License and Registration

Exploring the data behind police stops and searches


007

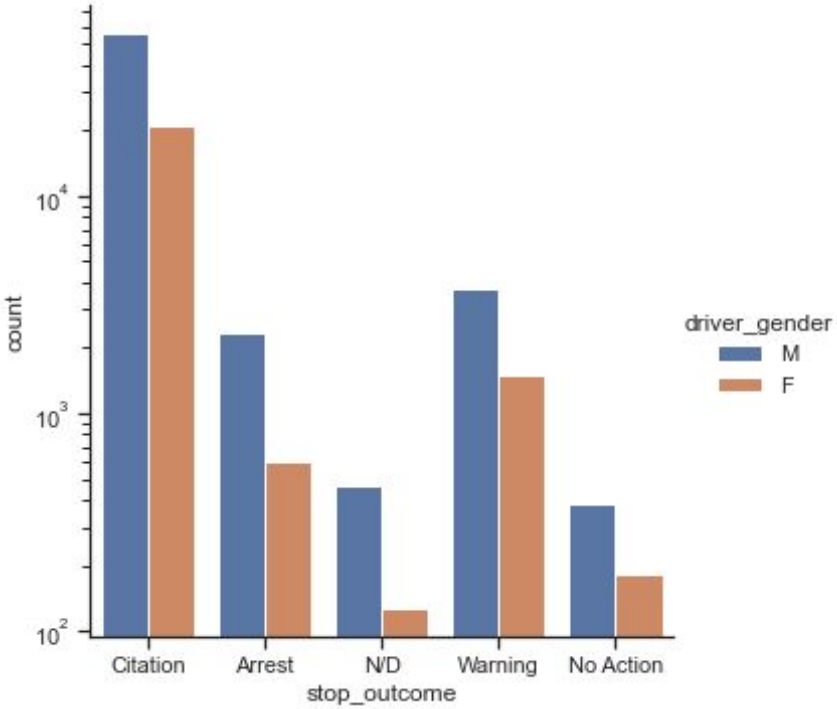
Importance

About the Data

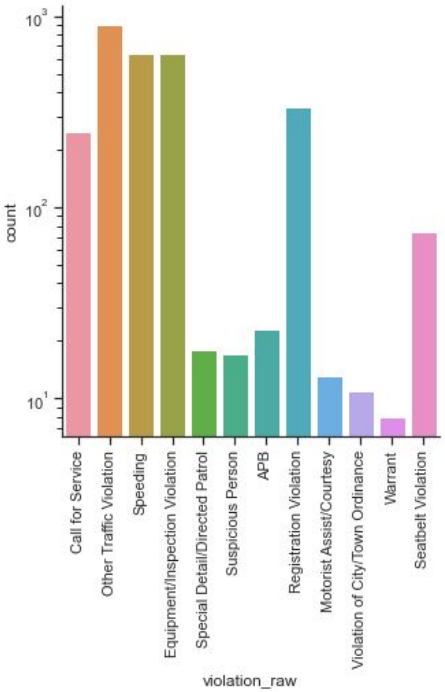
- Target variable is stop outcome
- 14 predictor variables
- 91,741 records

