Lydia Teinfalt

Individual Report

DATS 6103 Spring 2021

Updated: 4/30/2021

Introduction

In 2019, the Washington, DC area was ranked 5th for worst traffic in the United States (INRIX, 2020). During 2020 with the Covid-19 pandemic, the DC area saw 77% less congestion and improved its status by no longer being in the top 10 cities with worst traffic (INRIX, 2021). Our group analyzed DC Crash data available at the Open Data DC website from DC Government. With the introduction of the Covid-19 vaccine and more cars coming back to the roads, we wanted to understand the factors involved in traffic crash where major injuries or fatalities were incurred.

Description

Our group was going to use two datasets: Crashes in D.C and Crashes Details Table from the D.C. government—The District Department of Transportation (DDOT) and The Metropolitan Police Department (MPD). *Crashes in D.C.* dataset provides overall information about the crash event whereas *Crashes Details* dataset is a companion and provides information about the individuals involved in the crash.

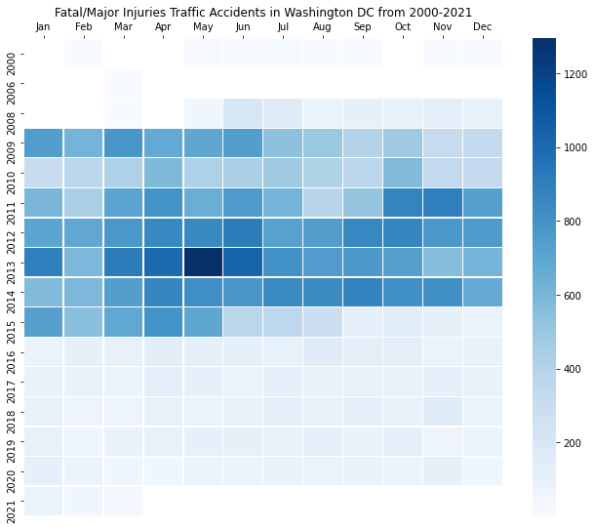
Table 1 Data Sample of Crashes Details DC from March 2020

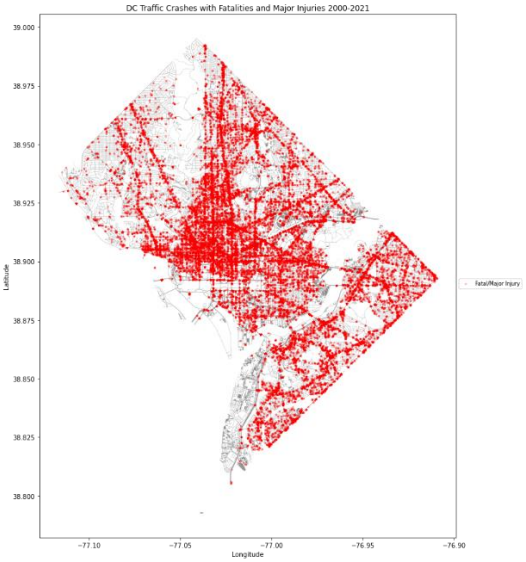
OBJECTID	CRIMEID	CCN	PERSONID	PERSONTY	AGE	FATAL	MAJORINJ	MINO	VEHICLEID	INVEHICLETYPE	TICKETISSU	LICENSEPL	IMPAIRED	SPEEDING
430455865	27615913	18042044	86838139	Driver	49	N	N	N	3766128	Large/heavy Truc	N	MD	N	N
430455866	27615913	18042044	86838245	Driver	59	N	N	Y	3766126	Passenger Car/at	Y	VA	N	N
430455867	27615913	18042044	86836893	Driver	61	N	N	N	3766127	Bus	N	PA	N	N
430455868	26873834	16035157	84968953	Driver	28	N	N	Y	2277107	Passenger Car/at	Y	VA	N	N
430455869	26873834	16035157	84921236	Passenger	33	N	N	N	2277107	Passenger Car/at	N	VA	N	N
430455870	26873834	16035157	84748308	Driver	63	N	N	N	2277106	Passenger Car/at	Y	DC	N	N
430455871	26873836	16035159	84962811	Driver	37	N	N	N	2277098	Passenger Car/at	Y	DC	N	N
430455872	26873836	16035159	84570868	Driver	45	N	N	N	2277099	Other Vehicle	Y	None	N	N
430455873	26873838	16035120	84584071	Driver		N	N	N	2277108	Passenger Car/at	N	DC	N	N
430455874	26873838	16035120	84936111	Driver	67	N	N	N	2277108	Passenger Car/at	N	DC	N	N
430455875	26873846	16035140	84956752	Driver	35	N	N	N	2277103	Passenger Car/at	N	MD	N	N

