Project to predict traffic accident severity by machine learning

Zhang Haibin

18-Oct-2020

1. Introduction

1.1 Background

Average number of car accidents in the U.S.A every year is 6 million, and 3 million people in the U.S.A are injured every year in car accidents, and more than 90 people die in car accidents everyday.The United States is among the countries with the highest rate of traffic-related fatalities per one million population.

Which kind of scenario is the worst accident you should avoid, and how to avoid the worst accident in traffic.To reduce traffic accidents and reduce the severity of car accidents is an important public safety challenge for U.S.A, also for around the world.

1.2 Problem

There are different major factors for different accidents, however, some commonalities exist in the accidents, which may explain the severity of traffic accidents. And accidents can be prevented by revealing hidden patterns in the data, such as: weather condition, Road condition, Light condition, address type and Junction type. Thus, one traffic accident severity prediction should be useful and instructive for accidents reduction.

1.3 Interest

Transportation department of government, traffic designers, police and respective drivers of vehicles should be interested at this kind of prediction. And the key points and inclusive result should be able to remind the drivers to be more vigilant and watchful, and let them avoid the accident and reduce severity level of accident.

2. Data Acquisition and Cleaning

2.1 Data Source

Data provided by the Seattle Department of Transportation (SDOT) , can be downloaded as below link: https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv

The data provided the key information that included Severity of the accident, Incident Date/Time,location of collision, number of vehicles and persons involved in accidents, Road type and conditions, Weather and light conditions.

This dataset has details about 194k accident details. Each accident is defined by 38 different attributes (features). The metadata of the features are given as below link: https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Metadata.pdf

Severity of the accident (feature name: SEVERITYCODE) will be the predictor.it is used to measure the severity of an accident from 0 to 3 within the dataset

2.2 Feature Selection and drop

As mentioned, SEVERITYCODE is the predictor ranged from 0 to 3, as below:

3—fatality; 2b—serious injury; 2—injury; 1—prop damage; 0—unknown

Attributes used to weigh the severity of an accident are selected:

‘ADDRTYPE’, JUNCTIONTYPE’, ‘WEATHER’, ‘ROADCOND’ and ‘LIGHTCOND’. And for exploratory data analysis, we would like to choose some of numerical variables such as (PERSONCOUNT, VEHCOUNT, PEDCOUNT, PEDCYLCOUNT). The rest features as below to be deleted: ‘X’, ‘Y’, ‘Coldetkey’, ‘Reportno’, ‘Intkey’, ‘Location’, ‘Exceptrsncode’, ‘Exceptrsndesc’, ‘Severitycode.1’, ‘Severitydesc’, ‘Incdate’,’Incdttm’, ‘, ‘Sdot_Colcode’, ‘Sdot_Coldesc’, ‘Inattentionind’, ‘Underinfl’, ‘Pedrownotgrnt’, Sdotcolnum’, ‘Speeding’, ‘St_Colcode’, ‘St_Coldesc’,’Seglanekey’, ‘Crosswalkkey’, ‘Hitparkedcar’, ‘Objectid’, ‘Collisiontype’,’Status’.

```
In [4]:
#Feature selection
df_sev = df_sev.drop(columns = ['OBJECTID', 'SEVERITYCODE.1', 'REPORTNO', 'INCKEY', 'COLDETKEY',
                'X', 'Y', 'STATUS',
                'INTKEY', 'LOCATION', 'EXCEPTRSNCODE',
                'EXCEPTRSNDESC', 'SEVERITYDESC', 'INCDATE',
                'INCDTTM', 'SDOT_COLCODE',
                'SDOT_COLDESC', 'PEDROWNOTGRNT', 'SDOTCOLNUM',
                'ST_COLCODE', 'ST_COLDESC', 'SEGLANEKEY',
                'CROSSWALKKEY', 'HITPARKEDCAR', 'PEDCOUNT', 'PEDCYLCOUNT',
                'PERSONCOUNT', 'VEHCOUNT', 'COLLISIONTYPE',
                'SPEEDING', 'UNDERINFL', 'INATTENTIONIND'])
```

2.3 Data Cleaning and handling

Based on the background and data understanding, the irrelevant columns, or minor factors on accident severity to be dropped.

