EXECUTIVE SUMMARY REPORT

Final Project

Daksh Shah

Northeastern University

College of Professional Studies

ALY6015 – Intermediate Analytics

Prof – Ajit Appari

February 19, 2022

- Business Problem Statement
  The Traffic Department of Maryland county has requested to perform data analysis and draw insights from it.
  Using the data set provided, suggestions are requested to make the roads of Maryland safe and sound for the people.
  The Analysis should showcase the complete view of the violations in the county.

- Data Description
- Data can be access using Link.
  - Agency: Agency issuing the traffic violation. (Example: MCP is Montgomery County Police)
  - Alcohol: If the traffic violation included an alcohol related
  - Belts: If traffic violation involved a seat belt violation.
  - Contributed To Accident: If the traffic violation was a contributing factor in an accident.
  - Date Of Stop: Date of the traffic violation.
  - Fatal: If traffic violation involved a fatality.
  - Gender: Gender of the driver (F = Female, M = Male)
  - HAZMAT: If the traffic violation involved hazardous materials.
  - Latitude: Latitude location of the traffic violation.
  - Longitude: Longitude location of the traffic violation.
  - Personal Injury: If traffic violation involved Personal Injury.
  - Property Damage: If traffic violation involved Property Damage.
  - Race: Race of the driver. (Example: Asian, Black, White, Other, etc.)
  - Subagency: Court code representing the district of assignment of the officer. R15 = 1st district, Rockville B15 = 2nd district, Bethesda SS15 = 3rd district, Silver Spring WG15 = 4th district, Wheaton G15 = 5th district, Germantown M15 = 6th district, Gaithersburg / Montgomery Village HQ15 = Headquarters and Special Operations
  - Time Of Stop: Time of the traffic violation.
  - Vehicle Type: Type of vehicle (Examples: Automobile, Station Wagon, Heavy Duty Truck, etc.)
  - Violation Type: Violation type. (Examples: Warning, Citation, SERO)
  - Work Zone: If the traffic violation was in a work zone.

- We have utilized the date of stop field and divided it into the year, hour, months to have a better view of the data.

- **Understanding some important columns in the Data set. (Proportion are in %)**

1. Sub Agency

Sr No	Subagency	Proportion
1	1st district, Rockville	11.33
2	2nd district, Bethesda	15.7
3	3rd district, Silver Spring	20.05
4	4th district, Wheaton	25.58
5	5th district, Germantown	11.51
6	6th district, Gaithersburg / Montgomery Village	12.49
7	Headquarters and Special Operations	3.34

2. Generalized Vehicle type for analysis.

Sr No	Car Type	Proportion
1	Two - Wheeler	1.05
2	Four - Wheeler	88.09
3	SUV	1.6
4	Truck	6.18
5	Special Vehicles	2.88
6	Bus	0.06
7	Service Vehicle	0.01
8	Trailer	0.13

3. Some other categories

Sr No	Categorical	Yes Proportion	No Proportion
1	Belts Violation	3.32	96.68
2	Alcohol Violation	0.17	99.83
3	Hazmat Violation	0.01	99.99
4	Commercial Vehicle Violation	0.25	99.75
5	Work zone Violation	0.02	99.98
6	Personal Injury	1.25	98.75
7	Property Damage	2.05	97.95
8	Was Violation Fatal?	0.02	99.98

4. Gender

Sr No	Gender	Proportion
1	Male	67.44
2	Female	32.56

5. Race

Sr No	Race	Proportion
1	Asian	5.78
2	Black	31.92
3	Hispanic	21.76
4	Native American	0.22
5	Other	5.41
6	White	34.9

6. Utilizing Date column
    a. Quarterly distribution

Sr No	Quarter	Proportion
1	Q1	26.45
2	Q2	22.27
3	Q3	26.29
4	Q4	24.98

    b. Yearly distribution

Sr No	Years	Proportion
1	2012	5.93
2	2013	12.63
3	2014	12.93
4	2015	22.52
5	2016	20.89
6	2017	17.96
7	2018	7.14

