Customers churn prediction

in the telecommunications company Vodafone

Yaroslav Bezrukavyi

BigDataLab

Problematics

Reduction of Income

Customers who leave the company **no longer generate revenue** for it. This can significantly impact the company's financial performance.

Increased Costs

According to various studies, acquiring new customers can be **5 times** more expensive than retaining existing ones.

Negative reputation

A high churn rate can **signal problems with the services** provided by the company, negatively impacting its reputation.

Assumption

Retention Cost

5 USD

The company's costs for retaining **existing customers**

Acquisition Cost

25 USD

The company's costs for acquiring **new customers**

Given that acquiring new customers is **5 times** more expensive than retaining existing ones

Goals

Prediction Model

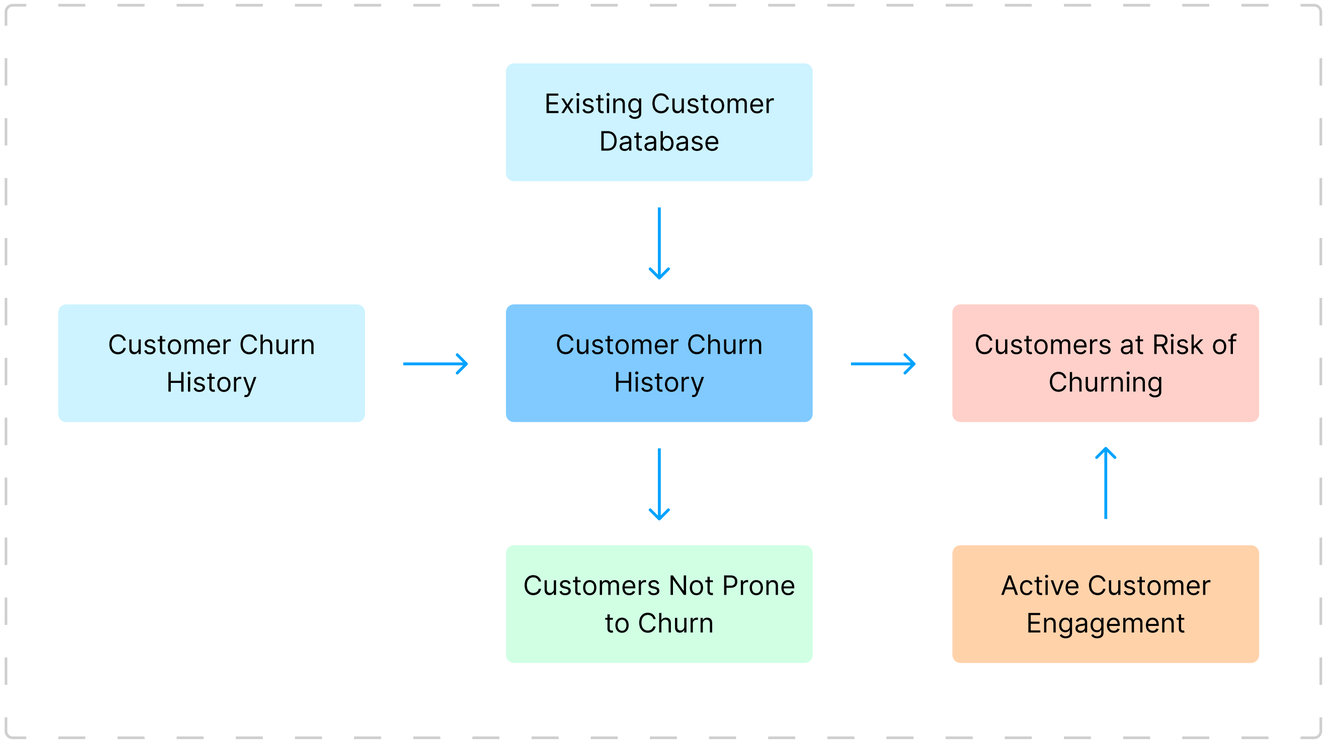
Development and implementation of a machine learning model for predicting customer churn probability

Churn Analysis

Analyze the obtained data and identify the factors influencing customer churn

Minimizing Company Costs

Use the obtained information to conduct more successful engagement campaigns with customers prone to churn and to optimize the company's service delivery processes



