Prediction model for Collision severity

Balaji Mohan

September 15, 2020

Introduction

Road accidents occur when a vehicle collides with another vehicle, pedestrian, animal, tree, or any stationary object like a building. Road accidents cause property damage or sometimes injury leading to either disability or even extreme case a threat to life.

Road accidents may occur due to various factors like visibility, traffic congestion on certain days, or weather. There are several statistics available in both open and closed domains for people to analyze and also create models to avoid road accidents in the future.

Once the models are created, it can either help reduce road accidents by warning the human beings or feeding these models into Autonomous vehicle systems to make clever decisions to avoid accidents.

The model can serve society in a very beneficial way of reducing property loss as well as the loss of life.

This model will particularly be helpful to the traffic department in warning the drivers by placing appropriate sign boards were accidents can be avoided. It can also serve the autonomous vehicle manufacturer to incorporate it into their vehicle system to make clever decisions based on the situations to avoid accidents.

Data

The dataset used in this study consists of 37 attributes and one target. The road accidents may be of any kind like a car collision, bicycle, other vehicle collisions. The data is collected from 2004 to present and provided by SPD and recorded by traffic records. The dataset is updated weekly based on the data collected every week.

The dataset was imbalanced with a lot of unknowns and null values. The dataset was cleaned by assigning the modal value of each attribute to the unknowns and null values. Then the attributes were encoded using a label encoder to assign integer values.

Methodology

The dataset after cleaning was split into training and test dataset as 70:30. The training dataset was used to train five different machine learning classifier algorithms. Then the testing dataset was used to evaluate the models by using the accuracy and F1 score.
The five machine learning classifier algorithms used are

- K Nearest neighbor
- Logistic Regression
- Support Vector machine

- Decision tree
- XGBoost

Results and discussion

The following subsections shows the confusion matrix for various models trained with 70% of the data from the dataset. The confusion matrix is plotted from the predicted and true data from the test dataset which remaining 30% of the collision dataset.

K-nearest Neighbor

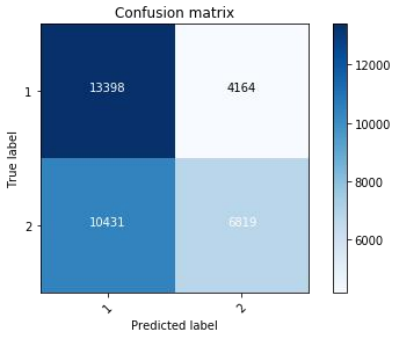

Fig. 1 Confusion matrix for K-nearest Neighbor model

Logistic regression

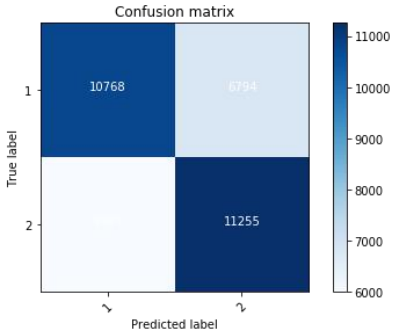

Fig. 2 Confusion matrix for logistic regression model

Support vector machine

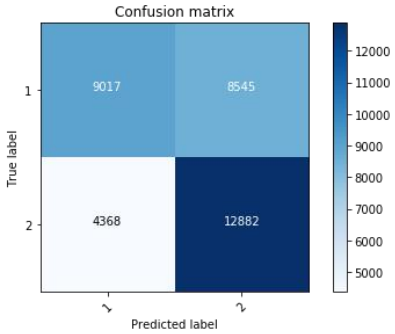

Fig. 3 Confusion matrix for support vector machine model

Decision tree classifier

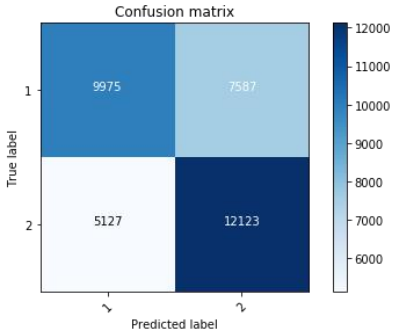

Fig. 4 Confusion matrix for Decision tree model

XGBoost

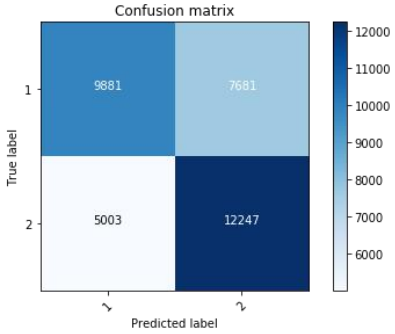

Fig. 5 Confusion matrix for XGBoost model

After training five different models, the k-nearest neighbor had an accuracy of 0.581 and an F1 score of 0.566, the Logistic regression model had an accuracy of 0.633 and an F1 score of 0.633, the SVM had an accuracy of 0.629 and an F1 score of 0.624, the decision tree also had an accuracy of 0.629 and F1 score of 0.624 and XGBoost had an accuracy of 0.629 and F1 score of 0.624.

The table 1 shows the accuracy score and F1 score for various models used in this study.

Table 1 Accuracy and F1 score for various models used in this study

	Score	F1 Score
KNN	0.581	0.566
Logistic Regression	0.633	0.633
SVM	0.629	0.624
Decision Tree	0.629	0.624
XGBoost	0.629	0.624

Conclusion

The models trained on the collision dataset provide a prediction to judge whether the collision would occur or not. The model also predicts the severity of collision in case of a collision. This model would help in reducing the collision and its severity.