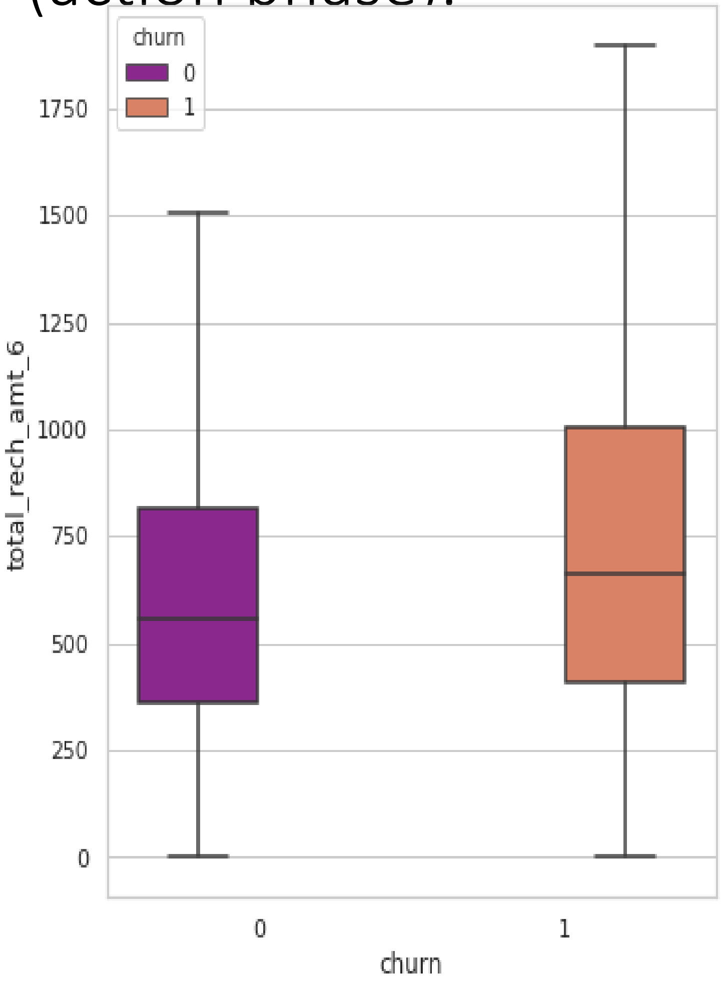
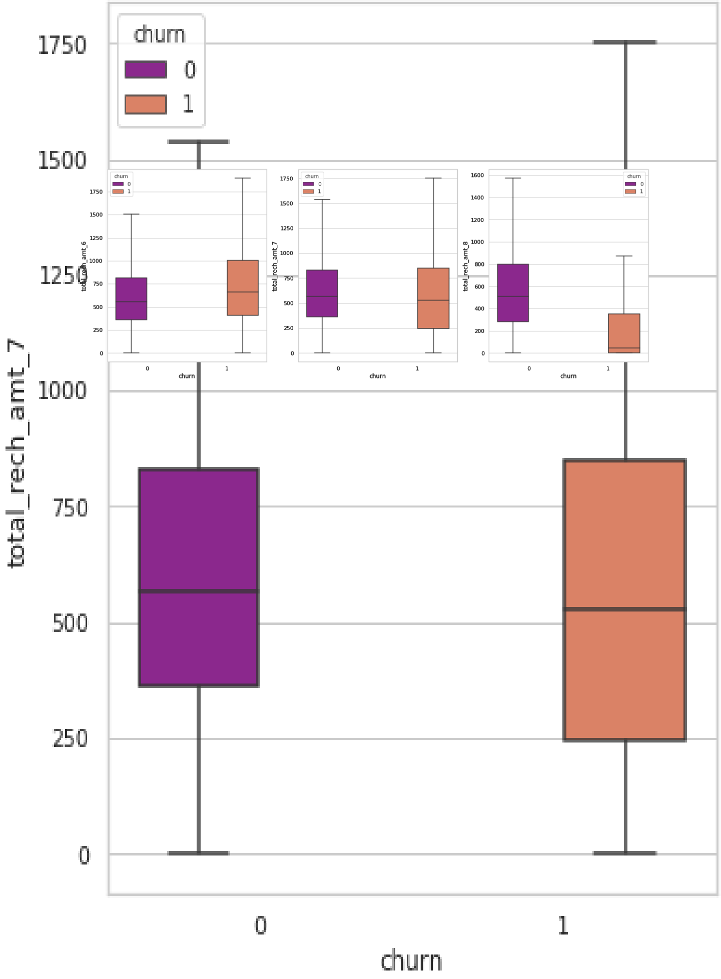
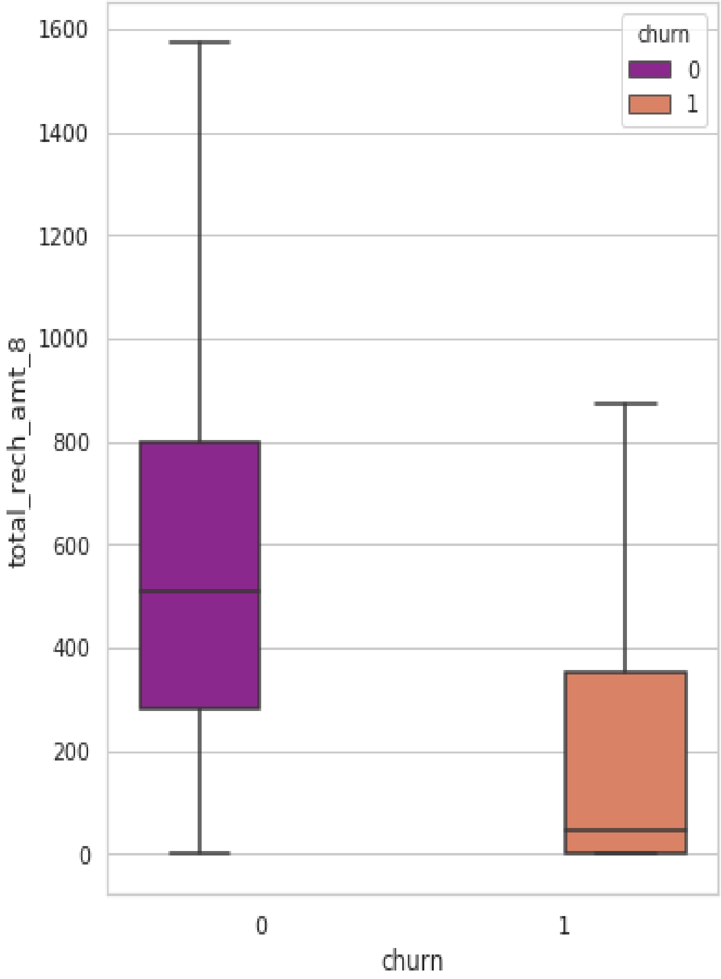
Case Study : Telecom Churn Case Study
Submitted by Ishan Kumar

EDA

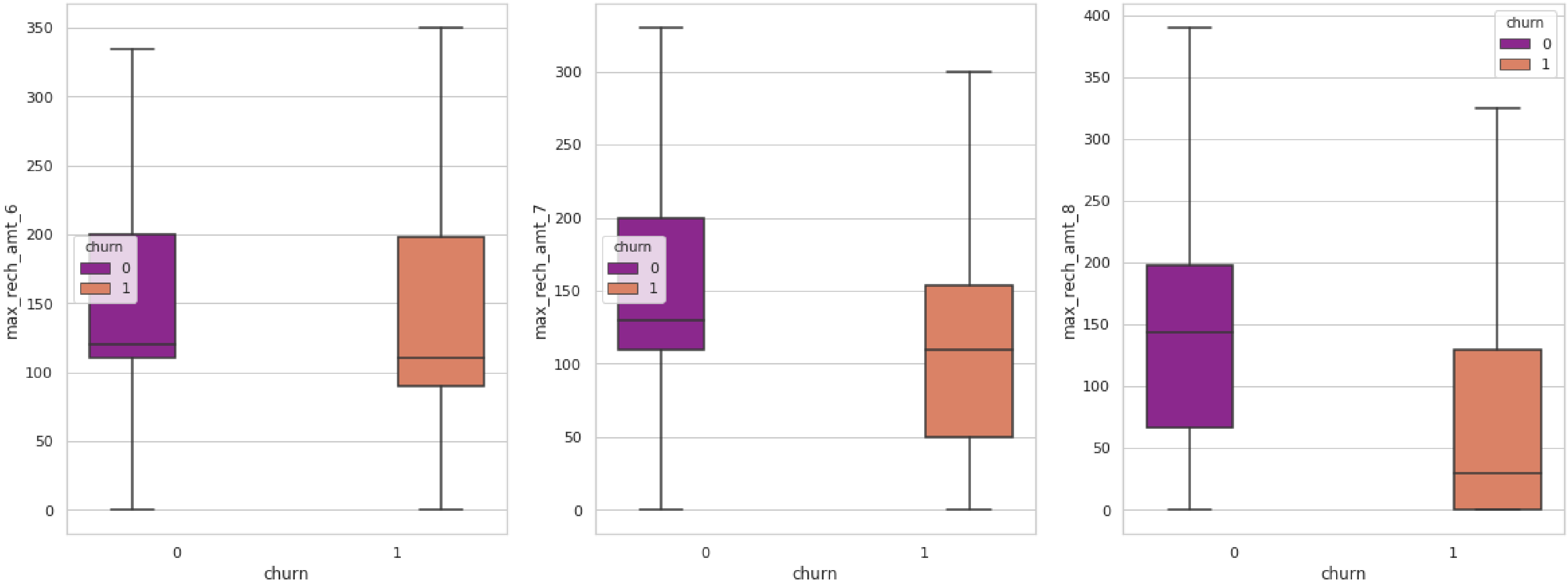
> Analysis:** In the eighth month (Action Phase), we can observe a decline in the overall recharge amount for churned consumers.
> Evaluation: We can see that there is a significant decrease in the total amount of data recharges for churned customers in the eighth month (action phase).



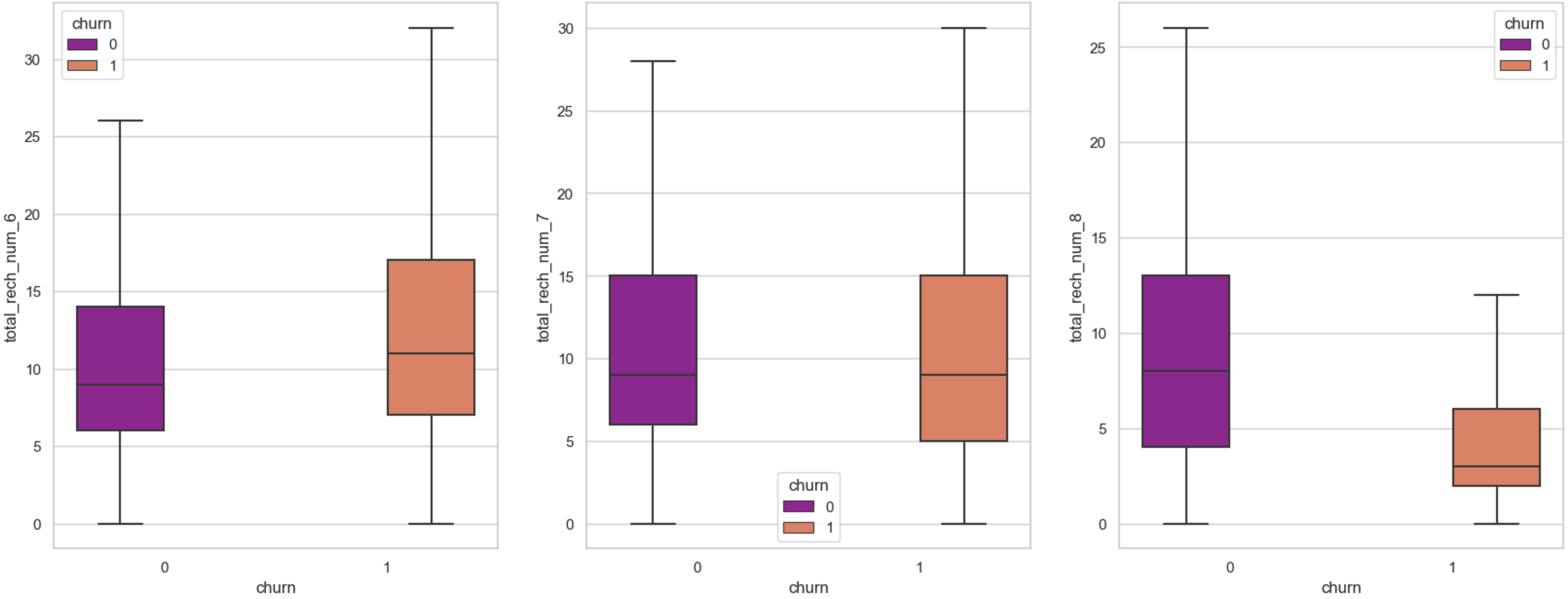




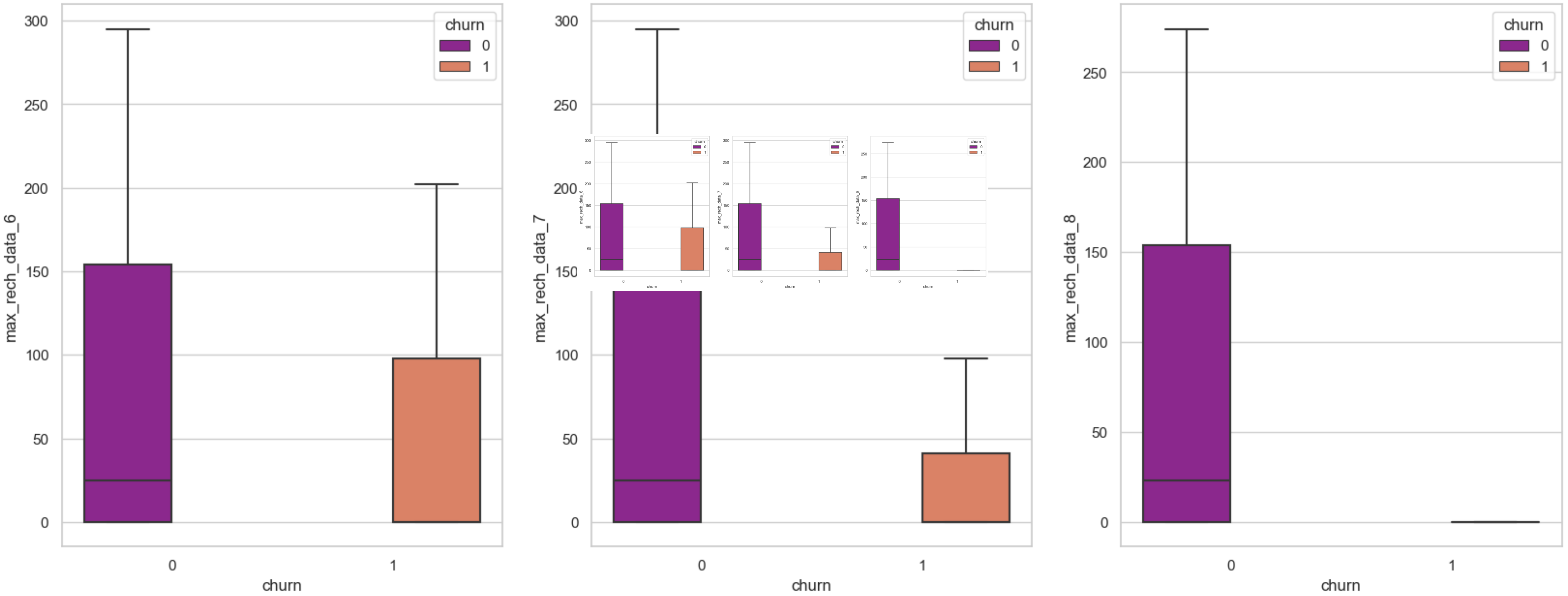
Analysis: For churned clients, the maximum data recharge amount has significantly decreased by the eighth month (action phase).



