CAR ACCIDENT SEVERITY

Tim Skates

IBM Data Science Capstone

INTRODUCTION

- Last year almost 39,000 people lost their lives in car accidents in the United States with 4.4 million people having to seek medical care after a collision

- What external factors may result in a higher frequency of crashes and/or a greater chance of injury?

- The goal of this project is to determine if we can predict the severity or risk of an accident and what factors may be at play.

DATA

- Data from the Seattle Department of Transportation and includes 194,673 accidents and 37 features such as location or address of accident, weather conditions, driver impairment, and collision type.

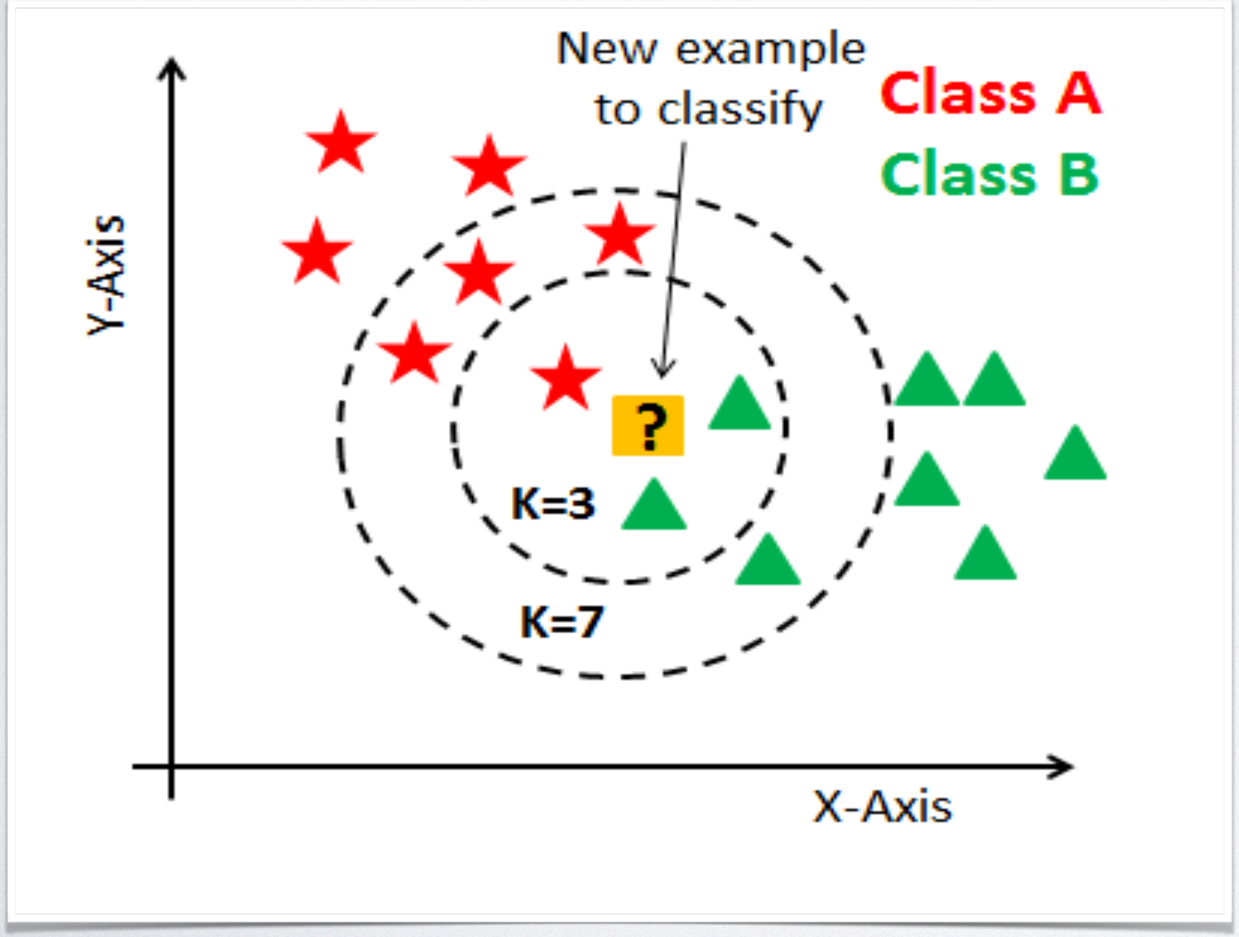
- Data set needed to be cleaned using Pandas function to remove all incomplete data.