There are some null values for ADDRTYPE, JUNCTIONTYPE, WEATHER, ROADCOND and LIGHTCOND attributes which are replaced with value ‘Others’. As these are expected to be among the features which influence the likelihood and severity of accidents, we must consider discarding these rows before training the model

```
In [5]:
#clean the data
df_sev['ADDRTYPE'].fillna('Other',inplace=True)
df_sev['JUNCTIONTYPE'].fillna('Other',inplace=True)
df_sev['WEATHER'].fillna('Other',inplace=True)
df_sev['LIGHTCOND'].fillna('Other',inplace=True)
df_sev['ROADCOND'].fillna('Other',inplace=True)
df_sev.isnull().sum()
```

```
Out[5]: SEVERITYCODE    0
        ADDRTYPE        0
        JUNCTIONTYPE    0
        WEATHER         0
        ROADCOND        0
        LIGHTCOND       0
        dtype: int64
```

3. Exploratory Data Analysis(EDA)

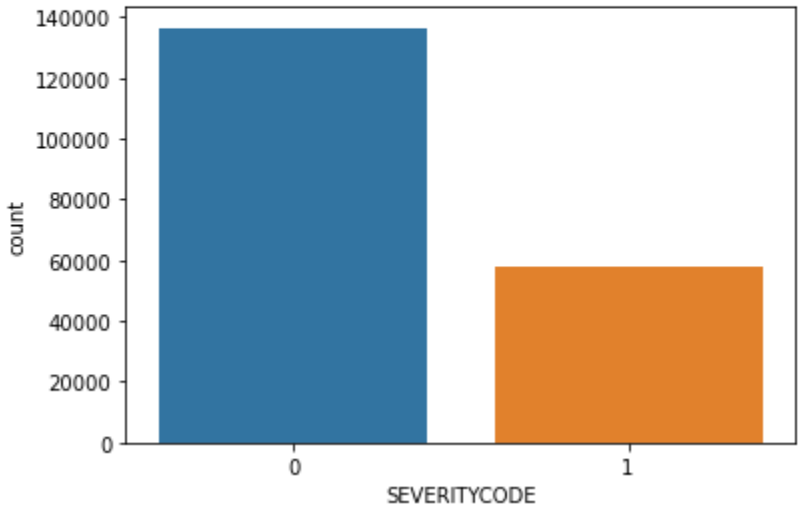
In order to understand the relationship between SEVERITYCODE with the selected features, exploratory data analysis has done as below:

3.1 Target variable

To get a good understanding of the dataset, and check the values of target variable. And the barchart shown there exist on categories on severitycode. “SEVERITYCODE” class 1 is bigger than number of rows in class2

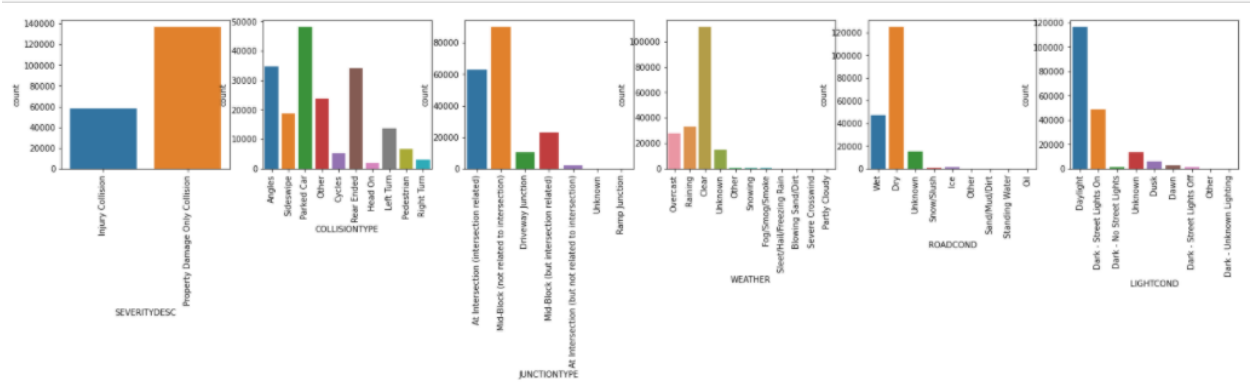
```
In [13]:  # to check the predictor values
          df_sev["SEVERITYCODE"].value_counts()
          sns.countplot(df_sev['SEVERITYCODE'],data=df_sev)

Out[13]: <matplotlib.axes._subplots.AxesSubplot at 0x265075eb370>
```



3.2 Categorical variables analysis

Logistic models do not handle categorical variables. For some features are of categorical data types, that need to be converted into numerical data types, and do the categorical variables analysis.

