Seattle Car Accident Severity Prediction Model

Applied Data Science Capstone Project

1. Introduction

1.1 Background

Seattle is a major city in the North West of the USA with headquarters of some major companies of the world like Microsoft, Boeing, Amazon, Costco, etc. The city also serves as a transit hub for people travelers and tourists moving to and from Canada. The city boasts a high concentration of technology workers, with median income more than most of the metro cities in the US. The city's real estate and housing stats are very high as well. With a population of three quarters of a million and the area of 217 Sq km, a higher standard of services is expected by the patrons of the city. The city does its best to improve services, especially by employing AI/ML based solutions. One such opportunity is the emergency management system to address traffic accidents. This project throws some light on to solving one problem, namely, trying to predict the severity of the accident as soon as an accident is reported to the city's emergency management system.

1.2 Problem

The city of Seattle emergency management system (911) would like to deploy a new model to predict the severity of a newly reported accident based on the information received. The city currently has the 911 service which generates the data by receiving a 911 call, which results in dispatching the police and ambulance to the accident site. The city would like have a predictive model to predict the severity of the accident. Based on the severity of the accident, the city administration wants manage its resources like emergency personnel, the traffic management and the trauma centers in the city.

1.3 Interest

Obviously, by saving lives and streamlining the city administration the city will not only save money on its operations, but with improved services, the rating of the city will improve, thereby attracting more businesses and families into the city.

2. Data Acquisition and Cleaning

2.1 Data Source

Seattle Department of Transportation (SDOT) compiles all collisions provided by the Seattle Police Department and recorded by traffic records. This compiled data is reported on their website as a csv file for the general public to consume. This data is refreshed every week with the latest records. The dataset include data from 2004.

2.2 Data Cleaning

Even though SDOT reported data has all types of SEVERITY codes (0,1,2,2b,3), the sample dataset provided on the course website has only two codes, 1(property damage) and 2(injury) only. Variables like SERIOUSINJURIES and FATALITIES that are reported in the metadata file are not reported in the actual data. Some other observation:

- EXCEPTRSNCODE is a matadata
- SEVERITYCODE is a duplicate column
- There are two collision codes, 1) State Code and 2) SDOT code. One is redundant
- Descriptions attributes are redundant. Metadata has detailed information.
- Location variable is redundant in light of X and Y.
- OBJECTID, INCKEY, COLDETKEY, and REPORTNO are index columns and cannot be u sed as attributes.

The shape of the dataset: 194673, 38, whereas the first row is the column names.

Severity Codes in the dataset:

- 1. 1-prop damage
- 2. 2-injury

Due to the presence of null values and the redundant variables as mentioned above, the data has to be preprocessed before going further.

3.Exploratory Data Analysis

3.1 Dummying the Target Variable

The target variable is a binary classification variable with values 1 and 2, where 1 represents property damage and 2 is injury. Since it's a binary variable, the obvious choice is the have a binary classification model. In order to achieve that, the target variable should be converted into a dummy variable, where 0 represents a non-event and 1 represents an event, where an event is an injury. The model will try to predict the occurrence of an injury due to the accident reported.

3.2 Binning and re-classifying the variables

Many of the variables are redundant variables as discussed in the Data section. After eliminating those, the final set of variables are as follows:

#	Variable
1	ADDRTYPE
2	COLLISIONTYPE
3	PERSONCOUNT
4	PEDCOUNT
5	PEDCYLCOUNT
6	VEHCOUNT
7	JUNCTIONTYPE
8	SDOT_COLCODE
9	INATTENTIONIND
10	UNDERINFL
11	WEATHER
12	ROADCOND
13	LIGHTCOND
14	PEDROWNOTGRNT
15	SPEEDING
16	HITPARKEDCAR

After collapsing the categories based on the Information Value statistic, the final binnings are as follows:

COLLISIONTYPE	Category	New Group
	NaN	Other
	Angles, Left Turn	Angles/Left Turn
	Cycles, Pedestrian	Cycles/Pedestrian
	Head On, Rear Ended	Head On/Rear Ended
	Parked Car	Parked Car
	Right Turn	Right Turn
	Sideswipe	Sideswipe

