12/12/2020

Predicting Severity and Causes of Road Accidents

IBM CAPSTONE PROJECT

Wijeratne, Suranga

1.0 Introduction

1.1 Background

Road accident are extremely common around the world. More server road accidents could end up with single or multiple causalities. According to the TAC (Transport Accident Commission) in Victoria, average of 250 loss their lives due to accident in every year. Also, these accidents lead to lots of property losses. There are many reasons for these accidents, but it would be great to be understand what the most common causes, in order to prevent them from happening. For this analysis, I am attempting to build a machine learning model to predict what are the common reasons for severe accidents in Victoria, Australia. Different features such as road conditions, weather, light conditions, number of passengers etc are fed into the model to see what are the most corelating feature for severe accidents. In future, depending on the feature status, model can be trained to predict and warn drivers a possibility of a crash.

1.2 Stakeholders

The reduction in severity of accidents can be beneficial to the public authorities to work towards improving those road factors and also car drivers themselves who may take precautions to reduce the severity of accidents.

2.0 Data Acquisition and Cleaning

2.1 Data Sources

The data set is taken from Data.Vic which is the place to discover and access Victorian government open data. Crash Stats data is provided to user by VicRoads for the purpose of supplying information about road crashes in Victoria for education purposes. This is an extensive data with over 14 years from 2006 to 2020 over 200,000 observation across Victoria. There are many files with different features. Before dive deep into the data, all the features were consolidated to one file.

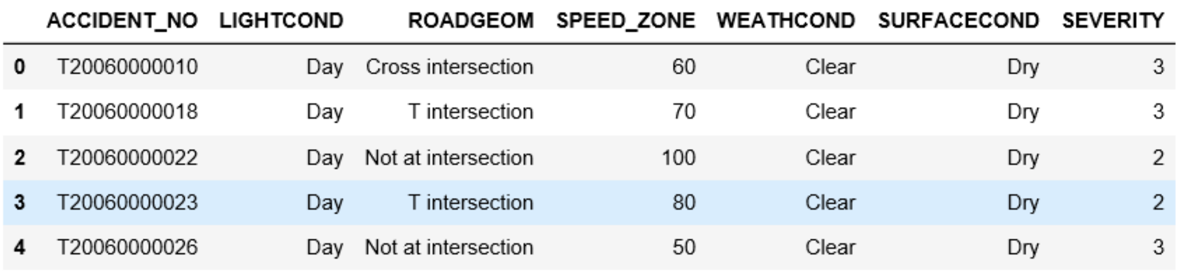
2.2 Data Cleaning

The data set consists of huge number of features which was identified as unrelated features for this exercise. Some of the categorical features are already encoded to numeric but there was no description provided. Out of all the features, below variables were chosen for the model.

1. Light Condition
2. Road Geometry
3. Speed Zone
4. Atmosphere Condition
5. Surface Condition

Once all above features including target variable “Severity” were taken to a new data frame, insight of data was carefully overlooked for further cleaning. There were unknow data in each column which were huge and deleting those data entirely would have led to a lot of loss of data which was not preferred. All “Unknown” and “Not Known” data were replaced with most common category rather than completely removing them from the data set. Also, all the column names were updated with meaningful names.

Feature Variable	Description
LIGHTCOND	Light conditions during the collision (Lights On/Dark)
ROADGEOM	Road geometry (Main road, intersection or by road)
WEATHCOND	Weather condition at the time of the accident (Rain, snow etc)
SURFACECOND	Surface condition of the road at the time of accident



	ACCIDENT_NO	LIGHTCOND	ROADGEOM	SPEED_ZONE	WEATHCOND	SURFACECOND	SEVERITY
0	T20060000010	Day	Cross intersection	60	Clear	Dry	3
1	T20060000018	Day	T intersection	70	Clear	Dry	3
2	T20060000022	Day	Not at intersection	100	Clear	Dry	2
3	T20060000023	Day	T intersection	80	Clear	Dry	2
4	T20060000026	Day	Not at intersection	50	Clear	Dry	3

2.3 Data Pre-Processing

All the features above selected are categorical variables which needed to encode to numerical. Some variables had many categories which were simplified to less numbers of categories. Ordinary encoding method used to convert categorical variables to numeric values.

