Predicting Collision Severity in Seattle

IBM Data Science Professional Certificate
Theo Hayeck

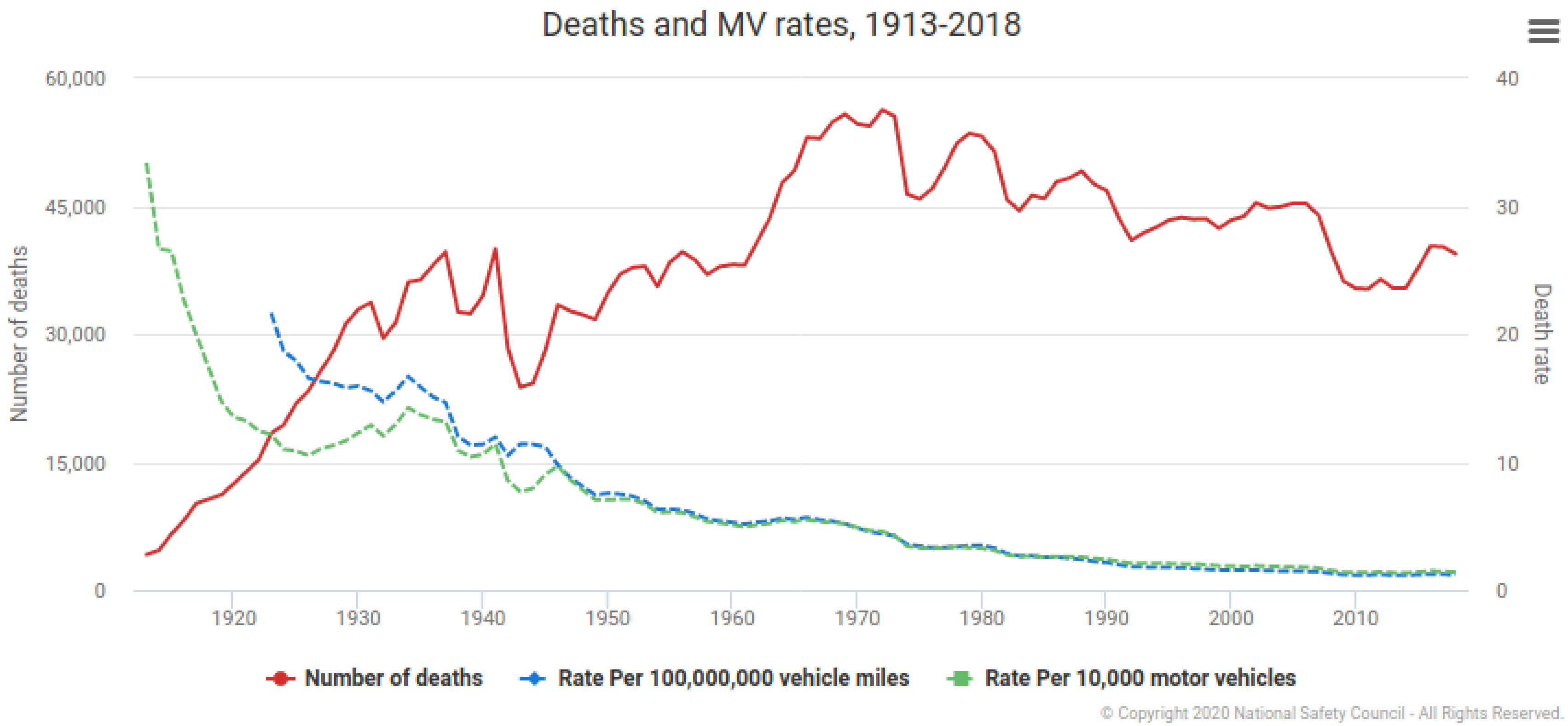
Introduction

Problem:

- **Collisions cause damage, injury or death.**
- **Since invention of the motor vehicle, there have been many improvements in safety features including:**
  - Seat Belts
  - Speed Limits
  - Airbags
  - Manufacturing Regulations
- **US Collisions at an all time low.**
- **However, approximately 5000 pedestrians are still killed each year as a result of a vehicle collision.**

Introduction

- **Decline in accident severity since mid 1920’s**



Introduction

Goal:

- **Create a Machine Learning Model that can:**
  - Predict the severity class of any given collision
  - Identify causes of severe collisions
  - Help reduce the severity and the total number of collisions.
- **Deployment Options:**
  - Electronic warning signs
  - Models built into car navigation systems
  - Mobile Application
  - Future: feed model into automated AI driven cars.

