Da

Jon Frampton

Abstract

Would it surprise you to know more accidents happen between intersections rather than in intersections? In Seattle, you are just as likely to suffer injury from a ‘rear-end’ collision between intersections as injury from a ‘head-on’ collision in an intersection.

IBM Applied Data Science Capstone

This is a capstone project to the IBM Applied Data Science Certificate Course.

Table of Contents

[1. Introduction 3](#_Toc50654796)

[1.1 Background 3](#_Toc50654797)

[1.2 Statement of Problem - Objective 3](#_Toc50654798)

[1.3 Interest 3](#_Toc50654799)

[2. "Business Understanding" - Data Acquisition and Initial Examination 3](#_Toc50654800)

[2.1 About the Dataset 3](#_Toc50654801)

[2.2 Initial Data Set Review and Discussion 3](#_Toc50654802)

[3 Initial Data Cleaning, Transformation and Feature Selection 4](#_Toc50654803)

[3.1 Data Cleaning 4](#_Toc50654804)

[3.2 Feature Selection 4](#_Toc50654805)

[4. Exploratory Data Analysis 6](#_Toc50654806)

[4.1 Relationship Between Collision Type, Location and Severity 6](#_Toc50654807)

[4.2 Exploring Time of Day, Location and Severity 7](#_Toc50654808)

[4.3 Exploring Human Factors, Location and Severity 7](#_Toc50654809)

[4.4 Exploring Environmental Factors, Location and Severity 8](#_Toc50654810)

[4.5 Exploring Contributing Factors, Location and Severity 8](#_Toc50654811)

[4.6 Interim Summary Data Findings 9](#_Toc50654812)

[5. Deeper Data Analysis with Maps 9](#_Toc50654813)

[5.1 Event Location Map to Explore Insights 9](#_Toc50654814)

[5.2 Clustered Drill Down Map 10](#_Toc50654815)

[5.4 Deeper Insights with Data Markers/Labels 11](#_Toc50654816)

[5.5 Summary Map Findings 11](#_Toc50654817)

[6 Methodology: Choosing a Predictive Model 12](#_Toc50654818)

[6.1 Classification 12](#_Toc50654819)

[6.2 Comparing the Different Models 12](#_Toc50654820)

[6.3 Information on these algorithms: 12](#_Toc50654821)

[6.4 Information on the scores: 13](#_Toc50654822)

[7 Discussion 13](#_Toc50654823)

[8 Conclusion 13](#_Toc50654824)

[9 Future Directions 14](#_Toc50654825)

Table of Figures

[Figure 1: Collision Type, Location and Severity 6](#_Toc50654826)

[Figure 2: Time of Day, Location and Severity 7](#_Toc50654827)

[Figure 3: Human Factors, Location and Severity 7](#_Toc50654828)

[Figure 4: Environmental Factors, Location and Severity 8](#_Toc50654829)

[Figure 5: Contributing Factors, Location and Severity 8](#_Toc50654830)

[Figure 6: Incident Location Map (limited to 5000 points for this figure). 10](#_Toc50654831)

[Figure 7: Clustered Incident Location Map (limited to 5000 points for this figure). 10](#_Toc50654832)

[Figure 8: Clustered Incident Location Map Drill Down (limited to 5000 points for this figure). 11](#_Toc50654833)

Table of Tables

[Table 1: Features Selected 4](#_Toc50654834)

[Table 2: Features Rejected 5](#_Toc50654835)

[Table 3: Features Transformed 6](#_Toc50654836)

[Table 4: Predictive Model Scores: 12](#_Toc50654837)

# 1. Introduction

## 1.1 Background

The city of Seattle, Washington has made public a dataset containing collision data from Jan 2004 through May 2020. Seattle makes their data openly available to the public in an apparent attempt to be more transparent. Additional information can be found:

<https://data-seattlecitygis.opendata.arcgis.com/datasets/5b5c745e0f1f48e7a53acec63a0022ab_0>

## 1.2 Statement of Problem - Objective

The objective is to predict traffic accident severity given the available data (severity values: 1 = Property Damage Only Collision, 2 = Injury Collision).

