factors that may bring about a higher risk of a vehicular accident occurring. In this study, the impact of the crash on the vehicle occupants is estimated based on environmental conditions such as weather conditions, lighting conditions, road infrastructure, etc. The methods and models considered in this study are various forms of regression and classification methods such as decision trees and multi-layer perceptron classifiers.**

***Index Terms*—accidents,injuries,driver,vehicle,weather**

I. Introduction

Travelling and transportation of goods are a part of our day to day lives. Measures must be taken to ensure safety and identify the leading causes that pose a danger to commuters as they travel. Road safety and traffic management programs lay the foundation to prevent traffic accidents. However, a detailed analysis of previous cases of vehicle crashes and the results obtained from these could further reduce the rate of occurrences of such accidents.

Accurate predictions estimated from statistical models can assist traffic personnel and other authorities to identify areas with higher risk and implement better accident management procedures. This would also help improve traffic safety and increase the efficiency of our current transportation systems.

An important step to be taken before building any analytical models is identifying and collecting good data to work with. The dataset should describe enough independent variables so results can be modeled correctly. Effective analysis can hence improve worldwide traffic accident statistics.

A. Problem Statement

Here we seek to solve this problem by predicting the severity of a crash from the perspective of environmental data. This models factors like weather conditions, speed limits, road infrastructure etc. which play a considerable role in causing road accidents. The approach that we take should consider factors like vehicle interactions, road users, road infrastructure, etc so as to arrive at an appropriate solution. The dataset used here is "Traffic Crashes-City of Chicago" from the Chicago data portal.

II. Previous Work

Techniques used to model traffic accidents data include different kinds of regression, time series analysis.

Logistic Regression is one type of regression that models a binary dependent variable using the logistic function. [1] presents the application of multiple logistic regression to investigate various factors that play a role in vehicle collisions and predict the odds of having a fatality in a vehicle. The author uses a stepwise attribute selection procedure, a combination of forwarding selection and backward elimination procedures, aimed at reducing AIC and BIC values. After selecting the attributes, the goodness of fit of the model was found by using the Pearson Chi-Square, Deviance, and Hosmer-Lemeshow. The ROC curve was also plotted to find the goodness of fit. The results found for the selected models were as follows - Chi-square is 5.295, Area under curve is 0.905, p values were 0.726 and the distance for the optimal cutoff point was 0.3413. Lastly, Cook's distance and the leverage of points were found along with VIF was also found to check for multicollinearity.

We can use an ordinal logit if the dependant variable has more than two categories, and the values of each category have a meaningful sequential order. In [3] the authors worked on a crash severity model using an ordered logistic regression, thus allowing a transformation of total crash counts into counts by severity. It was found that the safety effects of speed limits exhibit a convex relationship in terms of crash severity. They used FE and RE (Fixed effect and random effect) model for total crash per million. The RE estimated were expected to be biased as indicated by Hausman test.The R square statistic suggested that 90% of the variation in crash rates was explained by the model's control variables.The p values for most of the variables used were found to be 0.

Sajjakaj Jomnonkwao et al. worked on data compiled from different departments such as population data, economic data, transportation data. They explored four different methods to analyse traffic crashes [2]. Namely time series analysis models like exponential smoothing and Holt's linear trend technique were used to ascertain linear trend data without seasonal influences which resulted in a MAPE value of 8.1Curve estimation was be done with the independent variable as time (years), with road death rates per 100,000 population. The cubic model, quadratic model and the linear model with adjusted R2 values of 0.813,0.794, and 0.724, and MAPE values 11.2%, 10.2% and 12.6% were got respectively. Multiple regression

analysis was done with the death rate from road accidents as the dependent variable. Variable Selection based on forward selection, backward elimination and stepwise regression was used to build three models. The had an accuracy of 88.8%, F-test value of 51.144 and MAPE of 6.4%. They considered the number of registered cars over the number of registered vehicles, number of registered trucks over the number of registered vehicles, and the energy consumption of the transport sector over the number of registered vehicles and found how they affected the death rate from road accidents.

