Consolidated Report Capstone 1 - Traffic Violations

Hypothesis: Accident frequency can be predicted for automobiles based on their color and the month of the year.

Data Wrangling

Datasets: The dataset contains all electronic traffic violations issued in Montgomery County, Maryland. This is a freely accessed public dataset at https://data.montgomerycountymd.gov/Public-Safety/Traffic-Violations/4mse-ku6q.

Wrangling Steps Performed:

Traffic Accidents

This will include looking to see if particular colors of vehicles and timeframes of the year can be predictors for the probability of an automobile getting in an accident.

1. A csv file was downloaded and read into a normalized pandas dataframe, along with parsing dates so that data can be utilized for time series portions of the study.

2. This was also an opportunity to explore which columns to use for this particular project. At this stage in the game it is Vehicle Type, Color, and Accident.

3. Further investigation into the Vehicle Type column showed the counts for the types of vehicles. This led to a focus on only Automobiles because this had the most value and well over one million items. This set would appeal to the clientele of consumers, insurance companies, and car dealerships.

4. The data was checked and scanned for any null information present within the colors listed for automobiles. Those Automobile rows with nan for color were dropped. The data was rechecked to be sure the new dataset was clean, and all Automobiles were identified in a color category.

5. Two additional dataframes were created to have one with Automobile stops marked as being in an accident and the other were those not in an accident. This was the counts for each color were more apparent and easily viewed.

Automobile Data

A new dataframe was created of only Automobiles, with rows labeled by Color. There are total counts for the number of each color in total, the number in accidents, and the number not in accidents. Rather than continuously updating the study by query of the API, this set used contains information for all of the years 2012-2019. It was downloaded and read as the csv file and that allows for quicker retrieval in the future.

Features to be used: Vehicle Type, Color, Accident, Date of Stop, Alcohol

Dealing with Outliers

The outliers for more popular colors of Automobiles stood out. Those are not eliminated from the dataset because that may be of interest once viewing the likelihood of each color along with the other variables.

Data Story

Can accident frequency be predicted for automobiles based on particular factors? In this capstone project the data from Maryland's Montgomery county traffic stop database is used to look at variables that could potentially help predict increased accident likelihood.

-Can accidents be predicted based on the month?

-Is there a connection between certain colors of automobiles being in more accidents due to their color, or is it more a popularity of that color leading to more of them to have more accidents?

-Can a recommendation be made on certain colors being safer automobiles?

-What caused the year 2017 to have more accidents?

-Did alcohol affect the amount of accidents with any significance?

Future questions - Can accidents be predicted based on the day of the week (weekday, weekend)?

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 vs other car colors?

Question: Can missing data be identified and can we see where there may be missing values?

Answer: 1,632,871 is the total number of non-null objects recorded. Any category without this full amount is missing information.

Question: What vehicle types are listed in the data set? What would be a good representation for a general population to sample?

Answer: There are over one million automobiles in this data set. This would be primarily the focus for most consumers and the clients identified for this study.

Question: How many of each color automobile exists in the data set?

BLACK	302920	TAN	29503	BRONZE	3179
SILVER	277374	MAROON	23658	PURPLE	2683
WHITE	204921	BLUE, LIGHT	18778	CREAM	845
GRAY	170297	BEIGE	16738	MULTICOLOR	510
RED	112190	GREEN, DK	14368	COPPER	430
BLUE	107037	GREEN, LGT	7831	PINK	192
GREEN	50681	BROWN	6246	CHROME	40
GOLD	45461	ORANGE	5036	CAMOUFLAGE	25
BLUE, DARK	30384	YELLOW	4234		

Question: When comparing the percentage of the represented colors for all vehicles to represented colors for only automobiles, are they similar enough to give an accurate assessment for what this study is looking at to use only the automobiles as the sample dataset?

Answer: The percentage of the represented colors for all vehicles vs. represented colors for only automobiles have similarity to give an accurate assessment for what this study is looking at to use only the automobiles as the sample dataset. For example Black is 20.75% and 21.10%, Gray is 11.53% and 11.86%.

Question: What is noticed in the number of automobiles in accidents compared to those not in accidents?

Answer: There is quite a difference between the number of automobiles in accidents versus those not. Also interesting to note how many more vehicles of certain colors exist in this sample and then consider if this affects the outcomes of those having higher accident rates.

