Car Collision Severity Analysis in Seattle

IBM Applied Data Science Projects

Xiaoting (Theresa) Liu

October 23, 2020

Table of Contents

Business Understanding

Car accident happens every day due to many reasons, such as road condition, weather, and lighting, etc. There are some interesting facts from Driver Knowledger.com. There are 6 million car accidents in average in the U.S per year; more than 90 people involved in fatal accidents; 3 million people are injured every year in the U.S. Moreover, the car collisions result in 6% fatality, 27% non-fatality injury and 72% property damage. According to Colburn Law office, a crash occurred in Washington every 4.5 minutes in 2015, and Seattle was the 8th most dangerous city. In order to reduce the frequency of car collisions in Seattle, an analysis is conducted to predict the severity of a car collision that caused by the current weather, road and visibility conditions. The purpose of this analysis is to create a model that can alert drivers when the conditions are dreadful.

Data Understanding

The dataset used for this project is presented by Seattle government, which recorded the collisions have taken place in Seattle since 2004. A full description for each variables can be found here. There are 38 variables and 194,673 observations in this dataset. This project focuses on analyzing the severity of a car collision. Thus, SEVERITYCODE will be used as the dependent variable Y(collision type), and other important variables will be used as the independent variables X (collision cause).

Data Preparation

There are some unrelated variables will be removed first. Then, the missing values will be eliminated. Here are some important variables may have important impact to the target variable Severity Score:

Variables	Description
ADDRTYPE	Collision address type: Alley, Block, Intersection
LOCATION	Description of the general location of the collision
JUNCTIONTYPE	Category of junction at which collision took place helps identify where most collisions occur
WEATHER	A description of the weather conditions during the time of the collision
ROADCOND	The condition of the road during the collision
LIGHTCOND	The light conditions during the collision

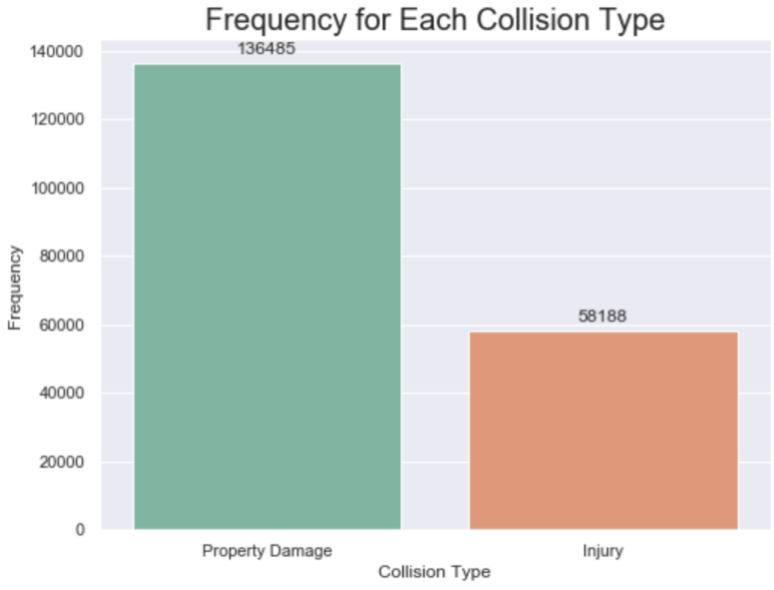
After clearing the missing values, there are total 182,679 records in the new dataset. Except the type of the SEVERITYCODE is integer, the types of other variables are all object.

```
Int64Index: 182679 entries, 0 to 194672
Data columns (total 7 columns):
ADDRTYPE        182679 non-null object
LOCATION        182679 non-null object
JUNCTIONTYPE    182679 non-null object
WEATHER         182679 non-null object
ROADCOND        182679 non-null object
LIGHTCOND       182679 non-null object
SEVERITYCODE    182679 non-null int64
dtypes: int64(1), object(6)
memory usage: 11.1+ MB
```

For further analysis, the dataset has to be encoded to change the categorical values to numerical values. Since the SEVERITYCODE is integer, it will be kept as its original type.

	ADDRTYPE	JUNCTIONTYPE	WEATHER	ROADCOND	LIGHTCOND	SEVERITYCODE
102163	0	4	1	0	5	1
122835	0	4	1	0	2	1
150937	1	1	6	8	2	1
110427	1	1	4	8	5	1
114240	0	4	1	0	5	1

Below is a bar plot to show the frequency of each collision type have been taken place in Seattle. Property damage collision is nearly three times of the injury collision, which means the dataset is unbalanced. The ultimate model will lead to an inaccurate prediction to collision type.



In order to increase the accuracy of prediction, oversampling method will be applied to adjust this issue.

