Data Comp - Team 322

April 5, 2020

1 Introduction

The New York Police Department, or NYPD, is one of the most active police departments in the US, where they go through many conflicts with criminals throughout the state with a multitude of different crimes ranging from murder, marijuana possession, robbery, arson, etc. Through many instances of arrests in the state, we can see some specific instances of discrimination of specific arrests where the perpetrator can be wrongly accused or wrongly acquitted from their crimes.

This project addresses the following question: Is there evidence of discrimination in NYPD arrests based on demographics?

The approach used will use machine learning algorithms to find insights into how the different variables in the given dataset of NYPD arrests interact with each other. In particular, our approach will use Random Forests to classify arrests based on various predictors.

2 Data and Problem Definition

2.1 Data Set

The given dataset contains variables of each arrest such as the age group, gender, precinct (city district), race, and other descriptive factors. In particular, each row contains information about the arrest key, arrest date, internal classification code, description of said code, a more general internal classification code, description of said code, law code charges, level of offense committed, the borough the arrest was made in, precinct where the arrest occurred, jurisdiction responsible for arrest, age group of the perpetrator, sex of the perpetrator, race of the perpetrator, and coordinates of the location of arrest.

In this project, we mainly use the following features: internal classification code, level of offense committed, age group of the perpetrator, sex of the perpetrator, and race of the perpetrator.

2.2 Problem Definition

In this particular approach, the problem is to find bias in classification in favor of or against a certain demographic.

Our approach will be as follows: Train a model using the data to predict a certain characteristic of the arrest and compare the predictions with the actual labels to see if the predictions are skewed.

3 Methodology

## 3.1 Exploratory Data Analysis

```
In [20]: import pandas as pd
         nypd = pd.read_csv("NYPD_Arrests_Data__Historic_.csv")

In [21]: nypd = nypd[['PD_DESC', 'LAW_CAT_CD', 'AGE_GROUP', 'PERP_SEX', 'PERP_RACE', 'ARREST_B(
         nypd.dropna(inplace=True)

In [22]: dict = {'V':'Violation', 'F':'Felony', 'M':'Misdemeanor', 'I':'Infraction'}
         for i in nypd.LAW_CAT_CD.unique():
             group = nypd[nypd.LAW_CAT_CD == i]
             counts = group.PERP_RACE.value_counts()
             counts = pd.DataFrame(counts)
             title = "Category of violation: " + dict[i]
             print(counts.plot(kind='bar', title = title))
```
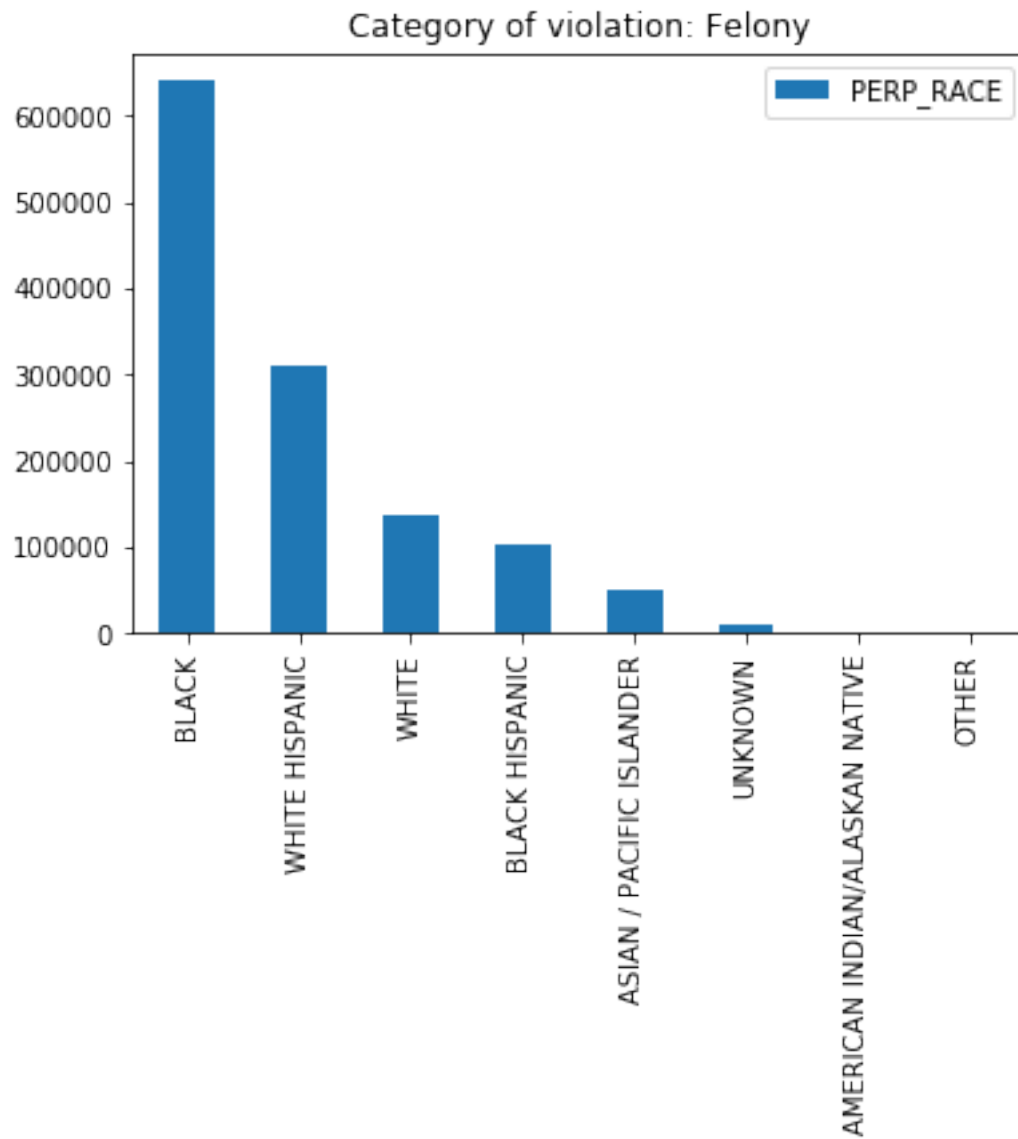
```
AxesSubplot(0.125,0.125;0.775x0.755)
AxesSubplot(0.125,0.125;0.775x0.755)
AxesSubplot(0.125,0.125;0.775x0.755)
AxesSubplot(0.125,0.125;0.775x0.755)
```