Initial Data

🔑 abon_id

FE - customers profile

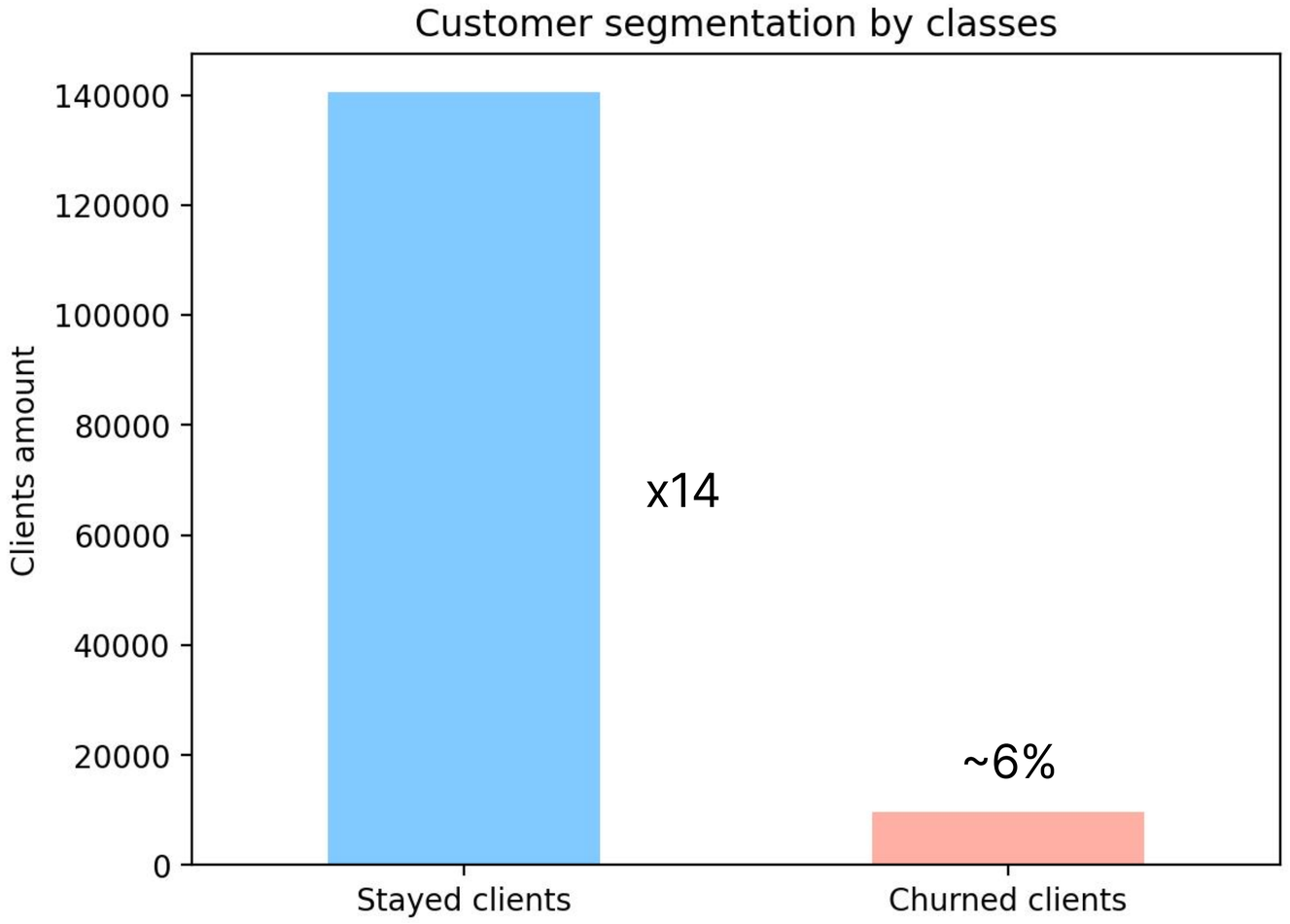
2x **150k** records
817 attributes

B_NUM - customers activity with other numbers

2x ~**669k** records
8 attributes

DPI - customers traffic by applications

2x ~**6.7m** records
6 attributes



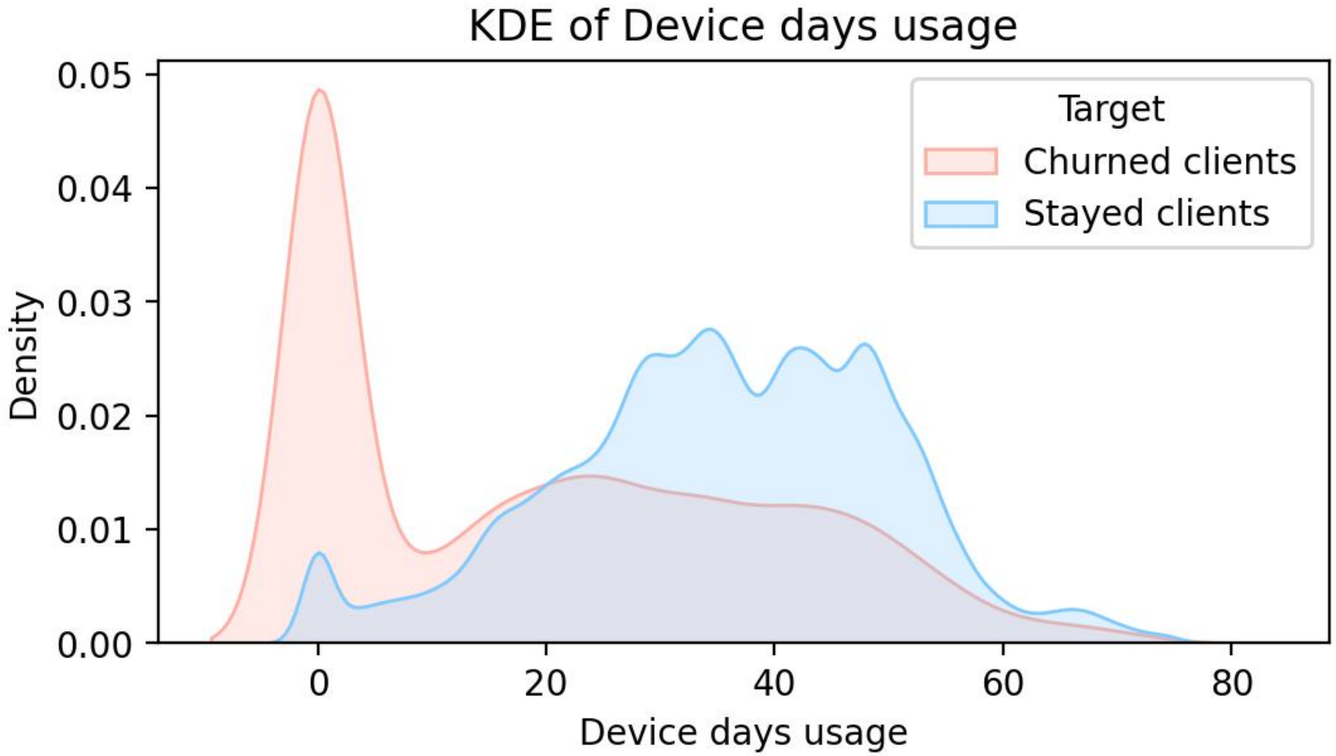
Data Analysis

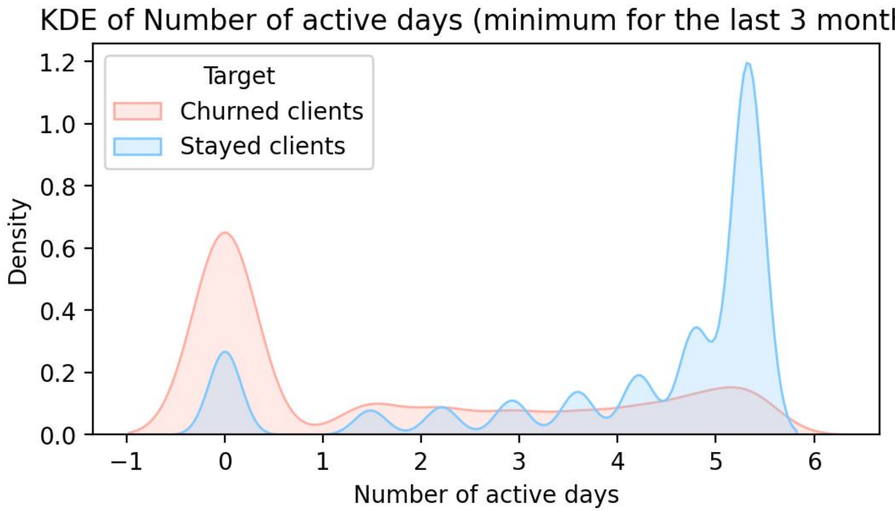
FE

Customers profile

FE - customers profile

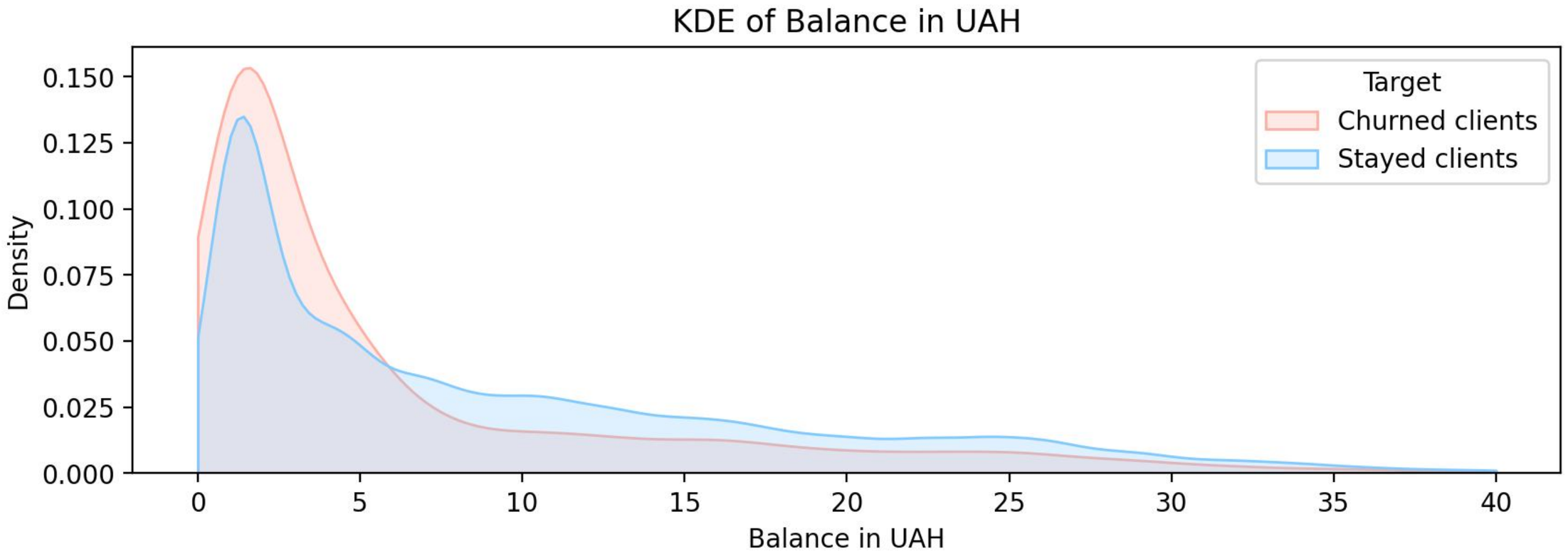
2x **150k** records
817 attributes





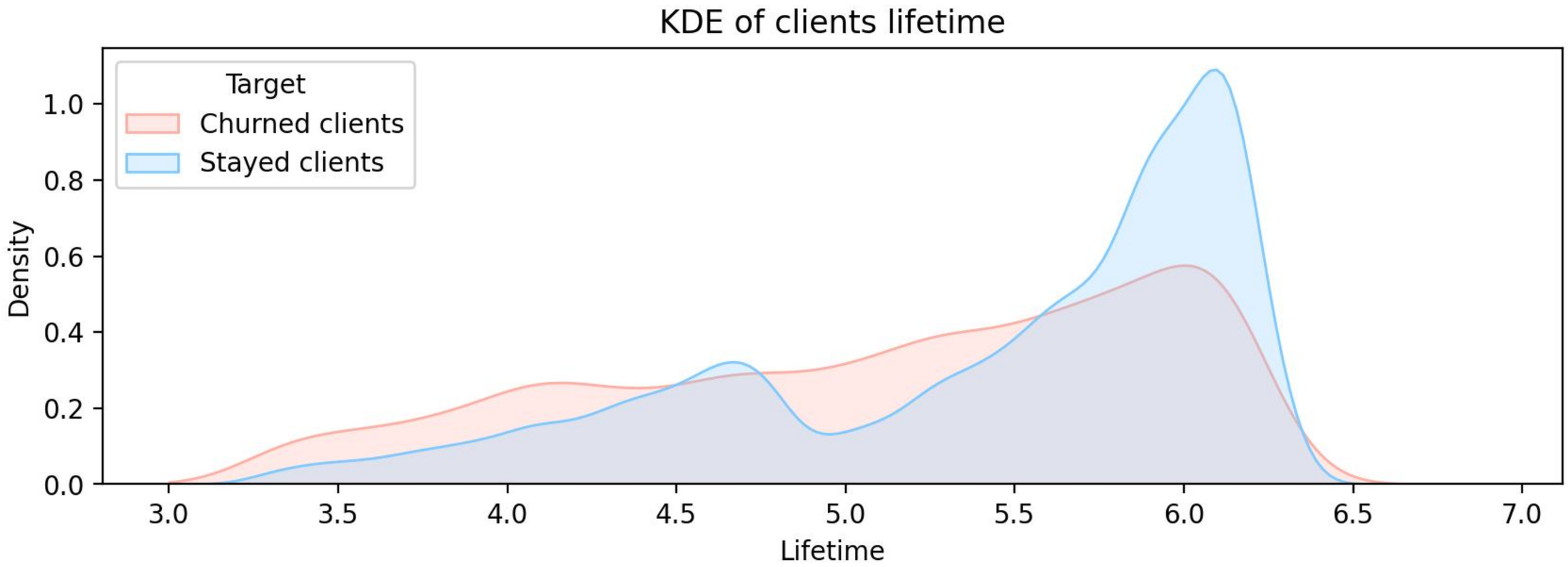
FE

Distribution of Customer Account Balances



FE

Distribution of Clients Lifetime

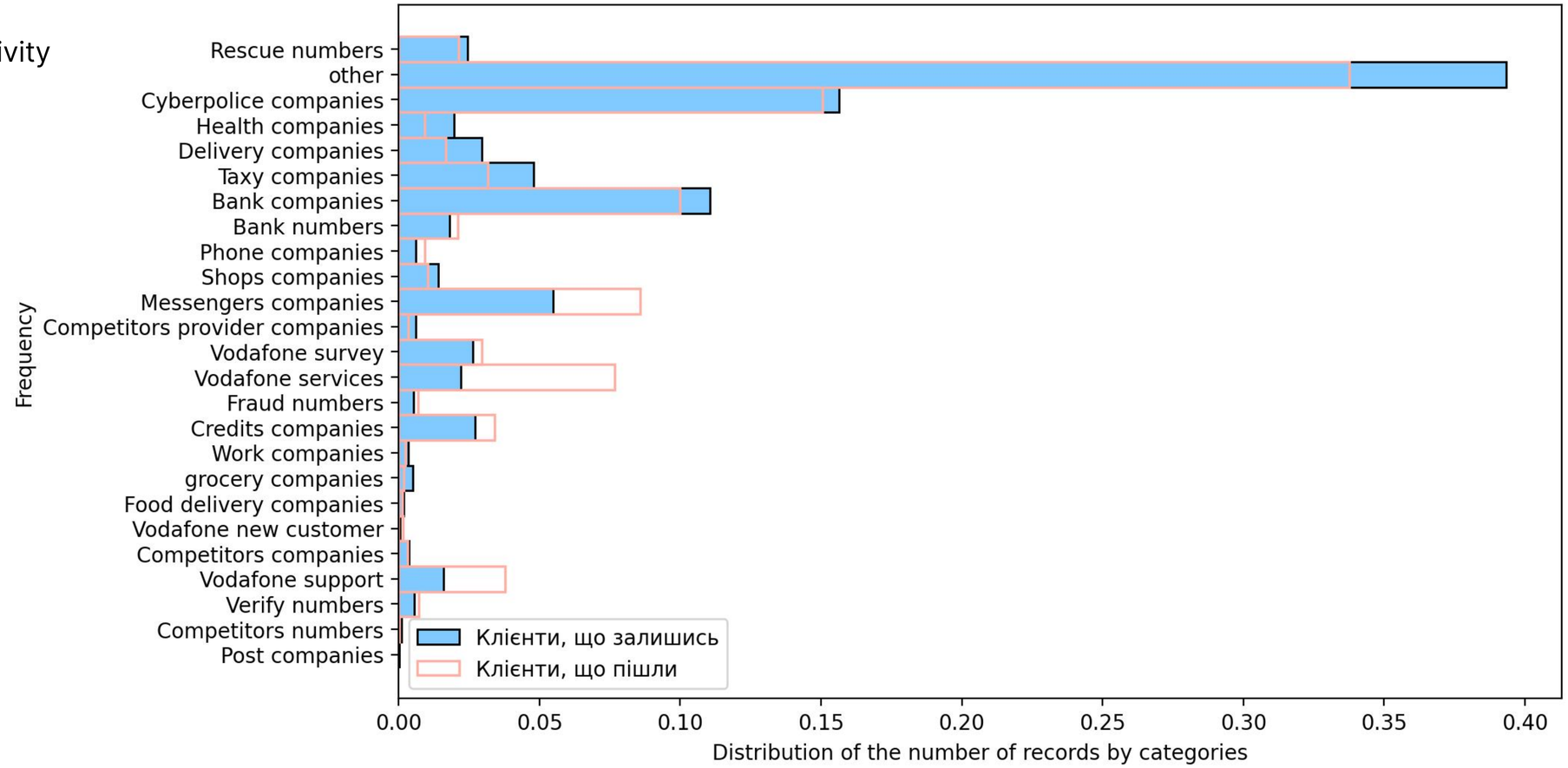


B_NUM

Interaction with Other Numbers

B_NUM - customers activity with other numbers

2x ~**669k** records
8 attributes

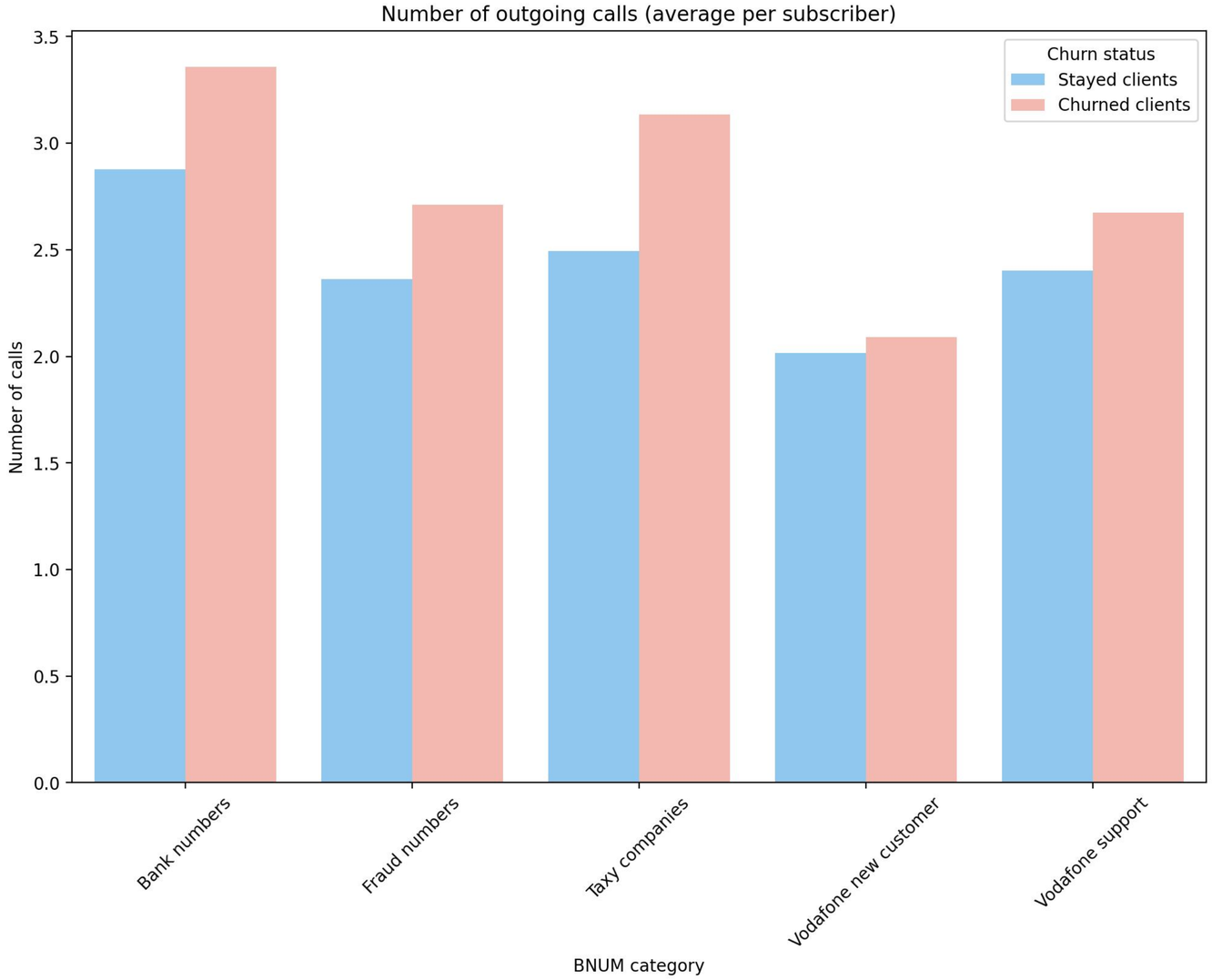


B_NUM

Number of Outgoing Calls
(Average per Customer)

On average, customers in the churn group interact more frequently with:

- Vodafone Support
- Taxi Services
- Fraud Numbers



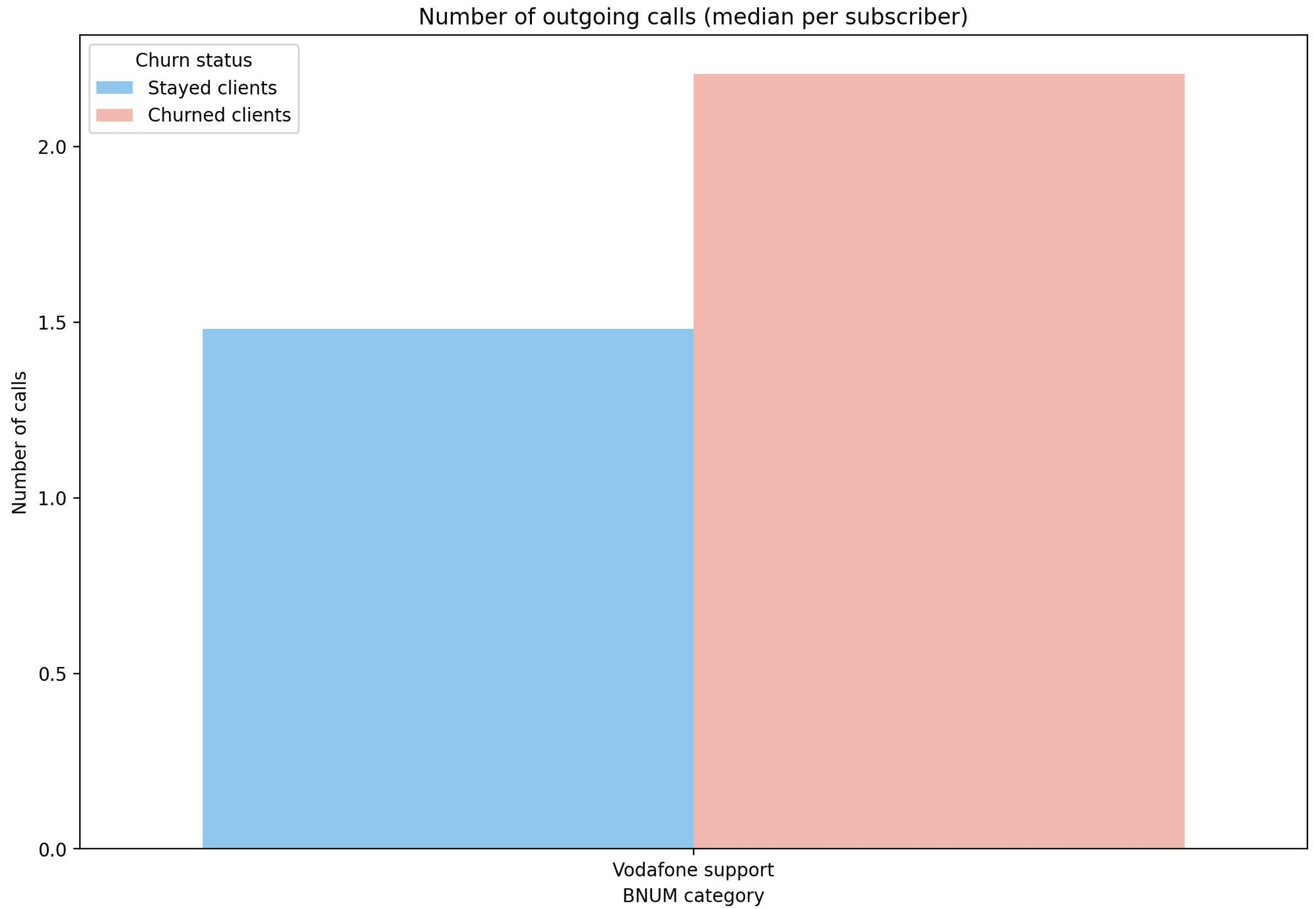
B_NUM

Number of Outgoing Calls (Median per Customer)

The median also indicates that customers in the churn group contacted Vodafone support more frequently

---

Recommendation: Investigate whether customers in the churn group encountered specific issues and if Vodafone support successfully resolved them

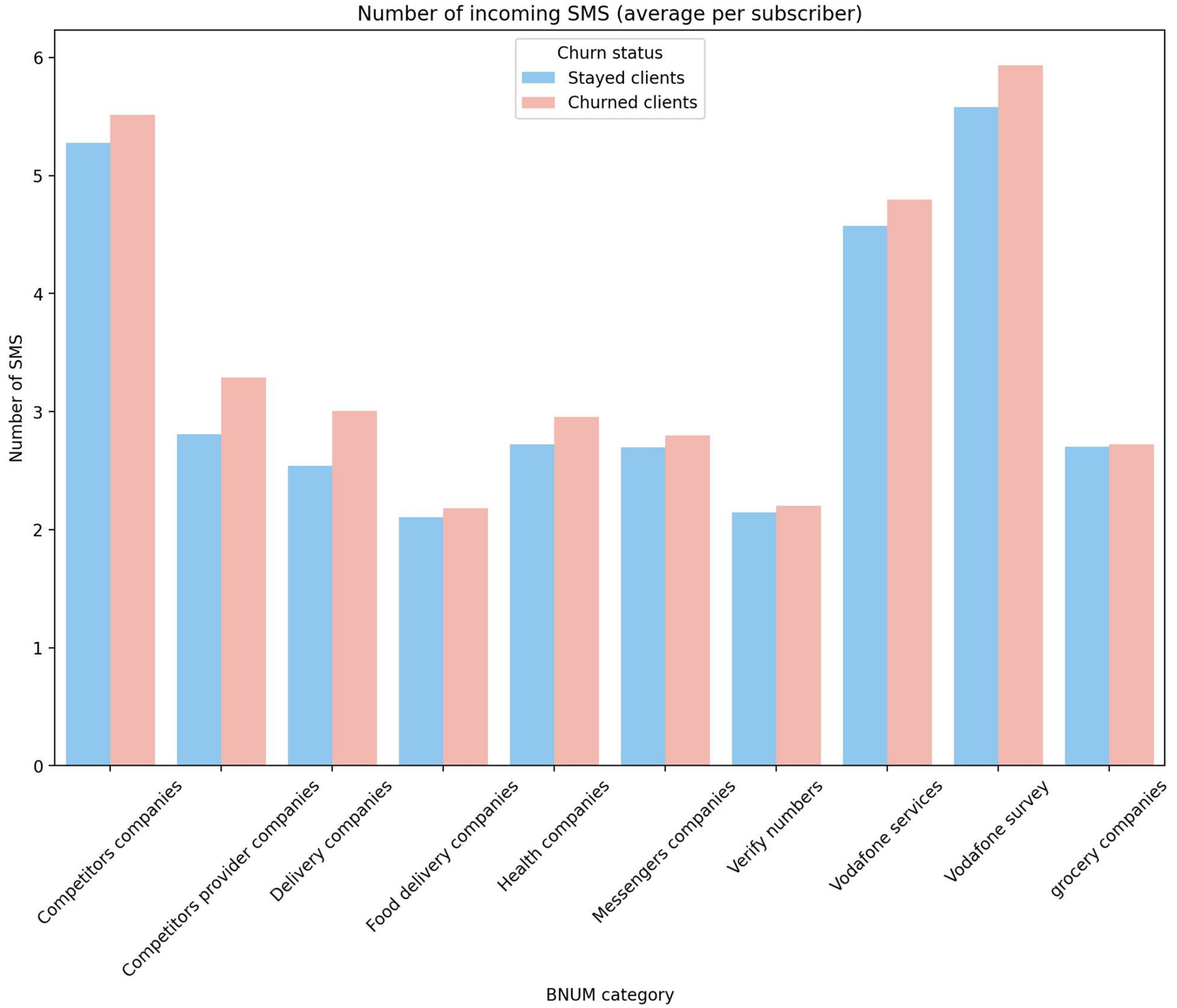


B_NUM

Number of Incoming SMS
(Average per Customer)