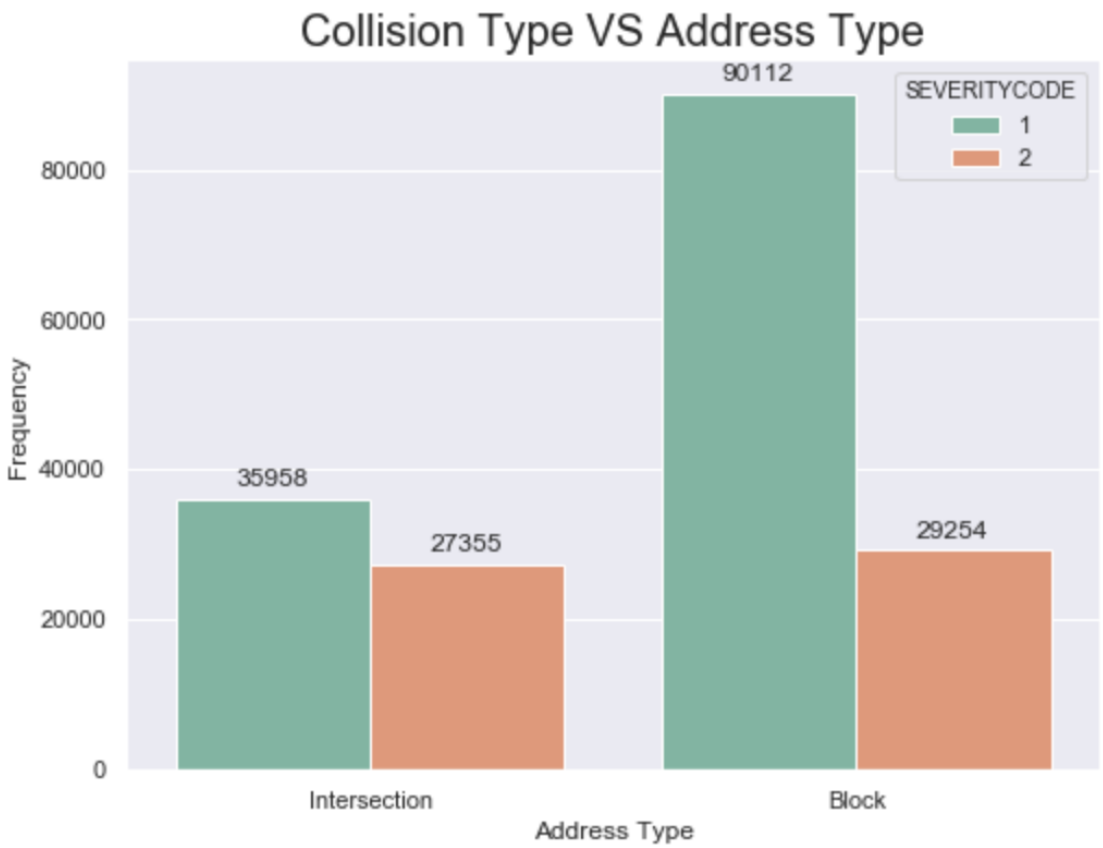
```
length of oversampled data is  85056
Number of property damage only collision in unsampled data is 42528
Number of injury collision is 42528
```

Exploratory Data Analysis

Recall that the Severity Code “1” indicates the Property Damage Only Collision, and the Severity Code “2” indicates the Injury Collision. This section uses Exploratory Data Analysis and some visualization plots to show the relationship between collision types and collision causes.

