TELCO churn Study case

CGI Group

Juan José CERVILLA

Exploratory Data Analysis

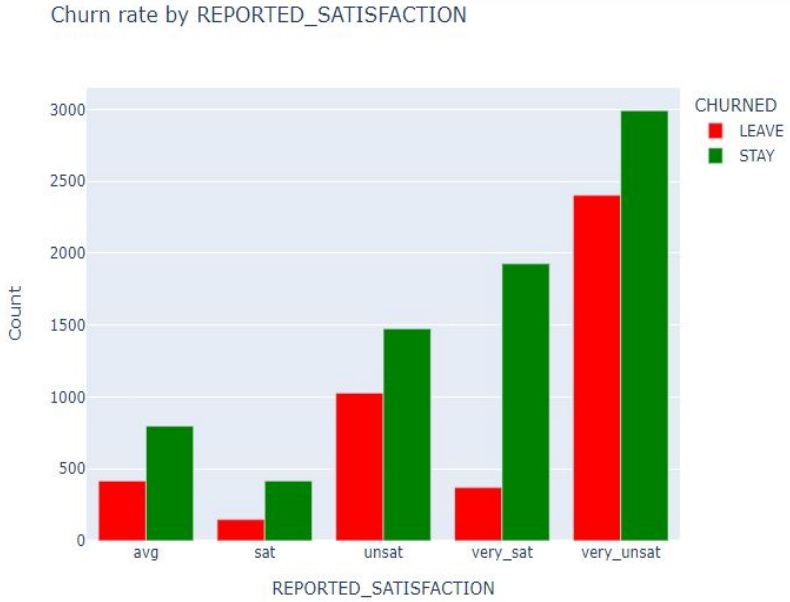
Exploratory Data Analysis

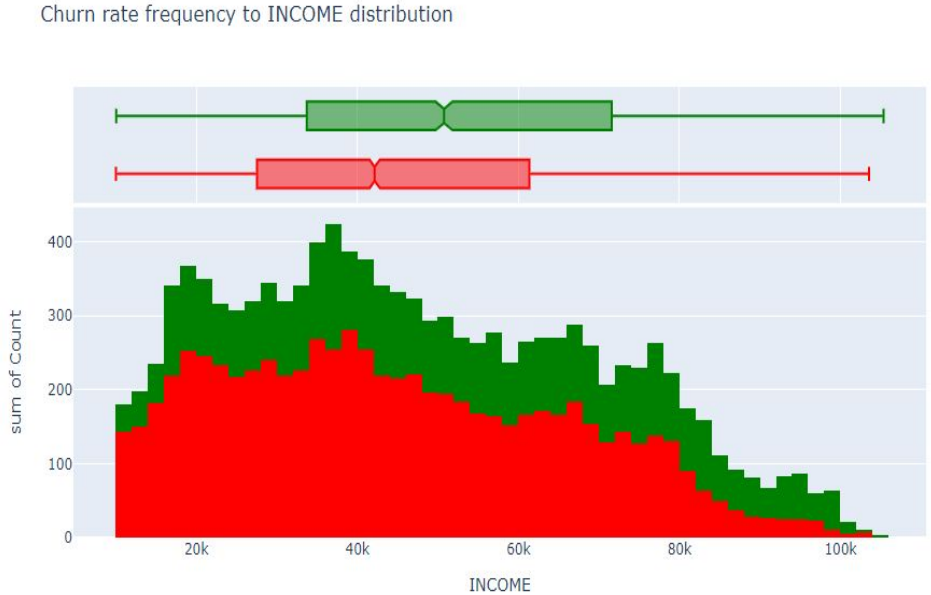
1) Understanding of problem and forecasting the main features
2) Data observation : info(), describe(), nan(), type of feature
3) Features deleted by common sense : CUSTOMER_ID, LESSTHAN600K

Variable	Explanation
CUSTOMER_ID	A technical unique identifier
COLLEGE	Is the customer college educated?
DATA	Monthly consumption of data (in Mo)
INCOME	Annual income (salary) of the client
OVERCHARGE	Average overcharge per year
LEFTOVER	Average number of lefover minutes per month
HOUSE	Estimated value of dwelling (from census tract)
LESSTHAN600k	Is the House value smaller or higher than 600k?
CHILD	Number of children
JOB_CLASS	Self reported type of job
REVENUE	Annual phone bill (excluding Overcharge)
HANDSET_PRICE	Cost of phone
OVER_15MINS_CALLS_PER_MONTH	Average number of long calls (>15 mins) per month
TIME_CLIENT	Tenure in years
AVERAGE_CALL_DURATION	Average duration of a call
REPORTED_SATISFACTION	Reported level of satisfaction
REPORTED_USAGE_LEVEL	Self reported usage level
CONSIDERING_CHANGE_OF_PLAN	Self reported consideration whether to change operator
CHURNED	Did the customer stay of leave

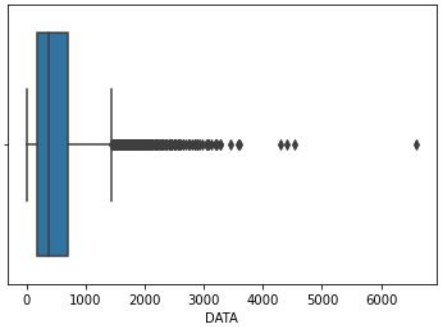
EDA: Visualization

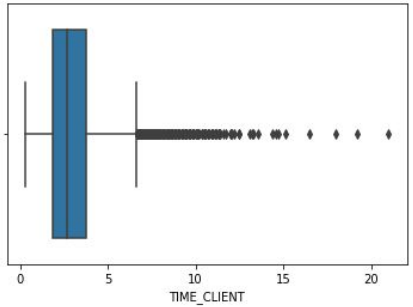
Value count of distribution of very_unsat, unsat, very_sat, avg & sat are 45.0%, 20.9%, 19.2%, 10.1% & 4.7% percentage respectively.

