Predicting Car Accident Severity in Seattle

Prepared by : **Suganthi Sathiyamoorthi**
For : **IBM DataScience Capstone Project**
Submitted On : **Oct 6th 2020**

Predicting Car Accident Severity in Seattle

1. Introduction

1.1 Background

According to the report published by US Department of Transportation, [NSHTA], there were 33,000 fatalities, 3.9 million injuries, 24 million vehicles damaged and 242 billion economic loss due to road accidents in US in 2010 alone. This economic loss includes productivity loss, medical costs for injuries, legal and court costs due to law suits, emergency service costs (EMS) for emergencies, insurance administration costs for the claims, congestion costs (including traffic delay, fuel wastage, greenhouse gas emissions etc.,), property damage, and workplace losses. There is a tremendous non-economic impact such as pain, emotional distress, PTSD and life valuations on societal harm amounts to $836 billion just for 2010 **(Blincoe, 2015, May)**. Hence, analyzing existing accident reports for patterns can help identify high accident zones, accident safety laws and predict possible accidents before it occurs. For our study we will be analyzing traffic accident reports of Seattle, Washington, US. Seattle is a very busy downtown and home of very big Tech companies like Microsoft, Facebook etc., Seattle downtown streets are very busy and prone to high accidents.

1.2 Problem

From the traffic accident report, information such as accident location, collision type, weather condition, road condition, light condition, severity type etc., are recorded and public can utilize this data to predict if a collision or injury will occur in Seattle given the suitable conditions. Hence, this data science project's goal is to predict if an accident will occur and what will be its severity when suitable conditions occur.

1.3 Interest

Such Machine Learning models will be useful for many companies.

1. These models can be used by city planning officials to plan road markers, traffic light lengths, traffic patterns and routing etc.,
2. Digital map providing companies and traffic advising companies like Google Maps, Waze etc., can utilize the prediction to warn their customers of perfect storm conditions of an accident when they are driving in those areas.
3. Vehicle Insurance companies can use these models to evaluate claims for vehicle accidents.

Predicting Car Accident Severity in Seattle

2. Data Preprocessing

2.1 Data Sources

Data provided with this project request is used for this analysis. A more raw data can from Seattle GeoData (DOT, 2018) where most of the data and attributes can be studied. Even though raw dataset has records from 2004 and has data for all severitycodes and studied for possible utilization for this project, course provided dataset with just 2 severity codes was utilized due to very unbalanced nature of the data. Models were evaluated for utilizing unbalanced data but could not use it because of heavily unbalanced data that requires further analysis and utilization of advanced techniques. The imported data has 194673 observations of various attributes such as severity code, severity code description, Address type, Junction type, collision type, Road Condition, Light Condition, Weather Condition, Speeding, Driving Under Influence indicator, Longitude, Latitude, Injury count, Fatality count, SDOT code, description, Data and Time of Accident etc., The attribute Types and descriptions can be found at Seattle DOT

2.2 Data Cleaning

Data downloaded from data source with nulls and Nan values were replaced with Unknown/Other values. Data formats were converted to other formats as required. A few columns such as Severity Code, Description & SDOT Code, Description & ST Code and Description were combined to new columns for analysis purposes. X and Y co-ordinates missing Latitude and Longitude were substituted with Seattle's Latitude and Longitude. Upon cleaning, it can be observed that the observations are available from April 1st 2004 to May 20th 2020.

2.3 Exploratory Data Analysis

As the source data has many attributes, all the attributes were studied for possible selection as feature for this project. Individual Attribute was checked for percentage valid data that might be a possible reason for the accident. Such attributes were further analyzed with respect to the target attribute of severity. The following shows the results of this analysis:

Total Data Observations : 194673

Attribute	Selected(Y)/Not Selected (X)	Reason
OBJECTID	X	Doesn't provide relevant information for Prediction
SHAPE	X	Doesn't provide relevant information for Prediction

Predicting Car Accident Severity in Seattle

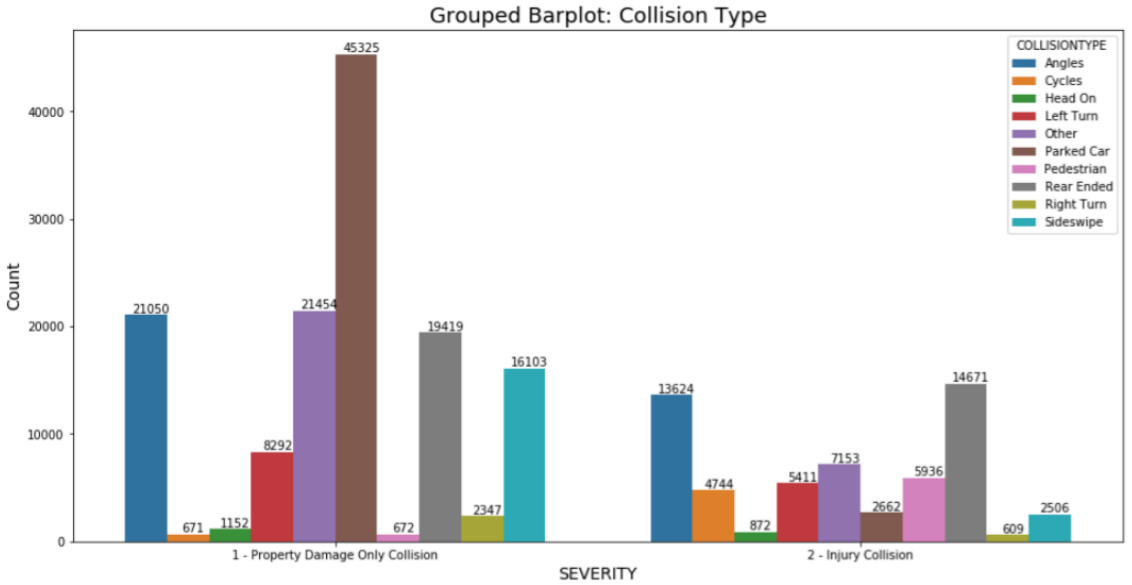
INCKEY	X	Doesn't provide relevant information for Prediction
COLDETKEY	X	Doesn't provide relevant information for Prediction
ADDRTYPE	X	Collision address type: Alley Block Intersection **Junction type provides more information than Address Type **
INTKEY	X	Doesn't provide relevant information for Prediction
LOCATION	X	Description of the general location of the collision , ** **X and Y co-ordinate provides more info than Location** **
EXCEPTRSNCODE	X	Doesn't provide relevant information for Prediction
EXCEPTRSNDESC	X	Doesn't provide relevant information for Prediction
SEVERITYCODE	Y	A code that corresponds to the severity of the collision: • 3—fatality • 2b—serious injury • 2—injury - Count - **58188** • 1—prop damage – Count - **136485** • 0—unknown
SEVERITYDESC	Y	A detailed description of the severity of the collision
COLLISIONTYPE	Y	Collision type Provides more information about the collision
PERSONCOUNT	X	The total number of people involved in the collision ** Not selected as this information pertains to particular accident **
PEDCOUNT	X	The number of pedestrians involved in the collision. This is entered by the state. ** Not selected as this information pertains to particular accident **
PEDCYLCOUNT	X	The number of bicycles involved in the collision. This is entered by the state. ** Not selected as this information pertains to particular accident **
VEHCOUNT	X	The number of vehicles involved in the collision. This is