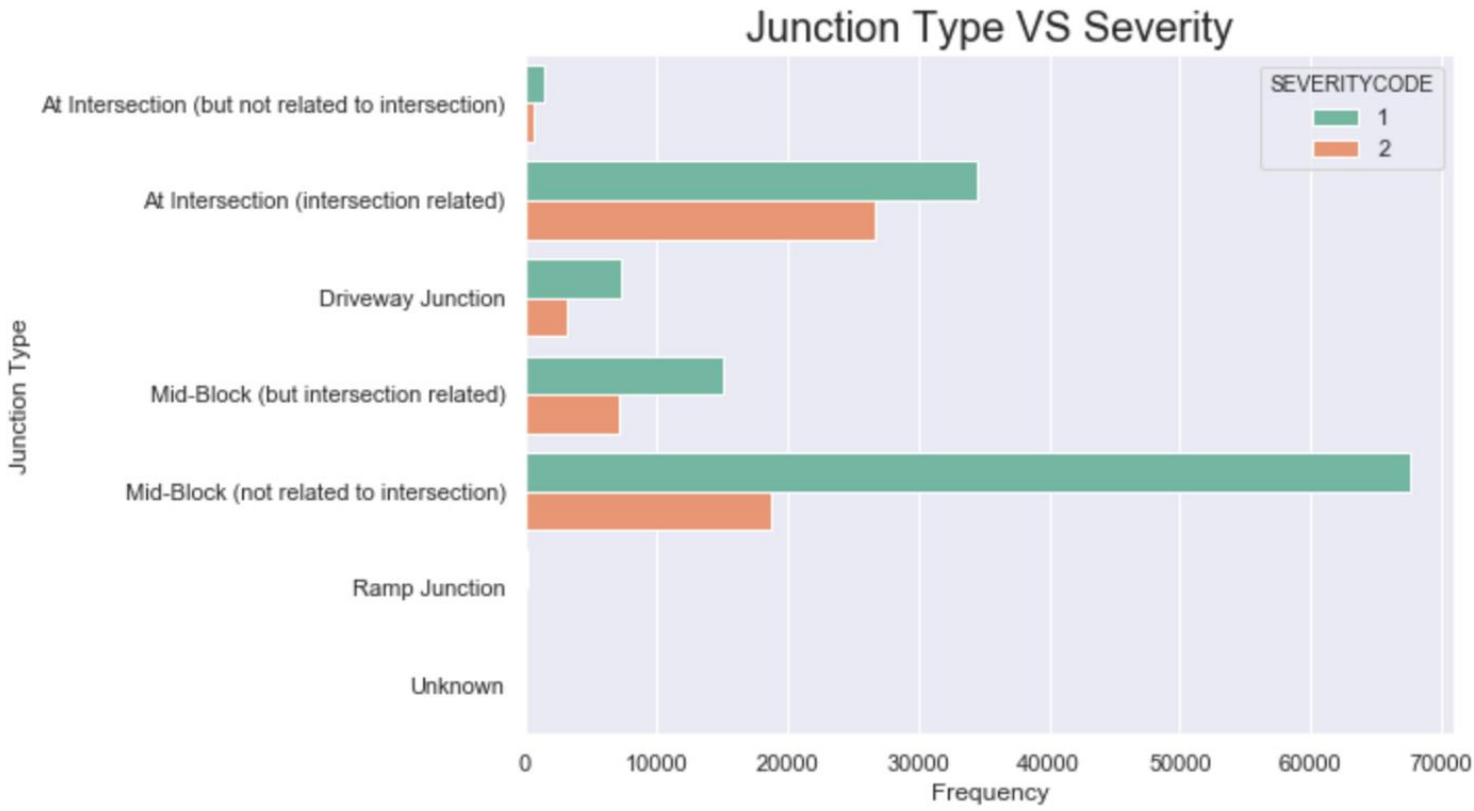
Address Type vs Collision Types

The first bar plot shows the relationship between collision types and address types. Property Damage Only collision occurred 35,958 times in the intersection compared to 90,112 times in the block; whereas Injury collision occurred 27,355 times in the intersection compared to 29,254 times in the block. More accidents took place in the block type address than the intersection type address.