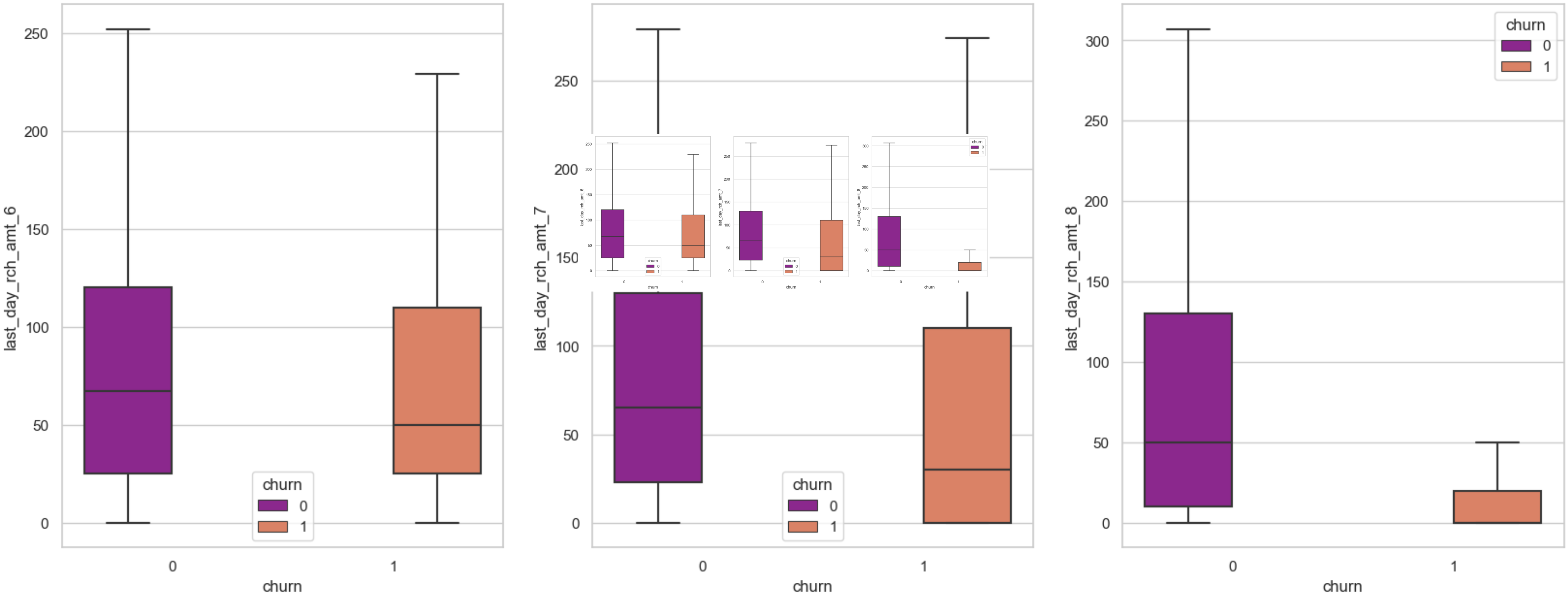
Analysis: It is clear that the number of total recharges for churned consumers significantly decreased in the eighth month (activity phase).



Analysis: It is clear that the maximum data recharge for churned consumers has significantly decreased in the eighth month (action phase).

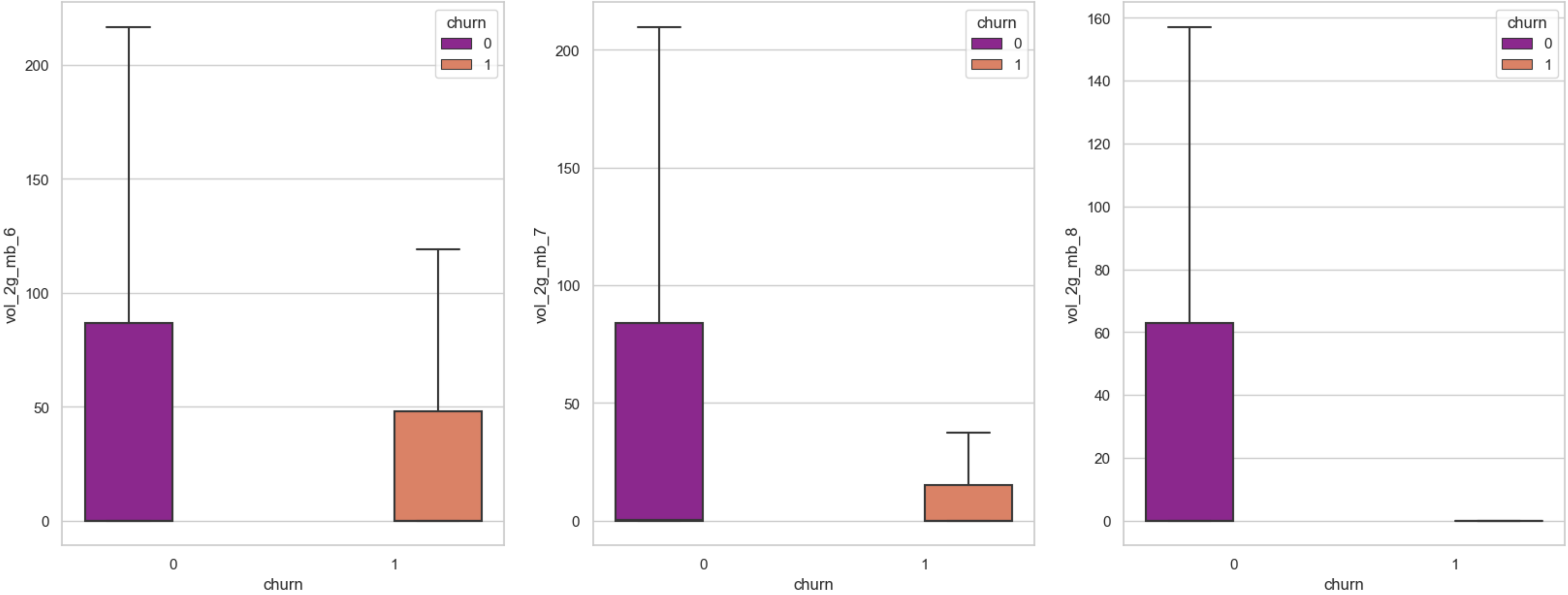


Analysis: For churned clients, the 8th month recharge amount has significantly decreased.



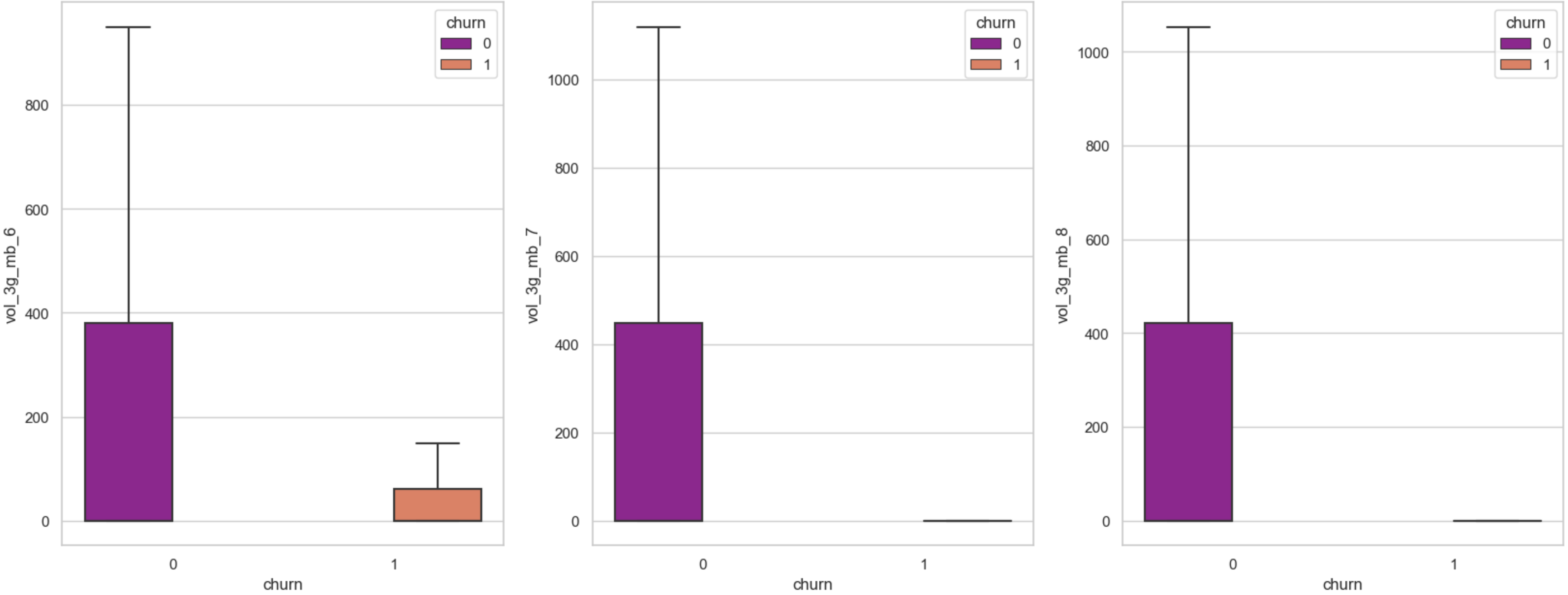
2G and 3G usage characteristics

- Analysis: For the average revenue generated by 2G/3G users and the number of recharges, more than 40% of the values are not accessible. Even if we have statistics on 2G and 3G volume utilisation, we can remove these columns.

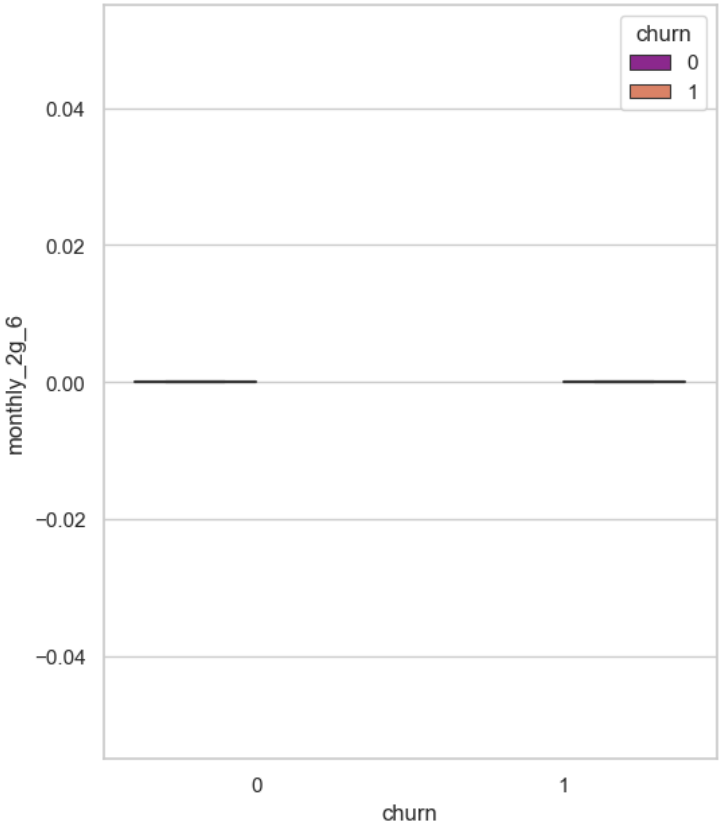

vol_2g_mb_6
0
50
100
150
200
churn
0
1
churn
0
1
vol_2g_mb_7
0
50
100
150
200
churn
0
1
churn
0
1
vol_2g_mb_8
0
20
40
60
80
100
120
140
160
churn
0
1
churn
0
1

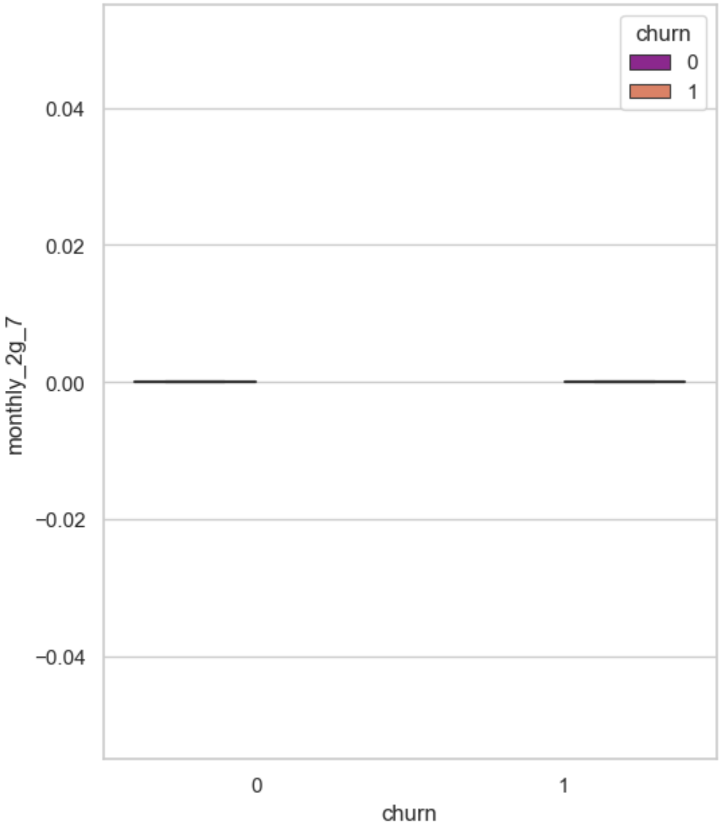
Evaluation: From the above, we can draw two conclusions:

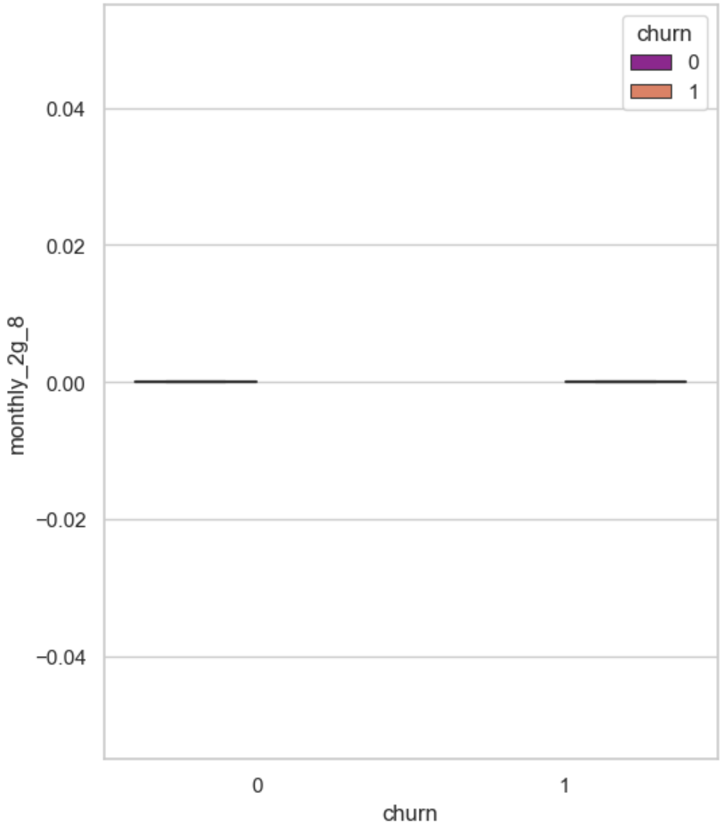
1) In the eighth month, less churned consumers are using 2G and 3G. 2) Additionally, we observe that non-churned consumers use 2G/3G services more frequently, suggesting that churned customers may be from regions where 2G/3G service is not properly provided.



Analysis: Because the value range is so narrow, the graph cannot provide a sufficient basis for argument. In order to perform analysis, examine the mean value.

