TELECOM CHURN CASE STUDY REPORT

HOANG NGOC TIEN – MASTER OF DATA IN DATA SCIENCE PROGRAME

PROBLEM & BUSINESS GOALS

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, and ***retaining high profitable customers is the number one business goal***

BUSINESS GOALS

To reduce customer churn, telecom companies need to **predict which customers are at high risk of churn.**

THE APPROACH

1. Data Understanding and Preparation

- Get familiar with the dataset's structure, variables and their meaning
- Filter high-value customers
- Tagging churn and remove unused columns
- Handle missing values using appropriate methods.

2. Exploratory Data Analysis (EDA)

- Identify outliers and anomalies.
- Examine distribution of features: Analyze the distributions of features with respect to churn status.
- Identify trends: Observe any patterns or trends that might be associated with lead conversion.

THE APPROACH

3. Data Preparation for Model Building

- Feature Engineering
- Splitting the Data into Train and Test sets.
- Class imbalance handling
- Scaling the Data

4. Building Model and Making Predictions

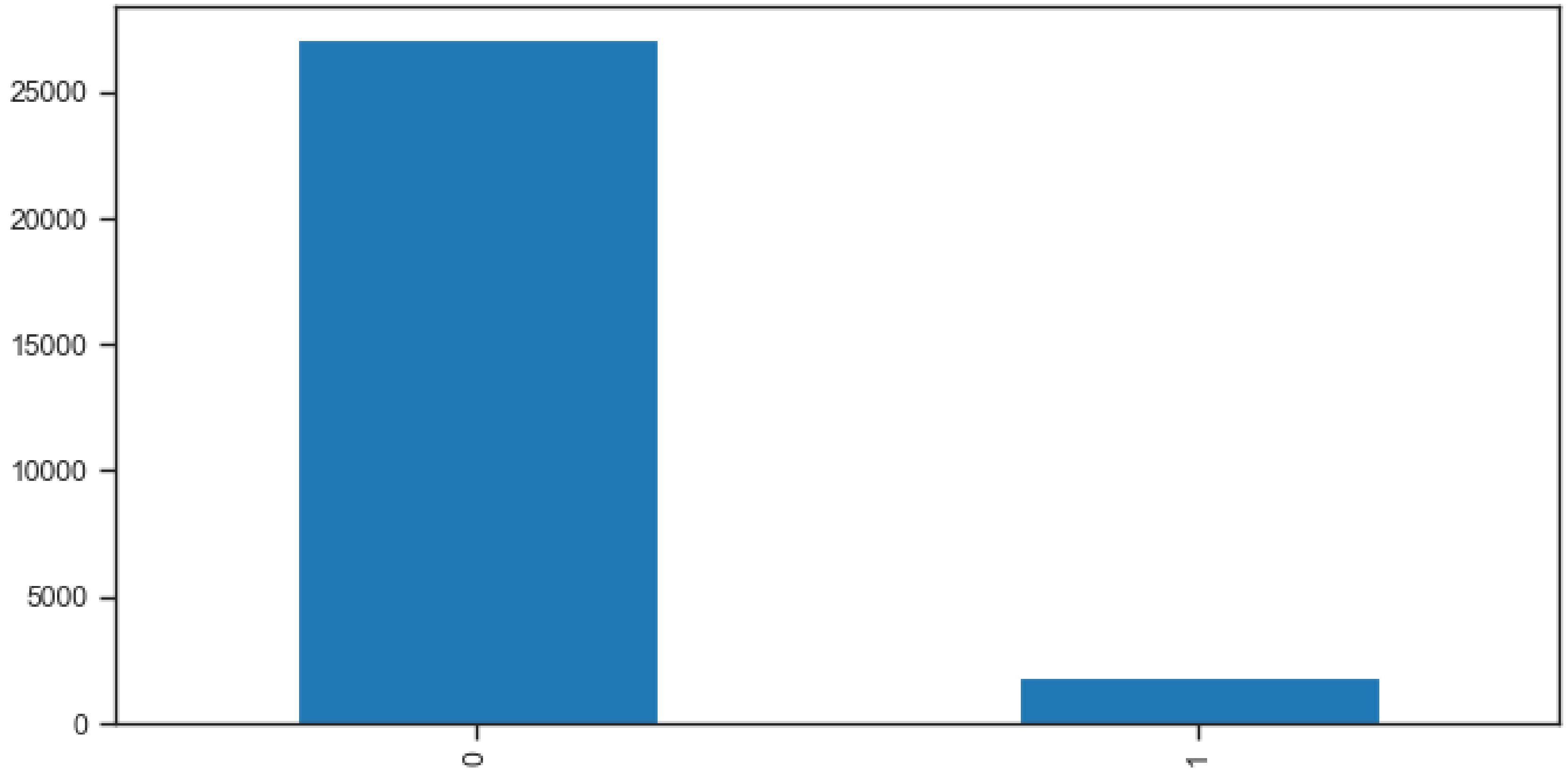
- Building models
- Tuning the models
- Making predictions.

5. Model Evaluation & Interpretation

- Evaluating model.
- Final model Interpretation

1. DATA UNDERSTANDING AND PREPARATION

Customer filtering and tagging churn

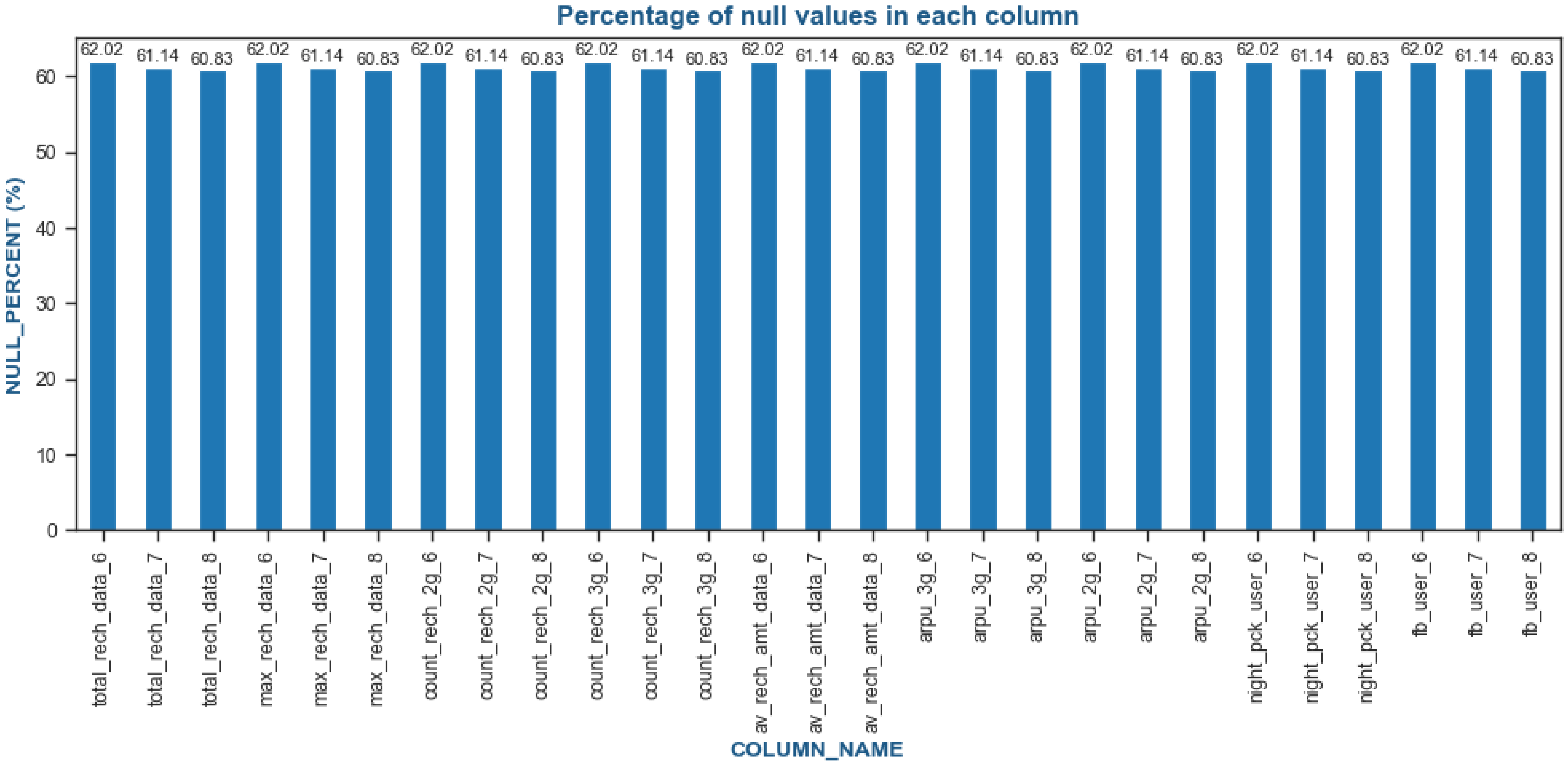


Results:

- Having dataset with 30011 rows after filtering

1. DATA UNDERSTANDING AND PREPARATION

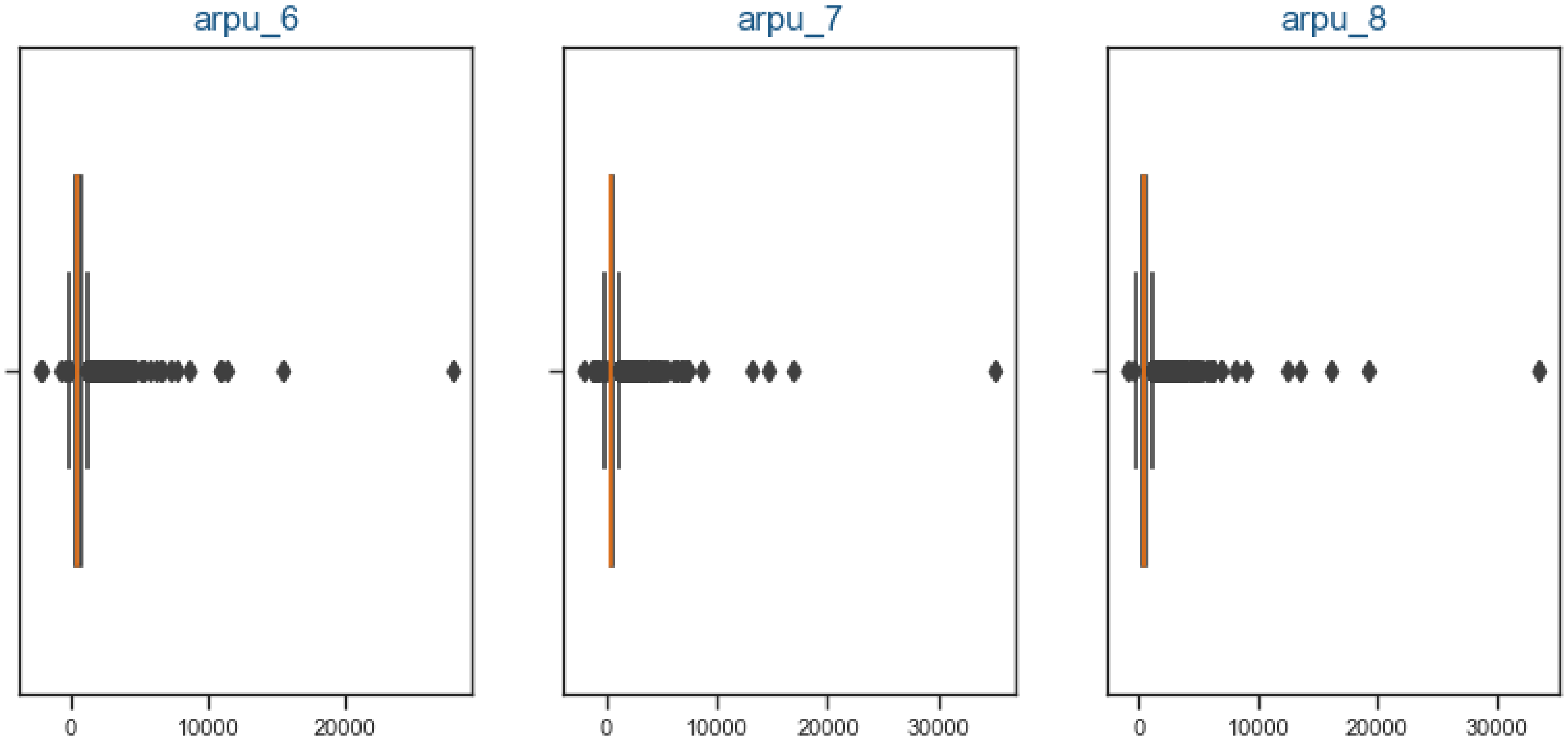
Handling missing values



Methods:

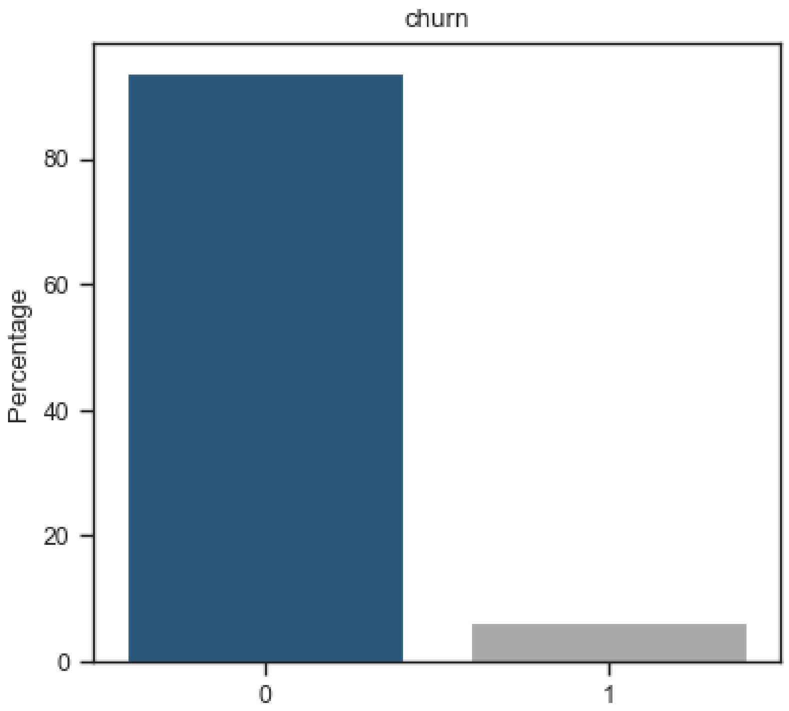
- Drop columns contain >40% missing values
- Remove rows having missing values: if small number of rows.

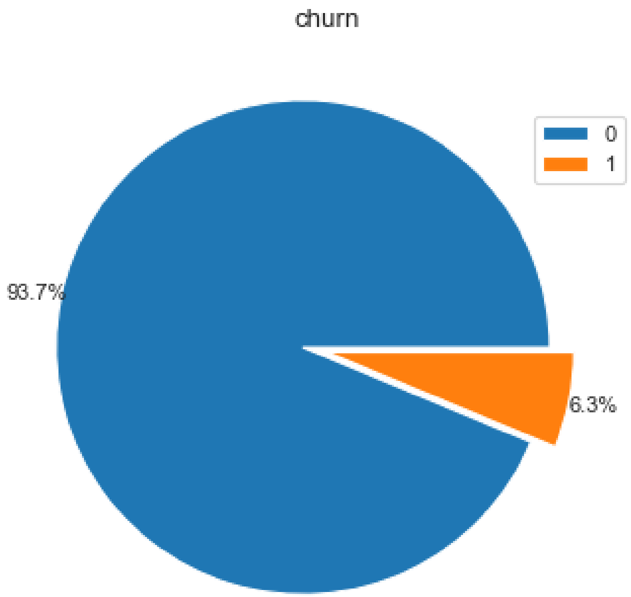
EDA - OUTLIER ANALYSIS



- All most all columns have outliers
- We replace outliers with max and min values of range based on 10th and 90th quantile values

EDA - DATA IMBALANCE



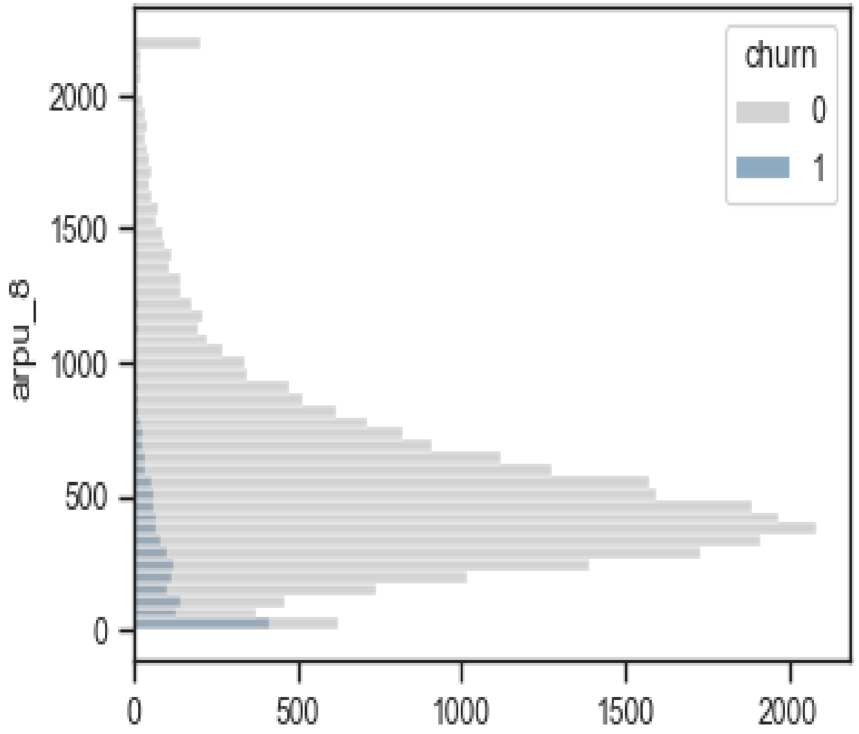


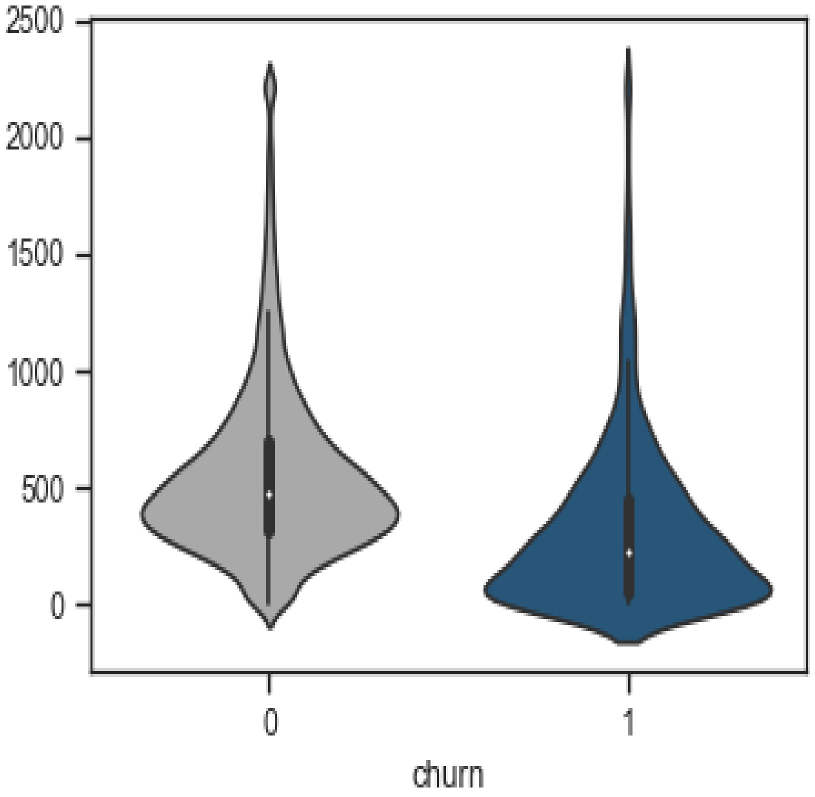
Target Variable:

- 6.3% of observations are churn customers
- The data is highly imbalanced and need to be handled before modeling.