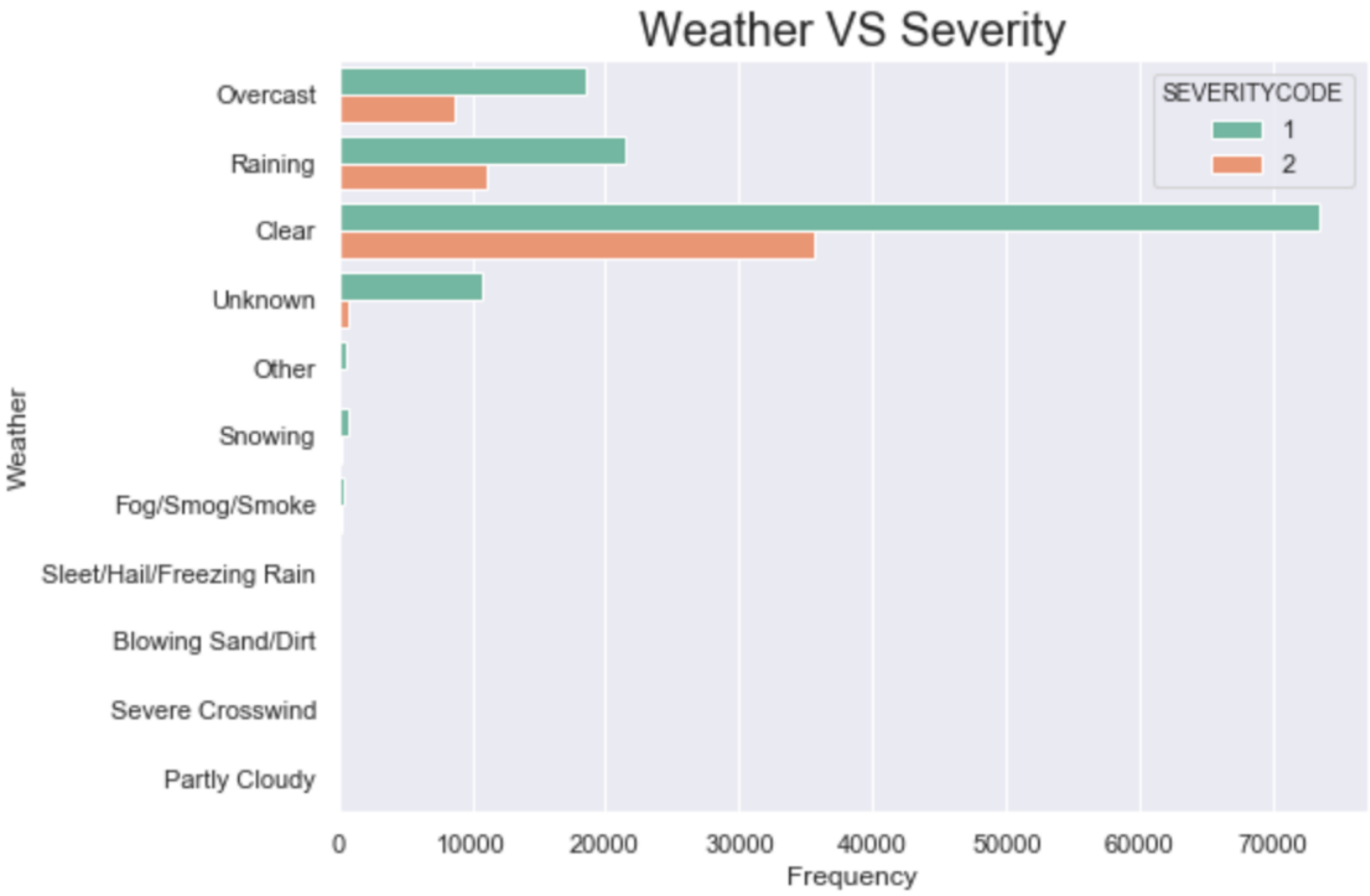
Junction Type vs Collision Type

The second plot indicates the relationship between junction types and collision types. Top three junction types with collisions taken in place are mid-block(not related to intersection), at intersection (intersection related) and mid-block(but intersection related). Among these junctions, Damage Only collision occurred more at mid-block(not related to intersection, compared to Injury collision occurred more at intersection.



