Predicting the probability of vehicle injury or fatality in the Seattle area

Deepak Mitra (Sept'20)

Problem background and understanding:

America has seen, in 2018, over 36,500 fatalities due to road/ Motor vehicle related incidents. (Source Wikipedia - https://en.wikipedia.org/wiki/Motor_vehicle_fatality_rate_in_U.S._by_year) This rate, though reducing, is still high and is over 11 fatalities per 100,000 people in the US.

In Seattle there are over 13,000 crashes a year with an average of 20 deaths and 150 people serious injuries every year. While the trend might have dipped in 2018, it has gone back up in 2019. (Source - https://sdotblog.seattle.gov/2019/12/10/slower-speeds-save-lives-a-path-to-end-traffic-deaths-serious-injuries-in-seattle/)

With the availability of historic and detailed data, coupled with modern computer capabilities in AI/ML - an attempt is being made to reduce avoidable fatalities on the roads by predicting conditions that result in high probabilities of road incidents resulting in injuries or even deaths.

Models can be built and tuned to use the SDOT historic data combined with current/ real-time data like weather, traffic conditions, time of day/ month/ year etc to advise authorities where and when (say) temporary speed limits need to be implemented or intersections need extra lights etc; or to advice commuters to avoid certain roads during certain conditions, or to advise pedestrians to navigate away from certain roads/ intersections.

The model so built should be accurate and consistent so that users gain trust in its capabilities and even a single life saved would be a success.

SDOT Data structure:

The data in provided in the course is from Seattle's DOT (SDOT) Traffic Management Division (Traffic records group) covering the periods from Jan'04 to May'20. It contains nearly 200,000 instances of incidence, measured across 38 different variables.

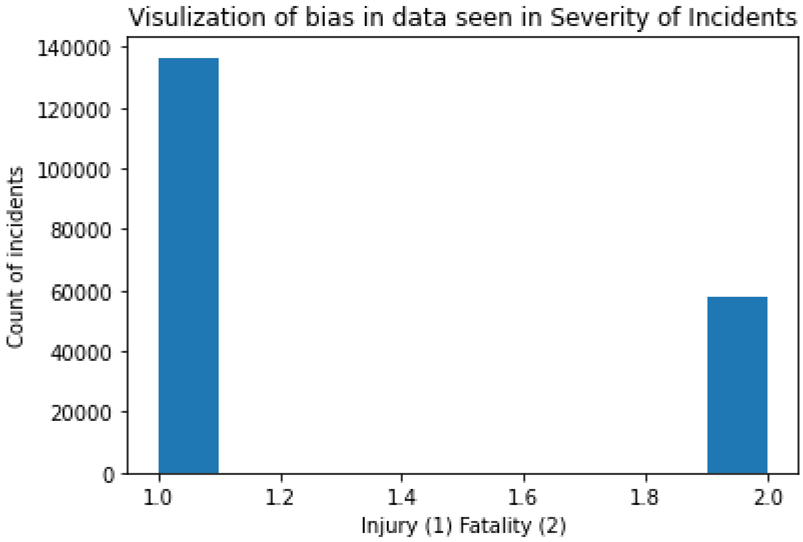
The data is labeled and largely has categorical attributes. This easily lends itself to a supervised, categorical Classification model like Decision Tree, Logistic Regression or SVM.

For this exercise, I intend to use the Logistic Regression model as it allows us to understand the impact an independent variable has on the dependent variable while controlling other independent variables.

Additionally, we can determine the probability of accuracy of the model's categorical outputs.

The target variable for the dataset is the "SEVERITYCODE" attribute which is binary - either 1 or 2 - indicating fatality in the incident (2) or injury (1). The feature variables selected for the model training are 'LOCATION','ROADCOND' (Road conditions),'LIGHTCOND' (Light conditions),'WEATHER','SPEEDING' (Speeding Yes/ No),'JUNCTIONTYPE','PERSONCOUNT' (Persons involved),'VEHCOUNT' (Vehicles involved).
There are a few outliers like one incident with 81 people involved in a single incident as well as a single incident involving 12 vehicles.

