Predicting car accident severity

Mousuf Shaikh

August 28, 2020

1. Introduction

Various types of Car Accidents in traffic lead to associated fatalities and economic losses every year worldwide and thus is an area of primary concern to society from loss prevention point of view. Modeling accident severity prediction and improving the model are critical to the effective performance of road traffic systems for improved safety. In accident severity modeling, the input vectors are the characteristics of the accident, such as driver behavior and attributes of vehicle, highway and environment characteristics while the output vector is the corresponding class of accident severity.

I will be using the Collisions dataset provided by SDOT Traffic Management Division, Traffic Records Group in order to analyse the accident severity using several parameters.

2. Data

The dataset in the SDOT(Seattle traffic department) dataset has 194673 rows of data and contains 38 columns including the severity(target) column. The target variable severity code has either 1 or 2 analysed by using .unique attribute after loading the dataset into a dataframe. All the columns in the dataset will be analysed before selecting the columns as model parameters. Various classification models like logistic regression, SVM,will be compared based on their accuracy and the best will be chosen as our final result. The models will be accessed based on several metrics like F1 score, Jaccard score, Logloss score. The dataset is unbalanced with severity 1 having 58188 while severity 2 has 136485 entries.

3. Methodology

The main idea here is to analyse the SDOT dataset and to check for various correlations between the attributes in the dataset and hence to predict the severity of the accident.

Firstly the variables are analysed and checked whether to include in the analysis or not. Then each of the remaining columns are analysed for number of null values which do effect our analysis and the columns with high number of null values are dropped. Some of the variables/attributes in the dataset are omitted and the dataset trimmed for ease of analysis and developing correlations.

Several visualizations are provided here which analyse the distribution of accident report with respect to time of the day, month and year.

Remaining columns are analysed and checked for rows containing null values and those rows are dropped if any. Afterwards the severity code distribution is checked. Similarly only most frequent type of collision is checked and distribution obtained. Other variables are checked

for label counts. Correlation matrix is developed. Number of vehicles involved in the accident is plotted.

Finally the data is prepared for some classification algorithms to be tested. The labels are encoded with the help of feature engineering tool and only some top frequent labels are included for analysis to reduce the number of features.

Logistic regression, decision tree and random forest algorithms is tried on the prepared dataset and accuracy measured.