On average, customers in the "churn" group received more SMS from:

- Vodafone (surveys and services)
- Competitor services



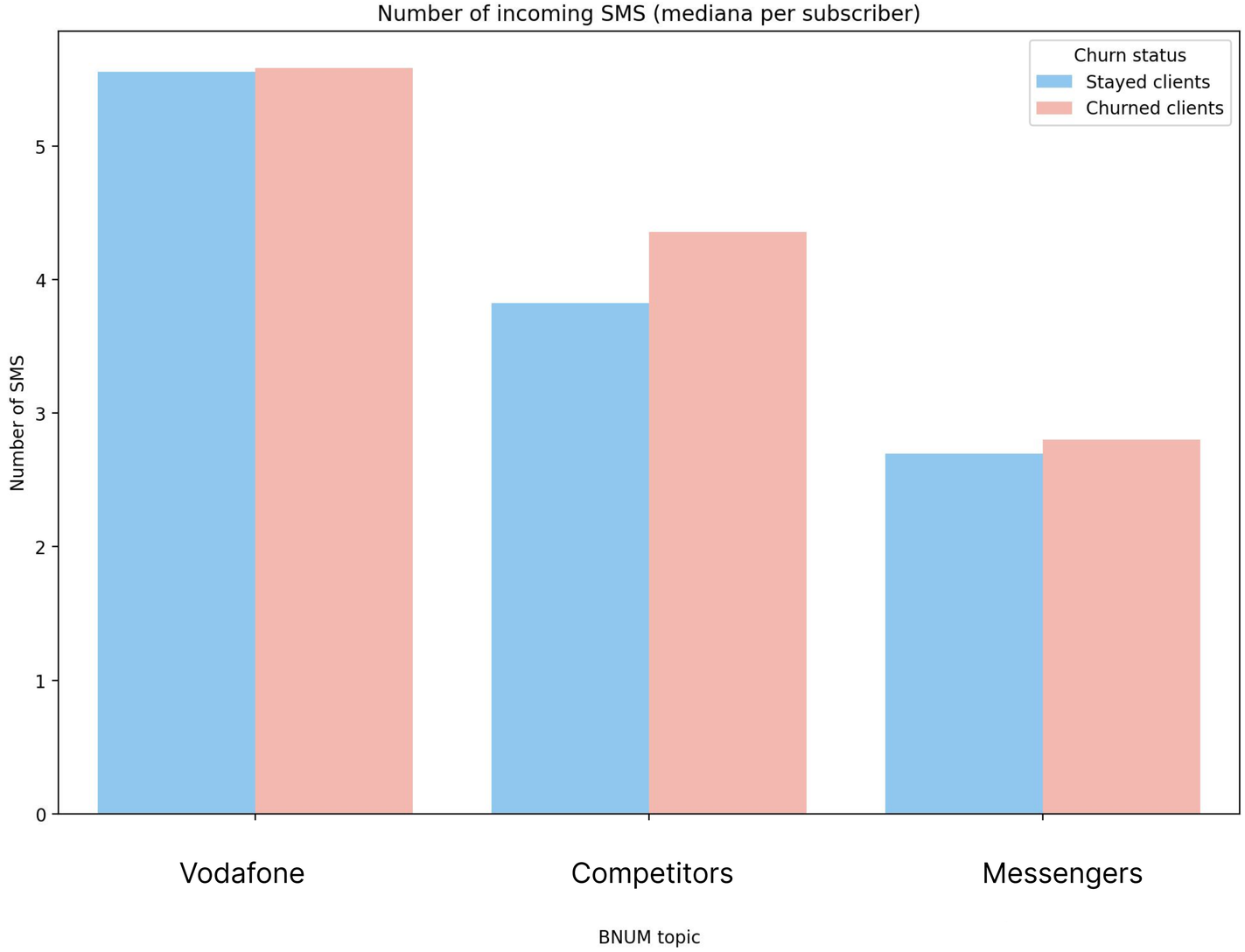
B_NUM

Number of Incoming SMS
(Median per Customer)

The median indicates that customers in the churn group received SMS from competitor companies more frequently

---

Based on the available data, it is worth investigating which services from competitors were offered to customers in the "churn" group



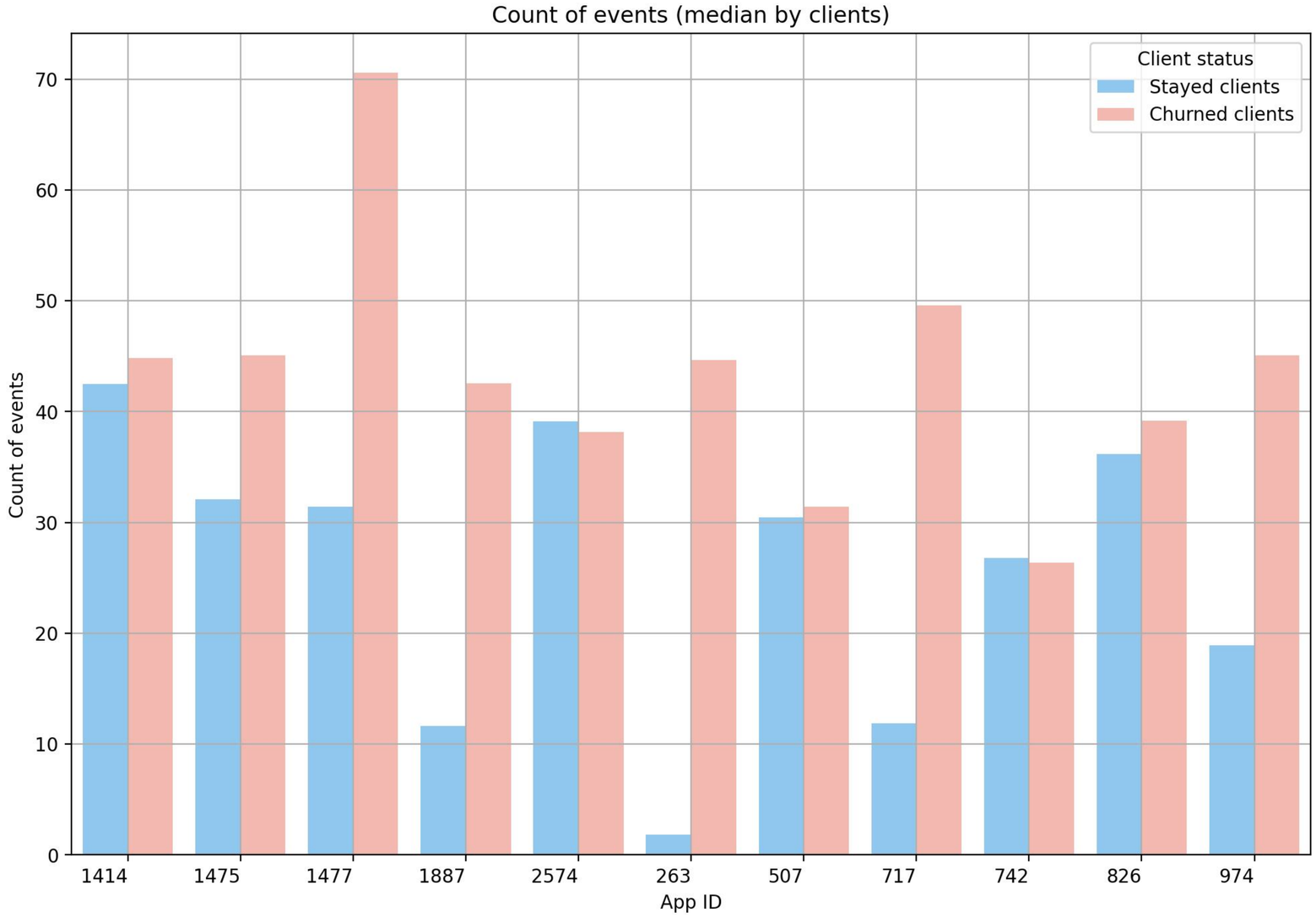
DPI

Customer Traffic by Applications

DPI - clients traffic by applications

2x ~**6.7m** records
6 attributes

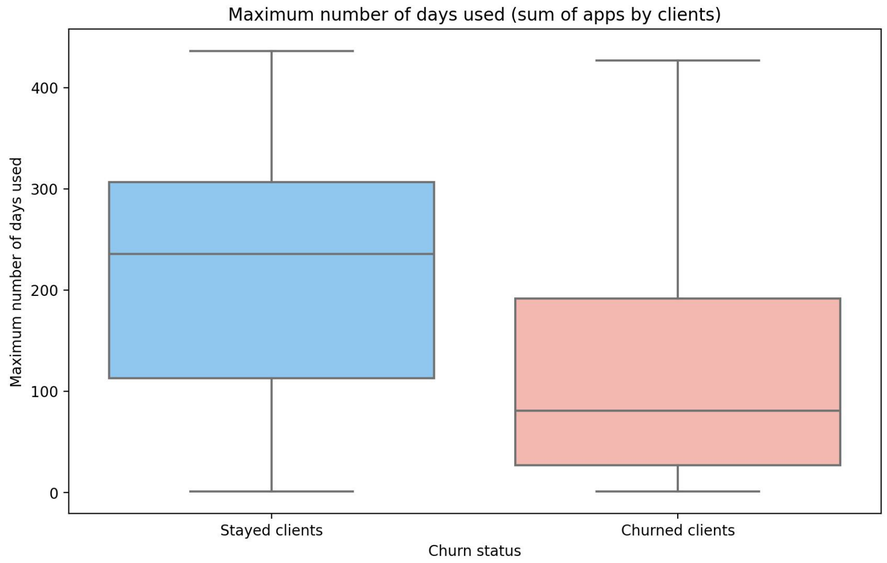
Certain applications have been noticed to be more prevalent among customers in the churn group

