Healthcare ML Analytics in the Covid-19 ERA

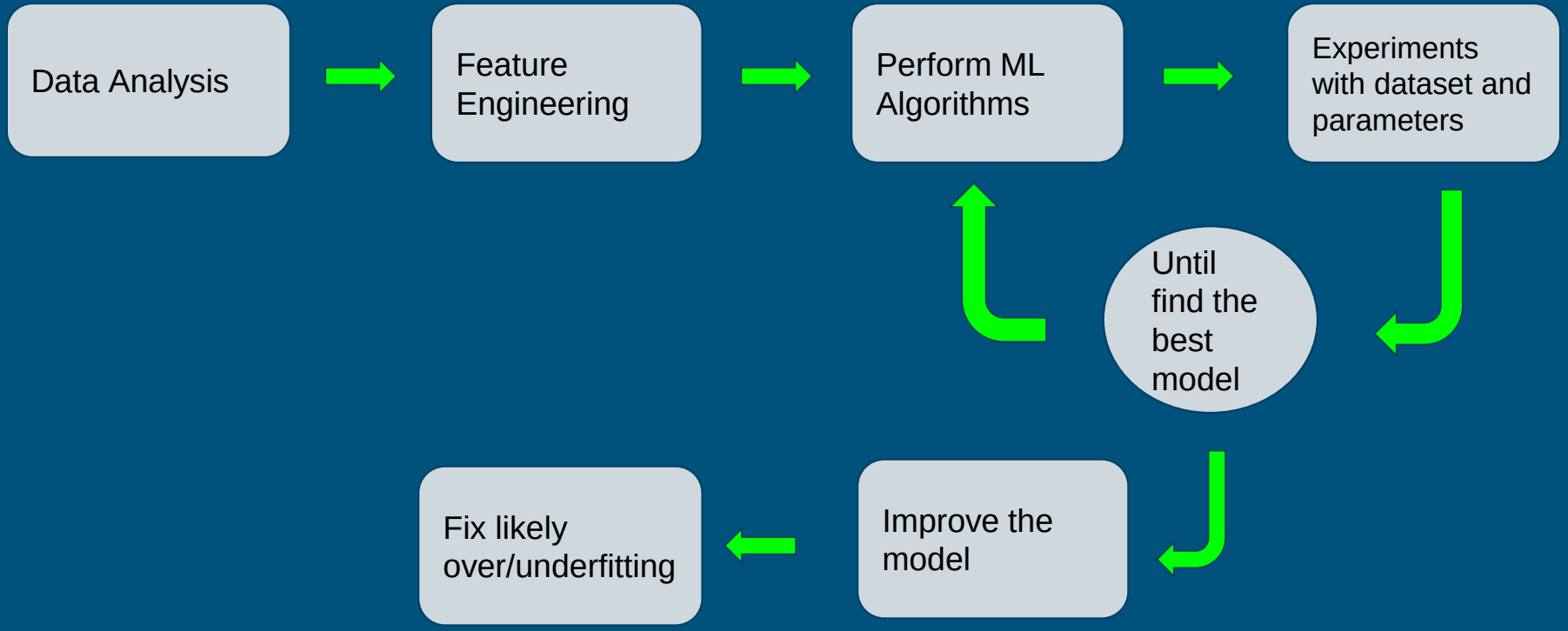
A collaboration between

Big Blue Data Academy

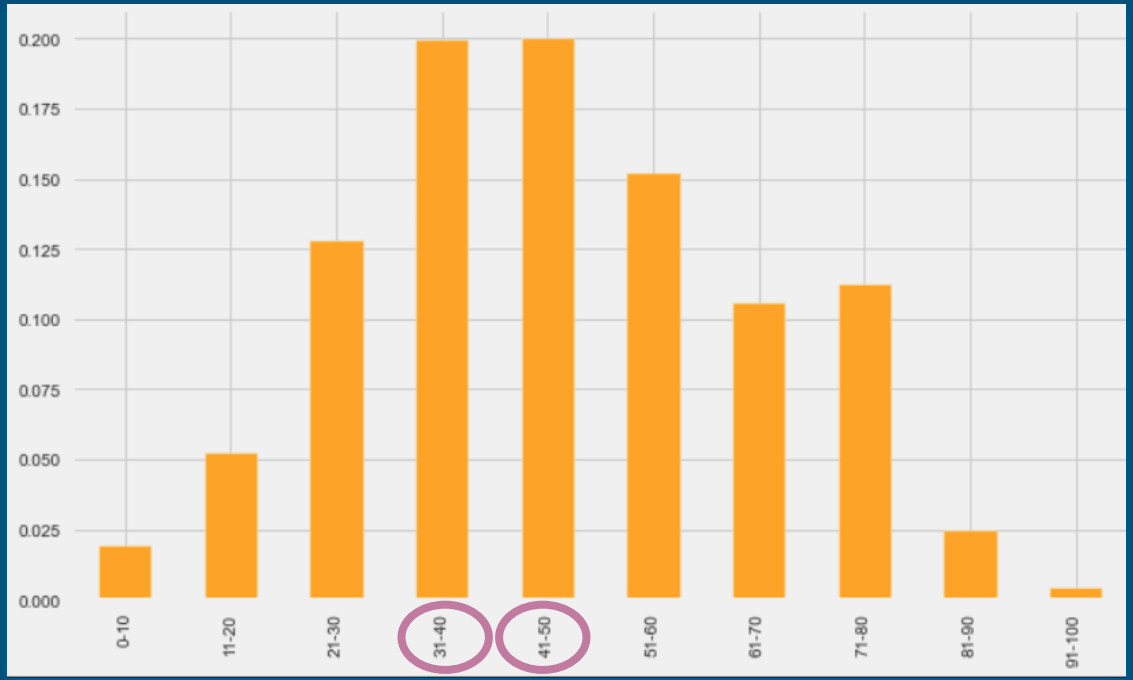
Health Human Institute

IT Department of Health Human Institute

Workflow


Data Analysis
Feature Engineering
Perform ML Algorithms
Experiments with dataset and parameters
Until find the best model
Improve the model
Fix likely over/underfitting

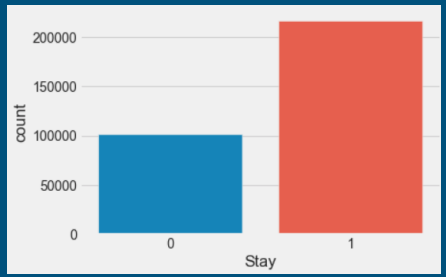
EDA - Hospital admissions / Age



Class - Target Analysis

Days stay in hospital

CLASS	TARGET
0	0 - 20
1	20 – above



Missing Data

Features	NAN values
Case_ID registered in Hospital	0
Unique code for the Hospital	0
Unique code for the type of Hospital	0
City Code of the Hospital	0
Number of Extra rooms available in the Hospital	0
Department overlooking the case	0
Code for the Ward type	0
Admission Type registered by the Hospital	113
Unique Patient Id	0
City Code for the patient	4532
Admission Type registered by the Hospital	0
Severity of the illness recorded at the time of..	0
Number of Visitors with the patient	0
Age of the patient	0
Deposit at the Admission Time	0
Stay Days by the patient	0

