**CAPSTONE PROJECT**

**Traffic Violations in US:  
Predicting Property damage during traffic violations**



## **MENTOR:** **TEAMMEMBERS:**

## Jatinder Bedi Bandaru Krishna Kanth

## Sudarsan

## Vijay Viswanathan

## Sandeep

## Madhumitha Raghuraman

CONTENTS

# CHAPTER 1……………………………………………………………………………………………………………………...……….4

# INTRODUCTION………………………………………………………………………………………………………………………..4

# 1.1 TRAFFIC VIOLATIONS….………………………………………………………………………………………………… 4

# 1.2 TRAFFIC TICKETS…………………………………………………………………………………………………….….… 4

# 1.3 BUSINESS CASE…………………..…………………………………………………………………………………….…….5

# 1.4 PROBLEM APPROACH……………………………………………………………………………………………………. 5

# 1.5 SCOPE OF THE PROJECT………………………………………………………………………………………………….6

# 1.6 VARIABLES CONSIDERED FOR ANALYSIS………………………………………………………………………...6

# CHAPTER 2………………………………………………………………………………………………………………...……………. 8

# DATA SAMPLING……………………...…………………………………………………………………………………………..….. 8

# 2.1 SAMPLE SELECTION FROM POPULATION………………………………………………..…………....……...... 8

# CHAPTER 3….......................................................................................................................................................................…10

# DATA CLEANING……………………………………………………………………………………………………………………..10

# 3.1 BRIEF DESCRIPTION ABOUT DATASET…………………………………………………………………..…..….10

# 3.2 UNIQUE VALUES IN FEATURES…………..………………………………………………………………………....10

# 3.3 NULL VALUES TREATMENT…………………..……………………………………………………………..………..10

# CHAPTER 4……………………………………………………………………………………………………………………………..12

# EXPLORATORY DATA ANALYSIS……………………………………………………………………………………...………12

# 4.1 INTRODUCTION…………………………………………………………………………………………………………… 12

# 4.2 BASIC EDA…………………………………………………………………………………………………………………… 12

# 4.2.1 YEAR OF DATE VERSUS PROPERTY DAMAGE…………...………………………………………..…………..12

# 4.2.2 SEASONS VERSUS PROPERT DAMAGE………………………………………….………………………………..13

# 4.2.3 TIME OF STOP VERSUS PROPERTY DAMAGE …………...………………………………………..…………. 14

# 4.2.4 DATE OF STOP VERSUS PROPERTY DAMAGE…………………………….…………………………….……. 14

# VEHICLE MAKE VERSUS PROPERTY DAMAGE…………...……………..………………………..…………..15

# 4.26 REASONS DESCRIPTION VERSUS PROPERTY DAMAGE.……………….…………………………….….. 15

# 4.2.7 RACE VERSUS PROPERTY DAMAGE …………...………………………...…………………………..……………16

# 4.2.8 COMMERCIAL VEHICLE VERSUS PROPERTY DAMAGE……………………………………………………16

# 4.2.9 SUB AGENCY VERSUS PROPERTY DAMAGE………………...……………………………………..…………...17

# 4.2.10 DL STATE VERSUS PROPERTY DAMAGE……………………………………….………………………………..18

# 4.2.11 ACCIDENT AREAS CLUSTERS VERSUS PROPERTY DAMAGE………….………………………………..18

# CHAPTER 5……………………………………………………………………………………………………………………………..20

# MACHINE LEARNING MODELS………………………………………………………………………………………...………20

# 5.1 INTRODUCTION…………………………………………………………………………………………………………… 20

# 5.2 LOGISTIC REGRESSION…..……………………………………………………………………………………………. 20

# 5.3 DECISION TREE CLASSIFIER………………………………………………………………………………………… 21

# 5.4 RANDOM FOREST………………………………………………………………………………………………………… 23

# 5.5 FEATURE IMPORTANCE………………………………………………………………………………………………..26

# 5.6 COMPARING MODEL SCORES………………………….……………………………………………………………..26

# 5.7 COMPARING ROC CURVES OF THREE MODELS…………………..…………………………………………..27

# CHAPTER 6……………………………………………………………………………………………………………………………..27

# 6.1 CONCLUSION…………………………………………………….…………………………………………………………. 27

# 6.2 RECOMMENDATIONS…..………………………………………………………………………………………………. 27

# 6.3 LIMITATIONS…………………………….……………...…………………………………………………………………. 27

# 6.4 CLOSING REFLECTIONS………………………….....…………………………………………………………………. 28

# 

# CHAPTER 1

# INTRODUCTION

## TRAFFIC VIOLATIONS

## Generally, In the United States federal criminal code, crimes are divided into two broad categories: Misdemeanors and Felonies. Misdemeanors are crimes that carry a maximum of one year of jail time. Felonies are crimes with punishments in excess of one year of imprisonment.

