Joshua Osko
Coursera Datascience Capstone
7 September 2020

Predicting Car Accident Severity Based on Traffic Conditions

Introduction

Automobile accidents are a responsible for over $800 billion [1] and over 30,000 casualties [2] in the United States each year. Identifying factors which most contribute to fatal accidents could help political officials better formulate policies which could reduce loss of life and save billions of dollars. In this paper, we will build a predictive supervised machine learning model to predict the potential severity of an automobile accident based on existing traffic conditions.

Data

For this paper, we will examine the data included in the example dataset provided by the Coursera Applied Data Science Capstone course. This dataset contains 194,673 accident observations and records 38 attributes for each accident.

Most of these columns are things which are clearly not causal to the accident and thus we will not consider as impactful in determining whether or not an automobile accident will be fatal. Here is the list of attributes to remove: X, Y, OBJECTID, INCKEY, COLDETKEY, REPORTNO, LOCATION, STATUS, INTKEY, EXCEPTRSNCODE, EXCEPTRSNDESC, SEVERITYCODE.1, SEVERITYDESC, COLLISIONTYPE, PERSONCOUNT, PEDCOUNT, PEDCYLCOUNT, VEHCOUNT', INCDATE, INCDTTM, SDOT_COLCODE, SDOT_COLDESC, SDOTCOLNUM, ST_COLCODE, ST_COLDESC, SEGLANEKEY, CROSSWALKKEY, HITPARKEDCAR, PEDROWNOTGRNT.

We will look to build a feature set useful for accurately predicting how severe an automobile accident will be given existing traffic conditions. The following features will be used to predict automobile accident severity:

Feature	Description
JUNCTIONTYPE	The type of junction where the accident occurred
INATTENTIONIND	Indicates cause of accident related to inattention
UNDERINFL	Whether or not the driver was under the influence
WEATHER	The weather conditions when the accident occurred
ROADCOND	The road conditions when the accident occurred
LIGHTCOND	The lighting conditions when the accident occurred
SPEEDING	Whether or not speeding was a cause of the accident

Table 1: Breakdown of the features

The target column SEVERITYCODE will be predicted based on these predictor variables.

The remaining attributes require a deal of cleaning to properly prepare it for training the model. All categorical variables will be converted into numerical integer values for model training. The following actions will be taken:

Feature	Action
JUNCTIONTYPE	Convert all values to a 0 if not intersection related or a 1 if intersection related
INATTENTIONIND	Convert all values to a 0 for cause of accident not due to inattention or a 1 if it was due to inattention
UNDERINFL	Convert all values into a 0 for not under the influence or a 1 for under the influence
WEATHER	Convert all values into a 0 for bad weather or a 1 for good weather
ROADCOND	Convert all values into a 0 for bad road conditions or 1 for good road conditions
LIGHTCOND	Convert all values into a 0 for poor visibility, 1 for moderate visibility, or 2 for good visibility
SPEEDING	Convert all values into a binary 1 for speeding or 0 for not speeding

Table 2: Data preparation actions

Following these preprocessing steps, the remaining rows with empty cells will be removed as we will not make any further assumptions of the missing data due to uncertainty and risk to disrupting desirable model results. To keep processing more pleasant, all attributes were finally converted into integers.

Methodology

Since we already have labeled data, we will leverage unsupervised learning to let the model work on its own to discover information and draw conclusions from unlabeled data for a generalized solution.

Exploratory data analysis to identify what characteristics have the most impact on the target of accident severity. Looking at correlation to the severity code yielded the following results:

```
WEATHER                -0.007544
ROADCOND               -0.003456
SPEEDING                0.026075
INATTENTIONIND          0.028435
UNDERINFL               0.031026
LIGHTCOND               0.033527
Intersection Related    0.166014
SEVERITYCODE            1.000000
```

Surprisingly weather and road conditions didn’t appear to be large factors in the severity of automobile accidents. This result seems to indicate a lot of the severity has to do with intersections and hitting other automobiles at angles.

Joshua Osko
Coursera Datascience Capstone
7 September 2020

To train our models the data was split into 30% test set and 70% training set. Since the attribute data is effectively categorical and not numerical, we will not use tools such as linear regression. The following models were considered for evaluation: K-Nearest Neighbors (KNN), Support Vector Machines, Decision Trees, and Logistic Regression.

To build a decision tree classifier, the approach was taken to iterate over different depths to see which one had the highest accuracy while minimizing entropy. This resulted in the following decision tree:

Figure 1: Decision Tree Classifier

After training and validating against the test data, this approach was found to have a 67.2% accuracy.