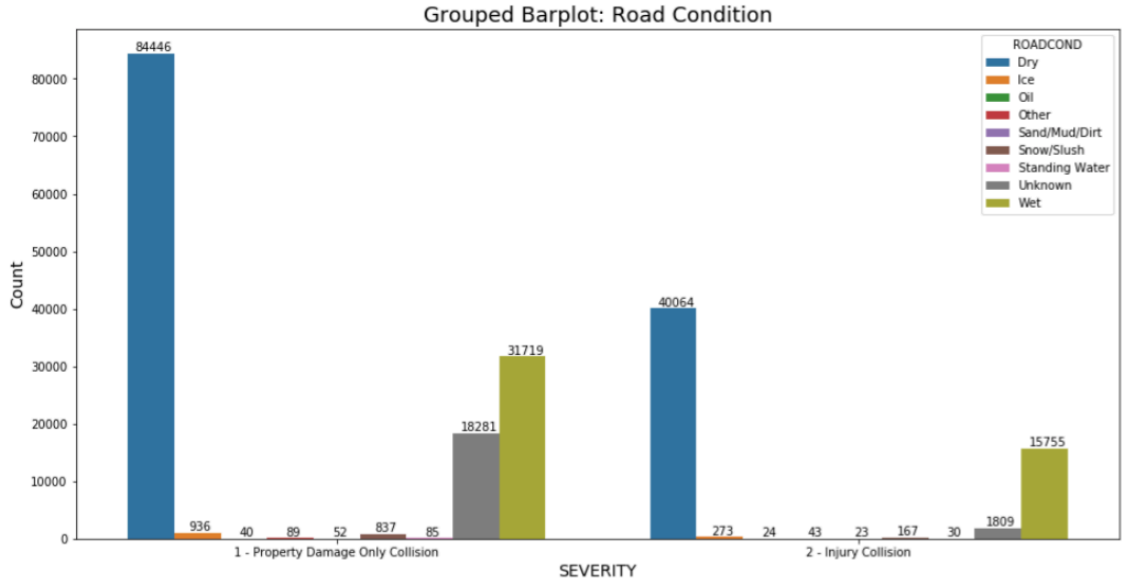
		entered by the state. ** Not selected as this information pertains to particular accident **
INJURIES	X	The number of total injuries in the collision. This is entered by the state. ** Not selected as this information pertains to particular accident **
SERIOUSINJURIES	X	The number of serious injuries in the collision. This is entered by the state. ** Not selected as this information pertains to particular accident and dataset doesnot contain this severity**
FATALITIES	X	The number of fatalities in the collision. This is entered by the state. ** Not selected as this information pertains to particular accident and dataset doesnot contain this severity**
INCDATE	X	The date of the incident.
INCDTTM	X	The date and time of the incident. **Incident Date and Time was used in evaluation but was not selected as time of the day was not available for more that 25000 of the cases. Day of the week was also evaluated but not significant difference was found in the day of the week**
JUNCTIONTYPE	Y	Category of junction at which collision took place
SDOT_COLCODE	Y	A code given to the collision by SDOT.
SDOT_COLDESC	Y	A description of the collision corresponding to the collision code.
INATTENTIONIND	X	Whether or not collision was due to inattention. (Y/N) ** Y datacount was very less **
UNDERINFL	X	Whether or not a driver involved was under the influence of drugs or alcohol. ** Y datacount was very less **
WEATHER	Y	A description of the weather conditions during the time of the

		collision.
ROADCOND	Y	The condition of the road during the collision.
LIGHTCOND	Y	The light conditions during the collision.
PEDROWNOTGRNT	X	Whether or not the pedestrian right of way was not granted. (Y/N)
SDOTCOLNUM	X	A number given to the collision by SDOT.
SPEEDING	X	Whether or not speeding was a factor in the collision. (Y/N) ** N count was very high and Y count very less . Y – `9333 count` **
ST_COLCODE	X	A code provided by the state that describes the collision. For more information about these codes, please see the State Collision Code Dictionary. ** Similar to SDOT **
ST_COLDESC	X	A description that corresponds to the state's coding designation. ** Similar to SDOT **
SEGLANEKEY	X	A key for the lane segment in which the collision occurred.
CROSSWALKKEY	X	A key for the crosswalk at which the collision occurred.
HITPARKEDCAR	X	Whether or not the collision involved hitting a parked car. (Y/N) ** Very Unbalanced datacontent **

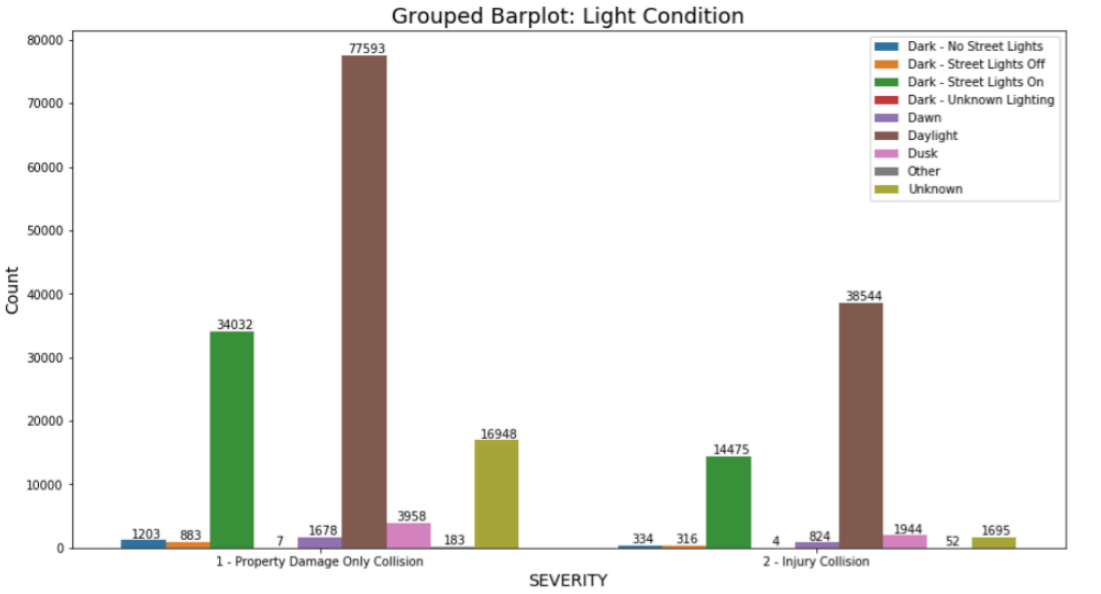
Predicting Car Accident Severity in Seattle