## Traffic Violations that comes under misdemeanors charges are:

## Driving under the influence of drugs or alcohol,

## Driving without a valid driver’s license,

## Driving without insurance (in some states),

## Failing to stop at the scene of an accident,

## Reckless driving that potentially endangers others.

## Traffic offenses that result in felony criminal charges usual involve severe injuries or serious property damage. Traffic violations that commonly result in felony charges include:

## Hit-and-run accidents,

## Repeat DUI convictions,

## Vehicular homicide,

## Repeat traffic offenses such as driving without a valid license.

## TRAFFIC TICKETS

A traffic ticket is a notice issued by a law enforcement official to a motorist or other road user, indicating that the user has violated traffic laws.Traffic tickets generally come in two forms, moving violation, such as exceeding the speed limit, or a non-moving violation, such as a parking violation, with the ticket also being referred to as a parking citation, or parking ticket. There are approximately 196,000,000 licensed drivers in America today. One in every six drivers will receive a speeding ticket this year, roughly 41 million speeding tickets, which is over $6 billion dollars each year on speeding ticket fines alone.

* 1. BUSINESS CASE

A lot of work has been done on predicting fatality in accidents, and it has helped in the following areas of road safety:

* Safe route planning
* Emergency vehicle allocation
* Roadway design
* Where to place additional signage (e.g. to warn for curves)

However, the gap found in this approach is that accidents that are not fatal, but nonetheless severe, are not given adequate focus.

In this project, we try to address this by predicting accidents that cause property damage (and hence are severe enough to be studied). This approach would not only aid in road safety planning, but also cater as one of the input variables for Used Car Inspection during resale.

Most countries do have a system in place for recording accidents linked to a vehicle. But there are also several cases where property damages are often not recorded, unless an insurance claim or a traffic ticket has been raised. Predicting the probability of property damage based on the vehicle, driver and location characteristics would enable Used car dealers to perform a closer inspection on the resale vehicle that share the characteristics of a damage-prone vehicle.

Thus, approaching the accident prediction challenge as a Classification problem, with property damage as the prediction variable, would provide a more cautious approach to both traffic planning and vehicle inspection.

* 1. PROBLEM APPROACH

1. Descriptive Analytics
2. Predictive Analytics

## SCOPE OF THE PROJECT

This project is a study that aims at assisting Law Enforcement Department in Montgomery County to know about factors which leads to property damage during traffic violations.

## VARIABLES CONSIDERED FOR ANALYSIS

The following variables are considered for the analysis:

|  |  |  |  |
| --- | --- | --- | --- |
| S.NO | VARIABLE | DESCRIPTION | DATA TYPE |
| 1 | Date Of Stop | Date of the traffic violation. | Calendar Date |
| 2 | Time Of Stop | Time of the traffic violation | Text |
| 3 | 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) | Text |
| 4 | Description | Text description of the specific charge. | Text |
| 5 | Location | Location of the violation, usually an address or intersection. | Text |
| 6 | Latitude | Latitude location of the traffic violation. | Number |
| 7 | Longitude | Longitude location of the traffic violation | Number |
| 8 | Accident | If traffic violation involved an accident. | Text |
| 9 | Belts | If traffic violation involved a seat belt violation. | Text |
| 10 | Personal Injury | If traffic violation involved Personal Injury. | Text |
| 11 | Fatal | If traffic violation involved a fatality. | Text |
| 12 | Commercial License | If driver holds a Commercial Driver’s License. | Text |
| 13 | HAZMAT | If the traffic violation involved hazardous materials. | Text |
| 14 | Commercial Vehicle | If the vehicle committing the traffic violation is a commercial vehicle. | Text |
| 15 | Alcohol | If the traffic violation included an alcohol related | Text |
| 16 | Work Zone | If the traffic violation was in a work zone. | Text |
| 17 | State | State issuing the vehicle registration. | Text |
| 18 | VehicleType | Type of vehicle (Examples: Automobile, Station Wagon, Heavy Duty Truck, etc.) | Text |
| 19 | Year | Year vehicle was made. | Text |
| 20 | Make | Manufacturer of the vehicle (Examples: Ford, Chevy, Honda, Toyota, etc.) | Text |
| 21 | Model | Model of the vehicle | Text |
| 22 | Color | Color of the vehicle. | Text |
| 23 | Violation Type | Violation type. (Examples: Warning, Citation, SERO) | Text |
| 24 | Contributed To Accident | If the traffic violation was a contributing factor in an accident. | Text |
| 25 | Race | Race of the driver. (Example: Asian, Black, White, Other, etc.) | Text |
| 26 | Gender | Gender of the driver (F = Female, M = Male) | Text |
| 27 | Driver City | City of the driver’s home address. | Text |
| 28 | Driver State | State of the driver’s home address. | Text |
| 29 | DL State | State issuing the Driver’s License. | Text |
| 30 | Arrest Type | Type of Arrest (A = Marked, B = Unmarked, etc.) | Text |
| **31** | **Property Damage** | **If traffic violation involved Property Damage**  **(Target variable)** | **Text** |

