Predicting- Car Accident Severity for Seattle.

Final Submission- S.S

Introduction

According to the National Safety Council report, approximately 38,300 people were killed and about 4.4 million injured in the road accidents United States. There are a variety of reasons that contribute to accidents. Some of the reasons are adverse Weather and Traffic conditions that cause accident prone situations. Predicting likelihood of vehicular crashes because of Weather and Traffic features would be a major step towards achieving better road safety.

- The target audience of the project is local Seattle government, police, rescue groups and first responders like emergency department

- The Traffic Records Group, Traffic Management Division, Seattle Department of Transportation, provides data for all collisions and crashes that have occurred in the state from 2004 to the present day. The objective is to exploit this data to extract vital features that would enable us to end up with a good model that would enable the prediction of the severity of future accidents that take place in the city.

Data acquisition

- Source- Data on collisions within the city of Seattle is available through an Open Data initiative hosted by Seattle Geo Data (Collisions, n.d.). It is noted this dataset is the example dataset provided for this capstone project.

  - The dataset is available as comma-separated values (CSV) files, KML files, and ESRI shapefiles that can be downloaded from the Seattle Open GeoData Portal.
  - Dataset contains 221,389 records and 40 variables.
  - The data contains several categorical fields and corresponding descriptions which could help us in further analysis. I have used SEVERITYCODE as our dependent variable Y, and try different combinations of independent variables X to get the result. Since the observations are quite large, we may need to filter out the missing value and delete the unrelated columns first. Then we can select the factor which may have more impact on the accidents, such as address type, weather, road condition, and light condition.

Dependent Variable- Description

The SEVERITYCODE field contains a code that corresponds to the severity of the collision. and SEVERITYDESC contains a detailed description of the severity of the collision.

From the data we can conclude that there were 349 collisions that resulted in at least one fatality, and 3,102 collisions that resulted in serious injuries.

The following table lists the meaning of each of the codes used in the SEVERITYCODE field

SEVERITYCODE VALUE	Meaning
1	Accidents resulting in property damage
2	Accidents resulting in injuries
2b	Accidents resulting in serious injuries
3	Accidents resulting in fatalities
0	Data Unavailable i.e. Blanks

Data Cleaning

The data collected is real world data and contained missing values. The missing values were encoded in a number of different ways, such as ‘Unknown’, ‘N/A’, ‘Not Reported’, or ’’.

The dataset in the original form is not ready for data analysis. In order to prepare the data, first, we need to drop the non-relevant columns. In addition, most of the features are of object data types that need to be converted into numerical data types.

These fields are OBJECTID, INCKEY, COLDETKEY, INTKEY, SEGLANEKEY, CROSSWALKKEY, and REPORTNO. Other fields suchs as EXCEPTRSNCODE, SDOT_COLCODE, SDOTCOLNUM and LOCATION and their corresponding descriptions (if any) are categorical but have a large number of distinct values that shall not be that much useful for analysis. The INCDATE and INCDTTM denote the date and the time of the incident but may not be of use in further analyses