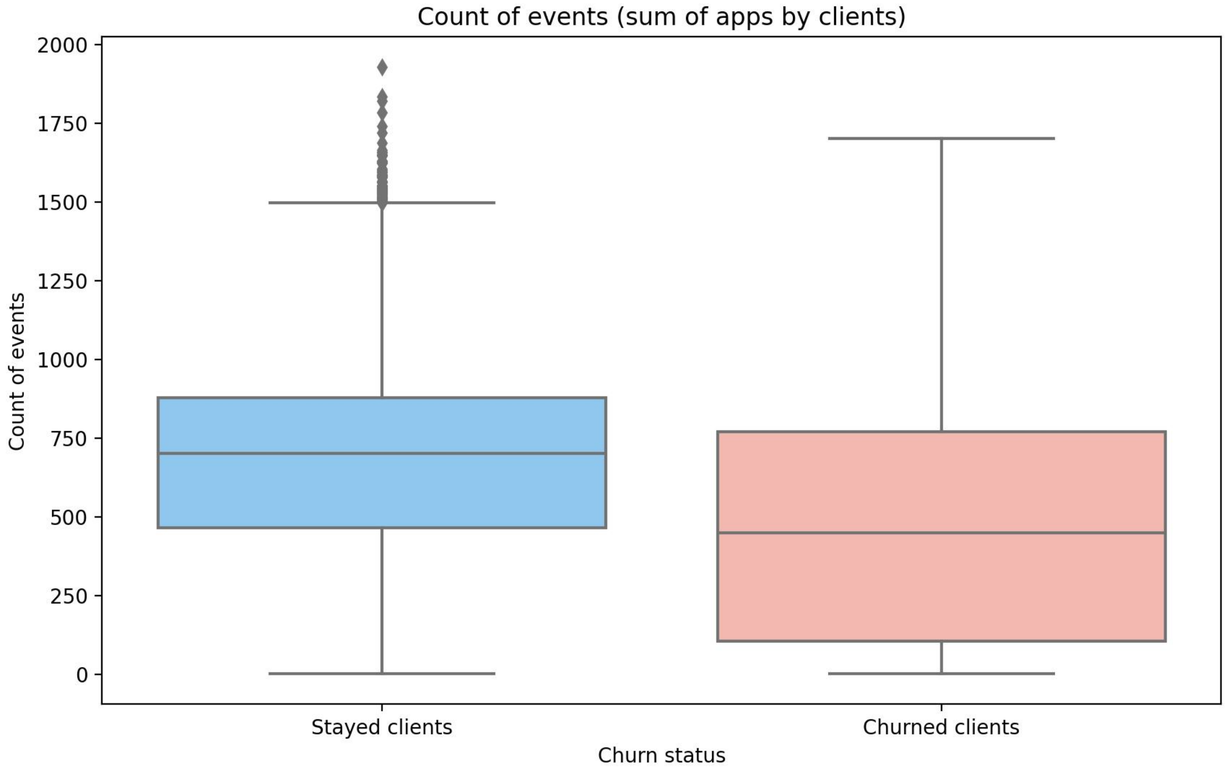


DPI

Number of Events Among Popular Applications

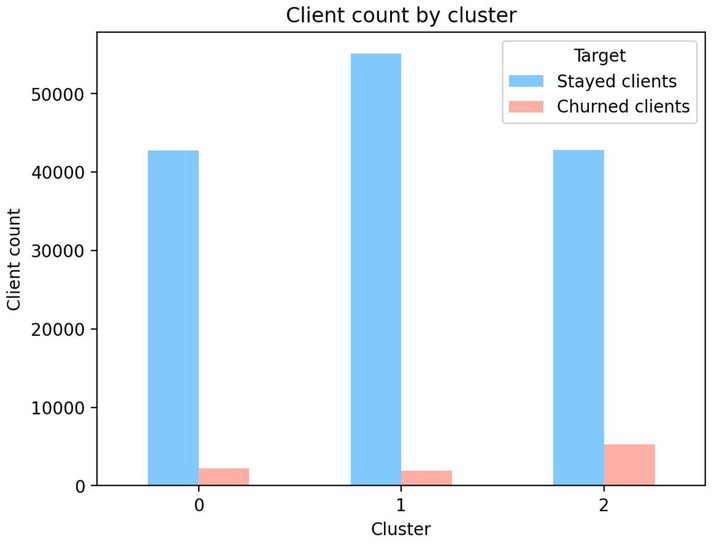
Popular applications are used less frequently by customers in the churn group.

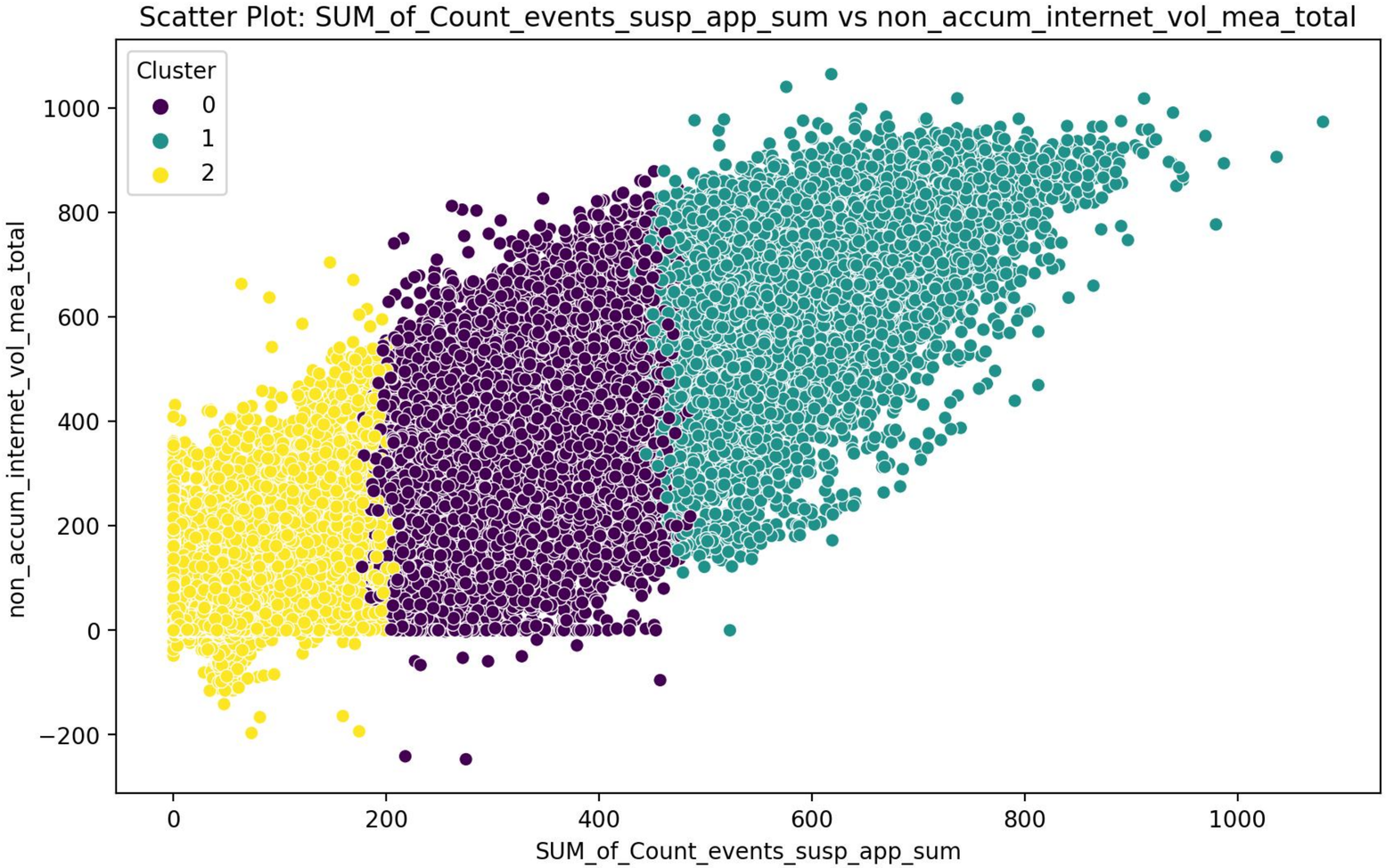




Clustering

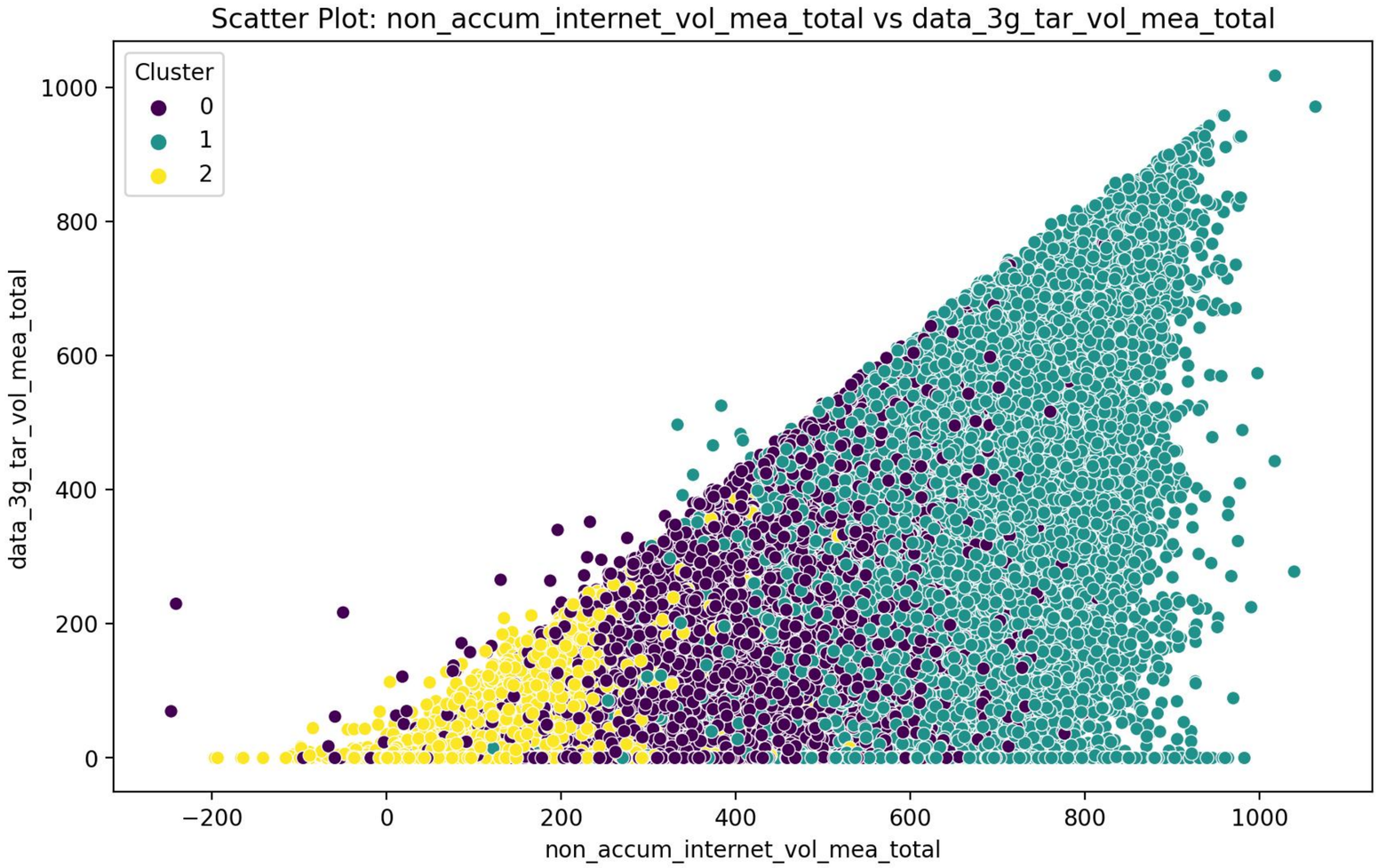
56% of customers in the churn group are located in "Cluster 2"





