Problem Statement

In the Indian and the southeast Asian market, approximately 80% of revenue comes from the top 20% customers (called high-value customers). Thus, if we can reduce churn of the high-value customers, we will be able to reduce significant revenue leakage.

Predict Churn In High-Valued Customers

About Dataset

The dataset contains customer-level information for a span of four consecutive months - June, July, August and September.

The business objective is to predict the churn in the last (i.e. the ninth) month using the data (features) from the first three months.

Identify the High-Valued Customers

High-valued customers are those who have recharged more than or equal to 70th percentile of the average recharge amount in the first two months (i.e. June, July)

HVC >= 70th percentile (Average of June, July)

Data Preparation

1. Derive new features to filter high-value customers

New Feature= Data greater than or equal to 70th percentile of the average recharge amount in June and July.

2. Tag churners and remove attributes of the churn phase.

Identifying ***churned high value customers*** based on the ***last month(September)***

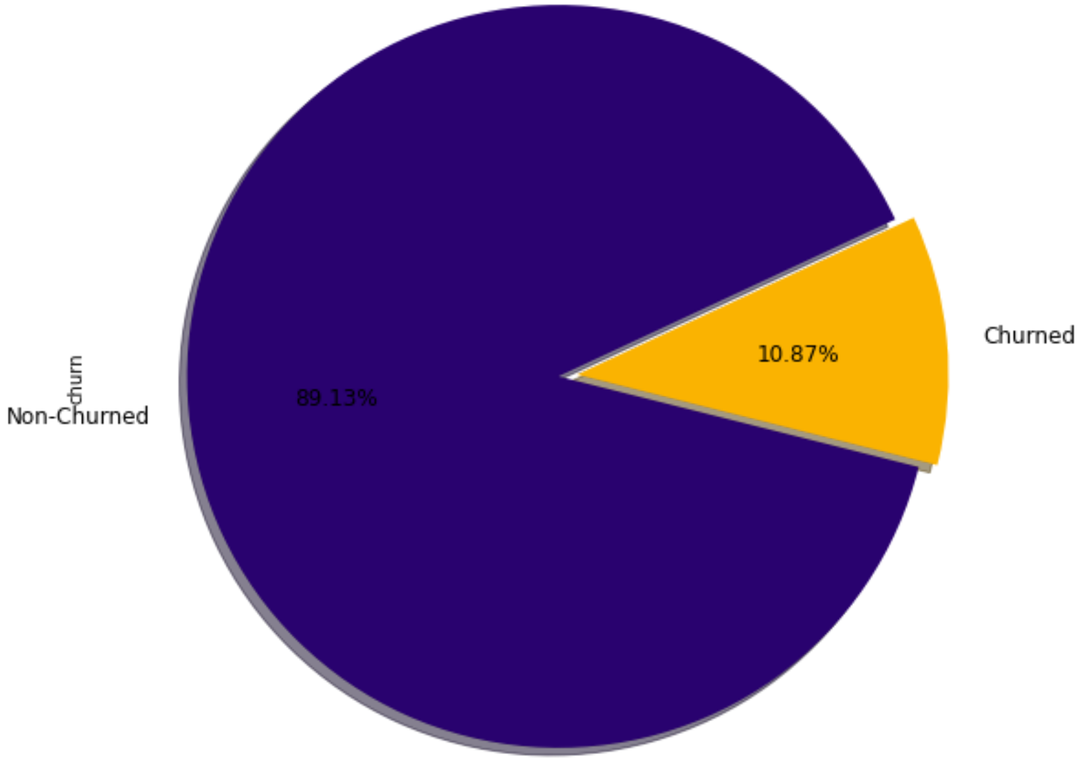
1=Churned Customers

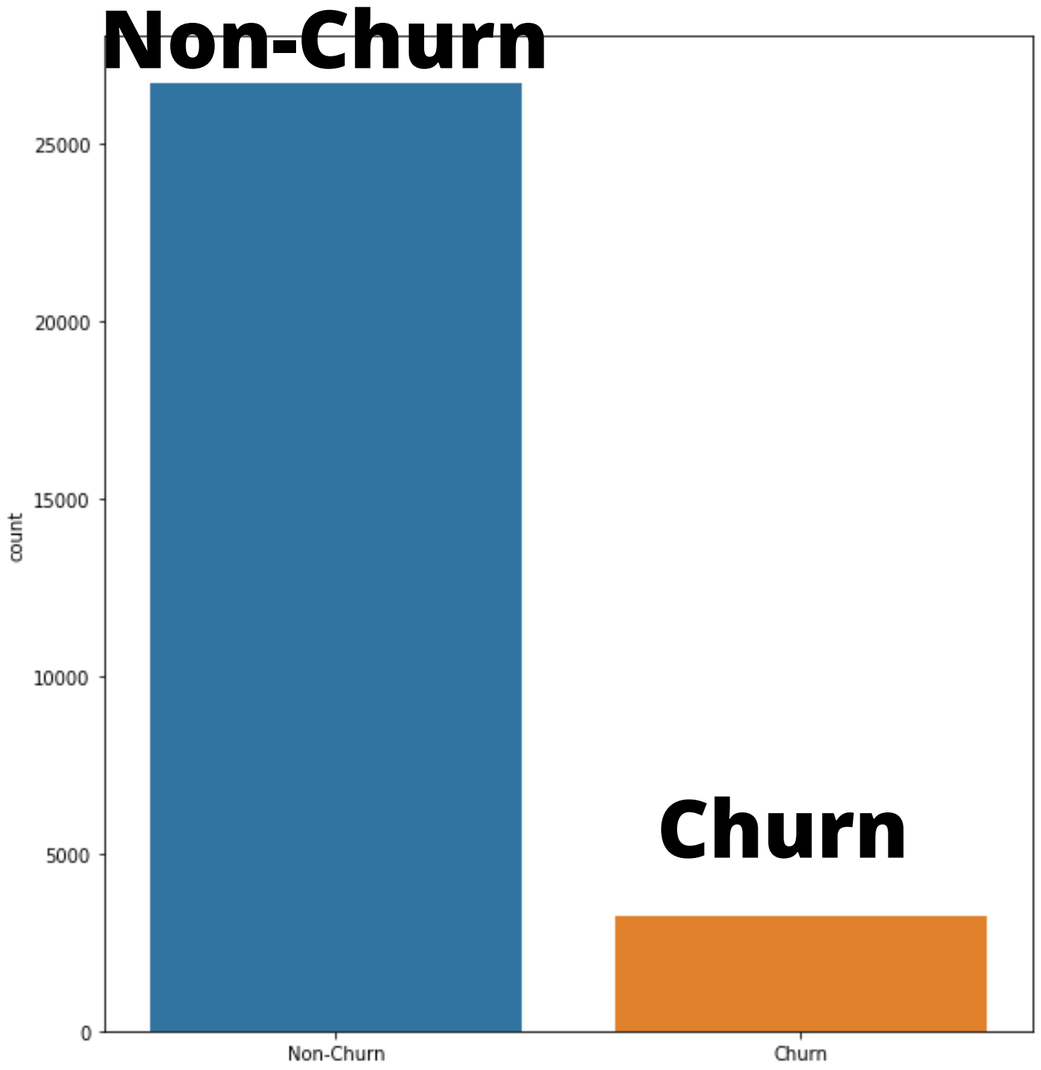
0=Non-Churned Customers

Model Building