# CHAPTER 2

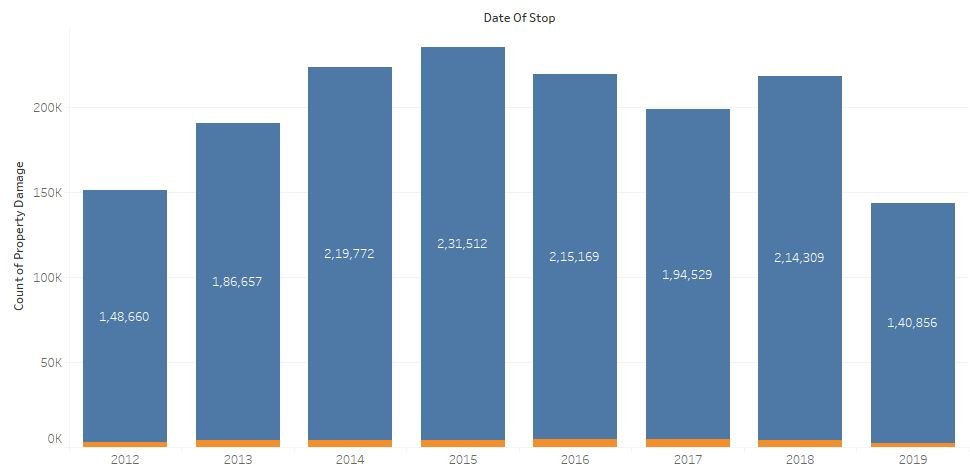
DATA SAMPLING

## SAMPLE SELECTION FROM POPULATION

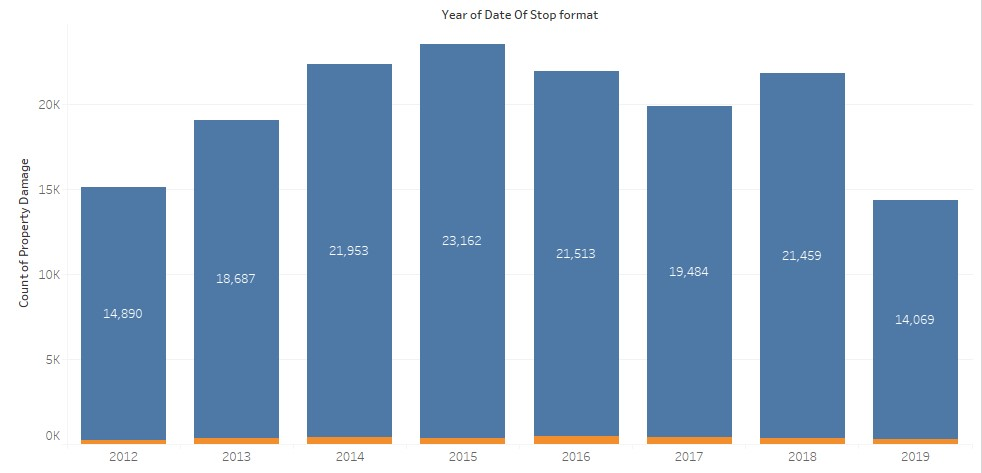
Now, performing **Test of proportion** for property damage:

**Null Hypothesis**: Proportions are equal in sample and population.

**Alternate Hypothesis:** Proportions are not equal in sample and population. After preforming the test in python, we get **p-value equals 0.9997**and since z Statistic lies in the acceptance region so we fail to reject the null hypothesis. We can infer that Value counts of property damage (yes) in sample are proportional to value counts of property damage(yes) in population.Similarly, the other class can also be tested.



**Population**



**Sample**

**As shown in the above two graphs, Sample is the true representation of the population**

Similarly, performing **Test of proportion** for accidents:

**Null Hypothesis**: Proportions are equal in sample and population.

**Alternate Hypothesis:** Proportions are not equal in sample and population.

After preforming the test in python, we get **p-value equals 0.9997** and since z Statistic lies in the acceptance region so we fail to reject the null hypothesis.

We can infer that Value counts of accidents (yes) in sample are proportional to value counts of accidents (yes) in population.

Similarly, the other class can also be tested.

**Population**

# 

# 

**Sample**

**As shown in the above two graphs, Sample is the true representation of the population**

# CHAPTER 3

DATA CLEANING

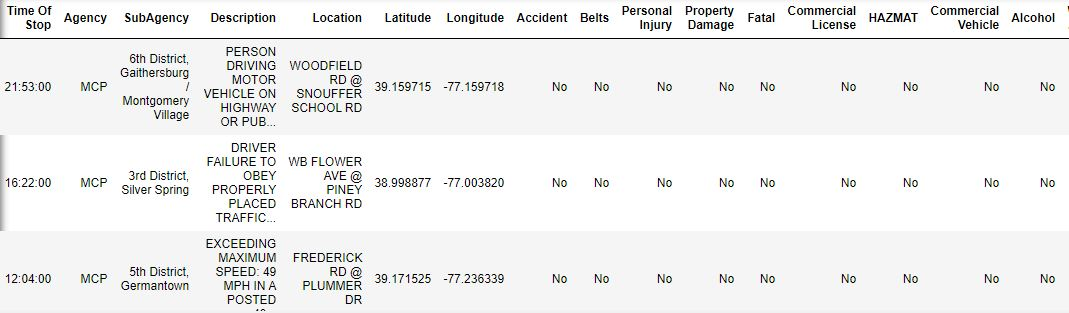
## BRIEF DESCRIPTION ABOUT DATASET

## **Source:**

## **https://data.montgomerycountymd.gov/api/views/4mseku6q/rows.csv?accessType=DOWNLOAD**

Montgomery County is the most populous county in the U.S. state of Maryland, located adjacent to Washington, D.C.