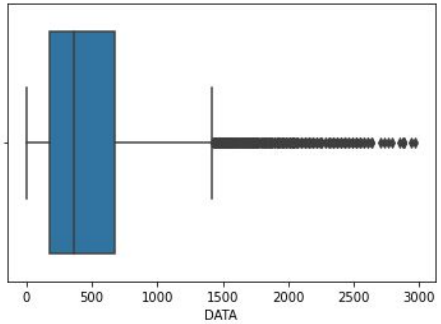


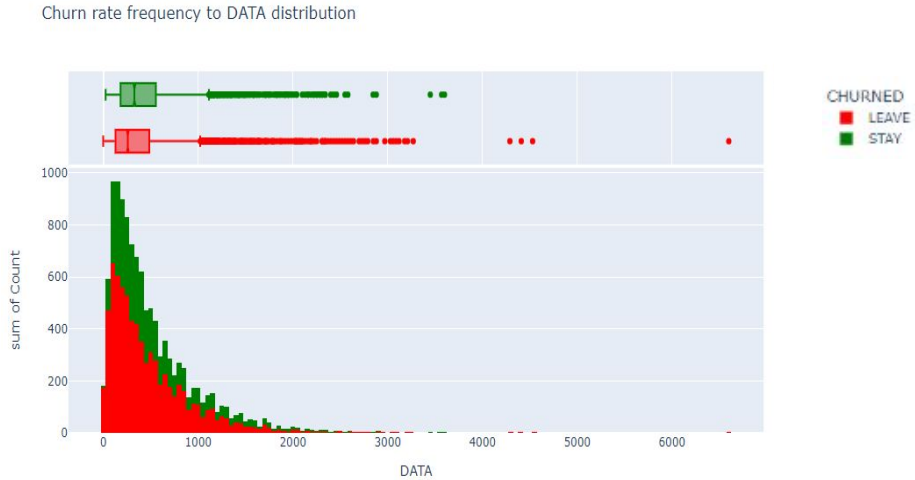


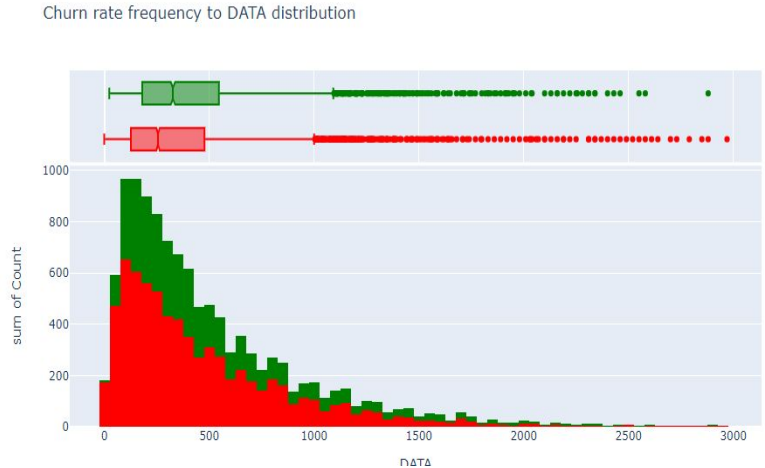
EDA : Outliers











EDA : NULL VALUES

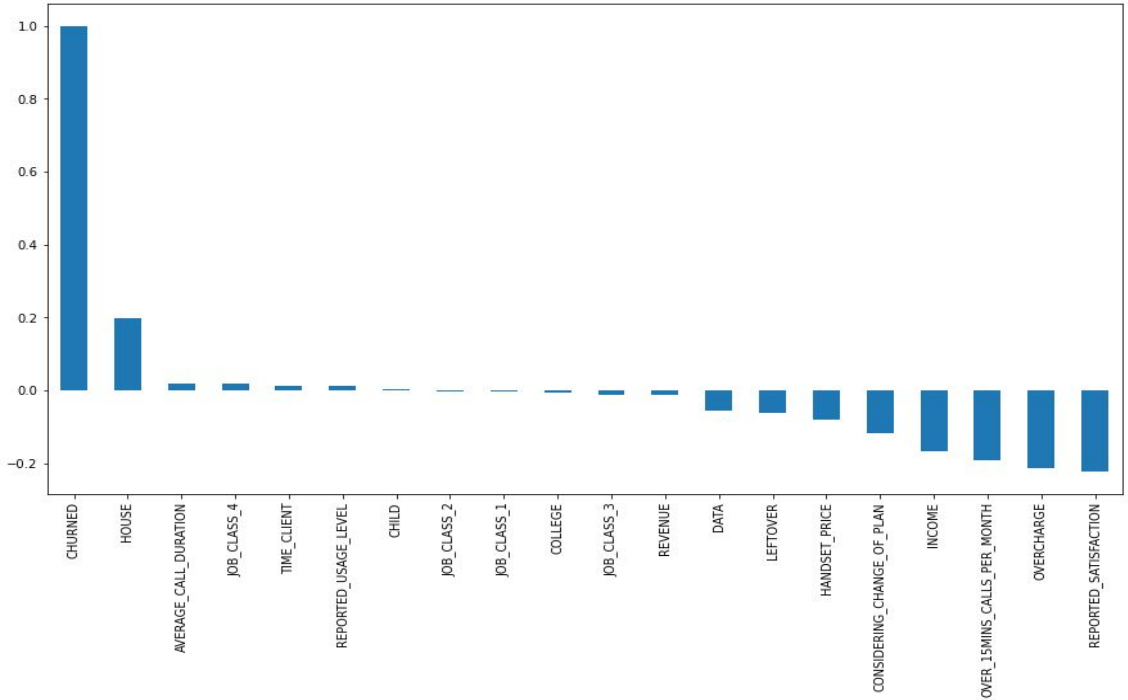
```
COLLEGE                          0
DATA                             0
INCOME                           0
OVERCHARGE                       0
LEFTOVER                         0
HOUSE                          635
CHILD                            0
JOB_CLASS                        0
REVENUE                          0
HANDSET_PRICE                    0
OVER_15MINS_CALLS_PER_MONTH      0
TIME_CLIENT                      0
AVERAGE_CALL_DURATION            0
REPORTED_SATISFACTION            0
REPORTED_USAGE_LEVEL             0
CONSIDERING_CHANGE_OF_PLAN       0
CHURNED                          0
dtype: int64
```

(11981, 19) → (11346, 17)

