Traffic Accident Severity Prediction

Santiago Andres Granda Bravo
sgranda996@gmail.com

October 2020

1 Introduction

1.1 Background

Traffic accidents cause not only a huge amount of deaths around the world but produces economic costs to governments. According to the authors in [1], "Traffic accidents are the 8th cause of mortality in different countries and are expected to rise to the 3rd rank by 2020." Analyzing the document and the data presented by the authors mentioned previously; car accidents are a concerning problem that keeps growing and taking several lives through its constant increment. The affected parts within traffic incidents are not only the actors involved directly in the accidents but the entire society, because as it was mentioned before accidents generate economic loses to governments, society, and at last the insurance companies. So, the general purpose of this capstone project is to create a machine learning model able to identify and classify when is secure to travel on a road trip and avoid traffic accidents.

1.2 Problem

After observing the problems caused by traffic accidents it is necessary to develop a solution that would help saving millions of lives that are involved in car accidents, by creating a model that allow people who is travelling from one point to another to be aware if it is secure to drive through a determine highway, road, etc., due to many factors such as, road condition, weather, vehicles affluence, etc.

1.3 Interest

Developing a system capable of predicting fatality of traffic accidents due to several conditions will become useful for anyone interested in saving lives. Governments will become benefited of implementing a system like this, because it will diminish the amounts of accidents on the road and economic loses caused by property damage. Private investors will benefit of the system because developing a recommendation system over this machine learning algorithm will help saving lives of all those who are on a car traveling to anyplace.

2 Data

2.1 Data Source

The dataset that will be analyzed for this project is the collisions dataset from Seattle from 2004 to present, shared by the instructors.

2.2 Dataset Description

The dataset contains information of the type of collision, location, weather, the severity of the collision, and many other attributes that occurred, which are reported and described on the dataset. The dataset contains approximately 200 thousand of samples with 37 features or columns. From the dataset, the “SEVERITYCODE” feature is the dependent variable. The values that can take the target feature are 1 or 2, where 1 indicates low severity or 2 high severity after a crash. All the other features must be analyzed to observe the correlation between the independent variables with the target variable. Examples of the independent variables are: “LOCATION” is a string type which indicates the general location of the collision such as: ‘5TH AVE NE AND NE 103RD ST’, “VEHCOUNT” is a numeric double type variable that indicates the number of vehicles involved in the collision such as: ‘2’, “WEATHER” is a string type variable and indicates the weather conditions during the time of collision such as: ‘Raining’, etc.

2.3 Feature Selection and Data Wrangling

To develop a model with the dataset, first, it must be cleansed and preprocessed to perform the classification algorithm. Applying this method, the dataset is going to have the necessary attributes and features to obtain a good an acceptable solution. Applying this feature selection method the features from the original dataset are reduced from thirty seven to fourteen features, where the severity of the accidents is the target variable, while the other thirteen features are the dependent variables used to find the values of the target feature. Due to the configuration and the types of variables of the dataset, some features must be transformed from an object variable into integers or floating point variables. Some other variables are labeled so they must be encoded as numbers.

3 Methodology

3.1 Data Analysis

In this section some features are visualized for further analysis. The following figure shows the total number of accidents that occurred each year since 2004 to the actual year.

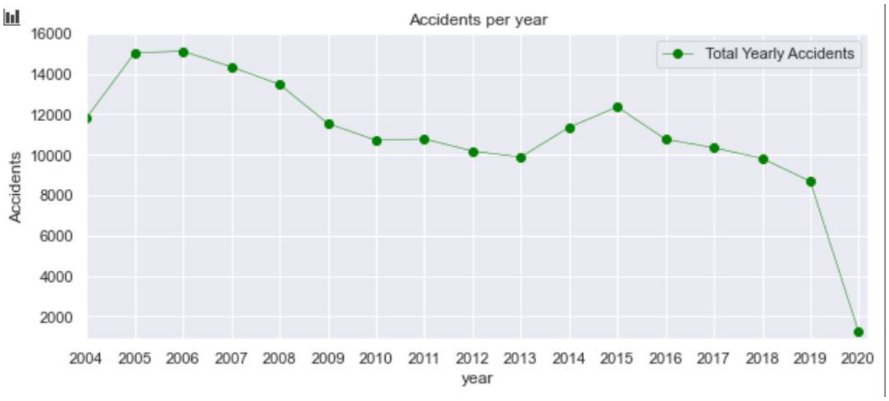


Figure 1: Accidents per year

Figure (1) shows the trend of the accidents through the years. We can observe that the accidents on 2020 diminished notoriously due to the pandemic.

