**Predicting the severity of car accidents**

Daria Siniatulova

September 30, 2020

# Introduction

## Background

According to World Health Organization, every year the lives of approximately 1.35 million people are cut short as a result of a road traffic crash. More than half of all road traffic deaths are among vulnerable road users: pedestrians, cyclists, and motorcyclists. 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. Road traffic injuries are the leading cause of death for children and young adults aged 5-29 years. Males are more likely to be involved in road traffic crashes than females. About 73% of all road traffic deaths occur among young males under the age of 25 years who are almost 3 times as likely to be killed in a road traffic crash as young females.

The information on road traffic crash cases can be used to understand what factors contribute to different types of accidents, predict their severity category and help to prevent them or reduce their damage.

## Problem

The goal of this project is to build several machine learning models to be able to predict severity of consequences for motor vehicles collisions in Seattle using the dataset of previously registered cases and their attributes. This project is based on City of Seattle data on traffic crash, also called a motor vehicle collisions, or car accidents. Motor vehicle collisions occurs when a vehicle collides with another vehicle, pedestrian or cyclist, animal or stationary obstruction.

Project dataset is hosted by the City of Seattle at the open data platform (<https://data.seattle.gov>), the dataset’s attributes description is published at <https://www.seattle.gov/Documents/Departments/SDOT/GIS/Collisions_OD.pdf>.

Different attributes of the car accidents will be used to build machine learning models and predict severity of accidents.

## Interest

Road traffic injuries can be prevented. Therefore, the possibility to accurately predict whether and how much people risk with their property, health or life while using roads (as car drivers, motorcyclists, passengers, cyclists, or pedestrians) can be used to make important decisions at many levels.

The results of this project could be of interest to governmental regulators and municipal public service officials (for example, in the decision-making process on roads and bridges construction planning), health insurance organizations and medical organizations that work with traffic accidents trauma cases, various businesses (banks, car insurance companies, car dealers, car parts suppliers, repair shops etc.), as well as mass public, media and individual drivers and passengers (for example, on additional precaution measures in particular parts of roads, or types of junctions, or during specific weather).

# Data acquisition and cleaning

## Data sources

The dataset is hosted by the City of Seattle at the open data platform. The CSV file about collisions can be obtained via link <http://data-seattlecitygis.opendata.arcgis.com/datasets/5b5c745e0f1f48e7a53acec63a0022ab_0.csv>

Description of different attributes of the collision cases will be used to build machine learning models and predict severity of accidents.

The file of that description (Collisions\_OD.pdf) is available via link <https://www.seattle.gov/Documents/Departments/SDOT/GIS/Collisions_OD.pdf>.

The dataset includes all types of collisions.

The dataset timeframe: 2004 to Present.

Update Cycle: Weekly.

*Note: I decided to replace the dataset suggested by course staff that was published at* [*https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv*](https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv) *to have complete data in terms of dates of accidents (the latest dates available), detailed severity codes (not only 1 and 2, but also 2b and 3), list of columns (as INJURIES, SERIOUSINJURIES and FATALITIES columns are missing in the suggested dataset), as I consider this attributes and details to be important for the given assignment.*

## Data cleaning

At this stage I import libraries, collect the dataset from CSV file and then proceed to determine the attributes (columns) that should be used to train machine learning models.

Exploring the data also implied assessing the condition of chosen attributes by looking for trends, patterns, skewed information, correlations, missing data and errors. Collisions have geospatial coordinates at the intersection or mid-block of a segment. I plot the map of Seattle to check if all of the dots are within the city borders (thhhaaat is corrects).

The dataset is imbalanced because number of majority cases is 2.2 times bigger than all injuries and fatalities number combined. The "Unknown" severity of accidents (21656 rows) should be dropped at the Data preparation stage.   
I decided not to perform under-sampling to reduce the number of majority cases (property damage), if my model would show decent results; and it did, so I skipped this step.