LIGHTCOND	ROADGEOM	WEATHCOND	SURFACECOND	SEVERITY
0 = Light	0 = Main Road	0 = Clear	0 = Dry	0 = Fatal & Series
1 = Medium	1 = Intersection	1 = Overcast/Cloudy	1 = Mushy	1 = Non-Series
2 = Dark	2 = By Road/Priv	2 = Windy	2 = Wet	
		3 = Rain/Snow		

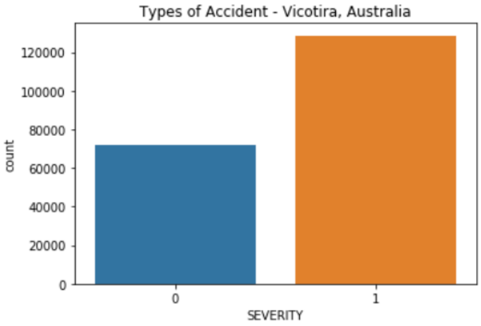
3.0 Methodology

3.1 Tools and Platforms

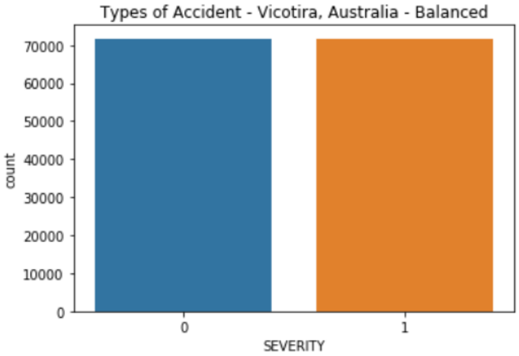
For implementing the Machine Learning solution, I have used Github as a repository and running Jupiter Notebook to pre-process data and build Machine Learning models. Regarding the coding, Python and its packages such as Pandas, NumPy, Sklearn were used.

3.2 Exploratory Data Analysis

It is observed that the target variable ("SERVERITY") in unbalanced. Majority incidents are Non-Serious/Fatal [0] and they are 75% higher than the Series incidents [1]. However this imbalance between the real-life occurrence of different accident outcomes may bias the model if not accounted for.



To create the least-biased model, re-sampling was carried out for the data set by selecting equal number of incidents for both cases in order to have an unbaiased classification model which is trained on equal instances of both the elements under severity of accidents.



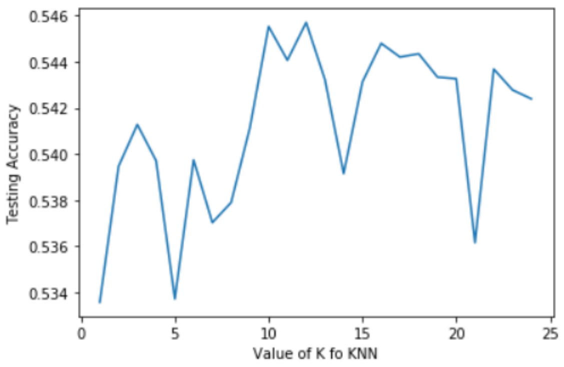
3.3 Machine Learning Model Selection

In order to develop a model for predicting accident severity, the re-sampled, cleaned dataset was split in to testing and training sub-samples (containing 20% and 80% of the samples, respectively) using the scikit learn package in python. Below machine learning models were used to evaluate the severity of accidents.

- K Nearest Neighbour (KNN)
- Decision Tree
- Logistic Regression

3.3.1 K Nearest Neighbour

K-Nearest neighbour classification was used from sklearn library to run the KNN machine learning classifier on the car accident severity data. The best K as shown in below for the model has the highest accuracy which was 12.



```
The best accuracy with  0.5457016412865456 with k =  12
```

3.3.2 Decision Tree

Decision Tree Classifier from the scikit-learn library was used to run the Decision Tree Classification model on the Car Accident Severity data. The criterion chosen for the classifier was ‘entropy’ and the max depth was ‘4’.

3.3.3 Logistic Regression

Logistic regression is a classification algorithm for categorical variables.

3.4 Machine Learning Model Evaluation

The final results of the model evaluations are summarize in the following table:

ML Model	Jaccard Score	F1 Score	Accuracy
K Nearest Neighbour	0.36655	0.54232	0.545701
Decision Tree	0.30725	0.53753	0.573743
Logistic Regression	0.33454	0.54255	0.546259

Based on the above evaluation results, KNN is the best model for Car Accident Severity prediction.

4.0 Conclusion

Based on the dataset provided for this capstone from weather, road surface, light conditions and road geometry pointing to certain classes, we can conclude that particular conditions have a somewhat impact on car accidents in Victoria. However, the model could have been performed more better if a few more things were presented and possible in data set.

- A balance dataset for the target variable
- Less unknown values within the dataset for all the variables
- More features such as drivers status at the time of driving (eg, after work, going for work, use phone etc)