Predicting the Car Accident Severity

Regina Rhodelia B. Monasterial
October 4, 2020

- **Introduction**

Road accidents can occur to anyone anytime anywhere. Fortunately, data on road accidents are continuously collected and are made available to everyone. We want to understand and maximize on the value of the available data in order to help us prevent road accidents – that when given information on location and other external conditions, the possibility and severity of an accident can be predicted and prevented thereby assisting local government units in publishing this information to travelers along their local areas.

- **Data**

The city of Seattle has a dataset on all collisions collected since 2004 to present, as provided by the Seattle Police District and recorded by Traffic Records. It has 194,673 observations with 37 attributes. It has been downloaded as a csv file –from this link:

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

The description of the attributes as lifted DataCollisions.csv from the metadata of the dataset.

Attribute	Data Type	Data Length	Description
ADDRTYPE	Text	12	Collision address type: Alley, Block, Intersection
COLDETKEY	Long		Secondary key for the incident.
COLLISIONTYPE	Text	300	Collision type.
CROSSWALKKEY	Long		A key for the crosswalk at which the collision occurred.
EXCEPTRSNCODE	Text	10	
EXCEPTRSNDESC	Text	300	
FATALITIES	Double		The number of fatalities in the collision. This is entered by the state.
HITPARKEDCAR	Text	1	Whether or not the collision involved hitting a parked car. (Y/N)
INATTENTIONIND	Text	1	Whether or not collision was due to inattention. (Y/N)
INCDATE	Date		The date of the incident.
INCDTTM	Text	30	The date and time of the incident.
INCKEY	Long		A unique key for the incident.
INJURIES	Double		The number of total injuries in the collision. This is entered by the state.
INTKEY	Double		Key that corresponds to the intersection associated with a collision.
JUNCTIONTYPE	Text	300	Category of junction at which collision took place.
LIGHTCOND	Text	300	The light conditions during the collision.
LOCATION	Text	255	Description of the general location of the collision.
OBJECTID	ObjectID		ESRI unique identifier.
PEDCOUNT	Double		The number of pedestrians involved in the collision. This is entered by the state.
PEDCYLCOUNT	Double		The number of bicycles involved in the collision. This is entered by the state.
PEDROWNOTGRNT	Text	1	Whether or not the pedestrian right of way was not granted. (Y/N)
PERSONCOUNT	Double		The total number of people involved in the collision.
ROADCOND	Text	300	The condition of the road during the collision.
SDOT_COLCODE	Text	10	A code given to the collision by SDOT.
SDOT_COLDESC	Text	300	A description of the collision corresponding to the collision code.

SDOTCOLNUM	Text	10	A number given to the collision by SDOT.
SEGLANEKEY	Long		A key for the lane segment in which the collision occurred.
SERIOUSINJURIES	Double		The number of serious injuries in the collision. This is entered by the state.
SEVERITYCODE	Text	100	A code that corresponds to the severity of the collision: [3 – fatality, 2b - serious injury, 2 – injury, 1 – prop, damage, 0 – unknown]
SEVERITYDESC	Text		A detailed description of the severity of the collision.
SHAPE	Geometry		ESRI geometry field.
SPEEDING	Text	1	Whether or not speeding was a factor in the collision. (Y/N)
ST_COLCODE	Text	10	A code provided by the state that describes the collision. For more information about these codes, please see the State Collision Code Dictionary.
ST_COLDESC	Text	300	A description that corresponds to the state's coding designation.
UNDERINFL	Text	10	Whether or not a driver involved was under the influence of drugs or alcohol.
VEHCOUNT	Double		The number of vehicles involved in the collision. This is entered by the state.
WEATHER	Text	300	A description of the weather conditions during the time of the collision.

To help in predicting the possibility and severity of an accident/collision [SEVERITYCODE], I decided to:

- use the information on the collision address type [ADDRTYPE] as source of location information, and
- select the following to form part of information on external conditions:
    - weather during the time of collision - WEATHER
    - light conditions during the collision - LIGHTCOND
    - the condition of the road during collision - ROADCOND
    - week of day as extracted from the date of collision - INCDATE

- **Methodology**

As the entire data from the DataCollisions.csv file is loaded into a dataframe, those columns which are not part of above-mentioned variables were dropped.