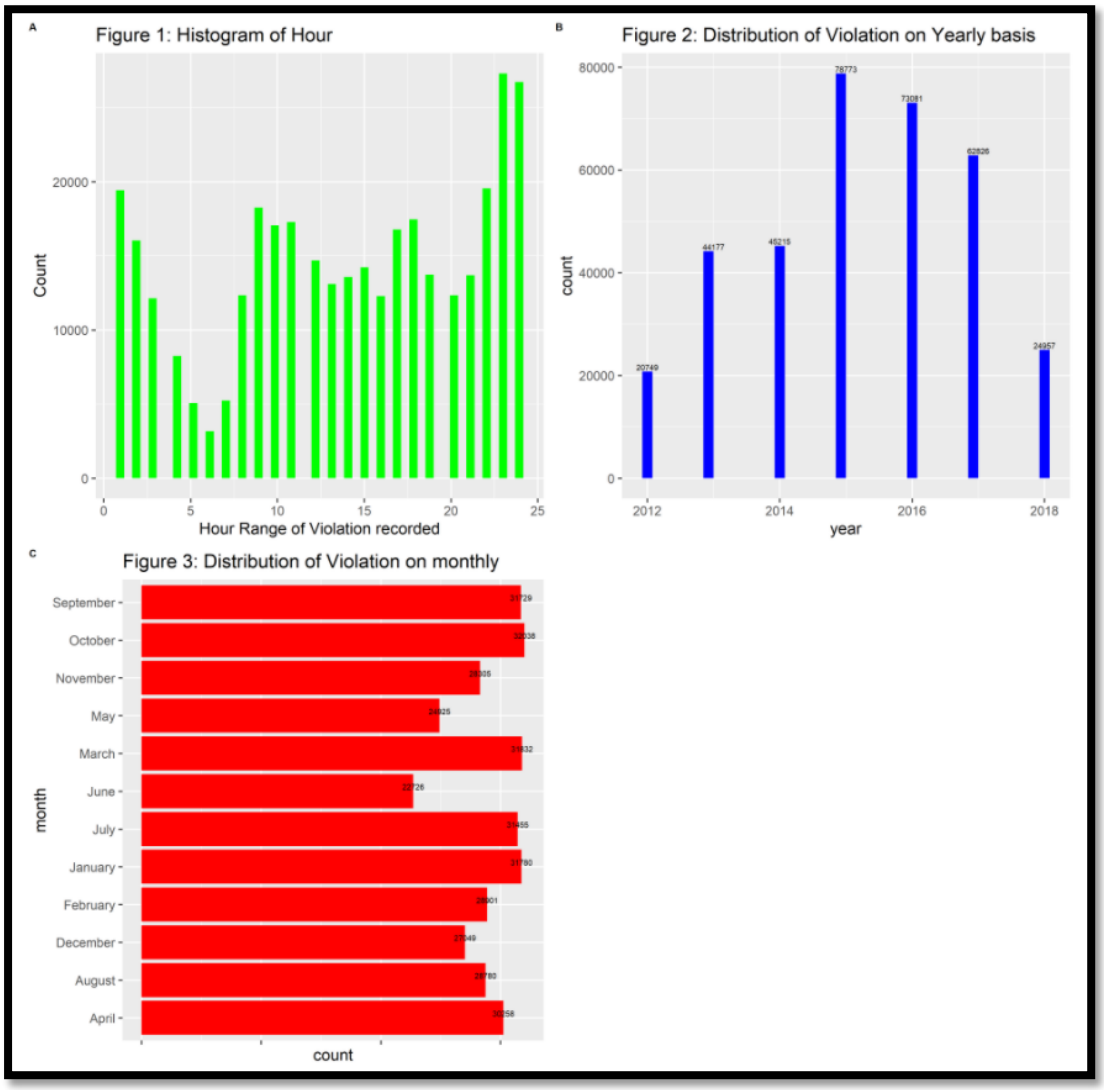
    c. Monthly distribution

Sr No	Months	Proportion
1	Jan	9.09
2	Feb	8.26
3	Mar	9.1
4	Apr	8.65
5	May	7.13
6	Jun	6.5
7	Jul	8.99
8	Aug	8.23
9	Sep	9.07
10	Oct	9.16
11	Nov	8.09
12	Dec	7.73

