IBM Data Sciences Capstone Project
Predicting Collision Severity in Seattle

Prepared by Mark Snuffin

October 1, 2020

1 Introduction

1.1 Background

The Seattle Department of Transportation (SDOT) Technology Division is preparing their 2021 budget and is working on how to allocate funds to meet the technical demands made by the director. A particular sticking point is that the SDOT Director wants to provide a mobile application to the citizens of Seattle to help improve driver safety, but it needs to be more than just general information. Instead it needs to provide actionable intelligence to help the public be safer on the roads, better yet indicate predict when a trip could be risky. The director understands this is ambitious but wants the Technology Division to innovate and demands that the team consider what is possible and get it into the budget.

The technology division knows how to build a mobile application but is not sure if they can create a capability that could perform this type of prediction therefore the team is hesitant to commit to this effort considering the risk of potentially wasting city resources. Time is of the essence though. They only have a limited amount of time to consider their options and they do not want to fall short of meeting the director's demands.

1.2 Problem

After considering their options the Technology Division Team decides they need a case study to determine what can be implemented within the resources they have allocated in their proposed budget. They believe leveraging Data Sciences and Machine Learning could be the foundation of a prediction service and decide the study must focus on predicting collision severity. **The problem to be answered by the study is given the currently available SDOT data is it possible to accurately predict collisions with an injury.** If the outcome of the study yields positive results they can move ahead with confidence and offer a predictive analytic to stakeholders in the new year.

2 Data Understanding

2.1 Data Sources

The study uses data from SDOT and is referred to as collisions (not accidents). The data is hosted by SDOT GIS Division and curated by the SDOT Traffic Division where collisions are collected from Seattle Police Department after a collision is reported. Instead of using the full data set this study uses a subset of the data provided by Coursera.

General characteristics of the subset -

- Format: CSV
- Timeframe: January 2004 to May 2020
- Columns: 38, 37 are unique that describe a collision
- Rows: 194,673 – each row represents a collision in Seattle
- Bounding Coordinates -- West Bounding Coordinate: -122.4754 -- East Bounding Coordinate: -122.2008 -- North Bounding Coordinate: 47.7582 -- South Bounding Coordinate: 47.4814

The list below represents the attributes to be considered as independent variables for modelling. Definitions and data types are defined in the SDOT Attribute Dictionary. The dependent variable is collision severity, defined as SEVERITYCODE. There are 37 possible independent variables.

Attribute details - https://www.seattle.gov/Documents/Departments/SDOT/GIS/Collisions_OD.pdf

Original Data Set - https://data-seattlecitygis.opendata.arcgis.com/datasets/5b5c745e0f1f48e7a53acec63a0022ab_0?geometry=-123.310%2C47.452%2C-121.352%2C47.776

Coursera Subset - https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv

ESRI Metadata
- https://www.arcgis.com/sharing/rest/content/items/5b5c745e0f1f48e7a53acec63a0022ab/info/metadata/metadata.xml?format=default&output=html

2.2 Data Source Attributes

- Severity
    - **SEVERITYCODE** - Text,100 - code that corresponds to the severity, e.g., 3—fatality, 2b—serious injury, 2—injury, 1—property damage, 0—unknown
    - **SEVERITYDESC** - Text,300 - Detailed description of the severity, e.g., Property Damage Only Collision, Injury Collision
- Location, Time
    - **INCDATE** - Date - date of the incident, ex: 2013/03/27 00:00:00+00
    - **INCDTTM** - Text,30 - date and time of the incident, ex: 3/27/2013 2:54:00 PM
    - **X** - float - longitude of the collision
    - **Y** - float - latitude of the collision
    - **LOCATION** - Text,255 - description of the general location of the collision, e.g. 5TH AVE NE AND NE 103RD ST
    - **ADDRTYPE** - Text,12 - collision address type, e.g, alley, block, intersection
    - **JUNCTIONTYPE** - Text,300 - category of junction where the collision occurred. e.g., At Intersection (intersection related)
    - **INTKEY** - Double - key corresponding to the intersection associated with a collision
    - **CROSSWALKKEY** - Long - key for the crosswalk at which the collision occurred
- Environmental Conditions
    - **WEATHER** - Text,300 - weather condition, e.g., Clear, Overcast, Raining, Snowing
    - **ROADCOND** - Text,300 - road condition, e.g., Wet, Dry
    - **LIGHTCOND** - Text,300 - light condition, e.g., Daylight, Dark, Dark - Street Lights On
- Counts

    - **VEHCOUNT** - Double - number of vehicles involved in the collision
    - **PERSONCOUNT** - Double - total number of people involved in the collision
    - **PEDCOUNT** - Double - number of pedestrians involved in the collision
    - **PEDCYLCOUNT** - Double - number of bicycles involved in the collision
- Human Factors
    - **INATTENTIONIND** - Text,1 - whether or not collision was due to inattention (Y/N)
    - **UNDERINFL** - Text,1 - whether or not a driver involved was under the influence of drugs or alcohol
- Collision Descriptors
    - **COLLISIONTYPE** - Text,300 - collision type, e.g., Angle, Sideswipe, Parked Car
    - **SPEEDING** - Text,1 - whether or not speeding was a factor in the collision (Y/N)
    - **SEGLANEKEY** - Long - key for the lane segment in which the collision occurred
    - **HITPARKEDCAR** - Text,1 - whether or not the collision involved hitting a parked car (Y/N)
    - **PEDROWNOTGRNT** - Text,1 - whether or not the pedestrian right of way was not granted (Y/N)
    - **SDOTCOLNUM** - Text, 10 - number given to the collision by SDOT
    - **SDOT_COLCODE** - Text,10 - code assigned to the collision by SDOT
    - **SDOT_COLDESC** - Text,300 - description corresponding to the collision code
    - **ST_COLCODE** - Text,10 - code provided by the state that describes the collision
    - **ST_COLDESC** - Text,300 - description that corresponds to the state's coding designation
- Miscellaneous
    - **OBJECTID** - Double - ESRI unique identifier
    - **INCKEY** - Long - A unique key for the incident
    - **COLDETKEY** - Long - Secondary key for the incident
    - **EXCEPTRSNCODE** - Text,10 - undefined
    - **EXCEPTRSNDESC** - Text,300 - A unique key for the incident
    - **REPORTNO** - Long - undefined
    - **STATUS** - Text - undefined

2.3 Data Preparation

Data preparation was straightforward with the exception of missing values and nulls. The most significant issue is that the dataset is missing values that are not trivial to replace. A significant effort needs to be undertaken to fix everything, for example comparing SDOT Collision Codes with Washington State Collision Codes, therefore data was cleaned with a limited application of business knowledge and primarily done with bulk changes in the notebook.

Most features with data issues were relatively easy to remedy, for example SPEEDING is a Yes/No column that was missing No values. However special encoding was used to repair categorical features that were missing values. Features include COLLISIONTYPE, WEATHER, ROADCOND, LIGHTCOND, ADDRTYPE and JUNCTIONTYPE. Check the features section below for specifics.

One Hot Encoding was used instead of Label Encoding. Any specific rules applied are listed in the sections below.

2.4 Dataset Balancing

The dataset contains 58,188 Injury Collisions and 136,485 Property Collisions. If left unchecked it will lead to biased model. For this study random under-sampling (RUS) was used to resolve the imbalance. RUS randomly deletes samples from the majority class until the number of samples matches the minority class. The downside of RUS is that it eliminates data from the majority class however this dataset is large therefore the effects of down-sampling were minor.

2.5 Feature Selection

After data cleaning, there are 194,673 samples and 31 features in the data. Upon examining the meaning of each feature, it was clear that there was redundancy in the features. For example, SDOT_COLCODE and ST_COLCODE are basically the same but have different coding schemes.

After discarding redundant features, the independent variables were inspected via a correlation matrix. Unfortunately, there were no highly correlated variable (Pearson correlation coefficient > 0.9) therefore further analysis was needed to find the right combination of variables to predict Injury Collisions.

After analysis the following are the remaining features used for modelling:

1. **SEVERITYCODE**
2. **ADDRTYPE**
3. **COLLISIONTYPE**
4. **PERSONCOUNT**
5. **VEHCOUNT**
6. **JUNCTIONTYPE**
7. **SDOT_COLCODE**
8. **INATTENTIONIND**
9. **UNDERINFL**
10. **WEATHER**
11. **ROADCOND**
12. **LIGHTCOND**
13. **PEDCOUNT**

14. **PEDROWNOTGRNT**
15. **SPEEDING**
16. **HITPARKEDCAR**

Note that there were several independent variables that offered a similar Pearson Correlation Coefficient and could have been included in the model. For completeness these variables are discussed below along with the feature list selected.

3 Methodology

Methodology section represents the main component of this report that discusses and describes all exploratory data analysis performed, any inferential statistical testing that performed, and what machine learnings were used and why.

3.1 Exploratory Data Analysis

3.1.1 Target Variable – Collision Severity

Collision Severity, or Severity, is used as the target/dependent variable to be predicted by the models. SDOT uses four types of Severity however this study uses two: Type 1 – Property Damage Collision (Property) and Type 2 – Injury Collision (Injury). The majority of collisions are Property (136,485), and the remaining are Injury (58,188). This is an expected distribution but will require balancing during the modelling stage.

Figure 1. charts the collisions represented in the data set. The data begins in 2004 and ends in 2020 where May 2020 is the newest period in the data set. Collisions in general have been on the decline however Injury Collisions have had a lesser decline and are still high as compared to Property Collisions with a total of 3,062 in 2019.

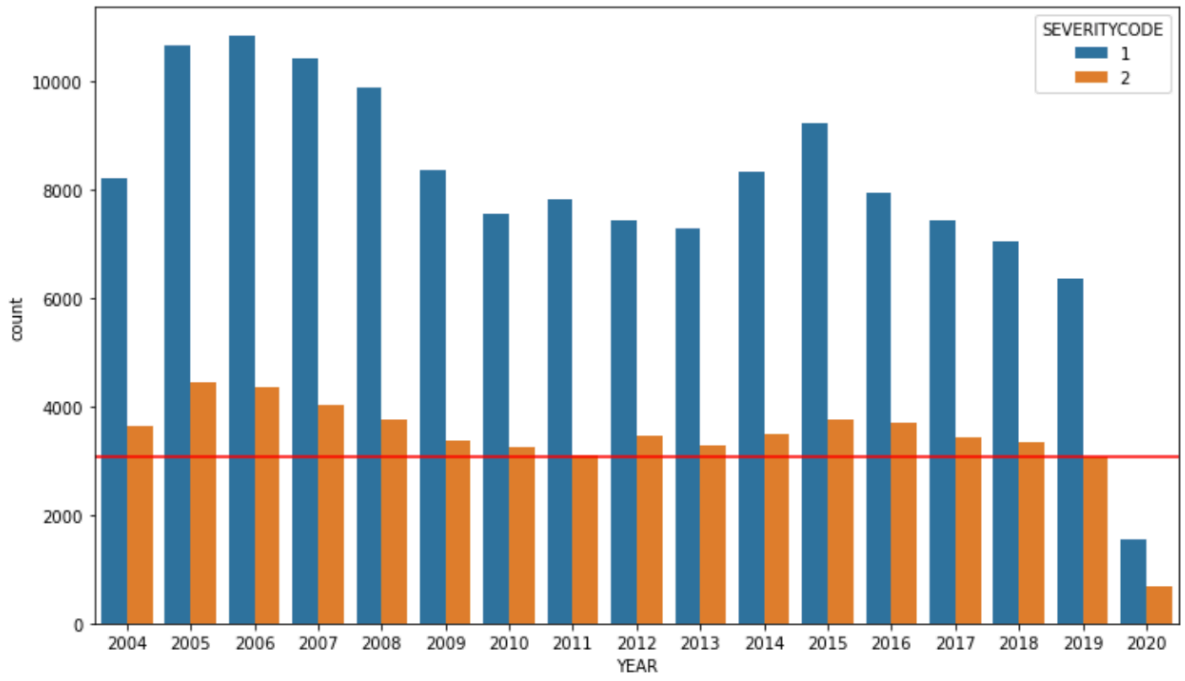


Figure 1 - Collisions by Year

3.1.2 Temporal Features and Severity

The collision date was analysed to determine if the date and time of the collision had any impact on the severity of the accident. Figure 2 displayed a snapshot of what was used for this analysis.