Data Acquisition and Cleaning

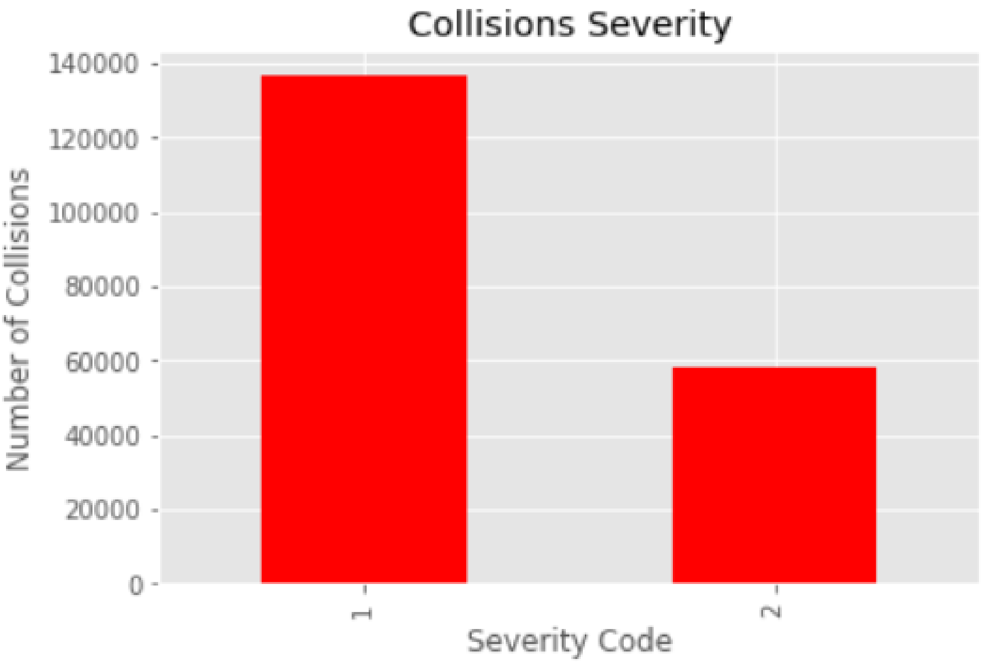
- **Data obtained from the Seattle Department of Transport (SDOT) Management Division from 2004 to present.**
- **Dataset available at:** **https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv**
- **In total there was 194,673 incidents recorded by the SDOT, with 38 columns corresponding to various features about each collision in the raw dataset.**
- **Duplicate, highly similar, or highly correlated features were dropped.**
- **Missing data was replaced by the mode for categorical features, and the mean value for continuous features.**
- **Cleaned data contains 12 features, including 1 feature label.**

Data Acquisition and Cleaning

Feature	Description
SEVERITYCODE	Code that corresponds to severity of collision
X	The longitudinal coordinate of the collision
Y	The latitudinal coordinate of the collision
JUNCTIONTYPE	The category of junction at which the collision took place
WEATHER	Weather conditions during the collision
ROADCOND	Road conditions during the collision
LIGHTCOND	Light conditions during the collision
INDTTME	Date and Time of the collision
INATTENTIONIND	Whether the driver was paying attention at the time of the collision
UNDERINFL	Whether the driver was under the influence of drugs or alcohol
SPEEDING	Whether the driver was speeding leading up to the collision
ST_COLCODE	A code issued by the state describing the collision

Using SEVERITYCODE as collision severity predictor

- **SEVERITYCODE comprises of 2 categorical codes.**
  - 1: Property damage only
  - 2: Collision resulting in injury or death.



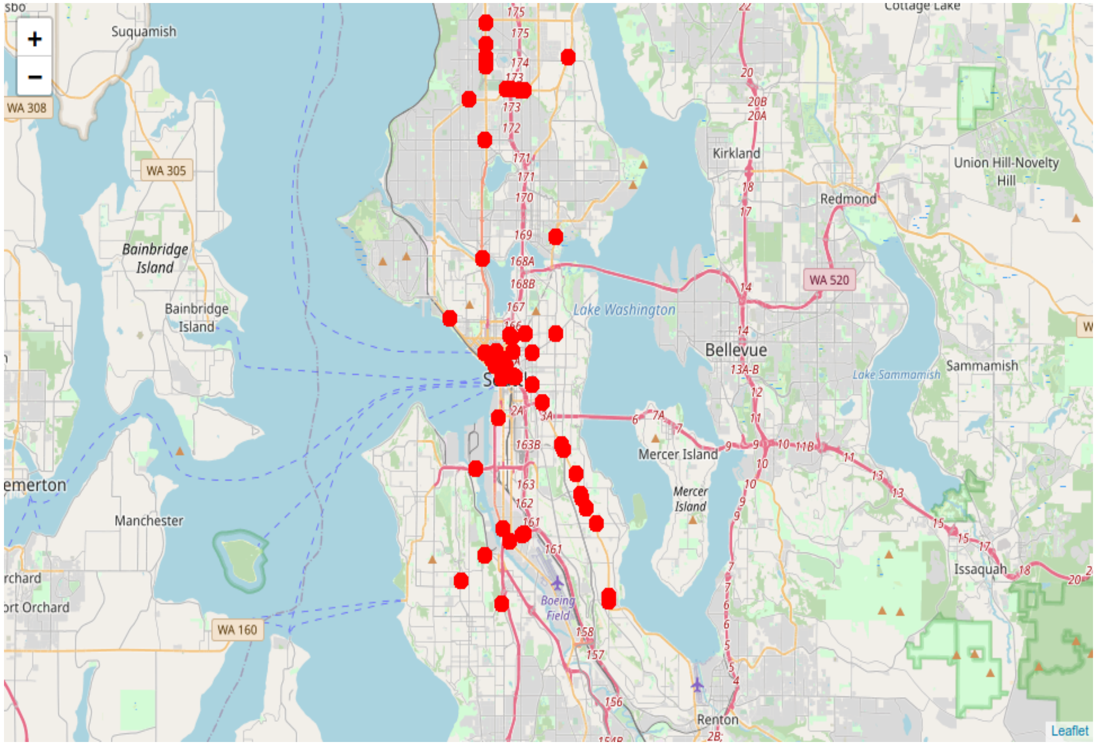
Exploratory Data Analysis

- **There are hotspots where a large amounts of collisions occur.**

Each red spot represents a location in Seattle where atleast 100 collisions (of class 1 and 2) have occurred.

