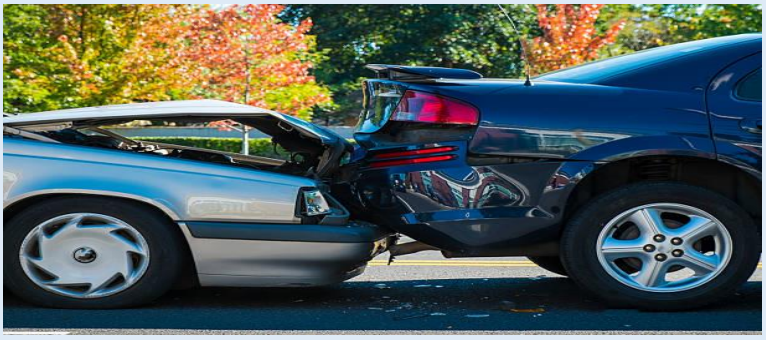


Capstone project- Seattle Accident Severity Data

October, 2020

Author: -Tuesy Bharadwaj

Contents

A. Introduction to the Business Problem

In today's time, in the light of modernized communities and the introduction of motorized vehicles with human lifestyle, land transportation has been at the edge of a great evolution. Increasing the number of cars, growing the traffic volume on the roads and the lack of safety have raised up the incidence and severity of traffic accidents. As per WHO reports, approximately 1.35 million people die each year as a result of road traffic crashes.

As per the Washington State Car Accident Statistics & Reports , In 2015, a crash occurred in Washington state every 4.5 minutes. Seattle is the 8th most dangerous city for accidents in the US. In this data set, we are analyzing the accident severity cases in Seattle from 2004 till May 2020 so that the city officials can take steps to make roadways safer for citizens. We will be analyzing factors that contributed to the collisions like weather, light, road conditions, speeding etc. and predict the severity of the cases based on them. This will help hospitals to anticipate the period during which the accidents increase and be prepared with the para medics, hospital staff and infrastructure for handling any crisis situation. Also, we could inform the Emergency services who can be well prepared.

The WHO sets the economic impact of road accidents in a developed country is at 2 to 3% of GDP, a significant figure for any country. Collaboration to reduce these losses has become an important issue of general interest, hence the it is important to analyze the data

This is the background to the business problem which we will try to solve with the help of machine learning algorithms

---

B. Description of Data

B.1 Data Source

This data have been collected and shared by the Seattle Police Department (Traffic Records) and are provided by Coursera for downloading through a link. I have uploaded the data in github repository

B.2 Data Location: https://github.com/TuesyBharadwaj/Coursera_Capstone

B.3 Data set name: Data-Collisions.csv It has 37 attributes and a little more than 0.1 million rows with the accidents' severity data. We will be looking into the below attributes to predict accident severity

The data describes two types of Severity for accidents

-> *Injury collision*

-> *Property Damage by collision*

We have to do exploratory data analysis on the data set to analyze patterns and relationships between factors in the table so that we can predict the severity of cases in the future.

We will be looking into the below attributes to predict accident severity

1. Road Condition (Wet, Dry, Sand, Oil, Mud, Snow,)

2. Weather Condition (Clear, Sun, Partly cloudy, rainy,)

3. Light Condition (Daylight, Dark, Dusk)

4. Inattention IND (If driver was distracted - Y/N)

5. UnderInfl (If driver was under influence of alcohol – Y/ N)

6. Speeding (Y/ N)

We will find out answers to the below questions

- What are the factors that have a high impact on road accidents?
- Is there a pattern to them?
- Is there any Correlation?
- Can we suggest any additional parameters for generating better insights

We will have to analyze the data to get a clearer picture and draw conclusions.

C. Objective

The objective is to define the problem, to find the factors that can have a relevant weight in the quantity and seriousness of the accidents, so that hospitals, emergency services, para medics etc. can be interested in reducing these figures, and focus the resources in points where these conditions converge.

In order to provide greater clarity, I will try to analyze the data, see if there are relationships or patterns, based on severity of accidents, so that measures can focus on these points as a first prevention strategy.

