Homework 3 – Report

Importing Data & Data Cleaning

- After importing the necessary modules from several libraries and the data set, I renamed the columns in accordance with the codebook and replaced missing values indicated in the questionnaire with NULL.
- Then, I tried to fill the NULL values:
    - I've dropped the columns with larger than 54% of NULL values (since I want to keep "**SOCIO ECONOMIC STATUS**", with 53.51%, which can be an important indicator of voting behavior).
    - I've also dropped "**RELIGIOUS DENOMINATION**" column, since its values does not correspond to the ones shown in questionnaire.
    - For the columns with less than 10% of NULL values, I've filled them with the most common value.
    - For the other columns, I've filled them with a value randomly chosen from that column, therefore, the values with more counts will have a higher chance to be selected.

Exploratory Data Analysis

- I've divided the data set column-wise according to whether a variable is categorical, ordinal, or numerical, by checking the values they can take from the questionnaire.

Categorical Columns:

- GENDER
- MARITAL STATUS
- UNION MEMBERSHIP OF RESPONDENT
- UNION MEMBERSHIP OF OTHERS IN HOUSEHOLD
- CURRENT EMPLOYMENT STATUS
- EMPLOYMENT TYPE - PUBLIC OR PRIVATE
- INDUSTRIAL SECTOR
- LANGUAGE USUALLY SPOKEN AT HOME
- REGION OF RESIDENCE
- RACE

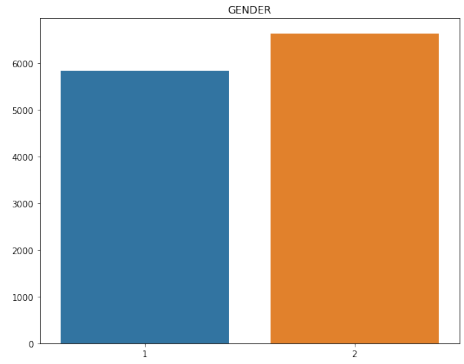
Ordinal Columns:

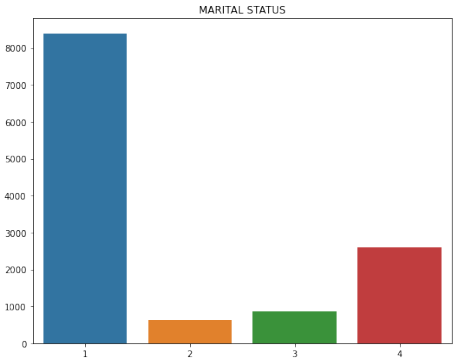
- EDUCATION
- SOCIO ECONOMIC STATUS
- HOUSEHOLD INCOME
- RELIGIOUS SERVICES ATTENDANCE
- RELIGIOSITY
- RURAL OR URBAN RESIDENCE
- NUMBER IN HOUSEHOLD IN TOTAL
- NUMBER OF CHILDREN IN HOUSEHOLD UNDER AGE 18
- NUMBER IN HOUSEHOLD UNDER AGE 6

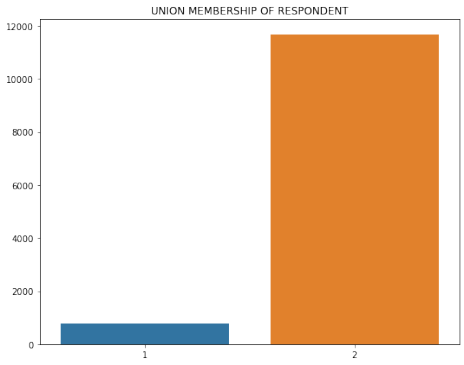
Numerical Columns:

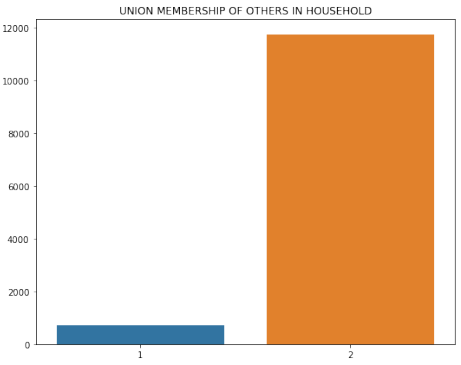
- AGE

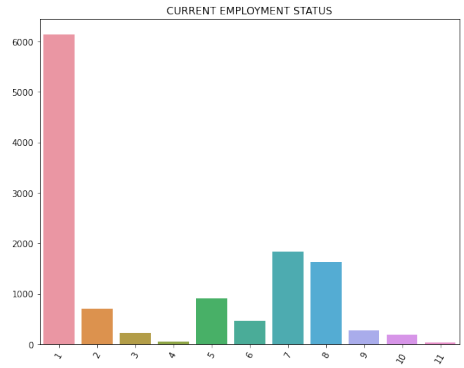
- Then, I've visualized each variable, by looking at the counts in categorical and ordinal variables, and distribution in the numerical variable. The plots in the following pages demonstrate these counts and distributions.
- I've also looked at the counts of the target variable, "VOTED", which indicated that the target variable is indeed imbalanced. Therefore, I've decided to look not only at the accuracy metric but also precision, recall, F1 score and ROC AUC metrics.
- Lastly, I have created a heatmap of the correlation between each of the variables. "AGE" variables seems to be the most correlated (0.26) with the target variable "VOTED", while "MARITAL STATUS" is the second mostly correlated with a negative correlation score (-0.23).