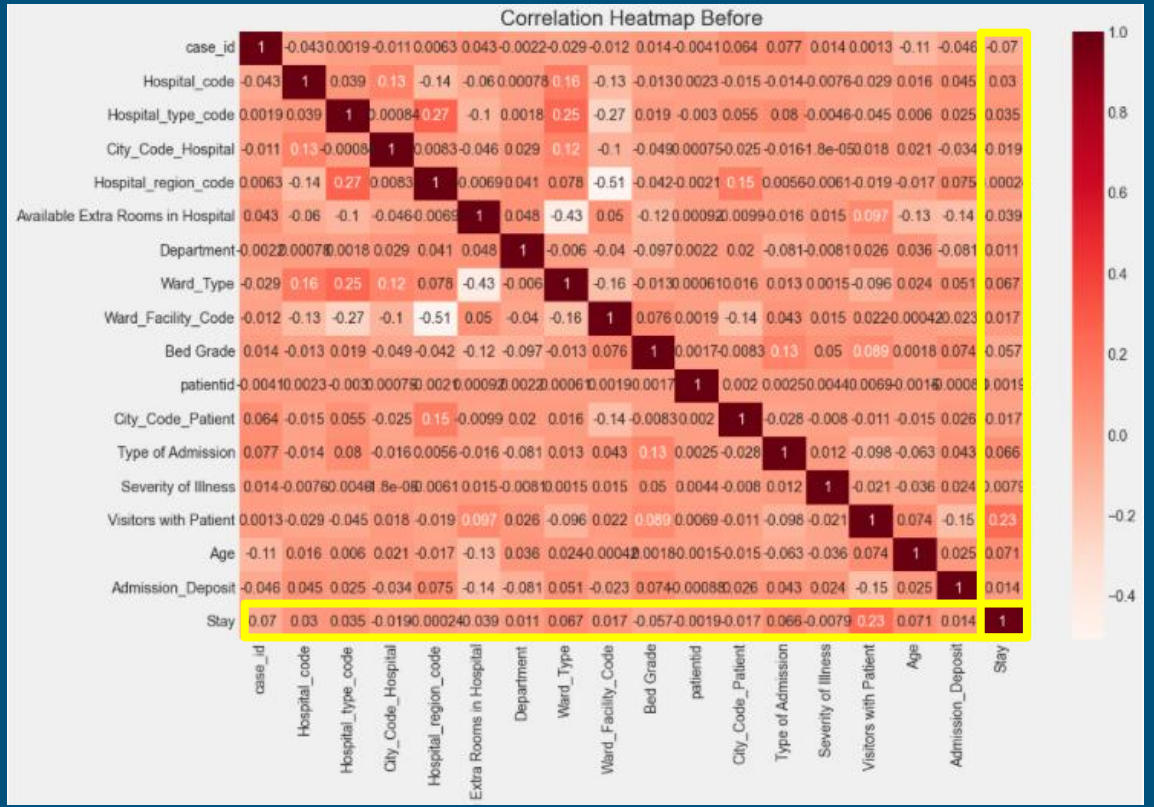
```
fillna(statistics.mode(inplace=True)
```

The Features

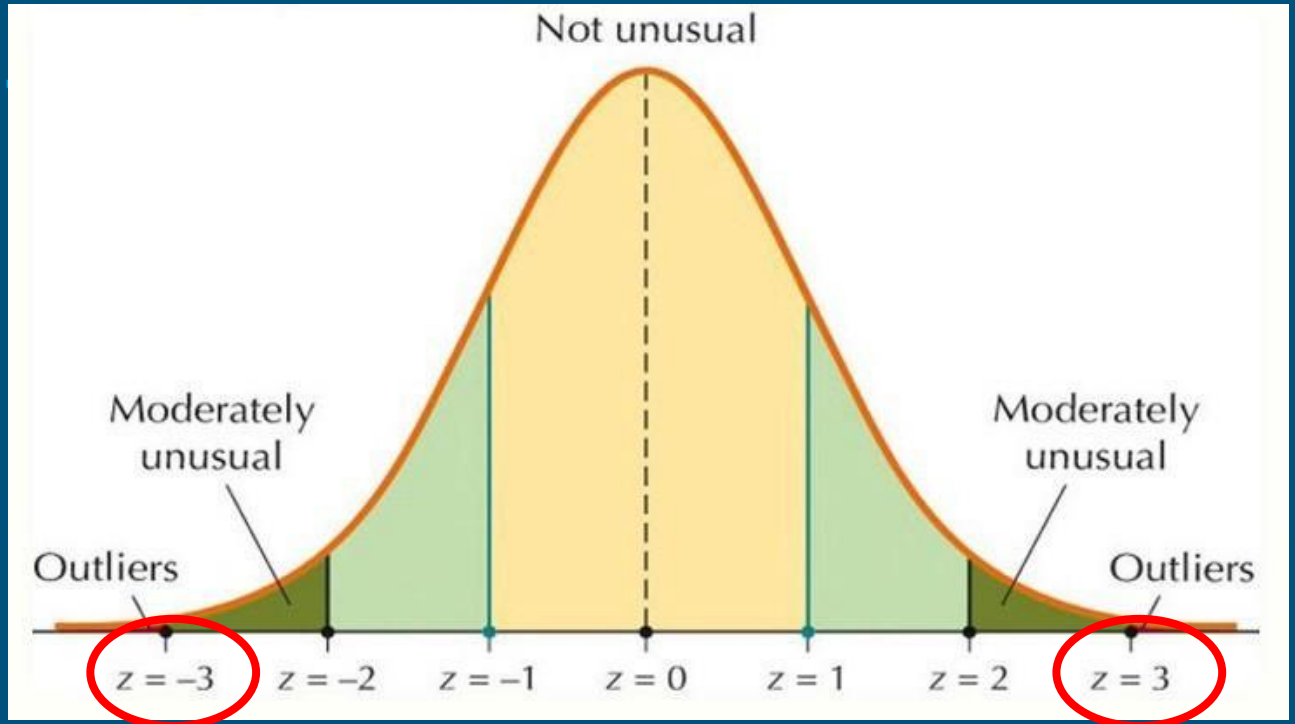
Index	Features
0	Case_ID registered in Hospital
1	Unique code for the Hospital
2	Unique code for the type of Hospital
3	City Code of the Hospital
4	Region Code of the Hospital
5	Number of Extra rooms available in the Hospital
6	Department overlooking the case
7	Code for the Ward type
8	Code for the Ward Facility
9	Condition of Bed in the Ward
10	Unique Patient Id
11	City Code for the patient
12	Admission Type registered by the Hospital
13	Severity of the illness recorded at the time o...
14	Number of Visitors with the patient
15	Age of the patient
16	Deposit at the Admission Time
17	Stay Days by the patient