	SEVERITYCODE	INCDTTM	MONTH	WEEKDAY	HOUR	PEAKOROFFPEAK	YEARWEEK
0	2	2013-03-27 14:54:00	3	2	14	1	13
1	1	2006-12-20 18:55:00	12	2	18	1	51
2	1	2004-11-18 10:20:00	11	3	10	1	47
3	1	2013-03-29 09:26:00	3	4	9	1	13
4	2	2004-01-28 08:04:00	1	2	8	1	5

Figure 2 - Collision Date and Time Permutations

Figure 3 shows the correlation between severity, month, weekday, hour, week of the year, and peak or off peak. Peak and Off Peak was engineered to possibly spot a trend in however this did not have a strong correlation, therefore it was not included. Figure 4 shows four charts that break down each of the date parts evaluated.

	SEVERITYCODE	MONTH	WEEKDAY	HOUR	PEAKOROFFPEAK	YEARWEEK
SEVERITYCODE	1.00	0.00	-0.02	0.03	0.04	0.01
MONTH	0.00	1.00	-0.00	0.01	0.00	0.98
WEEKDAY	-0.02	-0.00	1.00	-0.02	-0.08	-0.00
HOUR	0.03	0.01	-0.02	1.00	0.36	0.03
PEAKOROFFPEAK	0.04	0.00	-0.08	0.36	1.00	0.02
YEARWEEK	0.01	0.98	-0.00	0.03	0.02	1.00

Figure 3 - Date Parts Related to Severity

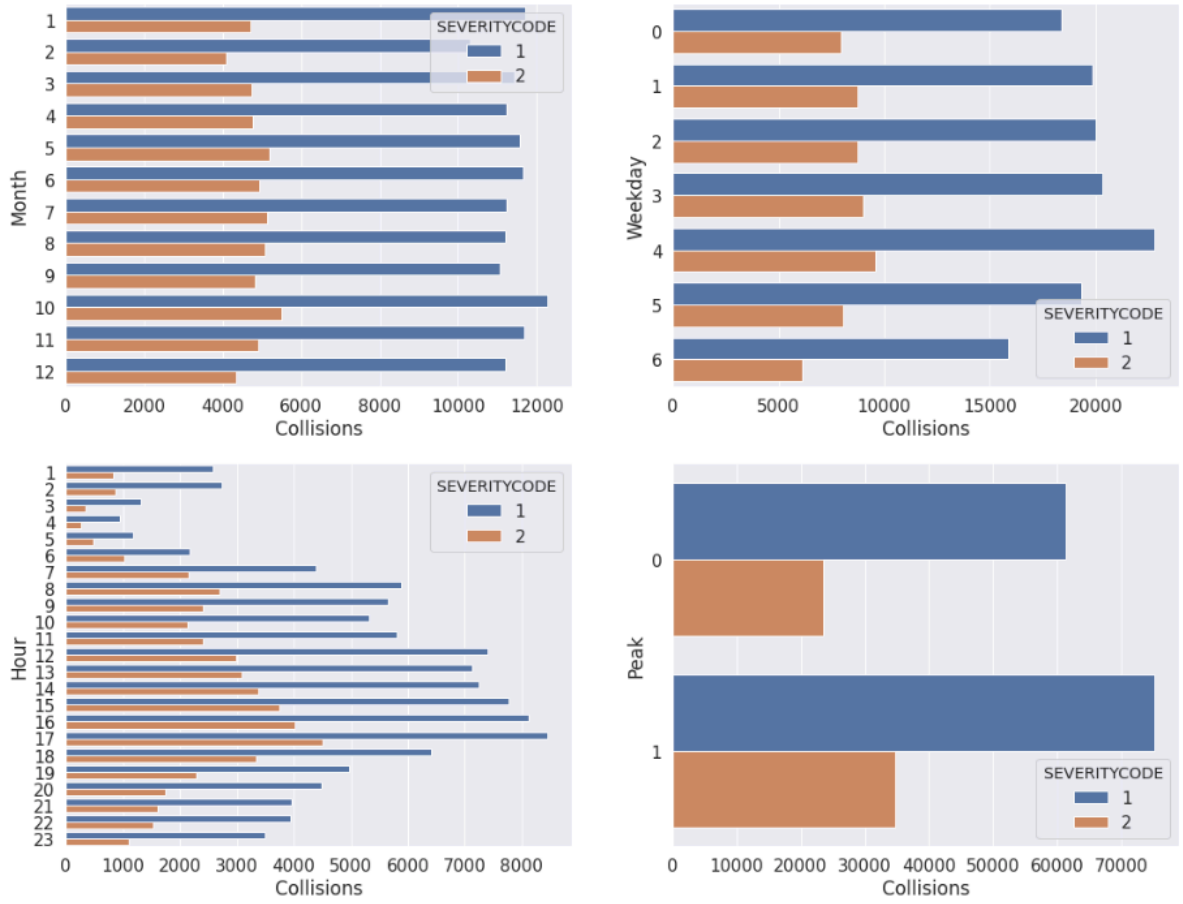


Figure 4 - Date Related to Severity

3.1.3 Location Features and Severity

The dataset has Latitide and Longtitude defined as the X and Y attributes. These attributes do not show a strong relation to severity based on the correlation matrix shown in Figure 5, and although it's clear there are collision hotspots throughout the region. While this data is compelling to use as features it was excluded for this iteration of the study.

	SEVERITYCODE	X	Y
SEVERITYCODE	1.00	0.01	0.02
X	0.01	1.00	-0.16
Y	0.02	-0.16	1.00

Figure 5 - Geocoordinates Related to Severity

The figures below chart collisions by severity type on Seattle. Figure 6 displays all collisions, Figure 7 displays Injury Collisions, and Figure 8 is for Property. Further examination shows that there numerous hotspots there is not clear evidence to sway the model to predict an Injury Collision.

