PREDICTION OF CAR ACCIDENT SEVERITY IN SEATTLE, WA

Coursera Capstone for IBM Data Science Professional Certificate

M. Sagianis

Introduction & Business Problem

Traffic accidents have been, and will continue to be, a problem plaguing society for the foreseeable future. With the world population expected to keep growing, we would expect there to be an increase in cars on the road, which will not be balanced equally by an increase in roadways. To follow that, we expect the numbers of accidents to also increase. However, with enough data, it may be possible to identify factors that increase the likelihood of accidents, and measures can be taken to prevent that. In some cases, measures may be infrastructure upgrades (signage, lightning, street bumps, speed limits, etc.), or they may be campaigns to spread awareness and change human behavior. With self-driving cars starting to become a reality in the upcoming decades, action must be taken now to create safe and reliable roadways. Government transportation departments need actionable data-driven recommendations to improve roadway conditions and reduce accidents and fatalities. Accident prevention will result in many positives for communities, such as less traffic, fewer emissions, improved public transportation reliability, and most importantly, safety. Information and conclusions drawn from this data will help with future urban planning and design, as well as traffic control, and can even be of benefit to car insurance related issues.

An immediately actionable resolution would be to be able to predict the severity of an accident based upon conditions surrounding the driver at the time of accident. In stressful situations such as a car accident, it may be hard to reliably assess how severe an accident can be based on a human explanation via a 911 call. However, if emergency responders were able to efficiently gauge the severity of an accident based upon objective facts (road conditions, weather conditions, collision type, etc.), then resources could be more effectively allocated, and first

responders would be more prepared for the situation they were entering. A model to predict this is only a first step to improving traffic conditions, however it is certainly a necessary one. This report will explore building a model to predict the severity of an accident based on data that is already being recorded by the city of Seattle.

Data

In order to solve the problem of traffic-related accidents, a publicly accessible data set from Seattle, Washington will be used to draw recommendations. The labeled dataset, accessible here, includes roughly 220,000 collisions from January 2004 to September 2020, courtesy of SPD and recorded by Traffic Records. Please note that this dataset is more exhaustive than the one that was supplied through Coursera. The dataset supplied through Coursera was only updated through May 2020, and additionally, only showed cases with accident severity of 0, 1, or 2. The dataset used for this project includes a more updated version that also included accidents of severity 2b and 3. These alphanumeric codes correspond to the following:

Severity Code	**Severity Description**
0	Unknown
1	Property Damage
2	Injury
2b	Serious Injury
3	Fatality

Table 1: Severity Code Description

This dataset includes all types of collisions, with collisions displayed at the intersection or mid-block of a segment. Each accident is assigned a unique identifier, with as many attributes to describe the accident as could be collected. Accidents here are defined as including, but not

limited to, personal vehicles, bicycles, and pedestrians. Other data attributes present include timestamps, location data, accident severity, injuries, collision description, weather conditions, and driver characterization, among others. There is a total of 37 potential data attributes which may be used for analysis and recommendations, although not all of them may be used. Other data will be brought in, such as Seattle road maps and geography, which will be cross referenced with location identifiers to view accident density across the city. Clustering accident information with location will serve to guide recommendations on what action is needed to be taken in which neighborhoods. The dataset will be used to evaluate accident severity and its key causes. A machine learning model will then be used in order to predict the severity of an accident based on a combination of multiple data attributes. A description of data attributes, as well as information of which were used, will be addressed next.

Methodology

The dataset accessed above was a raw data-dump, with missing, miscoded, and unnecessary values for the scope of this project. It would not be feasible to create a model based on all 37 data attributes, and missing values would skew the results. A cursory look at the data attributes shows a column “EXCEPTRSNDESC”, which was either blank, or contained “Not Enough Information, or Insufficient Location Information”. Immediately, any traffic incidents that had this value were dropped, as they would not be able to be used in modeling prediction. Additionally, all incidents were a Severity Code of 0 (“Unknown”), were dropped, as that would create problems with the accuracy of the model. Following this, data attributes were removed from the dataset if they were deemed irrelevant to the scope of this project, or many values were missing. Those data attributes are listed below, as well as their reason for removal:

Data Attribute	Description & Reason for Removal
OBJECTID	ESRI Unique Identifier – Assigned for internal reporting, not useful
INCKEY	Unique key – Assigned for internal reporting, not useful
COLDETKEY	Secondary key – Assigned for internal reporting, not useful
REPORTNO	Report Number – Assigned for internal reporting, not useful
STATUS	"Match"/ "Unmatched" – not useful
LOCATION	Description of location – Latitude and Longitude already reported
EXCEPTRSNCODE	Only blank values remain after removal of "Not Enough Information"
EXCEPTRSNDESC	Only blank values remain after removal of "Not Enough Information"
SEVERITYDESC	Severity Description – Severity code kept, no need for duplicate
INCDATE	Incident Date – DateTime already kept, duplicate info
SDOT_COLDESC	Description of collision code – SDOT code is kept, duplicate info
SDOTCOLNUM	Number given by SDOT -- Assigned for internal reporting, not useful
ST_COLDESC	Description of collision code – State code is kept, duplicate info
SEGLANEKEY	Key for lane segment of collision – Geolocation kept, duplicate
CROSSWALKKEY	Key for Crosswalk – Latitude/Longitude kept, duplicate info
INATTENTIONIND	Collision due to inattention – Many missing values
PEDROWNOTGRNT	Pedestrian right of way granted – Many missing values
SPEEDING	Speeding a collision factor – Many missing values

Table 2: Removed data attributes

Following the removal of data attributes, a couple of minor polishes were done to clean up the data. Any traffic incidents without a latitude and longitude associated with it were removed, as well as any accidents that had a missing or "Unknown" value for Weather, Road Condition, or Light Condition. Values for "UNDERINFL" (Under influence of drugs/alcohol), and "HITPARKEDCAR" (collision involving parked car) we're recoded for clarity, where 0 = No, and 1 = Yes. Finally, Severity code 2b was reclassified as 2.5, and collision codes were classified as -1 if left blank. This ensures that data would be properly handled and analyzed as needed. After all the cleanup, our dataset is left with 170,907 total incidents, and data attributes as stated below:

Data Attribute	Description & Reason for Removal
X	Longitude of location associated with the accident
Y	Latitude of location associated with the accident
ADDRTYPE	Collision Address type
INTKEY	Key for intersection associated with collision
SEVERITYCODE	Code for severity of collision
COLLISIONTYPE	Collision Type
PERSONCOUNT	Total number of people involved in the collision
PEDCOUNT	Number of pedestrians involved in the collision
PEDCYLCOUNT	Number of bicycles involved in the collision
VEHCOUNT	Number of vehicles involved in the collision
INJURIES	Number of total injuries in the collision
SERIOUSINJURIES	Number of serious injuries involved in collision
FATALITIES	Number of fatalities in the collision
INCDTTM	Date and time of incident
JUNCTIONTYPE	Category of junction at which collision took place
SDOT_COLCODE	A code given to collision by SDOT
UNDERINFL	Driver involved under the influence of drugs/alcohol
WEATHER	Weather conditions at time of the collision
ROADCOND	Condition of the road during the collision
LIGHTCOND	Light conditions during the collision
ST_COLCODE	A code given to collision by the State
HITPARKEDCAR	Parked car involved in collision

Table 3: Data attributes remaining

The remaining dataset was used as the starting point for visualization of traffic accidents, as well as the creation of a machine learning model to predict the severity of a traffic incident. K-Nearest Neighbors, Decision Trees, Support Vector Machine, and Logistic Regression models were all generated and compared against each other. K-Nearest Neighbors was chosen as an option since it is a supervised learning model, where data can be “trained” by data nearest to its classification. Decision Tree is an interesting choice, since we can use the historical data as a guideline to build a roadmap arriving to a classification decision. Support Vector Machine algorithm can be used to categorize data points where they would not be linearly separable, as is the case when we have many categorical variables. Finally, Logistic Regression can be used

because we are predicting a categorical value (Severity). These results will be shown below, and the benefits discussed.

Results

Below, a cursory understanding of the data will be shown through various graphs and models. To first understand the data, the range of accidents of various severities is shown below:

	SEVERITYCODE
1.0	112288
2.0	55359
2.5	2932
3.0	328

Table 4: Accident Severity Count

Of the roughly 170,000 cases, most of them are property damage only. This will potentially lead to a biased model, so downsampling will also occur in order to see the difference in accuracy. Following, the Collision Address Type is examined, whether the accident happened at an intersection, or mid-block, as well as the Collision type that occurred:

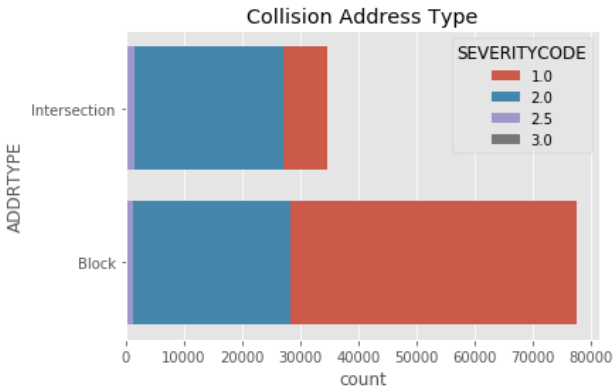


Chart 1: Collision Address Type

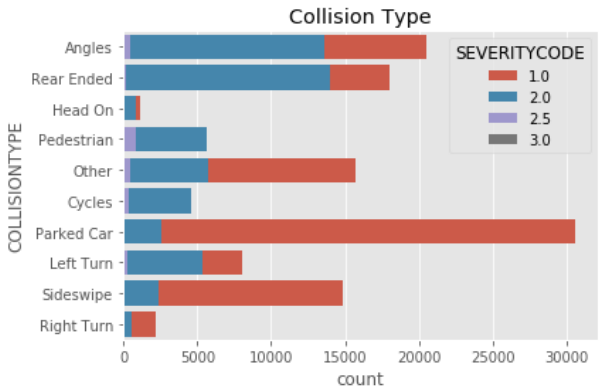


Chart 2: Collision Type

While there are over double the total number of accidents that happen mid-block, a significantly higher proportion of accidents at intersections involve personal injury or worse.

Additionally, while the highest amount of collisions involve a parked car, most are considered not serious, especially compared to Angles or Read Ended. Almost all accidents that involve a pedestrian or a cyclist involve injury. Below is a distribution of individuals involved in accidents, as recorded by the state of Washington:

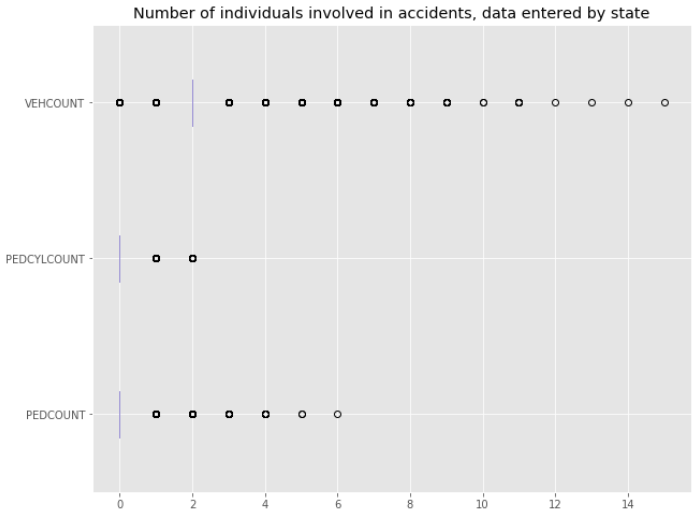


Chart 3: Individual Involvement

Generally, two vehicles are involved per incident, however there is a wider range of outliers. Cyclists and Pedestrians and rarely involved in incidents, but there is the exception. The data not entered by the state of total people involved in collisions show a similar story:

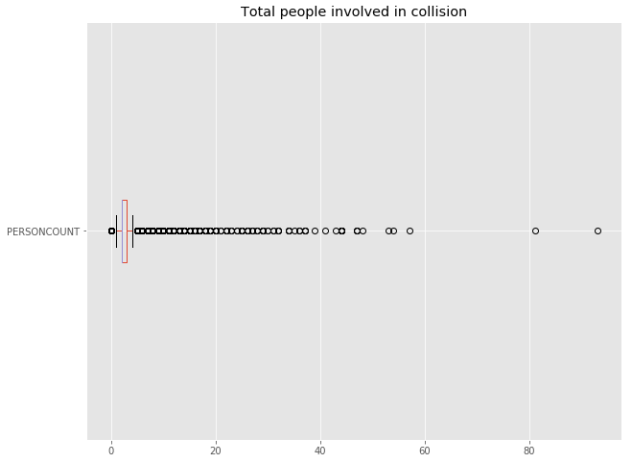


Chart 4: Total Person Involvement

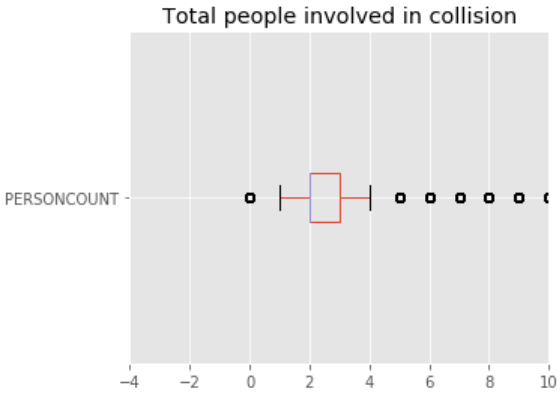


Chart 5: Total Person Involvement (Zoomed)

Over 75% of accidents involve less than 4 people, however there are many outliers, skewing up to the high 90s of people involved. A summary of collision codes, both by SDOT and Washington state are shown below:

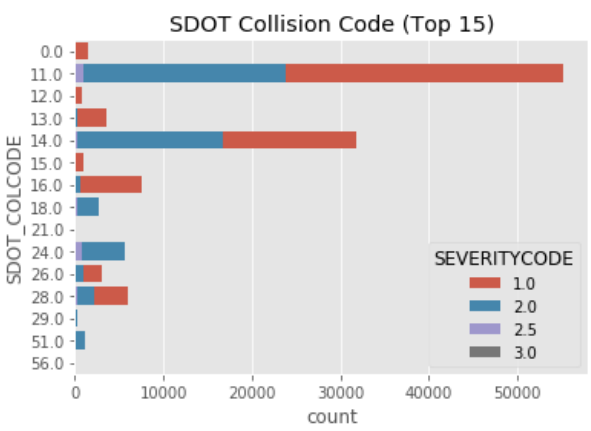


Chart 6: SDOT Collision Codes

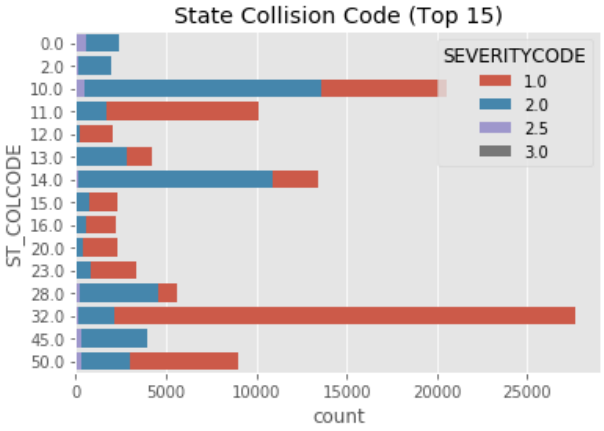


Chart 7: State Collision Codes

Most accidents were classified as 11 ("Struck Motor Vehicle front end (not head on)") or 14 ("Struck Motor Vehicle Read End"). The State collision codes were more descriptive, with the top 3 being 32 ("One Parked – One Moving"), 10 ("Entering at Angle") and 14 ("From Same Direction, Both Going Straight – One Stopped – Read End"). A full explanation of each collision code can be seen in the appendices for the dataset, available here. Finally, a depiction of Weather, Road Conditions, and Light Conditions at time of incident are displayed below:

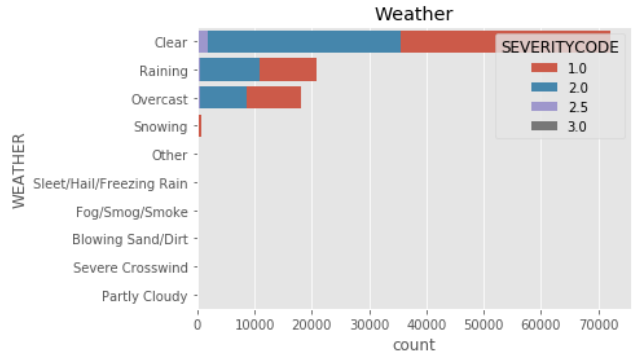


Chart 8: Weather Conditions

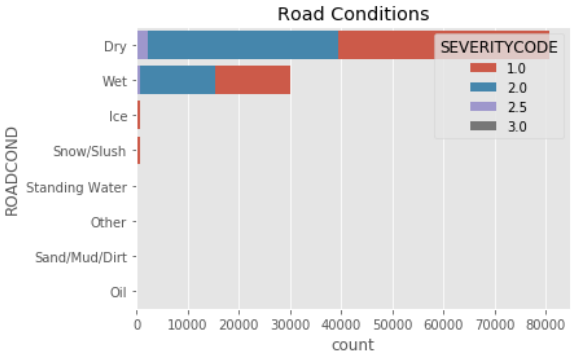


Chart 9: Road Conditions

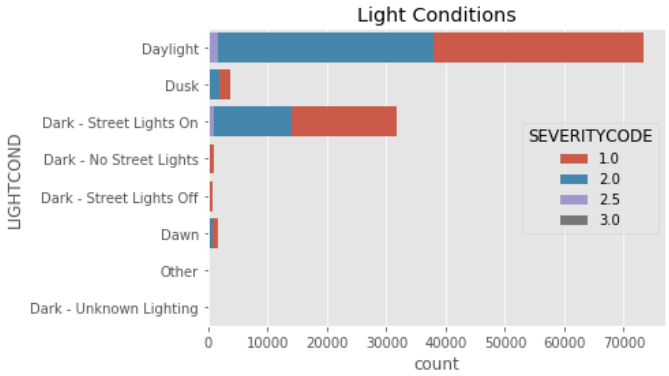
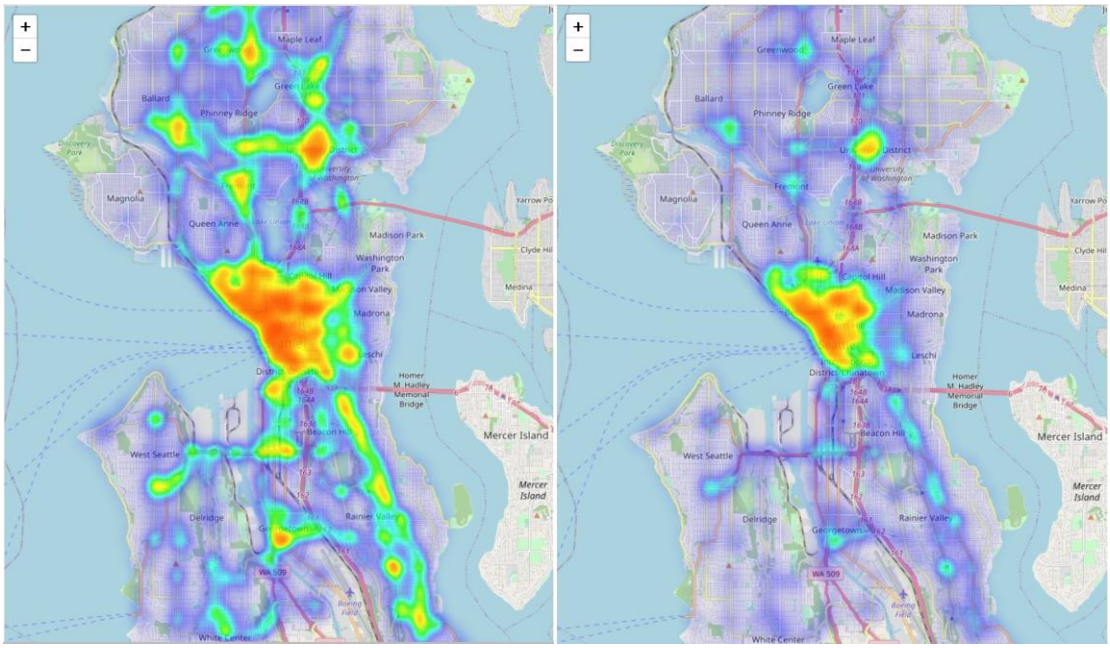
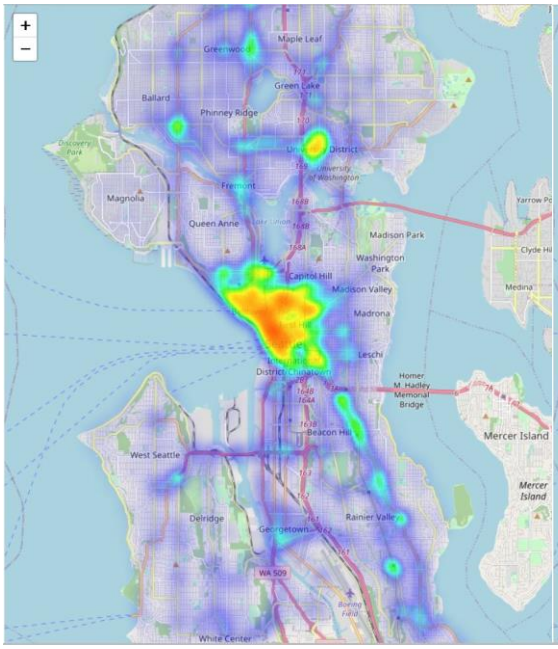


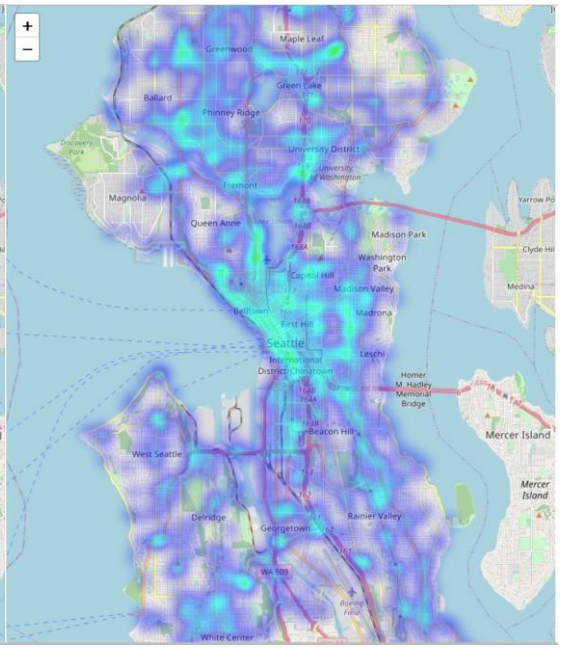
Chart 10: Light Conditions

Incident density was also visualized using the folium package. Heat maps of all incidents, as well as broken down by severity, can be seen below.