The data seems biased as the histogram below indicates. There are a total of 136,485 readings for injuries vs 58,188 readings for fatalities during the period of the data.



The model developed will predict with great probability the likelihood of a commuter injury or fatality - if the conditions related to weather, lighting, road conditions etc monitored in the model are met. This result can be obtained by users if the model is shared via an API and users can call for the model result automatically with current data attributes for the independent variables like weather or road conditions.

Methodology:

The pandas framework was used to process the data from SDOT and clean the data so that it could be readily used for model use. The data was first visually inspected and biases and outliers identified. There were 38 variables in the dataset, however, only 7 were selected based on their obvious correlation with road safety. The independent variables selected were checked for incomplete data – and any non a numeric (NaN) value was dropped from that respective data entry.

The Speeding variable had only Y for speeding – this is vital information and was not removed – instead, as this was a binary Y/N data entry – the entries for missing data was updated as a numeric – to prevent information loss.

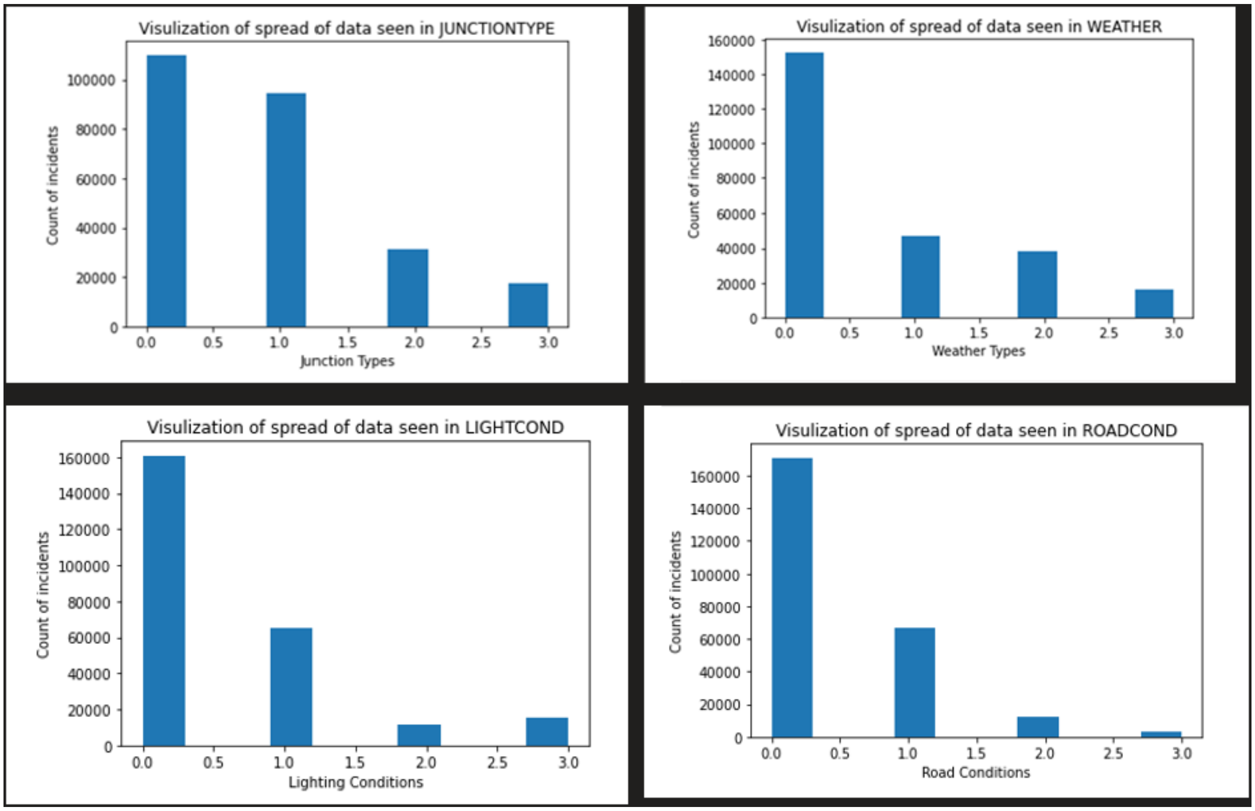
After this first level of cleaning, the data contained 183,196 entries (vs 194,673 entries in the original dataset) across 8 variables (vs 38 in the original dataset).

