Car accident severity

Business Problem:

For this project, I have chosen the data for all collisions and crashes for the city of Seattle provided by the Open data program. The data is available at the Seattle Open GeoData Portal.

The objective of the project is to understand and explore the data, to extract the important variables which would help us predict the severity of the accidents in future. This would enable the Department of Transportation to prioritise their SOPs and channel their energy to ensure that fewer fatalities result in automobile collisions.

Data:

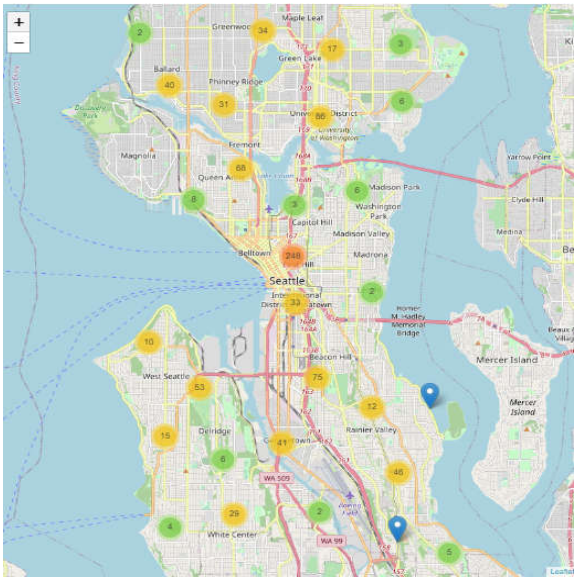
The dataset is available as

- comma-separated values (CSV) files,
- KML files, and
- ESRI

The data is also available from RESTful API services in formats such as GeoJSON.The dataset contains 221,389 records and 40 fields.The metadata of the dataset is found in the site Seattle Department of Transportation.

The data contains several categorical fields and corresponding descriptions. 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.



The **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. Blanks indicate cases where the vehicle was not speeding.

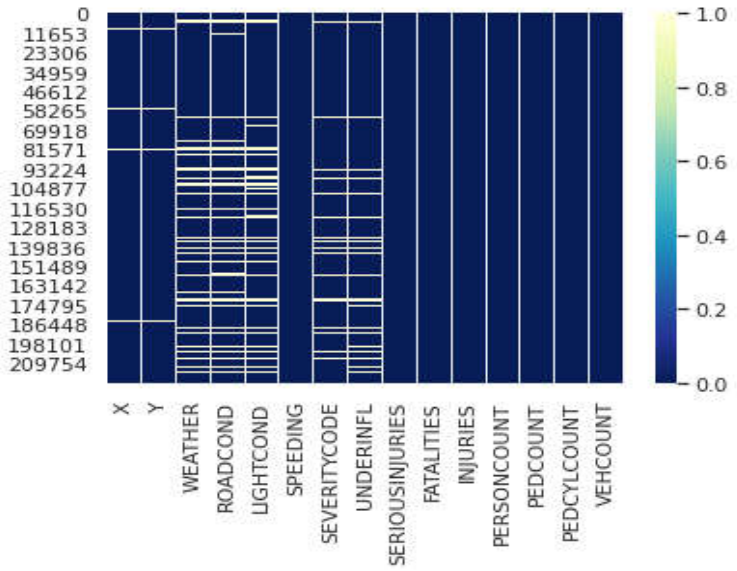
The **SEVERITYCODE** field contains a code that corresponds to the severity of the collision. and **SEVERITYDESC** contains a detailed description of the severity of the collision. We can conclude that there were 349 collisions that resulted in at least one fatality, and 3,102 collisions that resulted in serious injuries. The following table lists the meaning of each of the codes used in the **SEVERITYCODE** field:

SEVERITYCODE Value	Description
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

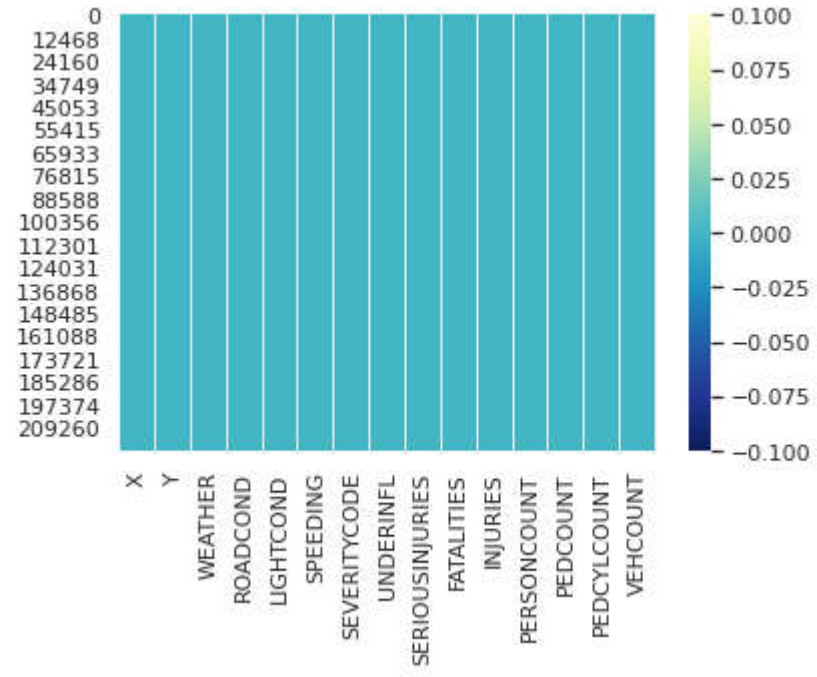
The **UNDERINFL** field describes whether a driver involved was under the influence of drugs or alcohol. The values **0** and **N** denote that the driver was not under any influence while **1** and **Y** that they were. The **PERSONCOUNT** and **VEHCOUNT** indicate how many people and vehicles were involved in a collision respectively.

As the dataset has possibly been sourced from a database table, 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 many 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.

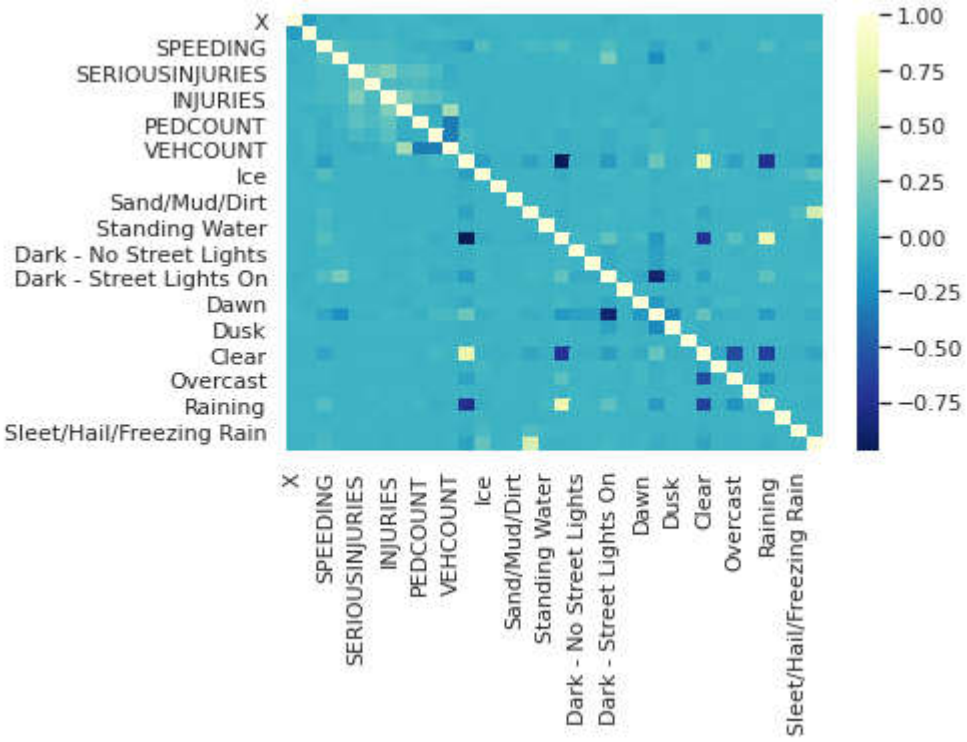
The data set before pre-processing can be visualized as:



After dropping irrelevant columns and null values and performing data cleaning, we got a dataset with 171,380 rows. The data set after removing the irrelevant columns look like this:



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



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.

Modelling

Creating Decision Tree

Decision Tree makes decision with tree-like model. It splits the sample into two or more homogenous sets 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).

Information gain for a decision tree classifier can be calculated either using the Gini Index measure or the Entropy measure, whichever gives a greater gain. A hyper parameter Decision Tree Classifier was used to decide which tree to use, DTC using entropy had greater information gain; hence it was used for this classification problem.

	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

Random Forest

Random Forest Classifier 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. Used for both classification and regression.

Similar to DTC, RFC 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. A hyperparameter was used to determine the best choices for the above mentioned parameters. RFC 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

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 LogRegClassifier

Neural Network

Neural networks can be used to capture non-linearity between features. We have used a Sequential ANN where there are 4 hidden layers. The **relu** and **sigmoid** activation functions are used. The loss function that is used is **categorical_crossentropy** as the target is integer-coded.

	precision	recall	f1-score	support
0	1.00	1.00	1.00	22504
1	1.00	1.00	1.00	11068
2	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 ANN

Conclusion

Initially, the classifiers had an prediction accuracy of 66%-71%, however, upon going back to the data preparation phase, minor tweaking and taking additional fields in the dataset improved the overall accuracy of all models.

The accuracy of the classifiers is excellent, i.e. 100%. 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.