Prediction of Car Accident Severity

Yongyou Hu, PhD

August 29, 2020

Overview

- Introduction
- Feature selection and data Cleaning
- Exploratory data analysis
- Machine learning models
- Results and discussion
- Conclusion

Introduction

- The Problem
  - ✓ Car accidents happen everyday and everywhere
  - ✓ High accident severity will involve injuries and even fatalities
  - ✓ Lack of a model to predict accident severity
- The Solution

  Extract insights from car accident data and build a prediction model to help people drive more safely

Feature Selection & Data Cleaning

- Feature selection

(1) 'VEHCOUNT', 'SPEEDING', 'ADDRTYPE', 'WEATHER', 'COLLISIONTYPE', 'PERSONCOUNT', 'INCDTTM' were selected based on meaning of each feature

(2) 'INCDTTM' was translated into 'WEEKEND' and 'TIMEOFDAY'.

(3) The following features were added with dummy ones:

"TIMEOFDAY": NIGHT, MORNING, AFTERNOON and EVENING

"ADDRTYPE": Alley, Block and Intersection

"COLLISIONTYPE": Angles, Cycles, Left Turn, Right Turn, Head on, Parked Car, Pedestrian, Rear Ended, Sidewipe and other

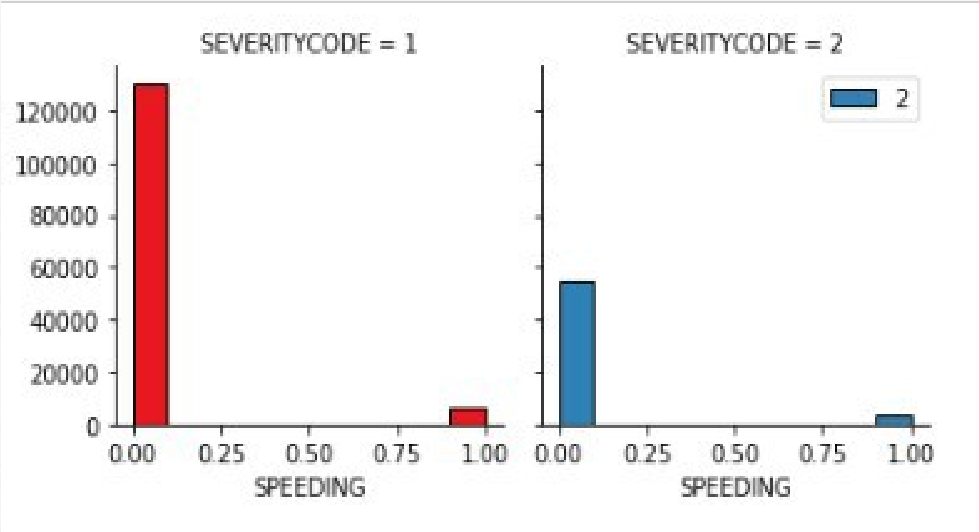
Feature Selection & Data Cleaning

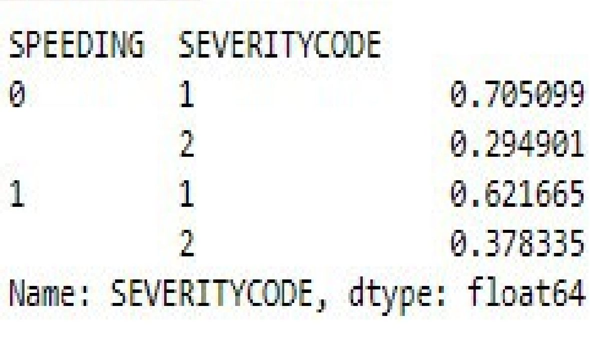
- Data cleaning

For simplicity, those data samples with missing values were deleted. After cleaning, the total data samples were reduced from 194673 to 187755.