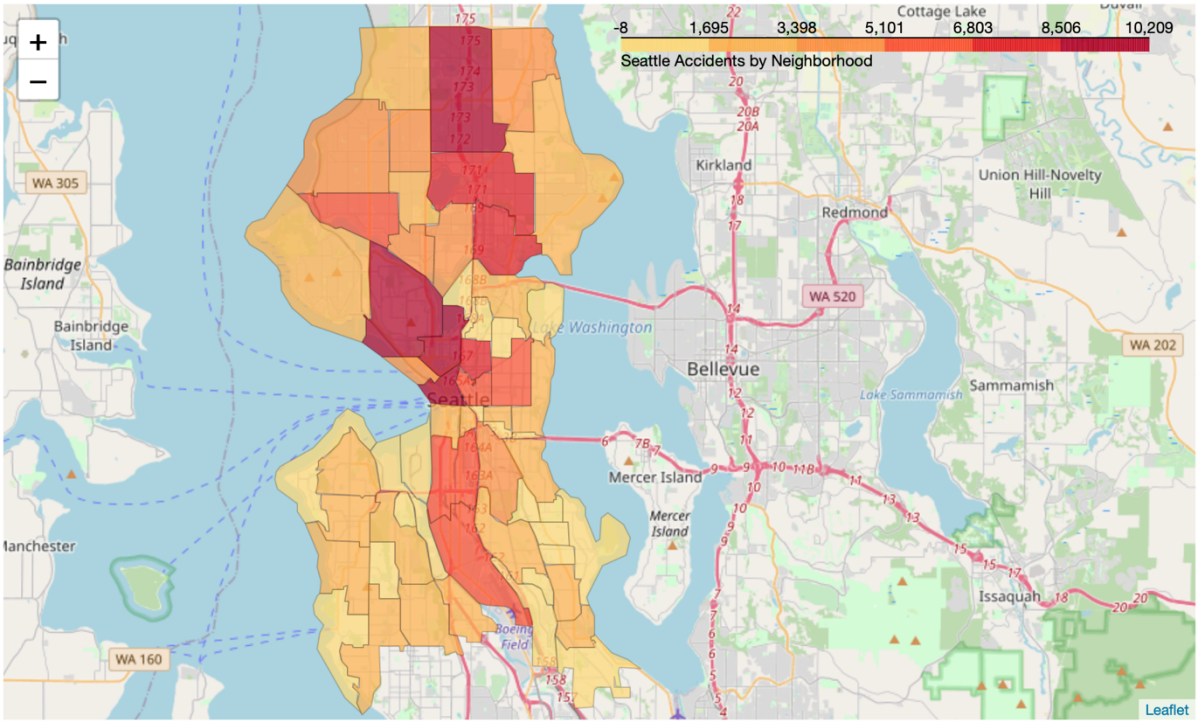


Figure 6 – Map with All Collisions in the Data Set

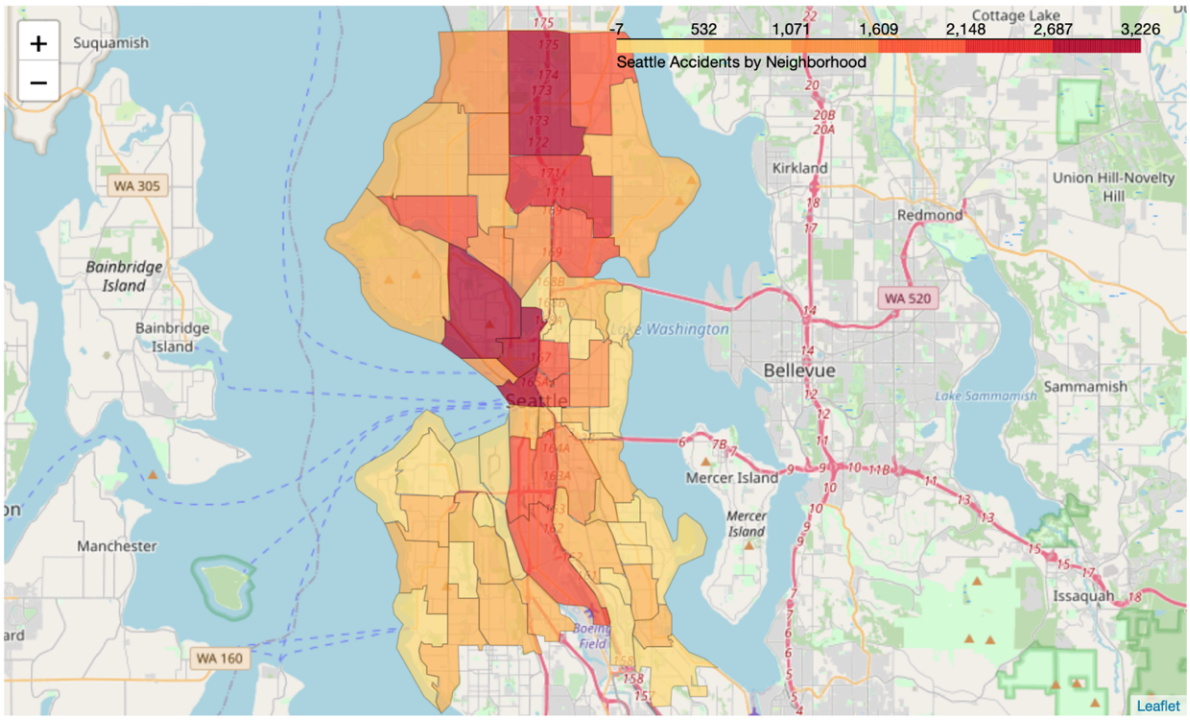


Figure 7 – Map with Injury Collisions in the Data Set

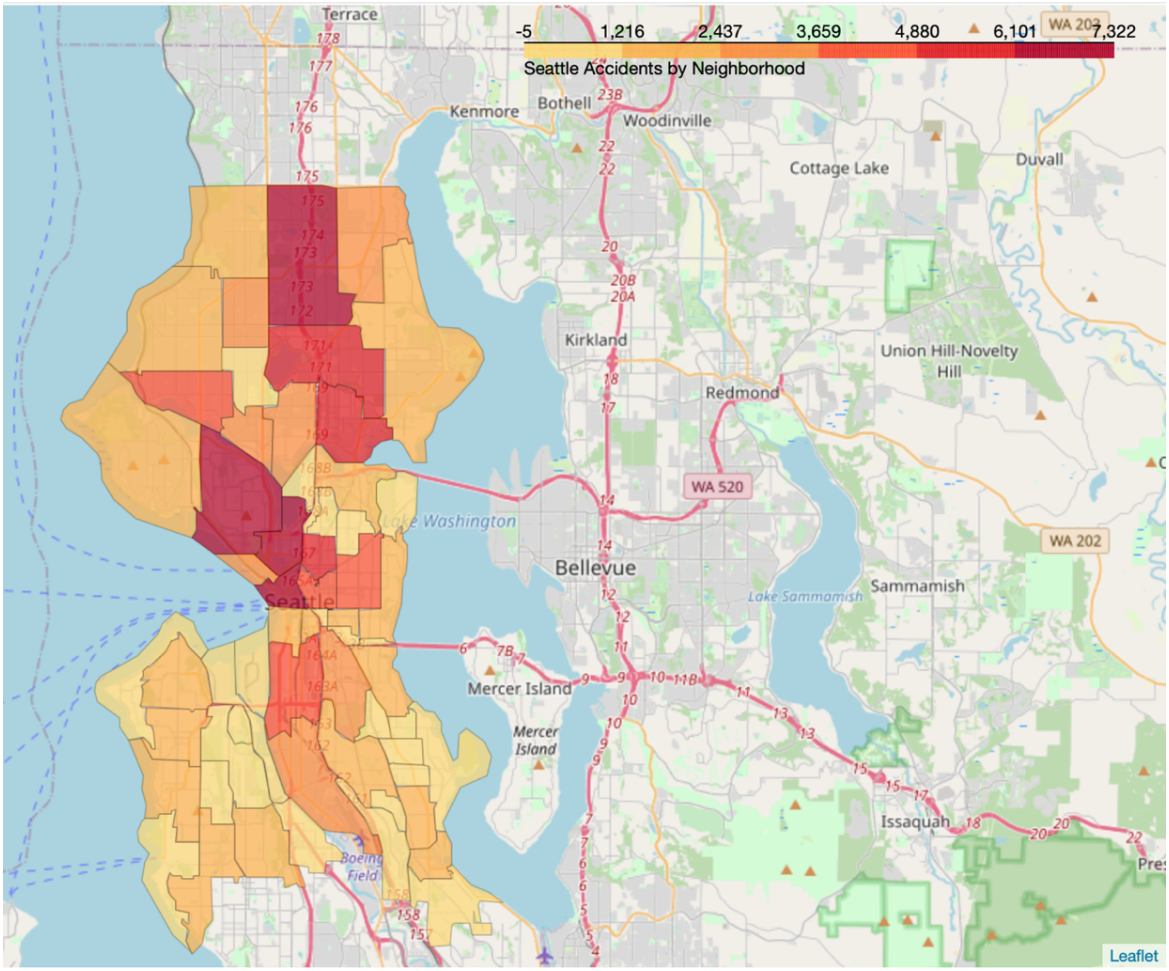


Figure 8 - Map Displaying Property Severity Collisions

At a higher level the dataset includes Junction Type and Address Type. As shown in Figure 9 there is solid correlation between severity and these features. There is some overlap though where ADDRTYPE and JUNCTIONTYPE both have an Intersection Type that appear similar but were kept for modelling. The other types did not have a strong correlation.

Figure 10 charts the ADDRTYPE Feature where Alley has the least significance and was mapped to Block. Figure 11 charts JUNCTIONTYPE where Mid-Block and Intersection have the greatest relation to severity. There is a lesser type of Intersection, "At Intersection (but not related to intersection)" that was excluded as a feature category. Ramp Junction and Unknown categories were removed.

	SEVERITYCODE
SEVERITYCODE	1.00
ADDRTYPE_Alley	-0.03
ADDRTYPE_Block	-0.18
ADDRTYPE_Intersection	0.20
JUNCTIONTYPE_At Intersection (but not related to intersection)	-0.00
JUNCTIONTYPE_At Intersection (intersection related)	0.20
JUNCTIONTYPE_Driveway Junction	0.00
JUNCTIONTYPE_Mid-Block (but intersection related)	0.02
JUNCTIONTYPE_Mid-Block (not related to intersection)	-0.17
JUNCTIONTYPE_Ramp Junction	0.00
JUNCTIONTYPE_Unknown	-0.00

Figure 9 - Correlation Matrix for ADDRTYPE and JUNCTIONTYPE

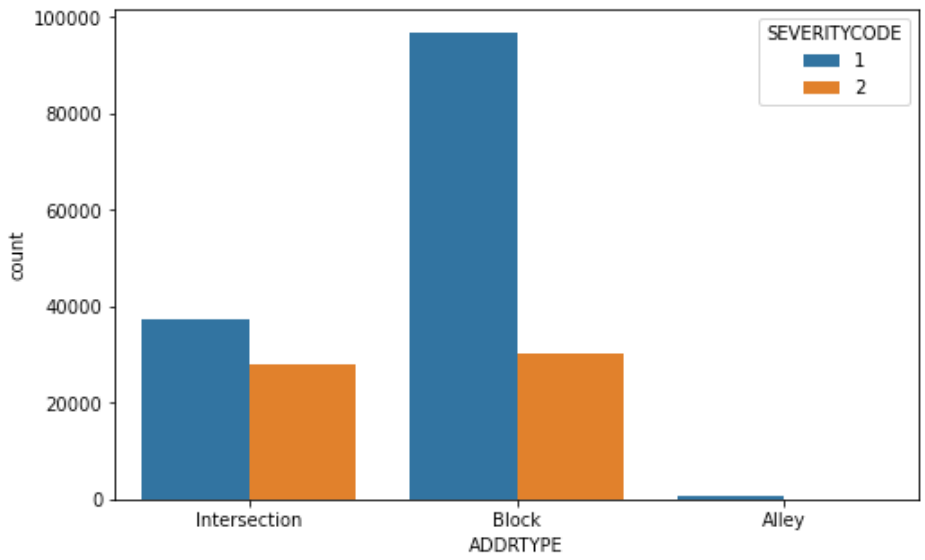


Figure 10 - Charting General Location of a Collision

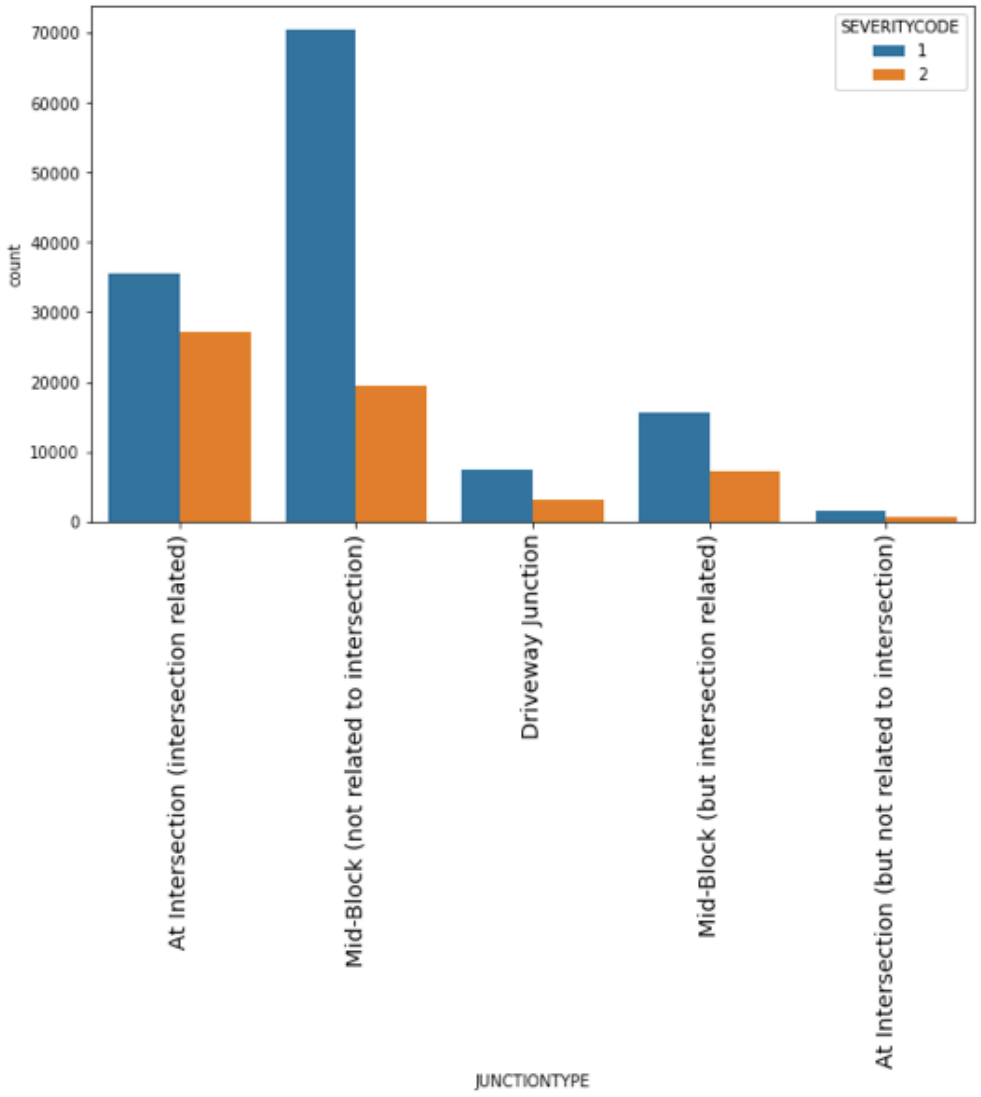


Figure 11 - Charting Type of Junction Where a Collision Occurred

3.1.4 Human Factor Features and Severity

Driving while intoxicated and distracted driving account for approximately 35% of Injury Collisions. There is substantially more collisions where the driver was distracted versus intoxicated yet both a similar correlation severity as indicated in Figure 12. Charts in Figure 13 show how this is represented in the dataset.

	SEVERITYCODE	INATTENTIONIND	UNDERINFL
SEVERITYCODE	1.00	0.05	0.04
INATTENTIONIND	0.05	1.00	-0.03
UNDERINFL	0.04	-0.03	1.00

Figure 12 - Correlation of Intoxicated and Distracted Drivers with a Collision

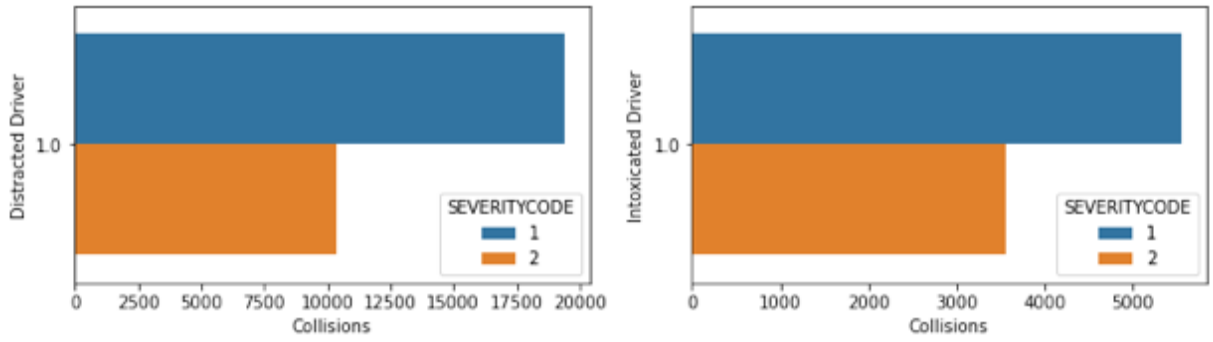


Figure 13 - Human Factors Affecting Severity

3.1.5 Count Features and Severity

Person, vehicle, pedestrian and cyclist counts have a varying degree of correlation with severity as shown in Figure 14. Collisions with a high number of vehicles, people, pedestrians, bicyclists are rare and will result in outliers in modelling. The charts in the Figure 15 show how the counts relate to collision severity. The most significant exclusion was Pedestrian Count that had a high Pearson Correlation Coefficient however was only present in a small number of collisions. This could be revisited in a future iteration to improve model performance.