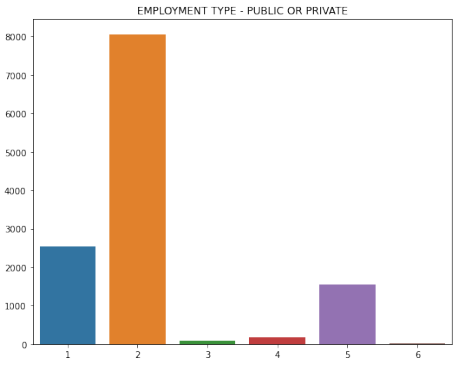

GENDER
0
1000
2000
3000
4000
5000
6000
1
2

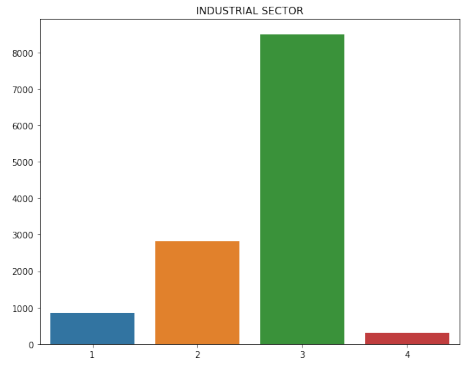

MARITAL STATUS
0
1000
2000
3000
4000
5000
6000
7000
8000
1
2
3
4

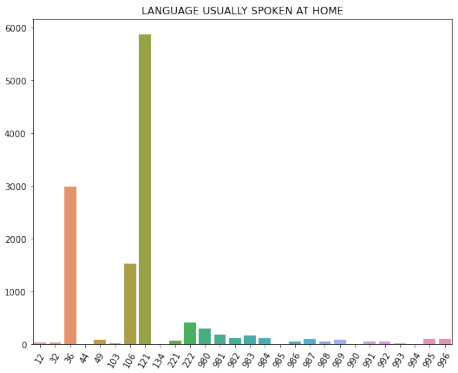

UNION MEMBERSHIP OF RESPONDENT
0
2000
4000
6000
8000
10000
12000
1
2

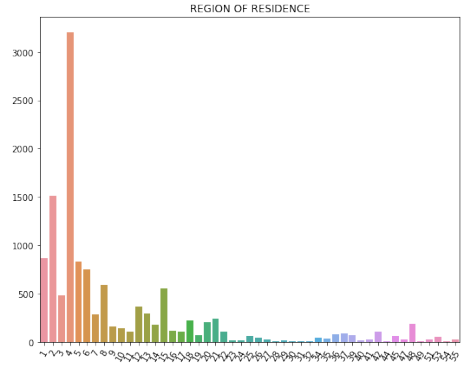

UNION MEMBERSHIP OF OTHERS IN HOUSEHOLD
0
2000
4000
6000
8000
10000
12000
1
2

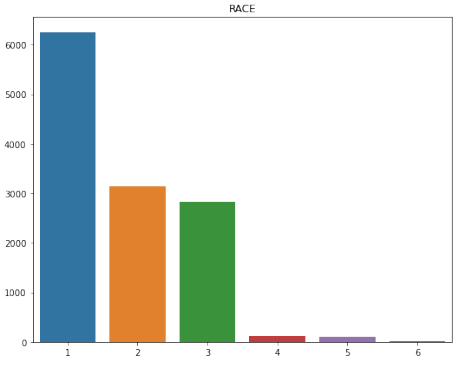

CURRENT EMPLOYMENT STATUS
0
1000
2000
3000
4000
5000
6000
1
2
3
4
5
6
7
8
9
10
11

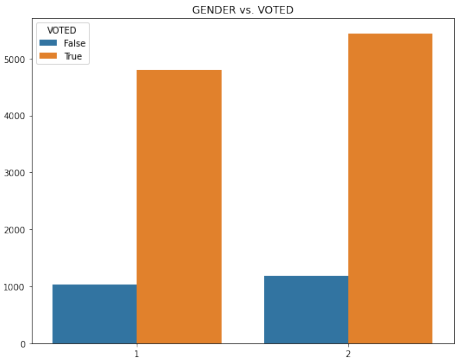

EMPLOYMENT TYPE - PUBLIC OR PRIVATE
0
1000
2000
3000
4000
5000
6000
7000
8000
1
2
3
4
5
6

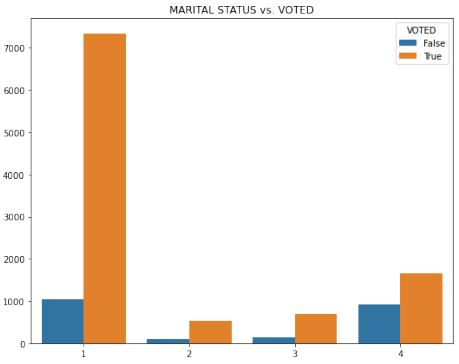

INDUSTRIAL SECTOR
0
1000
2000
3000
4000
5000
6000
7000
8000
1
2
3
4

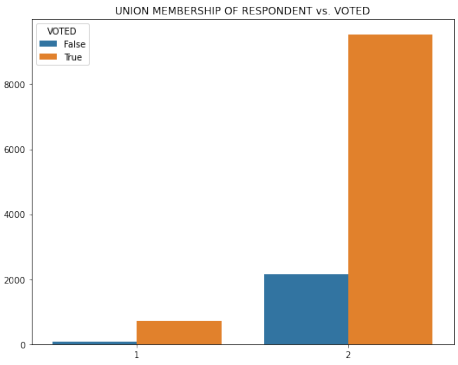

LANGUAGE USUALLY SPOKEN AT HOME
0
1000
2000
3000
4000
5000
6000
12
32
36
44
49
103
106
121
134
221
222
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994
995
998

