Capstone Project – Car Accident Severity

NC

Table of Contents:

1. Introduction

1.1 Background

According to WHO, every year the lives of approximately 1.35 Million people are cut short as a result of a road traffic accident. Between 20 and 50 Million more people suffer non-fatal injuries, with many incurring a disability as a result of their injury.

Road traffic injuries cause considerable economic losses to individuals, their families and to nations as a whole. These losses arise from the cost of treatment as well as lost productivity for those killed or disabled by their injuries, and for family members who need to take time off work or school to care for the injured. Road traffic accident caused most countries 3% of their gross domestic product (GDP).

In consideration of above description it is important to analyse the severity of accidents and their causes which will help in predicting severity of the actual accidents.

1.2 Business Problem

To predict the severity of accidents happening on the roads of Seattle based on multiple attributes like weather, road condition, light condition, vehicle count etc. These attributes will help in predicting the severity of the accidents.

1.3 Target Audience

The model generated in this project can be used by State Departments, drivers and cab companies to predict the severity of accidents that might happen given the conditions at hand and hence will allow them to be well prepared to handle such situations more effectively.

2. Data Acquisition and Cleaning

2.1 Data Source

Data used in this project was provided by Coursera course itself. It was provided as a CSV file. It can be downloaded from the link given below. The data was created by Seattle Department Of Transportation and Traffic Records Group. It contains the data from 2004 to Present.
The dataset is around 200,000 events.
https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv

2.2 Data Cleaning & Feature Selection

There are total 37 attributes in the chosen dataset including Severity as one of them. We can't use all the attributes to create a model in order to predict the severity of an accident. Hence, few relevant attributes had to be chosen. We had to choose the attributes which affects the accidents and has a potential to increase their severity. Hence they should be highly correlated (either positively or negatively) to the severity of the accident. The

following 4 attributes were chosen as they are the only ones who showed a negative correlation with the severity out of all the 37 attributes:

WEATHER: shows the weather condition
ROADCOND: shows the condition of the road
LIGHTCOND: shows the state of the natural light
VEHCOUNT: depicts the number of vehicles involved in the collision

The person correlation values for these 4 attributes:

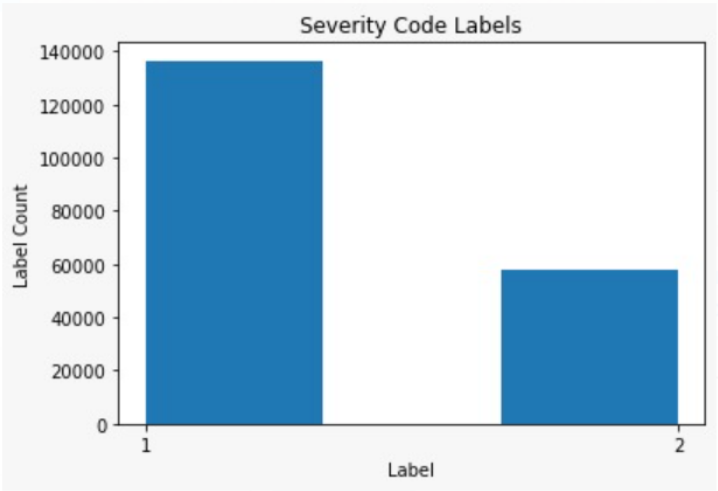
	SEVERITY
WEATHER	-0.1094
ROADCOND	-0.0493
LIGHTCOND	-0.0659
VEHCOUNT	-0.0546

Based on the above correlation values, these 4 attributes were chosen.

3. Methodology

3.1 Data Analysis

The dataset was highly imbalanced as visible from the figure below. It had much more rows for SEVERITY CODE 1 (approx. 70%) than the rows for SEVERITY CODE 2 (approx. 30%). Hence the total number of rows with SEVERITY CODE 1 had to be reduced to the same level as for SEVERITY CODE 2, by randomly choosing the rows to include in the balanced data set.