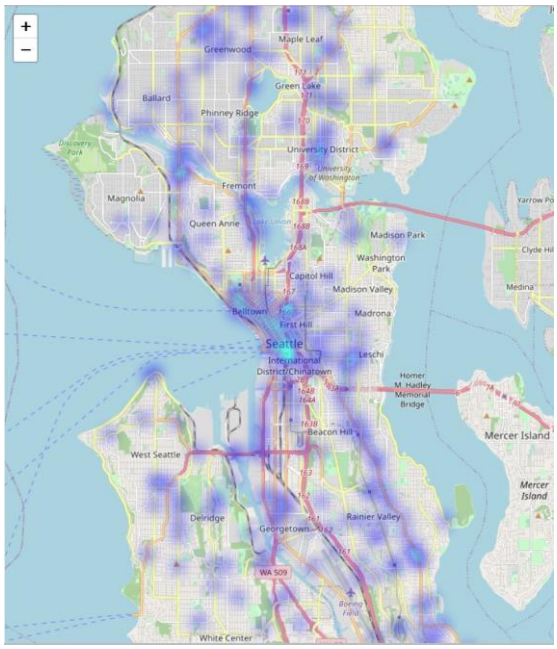
Map 1: All Severity Incidents (Top Left) *Map 2: Severity = 1 (Top Right)*





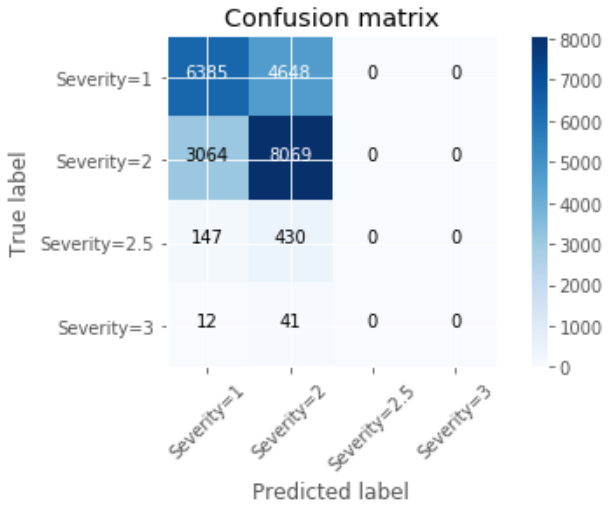
Map 3: Severity = 2 (Left) *Map 4: Severity = 2.5 (Right)*

Map 5: Severity = 3 (Below)



With the data properly visualized, the dataset was then balanced to prevent a biased Machine Learning model. Roughly 60,000 Severity Code = 1 incidents were removed via resampling in order to have an equal amount of Severity Code = 1 and Severity Code = 2 incidents. The unbalanced incident count was shown above in Table 4. For the Machine Learning

model, attributes were chosen which could easily and objectively be identified based on primary interaction from a 911 call. These attributes include “Collision Type”, “Address Type”, “Weather”, “Road Conditions”, “Light Conditions”. Weather conditions were grouped as “Visibility Unobstructed”, “Visibility Obstructed”, “Precipitation”, or “Other”. Road Conditions were grouped as “Dry – Road”, “Wet – Road”, “Wet – Slip”, “Dry – Slip”, or “Other”. Light Conditions were grouped as “Daylight”, “Dark”, “Dusk/Dawn”, or “Other”. These attributes were then converted to binary variables using one-hot encoding, then normalized, and then split into training data and test data with a split of 80% train & 20% test. Confusion Matrices for each model are shown below, with a summary following.



