Road accident severity prediction-Capston Project

HASLA M

September 2020

1 Introduction

1.1 Problem

One of the important safety challenge in the modern world is to prevent or reduce road accidents. Although road accidents have become a common issue, the time has exceeded to take more steps toward reducing the number of accidents, or atleast reduce the severity of it. So the main point here is to analyse and find the main cause of these accidents. If it is found, the upcoming accidents or the severity of an accident can be predicted and there by giving some warning messages or alerts, the number of accidents/fatalities can be reduced. For example, if it is a heavy rainy day, there can be a possibility of vehicles to skid from the certain surface (road), or the condition of some roads may be very poor. So there by creating a warning to take diversion or not to pass through a particular way, may reduce a chance of an accident. It will be helpful if we can predict possibility of an accident, also the severity of the accidents. initially we need to discover the causes of why these accident has occurred, so that we can generate warning to the public for taking particular precautions.

1.2 Interest

The work will be helpful to the public, especially, for the one who drives a vehicle. This system will alert the drivers about the possibility of an accident by predicting a severity condition based on the some features like weather condition, condition of the road, lighting condition etc.

2 Data Acquisition and Cleaning

2.1 Data source

In order to predict an accident severity, we need to have knowledge about the previous accidents in that place. Then only we can figure out the causes or reason behind it , and predict based on that causes, so that everyone can that precautions if the case is matched. The dataset used here includes all types of collisions in Seattle from 2004 onwards, by SDOT Traffic Management Division, Traffic Records Group, Seattle. The dataset is available here Collision.csv They also provide a metadata file, which contain information about each attributes in this dataset. The dataset contain 40 attributes like SEVERITY_CODE (code that corresponds to the severity of the collision), JUNCTION_TYPE, ROAD_COND(condition of the road, where collision happened), FATALITIES, INJURED etc.

As our aim is to predict the severity of accidents. Our target variable will be SEVERITY_CODE. We'll examine the relationship between the target variable and all other attribute, through correlation analysis, plotting etc. This will help in selecting the better attributes, which can give a better prediction.

2.2 Data Cleaning

The dataset contain 40 attributes and 213797 rows. The target variable SEVERITYCODE contain values 0,1,2,2b,and 3. It is encoded into :

- 0 : Very less to zero probability of accident (Clear)
- 1 : Low probability (Property damage)
- 2 : Mild probability (Chance of getting injured)
- 3 : High probability (Serious Injury)
- 4 : Very high probability (Chance of fatality)

There are also other attributes with datatype as object. Each values of such attributes are encoded into numerical values. For example WEATHER attribute is encoded as :

- Clear : 0
- Partly Cloudly : 1
- Fog/Smog/Smoke : 2
- Sleet/Hail/Freezing Rain : 3
- Raining : 4
- Overcast : 5
- Snowing : 6
- Severe Crosswind : 7
- Blowing Sand/Dirt : 8
- Other : 9

- Unknown : 10
- Blowing Snow : 11

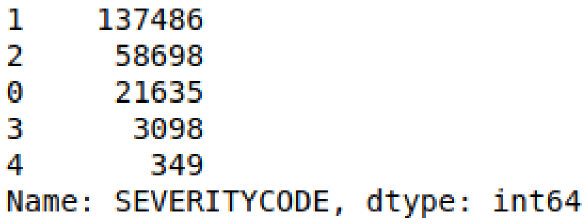
Simililarly other attributes like ADDRTYPE, COLLISIONTYPE, JUNCTIONTYPE, ROADCOND, LIGHTCOND, and UNDERINFL are also encoded with numeric values. Every missing values were enocded as '-1'. Correlation analysis of attributes were done, which will help in finding the best parameters.

2.3 Feature Selection

Initially we had 40 attributes. Many attributes doesn't contribute anything for prediction. also many attribute has missing values. So those attributes were removed. After correlation analysis, some of the attributes are selected. The selected features are

1. ADDRTYPE
2. COLLISIONTYPE
3. INJURIES
4. SERIOUSINJURIES
5. FATALITIES
6. JUNCTIONTYPE
7. SDOT_COLCODE
8. UNDERINFL
9. WEATHER
10. ROADCOND
11. LIGHTCOND
12. SEGLANEKEY
13. CROSSWALKKEY

