Predict Accident Severity

Introduction/ Business Problem

Road traffic accidents have a high cost of loss of life and economy of a city, state or country. Accidents may cause one or the other negative impacts in terms of property damage, medical treatment costs, insurance cost, legal issues etc. Identification of factors causing road accidents by studying past trends are of importance as it would provide guidance in setting up preventive measures and issuing guidelines to drivers/ general public which can benefit in reducing the occurrence of road accidents.

In this report, study is conducted on the dataset provided by Seattle Department of Transportation Traffic Management Division. The aim is to find the correlation between weather, location and road condition to predict severity of road accidents. In this report machine learning models are applied on past accident dataset to predict the accident severity which is divided into two levels 1-Property damage only collision and 2-Injury collision.

Overall, output of this study would help in better town-planning (by installation of traffic light, street lights etc at critical junctions/ roads) as well as reduce loss of life by setting up critical levels of emergency services responders across the city of Seattle.

Data Understanding

The dataset is provided by the Seattle Department of Transportation Traffic Management Division covers accidents in from Janauary'2004 to May'2020. Overall dataset covers 194,673 accidents capturing 37 features and each accident is given a severity code of 1 or 2 by the authorities. Dataset is unbalanced as 70% of accidents are marked with severity code = 1 (property damage only collision) while remaining 30% of accidents are marked with severity code = 2 (injuries collision). This would require balancing of dataset to achieve accurate predictive model.

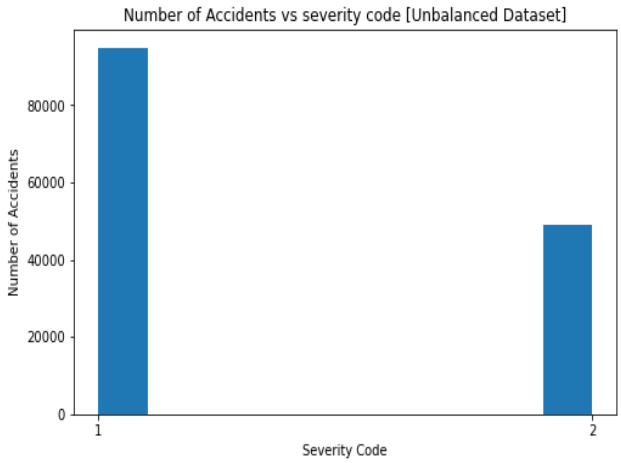
While the dataset is unbalanced it covers key features such as location of accident, weather conditions, road conditions, whether driver was under influence, whether accident was due to inattention etc. All these features/ parameters would play a pivotal role in predicting accident severity by using machine learning models. However, dataset comprises of features which have missing or unknown values. All such accidents/ data points would be excluded as part of data preparation for predictive modeling.

Data Pre-processing

The provided raw dataset is unbalanced i.e. 70% of records are marked under severity code = 1 while only 30% of records are marked under severity code = 2. This would require cleaning of data and generating a balanced dataset to remove any bias while applying machine learning models.

Below graphs depicts the unbalance dataset i.e. Number of accidents vs Severity Code

Unbalanced Dataset



Different steps are applied on the base dataset to clean the data for final implementation of machine learning models

Step 1 – Remove unwanted features/ columns from the dataset. Columns reduced from 37 to 19.

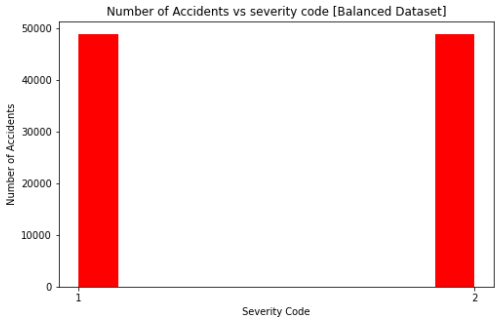
Following columns are removed from the dataset

'OBJECTID','INCKEY','LOCATION','COLDETKEY','REPORTNO','STATUS','INTKEY','EXCEPTRSNCODE', 'EXCEPTRSNDESC','SEVERITYDESC','INCDATE','SDOT_COLCODE','SDOT_COLDESC','SDOTCOLNUM','ST_COLCODE', 'ST_COLDESC','SEGLANEKEY','CROSSWALKKEY','INCDTTM'

Step 2 – Identify features/ columns which have blank, unknown or no data and replace the values with NaN. Post replacements drop all such records where one or the other feature is NaN.

This reduces the records from 194,673 to 143,747

Step 3 – Balance the dataset so that it has equal number of records for Severity code = 1 and Severity code 2. This is to normalize the data and reduce bias while applying machine learning models. 48,926 records each for severity code = 1 and 2. Total records in balanced dataset = 97,852.



Step 4 – Label encoding to convert categorical variables (WEATHER, ROADCOND, LIGHTCOND) to numerical variables for ML modeling

```
[13]: label_encoder = preprocessing.LabelEncoder()
      df_balanced ['WEATHER']= label_encoder.fit_transform(df_balanced['WEATHER'])
      df_balanced ['ROADCOND']= label_encoder.fit_transform(df_balanced['ROADCOND'])
      df_balanced ['LIGHTCOND']= label_encoder.fit_transform(df_balanced['LIGHTCOND'])
      df_balanced.head()
```

[13]:

COUNT	PEDCYLCOUNT	VEHCOUNT	JUNCTIONTYPE	UNDERINFL	WEATHER	ROADCOND	LIGHTCOND
0	0	2	Mid-Block (not related to intersection)	N	1	0	5
0	0	2	Mid-Block (not related to intersection)	N	1	0	5
0	0	2	Mid-Block (but intersection related)	N	3	6	2
0	0	2	At Intersection (intersection related)	N	1	0	5
0	0	2	At Intersection (intersection related)	N	3	6	5

Step 5 - Reduce the balanced dataset to only 3 key features which are to be used for machine learning modeling to predict accident severity. 3 key features are Weather, Road condition, Light condition

[19]:

	WEATHER	ROADCOND	LIGHTCOND
139054	1	0	5
145014	1	0	5
186440	3	6	2
191150	1	0	5
159954	3	6	5

```
[20]: Y = df_balanced['SEVERITYCODE'].values
      Y[0:5]
      #Y.shape
```

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

Step 6 – Preparing Train and test dataset. For this exercise, dataset is split into 70% train and 30% test data.