The data will be used so that we can determine which attributes are most common in traffic accidents in order to target prevention at these high-incidence points, facilitate resource allocation and mitigate probability of serious accidents given conditions

---

D. Methodology-Exploratory Data Analysis

D.1 Data Cleaning:

I. Converted the features like weather, road and light conditions which have string values into Numerical data

For example: Weather [‘Dry’, ‘Rainy’, ‘Snow’]

0: Dry

1: Rainy

2: Snow

II. For ATTENTIONIND attribute,
- Delete Nan values
- Replace blanks with N

III. For UNDERINFL attribute,

- Delete Nan values
- Replace blanks with N

IV. Redundant columns- There are no redundant columns or rows in the data set

V. Delete not related columns- I have deleted below attributes which are not related to the analysis

X
Y
OBJECTID
INCKEY
COLDETKEY
INTKEYCROSSWALKKEY

D.2 Extract year and month from date column and create a new column for analysis

```
df_data_1['year'] = pd.DatetimeIndex(df_data_1['INCDATE']).year
df_data_1.head()
```

```
df_data_1['month'] = pd.DatetimeIndex(df_data_1['INCDATE']).month
df_data_1.head()
```

E. Results

E.1 ACCIDENT COUNT ANALYSIS

Total number of accidents severity wise are 197643

SEVERITYDESC	Count
Property Damage Only Collision	136485
Injury Collision	58188

The number of incidents by above analysis show that Injury collision cases are 29.89% of total cases. In the rest 70.10%, only property was damaged

E.2 WEATHER CONDITION ANALYSIS

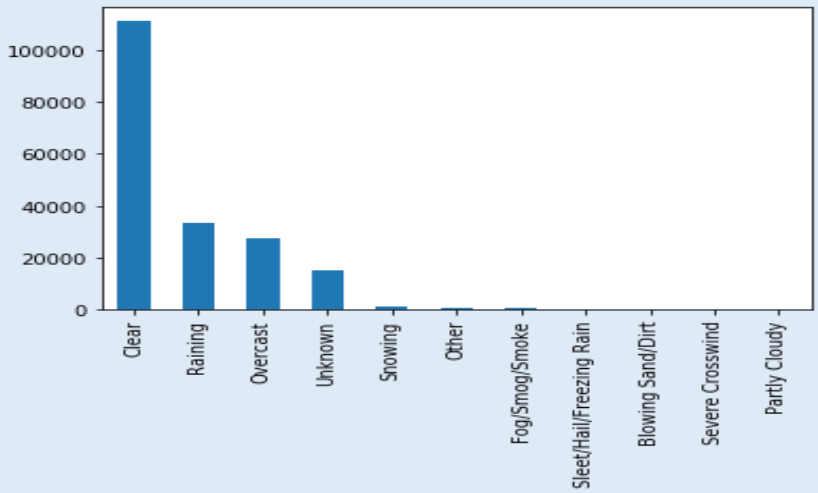


Figure1: Segregation of total number of accidents based on weather conditions

Figure 1 shows that in 57% cases when accident took place the weather was Clear followed by raining, overcast, snow and unknown reasons

E.3 ROAD CONDITION ANALYSIS

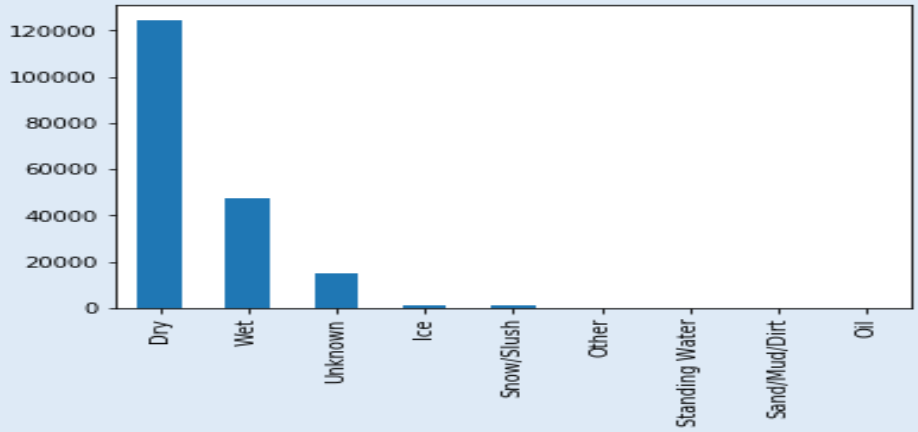


Figure 2: Segregation of total number of accidents based on road conditions

Figure 2 shows that in 70% accidents, the road condition was dry followed by wet, unknown ,ice and snow