Dataset before and after Cleaning

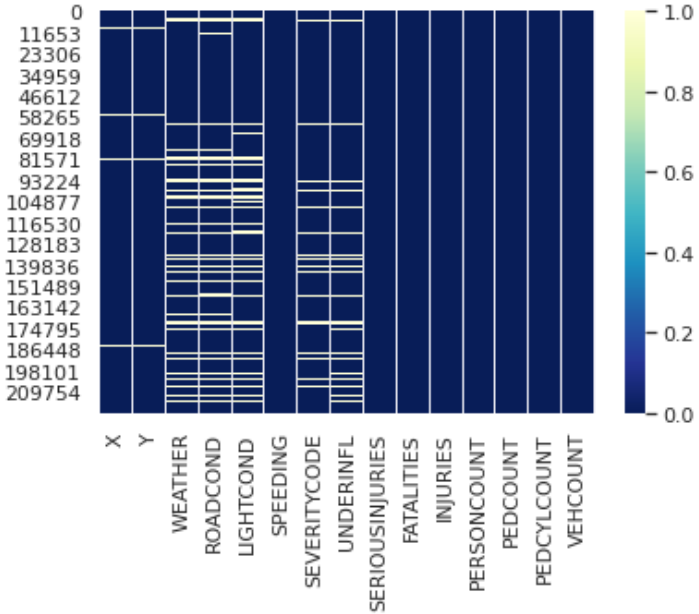
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 221389 entries, 0 to 221388
Data columns (total 40 columns):
 #   Column           Non-Null Count   Dtype
---  ------           --------------   -----
 0   X                213918 non-null  float64
 1   Y                213918 non-null  float64
 2   OBJECTID         221389 non-null  int64
 3   INCKEY           221389 non-null  int64
 4   COLDETKEY        221389 non-null  int64
 5   REPORTNO         221389 non-null  object
 6   STATUS           221389 non-null  object
 7   ADDRTYPE         217677 non-null  object
 8   INTKEY           71884 non-null   float64
 9   LOCATION         216801 non-null  object
 10  EXCEPTRSNCODE    100986 non-null  object
 11  EXCEPTRSNDESC    11779 non-null   object
 12  SEVERITYCODE     221388 non-null  object
 13  SEVERITYDESC     221389 non-null  object
 14  COLLISIONTYPE    195159 non-null  object
 15  PERSONCOUNT      221389 non-null  int64
 16  PEDCOUNT         221389 non-null  int64
 17  PEDCYLCOUNT      221389 non-null  int64
 18  VEHCOUNT         221389 non-null  int64
 19  INJURIES         221389 non-null  int64
 20  SERIOUSINJURIES  221389 non-null  int64
```

Sample data before cleaning- contains 40 rows

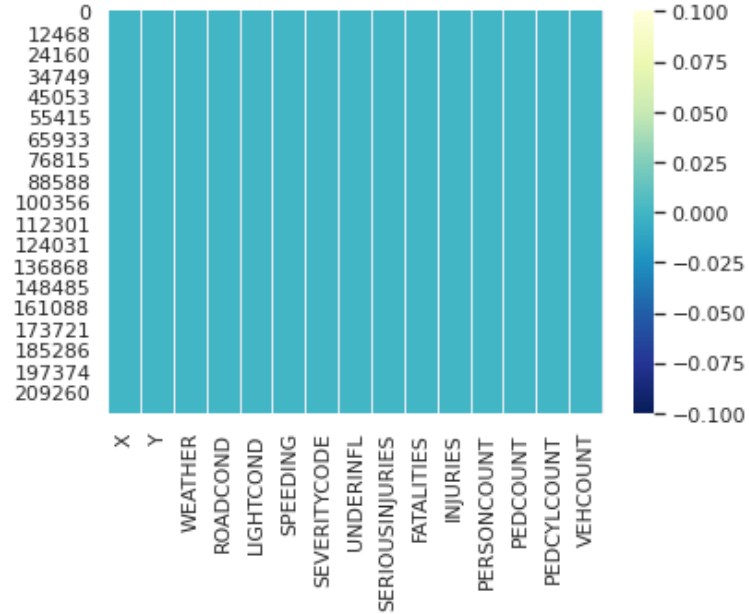
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 221389 entries, 0 to 221388
Data columns (total 15 columns):
 #   Column           Non-Null Count   Dtype
---  ------           --------------   -----
 0   X                213918 non-null  float64
 1   Y                213918 non-null  float64
 2   WEATHER          194969 non-null  object
 3   ROADCOND         195050 non-null  object
 4   LIGHTCOND        194880 non-null  object
 5   SPEEDING         9928 non-null    object
 6   SEVERITYCODE     221388 non-null  object
 7   UNDERINFL        195179 non-null  object
 8   SERIOUSINJURIES  221389 non-null  int64
 9   FATALITIES       221389 non-null  int64
 10  INJURIES         221389 non-null  int64
 11  PERSONCOUNT      221389 non-null  int64
 12  PEDCOUNT         221389 non-null  int64
 13  PEDCYLCOUNT      221389 non-null  int64
 14  VEHCOUNT         221389 non-null  int64
dtypes: float64(2), int64(7), object(6)
memory usage: 25.3+ MB
```

After cleaning- 15 columns

Data Cleaning scenarios before and After using heatmap



Before Data Cleaning- contains 221389 rows



After data cleaning- contains -171540 (removed blank and duplicate data.