As mentioned earlier, the data labels were unbalanced, and therefore had to balanced to remove bias. This was done next, in order to use the maximum information in the dataset before further processing and data cleaning removed information for the dataset.

Post upscaling of the dataset – the new dataset (df_data_3) contained 253,054 entries over 8 variables. This data set was now ready for further processing.

The selected feature set of independent variables unfortunately were mostly categorical in nature – like weather conditions, road conditions, lighting conditions etc. and therefore had to be processed so that the models could use the dataset and data could be trained on the model.

To do this, Label Encoding was done to group the categorical variables to numeric variables, without affecting the inherent nature of the data in each variable set or the number of entries in the dataset. Below is the reflection of the processing on the categorical variables.



The last pre-processing done on the dataset was the carving out of the feature set (X) from the dependent variable (y – severity) and applying the StandardScalar processing on the feature set.

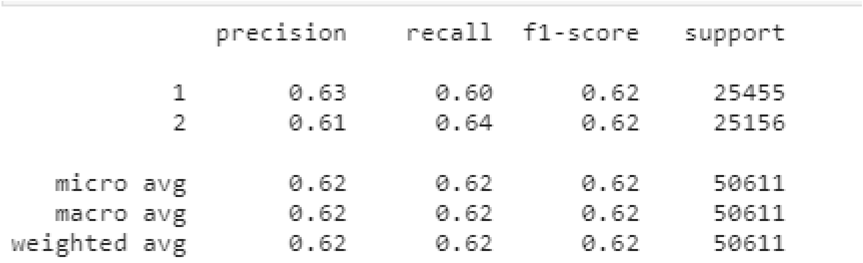
The model I selected for this exercise is the Logistic Regression model. As the out put of the model needed to be Classification and Categorical in nature – i.e. Sever or not sever – a Regression model, Decision Tree or Support Vector Machine (SVM) could be used. However, with the LR model the probability of the classification selected is additionally easily checked – this model was best suited for this dataset in my opinion. An additional advantage is that as it allows us to understand the impact an independent variable has on the dependent variable while controlling other independent variables, making the LR model flexible besides being accurate.

The dataset was split 80/20 between train and test data subsets and the “Liblinear” numerical optimizer was used.

Results:

The results for dependent variable predictions based on the LR model fitted to the selected variable feature set was fairly accurate.

The Jaccard Similarity score showed a value of 62% and F1 score too was a 62%. Below is the classification report for the model fitment.



```
              precision    recall  f1-score   support

           1       0.63      0.60      0.62     25455
           2       0.61      0.64      0.62     25156

   micro avg       0.62      0.62      0.62     50611
   macro avg       0.62      0.62      0.62     50611
weighted avg       0.62      0.62      0.62     50611
```

The probability of the model accurately predicting severity or non-severity was 65% based on the Log Loss methodology.

Observations and Recommendations:

Based on the dataset and the results observed, it can be deduced that under particular conditions – like under wet road conditions, in the night, in rainy weather and at intersections – the probability of a person meeting with a sever injury increases by over 50%.

Using this data, authorities/ users can monitor these variables based on other data sources and feed this LR model continuously/ regularly so as to pre-empt any motor vehicle incident.

Based on the model recommendation – the authorities/ users can take actions like lower speed limits on approaches to intersections or check for additional lighting in areas that have frequent accidents for poor lighting.

Perhaps the model can be further developed to include location data as a variable to provide a easily readable map of Seattle where motor vehicle related incidents are highlighted as the dependent variables change.

In Conclusion:

Based on the exhaustive data collection by the SDOT, applying the AI/ML tools of Logistic Regression with probabilistic assurance on the outcome and using pandas to manipulate the dataset an algorithm has been developed that can predict the severity of an accident at various times/ conditions in the city of Seattle.