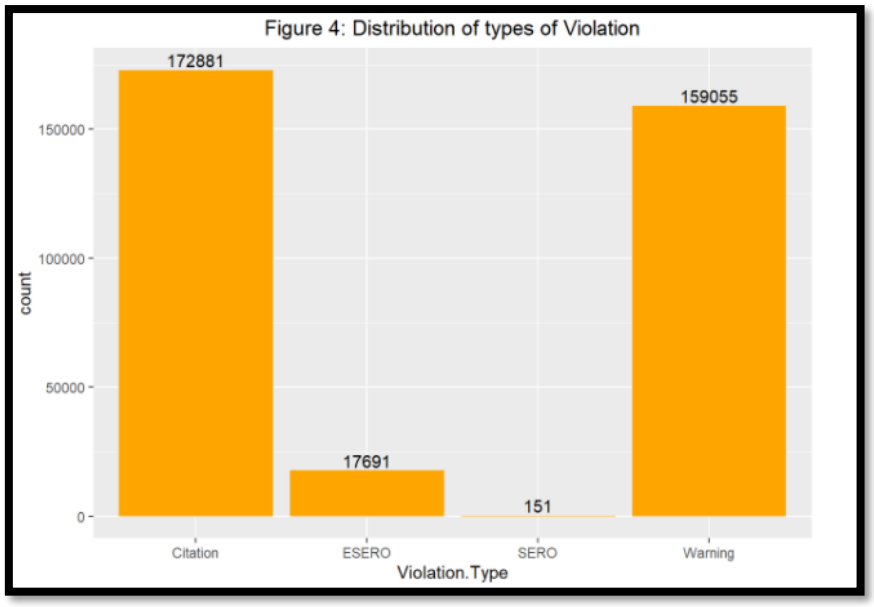
d. Hourly distribution

Sr No	Hour	Proportion
1	1	5.55
2	2	4.59
3	3	3.47
4	4	2.36
5	5	1.45
6	6	0.9
7	7	1.5
8	8	3.53
9	9	5.22
10	10	4.88
11	11	4.94
12	12	4.2
13	13	3.75
14	14	3.88
15	15	4.07
16	16	3.51
17	17	4.8
18	18	5
19	19	3.92
20	20	3.53
21	21	3.91
22	22	5.59
23	23	7.81
24	24	7.64

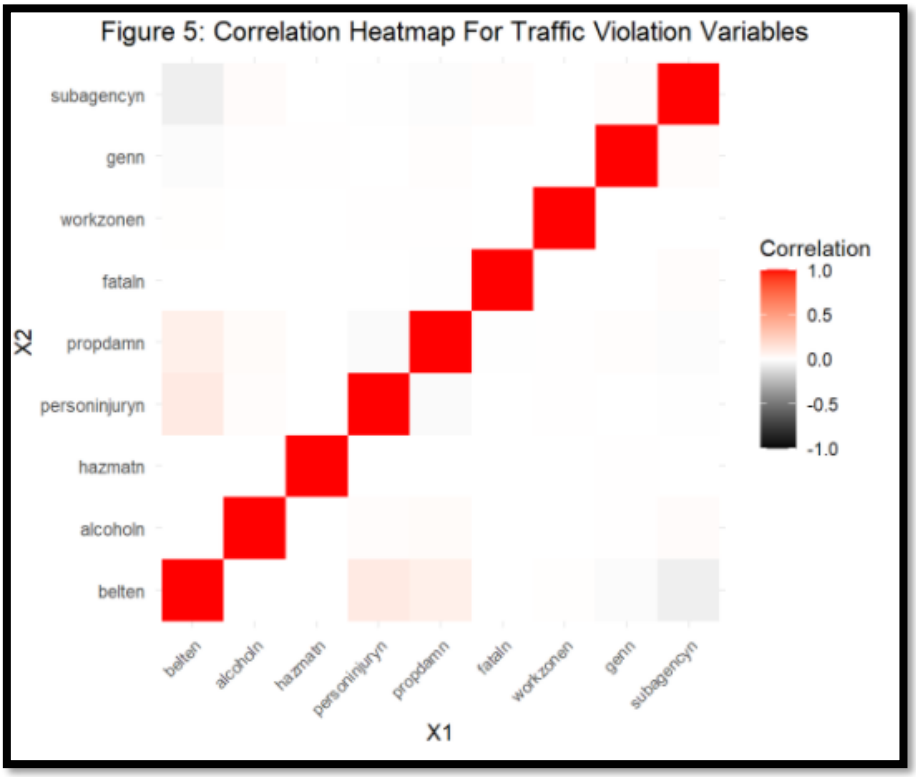
- Analysis:
  - EDA



- Study of violation over the years (2012 - 2018)
  - From Chart A - Histogram of Hour we understand that that most of the Violation are made between the time range 20PM to 24PM
  - From Chart B - Histogram of we understand that that most of the Violation on yearly basis
  - From Chart C - Bar plot we understand the distribution of violation based on months.