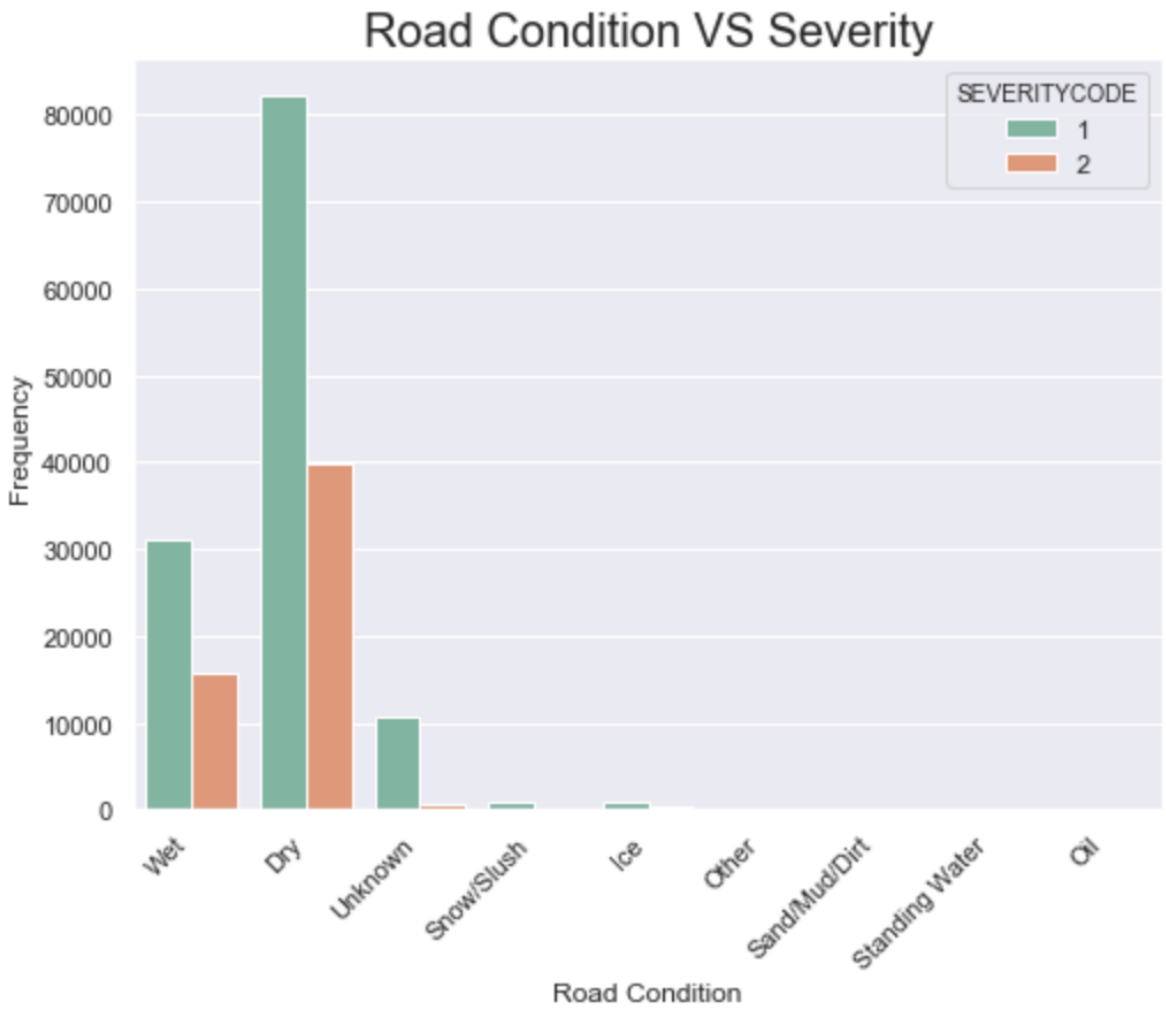
Weather Type vs Collision Type

The third plot indicated the relationship between weather types and collision types. Even though the weather was clear, collisions still occurred more times than overcast and raining days. Same as the previous collision causes, Damage Only collision occurred more than the Injury collision.



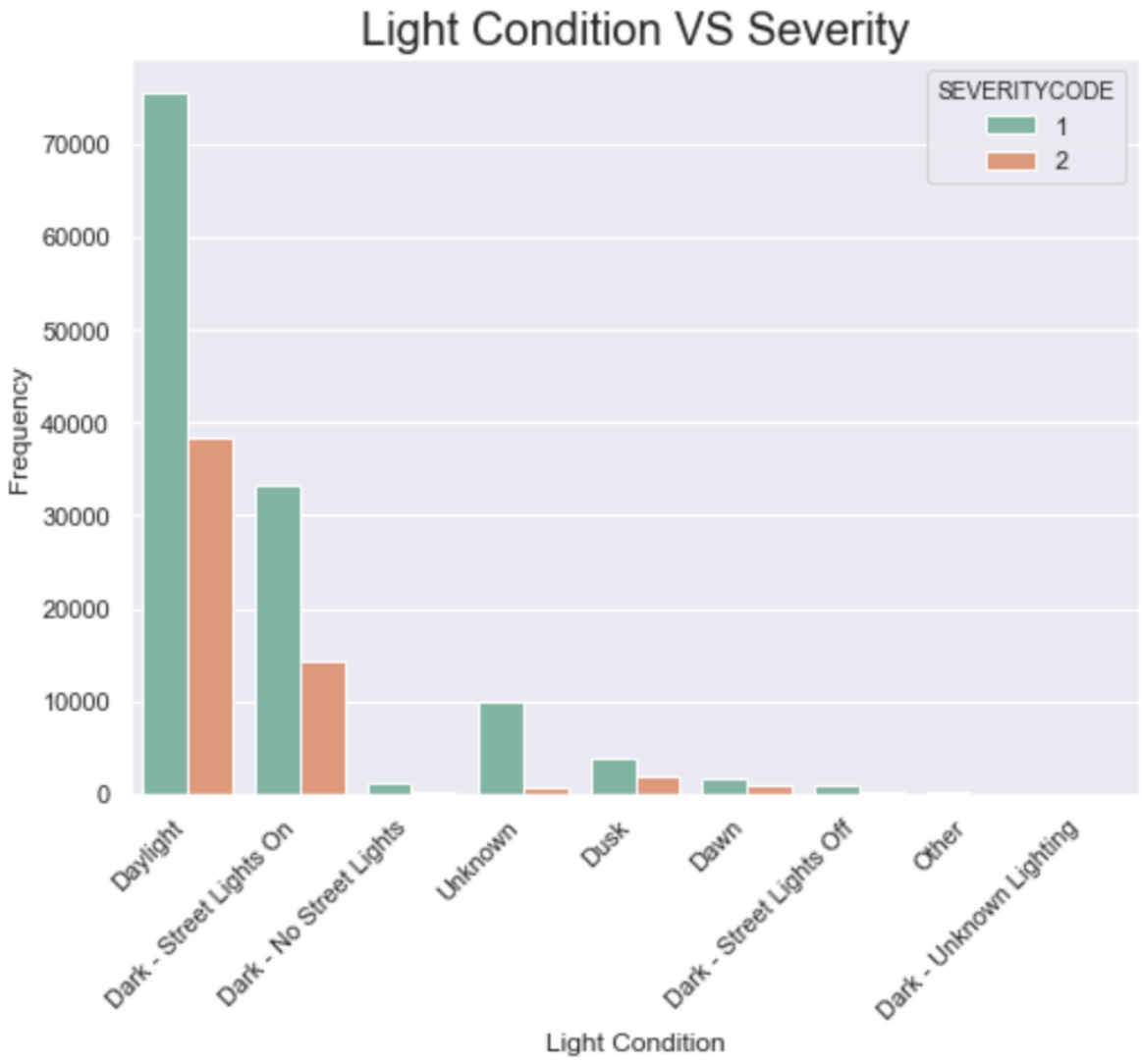
Road Condition vs Collision Type

The fourth plot indicated the relationship between road condition and collision types. Most of the collisions occurred during wet or dry weather. Some collisions were reported with unknown weather condition. Damage Only collisions occurred nearly double of the Injury collision.