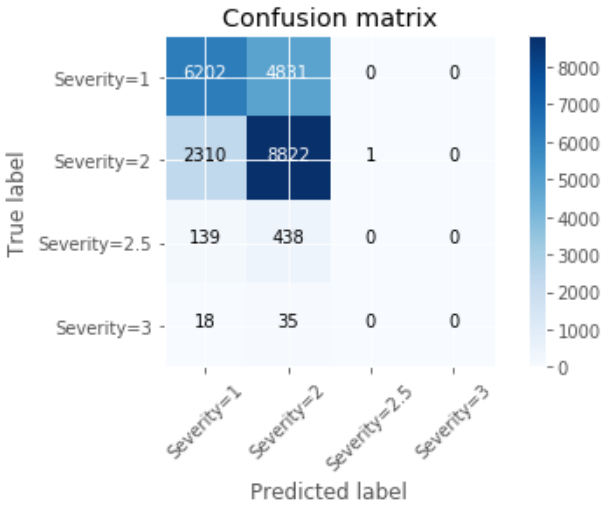


Chart 11: KNN Confusion Matrix (Top Left)

Chart 12: Decision Tree Confusion Matrix (Top Right)

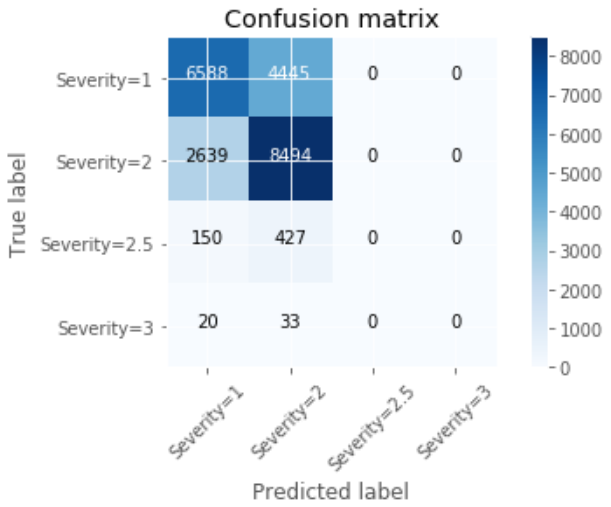


Chart 13: SVM Confusion Matrix (Bot Left)

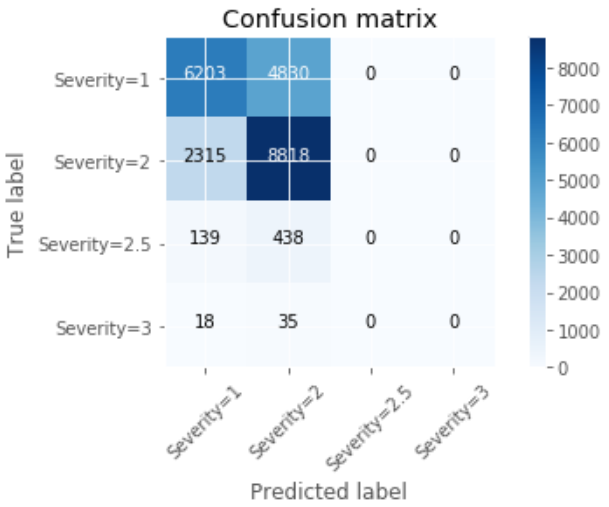


Chart 14: Logistic Regression Confusion Matrix (Bot Right)

Algorithm	Jaccard	F1-score	LogLoss
KNN	0.634	0.623	NA
Decision Tree	0.659	0.646	NA
SVM	0.662	0.65	NA
LogisticRegression	0.659	0.646	0.694

Table 5: Accuracy Scores

None of the higher severity cases were accurately predicted, likely due to the relatively low number of cases. Those incidents were removed, and the models were recreated.

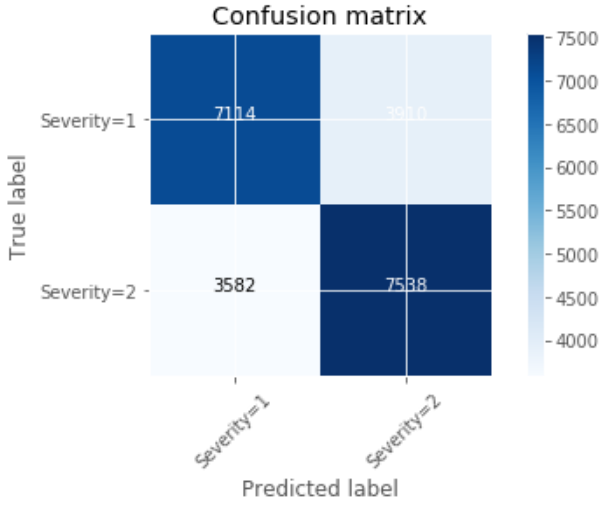


Chart 15: KNN Confusion Matrix (Top Left)