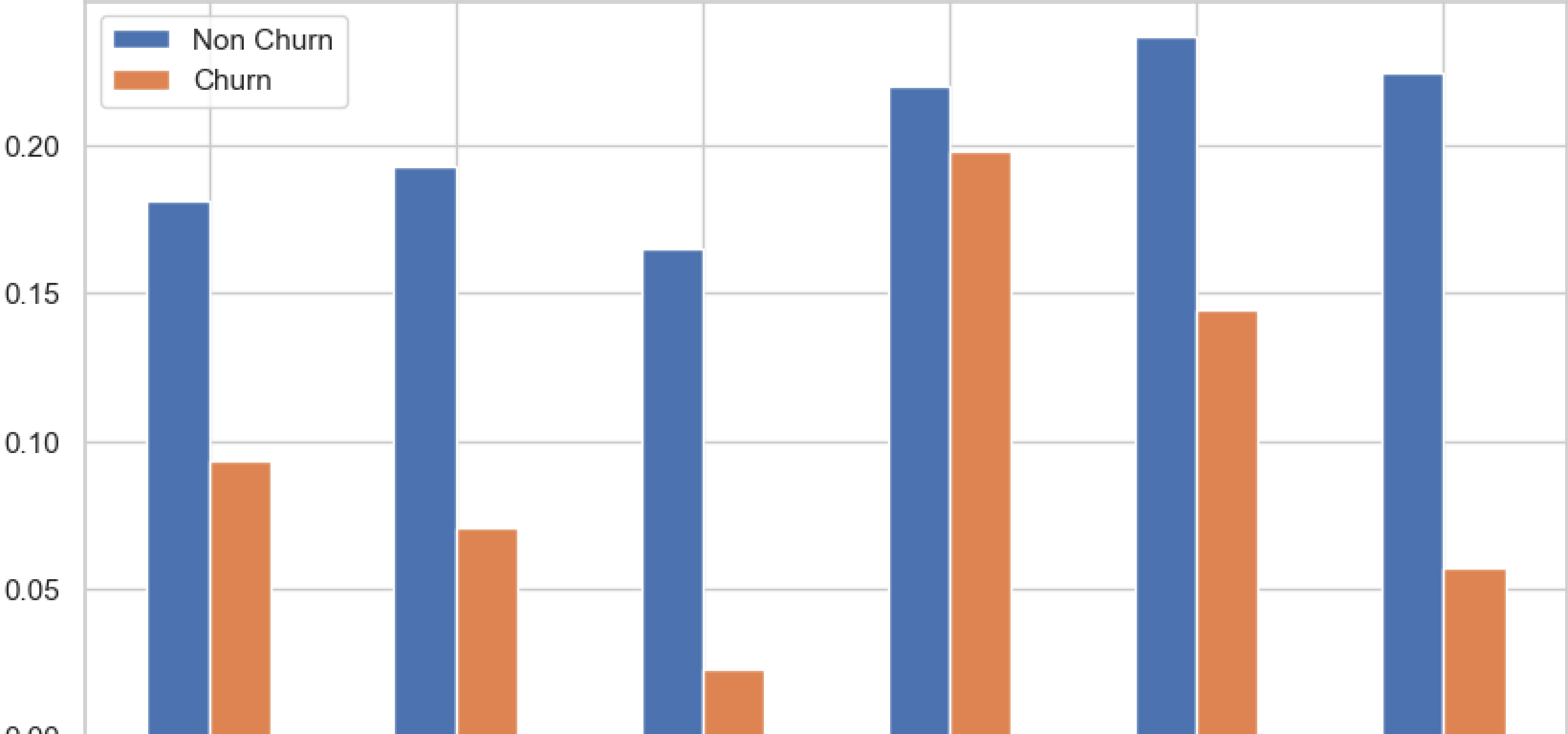


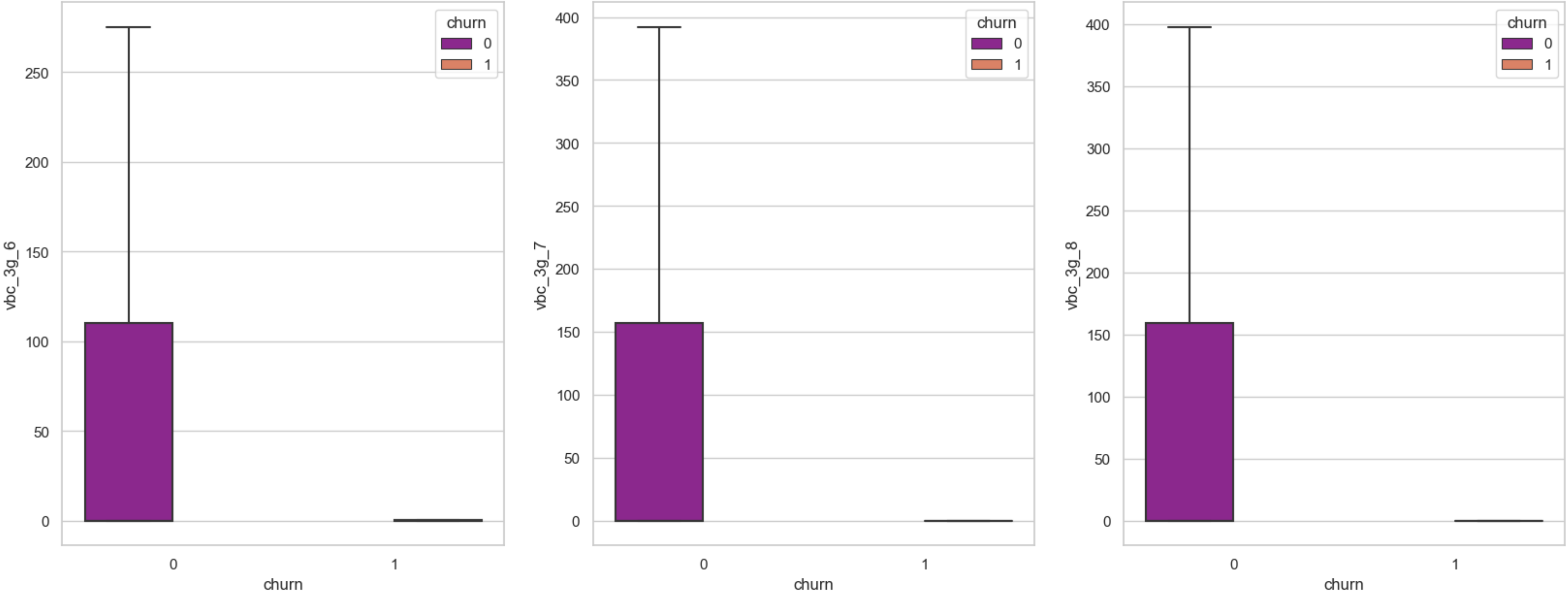




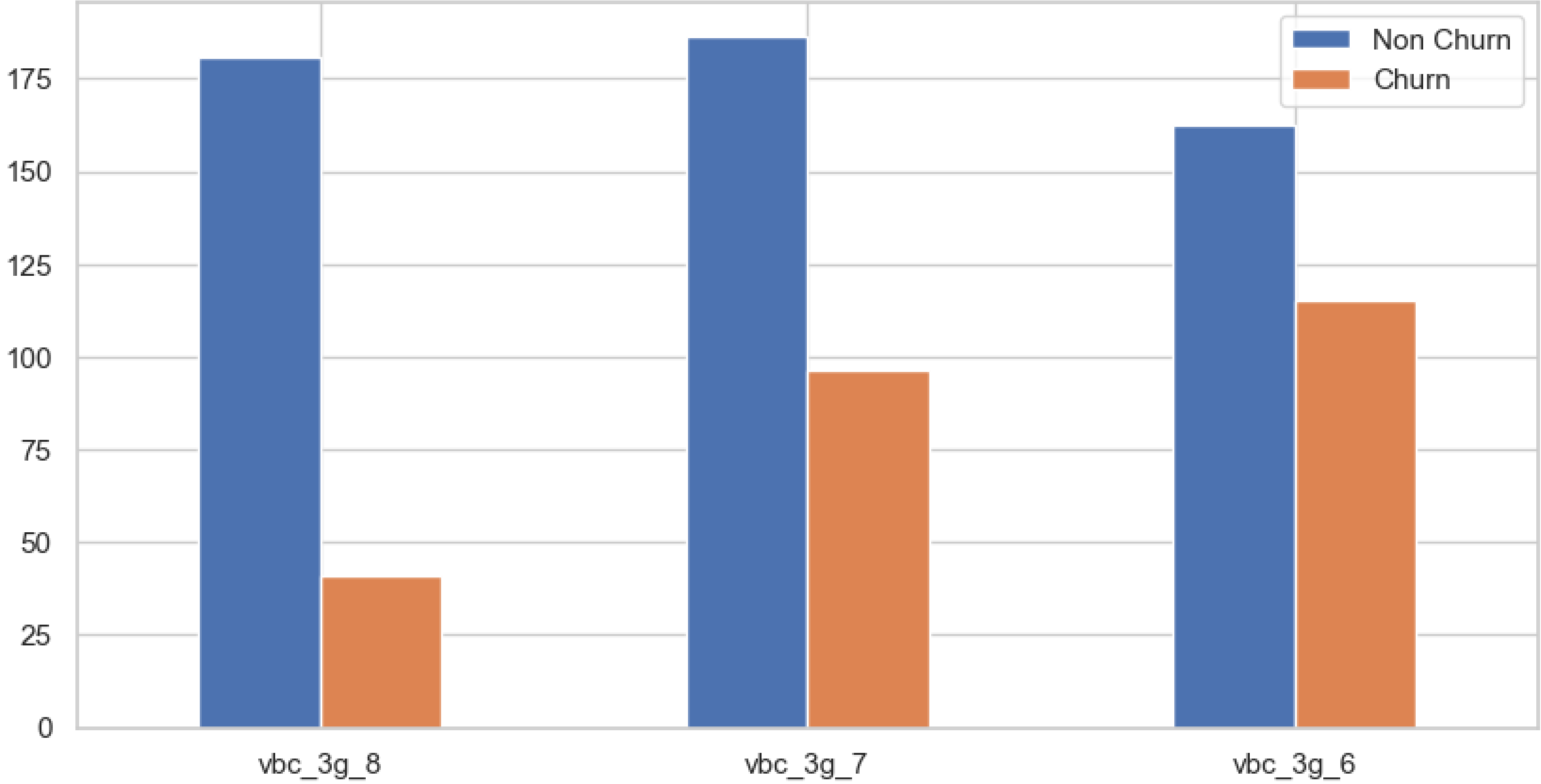
Analysis: In the eighth month, we can once more notice a decline in the number of subscribers on a monthly basis.

monthly_2g_6	monthly_2g_7	monthly_2g_8	monthly_3g_6	monthly_3g_7	monthly_3g_8	
Non Churn	0.18	0.19	0.17	0.22	0.24	0.22
Churn	0.09	0.07	0.02	0.20	0.14	0.06



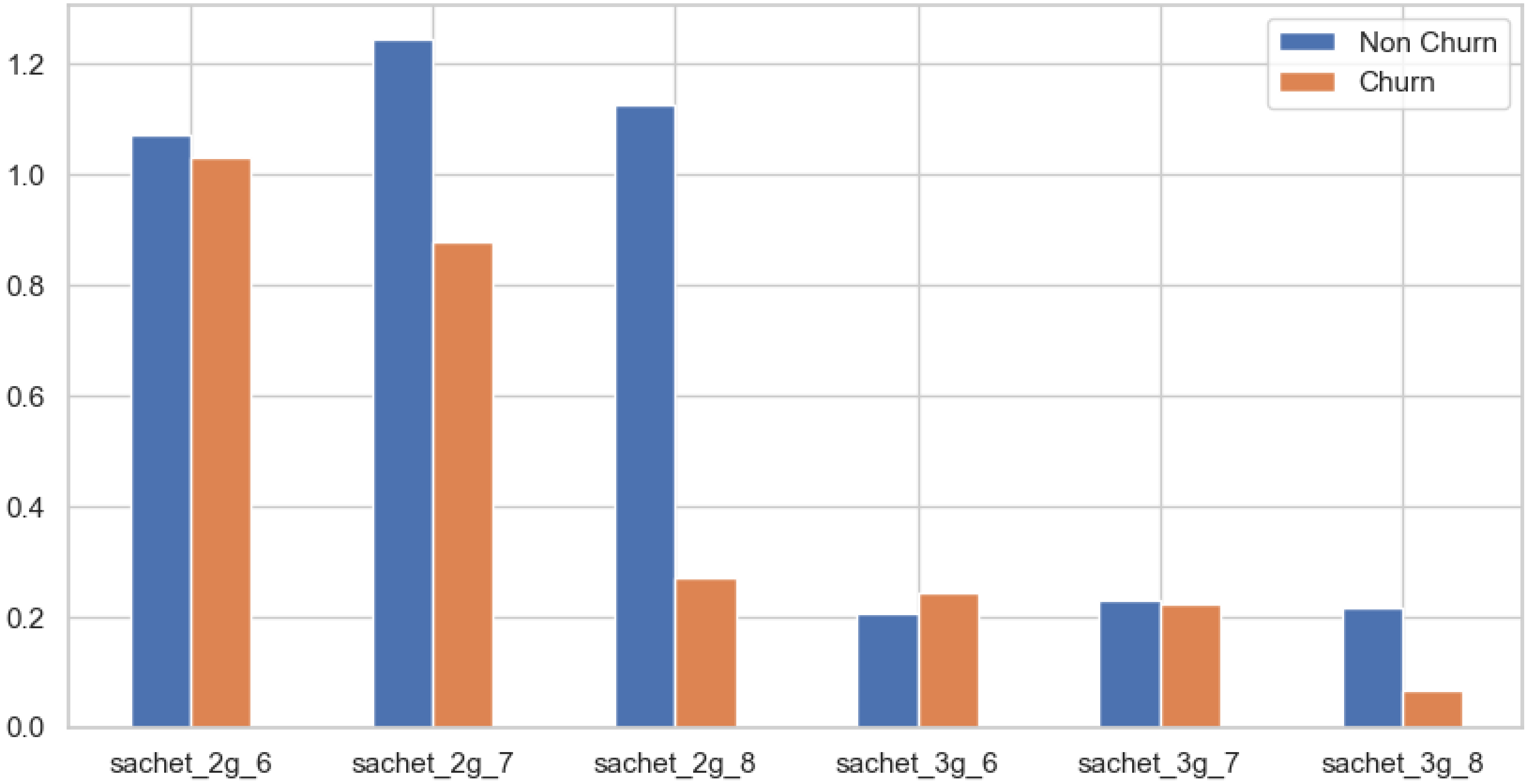

vbc_3g_6
0
50
100
150
200
250
churn
0
1
churn
0
1
vbc_3g_7
0
50
100
150
200
250
300
350
400
churn
0
1
churn
0
1
vbc_3g_8
0
50
100
150
200
250
300
350
400
churn
0
1
churn
0
1

	vbc_3g_8	vbc_3g_7	vbc_3g_6
Non Churn	180.62	186.37	162.56
Churn	40.94	96.34	115.46

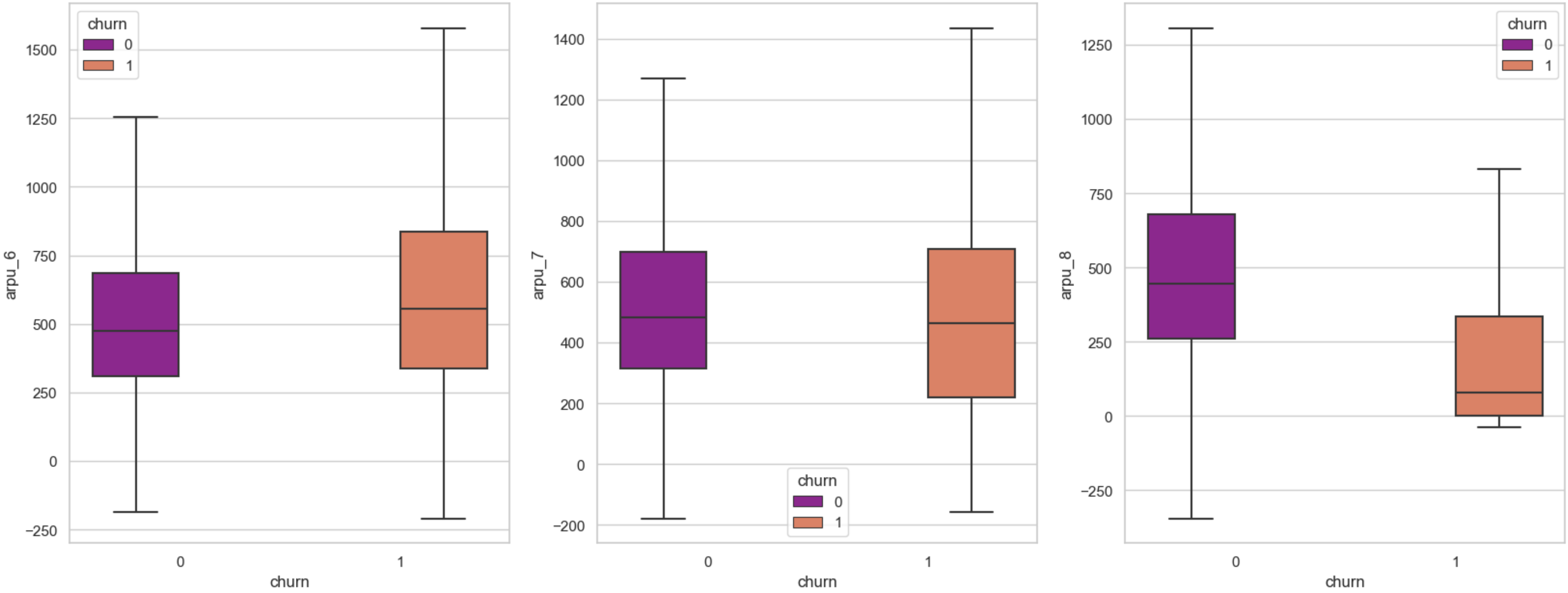


Analysis: Significantly it showing that volume based cost for 3G is much lower for Churned customers as compared to Non-Churned Customers and also there is a drop in vbc in 8th month

