Shahd AlShamsi
Applied Data Science Capstone
IBM / Coursera
October 2020

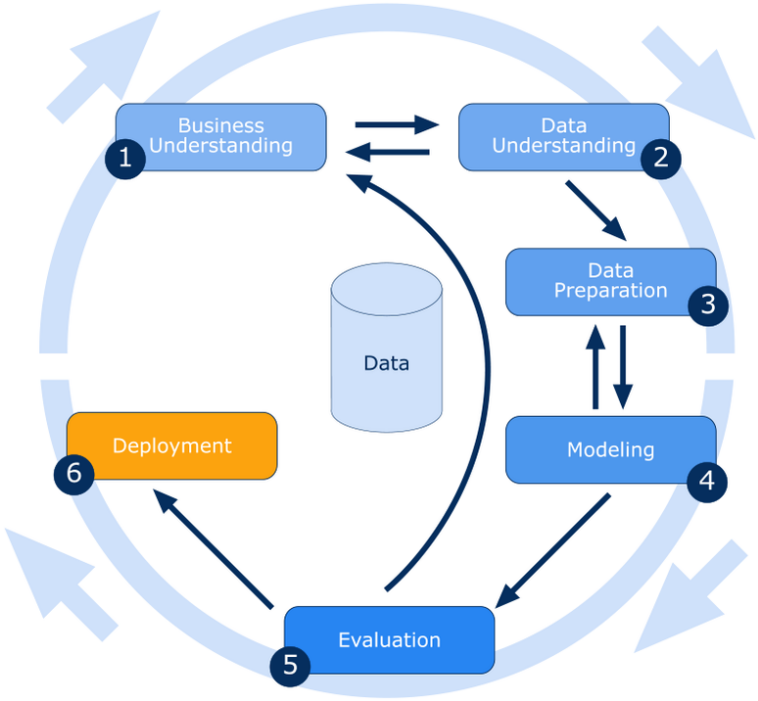
Car Accident Severity

Introduction

The following Applied Data Science Capstone project defines the last step required to receive the IBM Data Science Professional Certificate. It will allow us to apply all the concepts learned in the previous 8 courses and utilise the relevant tools to solve the problem at hand. The 8 courses completed prior to attempting this project are listed in order:

1. What is Data Science?
2. Tools for Data Science
3. Data Science Methodology
4. Python for Data Science and AI
5. Databases and SQL for Data Science
6. Data Analysis with Python
7. Data Visualisation with Python
8. Machine learning with Python

The purpose of this document is to report back on all of the steps taken to solve the project, which will follow the CRISP-DM framework shown below:



CRISP-DM Framework

Business Understanding

In an aim to minimise the uncertainty around car accidents in Seattle city, the purpose of this project is to develop a warning system solution that will warn Seattle residents of the potential severity of a getting into a car accident, based on several critical factors. In particular, the project aims to answer the following 6 questions:

1. What is the impact of light conditions on car accident severity?
2. What is the impact of weather on car accident severity?
3. What is the impact of road conditions on car accident severity?
4. What is the impact of junction type on car accident severity?
5. What is the impact of collision type on car accident severity?
6. What is the impact of speeding on car accident severity?

The insights gained will hopefully increase awareness within the community around car travel safety and could even cause the user to change his/her travel journey if possible. The final solution will be presented to the Seattle Department of Transportation and Seattle Police (target audience), with the end-goal of potentially developing a mobile application to help reduce car collisions in the future.

Data Understanding

To tackle this problem, we will be using the shared dataset CSV file provided by the course which can be accessed through the following link: https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv.

The dataset has been collected by the SDOT Traffic Management Division, Traffic Records Group. It captures all car collisions in Seattle from 2004 to present. It is also worthy to note that the dataset is updated on a weekly basis. The following provides a snapshot of the dataset:

	SEVERITYCODE	X	Y	OBJECTID	INCKEY	COLDETKEY	REPORTNO	STATUS	ADDRTYPE	INTKEY	...	ROADCOND	LIGHTCOND
0	2	-122.323148	47.703140	1	1307	1307	3502005	Matched	Intersection	37475.0	...	Wet	Daylight
1	1	-122.347294	47.647172	2	52200	52200	2607959	Matched	Block	NaN	...	Wet	Dark - Street Lights On
2	1	-122.334540	47.607871	3	26700	26700	1482393	Matched	Block	NaN	...	Dry	Daylight
3	1	-122.334803	47.604803	4	1144	1144	3503937	Matched	Block	NaN	...	Dry	Daylight
4	2	-122.306426	47.545739	5	17700	17700	1807429	Matched	Intersection	34387.0	...	Wet	Daylight