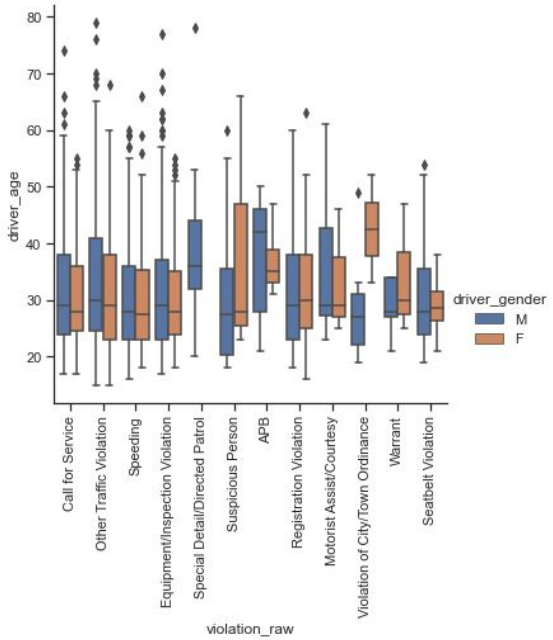
	38695	86340	80059	24908	22532
stop_date	2009-11-30	2015-05-28	2014-08-21	2008-04-05	2007-12-23
stop_time	07:54	10:36	23:56	16:22	15:20
county_name	NaN	NaN	NaN	NaN	NaN
driver_gender	M	M	F	M	M
driver_age_raw	1983.0	1992.0	1991.0	1944.0	1983.0
driver_age	26.0	23.0	23.0	64.0	24.0
driver_race	White	Black	Black	White	Hispanic
violation_raw	Speeding	Speeding	Registration Violation	Speeding	Speeding
violation	Speeding	Speeding	Registration/plates	Speeding	Speeding
search_conducted	False	False	False	False	False
search_type	NaN	NaN	NaN	NaN	NaN
stop_outcome	Citation	Citation	Citation	Citation	Citation
is_arrested	False	False	False	False	False
stop_duration	16-30 Min	0-15 Min	0-15 Min	0-15 Min	0-15 Min
drugs_related_stop	False	False	False	False	False

Raw Data Analysis

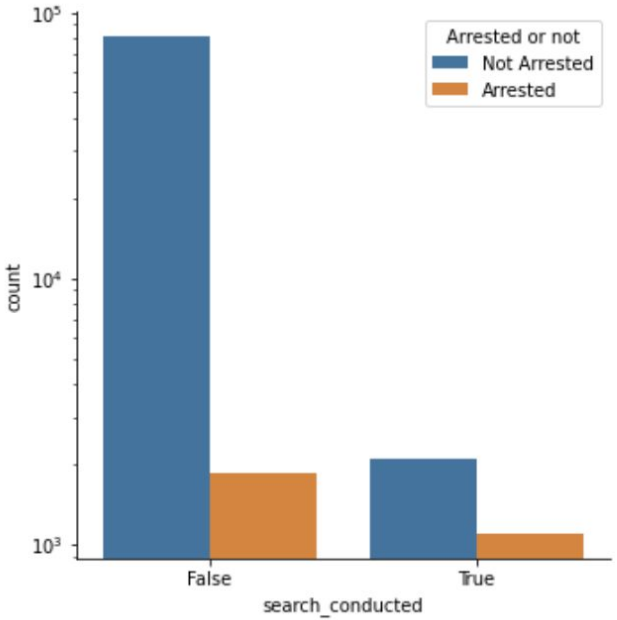


Citations account for ~70,000 instances of the stop outcome (target variable)

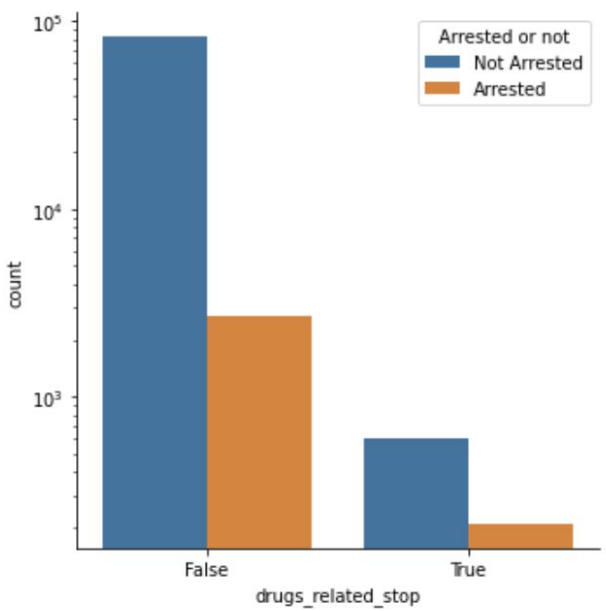




Isolating the data for the other stop outcomes, we find that oversampling techniques will yield accurate results



There is a higher proportion of people who have been searched getting arrested than people who have not been searched.