	SEVERITYCODE	VEHCOUNT	PERSONCOUNT	PEDCOUNT	PEDCYLCOUNT
SEVERITYCODE	1.00	-0.05	0.13	0.25	0.21
VEHCOUNT	-0.05	1.00	0.38	-0.26	-0.25
PERSONCOUNT	0.13	0.38	1.00	-0.02	-0.04
PEDCOUNT	0.25	-0.26	-0.02	1.00	-0.02
PEDCYLCOUNT	0.21	-0.25	-0.04	-0.02	1.00

Figure 14 - Counts Correlation with Severity

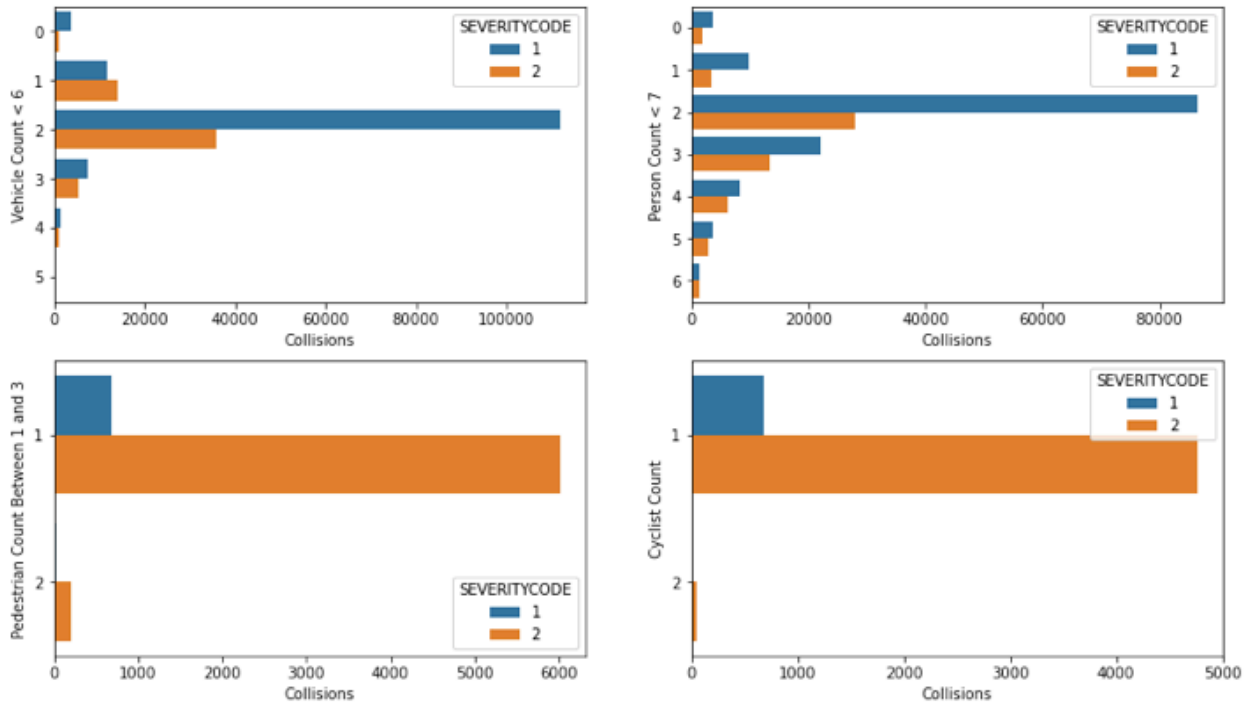


Figure 15 - Vehicle, Person, Pedestrian and Bicyclist Collisions

The charts indicate that when there is greater than two people involved in a collision, the likelihood of an Injury Collision increases. A collision involving any number of pedestrians or pedalcyclists has at least 87% chance to result in an Injury Collision but this is a small number of occurances.

3.1.6 Environmental Features and Severity

Inclement weather often contributes to collisions but the majority of collisions for both severity types overwhelmingly occur when the weather is clear. Rain is the most hazardous condition in the data set followed by overcast or cloudy skies which indicates lighting conditions could be a factor. A future consideration is to have only two weather conditions, e.g., Clear or Inclement, which might make the model more general.

Similarly, dry road conditions resulted in the most collisions of both types. Wet conditions attributed for the most after that and the rest do not have any impact to the severity. It's possible that it's difficult for the police to determine these conditions and perhaps an alternate source of weather data is needed to provide an additional support to the collision details. This could be considered in the future and could help solidify weather and road conditions.

Lighting conditions clearly show that the majority of collisions occur in daylight hours. There is a not a strong relation here but worth including in the model.

Figures 16 -18 chart these conditions and how each are related to severity.

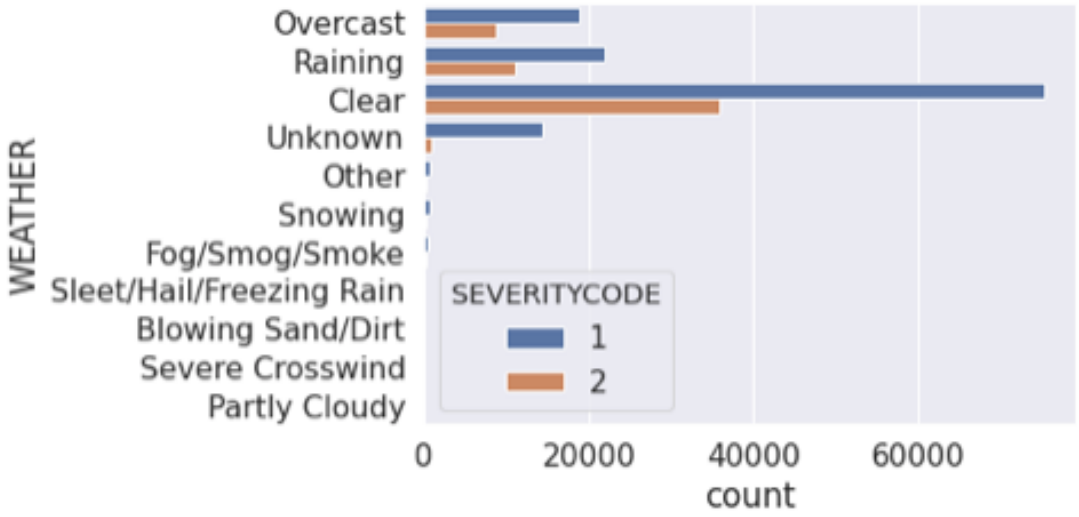


Figure 16 - Weather Conditions When a Collisions Occurred

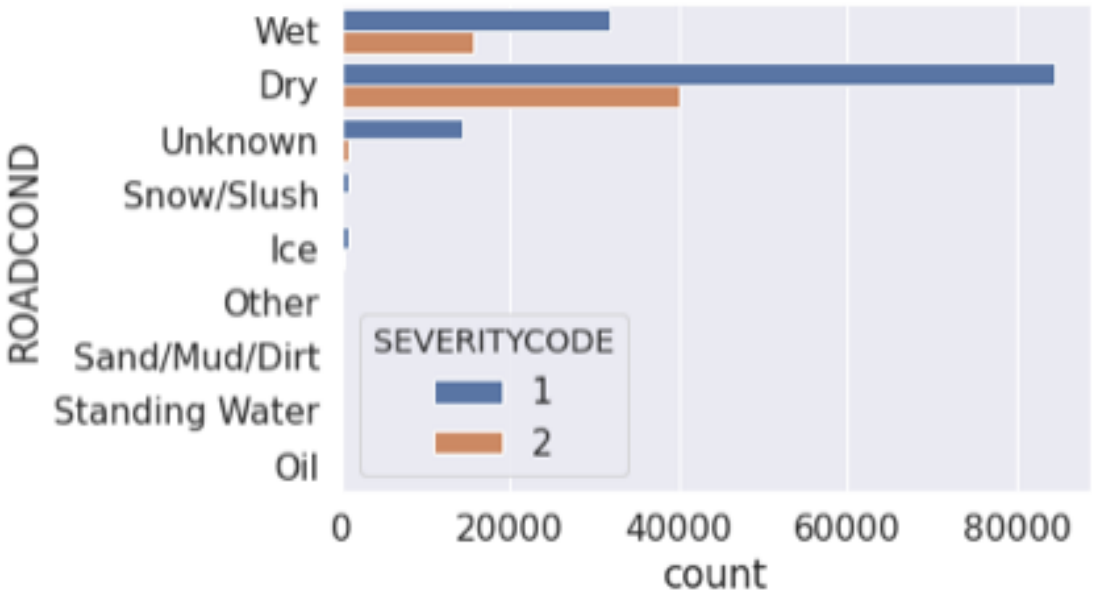


Figure 17 - Road Conditions When Collisions Occurred

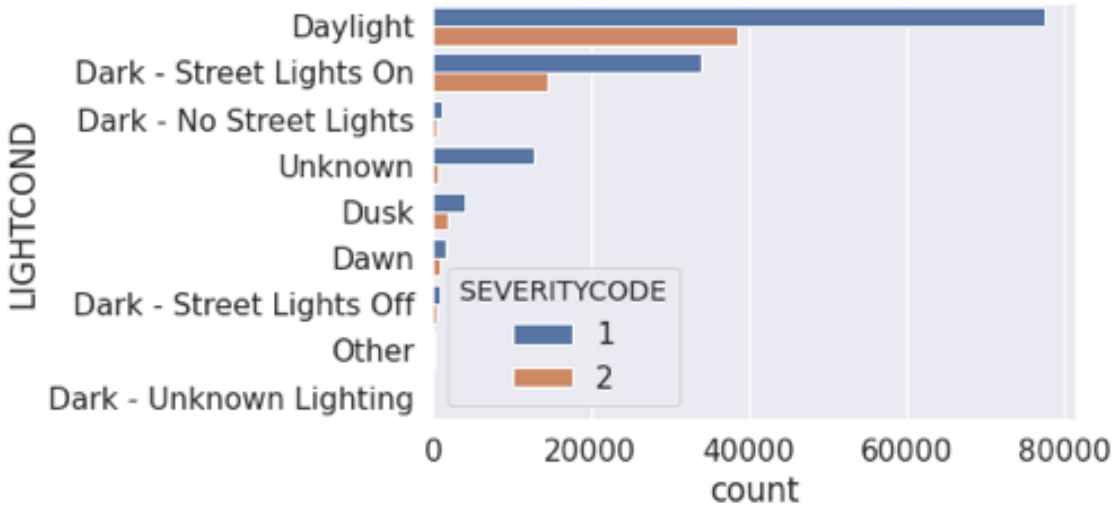


Figure 18 - Lighting Conditions When a Collisions Occurred

3.1.7 Collision Descriptor Features and Severity

The charts below in Figures 19 – 22 show the collision descriptors related to severity. Two of the stronger relations are "Pedestrian Not Granted Right of Way" and "Parked Car". SDOT Collision is difficult to interpret but clearly is related to severity, for the most part it is related to Property Collisions.