Dataset Snapshot

The dataset includes a total of **194,673** observations and **37** attributes. Since we will be looking at the car collision severity, our target/dependent variable will be the SEVERITYCODE:

Target: SEVERITYCODE

Data type: int64

Possible Values:

1 to indicate Property Damage Only Collision

2 to indicate Injury Collision

The following shows all the attributes available in the dataset:

```
Index(['SEVERITYCODE', 'X', 'Y', 'OBJECTID', 'INCKEY', 'COLDETKEY', 'REPORTNO',
       'STATUS', 'ADDRTYPE', 'INTKEY', 'LOCATION', 'EXCEPTRSNCODE',
       'EXCEPTRSNDESC', 'SEVERITYCODE.1', 'SEVERITYDESC', 'COLLISIONTYPE',
       'PERSONCOUNT', 'PEDCOUNT', 'PEDCYLCOUNT', 'VEHCOUNT', 'INCDATE',
       'INCDTTM', 'JUNCTIONTYPE', 'SDOT_COLCODE', 'SDOT_COLDESC',
       'INATTENTIONIND', 'UNDERINFL', 'WEATHER', 'ROADCOND', 'LIGHTCOND',
       'PEDROWNOTGRNT', 'SDOTCOLNUM', 'SPEEDING', 'ST_COLCODE', 'ST_COLDESC',
       'SEGLANEKEY', 'CROSSWALKKEY', 'HITPARKEDCAR'],
```

Since our target is to *predict* car accident severity with the above *labelled* data, we will be utilising **supervised machine learning algorithms** to solve our problem. As the dataset is quite large, we would need to scope down and focus only on the independent variables of interest. In reference to the problem questions discussed earlier, we will look at INCDTTM, WEATHER, ROADCOND, JUNCTIONTYPE, COLLISIONTYPE, and SPEEDING which are all object data types.

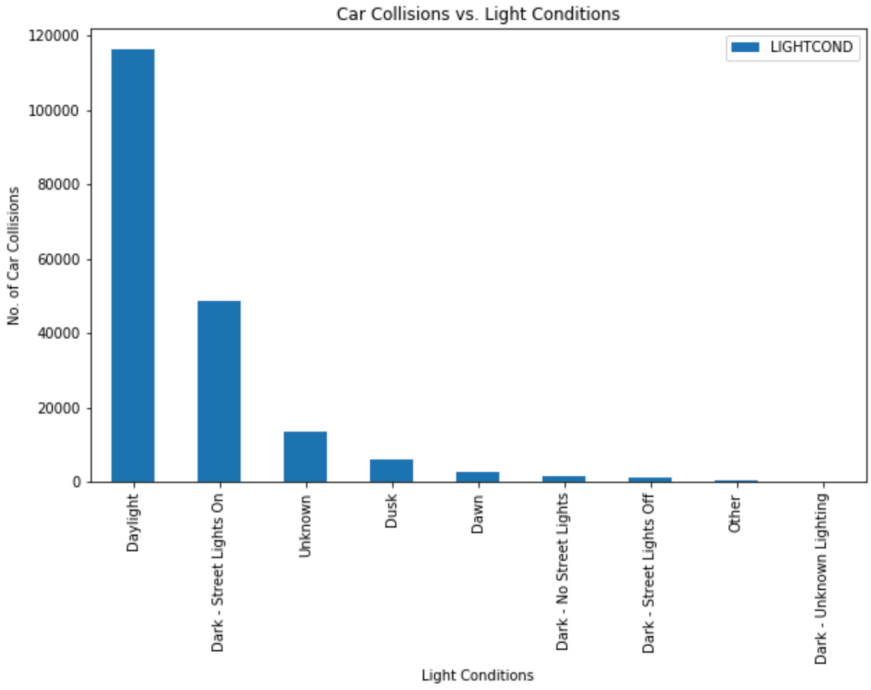
1. LIGHTCOND: The light conditions during the collision
2. WEATHER: A description of the weather conditions during the time of the collision
3. ROADCOND: The condition of the road during the collision
4. JUNCTIONTYPE: Category of the junction at which the collision took place
5. COLLISIONTYPE: Type of car collision
6. SPEEDING: Whether or not speeding was a factor in the car collision (Y/N)