PERSONCOUNT	Category	New Group
	<=2	2
	3 to 5	3 to 5
	>=6	6+

PEDCOUNT	Category	New Group
	0	0
	>0	1+

PEDCYLCOUNT	Category	New Group
	0	0
	>0	1+

VEHCOUNT	Category	New Group
	<=1	1
	2	2
	>=3	3+

JUNCTIONTYPE	Category	New Group
	NaN, Ramp Junction	Unknown
	At Intersection (but not related to intersection)	At Intersection (but not related to intersection)
	At Intersection (intersection related)	At Intersection (intersection related)
	Driveway Junction	Driveway Junction
	Mid-Block (but intersection related)	Mid-Block (but intersection related)
	Mid-Block (not related to intersection)	Mid-Block (not related to intersection)

SDOT_COLCODE	Category	New Group
	11	11
	14	14
	16	16
	14	14
	0, 13	1300
	26,28	2628
	all other	99

INATTENTIONIND	Category	New Group
	Y	Y
	NaN	N

ROADCOND	Category	New Group
	NaN, Unknown, Standing Water, Oil, Sand/Mud/Dirt	Other
	Ice, Snow/Slush	Ice/Snow/Slush
	Dry, Wet	Dry/Wet

LIGHTCOND	Category	New Group
	NaN, Unknown	Other
	Dark - Street Lights On	Dark - Street Lights On

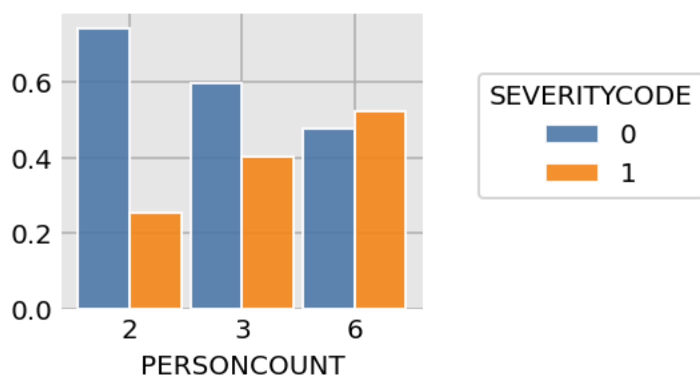
	Dark-No Lights/Off/Other	Dark-No Lights/Off/Other
	Dawn, Desk	Dawn/Desk
	Daylight	Daylight

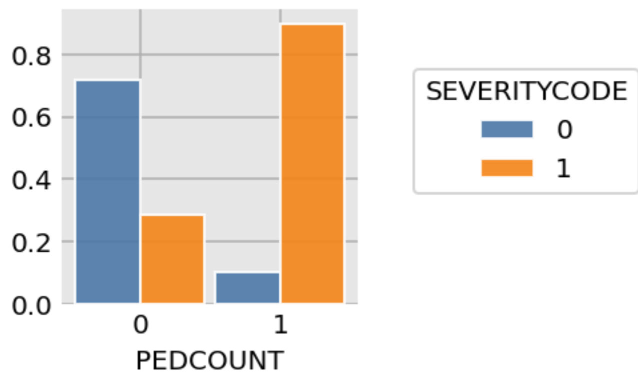
WEATHER	Category	New Group
	Clear	Clear
	Fog/Smog/Smoke	Fog/Smog/Smoke
	Overcast	Overcast
	Raining	Raining
	Snowing	Snowing
	NaN, Other	Snowing

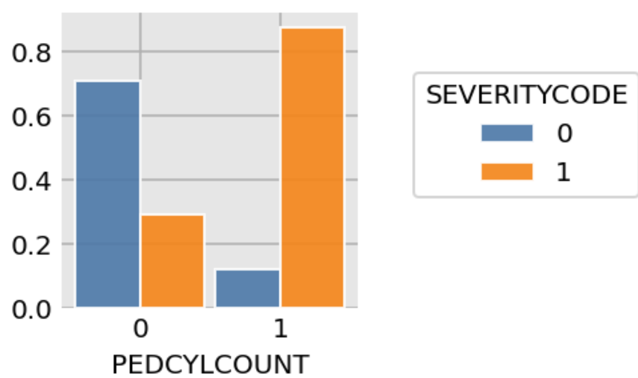
PEDROWNOTGRNT	Category	New Group
	NaN	N
	Y	Y