E.4 **LIGHT CONDITION ANALYSIS**

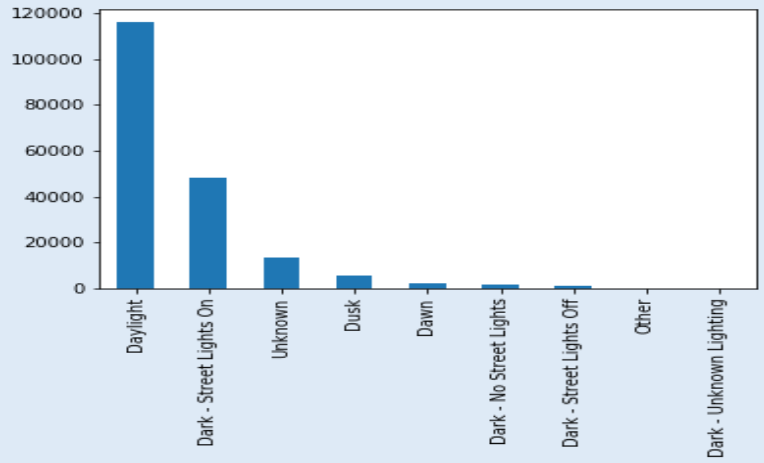


Figure 3: Segregation of total number of accidents based on light conditions

Figure 3 shows Close to 60% accident cases had light condition as broad daylight followed by dark street light on

E. 5 **UNDERINFLUENCE OF ALCOHOL**

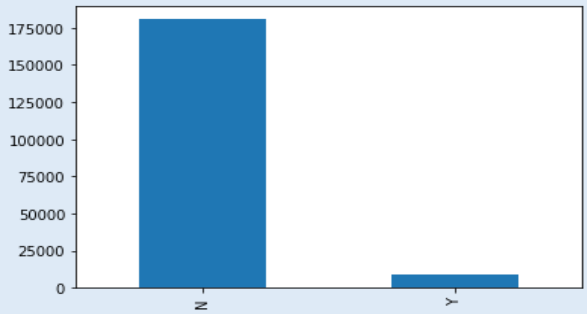


Figure 4: Accidents where the driver was under influence of alcohol

Figure 4 shows 92% cases the driver was not under the influence of alcohol

E. 6 COLLISION TYPE ANALYSIS

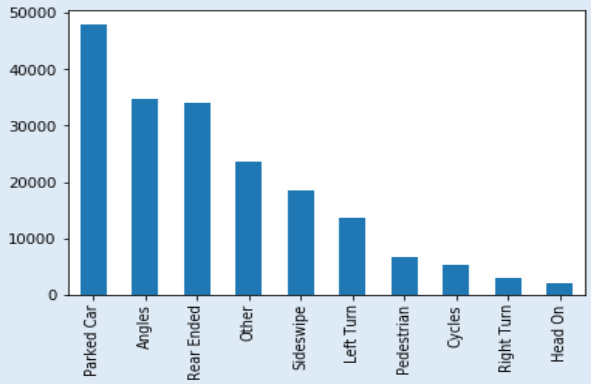


Figure 5: Segregation of total number of accidents based on collision type

Figure 5 shows Most collisions types were parked cars, angels, rear ended, sideswipe followed by left turn, pedestrian, cycle, right turn and head on

E. 7 SPEEDING

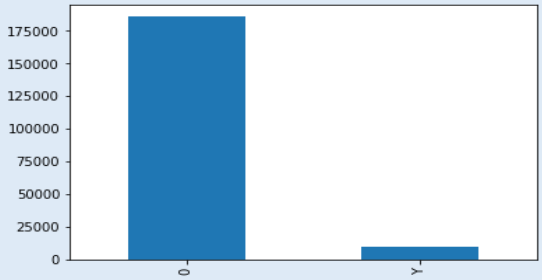


Figure 6: Count of accidents to check for speeding of driver

Figure 6 shows in most cases, the driver was not speeding.