Exploratory Data Analysis

- Relationship between SPEEDING and SEVERITY



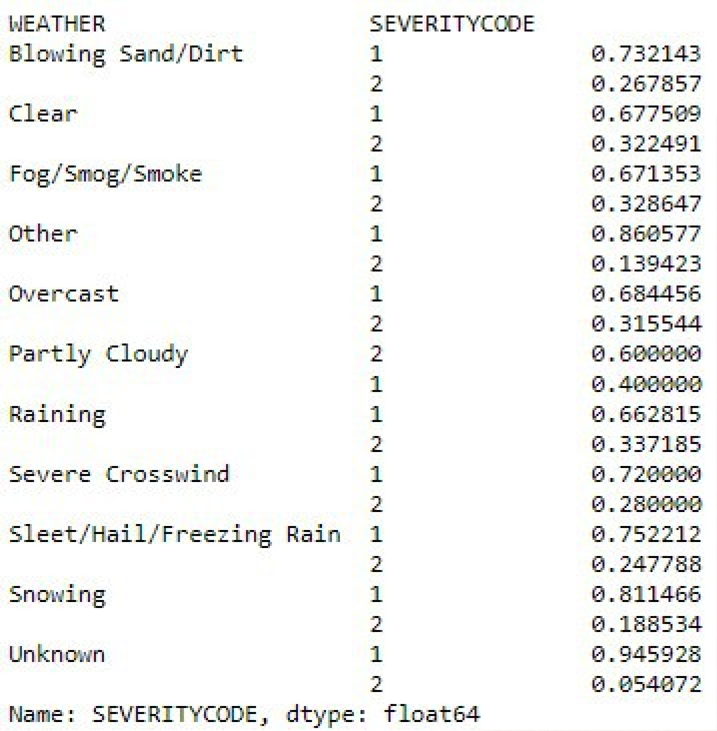


```
SPEEDING  SEVERITYCODE
0         1               0.705099
          2               0.294901
1         1               0.621665
          2               0.378335
Name: SEVERITYCODE, dtype: float64
```

August 29, 2020

Exploratory Data Analysis

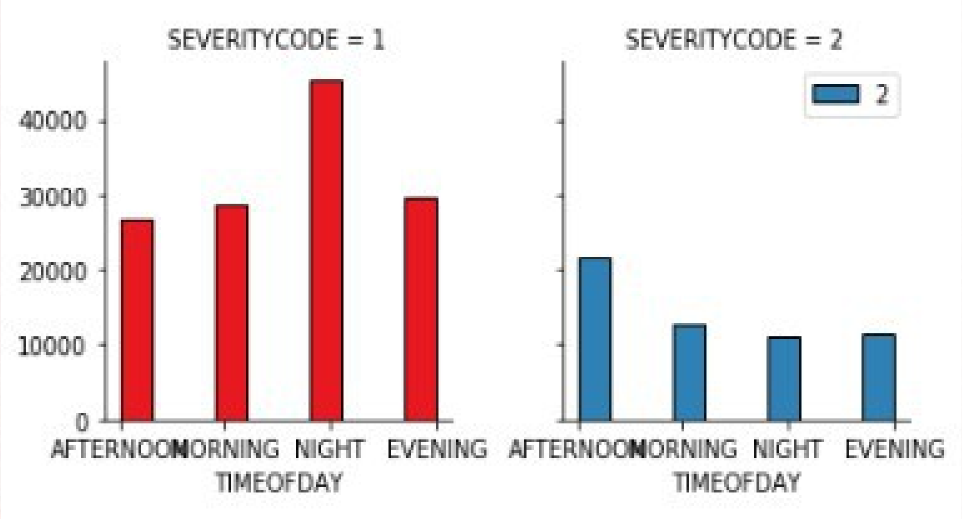
- Relationship between WEATHER and SEVERITY

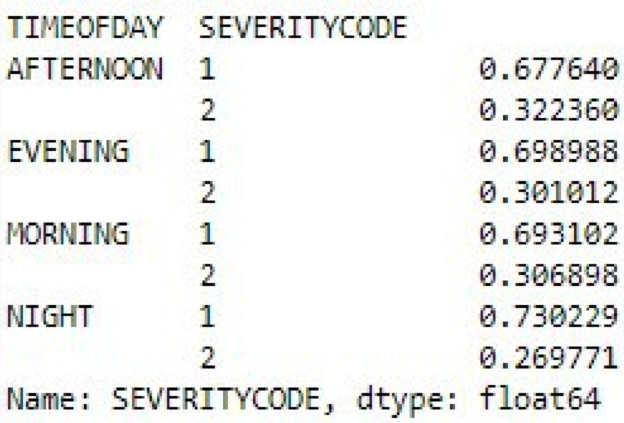


```
WEATHER                   SEVERITYCODE
Blowing Sand/Dirt         1               0.732143
                          2               0.267857
Clear                     1               0.677509
                          2               0.322491
Fog/Smog/Smoke            1               0.671353
                          2               0.328647
Other                     1               0.860577
                          2               0.139423
Overcast                  1               0.684456
                          2               0.315544
Partly Cloudy             2               0.600000
                          1               0.400000
Raining                   1               0.662815
                          2               0.337185
Severe Crosswind          1               0.720000
                          2               0.280000
Sleet/Hail/Freezing Rain  1               0.752212
                          2               0.247788
Snowing                   1               0.811466
                          2               0.188534
Unknown                   1               0.945928
                          2               0.054072
Name: SEVERITYCODE, dtype: float64
```