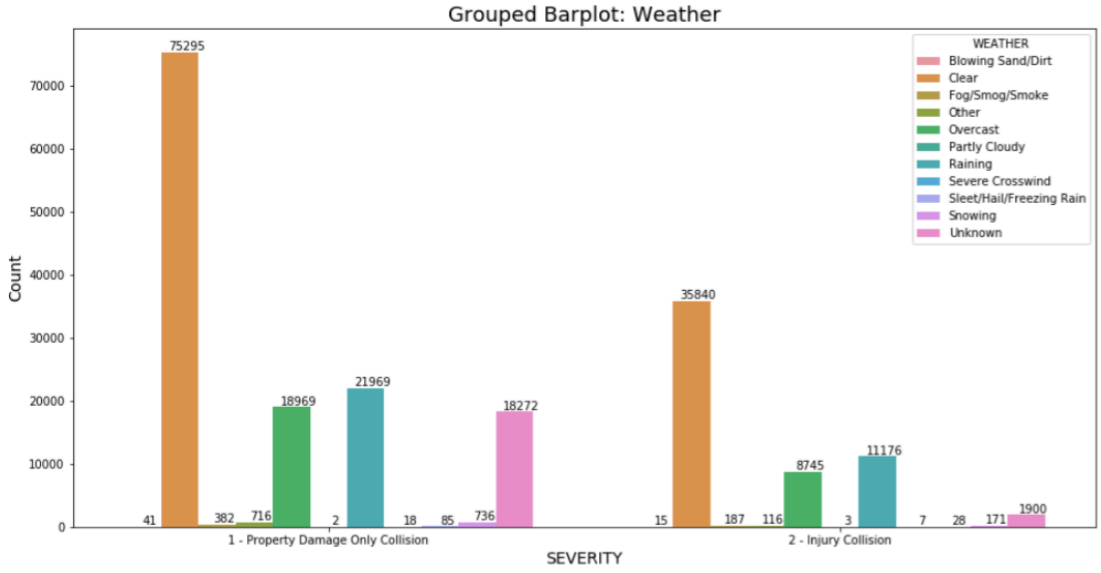
Selected Attributes were analyzed against severity as follows:



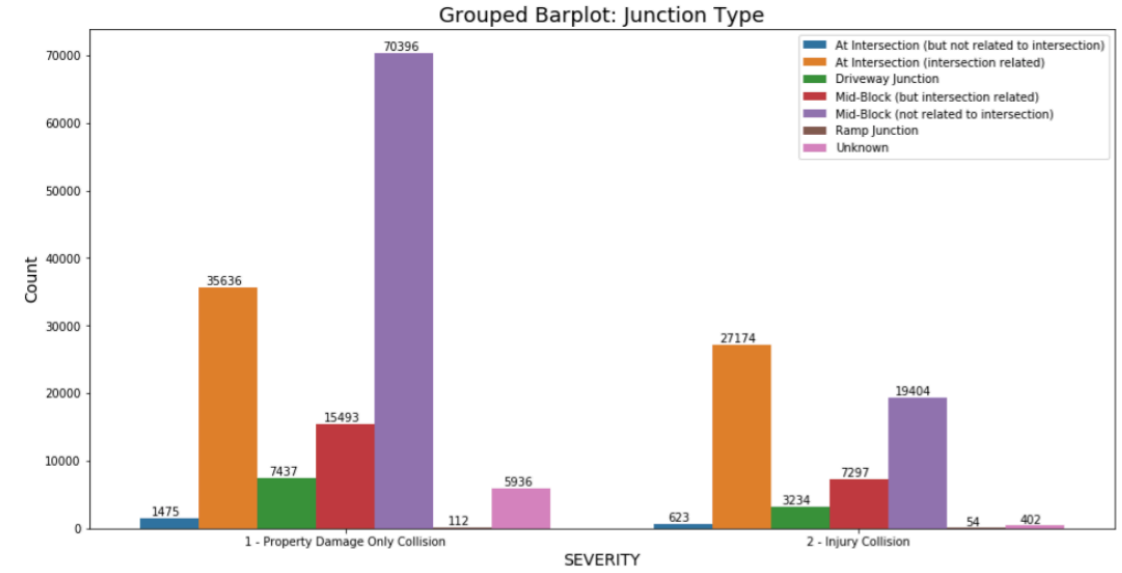


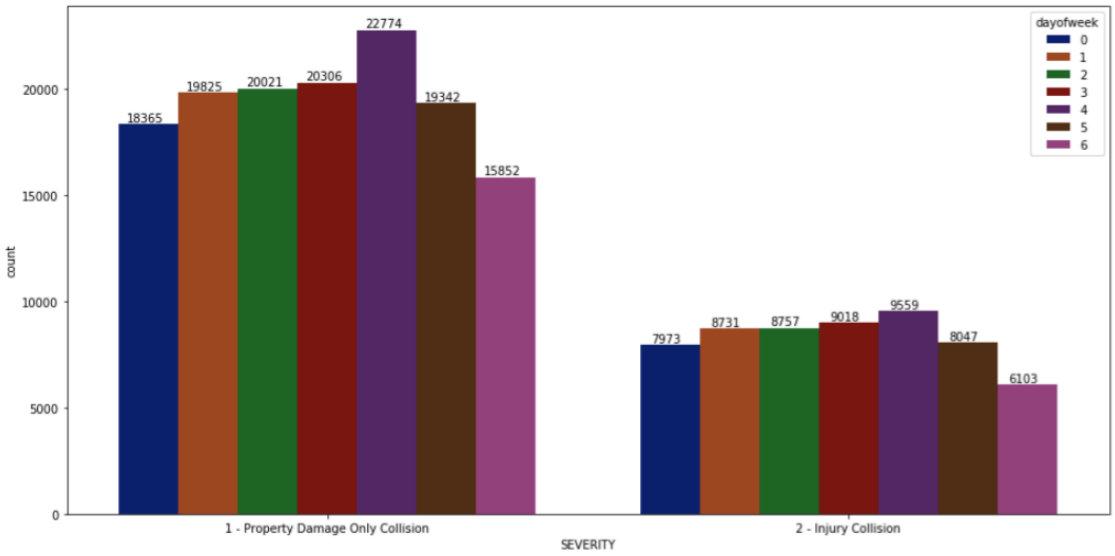
Predicting Car Accident Severity in Seattle