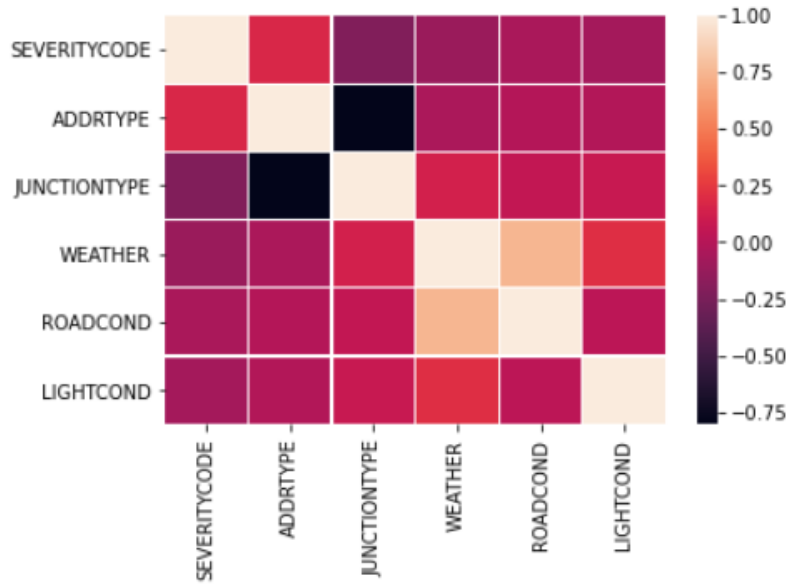


3.3 Relationship between “SEVERITYCODE” and selected features

After the variables converted to numerical variables, checked correlation matrix on numerical variables, to check if there exist correlation among them. heatmap is one of simple and straightforward visualization style to check variables relationship.

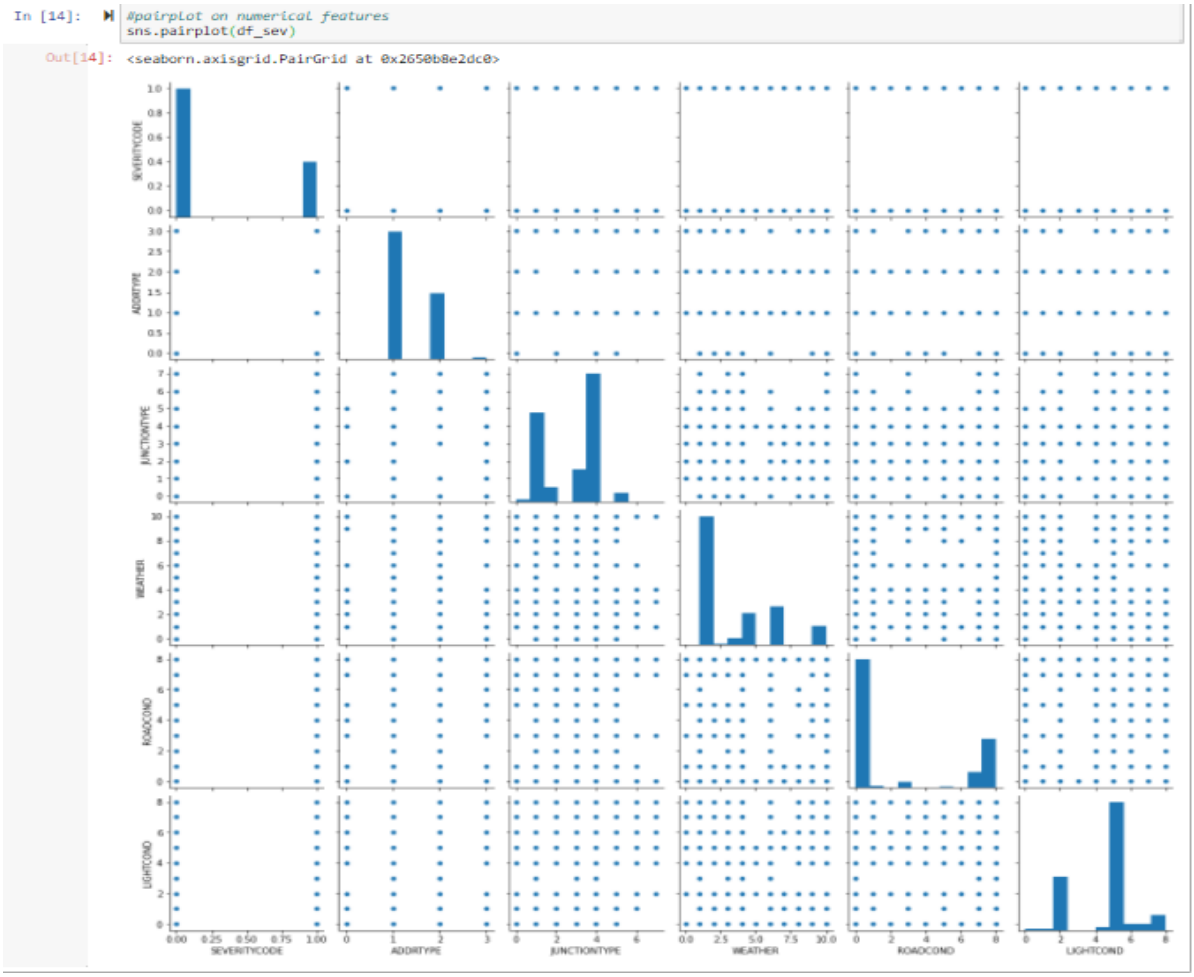
```
In [10]: #Show heatmap between features
         sns.heatmap(df_sev.corr(),linewidth=0.2,cbar_kws={"shrink":1})
```

```
Out[10]: <matplotlib.axes._subplots.AxesSubplot at 0x2651100fa30>
```