There are a lot of missing values in the dataset, because of lack of record keeping. I used different tactics to deal with this:

* if the number of missing values is not very high, I drop all the data for a particular observations (*dropna* method for axis=0 and particular subset);
* if the variable is not a very important attribute (predictor) for the target variable, it could be dropped completely; so when the data goes missing on 60-70 percent of the variable, dropping the variable (*drop* method for axis=1) should be considered;
* if the variable is important attribute (predictor) for the target variable, and such variable could be fixed using the conditions in another columns, I used masks with different conditions to fill out some missing values (*fillna* method for axis=0).

I used to\_datetime method to load the correct format of collision date (*Timestamp* instead of *string*), and then get the Year and Month from it as separate columns. Accidents grouped by year showed that in the first 7 years (2004 to 2011) number of accidents were gradually decreasing from ~15000 to ~ 12000 per year; average yearly number of collisions were higher than in the following years (2012-2019). In 2015 and 2016 there was a growth in number of accidents, but it was not that significant; in 2020 due to COVID-19 pandemic restrictions, traffic accidents number expected to be significantly lower than in 2019. Accidents grouped by month have nor revealed any significant seasonal influence.

After getting rid of missing values, I have corrected type *object* or *float* to type *integer*: for the columns that previously contained them (because NaN is considered as text data) for the following columns:

* ADDRTYPE,
* SEVERITYCODE,
* COLLISIONTYPE,
* JUNCTIONTYPE,
* SDOT\_COLCODE,
* UNDERINFL,
* WEATHER,
* ROADCOND,
* LIGHTCOND,
* HITPARKEDCAR.

Accidents grouped by address type demonstrate:

* accidents around the block happen 1.95 times more often than accidents at the intersection;
* accidents at the alley are extremely rare.

Accidents grouped by collision types demonstrate:

* 80.5% are incidents between cars, including:
* 24.9% involved parked car;
* 55.6% are collisions between moving cars.
* Only 3.9% involved pedestrians.
* Only 3.0% involved cyclists.

Accidents grouped by weather demonstrate:

* 64% of traffic accidents happen in clear weather, 19% - in rainy weather and about 16% - in overcast conditions;
* other weather conditions rarely lead to traffic accidents, probably because of both weather events happen not so often in Seattle or when they do happen the traffic is reduced, or the drivers are cautious due to difficult weather conditions.

Accidents grouped by road conditions demonstrate:

* 71.4% of accidents happen when the road is dry;
* 27.1% of accidents happen when the road is wet.

Accidents grouped by light conditions show that majority of cases happen during the daylight, and accidents during dark time of the day with street lights on are at the second place though are 3 times less frequent.

Accident that happened due to inattention cover only 30188 cases, due to driving under influence of drugs or alcohol – 9629 cases, due to pedestrian right of way was not granted – 5195, due to speeding – 9936, and involved hitting of parked car – 12089 cases.

In conclusion, violations like driver's inattention, speeding, DUI, not granting right of way to a pedestrian or hitting parked car are not necessarily correlated to number or severity of accidents. This could probably be related to limitations of the dataset, as police does not have all the information at the moment of accident registration. Besides, the registration of a collision is possible even without police officers present at the place of accident, as drivers can report them via mobile application. They would not even have to wait the police to arrive and get all the details of the accident for the records; for example, in cases “property damage only” they can simply exchange their insurance companies’ contacts and move on.

## Feature selection

As of September 30, 2020 the original dataset contains 221738 rows and 40 columns, and the columns' names are:

'X', 'Y', 'OBJECTID', 'INCKEY', 'COLDETKEY', 'REPORTNO', 'STATUS', 'ADDRTYPE', 'INTKEY', 'LOCATION', 'EXCEPTRSNCODE', 'EXCEPTRSNDESC', 'SEVERITYCODE', 'SEVERITYDESC', 'COLLISIONTYPE', 'PERSONCOUNT', 'PEDCOUNT', 'PEDCYLCOUNT', 'VEHCOUNT', 'INJURIES', 'SERIOUSINJURIES', 'FATALITIES', 'INCDATE', 'INCDTTM', 'JUNCTIONTYPE', 'SDOT\_COLCODE', 'SDOT\_COLDESC', 'INATTENTIONIND', 'UNDERINFL', 'WEATHER', 'ROADCOND', 'LIGHTCOND', 'PEDROWNOTGRNT', 'SDOTCOLNUM', 'SPEEDING', 'ST\_COLCODE', 'ST\_COLDESC', 'SEGLANEKEY', 'CROSSWALKKEY', 'HITPARKEDCAR'.