E. 8 ACCIDENT CASES YEAR ON YEAR

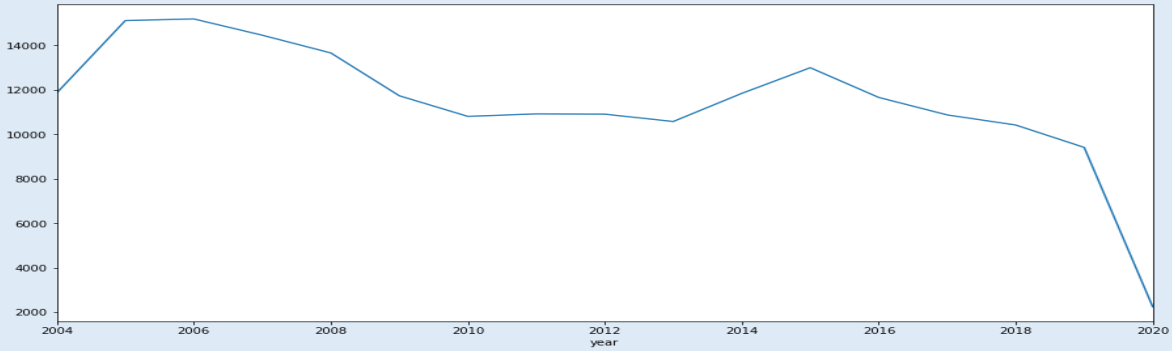


Figure 7 Line chart of Year On Year count of accidents

Figure 7 shows the year on year cases of accidents that happened from 2004 onwards till May 2020 in Seattle. There is a downward trend in the number of accidents

E. 9 ACCIDENT CASES MONTH ON MONTH

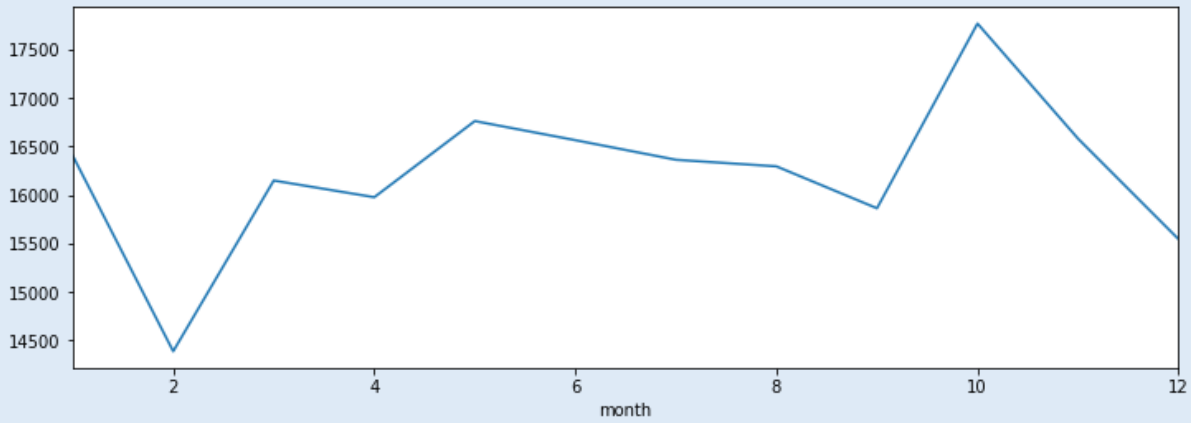


Figure 8 : Line chart of Month on month count of accidents

Figure 8 explains that during certain months (i.e. January, March, May, June, July, August, October) the accident count increases. This data can be shared with hospitals and emergency services so that can be prepared for future