EDA – INSIGHTS OF DATA

Arpu



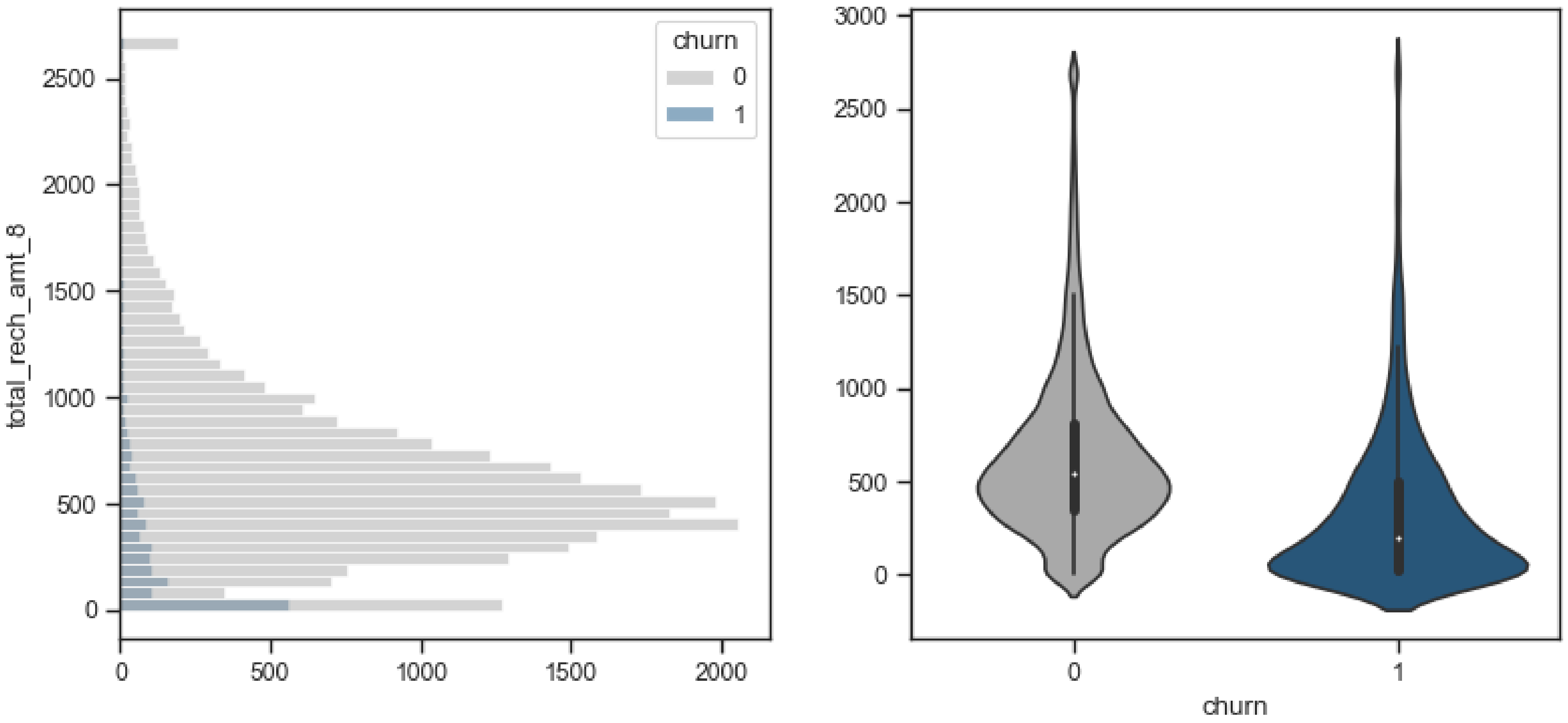


Inference:

- Drop down of arpu of churned customers in action phase

EDA – INSIGHTS OF DATA

Recharge Amount

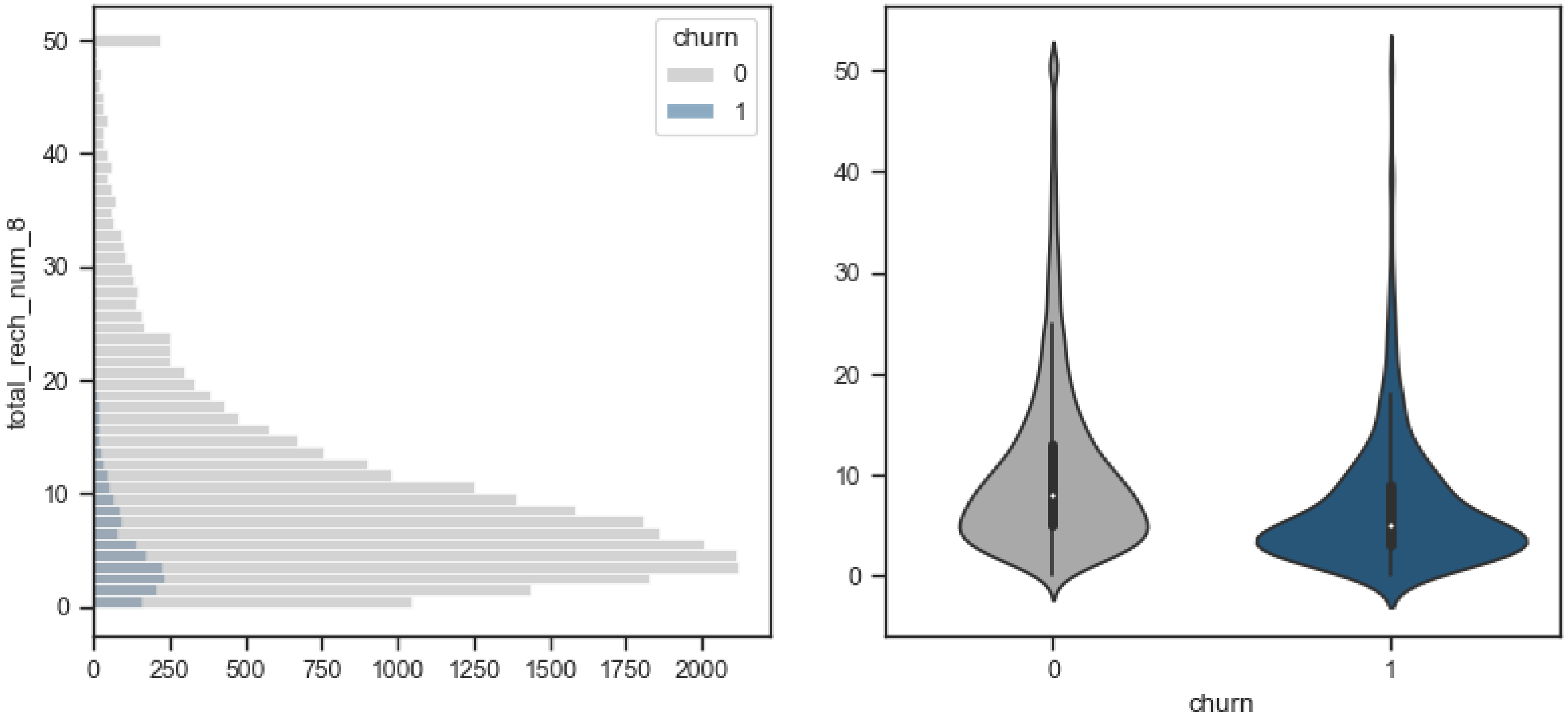


Inference:

- Drop down of recharge of churned customers in action phase

EDA – INSIGHTS OF DATA

Recharge num

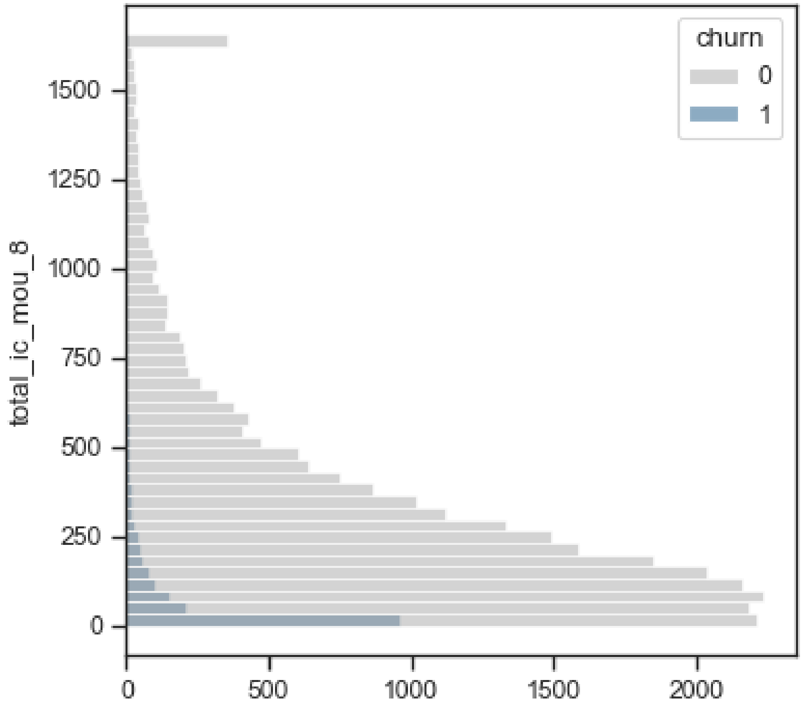


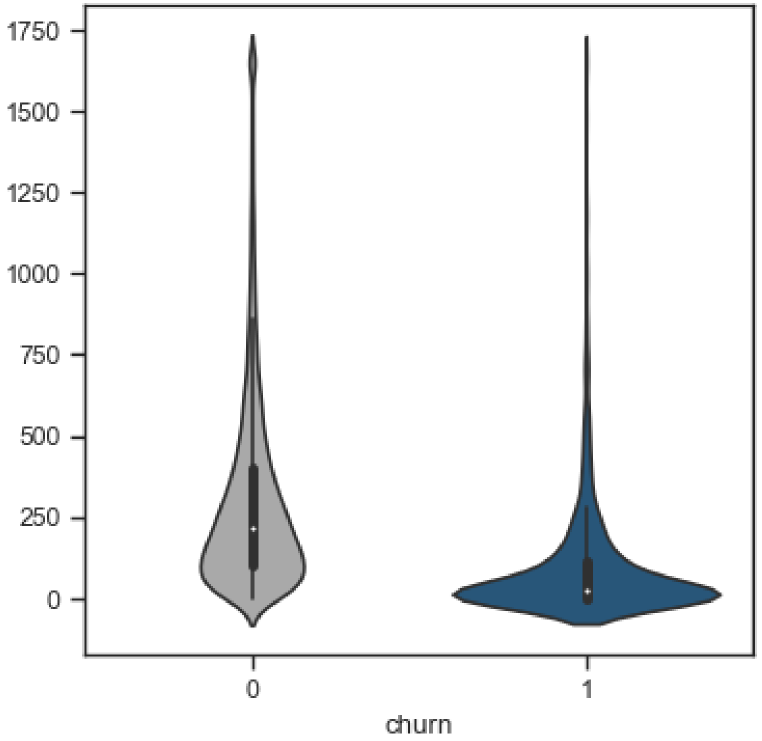
Inference:

- In good phase: recharge num of churn and not churn are about the same. But in action phase recharge num of churn customers drop down in comparation with not churn customers

EDA – INSIGHTS OF DATA

Incomming MOU



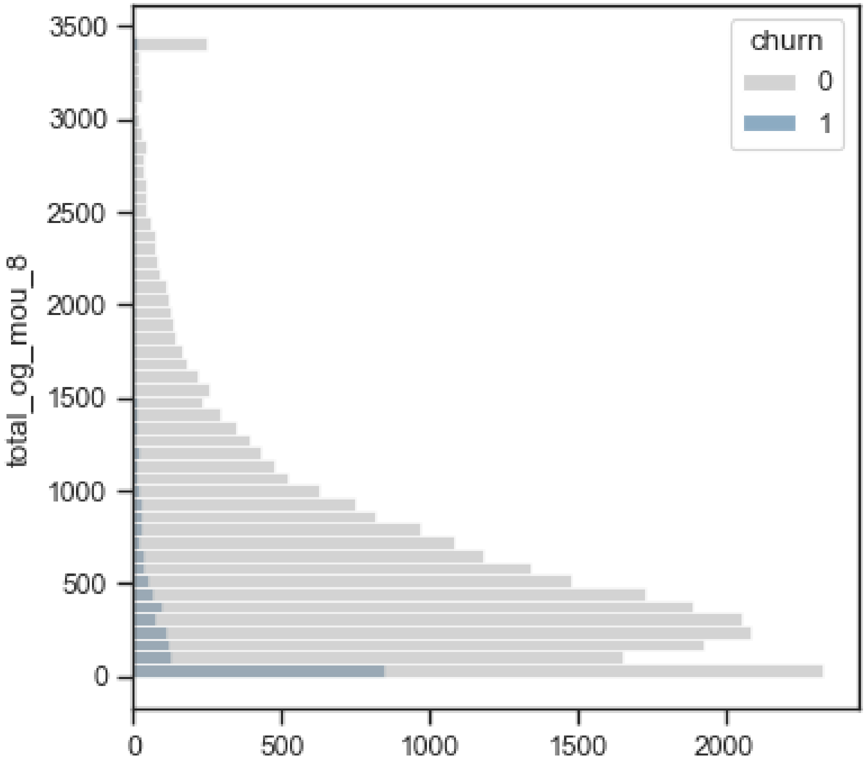


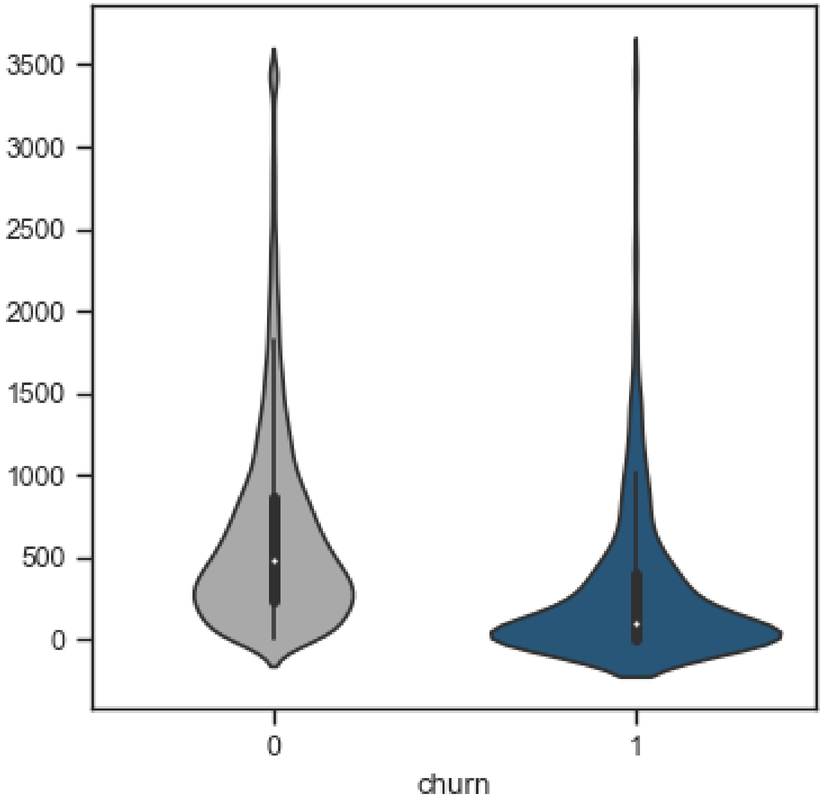
Inference:

- MOU are about the same in good phase.
- MOU of churns are drop down in action phase

