Seattle Traffic – Accident and Collision Analysis Report

Coursera Capstone Project – September 2020

Introduction / Business Understanding

Everyone who commutes daily to work would know that traveling can be a stressful and time wasting activity if not planned properly. Traveling time is one of the main factors that determine how pleasant your commute to and from work would be. One of the factors that needs to be taken into account is facing a Congestion/Traffic Jam because of an accident that took place on the route you take. The aim of this project is to see if we can build a model to be able to predict the severity of an accident taking place taking into account different environmental/traffic/geographical factors. This will be able to help both, law enforcing agencies as well as daily commuters.

- **Law enforcing agencies** stand to gain by being able to proactively avert such accidents if a certain set of conditions arrive and being able to take appropriate actions if and when it does, so that they can ensure minimum impact on traffic flow.

- **Commuters** stand to gain by being forewarned about the accidents and planning/rerouting their journey accordingly. They can also be more vigilant in certain conditions that are prone to accidents.

In the end, we all stand to gain collectively as a society as we will have less accidents, safer roads, less pollution (noise and air) due to less traffic jams and an over improvement in daily commute both in terms of time and stress.

Data

To realize the solution to the problem at hand, we needed an appropriate data source that contains data on past incidents, the conditions they took place in and outcomes, related to traffic related accidents. We got a data source from the Government of Seattle Website (https://data.seattle.gov/Land-Base/Collisions/9kas-rb8d) that contains the latest dataset for us to analyze and build a model to be able to predict the desired results.

For the date to make sense, we would also need to know what each attribute/column means and what data does it contain. The details were available at https://www.seattle.gov/Documents/Departments/SDOT/GIS/Collisions_OD.pdf.

There are a total of **40 Variables** and **221267 Data points/Observations**. Looking at the data set, we see some columns that look useful, including

- **LOCATION** - Description of the general location of the collision
- **SEVERITYCODE** - A code that corresponds to the severity of the collision (3—fatality, 2b—serious injury, 2—injury, 1—prop damage, 0—unknown)
- **SEVERITYDESC** - A detailed description of the severity of the collision
- **JUNCTIONTYPE** - Category of junction at which collision took place
- **UNDERINFL** - Whether or not a driver involved was under the influence of drugs or alcohol
- **INCDTTM** - The date and time of the incident (Time of the day might be of importance here)

- **WEATHER** - A description of the weather conditions during the time of the collision
- **ROADCOND** - The condition of the road during the collision
- **LIGHTCOND** - The condition of the road during the collision

Our Dependent/Predicted Variable will be **SEVERITYCODE** and during data processing and subsequent stages, we will go into in-depth analysis to see how each independent variable varies/is related to the dependent variable.

Note that data filtering will be needed to remove unwanted Columns/Variables and to remove and Null/Empty/Unwanted data observations. We will also need to do other data processing steps such a type casting, standardization, dummy variable creation etc.

Methodology

The following section will have details on the methodology used, including

- Exploratory data analysis
- Data Cleaning
- Feature Selection data analysis
- Model Development
- Accuracy Calculation

Exploratory data analysis, Data Cleaning and Feature Selection

One of the most crucial aspects of having an accurate and meaningful model is to be able to select the most meaningful inputs towards the prediction. We will analyze the data set to see:

- If there are observations with not enough data (invalid/empty data points)
- Go through the description of the features to see if we can remove/delete any unnecessary columns
- If the remaining data points have any correlation to our dependent/predictor variable

Let's start by taking a look at the dataset.

	X	Y	OBJECTID	INCKEY	COLDETKEY	REPORTNO	STATUS	ADDRTYPE	INTKEY	LOCATION	...	ROADCOND	LIGHTCOND	PEDROWNOTGRNT	SDOTCOLNUM	SPEEDING	ST_COLCODE	ST_COLDESC	SEGLANEKEY	CROSSWALKKEY	HITPARKEDCAR
0	-122.320757	47.609408	1	328476	329976	EA08706	Matched	Block	NaN	BROADWAY BETWEEN E COLUMBIA ST AND BOYLSTON AVE	...	Wet	Dark - Street Lights On	NaN	NaN	NaN	11	From same direction - both going straight - bo...	0	0	N
1	-122.319561	47.662221	2	328142	329642	EA06882	Matched	Block	NaN	8TH AVE NE BETWEEN NE 45TH E ST AND NE 47TH ST	...	Dry	Daylight	NaN	NaN	NaN	32	One parked--one moving	0	0	Y
2	-122.327525	47.604393	3	20700	20700	1181833	Unmatched	Block	NaN	JAMES ST BETWEEN 6TH AVE AND 7TH AVE	...	NaN	NaN	NaN	4030032.0	NaN	NaN	NaN	0	0	N
3	-122.327525	47.708622	4	332126	333626	M16001640	Unmatched	Block	NaN	NE NORTHGATE WAY BETWEEN 1ST AVE NE AND NE NOR...	...	NaN	NaN	NaN	NaN	NaN		NaN	0	0	N
4	-122.292120	47.559009	5	328238	329738	3857118	Unmatched	Block	NaN	M L KING JR ER WAY S BETWEEN S ANGELINE ST AND...	...	NaN	NaN	NaN	NaN	NaN		NaN	0	0	N