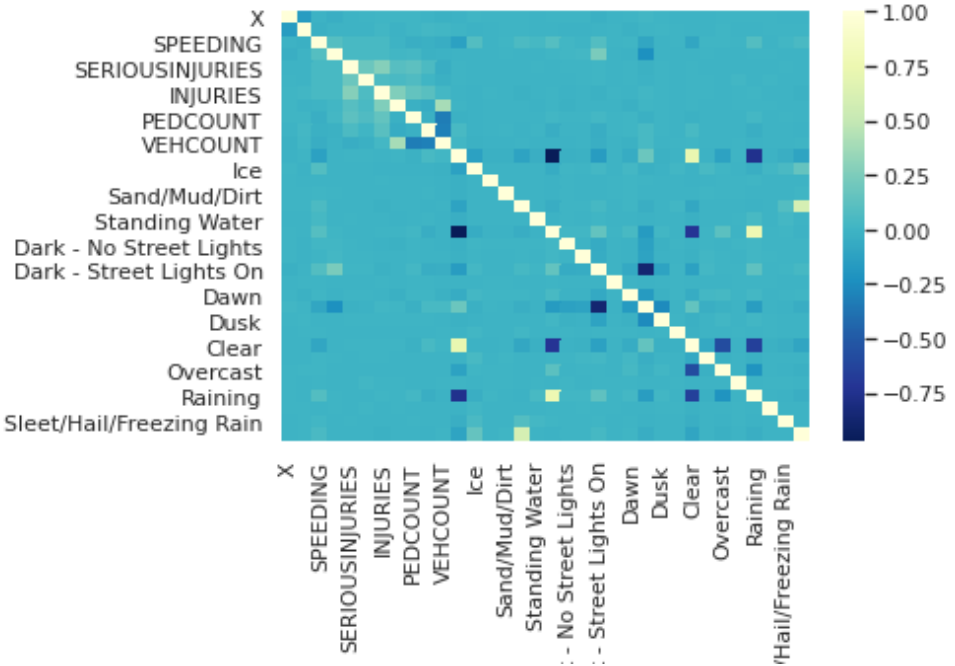
Data Preparation

Fixing the SPEEDING field by encoding it to 0 for the blanks and 1 for the Y values.

Records containing values as Unknown and Other can be considered as null values. Severity Code of 0 corresponds to unknown severity, which can also be treated as null.

Finding the correlation among the features of the dataset helps understand the data better. For example, in the heatmap shown below, it can be observed that some features have a strong positive / negative correlation while most of them have weak / no correlation.

We now do an one-hot encoding of the WEATHER, ROADCOND, and LIGHTCOND fields as they are categorical.



