Robin Kjär
21.10.2020
github.com/RobinKj/Coursera_Capstone

# PREDICTING TRAFFIC ACCIDENT SEVERITY IN SEATTLE, WA

IBM Data Science Professional Certificate
Capstone Project

# Table of Content

# 1. Introduction

## 1.1 Scope and Background

Road accidents are a growing issue worldwide, with millions of deaths and injuries every year all over the world. Road traffic injuries result in 1.35 million deaths on roadways each year globally. This effects people across all age groups with more than half of the victims being pedestrians, motorcyclists and cyclists. [1] This issue weighs even more heavily within the U.S. where people rely more on their cars, because of the limited public transport options.

Seattle is the **25th biggest city in the U.S.** and the biggest city in Washington State with almost 3,8 million people living in its metropolitan area. Its inhabitants like to sometimes call it rain city, as the Pacific Northwest is not particularly famous for its fair weather.

2019 the Seattle Times cite a Texas A&M report saying Seattle ranks **seventh** in the whole United States **in time stuck in traffic**. The report estimates that there were 167,384,000 hours of delay resulting from traffic in 2017. As the saying goes - time is money – meaning these hours also result in a total annual cost of $3.1 billion or $1,408 for each single commuter.[2]

The Washington State Department of Transportation (WSDOT) counted 6,252,554 registered vehicles in Washington State in 2015, with 59.7 billion vehicle miles being travelled within the state. There annual collision report emphasizes the incredible numbers of **a crash every 4.5 minutes and a fatal crash every 16 hours**.[3]

Sadly, these numbers don't seem to have been decreasing since.

## 1.2 The Objective

Which brings us to our objective: What if we could actively reduce the number of traffic accidents or their severity if we better understand the underlining factors that affect the severity of accidents.

We want to identify **which variables have the biggest impact on the severity of accidents**. Through analysis of the broad range of variables, from road and weather conditions to location data, we can train models to **accurately predict accident severity**.

The insight on which specific conditions can determine the severity can help improve vehicular and road safety, as well as potentially leading to early warning systems for traffic.

[1] https://www.cdc.gov/injury/features/global-road-safety/
[2] https://www.seattletimes.com/seattle-news/transportation/seattle-area-traffic-congestion-is-among-the-worst-in-the-country-study-shows/
[3] https://www.wsdot.wa.gov/mapsdata/crash/pdf/2015_Annual_Collision_Summary.pdf

These findings gained through the research on Seattle data can then be used to study other cities within the U.S. and worldwide.

## 1.3 Interest and Stakeholders

Learning more about what causes severe crashes should interest everyone, as it leads to making our roads safer.

Governments could use the findings to improve road safeness by counteracting the conditions affecting the severity of accidents. Installing early warning systems could inform drivers of dangerous conditions via radio, roadside signs or similar, leading to more careful driving. The police, ambulances and hospitals could be on alert near places where specific conditions are met.

Navigation apps leading through traffic could be enhanced by reacting to the underlying conditions found in our data analysis to avoid traffic and potential accidents. Even car manufactures and insurance companies could benefit by furthering their autonomous driving assistants and tailoring their products to the new information.

In general, the findings could save lives, time and money and therefore should concern everyone.

# 2. The Data

## 2.1 The Source

The data we used to study car accident severity in Seattle was provided through the course materials on Coursera[4], but can also be found on the open data platform of the city of Seattle[5].

It lists almost 200,000 samples of collisions between 2004 and today. The data comes from traffic records from the Seattle Department of Transportation (SDOT), containing all vehicular accidents between cars, pedestrians, cyclists and so forth.

It features the target variable *SEVERITYCODE* and 37 attributes describing all kinds of characteristics. From location data, type of surrounding of the accident, in which form the vehicles collided, to information about the state of the driver causing the accident and lighting, weather and road conditions.

[4] https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv
[5] https://data-seattlecitygis.opendata.arcgis.com/datasets/

The target variable has values between 1 and 2. The higher value is equivalent to a severe accident including injuries, while the value 0 equals an accident with only property damages.

We must take note of only 58,188 of the 194,673 samples being listed with a severity of 2, leading to the dataset being unbalanced.