E. 10 ACCIDENT CASES DAY ON DAY

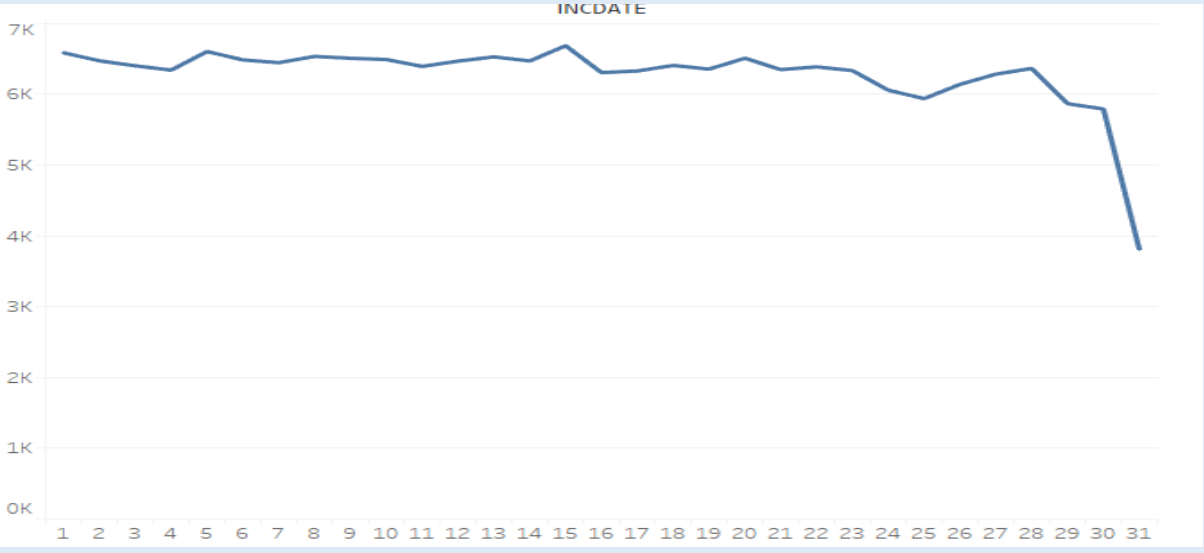


Figure 9: Line chart of Day wise segregation of the accidents

Figure 9 explains that generally accidents are happening throughout but with a slight increase on 27th, 28th, 29th of the month followed by a sharp decline towards the end of the month (i.e. 31st)

E. 11 WEEKDAY WEEKEND ANALYSIS

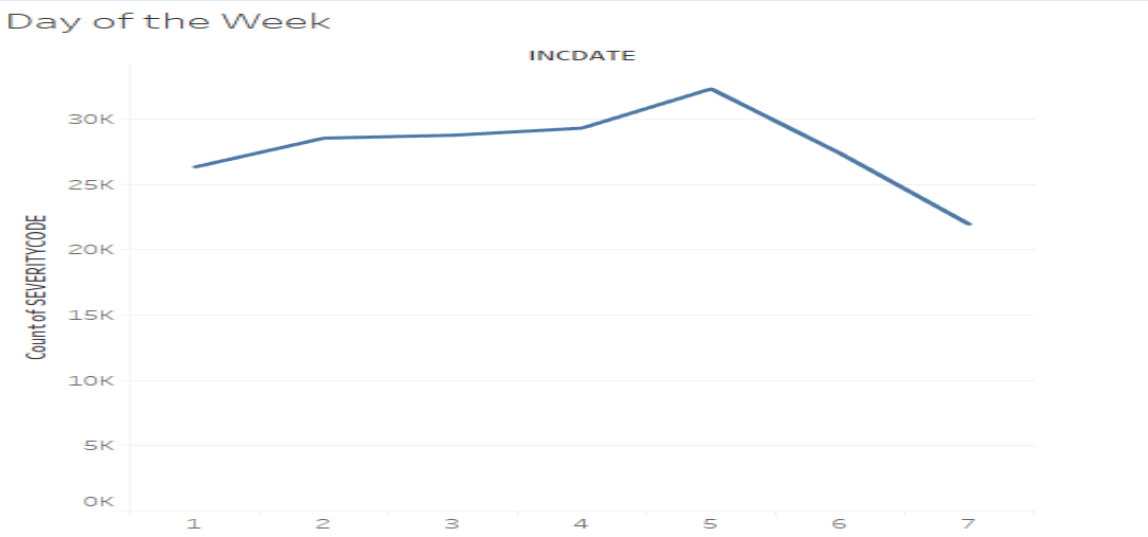


Figure 10: Line chart of Monday to Sunday weekdays wise accidents severity where in 1= Monday and 7= Sunday

