D.C. Crashes:
What factors contribute to major injuries and fatalities?

- DATS 6103 (SPRING 2021) GROUP 3:
- ARIANNA DUNHAM
- RYEANNE RICKER
- LYDIA TEINFALT

Crash-Details-Table

OBJECTID	CRIMEID	CCN	PERSONID	PERSONTY	AGE	FATAL	MAJORINJ	MINO	VEHICLEID	INVEHICLETYPE	TICKETISSI	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/au	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/au	Y	VA	N	N
430455869	26873834	16035157	84921236	Passenger	33	N	N	N	2277107	Passenger Car/au	N	VA	N	N
430455870	26873834	16035157	84748308	Driver	63	N	N	N	2277106	Passenger Car/au	Y	DC	N	N
430455871	26873836	16035159	84962811	Driver	37	N	N	N	2277098	Passenger Car/au	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/au	N	DC	N	N
430455874	26873838	16035120	84936111	Driver	67	N	N	N	2277108	Passenger Car/au	N	DC	N	N
430455875	26873846	16035140	84956752	Driver	35	N	N	N	2277103	Passenger Car/au	N	MD	N	N

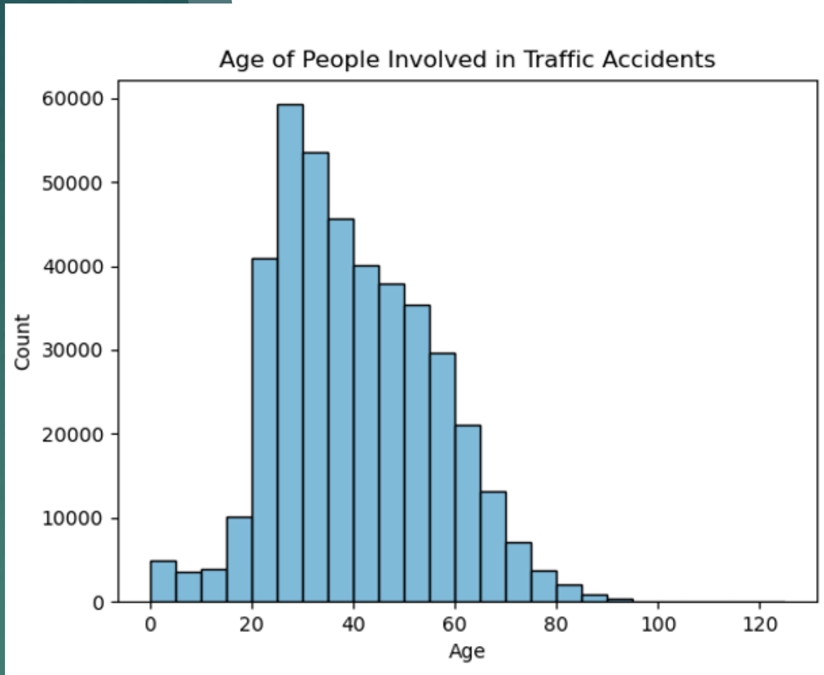
Crashes Details Table: https://opendata.dc.gov/datasets/crash-details-table

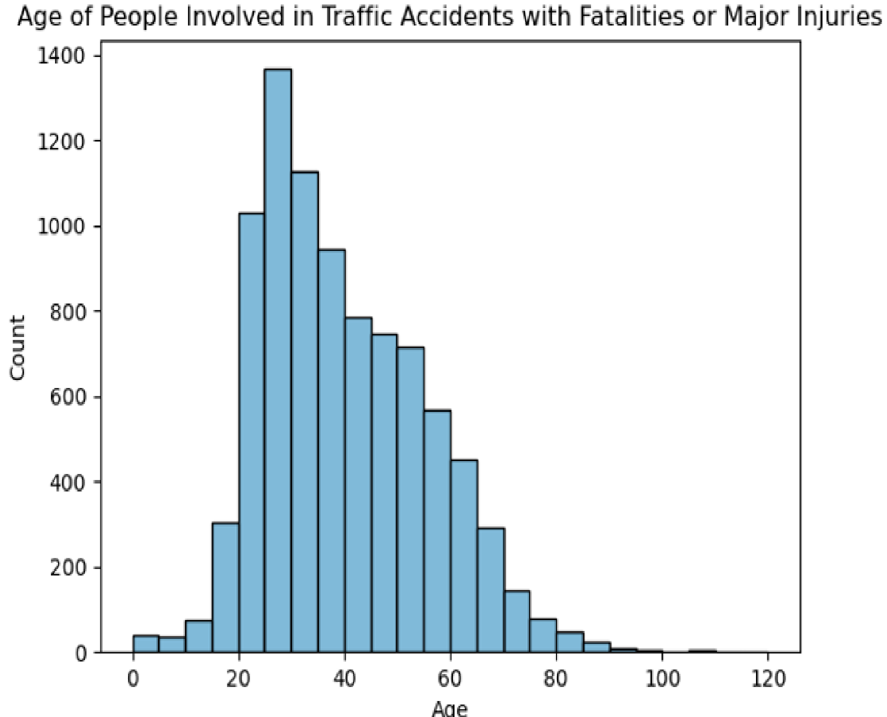
- Total number of rows: 599,670
- Total number of columns: 15
- Multiple people can be involved in a crash assigned a unique CRIMEID. Each person will have a unique PERSONID
- If an individual was involved in more than one crash, their PERSONID will have multiple data rows
- FATAL, MAJORINJURY, MINORINJURY, TICKETISSUED, IMPAIRED, and SPEEDING have Y/N values
- INVEHICLETYPE contain categorical data of vehicle type involved in the crash
- LICENSEPLATESTATE is a two-letter abbreviation of state where the plate was issued
- Original column names were capitalized. New column added with all capitals FATALMAJORINJURY = 1 if FATAL or MAJORINJURY = Y default value of column is 0