There are 599,670 observations and 15 columns in the Crashes Details table. I added a new column called “FATALMAJORINJURIES” to the dataset. This column is set to ‘1’ if the original two columns “FATAL" or "MAJORINJURY" is equal to Y (YES). The “FATALMAJORINJURIES” column is the target.

EDA

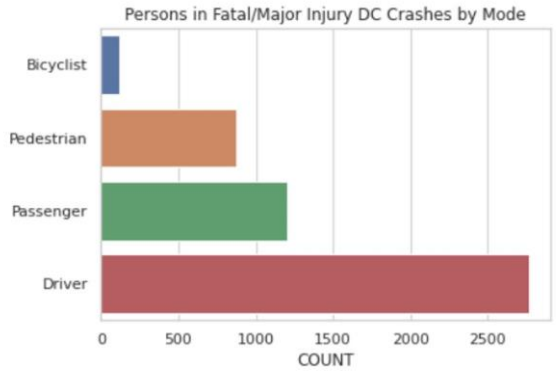
I created readdata.py class to read the two csv files and use the merge method on CRIMEID to combine the two datasets into a single pandas dataframe. The resulting table was a cartesian product of the two datasets -- Crashes in D.C. has a one-to-many relationship with Crashes Details table. The merge method used kept data from both dataset therefore, creating duplicate longitude and latitude coordinates with a single crash event. When we found this flaw with the way we were merging the two datasets, our group dropped the Crashes in DC dataset and focused on Crashes in Details dataset and so the following charts were not used in our EDA.





The left chart breaks down crashes by year and month. The upper portion of the chart shows a steady build-up of serious accidents from the early 2000s to a peak in May of 2013 with the decline later in 2015. The DC major introduced a project called "VISION ZERO" in 2015 which was designed to reduce accidents and make DC streets safer. The second chart is a map of DC and uses geopands to place a red dot to represent a crash resulting in fatality or major injuries from 2000-2021.

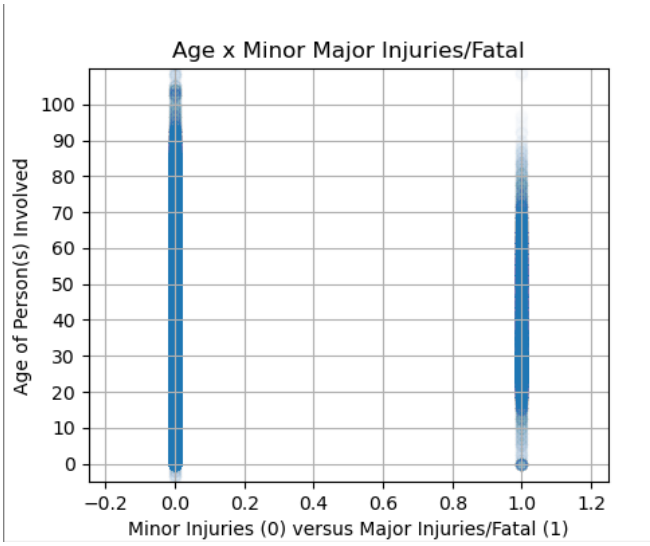
I created one chart describing Fatal/Major Injury crashes by the mode of the persons involved. A clear majority of the cases involved was the driver followed by passenger, pedestrian and bicyclist.