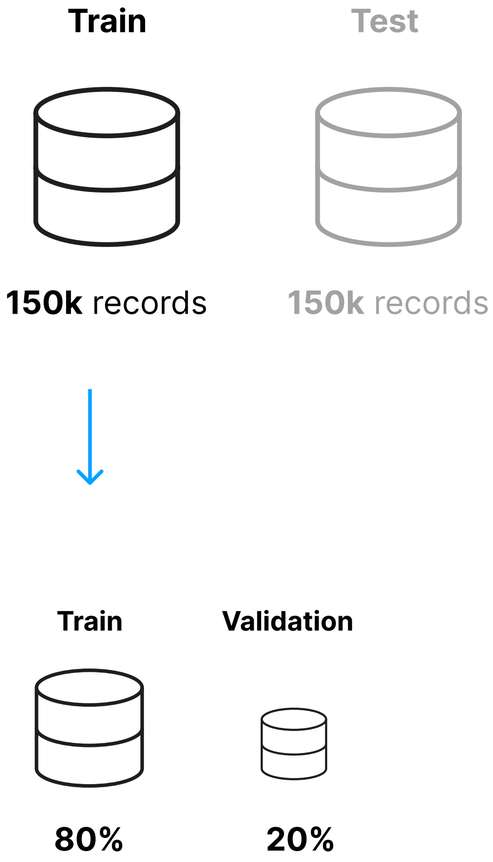
Clustering

In the future, this can be used in advertising campaigns to offer customers the most suitable tariffs and services based on their needs

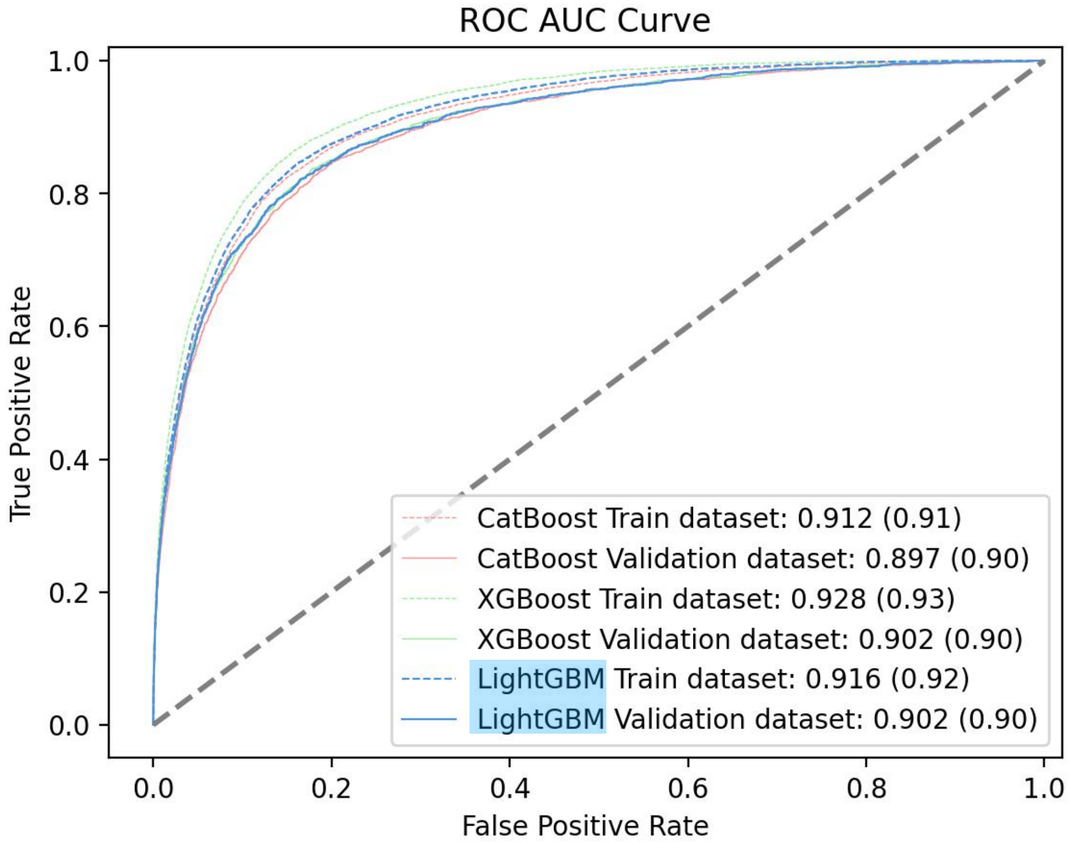


Model Selection and Training

Model Training



Randomly Split,
Maintaining Class Proportions

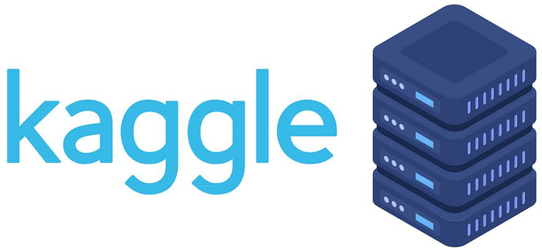


Metrics

- **ROC AUC** - Overall Model Performance
- **F score (beta = 0.5)** - Accuracy is the Priority

Hyperparameter Tuning

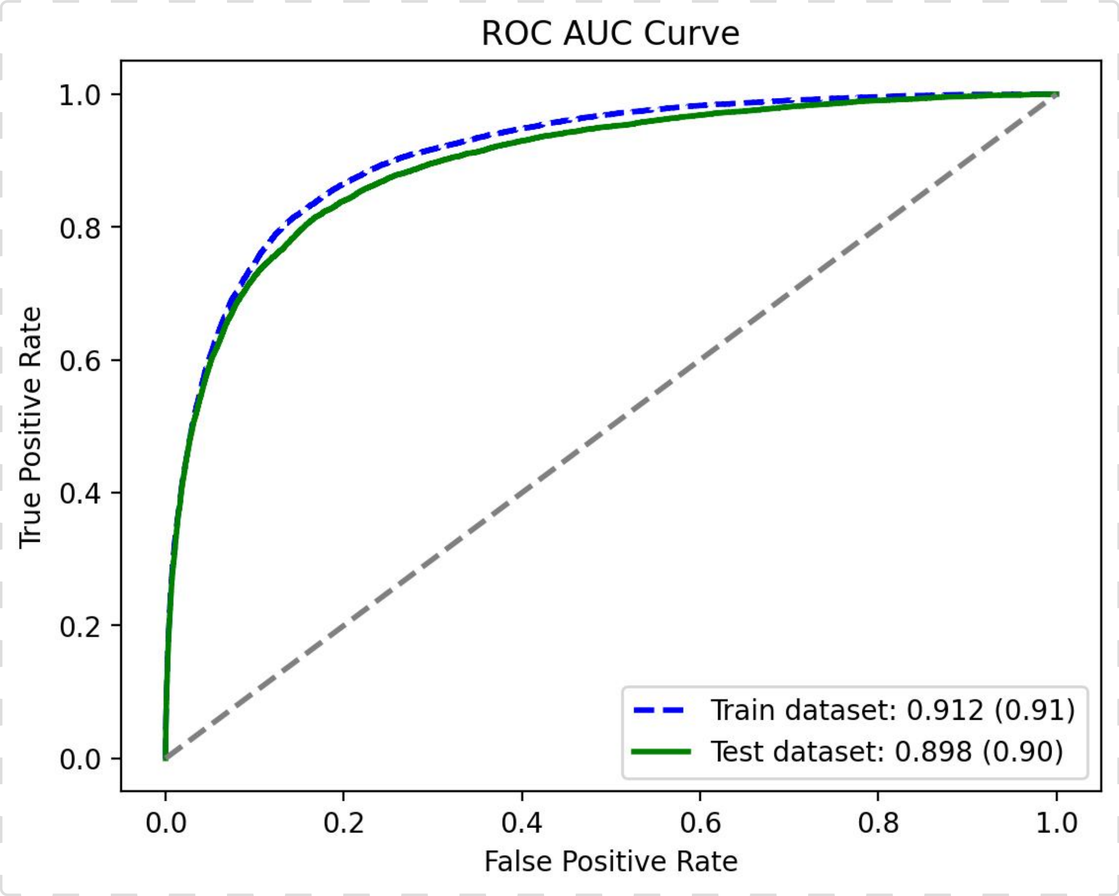
```
  model_params = {
-     "lambda_l1": 2.91,
-     "lambda_l2": 5.57,
-     "learning_rate": 0.075,
-     "num_leaves": 92,
-     "feature_fraction": 0.3888218905277871,
-     "bagging_fraction": 0.26715088946500626,
-     "max_depth": 17,
-     "n_estimators": 100,
+     "lambda_l1": 8,
+     "lambda_l2": 5,
+     "learning_rate": 0.018,
+     "num_leaves": 14,
+     "feature_fraction": 0.6803603979260223,
+     "bagging_fraction": 0.6735621254996546,
+     "max_depth": 11,
+     "n_estimators": 350,
      "random_state": 42,
      "seed": 42,
      "objective": "binary",
```




```
0.6 * np.mean(auc_scores)
+ 0.4 * np.mean(fbeta_scores)
- (10 * np.std(auc_scores))
- (10 * np.std(fbeta_scores))
```