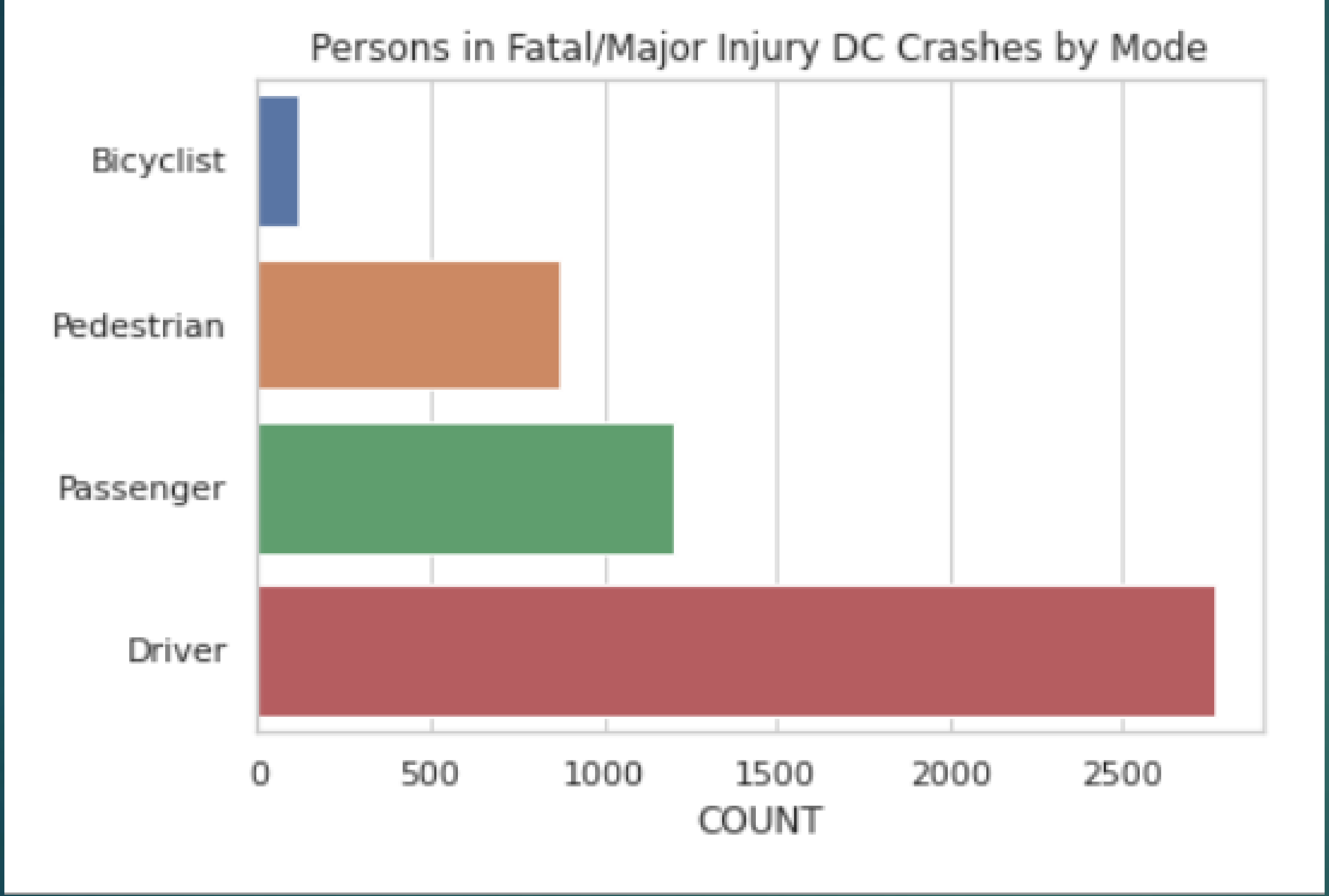
EDA

Age

- Mean age overall: 39
- Mean age with major injuries/fatalities: 34



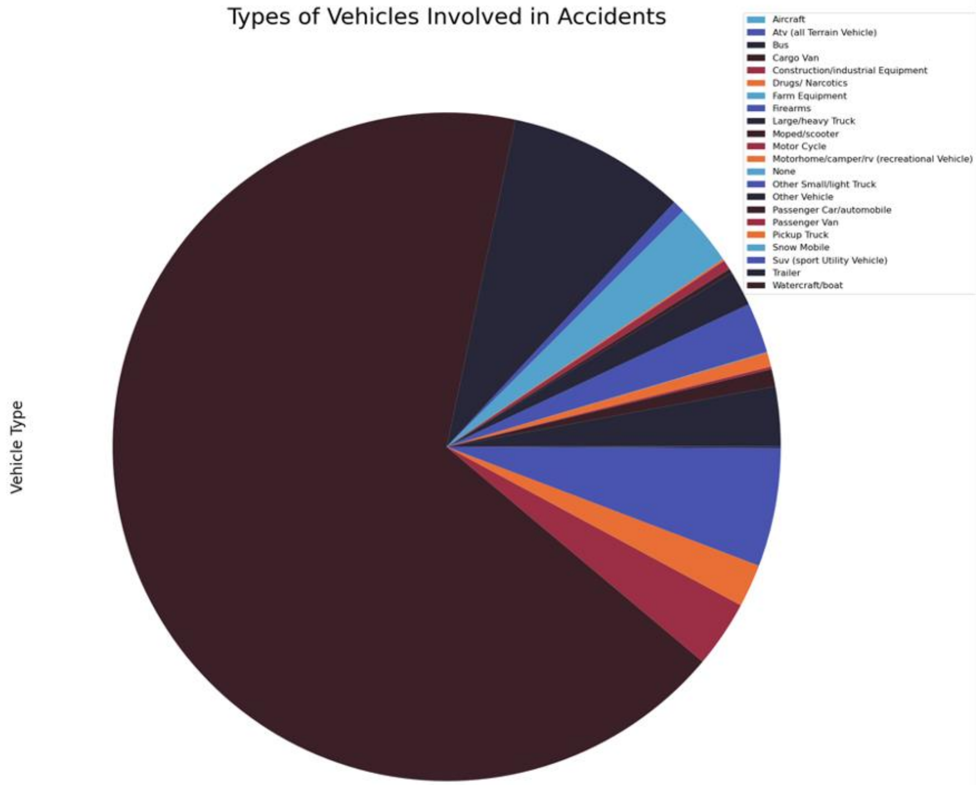


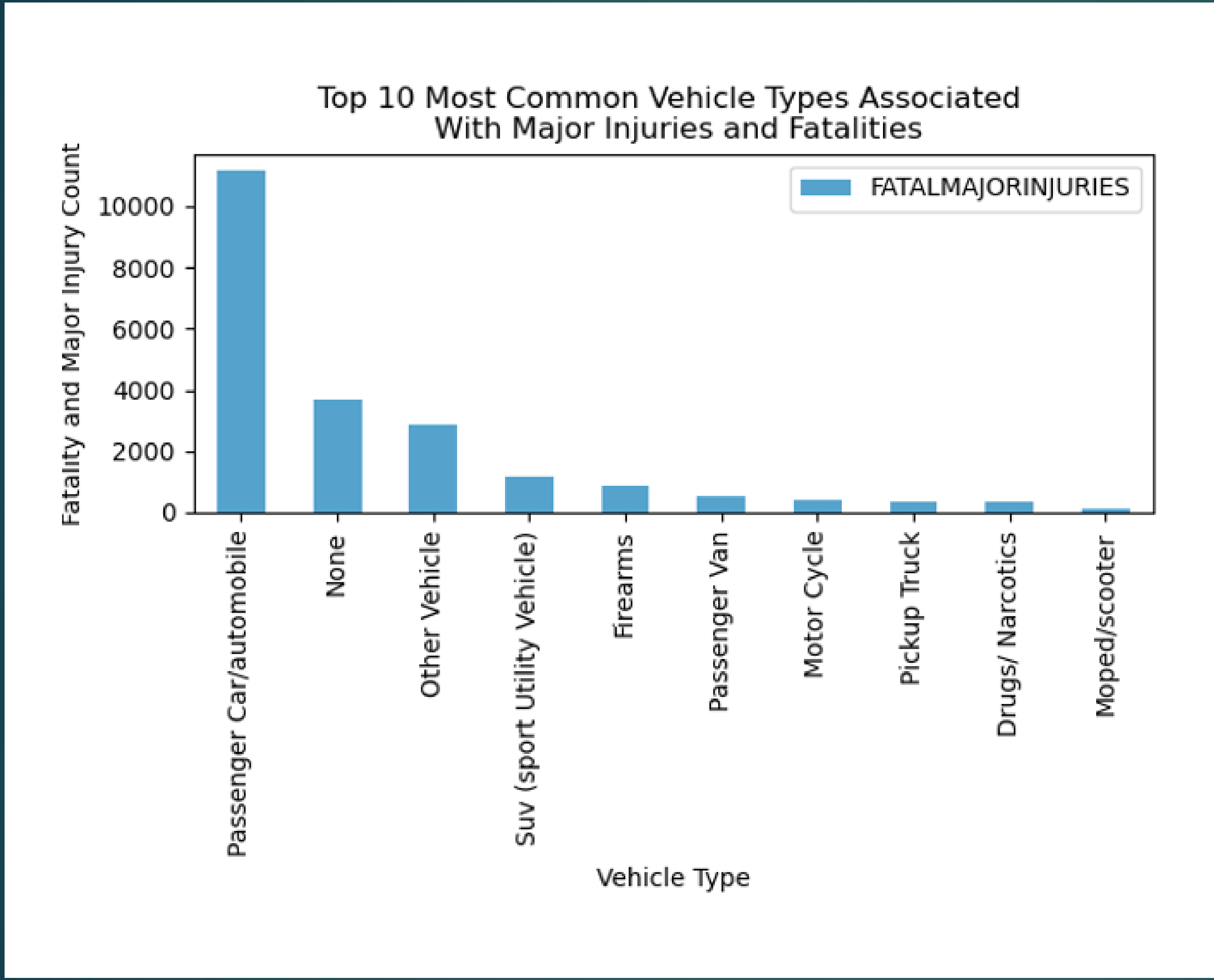


- Crashes by the mode of transportation for the persons involved
- Driver clearly majority of cases
- Passenger being second
- Pedestrians involved in crashes almost as many as passengers
- Bicyclists being safest mode of transportation

Vehicle Type

- 22 different vehicle types
- Passenger Car is most common is vehicle type





Vehicle Type

Statistics

Chi-Squared Test for Independence

Note:

>0.10 moderate

>0.15 strong

>0.25 very strong

	Fatal/Major Injury Occurrence
Speeding	0.02
Ticket Issued	0.07
Vehicle Type	0.18
License Plate State	0.10
Impaired	0.01
Person Type	0.18

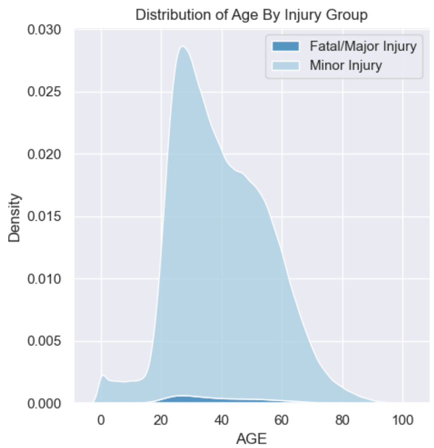
Summary Statistics	
Minimum	0.0
Median	38.0
Mean	39.75
Maximum	100.0
Standard Deviation	15.62

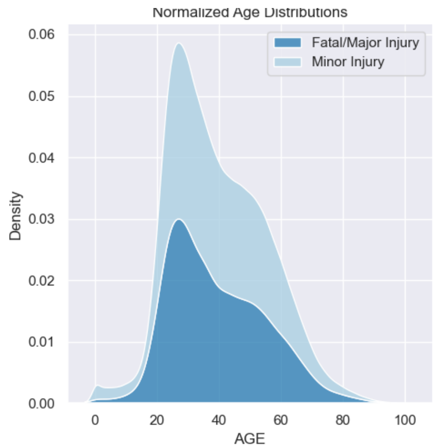
Quantitative Variable: Age

PDFs - Age

- *t*-test to compare the mean Age of those acquiring a major injury/fatality vs minor injury