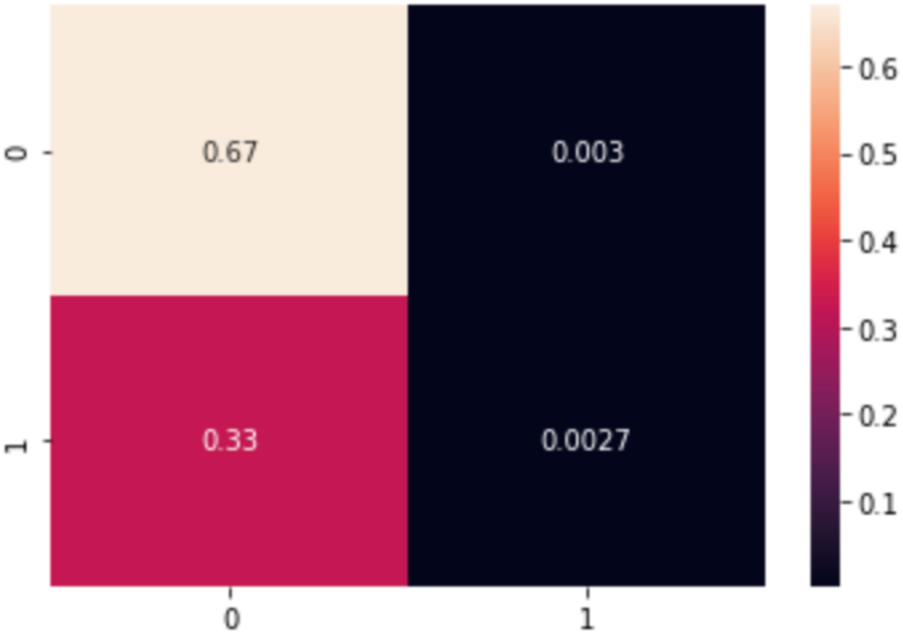

Figure 2: Confusion Matrix for Decision Tree Classifier

Joshua Osko
Coursera Datascience Capstone
7 September 2020

Results

To evaluate our models we looked at the Jaccard similarity score and The F1-Score. The Jaccard similarity score provides an index of the similarity between the test values and predicted values – the higher the score the more accurate the prediction model. The F1-Score comes from the confusion matrix and calculates the precision and recall of each class.

Building the models it was determined that the best depth for the decision tree classifier was 5 for an accuracy of 67.2%. The best K for the KNN model was 2 for an accuracy of 66.8%. The best kernel for SVMs was polynomial with an accuracy of 67.2%

The accuracy of the four models built are as follows:

Algorithm	Jaccard	F1-Score
Decision Tree	0.671	0.545
KNN	0.665	0.551
Logistic Regression	0.671	0.545
SVM	0.672	0.542

Table 3: Model Evaluation

The model accuracies are all very close, with none clearly outshining the others. As such we went with the Decision tree because it narrowly had the highest cumulative scores between its Jaccard similarly score and the F1-Score. So the decision tree classifier was chosen as the best predictor model.

Discussion

The highest accuracy any of the models produced was 67.2%, which is better than chance but still not very reliable in the general sense. Attempts were made to see if the accuracy of these models could be improved. Some options explored included eliminating some of the attributes which had a smaller correlation. However, no changes made seemed to significantly affect the change, leading me to believe there's just too many factors that could potentially be taken into account which could affect how severe an accident really could be. It is possible other information could be beneficial, such as precise time of day to see if rush hour has an impact. However, based on the results with a two out of three chance of being correct and seeing as intersections are the largest contributor, I think it's safe to say improvements can be made with regards to them. It would be valuable to know at these intersections for instance if there was a stop sign, or a traffic light, or what the speed limit was. Having more data can help us narrow down remedies to alleviate the scale of these accidents.

Joshua Osko
Coursera Datascience Capstone
7 September 2020

Conclusion

Cities can benefit greatly from leveraging machine learning to drive business decisions with regards to infrastructure plans that would reduce the number of traffic accidents and fatalities; however, more variables need to be weighed in than were obtained from this dataset to build a more accurate prediction model. From these results it seems indicative that intersections are a large cause of fatal traffic accidents. Perhaps improving transportation infrastructure by eliminating intersections and replacing them with roundabouts could help ease traffic congestion and limit the number of high fatalities that result as a consequence of automobile accidents.

References:

[1] https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812013
[2] https://www-fars.nhtsa.dot.gov/Main/index.aspx