- From the above chart - Barplot we understand the distribution of Violation type.



- According to the heat map above, there is a very weak association between those who violate the law by not wearing seat belts and personal injury, also with weak correlation between the belts no worn and Sub Agency, with no other correlation between other variables plotted in the heat map.

- Data Preparation:
    - We will perform a Logistic Regression with Family Binomial and Link as “Logit”.
    - The prediction and study will revolve around understating the behaviour of the driver based on various factors as provided in the data set.
    - The choice of Logistic model is based on Target variables, as the variables a binomial the logistic model is more useful and a right choice to make.
- Target Variable:
    - Belts (Was the violation recorded for belts)
    - Alcohol (Was the violation recorded for Alcohol)

- Sampling Technique:
    - We have sampled the data into Train and Test to run the model and perform predictions. 70% Training data and 30% Test data
- Encoding:
    - We have generalized the column Vehicle type to reduce the levels as that will be used in Logistic model
    - We have also encoded the target variables with Yes = 1 and No = 0.

- Model 1:
    - Target variable – Belts.

```
Call:
glm(formula = belten ~ Gender + Race + workzonen + cartype +
    SubAgency + Commercial.Vehicle + Quaterly_view2, family = binomial(link = "logit"),
    data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.8798  -0.3159  -0.2204  -0.1904   3.8533

Coefficients:
                                                    Estimate Std. Error z value Pr(>|z|)
(Intercept)                                        -3.110474   0.066193 -46.991  < 2e-16 ***
GenderM                                            -0.145149   0.024179  -6.003 1.93e-09 ***
GenderU                                            -1.549100   0.583540  -2.655  0.00794 **
RaceBLACK                                          -0.099683   0.052204  -1.909  0.05620 .
RaceHISPANIC                                        0.037356   0.053567   0.697  0.48557
RaceNATIVE AMERICAN                                -0.480106   0.323924  -1.482  0.13830
RaceOTHER                                           0.151548   0.066652   2.274  0.02298 *
RaceWHITE                                           0.126442   0.050887   2.485  0.01296 *
workzonen                                           1.816437   0.416380   4.362 1.29e-05 ***
cartype                                             0.180518   0.013479  13.393  < 2e-16 ***
SubAgency2nd district, Bethesda                    -1.243959   0.045199 -27.522  < 2e-16 ***
SubAgency3rd district, Silver Spring               -0.210184   0.034149  -6.155 7.51e-10 ***
SubAgency4th district, Wheaton                     -0.967587   0.036846 -26.260  < 2e-16 ***
SubAgency5th district, Germantown                  -3.262156   0.122257 -26.683  < 2e-16 ***
SubAgency6th district, Gaithersburg / Montgomery Village -0.075613   0.035461  -2.132  0.03299 *
SubAgencyHeadquarters and Special Operations       -3.156144   0.210372 -15.003  < 2e-16 ***
Commercial.VehicleYes                              -0.899524   0.282835  -3.180  0.00147 **
Quaterly_view2                                      0.005437   0.010172   0.535  0.59299
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 70536  on 244845  degrees of freedom
Residual deviance: 66708  on 244828  degrees of freedom
AIC: 66744

Number of Fisher Scoring iterations: 8
```

Figure A: Summary of Model 1

- The call Section tells us about the response and the predictor variable we fit in the model to run the logistic regression.
    - Response variable is Belts i.e., committing or booked for a Belt violation.
    - Predictor variables consist of Gender, Race, Subagency, Car type, Work zone, Commercial Vehicle, and Quarterly View.
- Deviance Residuals: The median value is near the 0 which indicates the residuals show symmetry.
- Gender
    - From the estimates we understand that the beta β estimate is -0.145 where the exp(coefficient) comes to 0.86 for Males and 0.21 for the unknown which means that on male gender basis there are 0.86 times of committing a belts violation then others.
- Race, Car type and Work zone
    - There are 1.163 and 1.134 times of odds a person of White Race or other Race will perform a belts violation. There are 6.15 and 1.19 that the violation was committed in a Work zone area and a particular car type is involved.

