TELECOM CHURN PREDICTION

HRITHIK KUMAR

HARSIMRAN SINGH

SRI HARSHA KODURI

PROBLEM STATEMENT

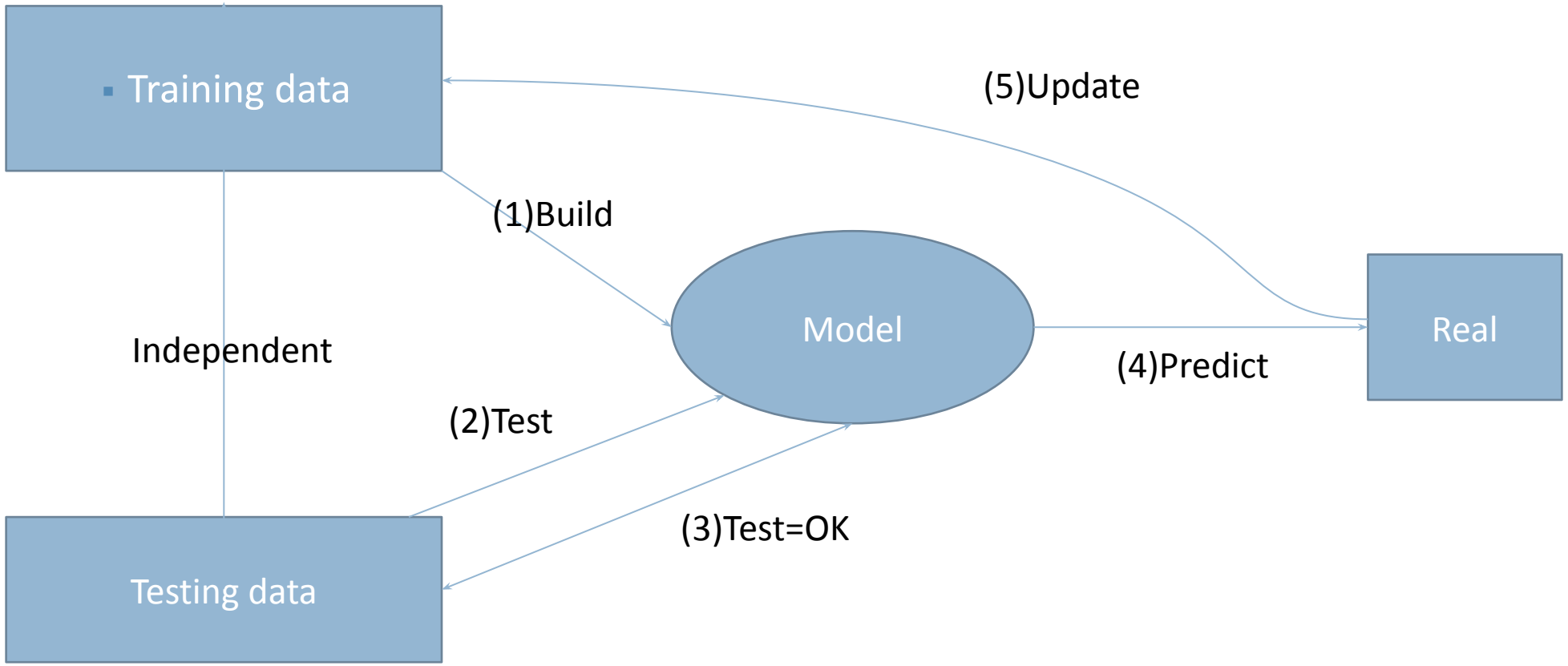
Telecom companies face a 15-25% annual churn rate and find it 5-10 times more expensive to acquire new customers than to retain existing ones.

Predict customer churn in the fourth month (September) using data from the first three months (June, July, August).

PROJECT OBJECTIVES

- To predict Customer Churn.
- Identify the primary variables and factors contributing to customer churn.
- Utilize various machine learning algorithms to develop predictive models.
- Assess the accuracy and performance of these models.
- Determine the most effective model for our business needs and provide an executive summary of our findings.

CHURN PREDICTION MODEL



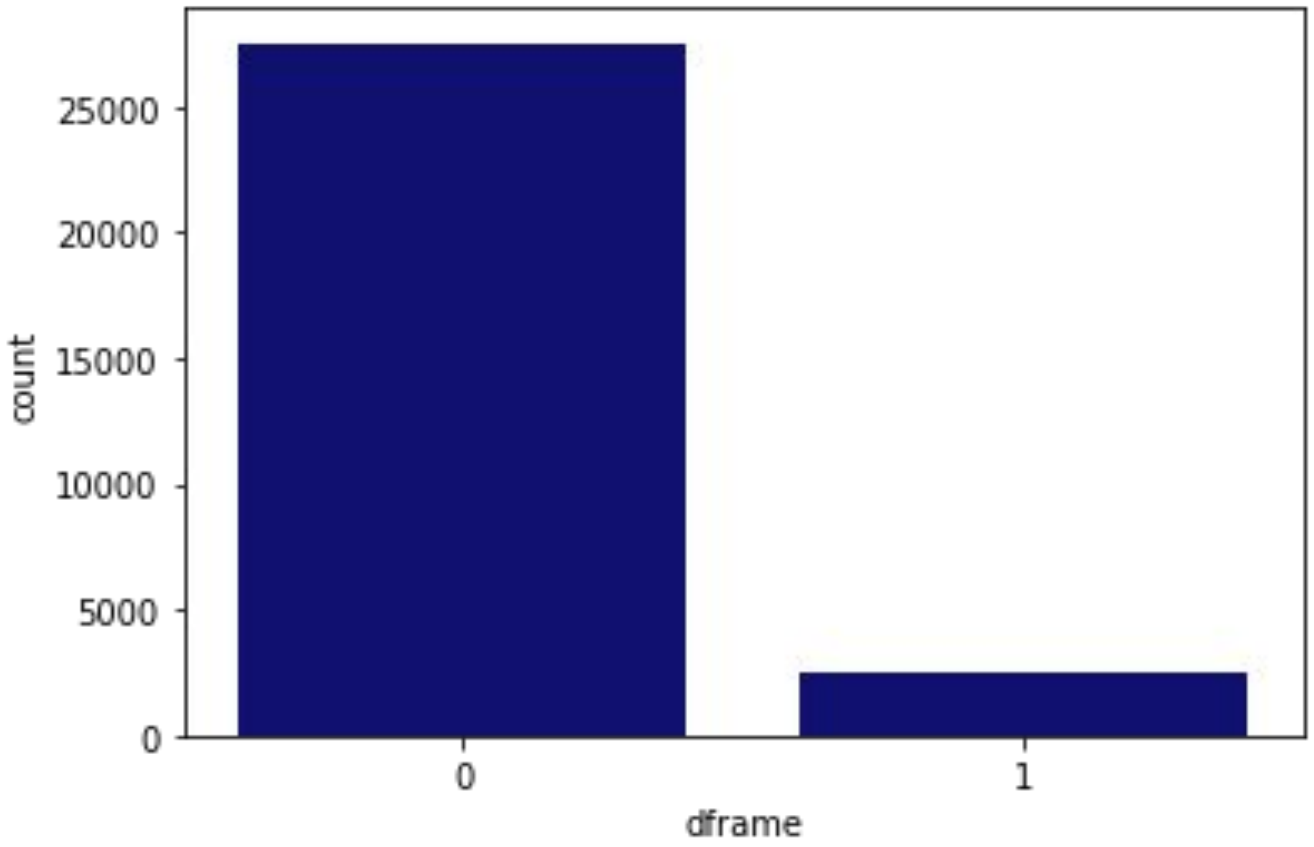
EXPLORATORY DATA ANALYSIS

- Data visualisation using seaborn and matplotlib
- Exploratory data analysis (EDA) is an approach to analyse data sets & to summarize their main characteristics, often with visual methods.
- A Statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modelling or hypothesis.