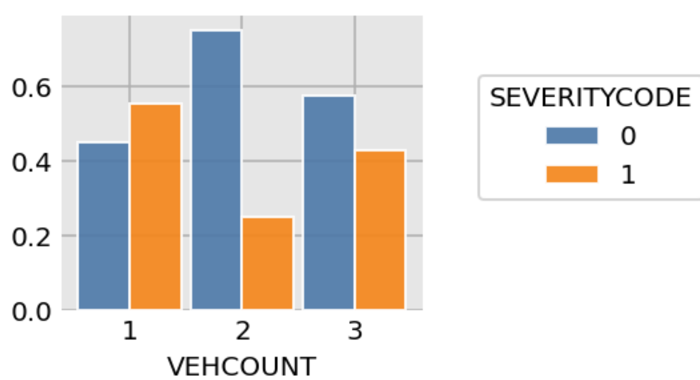
SPEEDING	Category	New Group
	NaN	N
	Y	Y

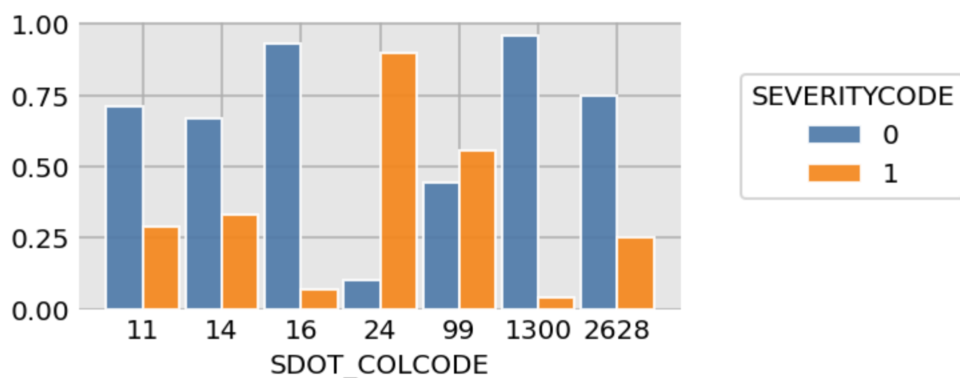
3.3 Weight of evidence visual charts

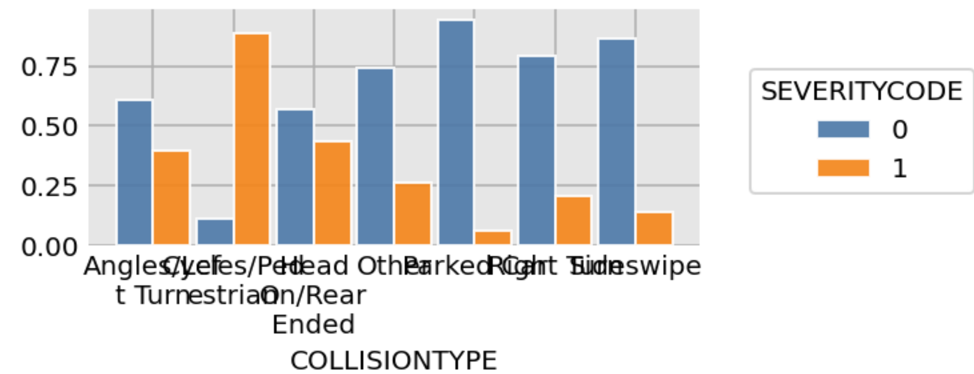



0.8
0.6
0.4
0.2
0.0
0
1
PEDCOUNT
SEVERITYCODE
0
1


0.8
0.6
0.4
0.2
0.0
0
1
PEDCYLCOUNT
SEVERITYCODE
0
1


0.6
0.4
0.2
0.0
1
2
3
VEHCOUNT
SEVERITYCODE
0
1


1.00
0.75
0.50
0.25
0.00
11
14
16
24
99
1300
2628
SDOT_COLCODE
SEVERITYCODE
0
1



3.4 Changing categorical variables into continuous variables

To fit the logistic regression model the features are converted into continuous variables using weight of evidence of the SEVERITYCODE on each of the variable and their classifications. This is the basis of the weights for each of the category. The final category is as follows: Note: only the variables and their final binning is shown here.

COLLISIONTYPE	Category	Value
	Angles/Left Turn	23
	Cycles/Pedestrian	49
	Head On/Rear Ended	25
	Other	17
	Parked Car	0
	Right Turn	14
	Sideswipe	9

PERSONCOUNT	Category	Value
	2	0
	3 to 6	7
	6+	12

PEDCOUNT	Category	Value
	0	0
	1+	31

PEDCYLCOUNT	Category	Value
	0	0
	1+	29

VEHCOUNT	Category	Value

	1	13
	2	0
	3+	8

SDOT_COLCODE	Category	Value
	11	22
	14	24
	16	5
	14	53
	1300	33
	2628	0
	99	20

4. Model Development

Since the target variable is a binary variable, the model employed is Logistic Regression. Logistic regression is widely employed by data scientists and researchers. Some of the added advantages include easy to implement, and interpret.

