1 Introduction

1.1 Background

A traffic collision, also called a motor vehicle collision, car accident, or car crash, occurs when a vehicle collides with another vehicle, pedestrian, animal, road debris, or other stationary obstruction, such as a tree, pole or building. Traffic collisions often result in injury, disability, death, and property damage as well as financial costs to both society and the individuals involved.[i]

Traffic collision affect the national economy as the cost of road injuries are estimated to account for 1.0% to 2.0% of the gross national product (GNP) of every country each year.
In 2013, 54 million people worldwide sustained injuries from traffic collisions.[1] This resulted in 1.4 million deaths in 2013, up from 1.1 million deaths in 1990 [ii] . About 68,000 of these occurred in children less than five years old.[iii]

1.2 Problem

Data about traffic collisions such as location, light condition, road condition, etc. is regularly collected by law enforcement agencies for statistical analysis purpose.

This project aim is to analyze historical data about car accidents in order to determine factors which mostly impact on accident severity. I will be focusing on light conditions, road condition and location type (i.e. intersection, block,..) features to figure out if those factors play a role in accident severity and which actions might be taken to reduce impact.

2 Data

The data was collected by the Seattle Police Department and share by Coursera for this work.
Dataset is publicly available at http://data-seattlecitygis.opendata.arcgis.com/datasets/5b5c745e0f1f48e7a53acec63a0022ab_0

It includes 221.144 accident records in the state of Seattle, from 2004 to the date it was issued, in which 37 attributes or variables are recorded and a codification of the type of accident is assigned among 84 available codes.

Records include relevant information such as accident severity, road condition, light condition, and address type (i.e. intersection, block, alley):

- Severity code: a code that corresponds to the severity of the collision:
    - 3—fatality
    - 2b—serious injury
    - 2—injury
    - 1—prop damage
    - 0—unknown

The dataset records show severity code values 1 or 2 only. No other value is reported.

- Road condition: the condition of the road during the collision
    - i.e. 'Dry', 'Wet', 'Unknown','Ice','Snow/Slush','Other','Standing Water','Sand/Mud/Dirt','Oil'
- Light condition: the light conditions during the collision.
    - 'Daylight', 'Dark - Street Lights On', 'Dusk', 'Dawn', 'Dark - No Street Lights', 'Dark - Street Lights Off', 'Dark - Unknown Lighting', 'Other', 'Unknown'
- Address type: Collision address type.
- 'Alley', 'Block', 'Intersection'

This project goal is to determine if address type, road and light conditions, can impact on accident severity.

2.1 Data cleaning

Data contains record with 'NaN' values, therefore I first dropped any record with 'NaN' value in any of aforementioned columns.

Furthermore I noticed dataset was imbalanced, we had the following occurrences for severity code values 1 and 2 respectively:

Severity code 1: 136485
Severity code 2: 58188

I obtained a balanced dataset by randomly under-sampling records with severity code to 1, until I got the dataset with the same amount of records for each adopted severity code:

Severity code 1 58188
Severity code 2 58188

---

[i] https://en.wikipedia.org/wiki/Traffic_collision

[ii] Global Burden of Disease Study 2013, Collaborators (22 August 2015). "Global, regional, and national incidence, prevalence, and years lived with disability for 301 acute and chronic diseases and injuries in 188 countries, 1990-2013: a systematic analysis for the Global Burden of Disease Study 2013". Lancet. 386 (9995): 743–800. doi:10.1016/s0140-6736(15)60692-4. PMC 4561509. PMID 26063472.

[iii] GBD 2013 Mortality and Causes of Death, Collaborators (17 December 2014). "Global, regional, and national age-sex specific all-cause and cause-specific mortality for 240 causes of death, 1990-2013: a systematic analysis for the Global Burden of Disease Study 2013". Lancet. 385 (9963): 117–71. doi:10.1016/S0140-6736(14)61682-2. PMC 4340604. PMID 25530442.