At this point, we can see some columns having NaN which means we have empty data. It would be worthwhile to see how many data points per feature are null/empty as a percentage of the total data points. Calculating for empty cells, we have the following distribution

```
PEDROWNOTGRNT       97.654807
SPEEDING            95.515586
EXCEPTRSNDESC       94.679501
```

```
INATTENTIONIND     86.364273
INTKEY             67.530455
EXCEPTRSNCODE      54.385268
SDOTCOLNUM         42.542312
LIGHTCOND          11.973946
WEATHER            11.933746
ROADCOND           11.897158
COLLISIONTYPE      11.847924
ST_COLDESC         11.847924
UNDERINFL          11.838890
JUNCTIONTYPE        5.407676
ST_COLCODE          4.251792
X                   3.374603
Y                   3.374603
LOCATION            2.072370
ADDRTYPE            1.676687
SDOT_COLCODE        0.000452
SEVERITYCODE        0.000452
SDOT_COLDESC        0.000452
OBJECTID            0.000000
INCKEY              0.000000
COLDETKEY           0.000000
REPORTNO            0.000000
STATUS              0.000000
HITPARKEDCAR        0.000000
SEVERITYDESC        0.000000
PERSONCOUNT         0.000000
PEDCOUNT            0.000000
PEDCYLCOUNT         0.000000
VEHCOUNT            0.000000
CROSSWALKKEY        0.000000
SERIOUSINJURIES     0.000000
FATALITIES          0.000000
INCDATE             0.000000
INCDTTM             0.000000
SEGLANEKEY          0.000000
INJURIES            0.000000
```

It would make no sense to use features that have NaN/missing data more than 40% of the sample points. Hence, we will drop these from the dataset.

The next step would be to go through the data description. We can see some columns are of no use towards our analysis since they contain with codes related to the accident or reporting related to the city laws. Therefore, we will drop them as well. The following is the list of features dropped.

- OBJECTID
- INCKEY
- COLDETKEY
- REPORTNO
- STATUS
- SDOT_COLCODE
- SDOT_COLDESC
- ST_COLCODE

- ST_COLDESC
- HITPARKEDCAR

We took another look at the data to see that we were left with 23 attributes.

	X	Y	ADDRTYPE	LOCATION	SEVERITYCODE	SEVERITYDESC	COLLISIONTYPE	PERSONCOUNT	PEDCOUNT	PEDCYLCOUNT	...	FATALITIES	INCDATE	INCDTTM	JUNCTIONTYPE	UNDERINFL	WEATHER	ROADCOND	LIGHTCOND	SEGLANEKEY	CROSSWALKKEY
0	-122.320757	47.609408	Block	BROADWAY BETWEEN E COLUMBIA ST AND BOYLSTON AVE	1	Property Damage Only Collision	Sideswipe	2	0	0	...	0	2020/01/22 00:00:00+00	1/22/2020 3:21:00 PM	Mid-Block (not related to intersection)	N	Raining	Wet	Dark - Street Lights On	0	0
1	-122.319561	47.662221	Block	8TH AVE NE BETWEEN NE 45TH E ST AND NE 47TH ST	1	Property Damage Only Collision	Parked Car	2	0	0	...	0	2020/01/07 00:00:00+00	1/7/2020 8:00:00 AM	Mid-Block (not related to intersection)	N	Clear	Dry	Daylight	0	0
2	-122.327525	47.604393	Block	JAMES ST BETWEEN 6TH AVE AND 7TH AVE	0	Unknown	NaN	0	0	0	...	0	2004/01/30 00:00:00+00	1/30/2004	Mid-Block (but intersection related)	NaN	NaN	NaN	NaN	0	0
3	-122.327525	47.708622	Block	NE NORTHGATE WAY BETWEEN 1ST AVE NE AND NE NOR...	0	Unknown	NaN	0	0	0	...	0	2016/01/23 00:00:00+00	1/23/2016	Mid-Block (not related to intersection)	NaN	NaN	NaN	NaN	0	0
4	-122.292120	47.559009	Block	M L KING JR ER WAY S BETWEEN S ANGELINE ST AND...	0	Unknown	NaN	0	0	0	...	0	2020/01/26 00:00:00+00	1/26/2020	Mid-Block (not related to intersection)	NaN	NaN	NaN	NaN	0	0