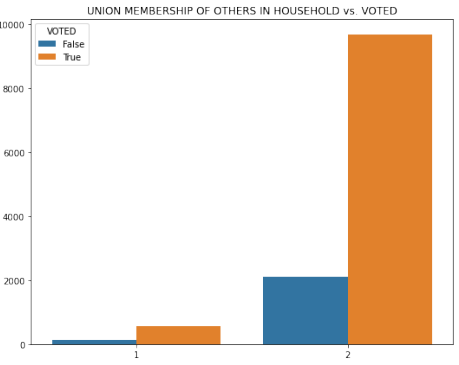

REGION OF RESIDENCE
0
500
1000
1500
2000
2500
3000

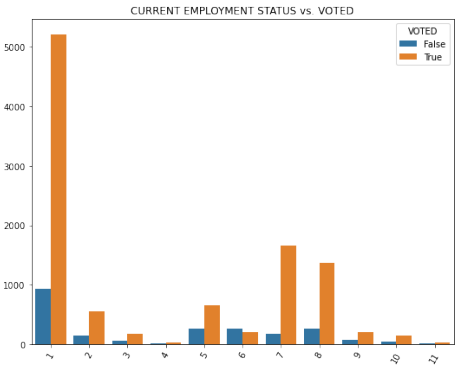

RACE
0
1000
2000
3000
4000
5000
6000
1
2
3
4
5
6

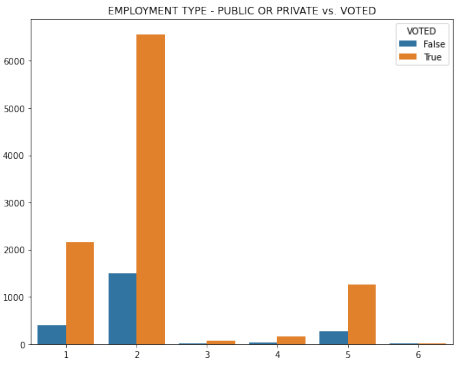

GENDER vs. VOTED
VOTED
False
True

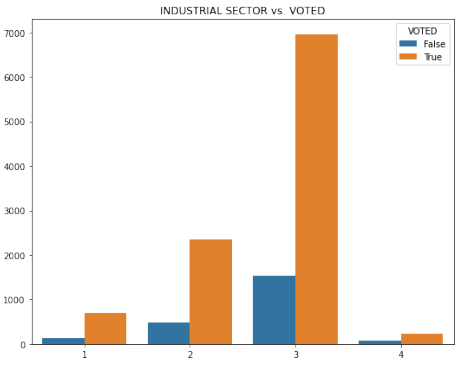

MARITAL STATUS vs. VOTED
VOTED
False
True

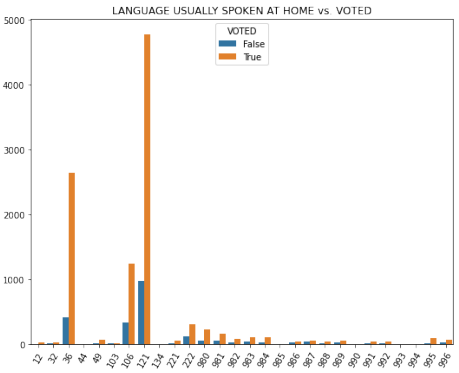

UNION MEMBERSHIP OF RESPONDENT vs. VOTED
VOTED
False
True

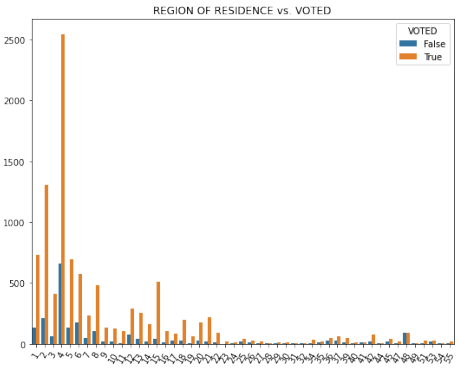

UNION MEMBERSHIP OF OTHERS IN HOUSEHOLD vs. VOTED
VOTED
False
True

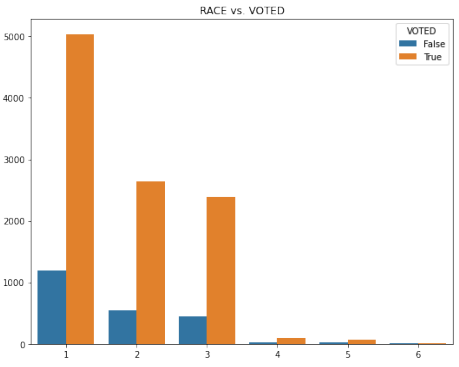

CURRENT EMPLOYMENT STATUS vs. VOTED
VOTED
False
True

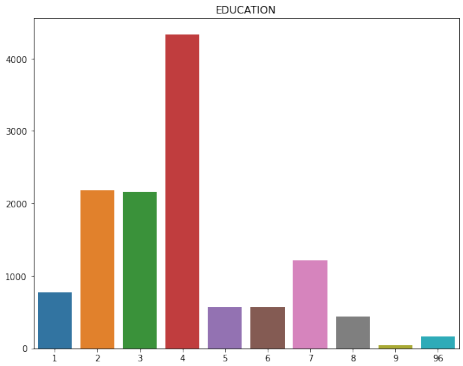

EMPLOYMENT TYPE - PUBLIC OR PRIVATE vs. VOTED
VOTED
False
True

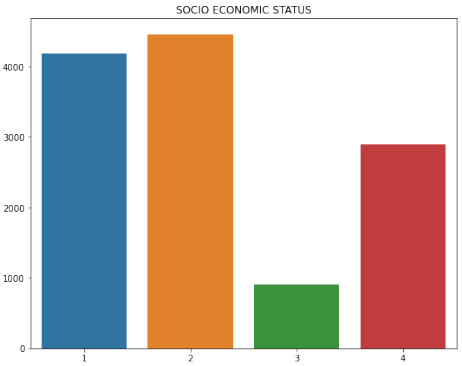

INDUSTRIAL SECTOR vs. VOTED
VOTED
False
True

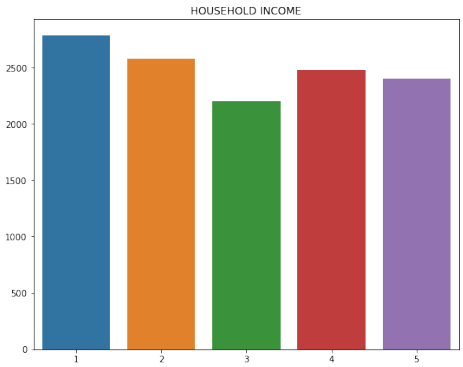

LANGUAGE USUALLY SPOKEN AT HOME vs. VOTED
VOTED
