**Car Accident Severity Prediction**

Aurindom Mahanti

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 a 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, if the driver was under influence of drugs or alcohol 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-Collisions.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 don’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. The feature set is selected by looking at the Pearson correlation coefficient or Chi-square coefficient. Here we have used Pearson correlation, and hence we convert the categorical values to variables of type int64, int32 or float.

COLLISION\_NUM -0.128429

WEATHER\_NUM -0.110431

VEHCOUNT -0.080807

LIGHTCOND\_NUM -0.061206

ROADCOND\_NUM -0.049446

LOCATION\_NUM -0.038109

SDOTCOLNUM 0.004528

INTKEY 0.005955

X 0.010369

Y 0.018359

OBJECTID 0.026527

INCKEY 0.029442

COLDETKEY 0.029465

SPEEDING\_NUM 0.038249

UNDERINFL 0.043761

INATTENTIONIND 0.044839

SEGLANEKEY 0.104412

PERSONCOUNT 0.129782

CROSSWALKKEY 0.175533

SDOT\_COLCODE 0.186127

PEDCYLCOUNT 0.215718

PEDCOUNT 0.248121

SEVERITYCODE.1 1.000000

SEVERITYCODE 1.000000

Looking at the above table we can conclude saying that the best features to consider can be the ones which have a coefficient <0.01

**LOCATION\_NUM, INTKEY and SDOTCOLNUM** are good feature variables but have too many values that makes it a bad data for our use.

Data like **SPEEDING\_NUM,UNDERINFL** and **INATTENTIONIND** have slightly lesser correlation but always have influence on such cases. Hence we consider them as well.

Target: **SEVERITYCODE**

Features selected for further steps:

* **VEHCOUNT**
* **LIGHTCOND\_NUM**
* **ROADCOND\_NUM**
* **WEATHER\_NUM**
* **COLLISION\_NUM**
* **UNDERINFL**
* **INATTENTIONIND**
* **SPEEDING\_NUM**

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 Features involved in the accident**

