**Data science – Capstone project**

**Business problem**

Car crashes is the highest cause for accidental deaths and injuries. The Severity of a car accident might be influenced by a number of parameters. The ability to predict the severity could help to prevent future car crashes or reduce their severity.

**Data section**

The label is the incident severity (1 or 2) which are two categories. The rest of the variables are the independent variables. Since there are two categories of severity, I will consider this problem as a classification problem. The independent variables include both numerical and categorical variables. I will use the variable “ADDRTYPE” to investigate the effect of the crash location (block vs intersection) on severity. I found that if a crash is conducted on a block there is 75% its severity is 1 whereas in an intersection it is 56%. I investigated the variable “Location” and found that there are locations with higher frequencies for crashes. Furthermore, there are locations that result in higher change for severe crash (e.g. “10TH AVE E AND E JOHN ST”) and locations that result in higher change for non-severe crash (e.g. “10TH AVE BETWEEN E PIKE ST AND E PINE ST”). From the “severity desc” column, I understand that the difference between the severity levels are “property damage” vs. “injury”. Analysis of the “Collision type” variable shows that different types result in different severities. For example, 90% of “cycle” accident result in severe crashes, while only 10% of “packed car” crashes results on injuries. I will investigate the effect of the day of the week in which the crash happened. I will use the variable “Incdate” and create a new variable “weekday” that captures the day of the week of each crash. I investigated the effect of “junction type” on crash severity and found that some type include high probability for severe crush like “At intersection” while other not like “Ramp junction”. For the value “UNDERINFL” I noticed that the values are N,Y,0,1. I processed the data by changing the 0 to N and the 1 to Y.

I will investigate the number of missing data for each variable and delete variables with more than 40% of missing data. I will also delete the column “severityCode.1” Since it is identical to the label (y) column. I will balance the data using under balancing method. I will standardize the data using the “standardScalar” method.