	LIGHTCOND	WEATHER	ROADCOND	JUNCTIONTYPE	COLLISIONTYPE	SPEEDING
count	189503	189592	189661	188344	189769	9333
unique	9	11	9	7	10	1
top	Daylight	Clear	Dry	Mid-Block (not related to intersection)	Parked Car	Y
freq	116137	111135	124510	89800	47987	9333

Some descriptive statistics in relation to the 6 attributes are highlighted below:

The above summary demonstrates the imbalance in the dataset, thus it will need to be cleaned, formatted and standardised to avoid any bias in our analysis and modelling. After dropping the missing values under the columns of interest, we will look at the observations with the highest car collision count for each attribute.

Light Conditions:



The above bar chart indicates that out of the observations that reported light conditions, the highest occurrence **during the daytime (Daylight).**

Weather:

The above bar chart indicates that out of the observations that reported weather conditions, the highest occurrence was on a **clear day**.

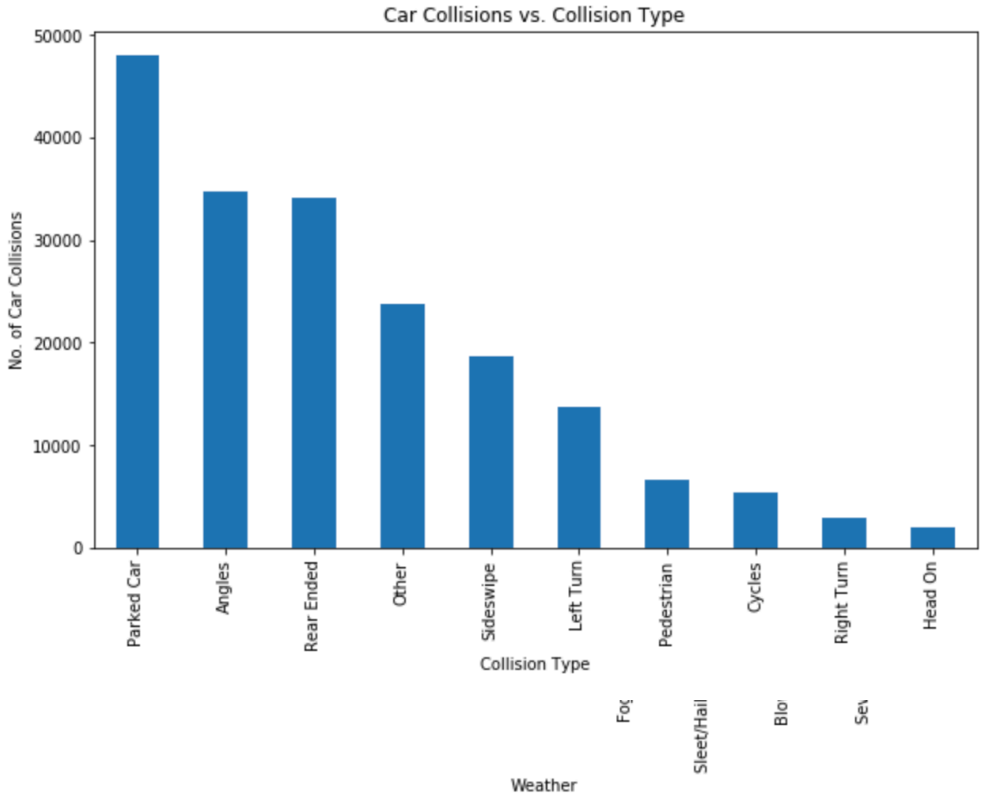
Road Conditions:

The above bar chart indicates that out of the observations that reported road conditions, the highest occurrence was on a **dry road**.

Junction Type:

The above bar chart indicates that out of the observations that reported junction type, the highest occurrence was on a **Mid-block (not related to intersection)**.

Collision Type:



The above bar chart indicates that out of the observations that reported collision type, the highest occurrence was with a **Parked Car**.