3.4 PairPlot on numerical variables

There are number of methods to use in EDA, one of the most effective and powerful starting tools is the pairs plot (also called a scatterplot matrix). A pairs plot allows us to see both distribution of single variables and relationships between two variables. Pairplot just show all variables paired with all the other variables. Below is the numerical variables pairplot, and specify with more correlation chart for the particular variables.



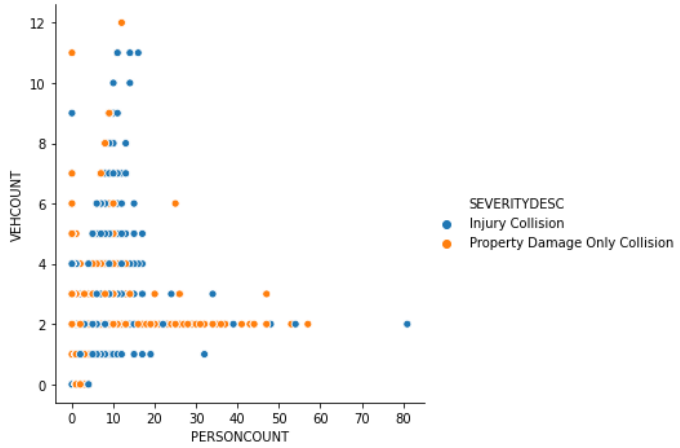
3.4.1 Relationship between VEHCOUNT, PERSONCOUNT based on SEVERITY

It seems there exist strong correlation between vehicle count and person count, two vehicles

Count is the most frequent accidents and involved more people based on the vehcount and personcount relationship graph.

```
In [19]: #Relplot
         sns.relplot(x='PERSONCOUNT',y='VEHCOUNT',hue='SEVERITYDESC',data=df)
```

```
Out[19]: <seaborn.axisgrid.FacetGrid at 0x2650d3d2250>
```