Category of violation: Violation

Category of violation: Misdemeanor

Category of violation: Felony

Category of violation: Infraction
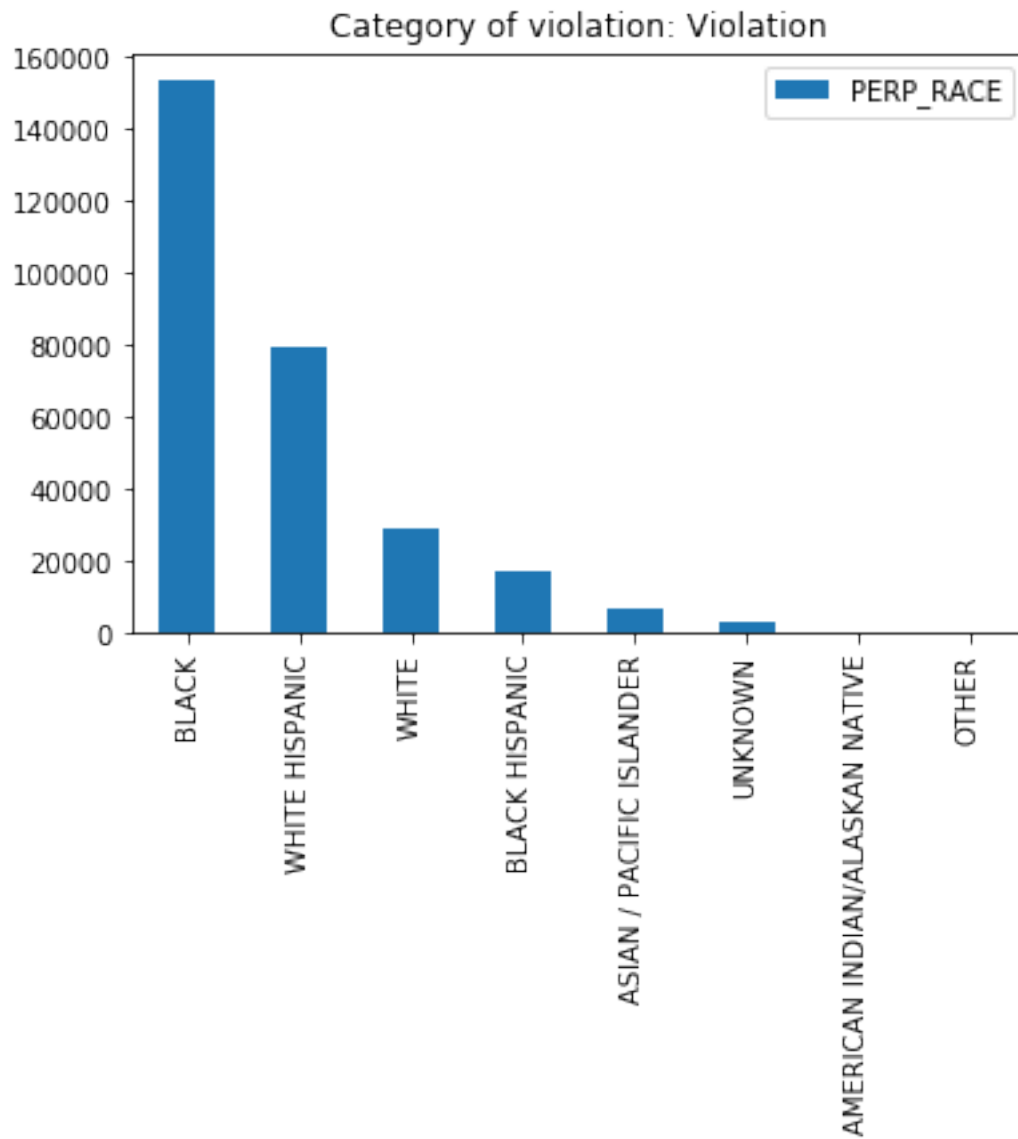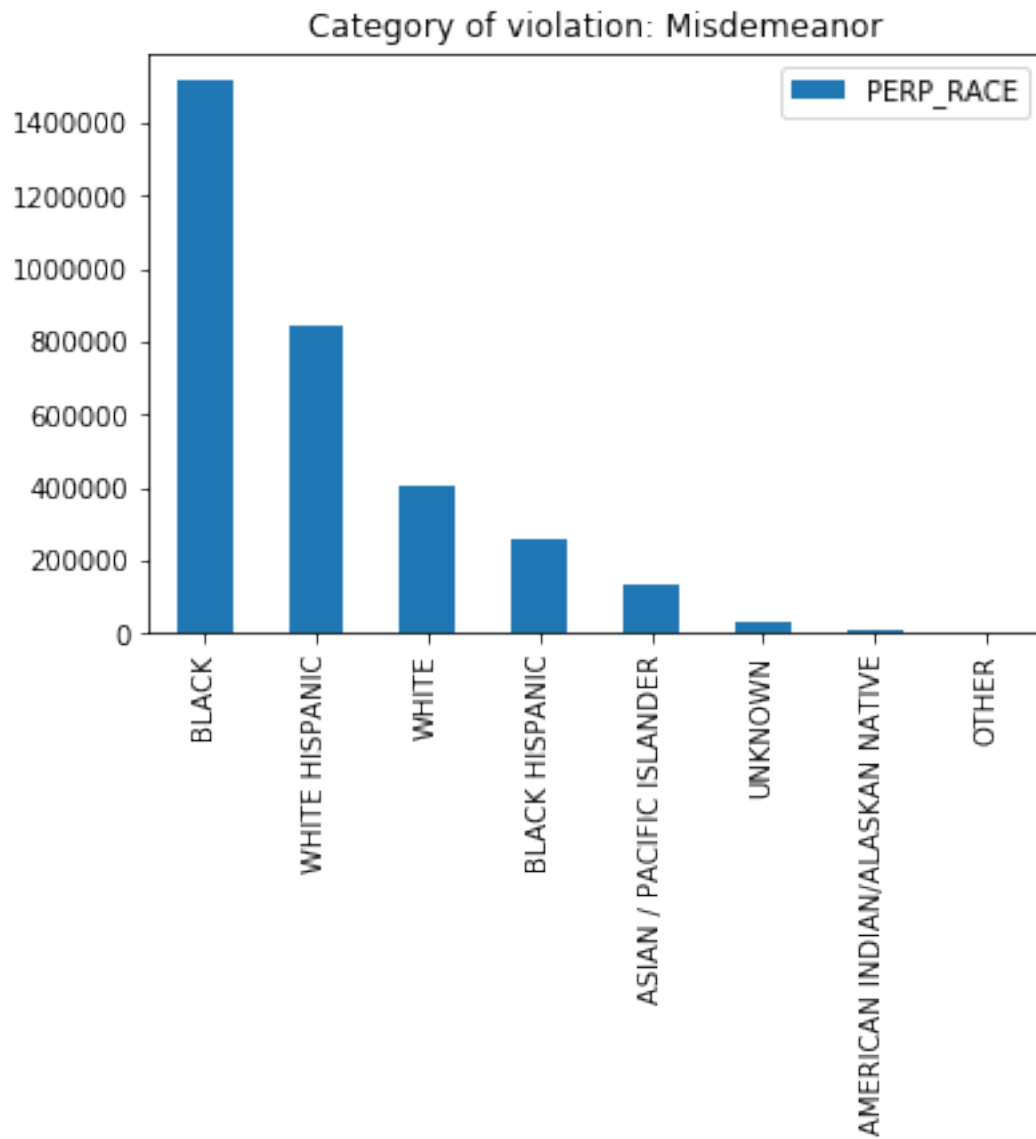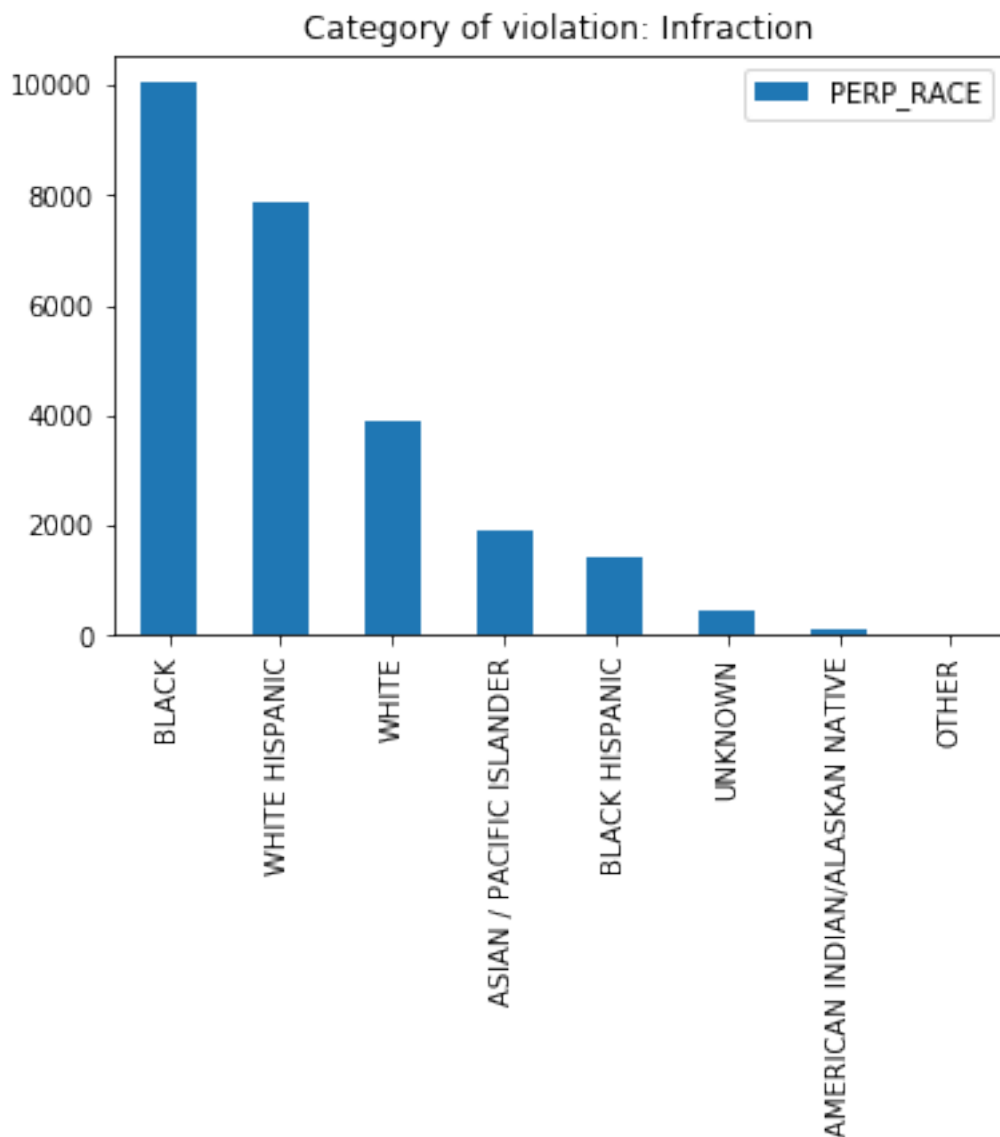
Using clustering, we get the following output:

Standardized Cluster Means: centroid pd_desc perp_race age_group arrest_boro perp_sex law_cat_cd ---------- ------------------------------------- -------------- ----------- ------------- ---------- ---------- -- 1 ASSAULT 3 BLACK 25-44 Q M M 2 THEFT OF SERVICES, UNCLASSIFIED BLACK 18-24 M M M 3 CONTROLLED SUBSTANCE, POSSESSION 7 WHITE HISPANIC 25-44 K M M 4 CONTROLLED SUBSTANCE,SALE 3 WHITE HISPANIC 25-44 M M F 5 NY STATE LAWS,UNCLASSIFIED VIOLATION BLACK 25-44 M M V 6 ASSAULT 2,1,UNCLASSIFIED WHITE HISPANIC 25-44 B F F

Centroid Statistics: centroid size within_cluster_sum_of_squares -- ---------- ----------- ------------- ------------------- 1 2.03914e+06 5.26579e+06 2 1.17129e+06 2.77919e+06 3 722192 1.5601e+06 4 416880 969848 5 235334 393414 6 213506 546798

3.2 Modeling

We start by using a Naive Bayes Classifier to predict the race of the perpetrator using their age group, sex, category of offense, and description of offense. Even though this is counter intuitive, we are trying to uncover a pattern that the model could learn about the race of the perpetrator in specific arrest scenarios.

```python
In [32]: from sklearn.model_selection import train_test_split
         import h2o
         from h2o.estimators.naive_bayes import H2ONaiveBayesEstimator

         h2o.init(max_mem_size = "10G")            #specify max number of bytes. uses all cor
         h2o.remove_all()


         data=nypd[['PD_DESC', 'AGE_GROUP', 'PERP_SEX', 'LAW_CAT_CD','PERP_RACE']]

         predictors = data.columns.values[:-1]
         target = data.columns.values[-1]


         train, test = train_test_split(nypd, test_size=0.3)

         h_train = h2o.H2OFrame(train)
         h_test = h2o.H2OFrame(test)

         model = H2ONaiveBayesEstimator(score_each_iteration=True,
                                        seed = 1234)
         model.train(training_frame = h_train,
                     validation_frame = h_test,
                     y = str(target),
                     x = list(predictors))
         print(model)
```

Checking whether there is an H2O instance running at http://localhost:54321 . connected.

H2O_cluster_uptime:	25 mins 51 secs
H2O_cluster_timezone:	America/Los_Angeles
H2O_data_parsing_timezone:	UTC
H2O_cluster_version:	3.30.0.1
H2O_cluster_version_age:	1 day
H2O_cluster_name:	H2O_from_python_varch_j63zsy
H2O_cluster_total_nodes:	1
H2O_cluster_free_memory:	9.97 Gb
H2O_cluster_total_cores:	4
H2O_cluster_allowed_cores:	4
H2O_cluster_status:	locked, healthy

```
H2O_connection_url:          http://localhost:54321
H2O_connection_proxy:        {"http": null, "https": null}
H2O_internal_security:       False
H2O_API_Extensions:          Amazon S3, Algos, AutoML, Core V3, TargetEncoder, Core V4
Python_version:              3.7.1 final
--------------------------   ----------------------------------------------------------


Parse progress: || 100%
Parse progress: || 100%
naivebayes Model Build progress: || 100%
Model Details
=============
H2ONaiveBayesEstimator :  Naive Bayes
Model Key:  NaiveBayes_model_python_1586062955647_1



Model Summary:


    number_of_response_levels  min_apriori_probability  \
0                         8.0                 0.000284


   max_apriori_probability
0                 0.485656




ModelMetricsMultinomial: naivebayes
** Reported on train data. **

MSE: 0.4683722936647309
RMSE: 0.684377303586794
LogLoss: 1.3171315620859425
Mean Per-Class Error: 0.8687579092919805

Confusion Matrix: Row labels: Actual class; Column labels: Predicted class


   AMERICAN INDIAN/ALASKAN NATIVE  ASIAN / PACIFIC ISLANDER      BLACK  \
0                            0.0                      14.0     6884.0
1                            2.0                     724.0   122904.0
2                            2.0                     502.0  1588126.0
3                            0.0                     200.0   256613.0
4                            0.0                       1.0      907.0
5                            1.0                      60.0    31237.0
6                            3.0                     364.0   373104.0
7                            1.0                     663.0   822796.0
```