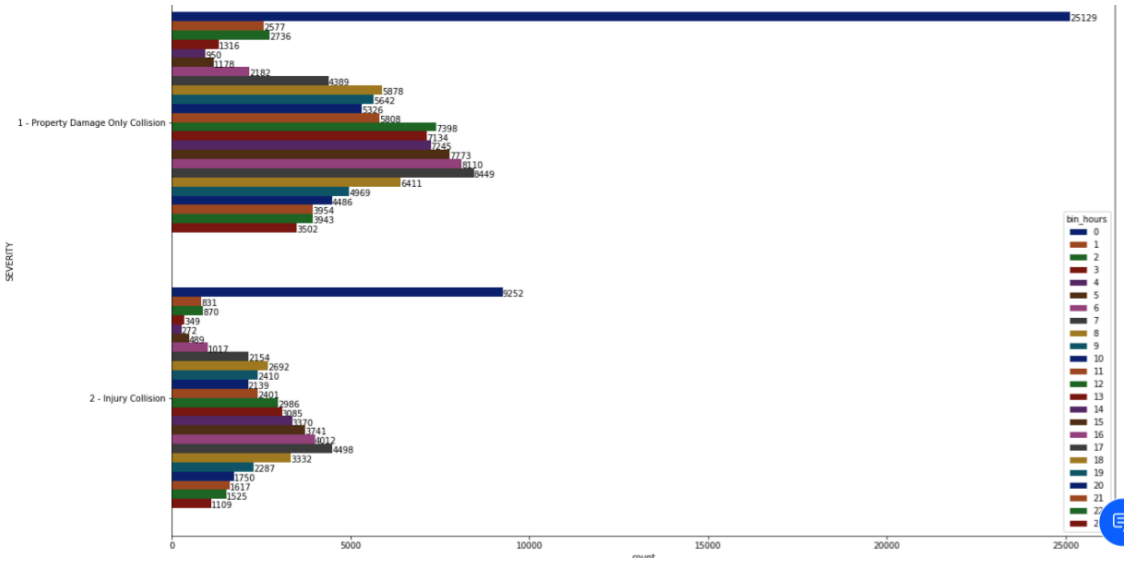


Predicting Car Accident Severity in Seattle

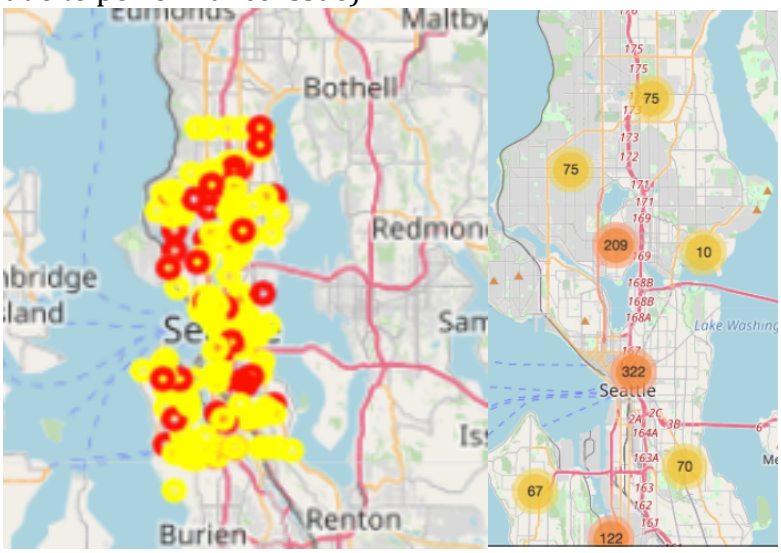




Predicting Car Accident Severity in Seattle



X and Y Co-ordinates were plotted using folium (1000 samples only were plotted due to performance issue)



Predicting Car Accident Severity in Seattle

3. Methodology

Based upon the exploratory data analysis, we can come to conclusion that most of the attributes are categorical variables. Hence, our models should be able to handle classification problems. Decision Trees are very good in handling such classification problem with categorical values. Models can be built and run for different combination of features to identify better performance.

3.1 Feature Selection

Hence, in feature engineering we are going to select the possible features and their relationship to the target variable Severity code. Based upon the above analysis, the following attributes are selected for feature engineering:

- SEVERITY CODE, SEVERITY CODE DESCRIPTION

- 1 - Property Damage Only Collision 136485
  2 - Injury Collision 58188

- X, Y,
- JUNCTION TYPE, COLLISION TYPE
- ROAD CONDITION, LIGHT CONDITION, WEATHER CONDITION
- SDOT CODE, SDOT DESCRIPTION
- PERSONCOUNT, VEHCOUNT, PEDCOUNT, PEDCYLCOUNT
- INCIDENT DATE and TIME

The following attributes were not selected based upon 194673 observations:

- SPEEDING – Only 9333 are with Y values
- DRIVING UNDER INFLUENCE – 9121 with Y or 1 values, 185552 with N or 0 values
- ADDRESS TYPE is similar to JUNCTION TYPE. JUNCTION TYPE has more information than ADDRESS TYPE.
- INATTENTIONIND was not selected due to very less Y count – 29285
- PEDROWNOTGRNT – very less Y count – 4667
- HITPARKEDCAR - Y count – 7216
- Keys and Ids were not selected for analysis

As there many features available we will be running our models against 4 feature sets. as follows:

Feature No:	Feature sets
1	• JUNCTION TYPE, COLLISION TYPE • ROAD CONDITION, LIGHT CONDITION, WEATHER CONDITION • SDOT
2	• JUNCTION TYPE, COLLISION TYPE • ROAD CONDITION, LIGHT CONDITION, WEATHER CONDITION
3	• JUNCTION TYPE, COLLISION TYPE • ROAD CONDITION, LIGHT CONDITION, WEATHER CONDITION • PERSONCOUNT, VEHCOUNT, PEDCOUNT, PEDCYLCOUNT
4	• JUNCTION TYPE, COLLISION TYPE • ROAD CONDITION, LIGHT CONDITION, WEATHER CONDITION • INCIDENT TIME BINS

3.2 Data Selection and Wrangling

3.2.1 Balancing the Data

As we see that the number of rows with Severity 1 – Property Damage and Collision (136485) is 3 times more than Severity 2 - Injury Collision - (58188), there is a need to balance the unbalanced data. There are different methods such as Upsample the minority, Downsample the majority data or utilize models with algorithms to handle unbalanced data. In this project, we will use Downsample as we have huge sample volume.