	Original	2.5 %	97.5 %
(Intercept)	0.04457983	0.03912865	0.05072203
GenderM	0.86489373	0.82493821	0.90695332
GenderU	0.21243917	0.05240445	0.56015538
RaceBLACK	0.90512470	0.81790528	1.00367869
RaceHISPANIC	1.03806285	0.93545424	1.15408012
RaceNATIVE AMERICAN	0.61871783	0.30625098	1.10509047
RaceOTHER	1.16363380	1.02111033	1.32610139
RaceWHITE	1.13478412	1.02815365	1.25518648
workzonen	6.14990443	2.49287295	13.07159228
cartype	1.19783822	1.16634949	1.22963993
SubAgency2nd district, Bethesda	0.28824087	0.26366821	0.31478741
SubAgency3rd district, Silver Spring	0.81043543	0.75802631	0.86660860
SubAgency4th district, Wheaton	0.37999869	0.35349973	0.40843423
SubAgency5th district, Germantown	0.03830573	0.02986943	0.04827327
SubAgency6th district, Gaithersburg / Montgomery Village	0.92717494	0.86491612	0.99391313
SubAgencyHeadquarters and Special Operations	0.04258965	0.02738150	0.06270291
Commercial.VehicleYes	0.40676305	0.22176257	0.67817134
Quaterly_view2	1.00545198	0.98560613	1.02570232

Figure B: Odds and Odds Ratio Table

- Subagency
    - There are 0.28 to 0.92 times odds that a person from each district listed will perform a belts violation.
- Commercial Vehicle
    - There are 0.407 times odd that a violation of belts is performed by owner/driver of Commercial vehicle.
- Quarterly View
    - There are 1.005 times odd that a belts violation is performed in a Quarter.

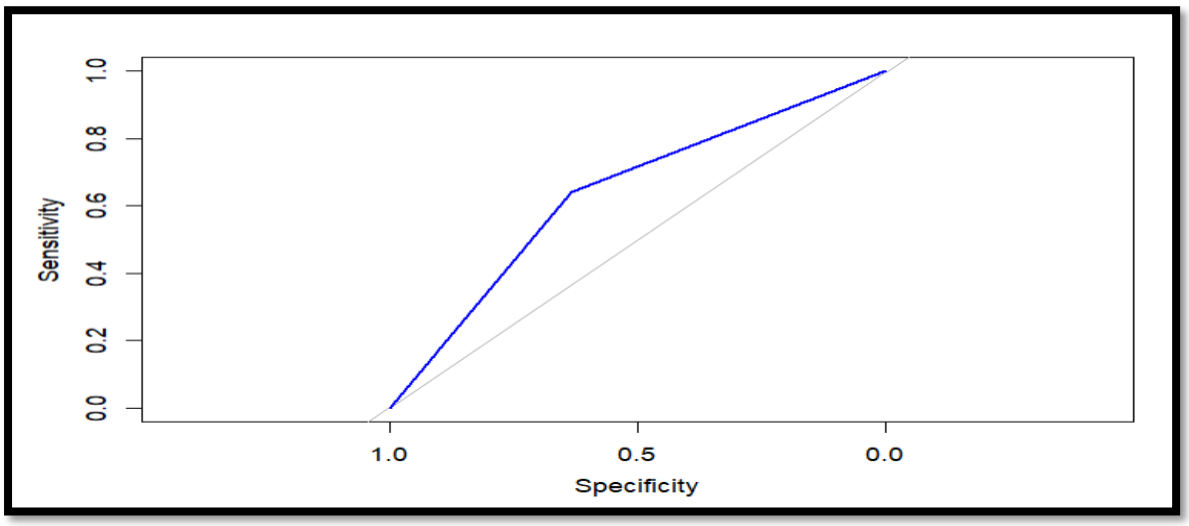


Figure C: ROC Curve

- AOC: Area under the curve: 0.6369
    - As per the ROC and AOC of the model we can conclude that the model designed is 63.69% fit for the prediction.