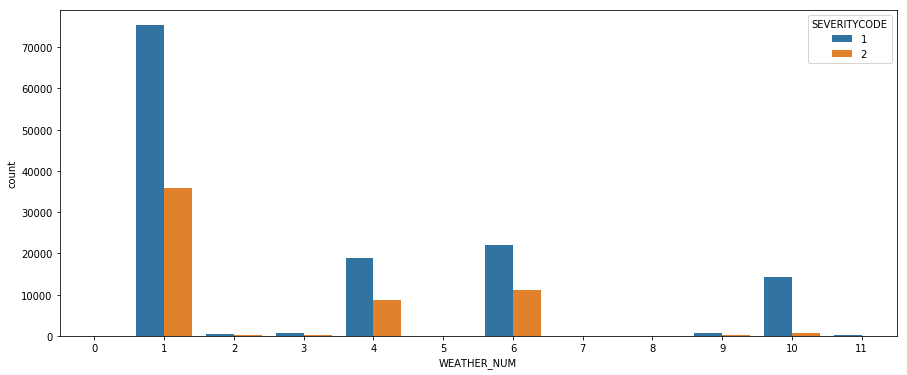


Fig1: Weather v/s severity of accident

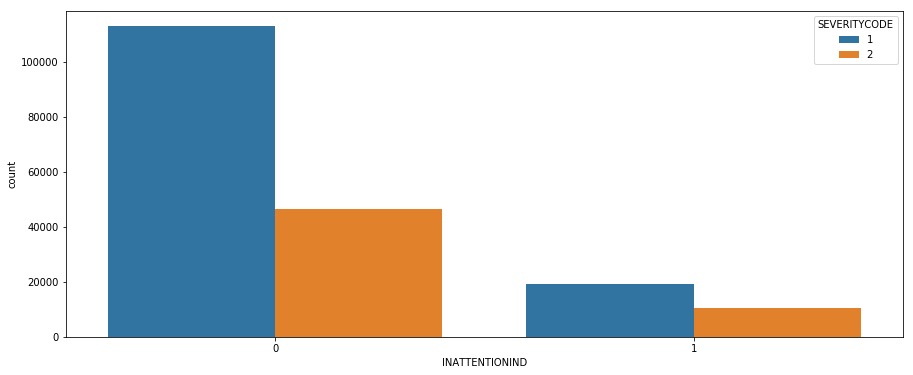


Fig2: INATTENTIONIND v/s severity of accident



Fig3: ROADCOND\_NUM v/s severity of accident

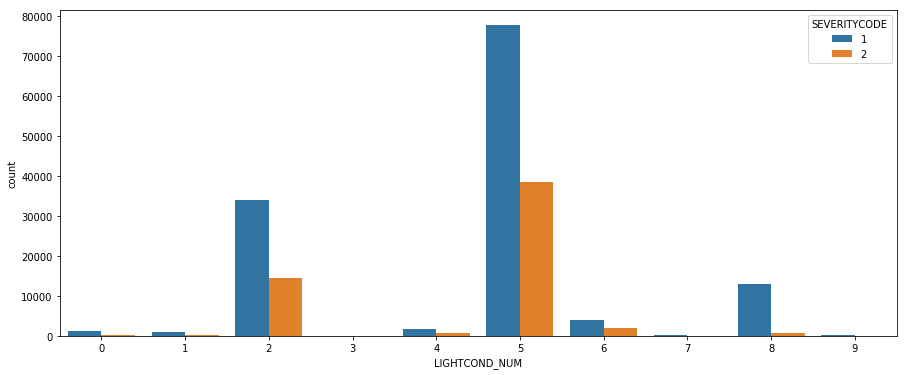


Fig4: LIGHTCOND\_NUM v/s severity of accident

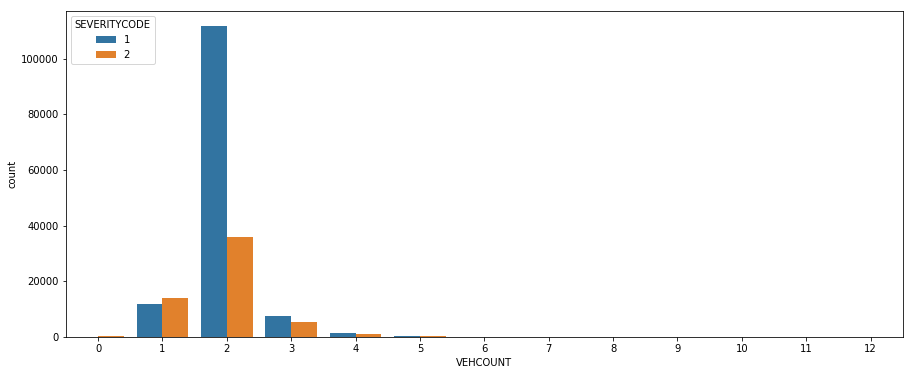
****

Fig5: VEHCOUNT v/s severity of accident

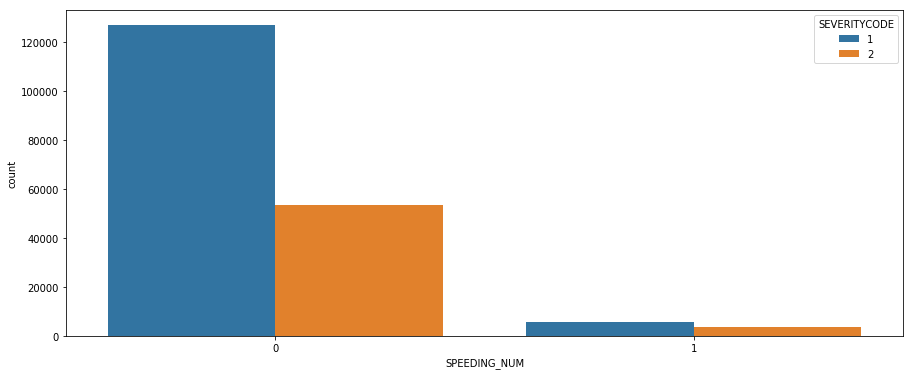


Fig6: SPEEDIND\_NUM v/s severity of accident

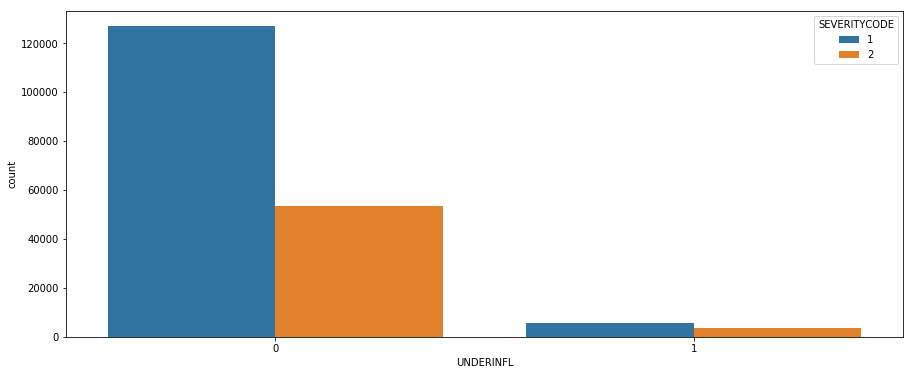


Fig7: UNDERINFL v/s severity of accident

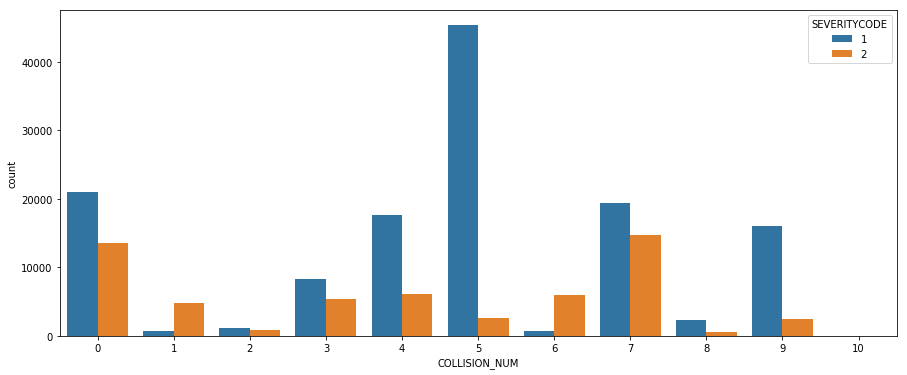


Fig8: COLLISION\_NUM v/s severity of accident

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.

After balancing the dataset, we standardize the data. After doing so we use K fold validation technique or ‘Train-Test-Split’ technique to split dataset according to test size. We train and test the data using classification techniques.

* 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 are evaluated using Precision, Recall, 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 testing set with the accuracy around 71% and K nearest neighbors was best when used upon test dataset with the accuracy around 70%.

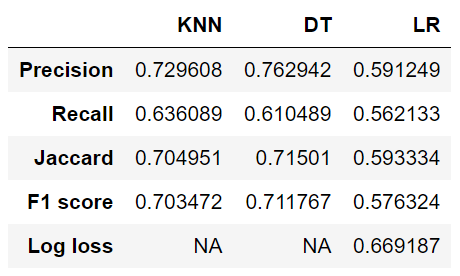


Table1: Test data accuracy report

1. **Conclusions**

Looking at the above table, ‘Table1’ we can conclude that Decision Tree is the most ideal technique to predict accident severity.

