**Improving the situations on road**

Suyash Prasad

1. **Introduction**
   1. **Background**

Traversing on road is the most common mode of transportation people do in day to day life on daily basis. It may be pedestrian crossing the road, someone cycling across the street or using some motor vehicles like car or motorbikes for personal use or maybe the use of public transport like bus or trucks for various purposes. But with accessibility and reach that these have, there are also many numbers of road accidents taking place while travelling on roads. It is causing a severe effect on well-being and prosperity of the country. So we must look into the reasons of this cause and try to improve the condition as much as possible.

* 1. **Problem & Interest**

The well-being and prosperity of citizens of a locality or country is responsibility of a government. So in regards to road safety there are a lot of factors that need to be taken care of like road condition, lighting condition, junction-type etc. So this data can be used by the local government body to improve the situations on road. This can also be used to provide better and fast aid to the people injured in the accident. People itself can also use it to decide whether to take that route or take some other route to travel on a particular day and time.

1. **Data acquisition and cleaning**
   1. **Data collect**

This dataset is the example dataset of the Coursera capstone project, which can also be accessed by this link (<https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collision.csv> ).

* 1. **Data Cleaning and feature selection**

The given data needs good amount of cleaning before getting used for the purpose of fatality prediction. There are a lot of missing row values in many columns that needs to be handled properly. In column “Status” two values are given (matched and unmatched) showing the status of given data after being verified. So the data that were not matching the record i.e. the rows marked “unmatched” had to be removed first.

Now there are a lot of columns that are needed to be removed because they doesn’t seem to have any meaning towards severity of the accident. Latitude(**X**) and longitude(**Y**) of the location of accident doesn’t seem to make any meaning towards severity of the accident, and for the same reason they had to be removed. **OBJECTID** is ESRI unique identifier, **REPORTNO, INCKEY** and **COLDETKEY** are unique identifier of the incident. **EXCEPTRSNCODE** and **EXCEPTRSNDESC** have less entries. **SEVERITYDEC** is just the description for the code **SEVERITYCODE** so basically both are same. Because of convenience **SEVERITYCODE** was selected out of two. **SDOT\_COLCODE** and **SDOT\_COLDESC.** In the similar fashion **ST\_COLCODE**, **SEGLANEKEY** and **ST\_COLDESC** doesn’t seem to be useful. **INCDATE** and **INCDTTM** are data and date-time of the accident that will just be helpful to the police and not to us for our classification.

Now from the rest of the feature we have to select the best feature that can be good for our analysis. **LOCATION**, **CROSSWALKKEY** is a categorical variable but has too many values that makes it a bad data for our use. **JUNCTIONTYPE** is a better description to tell about the junction of collision rather than **ADDRTYPE** so we’ll go with the former one. **COLLISIONTYPE** tells about the kind of collision that has occurred which can be a pretty good factor to predict collision. **PERSONCOUNT** and **VEHCOUNT** are very good identifier for our purpose. **PEDCOUNT** (no. of pedestrian involved in the collision) and **PEDCYLCOUNT** (no. of bicycle involved in the collision) are also good determining factor but in our data more that 95% of the cases involved 0 count in these columns. If the person was attentive or not at the time of accident and was he under the influence of any drugs or alcohol is also a good decider so we’ll also have **INATTENTIONIND** and **UNDERINFL** as our features for the analysis. Condition of road, light and weather condition are main physical features that can be very good determination for any accident so **ROADCOND**, **LIGHTCOND and** **WEATHER** will also be selected as the determining factors. **SPEEDING** is also a good feature seeing how speed of vehicle can lead to disasters. **PEDROWNOTGRANED** and **HITPARKEDACAR** couldhave been good for our classification purpose but our data is biased towards no that’ll not be proved good for the classification.