Final Metrics

Results

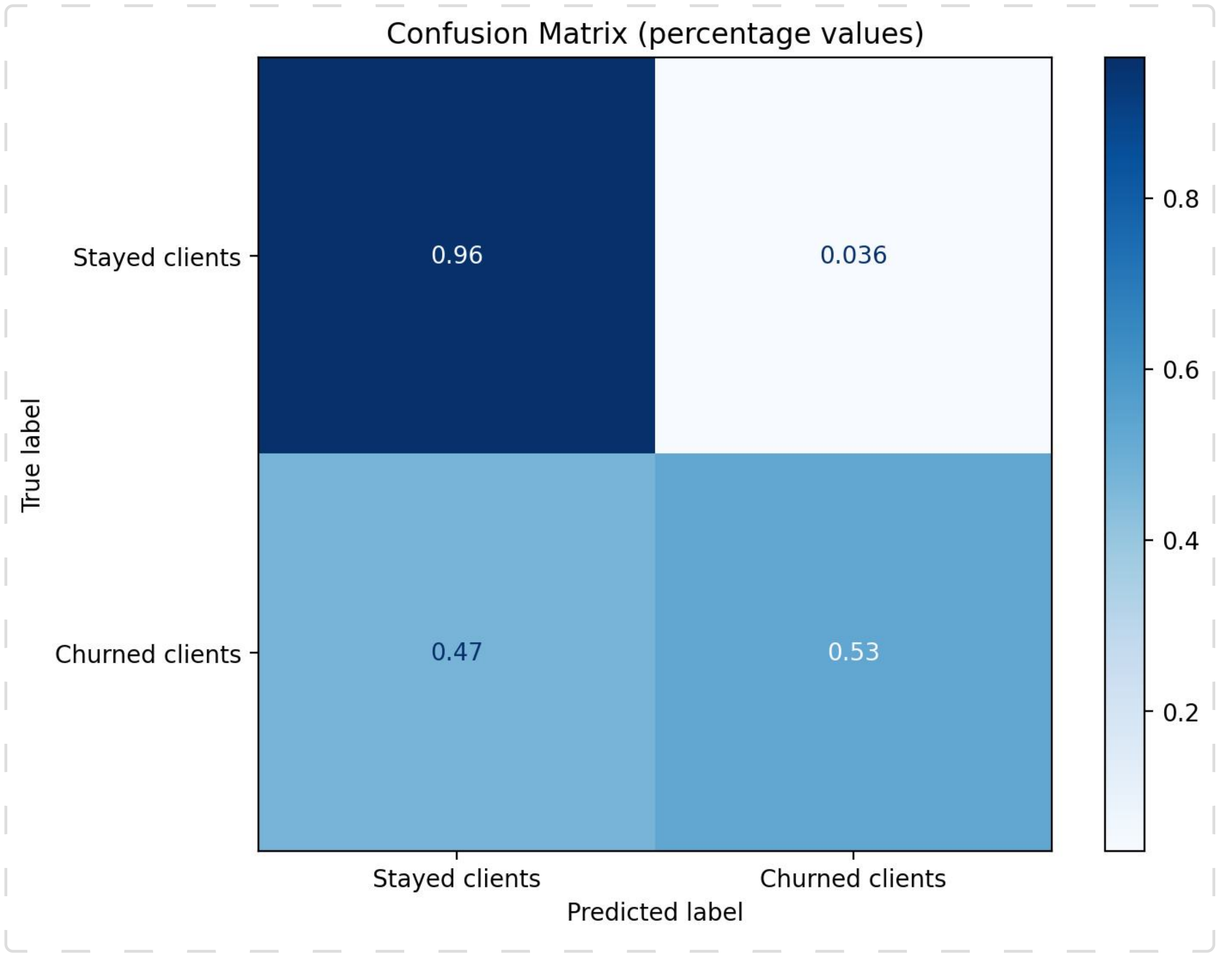


Train		Test
0.912 ROC AUC	>1%	**0.898** ROC AUC
0.560 F beta=0.5 score	>1%	**0.548** F beta=0.5 score
0.909 Threshold	>1%	**0.895** Threshold
0.95 Accuracy		**0.95** Accuracy

Probability Distribution

The model distinguishes quite well between customers who will stay and those who will leave based on churn probability.

Confusion Matrix



Test

0.94

Accuracy

0.49 / 0.53

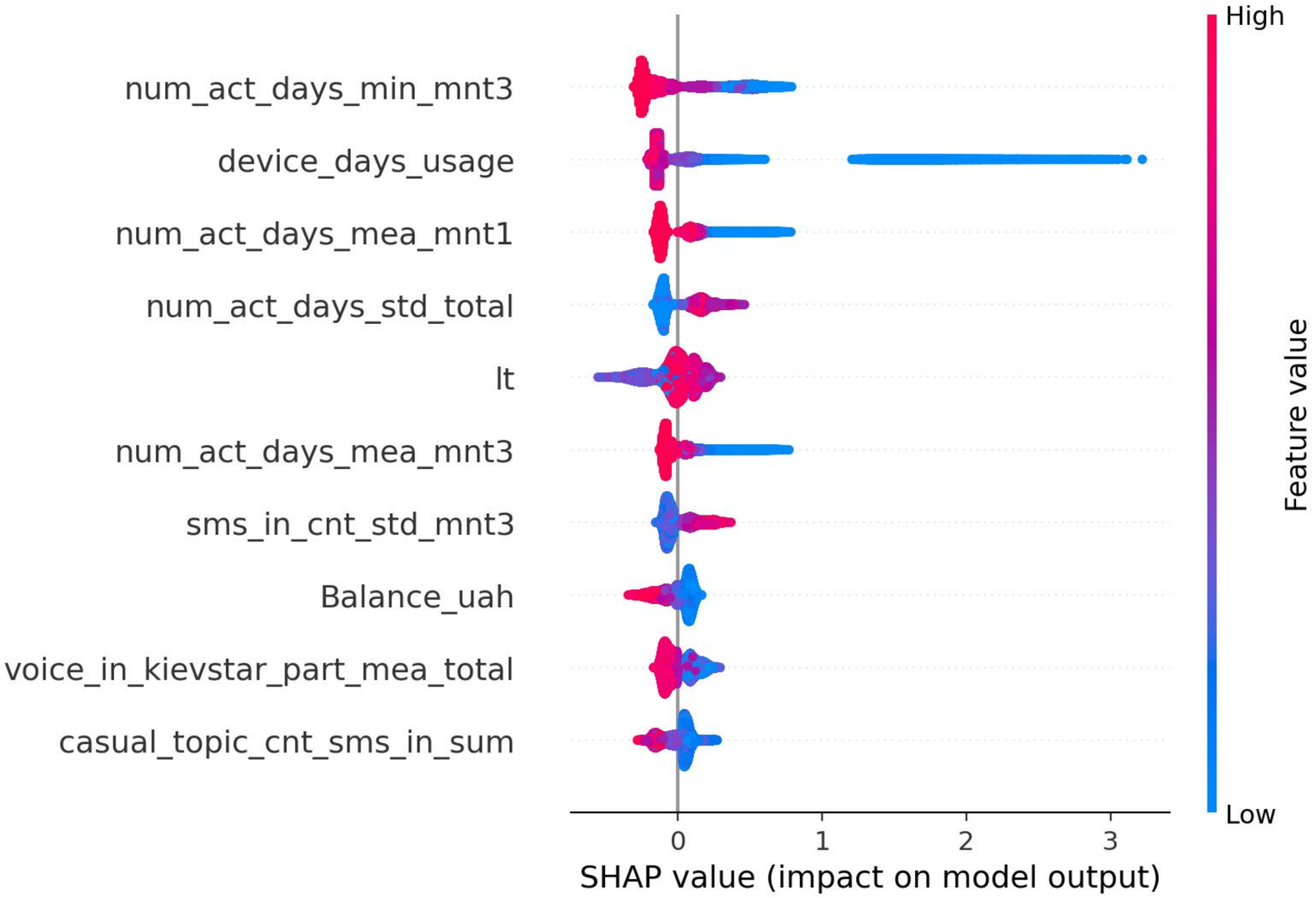
Precision / Recall for 1 class (Churn)

0.97 / 0.96

Precision / Recall for 0 class (Not Churn)

Threshold 0.83 (maximum for F1 score)

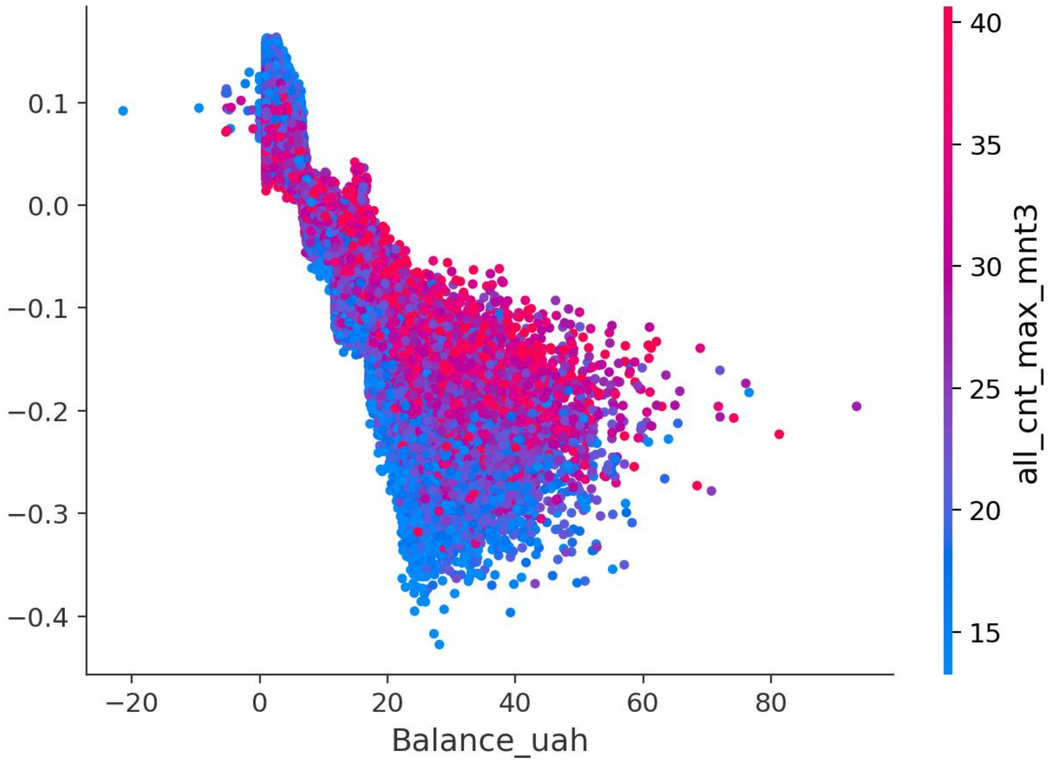
Feature Impact



Feature Importance

As the balance decreases, the probability of customer churn increases

A certain balance threshold can be monitored, and specific offers can be made to customers prone to churn