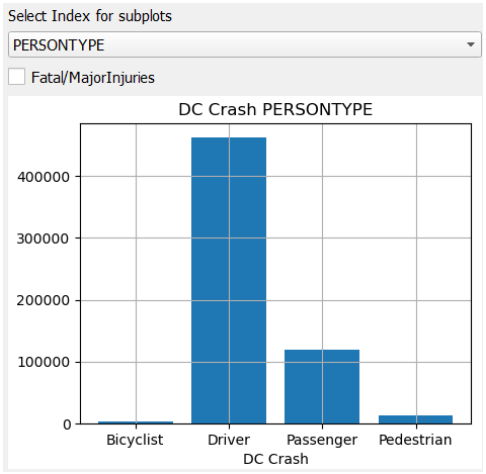
For the project, I did not implement models. My main task was to focus on re-using the PyQt5 demo you provided and customize it to showcase the models developed by Arianna and RyeAnne.

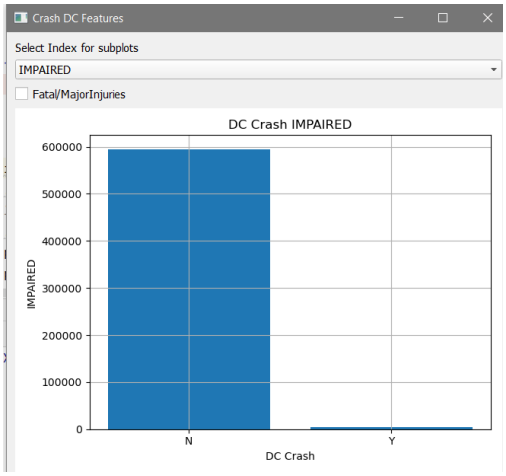
GUI – Incorporating EDA and Models

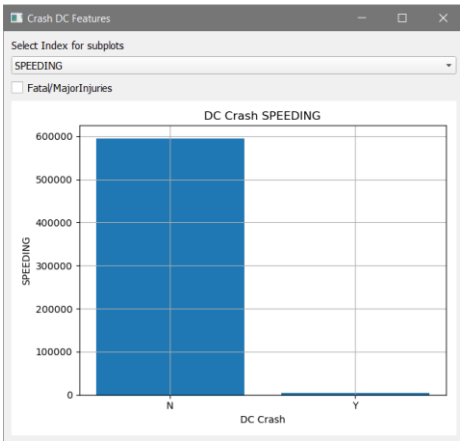
From the crash's GUI application, the Data menu – Sample displays the a few rows data from DC Crash Details table after we did an initial clean up of the data. Under EDA menu are three exploratory data results. The first is Age Histogram developed by Arianna Dunham which displays "Age of People Involved in Traffic Accidents with Fatalities or Major Injuries" where Age is on the x-axis and number of crashes on the y-axis. I developed the second EDA "Age Scatter" plot. The plot show that the dataset had negative ages and ages above 110 years old.

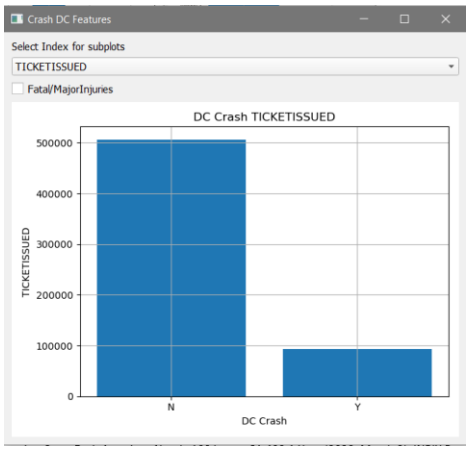


The third EDA menu option Crash Graphs allows user interaction where users can view the number of crashes based on impaired, speeding, ticket issued, or person type. Additional check mark will subset the result to crashes involving a fatality and/or major injury. The graphs debunk the myth that DC crashes were result of an impaired driver, speeding or where a ticket was issued.