Exploratory Data Analysis

- Relationship between TIMEOFDAY and SEVERITY

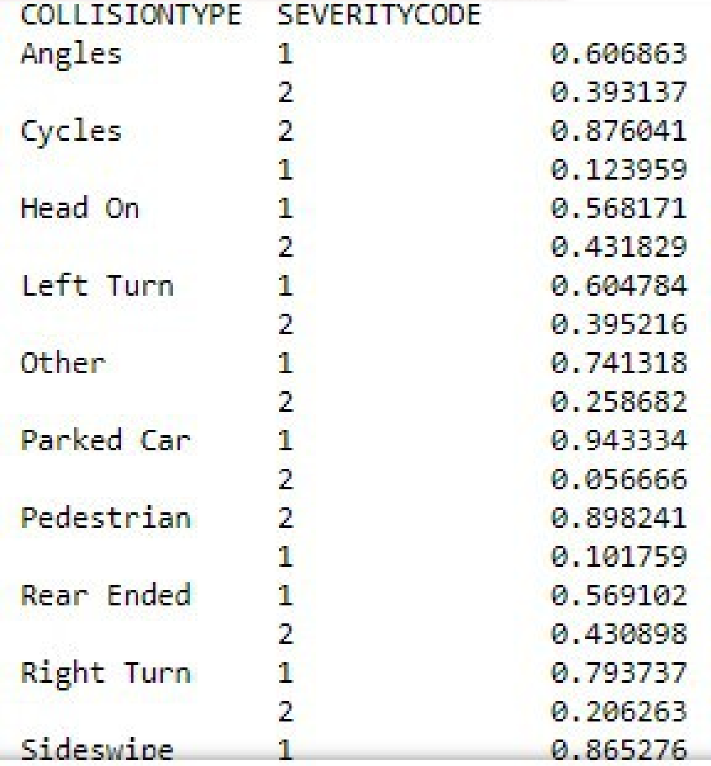




```
TIMEOFDAY  SEVERITYCODE
AFTERNOON  1               0.677640
           2               0.322360
EVENING    1               0.698988
           2               0.301012
MORNING    1               0.693102
           2               0.306898
NIGHT      1               0.730229
           2               0.269771
Name: SEVERITYCODE, dtype: float64
```

Exploratory Data Analysis

- Relationship between COLLISIONTYPE and SEVERITY



```
COLLISIONTYPE  SEVERITYCODE
Angles         1               0.606863
               2               0.393137
Cycles         2               0.876041
               1               0.123959
Head On        1               0.568171
               2               0.431829
Left Turn      1               0.604784
               2               0.395216
Other          1               0.741318
               2               0.258682
Parked Car     1               0.943334
               2               0.056666
Pedestrian     2               0.898241
               1               0.101759
Rear Ended     1               0.569102
               2               0.430898
Right Turn     1               0.793737
               2               0.206263
Sideswipe      1               0.865276
```

Machine Learning Models

- Three classification algorithms were investigated

1) Logistic Regression

The optimal regularization parameter, C was found to be 0.1 based on a few trials.

2) K-Nearest Neighbor (KNN)

The optimal neighbors, K was found to be 8 based on the max. accuracy

3) Support Vector Machines (SVM)

The optimal kernel was found to be radial basis function (rbf) based on results of all four kernels (linear, ploynomial, rbf and sigmoid)

Machine Learning Models

- Model performance

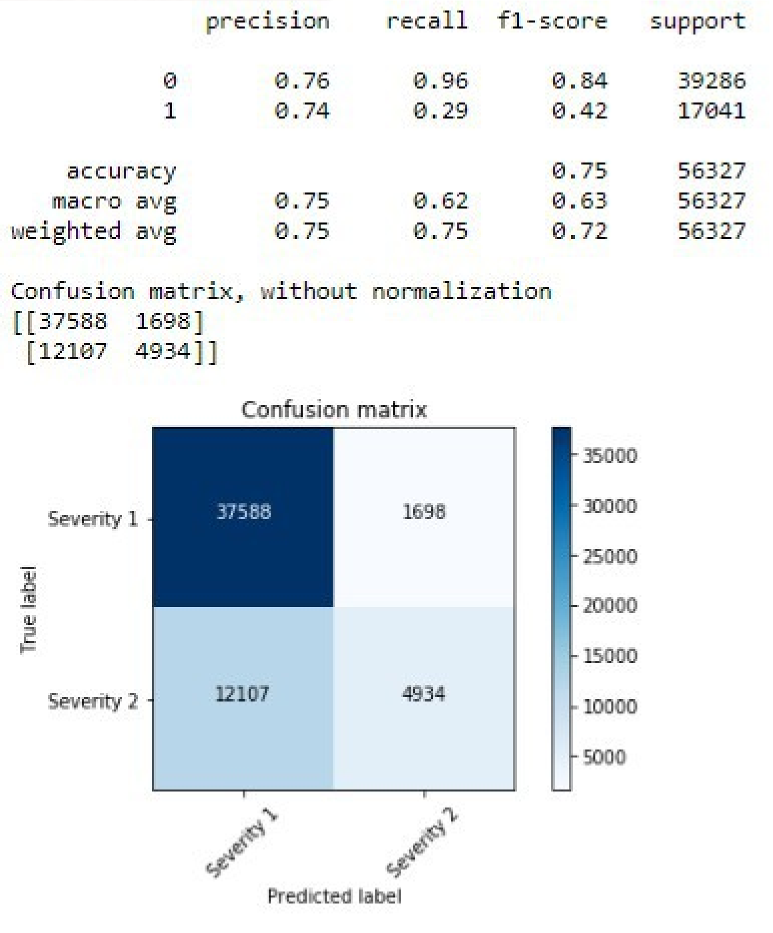
Algorithm Name	Jaccard	F1-Score	Log Loss
KNN	0.730	0.705	-
SVM	0.748	0.687	-
Logistic Regression	0.755	0.715	0.485