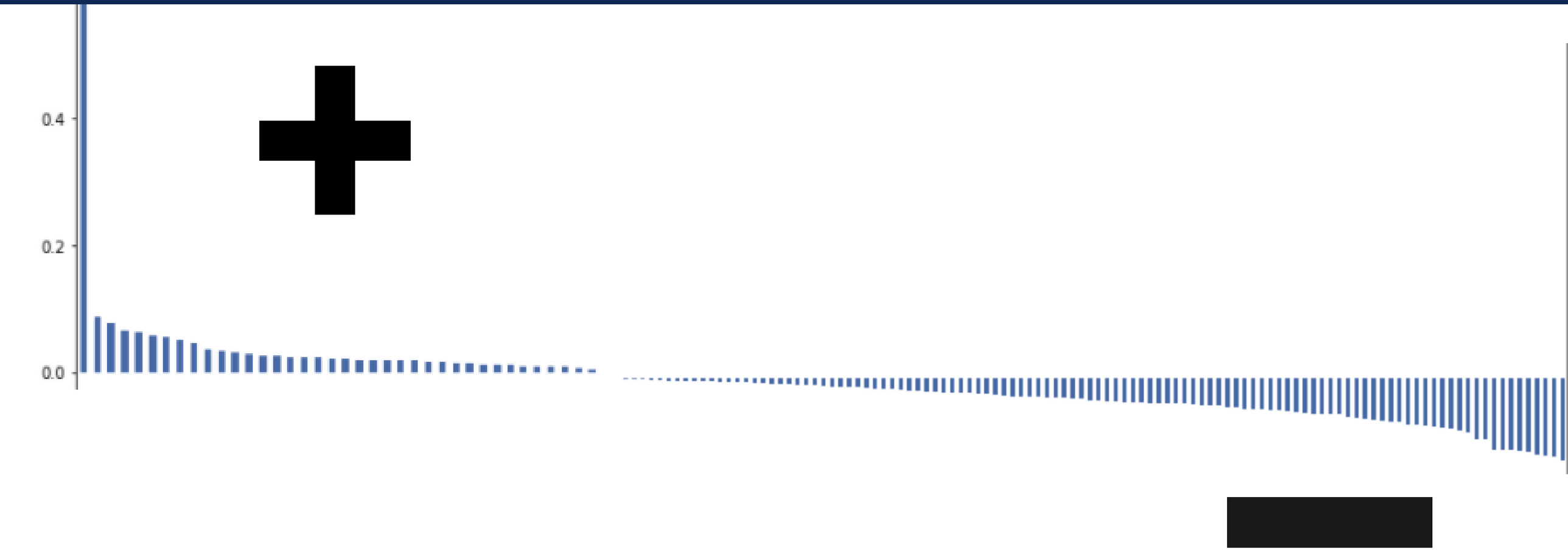
- Preprocess data
- Exploratory Data Analysis
- Derive new features.
- Reduce the number of variables using PCA
- Handle class imbalance
- Train different models and perform appropriate hyperparameters
- Evaluate the models using appropriate evaluation metrics
- Choose a model based on some evaluation metric

Further Exploratory Data





Exploratory Data Analysis: Correlation



Deriving New Features

[Out going call (June) + Out going call (July)]/2 = Average Out going call (June and July)

[3G Network (June) + 3G Network (July)]/2 = Average 3G Network (June and July)

Building The Model

Reducing Dimensions

(29953, 204)