DATA CLEANING STEPS

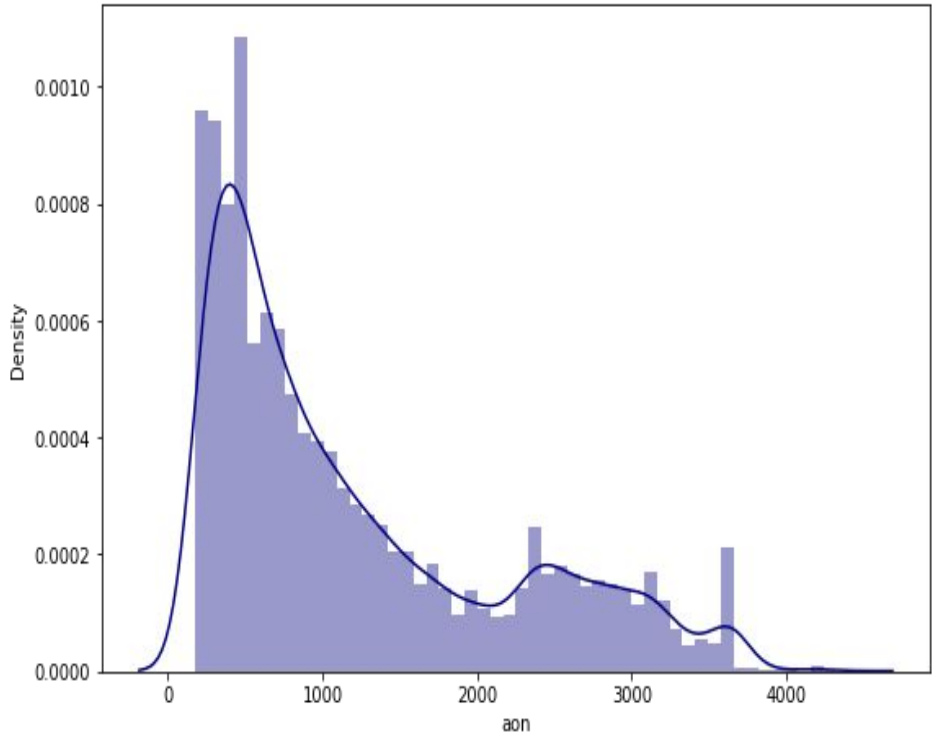
- Drop columns with greater than 40% of null values
- Check for duplicates. Drop any repetitive columns
- Categorize columns (Identifier, numerical, etc.)
- Impute missing values (using median, mode, etc.)
- Drop columns with unique identifiers
- Drop columns with constant value (zero variance)

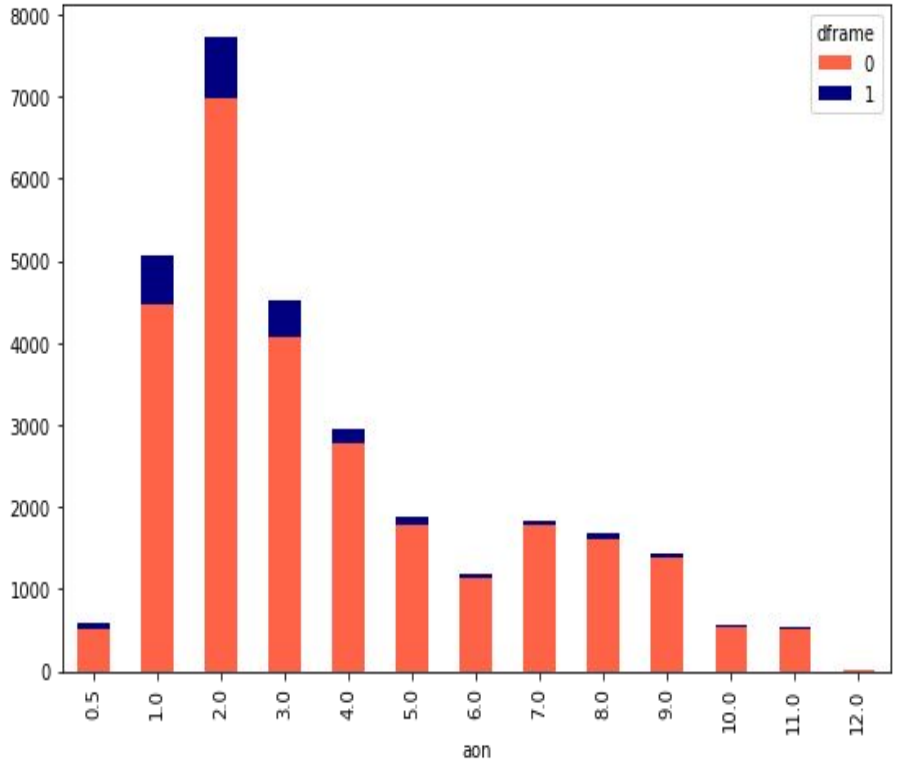
TARGET VARIABLE DISTRIBUTION



91.9% of the customers do not churn (TARGET VARIABLE = 0)