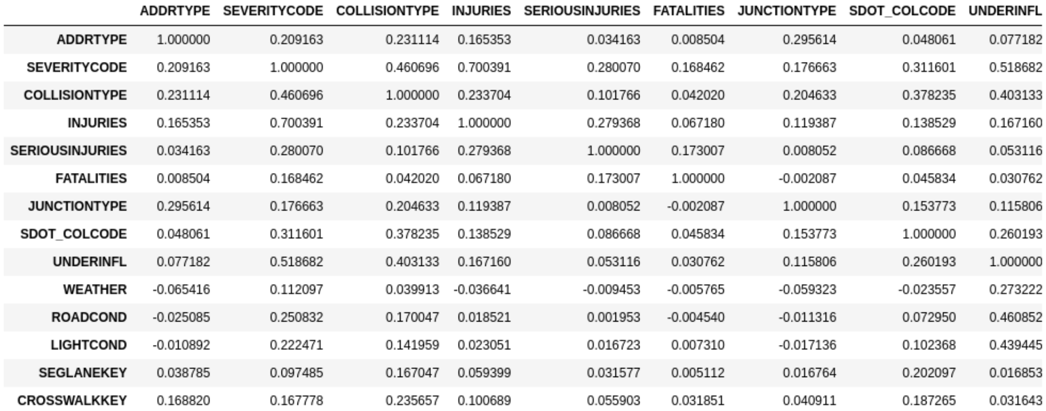
The target variable , SEVERITYCODE, values counts are



```
1    137486
2     58698
0     21635
3      3098
4       349
Name: SEVERITYCODE, dtype: int64
```

3 Exploratory Data Analysis

Correlation Analysis is statistical method that is used to discover if there is a relationship between two variables in the datasets, and how strong that relationship may be. The following is a correlation matrix :



	ADDRTYPE	SEVERITYCODE	COLLISIONTYPE	INJURIES	SERIOUSINJURIES	FATALITIES	JUNCTIONTYPE	SDOT_COLCODE	UNDERINFL
ADDRTYPE	1.000000	0.209163	0.231114	0.165353	0.034163	0.008504	0.295614	0.048061	0.077182
SEVERITYCODE	0.209163	1.000000	0.460696	0.700391	0.280070	0.168462	0.176663	0.311601	0.518682
COLLISIONTYPE	0.231114	0.460696	1.000000	0.233704	0.101766	0.042020	0.204633	0.378235	0.403133
INJURIES	0.165353	0.700391	0.233704	1.000000	0.279368	0.067180	0.119387	0.138529	0.167160
SERIOUSINJURIES	0.034163	0.280070	0.101766	0.279368	1.000000	0.173007	0.008052	0.086668	0.053116
FATALITIES	0.008504	0.168462	0.042020	0.067180	0.173007	1.000000	-0.002087	0.045834	0.030762
JUNCTIONTYPE	0.295614	0.176663	0.204633	0.119387	0.008052	-0.002087	1.000000	0.153773	0.115806
SDOT_COLCODE	0.048061	0.311601	0.378235	0.138529	0.086668	0.045834	0.153773	1.000000	0.260193
UNDERINFL	0.077182	0.518682	0.403133	0.167160	0.053116	0.030762	0.115806	0.260193	1.000000
WEATHER	-0.065416	0.112097	0.039913	-0.036641	-0.009453	-0.005765	-0.059323	-0.023557	0.273222
ROADCOND	-0.025085	0.250832	0.170047	0.018521	0.001953	-0.004540	-0.011316	0.072950	0.460852
LIGHTCOND	-0.010892	0.222471	0.141959	0.023051	0.016723	0.007310	-0.017136	0.102368	0.439445
SEGLANEKEY	0.038785	0.097485	0.167047	0.059399	0.031577	0.005112	0.016764	0.202097	0.016853
CROSSWALKKEY	0.168820	0.167778	0.235657	0.100689	0.055903	0.031851	0.040911	0.187265	0.031643

We have selected 11 most prominent attributes which give best correlation value with serverity code, as the features for building the model.