3.2.2 Data Transformation

As we have categorical values, these values need to be transformed to numeric values and normalized. Available observations also need to be split into Train, Test and

Evaluation datasets. In case of some of the models such as Knn and SVM, the number of samples in train and test datasets had to be reduced to 5000 records due to very poor performance. The Train, Test and Evaluation DataSets are as follows:

```
Full Sample Dataset:
Main set: 107180
Eval set: 2002

Partial Sample Dataset:
Main set: 10000
Eval set: 2002
```

3.2 Model Creation

Supervised classification models run well with the classification datasets. Hence, the following models will be created, fitted with training data and predicted with test data and evaluation datasets.

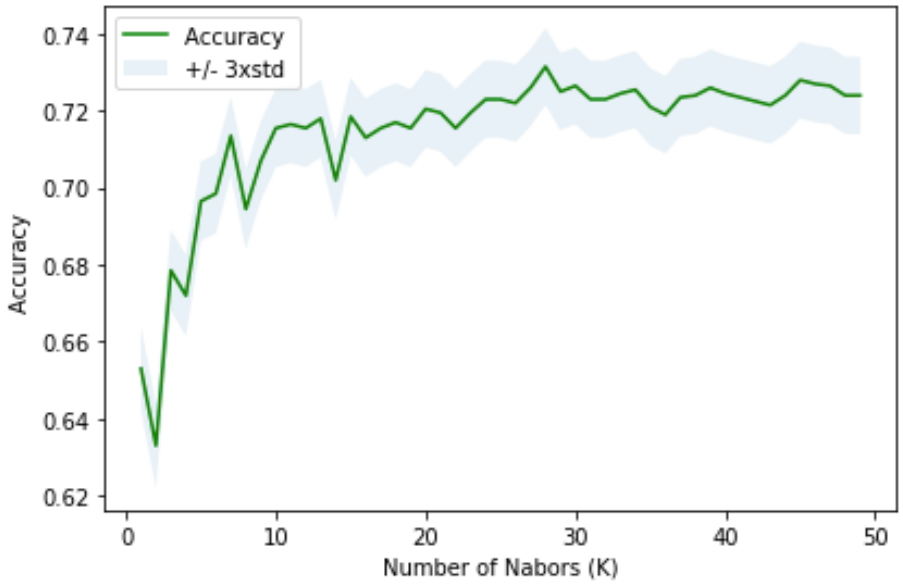
1. Knn
2. DecisionTree
3. SVM
4. Logistic Regression
5. Random Forest (Once to compare with decision tree)

Created models are fitted to training data and predicted with test data to evaluated various scores such as accuracy score, confusion matrix, jaccard score, f1score, ROC curve, Precision recall curve. These models are trained for 4 sets of featur

Predicting Car Accident Severity in Seattle

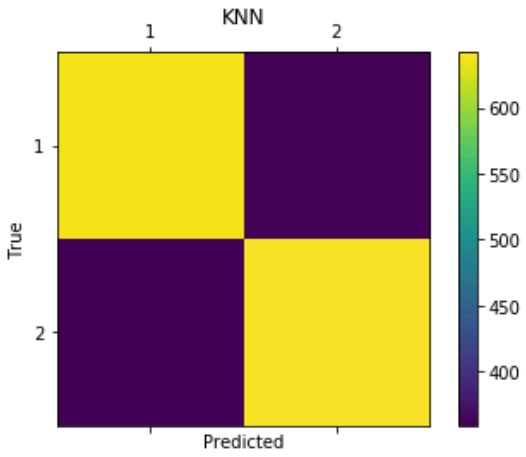
4. Results

All the models were built, fitted, predicted and evaluated as follows:
Knn: Knn was run for 50 iterations with k varying from 1 to 50.The best performance was on k=28 for the 4th feature set.



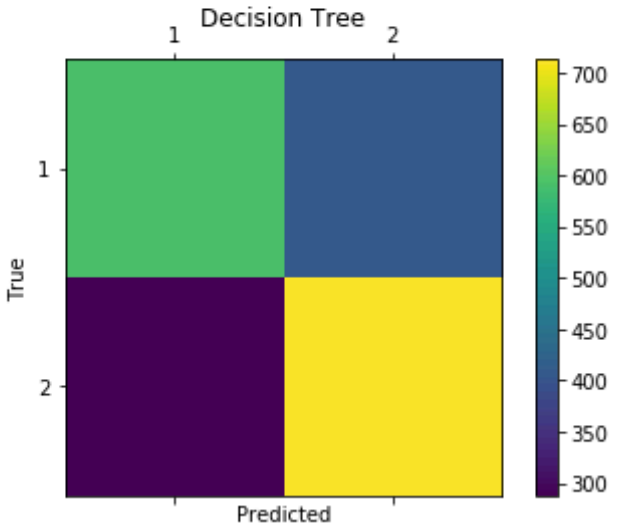
Predicting Car Accident Severity in Seattle

```
Knn Evaluation:
Accuracy: 63.94%
Jaccard: 63.94%
f1_score: 63.86%
precision: 44.97%
recall: 63.74%
              precision    recall  f1-score   support

           1       0.64      0.64      0.64      1001
           2       0.64      0.64      0.64      1001

   micro avg       0.64      0.64      0.64      2002
   macro avg       0.64      0.64      0.64      2002
weighted avg       0.64      0.64      0.64      2002

Confusion matrix:
 [[638 363]
 [359 642]]
```