Where ever there has been atleast 100 collisions, they have all had atleast 1, Class 2, collision occur.

These hotspots are particularly dangerous

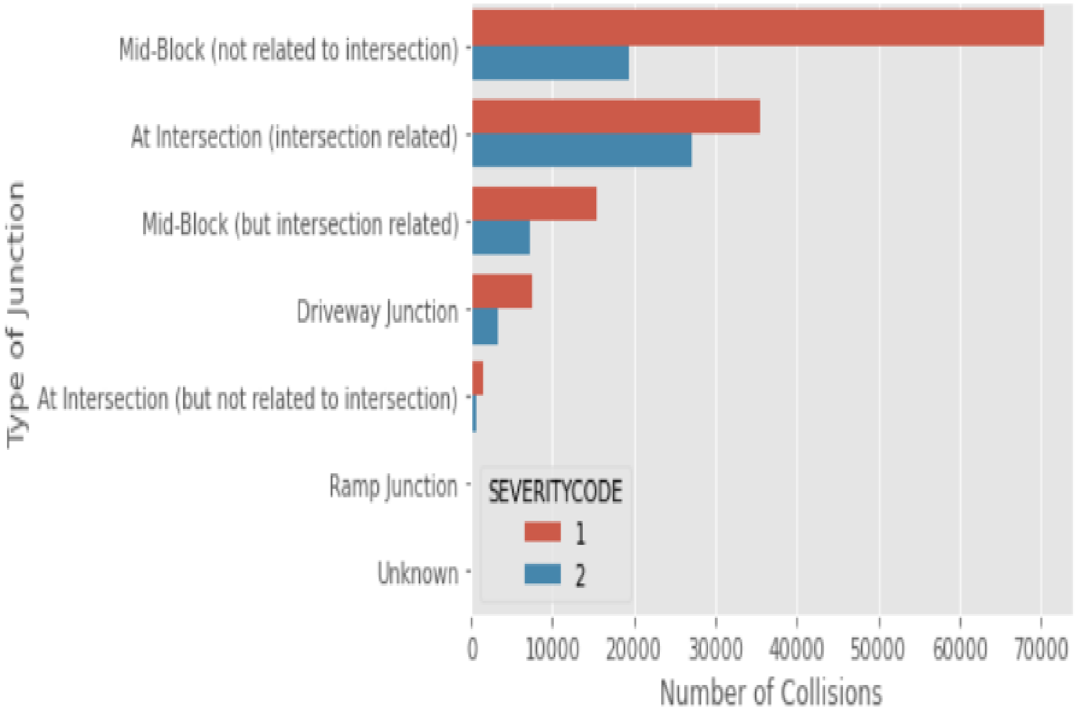


Exploratory Data Analysis

- **Mid-block (not related to intersection) and At Intersection (Intersection related) result in a higher ratio of class 2 collisions than other junction types.**

Although the greatest number of collisions occur at Mid-Blocks (not related to intersections), it also has the highest class 2:1 ratio, meaning that if a collision was to happen at this junction type, it will result in the highest chance of the collision being of class 2 than at any other junction. This is also true for At Intersection (intersection related).