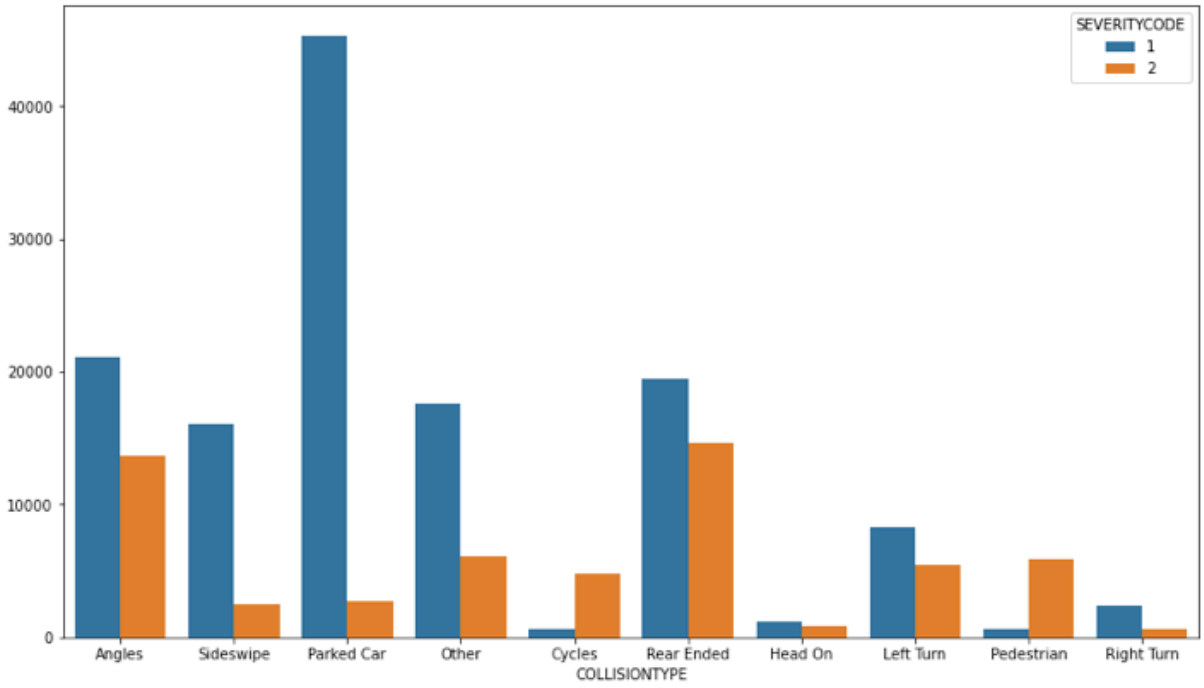


Figure 19 - Collision Type Related to Severity

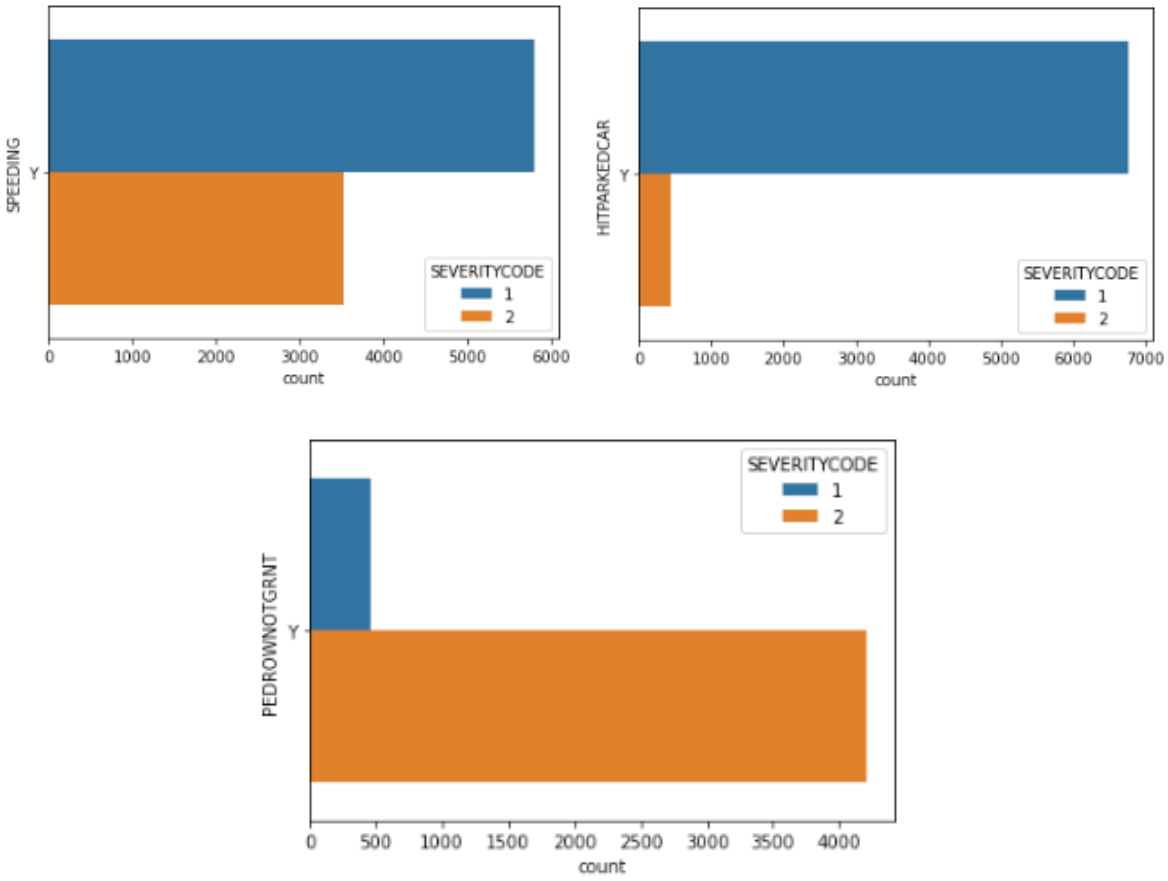


Figure 20 - Speeding, Hitting Parked Car, Pedestrian Not Granted Right of Way

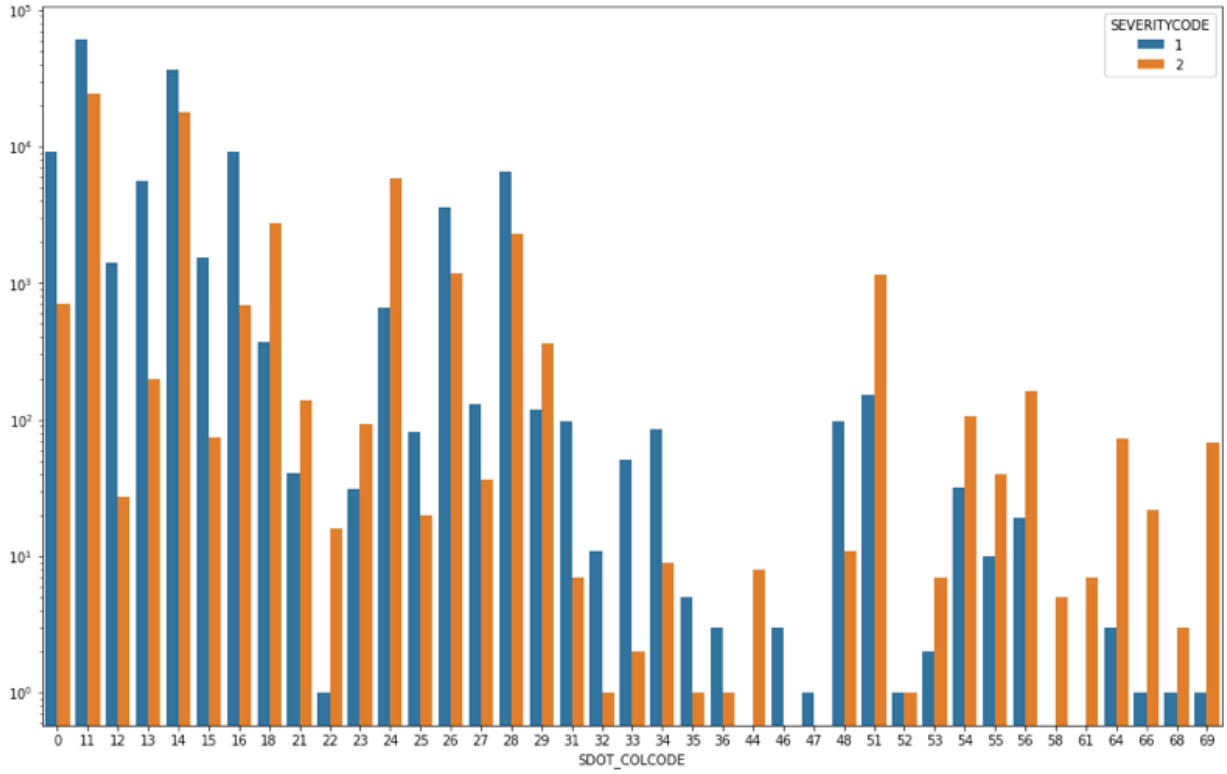


Figure 21 - SDOT Collision Code Related to Severity

3.1.8 Features Wrap-Up

Figure 22 shows the final correlation matrix for the features selected for modelling.

	SEVERITYCODE	ADDRTYPE	COLLISIONTYPE	PERSONCOUNT	VEHCOUNT	SDOT_COLCODE	INATTENTIONIND	UNDERINFL	PEDROWNOTGRNT	SPEEDING	HITPARKEDCAR
SEVERITYCODE	1.00	0.18	-0.13	0.13	-0.05	0.19	0.05	0.04	0.21	0.04	-0.10
ADDRTYPE	0.18	1.00	-0.46	0.06	-0.08	-0.07	-0.08	-0.04	0.15	-0.06	-0.12
COLLISIONTYPE	-0.13	-0.46	1.00	0.02	0.11	-0.00	0.12	0.00	-0.02	-0.00	0.03
PERSONCOUNT	0.13	0.06	0.02	1.00	0.38	-0.13	0.08	0.02	-0.03	-0.00	-0.05
VEHCOUNT	-0.05	-0.08	0.11	0.38	1.00	-0.37	0.08	0.01	-0.23	-0.03	0.05
SDOT_COLCODE	0.19	-0.07	-0.00	-0.13	-0.37	1.00	0.03	0.11	0.24	0.14	-0.10
INATTENTIONIND	0.05	-0.08	0.12	0.08	0.08	0.03	1.00	-0.03	-0.03	-0.05	0.01
UNDERINFL	0.04	-0.04	0.00	0.02	0.01	0.11	-0.03	1.00	-0.02	0.09	0.01
PEDROWNOTGRNT	0.21	0.15	-0.02	-0.03	-0.23	0.24	-0.03	-0.02	1.00	-0.03	-0.03
SPEEDING	0.04	-0.06	-0.00	-0.00	-0.03	0.14	-0.05	0.09	-0.03	1.00	-0.03
HITPARKEDCAR	-0.10	-0.12	0.03	-0.05	0.05	-0.10	0.01	0.01	-0.03	-0.03	1.00

Figure 22 - Correlation Matrix of Features

3.2 Modelling

The problem we are solving is to predict Injury Collisions with the highest degrees of accuracy. This is a binary classification problem thus several classification models are used with the addition of Logistic Regression. For each model Grid Search was used to perform cross validation and hyperparameter tuning. This step was critical to determine the right mix of parameters to most accurately predict Injury Collisions.

A number of scoring metrics where used to evaluate each model including a general accuracy score, Jaccard Score, F1 Score and Log Loss. This is further supported by Classification Reporting, Confusion Matrices and Receiver Operating Curve (ROC). As part of the Confusion Matrix True Positive and Negatives and False Positive and Negatives were used to further evaluation each model. The evaluation section below discusses the above metrics in greater detail.

In preparation for running each model the dataset was scaled used the Standard Scaler with no special options. The dataset was divided into test and training sets using 80% of the data for training and 20% for testing. The number of samples were 93,100 and 23,276 respectively. When creating the training and testing datasets Stratification was applied with a Random State set to ensure repeatable results.

In machine learning, classification is considered supervised learning. Supervised learning is the machine learning task of learning a function that maps an input to an output based on example input-output pairs. It infers a function from labeled training data consisting of a set of training examples.

Since there are only two classes of collision severity this study utilizes binary classification models. The assumption is that the permutation or combination of all independent features in the dataset will have recurring patterns that predict the classes. The classification algorithm will find the common pattern of combinations that correspond to either one of the dependent classes.

The following classification models were selected based on each model's ability to do well with binary classification and constraints within the dataset:

- XGBoost (XGB)
- Gradient Boosting Trees (GBT)
- Random Forest (RF)
- Decision Tree (DT)
- K-Nearest Neighbors (KNN) – results not reported due to poor performance
- Logistic Regression (LR)

4 Results

Figure 23 presents the modelling results with the best performance in each category highlighted in green. On the following page is a gallery of Confusion Matrices, one for each model, charted in Figure 24. Figure 25 displayed ROC chart for all models.

The results represent the best run with optimized hyperparameters. This study considered the number of True Positives as the deciding factor to determine which model was the best. Based on these metrics the best is the XGB which had solid accuracy statistics and a very good True Positive Score over 95%. The worst performing in terms of TP was RF although it had the best F1 Score.