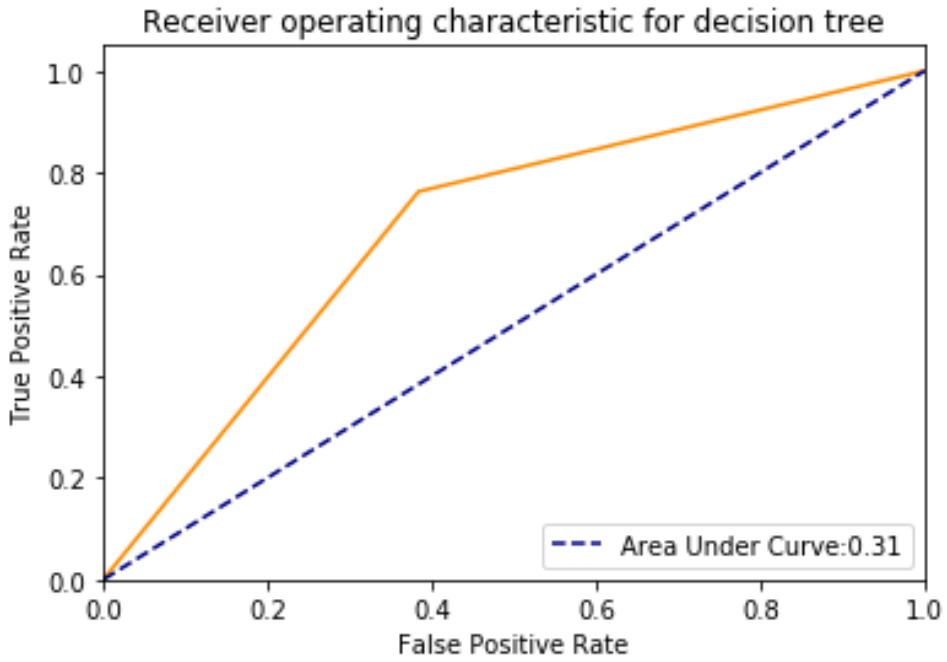
Predicting Car Accident Severity in Seattle

```
Decision Tree:
Accuracy: 65.23%
Jaccard: 65.23%
f1_score: 65.11%
              precision    recall  f1-score   support

           1       0.67      0.59      0.63      1001
           2       0.64      0.71      0.67      1001

   micro avg       0.65      0.65      0.65      2002
   macro avg       0.65      0.65      0.65      2002
weighted avg       0.65      0.65      0.65      2002

Confusion matrix:
 [[593 408]
 [288 713]]
```

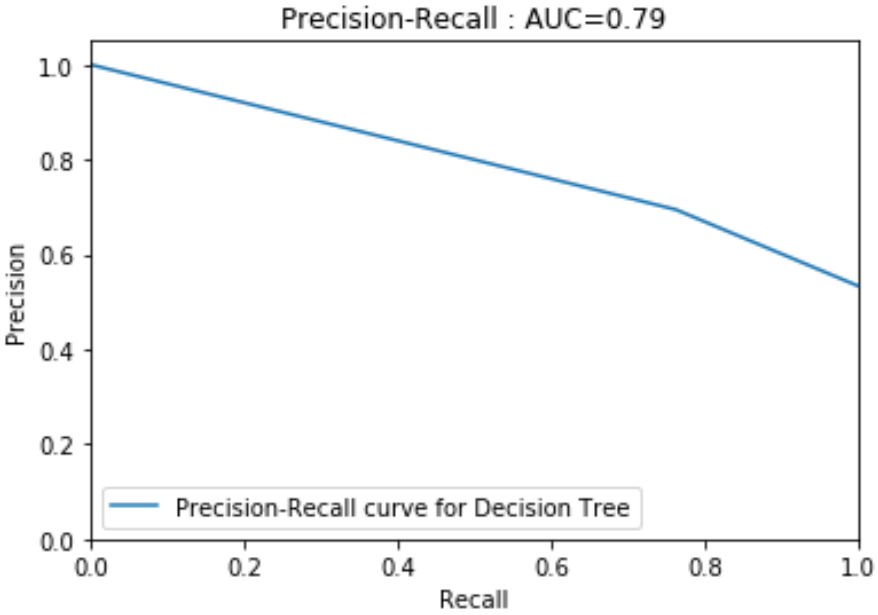


Predicting Car Accident Severity in Seattle

```
0.8101163850996853
Area Under Curve: 0.31
```

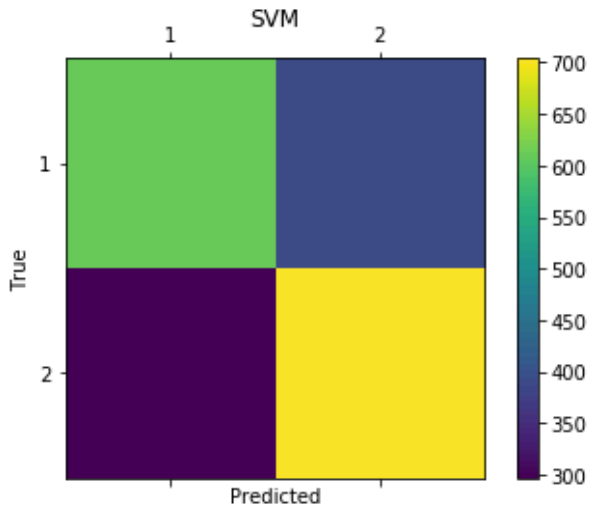



```
Decision Tree: Average precision-recall score: 0.40
Area Under Curve: 0.79
```



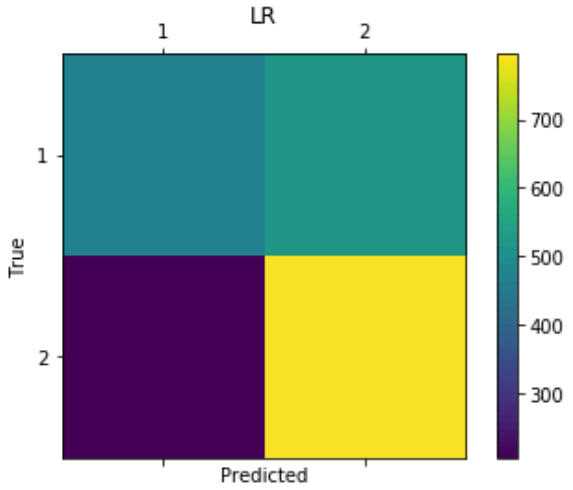
Predicting Car Accident Severity in Seattle

```
Support Vector Machine's Accuracy:  0.6568431568431569
SVM:
Jaccard: 65.68%
f1_score: 64.01%
              precision    recall  f1-score   support

           1       0.67      0.61      0.64      1001
           2       0.64      0.70      0.67      1001

   micro avg       0.66      0.66      0.66      2002
   macro avg       0.66      0.66      0.66      2002
weighted avg       0.66      0.66      0.66      2002

Confusion matrix:
 [[611 390]
 [297 704]]
```



