ADS_CapstoneProject

September 23, 2020

1 Prediction of car accidents

This notebook is created to be used in Coursera Applied Data Science Capstone Project.

1.1 Introduction/Business Problem

In our time, cars are used for both leisure purposes and transport. By 2019, there were a car per two persons in the most of the developed countries while in some of them it is even more. There are more and more cars on the roads as the demand increases.

With the increase of cars on the roads, the accident probability increases. Prediction of car accidents and their severity is important to decrease deaths, injuries and damage to vehicles. With an accurate predictive tool, people may plan their travels accordingly to avoid accidents. So, building a model to predict probability and severity of an accident with influencing parameters is necessary.

Drivers and passangers are the primary stakeholders that would be interested in knowing accident probability to make their transportation plans. Additionally, insurance companies and public authorities would be interested in such a model to avoid loss of money due to damages and hospital costs.

1.2 Data

To solve this problem, the data collected for Seattle city is used. The data includes all collisions provided by SPD and recorded by Traffic Records from 2004 to 2019. The dataset contains 7397574 entries for accident information with variable name of SEVERITY CODE and some related data such as location, road conditions etc in 38 columns (194673x38 table). The prediction will be done for SEVERITY CODE. First the attributes that will be used in the modeling will be selected and data will be prepared for model building. In my opinion, the attributes WEATHER, ROADCOND, LIGHTCOND are valuable for the prediction model. These attributes would mostly available for any time of the year; so, the model would work with their values. Also, some filtering might be done by using attributes such as UNDERINFL, which shows whether the driver was under influence of drugs etc.

1.3 Methods

In the final model, SEVERITY CODE is used as target while the attributes WEATHER describing the weather conditions during the time of the collision, ROADCOND showing the condition of the

road during the collision and LIGHTCOND showing the light conditions during the collision are used as predictors.

INATTENTIONIND showing whether or not collision was due to inattention, UNDERINFL showing whether or not a driver involved was under the influence of drugs or alcohol and SPEEDING showing whether or not speeding was a factor in the collision are used in the data exploration step.

First, the data table consisting only the variables that are used in the model and exploratory step is created. SEVERITY CODE values of 0 (0 = unknown) are planned to be excluded from the analysis since they do noy provide any information on accident severity . It is observed that the only severity codes included in the initial data were 1 and 2; so, no exclusion was made by SEVERITY CODE. 9333 rows with SPEEDING answered yes, 5126 (entry Y) + 3995 (entry 1) = 9121 rows with UNDERINFL answered yes and 29805 rows with INATTENTIONIND answered yes are excluded to avoid person-related accidents. The final data frame has shape of 144089x4 excluding 3 columns (SPEEDING, INATTENTIONIND and UNDERINFL) used in the preparation step and null entries for predictory variables.

Associations between the predictors between themselves and with target values are explored via using Cramer's V statistic since all variables are categorical. Moderate correlation between WEATHER and ROADCOND (0.48), and low correlation between WEATHER and LIGHTCOND (0.27) and between ROADCOND and LIGHTCOND (0.27) are found. None of the predictors are found to be highly correlated with target value (0.17-0.18 are the calculated correlation values).

The problem is a classification problem with two different output values for the target variable. Thus, classification methods are applied: KNN, decision trees, logistic regression and support vector machines. Before the method applications the dataset is divided into training (70%) and test sets (30%). The predictor variables are converted to dummy variables.

1.4 Results and Discussion

K nearest neighbors method gives best accuracy of 71.5% with k = 8. All methods are compared by using Jaccard score. There is no method which result in very good accuracy. KNN with k = 8 is selected as the final model with mean accuracy of about 50%.

1.5 Conclusion

The predictive model for accident severity using weather, road and light conditions provides good accuracy. The resulted model can be implemented in any algorithm used to calculate insurance rates. However, there should be some improvement of the model for low number of more severe accidents and maybe additional predictive variables.

1.6 Calculations

```
/Users/toz022/Library/Python/3.7/lib/python/site-
packages/IPython/core/interactiveshell.py:3072: DtypeWarning: Columns (33) have
mixed types.Specify dtype option on import or set low_memory=False.
  interactivity=interactivity, compiler=compiler, result=result)
```

[3]:

	SEVERITYCODE	X	Y	OBJECTID	INCKEY	COLDETKEY	REPORTNO	\
0	2	-122.323148	47.703140	1	1307	1307	3502005	
1	1	-122.347294	47.647172	2	52200	52200	2607959	
2	1	-122.334540	47.607871	3	26700	26700	1482393	
3	1	-122.334803	47.604803	4	1144	1144	3503937	
4	2	-122.306426	47.545739	5	17700	17700	1807429	

	STATUS	ADDRTYPE	INTKEY	…	ROADCOND	LIGHTCOND	\
0	Matched	Intersection	37475.0	…	Wet	Daylight	
1	Matched	Block	NaN	…	Wet	Dark - Street Lights On	
2	Matched	Block	NaN	…	Dry	Daylight	
3	Matched	Block	NaN	…	Dry	Daylight	
4	Matched	Intersection	34387.0	…	Wet	Daylight	