Preparing Test (30%) and Train (70%) data set

```
[22]: X_train, X_test, Y_train, Y_test = train_test_split(X,Y,test_size=0.3,random_state=0)
      print(X_train.shape,Y_train.shape)
      print(X_test.shape,Y_test.shape)
```

```
(68496, 3) (68496,)
(29356, 3) (29356,)
```

Machine Learning Models

For this study, 3 machine learning models are used

1. Logistic regression
2. K-nearest neighbor
3. Decision Tree

All the models are present with classification matrix, model accuracy scores and classification report

1. **Logistic regression**

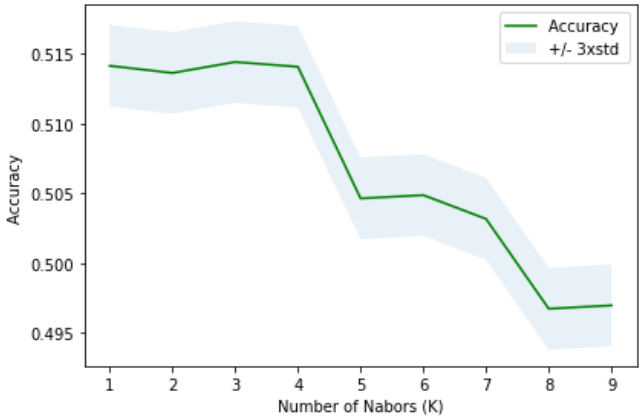
   Logistic Regression is a classifier that estimates discrete values (binary values like 0/1, yes/no, true/false) based on a given set of an independent variables. It basically predicts the probability of occurrence of an event by fitting data to a logistic function. Hence it is also known as logistic regression. The values obtained would always lie within 0 and 1 since it predicts the probability.

   Results of logistic regression model are:

```
[[ 4299 10444]
 [ 3848 10765]]

              precision    recall  f1-score   support

           1       0.53      0.29      0.38     14743
           2       0.51      0.74      0.60     14613

   micro avg       0.51      0.51      0.51     29356
   macro avg       0.52      0.51      0.49     29356
weighted avg       0.52      0.51      0.49     29356


0.5131489303719853
```

2. **K – Nearest Neighbours**

   K nearest neighbours algorithm used for both classification and regression problems. It basically stores all available cases to classify the new cases by a majority vote of its k neighbours.
   Highest accuracy for the given dataset is achieved for k=4. Hence, K=4 is used as classifier to implement K-Nearest neighbours model



Results of KNN model are:

```
[[ 4784  9959]
 [ 4306 10307]]
```

	precision	recall	f1-score	support
1	0.53	0.32	0.40	14743
2	0.51	0.71	0.59	14613
micro avg	0.51	0.51	0.51	29356
macro avg	0.52	0.51	0.50	29356
weighted avg	0.52	0.51	0.50	29356

```
0.5140686742062951
```

3. **Decision Tree classifier**

Decision Tree makes decision with tree-like model. It splits the sample into two or more homogenous sets (leaves) based on the most significant differentiators in the input variables. To choose a differentiator (predictor), the algorithm considers all features and does a binary split on them (for categorical data, split by category; for continuous, pick a cut-off threshold). It will then choose the one with the least cost (i.e. highest accuracy), and repeats recursively, until it successfully splits the data in all leaves (or reaches the maximum depth).

Results of Decision Tree model are:

```
[[ 3677 11066]
 [ 3095 11518]]
```

	precision	recall	f1-score	support
1	0.54	0.25	0.34	14743
2	0.51	0.79	0.62	14613
micro avg	0.52	0.52	0.52	29356
macro avg	0.53	0.52	0.48	29356
weighted avg	0.53	0.52	0.48	29356

```
0.5176113911977108
```

Machine Learning Results

Implementation of 3 machine learning models has yielded an accuracy score of approx. 51%.

Out of the 3 machine models, Decision Tree classifier has shown an accuracy of 51.7% which is marginally better as compared to accuracy of Logistic regression (51.3%) and KNN model (51.4%).

This shows that by using the 3 features of weather, road condition and light condition all the 3 models would predict the severity code of an accident with an accuracy of 51%.

Conclusion, Discussion & Future Suggestions

At the outset of this project 3 key features were picked up to predict the accident severity for Seattle Department of Traffic Management. On implementing the machine learning models on the dataset an accuracy score of 51% was achieved which is not great. This eludes to that the models were under fitted and would require more training dataset. Also part of data preparation, size of dataset was reduced due to removal of records that had many features as blank or unknown.

Keeping above findings in mind, it is suggested to increase the size of the dataset by collating complete data points for all the features which could assist in better correlation of important features with severity of accident. In addition to this, more features (features other than weather, road condition, and light condition) should be included for training the machine learning models for better prediction.