- p-value = 0.014
- Mean Age Fatality/Major Injury = 39.3 yo
- Mean Age Minor Injury = 39.7 yo

$$d = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$





Data Preprocessing

- Remove Identifiers
- Cleaning Data:
  - Removed rows with ages <0 and >100
  - Removed Drivers that were <10 yo
  - Removed Nonsense License Plate States (Ot, Ou, Vi, Pu, Un, Am, Di)
- Missing Data:
  - Removed 328 Empty Rows
  - Filled in missing Ages using the mean age
- Label Encoder
- Normalization of Age

Features Used

- Impaired: Categorical (Y/N)
- Age: Numerical
- Vehicle Type: Categorical (14 possibilities))
- Ticket Issued: Categorical (Y/N)
- Speeding: Categorical (Y/N)
- State of License Plate: Categorical
- Person Type: Categorical (Driver, Passenger, Pedestrian, Other)

Target

DID A MAJOR INJURY OR FATALITY OCCUR (Y/N)

INDIVIDUALS WITH A FATALITY/MAJOR INJURY: 21,772 OR 3.7%

INDIVIDUALS ACQUIRING A MINOR INJURY: 572,077 OR 96.3%

Machine Learning Algorithms Used

- Naïve Bayes
- Decision Trees:
  - Extreme Gradient Boosted DT
  - Random Forest
- Logistic Regression
- Voting Classifier