FEATURE SELECTION

- To create the most accurate model, we want to choose attributes in the data set that will affect the dependent variable (in the case it is Severity Code) and getting rid of any that don't apply.

- The attributes we selected were as follows: Address Type, Location, Junction type, Road Condition, Light Condition, and Speed.

K-NEAREST NEIGHBOR




```
from sklearn.neighbors import KNeighborsClassifier
k=23

kneighbors = KNeighborsClassifier(n_neighbors = k).fit(x_train, y_train)
k_Predict = kneighbors.predict(x_test)
k_Predict[0:5]
```

```
array([2, 1, 1, 1, 1])
```

```
j=jaccard_score(y_test,k_Predict)
f=f1_score(y_test, k_Predict, average = 'macro')
print("Jaccard Score: ",j)
print("F1 Score: ",f)
```

```
Jaccard Score:  0.42309313105773283
F1 Score:  0.6074617239040984
```

DECISION TREE

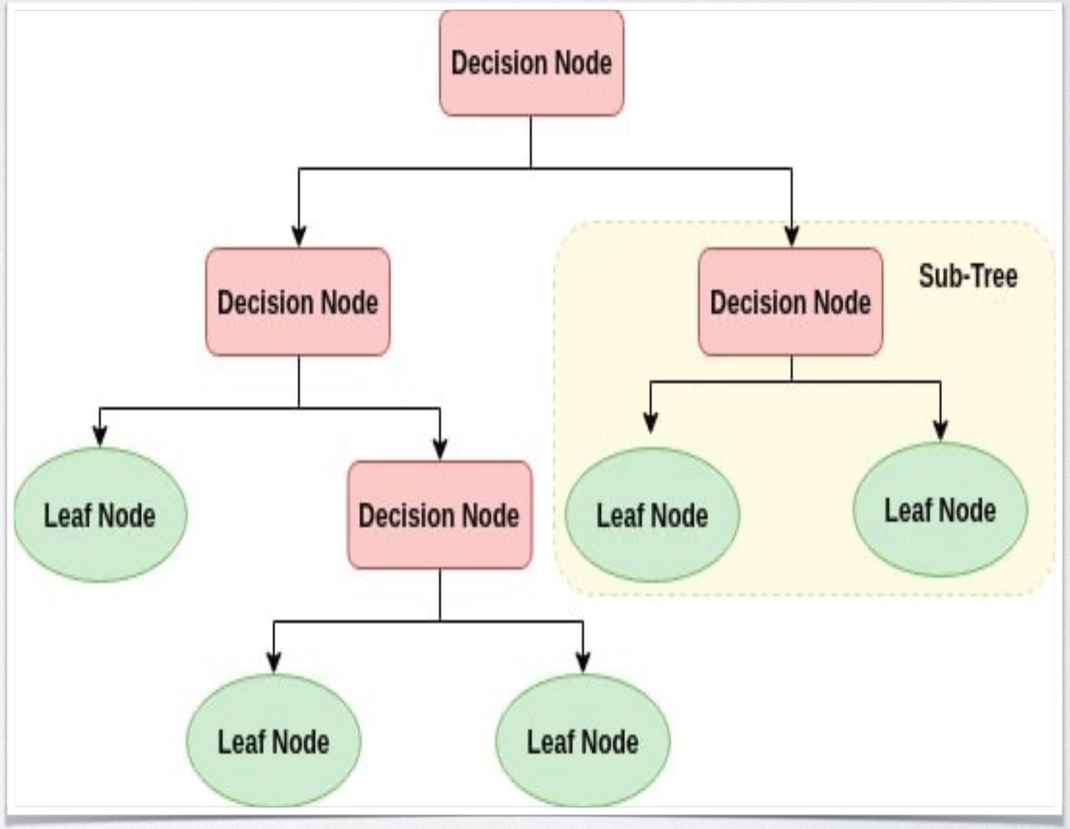
```
from sklearn.tree import DecisionTreeClassifier
dt = DecisionTreeClassifier (criterion = 'entropy', max_depth = 9)

dt.fit(x_train, y_train)
dt_predict = dt.predict(x_test)
dt_predict[0:5]
```

