Coursera IBM Data Science Capstone Project

Accident Severity Prediction Report

Ivan Salcedo Velazco

https://github.com/ivansalcedo/IBM_Coursera_Capstone_Project

1. Introduction

Car accidents are one of the leading causes of death in the world, especially among people between the ages of 20 and 45. Studying the factors involved in past accidents provided you can make an accurate prediction of the severity of future accidents.

These predictions could be used by emergency services, hospitals to have emergency personnel and material available to adequately attend. It can also be used by financial services such as auto and medical insurance to analyze their demand for future services and establish marketing strategies.

2. Data

In this project we will use the data of accidents that occurred in the city of Seattle that include automobile, bicycle and motorcycle accidents. Only 10,000 records will be selected out of the 194,674 randomly, as some algorithms take time and are not suitable for such a large data set.

The features selected for this project are :

```
0   SEVERITYCODE
1   ADDRTYPE
2   COLLISIONTYPE
3   PERSONCOUNT
4   VEHCOUNT
5   WEATHER
6   ROADCOND
```

After of deleting the records with incorrect data, they were taken randomly 10,000 records with the following severity data :

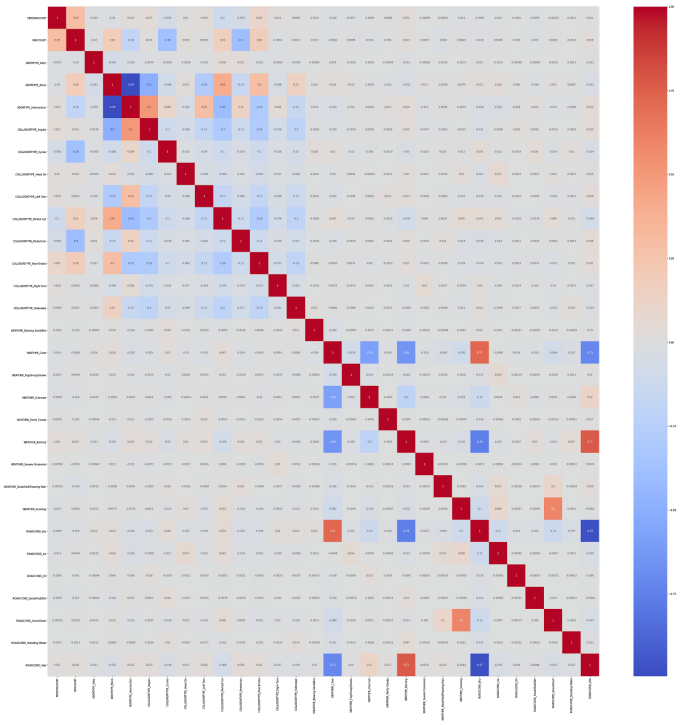
Severity Code	Count
1	6659
2	3341

The fields 'ADDRTYPE', 'COLLISIONTYPE', 'WEATHER', 'ROADCOND' was converted from categorical to numeric futures

3. Prediction Model

Before creating the models, were created the training and test dataset with percentage of 70% for training and 30% for test data set. Too, the selected characteristics must be normalized for performance in the algorithms.

With the data was generated the correlation study and following graph was constructed :

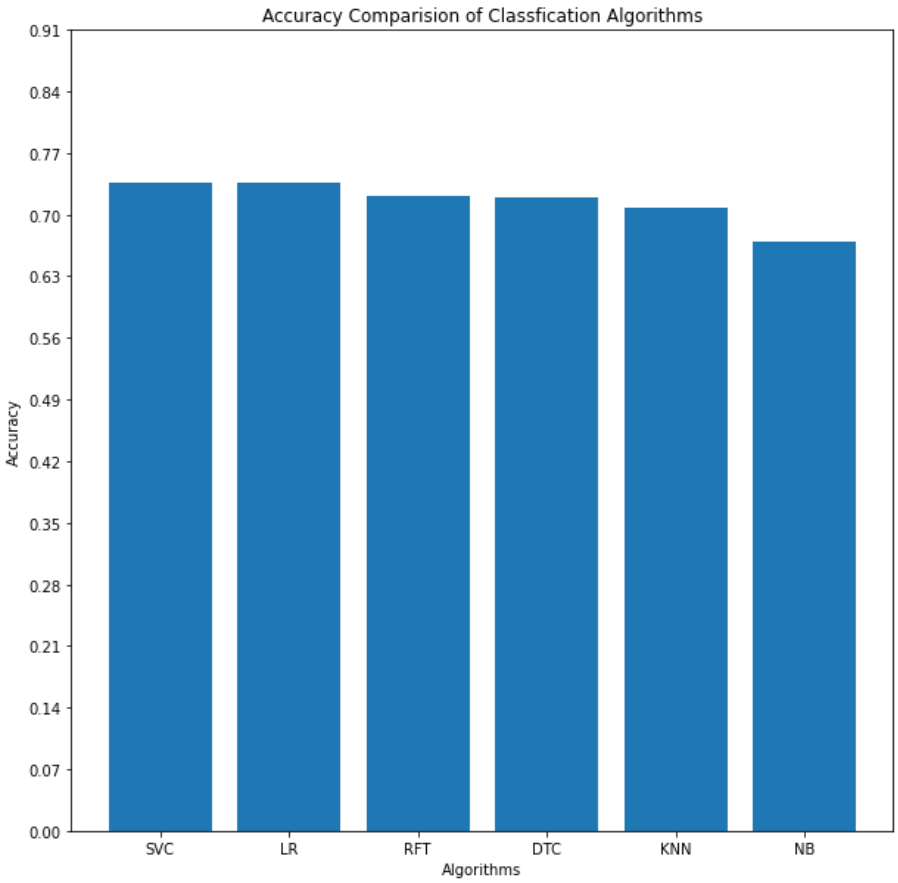


For accident prediction studied the following classification algorithms with the next accuracy

Algorithm	Accuracy
Logistic Regression	0.7360
Naive Bayes	0.6690
K Nearest Neighbours	0.7083
Decision Tree	0.7186
Random Forest Tree	0.7213
Support Vector Machine	0.7363

4. Results

For show the result of the accuracy of the algorithms can observed the next diagram.



Kernel Support Vector Machine and Logistic Regression are the best classifier based on accuracy of 73% while Random Forest and Decision Tree show a slightly lower performance with 72 and 72% respectively.

5. Conclusion

The models generated will not provide the desired predictions with these results. The maximum they provide is 75% and at least 85% or higher would be appropriate. This can be accomplished by selecting more data, searching for more scenarios, or experimenting with more variables.