Car Accident Severity – Seattle

IBM Data Science Capstone Project

Predicting the Car accident severity

1. Introduction

1.1 Background

Road traffic accidents (RTAs) have a significant impact on individuals, their families and the nation. In this report we are going to recognize the conditions that can cause future accidents in Seattle city as an example so as to caution the individuals with expectation to know and drive more cautiously. Further, the Seattle government is going to prevent avoidable car accidents by employing methods that alert drivers, health system, and police to remind them to be more careful in critical situations.

By and large, not giving enough consideration during driving, mishandling medications and liquor or driving at rapid are the fundamental cause of happening mishaps that can be forestalled by sanctioning harsher guidelines. Other than the previously mentioned reasons, climate, conceivability, or street conditions are the major wild factors that can be forestalled by uncovering shrouded designs in the information and declaring cautioning to the nearby government, police and drivers on the focused on streets.

The intended interest groups of this project are those individuals who truly care about the traffic records, particularly in the transportation division. Likewise, we need to make sense of the purpose behind accidents and help to lessen mishaps later on.

1.2 Problem

A portion of the elements which impact the probability and seriousness of a street auto collision include: the climate, nearby street conditions (for example parkways, metropolitan zones or rustic streets), season of day (and the presence or nonappearance of street-lamps), and the number and kind of vehicles in the region.

While it is instinctive that a mix of these components may be significant, instinct alone can't decide the general importance of these elements. Deciding the general importance of these contending factors is vital on the off chance that we are to completely comprehend the reasons for street car crashes and devise new techniques to limit their occurrence and seriousness.

1.3 Stakeholders

This research and report would be beneficial to the local government, people who live in Seattle, and also car insurance companies. By looking into road condition factors and address types such as intersection and block, we could see if there is any possible improvement in road condition and city planning.

Car Accident Severity – Seattle

2. Data

2.1 Dataset

Part of the project assignment we received a CSV Data File consisting of **194,673** records distributed into 38 header in this informational collection. Since we might want to recognize the elements that cause the mishap and the degree of seriousness, we will utilize **SEVERITYCODE** as our needy variable **Y**, and attempt various mixes of free factors **X** to get the outcome. Since the perceptions are very enormous, we may need to sift through the missing data values and erase the random segments first. At that point we can choose the factor which may have more effect on the accident, for example, address type, climate, street condition, and light condition.

The target/dependent variable is **SEVERITYCODE** which, in its original form, takes the values 1, 2.

The definitions of these severity codes are as follows:

- 1: Property Damage Only Collision
- 2: Injury Collision

2.2 Data Cleaning

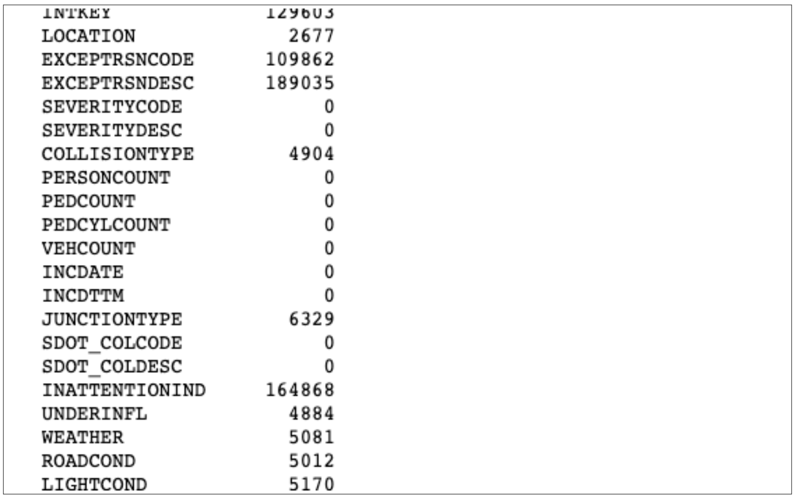
The original dataset is not suitable for quantitative analysis. There are many reasons for this, which are explained in the following subsections.

2.3 Missing Important data

The dataset contains missing entries, where one or more of the key predictor variables are absent or uninformative (e.g. 6.8% of accidents have “Unknown” listed in the **WEATHER** column). Including these data entries in the model not appropriate to produce an unbiased model, and so we drop the affected rows. Some of the columns (e.g. **ROADCOND**, **LIGHTCOND**) also have missing data.