Light Condition vs Collision Type

The last plot indicated the relationship between light condition and collision types. Most of collision occurred with daylight condition, followed by dark-street light on. There were small portion of collisions occurred in other light condition.



Methodology

Three machine learning algorithms chosen for this dataset are:

1. Binary Logistic Regression: It is a supervised learning classification algorithm used to predict the probability of a target variable.
2. Decision Tree: It is a supervised machine learning classification algorithm that can split the data into subset based on an attribute value test.
3. K-Nearest Neighbors: It relies on labeled input data to learn a function that produces an appropriate output when given new data. It helps to predict the severity code by finding the most similar to data point within k distance.

Some variables were removed for this analysis. Except the Severity Code. Other variables are categorical variables. For further analysis, it has to be encoded to numerical variable. Label encoded method was applied to encode these categorical variables.

	ADDRTYPE	JUNCTIONTYPE	WEATHER	ROADCOND	LIGHTCOND	SEVERITYCODE
102163	0	4	1	0	5	1
122835	0	4	1	0	2	1
150937	1	1	6	8	2	1
110427	1	1	4	8	5	1
114240	0	4	1	0	5	1

The dataset was split to 30% for testing and 70% for training.

```
Test Set (25517, 5) (25517,)
Tranning Set (59539, 5) (59539,)
```

Evaluation

Classification Report and Accuracy Score are used to evaluate each algorithm. In the Classification Report, “0” represents the Damage Only collision and “1” represents the Injury collision.

Logistic Regression

	precision	recall	f1-score	support
0	0.66	0.60	0.62	14240
1	0.54	0.61	0.57	11277
micro avg	0.60	0.60	0.60	25517
macro avg	0.60	0.60	0.60	25517
weighted avg	0.61	0.60	0.60	25517

```
Accuracy of logistic regression classifier on the test set is: 0.60
```

Decision Tree

	precision	recall	f1-score	support
0	0.57	0.62	0.59	11912
1	0.64	0.59	0.62	13605
micro avg	0.61	0.61	0.61	25517
macro avg	0.61	0.61	0.61	25517
weighted avg	0.61	0.61	0.61	25517

```
Accuracy of decision tree  classifier on the test set is: 0.61
```

K-Nearest Neighbors

	precision	recall	f1-score	support
0	0.64	0.59	0.61	13882
1	0.55	0.60	0.57	11635
micro avg	0.59	0.59	0.59	25517
macro avg	0.59	0.59	0.59	25517
weighted avg	0.60	0.59	0.59	25517

```
Accuracy of decision tree  classifier on the test set is: 0.59
```

Precision:

It is defined as the number of true positives divided by the number of true positives plus the number of false positives. True positives are data point classified as positive by the model that actually are positive, meaning the prediction is correct as the actual value. False positives, in our example, is the Severity Code the model classifies as “1 (Damage Only collision)” that are not.

Recall:

It is defined as the number of true positives divided by the number of true positives plus the number of false negatives, and false negatives are the Severity Code the model identifies as “2 (Injury collision) that actually are “1 (Damage Only collision)”.

F1-score:

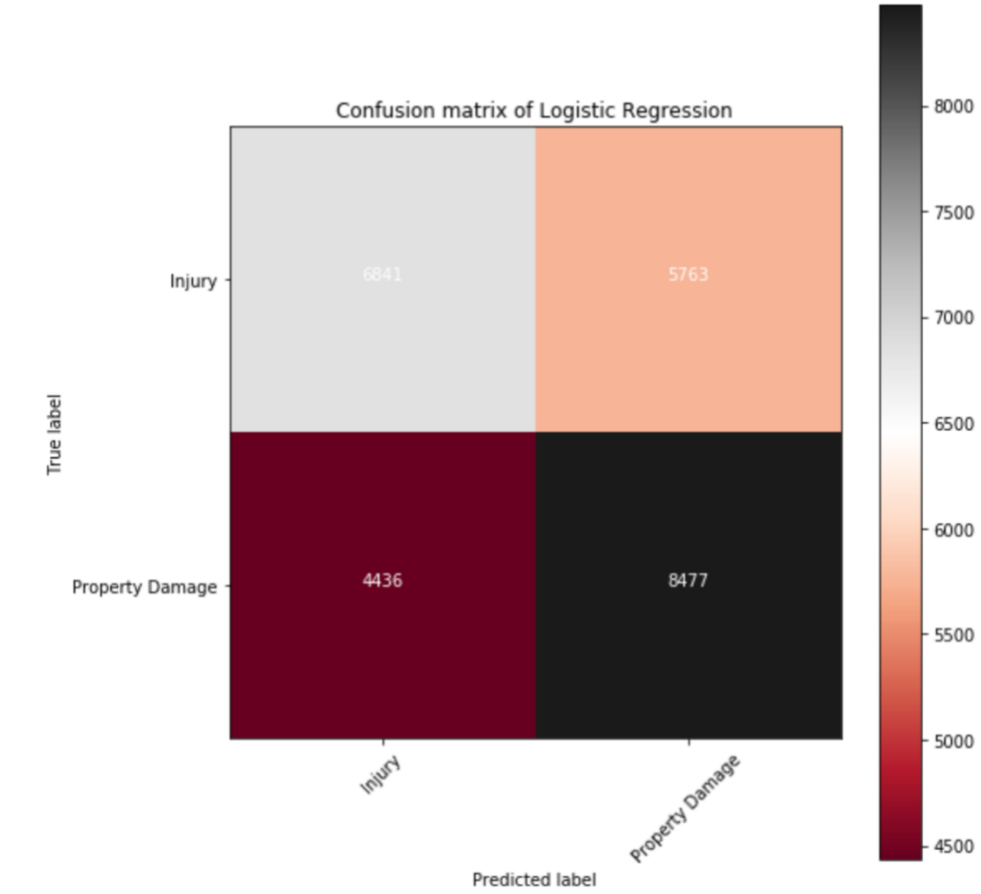
Sometimes, one classification model has better precision and the other classification model has better recall. In such a contradictory situation, F1 score is taken into account for the evaluation. The higher the F1 score, the better the classifier.