There is a higher proportion of people who did not have drug related stops getting arrested than those who have.

	mean		count	
is_arrested	False	True	False	True
drugs_related_stop				
False	34.13438	32.034267	82587	2714
True	27.85738	29.000000	603	211

1)Male drivers who are stopped because of drugs are generally older.

2)Among people who are stopped because of drugs, those that are not arrested are generally younger than those who are arrested on average.

3)Among people who are stopped because of other reasons, those that are arrested are generally younger than those that are not.

		mean		count	
	is_arrested	False	True	False	True
violation_raw	driver_gender				
APB	F	33.250000	37.000000	16.0	5.0
	M	36.743590	37.333333	39.0	18.0
Call for Service	F	33.191892	30.936508	370.0	63.0
	M	35.262687	32.675676	670.0	185.0
Equipment/Inspection Violation	F	31.603270	30.337500	2324.0	160.0
	M	31.882455	31.429474	8048.0	475.0
Motorist Assist/Courtesy	F	36.517857	33.333333	56.0	3.0
	M	36.708333	35.200000	120.0	10.0
Other Traffic Violation	F	34.091928	31.356250	3035.0	160.0
	M	36.846670	33.491299	12222.0	747.0
Registration Violation	F	32.964247	31.674157	923.0	89.0
	M	32.964977	31.171429	2170.0	245.0
Seatbelt Violation	F	30.145600	28.800000	625.0	10.0
	M	32.824678	31.093750	2253.0	64.0
Special Detail/Directed Patrol	F	36.794393	NaN	107.0	NaN
	M	43.784314	38.777778	2244.0	18.0
Speeding	F	32.540535	30.687500	15357.0	96.0
	M	34.061826	30.345656	32365.0	541.0
Suspicious Person	F	34.583333	39.000000	12.0	3.0
	M	33.000000	31.357143	25.0	14.0
Violation of City/Town Ordinance	F	34.022727	42.500000	44.0	2.0
	M	35.852564	28.555556	156.0	9.0
Warrant	F	24.000000	34.000000	1.0	3.0
	M	33.833333	28.800000	6.0	5.0

1) People who have violation are mostly around 35 years old on average, and people who violate for special detail and directed patrol are generally older.

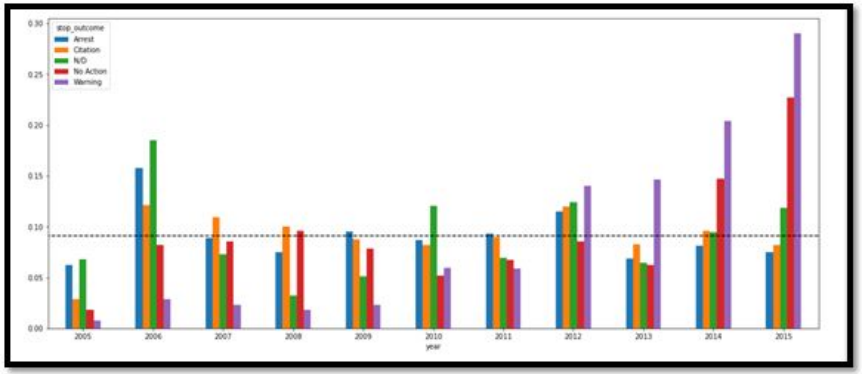
2) People who are caught for equipment or inspection violation generally correlate less to getting arrested.

3) In contrast, people who are caught because they are suspicious correlate more to getting arrested.