A null value or missing data snapshot:

FIG 01



```
INTKEY              129603
LOCATION              2677
EXCEPTRSNCODE       109862
EXCEPTRSNDESC       189035
SEVERITYCODE             0
SEVERITYDESC             0
COLLISIONTYPE         4904
PERSONCOUNT              0
PEDCOUNT                 0
PEDCYLCOUNT              0
VEHCOUNT                 0
INCDATE                  0
INCDTTM                  0
JUNCTIONTYPE          6329
SDOT_COLCODE             0
SDOT_COLDESC             0
INATTENTIONIND      164868
UNDERINFL             4884
WEATHER               5081
ROADCOND              5012
LIGHTCOND             5170
```

2.4 Relevant & Irrelevant columns

Various unwanted or irrelevant columns present in the dataset (for example they contain data which is inconsequential to the causes or seriousness of mishaps) or are excess (for example they basically imitate data which is as of now encoded in different segments). Instances of pointless segments

incorporate **objectid**, **inckey** and **coldetkey**, which all recognize the mishap records with regard to other information held by SDOT which are excluded from this dataset.

Ref: https://www.seattle.gov/Documents/Departments/SDOT/GIS/Collisions_OD.pdf

#SL No	Column Name	Data Type (Pandas)	Description	Drop?
1	**X**	float64	Longitude (deg.)	N
2	**Y**	float64	Latitude (deg.)	N
3	**OBJECTID**	int64	ESRI unique identifier	Y
4	**INCKEY**	int64	A unique key for the incident	Y
5	**COLDETKEY**	int64	Secondary key for the incident	Y
6	**REPORTNO**	object	Report number for incident.	Y
7	**STATUS**	object	'Matched', 'Unmatched'	Y
8	**ADDRTYPE**	object	Collision address type: • Alley • Block • Intersection	N
9	**INTKEY**	float64	Key that corresponds to the intersection associated with a collision	Y
10	**LOCATION**	object	Description of the general location of the collision	N
11	**EXCEPTRSNCODE**	object	No Definition Found	Y
12	**EXCEPTRSNDESC**	object	No Definition Found	Y
13	**SEVERITYCODE**	int64	A code that corresponds to the severity of the collision: 1, 2	N
14	**SEVERITYDESC**	object	A detailed description of the severity of the collision	Y
15	**COLLISIONTYPE**	object	Collision type	Y
16	**PERSONCOUNT**	int64	The total number of people involved in the collision	N
17	**PEDCOUNT**	int64	The number of pedestrians involved in the collision. This is entered by the state.	N
18	**PEDCYLCOUNT**	int64	The number of bicycles involved in the collision. This is entered by the state.	N
19	**VEHCOUNT**	int64	The number of vehicles involved	N

			in the collision. This is entered by the state.	
20	**INCDATE**	object	The date of the incident.	Y
21	**INCDTTM**	object	The date and time of the incident.	Y
22	**JUNCTIONTYPE**	object	Category of junction at which collision took place	Y
23	**SDOT_COLCODE**	int64	A code given to the collision by SDOT.	Y
24	**SDOT_COLDESC**	object	A description of the collision corresponding to the collision code.	Y
25	**INATTENTIONIND**	object	Whether or not collision was due to inattention. (Y/N)	Y
26	**UNDERINFL**	object	Whether or not a driver involved was under the influence of drugs or alcohol.	Y
27	**WEATHER**	object	A description of the weather conditions during the time of the collision.	N
28	**ROADCOND**	object	The condition of the road during the collision.	N
29	**LIGHTCOND**	object	The light conditions during the collision.	N
30	**PEDROWNOTGRNT**	object	Whether or not the pedestrian right of way was not granted. (Y/N)	Y
31	**SDOTCOLNUM**	float64	A number given to the collision by SDOT.	Y
32	**SPEEDING**	object	Whether or not speeding was a factor in the collision. (Y/N)	N
33	**ST_COLCODE**	object	A code provided by the state that describes the collision. For more information about these codes, please see the State Collision Code Dictionary.	Y
34	**ST_COLDESC**	object	A description that corresponds to the state's coding designation.	Y
35	**SEGLANEKEY**	int64	A key for the lane segment in which the collision occurred.	Y
36	**CROSSWALKKEY**	int64	A key for the crosswalk at which the collision occurred.	N
37	**HITPARKEDCAR**	object	Whether or not the collision involved hitting a parked car. (Y/N)	N