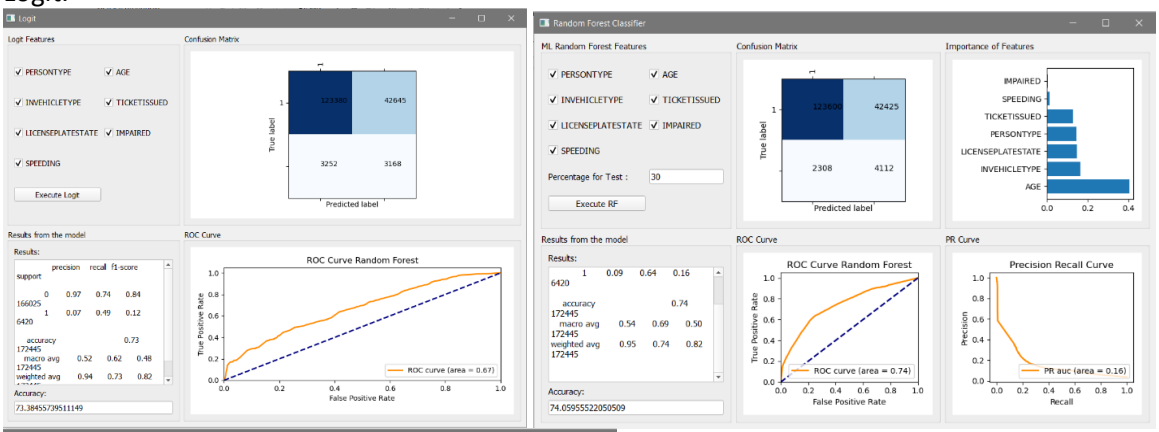






RESULTS

The ML Models menu has three models developed by the RyeAnne Ricker and Arianna Dunham: Logit, Random Forest Classifier and XGBoost Classifier. I only incorporated their code into classes for GUI (Main.py). The GUI incorporated results by displaying classification report, accuracy, confusion matrix, importance of features, ROC Curve and Precision Recall curve. For the sake of visualization, I added a Precision Recall (PR) curve for the Random Forest Classifier and the XGBoost DT. From the measure of AUC and PR AUC, XGBoost had the best value of 0.76 for AUC and PR AUC of 0.18 compared to RF or Logit.



Summary

The crashes details table describing traffic accidents in DC is an imbalanced dataset leaning heavily towards crashes with minor injuries. Out of a total of 599,670 crashes data rows, only 3.64% of represented ones with fatality or major injury. The crash details table did not contain any information about when the crash occurred, so one could not see if the occurrence of crashes increasing or decreasing over time. Age was the only quantitative feature in the dataset but we had bad data with negative age values or ages greater than 110 years old. Person type describes the mode of transportation that an individual was involved in the crash -- drivers were involved in most of the cases. For traffic accidents that resulted in fatalities or major injuries, the pedestrian count is greater than passenger count. I reused the logistic regression, random forest and XGBoost models created Arianna Dunham and RyeAnne Ricker for the desktop application/ GUI. For the random forest and XGBoost model, I introduced a Precision – Recall curve for the sake of visualization. The XGBoost had a slightly better PR AUC value of 0.18 than the random forest model (PR AUC = 0.16). Finally, I set up the GitHub

repo for our project and spent time helping my teammates clone the repo, push up changes, and troubleshoot any issues that arose during the project.

Percentage of Code

Lines of code from the internet (LI)= 1333

Modified lines of code (ML)= 512

New lines of code (NL)= 219

Percentage of code = (LI – ML)/(LI + NL) X 100 = 52.90%

References

INRIX: Congestion Costs Each American Nearly 100 hours, $1,400 A Year (2020, March 9). INRIX Press Release, https://inrix.com/press-releases/2019-traffic-scorecard-us/, (Accessed April 28, 2021)

INRIX: Downtown Travel Plummets 44% in 2020 amid COVID-19 Pandemic (2021, March 9). INRIX Press Release, https://inrix.com/press-releases/2020-traffic-scorecard-us/ , (Accessed April 28, 2021)

District Department of Transportation, Metropolitan Police Department, *Crashes Details Table*, Open Data DC, (District of Columbia): Vision Zero Data Planning Work Group, 2020. Accessed on: Mar. 14, 2020. [online]. Available: **https://opendata.dc.gov/datasets/crash-details-table**