For each model the top ten features where collected ranked by Gini Importance. The encoded Collision Type Feature "Parked Car" was the the top feature for just about every model including XGB. For XGB the top four features were Collision Types however this was not the case for the other models.

Model	Accuracy	Jaccard	F1-Score	Log Loss	True -	False +	False -	True +
Gradient Boosting Trees	**0.668843**	0.590327	**0.639449**	0.597680	**4461**	**7177**	531	11107
XGBoost	0.680014	**0.598685**	0.653896	**0.601270**	4717	6921	**527**	**11111**
Random Forest	**0.721430**	0.588787	**0.719798**	**0.528355**	**7508**	**4130**	**2354**	**9284**
Decision Tree	0.685513	0.593356	0.667581	0.551461	5275	6363	957	10681
Logistic Regression	0.710904	**0.581998**	0.708319	0.534020	7178	4460	2269	9369

Figure 23 - Model Results Summary

	Accuracy	Jaccard	F1-Score	Log Loss	True -	False +	False -	True +
count	5.000000	5.000000	5.000000	5.000000	5.000000	5.000000	5.000000	5.000000
mean	0.693341	0.590631	0.677809	0.562557	5827.800000	5810.200000	1327.600000	10310.400000
std	0.022003	0.006130	0.034792	0.034784	1418.953734	1418.953734	915.506308	915.506308
min	0.668843	0.581998	0.639449	0.528355	4461.000000	4130.000000	527.000000	9284.000000
25%	0.680014	0.588787	0.653896	0.534020	4717.000000	4460.000000	531.000000	9369.000000
50%	0.685513	0.590327	0.667581	0.551461	5275.000000	6363.000000	957.000000	10681.000000
75%	0.710904	0.593356	0.708319	0.597680	7178.000000	6921.000000	2269.000000	11107.000000
max	0.721430	0.598685	0.719798	0.601270	7508.000000	7177.000000	2354.000000	11111.000000

Figure 24 - Variance in Scores/Counts

Results Rubric -

- The most basic score is Accuracy that represents the number of all correct predictions for the total number of samples represented by a percentage. The higher the percentage the better.
- The Jaccard Score is a metric for calculating the dissimilarity between two sample sets, i.e. the predicted classes and the actual classes. The Jaccard Score is defined as the size of the intersection divided by the size of the union of the two sample sets. The higher the percentage the better.
- The F1-Score measures the balance between true positive and false positives. The F1-score is the harmonic mean of the precision and recall (ranging between 0 for worst and 1 for best). Precision is the number of true positive results divided by the number of all positive results, including false positives. The higher the percentage the better.
- True Negative, False Positive, False Negative and True Positive that are the underlying statistics for the confusion matrix (Figure 25). The confusion matrices in this study consist of number of Injury Collisions correctly predicted (True Positive), Injury Collisions incorrectly predicted (False Negative), Property Collisions correctly predicted (True Negative) and Property Collisions incorrectly predicted (False positive FP). In this study a higher True Positives count was the goal.

1st Place XGBoost

Notes

- Selected as best with the highest TP @ 95.5%
- Bad TN but not the worst
- Best Jaccard Score
- Did not have the best F1
- First four features where collision types, the only model that did this
- First feature Parked Car had over 4x importance then next feature
- It was by far the slowest model to run fit/predict take several minutes for each cycle

Scores

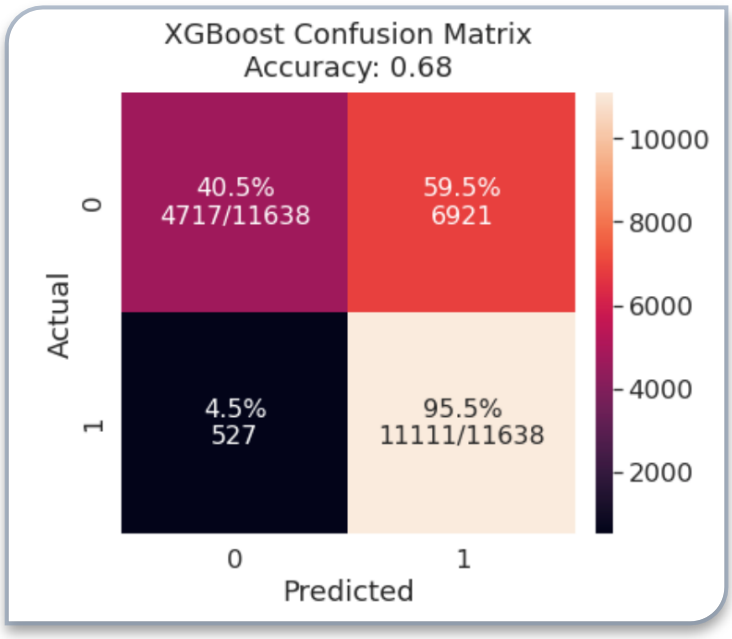
- General Accuracy - 0.680014
- Jaccard - 0.598685
- F1 - 0.653896
- Log Loss - 0.601270

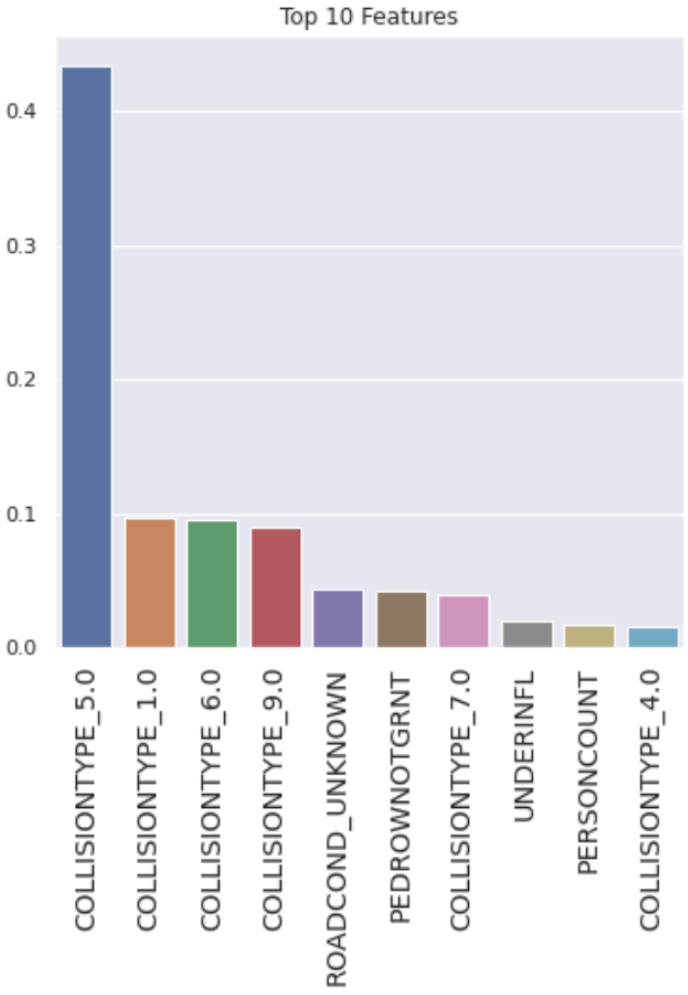
Key Parameters

- Scale Pos Weight = 2.7
- Learning Rate = .3
- Max Depth = 4
- Gamma = 4
- Min Child Weight = 5
- Col Sample by Tree = .4

Feature Map

- CT5 = Parked Car
- CT1 = Angles
- CT6 = Pedestrian
- CT9 = Sideswipe
- CT7 = Rear Ended
- CT4 = Other





2nd Place – Gradient Boosting Trees

Notes

- Second best TP, .1 worse than XGB
- Override balancing to assign weights based on original sample to best run
- Worst TN and FP
- Worst F1 Score
- Top feature was Parked Car with over twice the Gini Importance

Scores

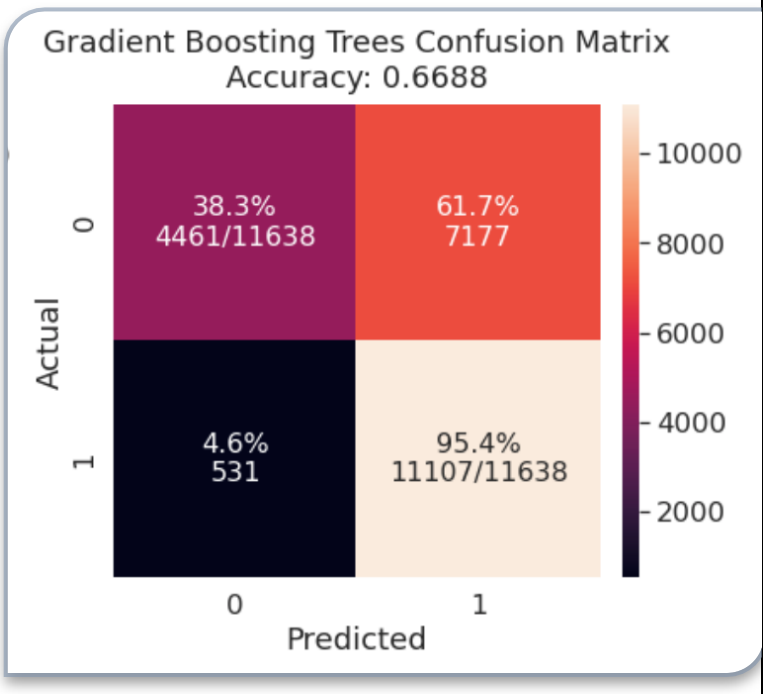
- General Accuracy - 0.680014
- Jaccard - 0.598685
- F1 - 0.653896
- Log Loss - 0.601270

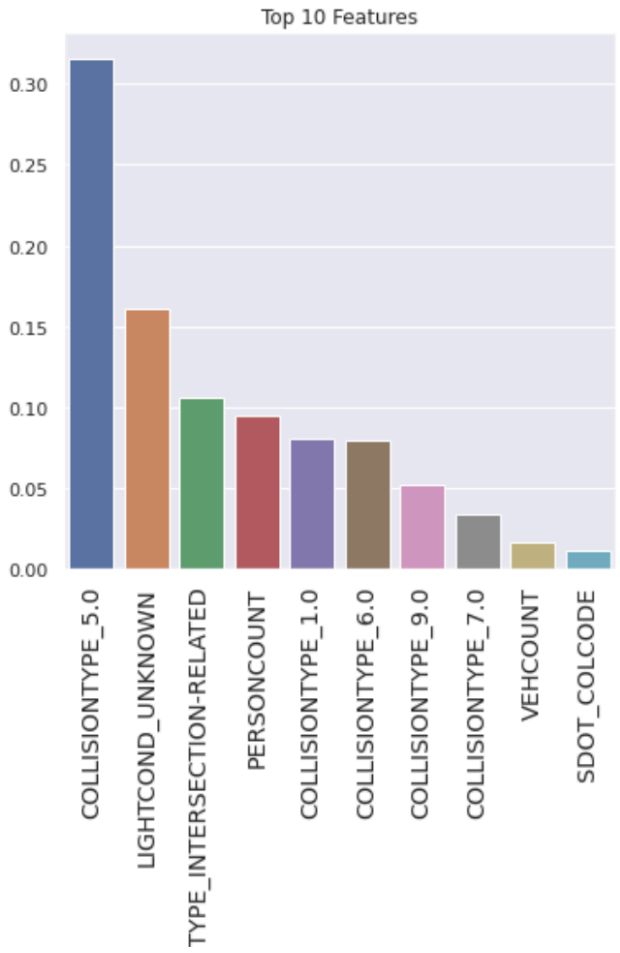
Key Parameters

- Estimators = 20
- Learning Rate = .75
- Max Depth = 2
- Sample Weight = .29/.70

Feature Map

- CT5 = Parked Car
- CT1 = Angles
- CT6 = Pedestrian
- CT9 = Sideswipe
- CT7 = Rear Ended