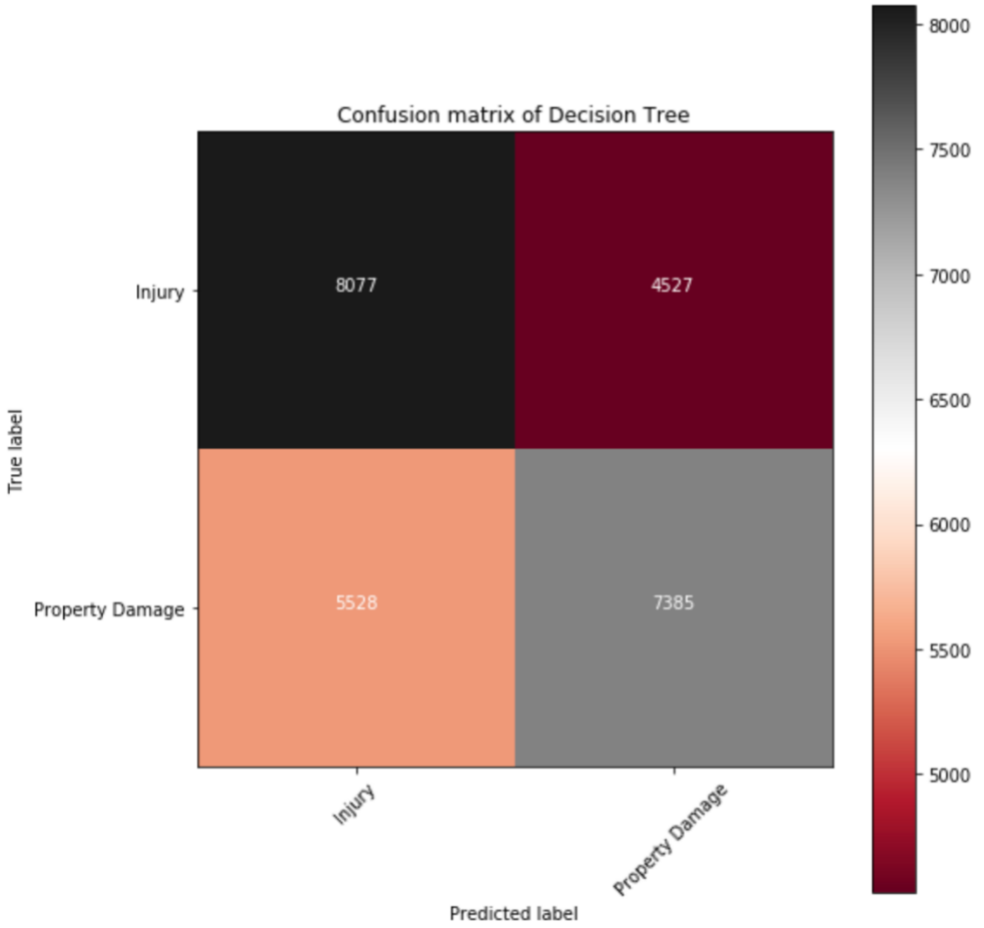
Accuracy Score:

Accuracy Score is most common evaluation index, and it is easy to understand that the number of correct predications divided by the number of samples, generally speaking, the higher the accuracy, the better the classifier.

Result

Looking at the Precision index and F-1 Score, Logistic Regression model has better prediction on Damage Only collision, while Decision Tree has better predication on Injury collision. Looking at the Recall index, all three models have similar prediction on both type collision. KNN is not the best model for prediction for this particular dataset. Accuracy Score indicates that Decision Tree has better prediction than Logistic Regression. Blow are the Confusion Matrix of Logistic Regression and Decision Tree. After evaluating the models, Decision Tree the most suitable model to predict the collision type.