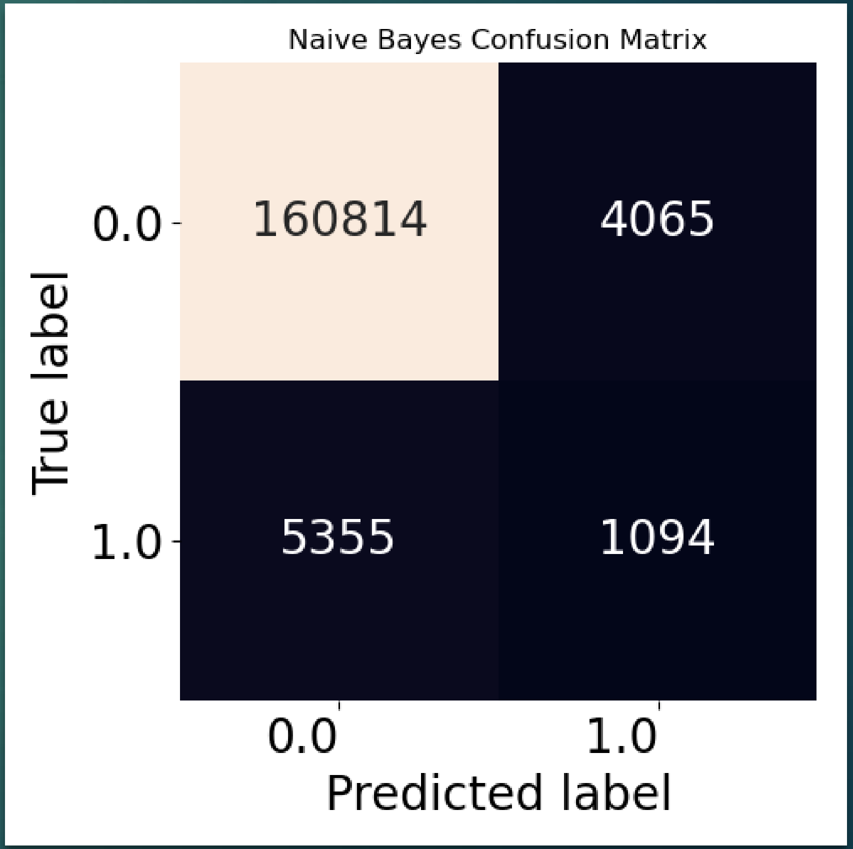
Naive Bayes

Overall Accuracy: 95.5%

AUC Accuracy: 0.70

Specificity: 97.5%

Sensitivity: 17.0%



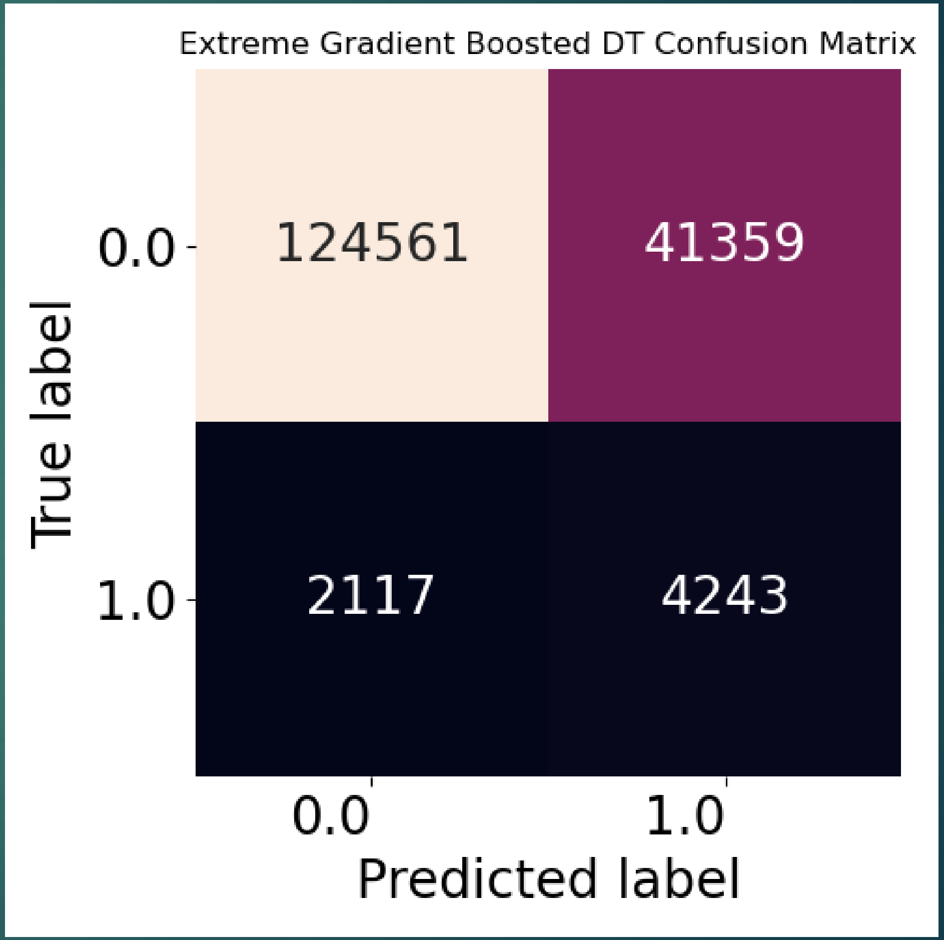
XGBoost Decision Tree

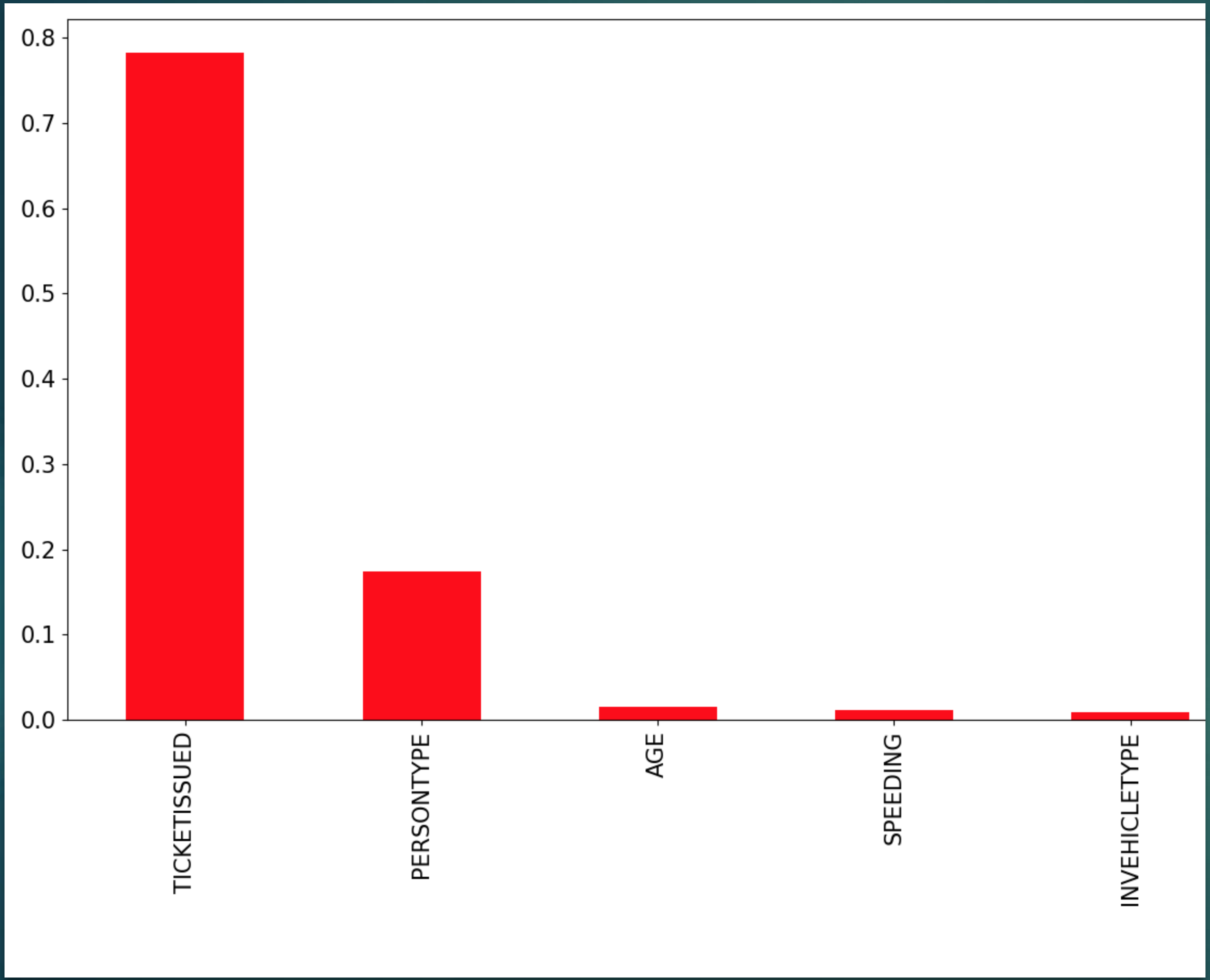
Overall Accuracy: 74.8%