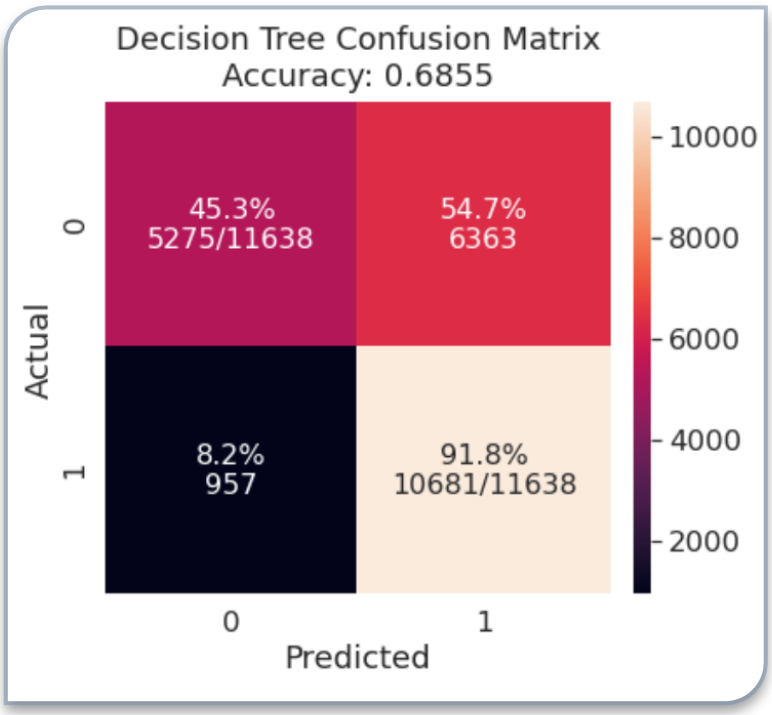
3rd Place – Decision Tree

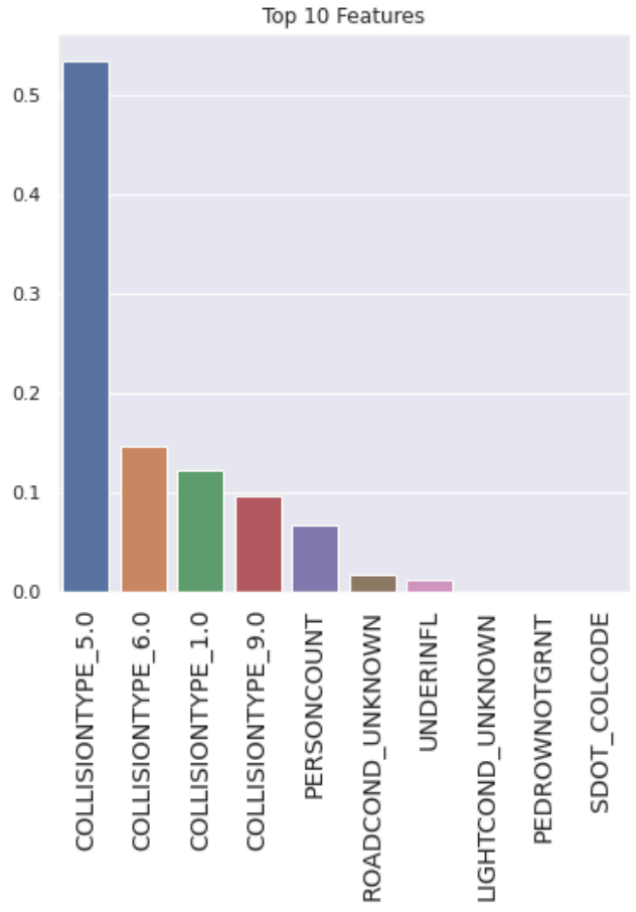
Notes
- Third best TP, over 90%
- Surprised by the overall performance
- Great to generate actual tree for interpretability
- Top feature Parked Car had over 4x Gini Importance

Scores
- General Accuracy - 0.685513
- Jaccard - 0.593356
- F1 - 0.667581

Key Parameters
- Log Loss - 0.551461
- Criterion – Entropy
- Max Depth – 4

Feature Map
- CT5 = Parked Car
- CT6 = Pedestrian
- CT1 = Angles
- CT9 = Sideswipe



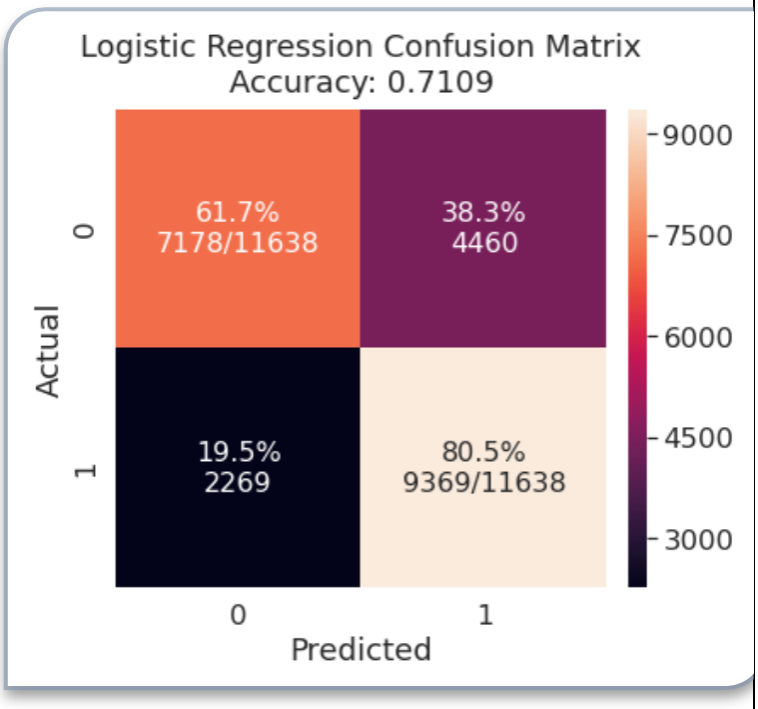


4th Place – Logistic Regression

Notes
- Good overall performance, shows great utility in supporting binary classification
- Worst Jaccard Score
- Second worst TP

Scores
- General Accuracy - 0.710904
- Jaccard - 0.581998
- F1 - 0.708319
- Log Loss - 0.534020

Key Parameters
- C = 0.1
- Solver = Liblinear
- Class Weight = balanced



Feature Importance Unavailable

5th Place – Random Forest

Notes

- Worst TP, FN make this the worst performer
- **But…** best scores for General Acc and F1
- Best TN and FP
- Top feature Parked Car had twice the Gini Importance than the next feature

Scores

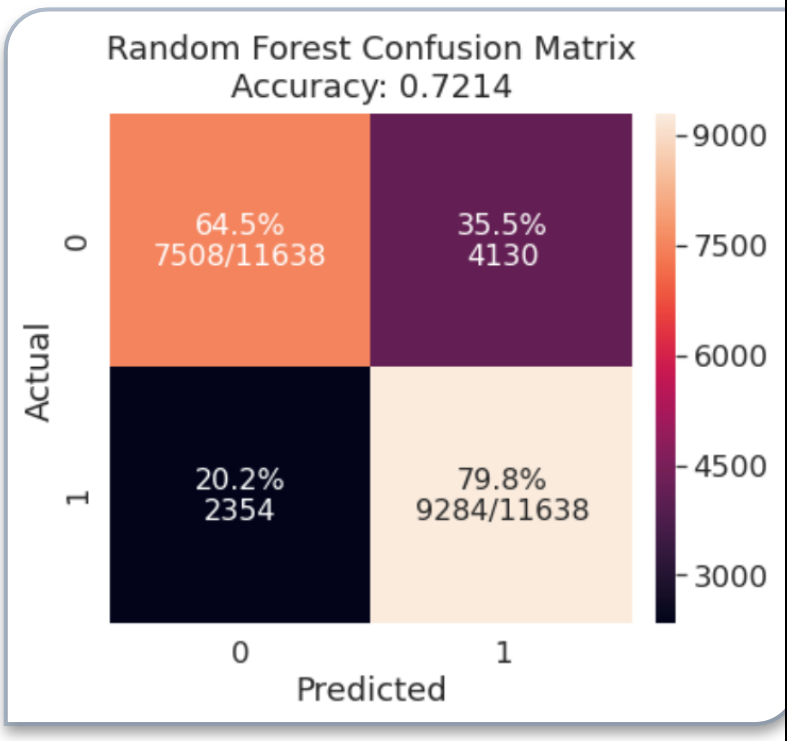
- General Accuracy - 0.721430
- Jaccard - 0.588787
- F1 - 0.719798
- Log Loss - 0.528355

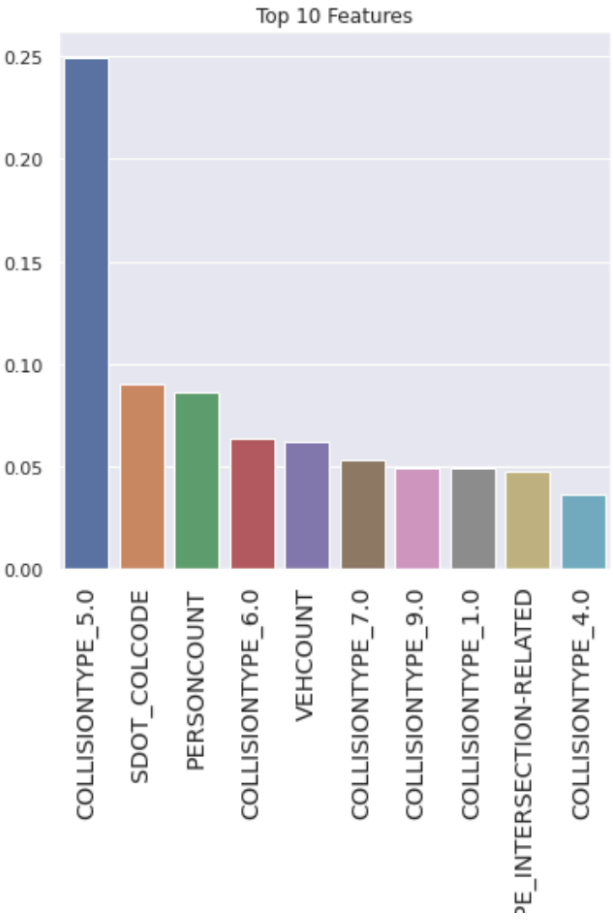
Key Parameters

- Estimators – 500
- Criterion – Gini
- Max Depth – 10
- Max Features - Auto

Feature Map

- CT5 = Parked Car
- CT6 = Pedestrian
- CT7 = Rear Ended
- CT9 = Sideswipe
- CT1 = Angles
- CT4 = Other





