Predicting Automobile Accidents in Montgomery County

Problem Statement

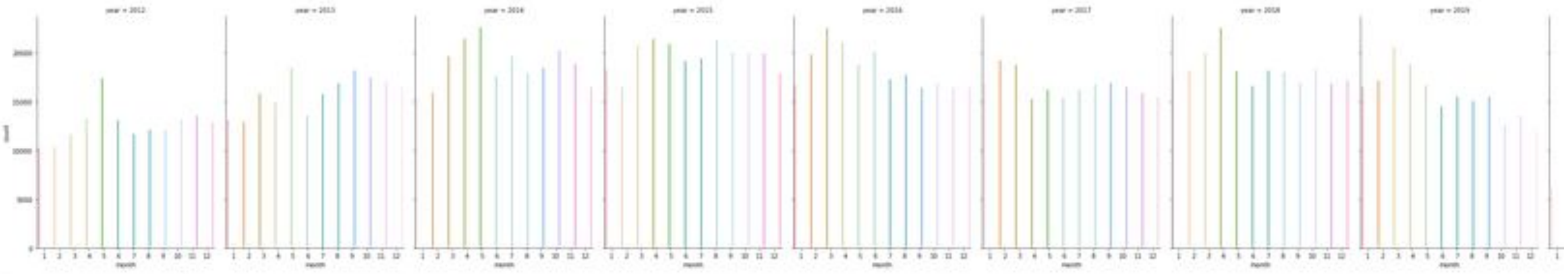
Automobile accidents are a part of society today. It is reasonable to believe if there is a way to predict what causes their increased likelihood, the overall society would benefit. The goal of this project is to look at connections in accident frequency for automobiles based on particular factors such the color of the automobile or the time of the year.

Exploratory Data Analysis

How does the time of year affect the number of accidents?

There is interest in looking over the total number of accidents reported in traffic stops each year broken down by month. A spike in 2017 and a low in 2012 is noted. Several months in 2012 are low compared to all months. September, October, and December of 2016 are significantly higher than most months recorded. The causes for this would require further investigation to see if this may have a correlation to being related to less vehicles on the road due to weather, the economy, or other factors. Later in the report, hypotheses are explored to include seeing if this can be predicted, leading to potentially reducing annual and monthly accidents.

	count							
Year	2012	2013	2014	2015	2016	2017	2018	2019
Month								
1	246.0	365.0	325.0	415.0	376.0	523.0	458.0	353.0
2	205.0	284.0	324.0	409.0	348.0	437.0	354.0	398.0
3	224.0	317.0	348.0	297.0	494.0	448.0	314.0	439.0
4	223.0	404.0	327.0	335.0	464.0	510.0	382.0	325.0
5	430.0	497.0	492.0	528.0	505.0	529.0	407.0	353.0
6	317.0	415.0	426.0	472.0	441.0	428.0	483.0	462.0
7	259.0	387.0	386.0	363.0	435.0	454.0	431.0	454.0
8	325.0	385.0	406.0	349.0	455.0	449.0	386.0	367.0
9	367.0	345.0	358.0	368.0	573.0	457.0	393.0	431.0
10	309.0	443.0	460.0	542.0	580.0	533.0	502.0	436.0
11	332.0	441.0	393.0	432.0	400.0	477.0	445.0	386.0
12	406.0	341.0	486.0	529.0	621.0	536.0	472.0	448.0



Can accidents be predicted based on the month?

Null Hypothesis: There is no statistical significance in the likelihood of an Automobile getting into an accident related to the month.

Alternative Hypothesis: Certain months of the year show a greater or reduced likelihood for an Automobile to get into an accident.

A Paired T-Test was run to discover if there is an association in the likelihood of an Automobile getting in an accident related to the month.

