Car accident severity

Applied Data Science Capstone by IBM/Coursera

Project Report

Pan Yongqi

Table of Contents

**No table of contents entries found.**

# Introduction: Business Problem

With the transportation getting more and more convenient, more and more people choose to travel by car. However, traffic accidents are happening more and more frequently as well. Traffic accidents usually result in varying degrees of death and injury, so it is important for people to know the severity of traffic accidents and know how to protect ourselves.

Collisions he most common type of traffic accident, this project is about trying to build a model that could predict the severity of collision. In particular, this project will be targeted to stakeholders who will be driving at high frequency and the traffic accident rescue personnel.

# Data

Based on definition of our problem, factors that will influence our prediction are:

* PERSONCOUNT
* PEDCOUNT
* PEDCYLCOUNT
* VEHCOUNT
* ROADCOND
* WEATHER
* LIGHTCOND

Since some of the above data are related to the injuries and deathes of people and vehicles and some are related to the situation of the accident, I decide to choose them to determine the severity code.But PEDCOUNT and PEDCYLCOUNT are not always noted down and are less relative, so I decide to drop these two statistics.

So the final factors we choose will be:

* PERSONCOUNT
* VEHCOUNT
* ROADCOND
* WEATHER
* LIGHTCOND

# Methodology

Firstly, we do some data cleaning and preprocessing to make our data suitable for modeling. Some techniques may include replace and dropna.

We replace the string value of the ROADCOND, WEATHER, LIGHTCOND into integers to fit to the machine learning models.

We also normalize the dataset and split it into train set and test set for fitting the model input.

Secondly, we build the machine learning models. I choose the Logistic Regression for this car accident severity report. The Logistic Regression algorithm classifies data based on its input, although only a binary variable can be predicted.

# Results

This chapter contains the results of the accuracy for the Logistic Regression.

* Jaccard similarity score: 0.6987958170486954
* log loss: 0.5851352839636487
* classification report:Table

  Description automatically generated

# Discussion

This chapter describes the limitations of the report and the analysis and other steps that could have been taken.

The dataset which contains the data about accidents in Seattle which occured since 2004 is a CSV file which has to be loaded into a data frame and prepared for evaluation and analysis. A large number of columns is dropped before the analysis due to redundancy reasons or because they are out of scope. So the accuracy of this report may not be as high as expected.

In this report, I used Linear regression for classification. However, other Machine Learning techniques can also be used for this. Decision trees can be used since it can have string input and may have higher accuracy.

Besides, clustering can also be a new way to explore this dataset.

# Conclusion

Predicting accident severity is a complicated task with many different features to be taken into account. However, with modern technology and built-in computing functions, models can be built for accident severity prediction in a practical way.

Besides, our accuracy for this classification can achieve 70%, which represents the efficiency and feasibility when in similar condition.