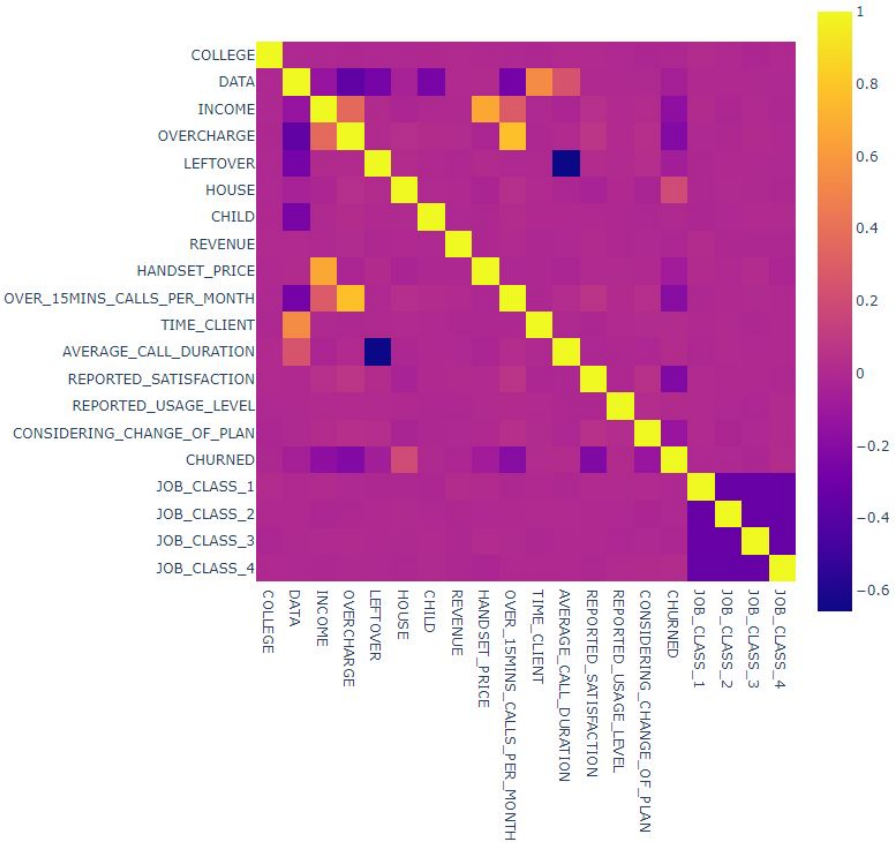
Data Cleaning

- No/Yes, LEAVE/STAY to 0/1
- Categorical features to numbers
- Dummy_feature

- Check Variance/Standard deviation

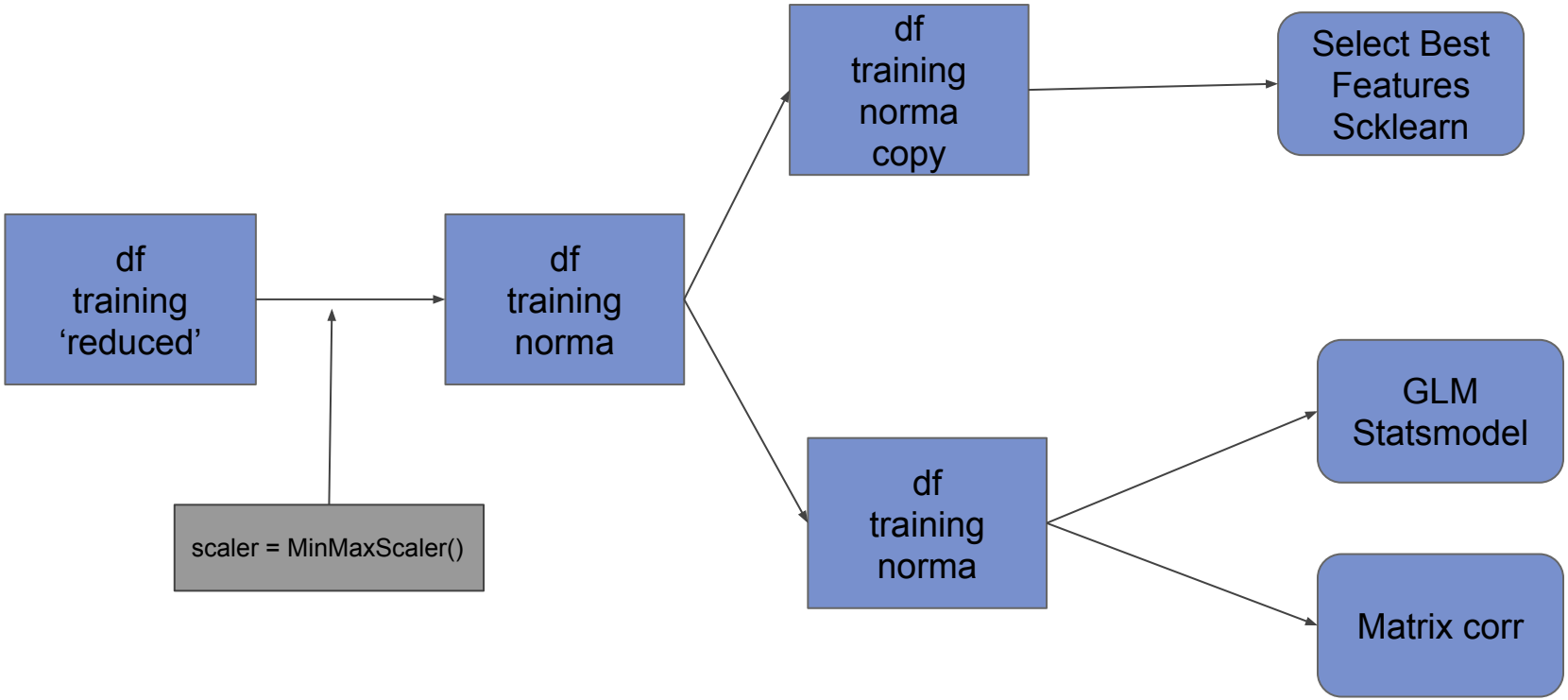


Features selection



- JOB_CLASS_1
- JOB_CLASS_2
- JOB_CLASS_3
- JOB_CLASS_4
- COLLEGE
- REPORTED_USAGE_LEVEL

Features selection



Features selection

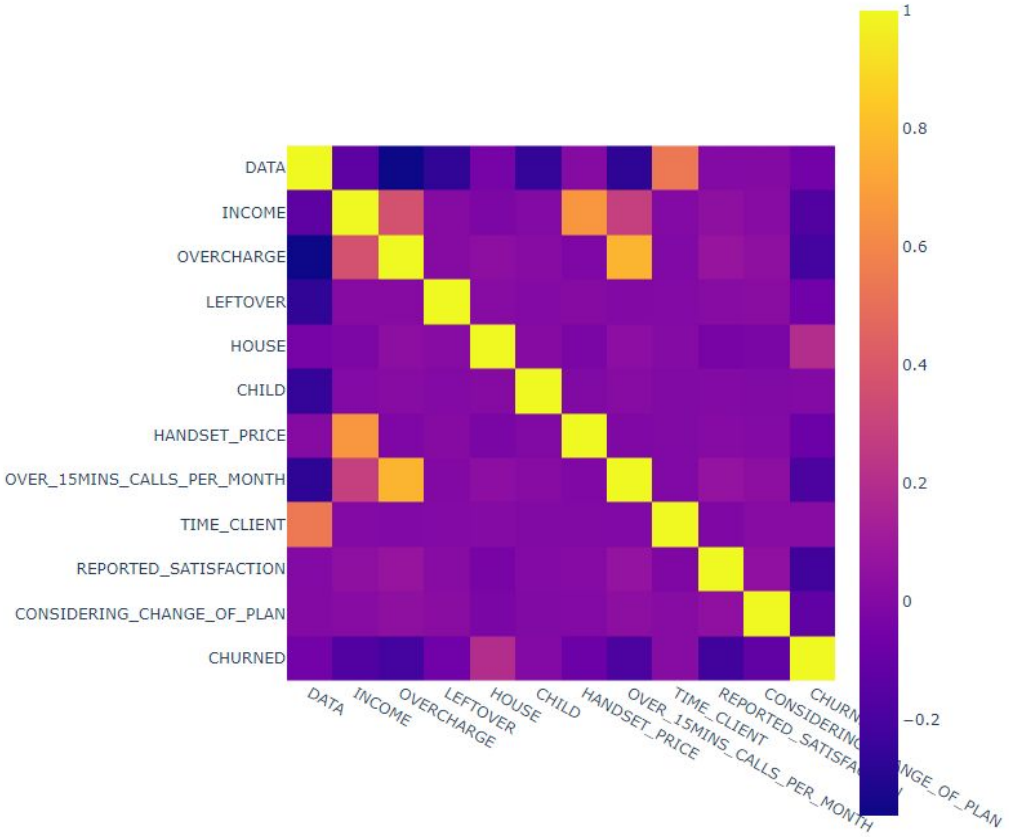
Generalized Linear Model Regression Results

Dep. Variable:	CHURNED	No. Observations:	11346
Model:	GLM	Df Residuals:	11332
Model Family:	Binomial	Df Model:	13
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-6230.7
Date:	Tue, 18 Oct 2022	Deviance:	12461.
Time:	21:09:13	Pearson chi2:	1.14e+04
No. Iterations:	5	Pseudo R-squ. (CS):	0.1931
