# Coursera

# IBM Applied Data Science Capstone Project

# Seattle Accident Classifier

Hossein Faghih

# Introduction

There are almost 200 million cars in United States that are being used every day. City of Seattle, as a small sample, has reported nearly 200 thousand reports of car accidents during 2005-2020. An accident can lead to property damage, injury, and loss of life. If we can predict the chances of accidents in different situations, we may be able to stop them at their origin or create other safety measures to reduce the number and severity of crashes happening every day.

For this aim, we are trying to use the accident log of the Seattle police department for the period of 2005-2020 to create a model that predicts the chance of a severe accident in different weather and road conditions. We will also get insights into how several human factors will affect crashes which might lead to policies to limit those incidents.

# Data

We have 194673 reported accidents with 38 columns of descriptive, administrative, and contextual data. As we go through the data, assuming that the context of an accident have a significant role on it, we try to choose the elements of the model.

We have data on weather, road and lighting conditions, type of the address, whether a parked car was hit and several human factors, e.g. whether the situation happened because of being under influence, speeding, inattention.

# Methodology

To do this analysis, we separated the contextual features, consistent with our goal and prepared and cleansed the data.

1. Weather

Encoding Weather Conditions(0 = Clear, 1 = Cloudy, 2 = Windy, 3 = Rain and Snow

1. Road condition

Encoding Road Conditions(0 = Dry, 1 = Dirty, 2 = Wet, 3 = Ice)

1. Light condition

Encoding Light Conditions(0 = Light, 1 = Medium, 2 = Dark)

1. Speeding

Encoding Speeding (0 = No, 1 = Yes)

1. Under influence

Encoding Under the influence (0 = No, 1 = Yes)

1. Inattention

Encoding in attention (0 = No, 1 = Yes)

Then, we substituted the unknown information with the median of that column.

# Results

We built two models based on our data.

1. Linear Regression Model
2. Decision Tree model

Comparing the results, Decision Tree is slightly working better and is more explainable due to the nature of the model and the problem we had at hand. Details can be seen in the notebook.

We tried to create a Support Vector model as well, however, the computations seemed to be outside the available computing power.

# Conclusion

It is astonishing that each element of the model is very weakly correlated with the outcome of the model, however, the model is producing an accuracy of 70% for the results.