1 Introduction Section

In this section we discuss the business problem and who would be interested in this project.

1.1.Introduction/Business Problem Car Accident Severity Prediction

1.1.1.Introduction/Background

This section defines the business problem wth risks in the mobility area. It is about to solve the decision problem, to drive or not to drive with a car under certain known conditions from the current location to a destination at a planned time in realation to the risk for having an accident.

1.1.2.Problem

In the following, this report is to predict the severity of an accident. The scenario could be described as follows: Say, you are driving to another city for work or to visit some friends. It is rainy and windy, and on the way, you come across a terrible traffic jam on the other side of the highway. Long lines of cars barely moving. As you keep driving, police car start appearing from afar shutting down the highway. Oh, it is an accident and there's a helicopter transporting the ones involved in the crash to the nearest hospital. They must be in critical condition for all of this to be happening. Now, wouldn't it be great if there is something in place that could warn you, given the weather and the road conditions about the possibility of you getting into a car accident and how severe it would be, so that you would drive more carefully or even change your travel if you are able to?

1.1.3.Interests

In the following this is handled as a data science problems that targets the car drivers audience in the first place, but is also meant to help all involved stakeholders to mitigate risk in the mobility, insurance, healthcare area, and at least for the family of the driver.
This is exactly what will be handled in this report: to predict the severity of a possible car accident on the base of available car accident data from the past and current driving conditions.

1.2.Introduction to the Car Accident Severity Data

1.2. Data acquisition and cleaning

Data
In this section the data that will be used to solve the problem is described. It contains the explanation, why the is date is adequate for the problem and is used. In the discussion part, examples of data is provided.

1.2.1 Data sources

To deal with accidents data the shared data for Seattle city is used as an example:
https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv
Metadata descriptions:
https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Metadata.pdf

1.2.2 Data cleaning

After open the CSV file and check what type of data is contained. The first column colored in
yellow is the labeled data. The remaining columns have different types of attributes. A selection of these attributes are used to train the model. Also most of the observations found are of sufficent quality to train and test the machine learning model to be build in this case.
The label for the data set used is severity, which describes the
fatality of an accident. It shall be notice that the shared data has unbalanced labels. So the data is balanced, otherwise, a biased ML model would be created.

1.2.3 Feature selection

The following is a list of attributes or features that are used. For description of each attribute, it is refered to the web link on the CSV file.
SEVERITYCODE, STATUS,
ADDRTYPE
• Alley
• Block
• Intersection
INTKEY,
with a collision
LOCATION,
SEVERITYCODE,
● 3—fatality
● 2b—serious injury ● 2—injury
● 1—prop damage
● 0—unknown
PERSONCOUNT,
- Collision address type:
- Key that corresponds to the intersection associated
- Description of the general location of the collision
- A code that corresponds to the severity of the collision:
- The total number of people involved in the collision
PEDCOUNT,
This is entered by the state.

PEDCYLCOUNT, entered by the state.
VEHCOUNT,
entered by the state.
INCDATE, INCDTTM, JUNCTIONTYPE, INATTENTIONIND,
- The number of pedestrians involved in the collision.
- The number of bicycles involved in the collision. This is - The number of vehicles involved in the collision. This is
- The date of the incident.
- The date and time of the incident.
- Category of junction at which collision took place
- Whether or not collision was due to inattention. (Y/N)
UNDERINFL, - Whether or not a driver involved was under the influence of drugs or alcohol.
WEATHER,
of the collision.
ROADCOND, LIGHTCOND,
PEDROWNOTGRNT, granted. (Y/N)
SPEEDING, (Y/N)
ST_COLCODE,
collision. For more information about these codes, please see the State Collision Code Dictionary. Codes: 0-5, 10-32, 40-57, 60-67, 71-74, 81-84.
SEGLANEKEY, occurred.
CROSSWALKKEY,
HITPARKEDCAR car. (Y/N)
- A key for the lane segment in which the collision
- A key for the crosswalk at which the collision occurred. - Whether or not the collision involved hitting a parked
- A description of the weather conditions during the time
- The condition of the road during the collision. - The light conditions during the collision.
- Whether or not the pedestrian right of way was not
- Whether or not speeding was a factor in the collision.
- A code provided by the state that describes the
In addition, there is probably need to do some feature engineering to improve the predictability
of the model as follows:
From the ST_COLCODE, a smaller set of cathegories could be defined. From INCDATE the day of the week could be calculated.
From INCDTTM the part of the day could be calculated: morning, afternoon, evening, night.
The target or label columns should be accident " severity" in terms of human fatality, traffic delay, property damage, or any other type of accident bad impact. These terms are cathegories and are constructed from the last attributes listed above.
Then, the built severity data set is applied building a machine learning model.