Cousera Capstone Project – severity prediction

**Introduction**

As car is the main transportation choice for people in the U.S., it is important to make driving experience as safe as possible. There are many features that can affect driving, including road condition, weather, light condition, etc. To find out potential danger and avoid it, we need a machine learning model to explore car accidents data and find out the potential reasons for danger. This project will construct a model to predict what features relate to accident rate and give suggestions to avoid danger.

**Method**

1. Data

The data used is downloaded from [the link](https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv). It contains 194674 car accident cases with description of location, time, weather condition, road condition, light condition, other features and severity. Using this data, we can find the relation between severity of accident and features. If the feature and severity is related, we can tell that the feature can cause potential danger for drivers.

Six features are selected as independent variable, they are: 'LIGHTCOND', 'ROADCOND','WEATHER', 'JUNCTIONTYPE', 'COLLISIONTYPE' and 'ADDRTYPE'. The dependent variable is “SEVERITYCODE”, which has two different values, “1” and “2”.

Since data are stored as uncalculatable format, the string, an encoder should be employed to transform data from string to numbers. After applying the encoder method provided by pandas data frame, all strings are converted to the calculatable integer format.

Training set takes 80% of the whole set, and test set takes 20% of the whole set.

1. Algorithm

To choose the algorithm that has the best performance in this classification, three algorithms, including Logistic Regression, Naïve Bayes and Support Vector machine, are employed. The performance of them are compared, and the best one of them chosen by their performance.

**Results**

1. Correlation analysis

A correlation analysis chart is shown as below:

A screenshot of a cell phone

Description automatically generated

From the above chart we can see that except “WEATHER” and “ROADCOND”, other independent variables are related to severity. We deleted these two variables from the dataset and continue to apply machine learning model on it.

1. Logistic Regression

The score of logistic regression is 0.699 with confusion matrix:

|  |  |  |
| --- | --- | --- |
|  | Actually positive | Actually negative |
| Predicted positive | 27030 | 395 |
| Predicted negative | 11317 | 193 |

The f1 scores is 0.588.

1. Naïve bayes

The score of logistic regression is 0.667 with confusion matrix:

|  |  |  |
| --- | --- | --- |
|  | Actually positive | Actually negative |
| Predicted positive | 21461 | 5964 |
| Predicted negative | 7006 | 4504 |

The f1 scores is 0.662.

1. Support Vector Machine

The score of logistic regression is 0.720 with confusion matrix:

|  |  |  |
| --- | --- | --- |
|  | Actually positive | Actually negative |
| Predicted positive | 26463 | 962 |
| Predicted negative | 9948 | 1562 |

The f1 scores is 0.650.

**Discuss**

From the experiment results, we can see that Support Vector Machine has the best performance to predict. It achieves 0.650 f1 score, and 0.720 accuracy rate.