**Below is a snapshot of the dataset.**



* The dataset has 15,80,153 records versus 43 features
* The dataset consists of 2 numerical, 2 ordinal and 39 categorical attributes.

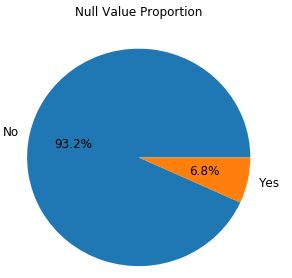
## UNIQUE VALUES IN FEATURES

* Date of Stop **- 258**
* Time of Stop **- 1448**
* Sub Agency **- 07**
* Description **- 2937**
* Location **- 30402**
* Latitude **- 55367**
* Longitude **- 58333**
* Accident **- 02**
* Belts **- 02**
* Personal Injury **- 02**
* Fatal **- 02**
* Commercial License **- 02**
* HAZMAT **- 02**
* Commercial Vehicle **- 02**
* Alcohol **- 02**
* Work Zone **- 02**
* State **- 58**
* VehicleType **- 24**
* Year **- 102**
* Make **- 927**
* Model **- 3653**
* Color **- 26**
* Violation Type **- 03**
* Contributed To Accident **- 02**
* Race **- 06**
* Gender **- 03**
* Driver City **- 2127**
* Driver State **- 59**
* DL State **- 63**
* Arrest Type **- 18**
* **Property Damage - 02**

## 3.3 NULL VALUES TREATMENT

Features like State, Year, Make, Model, Driver City, Driver State, DL State, Article, Color have null values proportion lesser than 0.5%. We can drop those records because even if we try to impute, it would create more noise in the data. All of the variables with missing values are categorical.

|  |  |  |  |
| --- | --- | --- | --- |
| **Column** | **Type** | **Null values** | **Null values (%)** |
| Time Of Stop | object | 0 | 0 |
| Agency | object | 0 | 0 |
| SubAgency | object | 0 | 0 |
| Description | object | 0 | 0 |
| Location | object | 0 | 0 |
| Latitude | float64 | 0 | 0 |
| Longitude | float64 | 0 | 0 |
| Accident | object | 0 | 0 |
| Belts | object | 0 | 0 |
| Personal Injury | object | 0 | 0 |
| Property Damage | object | 0 | 0 |
| Fatal | object | 0 | 0 |
| Commercial License | object | 0 | 0 |
| HAZMAT | object | 0 | 0 |
| Commercial Vehicle | object | 0 | 0 |
| Alcohol | object | 0 | 0 |
| Work Zone | object | 0 | 0 |
| State | object | 4 | 0.00252884 |
| VehicleType | object | 0 | 0 |
| Year | float64 | 999 | 0.631579 |
| Make | object | 6 | 0.00379327 |
| Model | object | 19 | 0.012012 |
| Color | object | 1875 | 1.1854 |
| Violation Type | object | 0 | 0 |
| Charge | object | 0 | 0 |
| Article | object | 7705 | 4.87119 |
| Race | object | 0 | 0 |
| Gender | object | 0 | 0 |
| Driver City | object | 44 | 0.0278173 |
| Driver State | object | 2 | 0.00126442 |
| DL State | object | 118 | 0.0746009 |
| Arrest Type | object | 0 | 0 |
| Date OfStop\_format | object | 0 | 0 |
| Contributed To Accident | bool | 0 | 0 |



# CHAPTER 4

EXPLORATORY DATA ANALYSIS

## INTRODUCTION

EDA is a general approach to exploring datasets by means of simple summary statistics and graphic visualizations in order to gain a deeper understanding of the data. Significance of Exploratory Data Analysis (EDA). Data science is ubiquitous to advanced statistical and machine learning techniques. As long as there is data to analyze, the need to explore is obvious. Yet, an important key component to any data science task frequently undervalued is the exploratory data analysis (EDA)

## BASIC EDA

## YEAR OF DATE VERSUS PROPERTY DAMAGE

## 

Observation:

2016 has the highest record of property damage incidents. Post 2016, there is a decline in the number of such accidents.

Insight:

* From external references( <https://www.pinderplotkin.com/motorcycle-accident-statistics-for-maryland-and-baltimore-county/>), we come to know that in 2016, 33% of traffic accidents which leads to property damage was due to Driving Under Influence (DUI).
* After 2016, Strict measures were taken like increasing check points and checking for driving under influence. Thus, enforcing legislations can be one of the solutions to reduce accidents.