Figure 10 explains that accidents are on a rise on Friday and there is a drop of accidents on the Weekend (i.e. Saturday and Sunday)

E. 12 HIT CAR PARKED ANALYSIS

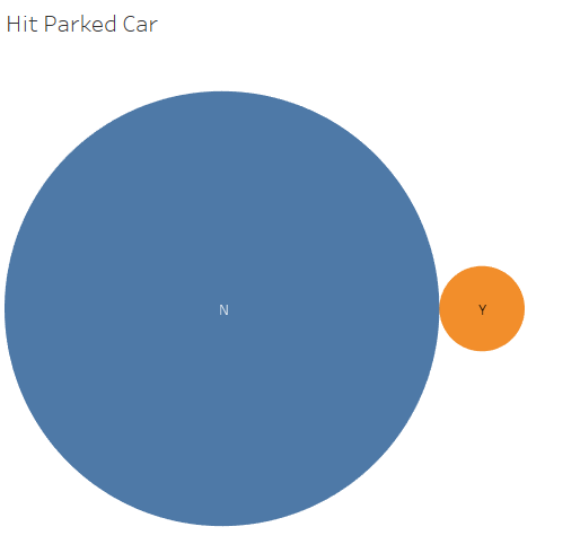


Figure 11: Bubble chart

Figure 11 shows that in 96% accidents, parked car was not hit. This suggests, that accidents that occurred were mostly of Severity 2 (i.e. Property Damaged) but not hitting the car while it was parked.

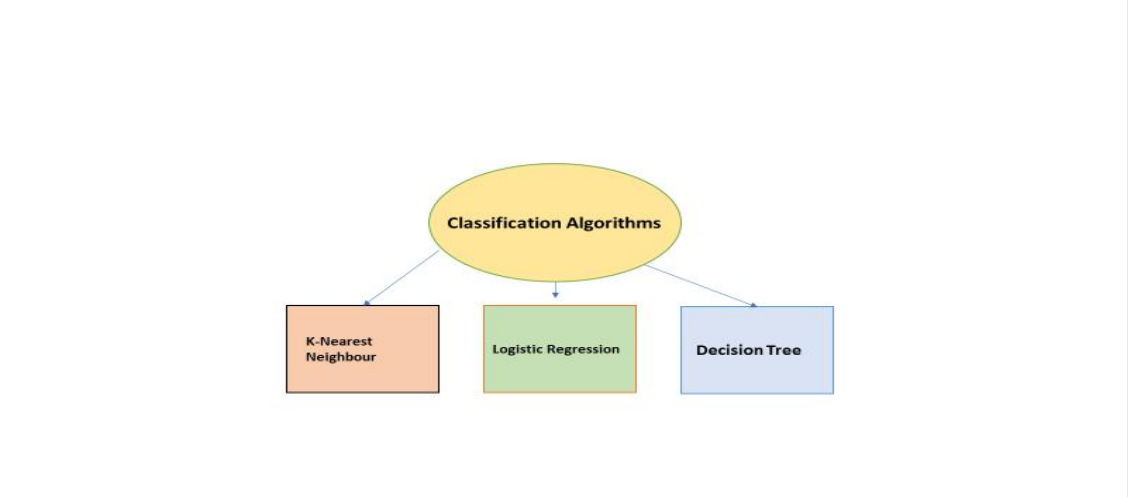
F. Discussion

After doing exploratory Data Analysis, I have concluded that most of the accidents are of the type (Property damage). The accidents are on the rise in particular months of the year like October and on particular days of the week like Friday. The weather, road and light conditions have effect on the accidents hence these factors have to be considered while doing the prediction analysis using Machine learning algorithms.

The hospital staff and emergency response team can be on the lookout for accidents based on the trend analysis done for Month on Month and Day on Day analysis. This data can be shared with the concerned authorities (hospitals, emergency services, para medics) for further course of action.