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	3.1833	0.129	24.662	0.000	2.930	3.436
DATA	-11.5441	0.499	-23.118	0.000	-12.523	-10.565
INCOME	-0.8015	0.144	-5.575	0.000	-1.083	-0.520
OVERCHARGE	-2.1033	0.135	-15.529	0.000	-2.369	-1.838
LEFTOVER	-1.1165	0.098	-11.378	0.000	-1.309	-0.924
HOUSE	1.5970	0.076	21.035	0.000	1.448	1.746
CHILD	-0.9136	0.112	-8.190	0.000	-1.132	-0.695
REVENUE	-0.0982	0.215	-0.458	0.647	-0.519	0.322
HANDSET_PRICE	-0.2207	0.114	-1.942	0.052	-0.443	0.002
OVER_15MINS_CALLS_PER_MONTH	-0.4882	0.109	-4.469	0.000	-0.702	-0.274
TIME_CLIENT	5.8785	0.375	15.656	0.000	5.143	6.614
AVERAGE_CALL_DURATION	0.0620	0.094	0.663	0.508	-0.121	0.245
REPORTED_SATISFACTION	-1.2952	0.063	-20.711	0.000	-1.418	-1.173
CONSIDERING_CHANGE_OF_PLAN	-0.7430	0.070	-10.565	0.000	-0.881	-0.605