8

```
8                          9.0                    2528.0  3202571.0

   BLACK HISPANIC  OTHER  UNKNOWN    WHITE  WHITE HISPANIC      Error  \
0            2.0    0.0      2.0     75.0           380.0  1.000000
1           30.0    0.0     63.0   1776.0          9264.0  0.994628
2          110.0    0.0    283.0   6639.0         27147.0  0.021372
3            0.0    0.0     73.0   1378.0          9221.0  1.000000
4            0.0    0.0      0.0      5.0            35.0  1.000000
5            7.0    0.0      0.0    485.0          2301.0  1.000000
6          108.0    0.0    165.0   7767.0         21939.0  0.980749
7           75.0    0.0    236.0   6164.0         40643.0  0.953315
8          332.0    0.0    822.0  24289.0        110930.0  0.510020

                  Rate
0         7,357 / 7,357
1     134,039 / 134,763
2     34,683 / 1,622,809
3     267,485 / 267,485
4             948 / 948
5       34,091 / 34,091
6     395,683 / 403,450
7     829,935 / 870,578
8  1,704,221 / 3,341,481


Top-8 Hit Ratios:


   k  hit_ratio
0  1   0.489980
1  2   0.746147
2  3   0.877586
3  4   0.951860
4  5   0.987473
5  6   0.997393
6  7   0.999696
7  8   1.000000



ModelMetricsMultinomial: naivebayes
** Reported on validation data. **

MSE: 0.46793905490871296
RMSE: 0.6840607099583436
LogLoss: 1.3156647265511612
Mean Per-Class Error: 0.8687512321177057
```

Confusion Matrix: Row labels: Actual class; Column labels: Predicted class

	AMERICAN INDIAN/ALASKAN NATIVE	ASIAN / PACIFIC ISLANDER	BLACK \
0	0.0	4.0	2853.0
1	0.0	313.0	52359.0
2	3.0	196.0	681506.0
3	0.0	71.0	109710.0
4	0.0	1.0	389.0
5	0.0	24.0	13433.0
6	1.0	153.0	160004.0
7	0.0	253.0	352407.0
8	4.0	1015.0	1372661.0

	BLACK HISPANIC	OTHER	UNKNOWN	WHITE	WHITE HISPANIC	Error \
0	0.0	0.0	1.0	36.0	178.0	1.000000
1	7.0	0.0	25.0	785.0	3960.0	0.994552
2	39.0	0.0	133.0	2900.0	11692.0	0.021484
3	7.0	0.0	32.0	575.0	4021.0	0.999939
4	0.0	0.0	0.0	1.0	24.0	1.000000
5	2.0	0.0	5.0	197.0	973.0	0.999658
6	40.0	0.0	83.0	3218.0	9171.0	0.981363
7	30.0	0.0	90.0	2635.0	17523.0	0.953014
8	125.0	0.0	369.0	10347.0	47542.0	0.509399

	Rate
0	3,072 / 3,072
1	57,136 / 57,449
2	14,963 / 696,469
3	114,409 / 114,416
4	415 / 415
5	14,629 / 14,634
6	169,452 / 172,670
7	355,415 / 372,938
8	729,491 / 1,432,063

Top-8 Hit Ratios:

	k	hit_ratio
0	1	0.490601
1	2	0.746526
2	3	0.878012
3	4	0.952205
4	5	0.987490

```
5   6    0.997445
6   7    0.999695
7   8    1.000000
```

Judging the model performance, we conclude that there is no clear pattern in the data with respect to race.

Next, we train a Naive Bayes Classifier on all demographic categorical variables to predict the level of offense.

```
In [35]: h2o.init(max_mem_size = "10G")               #specify max number of bytes. uses all cor
         h2o.remove_all()


         data=nypd[['AGE_GROUP', 'PERP_SEX', 'PERP_RACE', 'LAW_CAT_CD']]

         predictors = data.columns.values[:-1]
         target = data.columns.values[-1]


         train, test = train_test_split(nypd, test_size=0.3)

         h_train = h2o.H2OFrame(train)
         h_test = h2o.H2OFrame(test)

         model = H2ONaiveBayesEstimator(score_each_iteration=True,
                                        seed = 1234)
         model.train(training_frame = h_train,
                     validation_frame = h_test,
                     y = str(target),
                     x = list(predictors))
         print(model)
```

Checking whether there is an H2O instance running at http://localhost:54321 . connected.

| ------------------------- | --- |
H2O_cluster_uptime:	51 mins 56 secs
H2O_cluster_timezone:	America/Los_Angeles
H2O_data_parsing_timezone:	UTC
H2O_cluster_version:	3.30.0.1
H2O_cluster_version_age:	1 day
H2O_cluster_name:	H2O_from_python_varch_j63zsy
H2O_cluster_total_nodes:	1
H2O_cluster_free_memory:	9.97 Gb
H2O_cluster_total_cores:	4

```
H2O_cluster_allowed_cores:   4
H2O_cluster_status:          locked, healthy
H2O_connection_url:          http://localhost:54321
H2O_connection_proxy:        {"http": null, "https": null}
H2O_internal_security:       False
H2O_API_Extensions:          Amazon S3, Algos, AutoML, Core V3, TargetEncoder, Core V4
Python_version:              3.7.1 final
-------------------------  --------------------------------------------------------


Parse progress: || 100%
Parse progress: || 100%
naivebayes Model Build progress: || 100%
Model Details
=============
H2ONaiveBayesEstimator :  Naive Bayes
Model Key:  NaiveBayes_model_python_1586062955647_3


Model Summary:


     number_of_response_levels  min_apriori_probability  \
0                          4.0                 0.005317

   max_apriori_probability
0                 0.671163




ModelMetricsMultinomial: naivebayes
** Reported on train data. **

MSE: 0.27267301688953166
RMSE: 0.5221810192735195
LogLoss: 0.8098425328356452
Mean Per-Class Error: 0.750008137592646

Confusion Matrix: Row labels: Actual class; Column labels: Predicted class


      F    I         M     V    Error                  Rate
0   0.0  5.0   878035.0  15.0  1.000000      878,055 / 878,055
1   1.0  0.0    17767.0   0.0  1.000000       17,768 / 17,768
2  45.0  3.0  2242605.0  25.0  0.000033           73 / 2,242,678
3   2.0  1.0   202974.0   0.0  1.000000      202,977 / 202,977
4  48.0  9.0  3341381.0  40.0  0.328858  1,098,873 / 3,341,478
```

Top-4 Hit Ratios:


```
   k  hit_ratio
0  1   0.671142
1  2   0.933906
2  3   0.994660
3  4   1.000000
```


ModelMetricsMultinomial: naivebayes
** Reported on validation data. **

MSE: 0.2724527333255598
RMSE: 0.5219700502189372
LogLoss: 0.8097145524908334
Mean Per-Class Error: 0.7500033540457705

Confusion Matrix: Row labels: Actual class; Column labels: Predicted class

	F	I	M	V	Error	Rate
0	0.0	0.0	375656.0	2.0	1.000000	375,658 / 375,658
1	0.0	0.0	7875.0	0.0	1.000000	7,875 / 7,875
2	13.0	3.0	961820.0	8.0	0.000025	24 / 961,844
3	0.0	1.0	86684.0	1.0	0.999988	86,685 / 86,686
4	13.0	4.0	1432035.0	11.0	0.328367	470,242 / 1,432,063

Top-4 Hit Ratios:


```
   k  hit_ratio
0  1   0.671633
1  2   0.933949
2  3   0.994491
3  4   1.000000
```


From the results, we notice that the model classifies almost all instances as misdemeanors, indicating that it is simply favoring the most frequent level rather than exposing any structure within the data. This leads us to the conclusion that demographic variables do not accurately model the level of offense.

3.2.1 Random Forest

We now use a more sophisticated non-parametric model to use a selection of variables to predict the the level of offense. In this, we use a methodology of hiding the borough variable from the model after subsetting the data to a specific borough. Effectively, this results in implicitly borough-specific model. We then proceed to test each model agains data from different boroughs to check if there is a bias in the predictions of the levels of offense in any model against any different borough

```
In [37]: from h2o.estimators.gbm import H2OGradientBoostingEstimator
         from h2o.estimators.random_forest import H2ORandomForestEstimator

         boroughs = nypd['ARREST_BORO'].unique()

         nypd = nypd[['PD_DESC', 'PERP_RACE', 'AGE_GROUP', 'ARREST_BORO', 'PERP_SEX', 'LAW_CAT_

         for borough in boroughs:
             df_borough = nypd.loc[nypd['ARREST_BORO'] == borough]
             df_borough.drop(['ARREST_BORO'], axis=1, inplace=True)

             train, test = train_test_split(df_borough, test_size=0.2)

             predictors = train.columns.values[:-1]
             target = train.columns.values[-1]

             model = H2ORandomForestEstimator(
                 model_id="rf_"+borough,
                 ntrees=200,
                 stopping_rounds=7,
                 score_each_iteration=True,
                 max_depth=50,
                 seed=1000000)

             h_train = h2o.H2OFrame(train)
             h_test = h2o.H2OFrame(test)

             model.train(list(predictors), str(target), training_frame=h_train, validation_fram
             print(model)

             h2o.save_model(model=model, path="rf_"+borough, force=True)
             print(model.score_history())
Parse progress: || 100%
Parse progress: || 100%
drf Model Build progress: || 100%
Model Details
=============
H2ORandomForestEstimator :  Distributed Random Forest
Model Key:  rf_Q
```

Model Summary:

	number_of_trees	number_of_internal_trees	model_size_in_bytes \
0	27.0	108.0	386750.0

	min_depth	max_depth	mean_depth	min_leaves	max_leaves	mean_leaves
0	8.0	16.0	12.148149	28.0	298.0	133.24074

ModelMetricsMultinomial: drf
** Reported on train data. **

MSE: 0.013171771670983146
RMSE: 0.11476833914883994
LogLoss: 0.05353537307174825
Mean Per-Class Error: 0.14278233045171151

Confusion Matrix: Row labels: Actual class; Column labels: Predicted class

	F	I	M	V	Error	Rate
0	192838.0	0.0	4514.0	0.0	0.022873	4,514 / 197,352
1	0.0	4411.0	5176.0	143.0	0.546660	5,319 / 9,730
2	531.0	17.0	457625.0	0.0	0.001196	548 / 458,173
3	0.0	0.0	19.0	47408.0	0.000401	19 / 47,427
4	193369.0	4428.0	467334.0	47551.0	0.014593	10,400 / 712,682

Top-4 Hit Ratios:

	k	hit_ratio
0	1	0.985407
1	2	0.999572
2	3	0.999990
3	4	1.000000

ModelMetricsMultinomial: drf
** Reported on validation data. **

MSE: 0.013276702073372752
RMSE: 0.11522457235057439

LogLoss: 0.0543130314762207
Mean Per-Class Error: 0.13920517395963944

Confusion Matrix: Row labels: Actual class; Column labels: Predicted class

	F	I	M	V	Error	Rate
0	48443.0	0.0	1199.0	0.0	0.024153	1,199 / 49,642
1	0.0	1174.0	1287.0	42.0	0.530963	1,329 / 2,503
2	130.0	7.0	113998.0	0.0	0.001200	137 / 114,135
3	0.0	0.0	6.0	11885.0	0.000505	6 / 11,891
4	48573.0	1181.0	116490.0	11927.0	0.014991	2,671 / 178,171