These junction types are particularly dangerous

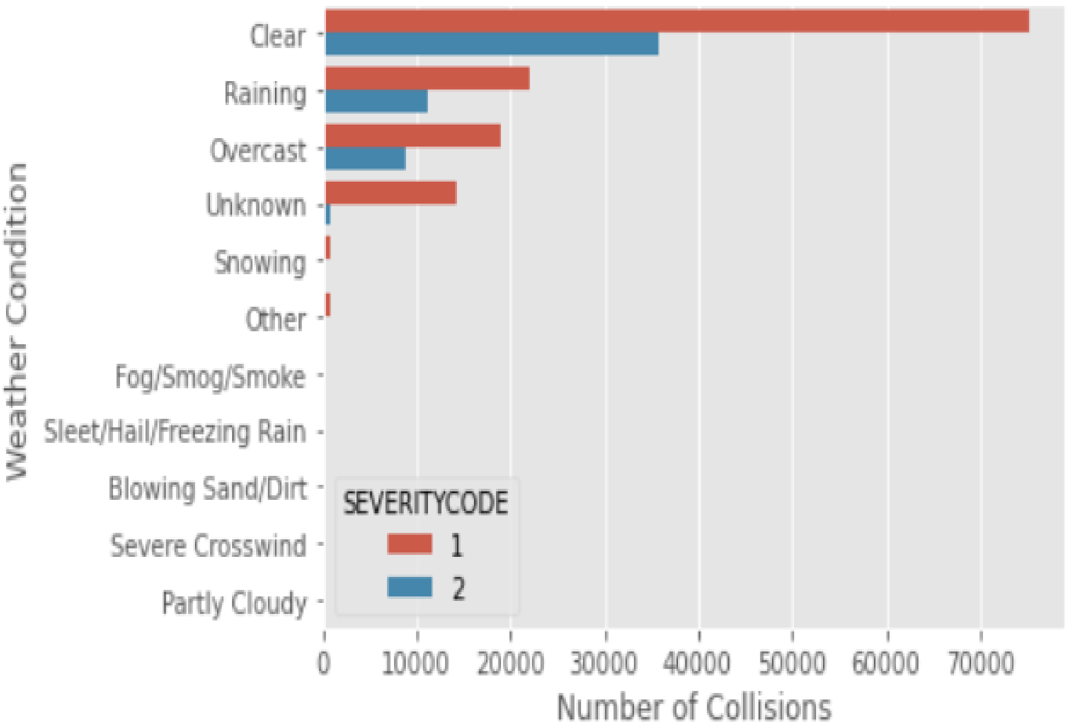


Exploratory Data Analysis

- **Rain and Fog/Smog/Smoke result in a higher ratio of class 2 collisions than other weather conditions**

Although the greatest number of collisions occur during clear conditions, Raining has the highest class 2:1 ratio, meaning that a higher proportion of class 2 collisions occur here than any other types of weather conditions. Fog/Smog/Smoke has the second highest class 2:1 ratio.

These weather conditions are particularly dangerous

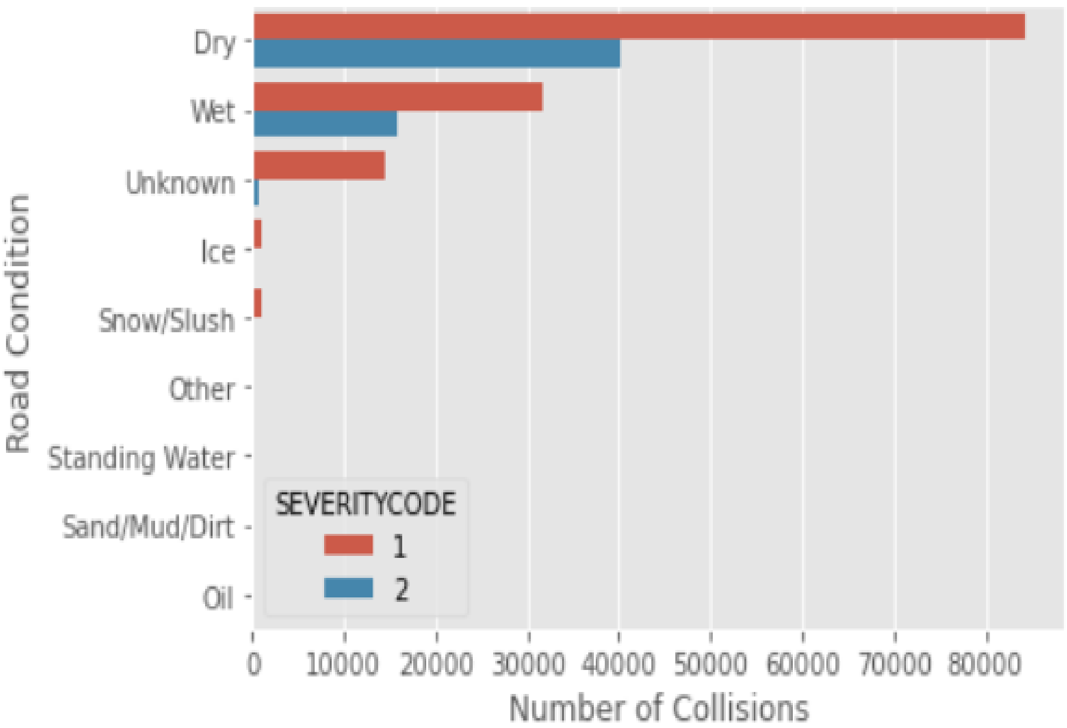


Exploratory Data Analysis

- **Oil on the road and wet result in a higher ratio of class 2 collisions than other road conditions.**

Oil on the road results in the greatest chance the collision being of class 2 should one occur out of all the road conditions, with 40 resulting collisions being of class 1 and 24 being of class 2, the largest class ratio out of all road conditions at 0.6000. This is followed by wet road conditions having a ratio of 0.4967