False
True

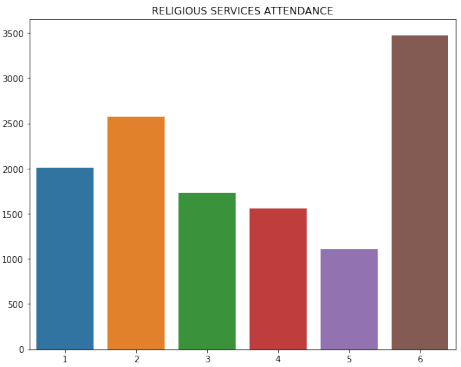

REGION OF RESIDENCE vs. VOTED
VOTED
False
True

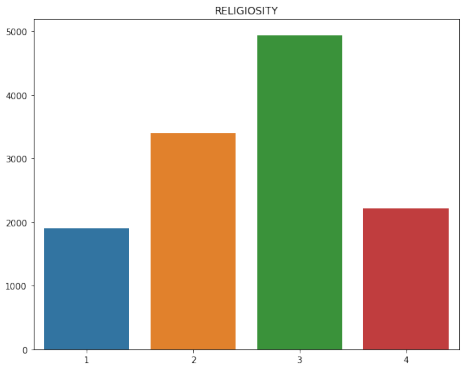

RACE vs. VOTED
VOTED
False
True

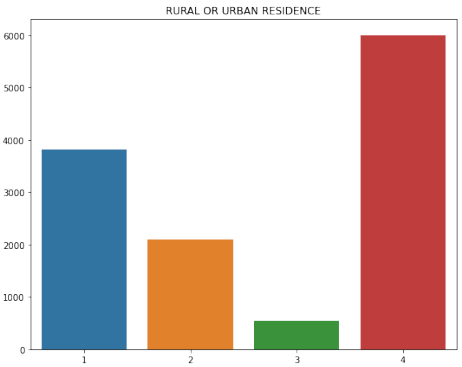

EDUCATION
0
1000
2000
3000
4000
1
2
3
4
5
6
7
8
9
96

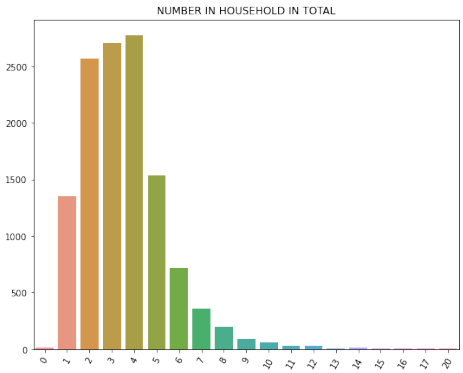

SOCIO ECONOMIC STATUS
0
1000
2000
3000
4000
1
2
3
4

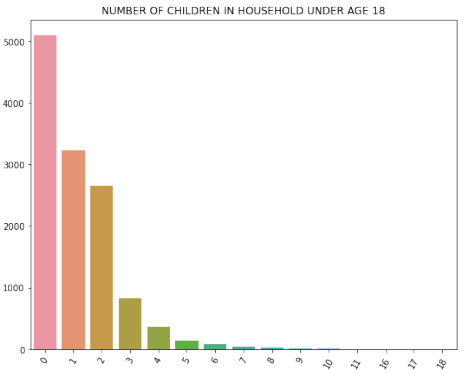

HOUSEHOLD INCOME
0
500
1000
1500
2000
2500
1
2
3
4
5

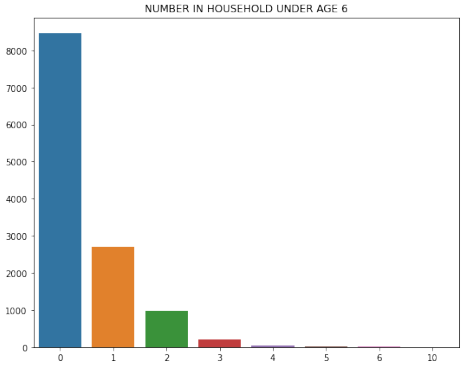

RELIGIOUS SERVICES ATTENDANCE
0
500
1000
1500
2000
2500
3000
3500
1
2
3
4
5
6

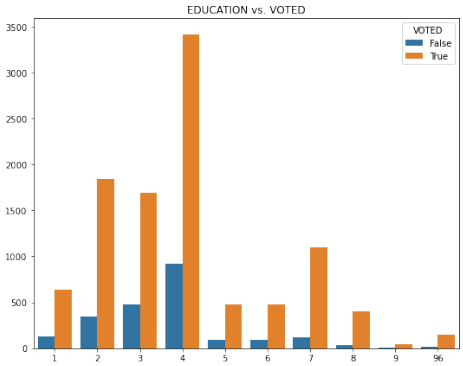

RELIGIOSITY
0
1000
2000
3000
4000
5000
1
2
3
4

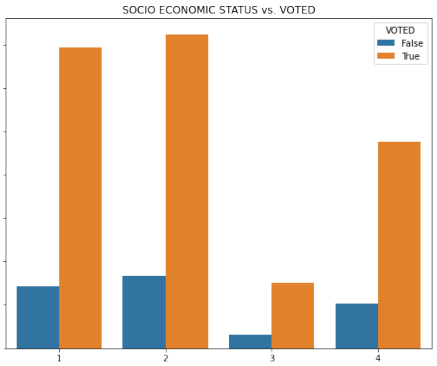

RURAL OR URBAN RESIDENCE
0
1000
2000
3000
4000
5000
6000
1
2
3
4

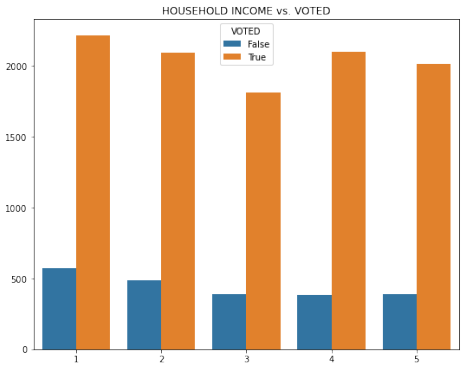

NUMBER IN HOUSEHOLD IN TOTAL
0
500
1000
1500
2000
2500
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
20