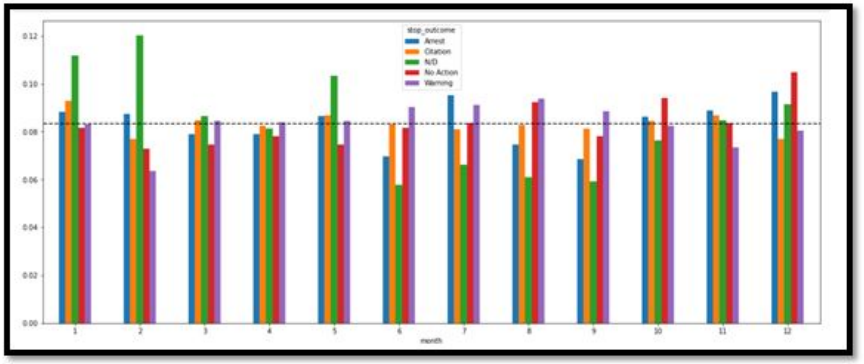
Time Data Wrangling

	stop_duration	stop_dt	DoW	month	year	is_weekend	bin_day	duration_btw_consecutive_mins	stop_greater_half_hour	half_hour_zone
189	0-15 Min	2005-10-06 16:32:00	3	10	2005	0	afternoon	37.0	1	100
190	0-15 Min	2005-10-06 17:42:00	3	10	2005	0	afternoon	70.0	1	101
191	0-15 Min	2005-10-06 17:50:00	3	10	2005	0	afternoon	8.0	0	101
192	0-15 Min	2005-10-06 18:45:00	3	10	2005	0	afternoon	55.0	1	102
193	0-15 Min	2005-10-06 22:30:00	3	10	2005	0	night	225.0	1	103
194	0-15 Min	2005-10-06 23:00:00	3	10	2005	0	night	30.0	0	103
195	0-15 Min	2005-10-06 23:10:00	3	10	2005	0	night	10.0	0	103
196	0-15 Min	2005-10-06 23:30:00	3	10	2005	0	night	20.0	0	103
197	0-15 Min	2005-10-07 00:30:00	4	10	2005	0	late_night	60.0	1	104
198	16-30 Min	2005-10-07 00:50:00	4	10	2005	0	late_night	20.0	0	104
199	0-15 Min	2005-10-07 01:52:00	4	10	2005	0	late_night	62.0	1	105
200	0-15 Min	2005-10-07 02:40:00	4	10	2005	0	late_night	48.0	1	106
201	30+ Min	2005-10-07 05:41:00	4	10	2005	0	late_night	181.0	1	107
202	16-30 Min	2005-10-07 06:30:00	4	10	2005	0	late_night	49.0	1	108
203	0-15 Min	2005-10-07 07:05:00	4	10	2005	0	morning	35.0	1	109
204	0-15 Min	2005-10-07 07:10:00	4	10	2005	0	morning	5.0	0	109

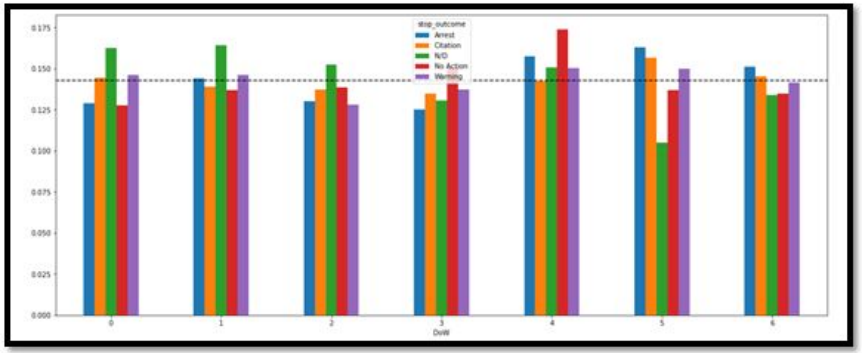
- Extracted columns associated with **year, month,** and **day_of_the_week**
- Generated **bin_day** column-
  - morning, afternoon, night, and late-night
  - 7-12,13-20,21-24,0-6 (bins based on hour)
- Computed the time difference between consecutive instances
  - **Indicator variable** suggesting whether the successive instances occurred in more than 30 mins or not
  - Used this Boolean to generate **zones of closely occurring instances**