The number of accidents through the months of the years were also analyzed, as we can observe on the next image. Figure (2) shows the number of accidents

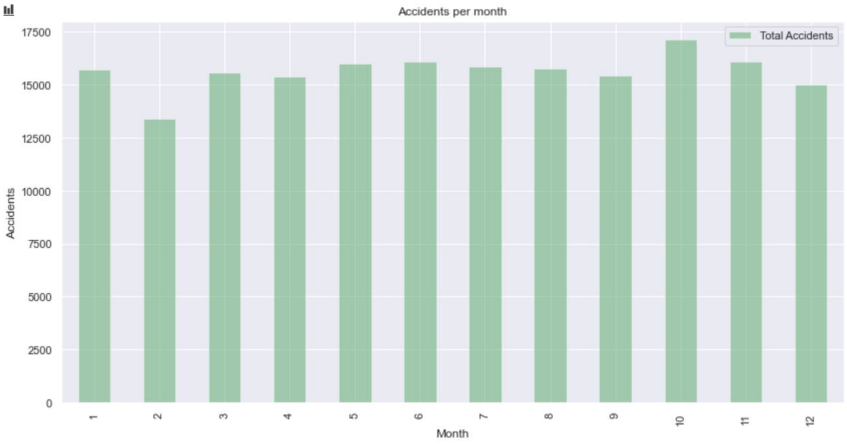


Figure 2: Accidents per month

on each month, we can observe that the values are similar on each month, having October the biggest amount of accidents with approximately 17500 total accidents.

Analyzing the severity of accidents on the following figure we can observe that there are more accidents of low severity than accidents of high severity. Figure (3) shows the number of accidents divided by severity. As we can observe on the figure there are approximately more than 120,000 accidents in the low severity category, and approximately 60,000 accidents on the high severity category.

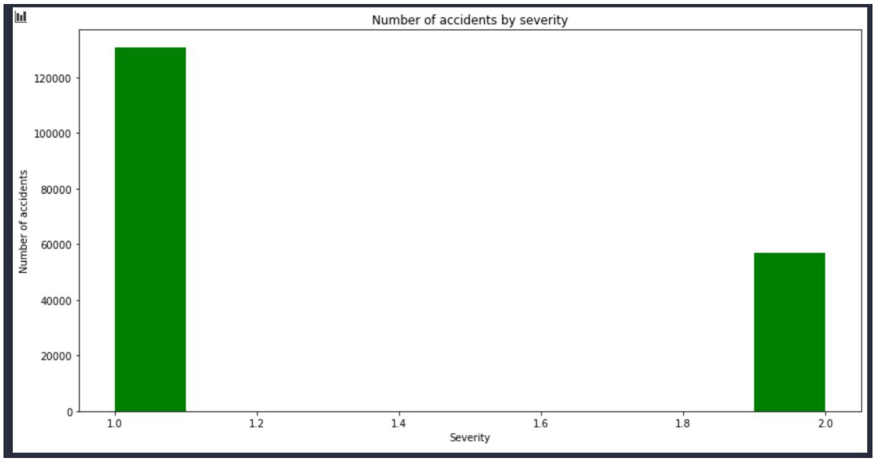
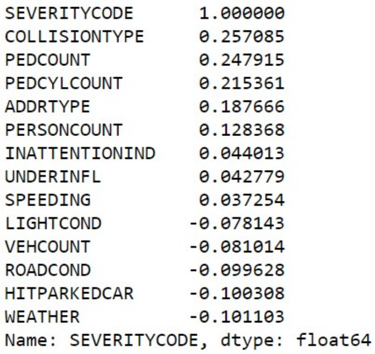


Figure 3: Number of accidents by severity

On this section, it was interesting to analyze the correlation of each independent variable with the target feature, so we can observe how much influence does each variable will have on the prediction. The figure (4) shows the correlation indexes of each variable with the 'SEVERITYCODE' variable or target feature. We can observe that the values are organized in ascending order.



```
SEVERITYCODE      1.000000
COLLISIONTYPE     0.257085
PEDCOUNT          0.247915
PEDCYLCOUNT       0.215361
ADDRTYPE          0.187666
PERSONCOUNT       0.128368
INATTENTIONIND    0.044013
UNDERINFL         0.042779
SPEEDING          0.037254
LIGHTCOND        -0.078143
VEHCOUNT         -0.081014
ROADCOND         -0.099628
HITPARKEDCAR     -0.100308
WEATHER          -0.101103
Name: SEVERITYCODE, dtype: float64
```

Figure 4: Correlation index

3.2 Predictive Modeling

For solving the proposed problem in this case classification algorithms were used. Classification is a supervised learning approach, which attempts to learn the relationship between a set of features and a target variable of interest. The target attribute in classification is a categorical variable with discrete categories or classes. In this project we are analyzing three classification algorithms which are:

1. Logistic Regression.
2. K-Nearest Neighbors.
3. Decision Tree.

To apply these classification methods, the clean dataset at first must be split into train and test datasets, for testing we use the 20% of the data, and all the remaining samples are used for the training stage.

3.2.1 Logistic Regression

Logistic regression is a statistical and machine learning technique for classifying records of a dataset, based on the values of the input fields. In logistic regression independent variables are used to predict a dependent variable. For logistic regression the value of C is fixed on 0.001, and the method used for training the model is the the LBFGS or Limited-memory BFGS which is an optimization algorithm in the family of quasi-Newton methods that approximates the Broyden–Fletcher–Goldfarb–Shanno algorithm (BFGS) using a limited amount of computer memory.