EDA – INSIGHTS OF DATA

Outgoing call



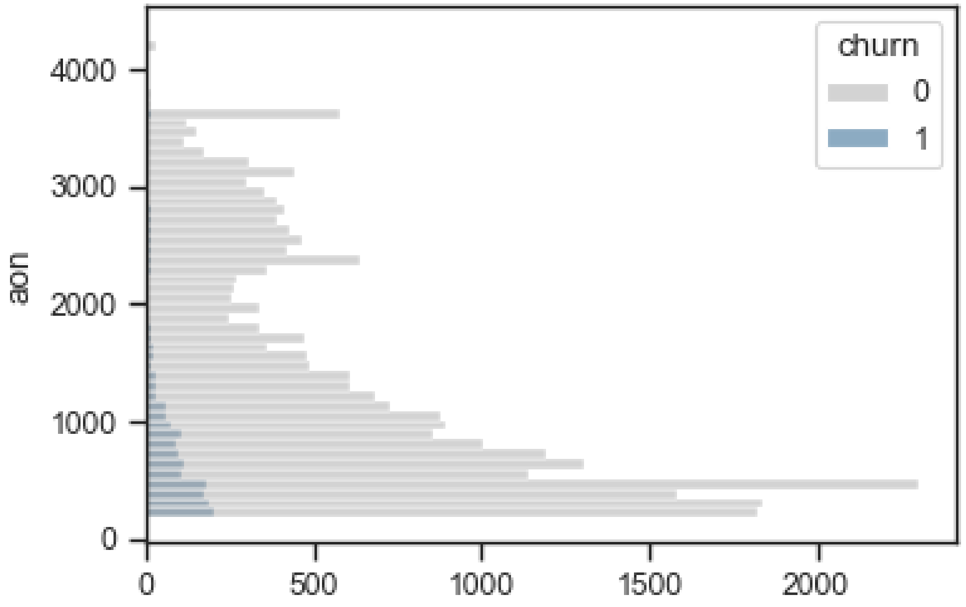


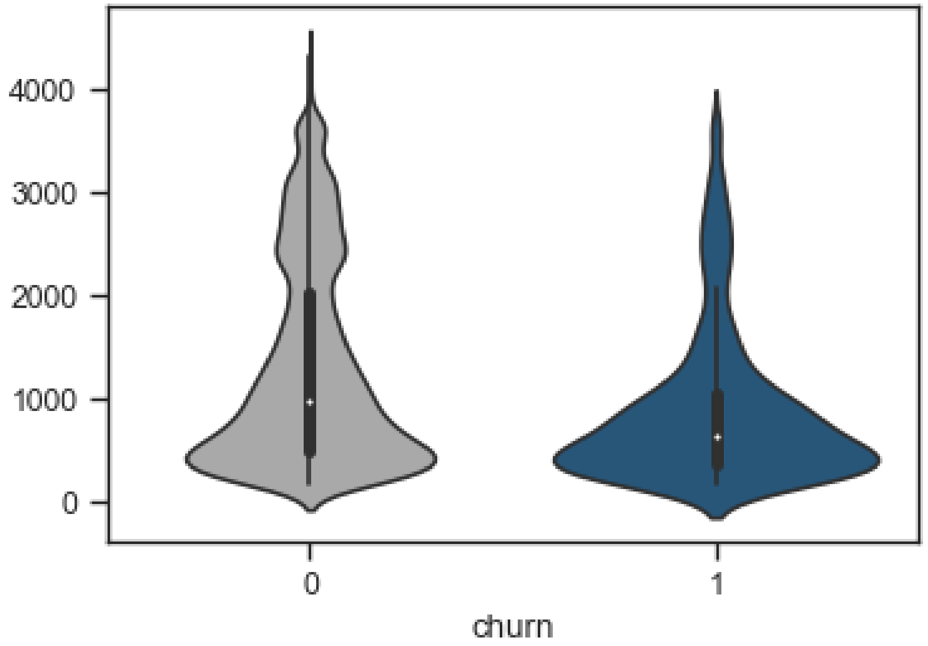
Inference:

- MOU are about the same in good phase.
- MOU of churns are drop down in action phase

EDA – INSIGHTS OF DATA

Age on network





Inference:

- Customers who have smaller AON are more likely to churn

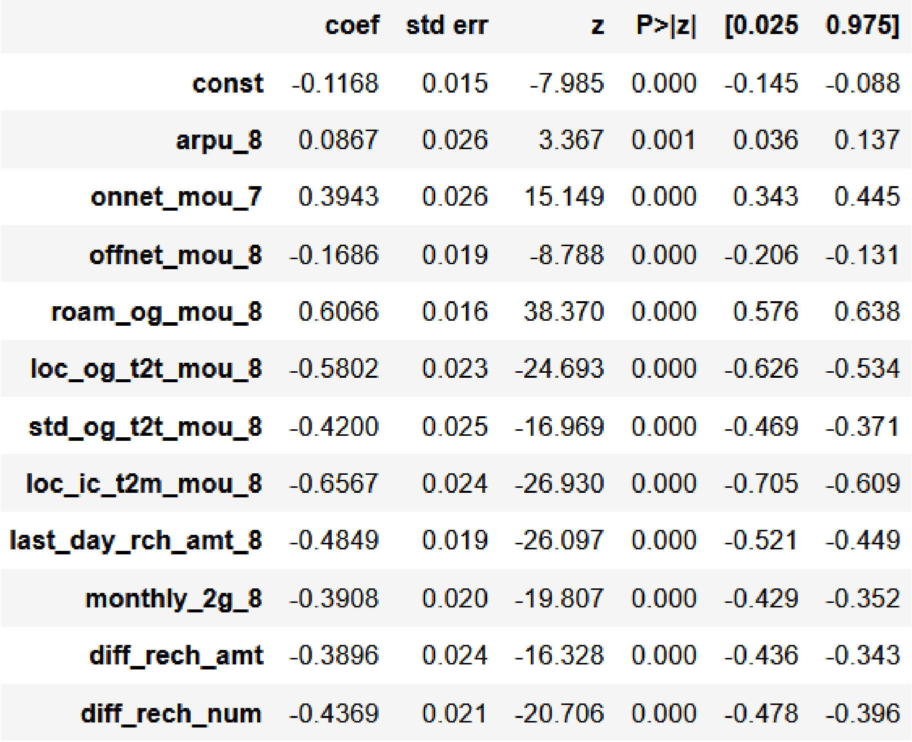
3. DATA PREPARATION FOR MODEL BUILDING

- Feature Engineering: all features are numerical
- Splitting the Data into Train and Test sets.
- Class imbalance handling: Using SMOTE
- Scaling the Data: Using Standardization technique

4. BUILDING MODEL AND MAKING PREDICTIONS

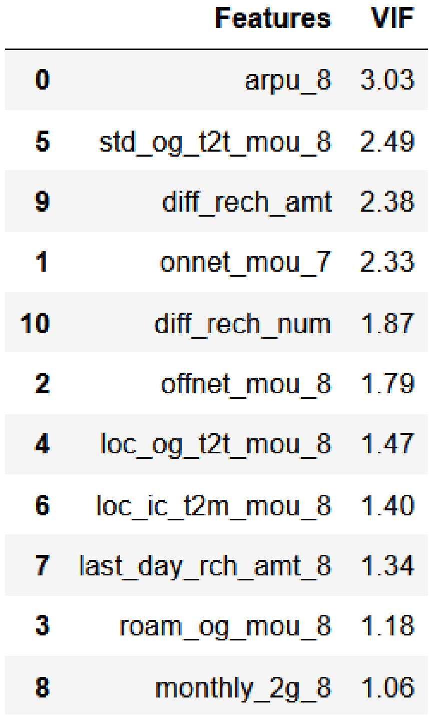
Building Model

Final Model



	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.1168	0.015	-7.985	0.000	-0.145	-0.088
arpu_8	0.0867	0.026	3.367	0.001	0.036	0.137
onnet_mou_7	0.3943	0.026	15.149	0.000	0.343	0.445
offnet_mou_8	-0.1686	0.019	-8.788	0.000	-0.206	-0.131
roam_og_mou_8	0.6066	0.016	38.370	0.000	0.576	0.638
loc_og_t2t_mou_8	-0.5802	0.023	-24.693	0.000	-0.626	-0.534
std_og_t2t_mou_8	-0.4200	0.025	-16.969	0.000	-0.469	-0.371
loc_ic_t2m_mou_8	-0.6567	0.024	-26.930	0.000	-0.705	-0.609
last_day_rch_amt_8	-0.4849	0.019	-26.097	0.000	-0.521	-0.449
monthly_2g_8	-0.3908	0.020	-19.807	0.000	-0.429	-0.352
diff_rech_amt	-0.3896	0.024	-16.328	0.000	-0.436	-0.343
diff_rech_num	-0.4369	0.021	-20.706	0.000	-0.478	-0.396

VIFs

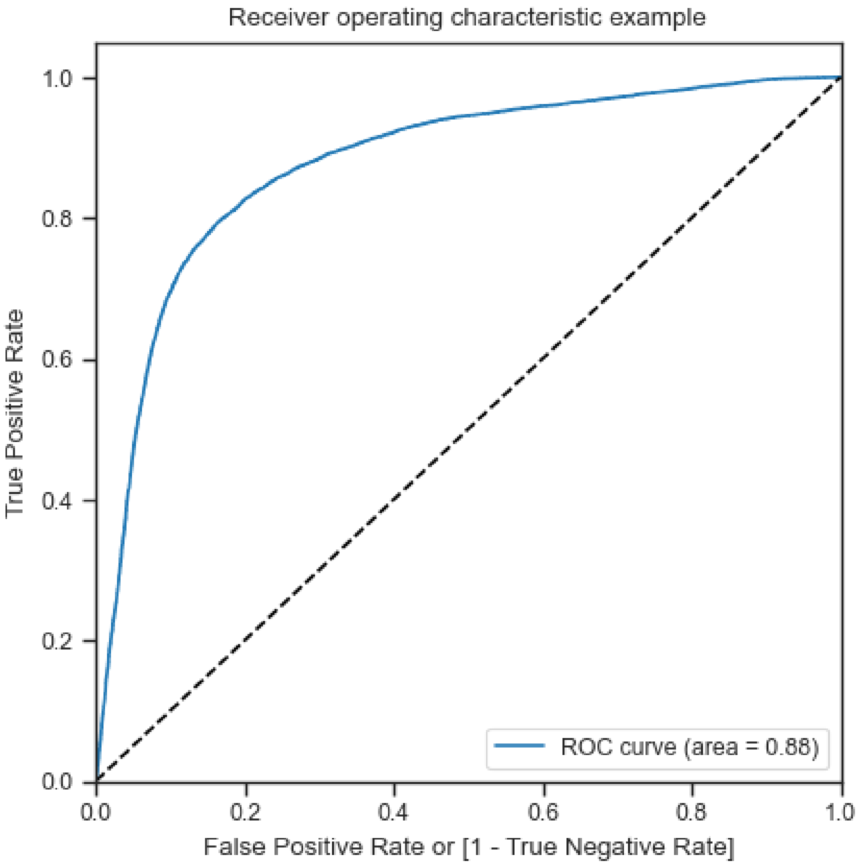


	Features	VIF
0	arpu_8	3.03
5	std_og_t2t_mou_8	2.49
9	diff_rech_amt	2.38
1	onnet_mou_7	2.33
10	diff_rech_num	1.87
2	offnet_mou_8	1.79
4	loc_og_t2t_mou_8	1.47
6	loc_ic_t2m_mou_8	1.40
7	last_day_rch_amt_8	1.34
3	roam_og_mou_8	1.18
8	monthly_2g_8	1.06

