Capstone Project 1 - In Depth Machine Learning Analysis

Overall Approach

Machine learning algorithms were employed in order to produce a more balanced sample set and predictions.

1. **Preprocess and Fitting Model to the data:** The full data set had a drastically higher number of automobiles not in accidents compared to those in accidents. Adjustments in what was sampled was in order for the model to work without being biased due to this. There was a random sample taken from the total automobiles not in accidents to match the sample size of each color of automobiles in accidents, which lead to a balanced overall sample set.
2. **Evaluate the Model:** Scikit learn was employed to build models with the given data. Logistic Regression and Decision Tree methods were used. Logistic Regression is used as the predictive analysis which looks to explain the relationship between the color of an automobile and likelihood to be in an accident. The Decision Tree is used to place a predicted value for each color of automobile.

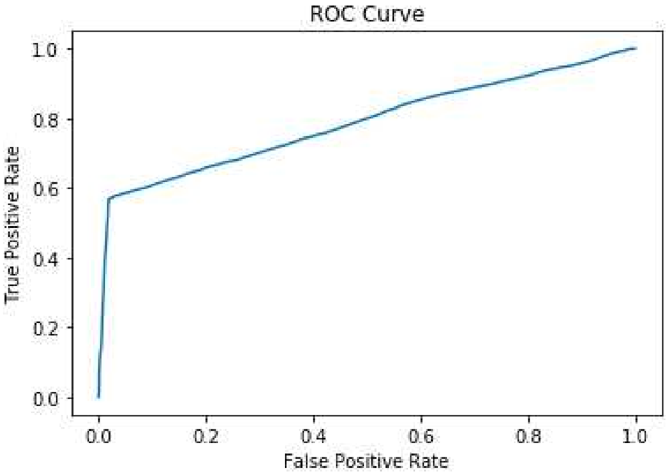
Regression

The sampling was adjusted to only include a total of 34,796 automobiles in accidents, and 34,796 automobiles not in accidents. 75% of the data was being used for the model training and 25% used for model testing. When running a logistic regression matrix, the resulting array came in with approximately a 77% accuracy rating.

```
array([[8396,  170],
       [3815, 5017]])
```

Classification

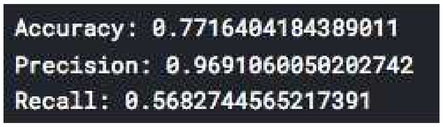
The ROC curve shows this model is doing better than a random model. It summarizes the performance plotting the True Positive Rate on the y-axis against the False Positive Rate on the x-axis.



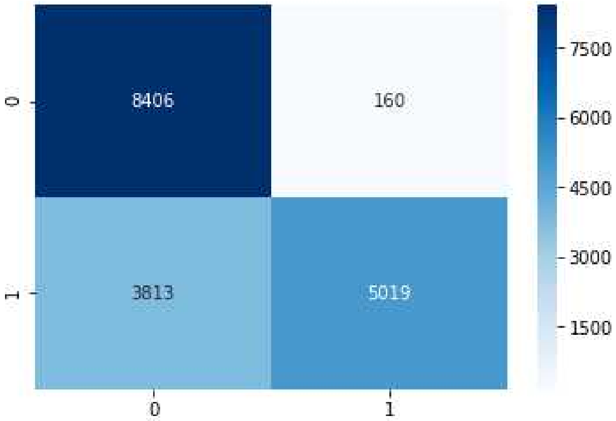
It would be better to understand this model's performance with a Confusion Matrix, shown below.

Confusion Matrix Plots

Visualizing the matrix as a heat map lot. The classification rate is around 77%, which is a good accuracy. The precision, or how often the model is correct, predicts that automobiles will or will not get in accidents 95.9% of the time. If there are automobiles that will or will not get in accidents, the Logistic Regression model can identify it 57.7% of the time.



```
Accuracy: 0.7716404184389011
Precision: 0.9691060050202742
Recall: 0.5682744565217391
```