5 rows × 23 columns

Date and time has no bearing on our analysis since light conditions have already been taken into account in the column "Light Condition". Further study of the data source documentation shows us that some columns are a result of the accident and not the cause. Therefore, it is safe to assume to delete them as well as they serve no purpose. The following columns were also removed

- INCDATE
- INCDTTM
- COLLISIONTYPE
- PERSONCOUNT
- PEDCOUNT
- PEDCYLCOUNT
- VEHCOUNT
- INJURIES
- SERIOUSINJURIES
- FATALITIES

SEVERITYCODE and SEVERITYDESC convey the same information. So does X, Y (Co-Ordinates) and LOCATION. Therefore, we will delete redundant columns.

We are now left with the following data.

	X	Y	ADDRTYPE	SEVERITYCODE	JUNCTIONTYPE	UNDERINFL	WEATHER	ROADCOND	LIGHTCOND	SEGLANEKEY	CROSSWALKKEY
0	-122.320757	47.609408	Block	1	Mid-Block (not related to intersection)	N	Raining	Wet	Dark - Street Lights On	0	0
1	-122.319561	47.662221	Block	1	Mid-Block (not related to intersection)	N	Clear	Dry	Daylight	0	0
2	-122.327525	47.604393	Block	0	Mid-Block (but intersection related)	NaN	NaN	NaN	NaN	0	0
3	-122.327525	47.708622	Block	0	Mid-Block (not related to intersection)	NaN	NaN	NaN	NaN	0	0
4	-122.292120	47.559009	Block	0	Mid-Block (not related to intersection)	NaN	NaN	NaN	NaN	0	0

At this point, we will do a deep dive to see what kind of data distribution do we have for each of the features left.

ADDRTYPE

```
Block           144917
Intersection     71884
Alley              876
Name: ADDRTYPE, dtype: int64
```

SEVERITYCODE

```
1     137596
2      58747
0      21594
2b      3102
3        349
Name: SEVERITYCODE, dtype: int64
```

JUNCTIONTYPE

```
Mid-Block (not related to intersection)              101632
At Intersection (intersection related)                69178
Mid-Block (but intersection related)                  24405
Driveway Junction                                     11496
At Intersection (but not related to intersection)      2495
Ramp Junction                                           190
Unknown                                                  21
Name: JUNCTIONTYPE, dtype: int64
```

UNDERINFL

```
N    103874
0     81676
Y      5399
1      4230
Name: UNDERINFL, dtype: int64
```

WEATHER

```
Clear                       114694
Raining                      34036
Overcast                     28543
Unknown                      15131
Snowing                        919
Other                          860
Fog/Smog/Smoke                 577
Sleet/Hail/Freezing Rain       116
Blowing Sand/Dirt               56
Severe Crosswind                26
Partly Cloudy                   10
Blowing Snow                     1
Name: WEATHER, dtype: int64
```

ROADCOND

```
Dry               128535
Wet                48734
```

```
Unknown            15139
Ice                 1232
Snow/Slush          1014
Other                136
Standing Water       119
Sand/Mud/Dirt         77
Oil                   64
Name: ROADCOND, dtype: int64
```

LIGHTCOND

```
Daylight                        119448
Dark - Street Lights On          50125
Unknown                          13532
Dusk                              6082
Dawn                              2608
Dark - No Street Lights           1579
Dark - Street Lights Off          1239
Other                              244
Dark - Unknown Lighting             23
Name: LIGHTCOND, dtype: int64
```

SEGLANEKEY

```
0        218353
6532         19
6078         19
12162        18
10336        15
10342        13
8985         12
10420        12
8816         12
10354        11
12179        11
10590         9
10368         9
8995          8
          ...
20933         1
10453         1
8651          1
13001         1
35934         1
21701         1
15688         1
17863         1
20038         1
9803          1
14281         1
4178          1
6355          1
9402          1
```

```
Name: SEGLANEKEY, Length: 2101, dtype: int64
```

CROSSWALKKEY

```
0         217147
523609        19
520838        15
524265        13
525567        13
523148        11
521707        10
523699        10
523735        10
521574         9
523109         9
521253         9
522891         9
521604         9
           ...
523295         1
631427         1
29369          1
522952         1
525111         1
523080         1
521033         1
523208         1
521927         1
Name: CROSSWALKKEY, Length: 2343, dtype: int64
```