G. Machine Learning Algorithms

This is a scenario for a classic classification problem. It has 2 discrete targets that needs to be predicted. We will be using the standard classification machine learning algorithms. The predictors used are: 'Weather', 'Roadcond', 'Lightcond'. These predictors were categorical in nature hence we can apply the below machine learning models to predict the target (i.e. Accident Severity in this scenario)



Let's start with the first algorithm

G.1 Classification -KNN

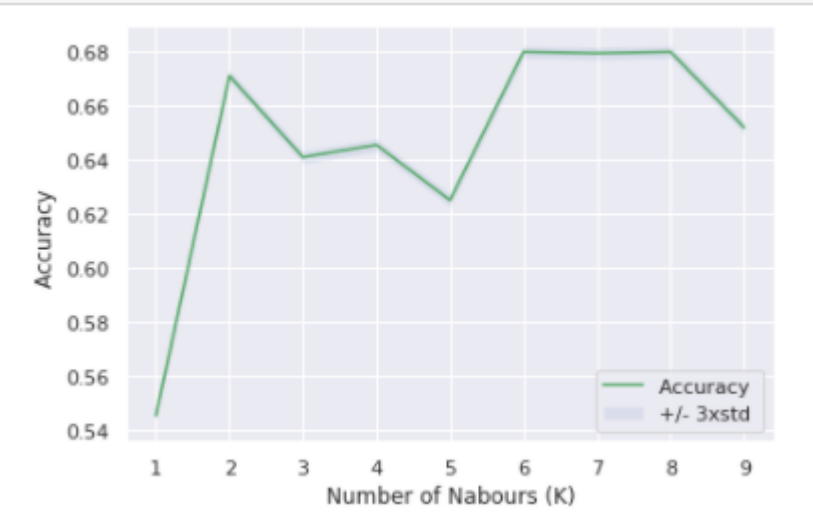
In the KNN model used, we have followed the below steps

1. Load the data
2. Normalize the data
3. Divide dataset into training and testing set
4. Predict using KNN algorithm; starting with k=4
5. Evaluate accuracy
6. Plot model accuracy
7. Calculate F1 and Jaccard score

```
We try to find accuracy of the model by taking K=4 initially
And check the accuracy. The accuracy is coming as below for both
Training and Testing
```

```
Train set Accuracy:  0.5418138155106653
Test set Accuracy:  0.5452420701168614
```

```
We then plot a graph to see for which K value the accuracy is the maximum
```



As per the above graph,

```
The best accuracy was with 0.6800308205984333 with k= 6
```

G. 2 Logistic regression- Liblinear

Logistic Regression is used when the dependent variable is categorical in nature. Here there are 2 outputs

SEVERITYCODE 1- Property damage only collision

SEVERITYCODE 2- Injury collision

Steps followed in LR model are as below: -

1. Load data
2. Data Normalization
3. Create Test and Training Set
4. Modeling using Scikit Learn (Use solver as 'Liblinear')
5. Predict Test set
6. Evaluation

Classification Report

	precision	recall	F1-score	support
1	0.70	1.00	0.83	27425
2	0.00	0.00	0.00	11510
micro avg	0.70	0.70	0.70	38935
macro avg	0.35	0.50	0.41	38935
weighted avg	0.50	0.70	0.58	38935

F1score is 0.58

Jaccard Score is 0.70

Log loss is 0.60

We can see from table above that this data set is unevenly distributed. Hence, the prediction for Class 2 is not showing correctly. This is because the data **is severely Imbalanced** with more cases for Type 1 Severity i.e. Property damage

We need to down sample the data and try again if the model is accurate to predict

Down Sampling Dataset

We create two data fame for Majority class and Minority class.