- Model 2:
    - Target Variable - Alcohol

```
Call:
glm(formula = alcoholn ~ Gender + Race + workzonen + cartype +
    Commercial.Vehicle + SubAgency + Quaterly_view2, family = binomial(link = "logit"),
    data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.2203  -0.0526  -0.0424  -0.0268   4.1348

Coefficients:
                                                    Estimate Std. Error z value Pr(>|z|)
(Intercept)                                        -9.46211    0.52781 -17.927  < 2e-16 ***
GenderM                                             0.37779    0.11498   3.286 0.001017 **
GenderU                                           -12.63505 1027.45315  -0.012 0.990188
RaceBLACK                                           1.36285    0.45868   2.971 0.002966 **
RaceHISPANIC                                        1.67785    0.46158   3.635 0.000278 ***
RaceNATIVE AMERICAN                               -12.57153  695.23748  -0.018 0.985573
RaceOTHER                                           0.92382    0.53277   1.734 0.082919 .
RaceWHITE                                           1.75366    0.45320   3.870 0.000109 ***
workzonen                                         -14.01824 2318.17903  -0.006 0.995175
cartype                                             0.12142    0.05673   2.140 0.032343 *
Commercial.VehicleYes                               0.05454    1.00992   0.054 0.956933
SubAgency2nd district, Bethesda                     0.49179    0.26192   1.878 0.060437 .
SubAgency3rd district, Silver Spring               -0.99252    0.35515  -2.795 0.005196 **
SubAgency4th district, Wheaton                      0.12050    0.25885   0.465 0.641574
SubAgency5th district, Germantown                   2.35206    0.22896  10.273  < 2e-16 ***
SubAgency6th district, Gaithersburg / Montgomery Village    0.65924    0.26602   2.478 0.013207 *
SubAgencyHeadquarters and Special Operations      -13.42826  193.76004  -0.069 0.944748
Quaterly_view2                                      0.07516    0.04435   1.695 0.090132 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5932.9  on 244845  degrees of freedom
Residual deviance: 5364.4  on 244828  degrees of freedom
AIC: 5400.4

Number of Fisher Scoring iterations: 19
```

Figure D: Summary Call for Model 2

- The call Section tells us about the response and the predictor variable we fit in the model to run the logistic regression.
    - Response variable is Alcohol i.e., committing or booked for a drinking Alcohol violation.
    - Predictor variables consist of Gender, Race, Subagency, Car type, Work zone, Commercial Vehicle, and Quarterly View.
- Deviance Residuals: The median value is near the 0 which indicates the residuals show symmetry.
- Gender
    - From the estimates we understand that the beta β estimate is 0.377 where the exp(coefficient) comes to 1.45 for Males and 3.25 for the unknown which means that on male gender basis there are 1.45 times of committing an alcohol violation then others.
- Race, Car type and Work zone
    - There is 3+ and 5+ and 5+ times of odds a person of Black, Hispanic and White Race will commit an Alcohol violation. There is 8+ times odds chance

that an Alcohol violation is committed in work zone area and 1.19 that the violation was committed involved car type.

	Original	2.5 %	97.5 %
(Intercept)	7.774231e-05	2.459246e-05	2.012919e-04
GenderM	1.459061e+00	1.168955e+00	1.835505e+00
GenderU	3.255883e-06	1.462062e-199	0.000000e+00
RaceBLACK	3.907330e+00	1.765178e+00	1.107786e+01
RaceHISPANIC	5.354019e+00	2.400827e+00	1.524235e+01
RaceNATIVE AMERICAN	3.469381e-06	5.471770e-137	5.141969e-229
RaceOTHER	2.518905e+00	9.333561e-01	7.923694e+00
RaceWHITE	5.775684e+00	2.646379e+00	1.624725e+01
workzonen	8.164976e-07	0.000000e+00	5.555747e+35
cartype	1.129095e+00	1.005312e+00	1.256152e+00
Commercial.VehicleYes	1.056053e+00	5.966091e-02	4.806066e+00
SubAgency2nd district, Bethesda	1.635234e+00	9.927385e-01	2.788045e+00
SubAgency3rd district, Silver Spring	3.706423e-01	1.800881e-01	7.347316e-01
SubAgency4th district, Wheaton	1.128057e+00	6.899466e-01	1.913950e+00
SubAgency5th district, Germantown	1.050719e+01	6.879932e+00	1.695760e+01
SubAgency6th district, Gaithersburg / Montgomery Village	1.933313e+00	1.162728e+00	3.318881e+00
SubAgencyHeadquarters and Special Operations	1.472931e-06	5.177602e-43	1.165614e-68
Quaterly_view2	1.078059e+00	9.884371e-01	1.176239e+00

