

Final Project Report

Cyclist Collisions

Sammar Abbas

157-597-170

**Overview**

This business report focusses on evaluation of Cyclist collisions data. It’s a general perception that any accident of collision is due to either alcohol, bad road or weather condition and driver negligence. My task is to analysis dataset and come out with the findings on how we can prevent cyclist from fatal or non-fatal collisions. The report also tries to analyze impact of weather and light while speeding accidents took place. I adopted Logistic Regression Model and Decision Tree Model to analyze my data with the help of various attributes.

**Dataset**

For this evaluation, the dataset I have used is ‘Cyclists.csv. The data focusses on collisions that took place in Toronto district wise between year 2006-2020. There is total 57 attributes and 1779 rows.

Graphical user interface, text, application

Description automatically generated

**Reasons of Collision**

Collisions are not relying on any specific reason. We can say the combination of different factors creating the effect into result of collision. In the Snow or Heavy rain cases of collisions are extremely low. Even in bad road condition number of cases are minimal. This is drawing the conclusion when all the conditions are good, group of factors are causing collisions rather any single factor like as, bad road condition, bad light, Alcohol, or speeding.

A picture containing shape

Description automatically generated

**Statically calculations**

I have analyzed the dataset statistically. I started generating simple graph like as number of collisions recorded yearly, where it is concluded that year 2012 – 2013 are the years of high collisions where is year 2020 is the year with least number of cases recorded.

Chart, histogram

Description automatically generated

But on the other hand, we can say year 2020 recorded high number of fatal cases as well.

Chart, line chart

Description automatically generated

Our report also contained that ‘Automobiles’ are highly engaged in the collisions; however, other types of vehicles are least involved in the collisions. We can say for the motorcycles are only on road in Summer or Fall season due to which very few numbers of cases recorded with motorcycles. But heavy vehicles like as Truck or Transit vehicles and emergency vehicles are also at least number. The only automobile typed vehicles are highly engaged in collisions with cyclist.

A picture containing shape

Description automatically generated

While working and analyzing the dataset, seasonal time series plot showed that number of accident cases are seasonal. Months near to June are having high number of cases however number of collisions in other months are dropping down.

Chart, line chart

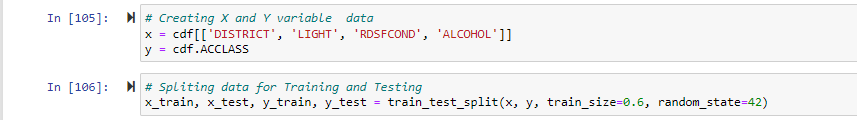
Description automatically generated

**Models:**

To carry on the process of analysis I used

1. Logistic Regression Model
2. Decision Tree Model

As my dataset containing high number of categorical values, I preferred to use above listed two models. Both models performed extremely well with the given values. For training purpose, I used 60% of data and remaining 40% of the data is used for testing purpose.



Logistic Regression model scored 0.96 and successfully passed the testing task.

Graphical user interface, text

Description automatically generated

Decision Tree model got the same score, this might be because both of the models were training with the same training dataset and also tested with the same testing dataset.

Graphical user interface, text, application, email

Description automatically generated

Prediction of both the models found exactly same. And because they both scored high and exactly same, I preferred to carry on with both models for Evaluation purpose.

Graphical user interface, text, application, email

Description automatically generated

**Evaluations:**

For evaluation purpose I used below listed techniques:

1. Confusion Matrix
2. MSE
3. RMSE

Both models exceptionally passed the evaluation process.

Below is the result of Logistic Regression Confusion Matrix

Graphical user interface, text, application, email

Description automatically generated

Here is the result of Decision Tree:

Graphical user interface, text, application, email

Description automatically generated

Decision Tree and Logistic Regression MSE and RMSE evaluation result

A picture containing text

Description automatically generated

**Conclusion**

From above analysis of cyclists’ collision dataset between 2006- 2020 indicates that there is not a single factor, that is acted as the major reason for collisions and fatal accidents. There can be hypnotized opinion or assumptions about different scenarios but based on the finding, different factors combining causing the major reason for collisions and fatal accidents.

YouTube link for video presentation:

<https://youtu.be/RZo2ciuevZo>