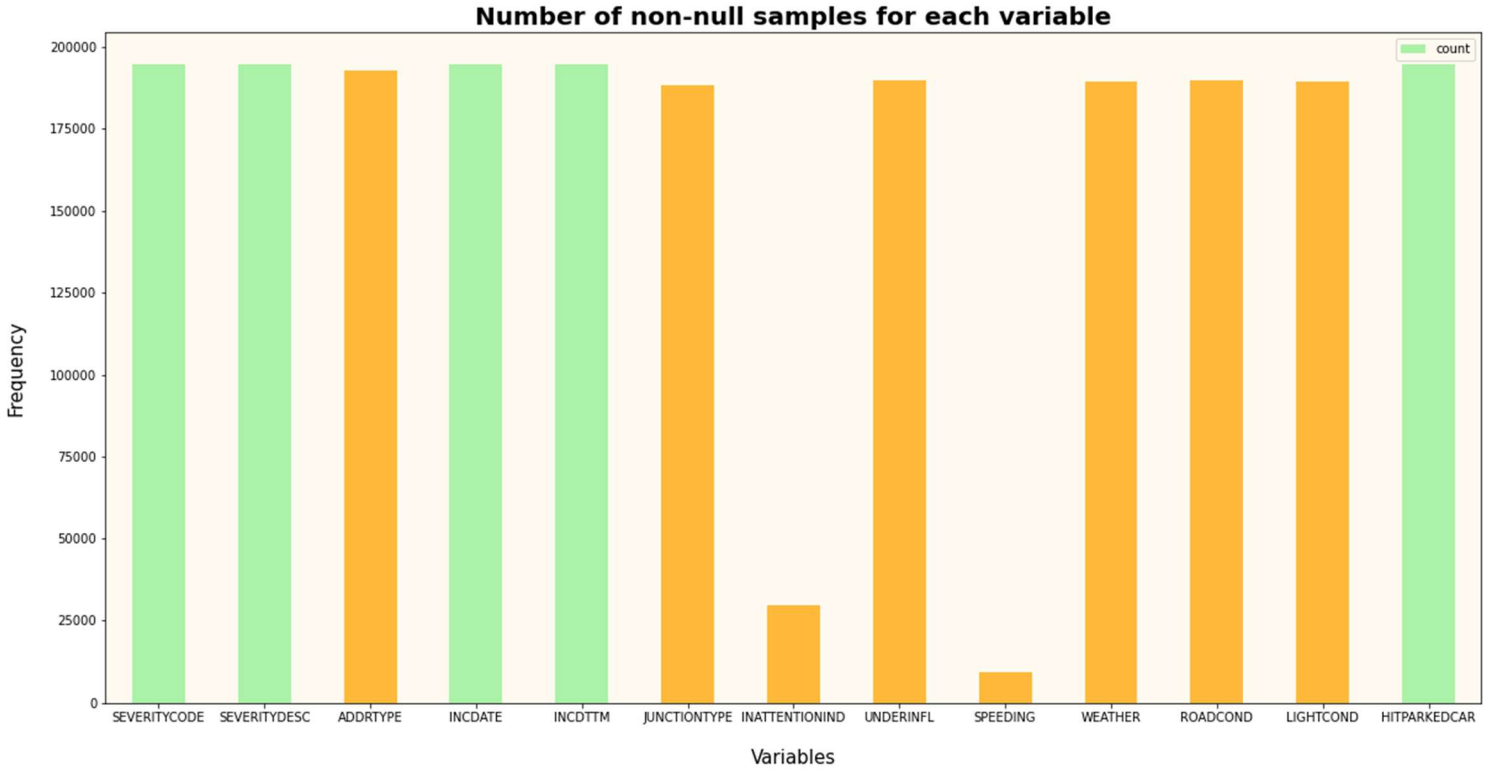
## 2.2 Data Cleaning

Our data has several problems regarding missing data as well as including a lot of unnecessary information for our specific purposes. We will replace missing values with the value of the highest frequency and eliminate some columns. Only then is the data suited for the machine learning algorithms.

The goal of our analysis and machine learning model is the prediction of the accident severity. This column is our target variable and for the sake of the machine learning algorithms we will transform the values from 1/2 to 0/1.

We also must react to the problem of an interchanging format in some columns. Those columns include values with the format Y/N, as well as values with the format 1/0. They contain the exact same information, but we must align the information to the same format. For easier computational practice, we chose the 0/1 format.

The biggest problem within the dataset is the big amount of missing data. The following graphic shows which of the columns we decided to work with are affected.



The green colored bars have zero missing data, while the orange bars are missing information. The variables about if the driver was speeding or paying attention are lacking a considerable amount of values, but we will optimize all variables listed in orange.

We replace missing values with the most common one, achieving a dataset with zero information left out. We also reduced the number of different categories in the categorical variables, by combing some categories to one. This makes it easier to plot information and get a clean overview of the dataset.

## 2.3 Feature Selection

As the dataset includes too much different variables, we decided on removing 25 columns that were deemed unnecessary for our work. This allows us to focus on the important factors in our models and graphical analysis. To utilize our planned machine learning algorithms, we need to reduce the number of variables again and do a feature selection. The features used in our machine learning models are therefore marked in orange.

| Variables | Description |
|---|---|
| **SEVERITYCODE** | The severity of the accident (1/0) |
| **SEVERITYDESC** | The severity of the accident (as a category) |
| **ADDRTYPE** | At what kind of location the accident took place (as a category) |
| **INCDATE** | Date of the accident |
| **INCDTTM** | Date of the accident |
| **JUNCTIONTYPE** | At what kind of junction the accident took place (as a category) |
| **INATTENTIONIND** | Whether or not the driver was paying attention (1/0) |
| **UNDERINFL** | Whether or not the driver was under the influence (1/0) |
| **SPEEDING** | Whether or not the driver was speeding (1/0) |
| **WEATHER** | Weather conditions during the accident |
| **ROADCOND** | Road conditions during the accident |
| **LIGHTCOND** | Light conditions during the accident |
| **HITPARKEDCAR** | Whether or not the accident involved a parked car (1/0) |