Car Accident Severity – Seattle

Ref:

drop_cols = ['OBJECTID', 'INCKEY', 'COLDETKEY', 'REPORTNO', 'STATUS', 'INTKEY', 'EXCEPTRSNCODE', 'EXCEPTRSNDESC', 'COLLISIONTYPE', 'INCDATE', 'INCDTTM', 'JUNCTIONTYPE', 'ST_COLCODE', 'ST_COLDESC', 'SEGLANEKEY']

2.5 Methodology

I used Anaconda & Jupyter Notebook to program using Python do the data analysis, cleaning and modelling. To generate the table and graph for the dataset, I imported Python libraries (Pandas, Numpy, Matplotlib, and Seaborn).

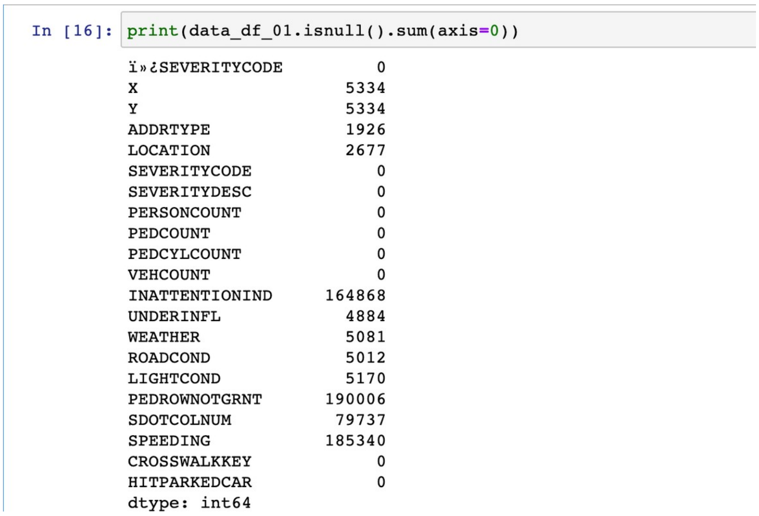
First I imported the data through **pd.read_csv**. I noticed that it had **194,673** rows and **38** columns. As I continued to analyse the dataset, we realize that only a few, around **12** columns are important for modelling:

(‘**SEVERITYCODE**’, ‘**X**’, ‘**Y**’, ‘**LOCATION**’, ‘**PERSONCOUNT’, ‘PEDCOUNT’, ‘PEDCYLCOUNT’, ‘VEHCOUNT’**, ‘**WEATHER**’, ‘**ROADCOND**’, ‘**LIGHTCOND**’, ‘**SPEEDING ‘**).

Give a run-down of the values present in the table for SEVERITYCODE:

```
1    136485
2     58188
Name: SEVERITYCODE, dtype: int64
```

Fig 2:



```
In [16]: print(data_df_01.isnull().sum(axis=0))

         ï»¿SEVERITYCODE          0
         X                     5334
         Y                     5334
         ADDRTYPE              1926
         LOCATION              2677
         SEVERITYCODE             0
         SEVERITYDESC             0
         PERSONCOUNT              0
         PEDCOUNT                 0
         PEDCYLCOUNT              0
         VEHCOUNT                 0
         INATTENTIONIND      164868
         UNDERINFL             4884
         WEATHER               5081
         ROADCOND              5012
         LIGHTCOND             5170
         PEDROWNOTGRNT       190006
         SDOTCOLNUM           79737
         SPEEDING            185340
         CROSSWALKKEY             0
         HITPARKEDCAR             0
         dtype: int64
```

Since most of the variable were categorical, it was hard to make the regression model. So, in this study, we focused more on the graphical data and the value count for different categories. There were around 135,000 (2/3) level 1 accidents and 60,000 (1/3) level 2 accidents.

After we cleaned first round, we tend to see the below dataset outstanding to further cleaning:

There are **15091** accidents with no weather information.

