Predict accident severity – Capstone project

Weigang Wei
20th September 2020

Introduction

- To predict the severity of an accident
- Prepared for transportation department
- Assist to set up diverted routes, and temporary speed limits etc.

Data description

- Data: 194673 rows and 38 columns.
- Unrelated columns removed:

    e.g. address, descriptions, coordinates, id

- Redundant columns removed:

    e.g. junction type, weather

- Data merged

    e.g. road condition → wet and dry

    light condition → bright and dark

Rows and columns for fitting

```
 #   Column            Non-Null Count   Dtype
---  ------            --------------   -----
 0   SEVERITYCODE          194673 non-null   int64
 1   PEDCOUNT              194673 non-null   int64
 2   VEHCOUNT              194673 non-null   int64
 3   SDOT_COLCODE          194673 non-null   int64
 4   SPEEDING_YES_NO       194673 non-null   int64
 5   Block                 194673 non-null   uint8
 6   Wet                   194673 non-null   uint8
 7   Dark                  194673 non-null   uint8
dtypes: int64(5), uint8(3)
```

Machine learning algorithms

- Algorithms

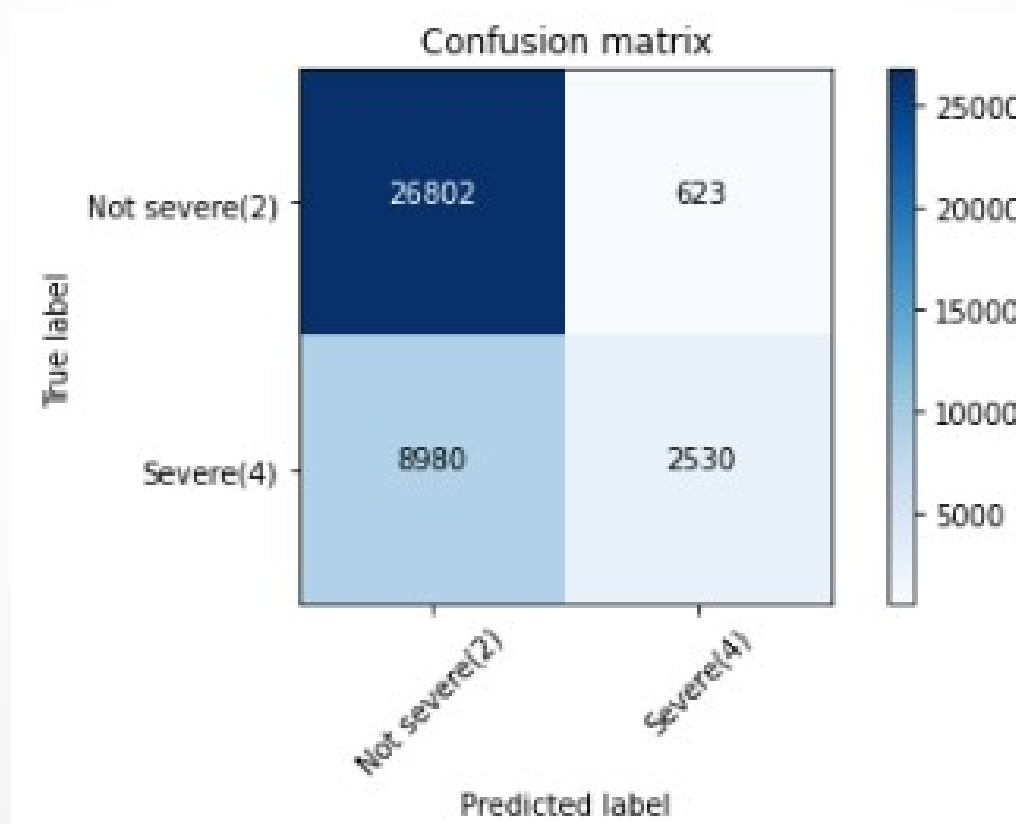   Knearest Neighbour, Decision Tree, Support Vector Machine, Logistic Regression

- Evaluation

    F1 score, Jaccard score, Log loss

Results

Table 2: Predicting accuracy

Algorithm	f1_score	jaccard_score	log_loss
KNN	0.70	0.73	--
Decision tree	0.68	0.73	--
SVM	0.70	0.74	--
Logistic regression	0.72	0.67	0.55

Confusion matrix

Discussion

- Severity code 1 is 2.3 times more than that of severity code 2. not evenly fitted possible and to predict more code 1 cases likely.

- The road conditions, and light conditions are simplified to binary values which may reduced the prediction accuracy

Conclusion

The dataset features were analysed and make it suitable for machine learning.

- Classify machine learning models were used and evaluated.

- Support Vector Machine is proved to be the best algorithm to predict the accident severity with F1 score of 0.70 and Jaccard score of 0.74.

- The accident severity may be predicted with relative high confidence.