Path Analysis was also done to determine the number and type of variables and the relative paths.Path Analysis was used as it identifies the details of bivaria te relationships between two variables and the weighting of values connected to these points. The path analysis results showed the following goodness-of-fit values: Error(MAPE) was 8.4%,Chi-square statistic was 17.706, RMSEA = 0.495,p value was 0.0005 and standardized root mean square residual (SRMR) = 0.078.

There are certain drawbacks of using these models, it is found that most of the techniques involved tried to predict whether accidents can be fatal under a set of conditions, it was found that data was not available at a single source and had to be collected from multiple sources making it non uniform, the variables being predicted were mostly dichotomous in nature and thus common practice was to use a probabilistic approach to regression and predicting whether the accident will be fatal. We also found that previous work done could not include all factors due to the dataset not being complete, like Sajjakaj et.al [2] tried to model the data using only vehicle type and GDP thus not including important driver information and weather. Similarly Kara et.al [3] do not use weather information in modeling accident data.

III. Proposed solution

Traffic Crashes dataset taken from the Department of Transportation, Chicago.

A. Preprocessing

This dataset describes in detail the environmental factors, time of occurrence of the crash, degree of severity of the crash, and injury. In line with the problem statement, environmental variables are focussed on. Consequently, the chosen attributes are Weather Conditions, Lighting conditions, Road Surface conditions, Road Defect, Posted Speed Limit, and Traffic control devices.

1) Missing Data: The rows containing the missing values are dropped. This is suitable given the large size of the dataset.

2) Attribute Selection: Several values within these attributes have been grouped together logically to be able to work more easily with fewer variables. For example, in Weather Conditions, values such as 'Freezing Rain' and 'Sleet/Hail' are combined. In Traffic Control Device, 'traffic signal' and 'Flashing Control Signal' are combined.

3) Encoding Categorical Variables: The majority of these attributes are qualitative in nature and hence are required to be converted into some numerical form so that they can be used while building models for the dataset. The general approach to this involves label encoding, i.e. assigning a numerical value for each unique value of the column. However, this could create problems wherein model results would be affected as some values are considered 'greater' than the others. Hence in such cases, the attributes are converted to dummy variables. Dummy variables allow us to take each feature, and incorporate them into one regression equation, by indicating their presence or absence using the values 1 and 0 respectively.

The target variable 'Most Severe Injury' has 5 categories (No sign of injury,Non Incapacitating injury, Incapacitating injury,Reported not evident, Fatal) and is of an ordinal nature. These have been label encoded according to the order of severity of the injury.

Once all attributes have been chosen the following approaches may be pursued to build a prediction model- As a prelim checking if treating the column data as nominal and performing classification will act as a good model. Using ordinal logistic regression, other multiple regression techniques to give a continuous variable as output which can be converted to the response variable.

B. Model Building

Our solution to the problem was to first perform Classification . The polychotomous response variable describes the type of injury, rather than classifying a binary value of injury vs no-injury.Thus we use multiclass classification techniques as given below-

1) Linear Regression: Linear Regression is performed as a baseline to our classification as it uses a simple function to fit the independent variables and return a continuous variable as the output, this continuous variable is rounded up and the integer value got is mapped to the dependent variable categories to make the prediction.

2) Decision Tree: It is a supervised learning algorithm, that use entropy, Gini index to calculate the information gain provided by splitting each feature at different levels and picks the attribute with highest information gain. This method can be used for multilable classification and was thus explored by us for this task.