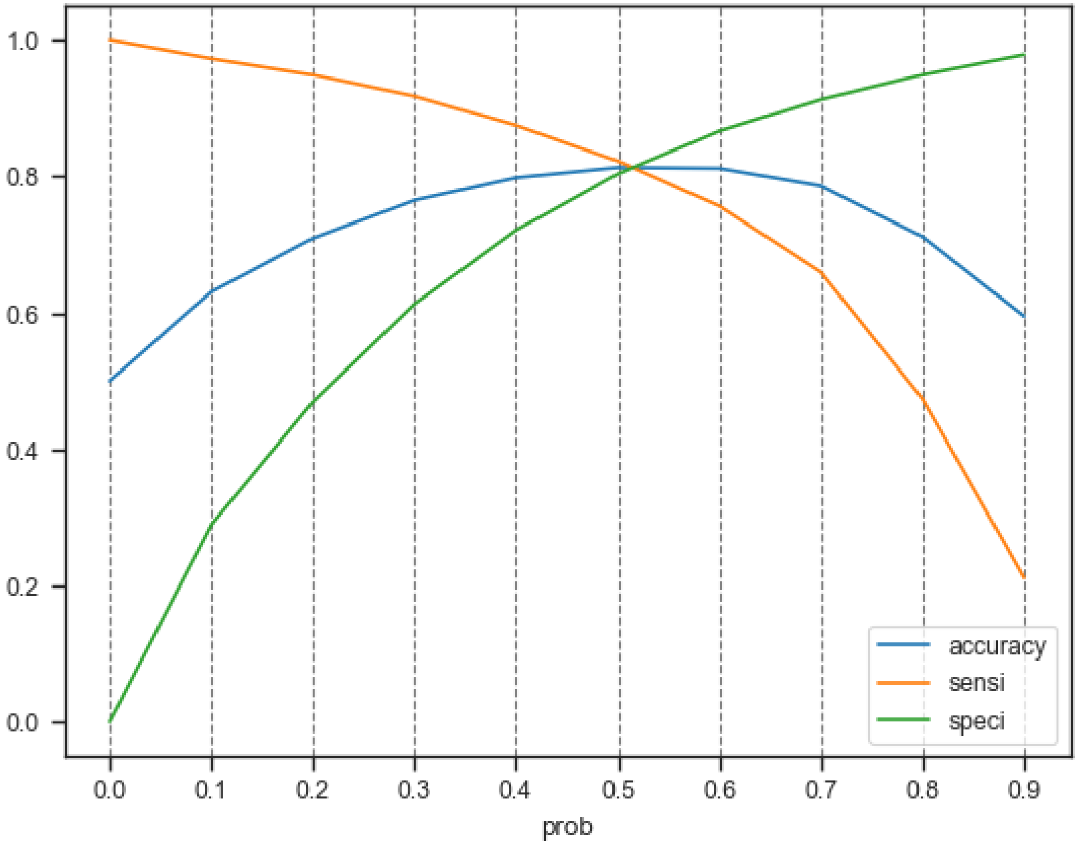
4. BUILDING MODEL AND MAKING PREDICTIONS

Making Predictions

ROC Curve

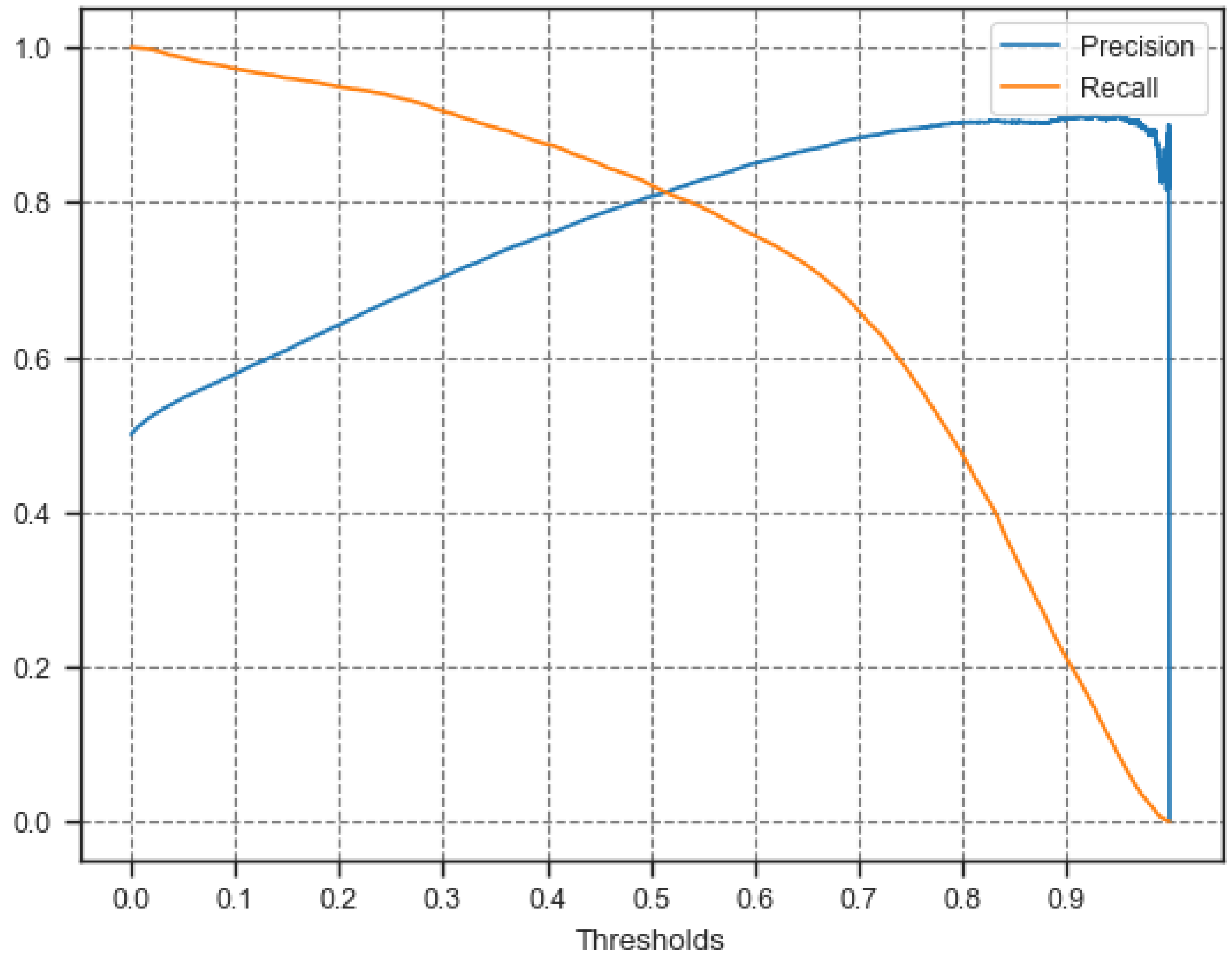


Optimal Cut – off Point: 0.52



4. BUILDING MODEL AND MAKING PREDICTIONS

Making Predictions



Precision and Recall Tradeoff:

- Precision increases, Recall decrease
- The balance point: 0.52

For the prolem:

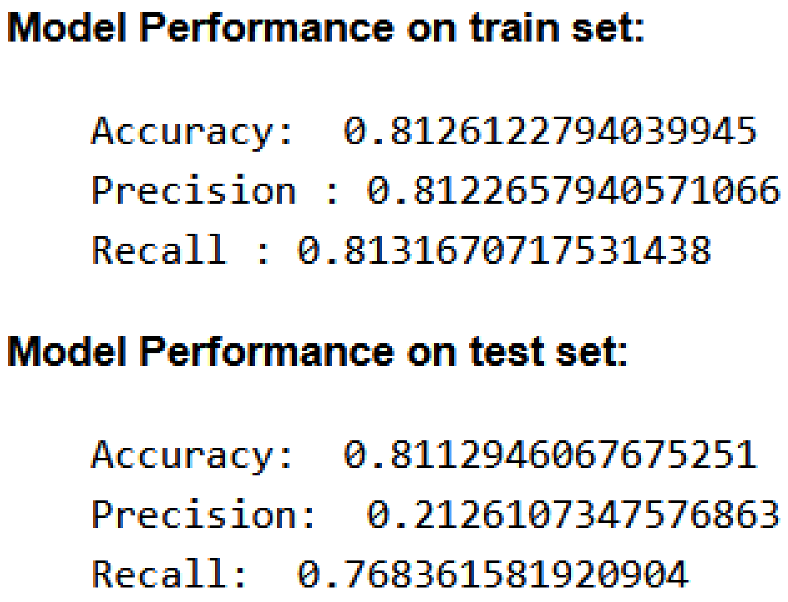
- We need high Recall because the main objective is customer retention => we can choose cutoff point <0.52.

5. MODEL EVALUATION AND INTERPRETATION

- Model Evaluation: based on recall metric
- Model Interpretation.