These road conditions are particularly dangerous

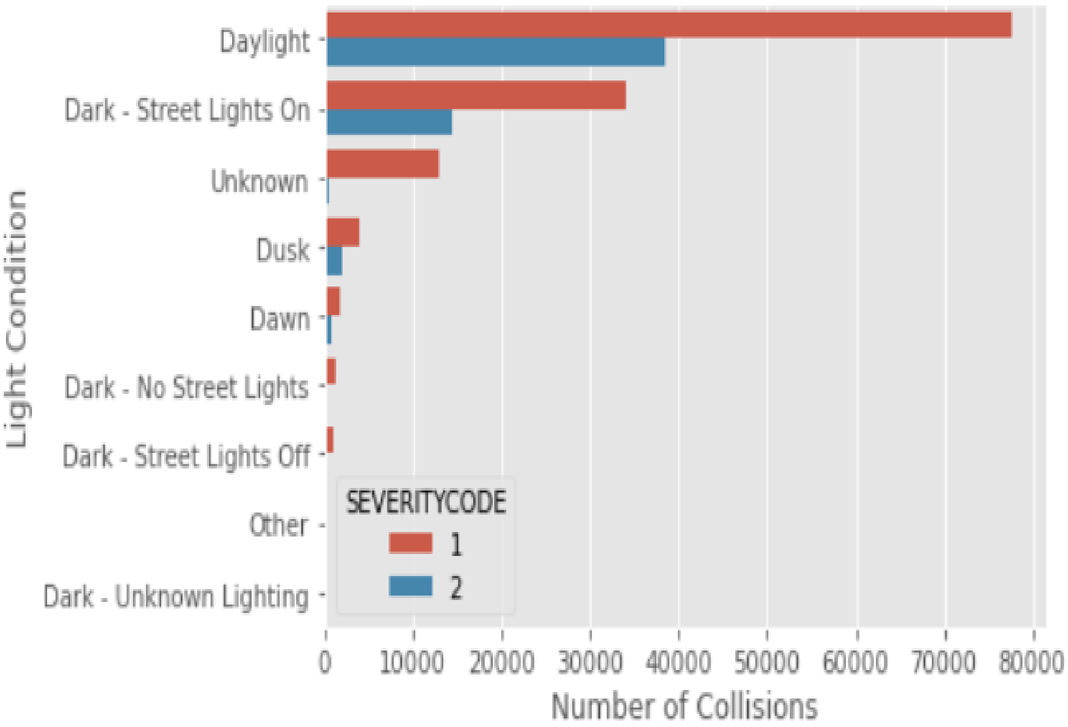


Exploratory Data Analysis

- **Dusk results in the higher ratio of class 2 collisions than other light conditions.**

Dusk results in the greatest chance of the collision being of class 2 if one was to occur out of all the light conditions, with a ratio of 0.491. It is to be noted that Dark Unknown did have a class 1:2 collision ratio of 0.571. However, due to it being unknown, it is less reliable to use

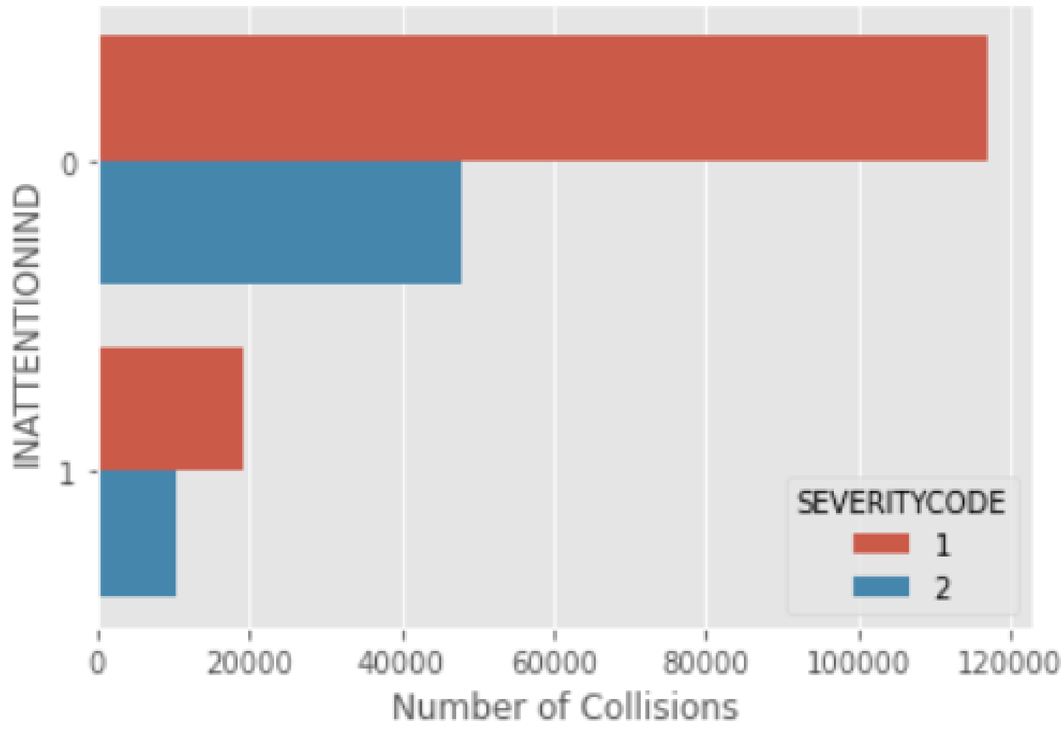
These light conditions are particularly dangerous



Exploratory Data Analysis

- **Drivers who were not paying attention have a higher chance of their collisions being of class 2.**

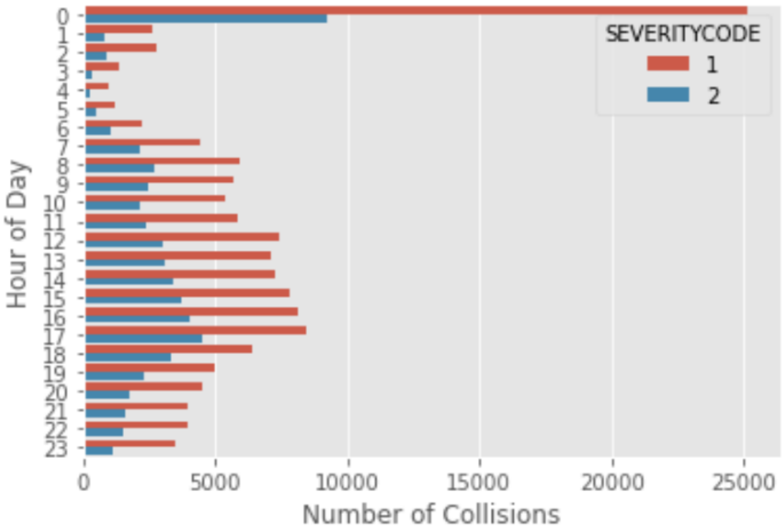
Even though more collisions occur when the driver was paying attention, as you can see from the chart, for all drivers who were **not** paying attention during the time of their collision, the greater chance of their collision being of class 2.