I can group these columns as follows:

* the columns 'X', 'Y', 'ADDRTYPE', 'LOCATION', 'JUNCTIONTYPE', contain attributes on the place of accidents, where 'Y' represents latitude and 'X' represents longitude;
* the columns 'INCDATE' and 'INCDTTM' contain data on time of the accidents;
* the columns 'OBJECTID', 'INCKEY', 'COLDETKEY', 'REPORTNO', 'STATUS', 'INTKEY', 'EXCEPTRSNCODE', 'EXCEPTRSNDESC', 'SDOTCOLNUM', 'SEGLANEKEY' and 'CROSSWALKKEY' contain some non-essential technical data like IDs and codes of reports of the dataset.
* the columns 'SDOT\_COLCODE', 'SDOT\_COLDESC', 'ST\_COLCODE' and 'ST\_COLDESC' code and describe the accidents themselves, like what type of participants were involved in the accident or how the collision between the participants happened;
* the columns 'INATTENTIONIND', 'UNDERINFL', 'PEDROWNOTGRNT', 'SPEEDING' and 'HITPARKEDCAR' contain the data on violations of rules that vehicle's driver was responsible for, like driving under influence of drugs or alcohol, or speeding;
* the columns 'WEATHER', 'ROADCOND', 'LIGHTCOND' describe weather, road and light conditions when the accidents happened;
* the columns 'SEVERITYCODE', 'SEVERITYDESC', 'COLLISIONTYPE', 'PERSONCOUNT', 'PEDCOUNT', 'PEDCYLCOUNT', 'VEHCOUNT', 'INJURIES', 'SERIOUSINJURIES', 'FATALITIES' contain the most valuable data piece on the consequences of these accidents, and the 'SEVERITYCODE' is the labeled data that I will need to predict using my machine learning model.

Based on definition of our problem and available data, main factors that influence my prediction models would be:

* number of registered accidents in Seattle;
* number of people and vehicles involved in the accidents (as majority of cases does not involve pedestrians nor cyclists);
* weather, road and light conditions (preferably changed from discrete to continuous values);
* severity codes of the accidents (preferably changed from discrete to continuous values).

Upon examining the meaning of each feature, it was clear that there was some redundancy in the features. For example, there was a feature of the severity code (as text strings), and another feature of the severity description (as text strings). These two features contained very similar information (consequences of the accident), with the difference being that the former feature could be easily transformed into numerical value, while the latter feature was simply a comment on that value, duplicating its meaning. In order to fix this, I decided to keep all features that were (or should be encoded to) numerical values that are important for the future machine learning models, and drop redundant ones.

Features (columns) 'INCDATE' and 'INCDTTM' related to the date and time of the accidents were dropped after getting months and years from them.

Features (columns) 'OBJECTID', 'INCKEY', 'COLDETKEY', 'REPORTNO', 'STATUS', 'INTKEY', 'EXCEPTRSNCODE', 'EXCEPTRSNDESC', 'SDOTCOLNUM', 'SEGLANEKEY', 'CROSSWALKKEY', 'SEVERITYDESC' were dropped completely as non-essential.

After data cleaning, there were 169,089 samples and 22 features in the dataset.

To train the models I used 18 attributes for 135,271 observations of accidents severity (Train set size: 135271, 18).

To test the models I used 18 attributes for 33,818 observations of accidents severity (Test set size: 33818, 18)