



IBM Data Science Professional Certificate

Applied Data Science Capstone

Prediction of the Car Accident Severity in Seattle City, Washington

Emre Can Okten

September 15th, 2020

1. Introduction

There are various factors contributing to the occurrence of a car accident in the highways. The main objective of this project is to create a machine learning model to predict the possibility of the car accident and how severe it would be according to given features such as weather and road conditions, so that people drive more carefully or even change their travel if they are able to.

This model can be utilized to warn people to be especially mindful when travelling and to notify relevant authorities to be prepared for an accident with a certain severity level in given conditions and take immediate action to mitigate the consequences of it.

2. Data

The required data for this project is acquired from Seattle Open Data Portal [1]. It includes all types of collisions in the city from 2004 to present with the label of accident severity which defines the fatality of an accident. Apart from the label column, the dataset consists of 37 total attributes providing various types of information about location, date/time, weather and light conditions.

Metadata form from the website provide a good description for all attributes. After investigating all the attributes in the dataset and performing the exploratory data analysis, it is concluded that initially selected attributes are the inattention contribution, the weather condition, the road condition, and the light condition of a car accident to train the machine learning model. Remaining attributes in the dataset will be removed to reduce the processing and memory load in the modelling.

[1] https://data-seattlecitygis.opendata.arcgis.com/datasets/5b5c745e0f1f48e7a53acec63a0022ab_0

3. Methodology

In this project, python has been used as a programming language to deal with basic data frame and array operations, fundamental data manipulation techniques, exploratory data visualizations, model development and model evaluation. The required libraries used in the analysis are following:

1. pandas, numpy for basic operations and data manipulation
2. seaborn, matplotlib for visualization
3. sklearn, xgboost for model development and evaluation

3.1 Exploratory data analysis

In order to have a deeper understanding of the dataset, fundamental data exploration techniques will be performed by means of shape, info, dtypes and describe methods in pandas library.

```

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 194673 entries, 0 to 194672
Data columns (total 38 columns):
SEVERITYCODE      194673 non-null int64
X                 189339 non-null float64
Y                 189339 non-null float64
OBJECTID          194673 non-null int64
INCKEY            194673 non-null int64
COLDETKEY         194673 non-null int64
REPORTNO          194673 non-null object
STATUS            194673 non-null object
ADDRTYPE          192747 non-null object
INTKEY            65070 non-null float64
LOCATION            191996 non-null object
EXCEPTRSNCODE   84811 non-null object
EXCEPTRSNDESC   5638 non-null object
SEVERITYCODE.1    194673 non-null int64
SEVERITYDESC       194673 non-null object
COLLISIONTYPE     189769 non-null object
PERSONCOUNT      194673 non-null int64
PEDCOUNT          194673 non-null int64
PEDCYLCOUNT        194673 non-null int64
VEHCOUNT           194673 non-null int64
INCDATE            194673 non-null object
INCDTTM            194673 non-null object
JUNCTIONTYPE      188344 non-null object
SDOT_COLCODE       194673 non-null int64
SDOT_COLDESC       194673 non-null object
INATTENTIONIND     29805 non-null object
UNDERINFL          189789 non-null object
WEATHER            189592 non-null object
ROADCOND           189661 non-null object
LIGHTCOND          189503 non-null object
PEDROWNOTGRNT      4667 non-null object
SDOTCOLNUM         114936 non-null float64
SPEEDING           9333 non-null object
ST_COLCODE         194655 non-null object
ST_COLDESC         189769 non-null object
SEGLANEKEY         194673 non-null int64
CROSSWALKKEY       194673 non-null int64
HITPARKEDCAR       194673 non-null object
dtypes: float64(4), int64(12), object(22)
memory usage: 56.4+ MB

```

Figure 1 Output of info() method

As the output of info method, it can be seen that dataset has 194.672 entries excluding headers and 38 columns in total. Since all 38 columns will not be used in the machine learning model, it must be decided which attributes to keep and which ones to remove before proceeding into data manipulation phase to reduce memory/data load from the system and have a cleaner dataset.

3.2 Data Wrangling

After investigating the attribute information in the metadata of the dataset and looking at the content in the cells it can be confidently said that the following attributes will not serve the goal of the project and will not be used in the feature set of our machine learning model since they include additional information such as registration information and ID codes.

'OBJECTID', 'INCKEY', 'COLDETKEY', 'REPORTNO', 'INTKEY', 'LOCATION', 'EXCEPTRSNCODE', 'STATUS', 'PEDROWNOTGRNT', 'EXCEPTRSNDESC', 'SEVERITYCODE. 1', 'SEVERITYDESC',

'INCDTTM', 'ST_COLCODE', 'HITPARKEDCAR', 'SDOT_COLCODE', 'SDOT_COLDESC', 'SDOTCOLNUM', 'ST_COLDESC', 'SEGLANEKEY', 'X', 'Y', 'CROSSWALKKEY'.

The new shape of the dataset is (194673, 15). Remaining attributes will serve as relevant information for our machine learning model and correlation will be investigated between them and the dependent variable.

Since dataset has an accident date attribute, we can convert it into the day of week and see if there is a correlation between the accident severity and the day of the week.

It is clearly seen that some attributes have significant amount of NaN values as well as inconsistent binary categorization labels such as having both 0-1 and N-Y. The missing data should be properly treated, and inconsistent values should be corrected. The steps for dealing with the missing data will be following:

1. Replacing 'Unknown' and 'Other' values with NaN in the attributes WEATHER, ROADCOND, LIGHTCOND, JUNCTIONTYPE.
2. Replacing NaN values with 0 and N in the attributes SPEEDING, INATTENTIONIND, UNDERINFL.
3. Replacing all N values with 0 and all Y values with 1 in the attributes UNDERINFL, SPEEDING, INATTENTIONIND.

Categorical variables that are more likely to be features for the model is investigated by grouping by the target variable. Results for the investigation of the categorical variables are following:



Figure 2 Investigation of ADDRTYPE attribute by SEVERITYCODE

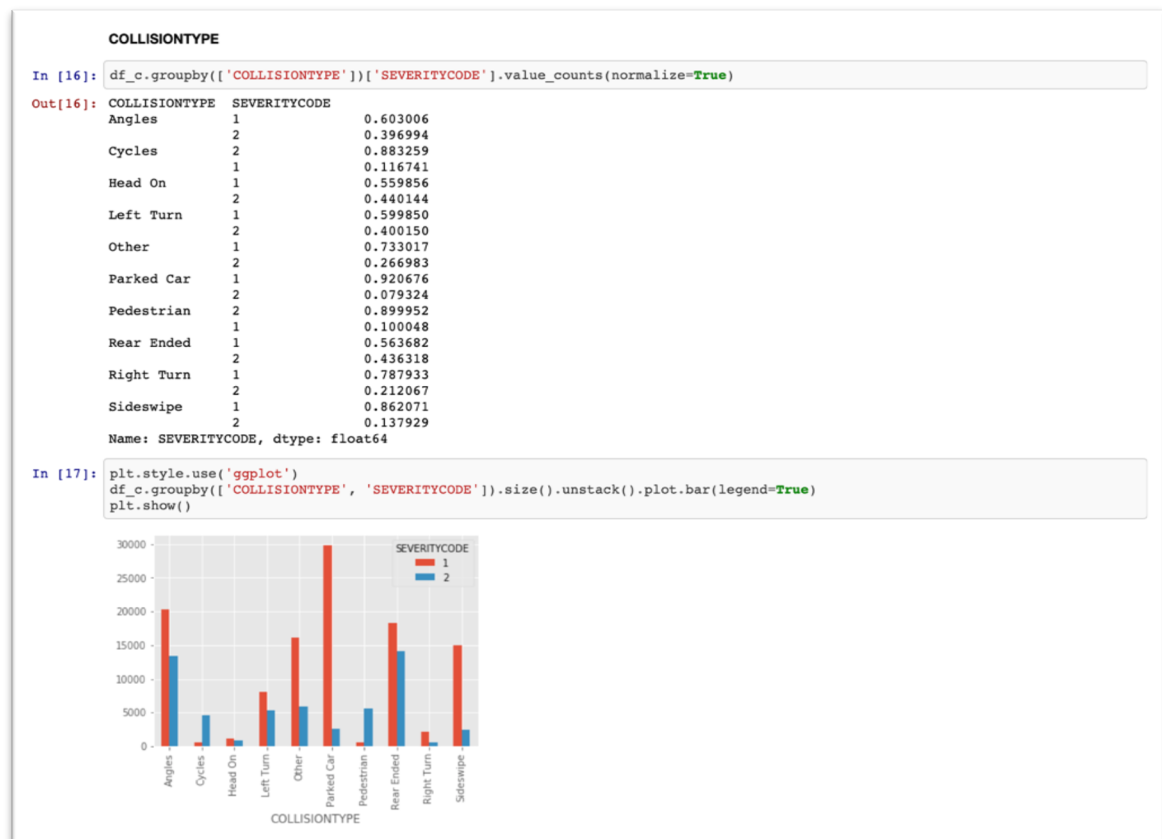


Figure 3 Investigation of COLLISIONTYPE attribute by SEVERITYCODE

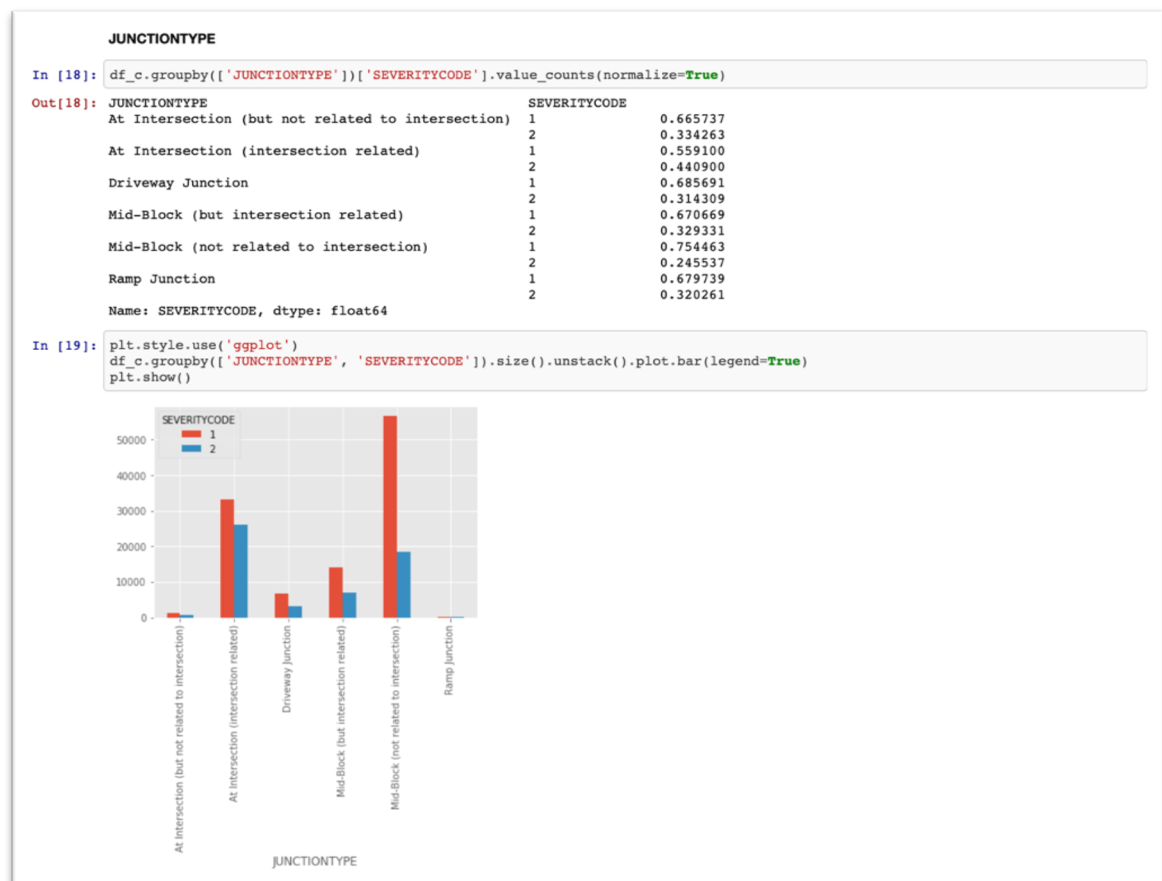


Figure 4 Investigation of JUNCTIONTYPE attribute by SEVERITYCODE

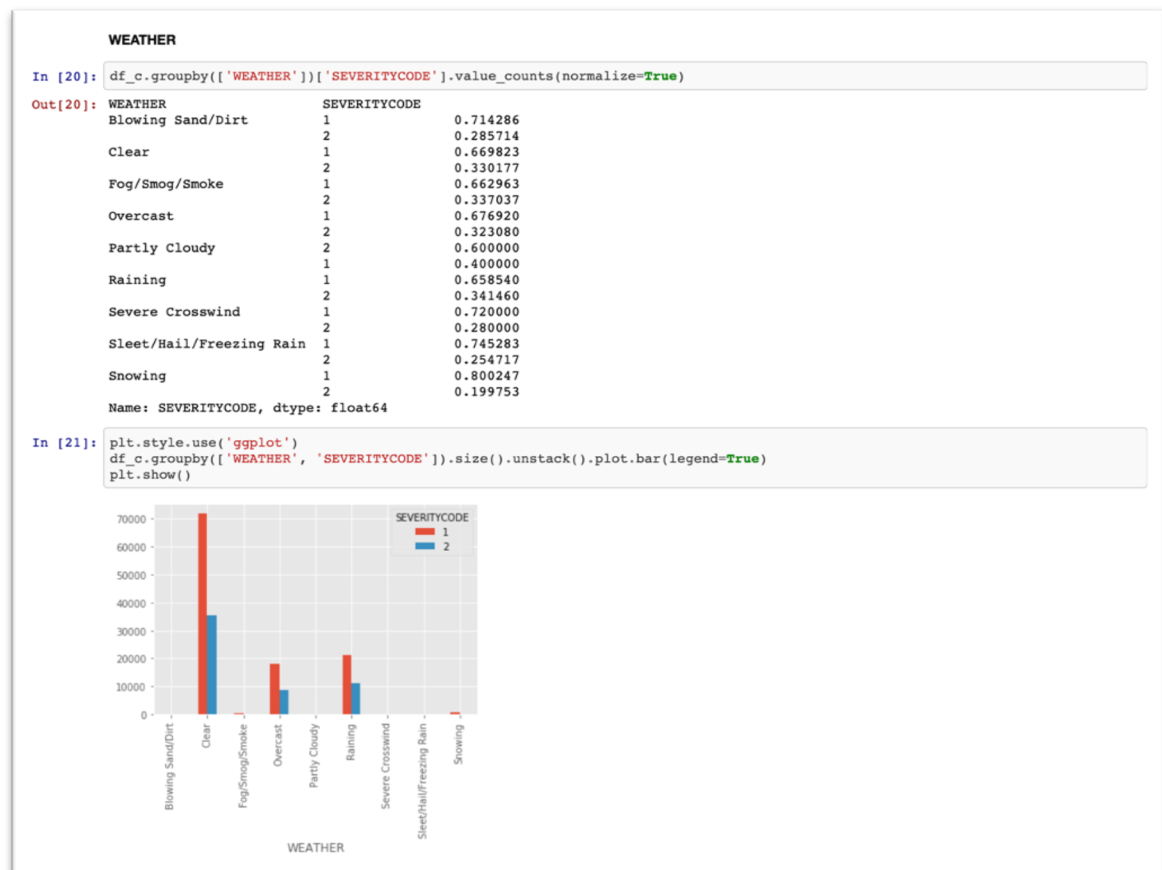


Figure 5 Investigation of WEATHER attribute by SEVERITYCODE

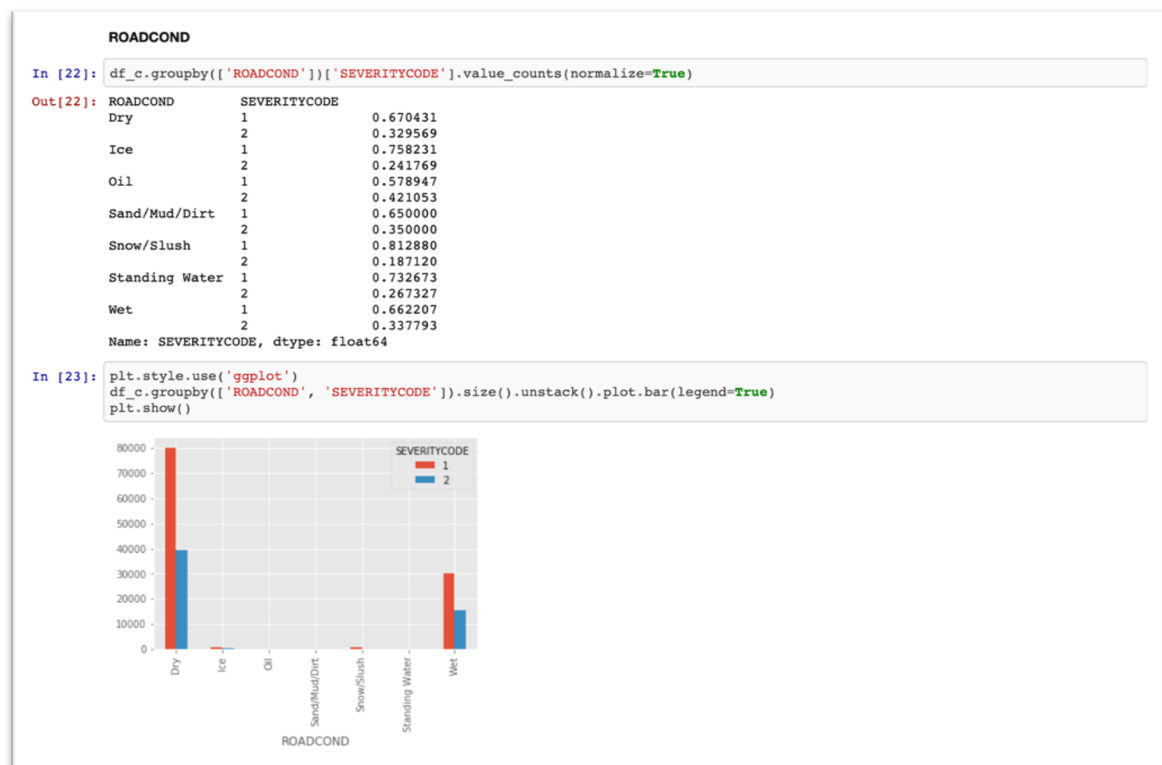


Figure 6 Investigation of ROADCOND attribute by SEVERITYCODE

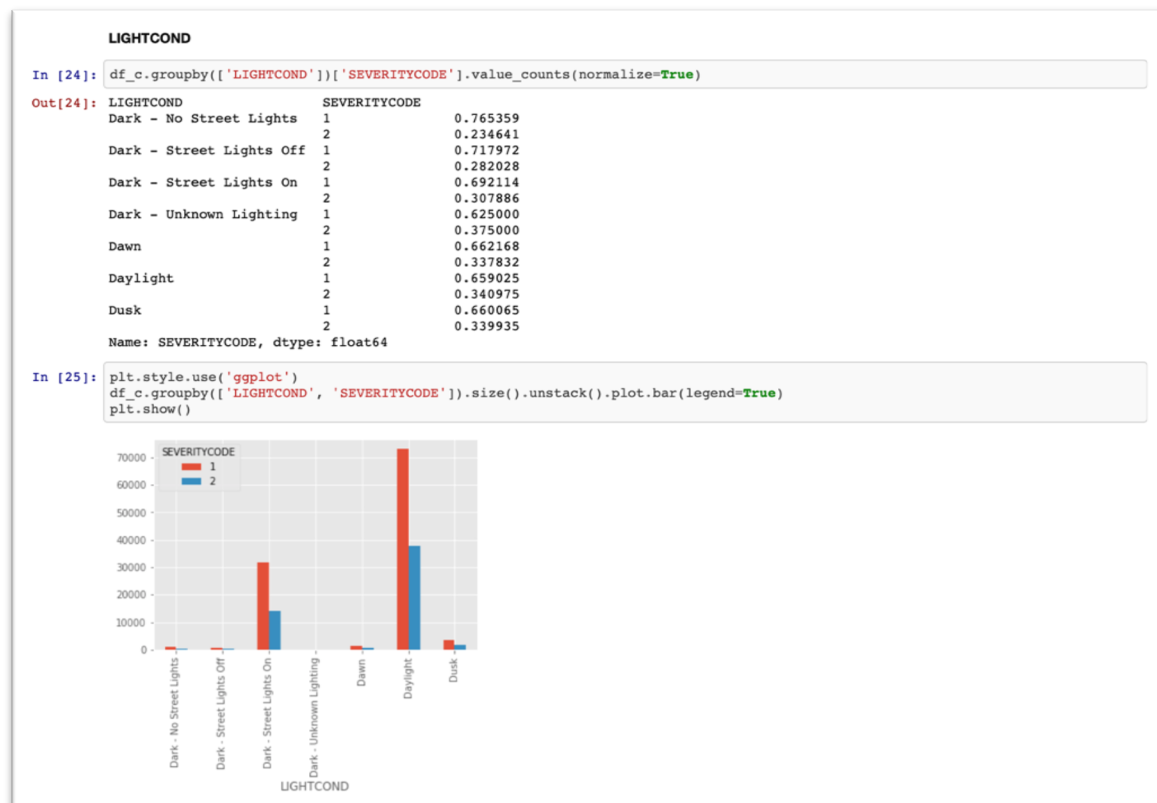


Figure 7 Investigation of LIGHTCOND attribute by SEVERITYCODE

The correlation between the accident severity and all of the categorical variables can be seen in the above figures, so these attributes will be used together with the binary categorical variables as independent variables in our machine learning model.

3.3 Modelling

Since there is a categorical data in the label column, supervised classification algorithm will be used for modelling. The data will be trained first and then the corresponding evaluation metrics will be calculated by means of the predicted values and the correct categorization labels in the test set.

In the feature set selection phase, the continuous variables and binary categorical variables are added into a new Feature data frame.

1. Continuous variables: 'PERSONCOUNT', 'PEDCOUNT', 'PEDCYLCOUNT', 'VEHCOUNT'
2. Binary Categorical variables: 'UNDERINFL', 'SPEEDING', 'DAY'

One-hot encoding technique is used to convert categorical variables to binary variables, appending them to the feature data frame. Since some of the attributes do not include significant amount of data after one-hot encoding, they will be dropped. This operation will remove the cluster from the data and save time in the modelling phase.

Label set is selected as SEVERITYCODE attribute.

Dataset is normalization before modelling in order to reduce data redundancy and improve data integrity. Then the feature set and label set are split into train and test set since we are using a supervised machine learning algorithm. Used classification methods are given below:

1. Decision Tree
2. XGBoost
3. Logistic Regression

4. Results

Below we can see the evaluation metrics for three different classification methods.

Method	Jaccard Index	F1-score	LogLoss
Decision Tree	0.7221	0.6535	NA
XGBoost	0.7334	0.6928	NA
Logistic Regression	0.7322	0.6915	0.5208

5. Discussion

Considering all the calculated evaluation metrics, we can say that the most accurate classification machine learning model to predict the severity of an accident in Seattle is XGBoost with a 74.34% accuracy with Jaccard Index.

Since we have a very close accuracy for Logistic Regression, which is used to calculate the probability of the binary classification, we also have the chance of guessing the probability of an accident severity with a 73.22% accuracy.

F1-score has been calculated to utilize the advantage of confusion matrix, which is that it shows the model's ability to correctly predict or separate the classes. Although it has a decent precision and recall scores, its weighted average is too low compared to Jaccard Index.

6. Conclusion

The main goal of this project was to analyse the Seattle car accident dataset and conclude a prediction model based on the given criteria such as day of the week, weather, road and light condition and different locations so that the travellers in the city has an idea about which factors and conditions to pay attention to before travelling.

At the end of the modelling part, we concluded with 3 different classification models all of which can predict the severity of a car accident with almost 73% accuracy. By the help of these models, drivers in Seattle will have the opportunity of paying more attention to given conditions and reduce both the frequency and the severity of a car accident.