2020

PREDICTING THE SEVERITY OF ACCIDENT COLLISIONS IN THE RAINY SEASON

VINCENT GYAN

Contents

INTRODUCTION

The CEO of Batogo Insurance Company is proposing to the Board of Directors to either increase the premiums for accident injury insurance or property damage insurance in the coming rainy season. He has consulted me to advise him on the basis of the road accident data available.

On my inspection of the data, I noted only 2 categories of the severity types in the metadata were recorded so far: 'Injury' and 'property damage'. Since this outcome is binary I decided to solve the problem by logistic regression based on their probabilities of them occurring.

About dataset

The dataset is about past road accidents. The Accident Collisions data set includes details of 19,4673 accident cases and 38 fields for all years. However after careful inspection of the data, I noted that the following fields are of more relevance to solve the problem at hand:

Field	Description
SEVERITYCODE	Codes coresponding to the severity of the accident:
WEATHER	Description of the weather condition at the time of collision
ROADCOND	Description of the road condition at the time of the collision
ST_COLCODE	A code describing the collision.

METHODOLOGY

Loading the Data

Below is a snapshot of the loaded original data showing the first 12 columns and first 5 rows

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

5 rows × 38 columns

Selecting the relevant fields

After a careful inspection of the fields in the data in light of the problem at hand, the following fields were selected for a logistic regression analysis:

First Five and Last Five Columns

	SEVERITYCODE	WEATHER	ROADCOND	SDOT_COLCODE
0	2	Overcast	Wet	11
1	1	Raining	Wet	16
2	1	Overcast	Dry	14
3	1	Clear	Dry	11
4	2	Raining	Wet	11
...	...	...	...	...
194668	2	Clear	Dry	11
194669	1	Raining	Wet	14
194670	2	Clear	Dry	11
194671	2	Clear	Dry	51
194672	1	Clear	Wet	14

194673 rows × 4 columns

As the CEO's problem is about possible increase in insurance claims in the coming rainy season because of the coming recurrent torrential rainfall season every 10 Years as predicted by the weather department, I selected the WEATHER and the ROADCOND fields for further exploration.

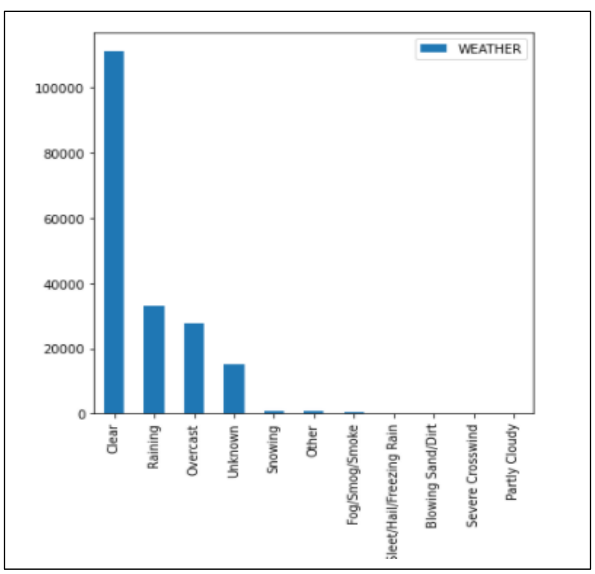
Below is the frequency of occurrence of factors in each field. 'Raining' and 'wet' conditions which will be prevalent in the rainy season are the second important factors respectively in the fields.

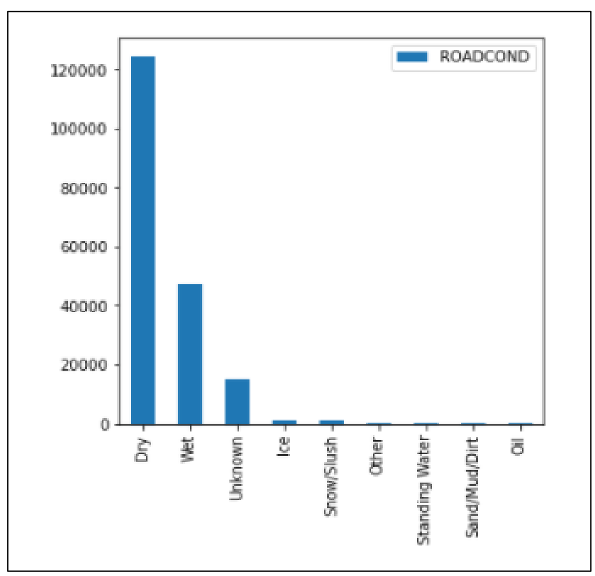
	SEVERITYCODE
1	136485
2	58188

	WEATHER
Clear	111135
Raining	33145
Overcast	27714
Unknown	15091
Snowing	907
Other	832
Fog/Smog/Smoke	569
Sleet/Hail/Freezing Rain	113
Blowing Sand/Dirt	56
Severe Crosswind	25
Partly Cloudy	5

	ROADCOND
Dry	124510
Wet	47474
Unknown	15078
Ice	1209
Snow/Slush	1004
Other	132
Standing Water	115
Sand/Mud/Dirt	75
Oil	64

Data visualization of the WEATHER and ROADCOND fields





Preprocessing the data

The SEVERITYCODE field was isolated from the data, converted into a binary outcome of 0 and 1 and stored as the label for the prediction.