## SEASONS VERSUS PROPERTY DAMAGE

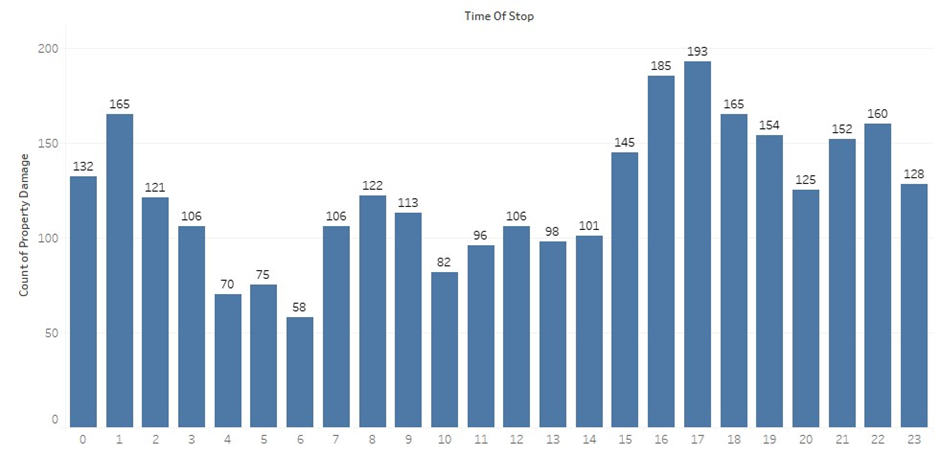
## 

Observation:

Spring season gets more property damage.

Insight:

* Besides the fact that Spring season has more months, this could be partly due to the increased number of travelers during the spring season.
* A lot of people on the roads would be from other towns, so they won’t be familiar with the area. Spring is also the time for college breaks and family vacations, creating more risk opportunities.
* One control measure would be to plan and install adequate sign boards and route maps in the area and increasing traffic control measures.
  + 1. TIME OF STOP VERSUS PROPERTY DAMAGE

****

Observation:

The evening rush hour 3-6pm, is the most common time for accidents. The accidents are maximum at 4-5 pm, and there is again a peak at 9-10 pm.

Insight:

This is expected, Since it's the time when most offices and schools in USA get over. Second shift of work in USA generally finishes at 9-10 pm. Higher the number of vehicles on road, more are the chances of accidents. Speed control limits, speed breakers, and lane management are necessary to control the traffic during rush hours.

## DATE OF STOP VERSUS PROPERTY DAMAGE

## 

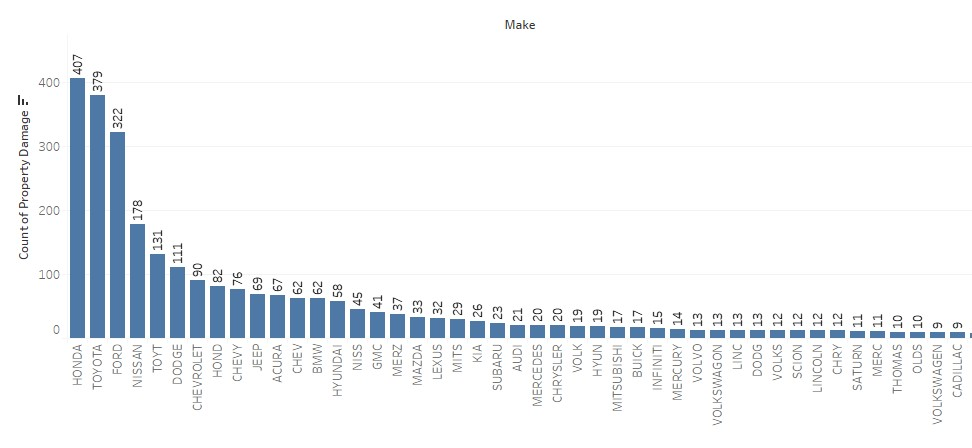
Observation:

Weekdays are more susceptible to accidents.

Insight:

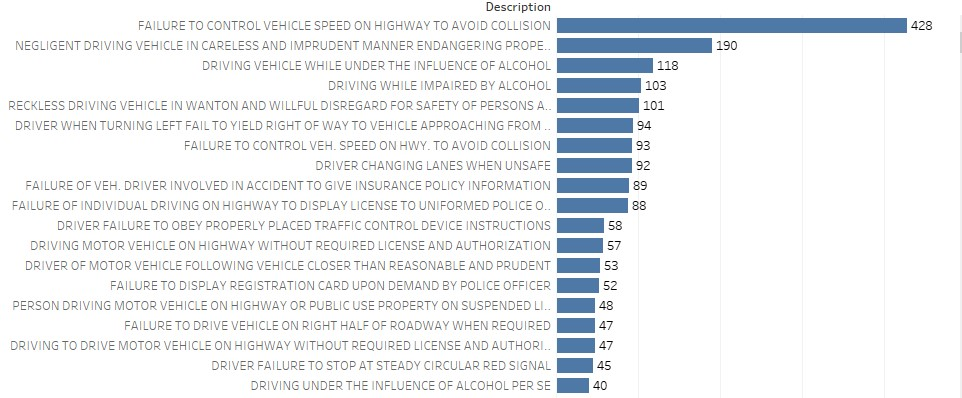
Work hour and school hour rush could be the reasons.