Target: **SEVERITYCODE**

Features selected for further steps:

* **JUNCTIONTYPE**
* **VEHCOUNT**
* **PERSONCOUNT**
* **COLLISIONTYPE**
* **UNDERINFL**
* **INATTENTION**
* **SPEEDING**

1. **Exploratory data analysis**
   1. **Target variable of the analysis**

Severity of the accident is the target variable for our analysis. There are two value 1 and 2. 1 corresponds to property damage only collision and 2 corresponds to injury involved during the collision.

* 1. **Relation between severity of accident and no. of vehicle involved in the accident**

As it’s generally thought as the number of vehicles involved in the collision increases the severity of the accident of the accident should increase.

We have got the same result by the analysis but with the exception when the vehicle count is 1. It may be the result of the fact that single vehicle could have got into accident due to bad road and person involved has got injured.

For the vehicle count of 1,2,3,4 and 5 we have injury ratios of .54, .24, .42, .44 and .49 respectively. Fig1. Depicts the same thing.

* 1. **Relation between severity of accident and no. of vehicle involved in the accident**

In this case also we think that as the number of person involved in the accident increases the chances of injury of people involved gets increases. And in fig 2. We can observe the same things without any kind of exception. In the fig 0 corresponds to the case when the count of people involved is greater than 5.

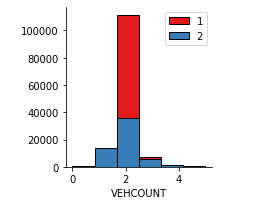


Fig1: Vehicle count v/s severity of accident

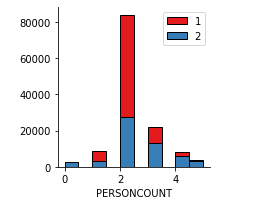


Fig2: Person count v/s the severity of accident

* 1. **Relation of junction type with the severity of accident**

As it can be predicted that at the intersection there is the most chances of injury, the same thing is predicted by the statistics. Mid-block (not related to intersection), at intersection (intersection related), mid-block (but intersection related), driveway junction, unknown, at intersection( but not related to intersection) and ramp junction are converted to whole number 0-6 in alphabetical order.

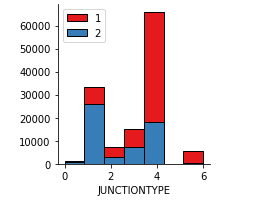


Fig3: Relation between junction type and severity of accident

* 1. **Relation between collision type and severity of accident**

As anyone will predict that if pedestrian or cyclist gets involved in a collision they are least protected to injury as they have either no protection or very less protection for themselves. And the injury will be least if the collision is with the parked car. That’s the exact result we have got from our statistics. In 89% of the cases when pedestrian was involved there was injury involved in the accident and with the cyclist it was 87%. And only 5.5% of the cases involved injury when the collision was with a parked car. Most of the collisions involved parked car, which reduces the overall injury ratio of our data. Fig 4. Gives the representation of this statistics.

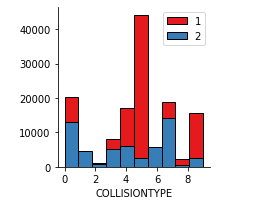


Fig4: Relation between collision type and severity of the accident

* 1. **Relation between severity of accident and if the driver involved in the accident was under influence of any drugs or alcohol**

As one would say chances of injury increases if the driver is under the influence of drugs owing to the fact that driver’s control over the vehicle reduces which can lead to very bad accident hence increasing the chances of the injury. The same is shown by our data but as expected chances of injury doesn’t change drastically if the driver is under the influence of ay drugs or alcohol. It only increases injury chances by 10% from 29% to 39%. Fig5. represents the same fact.

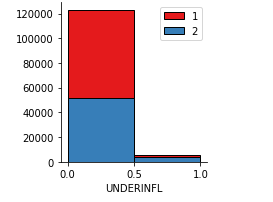


Fig5: Relation between driver being under influence and severity of the accident

* 1. **Relation between driver’s inattention and severity of the accident**

Here also our data show something uncommon from the general opinion. Inattention doesn’t increase the chances of injury very drastically. There is just a 5% increase in the chances of injury from 30% to 35% with the inattention. 0 represents that there was no inattention and 1 is represention of inattention. And in around 15% of the cases only there was inattention. Fig 6. Depicts the analytical finding.

