Predicting the severity of road collisions in Seattle

Submitted by: Thang Le Dinh

Date: 10th October 2020

# Introduction

## **Background**

Seattle is a seaport city on the West Coast of the United States. It is the largest city in both the state of Washington and the Pacific Northwest region of North America. According to U.S. Census data released in 2019, the Seattle metropolitan area's population stands at 3.98 million, making it the 15th-largest in the United States. In July 2016, Seattle was again the fastest-growing major U.S. city, with a 3.1% annual growth rate.

## **Problem**

Since Seattle is a big and dynamic city, one of the most important challenges for the government of Seattle is to prevent car collisions. For instance, hereafter is its 2015 collision clock[[1]](#footnote-1):

* A crash occurred every 4.5 minutes.
* A person died in a crash every 16 hours.
* A person was injured in a crash every 11 minutes.
* A motorcyclist was in a crash every 4 hours.
* A pedestrian or bicyclist was involved in a crash every 2 ½ hours.
* A pedestrian or bicyclist was killed in a crash every 4 days.
* A speeding driver was involved in a crash every 27 minutes.
* An inattentive/distracted driver was involved in a crash every 12 minutes.
* A person was killed by an impaired driver every 1 ½ days

## **Motivation**

This project aims at predicting the severity of car collisions for the Seattle Police Department (SPD) based on different factors, such as weather, road and light conditions as well as the number of peoples and vehicles involved.

# Data understanding

## **Data sources**

Since 2004, the Seattle Police Department (SPD) has collected data related to road collisions recorded by the Traffic Records.

The dataset includes the following key attributes:

* Collision address types
* Levels of the severity of the collision
* Numbers of people and pedestrians involved
* Numbers of bicycles and vehicles involved
* Numbers of injuries, serious injuries and fatalities in the collision
* Situations related to inattention, drugs or alcohol, speeding, and pedestrian right of way as well as conditions of weather, road, and light.

## Data cleaning

Firsly, the dataset is downloaded and read into a dataframe. Thus, this dataset is cleaned up to remove columns that are not informative to us for visualization.

The following columns have been removed: ["X", "Y", "INCKEY", "COLDETKEY", "REPORTNO", "STATUS", "INTKEY", "LOCATION", "EXCEPTRSNCODE", "EXCEPTRSNDESC", "SEVERITYCODE.1", "SEVERITYDESC", "COLLISIONTYPE", “UNDERINFL", "PEDCOUNT", "PEDCYLCOUNT", "INCDATE", "INCDTTM", "JUNCTIONTYPE", "SDOT\_COLCODE", "SDOT\_COLDESC", "PEDROWNOTGRNT", "SDOTCOLNUM", "ST\_COLCODE", "ST\_COLDESC", "SEGLANEKEY", "CROSSWALKKEY", "HITPARKEDCAR"].

Finally, columns and rows, where at least one element is missing, have been removed.

## Feature selection

As mentioned in the following Table, 10 key attributes were selected to determine certain patterns and correlations and then used for training a machine-learning model to predict the probability and severity of collisions.

Table 1: The feature selection.

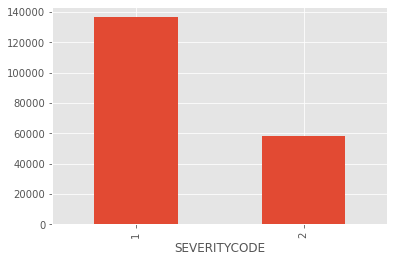
|  |  |
| --- | --- |
| Attributes | Description |
| OBJECTID | ESRI unique identifier |
| SEVERITYCODE | A code that corresponds to the severity of the collision |
| ADDRTYPE | Collision address type |
| PERSONCOUNT | The total number of people involved in the collision |
| VEHCOUNT | The number of vehicles involved in the collision |
| INATTENTIONIND | Whether or not collision was due to inattention |
| WEATHER | A description of the weather conditions during the time of the collision |
| ROADCOND | The condition of the road during the collision |
| LIGHTCOND | The light conditions during the collision |
| SPEEDING | Whether or not speeding was a factor in the collision |

For verifying the consistency, the types of the column labels have been examined to ensure that all column labels of type string.