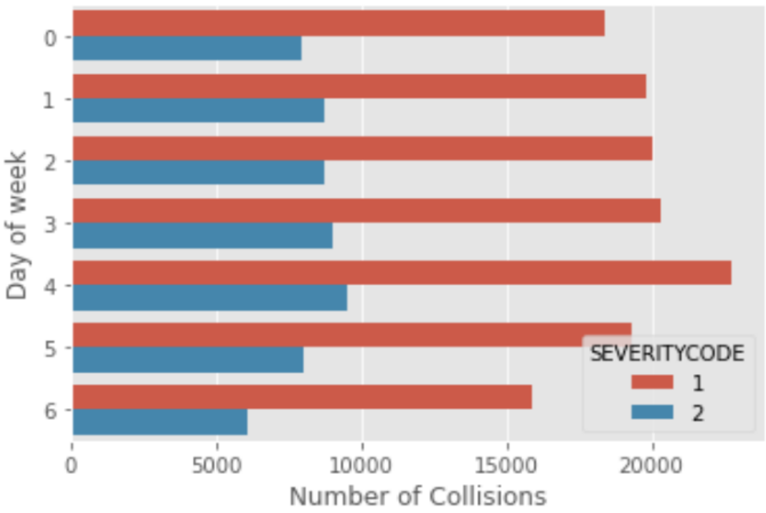


Innatention results in a greater chance of the collision class being of type 2 should one occur.

Exploratory Data Analysis

- **More collisions occur at around 17:00, with Thursday having the most collisions in every week.**





As shown above, you can see a bell shaped curve centered around the 17:00 hour of the day, resulting in the most number of crashes for each hour in the day. The fourth day, Thursday, also shows the most number of crashes per day in the week.

Exploratory Data Analysis

- **ST_COLCODES 45, 0, 1 and 2 result in the highest ratio of class 2 collisions**

From Figure 13 shown in my accompanying report, you can see that the ST_COLCODES with the highest class 2 collision ratio is 45, 0, 1 and 2. The description for each of these codes are as follows:

- 45: Bicycle
- 0: Vehicle Going Straight Hits Pedestrian
- 1: Vehicle Turning Right Hits Pedestrian
- 2: Vehicle Turning Left Hits Pedestrian

These ST_COLCODEs result in the highest chance of the collision being of type 2 should they occur.

Balancing The Dataset

- **Initially**
  - Total collisions = 194,673
  - Class 1 Collisions = 136,485
  - Class 2 Collisions = 58,188

- **This imbalance will bias the machine learning model to favour the majority class (Class 1).**
- **Random Under Sampling (RUS) have been used to balance the labels and reduce the number of Class 1 collisions to 58,188 (equal to that of Class 2)**
- **There are now 116,376 collisions in the dataset.**