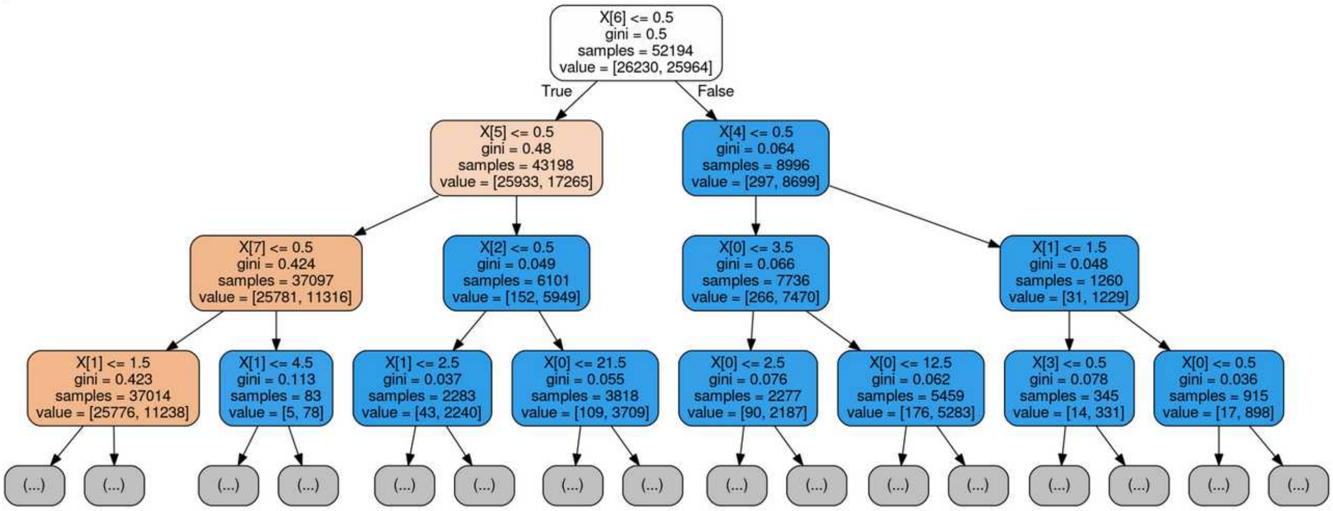


A Confusion Matrix is a performance measurement tool used in machine learning classification. It is useful for visualizing details of how well a classifier performs for one with any number of classes greater than 2. There are two classes here; “yes” the automobile will get into an accident or “no” the automobile will not get in an accident. The classifier made a total of 17,398 predictions. The prediction showed “yes” 5,179 times, and “no” 12,219. times. The true positive (TP), lower right corner, is when the prediction was to be an accident and there was an accident. The true negative (TN), upper left corner, is when the prediction was to not be an accident and there was not an accident. The false positive(FP), is a Type I error, where the prediction was to be an accident yet there was not. The false negative(FN), is a Type II error, where the prediction was for no accident to occur, and an accident did occur.

TN = 8406 FP= 160

FN = 3813 TP= 5019

A Decision Tree's root node(leaf) at the top is the most important feature for classification. The second level would be the second most features, the third level is the third most important, and finally the fourth is the fourth of importance.



X[0]= Color

X[1]= Race

X[2]= Gender

X[3]= Accident

X[4]= Alcohol

X[5]= Belts

X[6]= Personal Injury

X[7]= Property Damage

X[8]= Fatal

X[9]= Commercial License

X[10]= HAZMAT

X[11]= Commercial Vehicle

X[12]= Work Zone

In this instance, the Personal Injury appears to have the highest importance. That split then into Belts and Alcohol, and so on.

The Gini ratio measures the variance impurity of the node. Interesting to see that the nodes with the higher Gini ratio, also have the higher number of samples.

Conclusion

It was best to create balanced models for the predictions and drawing conclusions. This model had a good amount of accuracy and showed to be doing better than a random model. Using models such as this in other counties or statewide, could potentially give a little more insight into the likelihood of a particular automobile getting into an accident.

Future Work

1. Can accidents be predicted based on the day of the week? What about weekdays vs. weekends?
2. Can conclusions be made that driving a certain color car on a certain day of the week is more likely to get in an accident?
3. Go more in depth with this study and categories using more complex machine learning models.