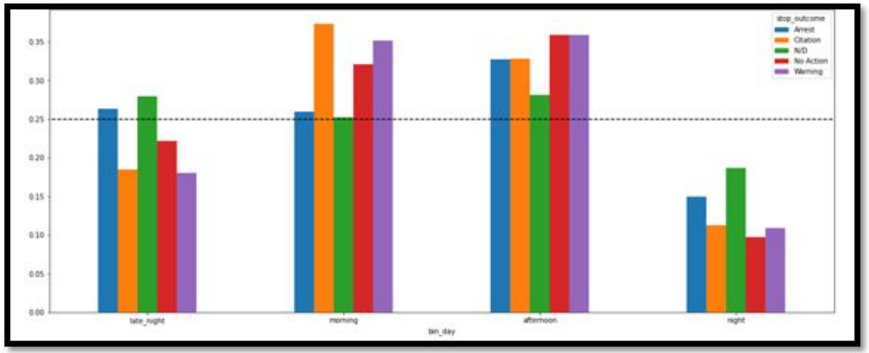
- **No-action and Warnings significantly increased** in later years from **2012 till 2015**
- Number of **arrests** reported in the year **2006** was exceedingly **high**



- **Normalized values** of each outcome, **except N/D** spread uniformly across all **months**



- **Citations and arrests** comparatively **more on weekends**

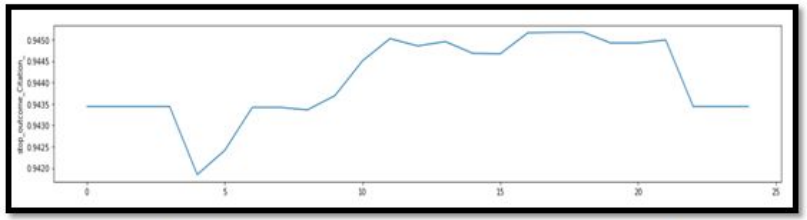


- **Warning and citation** more during **afternoon and day hours**
- **More arrests in night hours (0.26+0.14)**

Inferential Analysis

DATA with categorical variables → Dummy Variables → Explosion in number of columns (80) → Ideation → Large Simple Small complex experiment

- One vs. rest approach to perform this multiclass (5 classes) classification problem
  - Generated five indicator target columns
  - Trained the **stats model-based logit model**
- Sorted the predictors based on **their p values**
  - Selected the **top 25 columns for each model**
- Devised a **stepwise subset selection experiment**
  - Determined the combination of predictors that give **the best f1 score on hold out set data**
- **Analyzed** the best combinations for each class of target variable



Selection of best subsets for predicting Citation outcome using F1 score of test set data

	stop_outcome_Arrest_	stop_outcome_Citation_	stop_outcome_N/D_	stop_outcome_No Action_	stop_outcome_Warning_
0	search_conducted_1	search_conducted_1	violation_raw_Equipment/Inspection Violation	violation_raw_Motorist Assist/Courtesy	stop_duration_16-30 Min
1	search_type_Probable Cause	violation_raw_Speeding	bin_day_morning	violation_raw_Call for Service	stop_duration_30+ Min
2	stop_duration_30+ Min	violation_raw_Special Detail/Directed Patrol	bin_day_afternoon	violation_raw_Suspicious Person	driver_age
3	search_type_Reasonable Suspicion	violation_raw_Seatbelt Violation	search_conducted_1	search_conducted_1	search_type_Reasonable Suspicion
4	stop_duration_16-30 Min	violation_raw_Other Traffic Violation	violation_raw_Call for Service	stop_duration_16-30 Min	violation_raw_Equipment/Inspection Violation