Creating Training and Test set for Machine learning Models

- The datasets x and y are constructed. The set x contains all the training examples and y contains all the labels. Feature scaling of data is done to normalize the data in a dataset to a specific range.
- After normalization, they are split into x_train, y_train, x_test, and y_test. The first two sets shall be used for training and the last two shall be used for testing. Upon choosing a suitable split ratio, 80% of data is used for training and 20% of is used for testing

```
array([[-2.09945123,  1.19566934, -0.23626717, -0.23857092, -0.1063473 ,
        -0.04124978, -0.57973517, -0.35951016, -0.21022234, -0.18556125,
         0.0571671 , -1.59611376, -0.07952372, -0.01691142, -0.01839956,
        -0.06976108, -0.02391978,  1.6461363 , -0.08993329, -0.08074068,
        -0.61637295, -0.00996015, -0.12012905,  0.71362769, -0.18554957,
        -0.01565661, -1.34227426, -0.05648197, -0.43418077, -0.00763893,
         2.06989132, -0.01207874, -0.02522733, -0.06954962],
       [-1.00770154, -1.56411847, -0.23626717, -0.23857092, -0.1063473 ,
        -0.04124978, -0.57973517, -0.35951016, -0.21022234, -0.18556125,
         0.0571671 , -1.59611376, -0.07952372, -0.01691142, -0.01839956,
        -0.06976108, -0.02391978,  1.6461363 , -0.08993329, -0.08074068,
         1.62239436, -0.00996015, -0.12012905, -1.40129092, -0.18554957,
        -0.01565661, -1.34227426, -0.05648197, -0.43418077, -0.00763893,
         2.06989132, -0.01207874, -0.02522733, -0.06954962],
       [ 0.11216306, -1.61248912, -0.23626717, -0.23857092, -0.1063473 ,
        -0.04124978, -0.57973517, -1.05983925, -0.21022234, -0.18556125,
        -1.64493466, -1.59611376, -0.07952372, -0.01691142, -0.01839956,
        -0.06976108, -0.02391978,  1.6461363 , -0.08993329, -0.08074068,
         1.62239436, -0.00996015, -0.12012905, -1.40129092, -0.18554957,
        -0.01565661,  0.74500423, -0.05648197, -0.43418077, -0.00763893,
        -0.48311715, -0.01207874, -0.02522733, -0.06954962]])
```

Modelling and Evaluation

Models used for the project

1. Decision Tree Classifier
2. Random Forest Classifier
3. Logistic regression classifier

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 highest accuracy and repeats recursively, until it successfully splits the data in all leaves (or reaches the maximum depth). The accuracy of the model against the dependent variable (SEVERITY and its codes are shown below)

```
DecisionTreeClassifier(criterion='entropy', max_depth=5)
```

	precision	recall	f1-score	support
1	1.00	1.00	1.00	22504
2	1.00	1.00	1.00	11068
2b	1.00	1.00	1.00	633
3	1.00	1.00	1.00	71
accuracy			1.00	34276
macro avg	1.00	1.00	1.00	34276
weighted avg	1.00	1.00	1.00	34276

Classification Report for DTC

Random Forest Classifier (RFC)

RFC is an ensemble (algorithms which combines more than one algorithms of same or different kind for classifying objects) tree-based learning algorithm. RFC is a set of decision trees from randomly selected subset of training set. It aggregates the votes from different decision trees to decide the final class of the test object. Similar to Decision Tree classifier , RFT requires an input that specifies a measure that is to be used for classification, along with that a value for the number of estimators (number of decision trees) is required.

```
RandomForestClassifier(n_estimators=75)
```

A hyper parameter RFT was used to determine the best choices for the above mentioned parameters. RFT with 75 DT's using entropy as the measure gave the best accuracy when trained and tested on pre-processed accident severity dataset.

	precision	recall	f1-score	support
1	1.00	1.00	1.00	22504
2	1.00	1.00	1.00	11068
2b	1.00	1.00	1.00	633
3	1.00	0.99	0.99	71
accuracy			1.00	34276
macro avg	1.00	1.00	1.00	34276
weighted avg	1.00	1.00	1.00	34276

Logistic regression classifier

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. The chosen dataset has more than two target categories in terms of the accident severity code assigned, one-vs-one (OvO) strategy is employed

	precision	recall	f1-score	support
1	1.00	1.00	1.00	22504
2	1.00	1.00	1.00	11068
2b	1.00	0.99	1.00	633
3	1.00	0.99	0.99	71
accuracy			1.00	34276
macro avg	1.00	0.99	1.00	34276
weighted avg	1.00	1.00	1.00	34276

Classification Report for LogReg Classifier

Result and Discussion

The accuracies of all models applied was 100% which means we can accurately predict the severity of an accident. A bar plot is plotted below with the bars representing the accuracy of each model.

Discussion

From analysis of the model we can say that the model has trained well and fits the training data and performs well on the testing set as well as the training set. We can conclude that this model can accurately predict the severity of car accidents in Seattle.