## 1.3 Interest

Other cities may benefit by understanding various factors leading to accident severity, and may want to consider these findings in city planning. In addition, all levels of government, transportation manufacturers, innovators and insurance companies may be interested in these findings to improve current systems. People, in general, may benefit from understanding the risks.

# 2. "Business Understanding" - Data Acquisition and Initial Examination

## 2.1 About the Dataset

The initial dataset was obtained via: <https://www.coursera.org/learn/applied-data-science-capstone/supplement/Nh5uS/downloading-example-dataset>

The metadata was obtained via: <https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Metadata.pdf>

## 2.2 Initial Data Set Review and Discussion

Initial review indicates likely use of categorical supervised learning methods.

The accident "participants" are: Motor Vehicles, Driverless Vehicles, Pedacyclists, Pedestrians with several records not applicable. There are some incidental 'participants', namely objects, trains, etc.

The data from 2020 seems to have less rich data, appearing incomplete in general. Thus, for purposes of this capstone project and the prediction algorithm, it may make sense to limit the dates used to create the algorithm.

There appears to be some more interesting data which will help the analysis and model development. Some may possibly require transformation for categorical supervised learning:

* Environmental Factors (e.g., Weather, Road, Light). Note: Crosswalk may be a factor, but crosswalks are generally in intersections
* Human Factors (e.g., Driver Attentiveness, Speeding)
* #s Involved (e.g., People, Bikes, Vehicles)
* Time: Time is on the 24-hour scale, and time of day, specifically hour of day, may contribute
* Date: Weekend / Weekday / Holiday may, but likely not given the Objective and participants
* Address Type or Junction Type

Some data appears less relevant to any algorithm but may be generally informative:

* Indices/Keys, Lat/Long/Specific Address, Location/Intersection Key and similar data

And there is duplicative data between the city of Seattle and Washington State:

* Upon initial examination, the Seattle Department of Transportation (SDOT) data is more complete than the Washington State data with better descriptions. The Washington State (ST) data appears much less complete/accurate.
* Status appears related to the difference between ST and SDOT data. There is strong correlation between an unmatched status, lack of a time stamp, and lack of other amplifying information.

To complete this analysis, the more interesting data listed above will be explored. The remaining data may be initially examined to assess its potential use. Some data appears to be less useful and will be dropped going forward.

For the purposes herein, the following supervised learning models will be explored for prediction accuracy:

* K Nearest Neighbor (KNN)
* Decision Tree (DT)
* Support Vector Machine (SVM)
* Logistic Regression (LR)

# 3 Initial Data Cleaning, Transformation and Feature Selection

## 3.1 Data Cleaning

Many records had missing values. This was especially true for the year 2020, which seems to indicate the year is under review. Previous years contained more complete data. For this reason, all 2020 records were removed.

The records also had dual reporting between the city of Seattle and the state of Washington. The records contained a matched feature which indicated when these records had been completed and agreed upon by both agencies. This data inconsistency appeared, in general, to align with the 2020 timeline issue.

Several records had timestamps and other indicators of a non-standard input which created records with significantly incomplete data. These records were removed.

At this point, a count of 30483 records have been initially removed which had some issue (e.g., conflicting state vs Seattle data). The Seattle data appears more complete and closer to the source. Of the initial 194673 records, 164190 records remain. This should be enough for the analysis.

## 3.2 Feature Selection

Many features contained duplicate information. Multiple uses of the same data can bias outcomes. To preclude this, one data feature was used for a given variable. Several features appear useful for algorithm usage. This is the table of features selected.