# Exploratory Data Analysis

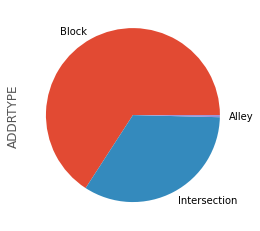
It is observed that the dataset includes different categorical variables, such as Severity Code, Address type, Weather, Road condition, and Light condition. Therefore, we can explore the data related to each variables and then explore the relationship between those variables.

## Severity code



In the dataset, the majority of collisions is related to the “1-prop damage”, which means that there is a damage of the vehicles. There is another category related to the “2-injury”, which means that there is injury. There is no collision, which are related to fatality and serious injury.

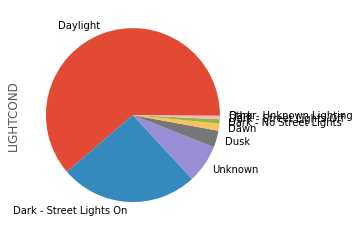
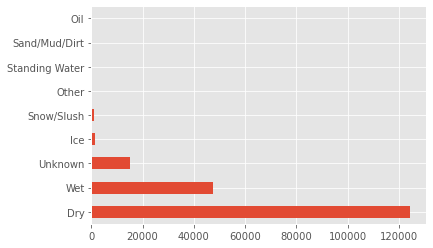
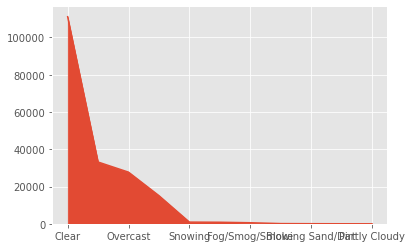
## Address type



There are three collision address types: Alley, Block and Intersection. The majority of collisions is related to block and then intersection.

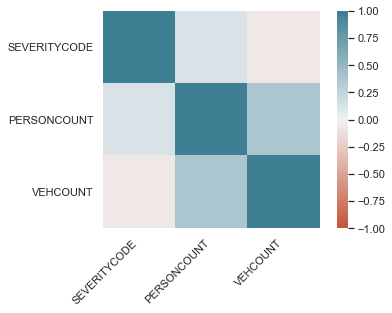
## Weather, Road and Light conditions

It is interesting to know that the majority of collisions is happened in the good driving conditions: clear vision (weather), dry (road) and day light (light condition).



## Relationships between the variables

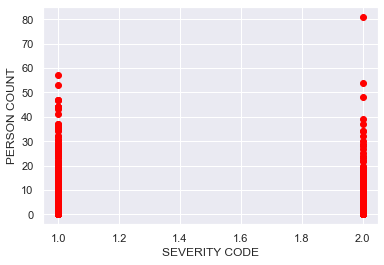
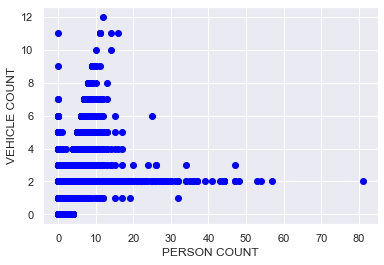
Let us focus on the following numerical variables: Severity code, Person count and Vehicle count.



To determine the potential relationships between those variables, a heat map is used to visualize the relationships depicted by color. It is observed that there may be relationships between *Person count* and *Vehicle count* as well as *Person count* and *Severity code*.

The simple scatter plots are have created for exploring about these two potential relationships.