Top 5 of the 20 predictors selected on the basis of p values

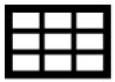
	stop_outcome_Arrest_	stop_outcome_Citation_	stop_outcome_N/D_	stop_outcome_No Action_	stop_outcome_Warning_
0	search_conducted_1	search_conducted_1	violation_raw_Equipment/Inspection Violation	violation_raw_Motorist Assist/Courtesy	stop_duration_16-30 Min
1	search_type_Probable Cause	violation_raw_Speeding	bin_day_morning	violation_raw_Call for Service	stop_duration_30+ Min
2	stop_duration_30+ Min	violation_raw_Special Detail/Directed Patrol	bin_day_afternoon	violation_raw_Suspicious Person	driver_age
3	search_type_Reasonable Suspicion	violation_raw_Seatbelt Violation	search_conducted_1	search_conducted_1	search_type_Reasonable Suspicion
4	stop_duration_16-30 Min	violation_raw_Other Traffic Violation	violation_raw_Call for Service	NaN	violation_raw_Equipment/Inspection Violation
5	search_type_Protective Frisk	violation_raw_Registration Violation	violation_raw_Motorist Assist/Courtesy	NaN	NaN
6	NaN	search_type_Probable Cause	NaN	NaN	NaN
7	NaN	violation_raw_Equipment/Inspection Violation	NaN	NaN	NaN
8	NaN	violation_raw_Violation of City/Town Ordinance	NaN	NaN	NaN
9	NaN	violation_raw_Call for Service	NaN	NaN	NaN

Subset of predictors giving the best f1 score on Hold out test set

Predictive Analysis

Data prep

One Hot Encoding



Drop Values

Normalization

Data Augmentation

Before Oversampling - Naive Bayes

```
              precision    recall  f1-score   support

      Arrest       0.37      0.37      0.37       583
    Citation       0.91      0.98      0.95     15349
     Warning       0.00      0.00      0.00      1060

    accuracy                           0.90     16992
   macro avg       0.43      0.45      0.44     16992
weighted avg       0.84      0.90      0.87     16992

[[  217   366     0]
 [  342 15007     0]
 [   24  1036     0]]
```

Naive Bayes

classifcation report:

	precision	recall	f1-score	support
Arrest	0.19	0.70	0.30	583
Citation	0.96	0.62	0.76	15349
Warning	0.14	0.62	0.22	1060
accuracy			0.63	16992
macro avg	0.43	0.65	0.42	16992
weighted avg	0.88	0.63	0.71	16992

F1 Score Average: 0.4242266507452612

Confusion Matrix:

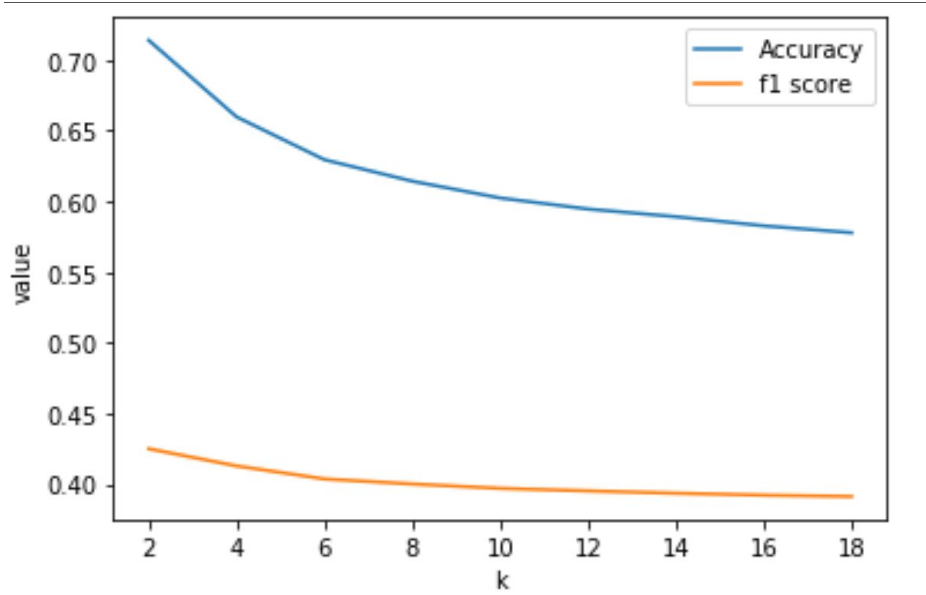
	Arrest	Citation	Warning
Arrest	409	72	102
Citation	1706	9565	4078
Warning	73	333	654

```
driver_race_Other                                          4.149220
search_type_Incident to Arrest,Protective Frisk           -3.899559
search_type_Incident to Arrest,Probable Cause             -3.693820
search_type_Incident to Arrest,Inventory                  -3.651215
search_type_Incident to Arrest,Inventory,Probable Cause   -3.494835
search_type_Incident to Arrest                            -3.414511
search_type_Incident to Arrest,Inventory,Protective Frisk -3.140678
violation_raw_Warrant                                     -2.940993
search_type_Incident to Arrest,Inventory,Reasonable Suspicion -2.646357
search_conducted                                          -2.561919
dtype: float64
```