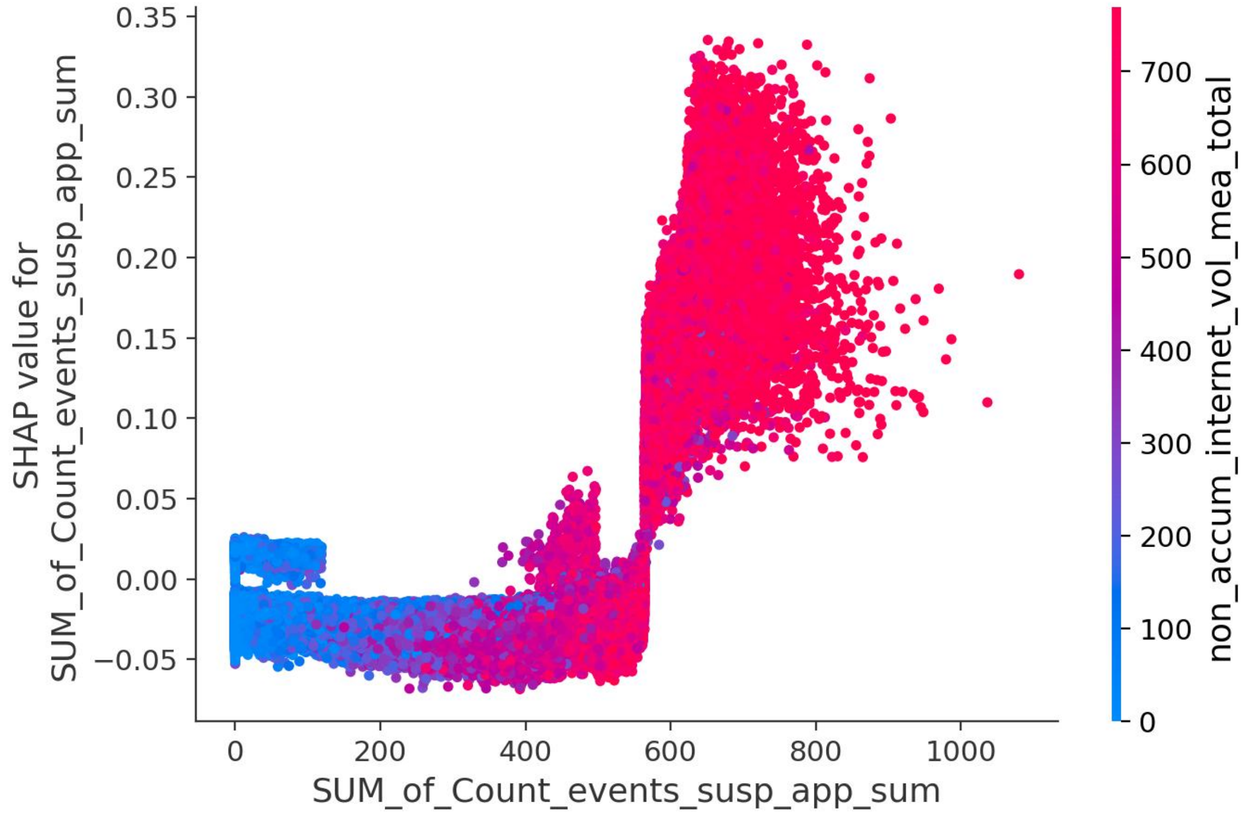


Feature Importance

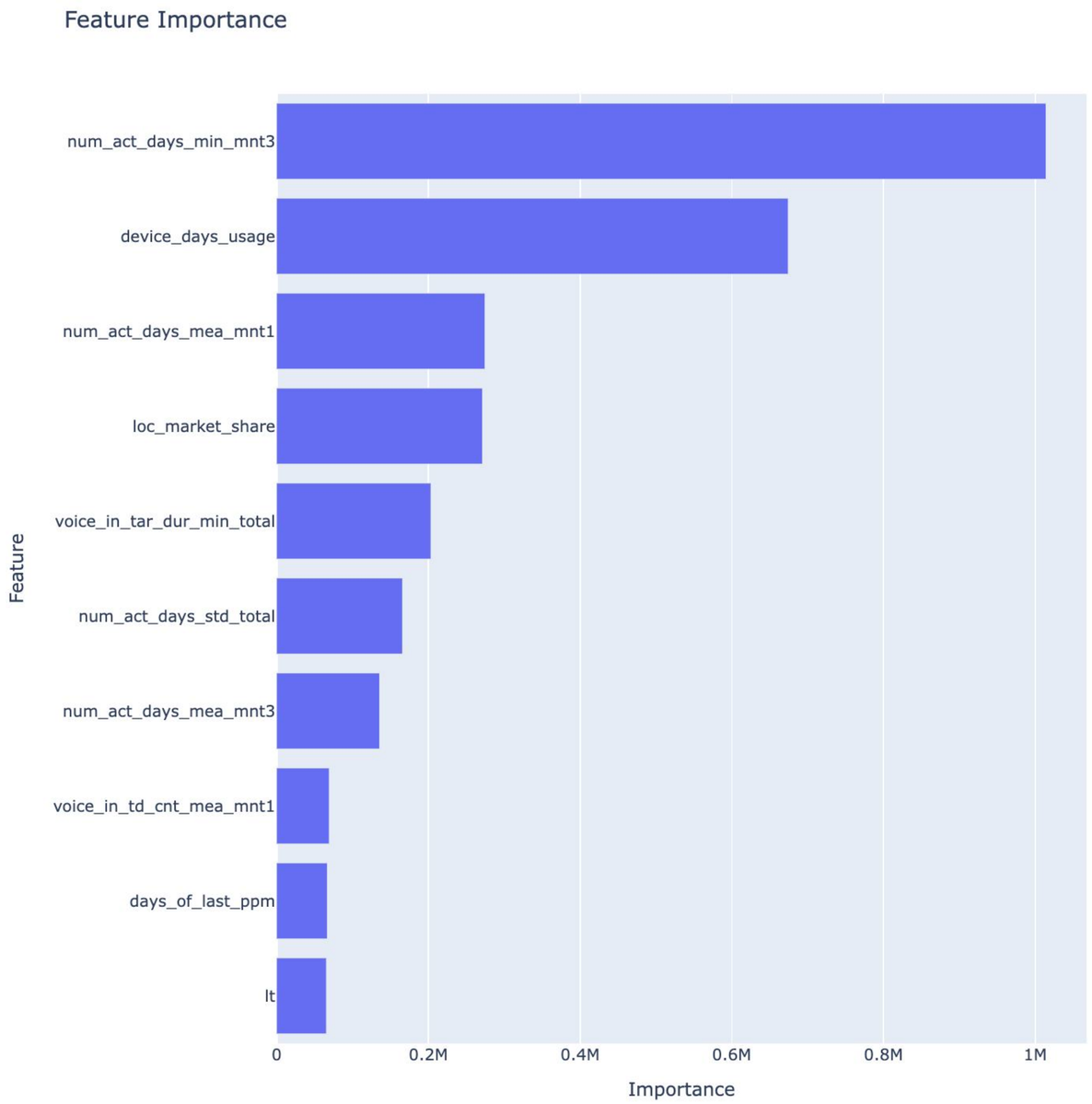
The number of events in popular applications among the churn group compared to data usage beyond the plan limits

High usage of certain applications increases the likelihood that a customer may churn

It can be hypothesized that some active users may be dissatisfied with their current plan, which in turn, may increase the likelihood of churn



Feature Importance



Economics

Результати за ймовірностю значень моделі

Rate	NOT Churned customers	Churned customers	Recall of churned customers	Recall of NOT churned customers	Precision (%)	Lift	Marketing Costs
5	3270	4230	45%	2%	56%	9	$166 750,00
10	9000	6010	64%	6%	40%	6	$159 800,00
15	15530	6970	74%	11%	31%	5	$173 250,00
20	22430	7570	80%	16%	25%	4	$195 750,00
25	29550	7950	85%	21%	21%	3	$223 750,00
30	36740	8260	88%	26%	18%	3	$253 500,00
35	44030	8470	90%	31%	16%	3	$285 750,00
40	51360	8640	92%	37%	14%	2	$319 000,00
45	58730	8770	93%	42%	13%	2	$353 250,00
50	66110	8890	95%	47%	12%	2	$387 750,00
55	73520	8980	96%	52%	11%	2	$423 000,00
60	80930	9070	96%	58%	10%	2	$458 250,00
65	88350	9150	97%	63%	9%	1	$493 750,00
70	95790	9210	98%	68%	9%	1	$529 750,00
75	103240	9260	98%	73%	8%	1	$566 000,00
80	110700	9300	99%	79%	8%	1	$602 500,00
85	118160	9340	99%	84%	7%	1	$639 000,00
90	125630	9370	100%	89%	7%	1	$675 750,00
95	133110	9390	100%	95%	7%	1	$712 750,00
100	140600	9400	100%	100%	6%	1	$750 000,00

64%

of churned customers found with **6%** false positives from the stayed group

x6

times better than random prediction

Economics

Input Data

140 000
Customers who stayed

10 000
Customers who churned

5 USD
Retention cost

25 USD
Acquisition cost

Randomness

75 000
customers need to be retained
75 000 * 5 USD = 375 000 USD

5 000
customers need to be acquired
5 000 * 25 USD = 125 000 USD

525 000 USD

Nothing

10 000
customers need to be acquired
10 000 * 25 USD = 250 000 USD

250 000 USD

Model

-34%

6 400 + 8 400
customers need to be retained
8 100 * 5 USD = 74 000 USD

3 600
customers need to be acquired
3 600 * 25 USD = 90 000 USD

6 400 - the model found
8 400 - the model was wrong
3 600 - the model did not find

164 000 USD

Summary

DATA COLLECTION

FE - Customers profile

2x **150 тис.** records
817 attributes

🔑 abon_id

B_NUM - traffic by numbers

2x ~**669 тис.** records
8 attributes

DPI - traffic by apps

2x ~**6.7 млн.** records
6 attributes

DATA ANALYSIS

Data preparation

❌ **-27** empty attributes
✅ **790** numeric attributes

↓

New attributes

FE **+526** attributes
BNUM **+201** attributes
DPI **+31** attributes

↓

High correlation Filter

Threshold: 95%

↓

989
numeric features

MODEL TRAINING

Model selection

🏆 LightGBM
🥈 XGBoost
🥉 CatBoost

↓↑

Model training

kaggle
LightGBM

↓↑

Feature selection

LGBM Feature Importance
SHAP Analysis

58
numeric features

0.90
ROC AUC for TEST data

0.91
ROC AUC for TRAIN data

~1%
percentage difference between Test and Train samples

40% Precision
64% Recall

Results

-34%
витрат в порівнянні з відсутністю стратегії утримання

64%
of churned customers found with **6%** false positives from the stayed group

x6
times better than random prediction

Recommendations

Investigate whether customers in the churn group encountered specific issues and if Vodafone support was able to resolve them

Check if the current tariff satisfies customers in the churn group by analyzing their app usage

Respond to low account balances by offering tailored proposals based on the customer's activity level

Thank you

Yaroslav Bezrukavyi

https://www.linkedin.com/in/bezrukavyi

https://github.com/bezrukavyi

yaroslav.bezrukavyi@gmail.com