### Relationship between SEVERITY CODE and PERSON COUNT

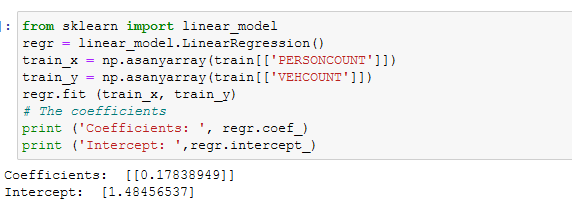
 

As observed in the scatter plots, there is no relationship between *Person count* and *Severity code*; however, there may be a relationship between *Vehicle count* and *Person count*.

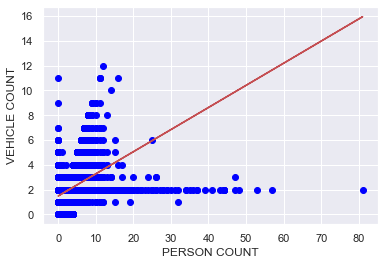
### Simple linear regress for representing the relationship between SEVERITY CODE and PERSON COUNT

Train/Test Split step splits the dataset into training and testing sets, which are mutually exclusive. We will split the dataset into train and test sets, 80% of the entire data for training, and the 20% for testing.

Thus, Coefficient and Intercept are used as the parameters of the fit line.



We can the plot the fit line over the data:

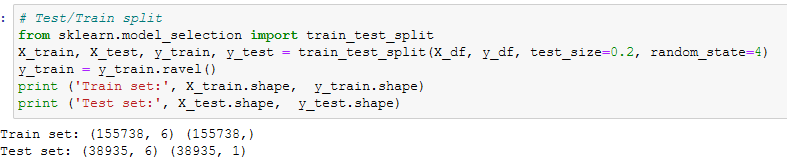


# Machine learning modeling

In this section, we will continue to apply different machine learning models to capture more insights from this dataset.

## Creating train and test dataset

Train/Test Split step splits the dataset into training and testing sets, which are mutually exclusive. We will split the dataset into train and test sets, 80% of the entire data for training, and the 20% for testing.

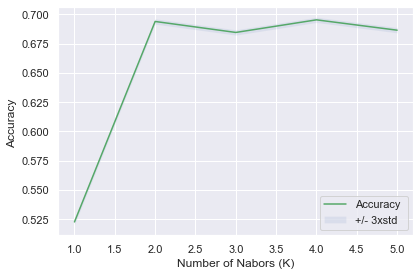


In the following, the three machine learning models such as K Nearest Neighbors, Decision Tree and Regression, which are often used to analyze the large dataset, will be selected for this project.

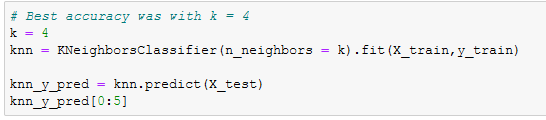
## K Nearest Neighbors analysis

The k-nearest neighbors (KNN) algorithm is an easy-to-implement supervised machine learning algorithm that can be used to solve both classification and regression problems.

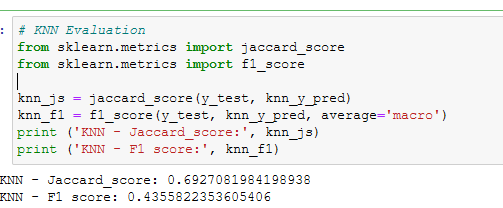
K in K Nearest Neighbors is the number of nearest neighbors to examine, which is supposed to be specified by the User. To choose right value for K, chose k =1 to calculate the accuracy of prediction using all samples in the test set and increase the k to find which k is the best for the model.



According to the above figure, the best accuracy was 0.6953127006549377, with K = 4.



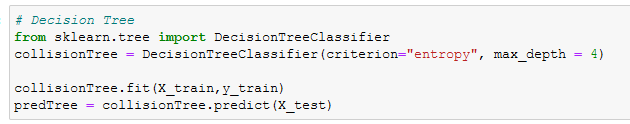
Continuing with K=4, we can train the data and evaluate the K Nearest Neighbours (KNN) using Jaccard score and F1 score. Jaccard is defined as the size of the intersection divided by the size of the union of two label sets. Meanwhile, the F1 score is the harmonic average of the precision and recall.