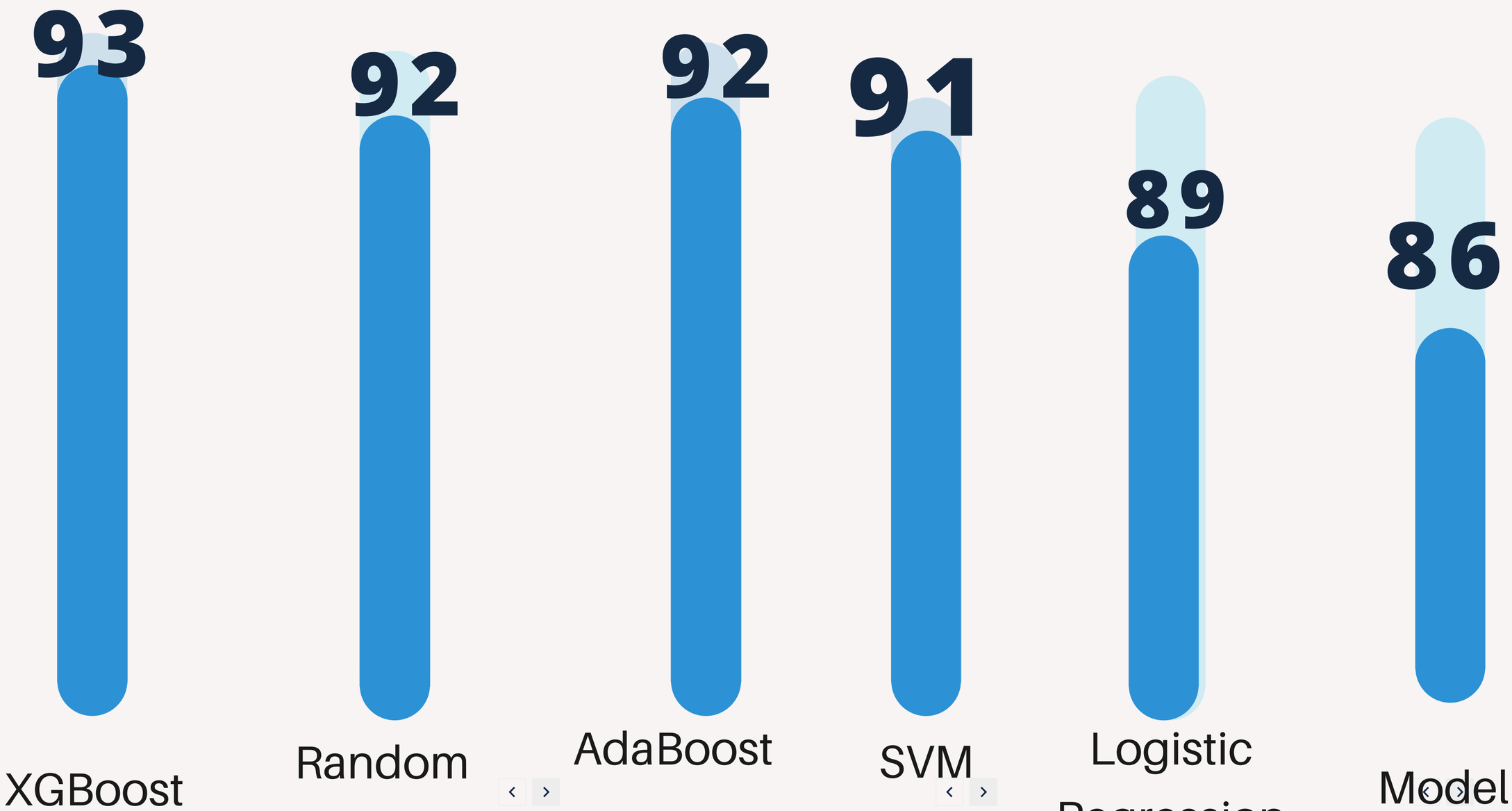
PCA

(5991, 54)