4 Model Building

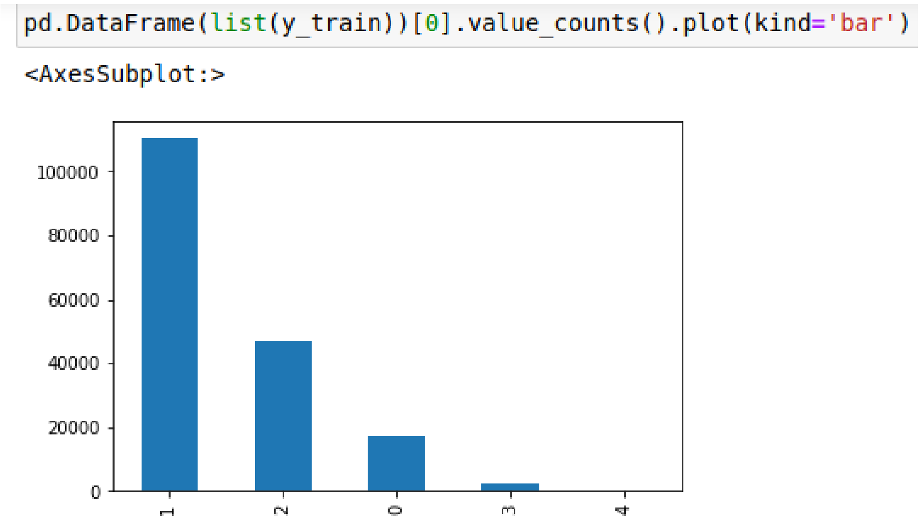
We have found the best attributes to build our model. We have standardized and normalized the data. Then the dataset is splitted into training and testing set (80:20). However, the data set is highly imbalanced. We need to balance it. SMOTE library is used for balancing the imbalanced training set. It is a method of over-sampling. It generates data in a way that resembles the underlying distribution of the real data. The following models were built and compared accuracy, f1-score, precision and recall of each models to find the best model.

1. K-Nearest Neighbor
2. Support Vector Machine
3. Logistic Regression

The train data before and after balancing using SMOTE are given in the figure:

4.1 KNN

The KNN algorithm assumes that similar things exist in close proximity. In other words, similar things are near to each other. We've built a KNN model



```
pd.DataFrame(list(y_train))[0].value_counts().plot(kind='bar')
```

```
<AxesSubplot:>
```

Figure 1: train data before balancing

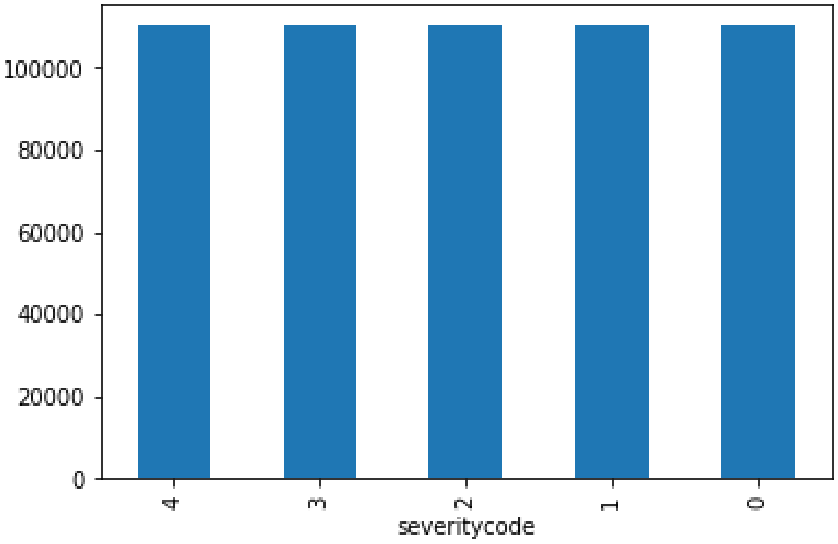
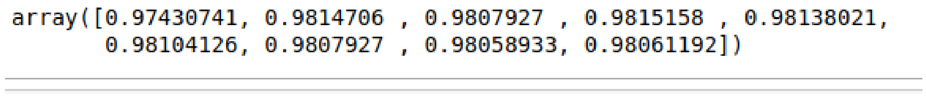


Figure 2: train data after balancing

with the upsampled training data. The most important task is to select the best K, as it give the best accuracy. The best value for k is 4.



```
array([0.97430741, 0.9814706 , 0.9807927 , 0.9815158 , 0.98138021,
       0.98104126, 0.9807927 , 0.98058933, 0.98061192])
```

4.2 Support Vector Machine

The objective of the support vector machine algorithm is to find a hyperplane in an N-dimensional space(N — the number of features) that distinctly classifies the data points. The SVM model which we've built also gives a better accuracy of 0.9823

4.3 Logistic Regression

Like all regression analyses, the logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables. THe logistic regression we've built using the upsampled training set, gives the best result than other two models. It gives an accuracy of 0.98316.

As we have built the models, we need to evalute the performance metrics of each model to finalize the best model for building the road accident severity prediction model.