## VEHICLE MAKE VERSUS PROPERTY DAMAGE



Toyota, Honda, Nissan and Ford form the bulk of usage in USA. Due to sheer high number the property damage instances too are higher.

## REASONS DESCRIPTION VERSUS PROPERTY DAMAGE



Failure to control vehicle speed on highway to avoid collision is the highest reason for the property damage

## RACE VERSUS PROPERTY DAMAGE

## 

## Observation:

## White, Hispanic and Black race Whites race account for 90% of accidents with property damage. Considering that Blacks race and Hispanics constitute only 29.72% and 3.22% of the population, their contribution to accidents is very high.

## As of 2016, number of female drivers are 2160894 while male drivers are 2103981. In spite of being higher in number than men, women are less prone to commit property damage then men

## COMMERCIAL VEHICLE VERSUS PROPERTY DAMAGE

## 

## Observation:

## Commercial vehicles are less prone to accident.

## Insight:

## Commercial vehicle drivers are professionals, and the size of the vehicles make it less incentive to drive rash.

## Also, the greater accountability of commercial vehicles brings in more responsibility on the driver's shoulders.

## SUB AGENCY VERSUS PROPERTY DAMAGE

## 

## Observations:

## Most accidents happen in 4th district of Wheaton and in Wheaton most of the accidents happen in Randolph road.

## Insights:

## The property cluster is high in the intersection where to neighboring areas combine

## DL STATE VERSUS PROPERTY DAMAGE

## 

## Peoples who belong to their own motherland (Montgomery County) commit more accidents than the outsiders

## ACCIDENT AREAS CLUSTERS VERSUS PROPERTY DAMAGE

## 

## Traffic Violations occur more in intersections areas of the Highway roads

### **Conclusions from Exploratory Data Analysis**

* + - * 2016, 33% of traffic accidents which leads to property damage was due to Driving Under Influence (DUI)
      * During 2012-2016, Most of the accidents occur in Maryland state routes 33.9 and 28% crashes occurred on country roads.
      * After 2016, Strict measures were taken like increasing check points and checking for driving under influence.
      * Spring season got more property damage because more months come under spring.
      * This could be partly due to the increased number of travelers during the spring season. Plus, a lot of people on the roads will be from out-of-town, so they won’t be familiar with the area.
      * High school and college students also celebrate spring break. However, many students take this time to let their hair down and party, often by consuming alcohol. Because of this, it’s safe to assume that there could be more drunk drivers on the road.
      * Families often take vacations during the spring so they can spend time together and create fun memories. They may have lost attentions to signs warning them of danger.
      * The evening rush hour 3-6pm, is the most common time for accidents and January with its combination of poor light and low sun along with icy and wet roads presents some of the most challenging driving conditions for motorists
      * Failure to control vehicle speed on highway to avoid collision is the highest reason for the property damage
      * Whites race contribute to property damage for 56.62%
      * Hispanics race contribute to property damage for 3.22%
      * Blacks race contribute to property damage for 29.72%
      * In spite of forming 3.22% in Maryland contribute much higher to property damage
      * This is due to gang and organized crimes.
      * Blacks too show higher levels of crimes even though they form 29.72% of population.
      * As of 2016, number of female drivers are 2160894 while male drivers are 2103981
      * In spite of being higher in number than men, women are less prone to commit property damage then men
      * Commercial vehicles are less prone to accident since the drivers there are professionals and the size of the vehicles make it less incentive to drive rash. Also, the greater accountability of commercial vehicles brings in more responsibility on the driver's shoulders
      * Peoples who belong to their own motherland (Montgomery County) commit more accidents than the outsiders
      * Traffic Violations occur more in intersections areas of the Highway roads.

# CHAPTER 5

MACHINE LEARNING MODELS

## INTRODUCTION

Four Machine Learning Models are used to predict the property damage classes.

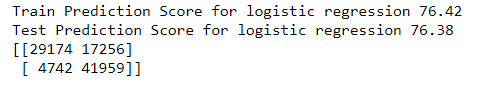
* Logistic Regression
* Decision Tree Classifier
* Random Forest
  1. LOGISTIC REGRESSION

The logistic regression is a predictive analysis.  Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

The general workflow is:

1. Get a dataset
2. Train a classifier
3. Make a prediction using such classifier

**Model Scores**



**Evaluation Metrics**

Precision (also called positive predictive value) is the fraction of relevant instances among the retrieved instances.