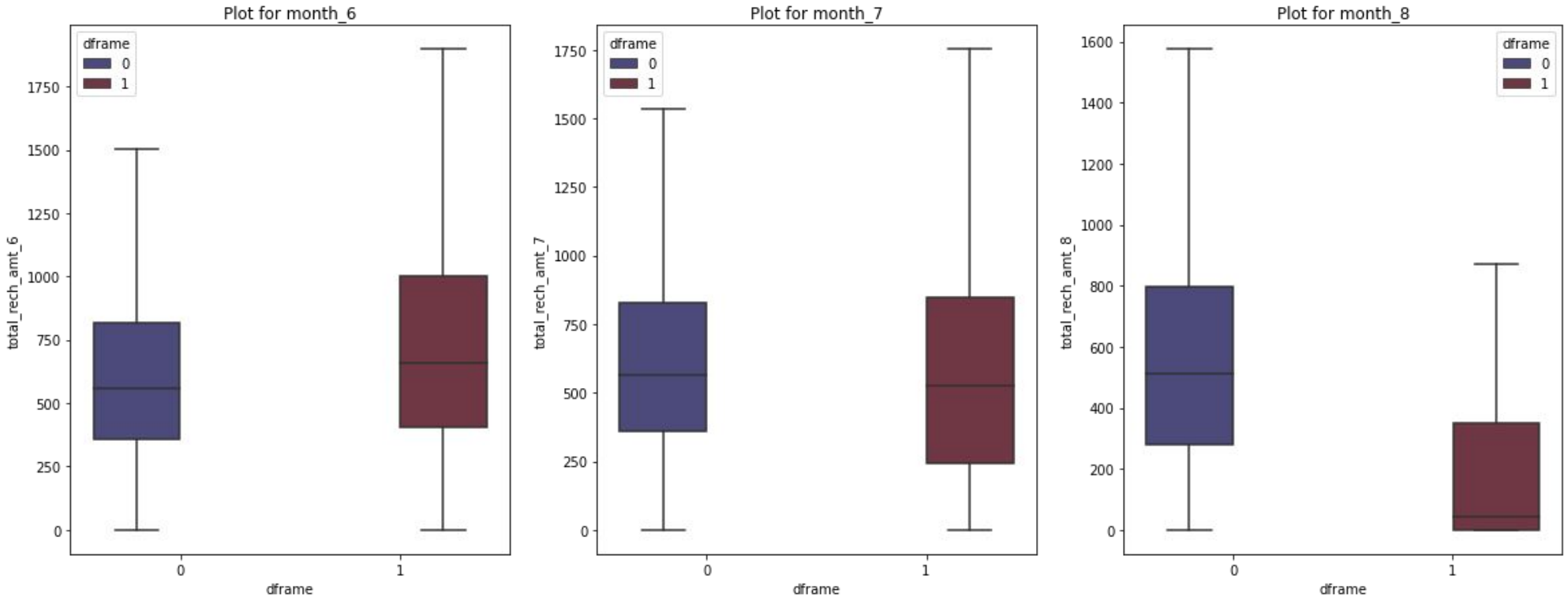
AGE ON NETWORK





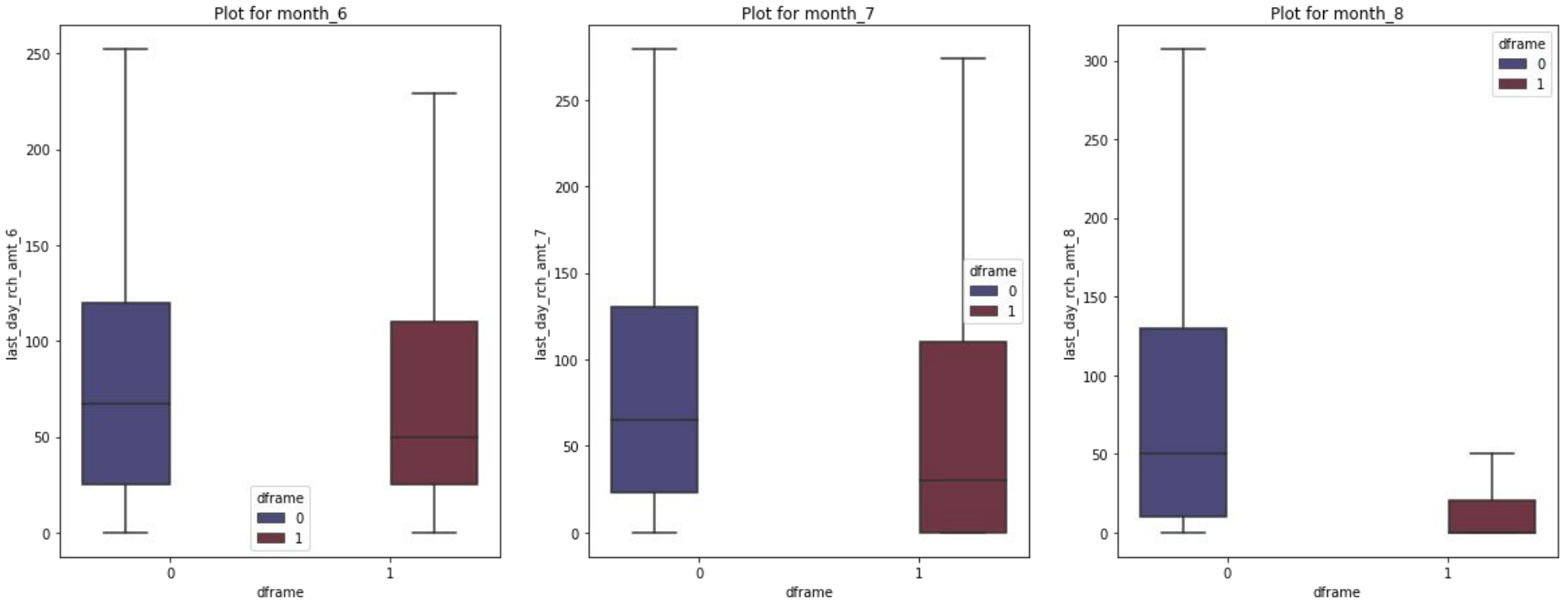
The longer the age on network, the lower the counts are.

TOTAL RECHARGE AMOUNT



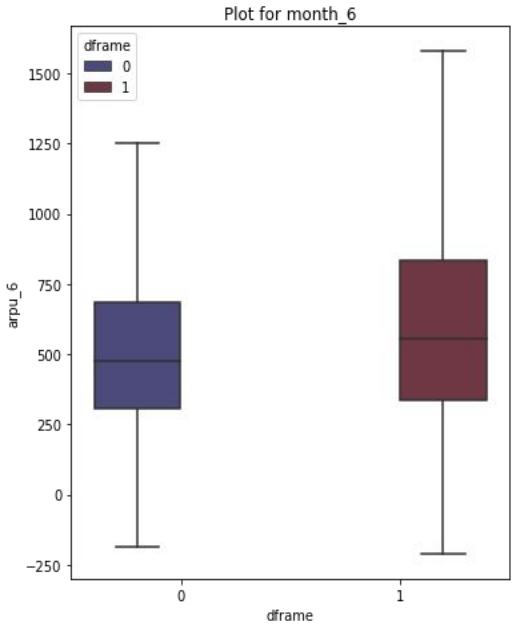
There is a drop in the maximum recharge amount for churned customers in action phase