Table 1: Features Selected

|  |  |  |
| --- | --- | --- |
| **FEATURE SELECTED** | **DESCRIPTION** | **WHY USED** |
| SEVERITYCODE  INCDTTM  STATUS  SDOT\_COLDESC  X  Y  ADDRTYPE  PERSONCOUNT  PEDCOUNT  PEDCYLCOUNT  VEHCOUNT  SPEEDING  INATTENTIONIND  WEATHER  ROADCOND  LIGHTCOND | Target Feature to Predict  Date and Time of Incident  Matched or Unmatched Record  Seattle Incident Description  Latitude of Incident  Longitude of Incident  Alley, Block or Intersection  Person Involved Count  Pedestrian Count  Pedcyclist Count  Vehicle Count  Speeding as a Factor  Known Driver Inattention  Weather Type  Road Condition Type  Lighting Condition Type | Feature to Predict  Potential Relevance  Initial Record Accuracy  Understand Incident Type  Location Info  Location Info  Location Type Indicator  Numbers Involved  Numbers Involved  Numbers Involved  Numbers Involved  Contributing Factor Potential  Contributing Factor Potential  Contributing Factor Potential  Contributing Factor Potential  Contributing Factor Potential |

Several features were rejected. This is the table of features rejected and why.

Table 2: Features Rejected

|  |  |  |
| --- | --- | --- |
| **FEATURE REJECTED** | **DESCRIPTION** | **WHY NOT USED** |
| OBJECTID  SHAPE  INCKEY  COLDETKEY  REPORTNO  INTKEY  LOCATION  EXCEPTRSNCODE  EXCEPTRSNDESC  SEVERITYCODE  COLLISIONTYPE  INCDATE  JUNCTIONTYPE  SDOT\_COLCODE  PEDROWNOTGRNT  UNDERINFL  SDOTCOLNUM  ST\_COLCODE  ST\_COLDESC  SEGLANEKEY  CROSSWALKKEY  HITPARKEDCAR | ESRI Unique Identifier  ESRI Geometry Field  Unique Key for Incident  Secondary Key for Incident  Assigned Report Number  Assigned Key for Intersection  Mailing Address Location  None  None  Severity of Accident  Collision Type  Date of Incident  Junction Category  Code Assigned by SDOT  Pedestrian Right of Way  Tire Inflation  Number Assigned by SDOT  Code by State  Description by State  Lane Incident Occurred  Crosswalk Key  Hit Parked Car (Y or N) | Database Key  Not Useful  Database Key  Database Key  Internal Report Number  Lookup Key  Lat/Long Used  Not Useful  Not Useful  Already Used  Data from SDOT\_COLDESC  Contained in INCDTTM  ADDRTYPE Used  Not Useful  Not Useful  Not Useful  Not Useful  State Data Lags  State Data Lags  Not Useful  ADDRTYPE Used  SDOT\_COLDESC Used |

Some of the data requires transformation to be used in the algorithms. The next several stages are the data transformation process.

Table 3: Features Transformed

|  |  |  |
| --- | --- | --- |
| **FEATURES TRANSFORMED** | **INTO** | **WHY** |
| INCDTTM | HOUR | Hour of Day Use |
| SPEEDING &  INATTENTIONIND | HUM\_FAC | Explore human factor aspect |
| WEATHER, ROADCOND, & LIGHTCOND | ENV\_FAC | Explore environmental factor aspect |
| HUM\_FAC & ENV\_FAC | CONTR\_FAC | Explore combined contributing factors aspect |

# 4. Exploratory Data Analysis

## 4.1 Relationship Between Collision Type, Location and Severity

In Figure 1, the relationship between three key variables is shown. Collision Type 17 is a ‘head-on’ motor vehicle collision. Collision Type 20 is a ‘rear-end’ motor vehicle collision. Given a basic understanding of physics, the figure does make sense, especially given traffic densities.

Intersection danger, traditionally, has resulted in many control devices being implemented (e.g., stop signs, lights, crosswalks). However, the high number of Block incidence came as a surprise, especially given the number of resulting injuries. This may come as a surprise to many.

Of note, approximately 2/3s of incidents occur in the Block environment. Of these, the ‘rear-end’ incidents cause the most injury. The count is similar to the Intersection count of injuries but is lower. The cause of these Block incidents is of interest.