Logistic Regression is the fastest and most accurate model

August 29, 2020

Machine Learning Models

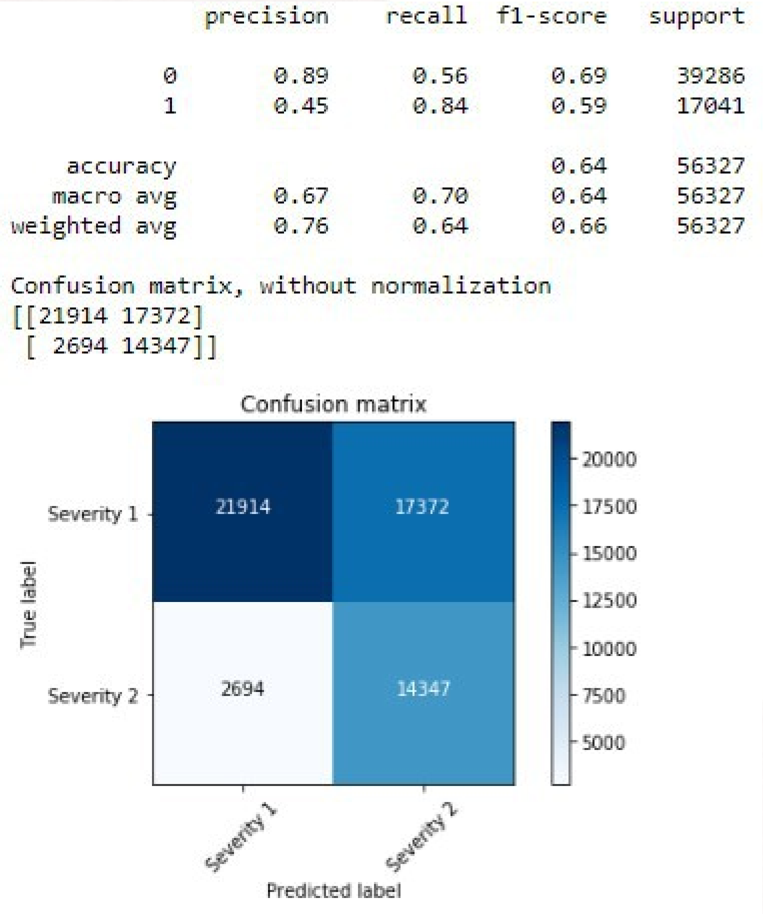
- Confusion Matrix of Logistic Regression Model with unbalanced labels



```
              precision    recall  f1-score   support

           0       0.76      0.96      0.84     39286
           1       0.74      0.29      0.42     17041

    accuracy                           0.75     56327
   macro avg       0.75      0.62      0.63     56327
weighted avg       0.75      0.75      0.72     56327

Confusion matrix, without normalization
[[37588  1698]
 [12107  4934]]
```

Machine Learning Models

- Confusion Matrix of Logistic Regression Model with balanced labels



```
              precision    recall  f1-score   support

           0       0.89      0.56      0.69     39286
           1       0.45      0.84      0.59     17041

    accuracy                           0.64     56327
   macro avg       0.67      0.70      0.64     56327
weighted avg       0.76      0.64      0.66     56327

Confusion matrix, without normalization
[[21914 17372]
 [ 2694 14347]]
```

Results & Discussion

- Logistic Regression Model is the best
- Weather and road conditions barely affected the accident severities
- Most accidents happened in the afternoon
- Accidents involving pedestrians and cyclists will most likely be Severity 2
- Drive more carefully at intersections and no speeding

Conclusion

- Driver should avoid speeding and driving in afternoon if possible
- This model will be helpful for drivers to decide if they should drive or not under certain conditions