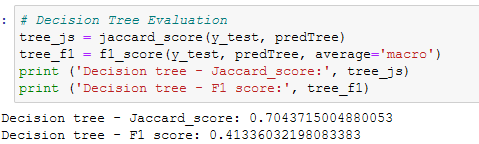
## Decision tree analysis

Decision tree is one of the machine learning approaches that uses a decision tree as a predictive model, which goes from observations about an item (represented in the branches) to conclusions about the item's target value (represented in the leaves).

We will first create an instance of the DecisionTreeClassifier called CollisionTree. We then fit the data with the training feature matrix X\_traint and training response vector y\_train. Finally, we make some predictions on the testing dataset and store it into a variable called predTree.

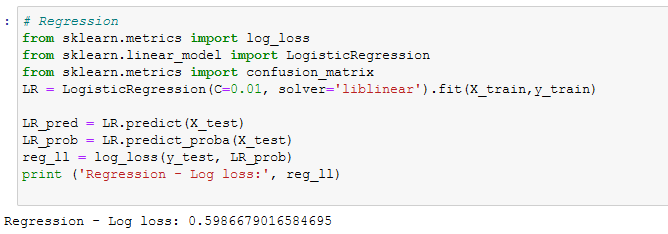


In the following, we import metrics from sklearn and check the accuracy of our model.

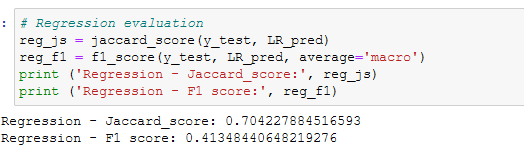


## Regression analysis

The logistic regression is selected to model the probability of an event existing. The current version of Logistic Regression in Scikit-learn, support regularization, which is a technique used to solve the overfitting problem (based on the C parameter indicating inverse of regularization strength). We then fit the model with the train set, predict using the test set and try log loss for evaluation.



Next, we try Jaccard index for accuracy evaluation and calculate the F1 scores for each label based on the precision and recall of that label.



# Discussion

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **K Nearest Neighbours** | **Decision Tree** | **Regression** |
| Jaccard score | 0.6927081984198938 | 0.7043715004880053 | 0.704227884516593 |
| F1 score | 0.4355822353605406 | 0.41336032198083383 | 0.41348440648219276 |
| Log loss | N/A | N/A | 0.5986679016584695 |

Firstly, the Jaccard score is used for accuracy evaluation. If the entire set of predicted labels for a sample strictly match with the true set of labels, then the subset accuracy is 1.0; otherwise it is 0.0. We can observe that accuracy of the three models is relative high (around 0.70). The best model in this case is Decision tree (highest Jaccard score = 0.7043715004880053).

Secondly, the F1 score is the harmonic average of the precision and recall. The F1 score reaches its best value at 1 (perfect precision and recall) and worst at 0. The best model based on this score is the K Nearest Neighbours (highest F1 score: 0.4355822353605406)

Thirdly, the Log loss (Logarithmic loss) measures the performance of a classifier where the predicted output is a probability value between 0 and 1. Since we can calculate only the Log loss for the Regression model, the only insight in that case is that the regression model is also an appropriate choice (Log loss = 0.5986679016584695 can be considered as an above average)

# Discussion

# Based on our analysis, hereafter are some important points:

# - The majority of collisions causes injury and vehicle damage

# - Most of the collision happened at certain specific address types: block and intersection.

# - Most of the collisions can be happened in a good driving condition such as vision clear, day light and dry weather.

# - When the collusions involved several vehicles, it may involve several persons.

# - Decision tree model could be a good model to build a decision process to warn drivers the possibility of collisions.

# - The K Nearest Neighbors model could be a good model to determine the patterns of collision to build an alert system.

1. 2015 Annual Collision Summary : <https://www.wsdot.wa.gov/mapsdata/crash/pdf/2015_Annual_Collision_Summary.pdf> [↑](#footnote-ref-1)