CAPSTONE PROJECT

ANALYZING THE SEVERITY OF CAR ACCIDENTS

TABLE OF CONTENTS

COVER PAGE

I always believe in the quote “The key to success is the dedication to lifelong learning”, and so I am here with this DATA SCIENCE Course, to gain knowledge and for the in-depth understanding of the Analytics.

I hold master’s degree in Applied Chemistry and in Human Resources with overall around 8 years of professional experience in a leading R&D solution provider in India.

As Data is the key asset of HR and HR team plays a vital role in providing Recruitment data, career progression data, training data, absenteeism figures, productivity data, personal development reviews, staff satisfaction data, scanning social media data, or analyzing the content of emails to gauge employee sentiment etc. I decided to enhance my knowledge and chose this course. Since all organizations are moving towards data-driven approach, I want to expand my area of expertise from a strategic HR approach to Data Driven HR.

I have learnt so many aspects of Data Science and have applied my learning in the capstone project titled “**ANALYZING THE SEVERITY OF CAR ACCIDENTS”.**

The source of the data is from IBM Coursera site and below is the link for the same.

https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv

The data is in CSV format and the same has been used for the analysis.

Using the above data set, I have tried to analyze the severity of car accidents and the detailed report can be viewed.

Author- Ramyashree Nuchin Veerappa

Date - 9/23/20

Introduction

Life is so busy in this 21st century and we all are often in fast-track!! Everybody rushes and always want to reach their destination in no time!! This hurry sometimes can be life threatening as accidents might happen. Now-a-days, road accidents are very common and most of the times they lead to loss of property, injuries and can even cause death. So, wouldn't it be great to try to understand the most common causes, in order to prevent them from happening?

In most cases, not paying enough attention during driving, abusing drugs and alcohol or driving at very high speed are the main causes of occurring accidents that can be prevented by enacting harsher regulations. Besides this, weather, visibility, or road conditions are the major uncontrollable factors that can be prevented by revealing hidden patterns in the data and announcing warning to the local government, police and drivers on the targeted roads. In order to understand these common factors that are causing accidents and the correlation between them, I am attempting to analyze the data from City of Seattle's police department showing all the collisions from 2004 till present.

In an effort to reduce the frequency of car collisions, an algorithm must be developed to predict the severity of an accident given the current weather, road and visibility conditions.

The target audience of the project is local Seattle government, police, rescue groups, and last but not least, car insurance institutes. The model and its results are going to provide some advice for the target audience to make insightful decisions for reducing the number of accidents in the city.

Data Understanding

Data understanding relies on business understanding. Data is collected at this stage of the process. The understanding of what the business wants and needs will determine what data is collected, from what sources, and by what methods. CRISP-DM combines the stages of Data Requirements, Data Collection, and Data Understanding from the Foundational Methodology outline.

Data set from the Seattle Police Department has 37 independent variables with 194,673 records collected since 2004 to present. As the main objective of this project is to analyze the severity of the accidents, our dependent/target variable will be 'SEVERITYCODE' and the attributes we will be using to measure the severity of accidents are 'WEATHER', 'ROADCOND' and 'LIGHTCOND'. Our target variable 'SEVERITYCODE' consists of numbers 1 & 2 with 1 being only 'Property Damage Only Collision' and 2 is 'injury collision'.

The ample amount of data collected from past 15 years can be used for the analysis of the severity of accidents only after pre-processing or cleaning the Data. That is, in the collected

dataset there are many irrelevant attributes that are not necessary for our analysis and can be dropped out from the data set.