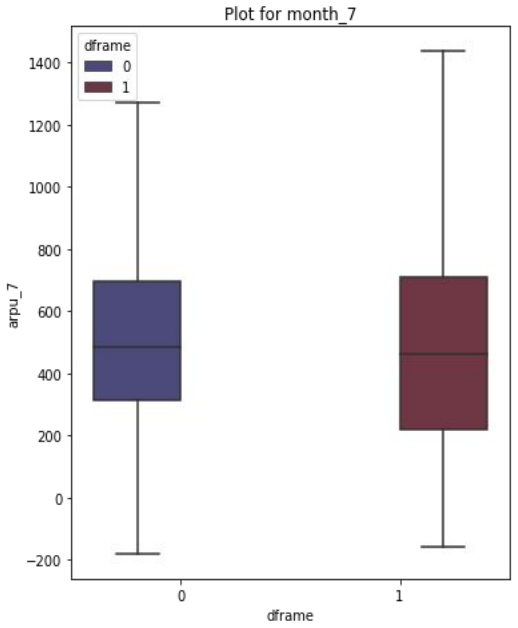
LAST DAY RECHARGE AMOUNT

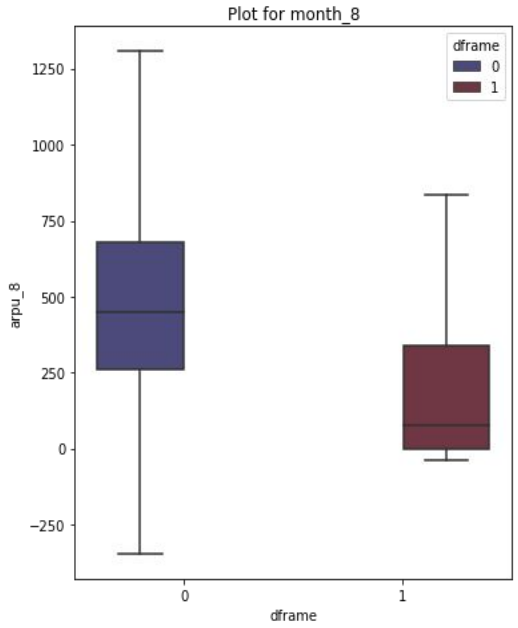


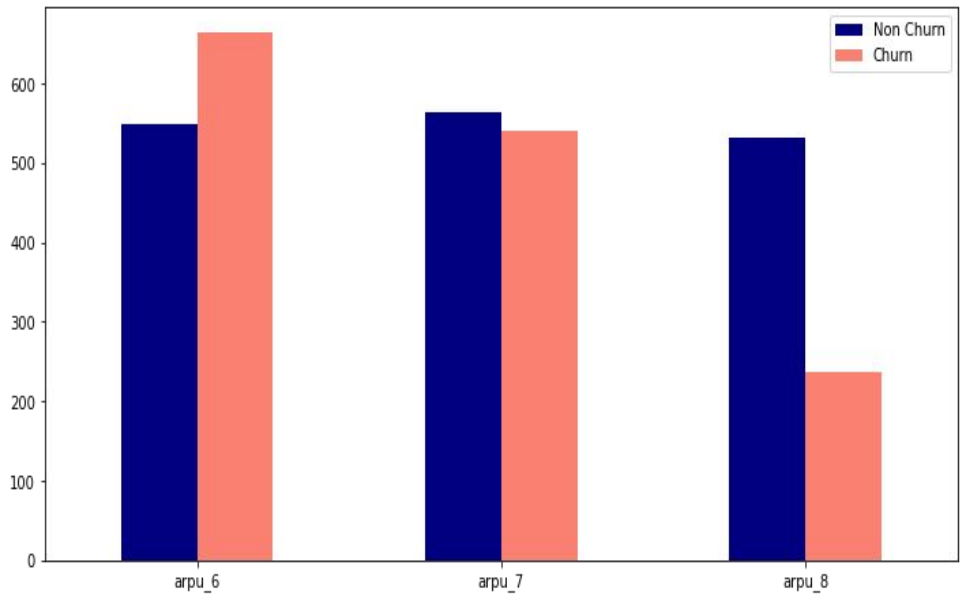
ThE drop off continues to increase for churned customers in action phase. The drop is more significant for last day recharge amount.

AVERAGE REVENUE PER USER



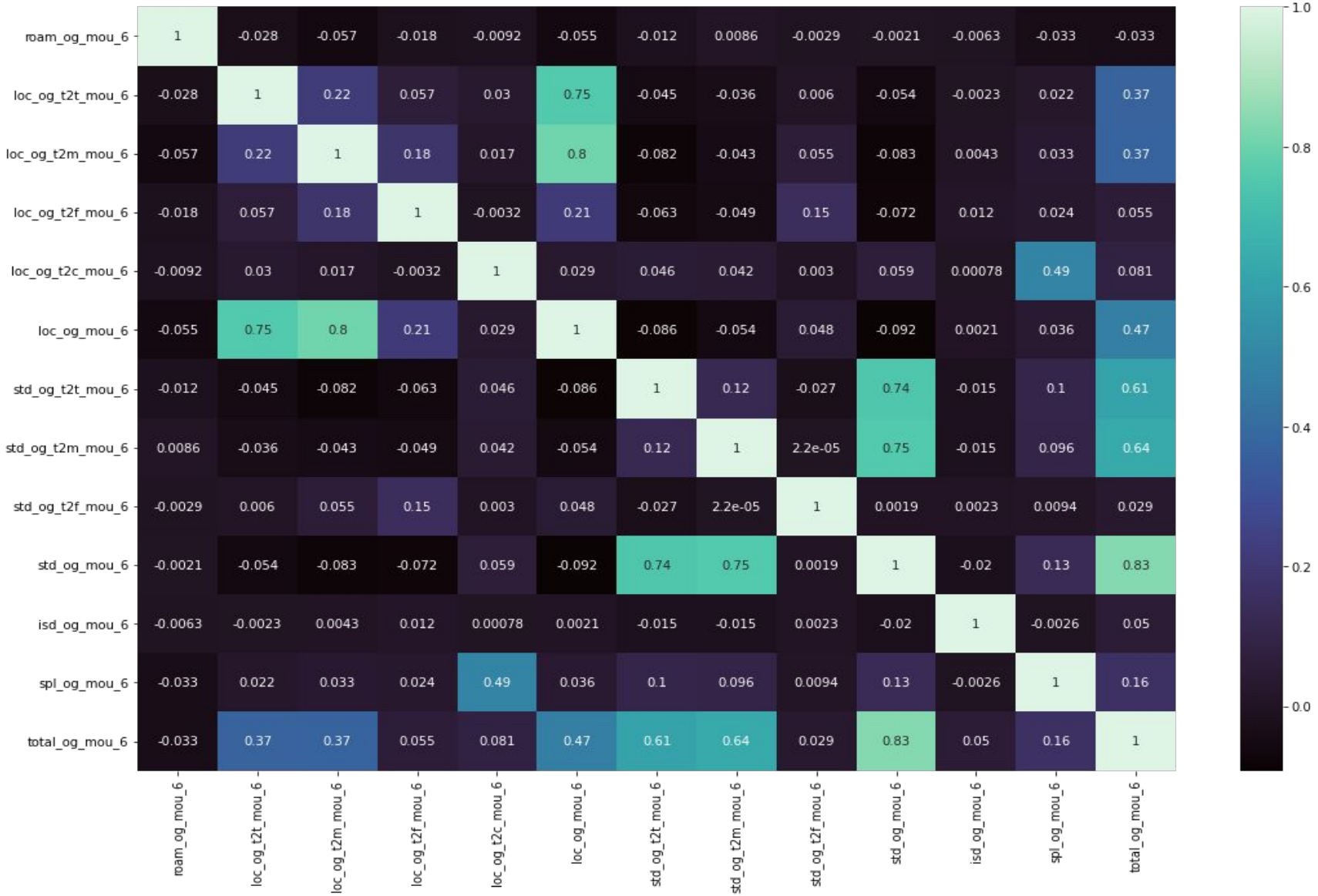






We can see that huge drops in Average revenue per user in the Action phase for churned customers