3) MLP Classifiers: Multilayer Perceptron, a neural network based classifier that is used to perform classifications into polychotomous classes. We used this to try to fit a non-linear classification line to our data to see how it performs on our data.This type of classifier requires standardization to

4) Ordinal Logistic Regression: Our final model was Ordinal regression which is an extension of the simple

logistic regression. While logistic regression can only be used with binary classification, this type of logistic regression has a few assumptions which were fulfilled by our dataset, like presence of little or no correlation between the independent variables, no outliers. This can also have a polychotomous dependent variable, and the dependent variable are of a ordinal nature.

IV. Experimental Results

We used an 80:20 split to split our data into test and train data. The train data had 318044 and the test data had 79511 rows. The train and test data was then partitioned into X train, Y train and X test, Y test with X train and test consisting of all independent columns taken into account to train our model ie.,Weather condition, Lighting conditions, Posted speed limit, Road surface condition, Traffic control device, Road defect. We trained separate models namely linear regression, Decision Trees, Multi layer Perceptron and Ordinal Regression using the X test and the Y test and made predictions on X test. The performance metrics we used for model evaluation were MAE, this kind of error measures errors between paired observations. MSE measures the mean of the squares of the errors. Accuracy is calculated by summing number of True positives and True negatives and dividing by Total number of observations.

The following table shows the error and accuracy values obtained for each model:

Model	MAE	MSE	Accuracy
Linear Regression	0.21	0.187	86.62
Decision Trees	0.152	0.19	86.63
Multilayer Perceptron	0.15	0.189	86.72
Ordinal Regression	0.15	0.189	86.72

By looking at the data we expected ordinal regression to perform better than classifying algorithms, but this wasn't seen in our observations. We have devised a few reasons for this- The independent variables we have taken are all of categorical type and model is generated by using dummies for them.The dataset is skewed with most of the data belonging to the dependent variable label of NONINCAPACITATING INJURY this produces a bias and causes the models to overfit to produce this label as output. The dependent variable looks like it is of the ordinal distribution but our models suggest that treating it as nominal can be done without affecting the result, this can be rooted back to our original data distribution.

Many other factors that could have played a role in modeling accidents were not present in the dataset and thus could not be mentioned in our model.These include driver related - driver age, physical and mental state etc, vehicle related - which vehicle was the driver driving (some vehicles have more safety measures built into them causing more severe accidents also to have lesser impact), other factors that could indirectly cause accidents (an accident in one place could cause another accident in some other place close by due to traffic build up). The time of the day, day of the week and week of the year can also be included to get a better fit as these factors will be able to explain the spike and drop in accident rates during festivals, peak hours and seasons.

V. Conclusions

From the above analysis, we have observed that Traffic Accidents can be modelled by considering numerous factors of which we have included environmental conditions such as Road Surface Conditions, Road Defects, Weather conditions, Posted Speed limits, Lighting conditions and Traffic Control Devices. Our models focused on classifying the severity of the injury in a crash, rather than counting the number of injuries or fatalities that occur over a time period. Future work as part of this project would involve including non-environmental factors such as driver condition, vehicle condition, time and day of occurrence of the crash.

References

[1] Mathis, Annabelle Marie, "Statistical Analysis of Fatalities Due to Vehicle Accidents in Las Vegas, NV" (2011). UNLV Theses, Dissertations, Professional Papers, and Capstones.

[2] Sajjakaj Jomnonkwao, Savalee Uttra & Vatanavongs Ratanavaraha, 2020. "Forecasting Road Traffic Deaths in Thailand: Applications of Time-Series, Curve Estimation, Multiple Linear Regression, and Path Analysis Models," Sustainability, MDPI, Open Access Journal, vol. 12(1), pages 1-17, January.

[3] Kara M. Kockelman, Clare Boothe, Luce Assistant and Professor Civil Engineering, "Crash modeling using clustered data from Washington State: Prediction of optimal speed limits"

Contributions :
Dataset selection - Abhishek,Shreya
Literature survey - Shreya, Mohit, Abhishek
EDA - Mohit,Shreya
Model Building - Shreya, Mohit
Report - Shreya, Mohit, Abhishek