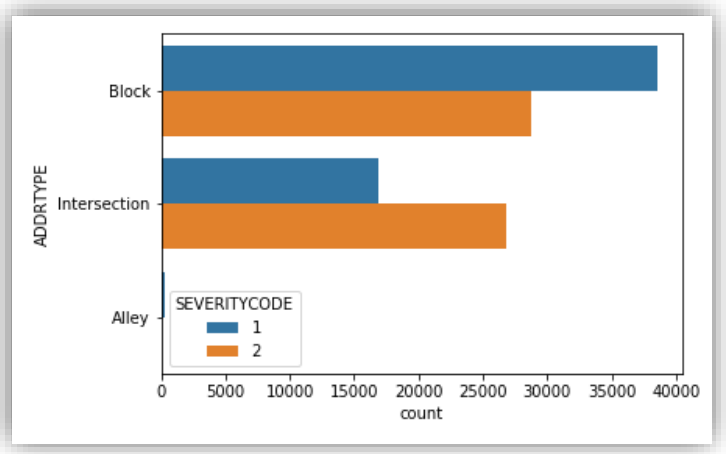
I observed that there are rows with NaN values. These rows were dropped for better processing of sklearn. On the other hand, there are unknown data on WEATHER, LIGHTCOND and ROADCOND as encoded, and these were dropped as well. These steps brought the total number of rows to 169,781.

Data on SEVERITYCODE only included those that resulted to property damage only [SEVERITY CODE 1] and injury [SEVERITYCODE 2]. Transforming the target variable SEVERITYCODE to balance the data prior to modelling processing resulted to our dataset with the following breakdown.

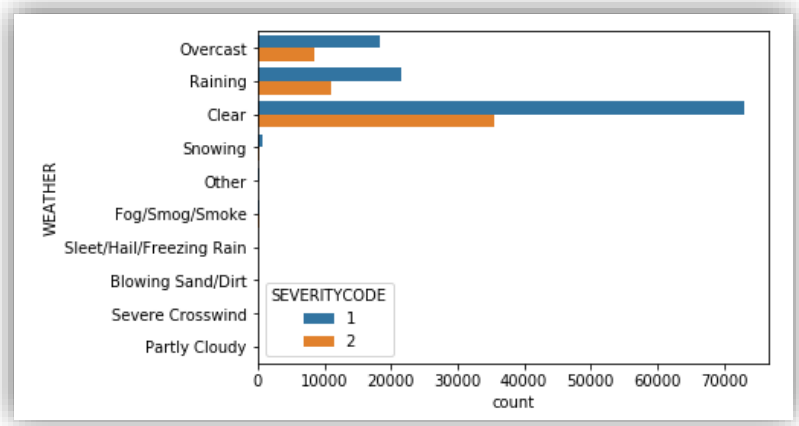
	SEVERITYCODE
2	55707
1	55707

Some observations while doing separate exploratory analysis on the independent variables:

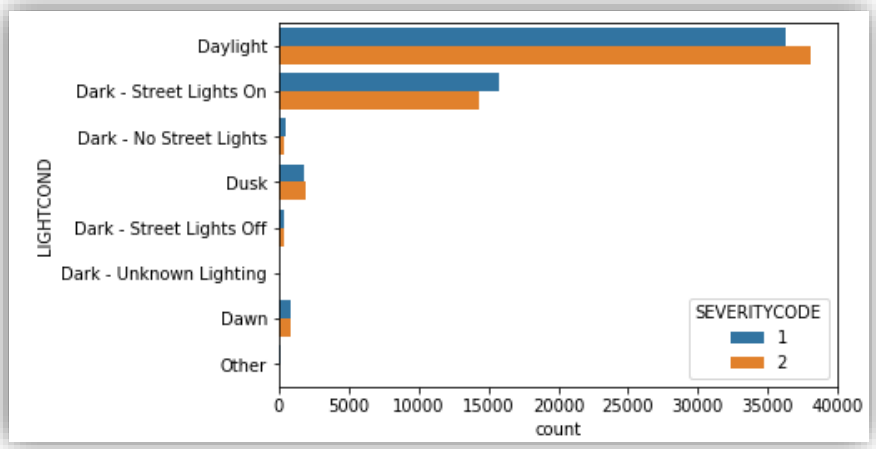
- There are more Severity Code 1 collisions along Blocks than on Intersections. Collisions of Severity Code 2 registered almost the same count on both Blocks and Intersections.



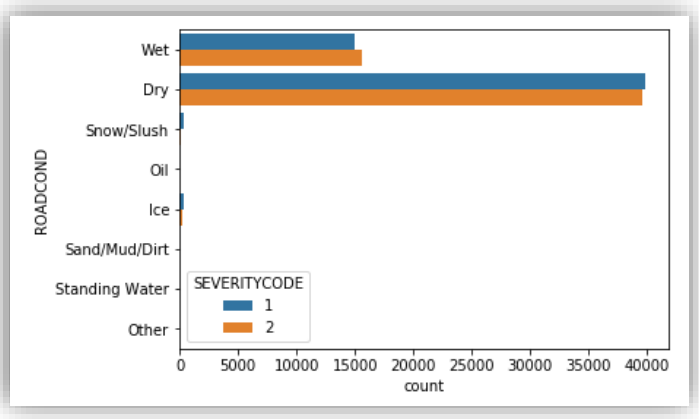
- Most number of collisions happen on Clear Weather.