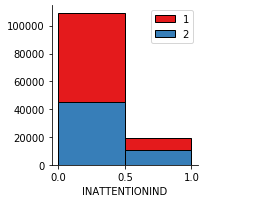


Fig 6: Relation between driver’s inattention and severity of the accident

* 1. **Relation between speeding of the vehicle and severity of the accident**

As expected there is more accident injuries occurred with the speeding of the vehicles. Chances of injury increased from 29.6% to 37.8% when the speeding was recorded in the accident cases. In fig7, 0 represents there was no speeding and one represents the speeding of the vehicle. Only in around 5% of the cases speeding was recorded.

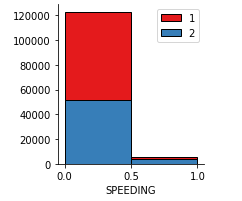


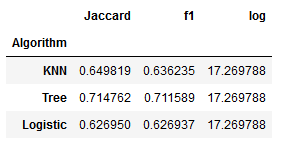
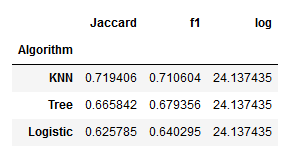
Fig7: relation between speeding and severity of the accident

1. **Predictive modelling**

For a good classification the more data is balanced it’s better for the prediction of result of the test data. But our dataset is biased towards 1. So the results we’ll get will be good while we have to predict but it won’t do the justice with 2 i.e. injury cases. So we need balance the training data first before putting it for prediction to that our accuracy of the prediction gets increased.

* 1. **Classification models**

Three classification models were used for making the predictions; K nearest neighbor, decision tree and logistic regression. And then accuracy of these models were evaluated using Jaccard similarity score, F1-score and log loss. Accuracy was checked in both train data and testing data. Decision tree was best while evaluating the training set with the accuracy around 71% and K nearest neighbors was best when used upon test dataset with the accuracy around 71%.

Fig 8: Training data accuracy report  Fig 9: Test data accuracy report

1. **Conclusions**

Form fig 10. it can be seen that we can expect more chances of getting correct prediction when severity code is 1. But when the severity code is 2 there is around 50-50 chances of getting the right prediction. This whole analysis can be used to improve the situations on the roads so that less accidents occur and people can feel safe while travelling the roads of Seattle.

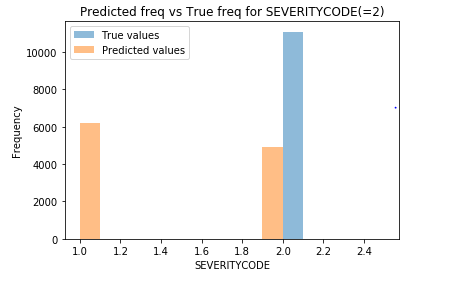
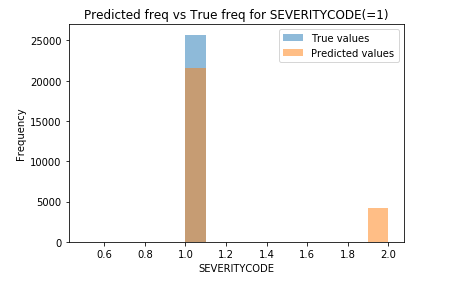


Fig 10: Correct and false prediction for different severity codes

1. **Future scope**

We can collect more data related to injuries and hope for better prediction of the model and also if we can have some more specifics like speed divided into various ranges and if the driver was under the influence of drugs or alcohol and was he/she under the sever influence or only a little bit of influence. Analyst can also consider some other features which might change the prediction quite considerably.