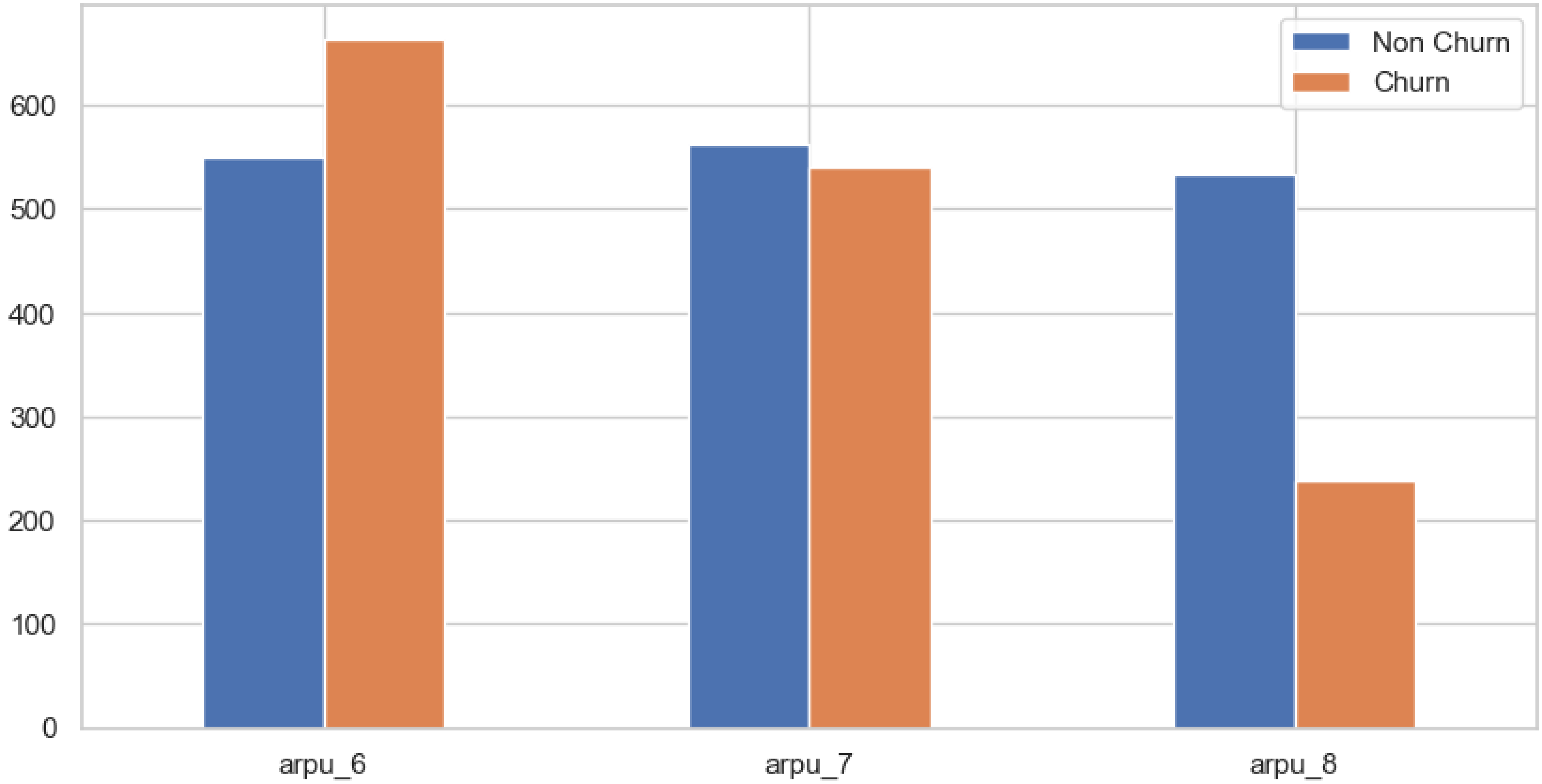
sachet_2g_6	sachet_2g_7	sachet_2g_8	sachet_3g_6	sachet_3g_7	sachet_3g_8	
Non Churn	1.07	1.25	1.13	0.21	0.23	0.2
Churn	1.03	0.88	0.27	0.24	0.22	0.0



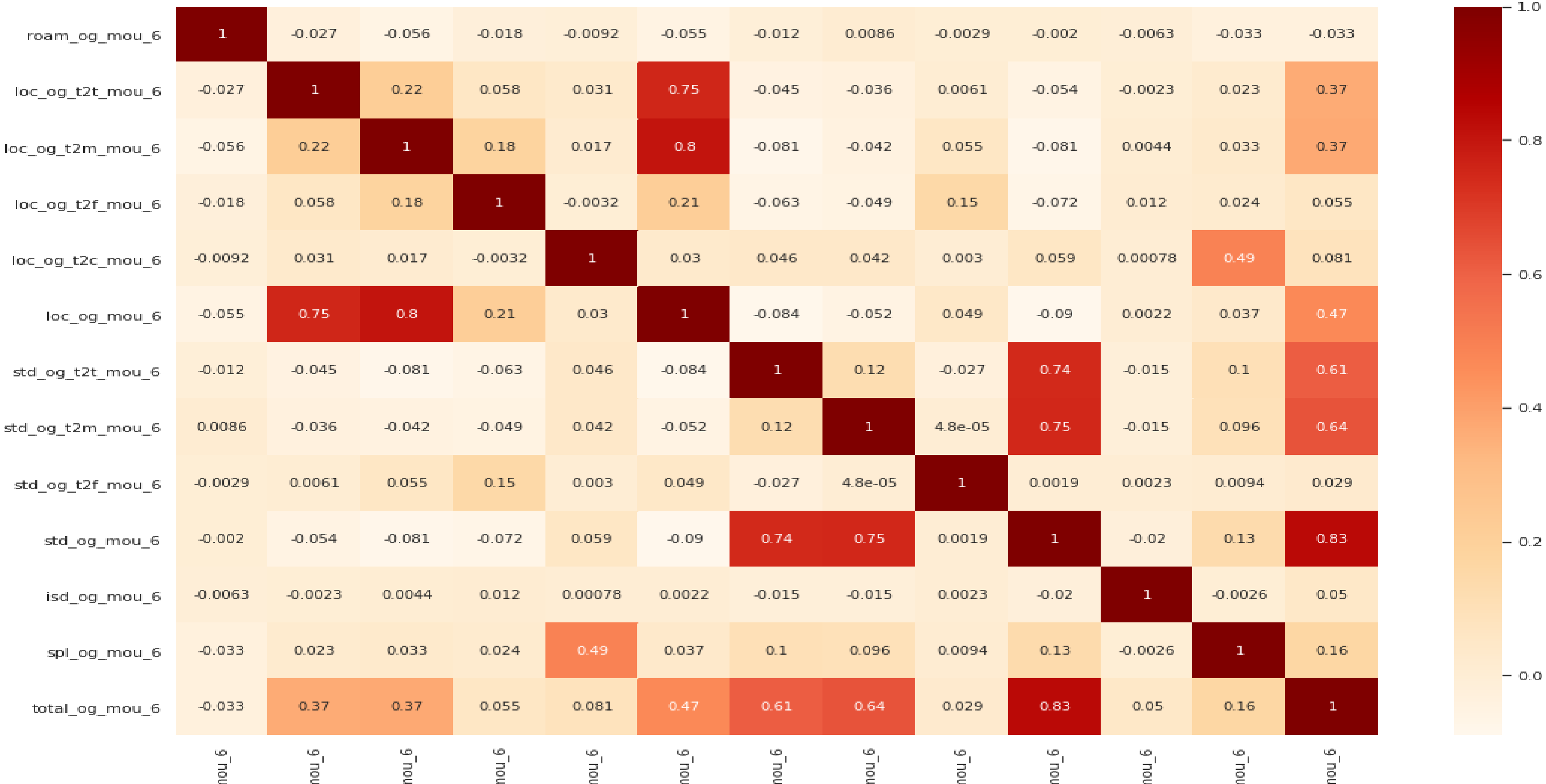
Analysis: We can see the drop in sachet services in 8th month for churned customers.


churn
0
1
1500
1250
1000
750
500
250
0
−250
arpu_6
0
1
churn
1400
1200
1000
800
600
400
200
0
−200
arpu_7
churn
0
1
0
1
churn
churn
0
1
1250
1000
750
500
250
0
−250
arpu_8
0
1
churn

arpu_6	arpu_7	arpu_8	
Non Churn	549.55	562.93	532.87
Churn	663.71	541.15	237.66

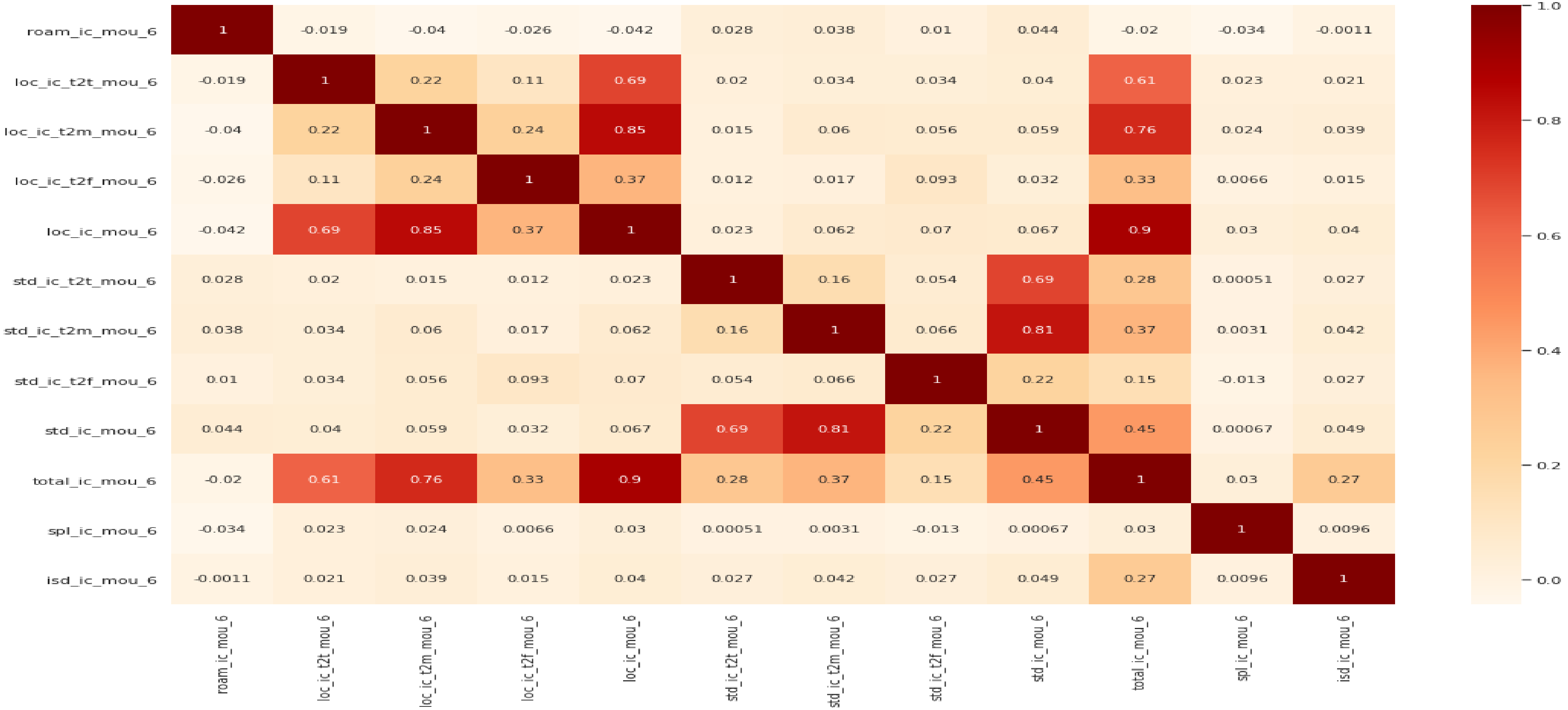


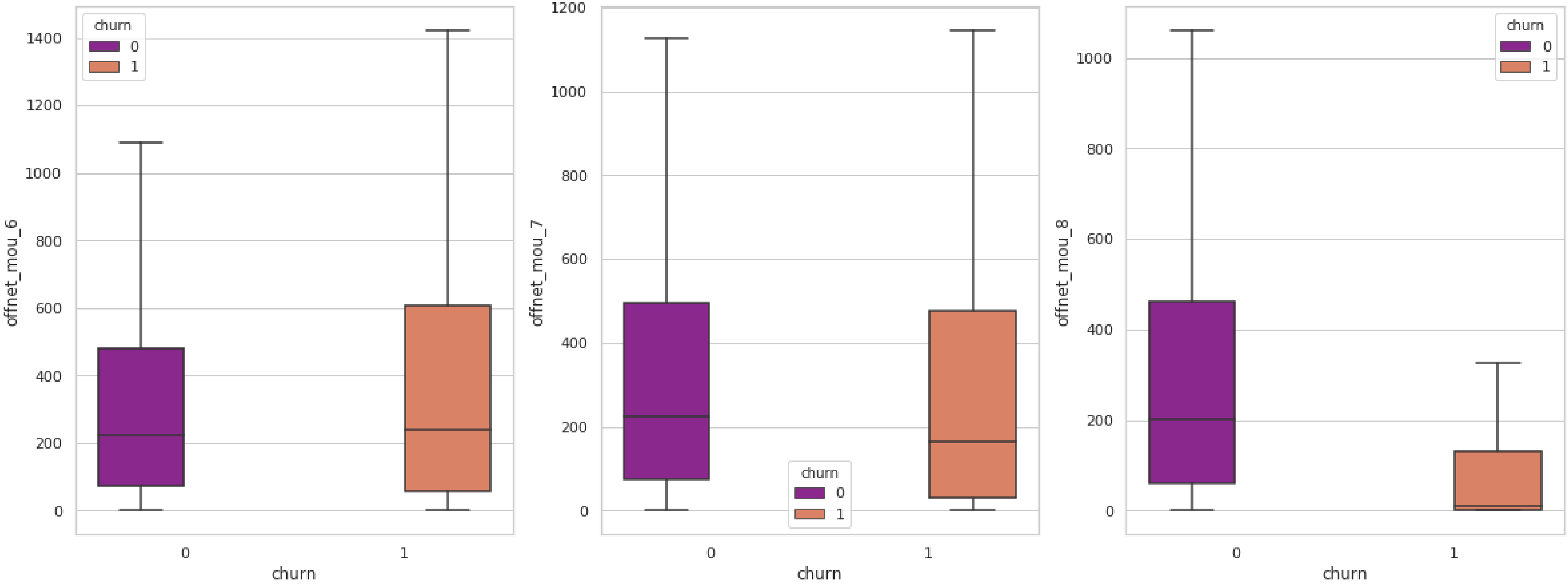
Analysis: It is clear that total_og_mou_6, standard_og_mou_6, and loc_og_mou_6 appear to have strong connection with other fields. These variables should be carefully examined to rule out multicolinearity problems.



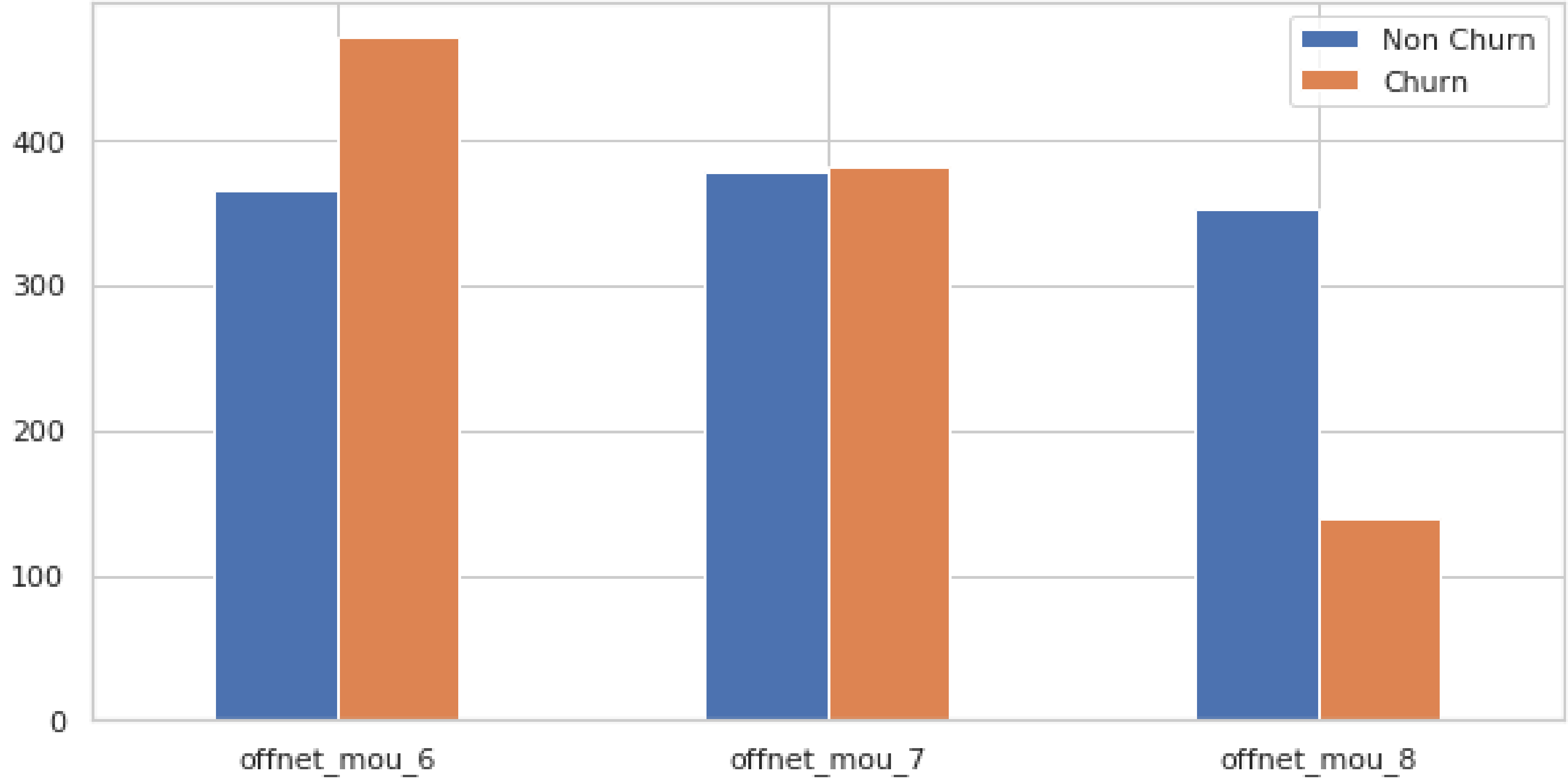
roam_og_mou_6	1	-0.027	-0.056	-0.018	-0.0092	-0.055	-0.012	0.0086	-0.0029	-0.002	-0.0063	-0.033	-0.033
loc_og_t2t_mou_6	-0.027	1	0.22	0.058	0.031	0.75	-0.045	-0.036	0.0061	-0.054	-0.0023	0.023	0.37
loc_og_t2m_mou_6	-0.056	0.22	1	0.18	0.017	0.8	-0.081	-0.042	0.055	-0.081	0.0044	0.033	0.37
loc_og_t2f_mou_6	-0.018	0.058	0.18	1	-0.0032	0.21	-0.063	-0.049	0.15	-0.072	0.012	0.024	0.055
loc_og_t2c_mou_6	-0.0092	0.031	0.017	-0.0032	1	0.03	0.046	0.042	0.003	0.059	0.00078	0.49	0.081
loc_og_mou_6	-0.055	0.75	0.8	0.21	0.03	1	-0.084	-0.052	0.049	-0.09	0.0022	0.037	0.47
std_og_t2t_mou_6	-0.012	-0.045	-0.081	-0.063	0.046	-0.084	1	0.12	-0.027	0.74	-0.015	0.1	0.61
std_og_t2m_mou_6	0.0086	-0.036	-0.042	-0.049	0.042	-0.052	0.12	1	4.8e-05	0.75	-0.015	0.096	0.64
std_og_t2f_mou_6	-0.0029	0.0061	0.055	0.15	0.003	0.049	-0.027	4.8e-05	1	0.0019	0.0023	0.0094	0.029
std_og_mou_6	-0.002	-0.054	-0.081	-0.072	0.059	-0.09	0.74	0.75	0.0019	1	-0.02	0.13	0.83
isd_og_mou_6	-0.0063	-0.0023	0.0044	0.012	0.00078	0.0022	-0.015	-0.015	0.0023	-0.02	1	-0.0026	0.05
spl_og_mou_6	-0.033	0.023	0.033	0.024	0.49	0.037	0.1	0.096	0.0094	0.13	-0.0026	1	0.16
total_og_mou_6	-0.033	0.37	0.37	0.055	0.081	0.47	0.61	0.64	0.029	0.83	0.05	0.16	1