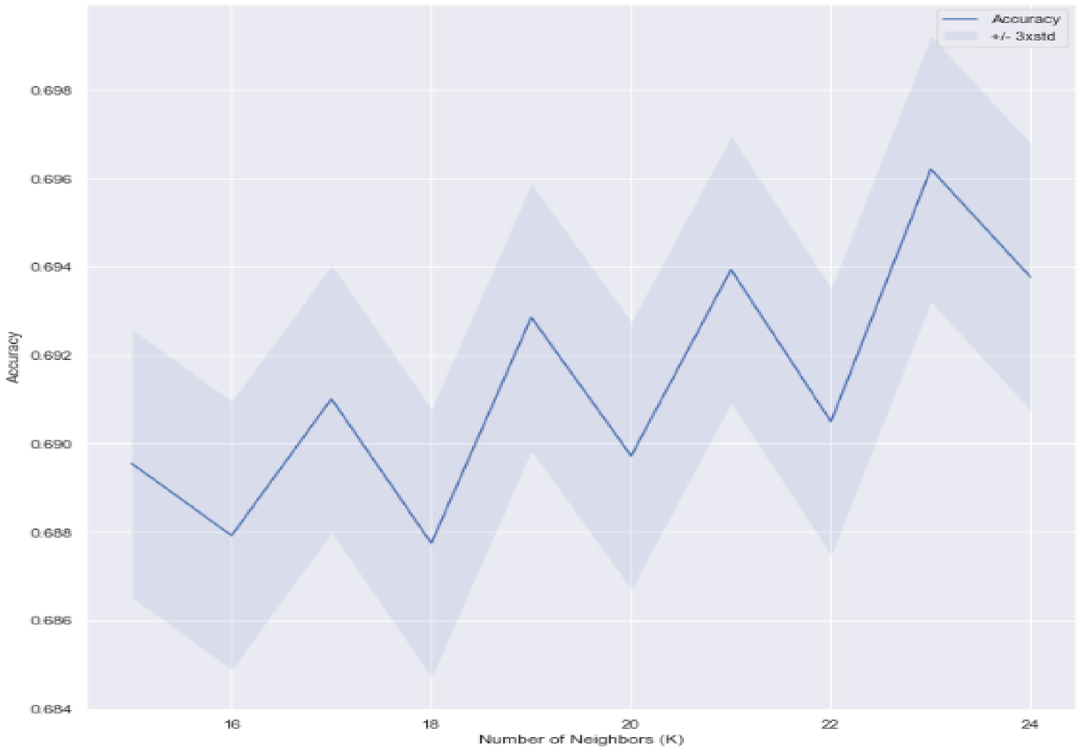
Construction of Machine Learning Models

- **The combination of features in the dataset will have recurring patterns that predict the severity of the collision.**
- **Machine Learning Models that we will employ:**
  - K-Nearest Neighbours
  - Decision Tree
  - Random Forest
  - Logistic Regression
  - Artificial Neural Networks

Machine Learning Model: K-Nearest Neighbour

- **Stores all cases and predicts new cases by comparing the similiarity of features to those it has already been trained on. Using distance metrics such as Euclidean and Minkowski to measure similarity between features.**

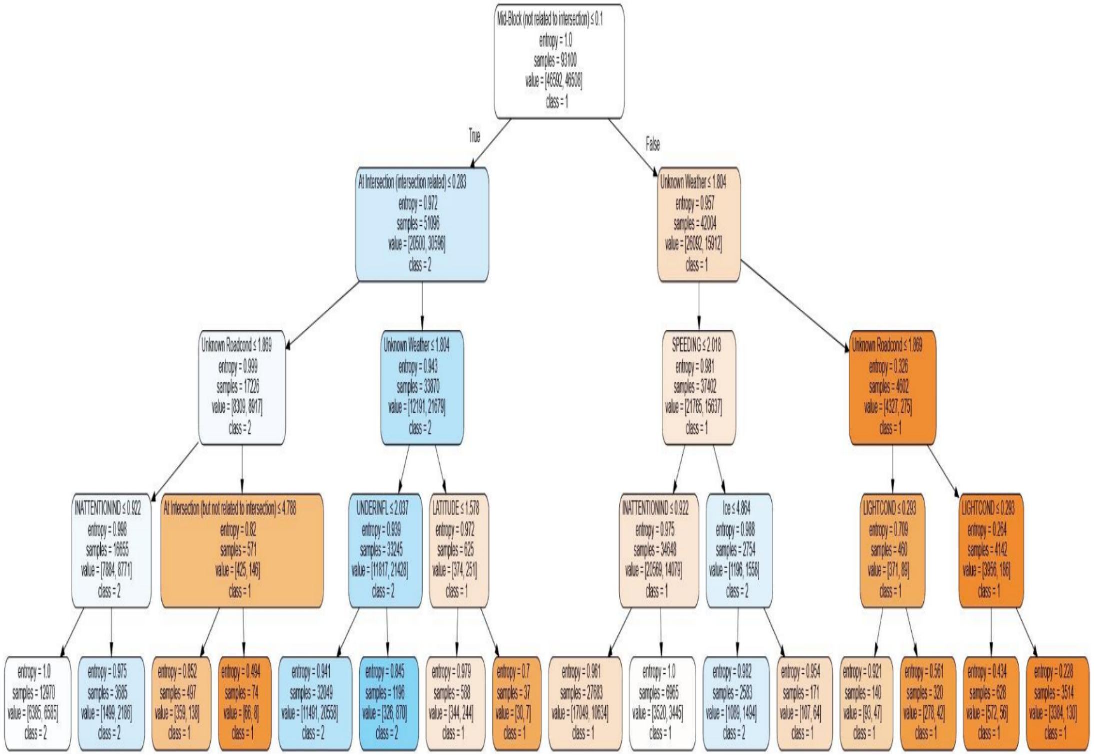
- **General Accuracy: 0.6987**
- **Jaccard-Score: 0.5181**
- **F1-Score: 0.6956**



Machine Learning Model: Decision Tree

- **The Decision Tree iterates through the features about a case (represented in the branches) to conclusions about the case's target value (represented in the leaves). Leafs represent class labels and branches represent conjunctions of features that lead to those class labels.**