Confusion matrix of Logistic Regression
Injury
Property Damage
True label
6841
5763
4436
8477
Injury
Property Damage
Predicted label
8000
7500
7000
6500
6000
5500
5000
4500

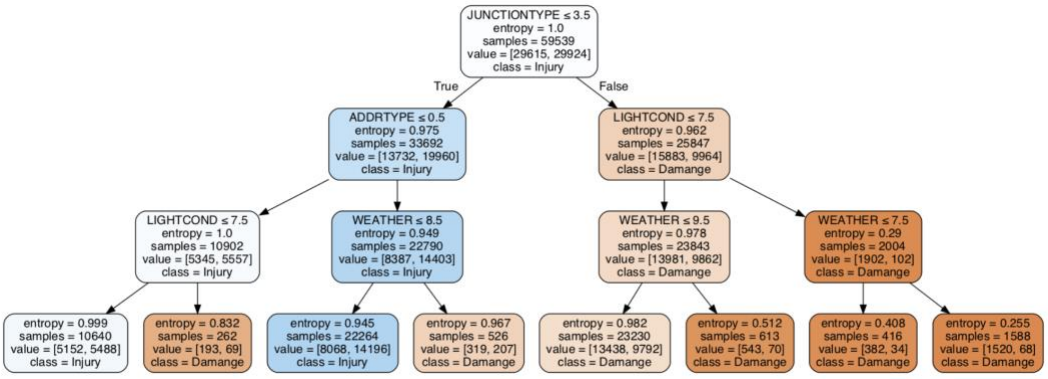

Confusion matrix of Decision Tree
Injury
Property Damage
True label
8077
4527
5528
7385
Injury
Property Damage
Predicted label
8000
7500
7000
6500
6000
5500
5000

The Logistic Regression Model shows that except the variable Light Condition, all other variables have a P-value less than 0.05. This indicates that Address Type, Junction Type, Weather and Road Condition have significant impact to the collision.

```
Optimization terminated successfully.
         Current function value: 0.666359
         Iterations 4
                         Results: Logit
=================================================================
Model:              Logit            Pseudo R-squared: 0.039
Dependent Variable: SEVERITYCODE     AIC:              113365.6796
Date:               2020-10-22 14:06 BIC:              113412.4350
No. Observations:   85056            Log-Likelihood:   -56678.
Df Model:           4                LL-Null:          -58956.
Df Residuals:       85051            LLR p-value:      0.0000
Converged:          1.0000           Scale:            1.0000
No. Iterations:     4.0000
-----------------------------------------------------------------
               Coef.   Std.Err.     z      P>|z|    [0.025   0.975]
-----------------------------------------------------------------
ADDRTYPE       0.7773    0.0178   43.5921  0.0000   0.7424   0.8123
JUNCTIONTYPE  -0.0304    0.0056   -5.4753  0.0000  -0.0413  -0.0195
WEATHER       -0.1064    0.0043  -24.5447  0.0000  -0.1149  -0.0979
ROADCOND       0.0401    0.0030   13.1610  0.0000   0.0341   0.0461
LIGHTCOND     -0.0069    0.0042   -1.6492  0.0991  -0.0152   0.0013
=================================================================
```

In Decision Tree chart, each internal node has a decision rule that splits the data. Entropy referred as measure of disorder, which measures the impurity of the node. A high level of disorder means a low level of purity. The original tree depth selected for the Decision Tree model was 6. Image can be viewed here.

To reduce the complexity of the model, the tree depth was changed to 3 while the accuracy score still remained the same.




```
Accuracy: 0.6032840851197241
```

Conclusion

This analysis focused on analyzing diffent type of variables that can cuase Danage Only collision or Injury collision. The dataset was not in balance at the beginning, oversampling method was applied to get a balance dataset. EDA analysis indicated that total cases of Damage Only collison were triple more than the Injury collision. More collision occurred in block type address, mid-block(not related to intersection), clear weather condition, dry road condition and daylight condition. After label encoding the categorical variables, there were three machine learning algorithms used for prediction analysis: Logistic Regression, Decision Tree and K-Nearest Neighbor. Classification report and confusion matrix were used for evaluation. KNN is not the best model in this situation compared to the other two models. Logistic Regression model had better prediction on Damage Only collision with 60% accuracy, and it indicated that the Light Condition did not have a big impact to the collision at all. Decision Tree had better predication on Injury collision with 61% accuracy, and thus it has better prediction than other two models.

To reduce the complexity of the model, the tree depth was reset to 3 and thus provided a better understanding when it comes to future planning and implementation.

Recommendation

Damage Only collision tends to happen more at mid-block (not related to intersection) and ramp junction type address. At Intersection (intersection related), mid-block (but intersection related), driveway junction and at intersection (but not related to intersection area, drivers should pay more attention to the weather condition. These are the places the Injury collision might happen. When the drivers are approaching in these areas, they should be alerted to drive slowly. Seattle city should implement more facilities to help reduce the Injury collision, especially in the block type address. For instance, this can be done by placing stop signs, speed reducing signs and traffic lights around these areas.

Reference

https:/ /www.driverknowledge.com/car-accident-statistics/

https://www.colburnlaw.com/seattle-traffic-accidents/