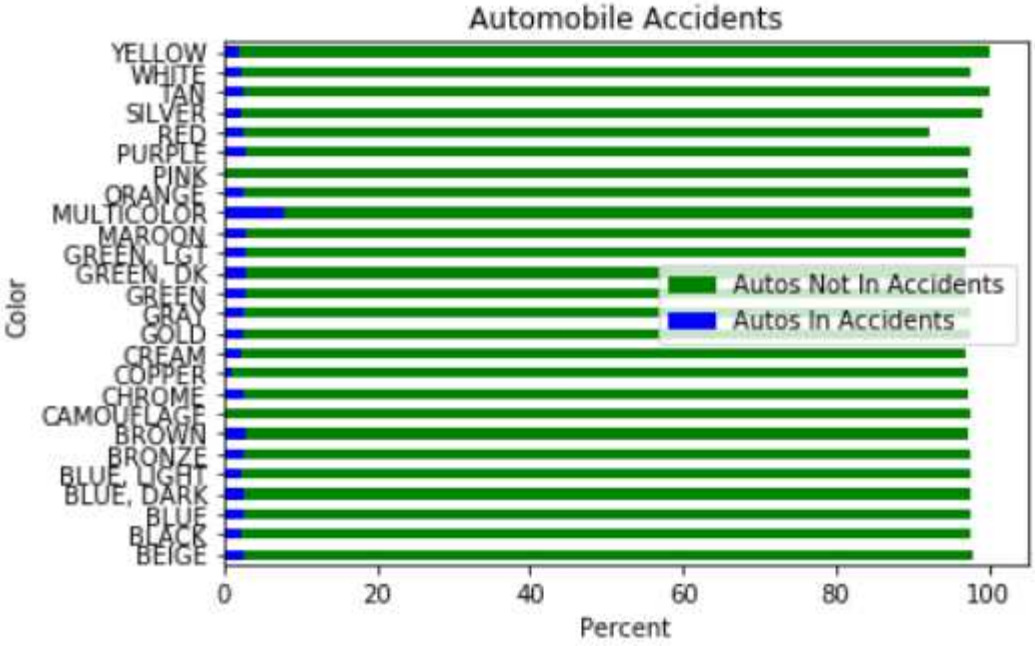
Question: Does the percentage of those more popular colors show a higher likelihood to be in an accident?

The chosen dataset of only the Automobiles from the database of vehicles that were stopped in Montgomery county. It is not surprising that the most popular colors are black, silver, white, gray, red, and blue. It is a consideration if drawing a conclusion that black cars get in more accidents than brown for example. Therefore, there was a need to look at the percentage of accidents of a color, out of the total vehicles of only that color.

Answer: According to the results, COPPER is the least likely to have been in an accident through the span in this sample, with multicolor being the most likely. This raises a flag to ask questions; What colors are in the category of MULTICOLOR? Why might COPPER be such an outlier?

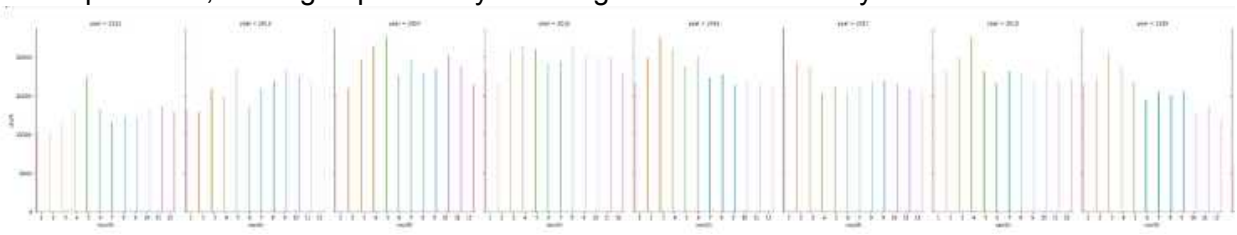
Question: When looking at automobile stops broken by color, are more vehicles in accidents compared to those not in accidents across the board or vice versa?

Answer: Visually looking at the difference between those in accidents vs. not in accidents tells overall most vehicle stops are not accident related across all colors.



Question: This is where we begin to see if not only color can help predict an automobiles chance of being in an accident. Are certain months more prone to have accidents? Further looking to see if certain years had more accidents? Can a cause for this spike be identified and is it possible it will occur again? On the reverse, if certain year had less accidents, can this cause be identified so that perhaps we can reduce accidents annually?

Answer: There is a spike in 2017 and a low in 2012. Several months in 2012 are low compared to all months. September, October, and December of 2016 are significantly higher than most months recorded. Time to consider what causes this. Are there less vehicles on the road due to weather, the economy, politics, etc.? Now to identify if these are event that could occcur again and 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

Question: What was the distribution for the number of automobile stops that were considered accidents each day over the entire span?

Answer: Below shows the maximum number of accidents in one day through the years was 56. The minimum being 1. The average, mean, per day over all years was about 12 accidents.

count	2908.000000
mean	12.017538
std	8.065481
min	1.000000
25%	6.000000
50%	10.000000
75%	16.000000
max	56.000000

Question: What was the distribution for the number of automobile stops that were considered accidents each month over the entire span?

Answer: Below shows the maximum number of accidents in one year through the years 22,601. The minimum being 7,372. The average, mean, per year over all years was about 16,932 accidents.

	count
count	97.000000
mean	408.164948
std	83.727320
min	203.000000
25%	353.000000
50%	409.000000
75%	458.000000
max	621.000000

Question: Does alcohol appear to have any significance in traffic stops for automobiles in accidents?

