Predicting Seattle accidents - Capstone assignment

Badrinath

Context:

As part of Capstone assignment we are required to explore Seattle transportation datasets containing accidents data set. Washington/Seattle city has a vision zero ambition where the government wants to reduce accidents to zero by 2030.

Idea of this exercise is to understand the patterns in data, assess if we can build a model to predict severity of accidents.

Business Description:

Need to understand Seattle accidents dataset, understand historical data and extract insights from the data. More specifically predict severity of accidents happening in Seattle city.

Objective of this exercise is to exploit this data to extract insights 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 state. This would further enable Seattle department of Transportation to prioritise their efforts and channel their energy to ensure that fewer fatalities result in automobile collisions.

Data Context:

Data is procured from Seattle city government website, Dataset contains 221525 rows and 40 columns. There were exploratory analysis done to understand the patterns from the past, data cleaning exercises were performed as well

Data Description:

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. The data is also available from RESTful API services in formats such as GeoJSON.

The data contains several categorical fields and corresponding descriptions which could help us in further analysis. We make an attempt at understanding the data in terms of the fields that we shall take into account for later stages of model building.

The X and Y fields denote the longitude and latitude of the collisions. We can visualize the first few non-null collisions on a map. We have used FOLIUM package to visualize accidents on the map

Attributes:

WEATHER
ROADCOND
LIGHTCOND
SPEEDING
SEVERITYCODE
SEVERITYDESC
UNDERINFL
PERSONCOUNT
VEHCOUNT

Detailed description of key fields:

WEATHER field contains a description of the weather conditions during the time of the collision. The ROADCOND field describes the condition of the road during the collision. The LIGHTCOND field describes the light conditions during the collision. The SPEEDING field classifies collisions based on whether or not speeding was a factor in the collision.

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.

The UNDERINFL field describes whether or not a driver involved was under the influence of drugs or alcohol. The PERSONCOUNT and VEHCOUNT indicate how many people and vehicles were involved in a collision respectively.

Additional Data Attributes:

OBJECTID
INCKEY
COLDETKEY
INTKEY
SEGLANEKEY
CROSSWALKKEY
REPORTNO
EXCEPTRSNCODE
SDOT_COLCODE SDOTCOLNUM
LOCATION

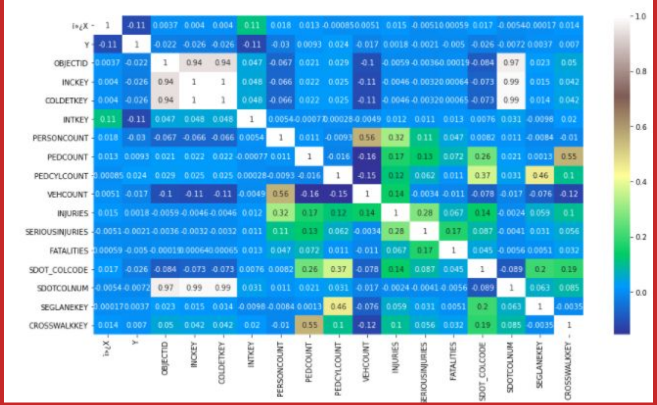
Data Attributes, which are irrelevant:

Several unique identifiers and spatial features are present in the database which may be irrelevant in further statistical analysis.

These fields are are OBJECTID, INCKEY, COLDETKEY, INTKEY, SEGLANEKEY, CROSSWALKKEY, and REPORTNO. Other fields such 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. The data needs to be pre-processed and cleaned up