 # Separate majority and minority classes

df_majority = df_pandas[df_pandas.SEVERITYCODE==1]

df_minority = df_pandas[df_pandas.SEVERITYCODE==2]

The Sample size now becomes as below

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

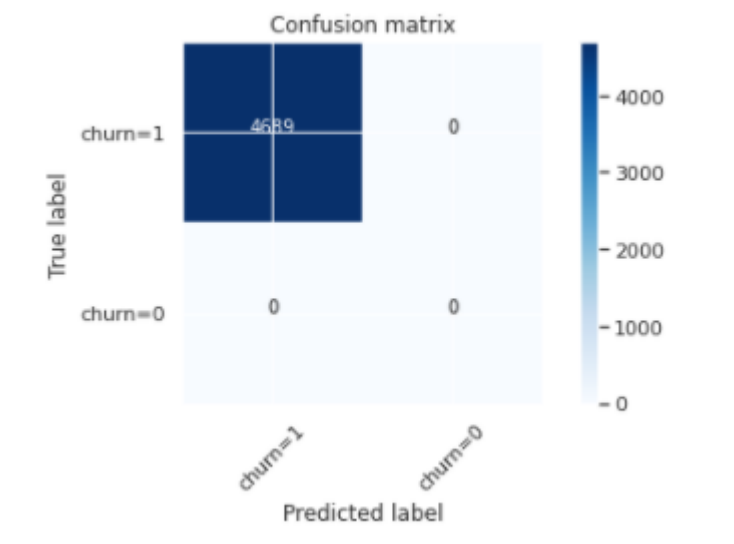
We model again using Logistic regression algorithm and below are the results :

After doing down sampling we get the below accuracy.

	precision	recall	f1-score	support
1	0.56	0.40	0.47	11612
2	0.54	0.69	0.60	11664
accuracy			0.55	23276
macro avg	0.55	0.55	0.54	23276
weighted avg	0.55	0.55	0.54	23276

We notice that the accuracy has improved slightly but its still not upto the mark

Confusion Matrix:



We can see from the Confusion Matrix that the model predicts True Positives correctly and not the rest (TN, FP, FN). The Precision and Recall probabilities are not that great either. I wouldn't say that the data set is problematic. It's just that this model is not right for this type of data sample.

We can conclude that this **model is not adequate for predicting correct outcome for our business problem**

G.3 Logistic Regression- SVM

We follow the same process for Support Vector Machine algorithm. The F1 score , Jaccard index and Log loss is similar to Liblinear. We can conclude that this model is not very effective to give correct results

G.4 Tree Model

Since, the data set is imbalanced, we can use Decision tree model for prediction. Decision trees often perform well on imbalanced datasets because their hierarchical structure allows them to learn signals from both classes.

We built the model with depth as 4 and the steps used to build the model were similar to KNN

The accuracy predicted is 0.699

```
from sklearn import metrics
import matplotlib.pyplot as plt
print("DecisionTrees's Accuracy: ", metrics.accuracy_score(y_testset, predTree))
```

```
DecisionTrees's Accuracy:  0.6994109790760591
```

I have tried to summarize the accuracy of all the 3 algorithms used on this data set for desired Outcome. Below is a snapshot of the results

Synopsis of accuracy table: -

Model	F1	Jaccard	Log Loss	Accuracy
KNN	0.612	0.652		0.68
Logistic Regression- Liblinear	0.577	0.701	0.60	0.54
Logistic Regression- SVM	0.577	0.701	0.60	0.54
Tree Model				0.699

H. Conclusion

The accuracy for KNN model is 0.68 and for Decision tree is 0.699 . The Accuracy for Logistic regression is 0.54 which is lesser than the remaining two. We have used the independent variables of Weather, Road and Light Conditions in the model.

By the analysis, we can interpret that the accident severity can depend on the variables we have taken into consideration and **KNN model & Decision tree will be ideal models to predict the Accident Severity**

I. Future direction

In future, we can suggest the traffic department at Seattle to capture some other parameters in this data set . **We can include Age of the driver and Accident zones**. This will help to analyze whether the accidents are occurring in a particular age group, and also if there are a specific zone where more accidents take place.

As the data set has more accidents of the Type Property damage, if accident zones are captured, the authorities can lay their focus on these areas.

Age of the driver is also an important parameter to be captured. We can analyze if the accidents are occurring in a particular age group which can generate some useful insights for the Traffic Department of Seattle

References

I have taken reference from the below article

Washington State Car Accident Statistics & Reports- https://www.colburnlaw.com/seattle-traffic-accidents/