5 Evaluation

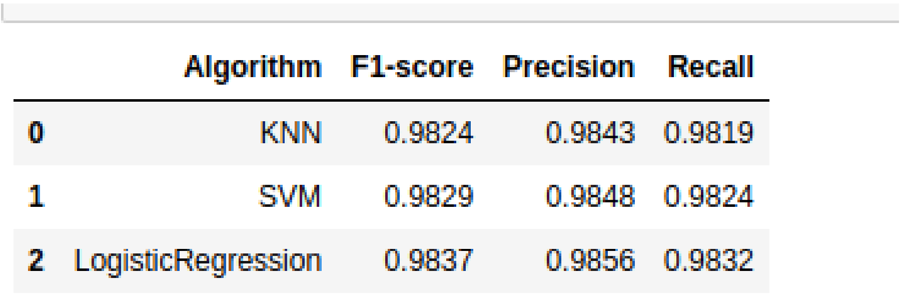
Evaluating a model is a core part of building an effective machine learning model. The evaluation metrics we are using for our models are:

- F1-score - It is the harmonic mean of precision and recall values for a classification problem.

- Precision and Recall - Precision-Recall is a useful measure of success of prediction when the classes are very imbalanced. In information retrieval, precision is a measure of result relevancy, while recall is a measure of how many truly relevant results are returned.

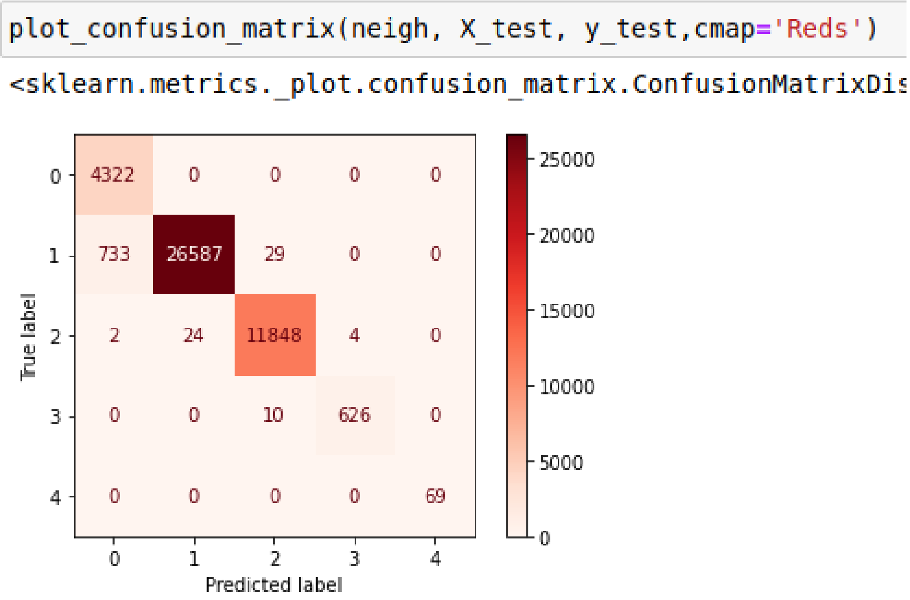
6 Result

The figure shows the result of evaluation in terms of performance metric of each models. It is clear that Logistic Regression model gives the higher F1-score, Precision and Recall, than other model, even though their performance are not bad.



	Algorithm	F1-score	Precision	Recall
0	KNN	0.9824	0.9843	0.9819
1	SVM	0.9829	0.9848	0.9824
2	LogisticRegression	0.9837	0.9856	0.9832