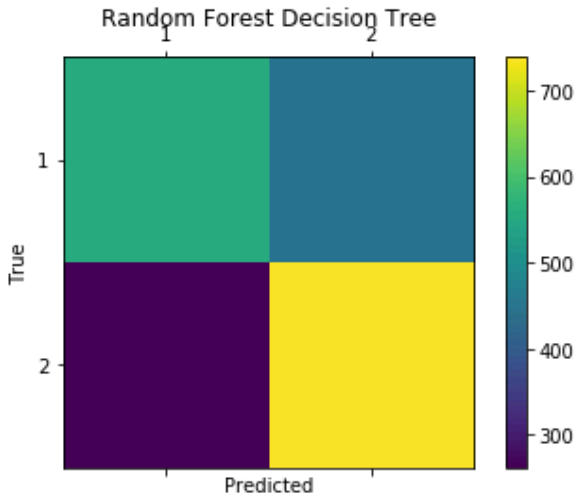
Predicting Car Accident Severity in Seattle

```
Logistic Regression:
Logistic Regression's Accuracy:  0.6358641358641358
Jaccard: 63.59%
f1_score: 62.64%
LogLoss: 61.17%
              precision    recall  f1-score   support

           1       0.70      0.48      0.57      1001
           2       0.60      0.80      0.69      1001

   micro avg       0.64      0.64      0.64      2002
   macro avg       0.65      0.64      0.63      2002
weighted avg       0.65      0.64      0.63      2002

Confusion matrix:
 [[477 524]
 [205 796]]
```



Predicting Car Accident Severity in Seattle

```
Random Forest Decision Tree:
Accuracy: 64.54%
Jaccard: 64.54%
f1_score: 64.23%
              precision    recall  f1-score   support

           1       0.68      0.55      0.61      1001
           2       0.62      0.74      0.68      1001

   micro avg       0.65      0.65      0.65      2002
   macro avg       0.65      0.65      0.64      2002
weighted avg       0.65      0.65      0.64      2002

Confusion matrix:
 [[553 448]
 [262 739]]
```



Predicting Car Accident Severity in Seattle

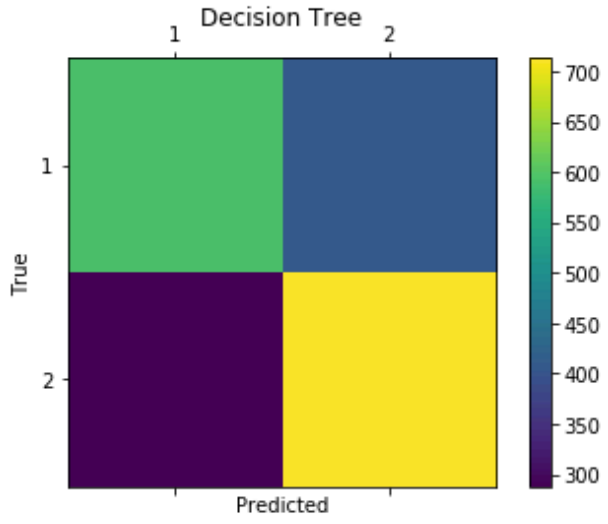
Feature Sets and Model Evaluation results are as follows:

Feature sets	Results			
JUNCTION TYPE, COLLISION TYPE, ROAD CONDITION, LIGHT CONDITION, WEATHER CONDITION, SDOT	**Algorithm**	**Jaccard**	**F1-score**	**LogLoss**
	KNN	0.65	0. 63	NA
	Decision Tree	0. 7008	0. 6985	NA
	SVM	0. 7130	0. 6491	NA
	Logistic Regression	0. 7041	0. 7022	0. 5524
JUNCTION TYPE, COLLISION TYPE ROAD CONDITION, LIGHT CONDITION, WEATHER CONDITION	**Algorithm**	**Jaccard**	**F1-score**	**LogLoss**
	KNN	0.65	0. 63	NA
	Decision Tree	0. 7008	0. 7008	NA
	SVM	0. 7130	0. 6491	NA
	Logistic Regression	0. 7041	0. 7022	0. 5524
JUNCTION TYPE, COLLISION TYPE, ROAD CONDITION, LIGHT CONDITION, WEATHER CONDITION, PERSONCOUNT, VEHCOUNT, PEDCOUNT, PEDCYLCOUNT	**Algorithm**	**Jaccard**	**F1-score**	**LogLoss**
	KNN	0.6428570	0. 631253	NA
	Decision Tree	0. 6233770	0. 623211	NA
	SVM	0. 6583420	0. 639621	NA
	Logistic Regression	0. 6708290	0. 666618	0. 594324
JUNCTION TYPE, COLLISION TYPE, ROAD CONDITION, LIGHT CONDITION, WEATHER CONDITION, INCIDENT TIME BINS	**Algorithm**	**Jaccard**	**F1-score**	**LogLoss**
	KNN	0.639361	0.638639	NA
	Decision Tree	0.652348	0.651094	NA
	SVM	0.656843	0.640126	NA
	Logistic Regression	0.635864	0.626378	0.611718
	Random Forest	0.6454	0.6423	NA