AUC: 0.768

Specificity: 75.1%

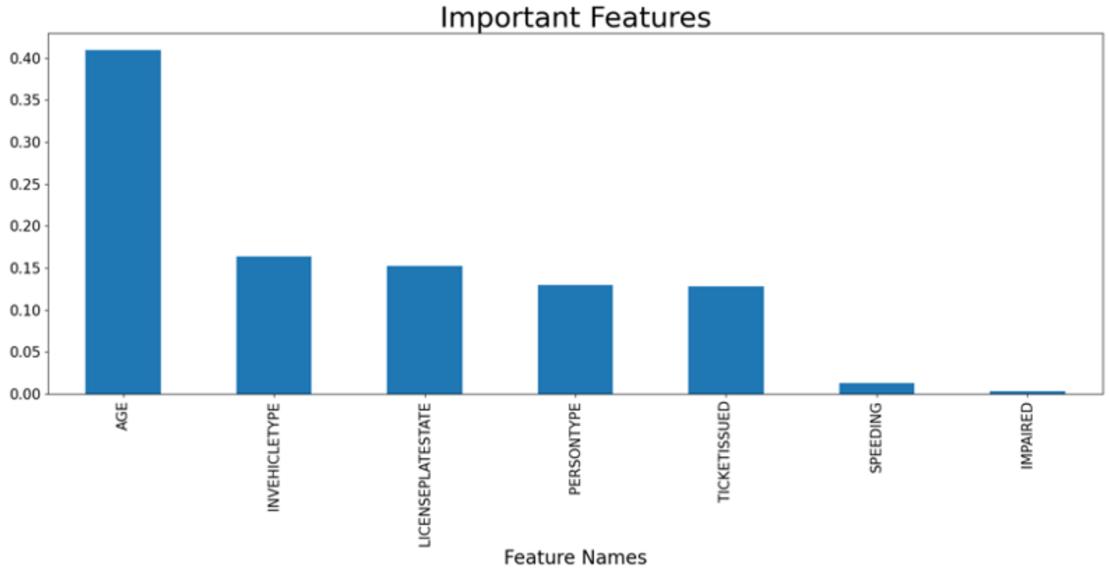
Sensitivity: 66.7%





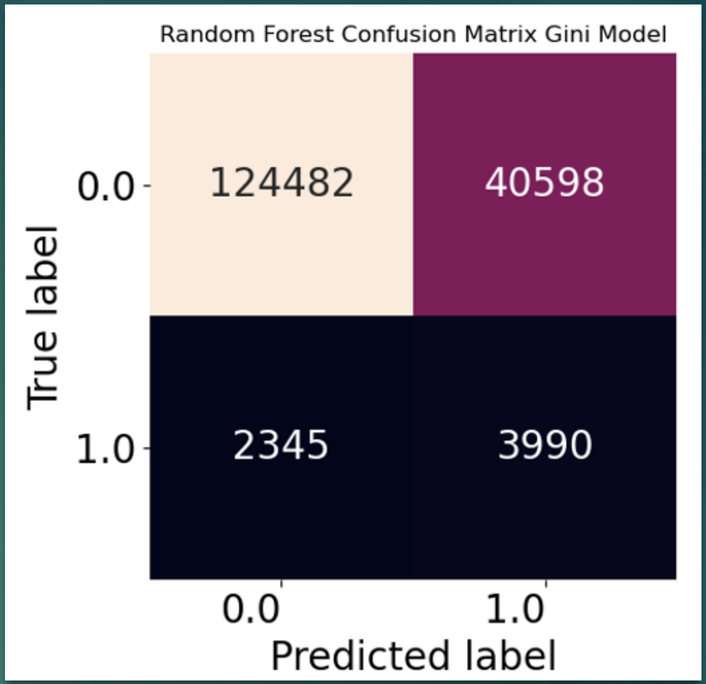
Feature Importance - XGBoost

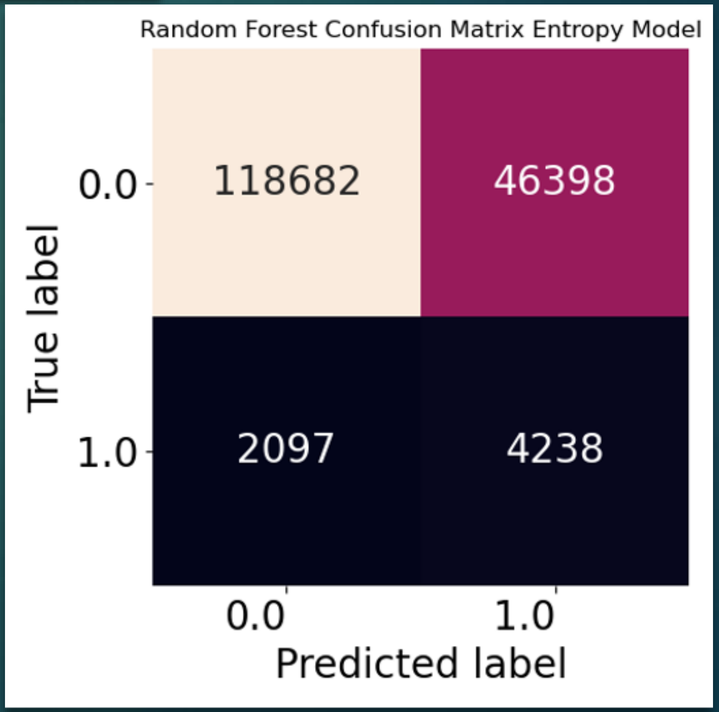
Random Forest



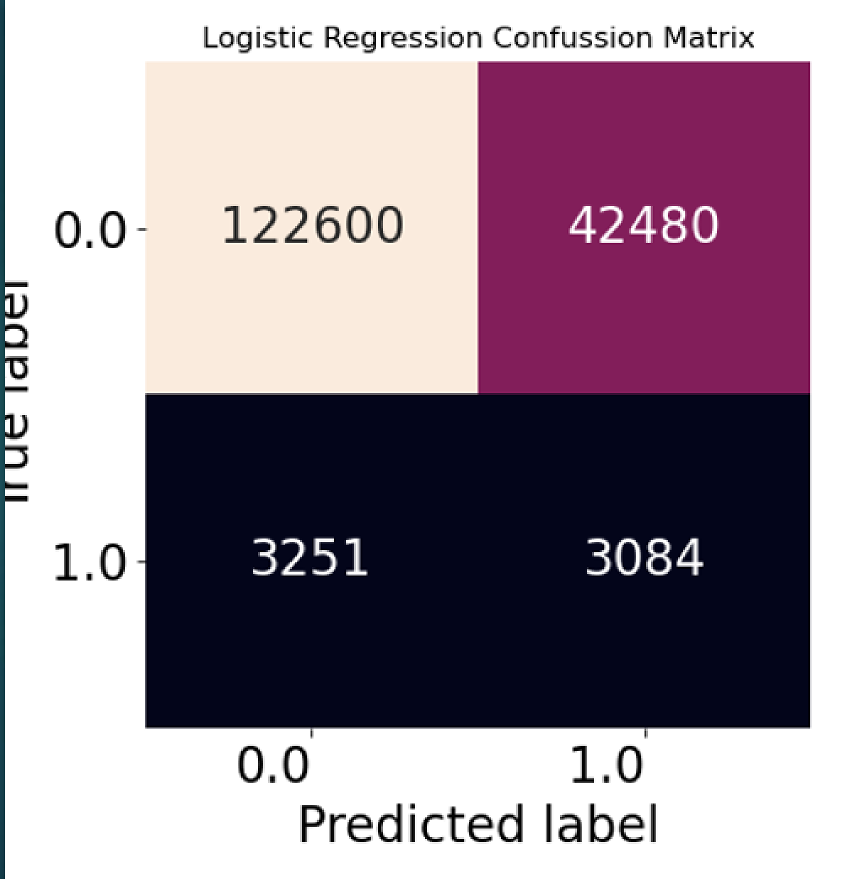
Random Forest

- Gini AUC: .74
- Gini Classification Accuracy: 74.95%
- Entropy AUC: .749
- Entropy Classification Accuracy: 71.71%
- Specificity: 63%
- Sensitivity: 75.41%





Logistic Regression



- AUC: .664
- Classification Accuracy: 73.32
- Sensitivity: 74.26
- Specificity: 48.68