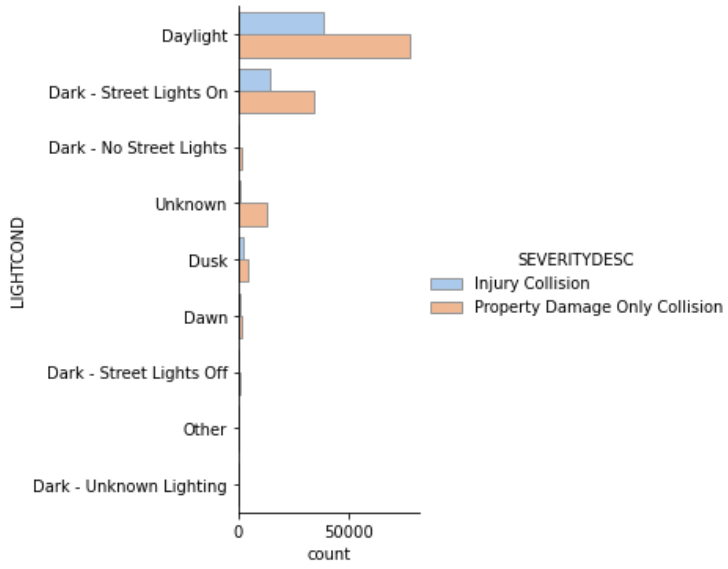


3.4.2 Relationship between LIGHTCOND by SEVERITY of Accidents

Based on below categorical plot between Lightcondition by Severity of accidents, we can say that more number of accidents related to property damage are reported during Daylight .The reason probably exist on the heavy traffic condition for the vehicles and cause the complicated sceniario for drivers.

```
In [20]: #Lightcondition with Severity
         sns.catplot(y='LIGHTCOND',hue='SEVERITYDESC',kind='count',palette='pastel',
         edgecolor='0.6',data=df)
```

```
Out[20]: <seaborn.axisgrid.FacetGrid at 0x2650b8e69d0>
```

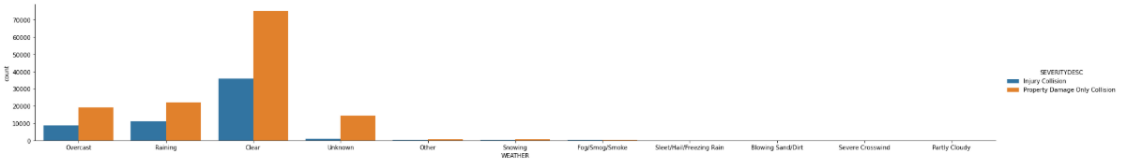


3.4.3 Relationship between WEATHER by SEVERITY of accidents

Based on below categorical plot between Weather by Severity of accidents, a lot of accidents happened at clear weather, it seems the weather is not the most important factor on accidents.

```
In [22]: sns.catplot(x='WEATHER',kind='count',hue='SEVERITYDESC',data=df,height=4,aspect=6)
```

```
Out[22]: <seaborn.axisgrid.FacetGrid at 0x27695a9dd30>
```

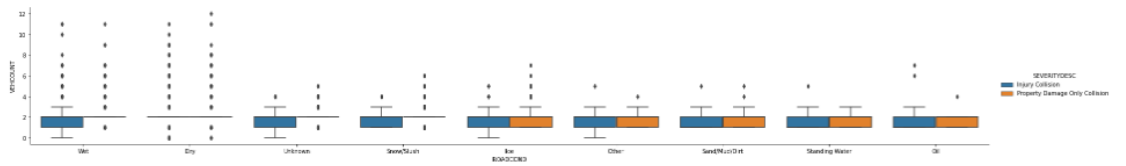


3.4.4 Relationship between ROADCOND, VEHCOUNT by SEVRITY of accidents

There exist some outliers of VEHCOUNT when road condition is Wet, Dry and ice.It means the abnormal road condition may cause sever accidents.

```
In [28]: #Relationship between ROADCOND, VEHCOUNT by SEVRITY of accidents
sns.catplot(x='ROADCOND',y='VEHCOUNT',kind='box',hue='SEVERITYDESC',data=df,height=4,aspect=6)
```

```
Out[28]: <seaborn.axisgrid.FacetGrid at 0x2650faa5820>
```

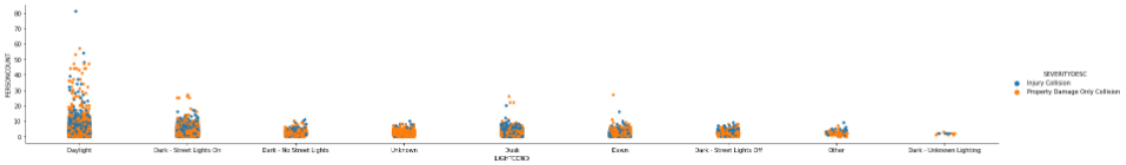


3.4.5 Relationship between LIGHTCOND, PERSONCOUNT by SEVERITY of accidents

Based on below categorical plot between LIGHTCOND, PERSONCOUNT by Severity of accidents, we can say that more number of people involved in accidents related to property damage and are reported during Daylight

```
In [25]: #Relationship between LIGHTCOND, PERSONCOUNT by SEVERITY of accidents
sns.catplot(x='LIGHTCOND',y='PERSONCOUNT',hue='SEVERITYDESC',data=df,height=4,aspect=6)
```

```
Out[25]: <seaborn.axisgrid.FacetGrid at 0x26501fdcb20>
```



