Car Collision Severity Report

Ruxian Li

Introduction

Background

Car collisions can be caused by a variety of elements by different scopes. With the dataset in this study, two important factors – weather and road conditions are dedicated to analyzing their impact on the severity of a car collisions based on logistic regression machine learning model. Afterwards, three evaluation metrics would be used to further evaluate respective accuracy level.

Problem

What things can lead to a car collision? There could be numerous reasons, such as human-led factors like concentration level and very importantly external factors including weather and road conditions. These two different factors from external environments are very often neglected in the consideration, but it is not any less crucial in the determination of the car collisions occurrence. Are there any models to study this assumption? It is the research question this study focuses on.

Interest

The study of the factors of car collisions occurrence could potentially help different stakeholders in the society. The government officials such as policy-makers and city planners are important roles to secure the proper construction of the external environment. The car drivers and passengers should also take good consideration of potential risks during the drive. Academic researchers could also have a fresh perspective of this social study. All in all, different people and roles in the society might be interested in the study of the severity of car collisions cases and the weather and road conditions.

Data and Methodology

Data Collection

This study focuses on the data retrieved from SDOT Traffic Management Division, Traffic Records Group (SDOT GIS Analyst, 2020). The data come from the official resources from the Seattle government and archived in the GISWEB. It is in the format of (csv.) which is easy for Python processing.

Data Understanding and Analysis

The full dataset compiles a broad spectrum of car collisions cases according to the detailed information including data, severity level, description, location, weather, road condition, etc. As this study focuses on the weather, road conditions, and severity level, the data of these three columns should be further analyzed. For severity level, the attribute “SEVERITYCODE” should be taken into account. A code that corresponds to the severity of the collision: 3—

fatality; 2b—serious injury; 2—injury; 1—prop damage; 0—unknown. In this study, there are only two levels of severity noted. They are 2—injury and 1—prop damage. Weather conditions are recorded under attribute “WEATHER” and road conditions under “ROADCOND”. With an understanding of these data, the study can further work on these items.

Methodology

IBM Watson Studio is used to run the Jupyter Notebook on Cloud computing for a seamless working. In the notebook, libraries such as pandas, matplotlib, scikit-learn, NumPy are deployed to run the codes. Initially, the study explores the dataset by the basic retrieval of the dataset information, headers, types, columns, and other attributes. Before modeling, the dataset went through data balancing and data standardization to ensure greater level of accuracy. Afterwards, it is split into train dataset and test dataset for model evaluation. Moreover, a machine learning models is deployed in this study. As the original severity code is already binary, using logistic regression is expected to be an ideal choice. In the end, three evaluation metrics such as Jaccard index, GF-1 score and Logloss would be used to evaluate the model.

Results

Initially it is to read the data frame and to scroll through the first few rows.

```
df=pd.read_csv("Data_Collision.csv")
df.head()
```

```
/opt/conda/envs/Python36/lib/python3.6/site-packages/IPython/core/interactiveshell.py:3020: DtypeWarning: Columns (33) have mixed types. Specify dtype option on import or set low_memory=False.
  interactivity=interactivity, compiler=compiler, result=result)
```

Out[98]:

	SEVERITYCODE	X	Y	OBJECTID	INCKEY	COLDETKEY	REPORTNO	STATUS	ADDRTYPE	INTKEY	...	ROADCOND	LIGHTCOND	PEDROWNOTGRNT	SDOTCO
0	2	-122.323148	47.703140	1	1307	1307	3502005	Matched	Intersection	37475.0	...	Wet	Daylight	NaN	
1	1	-122.347294	47.647172	2	52200	52200	2607959	Matched	Block	NaN	...	Wet	Dark - Street Lights On	NaN	635
2	1	-122.334540	47.607871	3	26700	26700	1482393	Matched	Block	NaN	...	Dry	Daylight	NaN	432
3	1	-122.334803	47.604803	4	1144	1144	3503937	Matched	Block	NaN	...	Dry	Daylight	NaN	
4	2	-122.306426	47.545739	5	17700	17700	1807429	Matched	Intersection	34387.0	...	Wet	Daylight	NaN	402

5 rows × 38 columns

It is indeed a big dataset with below 38 columns:

'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'.

Studying the main target severity code is very crucial. By analyzing its “value counts”, we got two very distinctive numbers, 136485 for code 1 and 58188 for code 2. Therefore, the study aims to down-sample the code 1 for later higher accuracy. The result achieves to be the same.

Before:

```
1    136485
2     58188
Name: SEVERITYCODE, dtype: int64
```

After:

```
2    58188
1    58188
Name: SEVERITYCODE, dtype: int64
```

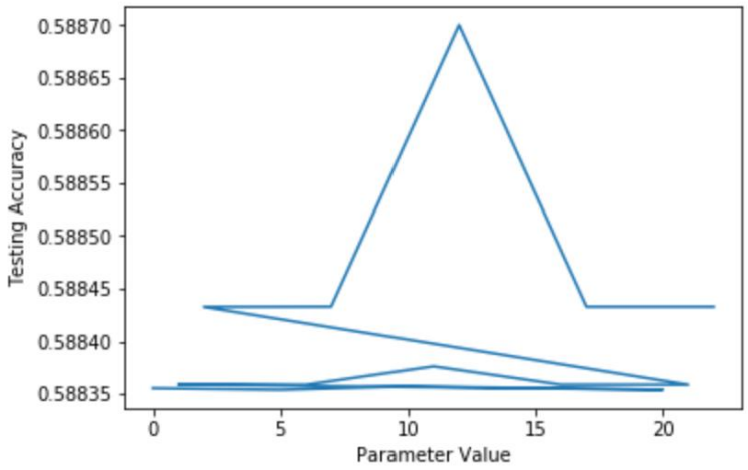
As the weather conditions and road conditions are in string format, to better study them, “get dummies” function is used to separate the features of both columns and their values would be represented by a number of 1. The first few rows and columns can be shown in below screenshot.