There are **15078** accidents with no road condition information.
There are **13473** accidents with no information about light conditions.
There are **18902** accidents without one or more of the above.

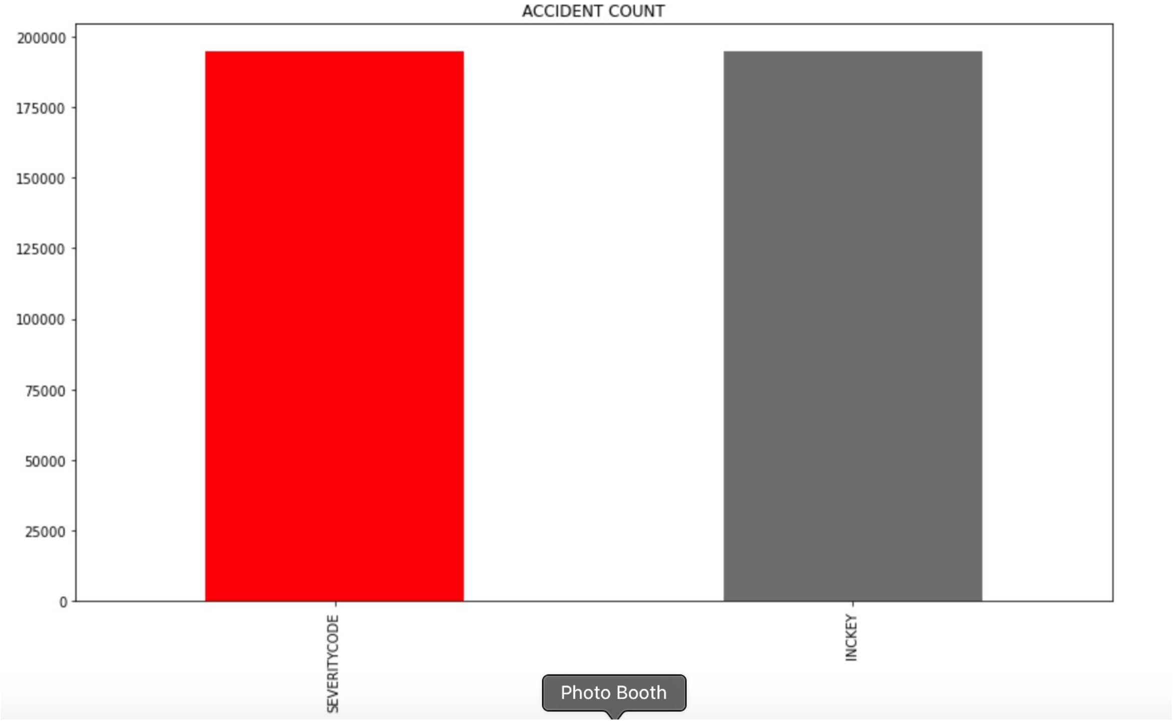
2.6 Data reset & remap

As I planned to cleaning the existing data and filter only the required dataset, I further realised that the existing data values needs to either reset with numeric values or remap to meaningful data.

For each accident there exists an **INCKEY,** hence there is no need for any cleaning or data value reset or remap.

After data reset and remapping, I kept the total dataset as **194673**

Fig: 01



Replace string values with Numeric, and NaN with 0

Column Name	Before Data	After Data
INATTENTIONIND	Y	1
UNDERINFL	Y, N	0, 1
SPEEDING	Y	0
LIGHTCOND	Daylight Dark - Street Lights On Dark - No Street Lights Dusk Dawn Dark - Street Lights Off Dark - Unknown Lighting Other	0 1 2 1 1 2 2 0

INATTENTIONIND	Y, N	1, 0
WEATHER	Clear Raining Overcast Unknown Snowing Fog/Smog/Smoke Sleet/Hail/Freezing Rain Blowing Sand/Dirt Severe Crosswind Partly Cloudy	0 3 1 0 3 2 3 2 2 1
	Dry Wet nan Unknown Snow/Slush Ice Other Sand/Mud/Dirt Standing Water Oil	0 2 0 0 1 2 0 1 2 2