	PEDROWNOTGRNT	SDOTCOLNUM	SPEEDING	ST_COLCODE	\
0	NaN	NaN	NaN	10	
1	NaN	6354039.0	NaN	11	
2	NaN	4323031.0	NaN	32	
3	NaN	NaN	NaN	23	
4	NaN	4028032.0	NaN	10	

	ST_COLDESC	SEGLANEKEY	\
0	Entering at angle	0	
1	From same direction - both going straight - bo…	0	
2	One parked--one moving	0	
3	From same direction - all others	0	
4	Entering at angle	0	

	CROSSWALKKEY	HITPARKEDCAR
0	0	N
1	0	N
2	0	N
3	0	N
4	0	N

[5 rows x 38 columns]

2 Data Preparation and Exploration

[5]:
```
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'],
      dtype='object')
```

[26]:
```
(194673, 7)
```

[27]:
```
(194673, 7)
```

[23]:
```
1    136485
2     58188
Name: SEVERITYCODE, dtype: int64
```

[28]:
```
SPEEDING
Y           9333
dtype: int64
```

[14]:
```
UNDERINFL
N            100274
0             80394
Y              5126
1              3995
dtype: int64
```

[15]:
```
INATTENTIONIND
Y                 29805
dtype: int64
```

[33]:
```
(149315, 4)
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 149315 entries, 0 to 194672
Data columns (total 4 columns):
 #   Column        Non-Null Count   Dtype
---  ------        --------------   -----
 0   SEVERITYCODE  149315 non-null  int64
 1   WEATHER       144281 non-null  object
 2   ROADCOND      144320 non-null  object
 3   LIGHTCOND     144206 non-null  object
dtypes: int64(1), object(3)
memory usage: 5.7+ MB
```

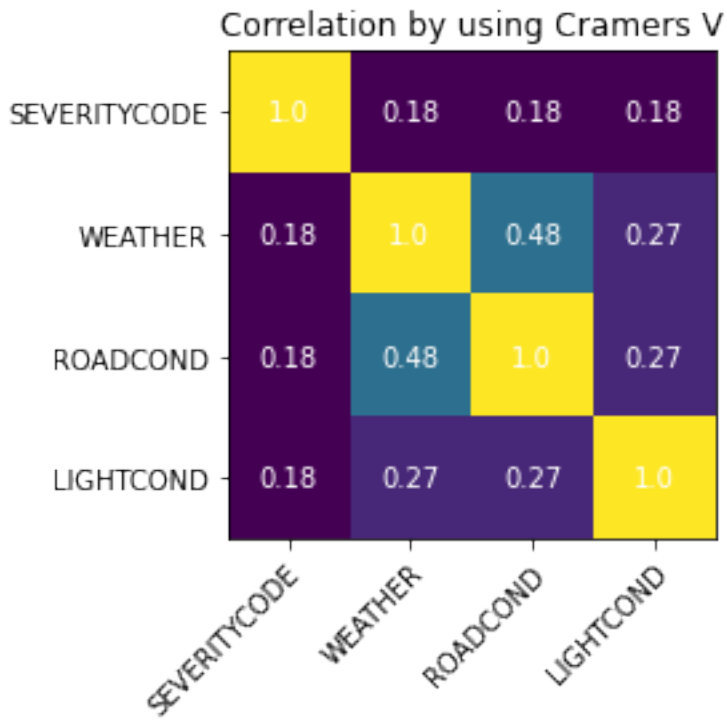
[39]:
```
(144089, 4)
```

[40]:
```
WEATHER
Clear                       83538
Raining                     23901
Overcast                    20579
Unknown                     14146
```

```
Other                          734
Snowing                        649
Fog/Smog/Smoke                 389
Sleet/Hail/Freezing Rain        82
Blowing Sand/Dirt               46
Severe Crosswind                20
Partly Cloudy                    5
dtype: int64
```

[41]:
```
ROADCOND
Dry               93944
Wet               34312
Unknown           14096
Ice                 771
Snow/Slush          704
Other                93
Standing Water       58
Oil                  57
Sand/Mud/Dirt        54
dtype: int64
```

[42]:
```
LIGHTCOND
Daylight                    89384
Dark - Street Lights On     33451
Unknown                     12682
Dusk                         4563
Dawn                         1859
Dark - No Street Lights      1121
Dark - Street Lights Off      827
Other                         191
Dark - Unknown Lighting        11
dtype: int64
```



3 Machine Learning

[96]:
```
array([['Overcast', 'Wet', 'Daylight'],
       ['Raining', 'Wet', 'Dark - Street Lights On'],
       ['Overcast', 'Dry', 'Daylight'],
       ['Clear', 'Dry', 'Daylight'],
       ['Raining', 'Wet', 'Daylight']], dtype=object)
```