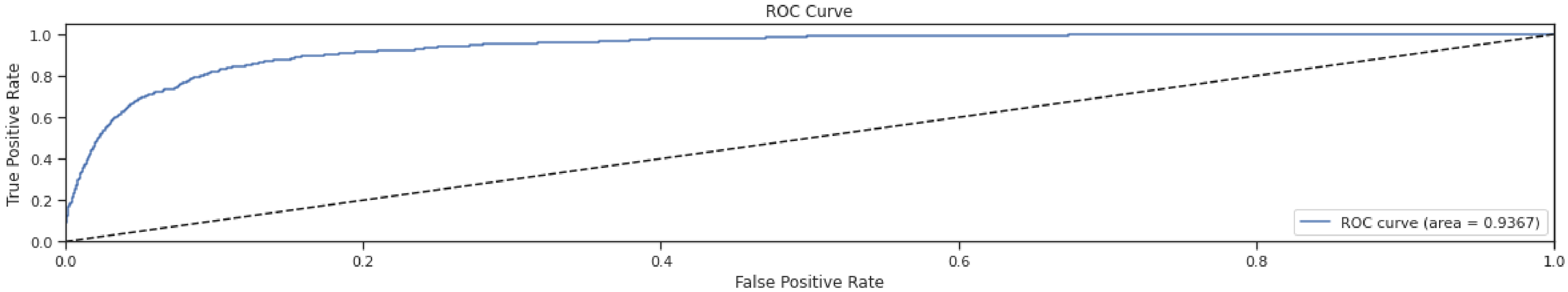
Imbalance Data

Upsampling
Downsampling

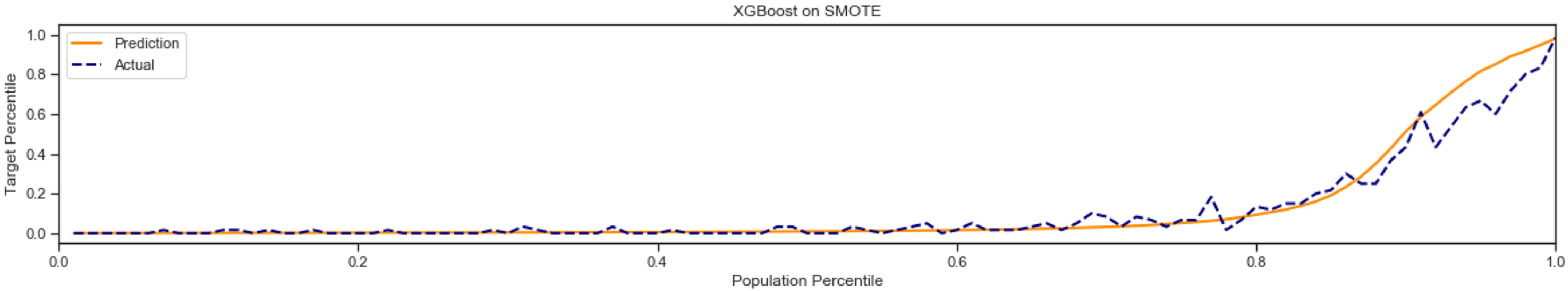
MODELS

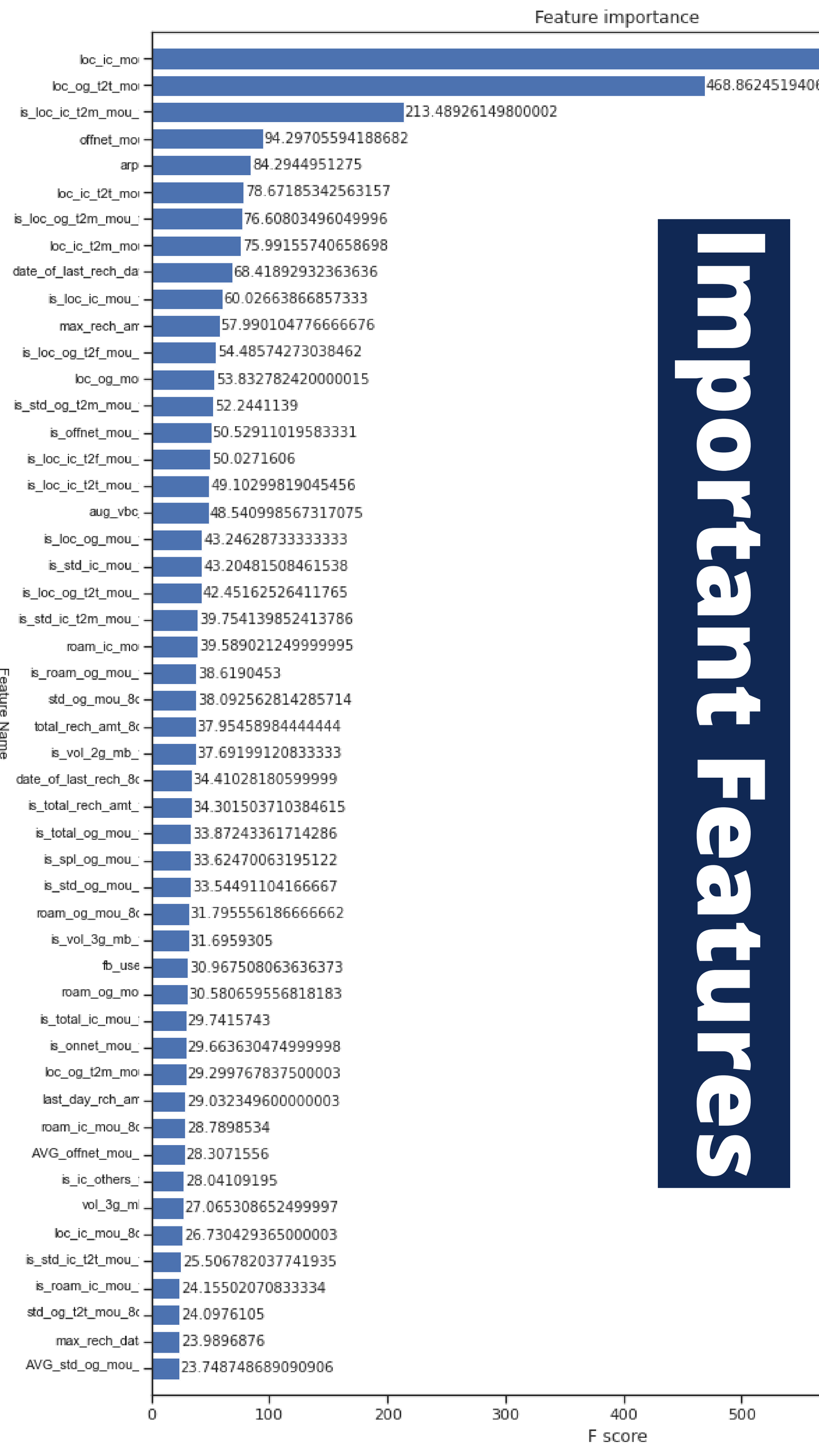

93
92
92
91
89
86
XGBoost
Random Forest
AdaBoost
SVM
Logistic Regression
Model

Evaluation Metrics



Evaluation Metrics




Feature importance
loc_ic_mo
loc_og_t2t_mo
468.8624519406
is_loc_ic_t2m_mou_
213.48926149800002
offnet_mo
94.29705594188682
arp
84.2944951275
loc_ic_t2t_mo
78.67185342563157
is_loc_og_t2m_mou_
76.60803496049996
loc_ic_t2m_mo
75.99155740658698
date_of_last_rech_da
68.41892932363636
is_loc_ic_mou_
60.02663866857333
max_rech_am
57.990104776666676
is_loc_og_t2f_mou_
54.48574273038462
loc_og_mo
53.832782420000015
is_std_og_t2m_mou_
52.2441139
is_offnet_mou_
50.52911019583331
is_loc_ic_t2f_mou_
50.0271606
is_loc_ic_t2t_mou_
49.10299819045456
aug_vbc
48.540998567317075
is_loc_og_mou_
43.24628733333333
is_std_ic_mou_
43.20481508461538
is_loc_og_t2t_mou_
42.45162526411765