Logistic regression uses a sigmoid function:

```
p = 1/1+e^-y
```

Properties of Logistic Regression:

- Dependent variable follows Bernoulli Distribution
- Estimation is done through maximum likelihood
- Model fitness is calculated through Concordance, KS-D static

4.1 Building

Model building in Scikit-learn - Here we are going to predict the severity of the accident using logistic regression classifier.

4.2 Testing and Training Data Sets

“from sklearn.cross_validation import train_test_split” The training data will be used to fit the model, and later the model will tested using the testing data. I used 85-15 split of training and testing data.

5. Results

There are three results for this model. They are:

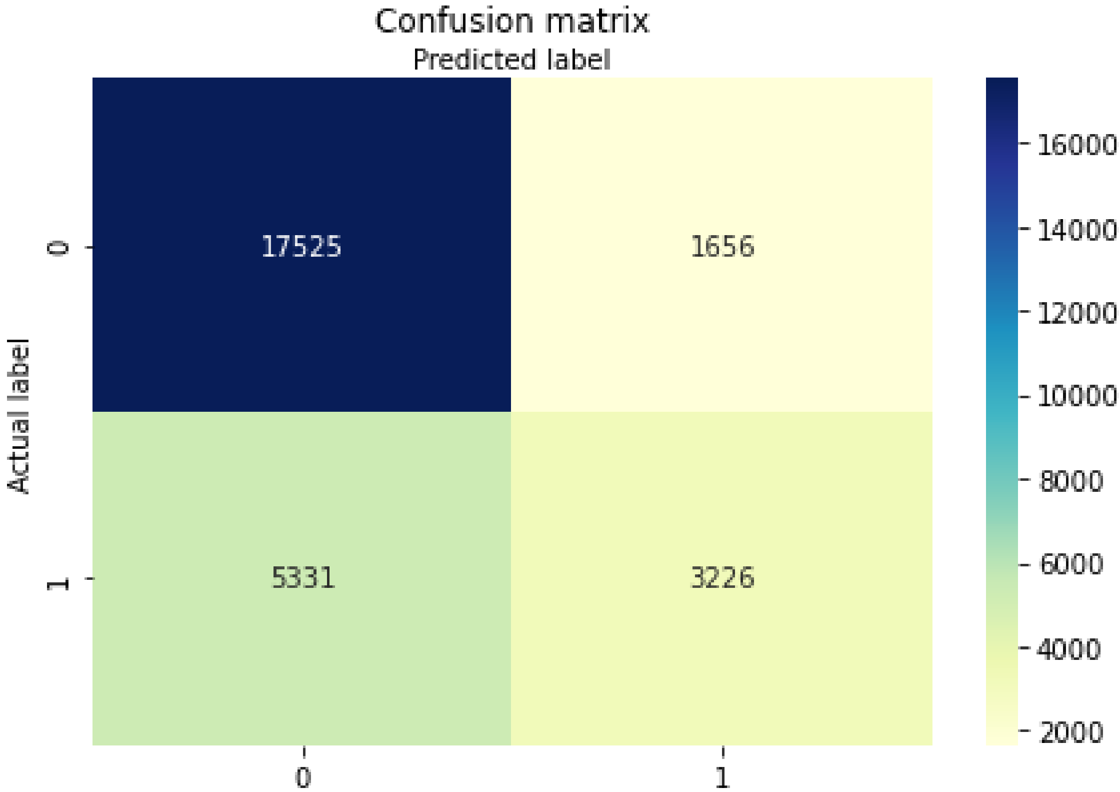
- Maximum likelihood and regression values table
- Confusion Matrix
- ROC curve