Exploratory analysis



	ï»¿X	Y	OBJECTID	INCKEY	COLDETKEY	INTKEY	PERSONCOUNT	PEDCOUNT	PEDCYLCOUNT	VEHCOUNT	INJURIES	ERIOUSINJURIES	FATALITIES	SDOT_COLCODE	SDOTCOLNUM	SEGLANEKEY	CROSSWALKKEY
ï»¿X	1	-0.11	0.0037	0.004	0.004	0.11	0.018	0.013	-0.00085	0.0051	0.015	-0.0051	0.00059	0.017	-0.0054	0.00017	0.014
Y	-0.11	1	-0.022	-0.026	-0.026	-0.11	-0.03	0.0093	0.024	-0.017	0.0018	-0.0021	-0.005	-0.026	-0.0072	0.0037	0.007
OBJECTID	0.0037	-0.022	1	0.94	0.94	0.047	-0.067	0.021	0.029	-0.1	-0.0059	-0.0036	0.00019	-0.084	0.97	0.023	0.05
INCKEY	0.004	-0.026	0.94	1	1	0.048	-0.066	0.022	0.025	-0.11	-0.0046	-0.0032	0.00064	-0.073	0.99	0.015	0.042
COLDETKEY	0.004	-0.026	0.94	1	1	0.048	-0.066	0.022	0.025	-0.11	-0.0046	-0.0032	0.00065	-0.073	0.99	0.014	0.042
INTKEY	0.11	-0.11	0.047	0.048	0.048	1	0.0054	-0.00077	0.00028	-0.0049	0.012	0.011	0.013	0.0076	0.031	-0.0098	0.02
PERSONCOUNT	0.018	-0.03	-0.067	-0.066	-0.066	0.0054	1	0.011	-0.0093	0.56	0.32	0.11	0.047	0.0082	0.011	-0.0084	-0.01
PEDCOUNT	0.013	0.0093	0.021	0.022	0.022	-0.00077	0.011	1	-0.016	-0.16	0.17	0.13	0.072	0.26	0.021	0.0013	0.55
PEDCYLCOUNT	-0.00085	0.024	0.029	0.025	0.025	0.00028	-0.0093	-0.016	1	-0.15	0.12	0.062	0.011	0.37	0.031	0.46	0.1
VEHCOUNT	0.0051	-0.017	-0.1	-0.11	-0.11	-0.0049	0.56	-0.16	-0.15	1	0.14	-0.0034	-0.011	-0.078	-0.017	-0.076	-0.12
INJURIES	0.015	0.0018	-0.0059	-0.0046	-0.0046	0.012	0.32	0.17	0.12	0.14	1	0.28	0.067	0.14	-0.0024	0.059	0.1
SERIOUSINJURIES	-0.0051	-0.0021	-0.0036	-0.0032	-0.0032	0.011	0.11	0.13	0.062	-0.0034	0.28	1	0.17	0.087	-0.0041	0.031	0.056
FATALITIES	0.00059	-0.005	-0.00019	0.00064	0.00065	0.013	0.047	0.072	0.011	-0.011	0.067	0.17	1	0.045	-0.0056	0.0051	0.032
SDOT_COLCODE	0.017	-0.026	-0.084	-0.073	-0.073	0.0076	0.0082	0.26	0.37	-0.078	0.14	0.087	0.045	1	-0.089	0.2	0.19
SDOTCOLNUM	-0.0054	-0.0072	0.97	0.99	0.99	0.031	0.011	0.021	0.031	-0.017	-0.0024	-0.0041	-0.0056	-0.089	1	0.063	0.085
SEGLANEKEY	0.00017	0.0037	0.023	0.015	0.014	-0.0098	-0.0084	0.0013	0.46	-0.076	0.059	0.031	0.0051	0.2	0.063	1	-0.0035
CROSSWALKKEY	0.014	0.007	0.05	0.042	0.042	0.02	-0.01	0.55	0.1	-0.12	0.1	0.056	0.032	0.19	0.085	-0.0035	1

Observations:

Univariate analysis of variables were done to understand the distribution.

Heatmap analysis was done to understand correlation between various variables. Some key findings are

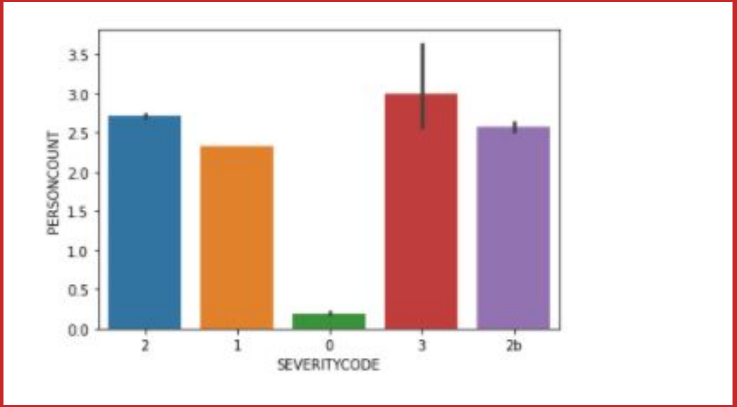
as follows:

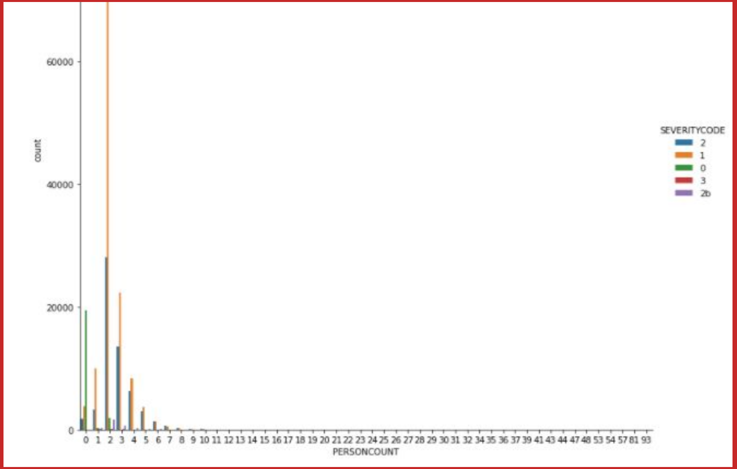
Positive Correlation:

(a) Number of pedestrians involved in the collision (PEDCOUNT) and Severit

(b) Number of people involved in the collision (PERSONCOUNT) and Severity

Exploratory analysis





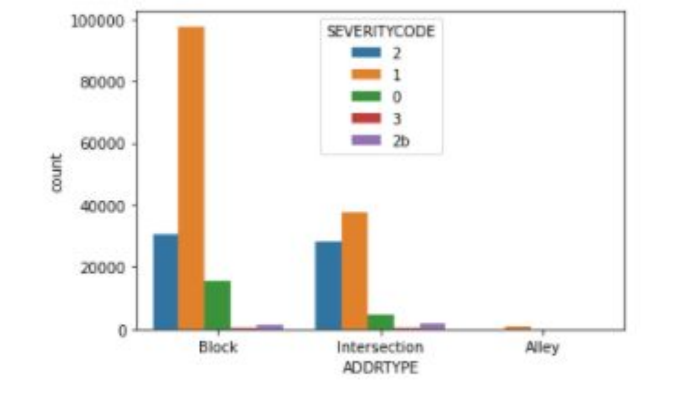
Various exploration analysis were performed between variables that are correlated to each other. Also to understand characteristics of other variables such as impact of Weather, Road condition, Light condition etc.,

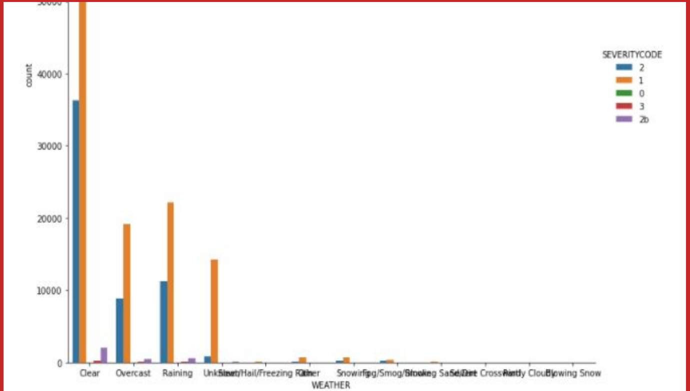
Observations:

Seems severity of accident is more as number of persons involved in accident is more

When person counts is in the range of 2-4 then it seems more accidents are happening. Largely when 2 or 3 passengers are traveling

Exploratory analysis





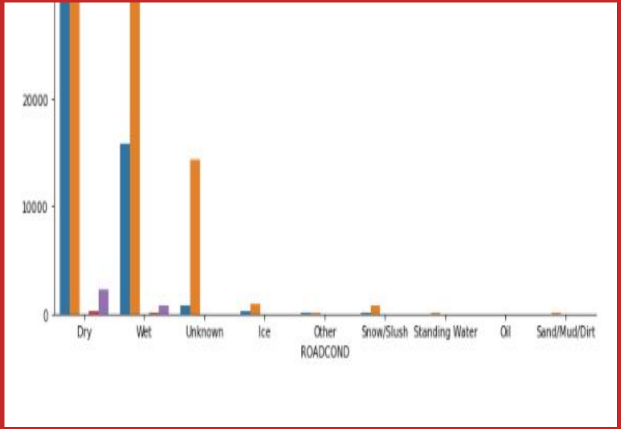
Observations:

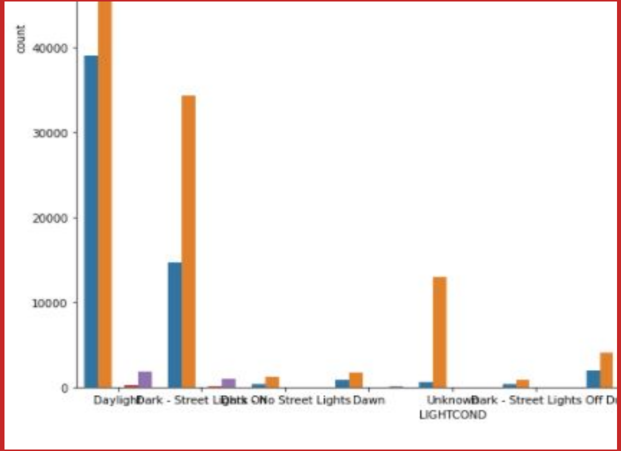
Most accidents are happening in BLOCK, less at intersection. Severity 2 is same on Block and on Intersection

Cycles and Pedestrian seems to be involved in Severity 2 accidents

Most accidents happening when weather is CLEAR, with slightly less accidents when weather is Overcast and Raining

Exploratory analysis





Observations:

Most accidents happening when Road condition is DRY, with slightly less accidents when weather is wet and unknown

For about 9300 car accidents, speeding seems to be the problem

Data Modeling Context:

Post initial exploration and data cleaning exercise with the updated datasets, we shall continue to construct training and test datasets and try out with various modeling experiments.

Next Steps:

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.

Model Experimentation:

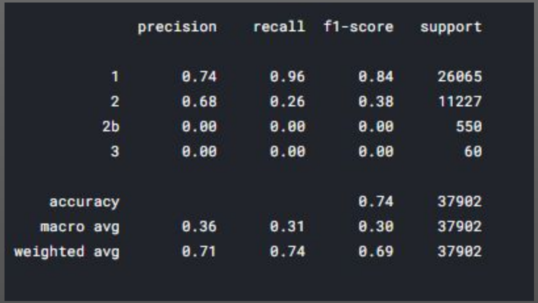
Following models were tried out

1) Decision Tree Classifier

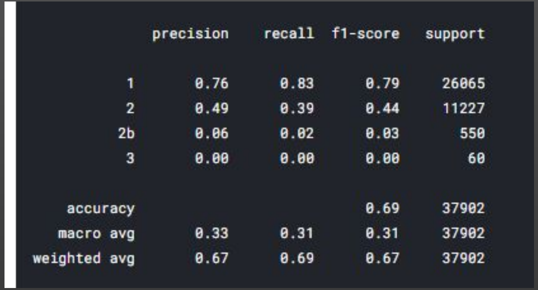
2) Random Forest Classifier

3) Logistic Regression Classifier

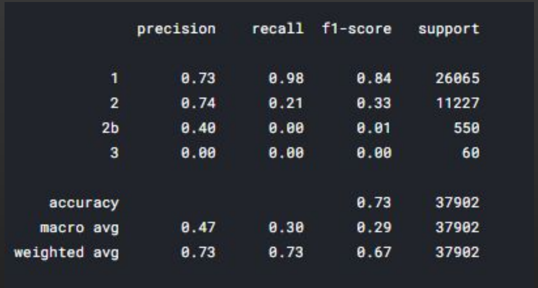
Precision, Recall and scores of the models are mentioned on the right side respectively.



```
              precision    recall  f1-score   support

           1       0.74      0.96      0.84     26065
           2       0.68      0.26      0.38     11227
          2b       0.00      0.00      0.00       550
           3       0.00      0.00      0.00        60

    accuracy                           0.74     37902
   macro avg       0.36      0.31      0.30     37902
weighted avg       0.71      0.74      0.69     37902
```



```
              precision    recall  f1-score   support

           1       0.76      0.83      0.79     26065
           2       0.49      0.39      0.44     11227
          2b       0.06      0.02      0.03       550
           3       0.00      0.00      0.00        60

    accuracy                           0.69     37902
   macro avg       0.33      0.31      0.31     37902
weighted avg       0.67      0.69      0.67     37902
```



```
              precision    recall  f1-score   support

           1       0.73      0.98      0.84     26065
           2       0.74      0.21      0.33     11227
          2b       0.40      0.00      0.01       550
           3       0.00      0.00      0.00        60

    accuracy                           0.73     37902
   macro avg       0.47      0.30      0.29     37902
weighted avg       0.73      0.73      0.67     37902
```

Conclusion:

As you can see all 3 models tried out had a prediction accuracy of 73%-76%,

This means 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.

If we continue to perform minor tweaking and take additional fields in the dataset - accuracy of these models will be greatly improved.