Top-4 Hit Ratios:

	k	hit_ratio
0	1	0.985009
1	2	0.999506
2	3	0.999977
3	4	1.000000

Scoring History:

	timestamp	duration	number_of_trees	training_rmse \
0	2020-04-04 23:04:49	0.112 sec	0.0	NaN
1	2020-04-04 23:04:53	3.980 sec	1.0	0.163007
2	2020-04-04 23:04:56	6.781 sec	2.0	0.138292
3	2020-04-04 23:04:58	8.925 sec	3.0	0.129223
4	2020-04-04 23:05:00	10.854 sec	4.0	0.124490
5	2020-04-04 23:05:02	12.713 sec	5.0	0.121484
6	2020-04-04 23:05:04	15.035 sec	6.0	0.119587
7	2020-04-04 23:05:06	17.313 sec	7.0	0.118227
8	2020-04-04 23:05:09	19.574 sec	8.0	0.117259
9	2020-04-04 23:05:11	21.875 sec	9.0	0.116538
10	2020-04-04 23:05:13	24.394 sec	10.0	0.115940
11	2020-04-04 23:05:16	26.951 sec	11.0	0.115600
12	2020-04-04 23:05:19	29.730 sec	12.0	0.115542
13	2020-04-04 23:05:22	32.544 sec	13.0	0.115182
14	2020-04-04 23:05:24	35.355 sec	14.0	0.114962
15	2020-04-04 23:05:28	38.541 sec	15.0	0.115219
16	2020-04-04 23:05:31	41.776 sec	16.0	0.114949
17	2020-04-04 23:05:34	45.065 sec	17.0	0.114733
18	2020-04-04 23:05:38	48.910 sec	18.0	0.114543

```
19     2020-04-04 23:05:42  52.722 sec                 19.0      0.114392

    training_logloss  training_classification_error  validation_rmse  \
0              NaN                            NaN              NaN
1         0.124956                       0.015037         0.163222
2         0.088431                       0.014741         0.132450
3         0.075707                       0.014691         0.123747
4         0.068961                       0.014758         0.120153
5         0.063396                       0.014701         0.118261
6         0.060611                       0.014688         0.117189
7         0.058778                       0.014663         0.116513
8         0.057572                       0.014613         0.116110
9         0.055820                       0.014637         0.115756
10        0.054638                       0.014632         0.115487
11        0.054029                       0.014639         0.115343
12        0.053913                       0.014633         0.115329
13        0.053047                       0.014637         0.115166
14        0.052763                       0.014629         0.115084
15        0.053969                       0.014625         0.115265
16        0.053403                       0.014623         0.115148
17        0.052874                       0.014616         0.115046
18        0.052404                       0.014608         0.114957
19        0.052007                       0.014606         0.114887

    validation_logloss  validation_classification_error
0              NaN                              NaN
1         0.125809                         0.015210
2         0.083278                         0.015008
3         0.070831                         0.014991
4         0.064559                         0.014980
5         0.060747                         0.014986
6         0.058372                         0.014963
7         0.056646                         0.014980
8         0.056152                         0.014963
9         0.055038                         0.014986
10        0.054155                         0.014986
11        0.053803                         0.014991
12        0.053995                         0.014991
13        0.053385                         0.014991
14        0.053170                         0.014986
15        0.054281                         0.014986
16        0.053809                         0.014986
17        0.053352                         0.014991
18        0.052935                         0.014986
19        0.052648                         0.014986
```

See the whole table with table.as_data_frame()

Variable Importances:

	variable	relative_importance	scaled_importance	percentage
0	PD_DESC	6.984099e+06	1.000000	0.995673
1	PERP_RACE	1.877215e+04	0.002688	0.002676
2	PERP_SEX	7.393794e+03	0.001059	0.001054
3	AGE_GROUP	4.187105e+03	0.000600	0.000597

	timestamp	duration	number_of_trees	training_rmse \
0	2020-04-04 23:04:49	0.112 sec	0.0	NaN
1	2020-04-04 23:04:53	3.980 sec	1.0	0.163007
2	2020-04-04 23:04:56	6.781 sec	2.0	0.138292
3	2020-04-04 23:04:58	8.925 sec	3.0	0.129223
4	2020-04-04 23:05:00	10.854 sec	4.0	0.124490
5	2020-04-04 23:05:02	12.713 sec	5.0	0.121484
6	2020-04-04 23:05:04	15.035 sec	6.0	0.119587
7	2020-04-04 23:05:06	17.313 sec	7.0	0.118227
8	2020-04-04 23:05:09	19.574 sec	8.0	0.117259
9	2020-04-04 23:05:11	21.875 sec	9.0	0.116538
10	2020-04-04 23:05:13	24.394 sec	10.0	0.115940
11	2020-04-04 23:05:16	26.951 sec	11.0	0.115600
12	2020-04-04 23:05:19	29.730 sec	12.0	0.115542
13	2020-04-04 23:05:22	32.544 sec	13.0	0.115182
14	2020-04-04 23:05:24	35.355 sec	14.0	0.114962
15	2020-04-04 23:05:28	38.541 sec	15.0	0.115219
16	2020-04-04 23:05:31	41.776 sec	16.0	0.114949
17	2020-04-04 23:05:34	45.065 sec	17.0	0.114733
18	2020-04-04 23:05:38	48.910 sec	18.0	0.114543
19	2020-04-04 23:05:42	52.722 sec	19.0	0.114392
20	2020-04-04 23:05:46	56.899 sec	20.0	0.114302
21	2020-04-04 23:05:50	1 min 1.436 sec	21.0	0.114465
22	2020-04-04 23:05:55	1 min 6.024 sec	22.0	0.115353
23	2020-04-04 23:06:00	1 min 10.781 sec	23.0	0.115177
24	2020-04-04 23:06:05	1 min 15.644 sec	24.0	0.115007
25	2020-04-04 23:06:10	1 min 20.825 sec	25.0	0.114877
26	2020-04-04 23:06:15	1 min 25.855 sec	26.0	0.114902
27	2020-04-04 23:06:20	1 min 31.240 sec	27.0	0.114768

	training_logloss	training_classification_error	validation_rmse \
0	NaN	NaN	NaN
1	0.124956	0.015037	0.163222
2	0.088431	0.014741	0.132450
3	0.075707	0.014691	0.123747

4	0.068961	0.014758	0.120153
5	0.063396	0.014701	0.118261
6	0.060611	0.014688	0.117189
7	0.058778	0.014663	0.116513
8	0.057572	0.014613	0.116110
9	0.055820	0.014637	0.115756
10	0.054638	0.014632	0.115487
11	0.054029	0.014639	0.115343
12	0.053913	0.014633	0.115329
13	0.053047	0.014637	0.115166
14	0.052763	0.014629	0.115084
15	0.053969	0.014625	0.115265
16	0.053403	0.014623	0.115148
17	0.052874	0.014616	0.115046
18	0.052404	0.014608	0.114957
19	0.052007	0.014606	0.114887
20	0.051816	0.014607	0.114852
21	0.052323	0.014605	0.114954
22	0.054680	0.014605	0.115476
23	0.054364	0.014602	0.115381
24	0.053991	0.014596	0.115287
25	0.053765	0.014590	0.115235
26	0.053827	0.014593	0.115303
27	0.053535	0.014593	0.115225

	validation_logloss	validation_classification_error
0	NaN	NaN
1	0.125809	0.015210
2	0.083278	0.015008
3	0.070831	0.014991
4	0.064559	0.014980
5	0.060747	0.014986
6	0.058372	0.014963
7	0.056646	0.014980
8	0.056152	0.014963
9	0.055038	0.014986
10	0.054155	0.014986
11	0.053803	0.014991
12	0.053995	0.014991
13	0.053385	0.014991
14	0.053170	0.014986
15	0.054281	0.014986
16	0.053809	0.014986
17	0.053352	0.014991
18	0.052935	0.014986
19	0.052648	0.014986
20	0.052435	0.014986
21	0.052937	0.014980

```
22          0.055460                    0.014980
23          0.055132                    0.014986
24          0.054756                    0.014986
25          0.054538                    0.015002
26          0.054601                    0.014991
27          0.054313                    0.014991
```
Parse progress: || 100%
Parse progress: || 100%
drf Model Build progress: || 100%
Model Details
=============
H2ORandomForestEstimator : Distributed Random Forest
Model Key: rf_K

Model Summary:


```
    number_of_trees  number_of_internal_trees  model_size_in_bytes  \
0              18.0                      72.0              253844.0

   min_depth  max_depth  mean_depth  min_leaves  max_leaves  mean_leaves
0        8.0       17.0   12.222222        19.0       301.0    135.45833
```


ModelMetricsMultinomial: drf
** Reported on train data. **

MSE: 0.007878494355481895
RMSE: 0.08876088302558675
LogLoss: 0.03252895137395644
Mean Per-Class Error: 0.14416619061137875

Confusion Matrix: Row labels: Actual class; Column labels: Predicted class


```
          F        I         M        V     Error                 Rate
0  301102.0      0.0    4935.0      0.0  0.016126      4,935 / 306,037
1       0.0   2613.0    2960.0    324.0  0.556893        3,284 / 5,897
2     926.0     16.0  671086.0      0.0  0.001402        942 / 672,028
3       0.0      2.0     180.0  80916.0  0.002244         182 / 81,098
4  302028.0   2631.0  679161.0  81240.0  0.008772  9,343 / 1,065,060
```