5.1 Maximum Likelihood

The Maximum Likelihood converged with function value: .49 The coefficients of all the variables employed are clearly displayed along with their std err and p-values. The Pseudo R-sqe also show the significance of the model

```
Optimization terminated successfully.
         Current function value: 0.489963
         Iterations 7
                           Logit Regression Results
==============================================================================
Dep. Variable:           SEVERITYCODE   No. Observations:               184920
Model:                          Logit   Df Residuals:                   184913
Method:                           MLE   Df Model:                            6
Date:                Fri, 11 Sep 2020   Pseudo R-squ.:                  0.2051
Time:                        06:29:13   Log-Likelihood:                -90604.
converged:                       True   LL-Null:                   -1.1399e+05
Covariance Type:            nonrobust   LLR p-value:                     0.000
==================================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
----------------------------------------------------------------------------------
const         -3.4351      0.030   -116.032      0.000      -3.493      -3.377
COLLISIONTYPE_  0.0865      0.001    107.005      0.000       0.085       0.088
PERSONCOUNT_   0.0907      0.002     58.796      0.000       0.088       0.094
PEDCOUNT_     -0.0075      0.002     -3.971      0.000      -0.011      -0.004
PEDCYLCOUNT_  -0.0047      0.002     -2.655      0.008      -0.008      -0.001
VEHCOUNT_      0.0309      0.001     22.679      0.000       0.028       0.034
SDOT_COLCODE_  0.0279      0.001     21.205      0.000       0.025       0.030
==================================================================================
```

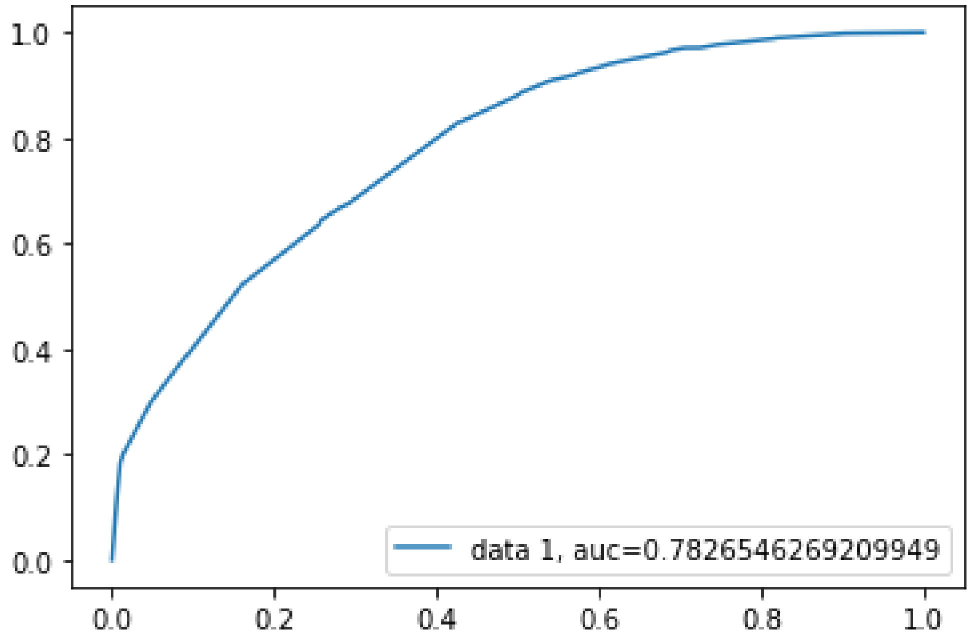
5.2 Confusion Matrix



Confusion Matrix provides Accuracy, Precision and mis-classification of the model. With the accuracy of 74% and precision of 66%, this model can significantly classify the severity.

- Accuracy: 0.7481072896387627
- Precision: 0.6607947562474396
- Recall: 0.3770012854972537

5.3 ROC Curve



The ROC curve shows the AUC at 78% which is significantly higher than 50%.

6. Observations and Discussions

6.1 Variables

This modeling exercise has opened up the world of road safety and its challenges to the world. It gave a window of opportunity into the Seattle PD data. There are some straight forward variables, but some intuitive variables happened to be very insignificant. Example: Under Influence is not a significant variable. Also, rush hour accidents are not injurious.

There are too many variables with missing values, else, the model would perform even better.

6.2 Opportunities

The data can be segmented to create multiple models. There is enough data to support that. Also, some of the verbiage from 2004 is little confusing. Example: Driverless in 2020 is an autonomous vehicle without a driver.

7. Conclusion

The deployment of the model is a completely different ball game. The deployment team should be well versed with the nuances and the treatment of the data, else the results will be highly predictive. Also, the model should be calibrated from time to time, and production team should keep track of the incoming data, and the variations in the variables from time to time. With this, I conclude the project.