The first few rows of data set is shown below for better understanding

```
df=pd.read_csv('Data-Collision.csv')
df.head()
```

```
/home/jupyterlab/conda/envs/python/lib/python3.6/site-packages/IPython/core/interactiveshell.py:3072: DtypeWarning: Columns (33) have mixed types.Specify dtype option on import or set low_memory=False.
  interactivity=interactivity, compiler=compiler, result=result)
```

	SEVERITYCODE	X	Y	OBJECTID	INCKEY	COLDETKEY	REPORTNO	STATUS	ADDRTYPE	INTKEY	...	ROADCOND	LIGHTCOND	PED
0	2	-122.323148	47.703140	1	1307	1307	3502005	Matched	Intersection	37475.0	...	Wet	Daylight	
1	1	-122.347294	47.647172	2	52200	52200	2607959	Matched	Block	NaN	...	Wet	Dark - Street Lights On	
2	1	-122.334540	47.607871	3	26700	26700	1482393	Matched	Block	NaN	...	Dry	Daylight	
3	1	-122.334803	47.604803	4	1144	1144	3503937	Matched	Block	NaN	...	Dry	Daylight	
4	2	-122.306426	47.545739	5	17700	17700	1807429	Matched	Intersection	34387.0	...	Wet	Daylight	

Data Preparation

Once the data has been collected, it must be transformed into a useable subset unless it is determined that more data is needed. Once a dataset is chosen, it must then be checked for questionable, missing, or ambiguous cases. Data Preparation is common to CRISP-DM and Foundational Methodology.

As we are clear with our target variable and the variables that can be used to measure severity, we can proceed with Data pre-processing step or cleaning the data.

The dataset in the original form is not ready for data analysis. In order to prepare the data, first, we need to drop the irrelevant columns. In addition, most of the features are of object data types that needs to be converted into numerical data types. Also, we can see that dataset has many null values that has to be expelled from the data set. Once, data is processed, then we can go ahead using the dataset for our analysis and to build model to prevent future accidents or to reduce severity. Below image shows the cleaned data for further processing.

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 169186 entries, 0 to 194672
Data columns (total 7 columns):
 #   Column          Non-Null Count   Dtype 
---  ------          --------------   ----- 
 0   SEVERITYCODE    169186 non-null  int64 
 1   LOCATION        169186 non-null  object
 2   SEVERITYCODE.1  169186 non-null  int64 
 3   SEVERITYDESC    169186 non-null  object
 4   WEATHER         169186 non-null  object
 5   ROADCOND        169186 non-null  object
 6   LIGHTCOND       169186 non-null  object
dtypes: int64(2), object(5)
memory usage: 10.3+ MB
```

```
df['SEVERITYCODE'].value_counts()
```

```
1    113556
2     55630
Name: SEVERITYCODE, dtype: int64
```

Now that the data is unbalanced we have to balance by downsampling

```
from sklearn.utils import resample
df_1 = df[df.SEVERITYCODE==1]
df_2 = df[df.SEVERITYCODE==2]

df_1_dsample = resample(df_1, replace= False, n_samples= 55630, random_state=100)

balanced_df= pd.concat([df_1_dsample, df_2])
balanced_df.SEVERITYCODE.value_counts()
```

