IBM DATA SCIENCE CAPSTONE PROJECT

CAR ACCIDENT SEVERITY

1. Introduction

The idea of this project is to analyze the severity of an accident in the United States of America. We're trying to engineer a model to predict the severity vehicle accidents throughout the States. Millions of people die every day due to accidents. This project could help solve this problem to a big extent.

Problem Statement: What is the magnitude of severity for an accident that occurs in USA?

Interest: Of course! Road accidents cab be prevented. The prediction aim for sustainable development, has set an ambitious target of halving the global number of deaths and injuries from road traffic crashes by 2021. Others, who are interested to reduce the accident impact, claims and to improve the Road safety such as Insurers, Organizations and Public Persons may also be interested.

2. Data

This is a countrywide car accident dataset, which covers 49 states of the USA. The dataset contains the driving conditions, number of people and vehicles involved in crash, and the severity of crash.

The data can be found at : https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv

Some entries were missing crucial data that were required. Some columns were filled with "Unknown" in the number of people injured. To rectify, I dropped the entire row as a mean of the crashes would not have been accurate.

3. Methodology

Exploratory Data Science

Examining the Environmental conditions:

First, I chose to examine the speeding conditions, road surface conditions, and weather conditions in which each accident occurred.

Speeding and Road Conditions:

- The proportion of L2 severity is higher when the driver speeds.
- The proportion of L2 severity is higher when the road condition is bad

This shows that Road Conditions and Speeding certainly have a big effect on accident severity.

4. Predictive Modeling

Machine Learning Models

Before model building, I have transformed all the categorical features using Label encoding method and normalized the datasets.

For modelling I have spilt the dataset into 80-20 ratio using train test split method. I.e. 80% as train set and 20% as test set.

Three classification algorithms were used and evaluated to predict the accident severity. Algorithms considered for classification were KNN algorithm, Decision Tree Classifier, Logistic Regression.

The first thing came to mind is that severity is based on different decisions like Road conditions and the speed. So, I tried ensemble methods as the data is very imbalanced.

KNN Test I

K nearest neighbours is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure. The final accuracy score cam out to be 0.696751

Decision Tree II

Decision tree builds regression or classification models in the form of a tree structure. It breaks down a dataset

into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. The final accuracy score came out to be 0.699679

Logistic Regression III

This model is basic and popular for solving classification problems. Logistics regression uses sigmoid function to deal with outliers. Class weight parameters sets the weights for imbalanced classes by adjusting weight inversely proportional to class frequency. The final accuracy score came out to be 0.699679

4. Results

	Algorithm	F1-score	Accuracy
0	KNN	0.591378	0.696751
1	Decision Tree	0.576051	0.699679
2	LogisticRegression	0.576051	0.699679

The expectation, going into this investigation, was that speeding while driving and in poor road conditions was dangerous, and that this would be evidenced by a pattern of higher crash rates has been supported by evidence. The coefficients have also come out to be positive which shows the imperative effect on the severity of accidents

	Intercept	Coef:SPEEDING	Coef:ROADCOND
0	-0.853729	0.067702	0.068295

Discussion of recommendation:

Road traffic injuries can be prevented. Governments need to take action to address road safety in a holistic manner. This requires involvement from multiple sectors such as transport, police, health, education, and actions that address the safety of roads, vehicles, and road users.

Effective interventions include designing safer infrastructure and incorporating road safety features into land-use and transport planning, improving the safety features of vehicles, improving post-crash care for victims of road crashes, setting and enforcing laws relating to key risks, and raising public awareness

5. Conclusion

The objectives of the investigation were met, but there are still many areas of the data which could be investigated further, such as how junction layout or vehicle type relate to collision rates in different conditions. This model provides empirical evidence against speeding and road conditions