For speeding, 9333 accidents were associated with a speeding factor; however, only 9333 observations were reported so there's not much information gain associated with this attribute. As such we will drop the speeding factor from further analysis.

Methodology

The project data analysis and modelling was done with a Jupyter Notebook. Following the preliminary data analysis that was referenced in the previous section, it was evident that we needed to do some encoding for the categorical variables of interest (note that similar categories were grouped to represent the same number):

Encoding

LIGHTCOND:

Daylight = 0 Dark - Street Lights On = 1 Dark - No Street Lights = 2 Dusk = 1 Dawn = 1 Dark - Street Lights Off = 2 Dark - Unknown Lighting = 2 Other = 999 Unknown = 999

WEATHER:

Clear = 0 Raining = 3 Overcast = 1 Other = 999 Unknown = 999 Snowing = 3 Fog/Smog/ Smoke = 2 Sleet/Hail/Freezing Rain = 3 Blowing Sand/Dirt = 2 Severe Crosswind = 2 Partly Cloudy = 1

ROADCOND:

Dry = 0 Wet = 2 Ice = 2 Snow/Slush = 1 Other = 999 Unknown = 999 Standing Water = 2 Sand/Mud/Dirt = 1 Oil = 2

JUNCTIONTYPE:

Mid-Block (not related to intersection) = 10 Mid-Block (but intersection related) = 11 At Intersection (but not related to intersection) = 20 At Intersection (intersection related)= 21 Driveway Junction = 3 Ramp Junction = 4 Unknown = 999

COLLISIONTYPE:

Parked Car = 0 Angles = 1 Rear Ended = 2 Sideswipe = 3 Left Turn = 4 Pedestrian = 5 Cycles = 6 Right Turn = 7 Head On = 8 Other = 999

Prior to moving the next step, we must normalise the data with X as the matrix containing the attribute data and y as the response vector. We also set up the Train/Test split.

```
X=df[[“LIGHTCOND","WEATHER","ROADCOND","JUNCTIONTYPE","COLLISIONTYPE"]].values

X[0:5]

y = df['SEVERITYCODE'].values

y[0:5]
```

The next stage consists of utilising the machine learning algorithms we’ve learned. Since our data is labelled, we will only be looking at supervised learning algorithms.

Results

Decision Tree

After setting up the train/test data, we fit the training data into a severityTree. We made some predictions on the test data and calculated the accuracy classification score of our model which was:

```
DecisionTrees's Accuracy:  0.752459226916656
```

K-Nearest Neighbours

Starting with k=2, we trained the model using the training data and used it to predict the severity code of test data. The following accuracy scores were reported:

```
Train set Accuracy:  0.7324673490092335
Test set Accuracy:  0.7345062283292667
```

k=3

```
Train set Accuracy:  0.7114448625255236
Test set Accuracy:  0.7114421471683575
```

k=4

```
Train set Accuracy:  0.7351577649642348
Test set Accuracy:  0.7368691408758187
```

k=5 resulted in a lower accuracy for both the train and test sets so following discussions will reference the k=4 model.

Logistic Regression

Predicting the car accident severity code with logistic regression resulted in the following:

Jaccard Similarity Score: 0.7043790933607295

Log Loss= 0.5754580122708407

Discussion

For discussion and recommendation purposes, we will refer to the Decision Tree predictive model as it yielded the highest accuracy score for predicting the car accident severity code given light condition, weather, road condition, junction type and collision type. As indicated before, we have dropped the speeding factor as a predictor of car accident severity as it doesn't provide much valuable information considering only 9333 YES observations were reported, which appx. 4.8% of the dataset.

In terms of correlation, the highest correlation was between road conditions and severity code. Indeed, correlation is not causation but this does give us a good predictor. The observations with the highest car accident frequencies (regardless of severity) were Daylight, on a Clear Day, Dry Road, which indicate that seemingly normal conditions do not guarantee a safe trip.

Indeed our observation and discussion is limited by how we scoped our problem by first selecting the independent variables of interest as opposed to conducting the analysis on all 38 attributes and dropping them as we move further with the analysis. We were also limited to two values for severity codes.

Conclusion

We recommend the solution of developing a Decision Tree Classifier system as a mobile application that can be used by drivers prior to their trip, which can predict that probability of getting into a car accident and its severity. They could select the input for different significant variables via a list of standardised options.