3 Data Analysis

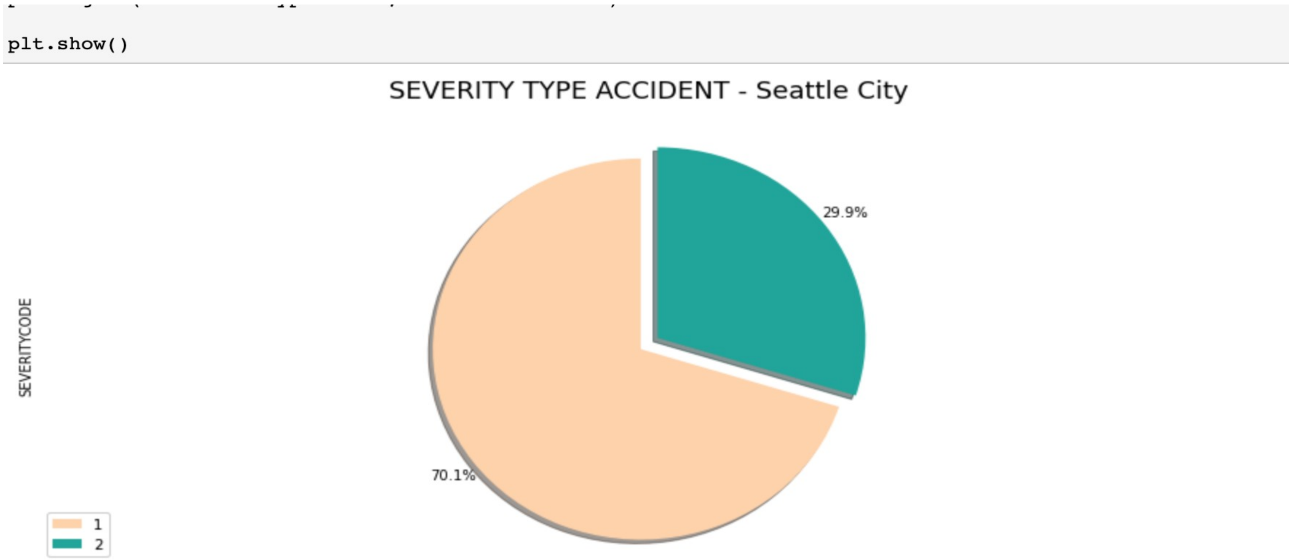
Taking into account that dataset and variables are all out factors with weather, road condition and light condition being an above level 2 categorical variables being an above level 2 clear cut factors whose qualities are restricted and normally dependent on a specific limited gathering whose connection may portray an alternate picture then what it really is. For the most part, considering the impact of these factors in fender benders are significant thus these factors were chosen. A couple of pictorial portrayals of the dataset were made so as to all the more likely comprehend the information.

After data analysis and feature selection, I explored **Accident Coun**t across different conditions and relative values to understand their distribution. The following two visualizations represent the dataset with/without missing values. Based on the different distribution, we can see that the unbalanced data of target variable 'severity' affects the number of collisions among different address types. So I decided to drop all the missing values and carry out my remaining modelling.

Let us look at the % of accident over **SEVERITYCODE** (1/2). We can see 70% and above of **SEVERITYCODE 1** accident and 30% and above of **SEVERITYCODE 2** accidents took place.