COLLINEARITY CHECK

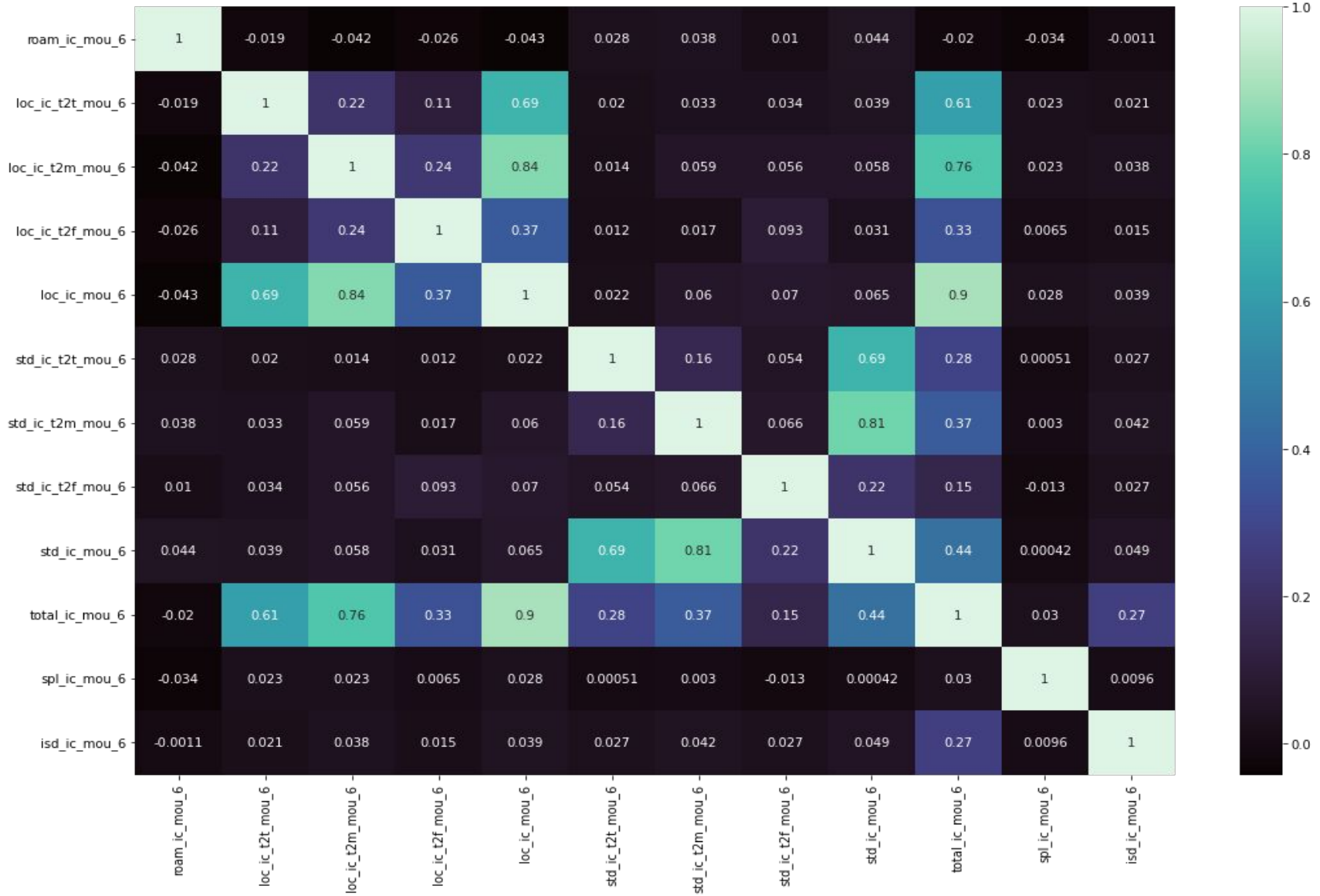


There are several pairs of variables with high positive correlations:

- loc_og_t2t_mou_6 and loc_og_mou_6
- loc_og_t2m_mou_6 and loc_og_mou_6
- std_og_t2t_mou_6 and std_og_mou_6
- std_og_t2m_mou_6 and total_og_mou_6

loc_og_t2m_mou_6 and total_og_mou_6 are negatively correlated. Increase in local outgoing calls to mobile numbers might be associated with a decrease in total outgoing minutes.

REMOVING DERIVED VARIABLES



We can see that total_ic_mou_6, std_ic_mou_6 and loc_ic_mou_6 seem to have a strong correlation with other fields

MODEL COMPARISON

	model_name	acc_score	roc_score	precision_score	recall_score
11	KNeighborsClassifier with PCA(Optimal Hyperpar...	87.95	88.01	80.54	99.93
9	KNeighborsClassifier	89.09	89.14	82.07	99.91
10	KNeighborsClassifier with PCA	89.14	89.19	82.14	99.89
3	RandomForestClassifier	97.17	97.17	96.27	98.11
5	RandomForestClassifier with PCA (Optimal Hyper...	95.80	95.81	93.80	98.04
4	RandomForestClassifier with PCA	95.84	95.85	94.24	97.60
6	DecisionTreeClassifier	92.55	92.56	91.83	93.34
16	GradientBoostingClassifier	93.12	93.12	92.96	93.23
8	DecisionTreeClassifier with PCA(Optimal Hyper...	89.70	89.71	89.05	90.43
12	AdaBoostClassifier	90.88	90.88	91.19	90.41
7	DecisionTreeClassifier with PCA	87.16	87.17	84.80	90.39
19	SVC_Kernel_rbf with PCA	86.67	86.68	85.29	88.48
18	SVC_Kernel_rbf	85.59	85.59	84.61	86.83
14	RidgeClassifier	83.46	83.47	81.58	86.24
1	LogisticRegression with PCA	79.55	79.58	76.69	84.65
17	GradientBoostingClassifier with PCA	83.16	83.16	82.38	84.15
2	LogisticRegression with PCA (Optimal Hyperpara...	81.56	81.57	80.46	83.14
15	RidgeClassifier with PCA	79.36	79.38	77.34	82.78
0	LogisticRegression	80.63	80.63	80.29	80.94
13	AdaBoostClassifier with PCA	78.87	78.88	78.11	79.93

RandomForestClassifier is chosen as it has the best parameters.

Accuracy = 97.17%
Precision = 96.27%
Recall = 98.11%.

MODEL 1

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.2847	0.016	-18.360	0.000	-0.315	-0.254
arpu_7	0.3714	0.020	18.769	0.000	0.333	0.410
onnet_mou_7	0.8109	0.088	9.181	0.000	0.638	0.984
offnet_mou_8	0.8111	0.064	12.661	0.000	0.686	0.937
loc_og_t2t_mou_7	-0.7961	0.044	-18.266	0.000	-0.881	-0.711
loc_og_t2m_mou_8	-0.8895	0.041	-21.722	0.000	-0.970	-0.809
std_og_t2t_mou_7	-1.0698	0.083	-12.898	0.000	-1.232	-0.907
std_og_t2m_mou_8	-0.8517	0.053	-16.202	0.000	-0.955	-0.749
loc_ic_t2m_mou_8	-0.4997	0.026	-19.487	0.000	-0.550	-0.449
spl_ic_mou_8	-0.3753	0.019	-20.266	0.000	-0.412	-0.339
total_rech_num_8	-0.4485	0.018	-24.761	0.000	-0.484	-0.413
last_day_rch_amt_8	-0.5480	0.016	-34.198	0.000	-0.579	-0.517
max_rech_data_6	0.4926	0.024	20.244	0.000	0.445	0.540
max_rech_data_8	0.2608	0.025	10.325	0.000	0.211	0.310
sachet_2g_6	0.4281	0.019	22.758	0.000	0.391	0.465
sep_vbc_3g	-0.6567	0.045	-14.515	0.000	-0.745	-0.568
night_pck_user_6_0.0	-0.6018	0.024	-24.593	0.000	-0.650	-0.554
fb_user_8_1.0	-1.1547	0.025	-46.578	0.000	-1.203	-1.106
avg_onnet_mou_av678	0.7391	0.076	9.670	0.000	0.589	0.889
avg_loc_og_t2t_mou_av678	-0.6827	0.036	-19.082	0.000	-0.753	-0.613
avg_std_og_t2t_mou_av678	-0.9277	0.071	-13.156	0.000	-1.066	-0.790

