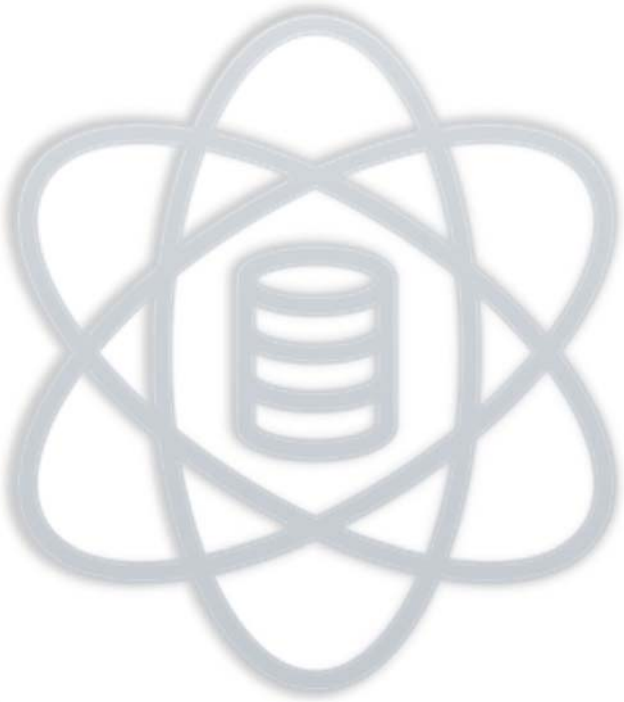
PREDICTING CAR ACCIDENT SEVERITY IN SEATTLE, WASHINGTON, US

A machine learning approach to the problem

By Kauvin Lucas

Sep 2020



IBM Developer

IBM Data Science Professional Certificate in Coursera

Introduction

Car road accidents are the number one problem in US road transportation, accounting for 99% of transportation injuries and representing an economic cost of 150 billion USD annually [1]. Deaths and injuries from the accidents are some of the issues concerned by insurance companies, healthcare institution and governments. It's very important for them to be able to predict the severity rate of an accident given some known variables. This project will attempt to employ classification models to predict severity rates and compare the accuracy of each built model.

Business problem

The seaport city of Seattle is one of the busiest cities in the Washington state, and cars accidents are very common in the place. The city administrators are making an effort to improve the road conditions, but serious injuries from the decreasing accidents rate are still a concern. The city has a goal of eliminating all traffic-related deaths by the year of 2030, and city officials are employing a data-driven approach to accomplish the goal.

Those who survive in a car accident can face hefty medical bills and thousand dollars in property damage. Discovering patterns in the data and making predictions may aid the decision-making processes of healthcare and insurance companies.

Analytic approach

Predicting car accidents severity is one of the biggest challenges faced by many actors. No single factor can help explain the severity of an accident, and the relationship between these factors, although intuitively positive, are mostly unclear from a statistical point of view.

But while each accident may be unique, accumulating insights from each accident may show macro trends and thus allow us to make accurate predictions. Taking these several seemly uncorrelated variables into a single accurate prediction model is a long process that requires finding generalizable predictive patterns. That can only be achieved by using machine learning algorithms.

Machine learning classification approaches and data science methodology will be employed to achieve the goal of this project. The following classification techniques will be used to predict and evaluate the model:

1. K Nearest Neighbors (KNN)
2. Decision Tree
3. Logistic Regression
4. Support Vector Machine (SVM)

Data collection

The data used was provided by the Washington State Department of Transportation (WSDOT) in a csv file. It's a public data that describes the accidents occurred between 2004 and 2020 in the city of Seattle. The information contained in the data goes through a month-long process that involves city and state transportation officials reviewing, comparing and analyzing reports from the Seattle Police Department.

Data understanding

The data had 37 features and 194. We are going to predict the severity code of an accident, which are labeled in the dataset as numbers between 0 and 3, from least to more severe.

Before the Feature Selection stage, two more features were added: "INCHOUR" and "ISHOLIDAY". The selected features were the following:

Feature name	Description	Comments
SEVERITYCODE	A code that corresponds to the severity of the collision:	This will be the output variable
ADDRTYPE	A description of the collision address type (intersection/alley/block)	Crash type may affect collision severity
INATTENTIONIND	Whether the person was not paying attention	Inattention and distraction may affect collision severity
UNDERINFL	Whether the person was driving under the influence of alcohol	Alcohol or drug impaired drivers may cause more severe collisions
WEATHER	A description of the weather conditions during the time of the collision	Poor weather conditions may affect driver visibility, and lead to more severe accidents
ROADCOND	The condition of the road during the collision	Poor road conditions may affect collision severity
LIGHTCOND	The light conditions during the collision	Poor road light conditions may increase accident frequency and fatality
SPEEDING	Whether or not speeding was a factor in the collision	Speeding vehicles may represent greater risk of huge material damage and bigger number of injuries
INCHOUR	Hour of the day when the collision has occurred	Newly created dimension. Severity of the accident may change depending on the hour of the day
ISHOLIDAY	Whether or not collision has occurred in holiday	It's known that holidays have an impact in the accident frequency and severity numbers

Data preparation