Telecom Churn Case Study

SUBJECT: TELECOM CHURN CASE STUDY – DOMAIN ORIENTED

PRESENTER: ARUNDHATHI V R

Problem Statement

In the telecom industry, customers are able to choose from multiple service providers and actively switch from one operator to another. In this highly competitive market, the telecommunications industry experiences an average of 15-25% annual churn rate. Given the fact that it costs 5-10 times more to acquire a new customer than to retain an existing one, **customer retention** has now become even more important than customer acquisition.

For many incumbent operators, *retaining high profitable customers is the number one business goal*. To reduce customer churn, telecom companies need to **predict which customers are at high risk of churn.**

In this project, you will analyse customer-level data of a leading telecom firm, build predictive models to identify customers at high risk of churn and identify the main indicators of churn.

Steps of analysis

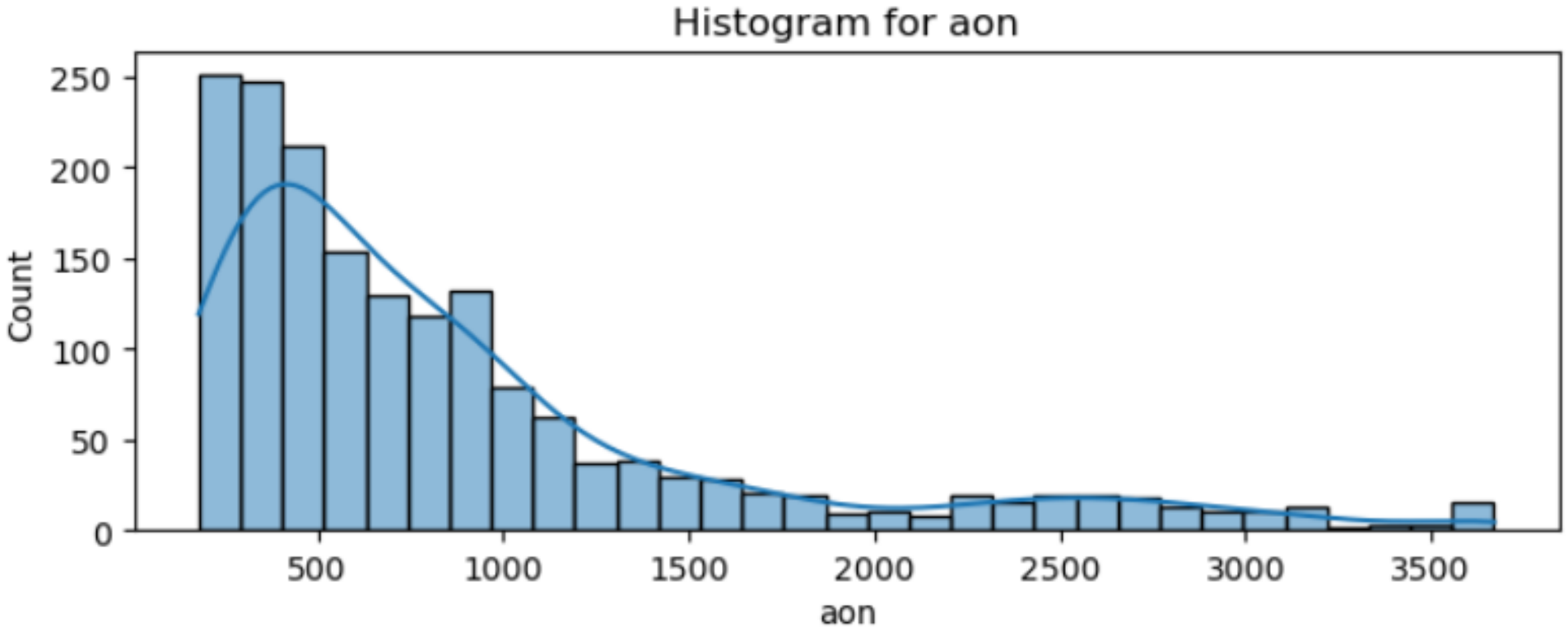
1. Reading and understanding data
2. Cleaning data
3. Defining target variable
4. EDA and feature engineering
5. Reducing dimensionality
6. Outlier management
7. Handling High Imbalance class in target variable
8. Data modelling - Logistic regression and Random forest
9. Feature importance

Reducing dimensionality

Below techniques are used to reduce Dimensionality of the data.

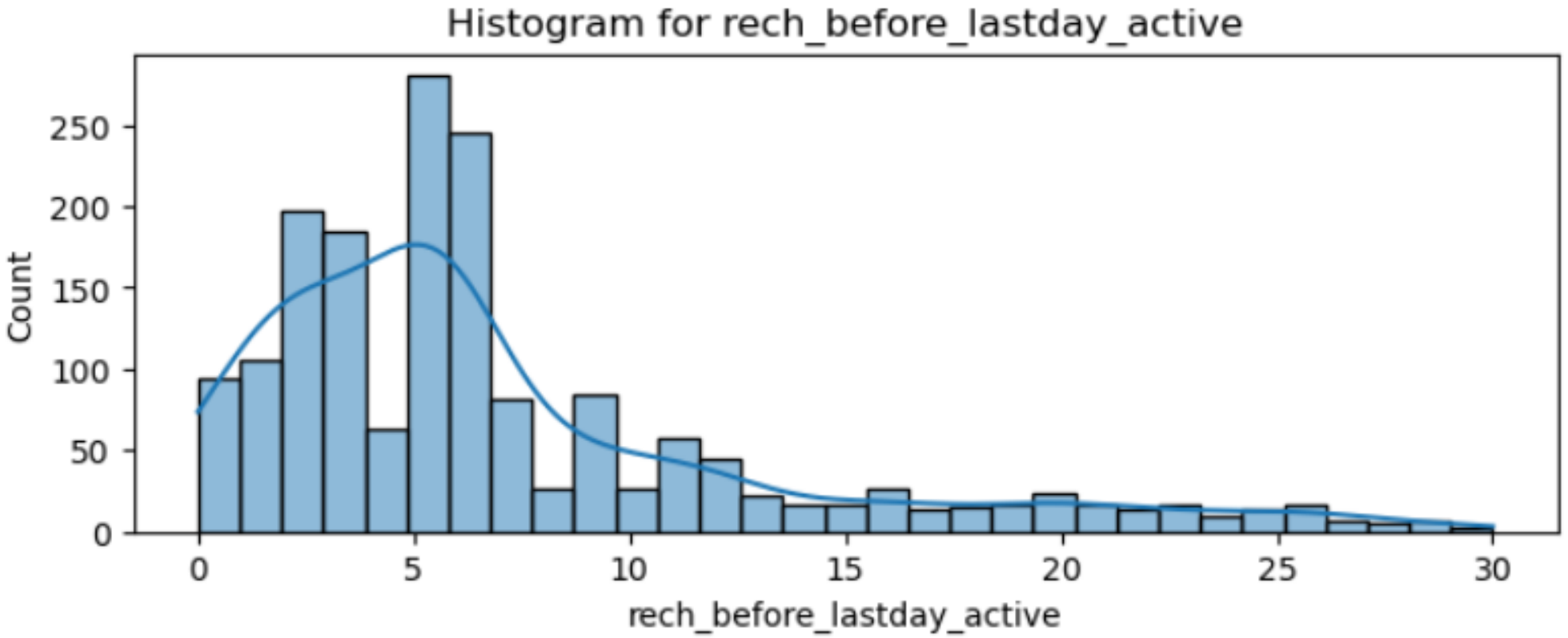
- Null value management
- 0 variance reduction
- Variables highly correlated to other variables removed
- Identified high value customers who contributes almost 90% of revenue
- Variables with low correlation to target variables removed
- Data shape before dimensionality – (99999, 226)
- Data shape before dimensionality – (28485, 25)

Visualization



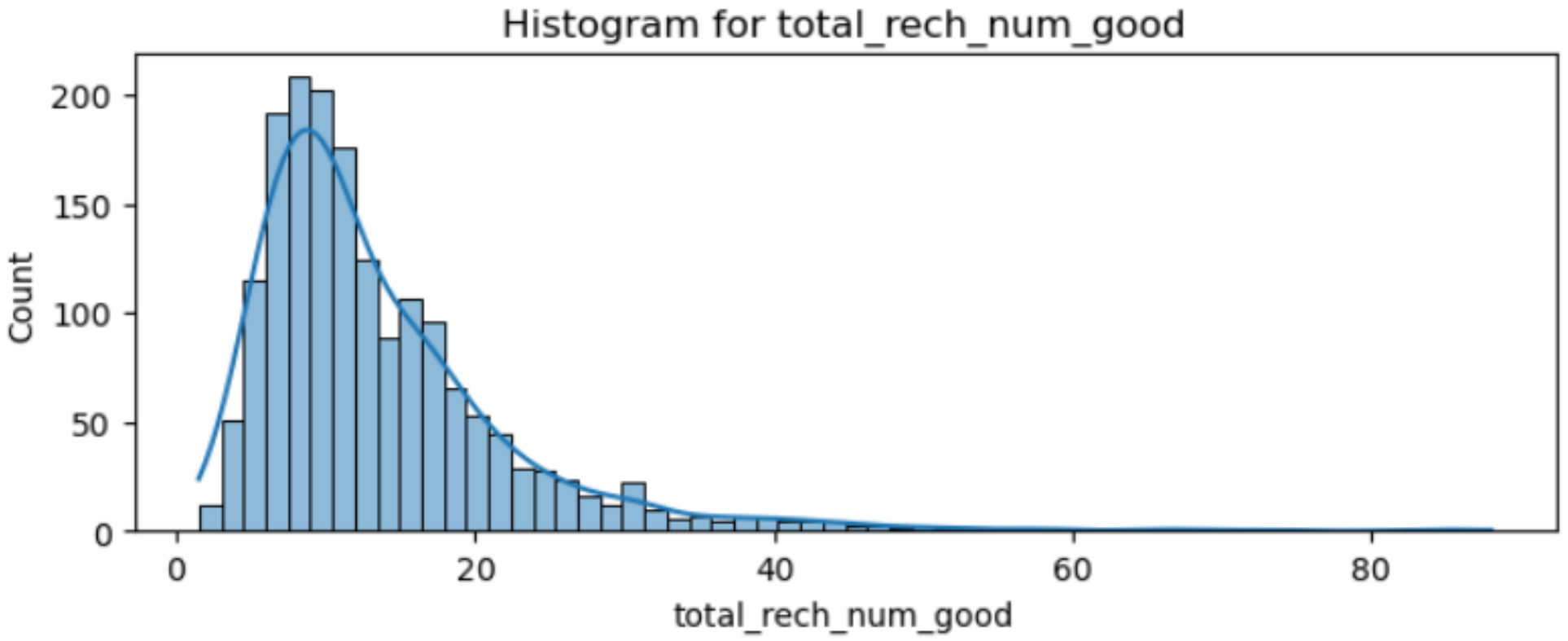
Churn rate is low in customers with longer tenure or age on network

Visualization



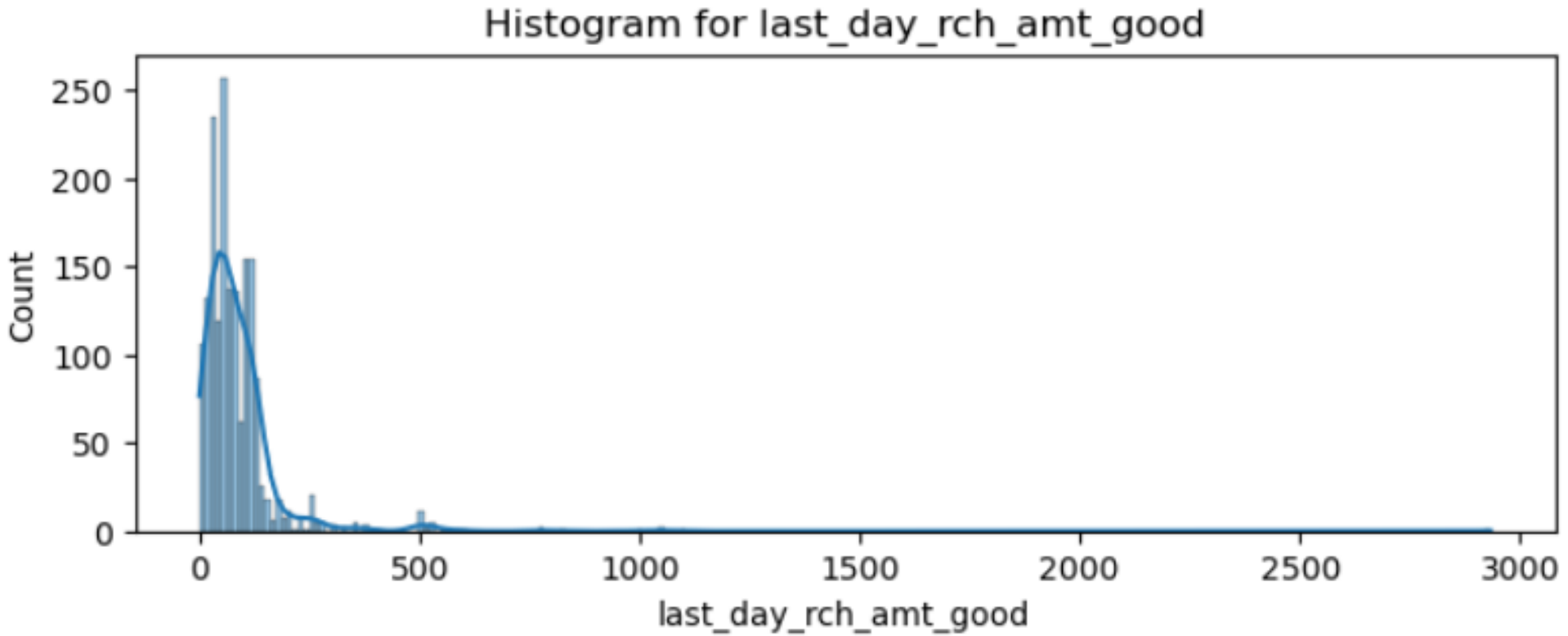
Churn rate is low as frequency of recharge in active month increases

Visualization



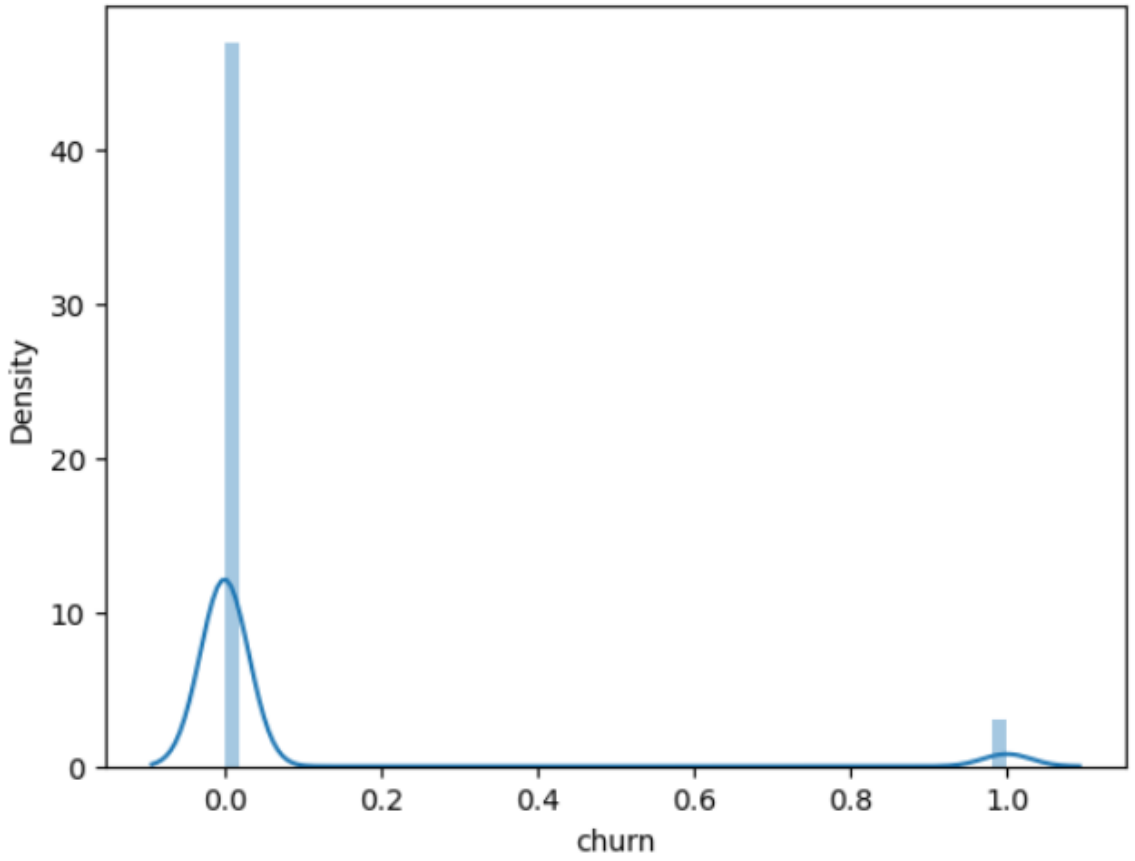
Churn rate is low as frequency of total recharge in good month increases

Visualization



Customers who tend to churn recharges with amt less than 200 on last recharge

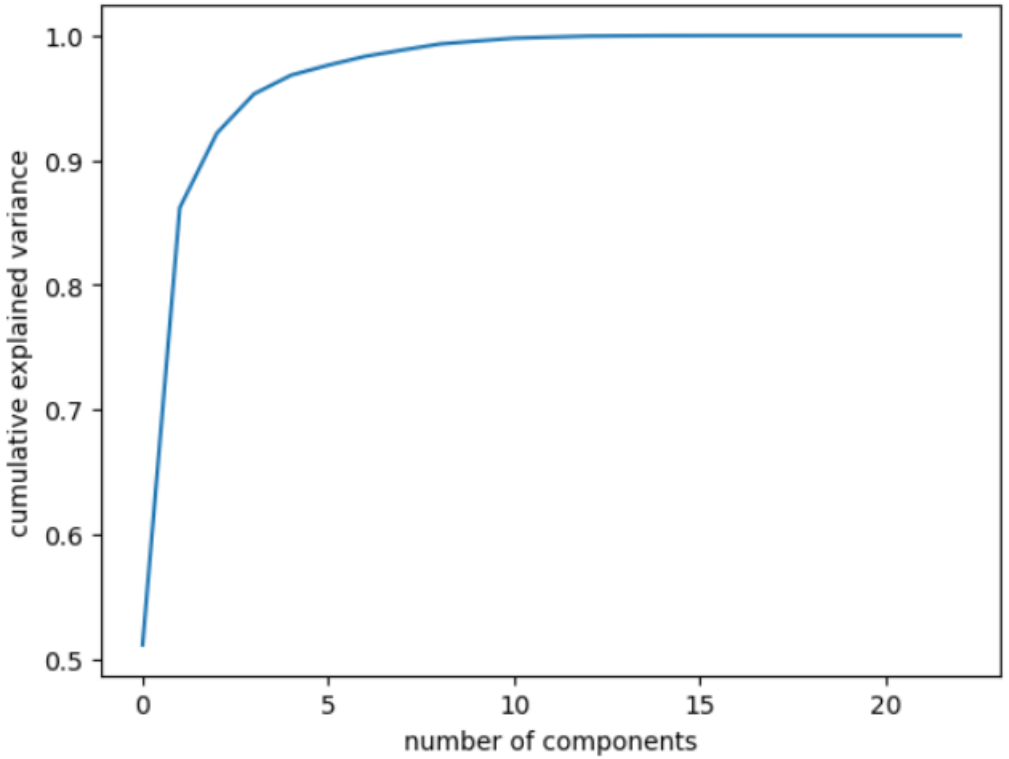
Data imbalance



Inferences:

- Target variable is highly imbalanced.
- Used SMOTE to handle High imbalanced data

Plotting PCA to explain variance



Inferences:

- 99% of variance is explained by 11 components

Model 1 - Logistic regression

Classification Report:

	precision	recall	f1-score	support
0	0.82	0.68	0.75	5624
1	0.73	0.85	0.78	5578
accuracy			0.77	11202
macro avg	0.77	0.77	0.77	11202
weighted avg	0.77	0.77	0.77	11202

Confusion Matrix:

```
[[3850 1774]
 [ 839 4739]]
```

Accuracy : 0.7958128326137163

True Positive Rate (Sensitivity, Recall): 0.8062449959967974
True Negative Rate (Specificity): 0.7951148963522416
False Positive Rate (Fall-Out): 0.2048851036477583
Precision (Positive Predictive Value): 0.2084023178807947
Negative Predictive Value: 0.9839586371470237

Model 2 - Random Forest

Accuracy: 0.944476982546562

Classification Report:

	precision	recall	f1-score	support
0	0.95	0.99	0.97	8037
1	0.57	0.20	0.30	500
accuracy			0.94	8537
macro avg	0.76	0.60	0.63	8537
weighted avg	0.93	0.94	0.93	8537

Confusion Matrix:

```
[[7962   75]
 [ 399  101]]
```

Model 3 - Random Forest

Accuracy: 0.9454140798875483

Classification Report:

	precision	recall	f1-score	support
0	0.95	0.99	0.97	8037
1	0.59	0.21	0.31	500
accuracy			0.95	8537
macro avg	0.77	0.60	0.64	8537
weighted avg	0.93	0.95	0.93	8537

Confusion Matrix:

```
 [[7964   73]
 [ 393  107]]
```

Model 4 - Random Forest

Accuracy: 0.9329975401194799

Classification Report:

	precision	recall	f1-score	support
0	0.97	0.96	0.96	8037
1	0.44	0.51	0.47	500
accuracy			0.93	8537
macro avg	0.70	0.73	0.72	8537
weighted avg	0.94	0.93	0.94	8537

Confusion Matrix:

```
[[7712  325]
 [ 247  253]]
```

Model evaluation

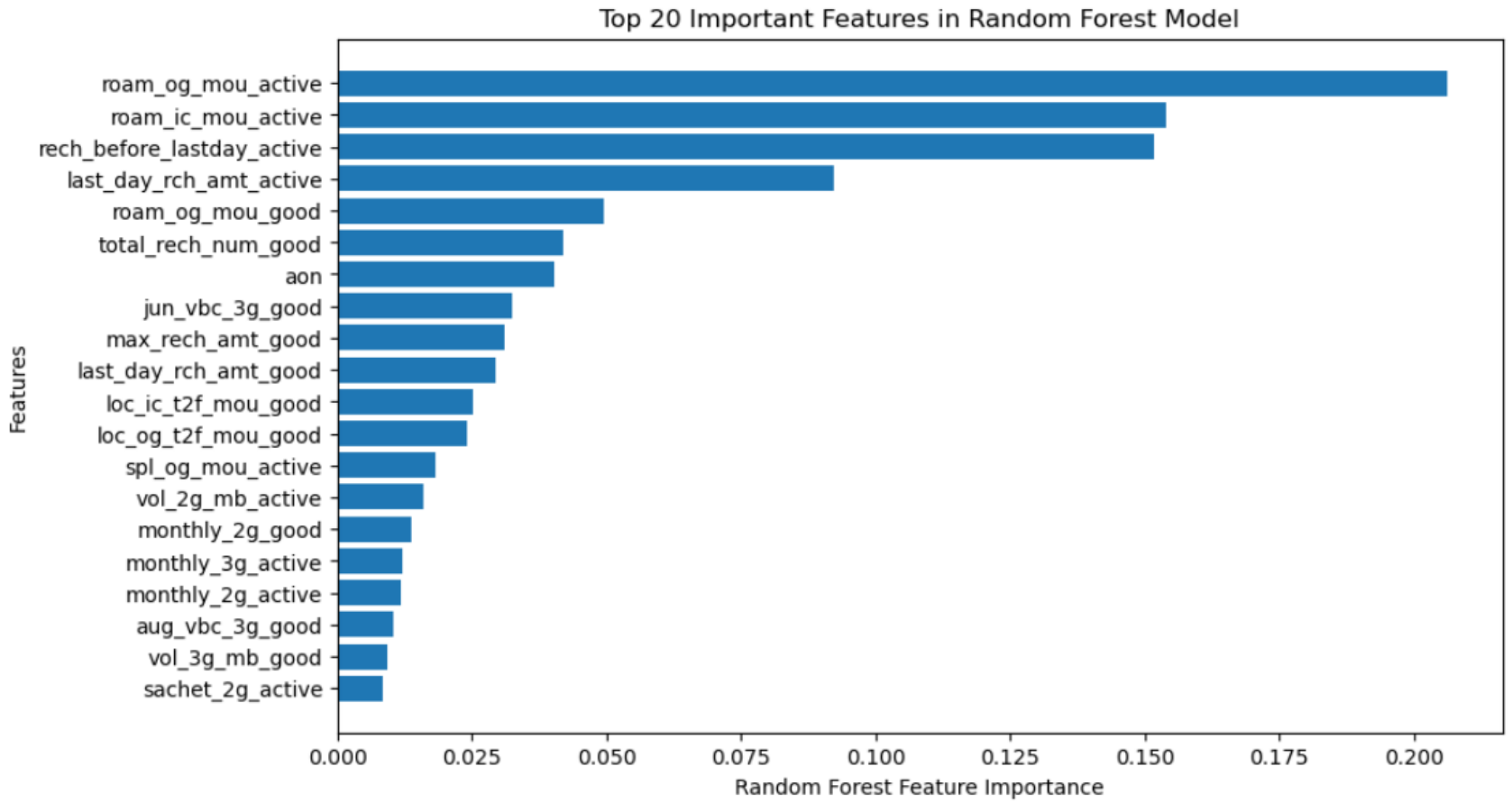
Logistic regression model perform better than Random forest since it captures the Churn customer data better

	precision	recall	f1-score	support
0	0.82	0.68	0.75	5624
1	0.73	0.85	0.78	5578

Variables used in final model:

spl_og_mou_active',
'spl_ic_mou_active',
'monthly_2g_active',
'sachet_2g_active',
'monthly_3g_active',
'rech_before_lastday_active',
'loc_og_t2f_mou_good',
'total_rech_num_good',
'monthly_2g_good'

Variable importance as per Random forest



Inferences from the Final Model

- Churn rate is low in customers with longer tenure or age on network
- Churn rate is low as frequency of recharge in active month increases
- Churn rate is low as frequency of total recharge in good month increases
- Customers who tend to churn recharges with amt less than 200 on last recharge

Business advise

- Company should focus on low tenure high value customers because they are highly prone to churn
- Concentrate on Customers who tend to churn recharges with amt less than 200 on last recharge
- Company should keep an eye on the drop of spl_og_mou, and spl_ic_mou, 'monthly_2g, monthly_3g, rech_before_lastday in active month
- There is an overall reduction of activity in active month for the customers who are likely to churn