	Features	VIF
1	onnet_mou_7	32.42
5	std_og_t2t_mou_7	28.86
17	avg_onnet_mou_av678	24.00
19	avg_std_og_t2t_mou_av678	20.61
2	offnet_mou_8	15.23
6	std_og_t2m_mou_8	10.39
3	loc_og_t2t_mou_7	6.17
4	loc_og_t2m_mou_8	5.88
18	avg_loc_og_t2t_mou_av678	3.99
12	max_rech_data_8	3.32
15	night_pck_user_6_0.0	3.22
16	fb_user_8_1.0	3.05
11	max_rech_data_6	2.97
7	loc_ic_t2m_mou_8	1.96
0	arpu_7	1.83
13	sachet_2g_6	1.73
9	total_rech_num_8	1.70
10	last_day_rch_amt_8	1.34
14	sep_vbc_3g	1.05
8	spl_ic_mou_8	1.05

Dropping onnet_mou_7 due to high VIF

MODEL 2

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.2881	0.016	-18.578	0.000	-0.318	-0.258
arpu_7	0.3905	0.020	19.799	0.000	0.352	0.429
offnet_mou_8	0.9209	0.063	14.596	0.000	0.797	1.045
loc_og_t2t_mou_7	-0.4943	0.028	-17.890	0.000	-0.548	-0.440
loc_og_t2m_mou_8	-0.9745	0.040	-24.230	0.000	-1.053	-0.896
std_og_t2t_mou_7	-0.3287	0.019	-17.748	0.000	-0.365	-0.292
std_og_t2m_mou_8	-0.9575	0.051	-18.614	0.000	-1.058	-0.857
loc_ic_t2m_mou_8	-0.5127	0.026	-19.923	0.000	-0.563	-0.462
spl_ic_mou_8	-0.3792	0.019	-20.461	0.000	-0.416	-0.343
total_rech_num_8	-0.4348	0.018	-24.167	0.000	-0.470	-0.400
last_day_rch_amt_8	-0.5430	0.016	-34.025	0.000	-0.574	-0.512
max_rech_data_6	0.4957	0.024	20.393	0.000	0.448	0.543
max_rech_data_8	0.2574	0.025	10.199	0.000	0.208	0.307
sachet_2g_6	0.4282	0.019	22.803	0.000	0.391	0.465
sep_vbc_3g	-0.6610	0.045	-14.565	0.000	-0.750	-0.572
night_pck_user_6_0.0	-0.6004	0.024	-24.566	0.000	-0.648	-0.553
fb_user_8_1.0	-1.1593	0.025	-46.807	0.000	-1.208	-1.111
avg_onnet_mou_av678	0.3799	0.063	6.052	0.000	0.257	0.503
avg_loc_og_t2t_mou_av678	-0.5688	0.033	-17.425	0.000	-0.633	-0.505
avg_std_og_t2t_mou_av678	-0.6018	0.058	-10.294	0.000	-0.716	-0.487

	Features	VIF
16	avg_onnet_mou_av678	20.13
18	avg_std_og_t2t_mou_av678	17.43
1	offnet_mou_8	14.74
5	std_og_t2m_mou_8	9.91
3	loc_og_t2m_mou_8	5.58
17	avg_loc_og_t2t_mou_av678	3.56
11	max_rech_data_8	3.32
14	night_pck_user_6_0.0	3.22
15	fb_user_8_1.0	3.05
10	max_rech_data_6	2.97
4	std_og_t2t_mou_7	2.06
6	loc_ic_t2m_mou_8	1.96
0	arpu_7	1.81
12	sachet_2g_6	1.73
8	total_rech_num_8	1.69
2	loc_og_t2t_mou_7	1.68
9	last_day_rch_amt_8	1.34
13	sep_vbc_3g	1.05
7	spl_ic_mou_8	1.05

Dropping onnet_mou_7 due to high VIF

MODEL 3

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.2885	0.015	-18.616	0.000	-0.319	-0.258
arpu_7	0.3772	0.020	19.277	0.000	0.339	0.416
offnet_mou_8	0.9641	0.063	15.406	0.000	0.841	1.087
loc_og_t2t_mou_7	-0.5157	0.028	-18.480	0.000	-0.570	-0.461
loc_og_t2m_mou_8	-0.9897	0.040	-24.666	0.000	-1.068	-0.911
std_og_t2t_mou_7	-0.3248	0.019	-17.539	0.000	-0.361	-0.288
std_og_t2m_mou_8	-0.9914	0.051	-19.409	0.000	-1.092	-0.891
loc_ic_t2m_mou_8	-0.5099	0.026	-19.855	0.000	-0.560	-0.460
spl_ic_mou_8	-0.3803	0.019	-20.540	0.000	-0.417	-0.344
total_rech_num_8	-0.4271	0.018	-23.873	0.000	-0.462	-0.392
last_day_rch_amt_8	-0.5397	0.016	-33.871	0.000	-0.571	-0.508
max_rech_data_6	0.4959	0.024	20.399	0.000	0.448	0.543
max_rech_data_8	0.2584	0.025	10.244	0.000	0.209	0.308
sachet_2g_6	0.4271	0.019	22.759	0.000	0.390	0.464
sep_vbc_3g	-0.6588	0.045	-14.557	0.000	-0.748	-0.570
night_pck_user_6_0.0	-0.5983	0.024	-24.494	0.000	-0.646	-0.550
fb_user_8_1.0	-1.1628	0.025	-46.974	0.000	-1.211	-1.114
avg_loc_og_t2t_mou_av678	-0.4602	0.027	-16.836	0.000	-0.514	-0.407
avg_std_og_t2t_mou_av678	-0.2662	0.018	-14.568	0.000	-0.302	-0.230

	Features	VIF
1	offnet_mou_8	14.47
5	std_og_t2m_mou_8	9.74
3	loc_og_t2m_mou_8	5.52
11	max_rech_data_8	3.32
14	night_pck_user_6_0.0	3.22
15	fb_user_8_1.0	3.05
10	max_rech_data_6	2.97
4	std_og_t2t_mou_7	2.06
6	loc_ic_t2m_mou_8	1.96
17	avg_std_og_t2t_mou_av678	1.81
0	arpu_7	1.80
12	sachet_2g_6	1.73
8	total_rech_num_8	1.69
2	loc_og_t2t_mou_7	1.67
16	avg_loc_og_t2t_mou_av678	1.50
9	last_day_rch_amt_8	1.34
13	sep_vbc_3g	1.05
7	spl_ic_mou_8	1.05