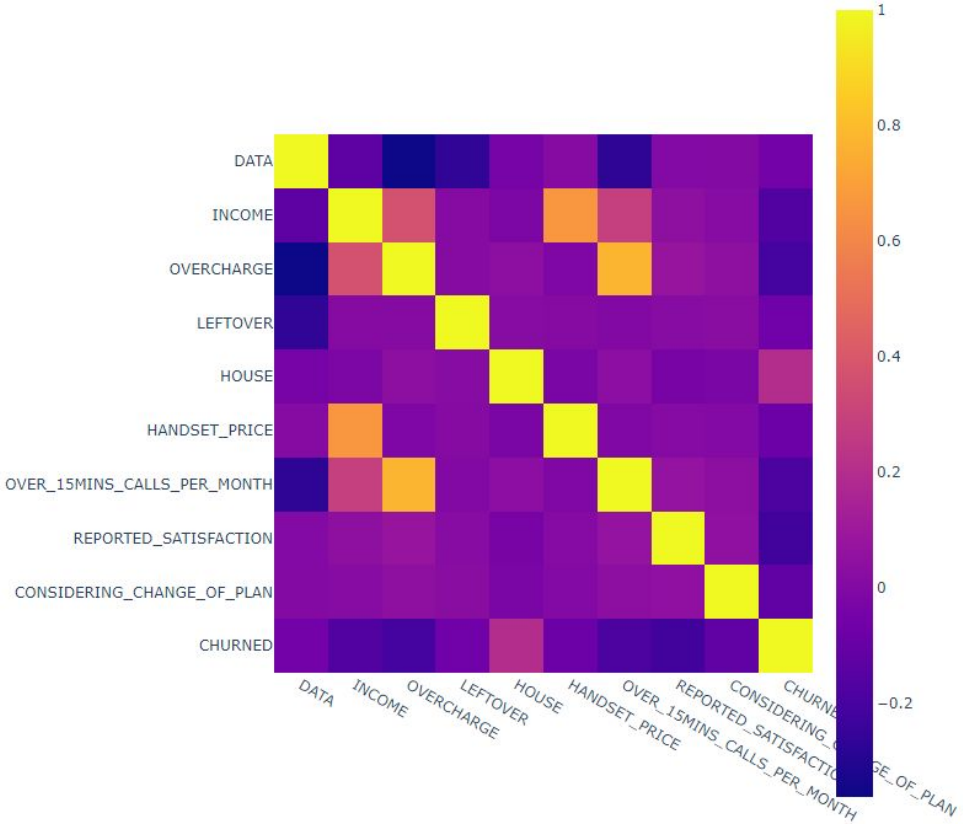
- REVENUE (!)
- AVERAGE_CALL_DURATION

Features selection



- CHILD
- TIME_CLIENT

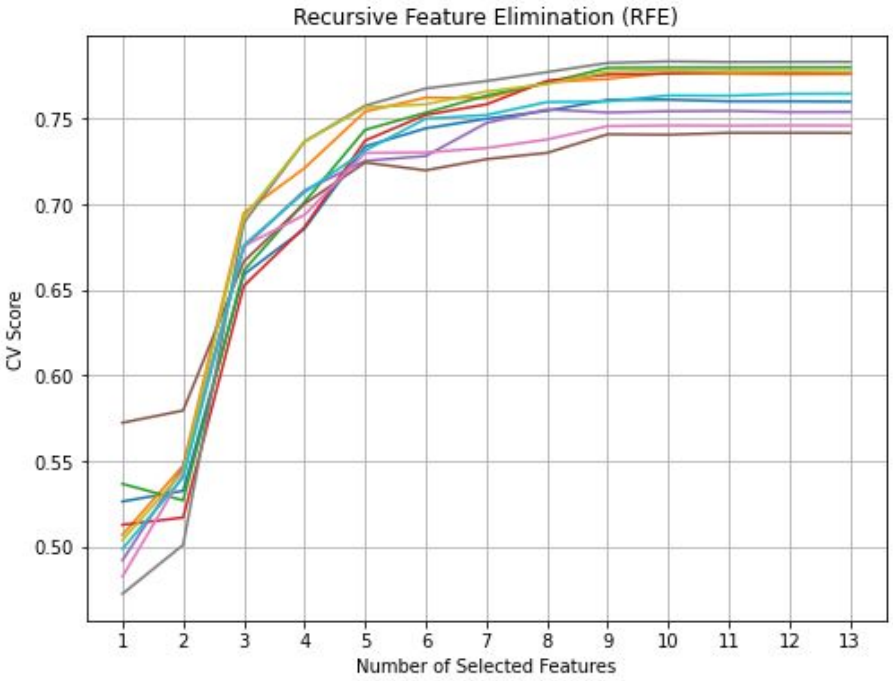
Selected features



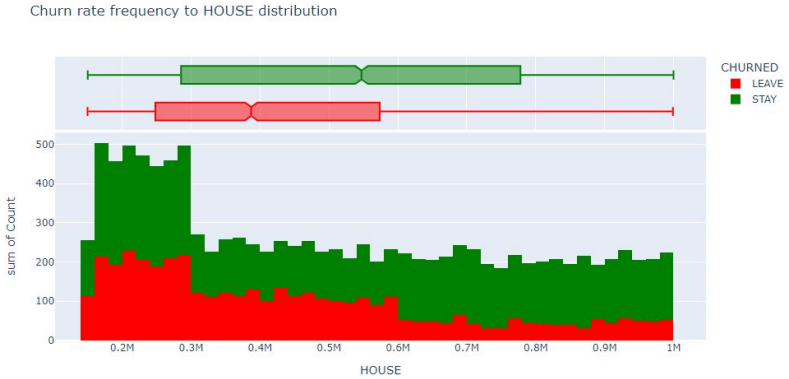
- DATA
- INCOME
- HOUSE
- OVERCHARGE
- OVER_15MINS_CALLS_PER_MONTH
- REPORTED_SATISFACTION
- CONSIDERING_CHANGE_OF_PLAN
- HANDSET_PRICE
- CHURNED