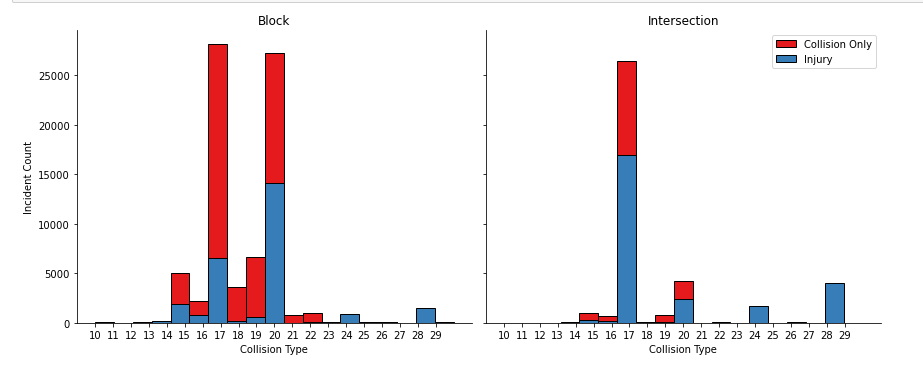


Figure 1: Collision Type, Location and Severity

## 4.2 Exploring Time of Day, Location and Severity

Figure 2 seems consistent with typical patterns of life. It is interesting to note the number seems to go up throughout the day. For example, why are there more incidents at 1700 as opposed to 1200 or 0800? Beyond this, the patterns seem to hold and nothing significant jumps out. The data at 0000 hour is indicative of a bad initial timestamp and is an outlier.

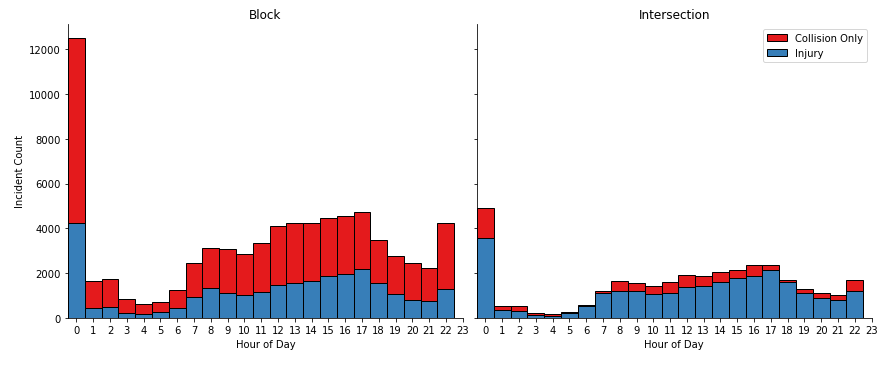


Figure 2: Time of Day, Location and Severity

## 4.3 Exploring Human Factors, Location and Severity

Figure 3 shows the relationship between speeding and driver inattentiveness (known), location and severity. In any location, the data show any count of speeding or driver inattention is a minor factor.

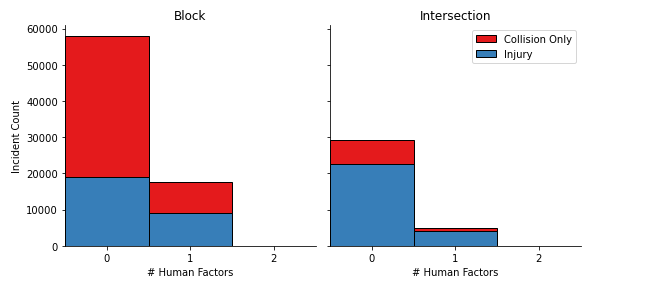


Figure 3: Human Factors, Location and Severity

## 4.4 Exploring Environmental Factors, Location and Severity

Figure 4 is a bit more interesting since the number of environmental factors (e.g., weather, lighting, road conditions) whether just one or more than one total is involved in more incidents than when none are present. Given the weather in Seattle, this may be normal.