Dropping offnet_mou_8 due to high VIF

MODEL 4

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.2967	0.015	-19.182	0.000	-0.327	-0.266
arpu_7	0.4672	0.019	24.932	0.000	0.430	0.504
loc_og_t2t_mou_7	-0.5404	0.028	-19.318	0.000	-0.595	-0.486
loc_og_t2m_mou_8	-0.5396	0.026	-20.609	0.000	-0.591	-0.488
std_og_t2t_mou_7	-0.3918	0.018	-21.690	0.000	-0.427	-0.356
std_og_t2m_mou_8	-0.2468	0.015	-16.787	0.000	-0.276	-0.218
loc_ic_t2m_mou_8	-0.5470	0.026	-21.051	0.000	-0.598	-0.496
spl_ic_mou_8	-0.3910	0.019	-21.102	0.000	-0.427	-0.355
total_rech_num_8	-0.4048	0.018	-22.987	0.000	-0.439	-0.370
last_day_rch_amt_8	-0.5110	0.016	-32.767	0.000	-0.542	-0.480
max_rech_data_6	0.4958	0.024	20.487	0.000	0.448	0.543
max_rech_data_8	0.2524	0.025	10.057	0.000	0.203	0.302
sachet_2g_6	0.4255	0.019	22.794	0.000	0.389	0.462
sep_vbc_3g	-0.6593	0.045	-14.710	0.000	-0.747	-0.571
night_pck_user_6_0.0	-0.5947	0.024	-24.416	0.000	-0.642	-0.547
fb_user_8_1.0	-1.1835	0.025	-48.015	0.000	-1.232	-1.135
avg_loc_og_t2t_mou_av678	-0.4532	0.027	-16.569	0.000	-0.507	-0.400
avg_std_og_t2t_mou_av678	-0.2857	0.018	-15.640	0.000	-0.322	-0.250

	Features	VIF
10	max_rech_data_8	3.32
13	night_pck_user_6_0.0	3.22
14	fb_user_8_1.0	3.04
9	max_rech_data_6	2.97
2	loc_og_t2m_mou_8	2.20
3	std_og_t2t_mou_7	1.98
5	loc_ic_t2m_mou_8	1.95
16	avg_std_og_t2t_mou_av678	1.80
11	sachet_2g_6	1.73
0	arpu_7	1.67
7	total_rech_num_8	1.67
1	loc_og_t2t_mou_7	1.66
15	avg_loc_og_t2t_mou_av678	1.50
4	std_og_t2m_mou_8	1.33
8	last_day_rch_amt_8	1.32
12	sep_vbc_3g	1.05
6	spl_ic_mou_8	1.05

Dropping max_rech_data_8 due to high VIF

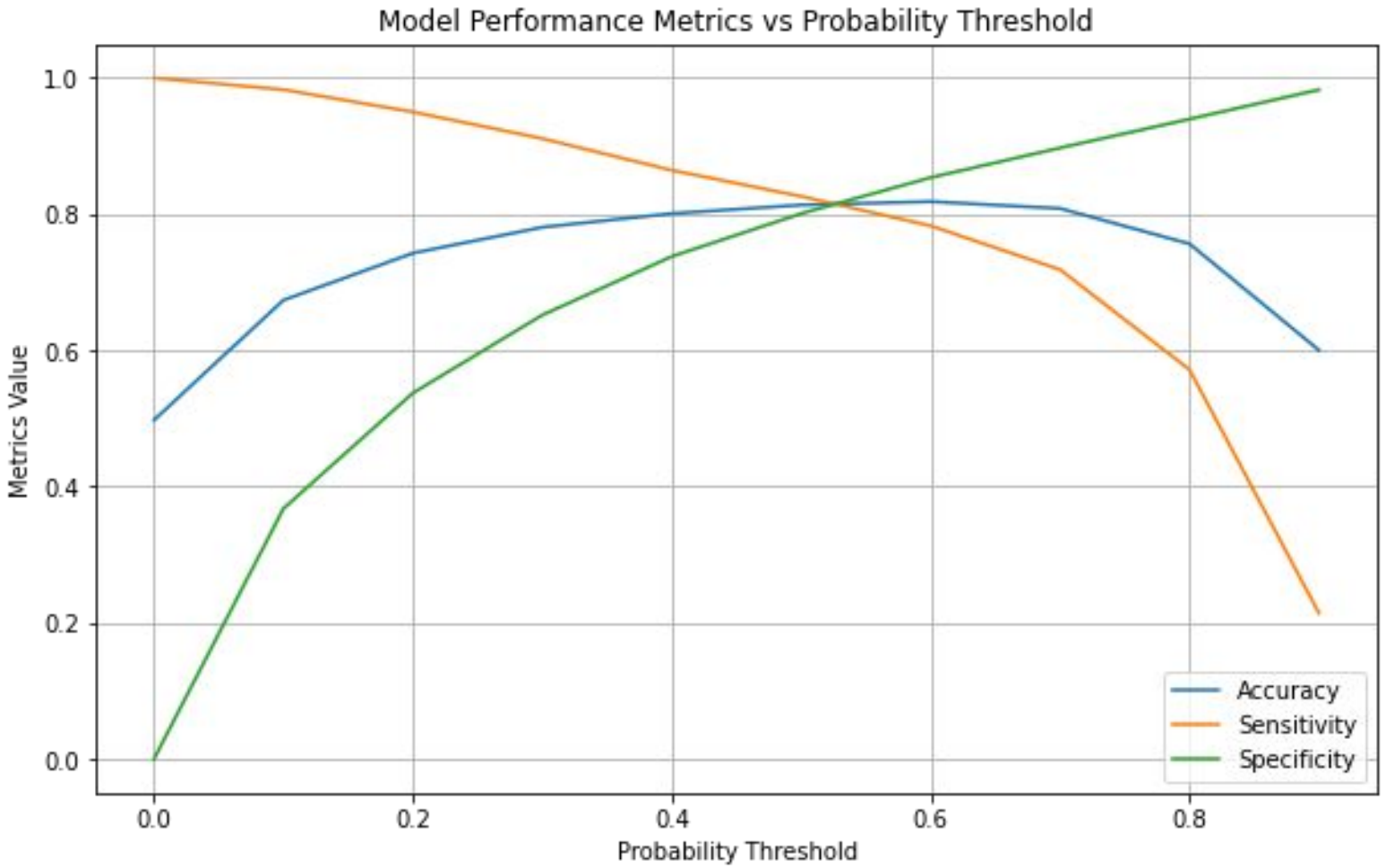
MODEL 5

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.2905	0.015	-18.819	0.000	-0.321	-0.260
arpu_7	0.4846	0.019	25.852	0.000	0.448	0.521
loc_og_t2t_mou_7	-0.5403	0.028	-19.314	0.000	-0.595	-0.485
loc_og_t2m_mou_8	-0.5415	0.026	-20.694	0.000	-0.593	-0.490
std_og_t2t_mou_7	-0.3918	0.018	-21.727	0.000	-0.427	-0.356
std_og_t2m_mou_8	-0.2452	0.015	-16.696	0.000	-0.274	-0.216
loc_ic_t2m_mou_8	-0.5511	0.026	-21.208	0.000	-0.602	-0.500
spl_ic_mou_8	-0.3931	0.019	-21.212	0.000	-0.429	-0.357
total_rech_num_8	-0.4172	0.018	-23.805	0.000	-0.452	-0.383
last_day_rch_amt_8	-0.4911	0.016	-31.625	0.000	-0.522	-0.461
max_rech_data_6	0.5890	0.023	25.996	0.000	0.545	0.633
sachet_2g_6	0.3974	0.018	21.924	0.000	0.362	0.433
sep_vbc_3g	-0.6678	0.045	-14.825	0.000	-0.756	-0.579
night_pck_user_6_0.0	-0.6196	0.024	-25.617	0.000	-0.667	-0.572
fb_user_8_1.0	-1.0111	0.017	-59.437	0.000	-1.044	-0.978
avg_loc_og_t2t_mou_av678	-0.4498	0.027	-16.451	0.000	-0.503	-0.396
avg_std_og_t2t_mou_av678	-0.2824	0.018	-15.493	0.000	-0.318	-0.247