Selected Features

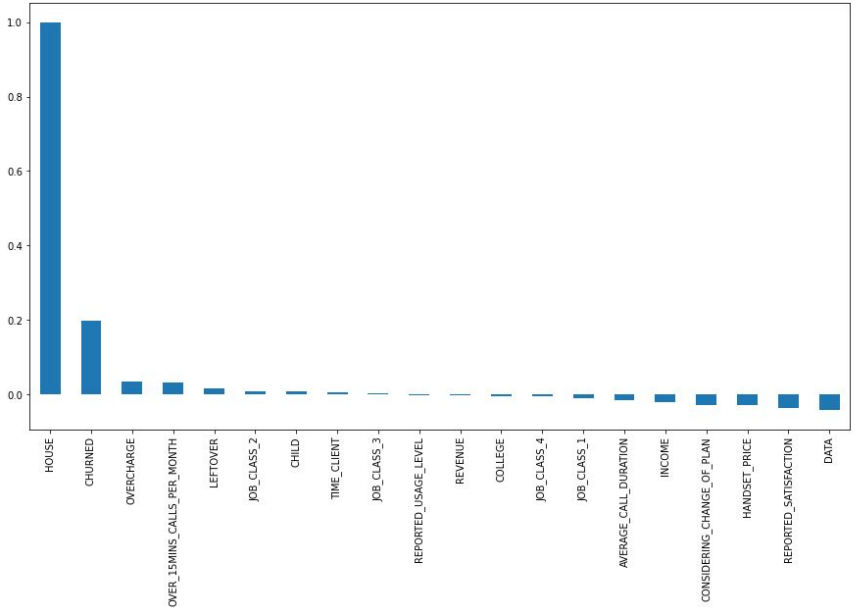
Logistic Regression



HOUSE ISSUE

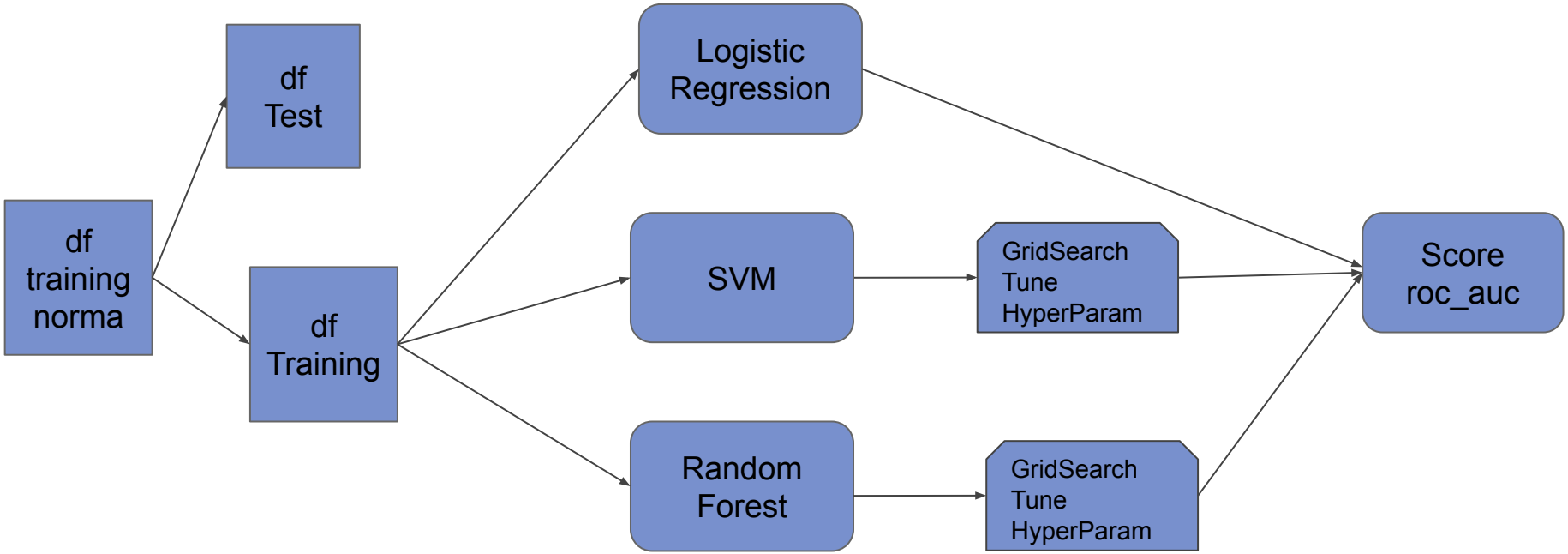

Churn rate frequency to HOUSE distribution
CHURNED
LEAVE
STAY
sum of Count
0
100
200
300
400
500
0.2M
0.3M
0.4M
0.5M
0.6M
0.7M
0.8M
0.9M
1M
HOUSE


Churn rate frequency to HOUSE distribution
CHURNED
0.0
1.0
sum of Count
0
100
200
300
400
500
0.2M
0.3M
0.4M
0.5M
0.6M
0.7M
0.8M
0.9M
1M
HOUSE


0.0
0.2
0.4
0.6
0.8
1.0
HOUSE
CHURNED
OVERCHARGE
OVER_15MINS_CALLS_PER_MONTH
LEFTOVER
JOB_CLASS_2
CHILD
TIME_CLIENT
JOB_CLASS_3
REPORTED_USAGE_LEVEL
REVENUE
COLLEGE
JOB_CLASS_4
JOB_CLASS_1
AVERAGE_CALL_DURATION
INCOME
CONSIDERING_CHANGE_OF_PLAN
HANDSET_PRICE
REPORTED_SATISFACTION
DATA

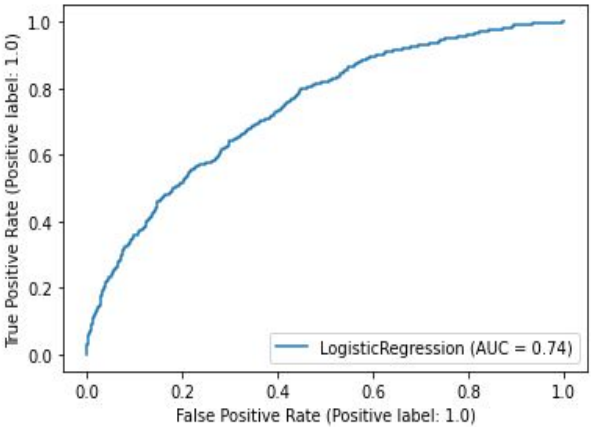
Choice of model

Choice of model

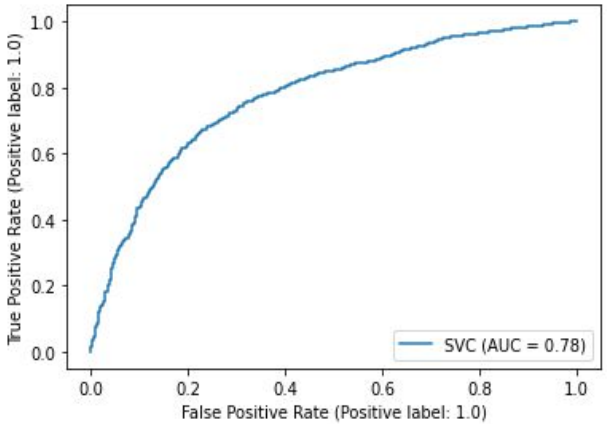


Choice of model

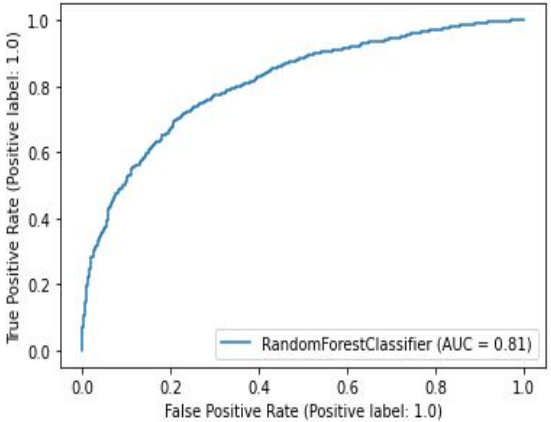
Logistic Regression



SVM



Random Forest



	LR	SVM	RF
score train	0.708	0.746	0.881
score test	0.711	0.724	0.751
roc_auc score	0.678	0.701	0.7334

Best Model = Random Forest

How could I select the CHURNED label ?

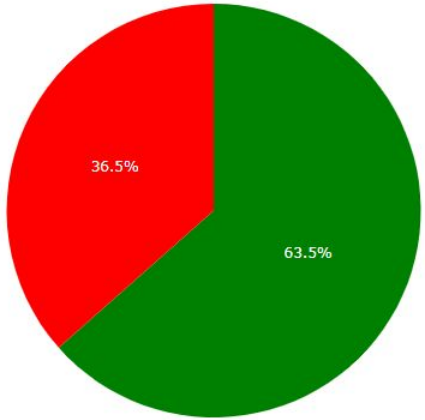
From what probability can we consider STAY or LEAVE?