Top-4 Hit Ratios:

```
      k  hit_ratio
0  1    0.991228
1  2    0.999744
2  3    0.999992
3  4    1.000000




ModelMetricsMultinomial: drf
** Reported on validation data. **

MSE: 0.0075037751187233
RMSE: 0.08662433329453854
LogLoss: 0.03159136187403362
Mean Per-Class Error: 0.14318723460345475

Confusion Matrix: Row labels: Actual class; Column labels: Predicted class


           F       I          M          V      Error              Rate
0  74772.0     0.0     1189.0        0.0   0.015653    1,189 / 75,961
1      0.0   629.0      710.0       71.0   0.553901       781 / 1,410
2    266.0     2.0   168119.0        0.0   0.001592     268 / 168,387
3      0.0     1.0        32.0    20542.0  0.001604        33 / 20,575
4  75038.0   632.0   170050.0    20613.0  0.008527    2,271 / 266,333



Top-4 Hit Ratios:


      k  hit_ratio
0  1    0.991473
1  2    0.999748
2  3    0.999989
3  4    1.000000




Scoring History:


                 timestamp      duration   number_of_trees   training_rmse  \
0     2020-04-04 23:07:10   0.018 sec                   0.0             NaN
1     2020-04-04 23:07:12   2.050 sec                   1.0        0.103695
2     2020-04-04 23:07:14   3.844 sec                   2.0        0.095218
3     2020-04-04 23:07:16   6.100 sec                   3.0        0.092595
4     2020-04-04 23:07:18   8.430 sec                   4.0        0.091148
5     2020-04-04 23:07:21  10.810 sec                   5.0        0.090288
```

```
6     2020-04-04 23:07:24  13.550 sec                 6.0          0.089748
7     2020-04-04 23:07:26  16.278 sec                 7.0          0.089364
8     2020-04-04 23:07:29  19.189 sec                 8.0          0.089033
9     2020-04-04 23:07:32  22.120 sec                 9.0          0.088683
10    2020-04-04 23:07:35  25.293 sec                10.0          0.088459
11    2020-04-04 23:07:39  28.614 sec                11.0          0.088302
12    2020-04-04 23:07:42  32.331 sec                12.0          0.089790
13    2020-04-04 23:07:46  35.909 sec                13.0          0.089455
14    2020-04-04 23:07:50  39.794 sec                14.0          0.089164
15    2020-04-04 23:07:54  44.109 sec                15.0          0.089371
16    2020-04-04 23:07:58  48.403 sec                16.0          0.089119
17    2020-04-04 23:08:03  52.771 sec                17.0          0.088915
18    2020-04-04 23:08:07  57.139 sec                18.0          0.088761

      training_logloss  training_classification_error  validation_rmse  \
0              NaN                            NaN              NaN
1         0.051007                       0.008943         0.102997
2         0.040597                       0.008827         0.092417
3         0.036993                       0.008882         0.089348
4         0.035597                       0.008855         0.088099
5         0.034112                       0.008850         0.087375
6         0.033708                       0.008828         0.087004
7         0.033000                       0.008837         0.086788
8         0.032466                       0.008828         0.086622
9         0.031747                       0.008813         0.086492
10        0.031275                       0.008811         0.086418
11        0.031022                       0.008824         0.086340
12        0.033337                       0.008801         0.087080
13        0.032987                       0.008783         0.086933
14        0.032583                       0.008779         0.086789
15        0.033356                       0.008776         0.086876
16        0.033082                       0.008769         0.086763
17        0.032760                       0.008769         0.086687
18        0.032529                       0.008772         0.086624

      validation_logloss  validation_classification_error
0              NaN                            NaN
1         0.049310                       0.008794
2         0.035531                       0.008613
3         0.032043                       0.008598
4         0.031082                       0.008591
5         0.030484                       0.008523
6         0.030347                       0.008512
7         0.030338                       0.008523
8         0.030145                       0.008516
9         0.029882                       0.008523
10        0.029681                       0.008527
11        0.029493                       0.008523
```

```
12          0.032044                    0.008531
13          0.031729                    0.008527
14          0.031474                    0.008527
15          0.032308                    0.008519
16          0.032066                    0.008516
17          0.031816                    0.008519
18          0.031591                    0.008527
```

Variable Importances:

```
    variable  relative_importance  scaled_importance  percentage
0    PD_DESC          7.348236e+06           1.000000    0.997538
1   PERP_SEX          8.802860e+03           0.001198    0.001195
2  AGE_GROUP          5.206346e+03           0.000709    0.000707
3  PERP_RACE          4.124638e+03           0.000561    0.000560
```

```
              timestamp     duration  number_of_trees  training_rmse  \
0   2020-04-04 23:07:10   0.018 sec               0.0            NaN
1   2020-04-04 23:07:12   2.050 sec               1.0       0.103695
2   2020-04-04 23:07:14   3.844 sec               2.0       0.095218
3   2020-04-04 23:07:16   6.100 sec               3.0       0.092595
4   2020-04-04 23:07:18   8.430 sec               4.0       0.091148
5   2020-04-04 23:07:21  10.810 sec               5.0       0.090288
6   2020-04-04 23:07:24  13.550 sec               6.0       0.089748
7   2020-04-04 23:07:26  16.278 sec               7.0       0.089364
8   2020-04-04 23:07:29  19.189 sec               8.0       0.089033
9   2020-04-04 23:07:32  22.120 sec               9.0       0.088683
10  2020-04-04 23:07:35  25.293 sec              10.0       0.088459
11  2020-04-04 23:07:39  28.614 sec              11.0       0.088302
12  2020-04-04 23:07:42  32.331 sec              12.0       0.089790
13  2020-04-04 23:07:46  35.909 sec              13.0       0.089455
14  2020-04-04 23:07:50  39.794 sec              14.0       0.089164
15  2020-04-04 23:07:54  44.109 sec              15.0       0.089371
16  2020-04-04 23:07:58  48.403 sec              16.0       0.089119
17  2020-04-04 23:08:03  52.771 sec              17.0       0.088915
18  2020-04-04 23:08:07  57.139 sec              18.0       0.088761
```

```
    training_logloss  training_classification_error  validation_rmse  \
0                NaN                            NaN              NaN
1           0.051007                       0.008943         0.102997
2           0.040597                       0.008827         0.092417
3           0.036993                       0.008882         0.089348
4           0.035597                       0.008855         0.088099
```

```
5           0.034112                      0.008850          0.087375
6           0.033708                      0.008828          0.087004
7           0.033000                      0.008837          0.086788
8           0.032466                      0.008828          0.086622
9           0.031747                      0.008813          0.086492
10          0.031275                      0.008811          0.086418
11          0.031022                      0.008824          0.086340
12          0.033337                      0.008801          0.087080
13          0.032987                      0.008783          0.086933
14          0.032583                      0.008779          0.086789
15          0.033356                      0.008776          0.086876
16          0.033082                      0.008769          0.086763
17          0.032760                      0.008769          0.086687
18          0.032529                      0.008772          0.086624


    validation_logloss  validation_classification_error
0               NaN                                  NaN
1          0.049310                             0.008794
2          0.035531                             0.008613
3          0.032043                             0.008598
4          0.031082                             0.008591
5          0.030484                             0.008523
6          0.030347                             0.008512
7          0.030338                             0.008523
8          0.030145                             0.008516
9          0.029882                             0.008523
10         0.029681                             0.008527
11         0.029493                             0.008523
12         0.032044                             0.008531
13         0.031729                             0.008527
14         0.031474                             0.008527
15         0.032308                             0.008519
16         0.032066                             0.008516
17         0.031816                             0.008519
18         0.031591                             0.008527
Parse progress: || 100%
Parse progress: || 100%
drf Model Build progress: || 100%
Model Details
=============
H2ORandomForestEstimator :  Distributed Random Forest
Model Key:  rf_M


Model Summary:


     number_of_trees  number_of_internal_trees  model_size_in_bytes  \
```

```
0                21.0                   84.0                   313167.0

    min_depth  max_depth  mean_depth  min_leaves  max_leaves  mean_leaves
0         8.0       15.0   12.059524        23.0       274.0    133.40475
```

ModelMetricsMultinomial: drf
** Reported on train data. **

MSE: 0.005186149758704385
RMSE: 0.07201492733249396
LogLoss: 0.027993786119493488
Mean Per-Class Error: 0.09385858053836384

Confusion Matrix: Row labels: Actual class; Column labels: Predicted class

```
           F        I          M          V     Error                    Rate
0   234508.0      1.0     3484.0        0.0  0.014643      3,485 / 237,993
1        0.0   2124.0      881.0      313.0  0.359855        1,194 / 3,318
2      313.0    110.0   707462.0        0.0  0.000598        423 / 707,885
3        0.0      0.0       29.0    85733.0  0.000338         29 / 85,762
4   234821.0   2235.0   711856.0    86046.0  0.004958    5,131 / 1,034,958
```

Top-4 Hit Ratios:

```
   k  hit_ratio
0  1   0.995042
1  2   0.999844
2  3   0.999993
3  4   1.000000
```

ModelMetricsMultinomial: drf
** Reported on validation data. **

MSE: 0.004743203790480386
RMSE: 0.06887092122572766
LogLoss: 0.028013911209936484
Mean Per-Class Error: 0.09571930927064987

Confusion Matrix: Row labels: Actual class; Column labels: Predicted class

```
          F        I           M          V      Error              Rate
0  58688.0      0.0       901.0        0.0  0.015120       901 / 59,589
1      0.0    478.0       214.0       63.0  0.366887          277 / 755
2     67.0     29.0    176930.0        0.0  0.000542      96 / 177,026
3      0.0      0.0         7.0    21381.0  0.000327         7 / 21,388
4  58755.0    507.0    178052.0    21444.0  0.004951    1,281 / 258,758
```

Top-4 Hit Ratios:

```
   k  hit_ratio
0  1   0.995049
1  2   0.999729
2  3   0.999969
3  4   1.000000
```

Scoring History:

	timestamp	duration	number_of_trees	training_rmse \
0	2020-04-04 23:08:58	0.022 sec	0.0	NaN
1	2020-04-04 23:09:00	2.374 sec	1.0	0.166039
2	2020-04-04 23:09:02	4.413 sec	2.0	0.123979
3	2020-04-04 23:09:05	6.668 sec	3.0	0.106279
4	2020-04-04 23:09:07	8.808 sec	4.0	0.095911
5	2020-04-04 23:09:09	11.122 sec	5.0	0.089497
6	2020-04-04 23:09:12	13.929 sec	6.0	0.085209
7	2020-04-04 23:09:15	16.881 sec	7.0	0.082022
8	2020-04-04 23:09:19	20.597 sec	8.0	0.079456
9	2020-04-04 23:09:22	23.562 sec	9.0	0.077304
10	2020-04-04 23:09:25	27.262 sec	10.0	0.075789
11	2020-04-04 23:09:29	31.196 sec	11.0	0.074580
12	2020-04-04 23:09:33	34.788 sec	12.0	0.073692
13	2020-04-04 23:09:37	38.699 sec	13.0	0.072847
14	2020-04-04 23:09:41	43.149 sec	14.0	0.072103
15	2020-04-04 23:09:46	47.820 sec	15.0	0.075292
16	2020-04-04 23:09:50	52.231 sec	16.0	0.074403
17	2020-04-04 23:09:55	57.234 sec	17.0	0.073639
18	2020-04-04 23:10:00	1 min 1.649 sec	18.0	0.072963
19	2020-04-04 23:10:04	1 min 6.498 sec	19.0	0.072374

	training_logloss	training_classification_error	validation_rmse \
0	NaN	NaN	NaN
1	0.108808	0.005319	0.166620
2	0.065191	0.005108	0.115016

3	0.050088	0.005063	0.095141
4	0.043323	0.005048	0.085312
5	0.038516	0.005017	0.079697
6	0.035525	0.005021	0.076265
7	0.033542	0.005008	0.074058
8	0.031699	0.004989	0.072515
9	0.029933	0.004977	0.071375
10	0.028702	0.004977	0.070537
11	0.027752	0.004976	0.069908
12	0.027185	0.004975	0.069485
13	0.026499	0.004973	0.069099
14	0.025806	0.004969	0.068774
15	0.030603	0.004967	0.070317
16	0.029843	0.004963	0.069909
17	0.029192	0.004961	0.069556
18	0.028481	0.004962	0.069253
19	0.027942	0.004957	0.069000

	validation_logloss	validation_classification_error
0	NaN	NaN
1	0.113532	0.005318
2	0.065646	0.005005
3	0.050401	0.004970
4	0.042527	0.004966
5	0.037757	0.004954
6	0.034504	0.004947
7	0.032223	0.004947
8	0.030603	0.004943
9	0.029191	0.004947
10	0.028107	0.004947
11	0.027241	0.004947
12	0.026671	0.004947
13	0.026059	0.004947
14	0.025484	0.004947
15	0.030435	0.004947
16	0.029722	0.004947
17	0.029041	0.004954
18	0.028420	0.004954
19	0.027890	0.004947