3.2.2 K-Nearest Neighbors

KNN algorithm is a classification method, that takes a bunch of labeled points and uses to learn how to label other points based on their similarity to other cases. For the purposes of this project due to computer processing time only 10 iterations can be processed taking during the interval the best value of K, as it can be observed in the figure (5) the best value of K takes place on 6 neighbors.

3.2.3 Decision Tree

The basic intuition of a decision tree is to map out all possible decision paths in the form of a tree. Decision trees are built by splitting the training set into distinct nodes. One node in a decision tree contains all or most of the categories of the data. For the implications of this project the method applied for solving the decision tree is the entropy.

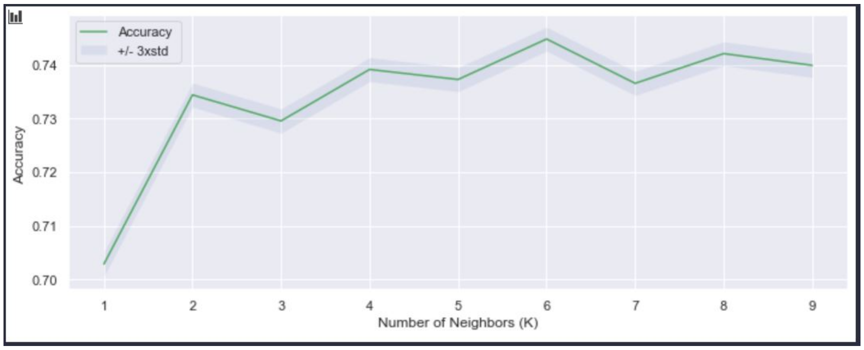


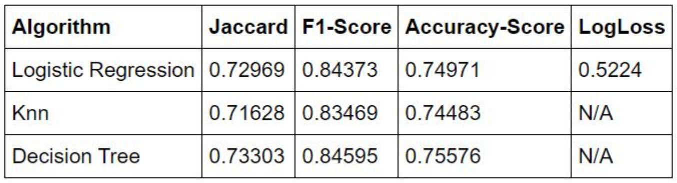
Figure 5: Number of Neighbors K

4 Results

On this section the results of the metrics of each model are analyzed. We analyze the following metrics:

- Jaccard Index.
- F1 - Score
- Accuracy Score
- Logarithmic Loss

The metrics enlisted before are used to compare the results and the quality for each of them between the three models analized in this project.



Algorithm	Jaccard	F1-Score	Accuracy-Score	LogLoss
Logistic Regression	0.72969	0.84373	0.74971	0.5224
Knn	0.71628	0.83469	0.74483	N/A
Decision Tree	0.73303	0.84595	0.75576	N/A

Figure 6: Metrics results report

On the figure (6) we can observe all the accuracy metrics measured for each algorithm. It can be mentioned that the lowest accuracy metric on each study is the KNN algorithm, also KNN method took the largest time to run to obtain the prediction. It can be narrowed down that the Decision tree has the best indexes for Jaccard analysis, F1-score, and Accuracy score. The logistic regression obtained results very close to those generated by the decision tree method, with a shorter implementation an run time which gives us a good approximation and results. On the following section a deeper analysis on the quality vs. computational time will be carried on.

5 Discussion

Observing the results showed on the previous section specifically on the figure (6) it can be observed that the results generated by the Decision tree model were the more accurate generating the best results, but analyzing the Logistic regression model results we observe that the accuracy scores were very similar to those obtained by the decision tree. Due to the extension of the dataset the computational time to apply the classification algorithms increased with the complexity of the model. Analyzing the computational time it can be mentioned that the Logistic regression was the method that took the less amount of time and gave a good approximation. With an accuracy score value of 0.749 against the 0.755 of accuracy score on the Decision tree model. All in all, the logistic regression algorithm presented a very good approximation generating good results having the smallest run-time of all the three analyzed methods.

6 Conclusion

Traffic Accidents are an important cause of deaths worldwide, so analyzing why and what conditions cause that an accident result in fatalities represent a good analysis, because there are several conditions, such as: illumination conditions, road conditions, etc., that influence directly on the fatality score of an accident. This is the reason of the analysis on this project, so anyone who is interested in using this model to predict the fatality of accidents due to several conditions can help saving lives that can be lost on the roads.

Even though the logistic regression model did not present the best accuracy scores, the quality vs. time results were a good metric to validate the quality of the results generated by the logistic regression model.

References

[1] Kambiz Masoumi, Arash Forouzan, Hassan Barzegari, Ali Asgari Darian, Fakher Rahim, Behzad Zohrevandi, and Somayeh Nabi. Effective factors in severity of traffic accident-related traumas; an epidemiologic study based on the haddon matrix. *Emergency*, 4(2):78, 2016.