Probability limit	accuracy_score
P = 0.5	0.75506
P = 0.43	0.75947
P = 0.53	0.758579

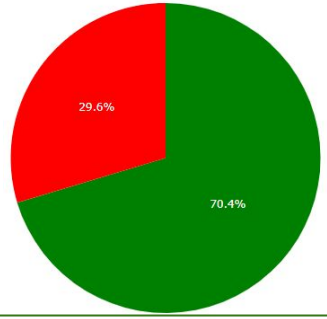
Choice of model

- CHURN PROBABILITY > THRESHOLD → 1 (STAY)
- CHURN PROBABILITY < THRESHOLD → 0 (LEAVE)

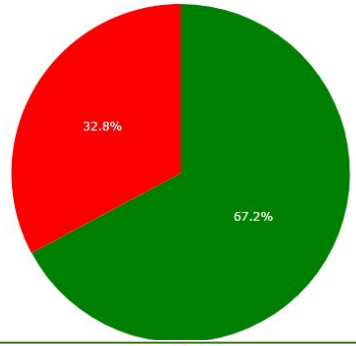
df training



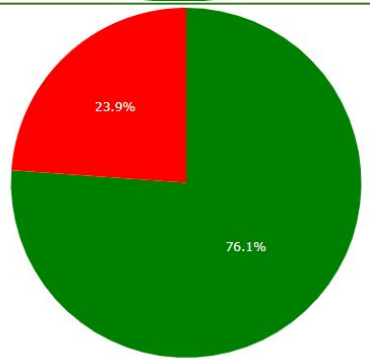
threshold = 0.5



threshold = 0.53



threshold = 0.43



Discount Strategy

Discount Strategy

INVOICE = REVENUE + OVERCHARGE

1

Expected_value = CHURN_PROBABILITY * INVOICE

2

OVERCHARGE' = OVERCHARGE * ((1- discount) / 100)

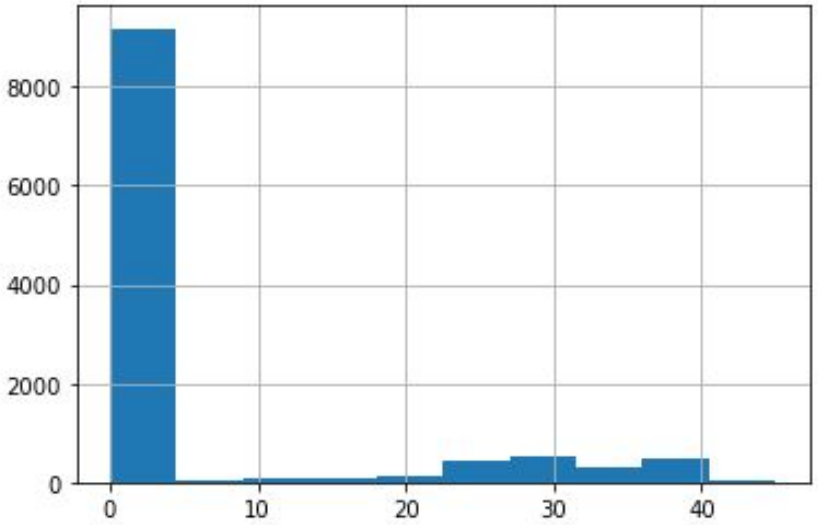
from RF_model → CHURN_PROBABILITY'

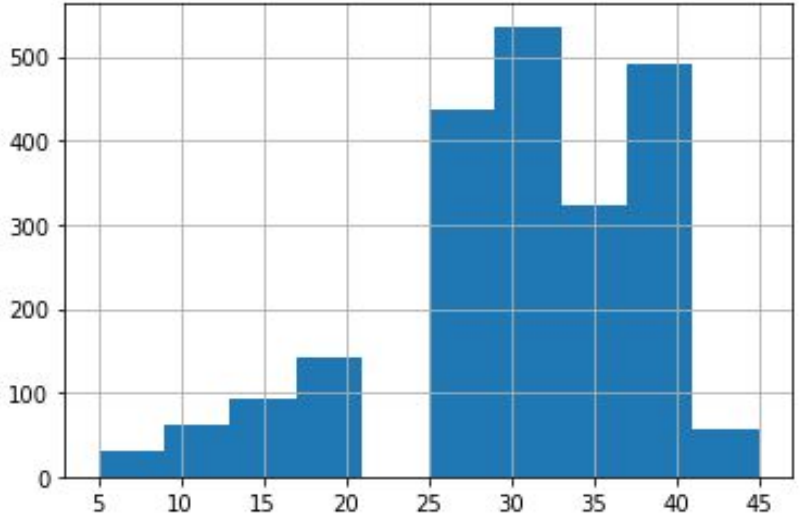
Expected_value' = CHURN_PROBABILITY' * INVOICE* ((1- discount) / 100) + 10

3

PROFIT = Expected_value' - Expected_value max BENEFIT between all discount [0% to 50%]