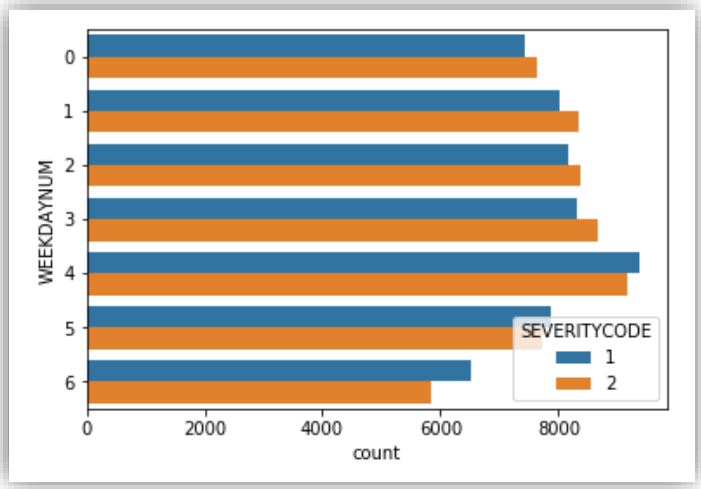
- Most of the collisions happen on daylight.



- Most of the collisions happen on dry roads.

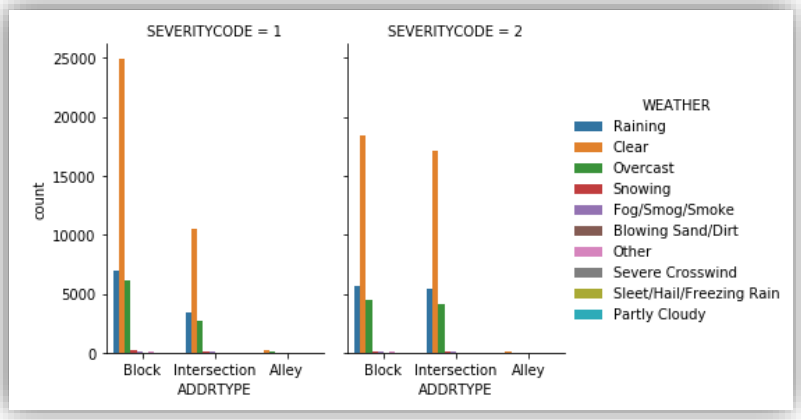


- There is less collisions during weekend and surprisingly, there is an upward trend on collisions on weekdays.

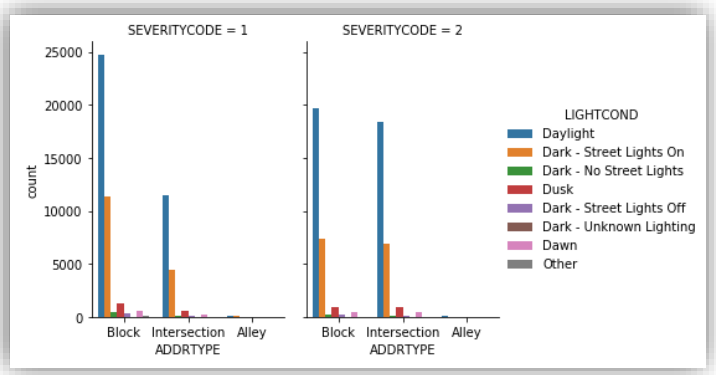


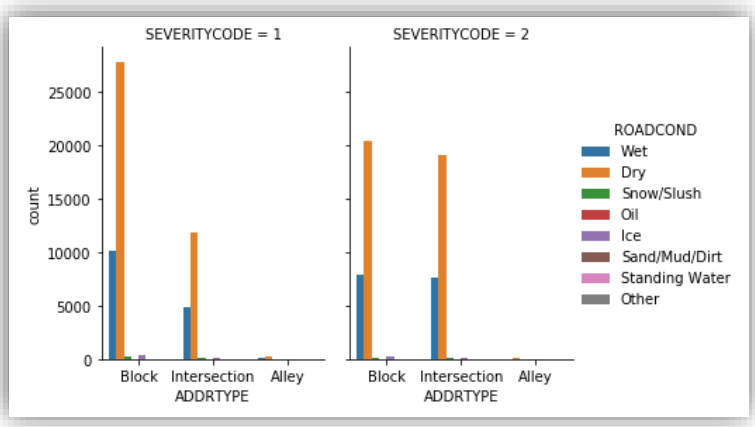
On the other hand, some observations while correlating the variables are as follows:

- Most collisions happen on blocks on a clear weather.