```
2    55630
1    55630
Name: SEVERITYCODE, dtype: int64
```

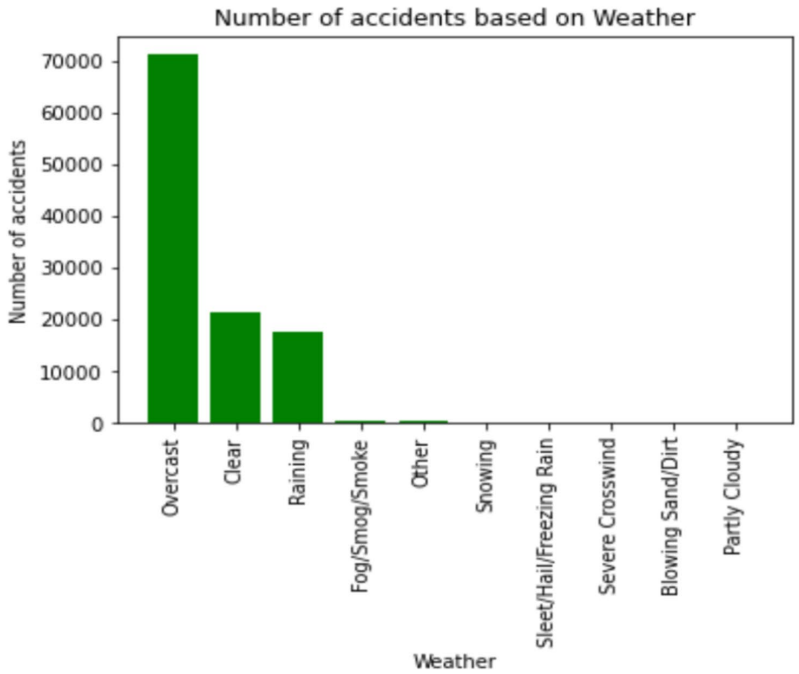
Data Visualization

Let us understand the data patterns or trends more clearly by data visualization method.

BAR GRAPH BASED ON WEATHER CONDITIONS:

From the bar graph, it is clearly understood that the number of accidents are high during overcast and lesser number of accidents during clear and raining weather conditions.

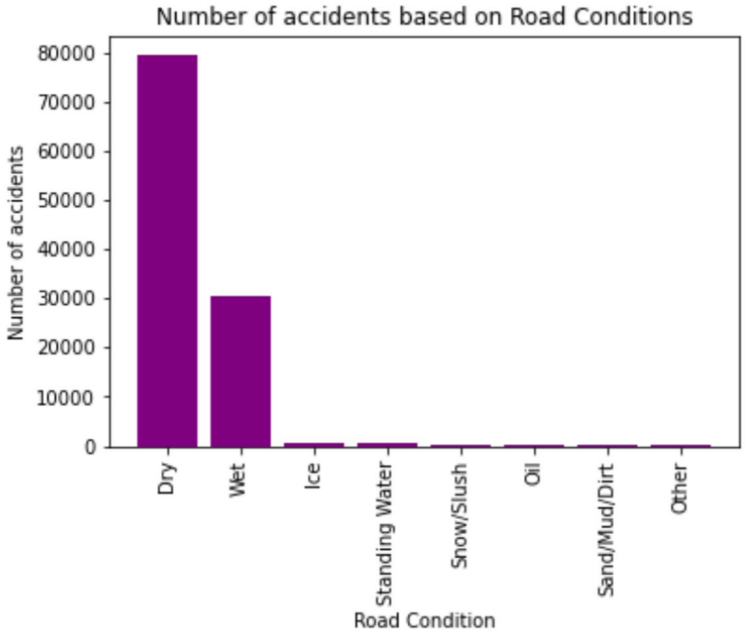
Accidents during all the other weather conditions are very negligible.



BAR GRAPH BASED ON ROAD CONDITIONS:

From the bar graph, it is clearly understood that the number of accidents are high during overcast and lesser number of accidents during clear and raining weather conditions.

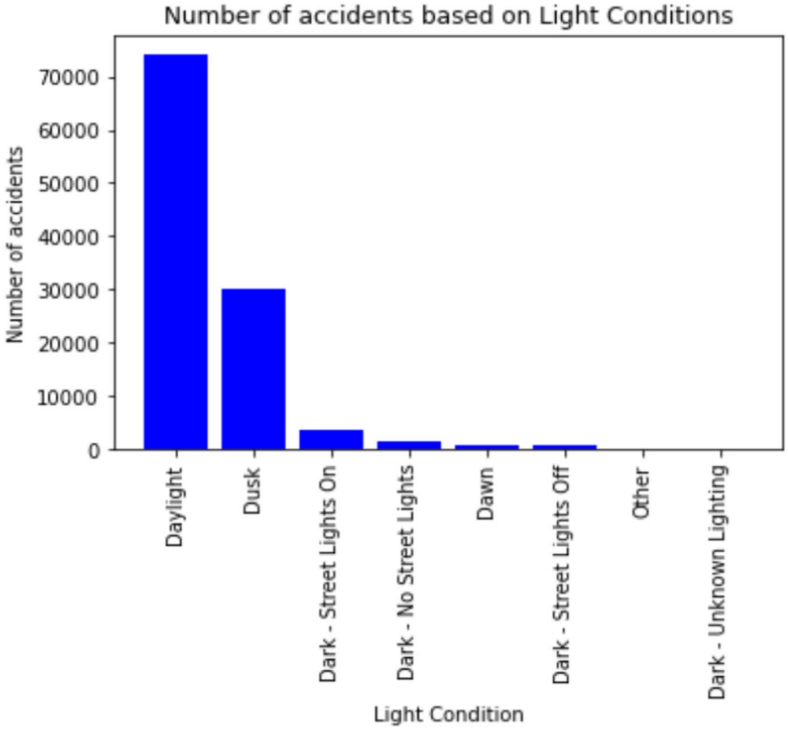
Accidents during all the other weather conditions are very negligible.



BAR GRAPH BASED ON LIGHT CONDITIONS:

From the bar graph, it is clearly understood that the number of accidents are high during daylight and lesser during dusk light condition.

Accidents during all the other light conditions are very negligible.



Executive Summary

For analyzing the data set, to preprocess data to build Machine Learning models I have used Jupyter Notebook and to run the code, I have used Python and its popular packages such as Pandas, NumPy and Sklearn. To share the Jupyter notebook, I have used Git Repository. I have selected the most important features to predict the severity of accidents in Seattle. Among all the features, the following features have the most influence in the accuracy of the predictions:

"WEATHER",

"ROADCOND",

"LIGHTCOND"

Target Variable is "SEVERITYCODE"

I have used Machine Learning models such as KNN, SVM, Decision Tree and Logistic Regression for my prediction and after the analysis I found out that the best model that can be used is KNN for predicting the severity of accidents.

Methodology

Once prepared for use, the data must be expressed through whatever appropriate models, give meaningful insights, and hopefully new knowledge. This is the purpose of data mining: to create knowledge information that has meaning and utility. The use of models reveals patterns and structures within the data that provide insight into the features of interest. Models are selected on a portion of the data and adjustments are made if necessary.

I have initially normalized the data by defining X and Y and then split the data into training and testing data sets for predicting the best model. Below is the screen shot of the same. I used KNN, SVM, Decision Tree and Logistic Regression Models for my prediction.

Below is the screen shot to understand that after data processing, data is cleaned and ready for analysis.

MODELING AND EVALUATION

Let us proceed with building machine learning models for evaluating the data set.

DATA NORMALIZATION

Let's define x and y