	Blowing Sand/Dirt	Clear	Fog/Smog/Smoke	Other	Overcast	Partly Cloudy	Raining	Severe Crosswind	Sleet/Hail/Freezing Rain	Snowing	Unknown	Dry	Ice	Oil	Other	Sand/Mud/Dirt	Snow/Slush	Standin Wat
0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	
1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	
2	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	
3	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	
4	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	

After standardization the data, the dataset is further split into test and train set (test size=0.2, random state=4) pending for data modeling and evaluation.

With all the data well preprocessed, the data should be ready to go through logistic regression modeling. Under the prediction and analysis, respective accuracy numbers generated from each test can be shown as below. By visualizing the result, the parameters can be more clearly displayed. As shown, all the tests’ accuracy locates at a moderate level. Therefore, it proves that the weather and road conditions have certain impact on the severity of car collision.

```
Text(0, 0.5, 'Testing Accuracy')
```



```
Test 1: Accuracy at C = 0.1 when Solver = lbfgs is : 0.5883553079503667
/opt/conda/envs/Python36/lib/python3.6/site-packages/sklearn/linear_model/sag.py:334: ConvergenceWarning: The max_iter was reached which
means the coef_ did not converge
  "the coef_ did not converge", ConvergenceWarning)
Test 2: Accuracy at C = 0.1 when Solver = saga is : 0.5883537706372713
Test 3: Accuracy at C = 0.1 when Solver = liblinear is : 0.588357072655139
Test 4: Accuracy at C = 0.1 when Solver = newton-cg is : 0.5883554096487158
/opt/conda/envs/Python36/lib/python3.6/site-packages/sklearn/linear_model/sag.py:334: ConvergenceWarning: The max_iter was reached which
means the coef_ did not converge
  "the coef_ did not converge", ConvergenceWarning)
Test 5: Accuracy at C = 0.1 when Solver = sag is : 0.58835389837573


Test 6: Accuracy at C = 0.01 when Solver = lbfgs is : 0.5883587534078697
/opt/conda/envs/Python36/lib/python3.6/site-packages/sklearn/linear_model/sag.py:334: ConvergenceWarning: The max_iter was reached which
means the coef_ did not converge
  "the coef_ did not converge", ConvergenceWarning)
Test 7: Accuracy at C = 0.01 when Solver = saga is : 0.588358608474673
Test 8: Accuracy at C = 0.01 when Solver = liblinear is : 0.5883760819749968
Test 9: Accuracy at C = 0.01 when Solver = newton-cg is : 0.5883587362791604
/opt/conda/envs/Python36/lib/python3.6/site-packages/sklearn/linear_model/sag.py:334: ConvergenceWarning: The max_iter was reached which
means the coef_ did not converge
  "the coef_ did not converge", ConvergenceWarning)
Test 10: Accuracy at C = 0.01 when Solver = sag is : 0.5883586614124134


Test 11: Accuracy at C = 0.001 when Solver = lbfgs is : 0.5884326122352971
Test 12: Accuracy at C = 0.001 when Solver = saga is : 0.5884328295178463
Test 13: Accuracy at C = 0.001 when Solver = liblinear is : 0.5886999961940418
/opt/conda/envs/Python36/lib/python3.6/site-packages/scipy/optimize/linesearch.py:313: LineSearchWarning: The line search algorithm did n
ot converge
  warn('The line search algorithm did not converge', LineSearchWarning)
/opt/conda/envs/Python36/lib/python3.6/site-packages/sklearn/utils/optimize.py:195: UserWarning: Line Search failed
  warnings.warn('Line Search failed')
Test 14: Accuracy at C = 0.001 when Solver = newton-cg is : 0.5884327483155763
Test 15: Accuracy at C = 0.001 when Solver = sag is : 0.5884327615905889
```

However, the data analysis has not yet finished. Three evaluation metrics were used to assess the classification accuracy. They are Jaccard, F1=score, and LogLoss. Jaccard has a result of 0.7, proving the classification is considerably accurate. F1-score as 0.58 shows medium accurate. The LogLoss score 0.59 has demonstrated a less positive result but it is surely not a denial of the classification accuracy.

Algorithm	Jaccard	F1-score	LogLoss
Logistic Regression	0.7	0.58	0.59

Evaluation and discussion

Based on the data preprocessing, understanding, analysis, modeling, and evaluation, the study has shown a considerable level of impact of weather and road conditions on the severity of car collision of the community. The target of this study is SEVERITY, which in this dataset only has the value of 1 and 2. Therefore, using the logistics regression, the statistical modeling intended for binary variable is an ideal choice for this analysis. The model in the end also shows a positive impact with a moderate level of accuracy, as also in the end proven by three metrics – Jaccard, F1-score, and LogLoss.

Although the results have shown that the weather and road conditions could potentially contribute to the car collision, there should be further analysis involving a broader spectrum of features to make this prediction more generalizable. This study has some limits, but can serve as a starting point to dig out more ideas related to this target. A more holistic approach to this topic can benefit not only the drivers but also many other different stakeholders in the society.

Conclusion

In conclusion, this study has shown that the weather and road conditions have an impact on car collision severity in the community. Based on the dataset provided by the Seattle government, the logistic regression model is used to prove this relationship and is evaluated as moderately correct by three evaluation metrics – Jaccard, F1-score, and LogLoss. As supplemented, the scope of this study is limited. Therefore, a further analysis concerning more diverse aspects would be conducive to understanding the target more comprehensively.

References

SDOT GIS Analyst. (2020). Collisions—All Years. Traffic Records Group