	Features	VIF
12	night_pck_user_6_0.0	3.17
9	max_rech_data_6	2.55
2	loc_og_t2m_mou_8	2.20
3	std_og_t2t_mou_7	1.98
5	loc_ic_t2m_mou_8	1.94
15	avg_std_og_t2t_mou_av678	1.80
10	sachet_2g_6	1.71
1	loc_og_t2t_mou_7	1.66
0	arpu_7	1.65
7	total_rech_num_8	1.65
13	fb_user_8_1.0	1.58
14	avg_loc_og_t2t_mou_av678	1.49
4	std_og_t2m_mou_8	1.33
8	last_day_rch_amt_8	1.28
6	spl_ic_mou_8	1.05
11	sep_vbc_3g	1.05

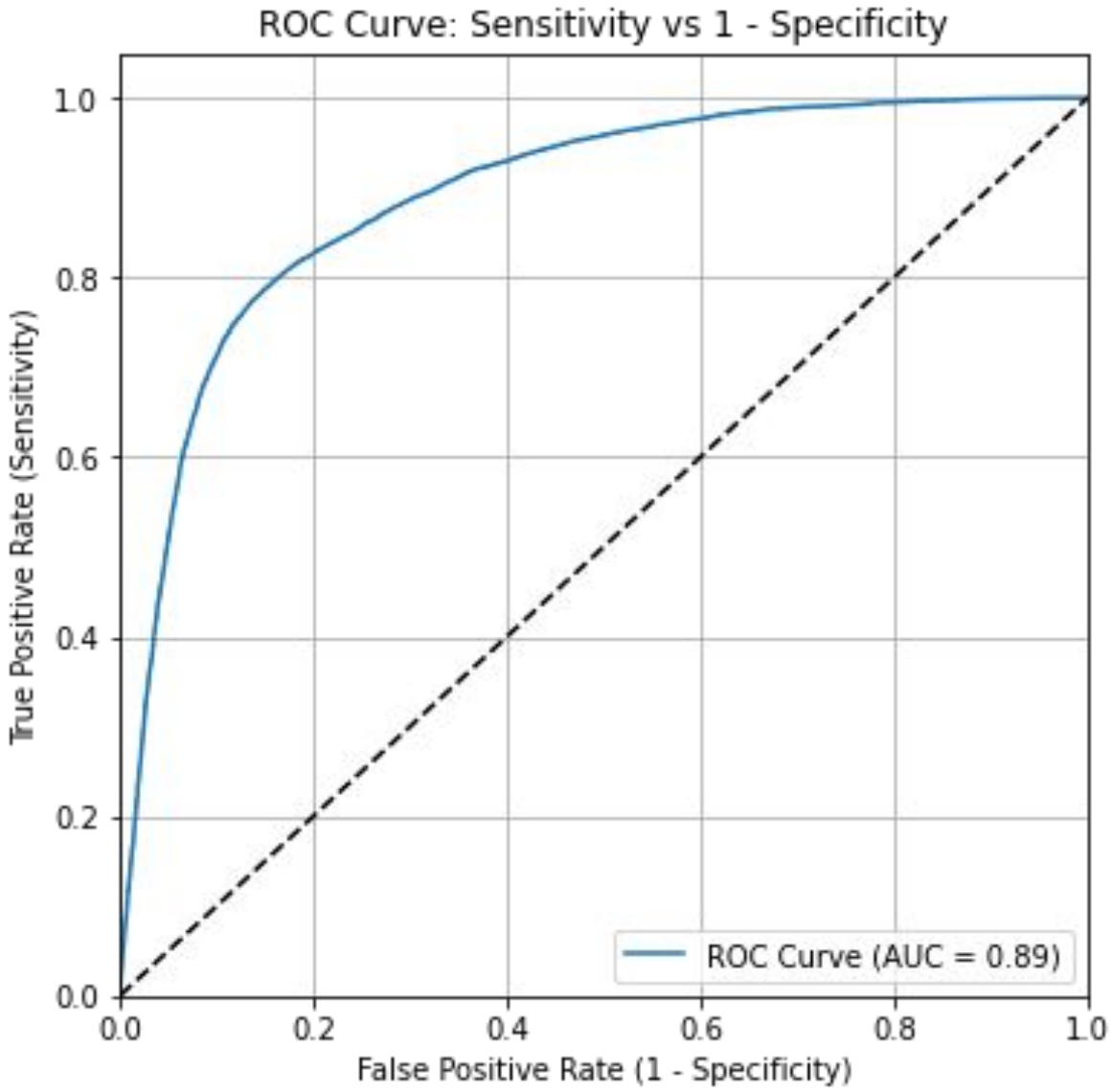
This model has P-Value < 0.05 & RFE < 5. So this is chosen as the final model.

MODEL PERFORMANCE



	prob	accuracy	sensi	speci
0.0	0.0	0.497604	1.000000	0.000000
0.1	0.1	0.673726	0.982757	0.367642
0.2	0.2	0.742709	0.950105	0.537291
0.3	0.3	0.780727	0.910896	0.651799
0.4	0.4	0.800831	0.864441	0.737827
0.5	0.5	0.813427	0.826103	0.800872
0.6	0.6	0.818425	0.782812	0.853697
0.7	0.7	0.808293	0.718472	0.897257
0.8	0.8	0.756424	0.571448	0.939635
0.9	0.9	0.600087	0.213794	0.982694

ROC CURVE: Sensitivity vs (1 - Specificity)



- Sensitivity : 0.78
- Specificity : 0.85
- False Positive Rate : 0.15
- Positive Predictive Value : 0.84
- Negative predictive value : 0.8

TOP 5 PREDICTORS

- **max_rech_data_6**: The maximum recharge done in June. Higher data recharge amounts in the June are positively correlated with churn
- **arpu_7**: Average Revenue Per User in July. Higher ARPU in July is positively related to churn
- **sachet_2g_6**: Usage of 2G sachet packs in June. Increased usage of these packs in the June positively impacts churn
- **std_og_t2m_mou_8**: Standard outgoing minutes to mobile numbers in action phase (August). This has a negative impact, indicating that higher usage negatively correlates with churn
- **avg_std_og_t2t_mou_av678**: Average standard outgoing minutes of usage to the same number type across June, July, August. This variable negatively correlates with the target, suggesting that consistent calling patterns across these months might reduce the likelihood of achieving the target

RECOMMENDED STRATEGIES

- Target retention for customers with less than 4 years of service duration. Offer loyalty programs or incentives to increase engagement.
- Improve ARPU: Look at high-ARPU customers who stay and try to their successful strategies for at-risk customers.
- Incoming & outgoing roaming calls during Action phase are strong indicators of churn.
- Revise Facebook user pack pricing if it contributes to churn, or offer more value-added services.
- Engage customers with decreased incoming local and outgoing ISD call usage, especially in August, with personalized offers.
- Monitor value-based cost increases during the action phase and offer targeted promotions, loyalty rewards etc.