is_std_ic_t2m_mou_
39.754139852413786
roam_ic_mo
39.589021249999995
is_roam_og_mou_
38.6190453
std_og_mou_8
38.092562814285714
total_rech_amt_8
37.95458984444444
is_vol_2g_mb_
37.69199120833333
date_of_last_rech_8
34.41028180599999
is_total_rech_amt_
34.301503710384615
is_total_og_mou_
33.87243361714286
is_spl_og_mou_
33.62470063195122
is_std_og_mou_
33.54491104166667
roam_og_mou_8
31.795556186666662
is_vol_3g_mb_
31.6959305
fb_use
30.967508063636373
roam_og_mo
30.580659556818183
is_total_ic_mou_
29.7415743
is_onnet_mou_
29.663630474999998
loc_og_t2m_mo
29.299767837500003
last_day_rch_am
29.032349600000003
roam_ic_mou_8
28.7898534
AVG_offnet_mou_
28.3071556
is_ic_others_
28.04109195
vol_3g_m
27.065308652499997
loc_ic_mou_8
26.730429365000003
is_std_ic_t2t_mou_
25.506782037741935
is_roam_ic_mou_
24.15502070833334
std_og_t2t_mou_8
24.0976105
max_rech_dat
23.9896876
AVG_std_og_mou_
23.748748689090906
Feature Name
0
100
200
300
400
500
F score
Important Features