```
[81]: x=np.asarray(balanced_df[['WEATHER','ROADCOND','LIGHTCOND']])
      x[0:5]
```

```
[81]: array([[4, 1, 7],
             [2, 1, 7],
             [2, 3, 7],
             [2, 1, 7],
             [5, 3, 7]])
```

```
[24]: y=np.asarray(balanced_df['SEVERITYCODE'])
      y[0:5]
```

```
[24]: array([1, 1, 1, 1, 1])
```

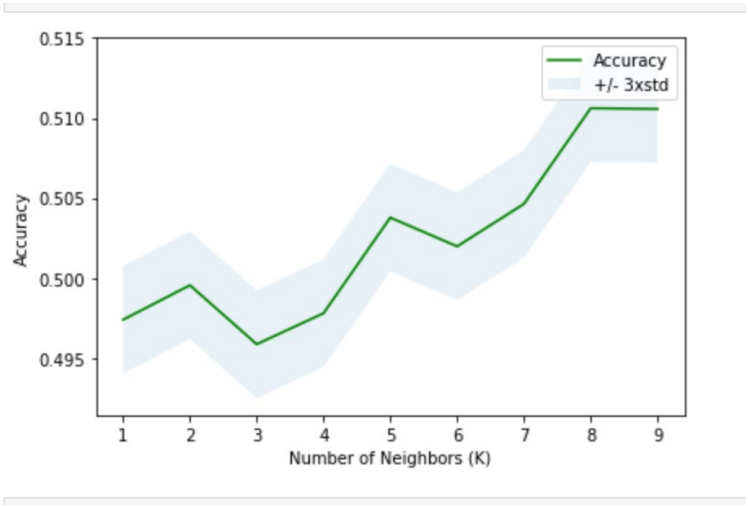
Normalize the data set

```
[89]: x = preprocessing.StandardScaler().fit(x).transform(x)
      x[0:5]
```