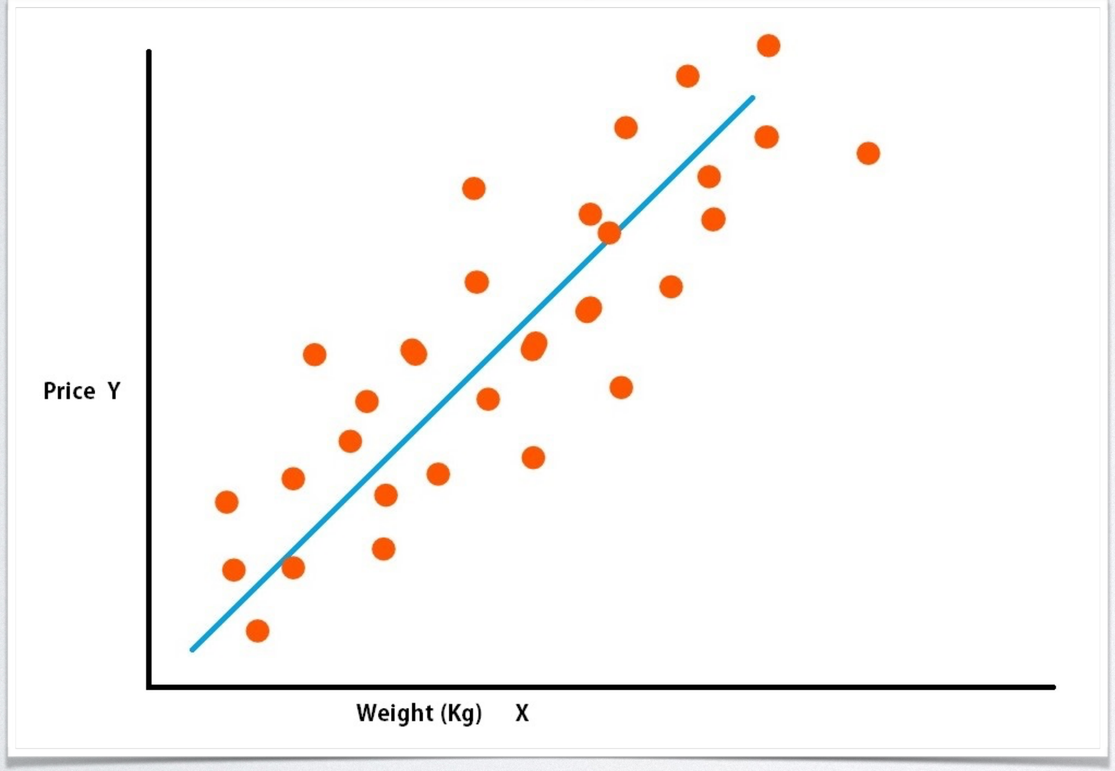
```
array([2, 1, 1, 1, 1])
```

```
j2=jaccard_score(y_test,dt_predict)
f2=f1_score(y_test,dt_predict, average = 'macro')
print("Jaccard Score: ",j2)
print("F1 Score: ",f2)
```

```
Jaccard Score:  0.4391772524606713
F1 Score:  0.617955163978534
```



LINEAR REGRESSION




```
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import confusion_matrix
LogReg = LogisticRegression(C = 6, solver = 'liblinear').fit(x_train, y_train)

logreg_predict = LogReg.predict(x_test)
logreg_prob= LogReg.predict_proba(x_test)
logreg_prob
```

```
array([[0.3285926 , 0.6714074 ],
       [0.65971659, 0.34028341],
       [0.58208952, 0.41791048],
       ...,
       [0.59150057, 0.40849943],
       [0.38589354, 0.61410646],
       [0.50987312, 0.49012688]])
```

```
j1=jaccard_score(y_test,logreg_predict)
f1=f1_score(y_test,logreg_predict, average = 'macro')
print("Jaccard Score: ",j1)
print("F1 Score: ",f1)
print("Log Loss: ",log_loss(y_test,logreg_prob))
```

```
Jaccard Score:  0.4585652934627008
F1 Score:  0.6090004494835676
Log Loss:  0.6554544694261182
```

MODELS

- Evaluation Metrics used were the Jaccard Score and F1 Score.
- Increase Accuracy by using different K-values, max depth, and C values.

	ML Model	Jaccard Score	F1 Score
0	KNN	0.423093	0.607462
1	Linear Regression	0.458565	0.609000
2	Decision Tree	0.439177	0.617955

CONCLUSION

- Linear Regression is the most accurate model for us to use to predict Accident Severity.
- We can use this model to determine what attributes contribute to a higher risk of being involved in a car accident and how severe the accident may be.
- Officials could use this information to make decisions to reduce risk and harm.