17 columns

318.438 rows



drop Outliers



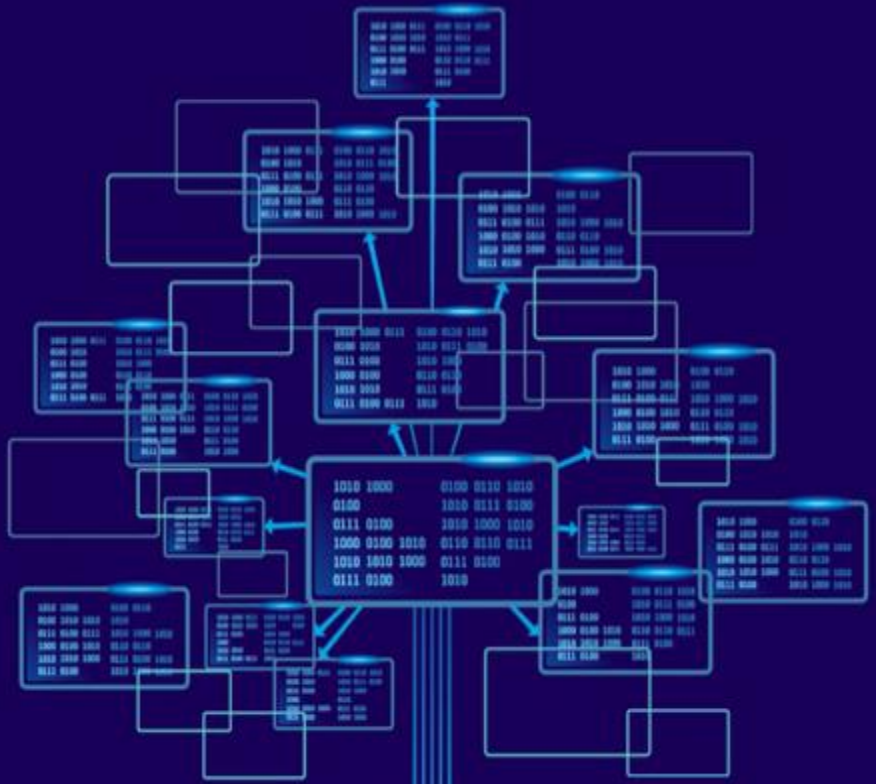
SCALING
StandardScaler( )