Figure E: Odds and Odds Ratio Table

- Subagency
    - There are 1 to 3 times odds that a person from each district listed will perform an alcohol violation.
- Commercial Vehicle
    - There are 1.1 times odd that a violation of alcohol is performed by owner/driver of Commercial vehicle.
- Quarterly View
    - There are 1.07 times odd that a Alcohol violation is performed in a Quarter.

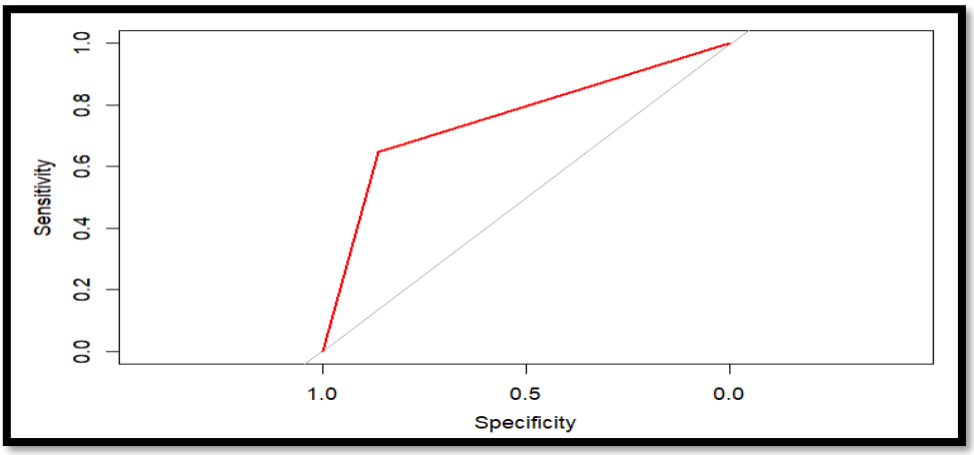


Figure F: ROC Curve

❖ Area under the curve: 0.7549

❖ As per the ROC and AOC of the model we can conclude that the model designed is 75.49% fit for the prediction.

- Combined result of the Models

Logistics regression models		
	Dependent variable:	
	Belts	Alcohol
	(1)	(2)
GenderM	-0.1***	0.4***
	(0.02)	(0.1)
GenderU	-1.5***	-12.6
	(0.6)	(1,027.5)
RaceBLACK	-0.1*	1.4***
	(0.1)	(0.5)
RaceHISPANIC	0.04	1.7***
	(0.1)	(0.5)
RaceNATIVE AMERICAN	-0.5	-12.6
	(0.3)	(695.2)
RaceOTHER	0.2**	0.9*
	(0.1)	(0.5)
RaceWHITE	0.1**	1.8***
	(0.1)	(0.5)
workzonen	1.8***	-14.0
	(0.4)	(2,318.2)
cartype	0.2***	0.1**
	(0.01)	(0.1)
SubAgency2nd district, Bethesda	-1.2***	0.5*
	(0.05)	(0.3)
SubAgency3rd district, Silver Spring	-0.2***	-1.0***
	(0.03)	(0.4)
SubAgency4th district, Wheaton	-1.0***	0.1
	(0.04)	(0.3)
SubAgency5th district, Germantown	-3.3***	2.4***
	(0.1)	(0.2)
SubAgency6th district, Gaithersburg / Montgomery Village	-0.1**	0.7**
	(0.04)	(0.3)
SubAgencyHeadquarters and Special Operations	-3.2***	-13.4
	(0.2)	(193.8)
Commercial.VehicleYes	-0.9***	0.1
	(0.3)	(1.0)
Quaterly_view2	0.01	0.1*
	(0.01)	(0.04)
Constant	-3.1***	-9.5***
	(0.1)	(0.5)
Observations	244,846	244,846
Log Likelihood	-33,354.1	-2,682.2
Akaike Inf. Crit.	66,744.1	5,400.4
Note:	*p<0.1; **p<0.05; ***p<0.01	

