# Predicting Severity of an Accident by Using Machine Learning

Report submitted by: Kuber Bisht

Dated: 14-Oct-2020.

ABSTRACT

As per World Health Organization (WHO), approximately 1.35 million people die in road accident and injuries from road accidents. The number road accident casualties and damage is much higher in developing countries than in developed countries. Various factors such as weather, light conditions, locations, speeding, inattention etc. are related to traffic accidents, some of those factors are more important in determining the accident severity than others. Machine learning techniques can be employed to determine and predict such influential factors among human, vehicle and environmental factors. In this project, the various classification techniques such as Random forest, K Nearest Neighbors (KNN), Decision Trees and Linear Regression algorithms are used for classifying the type of accident severity of various traffic accidents with the help of influential environmental features of road traffic accidents that can be used to build the prediction model. These techniques were tested using a shared Seattle city dataset. A decision system has been built using the model generated by different classification techniques that will help decision makers to enhance the decision making process by predicting the severity of the accident.

Keywords : Accidents Severity Prediction, Random Forest, K Nearest Neighbors (KNN), Decision Trees and Linear Regression algorithms.

INTRODUCTION

As traffic accidents are one of major threat to the human life worldwide. According to World Health Organization (WHO), there are 1.35 million deaths each year from road accidents and injuries related to such accidents. To control this ever-growing problem extensive research has been carried out into the prediction of traffic accidents in both developed and developing countries using various techniques. Realizing traffic accidents as a preventable problem developed countries have implemented different policies and measures to reduce this problem. These include enforcement, education, training and engineering improvements. Accidents or severity of accidents prediction is important for optimizing public transportation, enabling safer routes, and cost-effectively improving the transportation infrastructure, all in order to make the roads safer.

An accident occurs when a road vehicle collides with another vehicle, pedestrian, animal, or geographical or architectural obstacle. It can result in injury, property damage, and death. Traffic control system is the area, where critical data about the society is recorded and kept. Using this data, we can identify the risk factors for severity of vehicle accidents, injuries and fatalities and to make preventive measures to save the life. The severity of injuries causes an impact on the society.

Understanding the patterns of hidden data is very hard due to data accumulation. Apart from the gathering data it is important to get some knowledge out of it. Using machine learning techniques, the extraction of hidden predictive information from large data is possible as it is a powerful new technology with great potential. It is a useful tool to address the needs for shifting useful information such as hidden patterns from any given data.

The main objective of this project is to build a supervised machine learning model which can effectively predict accident severity using road traffic accidents data using Random Forest, K Nearest Neighbors (KNN), Decision Tree, Logistic regression and XGBoost algorithms.

DATA

The data that was used was the shared Seattle city data. The given dataset had 37 columns and 194673 rows having information about different accidents occurred from Year 2004 through 2020. As meta-data was also available with the data, it helped in understanding about each data element present in the data. After understanding the data, not all columns were needed to build model as these were not directly related to the reason for accident. After finalizing the columns to be used for building model, the data was cleaned. Few columns were updated for the missing information for data consistency. More information was extracted to be used for modeling from existing data. For example datetime type value was used to get Year, Month, Day, hourofday, Weekday etc. Also few column names were changed to give them meaningful name. The final dataset after preprocessing had 17 columns and 180067 rows.

METHODS AND MATERIAL

The main objective of the proposed methodology is to build supervised machine learning model for predicting severity of accident. I used the CRISP-DM methodology for building the supervised machine learning model. The various stages involved were

1. Business understanding

A machine learning model needs to be built to predict the severity of an accident. Ability to predict severity of an accident will help in reducing the number of accidents, loss of lives and development of safer routes and transportation.

1. Data understanding

I used the shared Seattle city data for accidents. It had a meta-data available that provided description about each data column provided in the comma separated (csv) file.

1. Data preparation

The final dataset was used for features selection to build the model.

1. Data classification/modeling

The given features in the data were analyzed and using one hot encoding technique, categorical variables were converted to binary variables and appended to the given data for better efficiency of model. After feature selection there were 140 column and 180067 rows of data. The given data was split in training set and test set.

1. Model Evaluation

Various machine learning classification techniques were used for training and testing the model for getting results. Different techniques were used such as Random Forest, K Nearest Neighbor, Decision Tree, Linear Regression and XGBoost.

1. Deployment/Results

The results of various model evaluations are

|  |  |  |  |
| --- | --- | --- | --- |
| S. No. | Modeling Technique | Training- Set Accuracy | Test-Set Accuracy |
| 1 | Random Forest | 99.99% | 73.73% |
| 2 | K Nearest Neighbors (KNN) (k=10) | 74.49% | 71.37% |
| 3 | Decision Tree (max\_depth=10) | 74.40% | 74.05% |
| 4 | Logistic Regression | 74.18% | 74.16% |
| 5 | XGBoost | 76.81% | 73.98% |

DISCUSSION

After using various modeling techniques, the logistic regression provided best results. The training set accuracy was highest for Random forest but test set accuracy for almost all models was approximately 73% to 74%. This might be due to unbalance in the given data as the number of severe accidents were very less in number as compared to other accidents. Also, the dataset was small and it might not had much data for severe accidents, that has resulted in lower accuracy as all features might not be available for predicting the severe accidents to predict severity with greater accuracy. We can use the same methodology on a larger dataset to get good results.

CONCLUSION

After analyzing and modeling the data, we can say most of the severe accidents happened on dry roads, clear weather and near intersection and were related to the intersection. A lot of severe accidents can be avoided if more precautions are used near intersection. For predicting severity of accidents with greater accuracy a large dataset is needed and it should be updated periodically and accurately. Model should also be updated with new data to increase the accuracy of predicting the severity of an accident.