KNN

```
classifcation report:
              precision    recall  f1-score   support

      Arrest       0.15      0.57      0.24       583
    Citation       0.93      0.74      0.83     15349
     Warning       0.14      0.33      0.19      1060

    accuracy                           0.71     16992
   macro avg       0.41      0.55      0.42     16992
weighted avg       0.86      0.71      0.77     16992

F1 Score Average: 0.41908612294837394
Confusion Matrix:
```

	Arrest	Citation	Warning
Arrest	335	189	59
Citation	1818	11391	2140
Warning	104	606	350



Random Forests

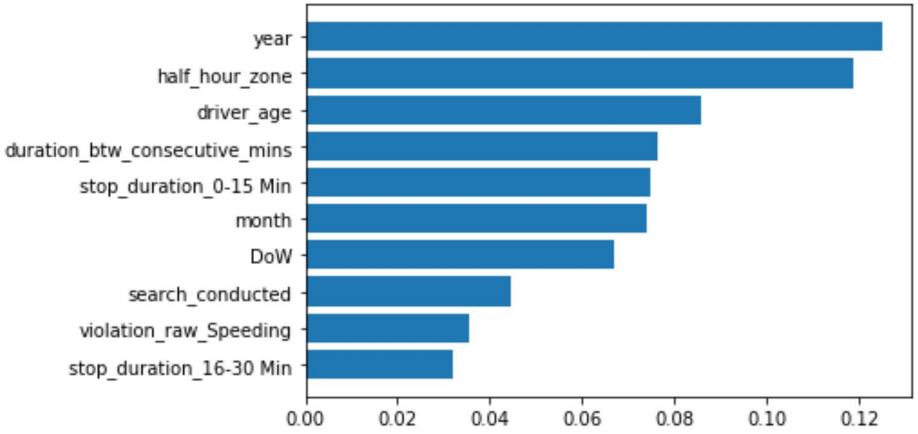
classifcation report:

	precision	recall	f1-score	support
Arrest	0.28	0.56	0.37	583
Citation	0.94	0.84	0.89	15349
Warning	0.19	0.36	0.25	1060
accuracy			0.80	16992
macro avg	0.47	0.59	0.50	16992
weighted avg	0.87	0.80	0.83	16992

F1 Score Average: 0.5028990738406974

Confusion Matrix:

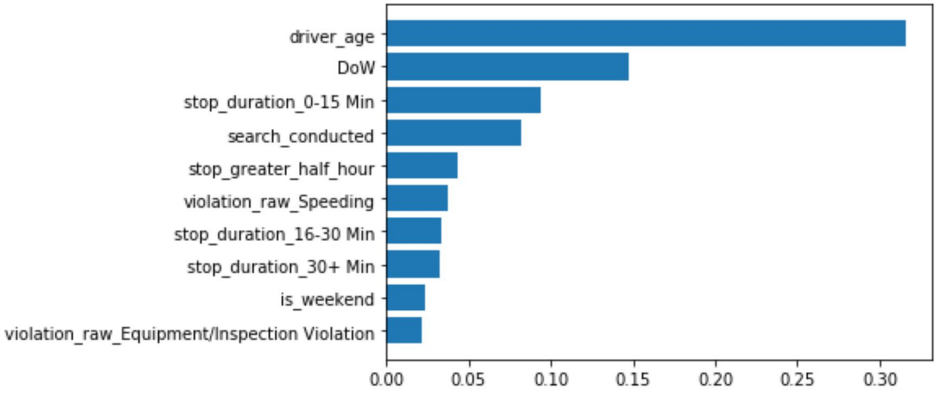
	Arrest	Citation	Warning
Arrest	329	208	46
Citation	808	12924	1617
Warning	47	627	386