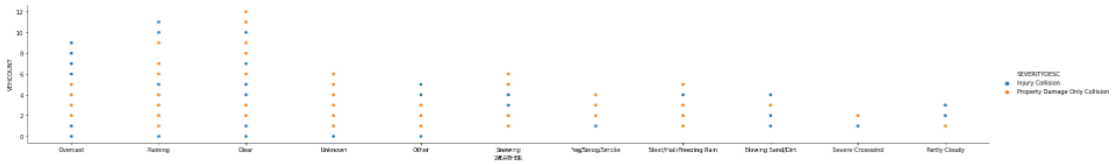
3.4.6 Relationship between WEATHER, VEHCOUNT by SEVERITY of accidents

The data shown clear and rain weather, more multiple vehicles were involved in accidents. It means clear day accidents perhaps to be related to careless.

However, the raining day, has the high risk to lead to more vehicles involved the accidents due to visibility and slippery road.

```
In [27]: #Relationship between WEATHER, VEHCOUNT by SEVERITY of accidents
         sns.relplot(x='WEATHER',y='VEHCOUNT',hue='SEVERITYDESC',data=df,height=4,aspect=6)

Out[27]: <seaborn.axisgrid.FacetGrid at 0x265060f39d0>
```



4. Predictive Modeling

The variables: ADDRTYPE, JUNCTIONTYPE, WEATHER, ROADCOND and LIGHTCOND were chosen for data analysis. After the preliminary analysis, these key factors will to be used to predict the traffic accident severity. And, here we will utilize the three popular Machine Learning models as below :

K Nearest Neighbor,

Decision tree

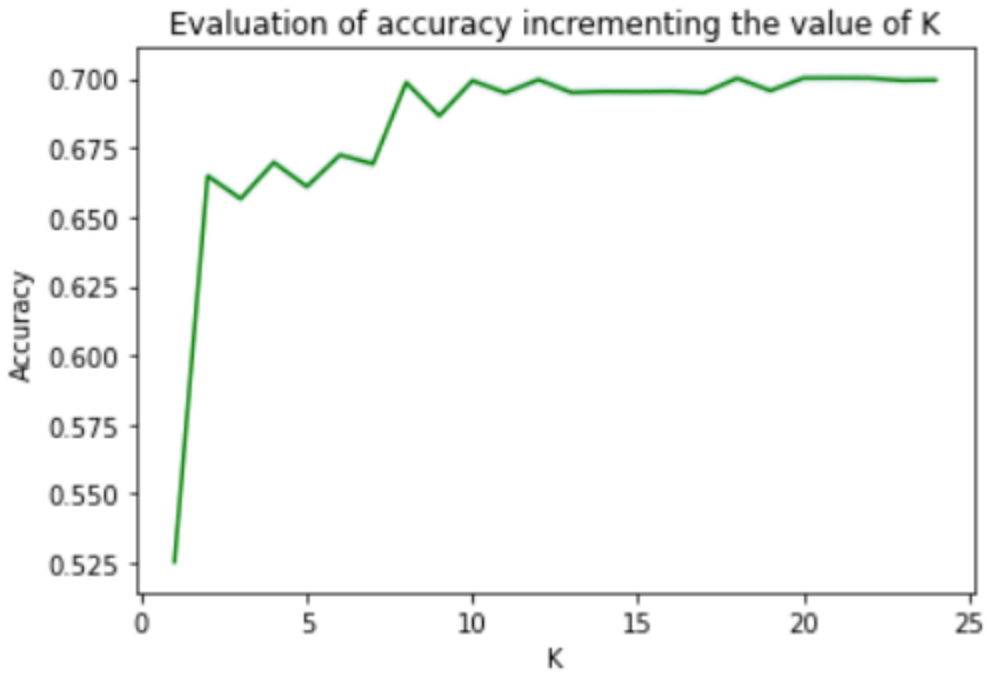
Logistic Regression

to predict the severity of traffic accident based on the provided data source.