One Hot Encoding of the categorical factors

As observed above, all the factors in the WEATHER and ROADCON fields are categorical and are therefore coded into numeric values as in the snapshot below:

	WEATHER_Blowing Sand/Dirt	WEATHER_Clear	WEATHER_Fog/Smog/Smoke	WEATHER_Other	WEATHER_Overcast	WEATHER_Partly Cloudy	WEATHER_Raining
0	0	0	0	0	1	0	0
1	0	0	0	0	0	0	1
2	0	0	0	0	1	0	0
3	0	1	0	0	0	0	0
4	0	0	0	0	0	0	1
5	0	1	0	0	0	0	0
6	0	0	0	0	0	0	1
7	0	1	0	0	0	0	0
8	0	1	0	0	0	0	0
9	0	1	0	0	0	0	0
10	0	0	0	0	1	0	0
11	0	1	0	0	0	0	0
12	0	0	0	0	0	0	1
13	0	0	0	0	0	0	1
14	0	1	0	0	0	0	0
15	0	0	0	0	0	0	0
16	0	0	0	0	1	0	0
17	0	0	0	0	1	0	0
18	0	1	0	0	0	0	0
19	0	0	0	0	0	0	0

The SDOT_COLCODE field is numerical and therefore remain same.

As the model must be tested on an unseen data, I split the data into training and test sets as follows:

```
X_train.shape
(155738, 20)

X_test.shape
(38935, 20)

y_train.shape
(155738,)

y_test.shape
(38935,)
```

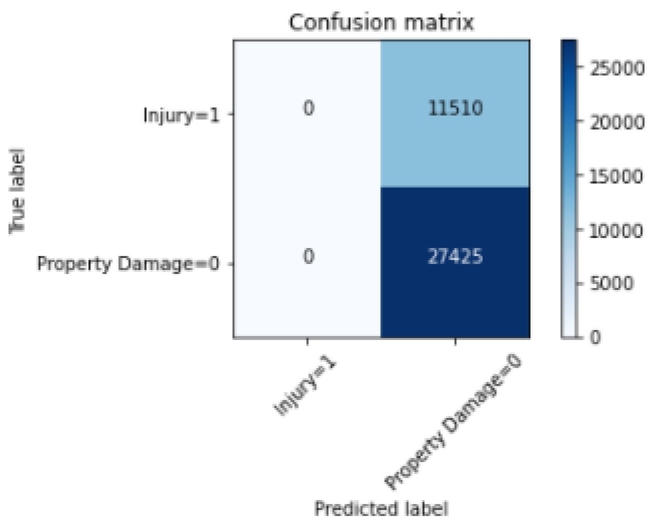
MODELING AND PREDICTION

The logistic regression algorithm was trained with the training set and tested on the test set with the outcomes below:

```
array([[0.66042117, 0.33957883],
       [0.67825467, 0.32174533],
       [0.67825467, 0.32174533],
       ...,
       [0.67567911, 0.32432089],
       [0.68347132, 0.31652868],
       [0.66042117, 0.33957883]])
```

Snapshots of resulting prediction, confusion matrix for evaluation and the precision accuracy measured and computation of precision.

```
Confusion matrix, without normalization
[[    0 11510]
 [    0 27425]]
```




```
              precision    recall  f1-score   support

           0       0.70      1.00      0.83     27425
           1       0.00      0.00      0.00     11510

    accuracy                           0.70     38935
   macro avg       0.35      0.50      0.41     38935
weighted avg       0.50      0.70      0.58     38935
```

RESULTS

Precision is a measure of the accuracy provided that a class label has been predicted. It is defined by: precision = TP / (TP + FP)
Recall is true positive rate. It is defined as: Recall = TP / (TP + FN)
So, I can calculate precision and recall of each class.
F1 score: Now I am in the position to calculate the F1 scores for each label based on the precision and recall of that label.
The F1 score is the harmonic average of the precision and recall, where an F1 score reaches its best value at 1 (perfect precision and recall) and worst at 0.

Precision is a measure of the accuracy provided that a class label has been predicted. It is defined by: precision = TP / (TP + FP)
Recall is true positive rate. It is defined as: Recall = TP / (TP + FN)
So, I can calculate precision and recall of each class.
F1 score: Now I am in the position to calculate the F1 scores for each label based on the precision and recall of that label.
The F1 score is the harmonic average of the precision and recall, where an F1 score reaches its best value at 1 (perfect precision and recall) and worst at 0.

CONCLUSION

Finally, I can tell the average accuracy for this classifier is the average of the F1-score for both labels, which is 0.83(0 for injury and 0.83 for property damage.
I therefore recommend that the CEO can go ahead to increase the premiums for property damage as it has the highest prediction of 83%.