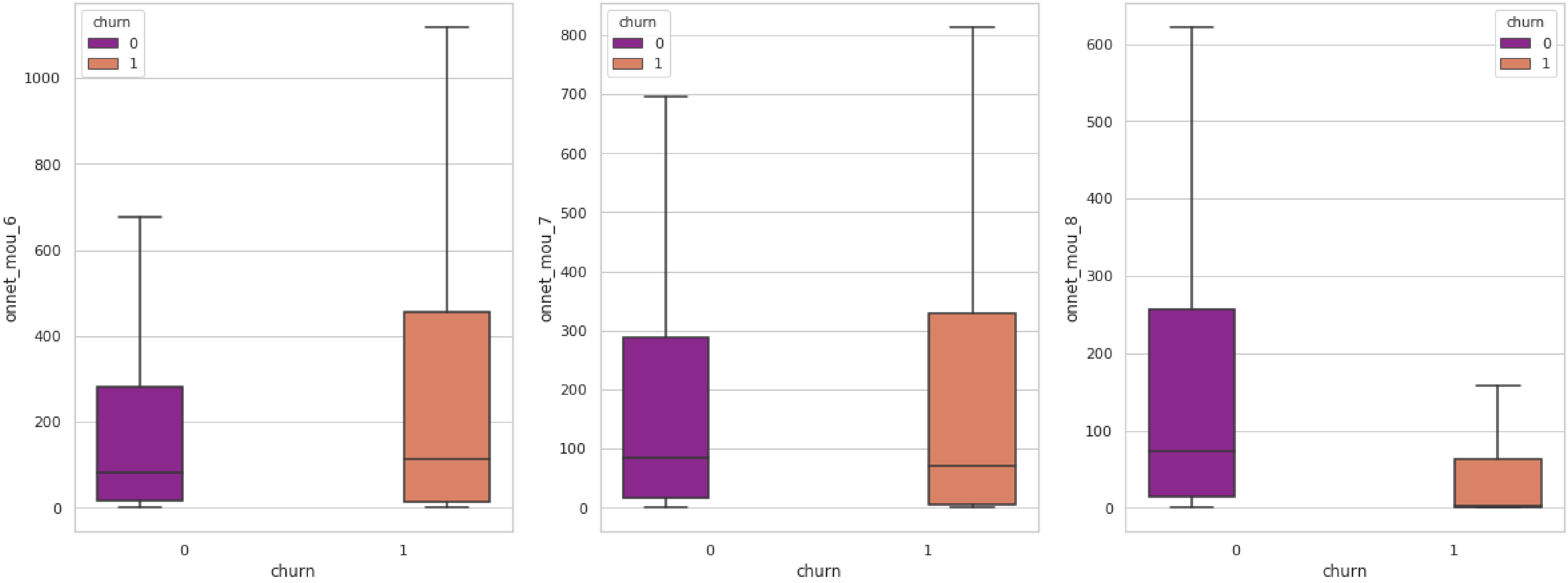
Analysis: It is clear that total_ic_mou_6, standard_ic_mou_6, and loc_ic_mou_6 appear to have a strong connection with other fields. These variables should be carefully examined to rule out multicolinearity problems.



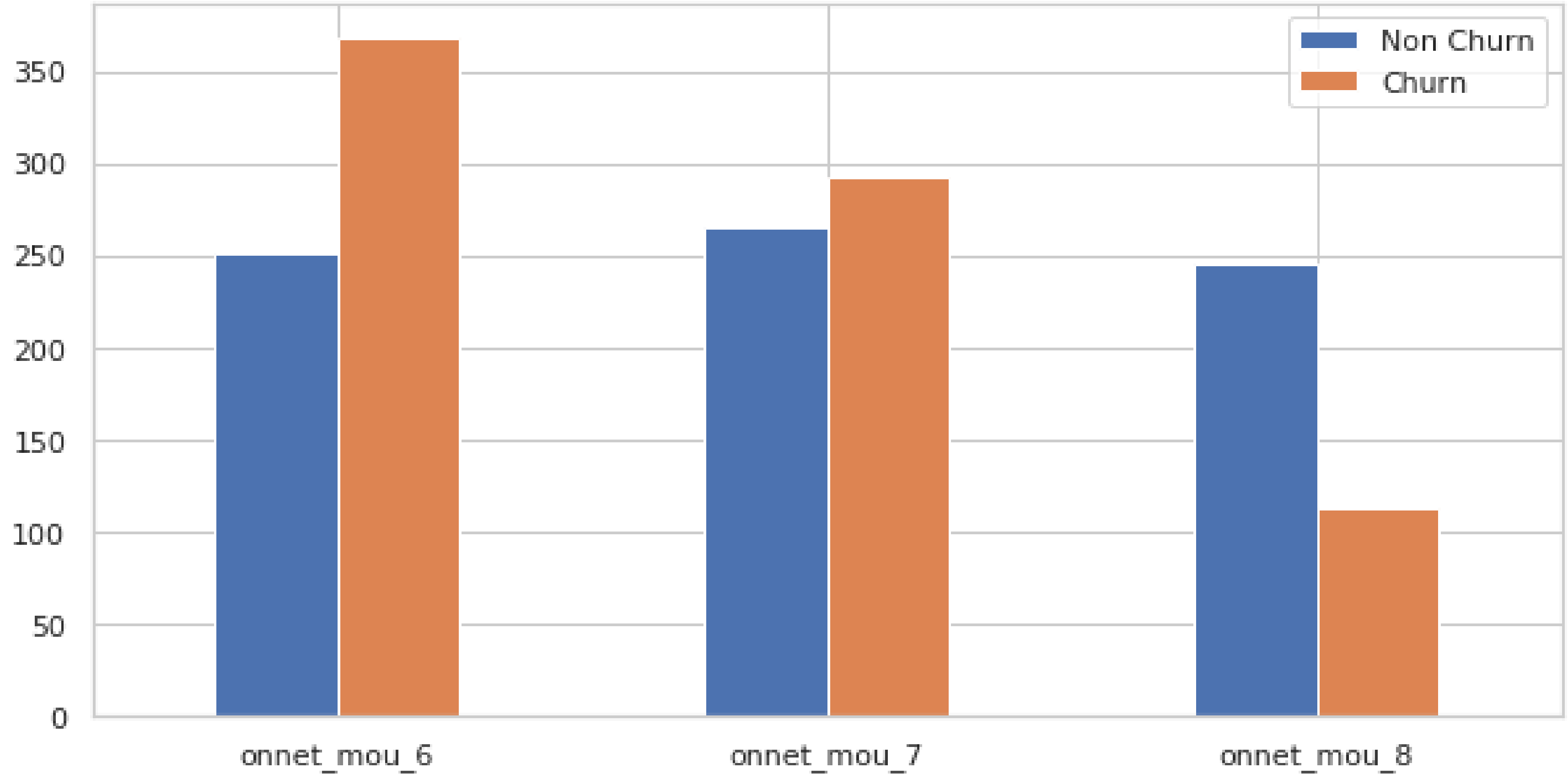

churn
0
1
offnet_mou_6
0
200
400
600
800
1000
1200
1400
0
1
churn
churn
0
1
offnet_mou_7
0
200
400
600
800
1000
1200
0
1
churn
churn
0
1
offnet_mou_8
0
200
400
600
800
1000
0
1
churn

offnet_mou_6	offnet_mou_7	offnet_mou_8	
Non Churn	365.12	377.88	352.50
Churn	471.95	382.28	138.52




churn
0
1
1000
800
600
400
200
0
onnet_mou_6
0
1
churn
churn
0
1
800
700
600
500
400
300
200
100
0
onnet_mou_7
0
1
churn
churn
0
1
600
500
400
300
200
100
0
onnet_mou_8
0
1
churn

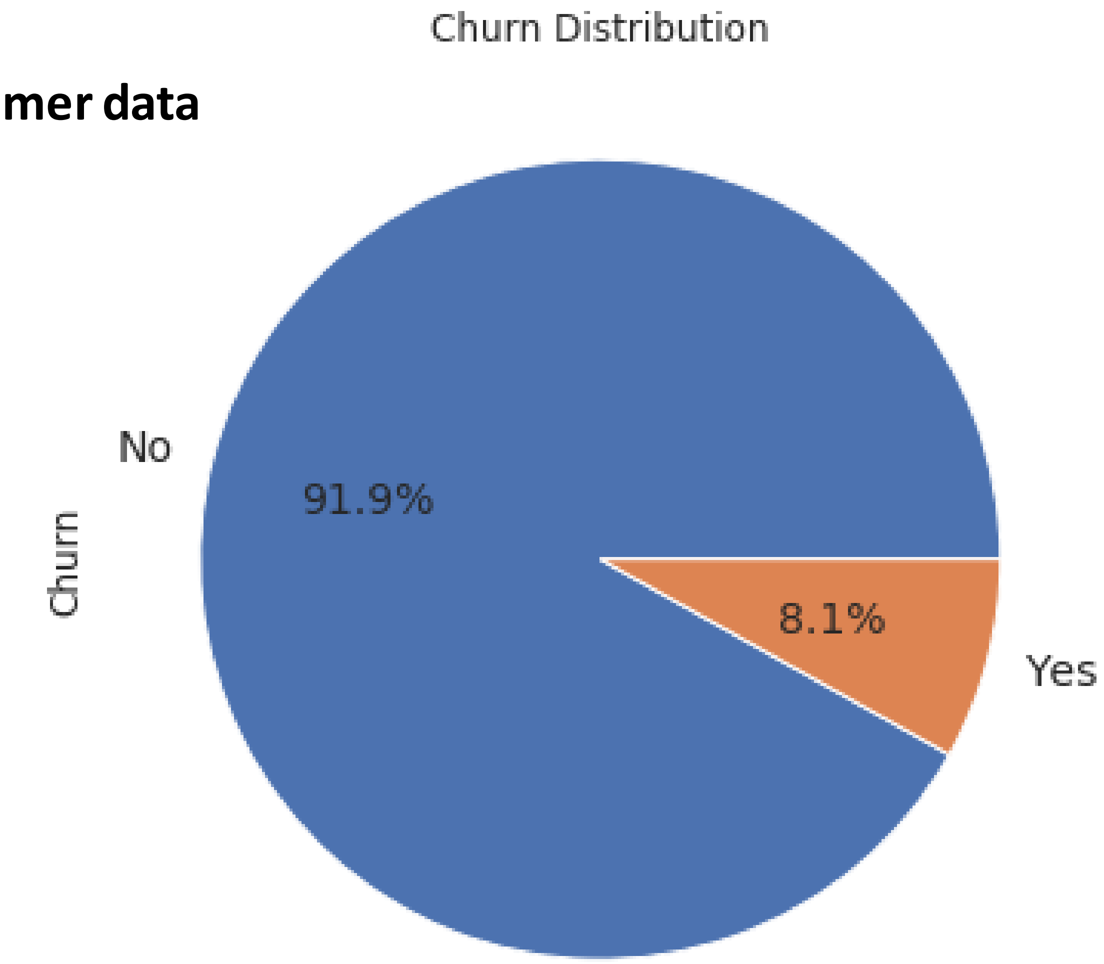
onnet_mou_6	onnet_mou_7	onnet_mou_8	
Non Churn	251.37	265.86	245.03
Churn	368.66	292.85	113.48



MODELLING

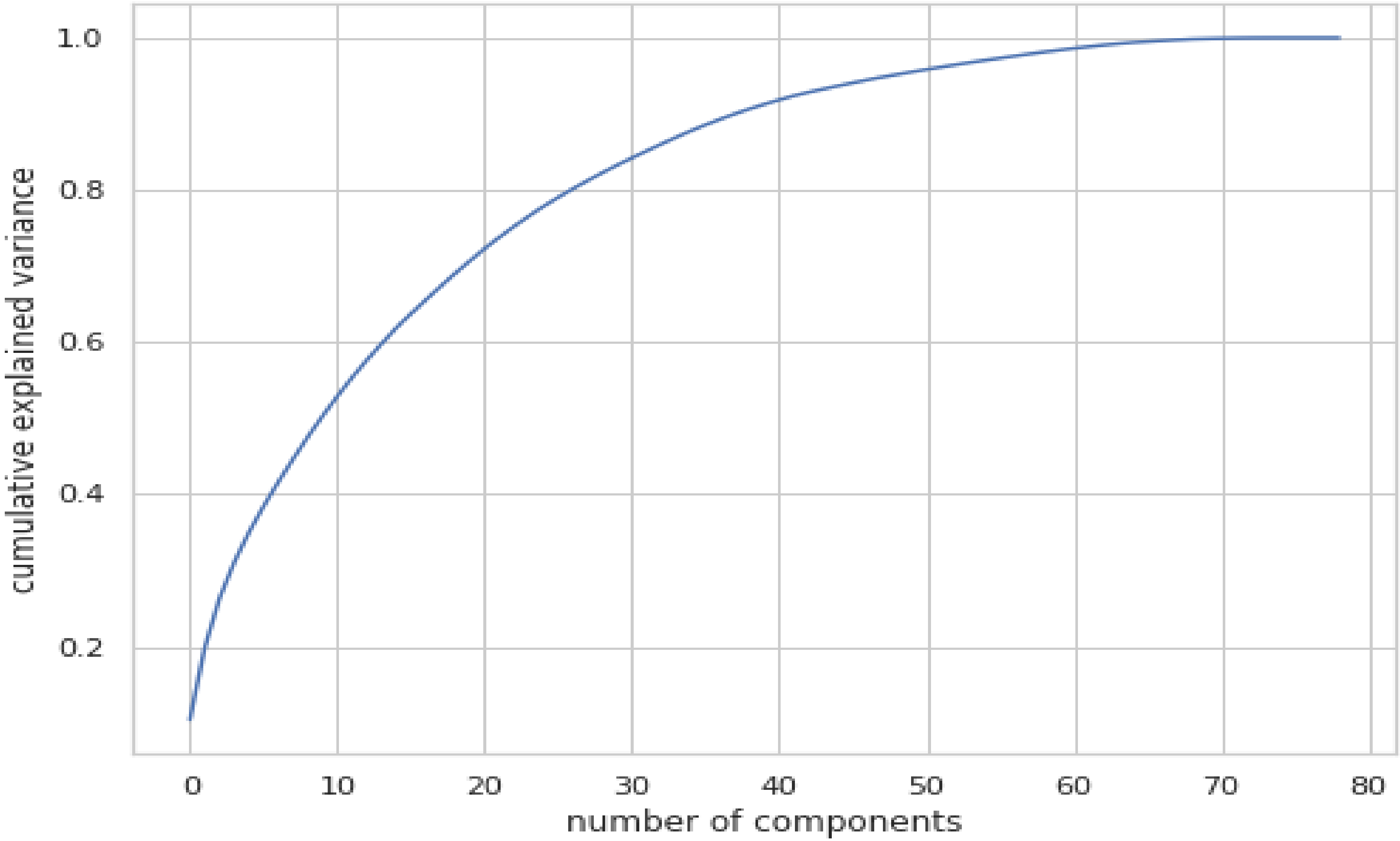
PCA : Principal Component Analysis

- **Percentage distribution of churn/non-churn customer data**



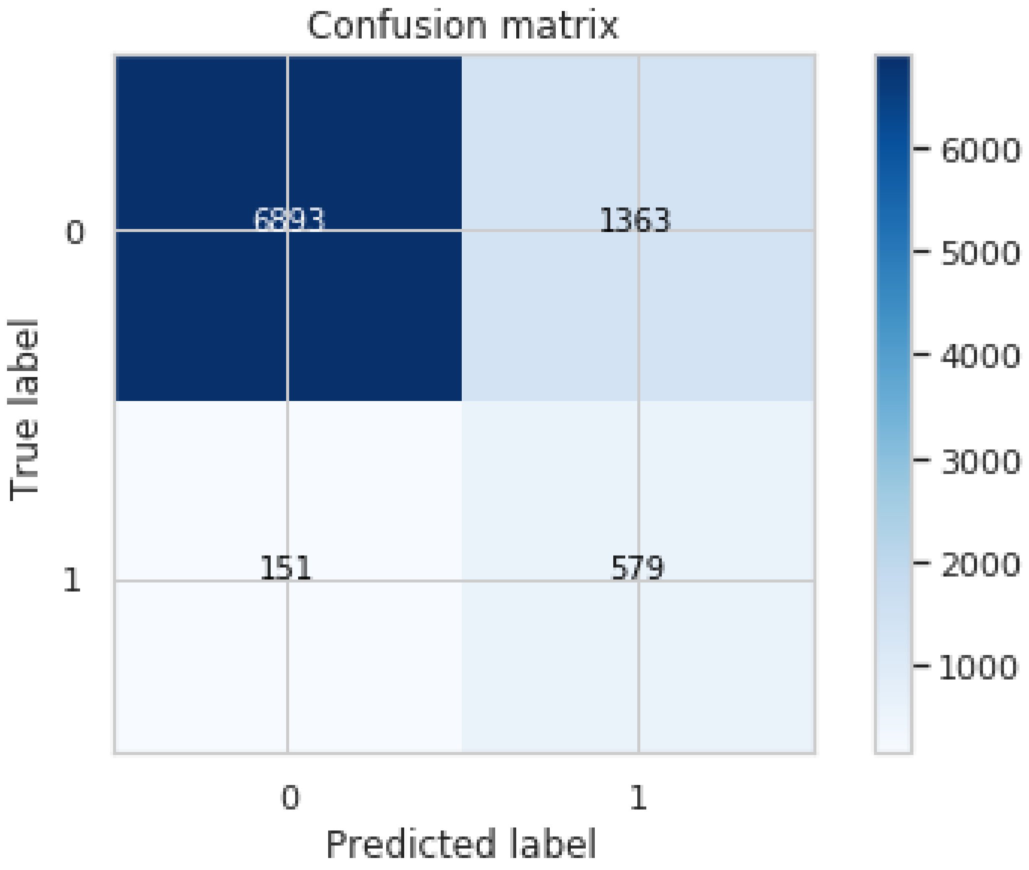
Scaling the data so as to normalize all the fields