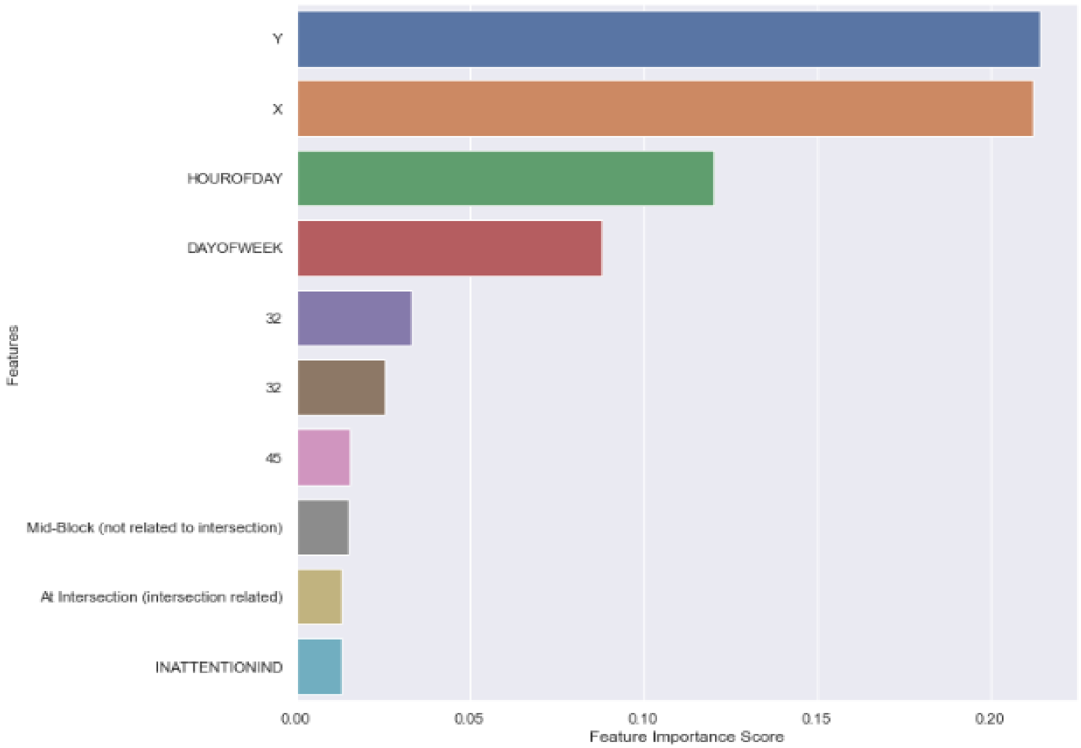
- **General Accuracy: 0.6310**
- **Jaccard-Score: 0.4644**
- **F1-Score: 0.6351**



Machine Learning Model: Random Forest

- **Operates by constructing numerous Decision Trees on various, random sub-samples of the dataset and then outputs the class that is the most common of the classes.**

- **General Accuracy: 0.6825**
- **Jaccard-Score: 0.5138**
- **F1-Score: 0.6873**



Machine Learning Model: Logistic Regression

- Logistic Regression converts a continuous variable given by a Linear Regression function to a, in this case binary, categorical variable. This is done by defining a threshold value, and assigning a class dependant on whether the continuous Linear Regression output falls below or above the threshold.

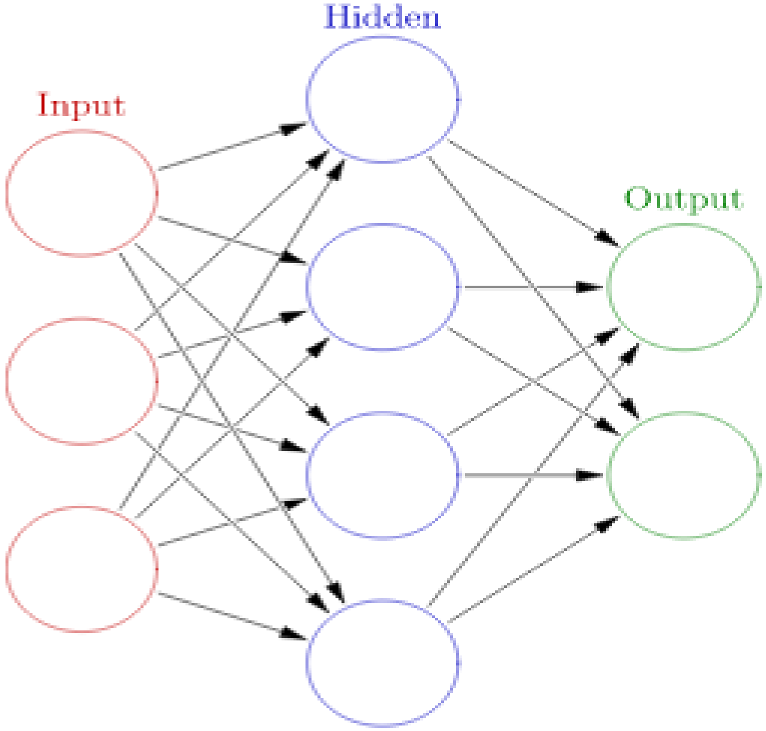
- General Accuracy: 0.7106
- Jaccard-Score: 0.5197

- F1-Score: 0.7084
- Log-Loss: 0.5413

Machine Learning Model: Artificial Neural Network

- The Decision Tree iterates through the features about a case (represented in the branches) to conclusions about the case's target value (represented in the leaves). Leafs represent class labels and branches represent conjunctions of features that lead to those class labels.

- General Accuracy: 0.6933
- Jaccard-Score: 0.5051
- F1-Score: 0.6955
- Log-Loss: 0.5714



Conclusion

- **Combination of location (X &Y) have the biggest impact on determining the severity of the collision. Followed by TIMEOFDAY, then DAYOFWEEK. STCOLCODEs, 32 and 45, being present and JUNCTIONTYPE: At Intersection (intersection related) or Mid-Block (not intersection related) also contribute to determining the severity of the collision, but to a lesser extent.**
- **Logistic Regression and Artificial Neural Network performs best in all evaluation categories. However, there is very little overall variation in the performance of all models and no single model significantly standouts as better, despite their very different approaches**
- **Top 3 Machine Learning Models:**
  - Artificial Neural Network
  - Logistic Regression
  - K-Nearest Neighbours