Random Forest (Cleaned Data)

```
classifcation report:
              precision    recall  f1-score   support

      Arrest       0.22      0.51      0.31       583
    Citation       0.93      0.77      0.84     15349
     Warning       0.11      0.32      0.17      1060

    accuracy                           0.73     16992
   macro avg       0.42      0.53      0.44     16992
weighted avg       0.86      0.73      0.78     16992

F1 Score Average: 0.43965986575143684
Confusion Matrix:
```

	Arrest	Citation	Warning
Arrest	298	223	62
Citation	990	11795	2564
Warning	62	658	340



Other Models - SVM

	precision	recall	f1-score	support
Arrest	0.20	0.64	0.31	583
Citation	0.96	0.60	0.74	15349
Warning	0.14	0.74	0.24	1060
accuracy			0.61	16992
macro avg	0.44	0.66	0.43	16992
weighted avg	0.89	0.61	0.69	16992

```
[[ 371  125   87]
 [1411 9222 4716]
 [  59  216  785]]
```

Other Models - Logistic Regression

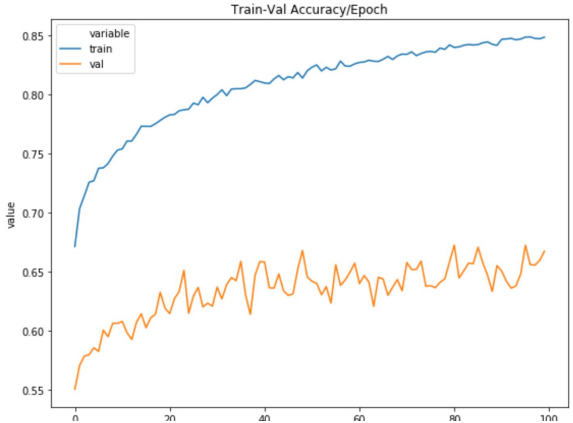
```
              precision    recall  f1-score   support

      Arrest       0.17      0.72      0.28       583
    Citation       0.97      0.57      0.72     15349
     Warning       0.14      0.75      0.24      1060

    accuracy                           0.59     16992
   macro avg       0.43      0.68      0.41     16992
weighted avg       0.89      0.59      0.68     16992

[[ 417   80   86]
 [1931 8789 4629]
 [  75  187  798]]
```

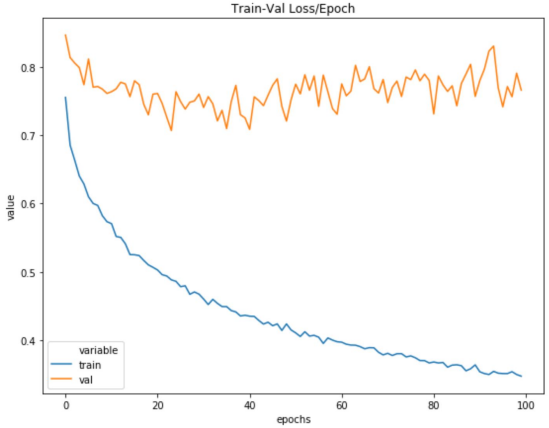
Other Models - Neutral Network (PyTorch)

	precision	recall	f1-score	support
0	0.19	0.54	0.28	585
1	0.95	0.68	0.80	15364
2	0.15	0.59	0.23	1043
accuracy			0.67	16992
macro avg	0.43	0.61	0.44	16992
weighted avg	0.87	0.67	0.74	16992

```
[[  314   190    81]
 [ 1274 10524  3566]
 [   65   358   620]]
```

+ Code





Conclusion

Takeaways

Data Insights

Model Results

Improvements

Thank you!