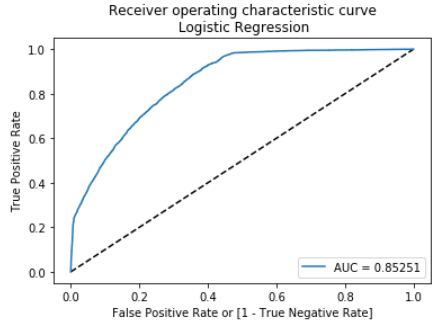


A table of confusion (sometimes also called a confusion matrix), is a table with two rows and two columns that reports the number of false positives, false negatives, true positives, and true negatives.



**ROC Curve for Logistic Regression**

A **Receiver Operating Characteristic curve**, or **ROC curve**, is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied.



* 1. DECISION TREE CLASSIFIER

Decision Tree algorithm belongs to the family of supervised learning algorithms.

The general motive of using Decision Tree is to create a training model which can use to predict class or value of target variables by learning decision rules inferred from prior data (training data).

The general workflow is:

1. Place the best attribute of the dataset at the root of the tree.
2. Split the training set into subsets. Subsets should be made in such a way that each subset contains data with the same value for an attribute.
3. Repeat step 1 and step 2 on each subset until you find leaf nodes in all the branches of the tree.

The popular attribute selection measures:

* Information gain
* Gini index

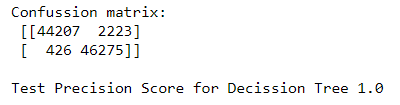
We have chosen ‘Gini index for selection.

Gini Index is a metric to measure how often a randomly chosen element would be incorrectly identified. It means an attribute with lower gini index should be preferred.

**Model Scores**

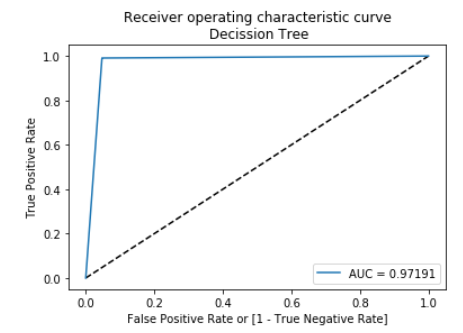


**Evaluation Metrics**

****

**ROC Curve for Decision Tree Classifier**

A **Receiver Operating Characteristic curve**, or **ROC curve**, is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied.

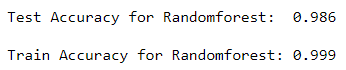
****

* 1. RANDOM FOREST

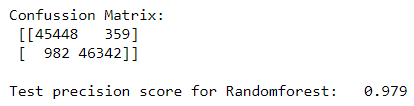
The random forest approach is a bagging method where deep trees, fitted on bootstrap samples, are combined to produce an output with lower variance.

However, random forests also use another trick to make the multiple fitted trees a bit less correlated with each other’s: when growing each tree, instead of only sampling over the observations in the dataset to generate a bootstrap sample, we also sample over features and keep only a random subset of them to build the tree.