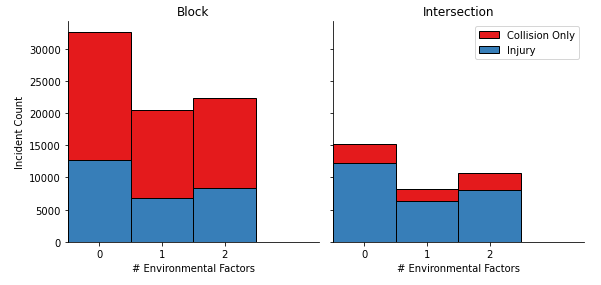


Figure 4: Environmental Factors, Location and Severity

## 4.5 Exploring Contributing Factors, Location and Severity

Taken in total, in approximately 65% of incidents, some contributing factor is present, whether human factors or environmental factors. Figure 5 provides some insight into the counts.

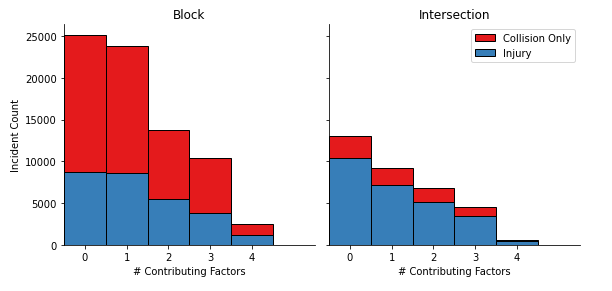


Figure 5: Contributing Factors, Location and Severity

## 4.6 Interim Summary Data Findings

Based on this initial data relationship exploration, here is a list of the interim findings:

* Intersection incidents tend to be 'head-on' and Block incidents tend to be 'rear-end.'
* Intersections: ~1/3 of total incidents occur in intersections with ~66% of these 'head-on' incidents resulting in injury
  + Human factors (i.e., SPEEDING and INATTENTIVENESS) seem to be a minor factor.
  + Environmental factors (i.e., LIGHTING, WEATHER, etc.) appear to contribute significantly
* Block: ~2/3s of total incidents are within the block infrastructure:
  + ~50% of these are 'rear-end' events with an ~50% injury rate
  + ~50% of these are 'head-on' events with ~25% injury rate
  + Human factors (i.e., SPEEDING and INATTENTIVENESS) seem to be a minor factor.
  + Environmental factors (e.g., reduced visibility) seem to contribute in a major way
* Regardless of location, when motor vehicles impact pedacyclists or pedestrians, the result is almost always injury.
* Some Contributing Factor (e.g., Human or Environmental) is present in ~2/3s of all incidents.
* The severity code ratios appear to hold across time of the day. Though there does appear to be more incidents as the day goes on until 1700.

# 5. Deeper Data Analysis with Maps

## 5.1 Event Location Map to Explore Insights

The latitude and longitude of incidents was plotted to get a sense of location and/or density issues. This map was limited to 5000 point-plots vice putting all 164190 points on the map. There are some insights which emerge. For example, many of the incidents occur along major roadways.