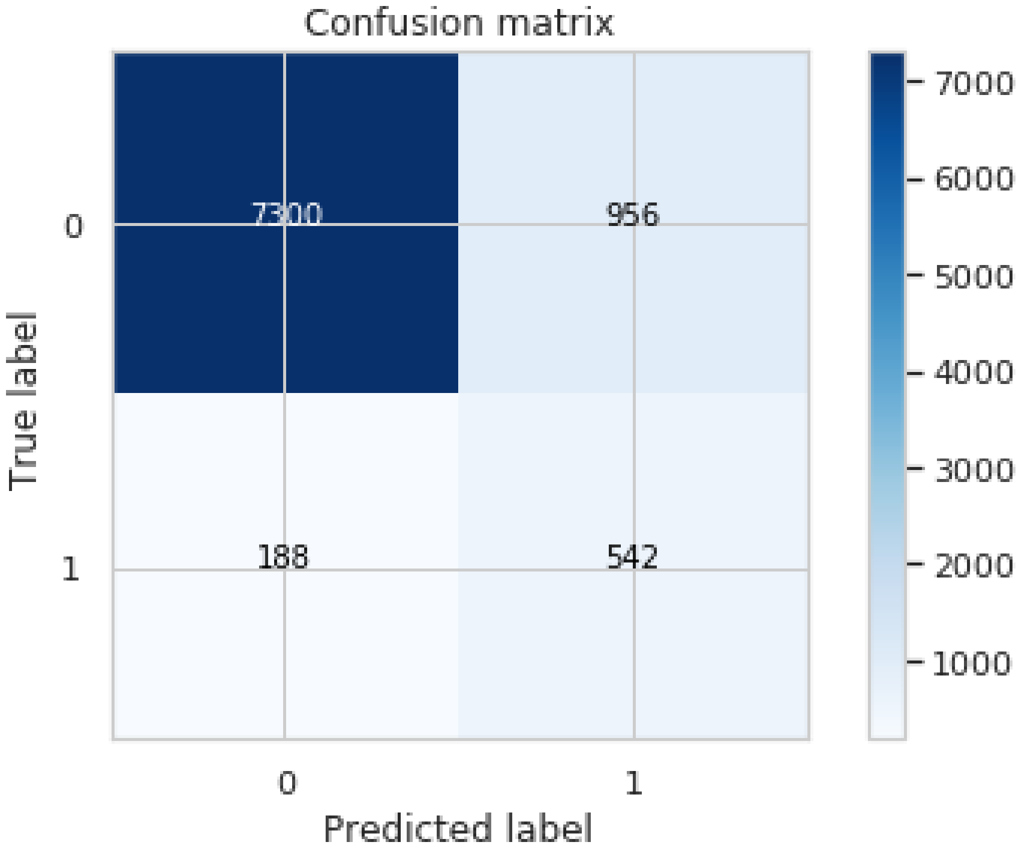
Analysis: Looks like 60 components are enough to describe 95% of the variance in the dataset. We'll choose **60** components for our modeling



Analysis: SMOTE bloated the dataset and balanced it by adding skewed data values.

SVM Regression Modelling

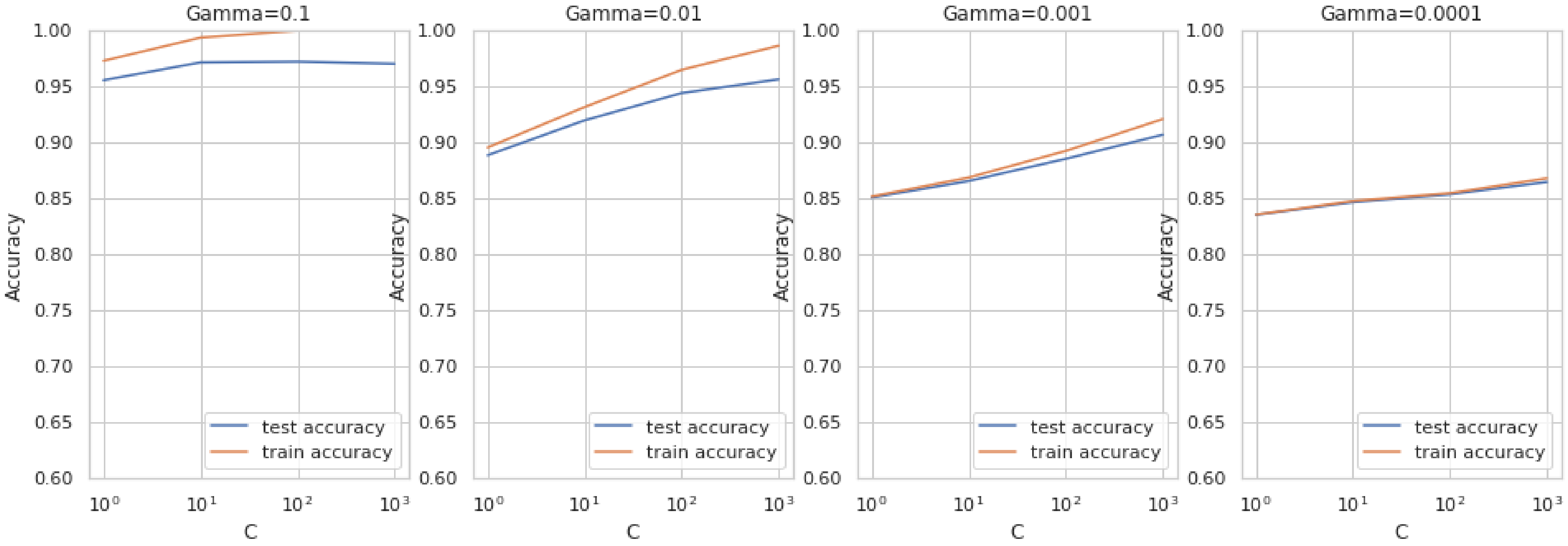




Analysis: The non-linear model gives approx. 87% accuracy. Thus, going forward, let's choose hyperparameters corresponding to non-linear models.

Grid Search: Hyperparameter Tuning

Let's now tune the model to find the optimal values of C and gamma corresponding to an RBF kernel. We'll use 5-fold cross validation.



Analysis: The plots above show some useful insights:

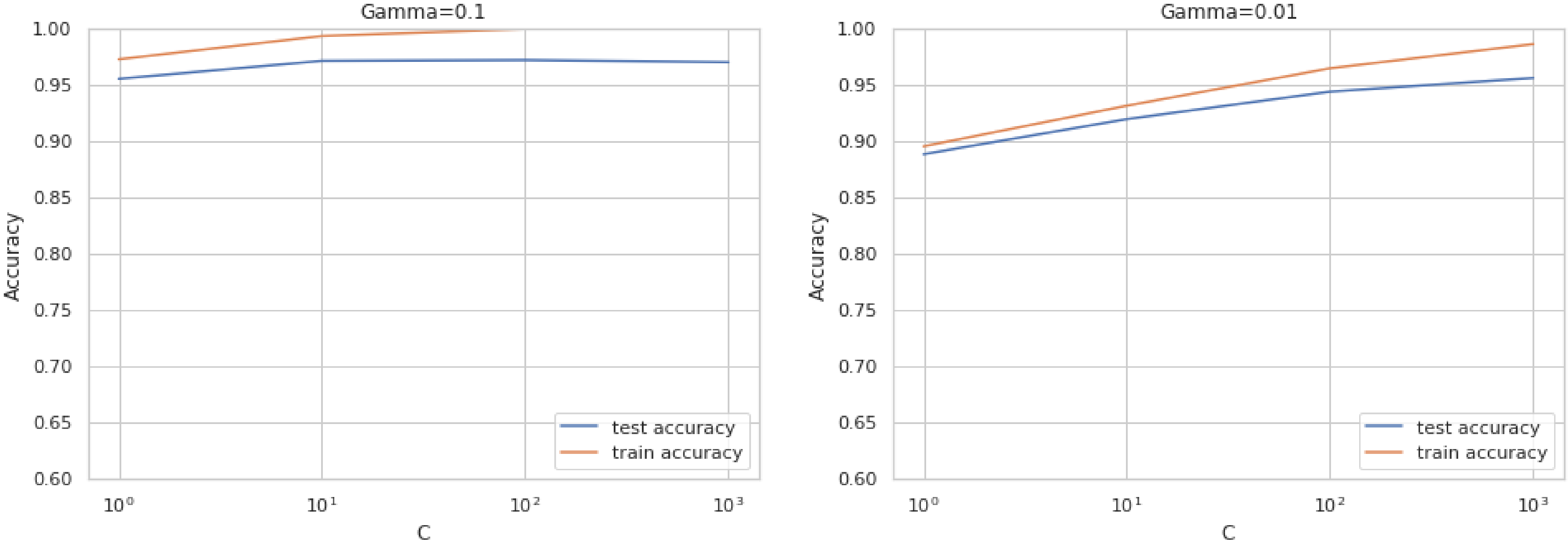
Analysis: The charts above provide the following key information:

Model with gamma=0.1 appears to be overfitting and should not be utilised based on the given curves and data.

The model was chosen with the best values of "C": 100 and "gamma": 0.1, although it also appears to be overfitting.

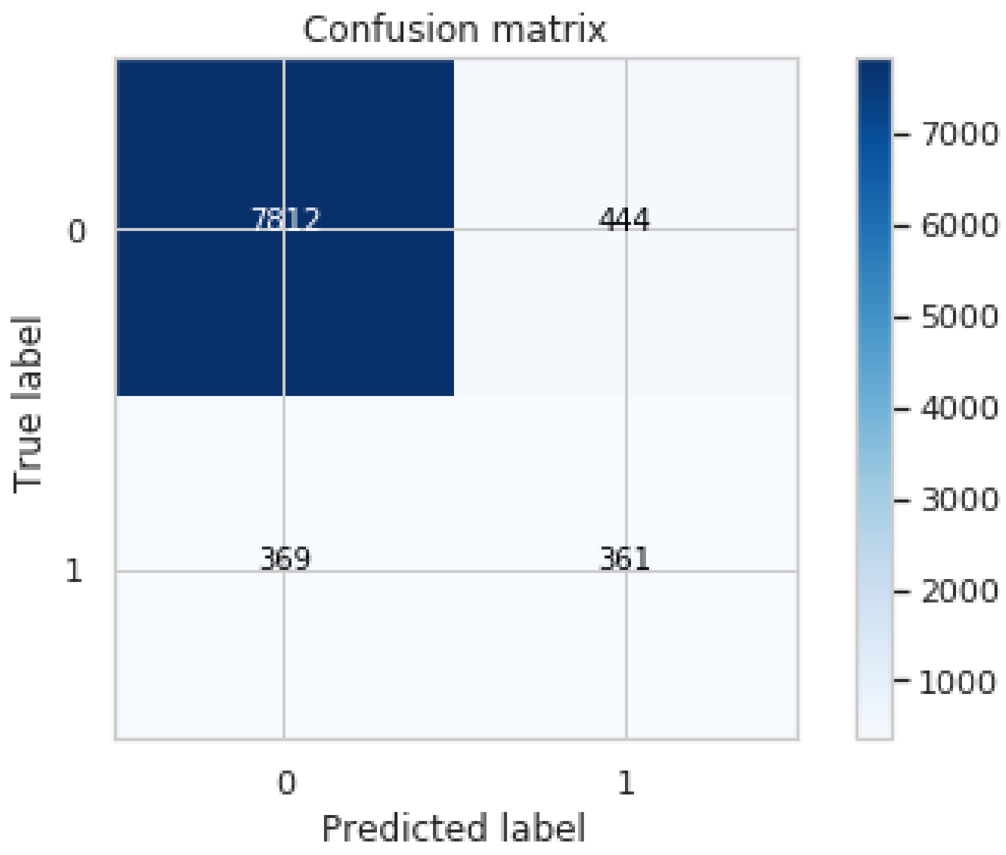
However, a model chosen with the values "C": 10 and "gamma": 0.1 ought to yield superior outcomes.

These values will be chosen for the final modelling.



FINAL MODELLING

Tree Model Regression

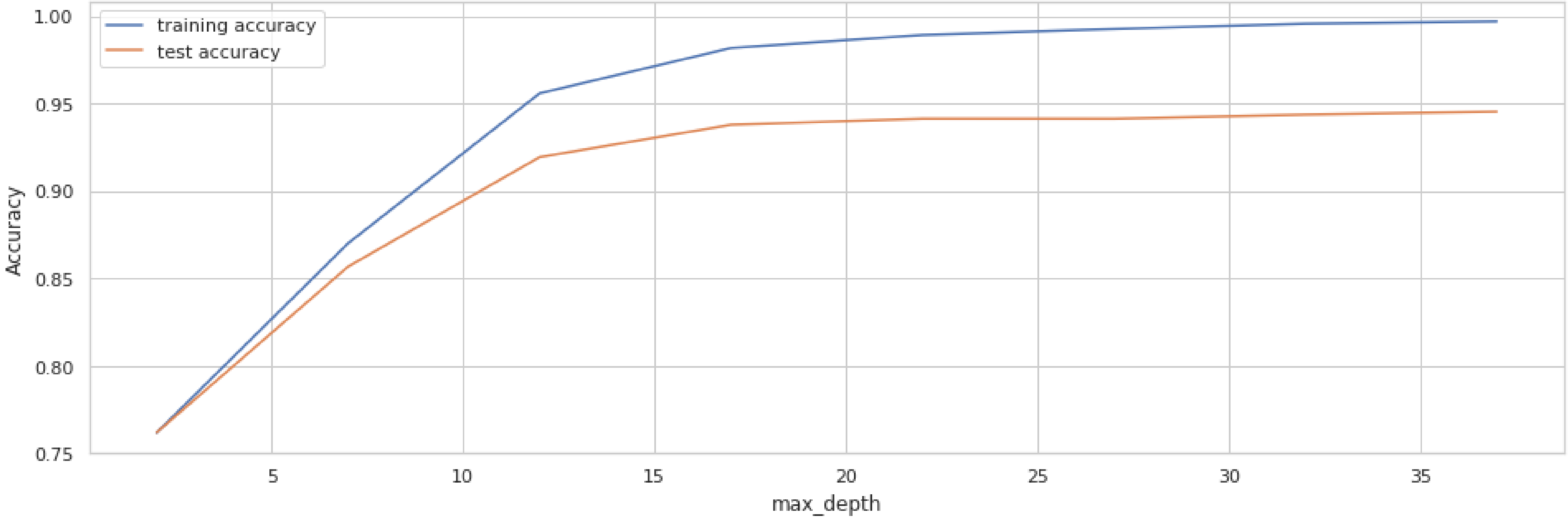


Analysis: Everything is going well so far. Let's look at the list of hyperparameters we may adjust to enhance model performance.