5. MODEL EVALUATION AND INTERPRETATION

Model Evaluation



Model Performance on train set:

```
Accuracy:  0.8126122794039945
Precision : 0.8122657940571066
Recall : 0.8131670717531438
```

Model Performance on test set:

```
Accuracy:  0.8112946067675251
Precision:  0.212610734757686З
Recall:  0.768361581920904
```

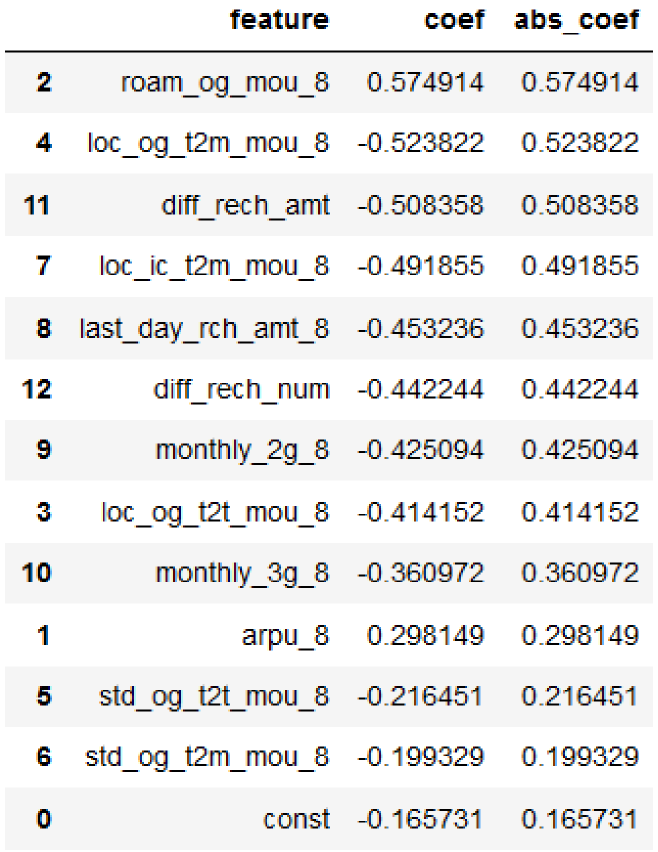
With optimal cut-off point:

- The model performance is still good on test datasets
- The model is good in accuracy
- The model is stable

5. MODEL EVALUATION AND INTERPRETATION

Model Interpretation

Feature coefs



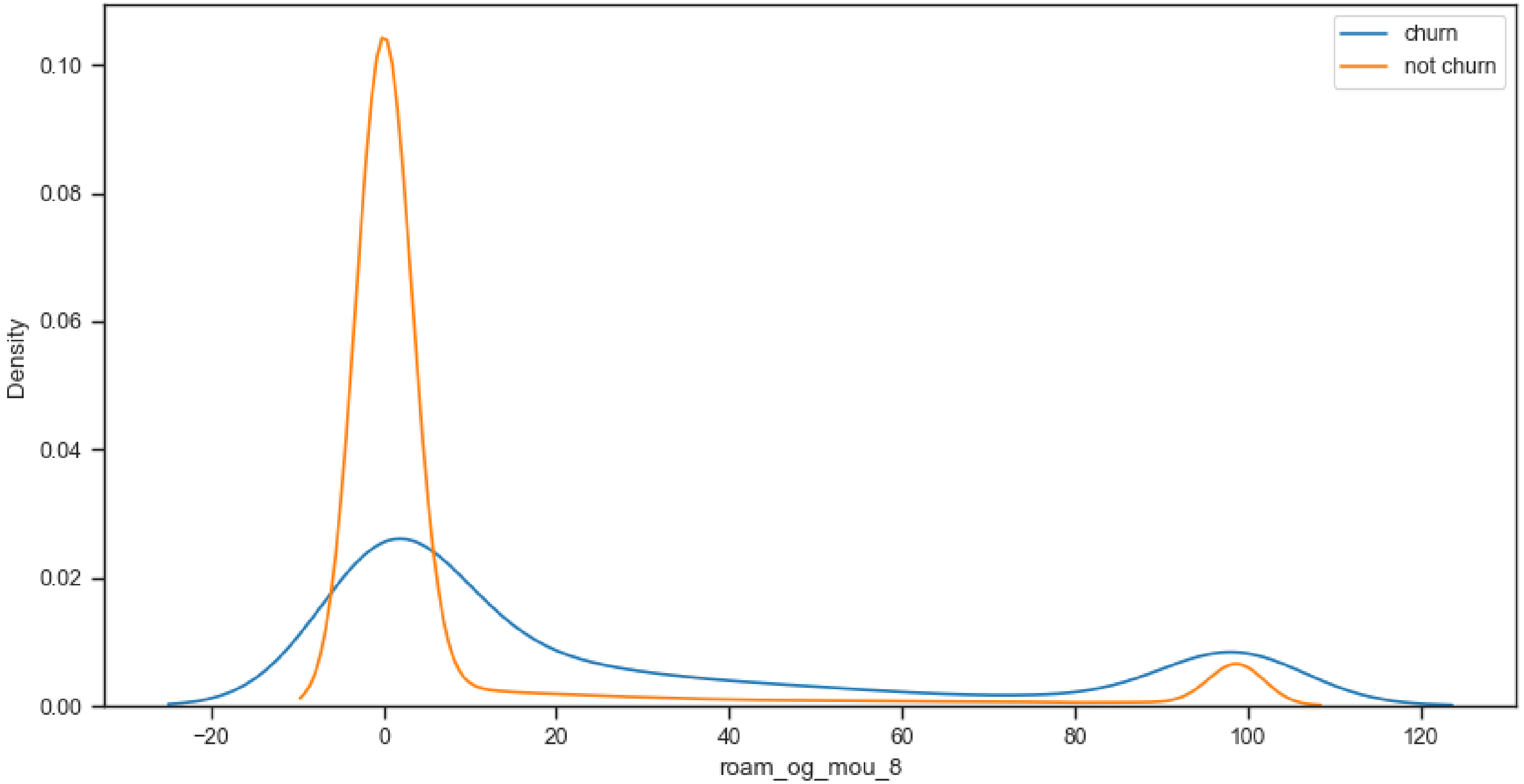
	feature	coef	abs_coef
2	roam_og_mou_8	0.574914	0.574914
4	loc_og_t2m_mou_8	-0.523822	0.523822
11	diff_rech_amt	-0.508358	0.508358
7	loc_ic_t2m_mou_8	-0.491855	0.491855
8	last_day_rch_amt_8	-0.453236	0.453236
12	diff_rech_num	-0.442244	0.442244
9	monthly_2g_8	-0.425094	0.425094
3	loc_og_t2t_mou_8	-0.414152	0.414152
10	monthly_3g_8	-0.360972	0.360972
1	arpu_8	0.298149	0.298149
5	std_og_t2t_mou_8	-0.216451	0.216451
6	std_og_t2m_mou_8	-0.199329	0.199329
0	const	-0.165731	0.165731

Some most important features:

- roam_og_mou_8
- loc_og_t2m_mou_8
- diff_rech_amt
- loc_ic_t2m_mou_8
- last_day_rch_amt_8

5. MODEL EVALUATION AND INTERPRETATION

Model Interpretation

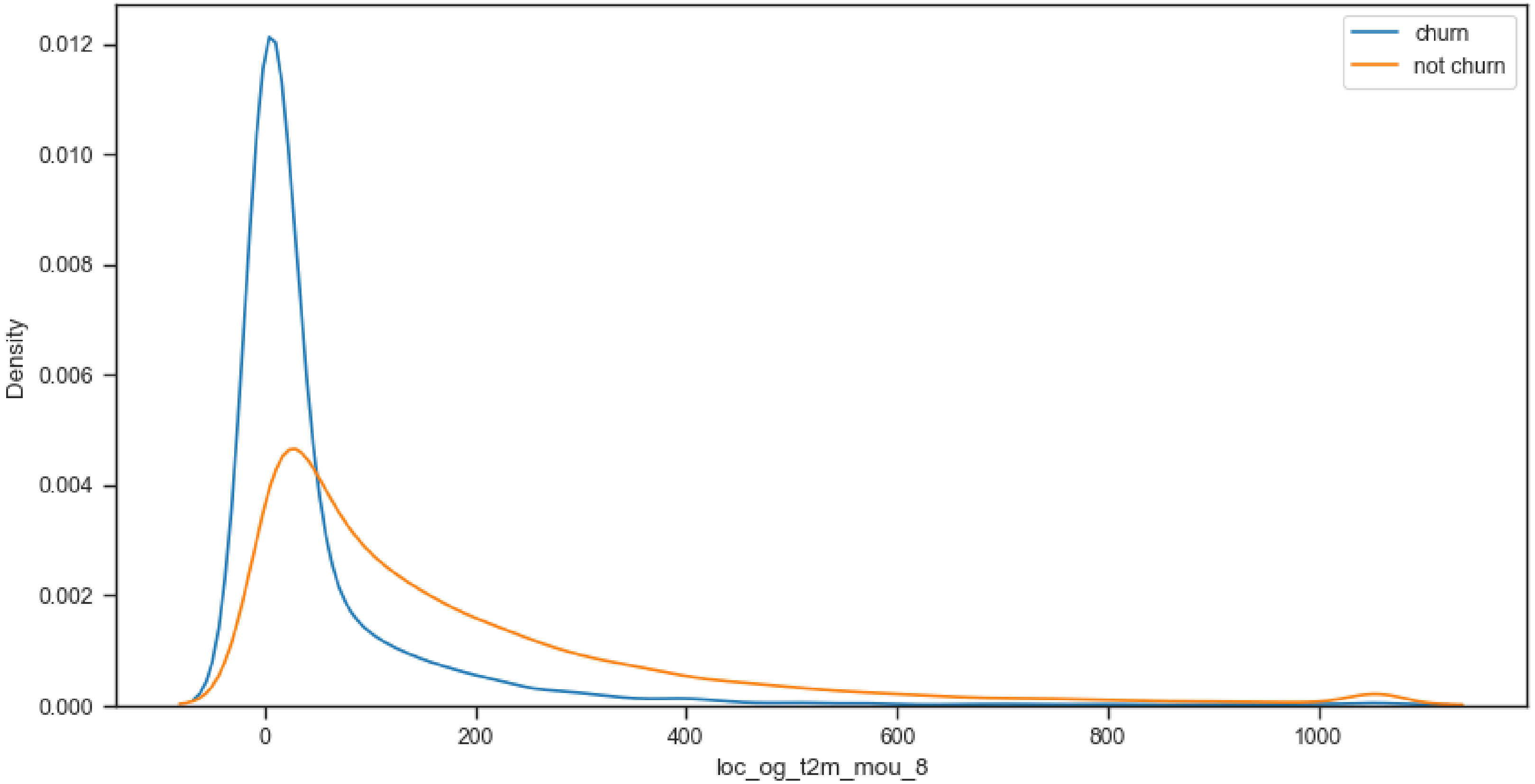


Inference:

- Churn customers have more MOU of roaming

5. MODEL EVALUATION AND INTERPRETATION

Model Interpretation

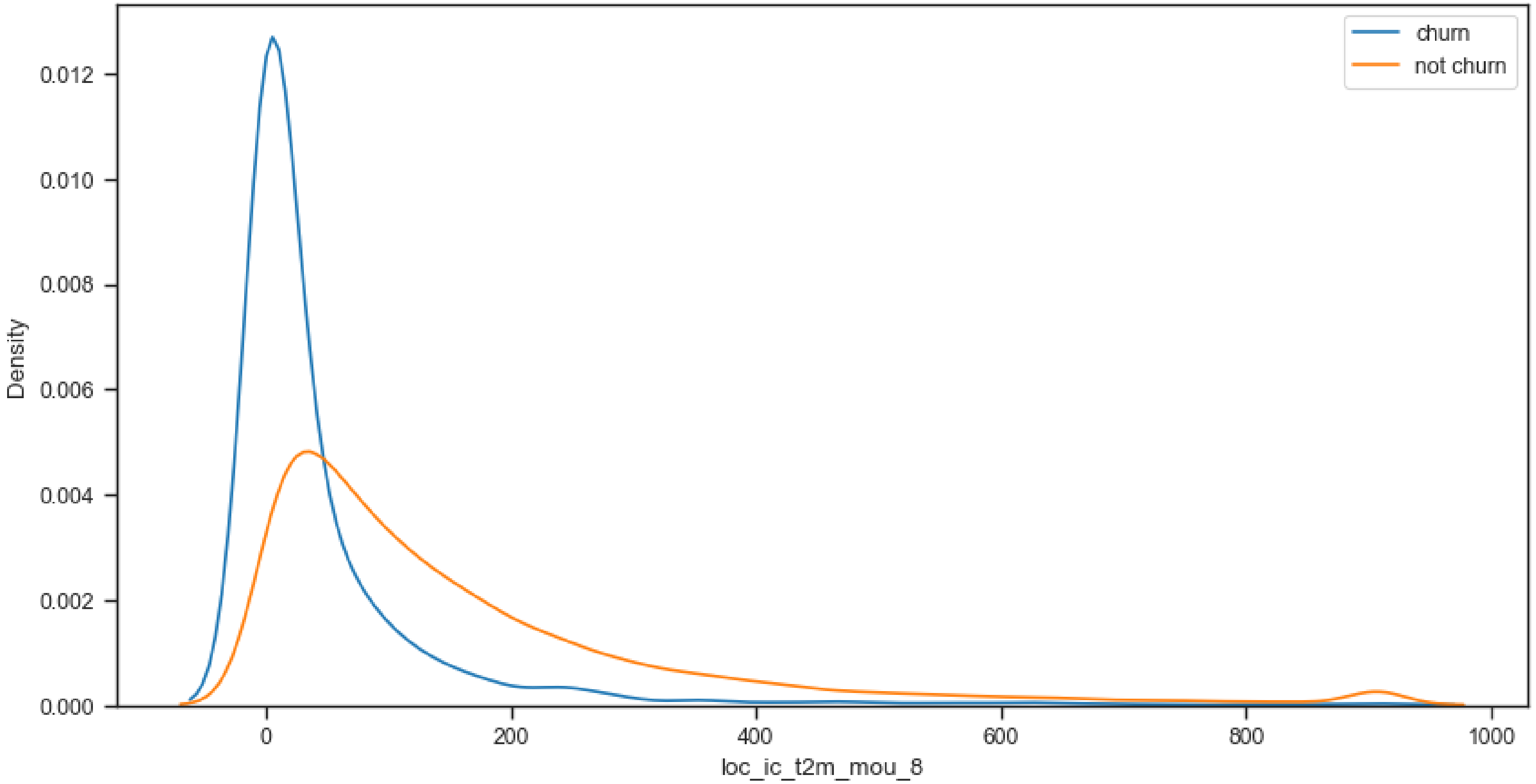


Inference:

- Almost all churn customers have values around 0, where not churn customers have values >0

5. MODEL EVALUATION AND INTERPRETATION

Model Interpretation

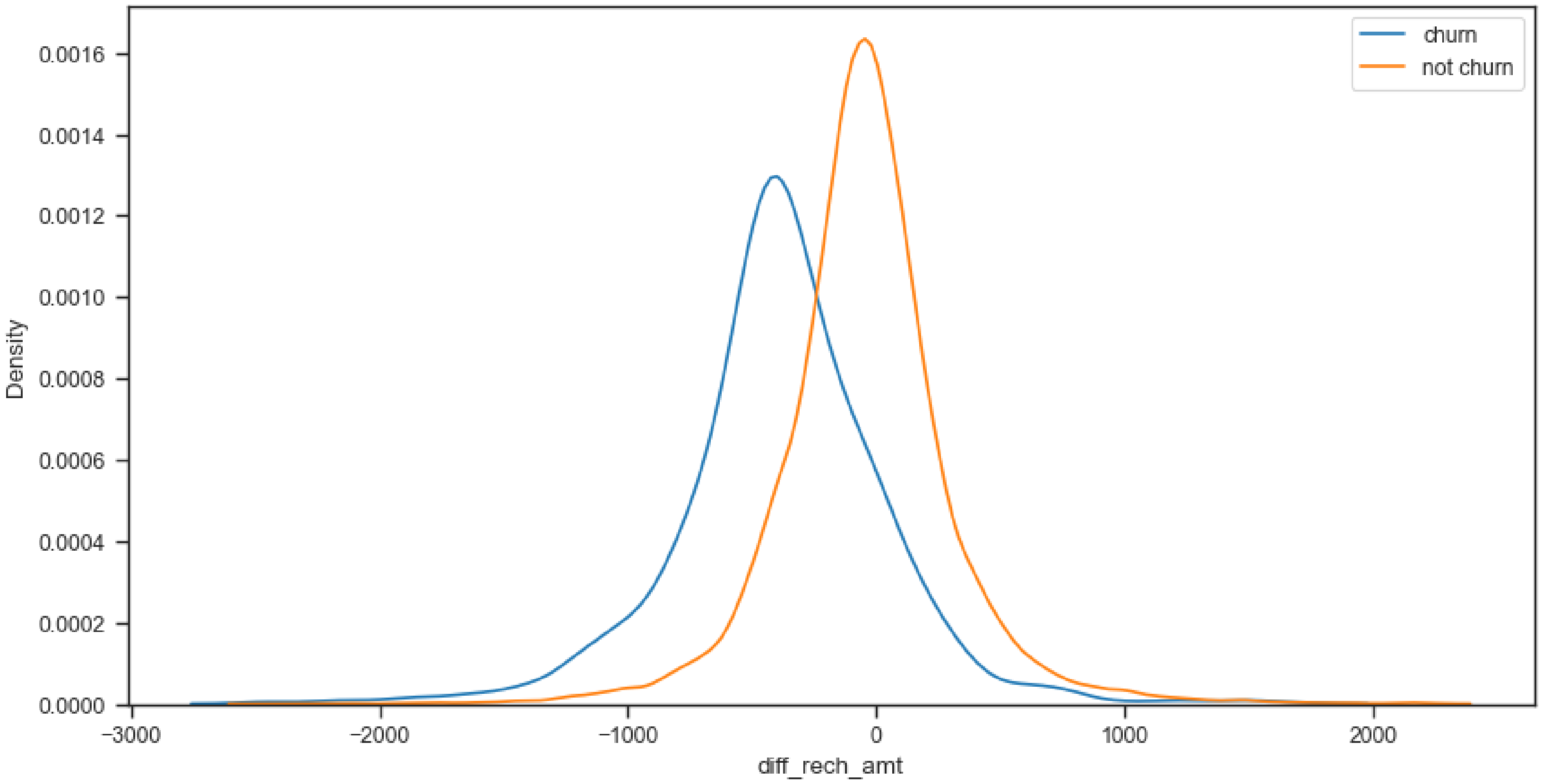


Inference:

- Almost all churn customers have values around 0, where not churn customers have values >0

5. MODEL EVALUATION AND INTERPRETATION

Model Interpretation



Inference:

- Almost all churn customers have values <0, that means dropping down of recharge amount in action phase in comparation with good phase

RECOMMENDATION

- The company should focus on **customer who have roaming call** in action phase as they are more likely to churn. There may be some reasons like quality of roaming service, the price too high,...
- The company should focus on customers who have **less MOU in local t2m outgoing/incomming call** in action phase. The dropping down of MOU may lead to churn.
- The company should focus on customers who have **less MOU in standard outgoing call** in action phase. The dropping down of MOU may be the sign of churn.
- The company should focus on customers who have **less use of monthly 2g and 3g** in action phase. The dropping down of number may be the sign churn.
- The company should focus on customers who have drastically **dropped down in recharge amount/recharge number from good phase to action phase**, since it's the sign of churn.
- The company should focus on customers who have **last recharge with small amount** in action phase in comparation to others, since it's the sign of churn.