Voting Classifier – Logistic Regression, Random Forest, XGBoost

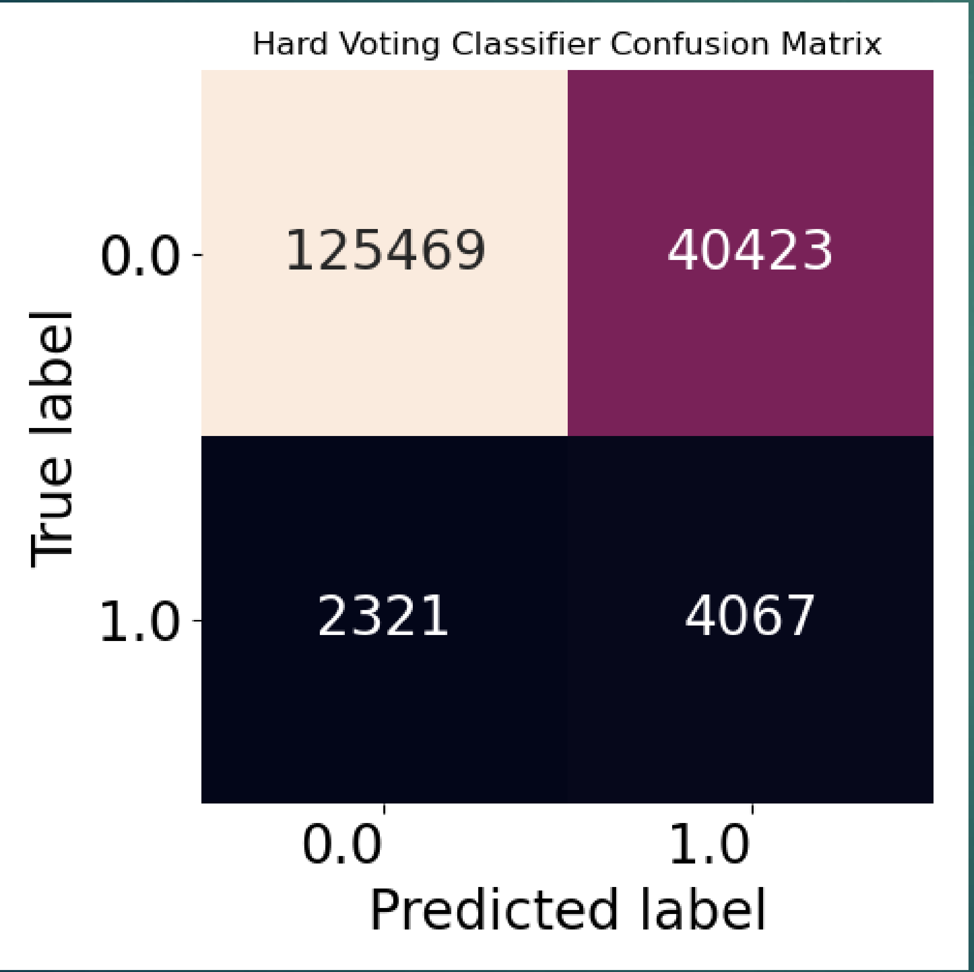
Hard Voting

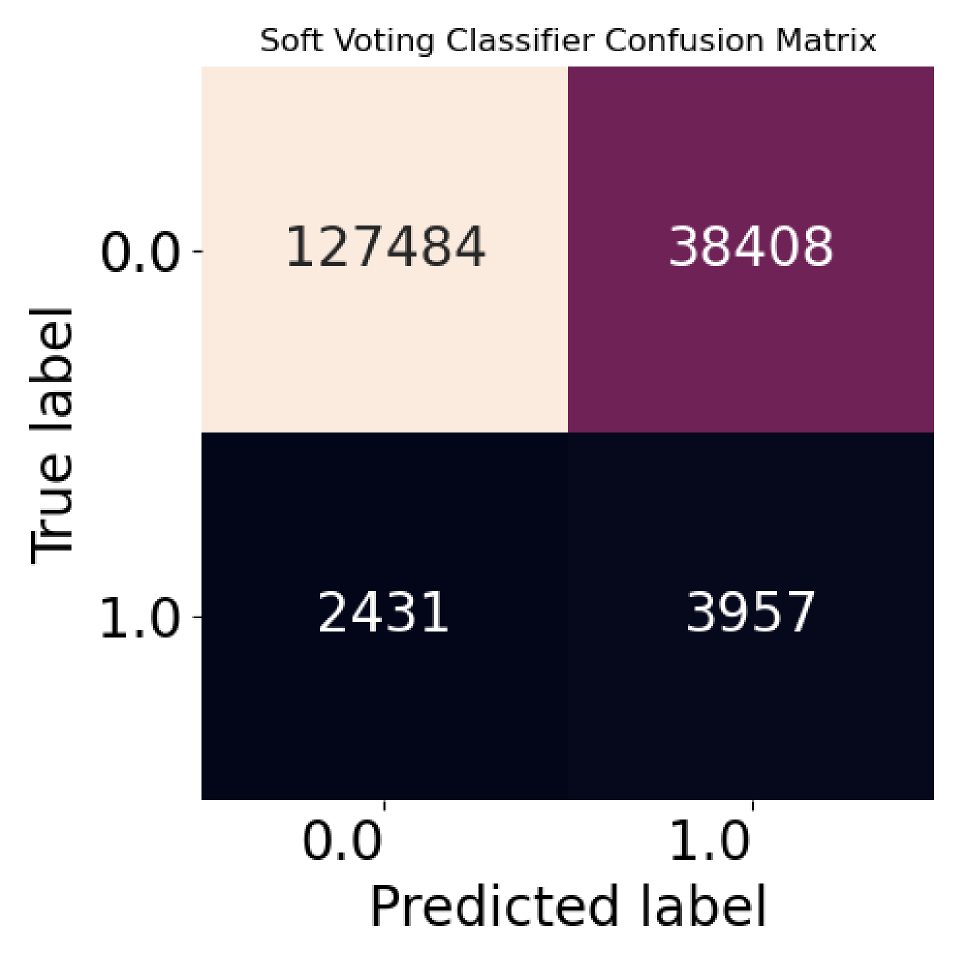
- Accuracy: 75.2%
- Specificity: 75.6%
- Sensitivity: 63.7%

Soft Voting

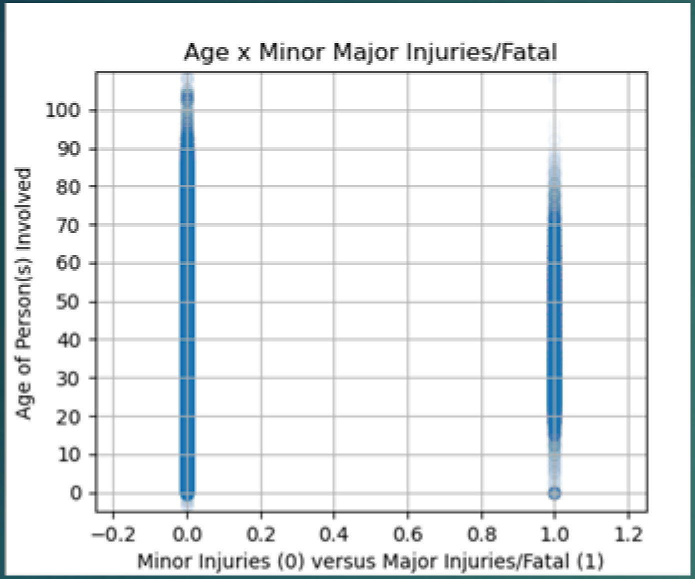
- Accuracy: 76.3%
- Specificity: 76.8%
- Sensitivity: 61.9%