Model	Accuracy	Precision	Recall	AUC	F1	
0	Random Forest (Default)	0.91	0.49	0.45	0.72	0.47

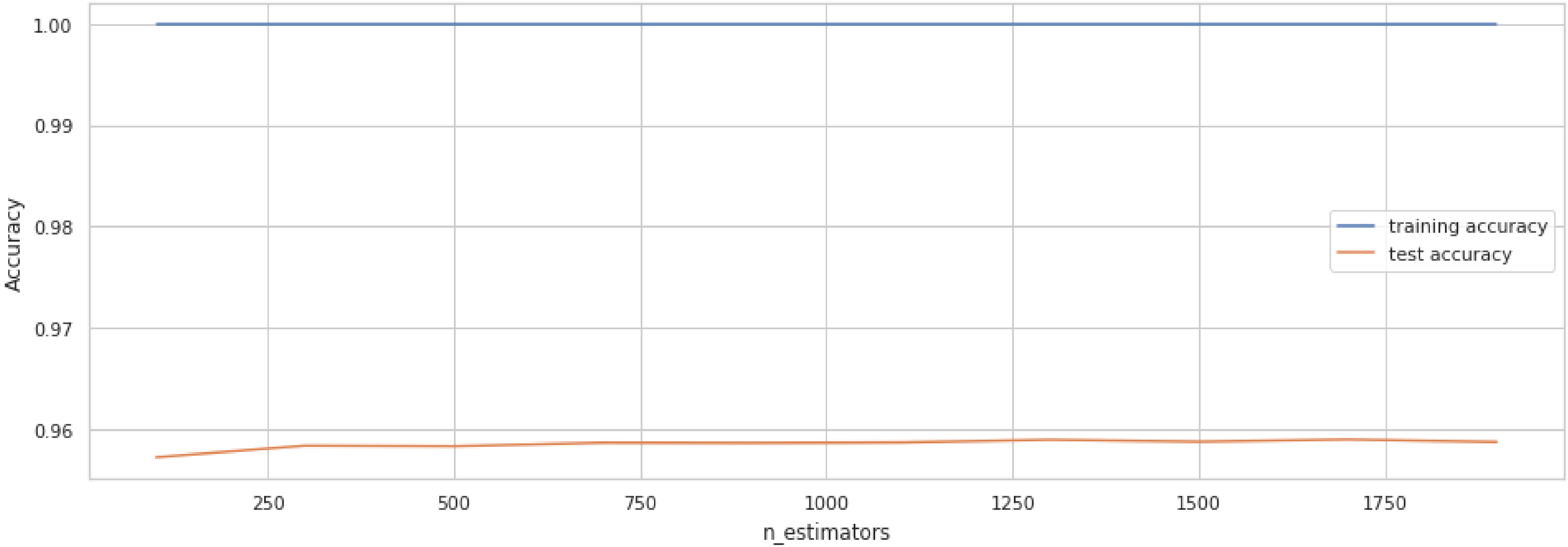
TUNING OF HYPERPARAMETER

- **Tuning max_depth**
- Let's try to find the optimum values for `max_depth` and understand how the value of max_depth impacts the overall accuracy of the ensemble.



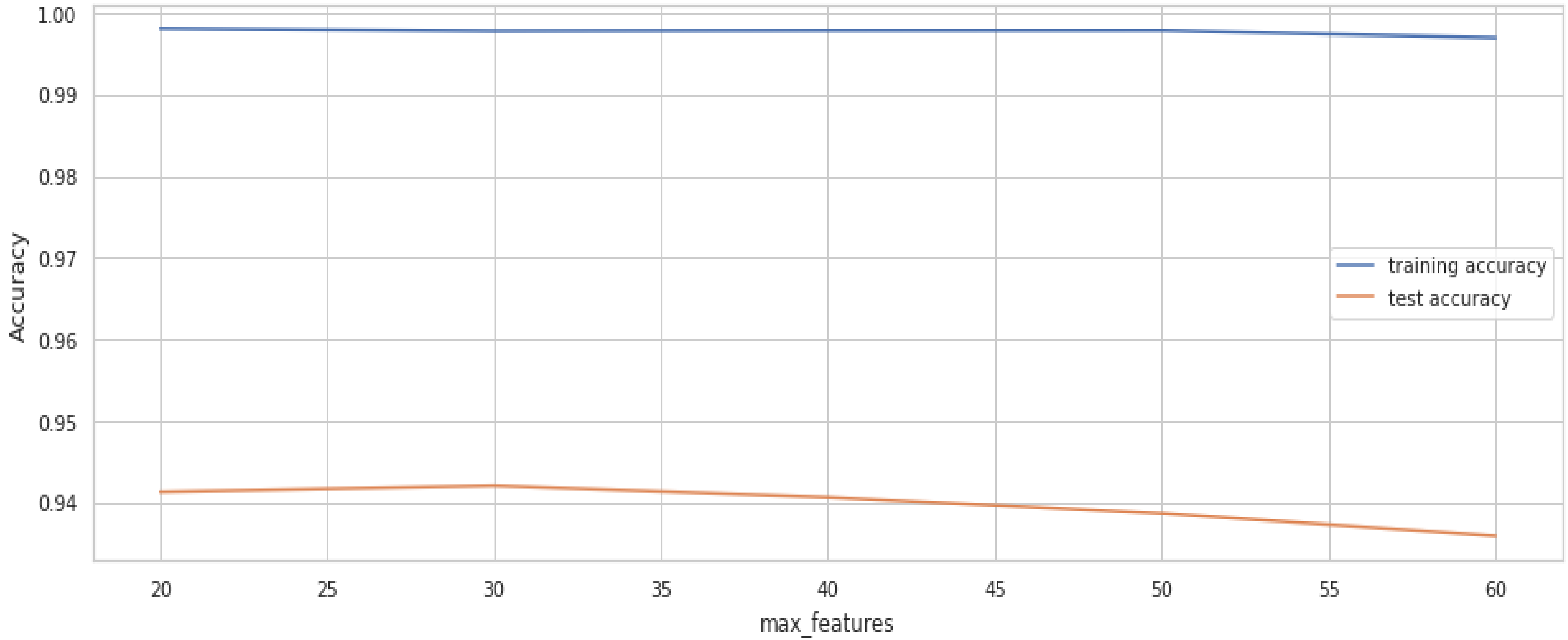
Analysis: As we increase the value of max_depth, we can see that both the train and test scores rise up to a certain point, but then the test score becomes stagnant. As we raise the max_depth, the ensemble tries to overfit. In order to lessen overfitting in the forest, it is possible to adjust the depth of the constituent trees. Peak convergens values of 12 and 18 can be used for grid view search.

Estimators for Tuning

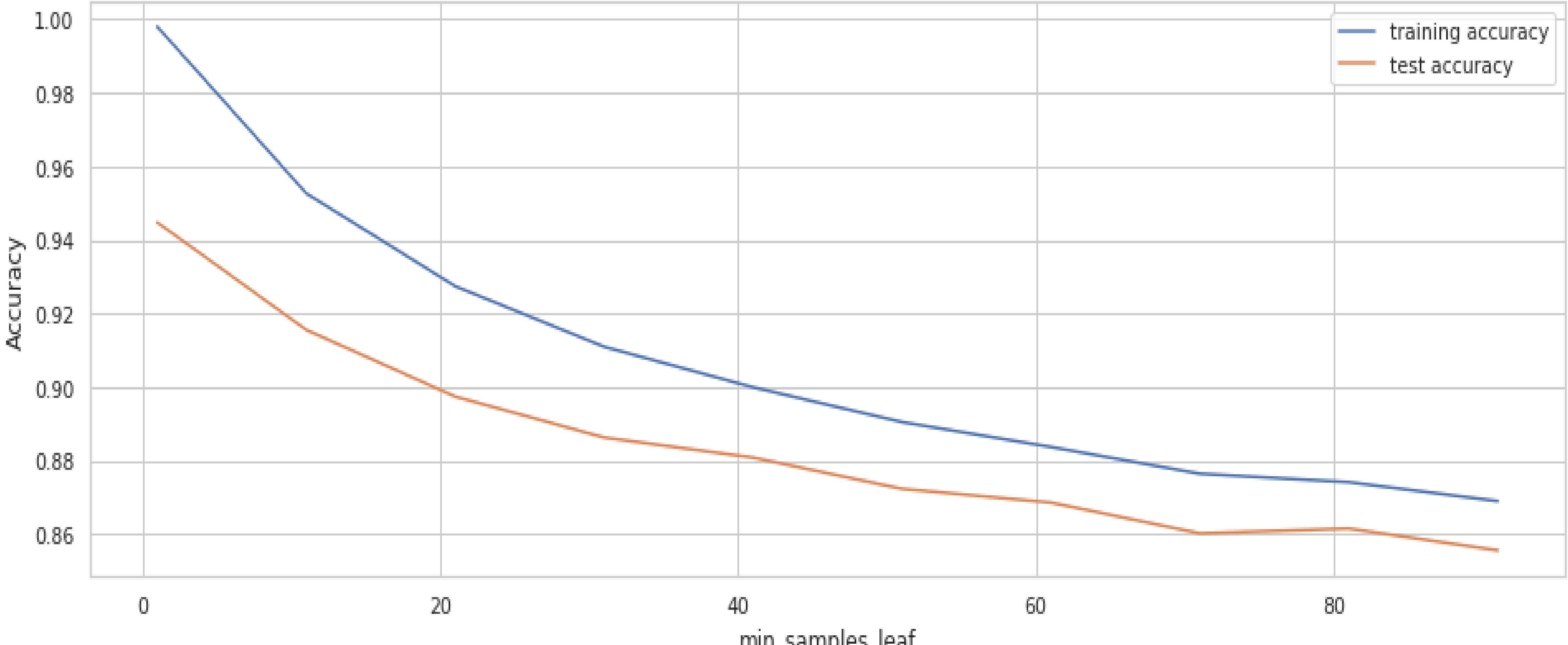


Analysis: The score nearly held steady throughout the range with very little dipping. To search the grid view, we'll utilize 200.

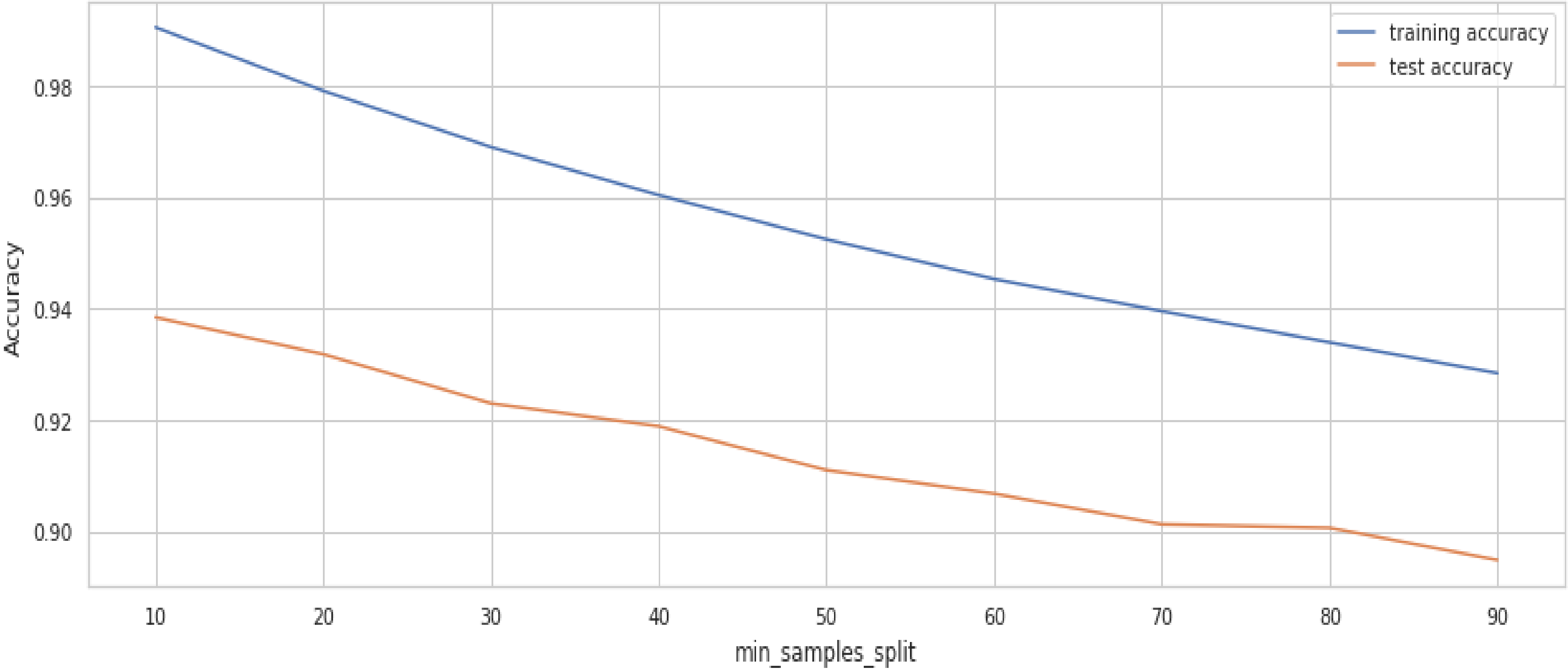
Analysis: It appears that training accuracy is steady, and test scores improve until they reach 30 before declining. We shall employ the rise on 40 that we notice once again.



Analysis: As the value of min_samples_leaf decreases, we can see that the model begins to overfit. The grid search will utilize the range of 10 to 20, which seems to be a good one.

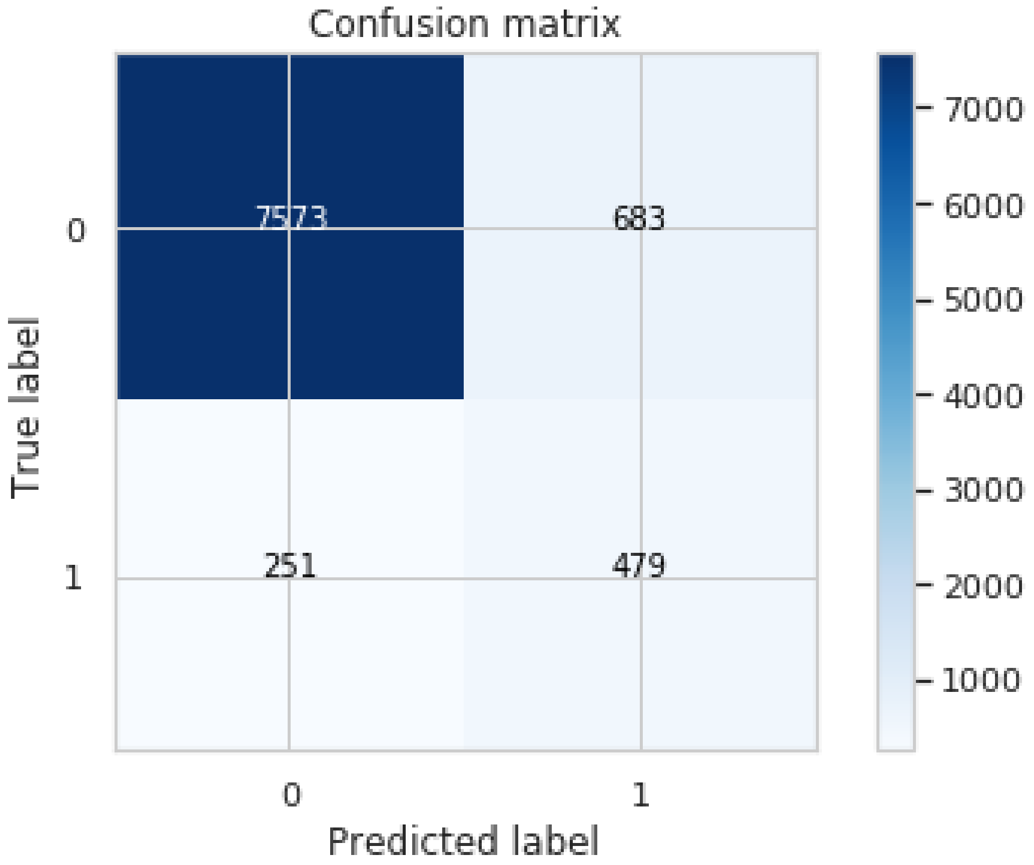


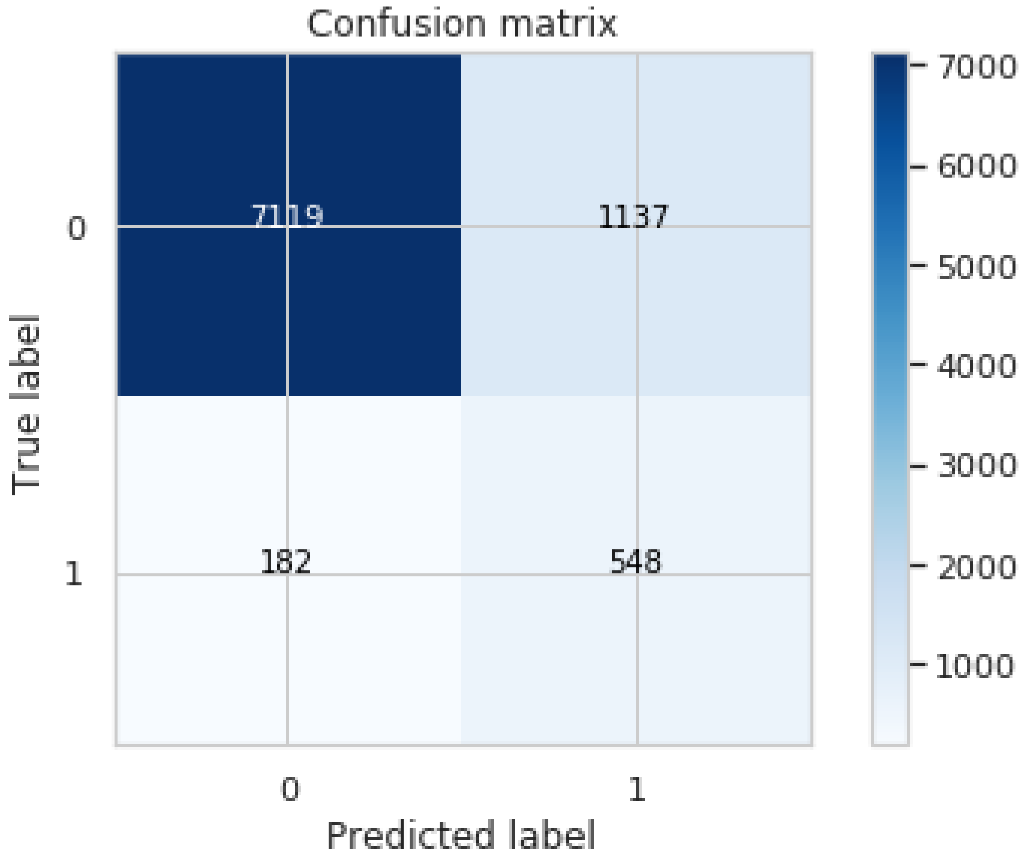
Range **10 to 30** is optimal with good accuracy.