- Collisions happen on dry roads that are well lighted, mostly on blocks.





Then, all selected columns are converted to the format that can be fed into machine learning models, using one-hot encoding to label them - converting categorical features to numerical values. The feature set and the label were then defined. Normalization of the data are performed before feeding them to our algorithm processing.

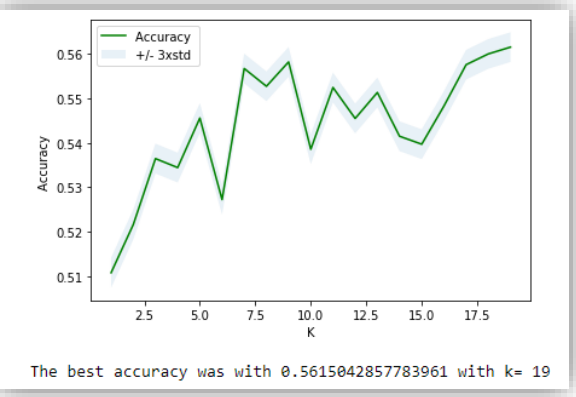
Twenty percent of the data was used as test data, and the rest as training data. Test set has 22,283 rows, while training set has 89,131 rows.

```
from sklearn.model_selection import train_test_split
X_train, X_test, Y_train, Y_test = train_test_split( X, Y, test_size=0.2, random
_state=4)
print ('Train set:', X_train.shape,  Y_train.shape)
print ('Test set:', X_test.shape,  Y_test.shape)
```

```
Train set: (89131, 5) (89131,)
Test set: (22283, 5) (22283,)
```

K-Nearest Neighbor, Support Vector Machine, Decision Tree and Logistic Regression are selected for our machine learnings to predict the severity. Training the model then proceeded.

The best accuracy for the KNN algorithm processing was with K=19 using the test data with the accuracy given at 0.56.




```
KNeighborsClassifier(algorithm='auto', leaf_size=30, metric='minkowski',
           metric_params=None, n_jobs=None, n_neighbors=19, p=2,
           weights='uniform')
```

Using the Decision Tree on the training set yielded the result:

```
SVC(C=1.0, cache_size=200, class_weight=None, coef0=0.0,
  decision_function_shape='ovr', degree=3, gamma='auto_deprecated',
  kernel='rbf', max_iter=-1, probability=False, random_state=None,
  shrinking=True, tol=0.001, verbose=False)
```

While that of Logistic Regression would show:

```
LogisticRegression(C=0.01, class_weight=None, dual=False, fit_intercept=True,
          intercept_scaling=1, max_iter=100, multi_class='warn',
          n_jobs=None, penalty='l2', random_state=None, solver='liblinear',
          tol=0.0001, verbose=0, warm_start=False)
```

- **Results**

Plotted are the results of our model evaluation using the test data.

Algorithm	Jaccard	F1-score	LogLoss
KNN	0.5615	0.5605	NA
Decision Tree	0.5896	0.5844	NA
SVM	0.5898	0.5847	NA
LogisticRegression	0.5896	0.5845	0.675

Using the F1-score of our algorithms, SVM gave the highest result, though not that significant difference from the other F1-scores.

The accuracy of the Logistic Regression is based on the Logistic Loss of 0.675.

The result is not as good as we expected because the accuracy of the models is not very high.

- **Discussion**

Having selected the collision address type, weather condition, road condition, light condition, date of the incident as our independent variables to predict the severity of a collision gave us valuable insight on how most of the collisions occur on the following conditions, though non-inclusive:

- Clear Weather
- Dry Road
- Daylight

Analysing the time of the incident would tell us that there is less collisions during weekend and surprisingly, there is an upward trend on collisions on weekdays.

Looking at the address/location of the collision, there are more Severity Code 1 collisions along Blocks than on Intersections. Collisions of Severity Code 2 registered almost the same count on both Blocks and Intersections.

With these observations, we cannot overemphasize the importance of safety and vigilance even at the most ideal driving situations, encouraging local governments in ensuring that protection and order on roads are in place.

- **Conclusion**

We were able to show how accident can be predicted by using collected and available data on collision. Although our analysis has given us some good insights, the accuracy of our models is not that optimum. Perhaps this can be improved by considering other features for analysis.