```
[89]: array([[ 0.82254307, -0.61044261,  0.66564029],
             [-0.72095207, -0.61044261,  0.66564029],
             [-0.72095207,  1.43198435,  0.66564029],
             [-0.72095207, -0.61044261,  0.66564029],
             [ 1.59429064,  1.43198435,  0.66564029]])
```

The selected models must be tested. This is usually done by having a pre-selected test, set to run the trained model on. This will allow us to see the effectiveness of the model/models on a set it sees as new. Results from this are used to determine efficacy of the model and foreshadows its role in the next and final stage.

Prediction by KNN: Found that K=8 would be the best value for the model.



Code used for other models is as shown below.

Prediction by Decision Tree

```
from sklearn.tree import DecisionTreeClassifier
LoanTree = DecisionTreeClassifier(criterion="entropy", max_depth = 4)
LoanTree.fit(x_train,y_train)
predTree = LoanTree.predict(x_test)
print("DecisionTrees's Accuracy:", metrics.accuracy_score(y_test, predTree))
print("DT Jaccard index: %.2f" % jaccard_similarity_score(y_test, yhat1))
print("DT F1-score: %.2f" % f1_score(y_test, yhat1, average='weighted'))
```

```
DecisionTrees's Accuracy: 0.5216909341482413
DT Jaccard index: 0.52
DT F1-score: 0.48
```

Prediction by Logistic Regression

```
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import log_loss
LR = LogisticRegression(C=0.01, solver='liblinear').fit(x_train,y_train)
yhat_lr = LR.predict(x_test)
yhat_prob = LR.predict_proba(x_test)
print('Accuracy of logistic regression classifier on test set: {:.2f}'.format(LR.score(x_test, y_test)))
print("LR Jaccard index: %.2f" % jaccard_similarity_score(y_test, yhat_lr))
print("LR F1-score: %.2f" % f1_score(y_test, yhat_lr, average='weighted') )
print("LR LogLoss: %.2f" % log_loss(y_test, yhat_prob))
```

```
Accuracy of logistic regression classifier on test set: 0.52
LR Jaccard index: 0.52
LR F1-score: 0.50
LR LogLoss: 0.69
```

Results

Based on the Jaccard and F1 Score we can clearly conclude that KNN is the best model for predicting the severity of car accidents.

Report of the accuracy of the built model using different evaluation metrics:

Algorithm	Jaccard	F1 Score	Log Loss
KNN	0.52	0.50	NA
SVM	0.52	0.48	NA
Decision Tree	0.52	0.48	NA
Logistic Regression	0.52	0.50	0.69

Discussions

In this analysis, our main objective was to analyze the severity of car accidents based on weather conditions, road conditions, and many other factors. Even though our data was of good size, there were number of missing elements and we needed to clean the data in order to get a good result. We had to drop many columns and unfortunately the column 'SPEEDING' because of too many missing entries and it was one of the important factors that should have been considered for the analysis of severity of accidents and to increase the efficiency of the machine learning models.

From the analysis, it is clear that most accidents are caused during overcast, in the daylight with dry/wet roads and are minor in nature. This could be helpful to the police department in understanding where to install more stop signs or speed bumps to avoid speeding specially during turns to overcome bad road conditions, installing streetlights at regular distances to overcome bad light conditions.

Based on the above report obtained KNN could be the best model to proceed with predicting severity of car accidents.

Conclusion

Although this analysis has given us some good insights, a much closer inspection would have been required to understand the impact of other important variables. From the above analysis it is clearly understood that accidents occuring are minor and the public can be alerted with the above accident causing conditions and accidents can be avoided with precautionary measures. Also, as the accidents involved considerable amount of loss of property or injuries, our findings can be helpful to the Seattle Police Department in enforcing some new measures to prevent future accidents.

References

Previous courses of Coursera that involves detailed explanation of Machine Learning Models and Skills Network Labs for hands on experience and for the creation, utilization of Jupyter Notebook.

Acknowledgement

I thank my family and Coursera teachers and everyone who have supported me directly or indirectly to learn Data Science.