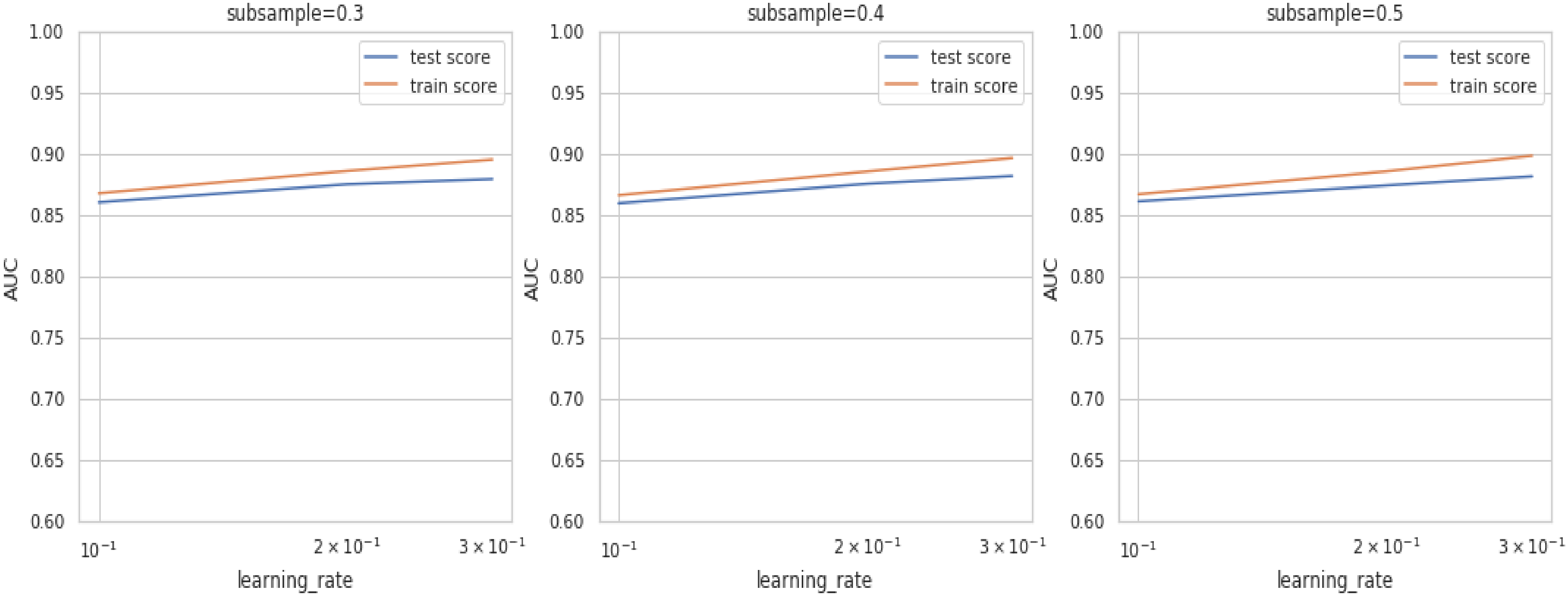
Final Model for Random Forest

XGBOOST: Analysis: Using the default hyperparameters, the roc_auc in this instance is roughly 85%.

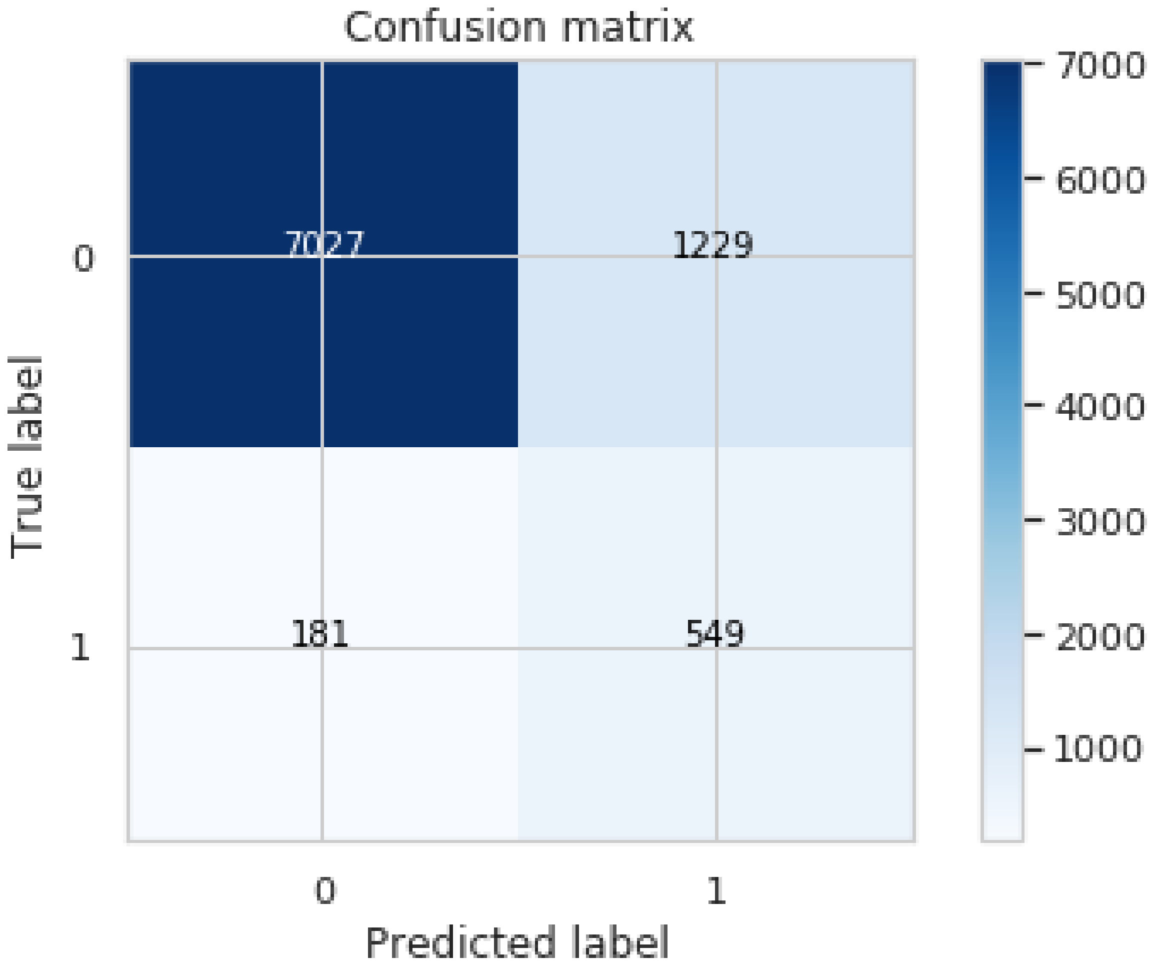


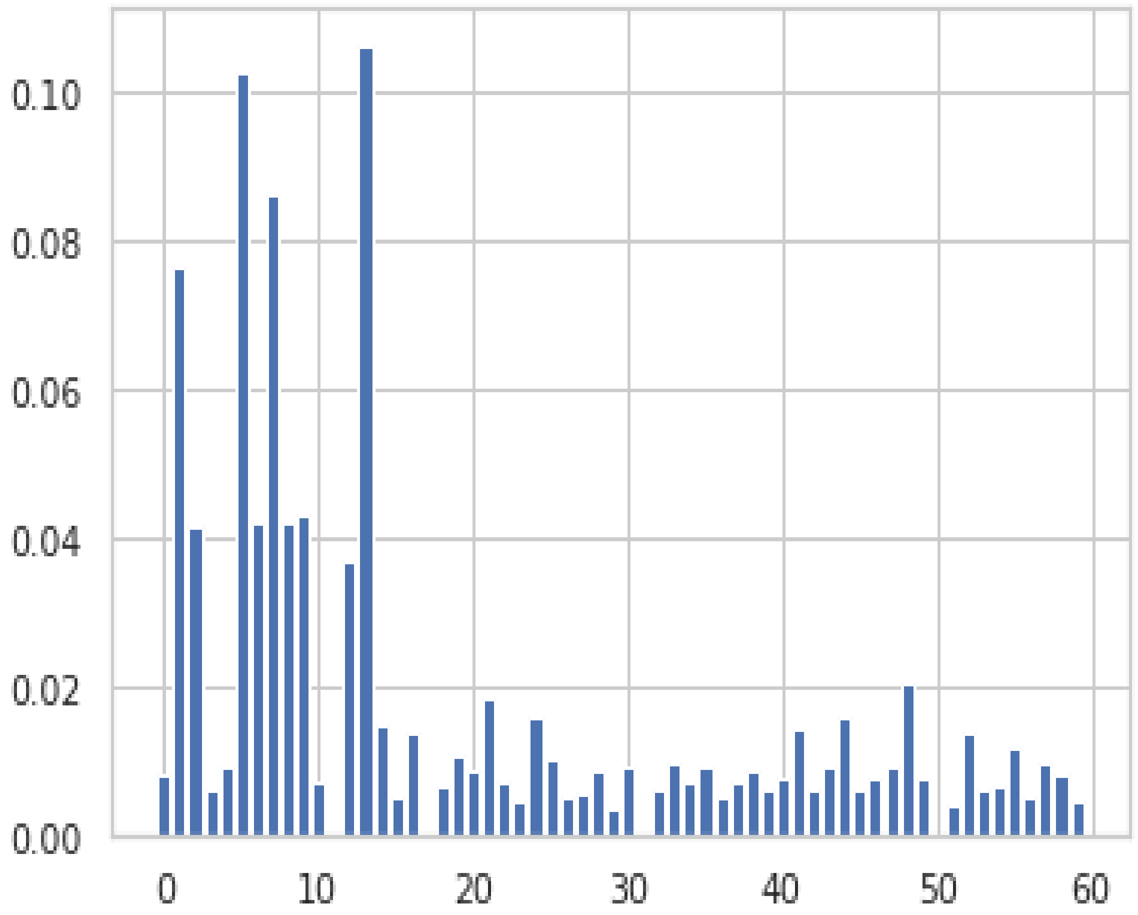


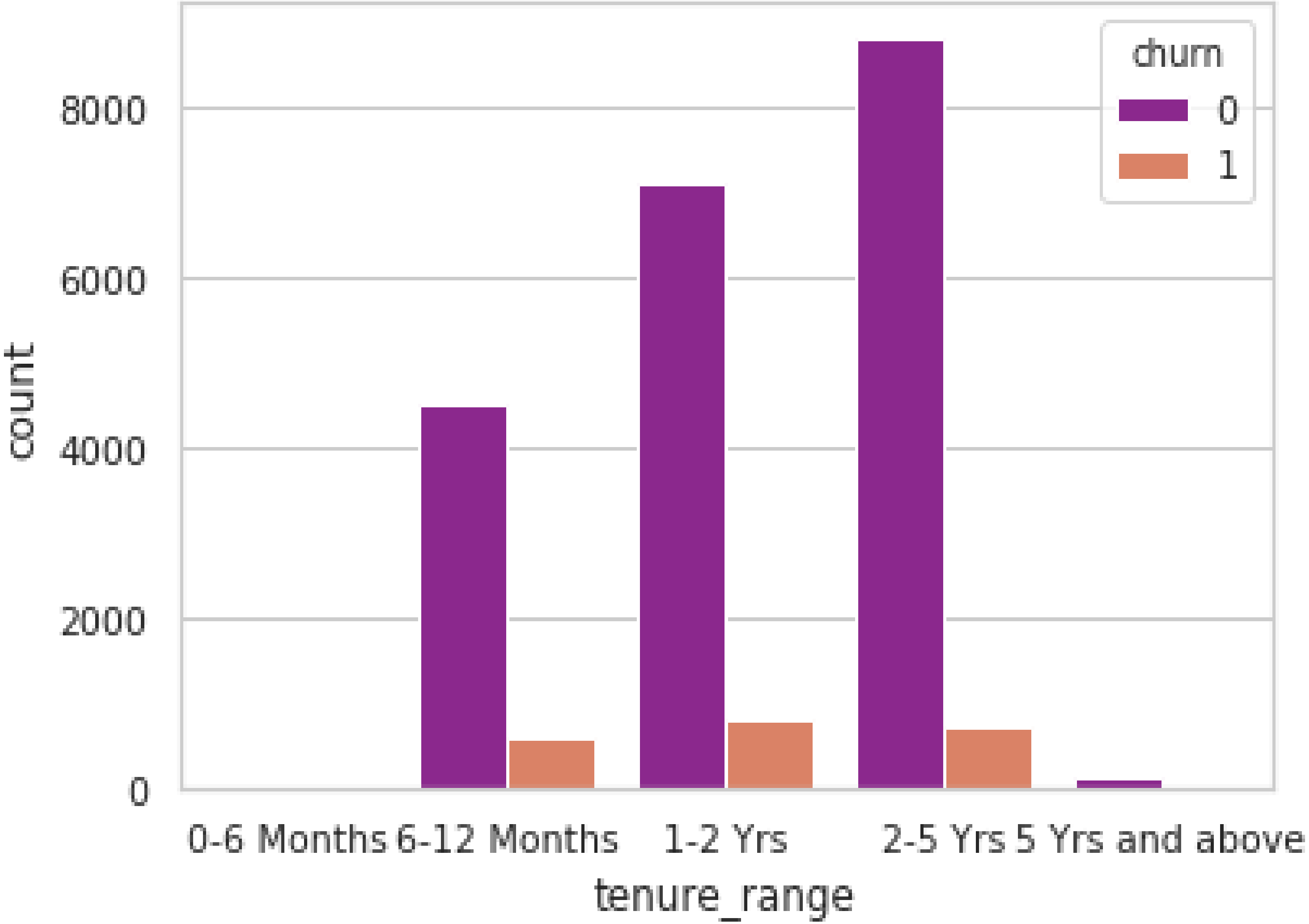
Analysis: Based on the findings, a subsample size of 0.5 and a learning rate of roughly 0.3 appear to be the most effective. Additionally, XGBoost produced the highest ROC AUC (across a variety of hyperparameters) results.



Final model with the chosen hyperparameters.






churn
0
1
8000
6000
4000
2000
0
count
0-6 Months
6-12 Months
1-2 Yrs
2-5 Yrs
5 Yrs and above
tenure_range

BUSINESS INSIGHTS

- Less high value customers are purchasing, but no new high value customers have been added in the last six months, which is worrying and needs the company's attention.
- Customers with tenures of less than four years are more likely to churn, thus the business should focus more on them by introducing new strategies to them.
- In predicting turnover, average revenue per user appears to be the most crucial factor.
- Strong predictors of churn behaviour include incoming and outgoing calls on romaing during the eighth month.
- Local Outgoing calls to landlines, fixed lines, mobile phones, and call centres are a reliable measure of customer attrition.
- Better 2G/3G area coverage is a key determinant of churn behaviour where 2G/3G services are subpar.

MODEL INSIGHTS

	Model	Accuracy	Precision	Recall	AUC	F1
0	SVM (Default)-linear	0.83	0.79	0.30	0.81	0.43
1	SVM (Default)-rbf	0.87	0.74	0.36	0.81	0.49
2	SVM(rfb) [Hyper]	0.92	0.48	0.49	0.72	0.49
3	RandomForest (Default)	0.91	0.49	0.45	0.72	0.47
4	RandomForest (Hyper)	0.90	0.66	0.41	0.79	0.51
5	XGBoost (Default)	0.85	0.75	0.33	0.81	0.45
6	XGBoost (Hyper Tuned)	0.84	0.75	0.31	0.80	0.44

MODEL INSIGHTS

- On this dataset, SVM with optimised hyperparameters produces the best results with an accuracy of 0.92.

- With a default overfit model score of 0.91 and an adjusted hyperparameter score of 0.90, random forest also produces good accuracy.

- With adjusted hyperparameters, XGBoost also produces appropriate accuracy of 0.85 and 0.86 (default overfit model).

- SVM and Random forest are the models that, according to our investigation, give the best accuracy for predicting churn data for a future dataset or output.