Sample Important Features

Data of last recharge in the 8th month

The last day recharge amount in 8th month

Total In-coming and out-going calls

Last day recharge amount in the 6th month

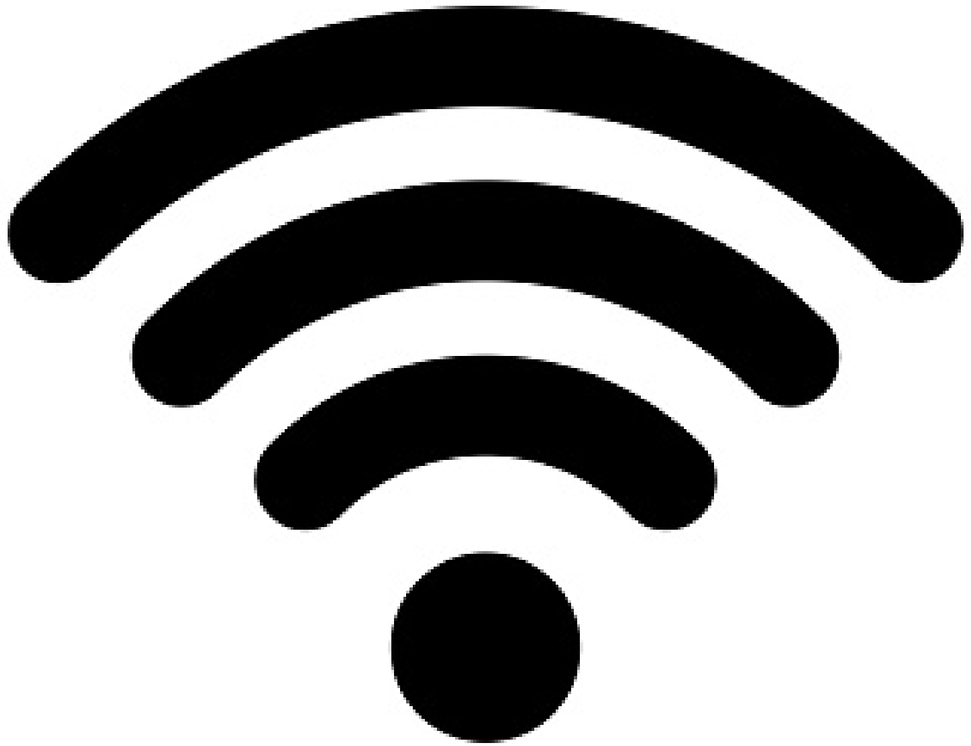
Average Revenue Per User

Recommendations to Business

We can infer from our analysis that there are certain features which are high indicaters of whether a customer will churn or not. In order for Telecom to reduce the number of churn and maintain these high-valued customers, it has to take note of these features and monitor it quite consistently in order to identify the stage those customers that are likely to churn and address their pain-point as soon as possible to prevent them from churning.

Another key factor to be consider is that if the recharge amount of the customer starts to reduce at the month where the customer has shown consistenet activeness, then it's a sign of churn and must be address as soon as possible.

Age on network is also a key indicator for identifying the churn, if age is less than 500 days and their usage is reduce then the customer is likely to churned. The other key indicater that the company should consider is data usage & amount.



Thank You