SHAPE before	SHAPE after
318.438	283.523
z-scores	z-scores < |3|

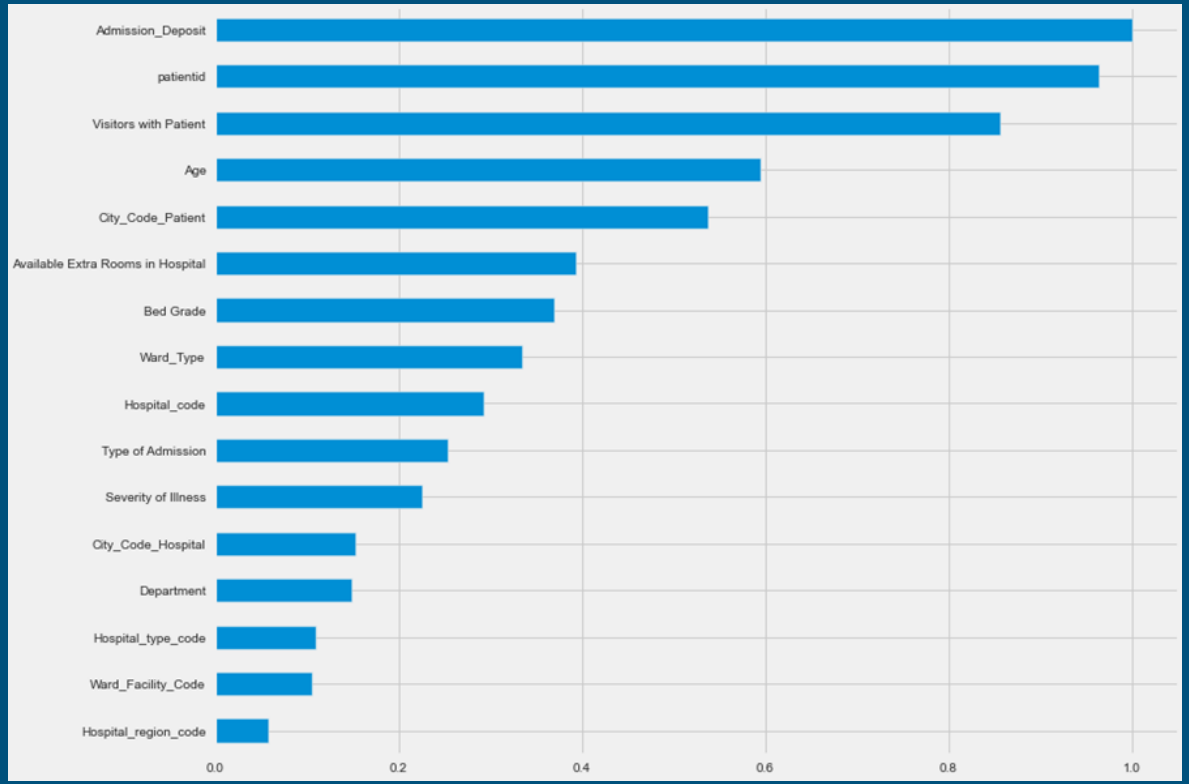
Testing Models

Algorithms	Balanced Accuracy	macro F1	Fit time
Logistic Regression	0.54	0.51	6.65
GaussianNB	0.59	0.59	0.15
KNN	0.56	0.56	14.75
Decision Tree	0.54	0.51	0.33
Random Forest	0.61	0.61	99.06
Ada Boost	0.57	0.56	110.61
Gradient Boosting	0.62	0.60	13.58
XGBoost	0.62	0.60	204.33

Random Forest Classifier

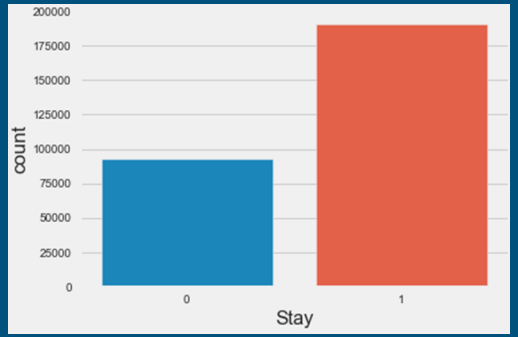


Feature Importance



Experiments with dataset and RF evaluation

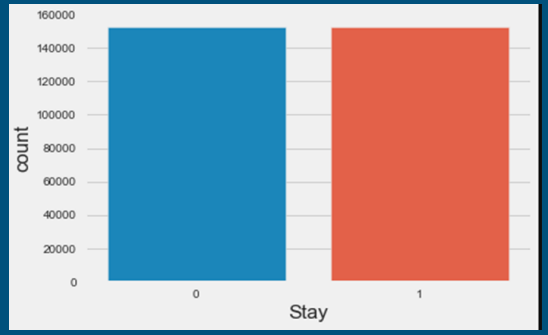
	precision	recall	f1 score	support
class 0 (0-20)	0.59	0.43	0.50	18463
class 1 (20 -)	0.76	0.85	0.80	38242
accuracy			0.72	56705
macro avg	0.67	0.64	0.65	56705
weight. avg	0.70	0.72	0.70	56705