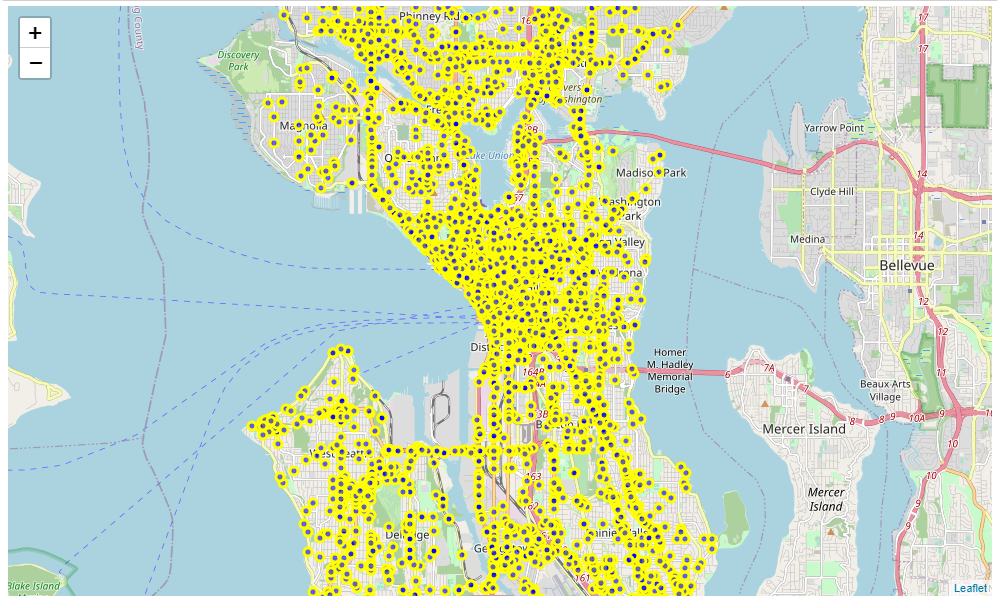


Figure 6: Incident Location Map (limited to 5000 points for this figure).

## 5.2 Clustered Drill Down Map

Figure 7 clusters the incidents by zones showing the counts with an ability to drill down to specific incidents as shown in Figure 8. One can see the majority of plotted incidents tend to occur in and around downtown.

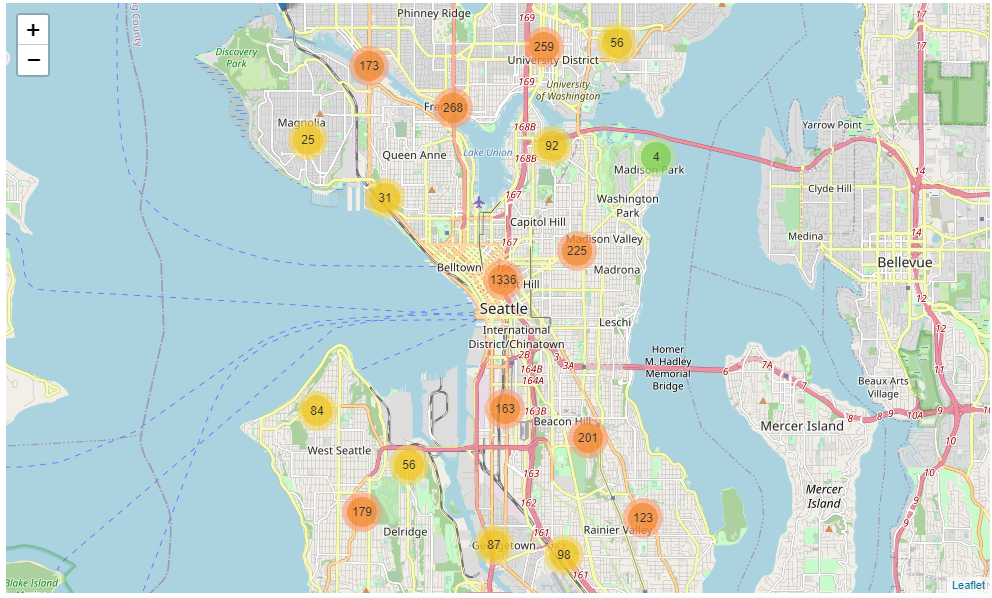


Figure 7: Clustered Incident Location Map (limited to 5000 points for this figure).

## 5.4 Deeper Insights with Data Markers/Labels

This figure is an example of the ability to drill down to specific incident locations. It also helps see numbers of incidents in each area. In general, the closer to downtown, the higher in the number of incidents.