**Model Scores**

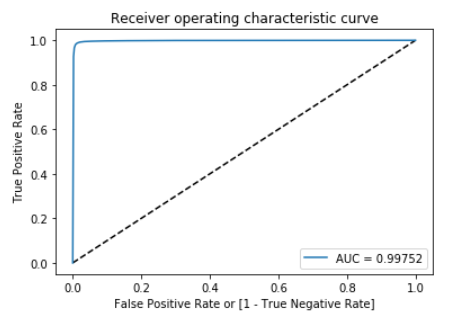


**Evaluation Metrics**

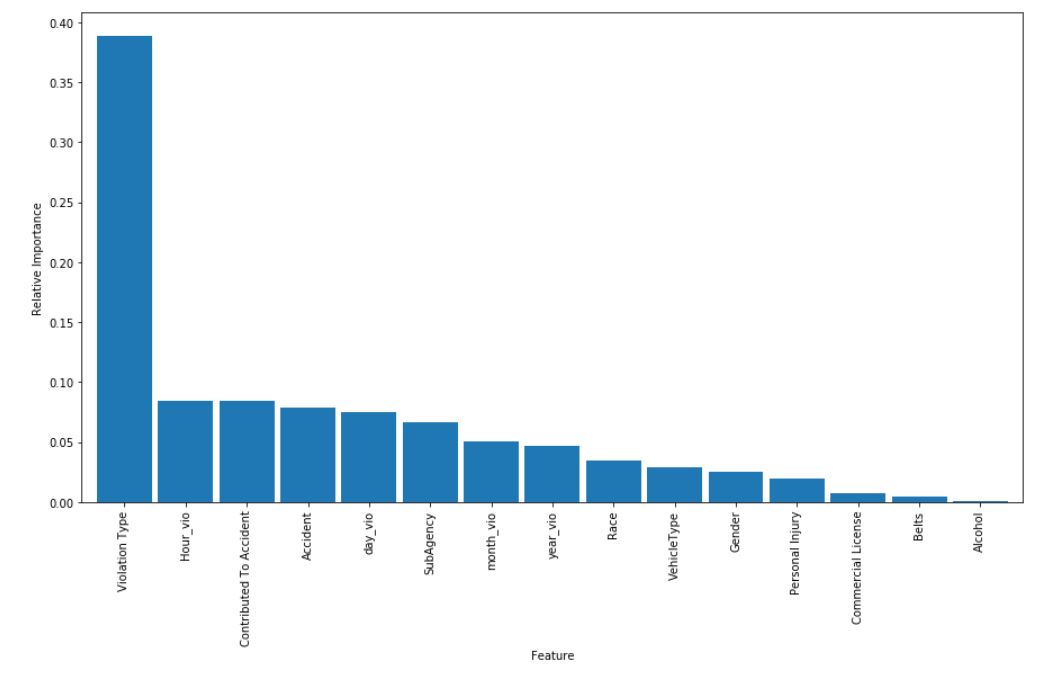
****

**ROC Curve for Random Forest**

A **Receiver Operating Characteristic curve**, or **ROC curve**, is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied.



## FEATURE IMPORTANCE

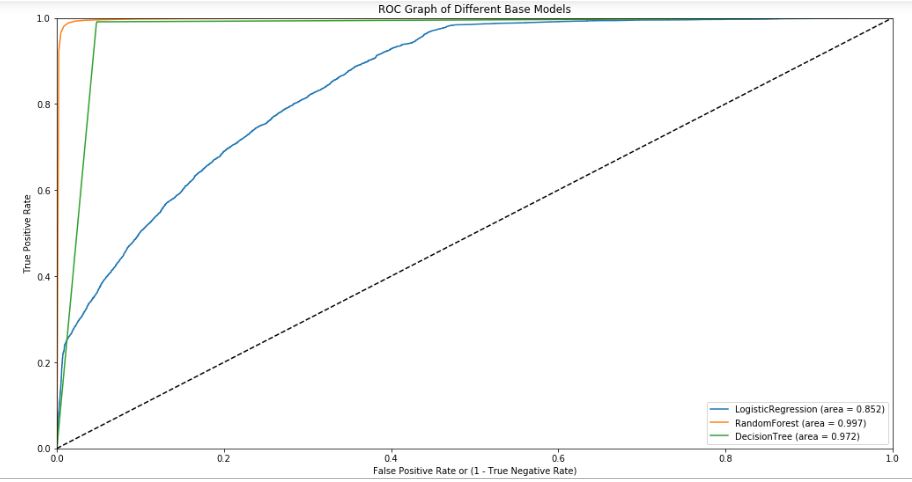


These are the important features that random forest model has predicted. Violation type is the most pertinent feature to classify accidents prone to property damage.

* 1. COMPARING MODEL SCORES

|  |  |  |
| --- | --- | --- |
| Logistic Regression | Decision Tree Classifier | Random Forest |
| Train Prediction Score  76.42 | Train Prediction Score  99.95 | Train Prediction Score  98.6 |
| Test Prediction Score  76.42 | Test Prediction Score  97.16 | Test Prediction Score  99.9 |
| [29174 17256]  [ 4742 41959] | [44207 2223]  [ 426 46275] | [45448 359]  [ 982 46342] |
| AUC: 0.853 | AUC: 0.972 | AUC: 0.998 |

5.7 COMPARING ROC CURVES OF THREE MODELS



We can infer that Random Forest has a better classification capability than other models.

# CHAPTER 6

6.1 CONCLUSION

* Based on the business use case, the independent variable was selected to be ‘Property damage’. The objective was to predict the vulnerability of a vehicle to property damage, using features like Geographical info, Season, features of the Car.
* Data Sampling was done and sample and population characteristics were statistically tested.
* EDA on the dataset gave us insights on the probable factors that contribute to property damage.
* Classification models were built using Logistic Regression, Decision Tree Classifier and Random Forest.
* Through model evaluation Random Forest Classifier was selected.

6.2 RECOMMENDATIONS:

* The project is aimed at assisting Law Enforcement Department in Montgomery County to know about factors which leads to property damage during traffic violations.
* Based on the insights from EDA, accidents prone to damages have a pattern with respect to time, geography and driving errors. The traffic control team could focus on mitigating these factors through road safety planning and enforcing regulations.
* The results from the model prediction can be used as a litmus test during Used Car Inspection for resale. If the model predicts that the vehicle could have been susceptible to property damages, stringent inspection can be done to ensure the vehicle condition.

6.3 LIMITATIONS:

* ‘Accident’ is a chance occurrence, and this project aims at predicting the outcome of such an occurrence with known factors. Hence, the output of this model cannot be taken as a standalone solution. Rather it gives a caution on various risk factors.
* We believe that with more features, especially real time traffic information, construction work details, important events, weather conditions, *etc.*  we can improve this model significantly for real time use cases.

6.4 CLOSING REFLECTIONS:

* This project was our first end-to-end Machine learning case study, and we got to know the kind of issues that crop up during data understanding, data cleaning, model selection and testing.
* It was a team effort that required regular communication, several trial and errors, and brainstorming at each stage.
* What would we do differently next time? Now that we have some experience in data science projects, we would change the thought process during dataset selection. We would first validate the attributes in the dataset to determine if they are sufficient for real time use cases, and then go for the problem approach.