3.2 Data Preparation and Normalization

3 out of the 4 chosen attributes had values which were object type and not integer. So they needed to be converted into integer in order to make them usable for modelling later on. So I chose integer codes for each of the String value for each attribute. The rows where there were no values available, we added a new integer code under the label Unknown. The corresponding integer codes for each of the String values for each such attribute are given below:

WEATHER	
Dry	0
Ice	1
Oil	2
Other	3
Sand/Mud/Dirt	4
Snow/Slush	5
Standing Water	6
Unknown	7
Wet	8

ROADCOND	
Blowing Sand/Dirt	0
Clear	1
Fog/Smog/Smoke	2
Other	3
Overcast	4
Partly Cloudy	5
Raining	6
Severe Crosswind	7
Sleet/Hail/Freezing Rain	8
Snowing	9
Unknown	10

LIGHTCOND	
Dark - No Street Lights	0
Dark - Street Lights Off	1
Dark - Street Lights On	2
Dark - Unknown Lighting	3
Dawn	4
Daylight	5
Dusk	6
Other	7
Unknown	8

Now all the 4 chosen attributes have integer values. After this, the dataset was normalized transforming the values of all the attributes into a similar range. After normalization, the dataset was split into training and testing sets using a 75%-25% split. 75% of the dataset was randomly selected to create a training dataset and 25% of it was randomly selected to create a testing dataset.

The training dataset will be used to build and train the model. The testing dataset will be used to assess the performance of a predictive model.

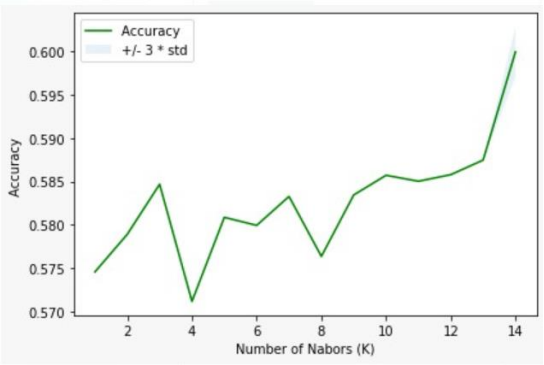
3.3 Modelling/ Classification

I used the below mentioned modelling techniques using the above created training dataset and evaluated the performance of each of them using the above created testing dataset:

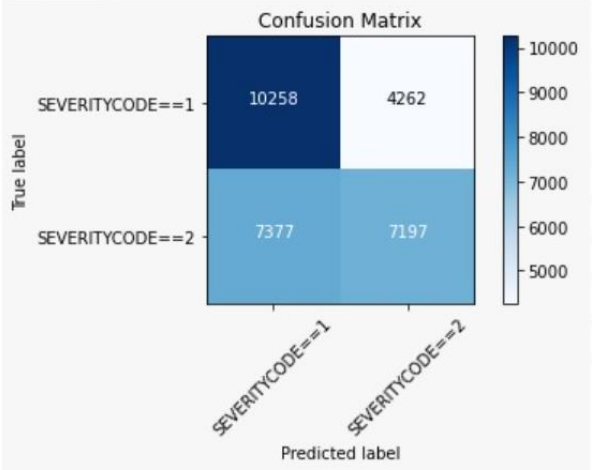
- K-Nearest Neighbours (KNN)
- Decision Tree (DT)
- Logistic Regression (LR)

K-Nearest Neighbour:

I started by choosing a value of Ks = 15. This value was chosen as this is a good mid value to avoid overfitting or overly generalized model. I evaluated the accuracy of the model at different values of K and found that K=14, provides the highest accuracy, as shown below:



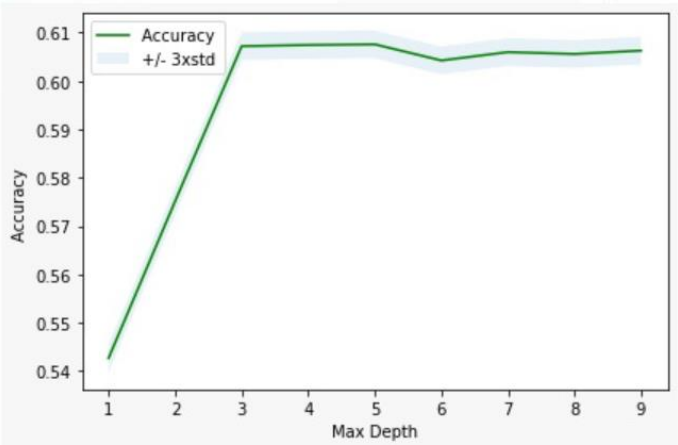
Then I calculated the accuracy of the model using the confusion matrix as well as the F1-score and Jaccard similarity score, as shown below:



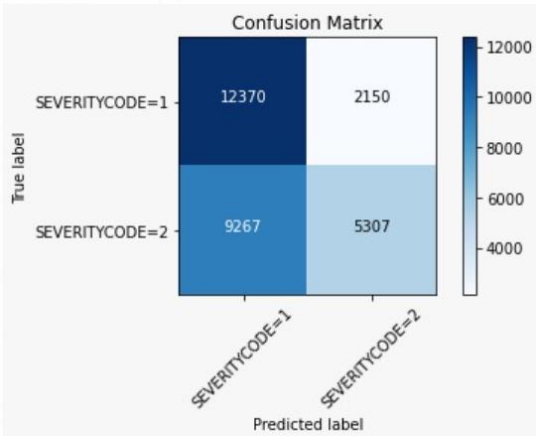
Train set Accuracy: 0.5995050525881626
Test set Accuracy: 0.599951880112738
The F1-score is 0.5953950705187826
The Jaccard similarity score is 0.4684659999086633

Decision Tree:

I started with a MAX DEPTH value of 10, to find out the right depth for our model. The depth with the highest accuracy came out to be 5, as shown below:



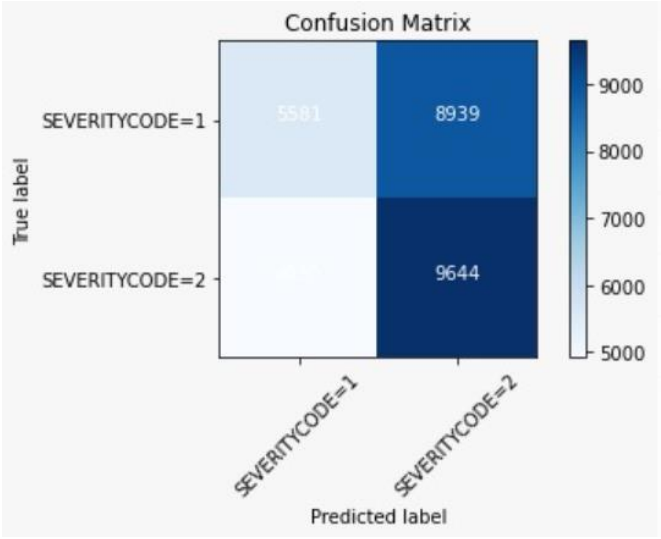
Then I calculated the accuracy of the model using the confusion matrix as well as the F1-score and Jaccard similarity score, as shown below:



Train set Accuracy: 0.6059439517884558
Test set Accuracy: 0.607582319378566
DecisionTrees's Accuracy: 0.607582319378566
The F1-score is 0.5828190410528927
The Jaccard similarity score is 0.5200319502249128

Logistic Regression:

The LR model was created using the training data set and the accuracy of the LR model was calculated using confusion matrix as well as F1-score, jaccard similarity score and Log Loss:



Train set Accuracy: 0.5257785110332027
Test set Accuracy: 0.5233037739740153
LR Accuracy: 0.5233037739740153
The F1-score is 0.5139481843676582
The Jaccard similarity score is 0.2869408740359897
The logLoss is: 0.6811191407676671

4. Result and Discussion

After studying & comparing the accuracy metrics of different models, the best model is the one which gives a combination of high F1-score, high jaccard similarity score and the smallest log loss. Based on this criteria's we can see that the best performance is observed in Decision Tree model:

	KNN	DT	LR
Accuracy	0.5999	0.6075	0.5233
F1-Score	0.5953	0.5828	0.5139
Jaccard similarity score	0.4684	0.5200	0.2869
Log Loss	-	-	0.6811

5. Conclusion

This decision tree can now be used to predict the severity of the car accidents that will happen in Seattle based on the known conditions like weather, lighting conditions, road condition and vehicle count.

The model generated in this project can be used by state department, drivers, cab companies, logistic companies to predict the severity of the accidents that might happen given the conditions at hand and hence will allow them to be well prepared to handle such situation more effectively.