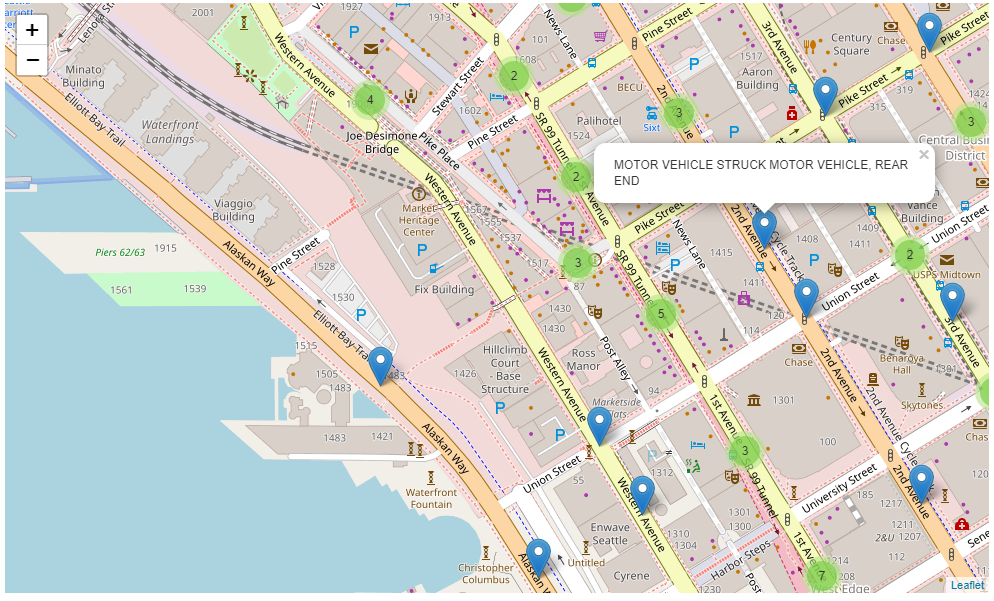


Figure 8: Clustered Incident Location Map Drill Down (limited to 5000 points for this figure).

## 5.5 Summary Map Findings

By examining the maps, the majority of incidents seem to occur in downtown area, which is to be expected. Incidents in congested areas likely cause greater congestion and delays.

# 6 Methodology: Choosing a Predictive Model

## 6.1 Classification

For the purposes herein, the following supervised learning models will be explored for prediction accuracy

* K Nearest Neighbor(KNN)
* Decision Tree
* Support Vector Machine
* Logistic Regression

## 6.2 Comparing the Different Models

Looking at the various predictive models and the associated scores, Logistic Regression is the chosen means to predict the severity of an incident in Seattle with the given information. Several permutations were tried with these results being most appropriate without over-fitting or under-fitting the data.

This paper doesn’t go through the mechanics of determining Jaccard Scores, F1\_Scores or LogLoss. The means of how these were derived can be found at: <https://github.com/API4Fun/Coursera_Capstone/blob/master/Coursera%20IBM%20Capstone.ipynb>

Table 4: Predictive Model Scores:

| **Algorithm** | **Jaccard** | **F1\_Score** | **LogLoss** |
| --- | --- | --- | --- |
| **KNN** | 0.688673 | 0.686588 | NA |
| **Decision Tree** | 0.703648 | 0.664863 | NA |
| **SVM** | 0.705951 | 0.655852 | NA |
| **Logistic Regression** | 0.700761 | 0.779769 | 0.575659 |

## 6.3 Information on these algorithms:

* ***KNN:*** <https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm>
* ***Decision Tree:*** [https://en.wikipedia.org/wiki/Decision\_tree#:~:text=A%20decision%20tree%20is%20a%20flowchart%2Dlike%20structure%20in%20which,taken%20after%20computing%20all%20attributes).](https://en.wikipedia.org/wiki/Decision_tree%23:~:text=A%20decision%20tree%20is%20a%20flowchart%2Dlike%20structure%20in%20which,taken%20after%20computing%20all%20attributes).)
* ***SVM:***  [https://en.wikipedia.org/wiki/Support\_vector\_machine](%20https:/en.wikipedia.org/wiki/Support_vector_machine)
* ***Logistic Regression:***[https://en.wikipedia.org/wiki/Logistic\_regression#:~:text=Logistic%20regression%20is%20a%20statistical,a%20form%20of%20binary%20regression).](https://en.wikipedia.org/wiki/Logistic_regression%23:~:text=Logistic%20regression%20is%20a%20statistical,a%20form%20of%20binary%20regression).)