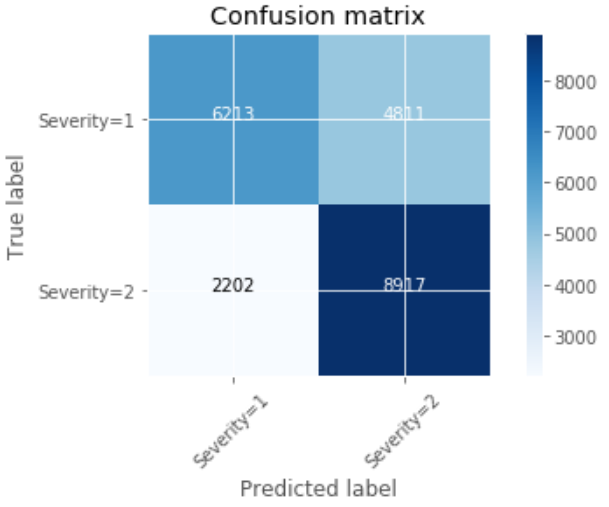


Chart 16: Decision Tree Confusion Matrix (Top Right)

Chart 17: SVM Confusion Matrix (Bot Left)

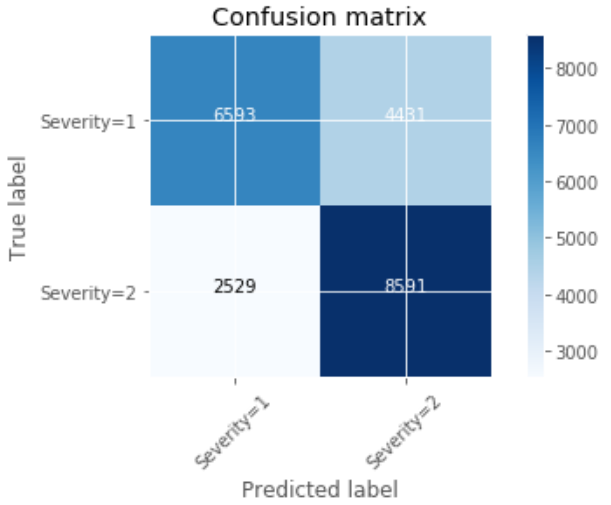
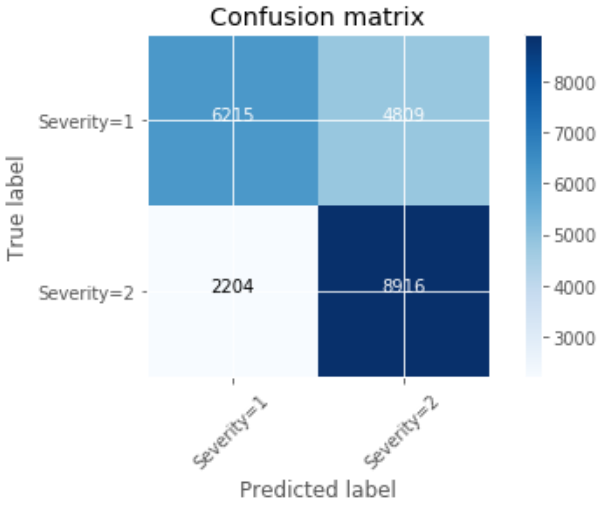


Chart 17: Logistic Regression Confusion Matrix (Bot Right)



Algorithm	Jaccard	F1-score	LogLoss
KNN	0.662	0.662	NA
Decision Tree	0.683	0.679	NA
SVM	0.686	0.683	NA
LogisticRegression	0.683	0.679	0.578

Table 6: Accuracy Scores

Discussion

Each of the 4 models that were created had a similar accuracy score, which is a good sign for the consistency of the model. However, it was determined during the original model creation that additional data was needed. None of the higher severity incidents (Severity = 2.5 or 3) we're correctly predicted by any of the models, which is a flaw in the model. This is most likely due to the low number of cases of higher severity, so the models could not be accurately trained on those cases. Even with down sampling the data, the model did not improve in these instances. To see if the model results would improve, those high severity incidents were removed, and the models were retrained with only low severity cases. We see that the accuracy of each model does improve, but it was not as large a jump as expected. With only 60-70%

accuracy, the models are a good starting point, however they are not perfect. It is possible that other data attributes can be used to improve the model, however that is further than the scope of this project. Other data attributes may not be easily determined by first responders, but this model is helpful as a first line effort.

Conclusion

This report analyzes a data set from Seattle, Washington which represents a comprehensive report of the traffic incidents for the last 16 years. The dataset is extremely thorough and highly valuable for many purposes but does lack information about higher severity traffic accidents (perhaps a good thing). The models created and trained would likely be improved if the dataset was resampled further, so that higher severity incidents took up a larger proportion of total accidents, however this runs the risk of reducing total data which could potentially pose more issues. The report generated was successful as a starting point, however more research could be done. It is likely that there is a strong correlation between location of the incident as well as traffic severity (as seen by hotspots on the heatmaps generated), but that is outside the scope of this project.