Predicting Car Accident Severity in Seattle

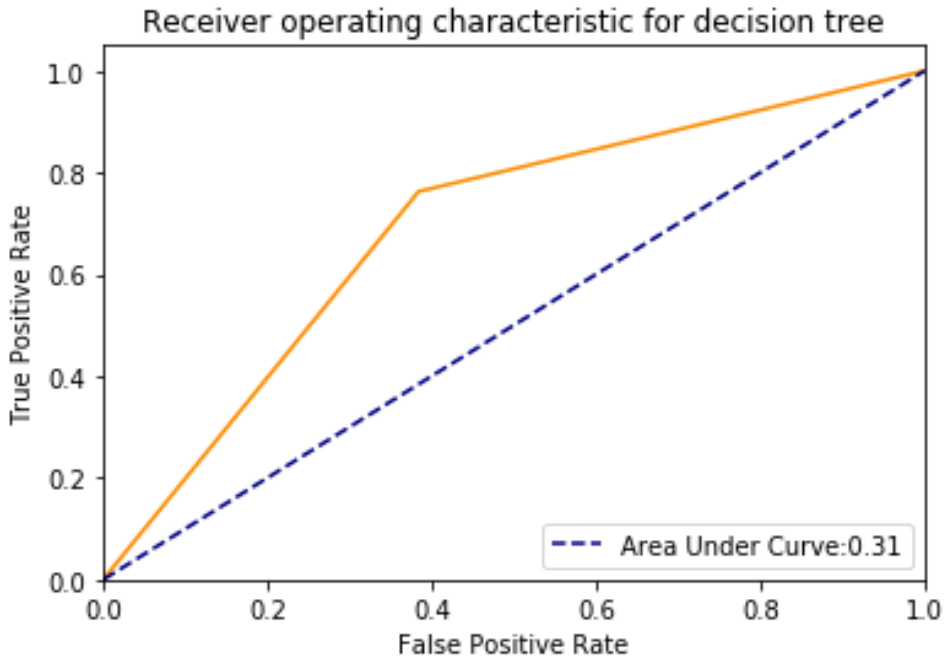
Models were also evaluated with Confusion Matrix, ROC and Precision Recall Scores and Curves. Samples of the graphs are as follows:

```
Decision Tree:
Accuracy: 65.23%
Jaccard: 65.23%
f1_score: 65.11%
              precision    recall  f1-score   support

           1       0.67      0.59      0.63      1001
           2       0.64      0.71      0.67      1001

   micro avg       0.65      0.65      0.65      2002
   macro avg       0.65      0.65      0.65      2002
weighted avg       0.65      0.65      0.65      2002

Confusion matrix:
 [[593 408]
 [288 713]]
```

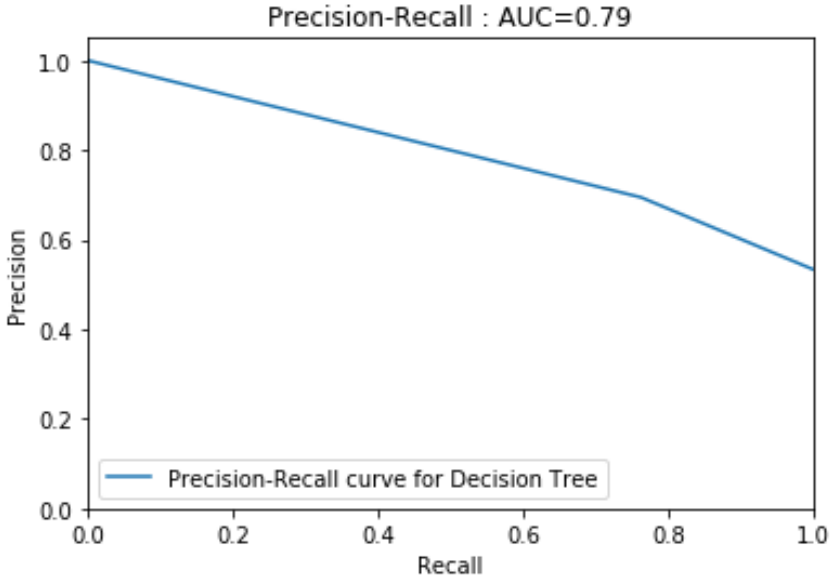


Predicting Car Accident Severity in Seattle

```
0.8101163850996853
Area Under Curve: 0.31
```




```
Decision Tree: Average precision-recall score: 0.40
Area Under Curve: 0.79
```



Predicting Car Accident Severity in Seattle

5. Discussion on Predictive models built and Evaluation

The models built for various feature sets provide a good insight into how the models behave for various features. The models provide better performance when the features are less. As the features become more, performance of the models degraded. Performance of Knn and SVM was very poor for 100,000 records (Train set: (85744) Test set: 21436). Hence, the sample dataset for these models had to be scaled down to 10,000 (8000 Training set, 2000 Testing set). The Evaluation dataset was restricted to 2000 records (1000 in each class).

Original data on the SDOT website has data available for other classes also. Building and running multiclass models with heavily unbalanced data should be tried as this dataset emulates real world scenarios. Handling of such data for machine learning requires better algorithms, techniques like SMOTE and better modeler environments such as SPSS. Newer softwares such as AutoAI can be used to automate prediction of better feature sets for better performance of the models. Gradient Boosting algorithms can be utilized for training the models become better in performance.

We can also infer from the ROC and Precision recall curves that the models are able to predict true positive cases 70 to 80% of the time. We can also see that the performance of the models degrades with new evaluation dataset. Hence, extensive training for the models are imperative for better performance. Models such as Gradient Boosting / XGBoost etc., that can learn from their previous train and test data can be utilized for continuous increase in performance.

6. Conclusion

I am confident such Machine learning models can be deployed in daily activities to improve road safety, vehicle safety and overall people's safety and hence reduce accidents before they occur and also reduce the severity of the accidents. Such machine learning models are already analyzing the roads from traffic cameras to identify dangerous spots and mitigate the risk of accidents (Fucoloro, 2017)

.

Predicting Car Accident Severity in Seattle

Works Cited

Blincoe, L. J. (2015, May). *The economic and societal impact of motor vehicle crashes, 2010. (Revised) (Report No. DOT HS 812 013). .* Washington, DC : National Highway Traffic Safety Administration .

DOT, S. (2018, 4 4). *data-seattlecitygis.opendata.arcgis.com.* Retrieved Oct 2, 2020, from data-seattlecitygis.opendata.arcgis.com: https://data-seattlecitygis.opendata.arcgis.com/datasets/5b5c745e0f1f48e7a53acec63a0022ab_0

Fucoloro, T. (2017, June 2). *Seattlebikeblog.* Retrieved Oct 5, 2020, from www.seattlebikeblog: https://www.seattlebikeblog.com/2017/06/02/using-machine-learning-to-predict-traffic-collisions-in-bellevue-and-how-you-can-help/

References

https://scikit-learn.org/stable/
https://pandas.pydata.org
https://machinelearningmastery.com/roc-curves-and-precision-recall-curves-for-classification-in-python/
https://www.kaggle.com/lct14558/imbalanced-data-why-you-should-not-use-roc-curve
https://stats.stackexchange.com/questions/204589/how-to-make-predictions-using-multiclass-unbalanced-data
https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5
https://elitedatascience.com/imbalanced-classes
https://medium.com/digital-catapult/dealing-with-imbalanced-data-8b21e6deb6cd
https://heartbeat.fritz.ai/working-with-geospatial-data-in-machine-learning-ad4097c7228d