Discount Strategy

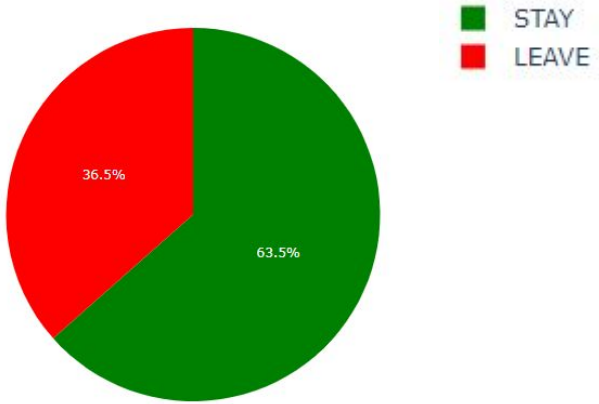


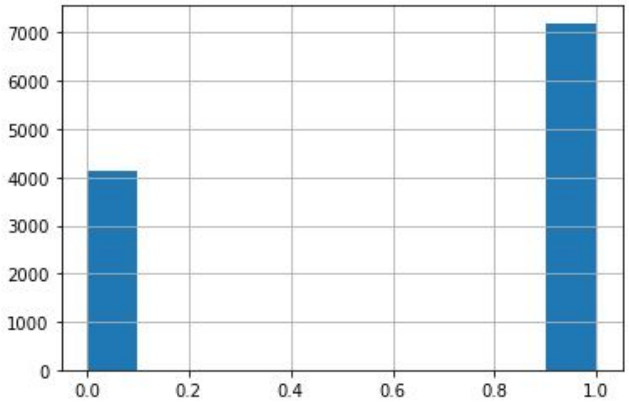


Discount Strategy

df training

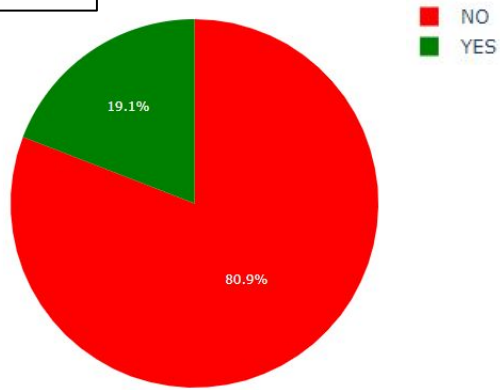
CHURNED feature





result of training

Contact Client

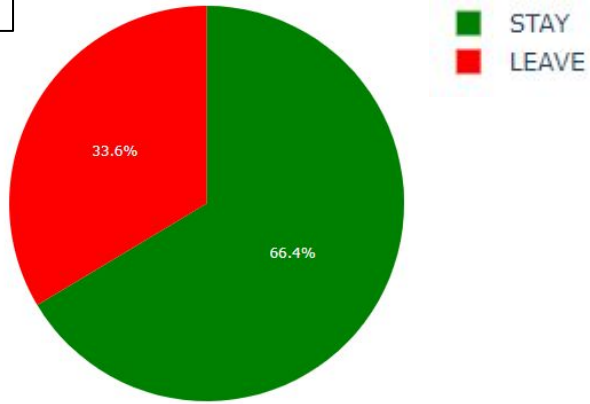


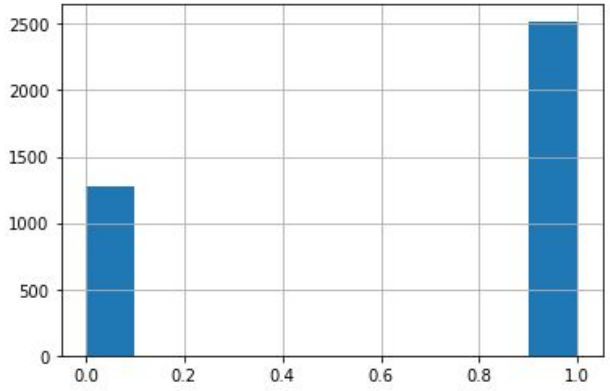
Results

Results

df validation

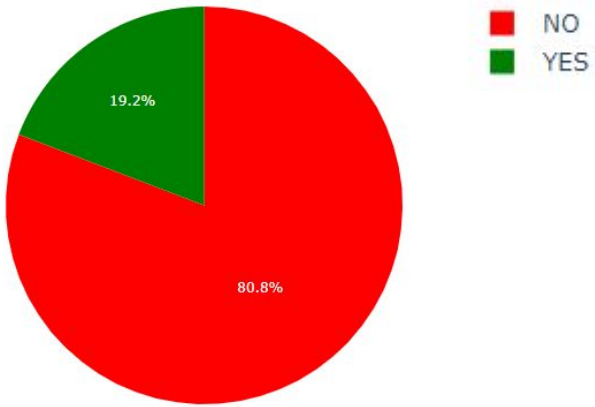
CHURN_LABEL feature

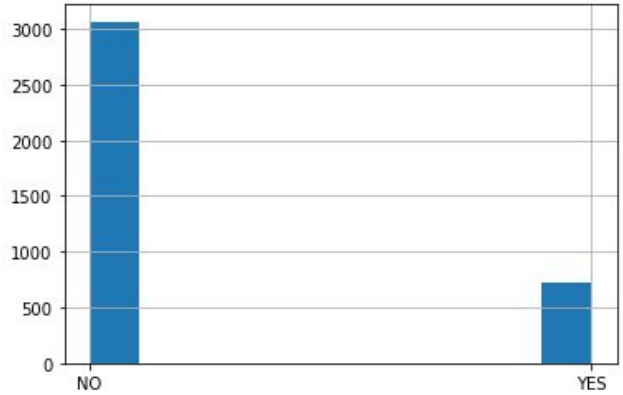




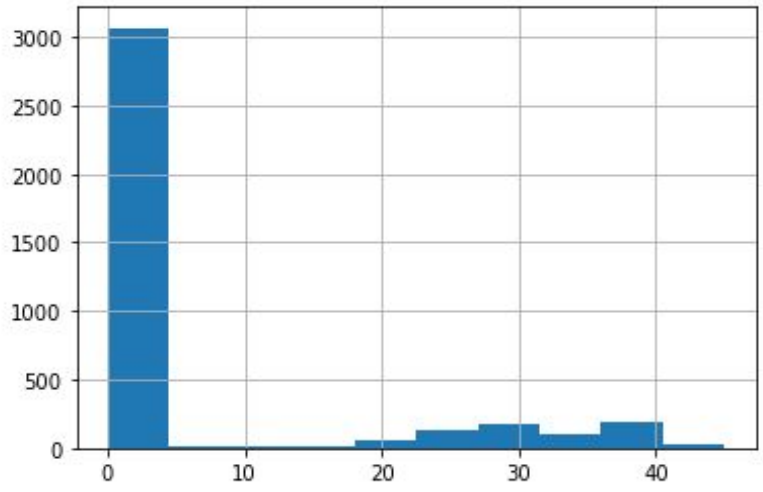
result of validation

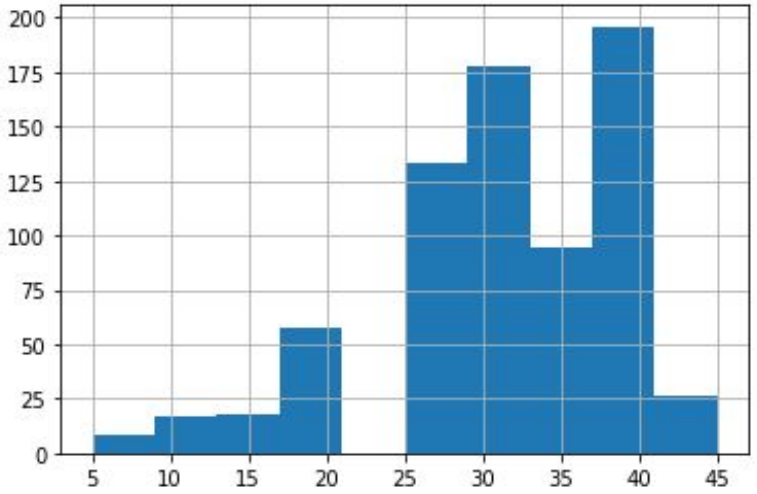
Contact Client





Results


3000
2500
2000
1500
1000
500
0
0
10
20
30
40


200
175
150
125
100
75
50
25
0
5
10
15
20
25
30
35
40
45