See the whole table with table.as_data_frame()

Variable Importances:

    variable relative_importance scaled_importance percentage

```
0     PD_DESC        7.800080e+06            1.000000    0.997014
1     PERP_SEX       1.110029e+04            0.001423    0.001419
2    PERP_RACE       8.960558e+03            0.001149    0.001145
3    AGE_GROUP       3.299111e+03            0.000423    0.000422
```

```
                 timestamp           duration  number_of_trees  training_rmse  \
0      2020-04-04 23:08:58        0.022 sec               0.0            NaN
1      2020-04-04 23:09:00        2.374 sec               1.0       0.166039
2      2020-04-04 23:09:02        4.413 sec               2.0       0.123979
3      2020-04-04 23:09:05        6.668 sec               3.0       0.106279
4      2020-04-04 23:09:07        8.808 sec               4.0       0.095911
5      2020-04-04 23:09:09       11.122 sec               5.0       0.089497
6      2020-04-04 23:09:12       13.929 sec               6.0       0.085209
7      2020-04-04 23:09:15       16.881 sec               7.0       0.082022
8      2020-04-04 23:09:19       20.597 sec               8.0       0.079456
9      2020-04-04 23:09:22       23.562 sec               9.0       0.077304
10     2020-04-04 23:09:25       27.262 sec              10.0       0.075789
11     2020-04-04 23:09:29       31.196 sec              11.0       0.074580
12     2020-04-04 23:09:33       34.788 sec              12.0       0.073692
13     2020-04-04 23:09:37       38.699 sec              13.0       0.072847
14     2020-04-04 23:09:41       43.149 sec              14.0       0.072103
15     2020-04-04 23:09:46       47.820 sec              15.0       0.075292
16     2020-04-04 23:09:50       52.231 sec              16.0       0.074403
17     2020-04-04 23:09:55       57.234 sec              17.0       0.073639
18     2020-04-04 23:10:00  1 min  1.649 sec              18.0       0.072963
19     2020-04-04 23:10:04  1 min  6.498 sec              19.0       0.072374
20     2020-04-04 23:10:09  1 min 11.361 sec              20.0       0.071883
21     2020-04-04 23:10:15  1 min 17.112 sec              21.0       0.072015
```

```
     training_logloss  training_classification_error  validation_rmse  \
0              NaN                            NaN              NaN
1         0.108808                       0.005319         0.166620
2         0.065191                       0.005108         0.115016
3         0.050088                       0.005063         0.095141
4         0.043323                       0.005048         0.085312
5         0.038516                       0.005017         0.079697
6         0.035525                       0.005021         0.076265
7         0.033542                       0.005008         0.074058
8         0.031699                       0.004989         0.072515
9         0.029933                       0.004977         0.071375
10        0.028702                       0.004977         0.070537
11        0.027752                       0.004976         0.069908
12        0.027185                       0.004975         0.069485
13        0.026499                       0.004973         0.069099
14        0.025806                       0.004969         0.068774
15        0.030603                       0.004967         0.070317
```

16	0.029843	0.004963	0.069909
17	0.029192	0.004961	0.069556
18	0.028481	0.004962	0.069253
19	0.027942	0.004957	0.069000
20	0.027438	0.004959	0.068784
21	0.027994	0.004958	0.068871

	validation_logloss	validation_classification_error
0	NaN	NaN
1	0.113532	0.005318
2	0.065646	0.005005
3	0.050401	0.004970
4	0.042527	0.004966
5	0.037757	0.004954
6	0.034504	0.004947
7	0.032223	0.004947
8	0.030603	0.004943
9	0.029191	0.004947
10	0.028107	0.004947
11	0.027241	0.004947
12	0.026671	0.004947
13	0.026059	0.004947
14	0.025484	0.004947
15	0.030435	0.004947
16	0.029722	0.004947
17	0.029041	0.004954
18	0.028420	0.004954
19	0.027890	0.004947
20	0.027406	0.004947
21	0.028014	0.004951

Parse progress: || 100%
Parse progress: || 100%
drf Model Build progress: || 100%
Model Details
=============
H2ORandomForestEstimator : Distributed Random Forest
Model Key: rf_S

Model Summary:

	number_of_trees	number_of_internal_trees	model_size_in_bytes	\
0	72.0	288.0	559503.0	

	min_depth	max_depth	mean_depth	min_leaves	max_leaves	mean_leaves
0	4.0	16.0	11.017361	7.0	176.0	77.21875

```
ModelMetricsMultinomial: drf
** Reported on train data. **

MSE: 0.007346465050413919
RMSE: 0.08571152227334385
LogLoss: 0.0384756856978329
Mean Per-Class Error: 0.10515733899416524

Confusion Matrix: Row labels: Actual class; Column labels: Predicted class
```

	F	I	M	V	Error	Rate
0	38371.0	0.0	766.0	0.0	0.019572	766 / 39,137
1	3.0	260.0	154.0	8.0	0.388235	165 / 425
2	112.0	8.0	92113.0	0.0	0.001301	120 / 92,233
3	4.0	0.0	6.0	858.0	0.011521	10 / 868
4	38490.0	268.0	93039.0	866.0	0.007998	1,061 / 132,663

```
Top-4 Hit Ratios:
```

	k	hit_ratio
0	1	0.992002
1	2	0.999812
2	3	0.999955
3	4	1.000000

```
ModelMetricsMultinomial: drf
** Reported on validation data. **

MSE: 0.007450859906088194
RMSE: 0.08631836366665087
LogLoss: 0.03921656968285433
Mean Per-Class Error: 0.14390466445553032

Confusion Matrix: Row labels: Actual class; Column labels: Predicted class
```

	F	I	M	V	Error	Rate
0	9516.0	0.0	191.0	0.0	0.019677	191 / 9,707
1	0.0	45.0	51.0	4.0	0.550000	55 / 100
2	16.0	1.0	23150.0	0.0	0.000734	17 / 23,167
3	1.0	0.0	0.0	191.0	0.005208	1 / 192

```
4  9533.0  46.0  23392.0  195.0  0.007960  264 / 33,166
```

Top-4 Hit Ratios:

```
   k  hit_ratio
0  1   0.992040
1  2   0.999879
2  3   0.999940
3  4   1.000000
```

Scoring History:

```
                     timestamp      duration   number_of_trees   training_rmse  \
0    2020-04-04 23:10:55    0.056 sec               0.0             NaN
1    2020-04-04 23:10:55    0.296 sec               1.0        0.226961
2    2020-04-04 23:10:56    0.561 sec               2.0        0.165825
3    2020-04-04 23:10:56    0.849 sec               3.0        0.140145
4    2020-04-04 23:10:56    1.282 sec               4.0        0.125124
5    2020-04-04 23:10:57    1.796 sec               5.0        0.114856
6    2020-04-04 23:10:58    2.497 sec               6.0        0.108226
7    2020-04-04 23:10:58    3.192 sec               7.0        0.104147
8    2020-04-04 23:10:59    3.829 sec               8.0        0.100035
9    2020-04-04 23:11:00    4.449 sec               9.0        0.096980
10   2020-04-04 23:11:00    5.151 sec              10.0        0.094922
11   2020-04-04 23:11:01    5.863 sec              11.0        0.093135
12   2020-04-04 23:11:02    6.513 sec              12.0        0.096629
13   2020-04-04 23:11:02    7.263 sec              13.0        0.094786
14   2020-04-04 23:11:03    7.982 sec              14.0        0.093312
15   2020-04-04 23:11:04    8.749 sec              15.0        0.092905
16   2020-04-04 23:11:05    9.535 sec              16.0        0.091924
17   2020-04-04 23:11:06   10.380 sec              17.0        0.090973
18   2020-04-04 23:11:06   11.271 sec              18.0        0.090195
19   2020-04-04 23:11:07   12.177 sec              19.0        0.089566

    training_logloss   training_classification_error   validation_rmse  \
0             NaN                            NaN               NaN
1        0.219991                       0.047796          0.227545
2        0.130328                       0.023896          0.142502
3        0.102013                       0.016387          0.116691
4        0.089853                       0.012812          0.105161
5        0.074680                       0.010814          0.098635
6        0.065567                       0.009814          0.095051
7        0.062841                       0.009299          0.092862
```

8	0.056350	0.008776	0.091060
9	0.051707	0.008500	0.089832
10	0.049436	0.008367	0.088979
11	0.047377	0.008238	0.088421
12	0.052354	0.008182	0.090589
13	0.050047	0.008080	0.089795
14	0.048098	0.008027	0.089139
15	0.047328	0.008000	0.089227
16	0.046395	0.008004	0.088834
17	0.044568	0.007993	0.088408
18	0.043259	0.007977	0.088025
19	0.042452	0.007984	0.087722

	validation_logloss	validation_classification_error
0	NaN	NaN
1	0.223925	0.049056
2	0.110307	0.008382
3	0.084242	0.008231
4	0.069287	0.008201
5	0.058296	0.008171
6	0.054095	0.008171
7	0.050323	0.008171
8	0.047905	0.008141
9	0.044297	0.008141
10	0.042820	0.008141
11	0.041729	0.008141
12	0.047622	0.008141
13	0.046254	0.008141
14	0.045070	0.008141
15	0.045226	0.008081
16	0.044318	0.008081
17	0.043466	0.008081
18	0.042697	0.008081
19	0.042069	0.008020

See the whole table with table.as_data_frame()

