Car Collision Severity Prediction Model

Naveen Kumar Mangal

September 12, 2020

1. Introduction

a. **Background:** Car accidents severity prediction is important to alert the people which is based on the important attributes. It will help to choose proper driving conditions for safe travel. Some bad conditions and wrong decisions can lead to road accidents of different severity. Safe driving can be achieved if we have a system which can predict the severity conditions with the help of input parameters given by users.

b. **Business Problem:** Our objective is to build a model which can predict the road accident severity and alerts the driver to be more careful while driving. It will reduce the chances of accidents and severity.

c. **Target Audience:** This project is particularly useful for the for the people who are looking in advance for their safety during road travel. The prediction model will decide the collision severity based on the attributes given to the model. Also, this project will be helpful for the insurance companies for after collision claim settlement by analyzing the all the conditions related to driving. With the help of the prior data, this model gives the output in terms of severity associated with incident.

2. Dataset Description:

We have used the weekly updated traffic dataset provided by Seattle government that contains road accident and corresponding severity based on the different attributes. The link of the dataset is given below: https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv

a. **Data Pre-processing:** The above dataset in its original form is not ready for analysis. In data pre-processing steps, first, we need to eliminate the non-relevant columns. Additionally, most of the attributes are of object data types that need to be converted into numerical data types. After analyzing the data set, I have chosen four attributes namely, severity, weather conditions, road conditions, and light conditions, among others. Since I have checked different values in the attributes for good understanding of dataset. the analysis results show that target feature is imbalance, so I use a simple statistical technique to balance it. I simply down sampled the class numerical values which are higher in number to balance the dataset (please see the notebook section for details)

b. **Attribute Selection:** The above dataset has 37 attributes which act as independent variable for us to build the prediction model. The target variable will be the collision severity code which ranges from 0 to 3 namely; 0 for unknown, 1 for prop damage, 2 for injury, 2b for serious injury, and 3 for fatality. Here we will choose the attributes which strengthen our prediction model. In term of Data Science, we should select the attributes which column has fewer empty cells. After uploading the dataset, I have used Dataframe, used 'dtypes' attribute to check the feature names and their data types. Then I have selected the most important features to predict the severity of accidents in Seattle. Among all the features, the following features have the most influence in the accuracy of the predictions:
   i. "WEATHER",
   ii. "ROADCOND",
   iii. "LIGHTCOND"

3. Methodology:

a. **Exploratory Data Analysis:** In this approach, we do initial investigation of the dataset to summarize its patterns, spot anomalies and characteristics. It consists of visual methods (Data visualization) which tell us about the data and help us in modelling. The steps are as follows:
   i. I have used head() function of pandas library to observer the first 5 rows of the car collision dataset.
   ii. I have used shape() function to see the size of dataset in terms of rows and columns.
   iii. I have used describe() function, which told about count, mean, standard deviation, minimum and maximum values and quantiles of the data.
   iv. The mean values of each columns are greater than the median values (represented by 50% row)
   v. Few of columns have noticeable difference between 75% and the max value, which shows a smaller number of outliers in the dataset.
   vi. I examined the dependent/target variable "SEVERITYCODE" and results shows that it is discrete and categorical in nature.
   vii. I have used corr() function using seaborn library to visualize the heatmap. It is good practice to remove the correlated variables during the attribute selection.

viii. Box-plot shows the distribution of the quantitative data which helps in the comparison of variables. It shows the five number summery namely, minimum, first quartile, median, third quartile, and maximum.

b. Machine Learning Algorithms selection: For car collision severity prediction model development, I have used Github as a repository and running Jupyter Notebook to preprocess data and build Machine Learning models. In coding part , I have used Python and its popular packages such as Pandas, NumPy and Sklearn. Three machine learning models namely, **K Nearest Neighbour (KNN), Decision Tree, Logistic Regression**. After importing necessary packages, I defined attributes and target variables, splitted preprocessed data into test and train sets, for each machine learning model,

4. **Results:** Further, I have built and evaluated the model (please see jupyter notebook for details) and analysis results are as follows:

Model	**Jaccard Score**	**F1 Score**	**Accuracy**
kNN	0.560	0.514	0.560
Decision Tree	0.563	0.532	0.563
Logistic Regression	0.53	0.517	0.532

5. **Discussion:** In the beginning of this notebook, we had categorical data that was of type 'object'. This is not a data type that we could have fed through an algorithm, so label encoding was used to created new classes that were of type int8; a numerical data type. After addressing and solving the above, we were presented with another - imbalanced data. As mentioned earlier, class 1 was nearly three times larger than class 2. The solution to this was down sampling the majority class with sklearn's resample tool. We down sampled to match the minority class exactly with 58188 values each. Once we analyzed and cleaned the data, it was then fed through three ML models; K-Nearest Neighbor, Decision Tree and Logistic Regression. Although the first two are ideal for this project, logistic regression made most sense because of its binary nature. Evaluation metrics used for our models were **jaccard index, f-1 score, and accuracy**. The selection of different k, max depth and hyparameter C values helped to improve our accuracy to be the best possible.

6. **Conclusion:** Based on historical data from weather conditions pointing to certain classes, we can conclude that particular weather conditions have a somewhat impact on whether or not travel could result in property damage (class 1) or injury (class 2).