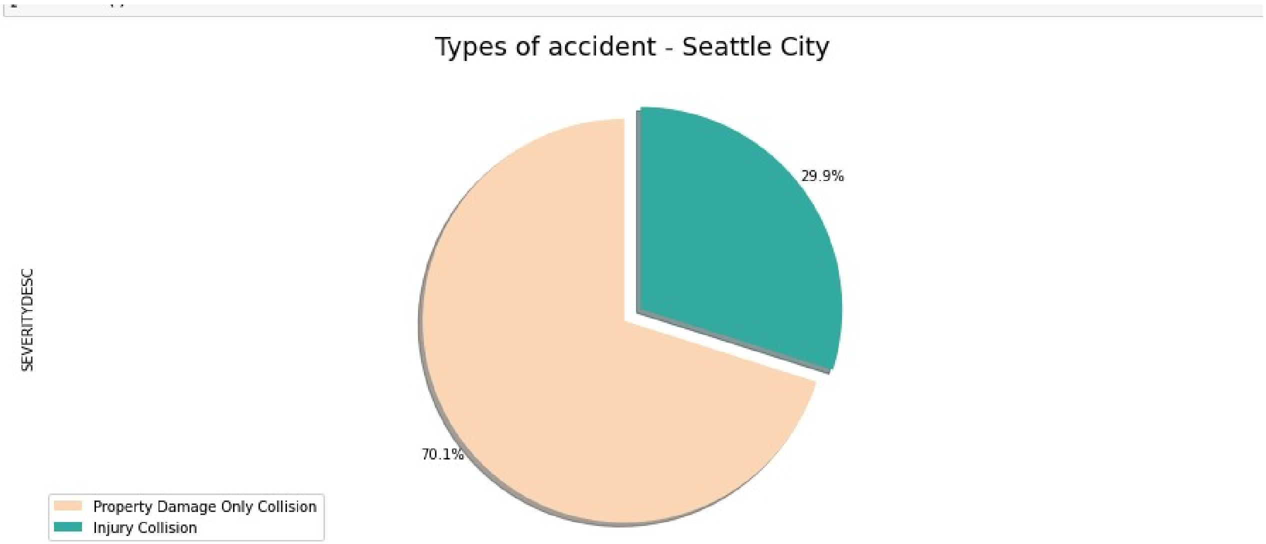
Car Accident Severity – Seattle

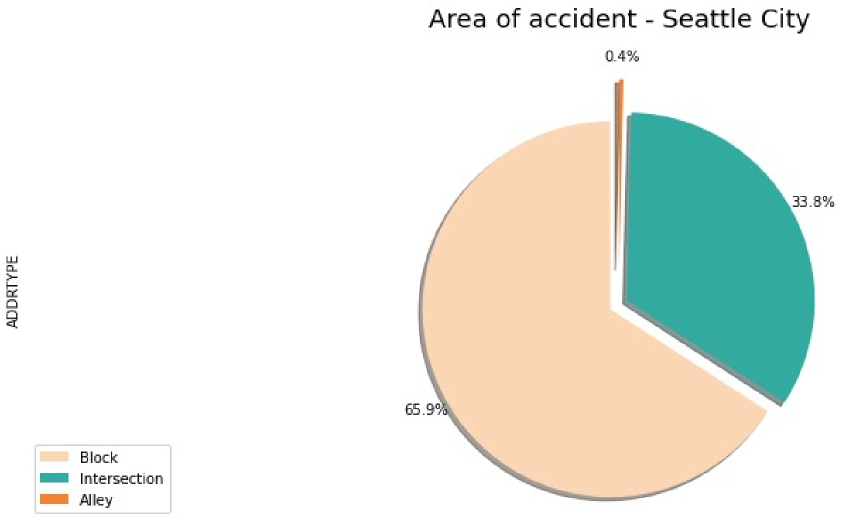
Fig: 02



```
plt.show()
```

Further, an insights of Types of accidents helped me understand Physical and Property damaged in percentile.





4 Modelling, Evaluation and Deployment

The models Logistic Regression, Decision Tree Analysis and k-Nearest Neighbor are used to predict the result. Calculated relapse is a measurable model that in its essential structure utilizes a strategic capacity to demonstrate a twofold reliant variable.

The Decision Tree Analysis separates an informational collection into littler subsets while simultaneously a related choice tree is steadily evolved. The eventual outcome is a tree with choice hubs and leaf hubs.

K closest neighbors is a basic calculation that stores every accessible case and characterizes new cases dependent on a similitude measure (in view of separation). The motivation behind why Decision Tree Analysis, Logistic Regression and k-Nearest Neighbor arrangement strategies were picked is on the grounds that the Support Vector Machine (SVM) model is off base for huge informational indexes, while this informational index has in excess of 180,000 lines loaded up with information. Besides, SVM works best with dataset loaded up with text and pictures.