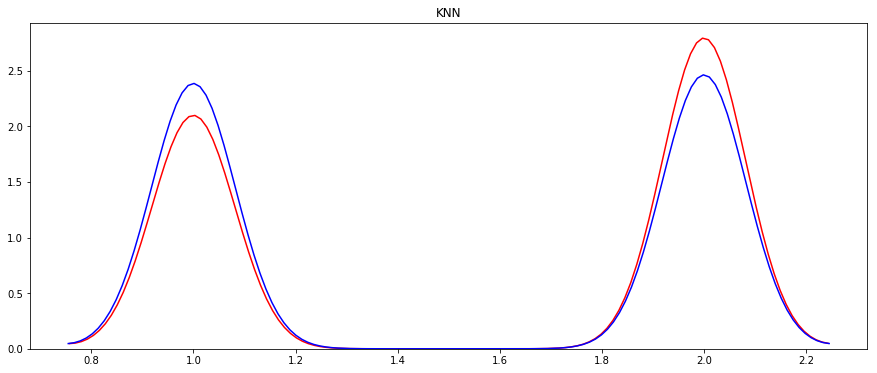


Fig9: Distribution plot for KNN.

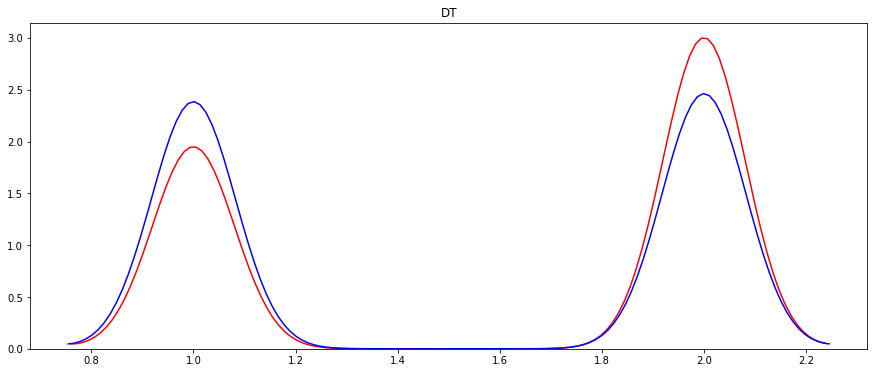


Fig10: Distribution plot for Decision Tree

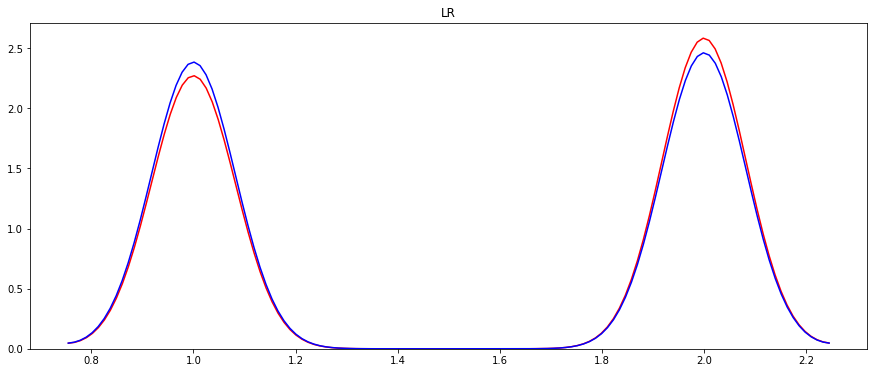


Fig11: Distribution plot for Logistic Regression

Legend:

Red line- Predicted Values

Blue line- Actual/True Values

Looking at the above distribution plots we can infer that Logistic Regression even though has lower accuracy and precision, fits the curve better than the rest of them. This is most likely because the target value is binary and evaluation parameter log loss is useful.

Hence for this model we consider **Logistic Regression** to be the best optimum technique rather than Decision Tree.