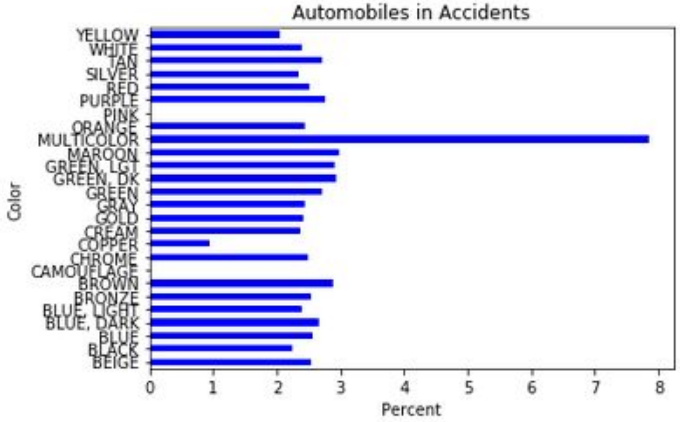
```
6.672627141580349e-69
```

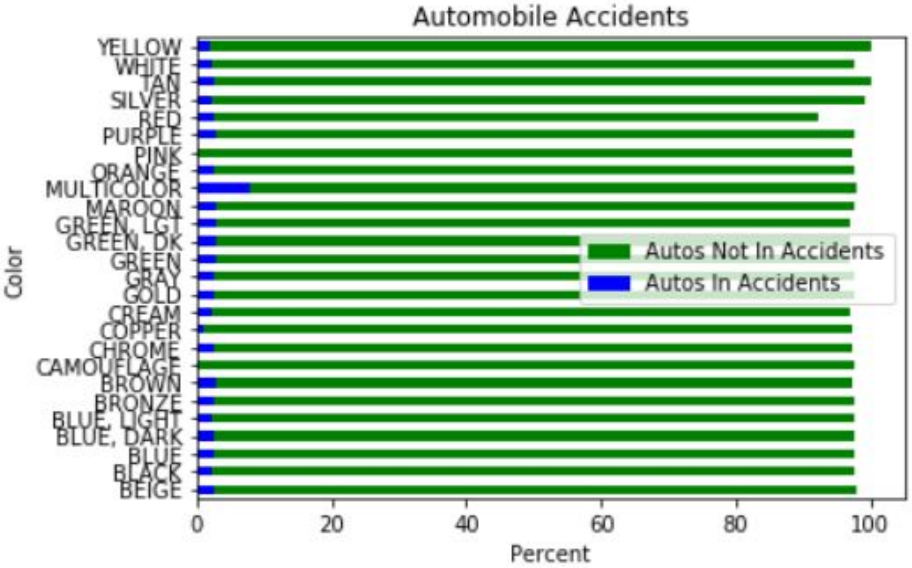
Comment:
The test shows to reject the null hypothesis. Therefore certain months of the year do show a greater or reduced likelihood for an Automobile to get in an accident.

Conclusion:
Certain months of the years showed a variance in accidents that could vary by more than 200 compared to other months in the same year. This merits further investigation by a client or company who wishes to use this information. The weather may have played a factor or perhaps the economy.

Does the color of the automobile affect accident probability?

The percentages of Automobile Accidents categorized by color, as reported in traffic stops over a period of five years from 2012-2019. It is unknown what automobiles are placed into the multicolor category that appears to be quite an outlier. It is interesting to see that pink and camouflage had no reported accidents.





Percent of Each Color in an Accident

BEIGE	2.527184
BLACK	2.240856
BLUE	2.552388
BLUE, DARK	2.649421
BLUE, LIGHT	2.385771
BRONZE	2.547971
BROWN	2.881844
CAMOUFLAGE	NaN
CHROME	2.500000
COPPER	0.930233
CREAM	2.366864
GOLD	2.408658
GRAY	2.443378
GREEN	2.720941
GREEN, DK	2.930122
GREEN, LGT	2.898736
MAROON	2.984191
MULTICOLOR	7.843137
ORANGE	2.442415
PINK	NaN
PURPLE	2.758107
RED	2.517158
SILVER	2.334393
TAN	2.708199
WHITE	2.392629
YELLOW	2.054795

Is there a connection between certain colors of automobiles being in more accidents due to their color?

Null Hypothesis: There is no statistical significance in the likelihood of an Automobile getting into an accident related to color.

Alternative Hypothesis: **C**ertain colors of Automobiles show a higher likelihood for getting in an accident.

A Chi-Square Test was run to discover if there is a significant association between the color of an automobile and its likelihood for being in an accident.