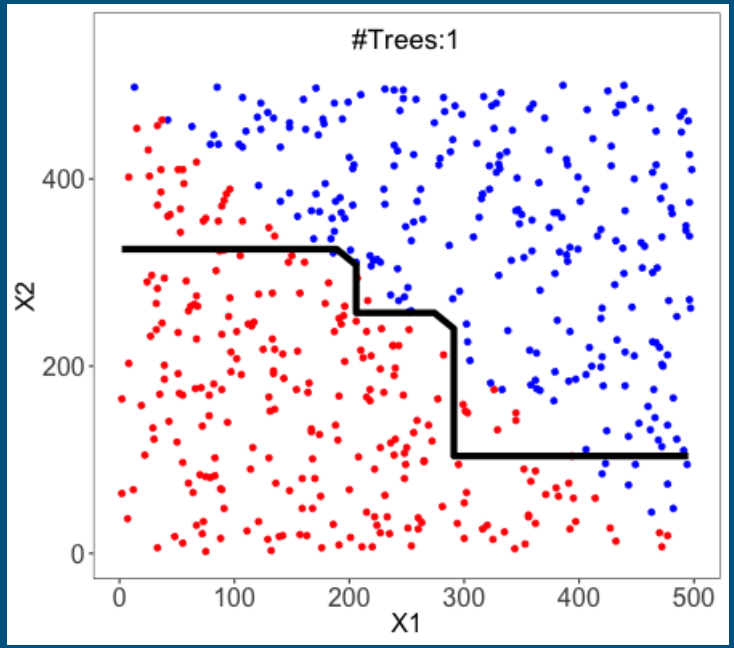
SMOTE

Synthetic Minority Over-Sampling Technique

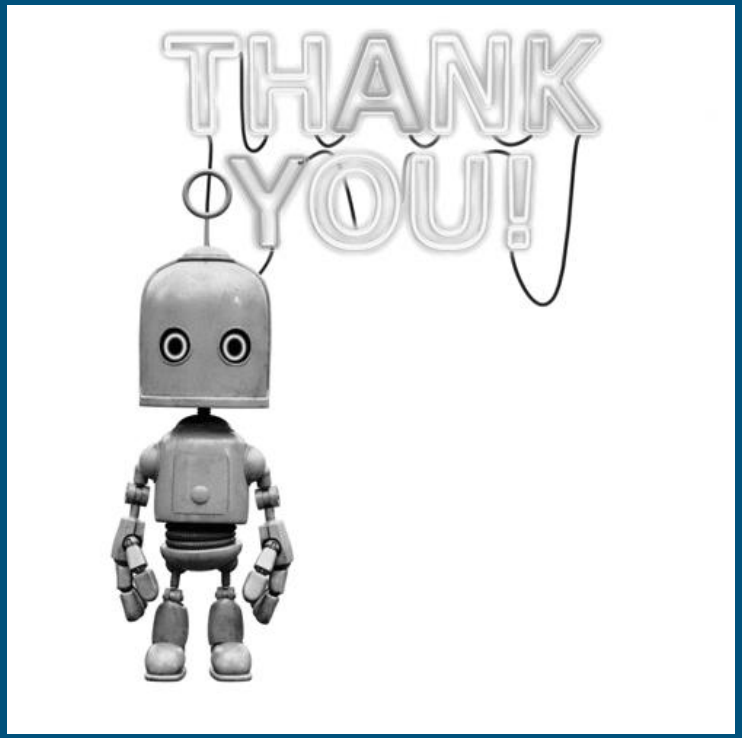
	precision	recall	f1 score	support
class 0 (0 - 20)	0.57	0.50	0.53	18463
class 1 (20 -)	0.77	0.82	0.79	38242
accuracy			0.71	56705
macro avg	0.67	0.66	0.66	56705
weight. avg	0.71	0.71	0.71	56705



Improve RF model
RANDOM_FOREST



Best Parameters	default Parameters of RANDOM_FOREST			
	• n_estimators = 200			
	precision	recall	f1 score	support
class 0 (0-20)	0.57	0.50	0.53	18463
class 1 (20-)	0.77	0.82	0.79	38242
accuracy			0.71	56705
macro avg	0.67	0.66	0.66	56705
weighted avg	0.71	0.71	0.71	56705



Questions?