4.1 KNN

The k-nearest neighbors (KNN) algorithm is a simple, easy-to-implement supervised machine learning algorithm that can be used to solve both classification and regression problems. The KNN algorithm hinges on this assumption being true enough for the algorithm to be useful. KNN captures the idea of similarity (sometimes called distance, proximity, or closeness) with some mathematics.

As the key factors have the similarity to lead to traffic accidents, it is suitable for KNN to predict the severity of traffic accident based on the data set.



4.2 Decision Tree

Decision Trees is one of very powerful Machine Learning model to provide high accuracy prediction. The “knowledge” learned by a decision tree through training is directly formulated into a hierarchical structure. This structure holds and displays the knowledge in such a way that it can easily be understood, even by non-experts.We will load in our dataset and initialize our decision tree for classification, and train for prediction

In [40]:
```
#Summary on Decision Tree

Dyhat=tree.predict(X_test)
DT_jacc=jaccard_score(y_test,Dyhat,average='macro')
DT_f1=f1_score(y_test,Dyhat,average='macro')
DT_acc=accuracy_score(y_test,Dyhat)

print('DT_jacc=',DT_jacc,"DT_f1=",DT_f1,"DT_acc=",DT_acc)
```

```
DT_jacc= 0.35397434583212195 DT_f1= 0.4182988472927632 DT_acc= 0.7026300469162015
```

In [41]:
```
severityTree=DecisionTreeClassifier(criterion='entropy',max_depth=max_depth).fit(X_train,y_train)
severityTree
```

Out[41]:
```
DecisionTreeClassifier(criterion='entropy', max_depth=9)
```

4.3 Logistic Regression

Logistic regression is one of the most common and useful classification algorithms in machine learning. It is a classification algorithm, used when the value of the target variable

is categorical in nature. Logistic regression is most commonly used when the data in question has binary output

In [45]:
```
#Summary on Logistic Regression
LRyhat=LR.predict(X_test)
LR_jacc=jaccard_score(y_test,LRyhat,average='macro')
LR_f1=f1_score(y_test,LRyhat,average='macro')
LR_acc=accuracy_score(y_test,LRyhat)
print("LR_jacc=",LR_jacc,"LR_f1=",LR_f1,"LR_acc=",LR_acc)
#log_loss
yhat_prob = LR.predict_proba(X_test)
LR_logloss=log_loss(y_test, yhat_prob)
print("Log_loss=",LR_logloss)
```

```
LR_jacc= 0.3527524426776879 LR_f1= 0.4154650876236762 LR_acc= 0.702989623643026
Log_loss= 0.5852487663216132
```

5.Result and Evaluation

Evaluation metrics used to test the accuracy of our models were jaccard index, F-1 score,accuracy score and log-loss for logistic regression.Based on three Jaccard/F1 and accuracy scores of ML(machine learning) models; K-Nearest Neighbor(KNN), Decision Tree and Logistic Regression.

Overall,the three models got similar accuracy score. KNN achieved better Jaccard and F1 scores, however, KNN need spent much more time for the computation. In addition, logistic regression made the most sense because of its binary nature.

Choosing different k, max depth and hyperamater C values helped to improve our accuracy to be the best possible.

Out[46]:

	Jaccard Score	F1-score	Subset Accuracy Score	Log Loss
K Nearest Neighbors	0.371912	0.456703	0.699514	NA
Decision Tree	0.353974	0.418299	0.702630	NA
Logistic Regression	0.352752	0.415465	0.702990	0.585249

6.Conclusion

KNN,Decision Tree and Logistic Regression have consistent accuracy to predict the severity of traffic accident. Based on machine learning modeling results, the features of ADDRTYPE, JUNCTIONTYPE, WEATHER, ROADCOND and LIGHTCOND have great impact on accident severity, however, not the single factor to affect the severity of traffic accident, due to most of accident happened at daylight and clear weather, and the drivers attention and physical

condition maybe the other factors which not touched on in existed dataset. Anyway, it is meaningful to consider these major factors into traffic design, traffic accident alert and drivers travel planning, to avoid serious traffic accident.