The outcomes were:

- Significance level: 0.05
- Degree of Freedom: 1
- chi-square statistic: 488.4484795288744
- critical_value: 3.841458820694124
- p-value: 0.0

Comment:

Therefore the Null Hypothesis is rejected. Accepting the Alternative Hypothesis that certain colors show a greater or reduced likelihood for an Automobile to get into an accident. (Detailed testing can be found in Capstone_1_ Data_Story notebook)

Conclusion:

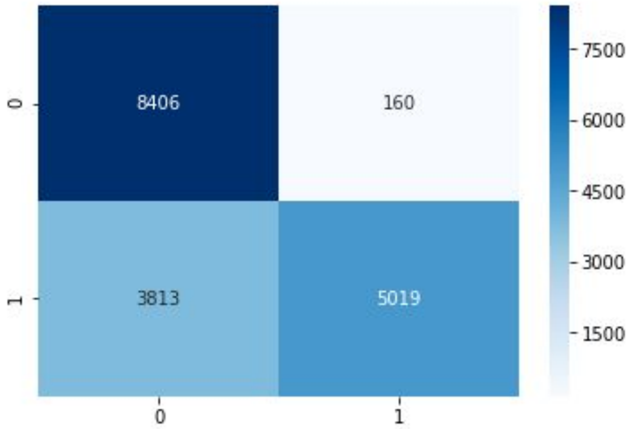
Certain colors of vehicles have a slightly higher risk for being in an accident when looking at the percentages related to each individual color. The risks overall vary by <1%, and therefore it doesn't seem a strong statement to make for clients to base decisions from.

Machine Learning - Supervised

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



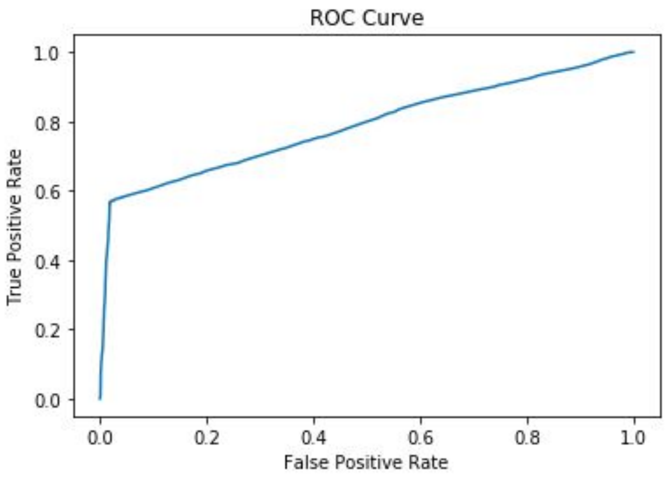
Visualizing the matrix as a heat map lot. The classification rate is around 77%, which is a good accuracy. The precision, or how often the model is correct, predicts that automobiles will or will not get in accidents 95.9% of the time. If there are automobiles that will or will not get in accidents,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

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

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]])
```

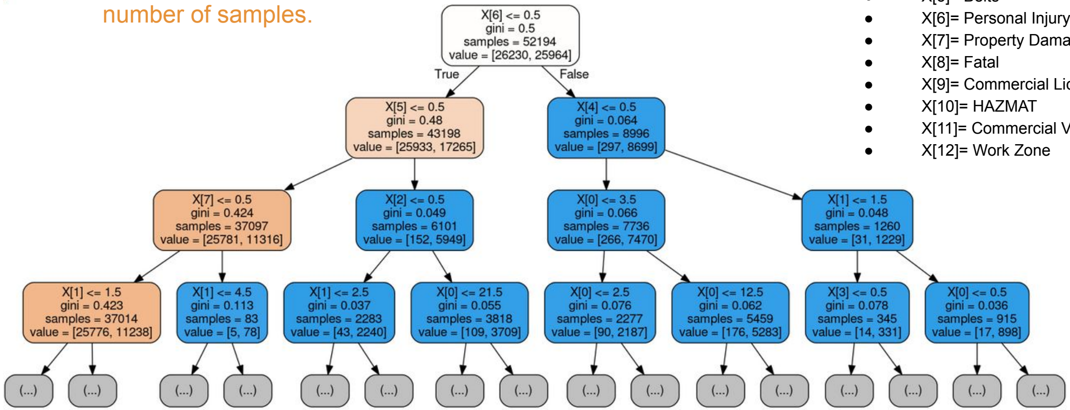
The ROC curve shows this model is doing better than a random model. It summarizes the performance plotting the True Positive Rate on the y-axis against the False Positive Rate on the x-axis.



The decision tree below is a type of supervised machine learning used to make conclusions about a target variable. The inputs are listed in the column.

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.



- 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

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.