## 6.4 Information on the scores:

* ***Jaccard:*** <https://en.wikipedia.org/wiki/Jaccard_index>
* ***F1-Score:*** <https://en.wikipedia.org/wiki/F1_score>
* ***LogLoss:*** <https://www.kaggle.com/dansbecker/what-is-log-loss>

# 7 Discussion

Re-examining the Summary Data and Mapping Findings:

Intersection incidents tend to be 'head-on' and Block events tend to be 'rear-end'. The data seem to indicate significance with 'head-on' and 'rear-end' collisions leading to injury vice impacts from other angles.

Intersection Incidents: ~1/3 of total incidents occur in intersections with ~66% of these 'head-on' incidents resulting in injury. a. Human factors (i.e., SPEEDING and INATTENTIVENESS) seem to be a minor factor. b. Environmental factors (i.e., LIGHTING, WEATHER, etc.) appear to contribute significantly

Block: ~2/3s of total incidents are within the block infrastructure: a. ~50% of these are 'rear-end' events with an ~50% injury rate b. ~50% of these are 'head-on' events with ~25% injury rate c. Human factors (i.e., SPEEDING and INATTENTIVENESS) seem to be a minor factor. d. Environmental factors (e.g., reduced visibility) seem to contribute in a major way

Regardless of location, when motor vehicles impact pedacyclists or pedestrians, the result is almost always injury.

A full ~65% of incidents had some sort of Contributing Factor. Human factors (i.e., speeding and driver inattentiveness) along with environmental factors (i.e., weather, lighting, and road conditions) appear to contribute to a significant number of incidents. These should be considered as necessary in city planning, transportation design, insurance, government research, etc. For example, reducing speed when any of these conditions are present may help.

The severity code ratios appear to hold across time of the day.

While more incidents occur in 'Block' environments, the injury counts are similar for 'Block' and 'Intersection' incidents. From a city planning perspective, much attention has been given to intersection design. It appears it may be worth exploring options for "Block" traffic control to lower the overall number of injuries.

A clear majority of incidents occur in the downtown area, meaning locations with higher traffic congestion counts. These incidents likely lead to higher levels of congestion and delay.

Based upon the available data, it appears a Logistic Regression algorithm can predict with ~70% accuracy whether or not an injury occurred. Several permutations were attempted. The prediction model could be likely improved with better data collection regarding safety devices (e.g., airbags, lane change sensors, collision warning sensors, automatic braking).

# 8 Conclusion

In this assessment, the ability to predict the outcome from a collision incident achieved an ~70% accuracy rate. Key factors arising from the assessment include 'head-on' collisions in the "Intersection" environment and 'rear-end' collisions in the "Block" environment. More incidents occur in congested areas. While these factors are key in city planning, other considerations like human factors (e.g., speeding and driver attentiveness) and environmental factors (e.g., visibility) play a role.

# 9 Future Directions

Several future efforts may help.

For City Planners: Some means of educating the public on the dangers posed by 'rear-end' incidents while 'on the Block.' Control methods may need researched and explored regarding the findings above to reduce risks.

For Transportation Manufacturers: Exploration of sensors to aid in assisted braking, lane control, attention monitoring, etc. may reduce risks.

For Innovators: Pursuit of developmental technologies to support City Planners, Transportation Manufacturers and Vehicle Operators in risk reduction.

For Insurers: Incentivizing City Planners, Transportation Manufacturers and vehicles operators to purchase supporting technologies and promoting driver education programs.

For Government Entities at all levels: Research investment into methods, technologies, training or other means to mitigate the risk findings is needed. It would be useful to determine if these finds hold true across the vast city population.