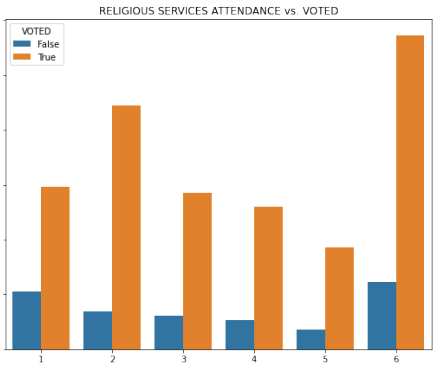

NUMBER OF CHILDREN IN HOUSEHOLD UNDER AGE 18
0
1000
2000
3000
4000
5000
0
1
2
3
4
5
6
7
8
9
10
11
16
17
18

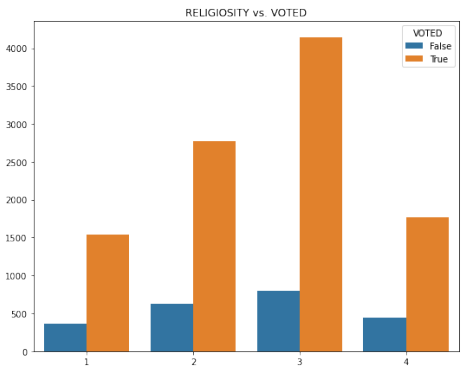

NUMBER IN HOUSEHOLD UNDER AGE 6
0
1000
2000
3000
4000
5000
6000
7000
8000
0
1
2
3
4
5
6
10

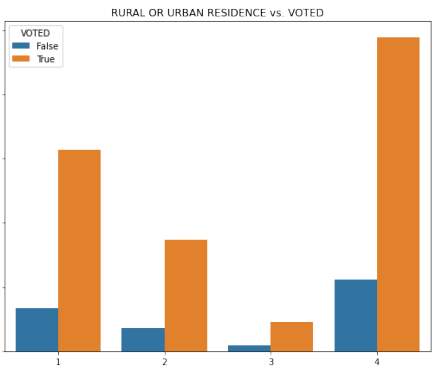

EDUCATION vs. VOTED
VOTED
False
True

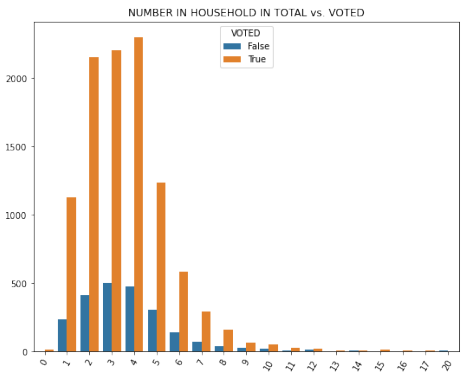

SOCIO ECONOMIC STATUS vs. VOTED
VOTED
False
True

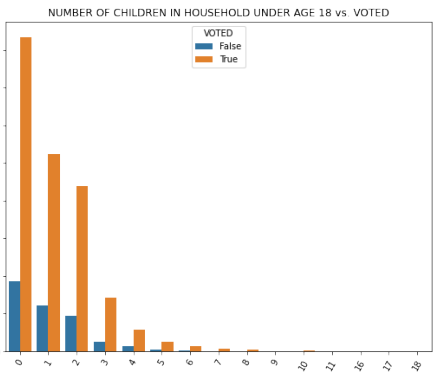

HOUSEHOLD INCOME vs. VOTED
VOTED
False
True

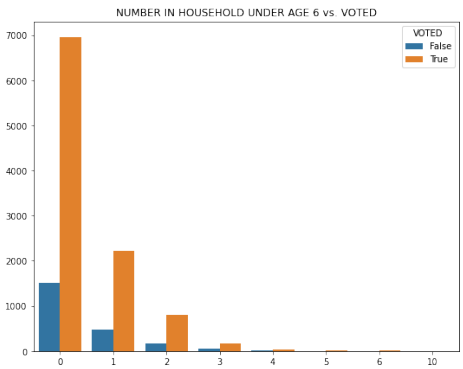

RELIGIOUS SERVICES ATTENDANCE vs. VOTED
VOTED
False
True

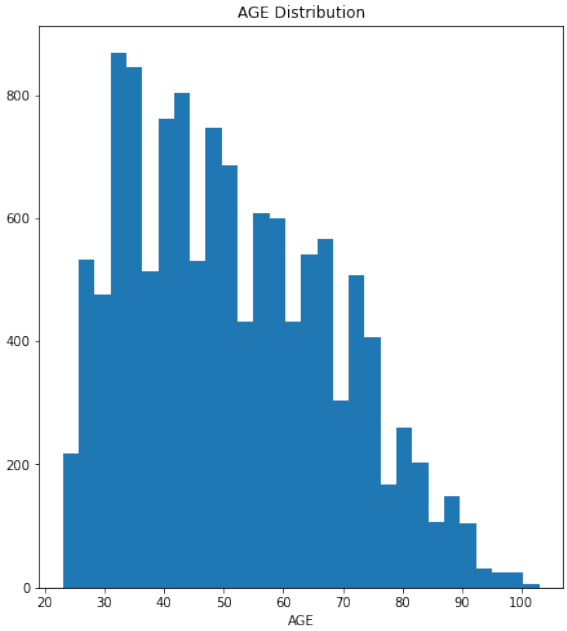

RELIGIOSITY vs. VOTED
VOTED
False
True

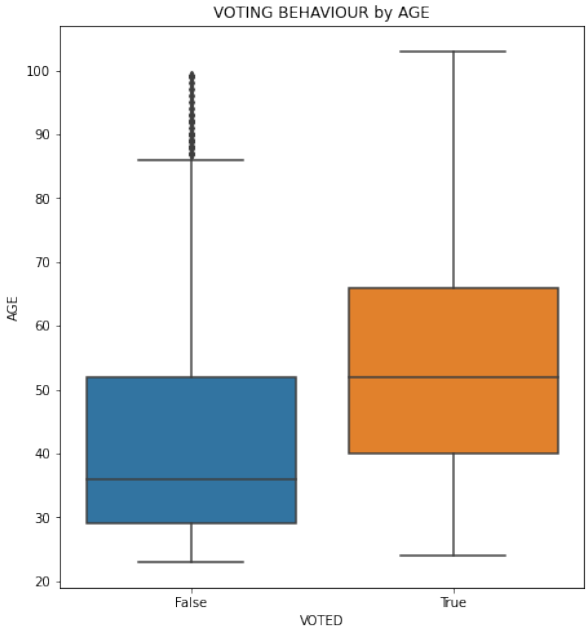

RURAL OR URBAN RESIDENCE vs. VOTED
VOTED
False
True

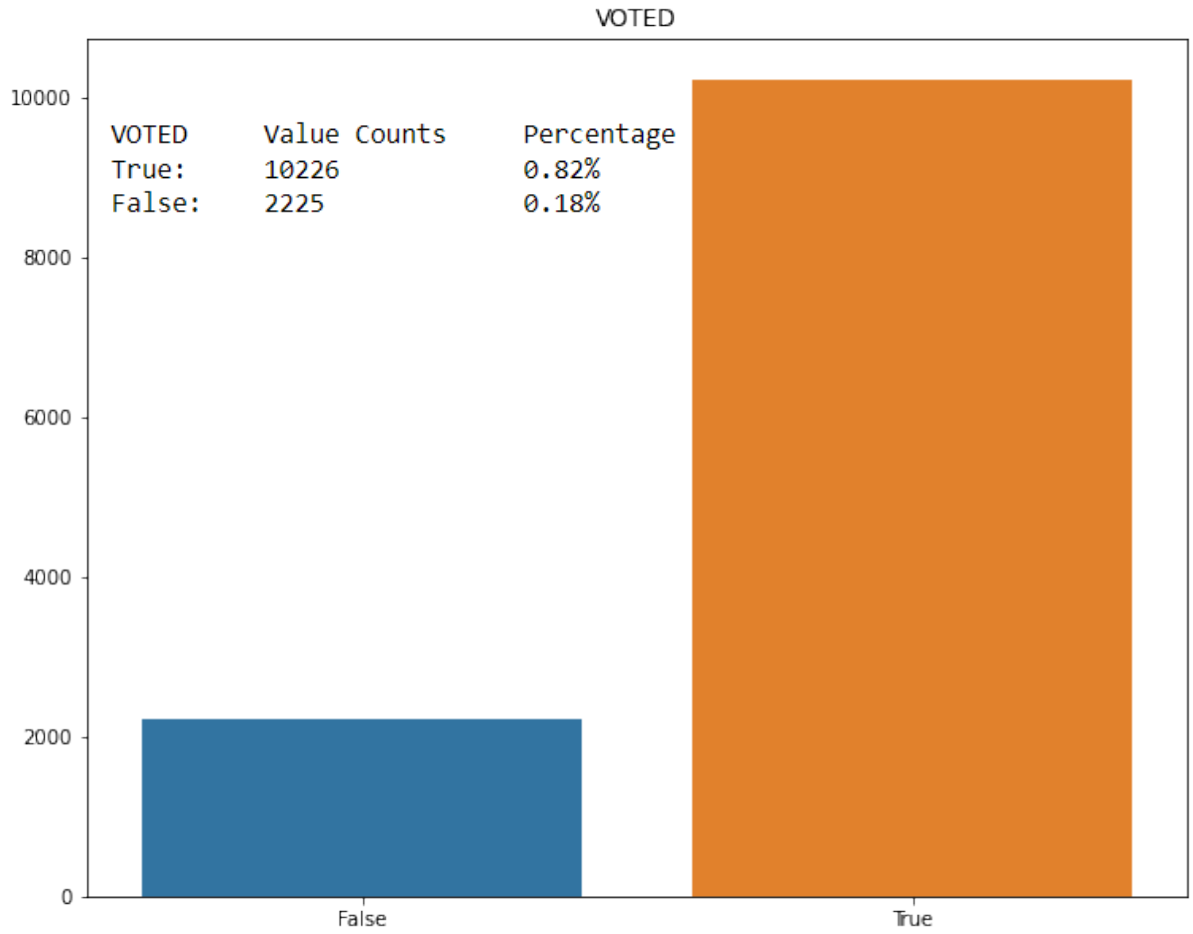

NUMBER IN HOUSEHOLD IN TOTAL vs. VOTED
VOTED
False
True

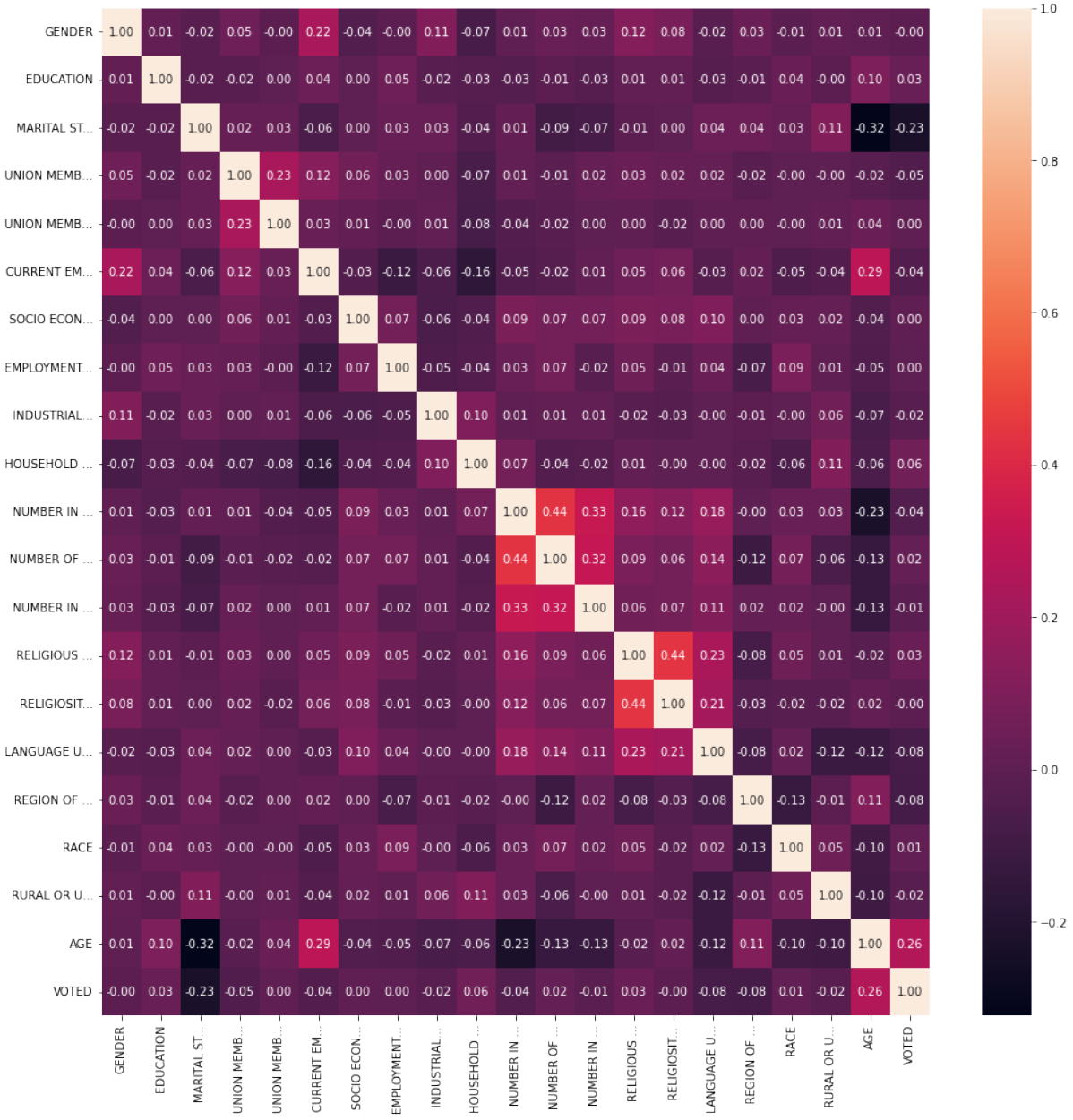

NUMBER OF CHILDREN IN HOUSEHOLD UNDER AGE 18 vs. VOTED
VOTED
False
True


NUMBER IN HOUSEHOLD UNDER AGE 6 vs. VOTED
VOTED
False
True


AGE Distribution
800
600
400
200
0
20
30
40
50
60
70
80
90
100
AGE


VOTING BEHAVIOUR by AGE
100
90
80
70
60
50
40
30
20
AGE
False
True
VOTED


VOTED
VOTED Value Counts Percentage
True: 10226 0.82%
False: 2225 0.18%
10000
8000
6000
4000
2000
0
False
True


GENDER 1.00 0.01 -0.02 0.05 -0.00 0.22 -0.04 -0.00 0.11 -0.07 0.01 0.03 0.03 0.12 0.08 -0.02 0.03 -0.01 0.01 0.01 -0.00
EDUCATION 0.01 1.00 -0.02 -0.02 0.00 0.04 0.00 0.05 -0.02 -0.03 -0.03 -0.01 -0.03 0.01 0.01 -0.03 -0.01 0.04 -0.00 0.10 0.03