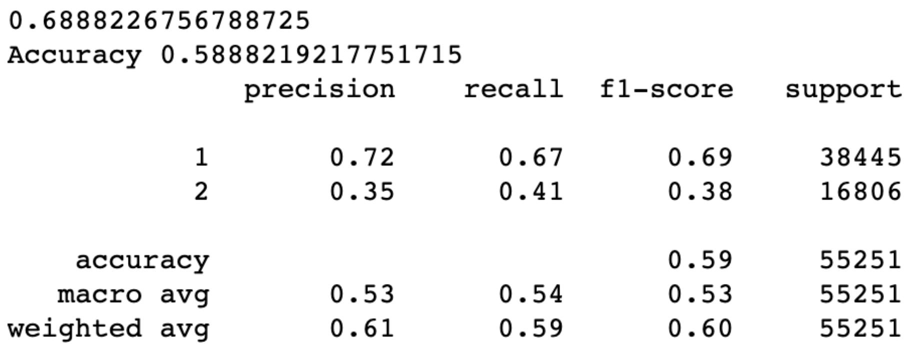
4. Conclusion

The scikit-learn library's Decision Tree Classifier used to run the Decision Tree Classification model on the Car Accident Severity data. The criterion chosen for the classifier was 'entropy' and the max depth was '6'. The post-SMOTE balanced data was used to predict and fit the Decision Tree Classifier.

4.1 Decision Tree Analysis

Decision Tree Classifier from the scikit-learn library was used to run the Decision Tree Classification model on the Car Accident Severity data. The criterion chosen for the classifier was 'entropy' and the max depth was '6'. The post-SMOTE balanced data was used to predict and fit the Decision Tree Classifier.

4.1.1 Classification Report

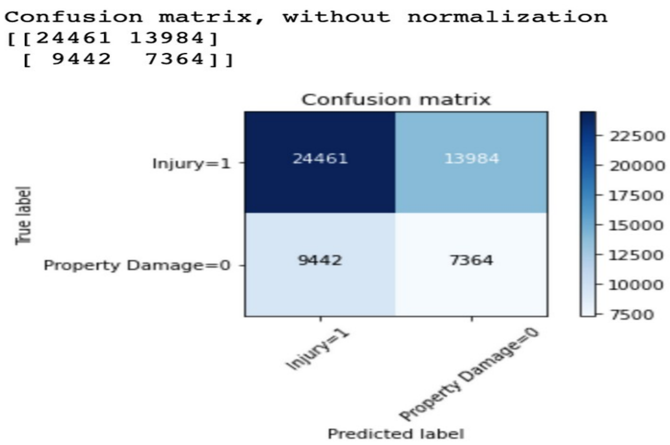


```
0.6888226756788725
Accuracy 0.5888219217751715
              precision    recall  f1-score   support

           1       0.72      0.67      0.69     38445
           2       0.35      0.41      0.38     16806

    accuracy                           0.59     55251
   macro avg       0.53      0.54      0.53     55251
weighted avg       0.61      0.59      0.60     55251
```

4.1.2 Confusion Matrix

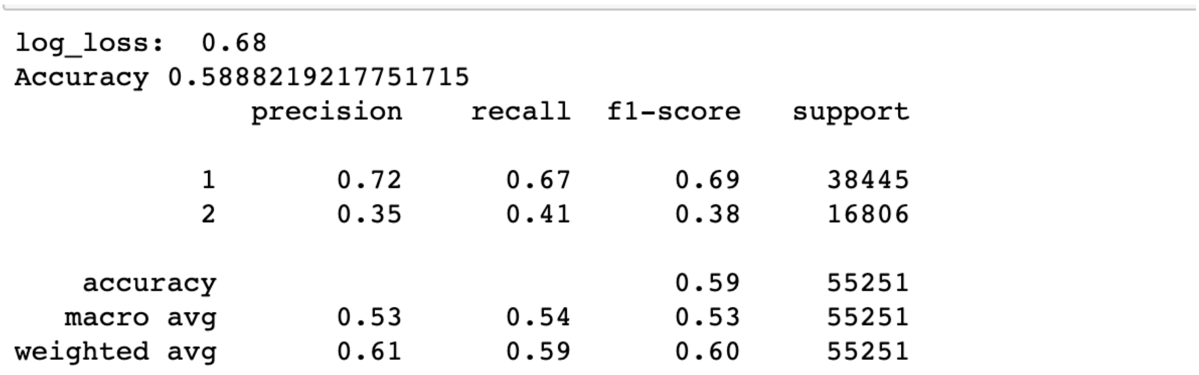


```
Confusion matrix, without normalization
[[24461 13984]
 [ 9442  7364]]
```

4.2 Logistic Regression

Logistic Regression from the scikit-learn library was used to run the Logistic Regression Classification model on the Car Accident Severity data. The C used for regularization strength was ‘0.01’ whereas the solver used was ‘liblinear’. The post-SMOTE balanced data was used to predict and fit the Logistic Regression Classifier.