Variable Importances:

	variable	relative_importance	scaled_importance	percentage
0	PD_DESC	3.048465e+06	1.000000	0.996221
1	PERP_RACE	7.735905e+03	0.002538	0.002528
2	PERP_SEX	2.010547e+03	0.000660	0.000657
3	AGE_GROUP	1.817694e+03	0.000596	0.000594

```
              timestamp              duration  number_of_trees  training_rmse  \
0    2020-04-04 23:10:55          0.056 sec              0.0            NaN
1    2020-04-04 23:10:55          0.296 sec              1.0       0.226961
2    2020-04-04 23:10:56          0.561 sec              2.0       0.165825
3    2020-04-04 23:10:56          0.849 sec              3.0       0.140145
4    2020-04-04 23:10:56          1.282 sec              4.0       0.125124
5    2020-04-04 23:10:57          1.796 sec              5.0       0.114856
6    2020-04-04 23:10:58          2.497 sec              6.0       0.108226
7    2020-04-04 23:10:58          3.192 sec              7.0       0.104147
8    2020-04-04 23:10:59          3.829 sec              8.0       0.100035
9    2020-04-04 23:11:00          4.449 sec              9.0       0.096980
10   2020-04-04 23:11:00          5.151 sec             10.0       0.094922
11   2020-04-04 23:11:01          5.863 sec             11.0       0.093135
12   2020-04-04 23:11:02          6.513 sec             12.0       0.096629
13   2020-04-04 23:11:02          7.263 sec             13.0       0.094786
14   2020-04-04 23:11:03          7.982 sec             14.0       0.093312
15   2020-04-04 23:11:04          8.749 sec             15.0       0.092905
16   2020-04-04 23:11:05          9.535 sec             16.0       0.091924
17   2020-04-04 23:11:06         10.380 sec             17.0       0.090973
18   2020-04-04 23:11:06         11.271 sec             18.0       0.090195
19   2020-04-04 23:11:07         12.177 sec             19.0       0.089566
20   2020-04-04 23:11:08         13.154 sec             20.0       0.088963
21   2020-04-04 23:11:09         14.204 sec             21.0       0.088840
22   2020-04-04 23:11:11         15.338 sec             22.0       0.089836
23   2020-04-04 23:11:12         16.520 sec             23.0       0.089381
24   2020-04-04 23:11:13         17.777 sec             24.0       0.088905
25   2020-04-04 23:11:14         19.139 sec             25.0       0.088586
26   2020-04-04 23:11:16         20.431 sec             26.0       0.088806
27   2020-04-04 23:11:17         21.793 sec             27.0       0.088492
28   2020-04-04 23:11:18         23.319 sec             28.0       0.088252
29   2020-04-04 23:11:20         24.847 sec             29.0       0.087899
.. ..                     ...                  ...              ...            ...
43   2020-04-04 23:11:49         54.095 sec             43.0       0.086432
44   2020-04-04 23:11:52         56.453 sec             44.0       0.086342
45   2020-04-04 23:11:54         58.809 sec             45.0       0.086196
46   2020-04-04 23:11:56   1 min  1.230 sec             46.0       0.086126
47   2020-04-04 23:11:59   1 min  4.074 sec             47.0       0.085994
48   2020-04-04 23:12:02   1 min  6.671 sec             48.0       0.085878
49   2020-04-04 23:12:04   1 min  9.210 sec             49.0       0.085822
50   2020-04-04 23:12:07   1 min 11.674 sec             50.0       0.085776
51   2020-04-04 23:12:09   1 min 14.219 sec             51.0       0.085716
52   2020-04-04 23:12:12   1 min 16.817 sec             52.0       0.085619
53   2020-04-04 23:12:15   1 min 19.419 sec             53.0       0.085674
54   2020-04-04 23:12:17   1 min 22.098 sec             54.0       0.085575
55   2020-04-04 23:12:21   1 min 25.810 sec             55.0       0.085489
56   2020-04-04 23:12:25   1 min 29.489 sec             56.0       0.085446
57   2020-04-04 23:12:28   1 min 33.021 sec             57.0       0.085402
```

```
58     2020-04-04 23:12:31   1 min 35.983 sec        58.0       0.085330
59     2020-04-04 23:12:34   1 min 38.876 sec        59.0       0.085260
60     2020-04-04 23:12:37   1 min 41.838 sec        60.0       0.085199
61     2020-04-04 23:12:40   1 min 44.953 sec        61.0       0.085130
62     2020-04-04 23:12:43   1 min 48.074 sec        62.0       0.085073
63     2020-04-04 23:12:47   1 min 51.375 sec        63.0       0.085061
64     2020-04-04 23:12:50   1 min 54.607 sec        64.0       0.085033
65     2020-04-04 23:12:53   1 min 57.988 sec        65.0       0.085126
66     2020-04-04 23:12:57   2 min  1.391 sec        66.0       0.085480
67     2020-04-04 23:13:00   2 min  4.909 sec        67.0       0.085414
68     2020-04-04 23:13:04   2 min  8.418 sec        68.0       0.085732
69     2020-04-04 23:13:08   2 min 12.362 sec        69.0       0.085720
70     2020-04-04 23:13:11   2 min 16.243 sec        70.0       0.085655
71     2020-04-04 23:13:16   2 min 20.442 sec        71.0       0.085585
72     2020-04-04 23:13:19   2 min 24.075 sec        72.0       0.085712

    training_logloss  training_classification_error  validation_rmse  \
0            NaN                            NaN              NaN
1       0.219991                       0.047796         0.227545
2       0.130328                       0.023896         0.142502
3       0.102013                       0.016387         0.116691
4       0.089853                       0.012812         0.105161
5       0.074680                       0.010814         0.098635
6       0.065567                       0.009814         0.095051
7       0.062841                       0.009299         0.092862
8       0.056350                       0.008776         0.091060
9       0.051707                       0.008500         0.089832
10      0.049436                       0.008367         0.088979
11      0.047377                       0.008238         0.088421
12      0.052354                       0.008182         0.090589
13      0.050047                       0.008080         0.089795
14      0.048098                       0.008027         0.089139
15      0.047328                       0.008000         0.089227
16      0.046395                       0.008004         0.088834
17      0.044568                       0.007993         0.088408
18      0.043259                       0.007977         0.088025
19      0.042452                       0.007984         0.087722
20      0.041777                       0.007976         0.087448
21      0.041698                       0.007945         0.087403
22      0.044085                       0.007998         0.088103
23      0.043571                       0.008013         0.087859
24      0.042986                       0.007975         0.087593
25      0.042521                       0.007983         0.087426
26      0.043367                       0.008013         0.087668
27      0.042514                       0.008043         0.087548
28      0.042086                       0.008043         0.087419
29      0.041386                       0.008035         0.087227
..           ...                            ...              ...
```

43	0.038706	0.008035	0.086537
44	0.038488	0.008043	0.086484
45	0.038275	0.008058	0.086420
46	0.038073	0.008066	0.086393
47	0.037857	0.008050	0.086324
48	0.037644	0.008020	0.086266
49	0.037469	0.008035	0.086235
50	0.037311	0.008050	0.086210
51	0.037144	0.008073	0.086187
52	0.036964	0.008050	0.086125
53	0.037193	0.008050	0.086175
54	0.037008	0.008050	0.086104
55	0.036840	0.008035	0.086050
56	0.036712	0.008035	0.086029
57	0.036548	0.008035	0.086003
58	0.036415	0.008035	0.085975
59	0.036272	0.008028	0.085932
60	0.036151	0.007998	0.085883
61	0.036007	0.007983	0.085839
62	0.035878	0.007975	0.085802
63	0.035813	0.007990	0.085798
64	0.035699	0.007983	0.085776
65	0.036153	0.007983	0.085839
66	0.037270	0.007990	0.086103
67	0.037127	0.007990	0.086064
68	0.038020	0.007990	0.086310
69	0.038018	0.008005	0.086304
70	0.037869	0.007990	0.086261
71	0.037726	0.007983	0.086220
72	0.038476	0.007998	0.086318

	validation_logloss	validation_classification_error
0	NaN	NaN
1	0.223925	0.049056
2	0.110307	0.008382
3	0.084242	0.008231
4	0.069287	0.008201
5	0.058296	0.008171
6	0.054095	0.008171
7	0.050323	0.008171
8	0.047905	0.008141
9	0.044297	0.008141
10	0.042820	0.008141
11	0.041729	0.008141
12	0.047622	0.008141
13	0.046254	0.008141
14	0.045070	0.008141
15	0.045226	0.008081

16	0.044318	0.008081
17	0.043466	0.008081
18	0.042697	0.008081
19	0.042069	0.008020
20	0.041465	0.007930
21	0.041768	0.007960
22	0.043419	0.008081
23	0.042888	0.007990
24	0.042291	0.007960
25	0.041864	0.007960
26	0.042682	0.007960
27	0.042289	0.008020
28	0.041837	0.008111
29	0.041383	0.007960
..	...	...
43	0.039306	0.008020
44	0.039081	0.008020
45	0.038902	0.008020
46	0.038708	0.008020
47	0.038495	0.008020
48	0.038300	0.007960
49	0.038137	0.007990
50	0.037976	0.008020
51	0.037829	0.008020
52	0.037646	0.008020
53	0.037897	0.008020
54	0.037721	0.008020
55	0.037560	0.007960
56	0.037426	0.008020
57	0.037280	0.008020
58	0.037139	0.007990
59	0.037001	0.007960
60	0.036860	0.007960
61	0.036723	0.007960
62	0.036596	0.007960
63	0.036532	0.007960
64	0.036409	0.007960
65	0.036860	0.007960
66	0.037969	0.007960
67	0.037834	0.007960
68	0.038707	0.007960
69	0.038698	0.007960
70	0.038558	0.007960
71	0.038423	0.007960
72	0.039217	0.007960

[73 rows x 10 columns]
Parse progress: || 100%

```
Parse progress: || 100%
drf Model Build progress: || 100%
Model Details
=============
H2ORandomForestEstimator :  Distributed Random Forest
Model Key:  rf_B


Model Summary:


     number_of_trees  number_of_internal_trees  model_size_in_bytes  \
0               18.0                      72.0             225684.0


   min_depth  max_depth  mean_depth  min_leaves  max_leaves  mean_leaves
0        6.0       16.0   11.444445        18.0       281.0   114.388885




ModelMetricsMultinomial: drf
** Reported on train data. **

MSE: 0.004301819027950368
RMSE: 0.06558825373457025
LogLoss: 0.01762195615523751
Mean Per-Class Error: 0.0763673557544016

Confusion Matrix: Row labels: Actual class; Column labels: Predicted class


          F       I         M         V     Error              Rate
0  218978.0     0.0    3555.0       0.0  0.015975  3,555 / 222,533
1       0.0   855.0     283.0      61.0  0.286906       344 / 1,199
2     280.0    36.0  632586.0       0.0  0.000499     316 / 632,902
3       0.0     1.0      33.0   16240.0  0.002089      34 / 16,274
4  219258.0   892.0  636457.0   16301.0  0.004868  4,249 / 872,908



Top-4 Hit Ratios:


   k  hit_ratio
0  1   0.995132
1  2   0.999968
2  3   0.999992
3  4   1.000000
```

```
ModelMetricsMultinomial: drf
** Reported on validation data. **

MSE: 0.004327694745189184
RMSE: 0.0657852167678209
LogLoss: 0.017743214393310787
Mean Per-Class Error: 0.08434717159805417

Confusion Matrix: Row labels: Actual class; Column labels: Predicted class


          F        I          M        V      Error              Rate
0   54689.0      0.0      909.0      0.0   0.016350       909 / 55,598
1       0.0    207.0       84.0     13.0   0.319079             97 / 304
2      72.0     10.0   158137.0      0.0   0.000518       82 / 158,219
3       0.0      1.0        5.0   4155.0   0.001442          6 / 4,161
4   54761.0    218.0   159135.0   4168.0   0.005012   1,094 / 218,282



Top-4 Hit Ratios:


   k  hit_ratio
0  1   0.994988
1  2   0.999973
2  3   0.999995
3  4   1.000000



Scoring History:


                   timestamp      duration  number_of_trees  training_rmse  \
0      2020-04-04 23:14:10    0.032 sec                 0.0            NaN
1      2020-04-04 23:14:12    1.663 sec                 1.0       0.067871
2      2020-04-04 23:14:14    3.524 sec                 2.0       0.066201
3      2020-04-04 23:14:16    5.655 sec                 3.0       0.065632
4      2020-04-04 23:14:18    7.584 sec                 4.0       0.065524
5      2020-04-04 23:14:20    9.667 sec                 5.0       0.065255
6      2020-04-04 23:14:22   11.958 sec                 6.0       0.065056
7      2020-04-04 23:14:24   14.322 sec                 7.0       0.065896
8      2020-04-04 23:14:27   17.144 sec                 8.0       0.065649
9      2020-04-04 23:14:30   19.809 sec                 9.0       0.065457
10     2020-04-04 23:14:33   22.789 sec                10.0       0.065246
11     2020-04-04 23:14:36   25.878 sec                11.0       0.065151
12     2020-04-04 23:14:39   29.337 sec                12.0       0.065361
```

```
13    2020-04-04 23:14:43  32.769 sec              13.0           0.065231
14    2020-04-04 23:14:46  36.291 sec              14.0           0.065119
15    2020-04-04 23:14:50  40.176 sec              15.0           0.065969
16    2020-04-04 23:14:54  44.274 sec              16.0           0.065860
17    2020-04-04 23:14:58  48.253 sec              17.0           0.065706
18    2020-04-04 23:15:02  52.359 sec              18.0           0.065588


    training_logloss  training_classification_error  validation_rmse  \
0            NaN                           NaN                  NaN
1       0.028939                      0.005201             0.068914
2       0.022993                      0.005039             0.066048
3       0.020637                      0.004975             0.065604
4       0.019811                      0.004963             0.065573
5       0.018931                      0.004939             0.065458
6       0.018174                      0.004911             0.065472
7       0.018541                      0.004902             0.065921
8       0.018006                      0.004895             0.065786
9       0.017551                      0.004886             0.065689
10      0.017039                      0.004874             0.065639
11      0.016750                      0.004871             0.065582
12      0.017699                      0.004873             0.065683
13      0.017518                      0.004866             0.065628
14      0.017173                      0.004861             0.065588
15      0.018110                      0.004865             0.065948
16      0.017946                      0.004872             0.065898
17      0.017757                      0.004868             0.065840
18      0.017622                      0.004868             0.065785


    validation_logloss  validation_classification_error
0             NaN                              NaN
1        0.031050                         0.005388
2        0.019038                         0.005117
3        0.016551                         0.005044
4        0.016561                         0.005053
5        0.016279                         0.005021
6        0.016074                         0.005021
7        0.016771                         0.005016
8        0.016735                         0.005012
9        0.016599                         0.004998
10       0.016431                         0.005007
11       0.016350                         0.005012
12       0.017345                         0.005012
13       0.017236                         0.005012
14       0.017111                         0.005012
15       0.018171                         0.005012
16       0.018028                         0.005012
17       0.017888                         0.005012
18       0.017743                         0.005012
```