MARITAL ST... -0.02 -0.02 1.00 0.02 0.03 -0.06 0.00 0.03 0.03 -0.04 0.01 -0.09 -0.07 -0.01 0.00 0.04 0.04 0.03 0.11 -0.32 -0.23
UNION MEMB... 0.05 -0.02 0.02 1.00 0.23 0.12 0.06 0.03 0.00 -0.07 0.01 -0.01 0.02 0.03 0.02 0.02 -0.02 -0.00 -0.00 -0.02 -0.05
UNION MEMB... -0.00 0.00 0.03 0.23 1.00 0.03 0.01 -0.00 0.01 -0.08 -0.04 -0.02 0.00 0.00 -0.02 0.00 0.00 -0.00 0.01 0.04 0.00
CURRENT EM... 0.22 0.04 -0.06 0.12 0.03 1.00 -0.03 -0.12 -0.06 -0.16 -0.05 -0.02 0.01 0.05 0.06 -0.03 0.02 -0.05 -0.04 0.29 -0.04
SOCIO ECON... -0.04 0.00 0.00 0.06 0.01 -0.03 1.00 0.07 -0.06 -0.04 0.09 0.07 0.07 0.09 0.08 0.10 0.00 0.03 0.02 -0.04 0.00
EMPLOYMENT... -0.00 0.05 0.03 0.03 -0.00 -0.12 0.07 1.00 -0.05 -0.04 0.03 0.07 -0.02 0.05 -0.01 0.04 -0.07 0.09 0.01 -0.05 0.00
INDUSTRIAL... 0.11 -0.02 0.03 0.00 0.01 -0.06 -0.06 -0.05 1.00 0.10 0.01 0.01 0.01 -0.02 -0.03 -0.00 -0.01 -0.00 0.06 -0.07 -0.02
HOUSEHOLD ... -0.07 -0.03 -0.04 -0.07 -0.08 -0.16 -0.04 -0.04 0.10 1.00 0.07 -0.04 -0.02 0.01 -0.00 -0.00 -0.02 -0.06 0.11 -0.06 0.06
NUMBER IN ... 0.01 -0.03 0.01 0.01 -0.04 -0.05 0.09 0.03 0.01 0.07 1.00 0.44 0.33 0.16 0.12 0.18 -0.00 0.03 0.03 -0.23 -0.04
NUMBER OF ... 0.03 -0.01 -0.09 -0.01 -0.02 -0.02 0.07 0.07 0.01 -0.04 0.44 1.00 0.32 0.09 0.06 0.14 -0.12 0.07 -0.06 -0.13 0.02
NUMBER IN ... 0.03 -0.03 -0.07 0.02 0.00 0.01 0.07 -0.02 0.01 -0.02 0.33 0.32 1.00 0.06 0.07 0.11 0.02 0.02 -0.00 -0.13 -0.01
RELIGIOUS ... 0.12 0.01 -0.01 0.03 0.00 0.05 0.09 0.05 -0.02 0.01 0.16 0.09 0.06 1.00 0.44 0.23 -0.08 0.05 0.01 -0.02 0.03