Figure 3: Evaluation result

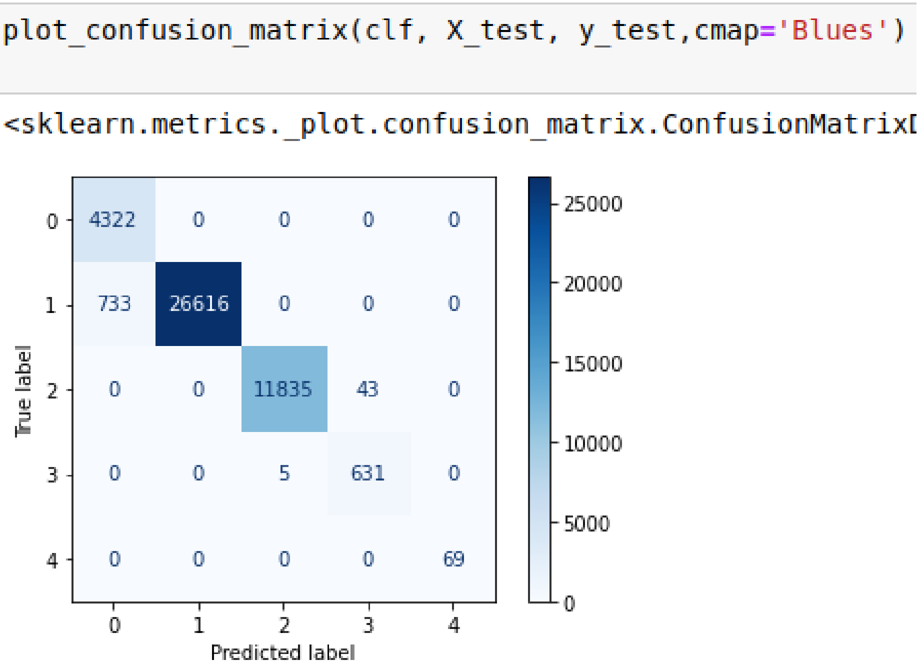


```
plot_confusion_matrix(neigh, X_test, y_test,cmap='Reds')
```

```
<sklearn.metrics._plot.confusion_matrix.ConfusionMatrixDis
```

Figure 4: Confusion matrix of KNN

Figure 5: Confusion matrix of SVM

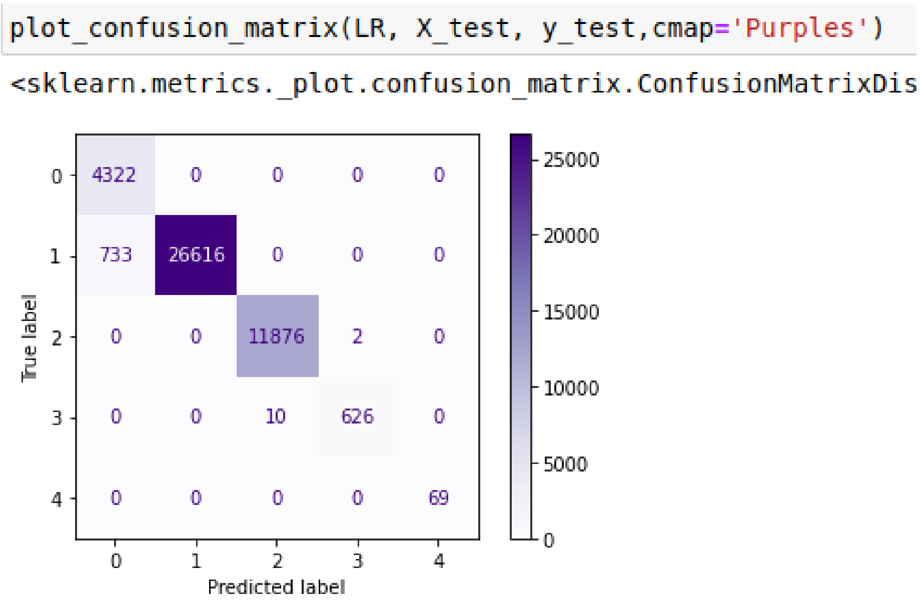


```
plot_confusion_matrix(clf, X_test, y_test,cmap='Blues')
```

```
<sklearn.metrics._plot.confusion_matrix.ConfusionMatrixD
```

	0	1	2	3	4
0	4322	0	0	0	0
1	733	26616	0	0	0
2	0	0	11835	43	0
3	0	0	5	631	0
4	0	0	0	0	69

We've evaluated the models using confusion matrix. The result shows that out of the 3 models we see that logistic regression model gives a more better f1-score, Precision and Recall as compared to the other 2 models. Also the KNN and SVM model also gives a good result. From the confusion matrix we can see that Lr and KNN model wrongly predicts 10 samples as label 2 instead of 3, where as in SVM it is only 5 . The misclassification is less in SVM as compared to KNN for true label 2. Misclassification is more in KNN considering all the labels.



```
plot_confusion_matrix(LR, X_test, y_test,cmap='Purples')
```

```
<sklearn.metrics._plot.confusion_matrix.ConfusionMatrixDis
```

Figure 6: Confusion matrix of LR

7 CONCLUSION

We have developed a road accident severity prediction model using collision dataset provided by Seattle Department of Transportation. Although the dataset contains 40 attribute , we've considered only 11 best attributes, and the target as severity code. It can be infered from the analysis that the consumption of alcohol, drugs, the location, the weather have an impact on accidents occured. The model can be used for predicting the possibility of accident, and alerting the public about the severity, which will help to be cautious and therbey reducing collisions.