4. Visualization of attributes in the dataset

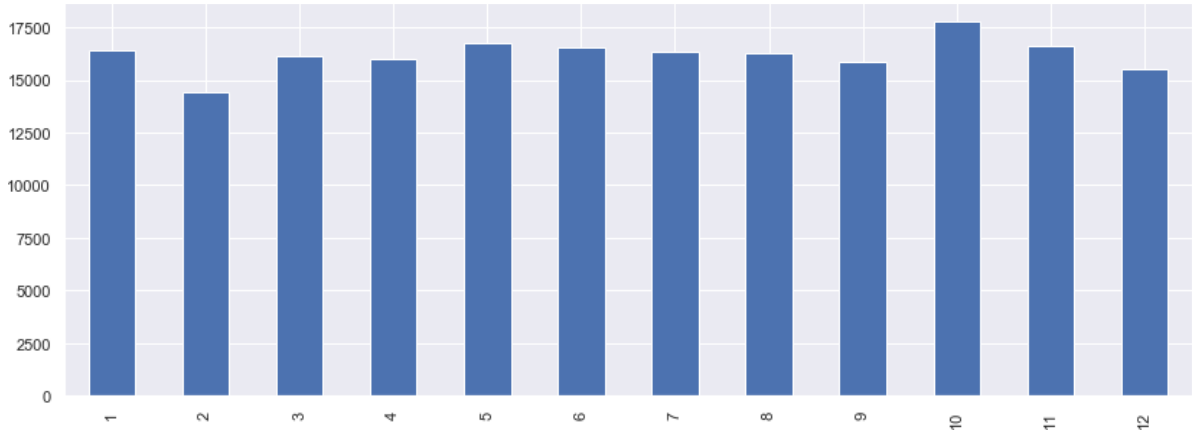

Figure 1: Distribution of accident report by month

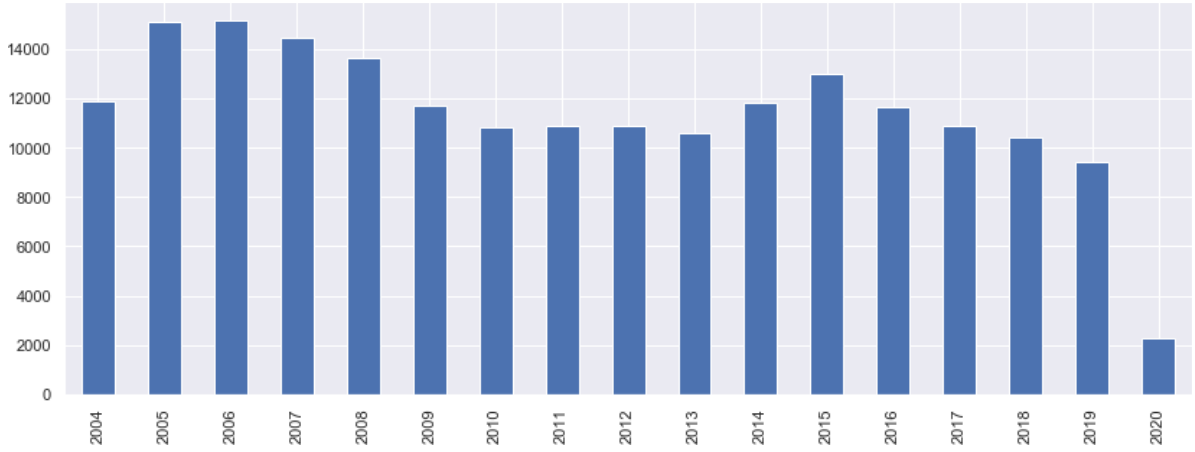

Figure2: Distribution of accident reporting by year

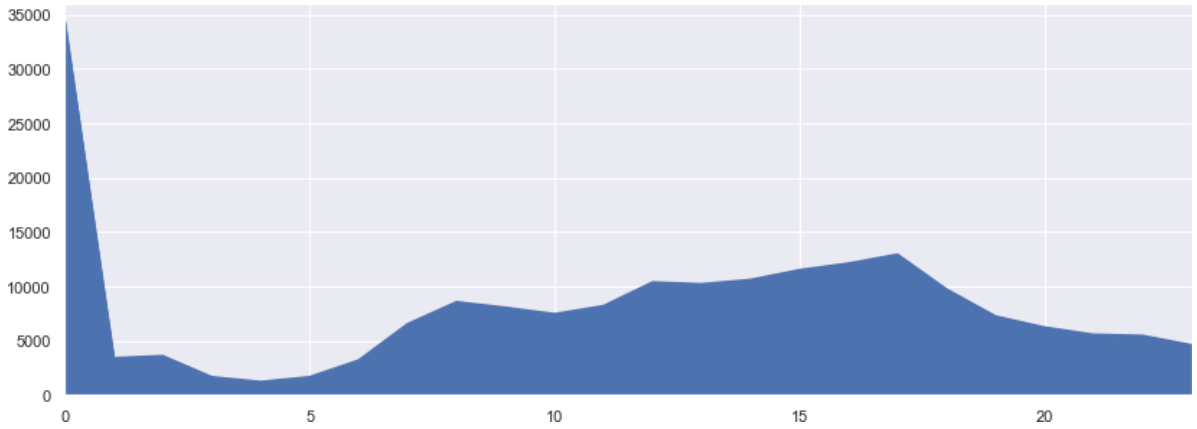

Figure 3: Distribution of accident reporting by time of the day:

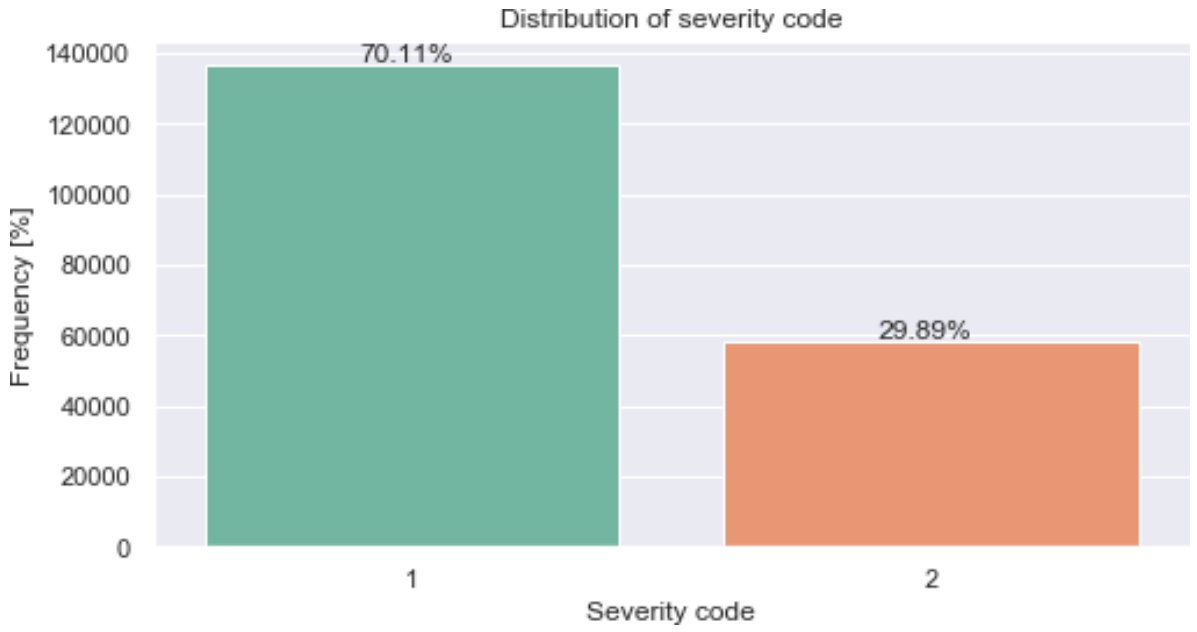

Figure 4: Distribution of target variable

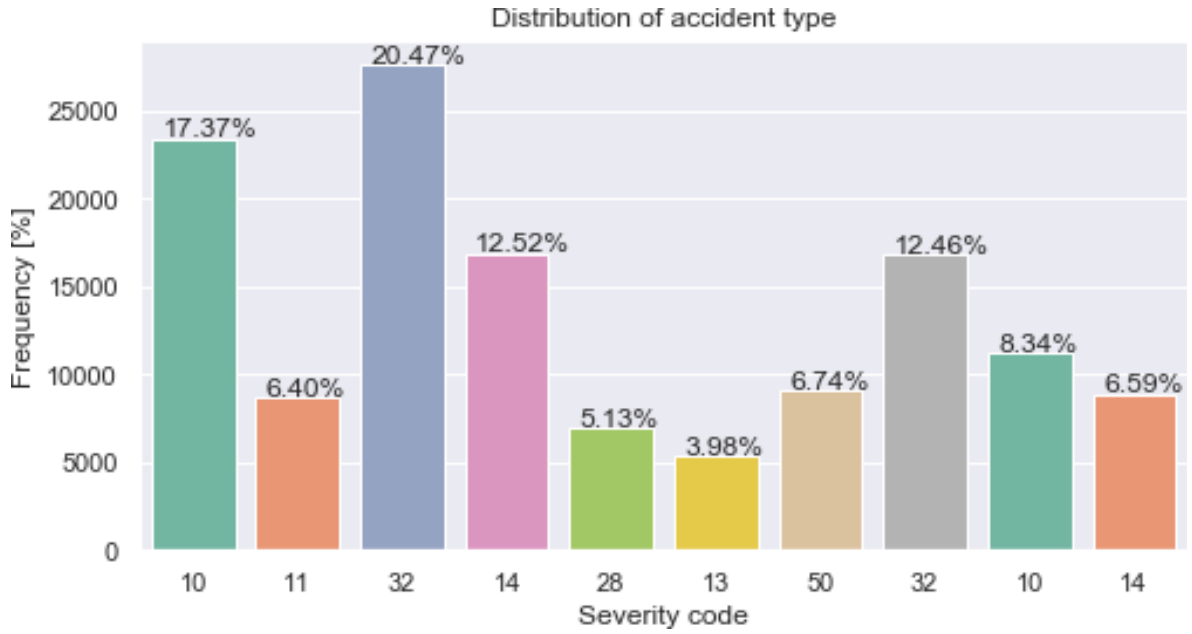

Figure 5: Distribution of severity code types:

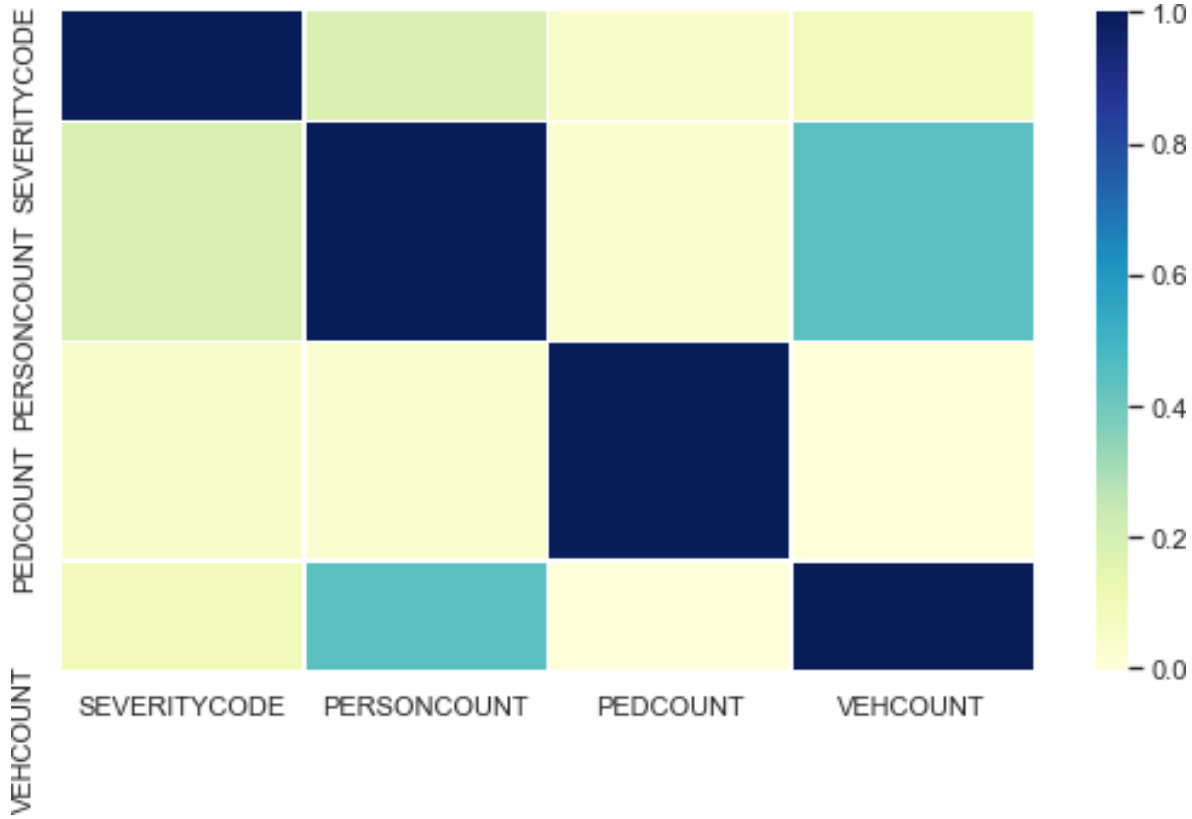

Figure 6: Correlation matrix:

5. Conclusion section:

As seen the visualizations give a certain understanding of variables in the columns. The distribution of report incident suggests that most of the accidents are reported after midnight hours. It also suggests that frequency of the accidents is more for the month of October. Distribution of target variables suggest that the tags are unbalanced with severity 1 having higher frequency that severity 2. Severity code distribution of the most frequent accidents suggest that accidents with severity code 32 i.e. One Parked - One Moving is the most common followed by severity code 10 i.e. Entering At Angle.

The output of the classification models like logistic regression, decision trees and random forest give almost perfect prediction with precision and recall of 1. This might be because of over defining the problem. Many features like vehicle count and person count are heavily correlated which in turn represent the severity of the accident. Hence we are getting perfect accuracy scores using these classification methods.