Some of the observations were

- The features SEGLANEKEY and CROSSWALKKEY have skewed data since major samples lie in the bracket "0". Therefore, it will not be helpful and will be deleted
- We can see some data points as "Null", "other" and "Unknown" (SEVERITYCODE = 0 is also unknown) and need to handle them in our dataset. Since the data is critical, using any data interpolation methods might skew the data. Therefore, considering the criticality of the task at hand, I would prefer to drop these data points instead
- We also realize that UNDERINFL has two data filling conventions. N meaning 0 and Y meaning 1. Therefore, we will replace 0 and 1 with N and Y to have data consistency

After the above steps were done, the date size was reduced from **(221525, 9)** to **(169906, 9)**. The data set now looks like

	X	Y	ADDRTYPE	SEVERITYCODE	JUNCTIONTYPE	UNDERINFL	WEATHER	ROADCOND	LIGHTCOND
0	-122.320757	47.609408	Block	1	Mid-Block (not related to intersection)	N	Raining	Wet	Dark - Street Lights On
1	-122.319561	47.662221	Block	1	Mid-Block (not related to intersection)	N	Clear	Dry	Daylight
5	-122.374194	47.564076	Block	1	Mid-Block (not related to intersection)	N	Clear	Dry	Daylight
6	-122.290734	47.709276	Block	1	Mid-Block (but intersection related)	N	Clear	Wet	Daylight
8	-122.336565	47.590398	Intersection	1	At Intersection (intersection related)	N	Overcast	Dry	Daylight

The data types are as follows

```
X                  float64
Y                  float64
ADDRTYPE            object
SEVERITYCODE        object
JUNCTIONTYPE        object
UNDERINFL           object
WEATHER             object
ROADCOND            object
LIGHTCOND           object
```

We are now happy with the data we have to build our Models. We will now move on to the important step of building our models.

Model Development and Accuracy Calculation

Since the problem at hand needs a **Supervised Machine Learning Algorithm**, we will look at the options we have. These are

- Classification Models (Used for Categorical Values)
- Regression Models (Used for continuous values)

As we have already seen above, out Dependent/Target is a **categorical variable**. Hence, we will be using **classification models** for prediction, namely

- K-Nearest Neighbor (KNN)
- Decision Tree

We will now do Pre-Processing for model development.

Pre-Processing Data

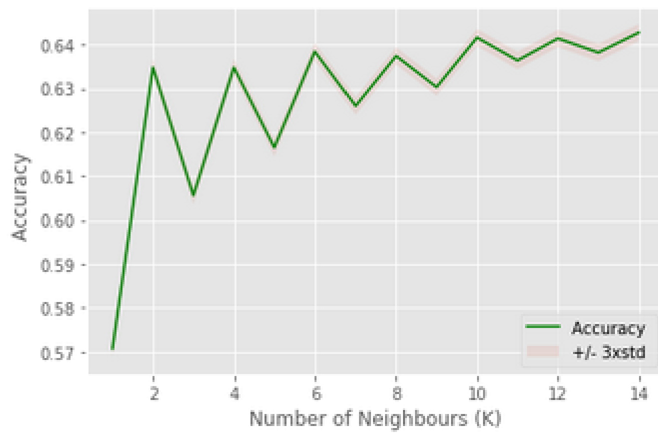
Pre Processing Data includes the following steps

- Dividing our data to X (**Independent Variables**) and Y (**Dependent variable**)
- Transforming our data using Label Encoder from categorical to numerical values
- Standardization and Normalization, as standard practice to avoid data of different ranges of each field form affecting the accuracy of the model

- Diving X and Y into **Training** and **Testing** sets. We will be using a **split of 70:30**. This will be used to calculate the accuracy of our model

K Nearest Neighbor (KNN)

KNN prediction was run for multiple values of K to see what will give us the best results.



Keeping the accuracy vs complexity trade-off in mind, a value of **K = 10** was used. This gave us a Training and Test accuracy of

```
Train set Accuracy:  0.6928368935194381
Test set Accuracy:  0.6391301431876866
```

Decision Tree

Decision Tree was made using a depth of 4. This gave us a Training and Test accuracy of

```
Train set Accuracy:  0.6569671909138514
Test set Accuracy:  0.6554776835116193
```

Results and Discussion

To assess the accuracy of our models, we will be using the following metrics

- F1 Score
- Jaccard Score

Following is the table that we get as a result.

Algorithm	Jaccard	F1-score
KNN	0.6416	0.5867

Decision Tree	0.6563	0.5310

We were not able to use any visual methods of data exploratory analysis owing to the nature of the data. Therefore, statistical methods were relied upon.

Conclusion

To summarize, we realized that both our models were able to predict the outcome with considerable accuracy. The decision tree was able to outperform KNN by a slight margin when it comes to Jaccard score but F1 was better for KNN. The prediction accuracy can be further improved by using other complex models available.

The Seattle department can now take it account these findings to be able to improve safety standards on the roads of Seattle.