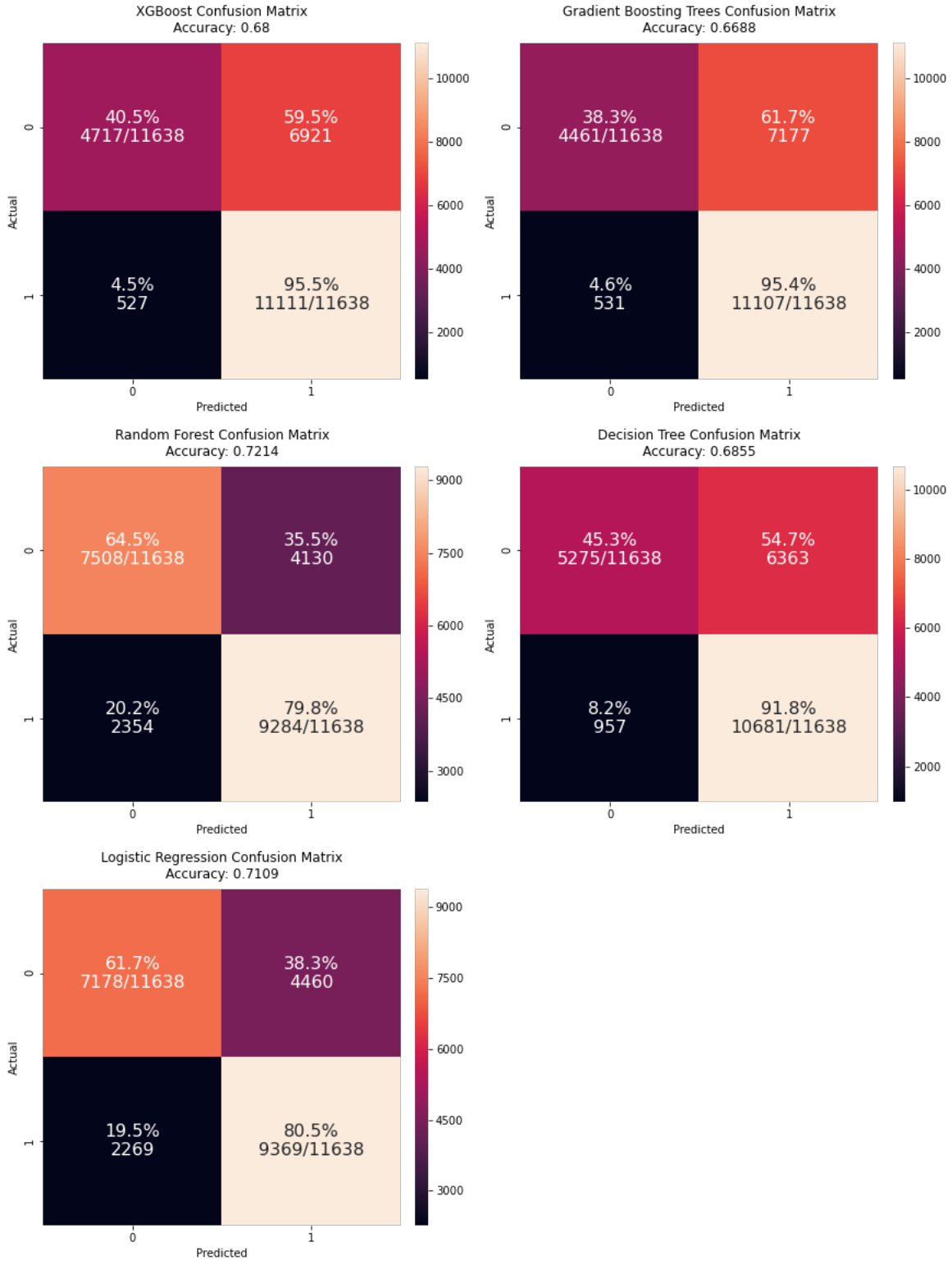


Figure 25 - Confusion Matrix Gallery for all Models

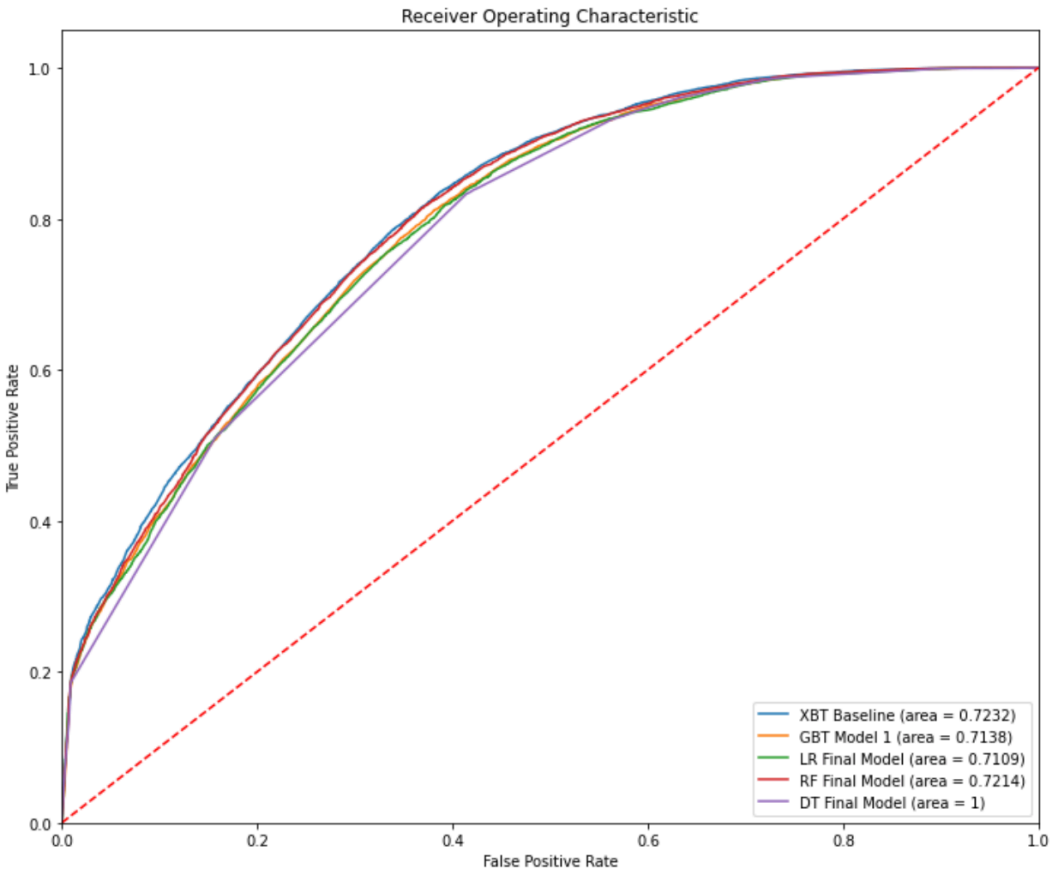


Figure 26 - Receiver Operating Characteristic by Model

Models were evaluated using the associated ROC curves as shown in Figure 26. In this study a higher True Positive Rate is more important than lower False Positive Rate. Meaning that it's more important for this study to accurately predict Injury Collisions whereby it would be more acceptable to wrongly predict a collision as an Injury Collision versus incorrectly predicting a collision as a Property Collision when it actually had an injury involved. In the ROC curves all of the models had a high True Positive rate with XBT nosing out the others.

5 Discussion

The problem this study aims to solve is the ability to predict Injury Collisions and results show that using the current features it's possible to do so with over a 95% accuracy using XGB. However, reaching this accuracy resulted in a compromise of overall predictive performance. In the real world first responders being sent to a collision scene that was just a minor collision with no injuries is obviously not desired. Nonetheless this study set out with the goal to determine how well a Machine Learning Model could predict Injury Collisions using the SDOT dataset without any additional data to improve the results.

Feature importance was relatively consistent where XGB selected six Collision Types in the top ten as ranked by Gini Importance. The top feature was Collision Type = Parked Car which was consistently selected by most of models. However, XGB was the only model to select Collision Type Categories as the top five features. GBT had two Collision Type Categories in the top five with the top feature going to Parked Car again. Decision Trees selected four Collision Type Categories in the top five and the lowest performer RF only had two Collison Types.

All of the classification models performed well and with additional tuning and feature engineering it appears that RF would eventually outperform the other models. Tuning was a critical step as it was often surprising that certain configurations proved better than others especially if one is unfamiliar with the innerworkings of a particular classifier. Tuning using a combination of cross validation and grid searching was the most helpful aspect of this effort although it often required a long time and patience to find the right balance.

Note that XGB did have the best Jaccard Score at .59865 but all of the models share a similar score (within .00610 standard deviation). A key factor was applying positive scoring weights to the XBG-based model. Without assigning the weights the performance was poor and well below the final results. In the study a positive weight score of 2.7 proved to be the best parameter however manually calculating the positive weight score was 2.3. This seems to underscore the need to perform tuning on each model using a cross-validated grid-search as it was key to obtain the best combination of hyperparameters to maximize TP. The downside was that some of the models, especially KNN, took a very long time to optimize using this technique.

Excluding True Positive from the evaluation criteria RF was the best overall performer. The model using RF had a F1-Score of .719798 and also scored best with True Negatives and False Positives, 7508 and 4130 respectively. General Accuracy was also the highest at .721430. Where RF fell flat on its face was TP and FN and it would be worth investigating further why this was the case.

It was surprising how well Logistic Regression performed and shows the flexibility of LR on binary classification problems. A balanced dataset played a critical role for these models performing well, especially with LR. With an unbalanced dataset scores across the board were significantly lower especially with LR.

And some words about balancing the data set. Random under-sampling was used to balance the dataset prior to modelling which made a significant difference. However, this was the only approach attempted due to time constraints. It would be worth exploring other approaches such as Synthetic Minority Over-sampling Technique or SMOTE. SMOTE sampling is combination of over-sampling the minority class and under-sampling the majority class that can achieve better classifier performance (in ROC space) than only under-sampling the majority class.

Using the correlation matrix was critical as part of feature selection used to pick the correct set of attributes that would provide the best results. The selection can be difficult at times and trial and error was the best approach to find the right combination. For example, the exclusion of Peak or Off Peak in the feature set needs further investigation because it did have a stronger correlation then several of the other features used. Unfortunately, none of the features had a Pearson Correlation Coefficient greater than 0.21 (Failed to grant pedestrian the right of way feature) so there was a slim picking from the start. Also, when the study started there was an inclination that latitude and longitude could be a strong candidate as a feature however it was proved false.

The interpretation of weather and road conditions as features is that these could be much more helpful. It's possible that police officer completing the collision report does not accurately report this information. Regardless of that the overwhelming number of collisions irrespective of severity occur in clear weather and good road conditions. The person count is a bit misleading as most collisions will have two parties however there was a significant increase in both severities with three people involved indicating that having a second person (or more) in a vehicle could translate to a higher risk due to distractions in the vehicle. Similarly, distracted driving is clearly risky and is trending up suggesting this factor is here to stay and needs to be addressed to ensure public safety.

6 Conclusion

The study results conclude that using machine learning can predict an Injury Collision with 95% accuracy. This predictive capability would be an invaluable tool for police and first responders to better prepare for a collision instead of having to assess the situation at the scene. Unfortunately, several of the features used are not known until well after the collision, for example Collision Type, therefore it's not likely this would have any use outside of this case study. Nonetheless applying machine learning techniques in public safety has incredible potential especially considering the vast amount of data that is already held by governmental organizations.

The One Hot Encoded Collision Type proved to be the most important feature in in all of the models with "Parked Car" being the top feature by far. XGB has six Collision Types in it's top ten followed by Unknown Road Condition, Pedestrian Right of Way, Driving While Intoxicated and Person Count. XGB and RF were the only models to have six Collision Types in its top ten features. RF had a lower Gini Importance for the Collision Types which could account for the lower performance. Outside of Collision Type the majority of features did not have a significant Gini Importance including SDOT Collision Code.

With respect to data, the first point is that how frequently the data is updated and validated. The dataset was last updated in May 2020 and while other factors could have caused this having up-to-date collision statistics would be important to not just predict severity but to also predict recent trends, e.g. a new intersection is resulting in an abnormal number of collisions. Furthermore, the systems used for collecting collision data need to be improved as the validity of the current dataset is suspect and in numerous incidents information was missing and/or undefined. Also, a state-wide standard should be considered instead of having multiple coding schemes to classify collisions. Currently SDOT and Washington State classify collisions in different ways. A unifying standard would appear to benefit the public and government in leveraging this information in the best way possible.

Also missing from the data are details about the vehicles involved as well as driver information, e.g., types of vehicles involved, age of the drivers. This information no doubt exists in the police collision report. This information would be helpful to improving the accuracy of the model but implies more comprehensive data collection and supporting application. However, in the quest for public safety this is well worth it.

Another place for improvement is weather. It is possible to collect weather data across the city that can include precipitation, temperature, windspeed, solar radiance and dew point. Collecting this data and correlating it with collision data would give a much better picture of environmental conditions at the time of the collision and greatly expand on the simple category that is used now.

Another example of improvement is that street names that are captured in text and concatenated. While it's possible to obtain this information from geocoordinates having this information properly attributed in the data set would simplify and enhance analysis.

Finally, this study did not meet the initial goal to provide a capability that could be integrated with an online application permitting the public to determine whether or not a trip within Seattle could pose higher risk than normal based on the current conditions. For example, a user of the application could enter their starting point and destination along with date and time and get a risk rating. It would be within reach to provide this capability but would require significant effort to operationalize such a system. In any case this sort of application is coming soon as all of the data is out there, it's just a matter of time and data scientists teaming with public servants to make it happen.