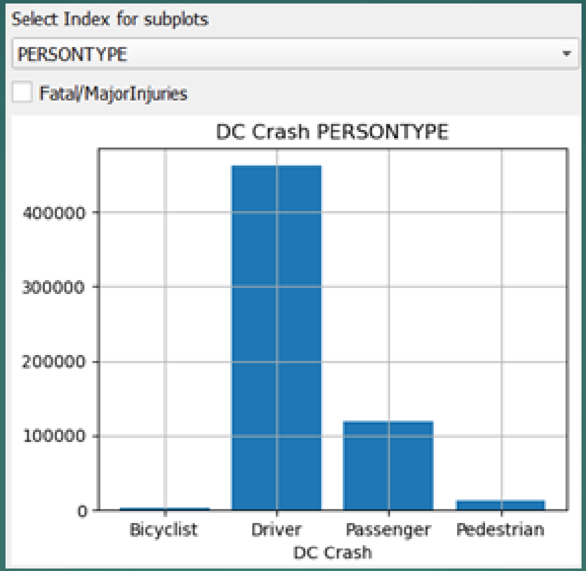
Voting Classifier Confusion Matrix

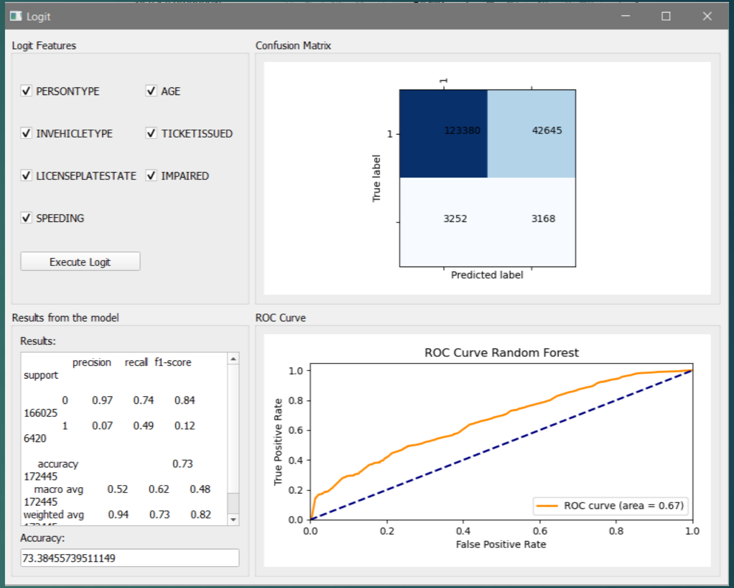


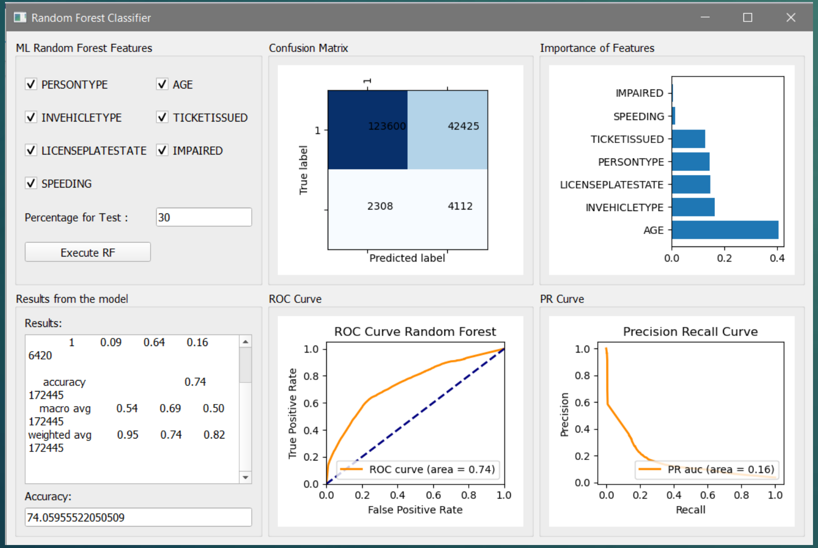


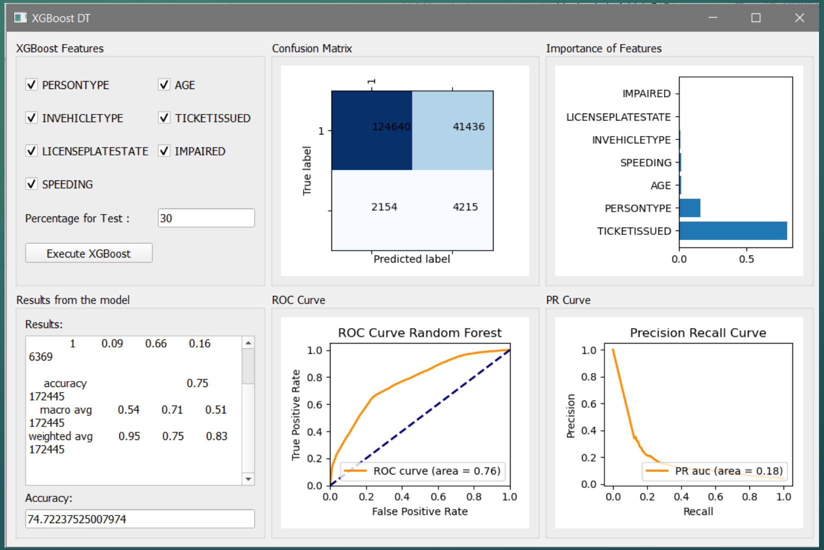
GUI











- Simple application running on Windows or Mac OS
- Users can interactively view data, run EDA and execute models.

Best Classifier

	Naïve Bayes	XGBoost	Random Forest	Logistic Regression	Voting Classifier
Overall Accuracy	95.5%	74.8%	74.9%	77.3%	76.3%
AUC	0.70	0.768	0.74	0.664	---
Sensitivity	17.0%	66.7%	74.7%	74.3%	61.9%
Specificity	97.5%	75.1%	63.5%	48.7%	76.8%

Conclusions

- Models predict whether an individual will experience a major injury or fatality better than a random guess
- Most important parameter: Class Weights
- Highest AUC = XGBoost
  - BUT it is costly in time
- Most Sensitive: Random Forest
- Most Specific: Voting Classifier (ignoring Naïve Bayes due to low sensitivity)
- Worst overall model = Naïve Bayes
- Voting Classifier did not substantially increase the accuracy of the model in terms of sensitivity or specificity

References

- 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: March. 11, 2021. [online]. Available: **https://opendata.dc.gov/datasets/crash-details-table**
- "VotingClassifier". *Sklearn.* [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.VotingClassifier.html
- "ML Voting Classifier Using Sklearn". *GeeksforGeeks.* Nov 25, 2019. [Online]. https://www.geeksforgeeks.org/ml-voting-classifier-using-sklearn/
- Navlani, Avalash. "Naive Bayes Classification Using Scikit-Learn". *DataCamp.* Dec. 4, 2018. [Online]. https://www.datacamp.com/community/tutorials/naive-bayes-scikit-learn?utm_source=adwords_ppc&utm_campaignid=1565261270&utm_adgroupid=67750485268&utm_device=c&utm_keyword=&utm_matchtype=b&utm_network=g&utm_adpostion=&utm_creative=332661264374&utm_targetid=aud-299261629574:dsa-429603003980&utm_loc_interest_ms=&utm_loc_physical_ms=9007810&gclid=Cj0KCQjwvYSEBhDjARIsAJMn0lj1DfpdDWQ5NbCTjk8GlsSJ21kKd8WcdrU5FLhU1Yy7NYkOM3vHUikaAuUREALw_wcB
- Scikit-learn: Machine Learning in Python, Pedregosa *et al.*, JMLR 12, pp. 2825-2830, 2011.[Online] https://scikit-learn.org/stable/auto_examples/model_selection/plot_precision_recall.html. Acessed Apr. 25, 2021.
- Pathak, Manish. "Using XGBoost in Python". *DataCamp.* Nov. 8, 2018. [Online]. https://www.datacamp.com/community/tutorials/xgboost-in-python