RELIGIOSIT... 0.08 0.01 0.00 0.02 -0.02 0.06 0.08 -0.01 -0.03 -0.00 0.12 0.06 0.07 0.44 1.00 0.21 -0.03 -0.02 -0.02 0.02 -0.00
LANGUAGE U... -0.02 -0.03 0.04 0.02 0.00 -0.03 0.10 0.04 -0.00 -0.00 0.18 0.14 0.11 0.23 0.21 1.00 -0.08 0.02 -0.12 -0.12 -0.08
REGION OF ... 0.03 -0.01 0.04 -0.02 0.00 0.02 0.00 -0.07 -0.01 -0.02 -0.00 -0.12 0.02 -0.08 -0.03 -0.08 1.00 -0.13 -0.01 0.11 -0.08
RACE -0.01 0.04 0.03 -0.00 -0.00 -0.05 0.03 0.09 -0.00 -0.06 0.03 0.07 0.02 0.05 -0.02 0.02 -0.13 1.00 0.05 -0.10 0.01
RURAL OR U... 0.01 -0.00 0.11 -0.00 0.01 -0.04 0.02 0.01 0.06 0.11 0.03 -0.06 -0.00 0.01 -0.02 -0.12 -0.01 0.05 1.00 -0.10 -0.02
AGE 0.01 0.10 -0.32 -0.02 0.04 0.29 -0.04 -0.05 -0.07 -0.06 -0.23 -0.13 -0.13 -0.02 0.02 -0.12 0.11 -0.10 -0.10 1.00 0.26
VOTED -0.00 0.03 -0.23 -0.05 0.00 -0.04 0.00 0.00 -0.02 0.06 -0.04 0.02 -0.01 0.03 -0.00 -0.08 -0.08 0.01 -0.02 0.26 1.00
GENDER
EDUCATION
MARITAL ST...
UNION MEMB...
UNION MEMB...
CURRENT EM...
SOCIO ECON...
EMPLOYMENT...
INDUSTRIAL...
HOUSEHOLD ...
NUMBER IN ...
NUMBER OF ...
NUMBER IN ...
RELIGIOUS ...
RELIGIOSIT...
LANGUAGE U...
REGION OF ...
RACE
RURAL OR U...
AGE
VOTED
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

Feature Engineering

- I've only one-hot-encoded the categorical variables, and concatenated all the data with different variable types.
- I decided to use tree-based algorithms. Therefore, I decided not to normalize the data since tree-based algorithms do not require normalization.
- I've splitted the data into train and test sets.