Answer: The percentage of vehicles in accidents that involved alcohol vs. vehicles in accidents in each color. This is a small percentage.

BLACK	0.135073
BLUE	0.045982
GRAY	0.045982
TAN	0.040235
MAROON	0.037361
WHITE	0.020117
BLUE, LIGHT	0.002874
GOLD	0.002874

For example, 47 out of 6,788 black automobiles in accidents involved alcohol. There was a total of 302,920 black automobiles overall.

Statistical Data Analysis

After the data is wrangled and cleaned, and the preliminary EDA complete, the hypothesis testing begins. Questions arise through exploring the data.

-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:

Certain 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)

-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`
reject null hypothesis

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 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.

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.

In-Depth 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.

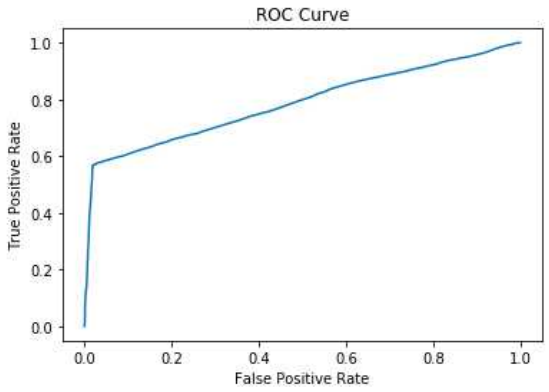
Regression

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

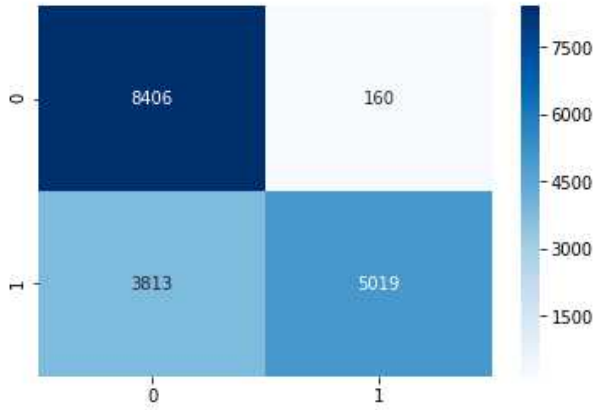
Visualizing the matrix as a heat map lot. The classification rate is around 77%, which is a good accuracy. The precision, or how often the model is correct, predicts that automobiles will or will not get in accidents 95.9% of the time. If there are automobiles that will or will not get in accidents,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

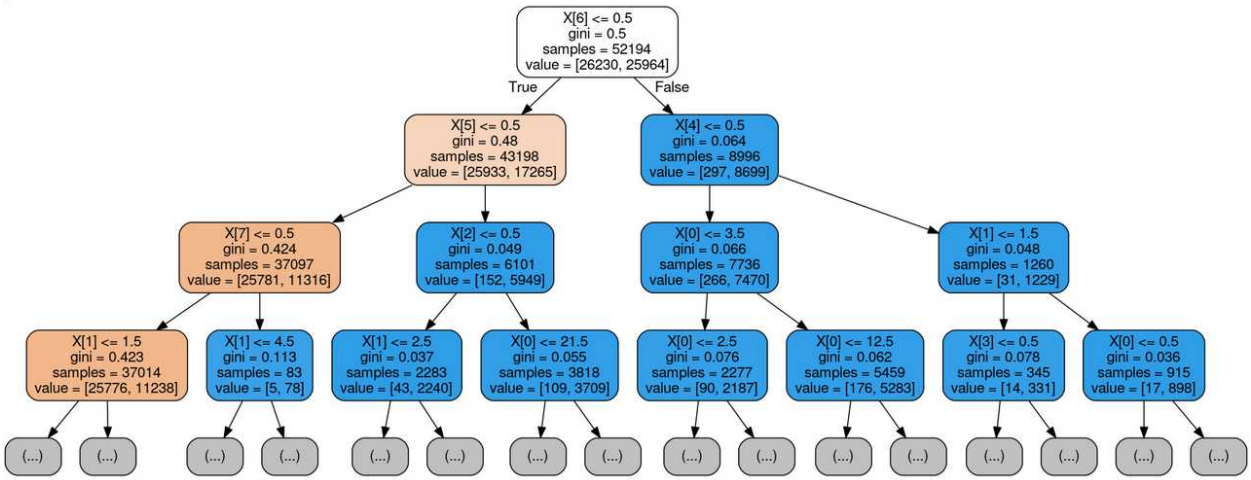
FN = 3813 TP= 5019



Decision Tree

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 importance, and finally the fourth is the fourth of importance. 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.



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.