IBM Capstone Project 2020

Predicting Car Accident Severity

Yassine Benider

September 29, 2020

Introduction

According to WHO, every year the lives of approximately 1.35 million people are cut short as a result of a road traffic crash. 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 crashes cost most countries 3% of their gross domestic product.

To help reduce deaths caused by car accidents, governements must think of innovative ways to prevent such problems. Luckily, we live in an era where computers have the power not only to do high computations, bt also help predict the likelyhood of certain outcomes with the help of Machine Learning.

This project will be a response to the car accidents problem and, hopefully, a first step to reduce road death with the help of machines given the current weather, road and visibility conditions. Our target audience is governements (in Seattle, and hopefully the rest of the world)

Data

The data we will use is the seatlle accident data given by the course.

The Data can be found in the following Cognitiveclass Data set Click here

Here are the target & features we will use to conduct our study:

the attributes that I'll use are :

SPEEDING : Whether or not speeding was a factor in the collision
WEATHER : weather at the time of crash
ROADCOND : road condition at the time of crash
LIGHTCOND : light conditions at the time of crash
INATTENTIONIND : whether the driver was distracted
UNDERINFL : whether the driver was under the influence of drugs or alcohol

And of course, the target values is :
SEVERITYCODE : corresponds to the severity of the collision

Methodology

0.1 Data Cleaning

The data cleaning is the process of giving a proper format to the data for its further analysis. The first step was to deal with missing values and outliers. Initially a lot of feature were dropped from the dataframe except those described above. Speeding was also droped since more than 80% of its values where NaN.

Now the columns in the dataframes are objects describing the nature of things. For example, ROADCOND has wet, dry ... to describe the road condition. For those features, I filled missing values with most frequent value of each column.

Next on, I transfromed the categorical data into integers using pandas get_dummies. I was all set to begin my machine learning modelisation.

0.2 Machine Learning Modeling

Different classification algorithms have been used and built for the prediction of the level of accident severity. These algorithms are a supervised learning approach. The best suited algorithm for this specific problem was compared using Accuracy, f1 Score and Jaccard Score.

Firstly, the 194672 rows where split 70/30 between the training and test sets. Since the data was already standardized by get dummies to all features, I moved right to modeling.

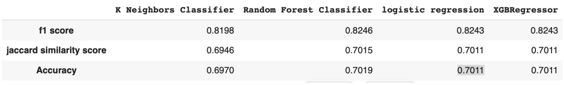
Four different approaches were used:

- Random Forest
- Logistic Regression
- K-Nearest Neighbour
- XGBoost

The same modus operandi was performed with each algorithm. Using the train set to train the models, and the test set to determine the accuracy for the development of the models were calculated.

Results

The metrics used to compare the accuracy of the models are the Jaccard Score, f1-score and Accuracy. This table reports the results of the evaluation of each model.



	K Neighbors Classifier	Random Forest Classifier	logistic regression	XGBRegressor
f1 score	0.8198	0.8246	0.8243	0.8243
jaccard similarity score	0.6946	0.7015	0.7011	0.7011
Accuracy	0.6970	0.7019	0.7011	0.7011

Figure 1: This table shows the end results for our models using the metrics described above).

We have succeded to make a model that predicts accidents severity with good accuracy and f1 score.

The logistic regression and XGboost models have similar accuracy f1 score, however the computational time from the regression is far better than the other two models. With no doubt both models are the best models.

Logistic regression is the best model to use !.

Discussion

I was able to obtain an accuracy of 70.1% in the logistic regression model and an f1 score of 82.4%. However, there was still a lot to be done to improve the accuracy of values predicted by the models in this study.

A lot of feature, such as speed and inattention had a lot of missing values. Wich is understandable given the difficulty to get such data. But I remain optimistic because such characteristics that are impossible to know at the moment, can be in the near future given the incredible rate in wich technology is evolving. Cars would soon be able to track them so that emergency services can use them.

Conclusions

In this study, I analyzed the relationship between severity of an accident and some features which describe the situation that involved the accident.

I built and compared 4 different classification models to predict whether an accident would have a high or low severity.

This study aims to reduce car accident numbers and also raise awareness by identifying the features that cause the most the gravity of an accident, these could be tackled easily by governements once discovered.

References

[1] https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries/