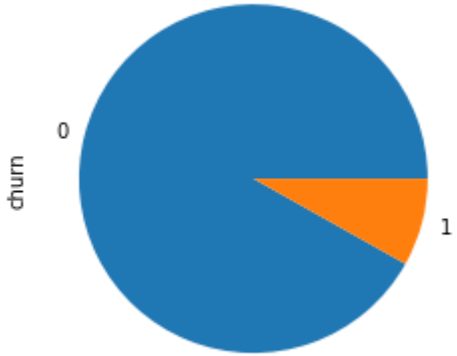
Telecom churn

Analyzing the churn Data

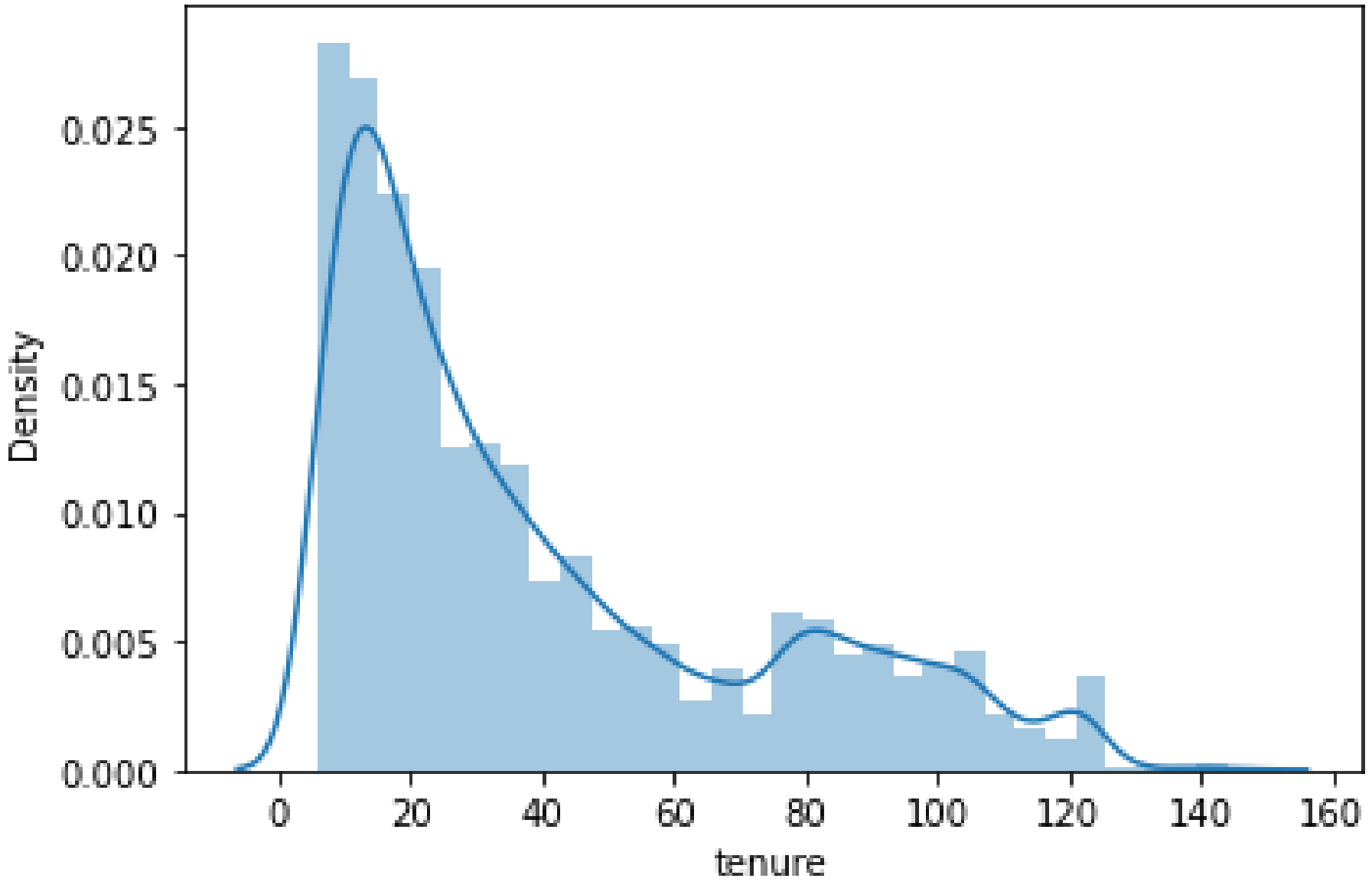
Content

1. Statistical Analysis of the Data
2. Handling missing values
3. Handling the other attributes with higher missing value percentage
4. Filtering the High Value Customer from Good Phase
5. *Defining Churn variable*
6. Deriving new variables to understand the data
7. Precision and recall trade-off
8. Metrics Evaluation



As we can see that 91% of the customers do not churn, there is a possibility of class imbalance

0 2-5 Yrs 7 2-5 Yrs 8 6-12 Months 21 1-2 Yrs 23 1-2 Yrs



0
2-5 Yrs

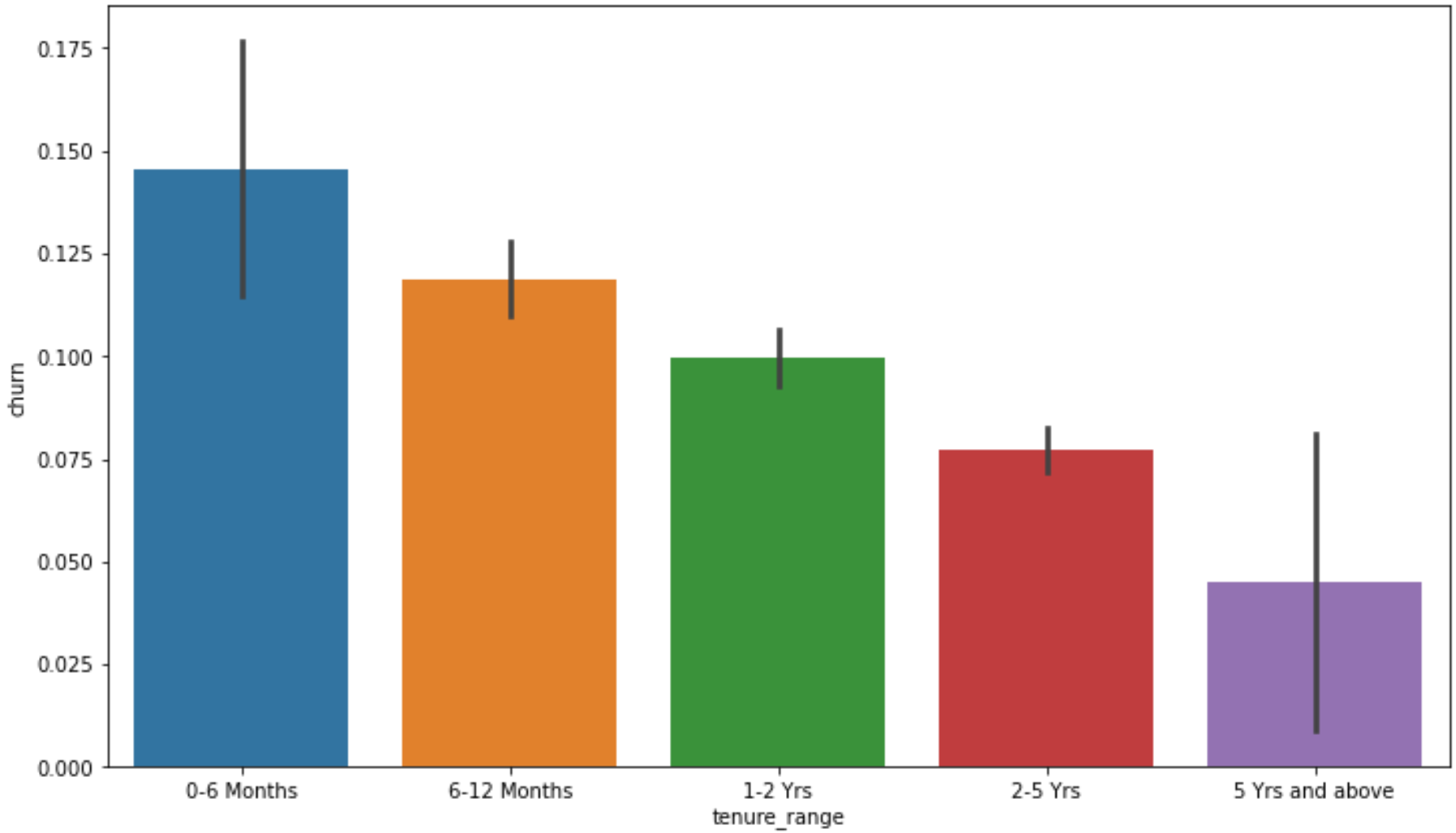
7
2-5 Yrs

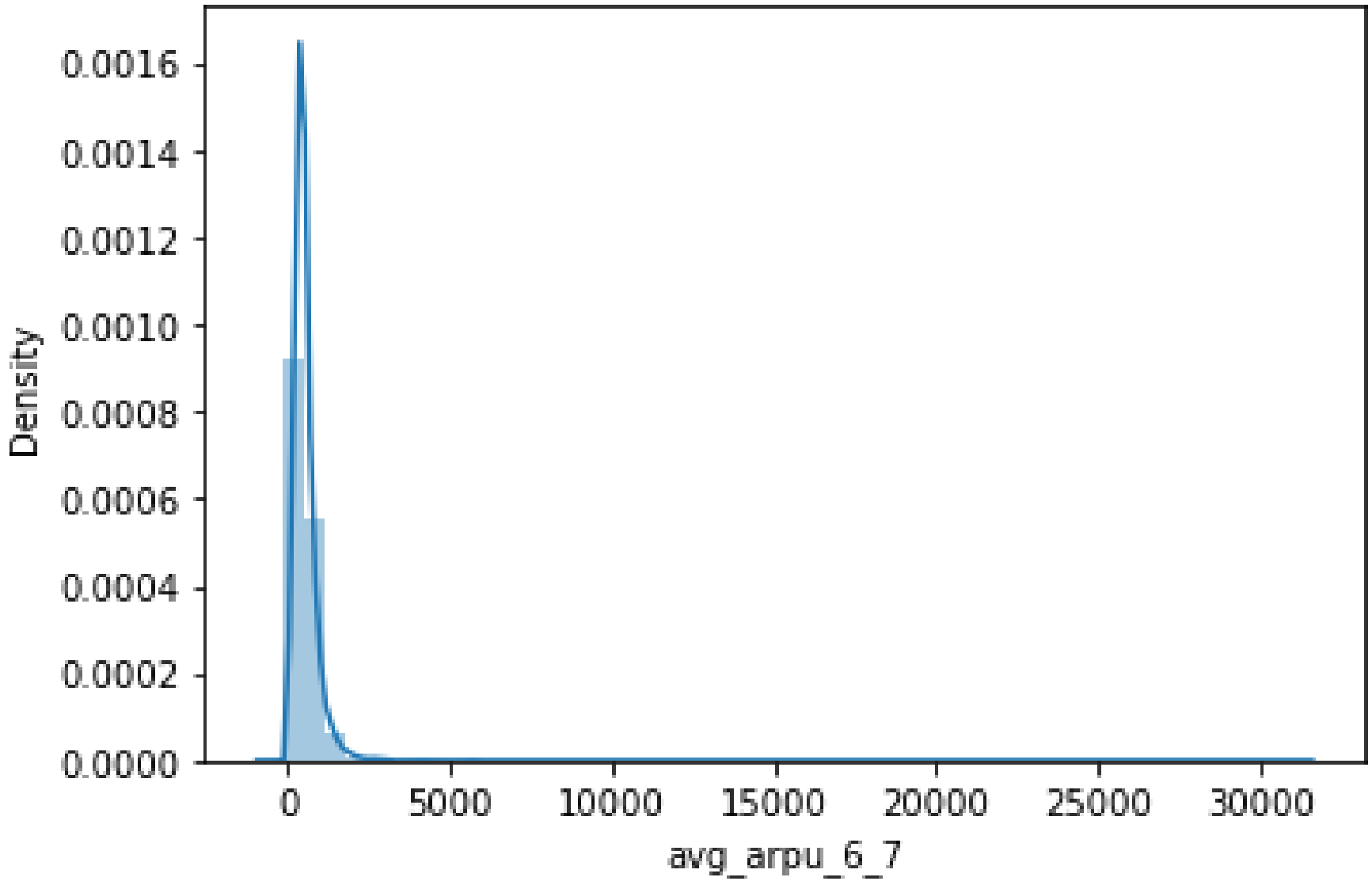
8
6-12 Months

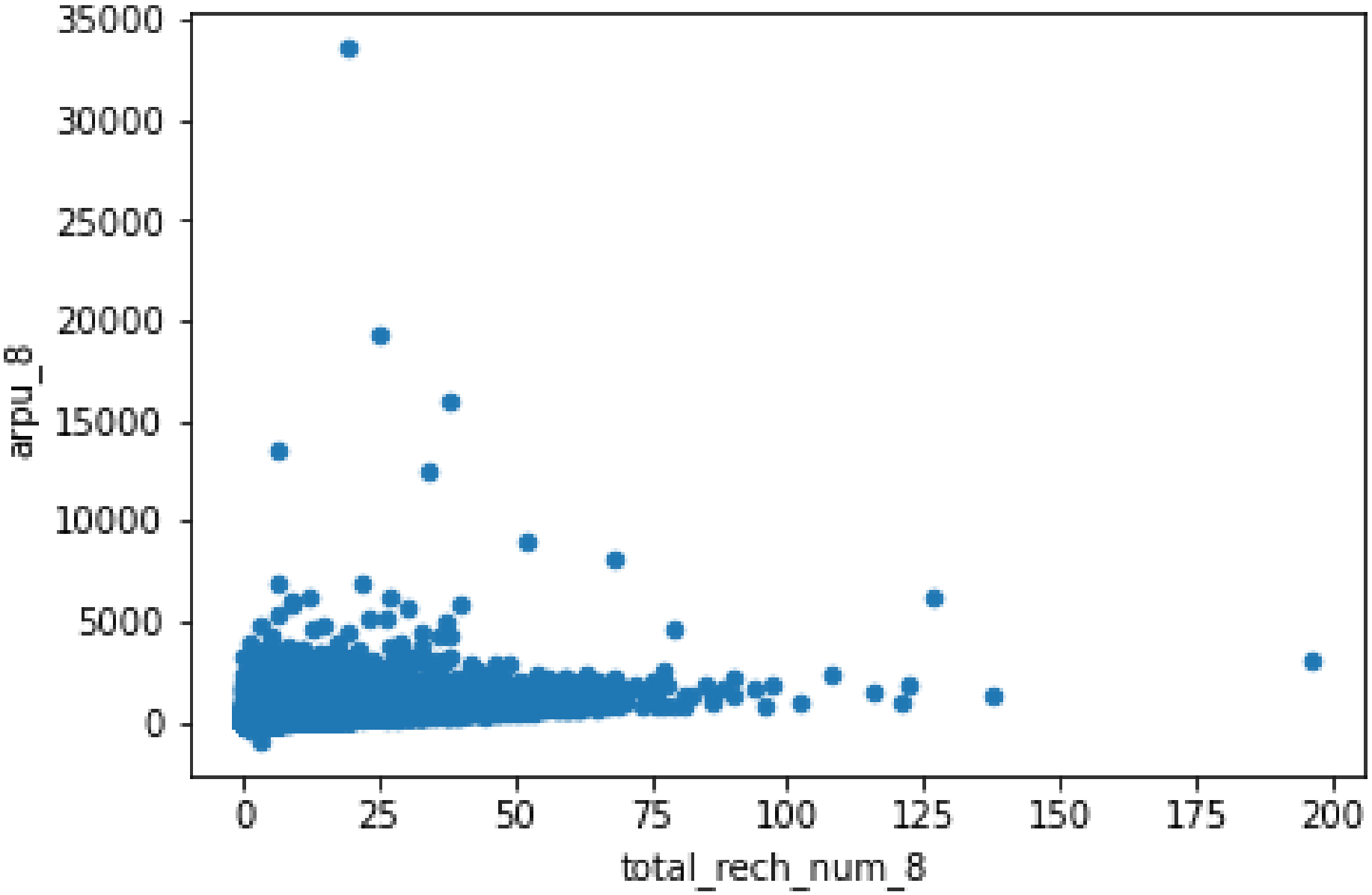
21
1-2 Yrs

23
1-2 Yrs

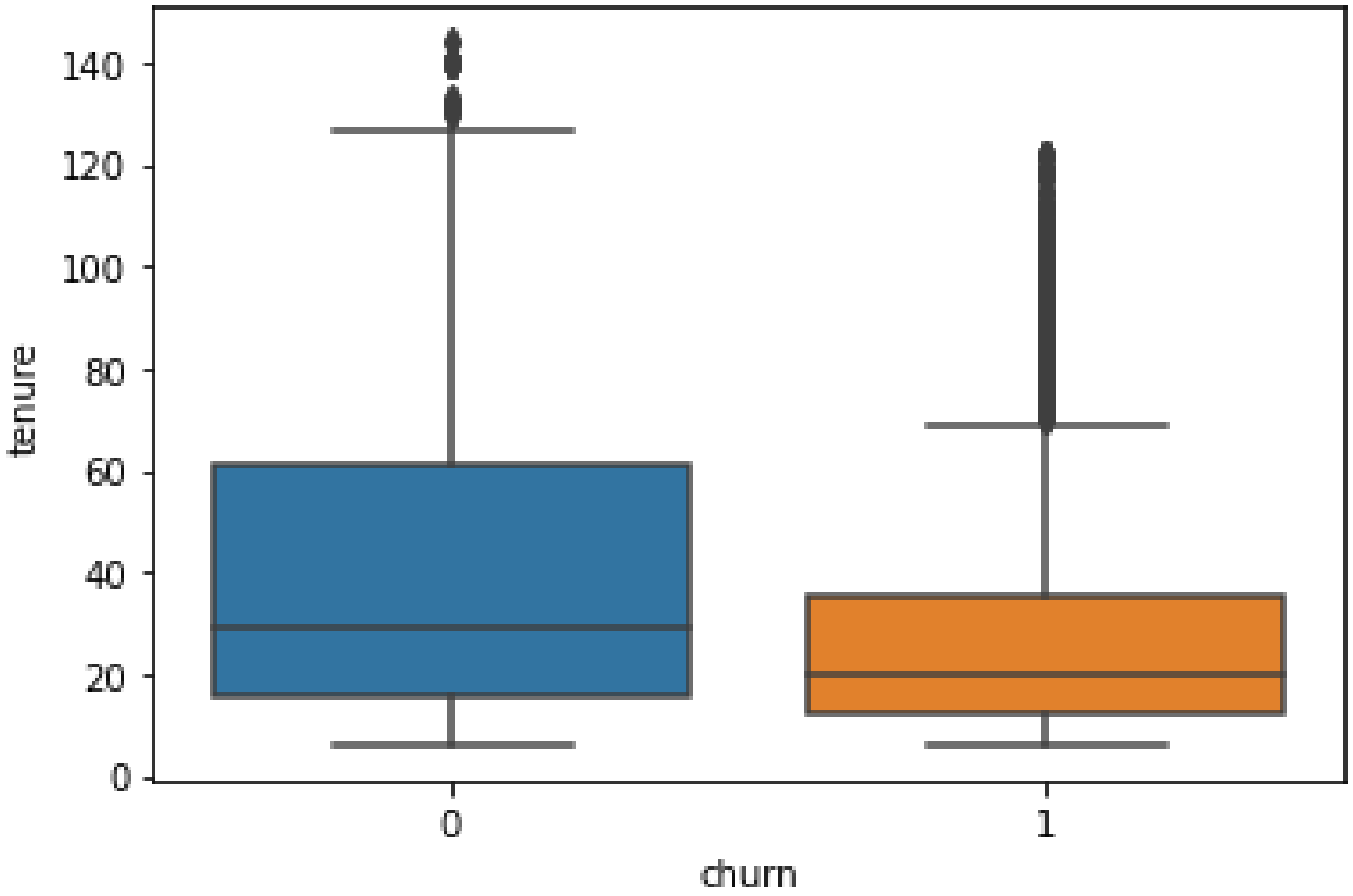
It can be seen that the maximum churn rate happens within 0-6 month, but it gradually decreases as the customer retains in the network.

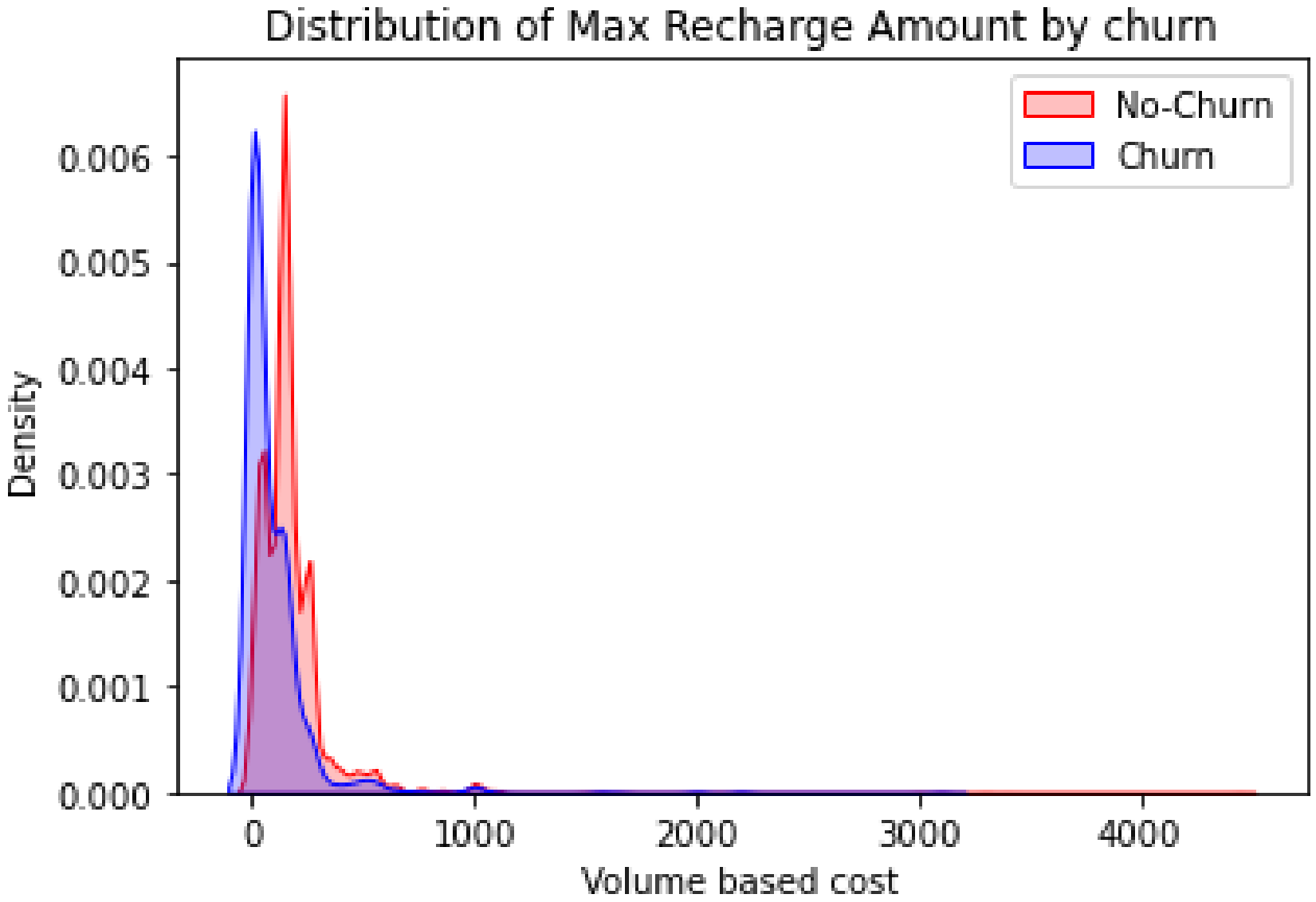


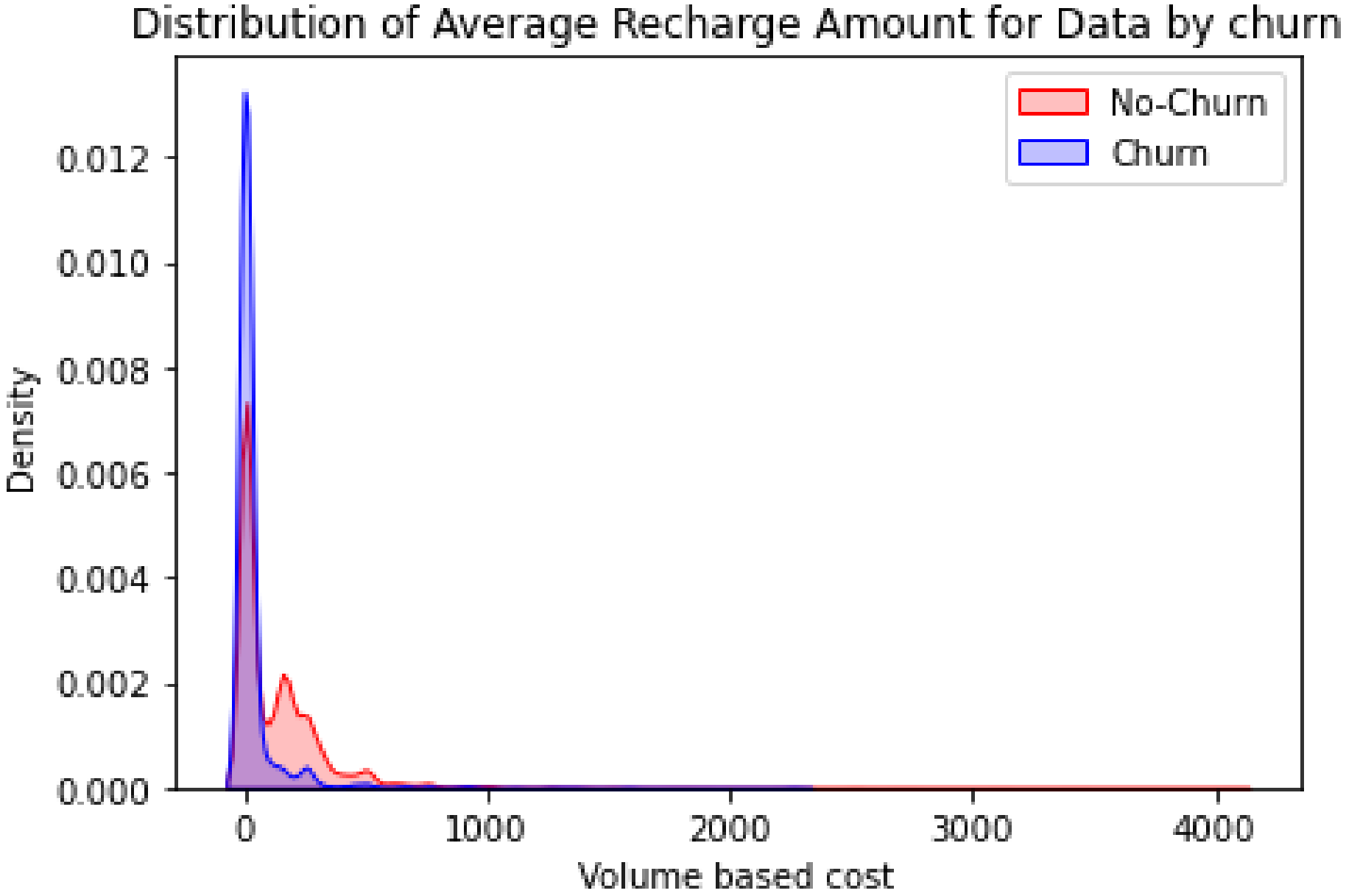

0.0016
0.0014
0.0012
0.0010
0.0008
0.0006
0.0004
0.0002
0.0000
Density
0
5000
10000
15000
20000
25000
30000
avg_arpu_6_7

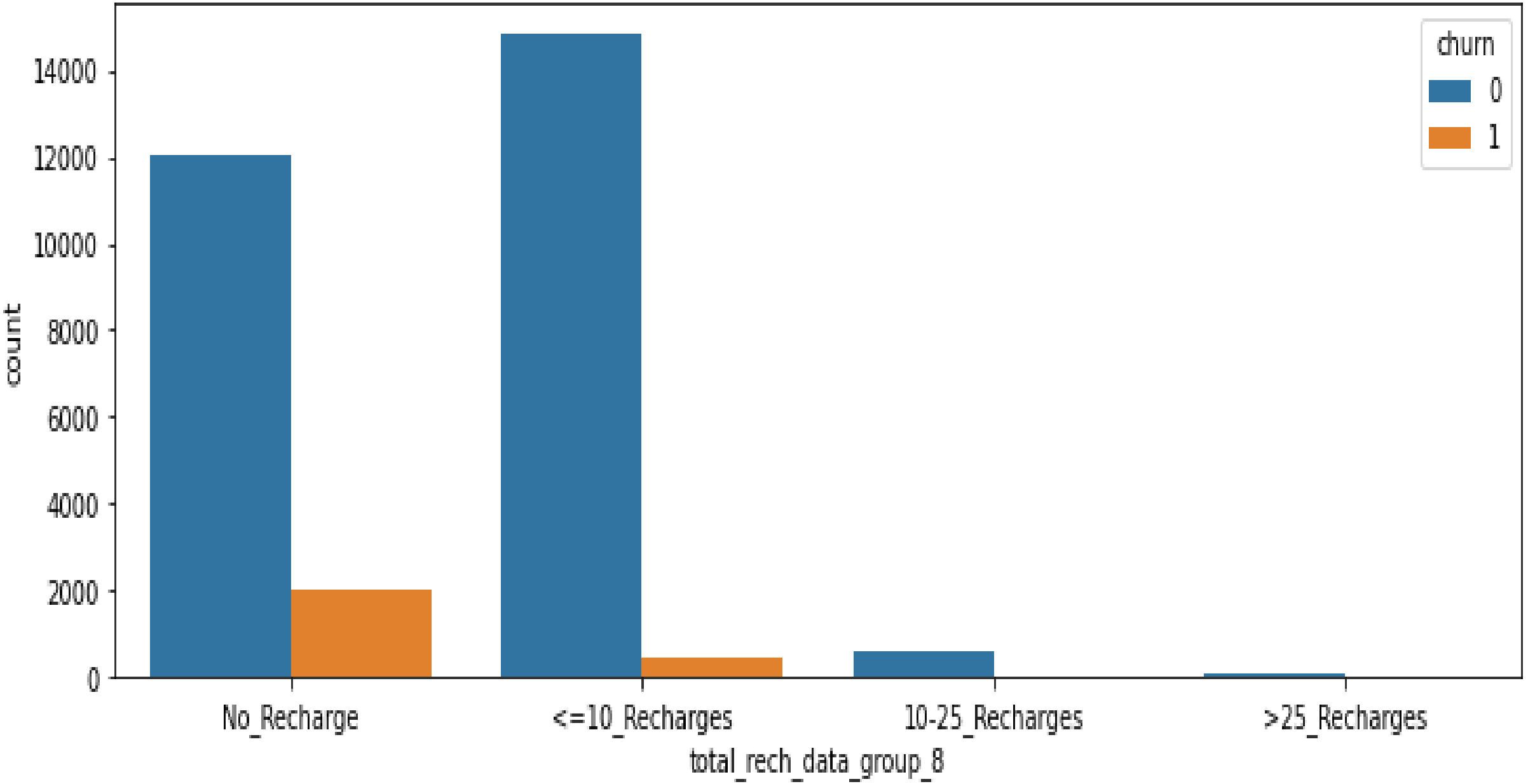

35000
30000
25000
20000
15000
10000
5000
0
arpu_8
0
25
50
75
100
125
150
175
200
total_rech_num_8

From the above plot , its clear tenured customers do no churn and they keep availing telecom services

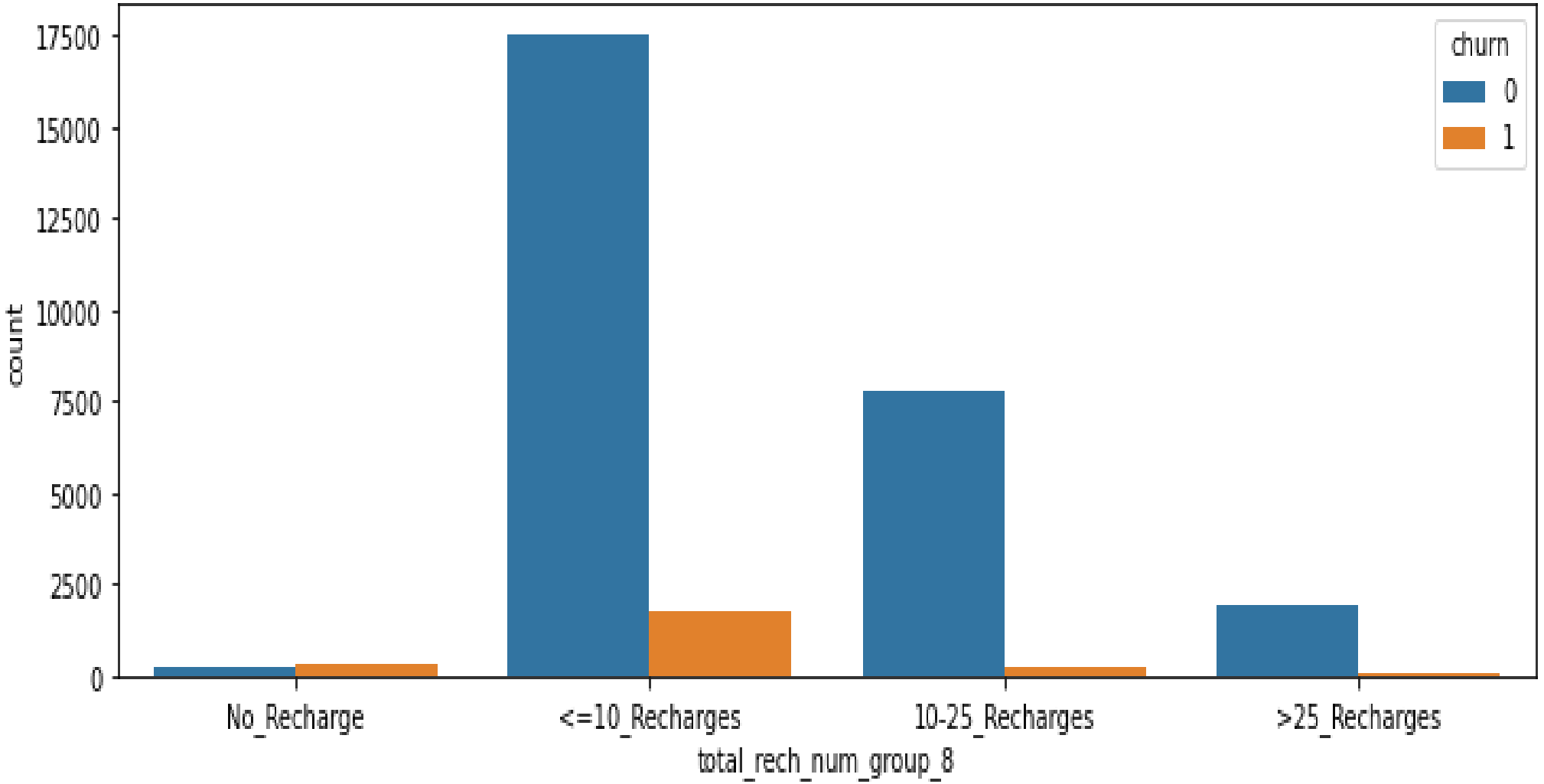


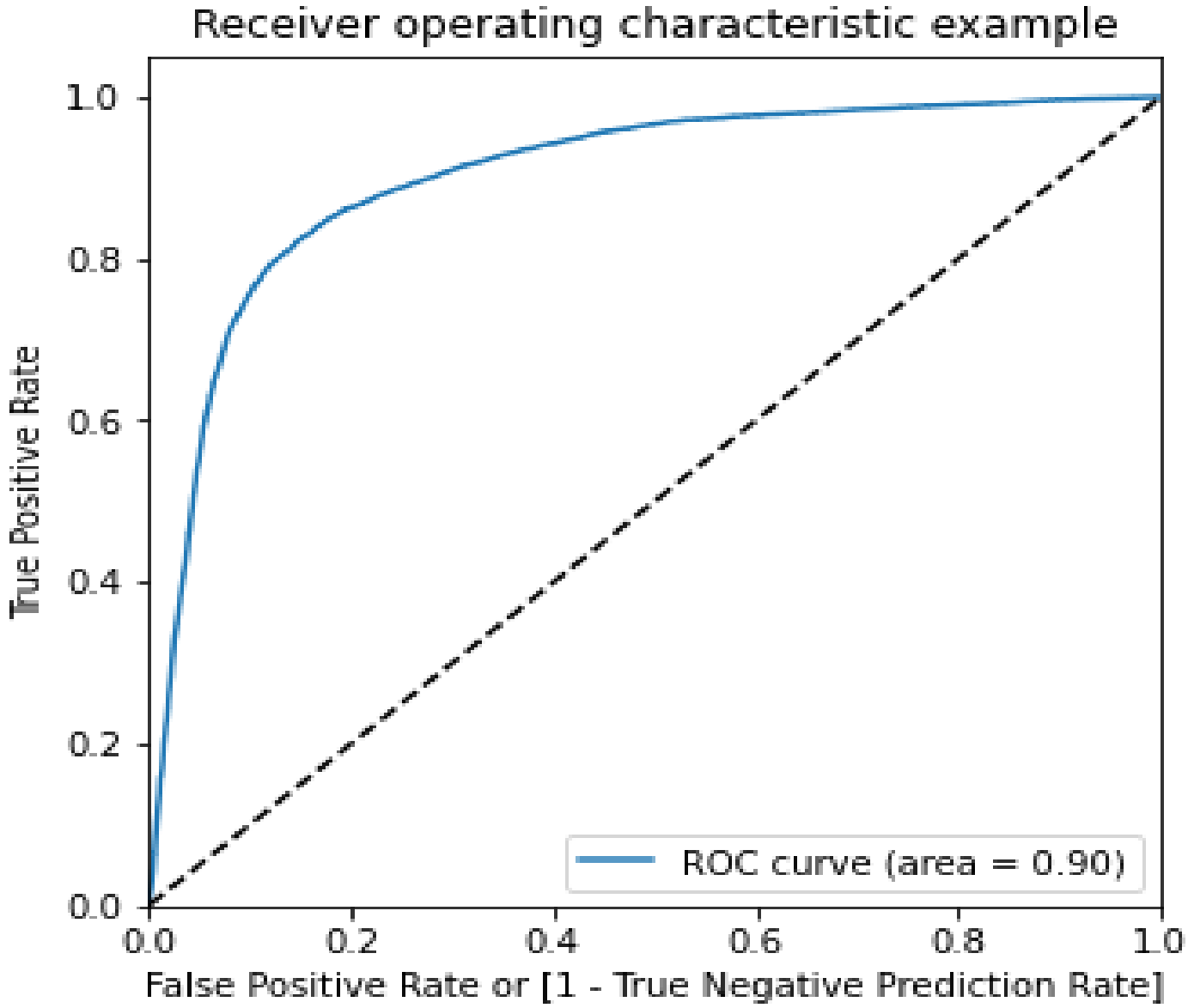

Distribution of Max Recharge Amount by churn
No-Churn
Churn
Density
0.006
0.005
0.004
0.003
0.002
0.001
0.000
0
1000
2000
3000
4000
Volume based cost

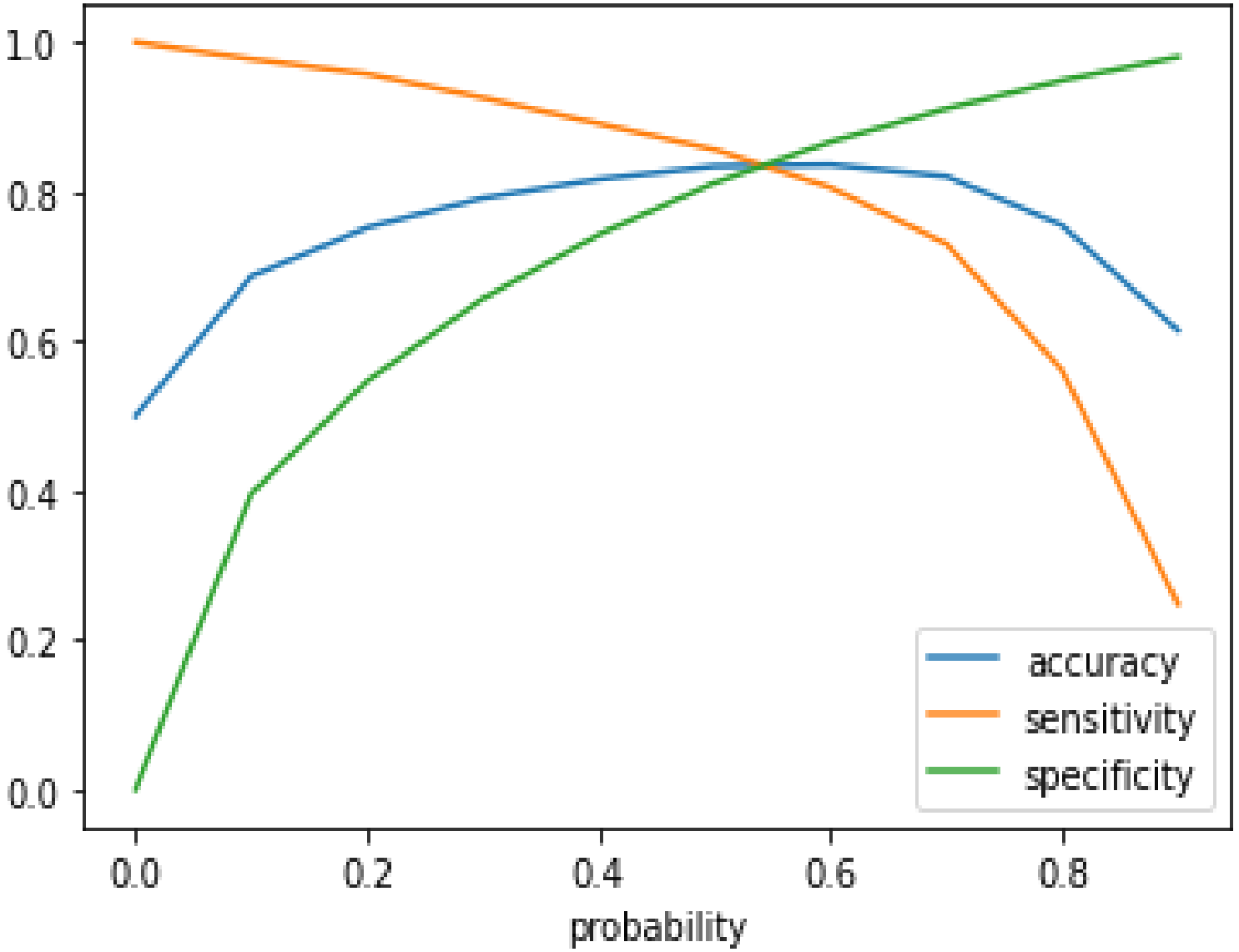

Distribution of Average Recharge Amount for Data by churn
No-Churn
Churn
Density
0.012
0.010
0.008
0.006
0.004
0.002
0.000
0
1000
2000
3000
4000
Volume based cost


churn
0
1
14000
12000
10000
8000
6000
4000
2000
0
count
No_Recharge
<=10_Recharges
10-25_Recharges
>25_Recharges
total_rech_data_group_8

As the number of recharge rate increases, the churn rate decreases clearly.

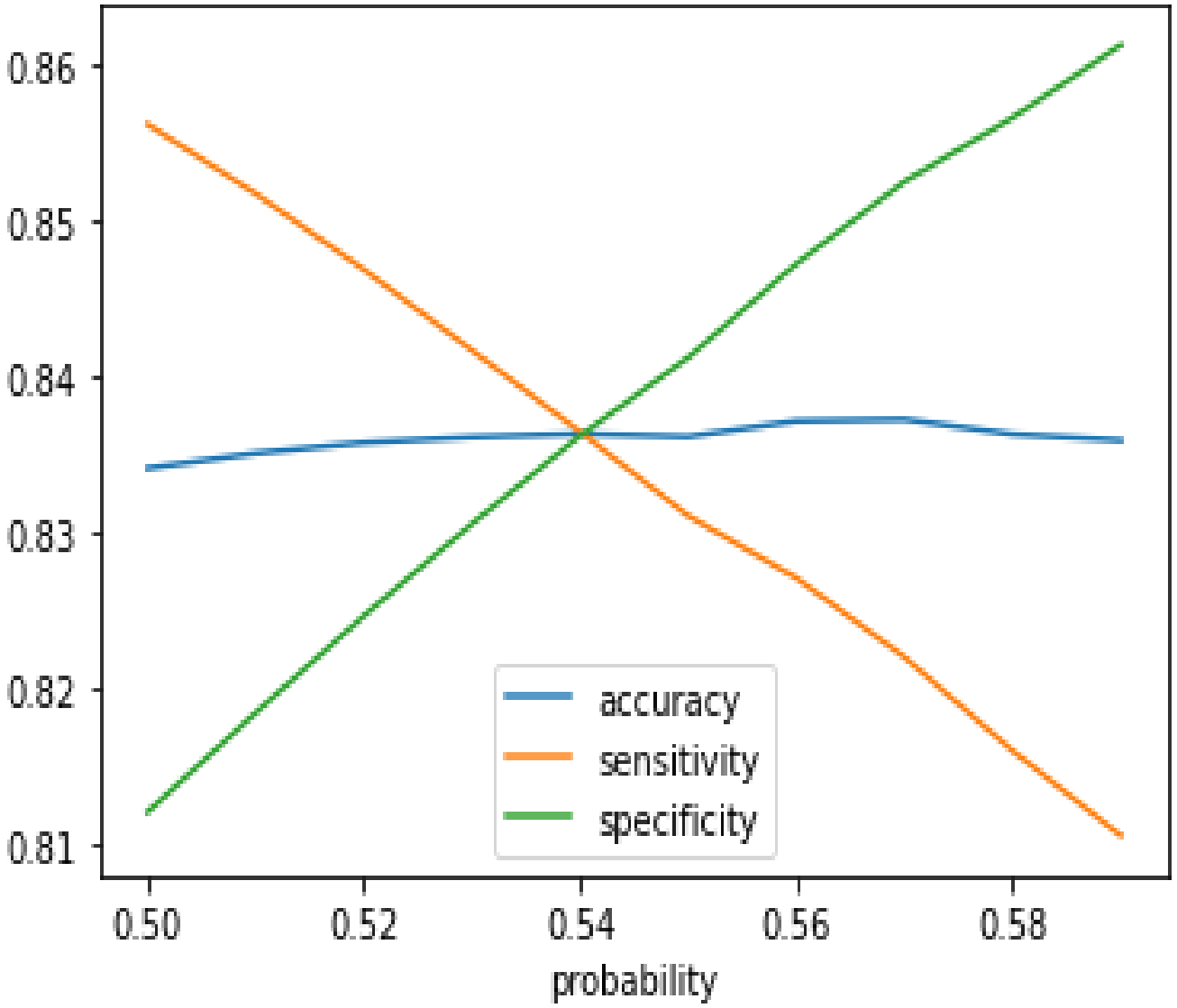



Receiver operating characteristic example
True Positive Rate
False Positive Rate or [1 - True Negative Prediction Rate]
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (area = 0.90)



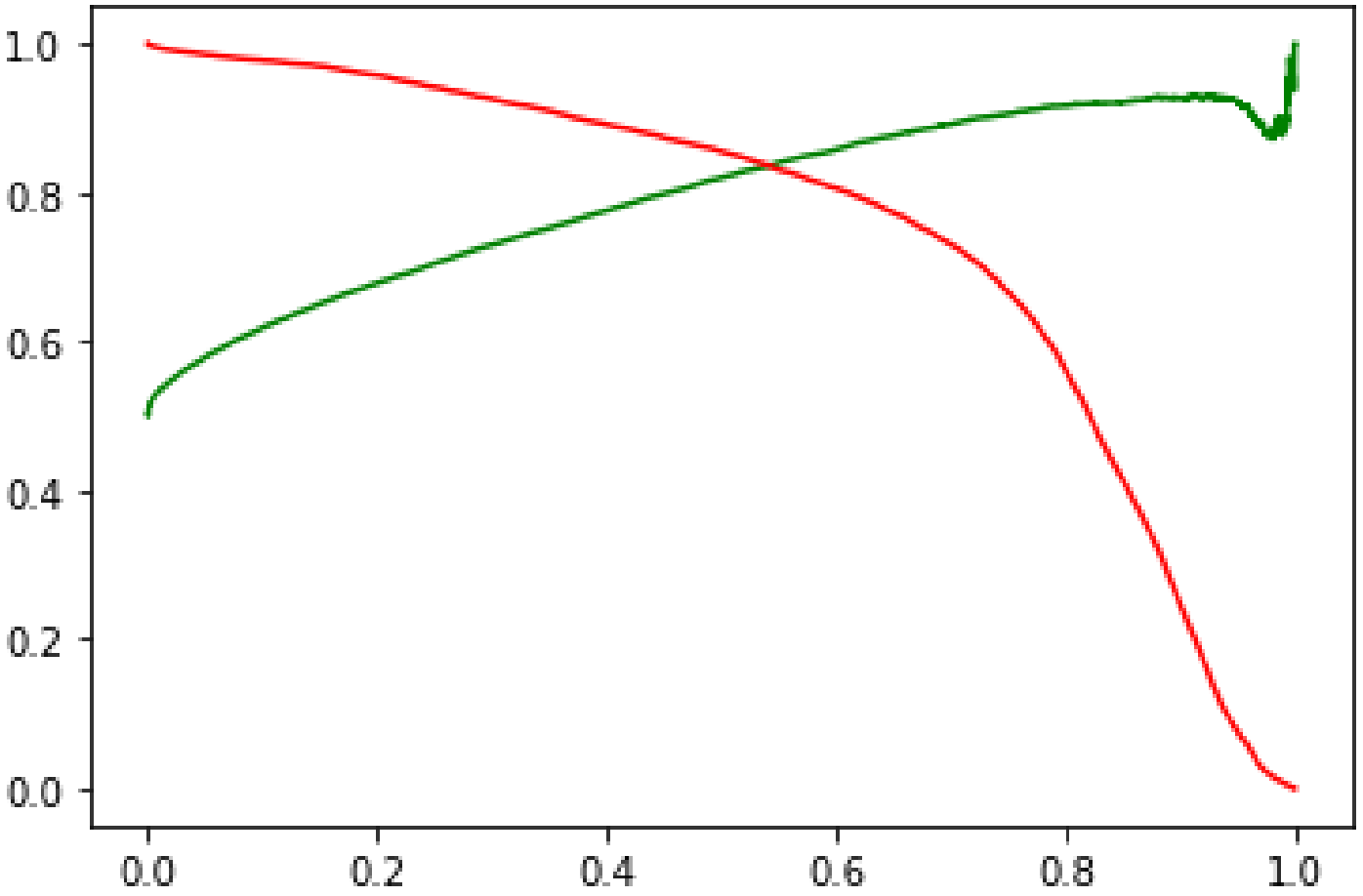
Initially we selected the optimm point of classification as 0.5.

From the above graph, we can see the optimum cutoff is slightly higher than 0.5 but lies lower than 0.6. So lets tweek a little more within this range.



From the above graph we can conclude, the optimal cutoff point in the probability to define the predicted churn variable converges at 0.54

Precision and recall trade-off



Confusion Matrix

[[6860 1412]

[145 584]]

Sensitivity=0.801097393689986 3

Specificity = 0.829303675048355 9

False Positive Rate = 0.170696324951644 1

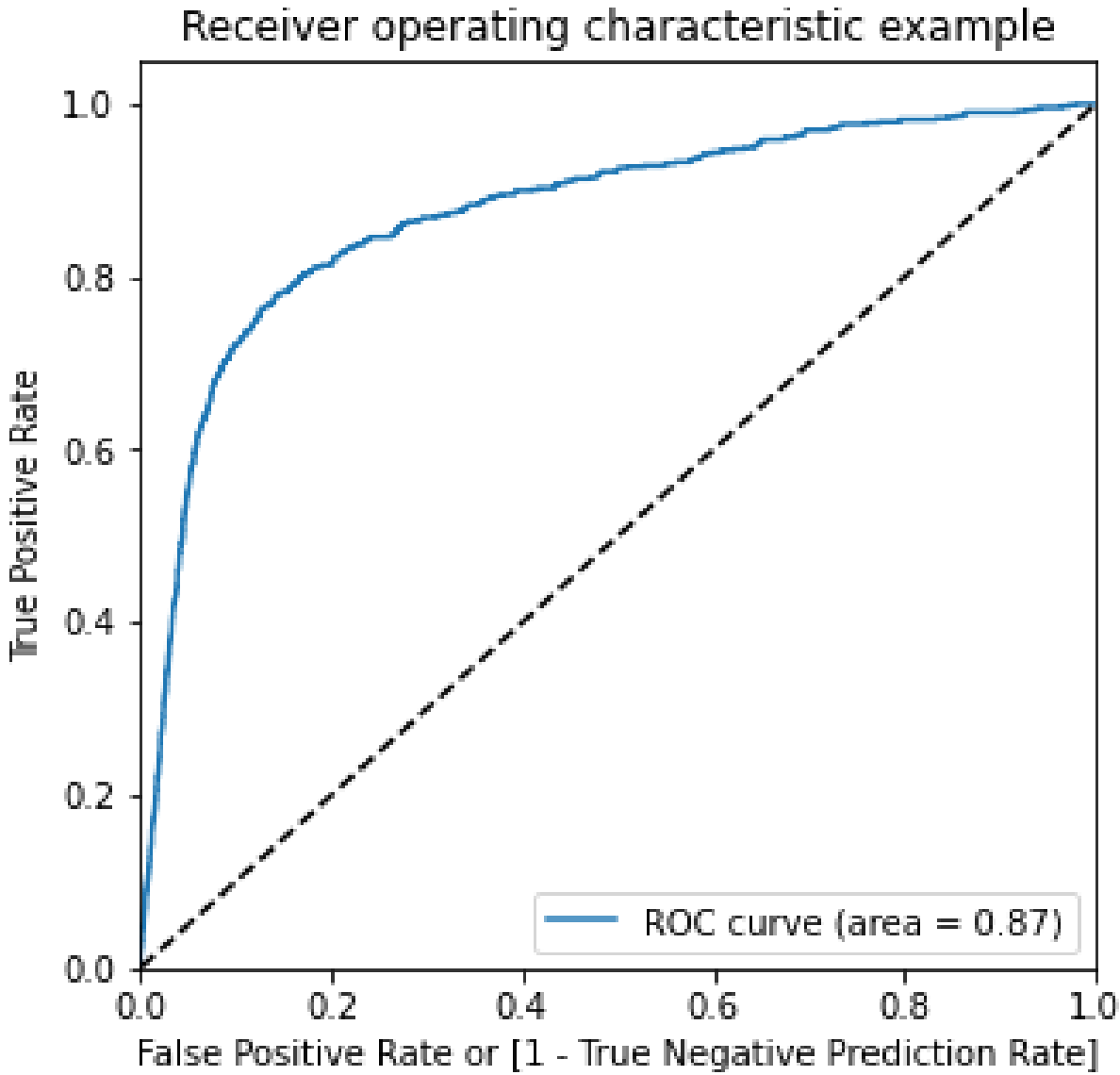
Precision = 0.292585170340681 4 True Negative