- **Conclusion:**
    1. We learned through the data exploration and summary that people in general commit 4 to 5 sorts of violations, such as not wearing seat belts, drinking alcohol while driving, transporting hazardous materials in the vehicle, and parking in a work zone area. Although most of the violations are not lethal, they do cause personal and property damage. The infractions have been reported in several districts and include all types of Vehicles out of which "Automobile vehicle type" car owners commit most of the violation. The violation is committed by people of all genders and races. From 2012 to 2018, the violations were registered. The number of violations registered in 2015 and 2016 is at an all-time high, with most of them occurring at night and midnight.
    2. 2. The variables have a very weak link; for example, there is a correlation between the belts and property damage variable and the belts and subagency variable. A hypothesis test was also conducted to determine the link between Gender, Race, and Violation Type variables, which revealed a dependency. 2 Logistic regression models were created to forecast and explain the impact of various predictor variables on the likelihood of a Belts violation or an Alcohol violation occurring. Males are shown to be more likely than females to violate belts and alcohol laws. Most races impact the conduct of violations, however White individuals and people categorized as "other" races are more likely to commit belt violations, while people of Black, White, and Hispanic races are more likely to conduct alcohol violations. Commercial vehicle owners/drivers are more likely to break the law by not wearing their seatbelts. The Quarterly view variable reveals that the violation will continue to occur, indicating that certain steps should be done to avoid and reduce the number of violations.
    3. Based on the findings of the study and analysis, we can conclude that in the county of Maryland, every person, regardless of gender or race, will commit violations. Below are some recommendations from the Maryland Traffic Police and Traffic Management Department to help reduce violations.
        a. a. Fine penalties for all types of violations can be increased, encouraging drivers to be more proactive.
        b. All districts should establish general driver guidelines, and awareness campaigns can be launched to make everyone aware of the problem.
        c. Traffic police patrols during the night and at midnight might be increased, which would help to prevent some offences.
        d. Automobile manufacturers may be asked to make wearing seat belts mandatory so that drivers do not violate the law.

❖ Bibliography:

- Traffic Violations in Maryland County. (2018, May 9). Kaggle. https://www.kaggle.com/rounak041993/traffic-violations-in-maryland-county
- ANOVA Test: Definition & Uses (Updated 2022). (2022, January 18). Qualtrics. Retrieved January 25, 2022, from https://www.qualtrics.com/experience-management/research/anova/#:%7E:text=You%20would%20use%20ANOVA%20to,are%20unequal%20(or%20different).
- Chi-square test of independence in R. (n.d.). Stats and R. Retrieved January 25, 2022, from https://statsandr.com/blog/chi-square-test-of-independence-in-r/
- R CODER. (2021, December 5). The ggplot2 package | R CHARTS. R CHARTS | A Collection of Charts and Graphs Made with the R Programming Language. https://r-charts.com/ggplot2/
- Admin, T. S. (2020, October 15). Traffic Violations: Definition, Types, Consequences and Examples. School Bus Tracking System with TrackSchoolBus Apps. https://www.trackschoolbus.com/traffic-violations/
- C. (2021, December 14). An Introduction to Logistic Regression for Categorical Data Analysis. Medium. https://towardsdatascience.com/an-introduction-to-logistic-regression-for-categorical-data-analysis-7cabc551546c
- Narkhede, S. (2021, June 15). Understanding AUC - ROC Curve - Towards Data Science. Medium. https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5
- Logit Regression | R Data Analysis Examples. (n.d.). UCLA Logistics. Retrieved February 19, 2022, from https://stats.oarc.ucla.edu/r/dae/logit-regression/
- Practical Guide to Logistic Regression Analysis in R Tutorials & Notes | Machine Learning. (2017, April 10). HackerEarth. https://www.hackerearth.com/practice/machine-learning/machine-learning-algorithms/logistic-regression-analysis-r/tutorial/