ML Algorithm

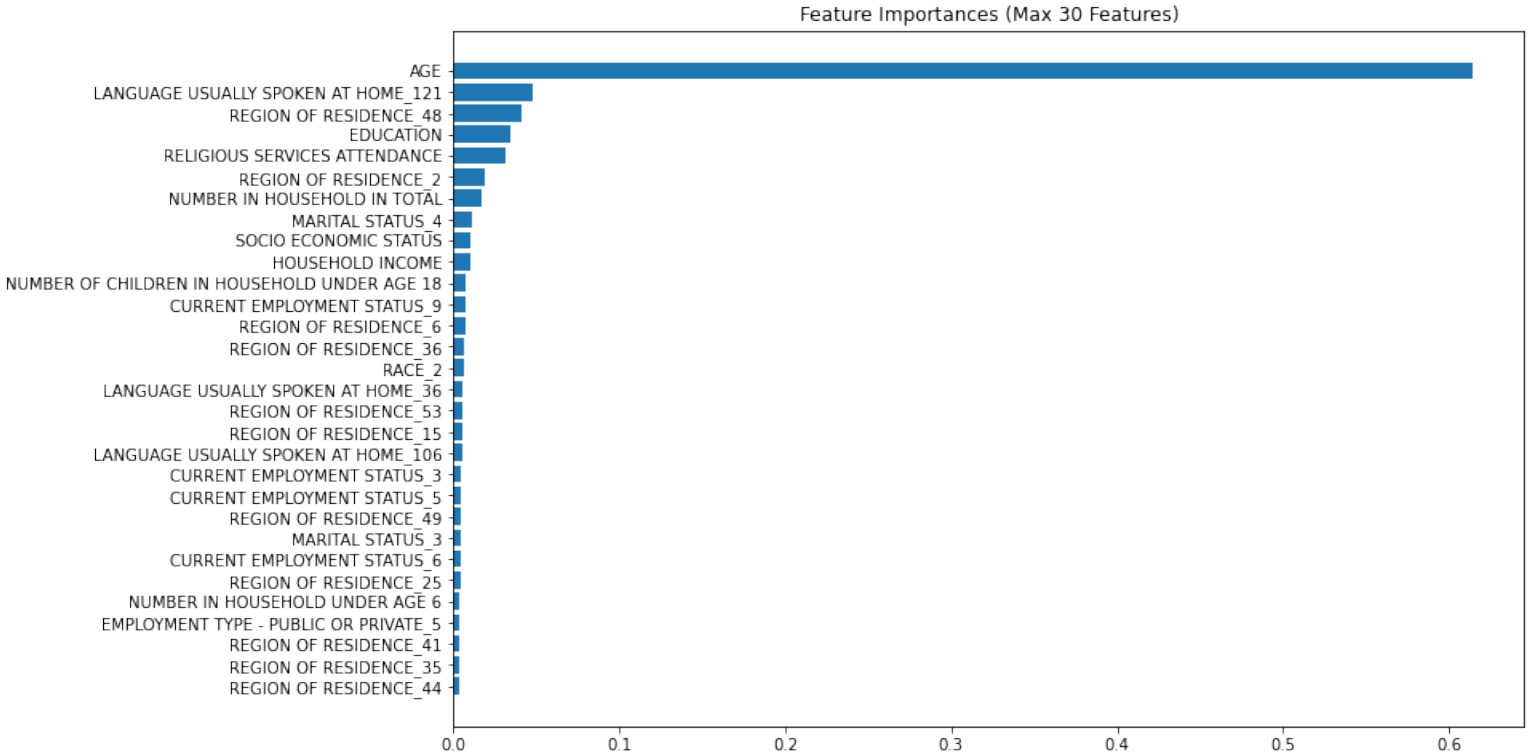
- Initially, I've tried the following four algorithms from Scikit-Learn: Decision Tree Classifier, Extra Trees Classifier, Random Forest Classifier, Gradient Boosting Classifier.

	DecisionTreeClassifier	ExtraTreesClassifier	RandomForestClassifier	GradientBoostingClassifier
Train Accuracy	1.000000	1.000000	1.000000	0.871014
Test Accuracy	0.792650	0.846435	0.859333	0.861037
Precision	0.874041	0.978171	0.981711	0.974926
Recall	0.874557	0.856184	0.865765	0.871802
F1	0.874299	0.913121	0.920100	0.920485
ROC AUC	0.641162	0.771312	0.815901	0.802253

- All models except the Gradient Boosting Classifier resulted in overfitting.

Feature Selection

- Using Gradient Boosting Classifier, I've found feature importances. The following plot shows the most important 30 features.



- Then, I have run the Gradient Boosting Classifier on the data set using [5, 10, 15, … 105] features each time. I got the best result with **75 features**. After this point, I have continued with those 75 features.

Hyperparameter Tuning

- I've used Scikit-Learn's GridSearchCV with the following parameters and 3-fold cross-validation:

```
params = {
    'learning_rate': [0.1, 0.3, 0.5, 0.7, 0.9],
    'max_depth': [1,3,5,7,9],
    'min_samples_split': [0.1, 0.3, 0.5, 0.7, 0.9],
    'min_samples_leaf': [0.1, 0.3, 0.5, 1],
}
```

- Finally, I have compared the initial results of Gradient Boosting Classifier, the results after reducing to 45 variables and the results of it with hyperparameters tuned.

	GBC_45_features	Initial_GBC	GBC_tuned
Train Accuracy	0.866579	0.867538	0.865020
Test Accuracy	0.859577	0.859577	0.859577
Precision	0.976991	0.976106	0.869451
Recall	0.869063	0.869645	0.976401
F1	0.919872	0.919805	0.919828
ROC AUC	0.803659	0.801599	0.642582