4.2.1 Classification Report

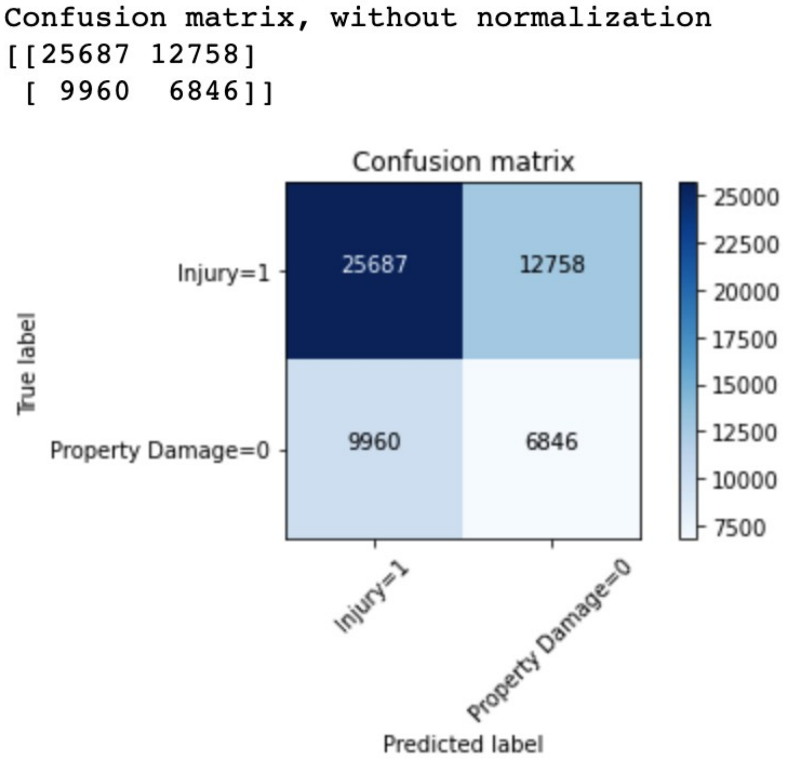


```
log_loss:  0.68
Accuracy 0.5888219217751715
              precision    recall  f1-score   support

           1       0.72      0.67      0.69     38445
           2       0.35      0.41      0.38     16806

    accuracy                           0.59     55251
   macro avg       0.53      0.54      0.53     55251
weighted avg       0.61      0.59      0.60     55251
```

4.2.2 Confusion Matrix



```
Confusion matrix, without normalization
[[25687 12758]
 [ 9960  6846]]
```

4.3 k-Nearest Neighbor

k-Nearest Neighbor classifier was utilized from the scikit-learn library to run the k-Nearest Neighbor AI classifier on the Car Accident Severity information. The best K, as demonstrated as follows, for the model where the most elevated elbow twist exists is at 4. The post-SMOTE adjusted information was utilized to anticipate and fit the k-Nearest Neighbor classifier.

5 Recommendations

As I performed the dataset analysis and post surveying the information and the yield of the Machine Learning models, a couple of suggestions can be made for the partners. The formative body for Seattle city can evaluate the amount of these mishaps have happened in a spot where street or light conditions were not ideal for that particular territory and could dispatch improvement anticipates for those regions where most extreme mishaps occur so as to limit the impacts of these two components. Though, the vehicle drivers could likewise utilize this information to evaluate when to play it safe out and about under the given conditions of light condition, street condition and climate, so as to evade a serious mishap, assuming any.

References

1. https://www.pbs.org/newshour/nation/motor-vehicle-crashes-u-s-cost-871-billion-year-federal-study-finds

2. https://data-seattlecitygis.opendata.arcgis.com/datasets/5b5c745e0f1f48e7a53acec63a0022ab_0