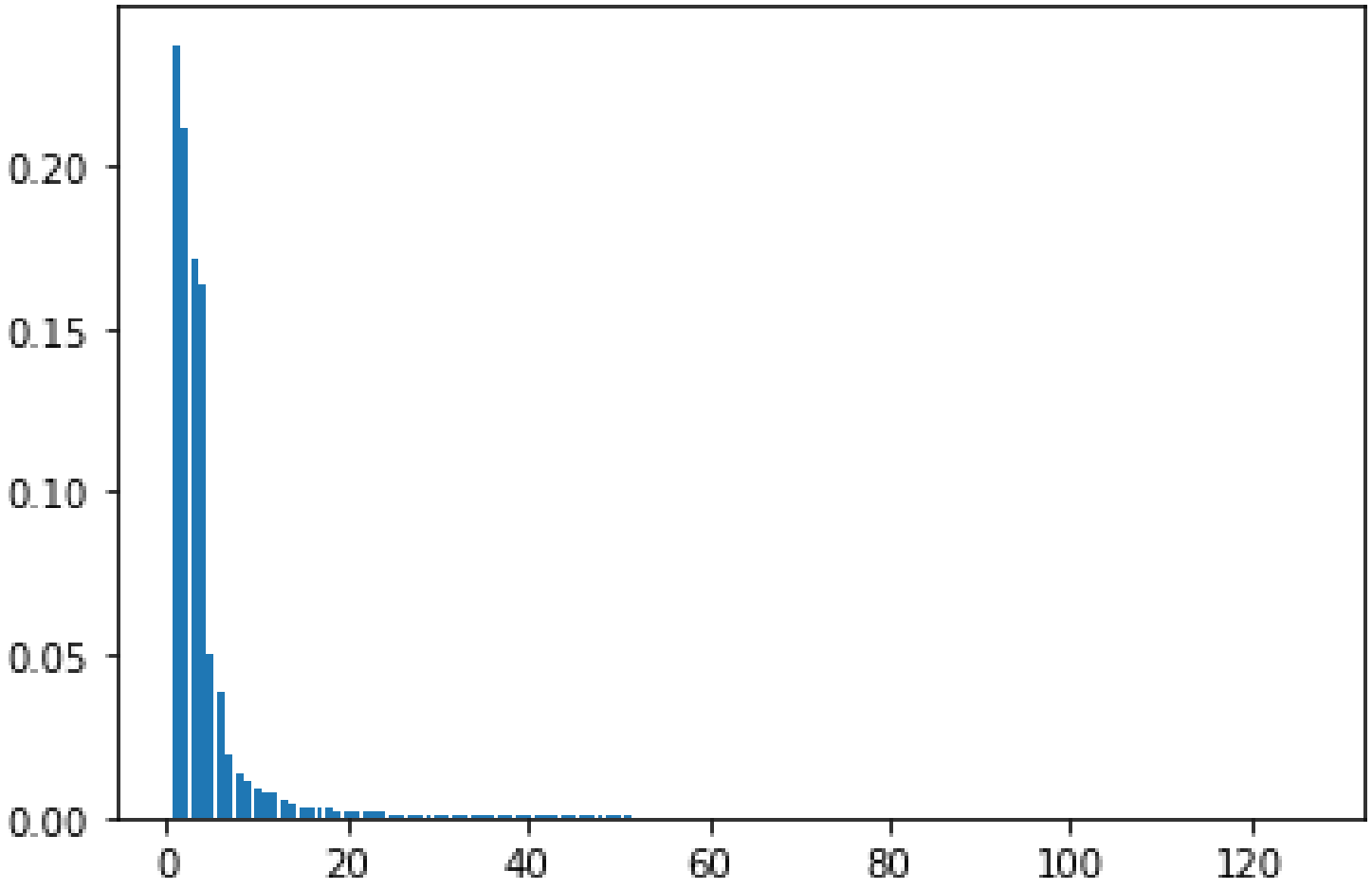
Prediction Rate = 0.979300499643112

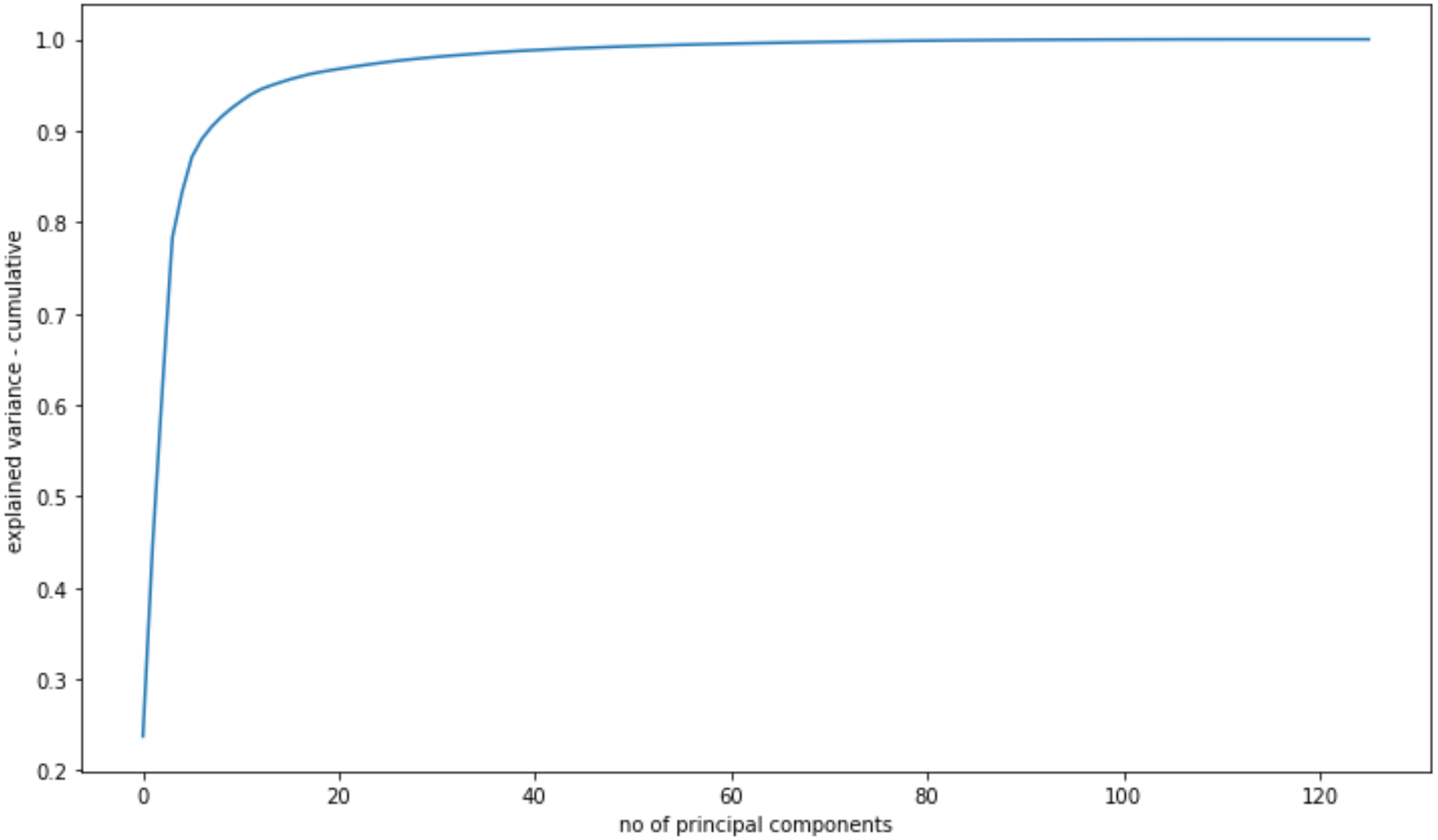
Results

The accuracy of the predicted model is: 83.0 % The sensitivity of the predicted model is: 80.0 % As the model created is based on a sensitivity model, i.e. the True positive rate is given more importance as the actual and prediction of churn by a customer

- The AUC score for train dataset is 0.90 and the test dataset is 0.87.
- This model can be considered as a good model.




0.20
0.15
0.10
0.05
0.00
0
20
40
60
80
100
120



Confusion Matrix for y-test & y - pred

[[6250 2022]
[185 544]]

Accuracy of the logistic regression model with PCA:
0.7548050216642596