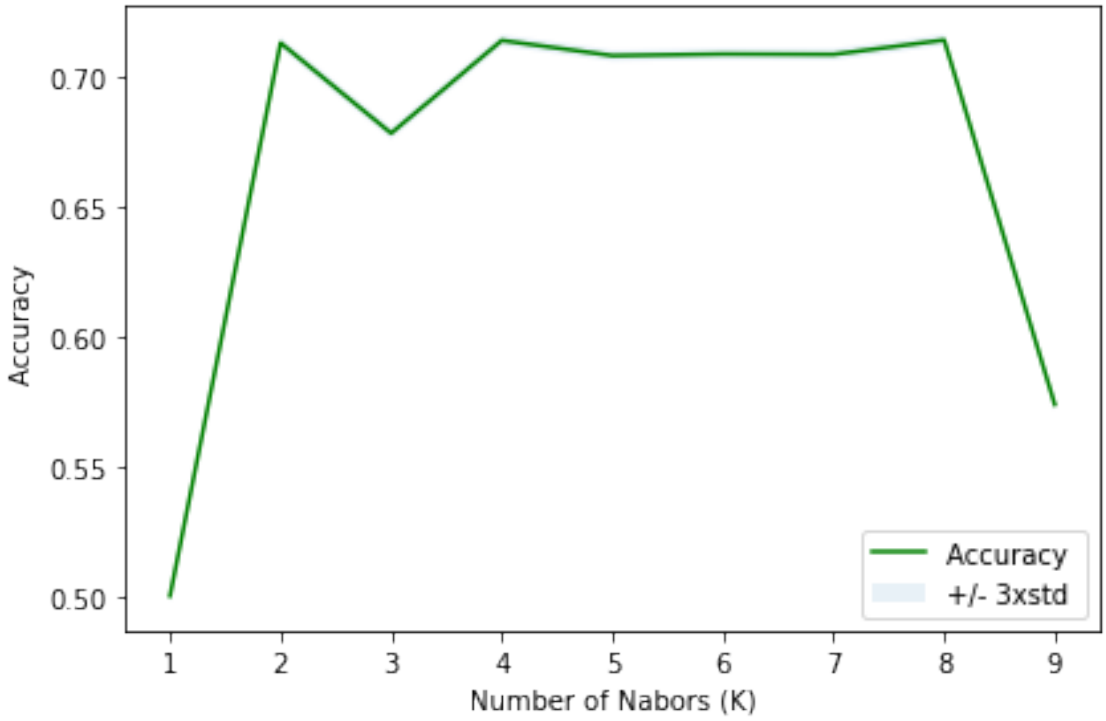
[97]:
```
array([[4, 8, 5],
       [6, 8, 2],
       [4, 0, 5],
       [1, 0, 5],
       [6, 8, 5]], dtype=object)
```

[98]:
```
0    2
1    1
2    1
3    1
4    2
Name: SEVERITYCODE, dtype: int64
```

```
Train set: (100862, 3) (100862,)
Test set: (43227, 3) (43227,)
```

3.1 KNN

[101]:
```
array([0.500, 0.713, 0.679, 0.714, 0.709, 0.709, 0.709, 0.715, 0.574])
```




```
The best accuracy was with 0.7145765378120157 with k= 8
```

[132]:
```
0.714470852329268
```

	precision	recall	f1-score	support
1	0.72	1.00	0.83	30917
2	0.27	0.00	0.00	12310
accuracy			0.71	43227
macro avg	0.49	0.50	0.42	43227
weighted avg	0.59	0.71	0.60	43227

4 Logistic regression

[104]:
```
LogisticRegression(C=0.01, solver='liblinear')
```

[116]:
```
0.7152242811205959
```

	precision	recall	f1-score	support
1	0.72	1.00	0.83	30917
2	0.00	0.00	0.00	12310
accuracy			0.72	43227
macro avg	0.36	0.50	0.42	43227
weighted avg	0.51	0.72	0.60	43227

```
/Users/toz022/Library/Python/3.7/lib/python/site-
packages/sklearn/metrics/_classification.py:1221: UndefinedMetricWarning:
Precision and F-score are ill-defined and being set to 0.0 in labels with no
predicted samples. Use `zero_division` parameter to control this behavior.
  _warn_prf(average, modifier, msg_start, len(result))
```

5 Decision trees

[113]:
```
0.7152242811205959
```

6 SVM

[117]:
```
0.7152242811205959
```

	precision	recall	f1-score	support
1	0.72	1.00	0.83	30917
2	0.00	0.00	0.00	12310
accuracy			0.72	43227
macro avg	0.36	0.50	0.42	43227
weighted avg	0.51	0.72	0.60	43227

```
/Users/toz022/Library/Python/3.7/lib/python/site-
packages/sklearn/metrics/_classification.py:1221: UndefinedMetricWarning:
Precision and F-score are ill-defined and being set to 0.0 in labels with no
predicted samples. Use `zero_division` parameter to control this behavior.
  _warn_prf(average, modifier, msg_start, len(result))
```