Variable Importances:

	variable	relative_importance	scaled_importance	percentage
0	PD_DESC	4.853892e+06	1.000000	0.998627
1	PERP_SEX	2.881350e+03	0.000594	0.000593
2	AGE_GROUP	2.620734e+03	0.000540	0.000539
3	PERP_RACE	1.169469e+03	0.000241	0.000241

	timestamp	duration	number_of_trees	training_rmse \
0	2020-04-04 23:14:10	0.032 sec	0.0	NaN
1	2020-04-04 23:14:12	1.663 sec	1.0	0.067871
2	2020-04-04 23:14:14	3.524 sec	2.0	0.066201
3	2020-04-04 23:14:16	5.655 sec	3.0	0.065632
4	2020-04-04 23:14:18	7.584 sec	4.0	0.065524
5	2020-04-04 23:14:20	9.667 sec	5.0	0.065255
6	2020-04-04 23:14:22	11.958 sec	6.0	0.065056
7	2020-04-04 23:14:24	14.322 sec	7.0	0.065896
8	2020-04-04 23:14:27	17.144 sec	8.0	0.065649
9	2020-04-04 23:14:30	19.809 sec	9.0	0.065457
10	2020-04-04 23:14:33	22.789 sec	10.0	0.065246
11	2020-04-04 23:14:36	25.878 sec	11.0	0.065151
12	2020-04-04 23:14:39	29.337 sec	12.0	0.065361
13	2020-04-04 23:14:43	32.769 sec	13.0	0.065231
14	2020-04-04 23:14:46	36.291 sec	14.0	0.065119
15	2020-04-04 23:14:50	40.176 sec	15.0	0.065969
16	2020-04-04 23:14:54	44.274 sec	16.0	0.065860
17	2020-04-04 23:14:58	48.253 sec	17.0	0.065706
18	2020-04-04 23:15:02	52.359 sec	18.0	0.065588

	training_logloss	training_classification_error	validation_rmse \
0	NaN	NaN	NaN
1	0.028939	0.005201	0.068914
2	0.022993	0.005039	0.066048
3	0.020637	0.004975	0.065604
4	0.019811	0.004963	0.065573
5	0.018931	0.004939	0.065458
6	0.018174	0.004911	0.065472
7	0.018541	0.004902	0.065921
8	0.018006	0.004895	0.065786
9	0.017551	0.004886	0.065689
10	0.017039	0.004874	0.065639
11	0.016750	0.004871	0.065582

12	0.017699	0.004873	0.065683
13	0.017518	0.004866	0.065628
14	0.017173	0.004861	0.065588
15	0.018110	0.004865	0.065948
16	0.017946	0.004872	0.065898
17	0.017757	0.004868	0.065840
18	0.017622	0.004868	0.065785

	validation_logloss	validation_classification_error
0	NaN	NaN
1	0.031050	0.005388
2	0.019038	0.005117
3	0.016551	0.005044
4	0.016561	0.005053
5	0.016279	0.005021
6	0.016074	0.005021
7	0.016771	0.005016
8	0.016735	0.005012
9	0.016599	0.004998
10	0.016431	0.005007
11	0.016350	0.005012
12	0.017345	0.005012
13	0.017236	0.005012
14	0.017111	0.005012
15	0.018171	0.005012
16	0.018028	0.005012
17	0.017888	0.005012
18	0.017743	0.005012

Here, we notice a relatively accurate model with an overall error of a mere 1.4593%, but a significant error in the prediction of Infractions.

Note that when we included the law code as a predictor (not shown in report), this error of infractions was also eliminated in a random forest model, result in a negligibly small error on all classifications. The following was the output obtained:

Parse progress: | | 100% Parse progress: | | 100% drf Model Build progress: | | 100%

4   Model Details

H2ORandomForestEstimator : Distributed Random Forest Model Key: rf_covType_v1

Model Summary: number_of_trees number_of_internal_trees model_size_in_bytes min_depth max_depth mean_depth min_leaves max_leaves mean_leaves -- ----------------- ---- ----------------------- -------------------- ---------- ---------- ----------- ----------- ----------- ------------- 51 204 533899 4 18 11.5882 7 237 80.7157

ModelMetricsMultinomial: drf ** Reported on train data. **

MSE: 0.0001734852476826297 RMSE: 0.013171379870105854 LogLoss: 0.003255583616165607 Mean Per-Class Error: 0.00033000525330283565

Confusion Matrix: Row labels: Actual class; Column labels: Predicted class F I M V Error Rate ------ ---- ----------- ------ ----------- -------------- 513090 0 28 0 5.45683e-05 28 / 513,118 0 7954 8 2 0.00125565 10 / 7,964 1 1 1.21166e+06 0 1.65062e-06 2 / 1,211,665 1 0 0 122674 8.15162e-06 1 / 122,675 513092 7955 1.2117e+06 122676 2.20974e-05 41 / 1,855,422

Top-4 Hit Ratios: k hit_ratio --- ----------- 1 0.999978 2 0.999999 3 1 4 1

ModelMetricsMultinomial: drf ** Reported on validation data. **

MSE: 0.00016996249120756354 RMSE: 0.013036966334525972 LogLoss: 0.0035461151687810766 Mean Per-Class Error: 0.0017149228257738725

Confusion Matrix: Row labels: Actual class; Column labels: Predicted class F I M V Error Rate ------ ---- ------ ----- ----------- ------------ 127951 0 14 0 0.000109405 14 / 127,965 0 2061 14 0 0.00674699 14 / 2,075 0 1 303173 0 3.29844e-06 1 / 303,174 0 0 0 30642 0 0 / 30,642 127951 2062 303201 30642 6.25194e-05 29 / 463,856

Top-4 Hit Ratios: k hit_ratio --- ----------- 1 0.999938 2 0.999963 3 0.999998 4 1

Variable Importances: variable relative_importance scaled_importance percentage -------- --- --------------------- ------------------- ------------ PD_DESC 3.08297e+07 1 0.853075 LAW_CODE 5.21138e+06 0.169038 0.144202 ARREST_BORO 50738.3 0.00164576 0.00140396 AGE_GROUP 27099.1 0.000878995 0.000749848 PERP_SEX 20594.9 0.00066802 0.000569871

Finally, we test the model trained on the Bronx borough (notorious for being relatively unsafe) against the data for the Manhattan borough (the most prestigious borough). If there were a bias, we would expect to see a lot of misclassifications, skewed towards more serious levels of offenses like Felony or Violation as opposed to Infraction or Misdemeanor.

```
In [53]: saved_model = h2o.load_model("C:/Users/varch/Desktop/Github/NYPD-Data-Comp/rf_B/rf_B")

In [54]: testingset = h2o.H2OFrame(nypd.loc[nypd['ARREST_BORO'] == 'M'])
         saved_model.model_performance(testingset)

Parse progress: || 100%

ModelMetricsMultinomial: drf
** Reported on test data. **

MSE: 0.0048710613504178765
RMSE: 0.06979298926409354
LogLoss: 0.023246667485524727
Mean Per-Class Error: 0.10366541412714929

Confusion Matrix: Row labels: Actual class; Column labels: Predicted class
```

	F	I	M	V	Error	Rate
0	293200.0	0.0	4404.0	0.0	0.014798	4,404 / 297,604
1	0.0	2458.0	1243.0	372.0	0.396514	1,615 / 4,073
2	387.0	75.0	884499.0	0.0	0.000522	462 / 884,961
3	0.0	180.0	123.0	106848.0	0.002828	303 / 107,151
4	293587.0	2713.0	890269.0	107220.0	0.005244	6,784 / 1,293,789

```
Top-4 Hit Ratios:


    k  hit_ratio
0   1   0.994756
1   2   0.999790
2   3   0.999930
3   4   1.000000
```

Out[54]:

However, we still get great predictions with an overall error rate of just 0.5244%

5 Conclusions

From all of our modeling and data exploration, we can conclude that the data does not represent discrimination on the basis of ethnographic variables. We reconcile this notion with the understanding that the data set is simply unable to capture the instances where discrimination occurs. Our recommendation for an accurate analysis of discrimination is to additionally capture two sets of occurences - before and after the arrests:

1. The stopping point before an arrest is actually made on the suspicion of illegal activity (arbitrary profiling)
2. The varying harshness of judgements given to individuals having committed the same crime but belonging to different demographics

Data representing these factors, when incorporated into our existing random forest model, is very likely to discern any patterns of discrimination present, considering the extreme accuracy with which our model currently provides predictions based on the provided data.