Telecom Churn Case Study
Advance Machine Learning

Submitted By -

Prateek Verma

Harshal K L

Challa Prasanth

Summary:

- Total around 30K customers are filtered as High value customers.
- In can be noticed Minutes of usage - voice calls for all the types of calls for the month of JUN, JUL, AUG, SEP, have missing values together
- As it can be seen, in month of septembers all columns with mou has missing value together. Removing those values.
- As it can be seen, in month of June all columns with mou has missing value together. Removing those values.
- As it can be seen, in month of August all columns with mou has missing value together. Removing those values.

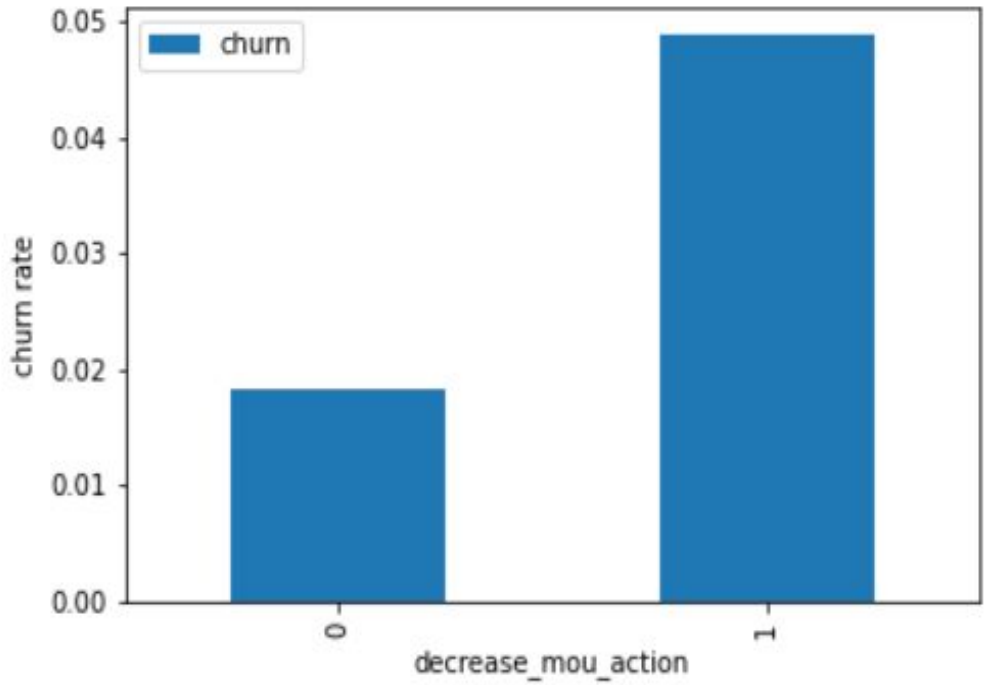
- As it can be seen, in month of July all columns with mou has missing value together. Removing those values.
- Now dataframe does not have null values, so we can proceed further.
- We can see that we have lost almost 7% records. But we have enough number of records to do our analysis.
- There is very little percentage of churn rate. We will take care of the class imbalance later.
- Outliers from all the Numerical collumns has been removed.

- Decrease MOU Action column indicates whether the minutes of usage of the customer has decreased in the action phase than the good phase.
- Decrease Rech Num Action column indicates whether the number of recharge of the customer has decreased in the action phase than the good phase.
- Decrease Rech Amt Action column indicates whether the amount of recharge of the customer has decreased in the action phase than the good phase.
- Decrease Arpu Action column indicates whether the average revenue per customer has decreased in the action phase than the good phase.
- Decrease VBC Action column indicates whether the volume based cost of the customer has decreased in the action phase than the good phase.

EDA – Univariate Analysis

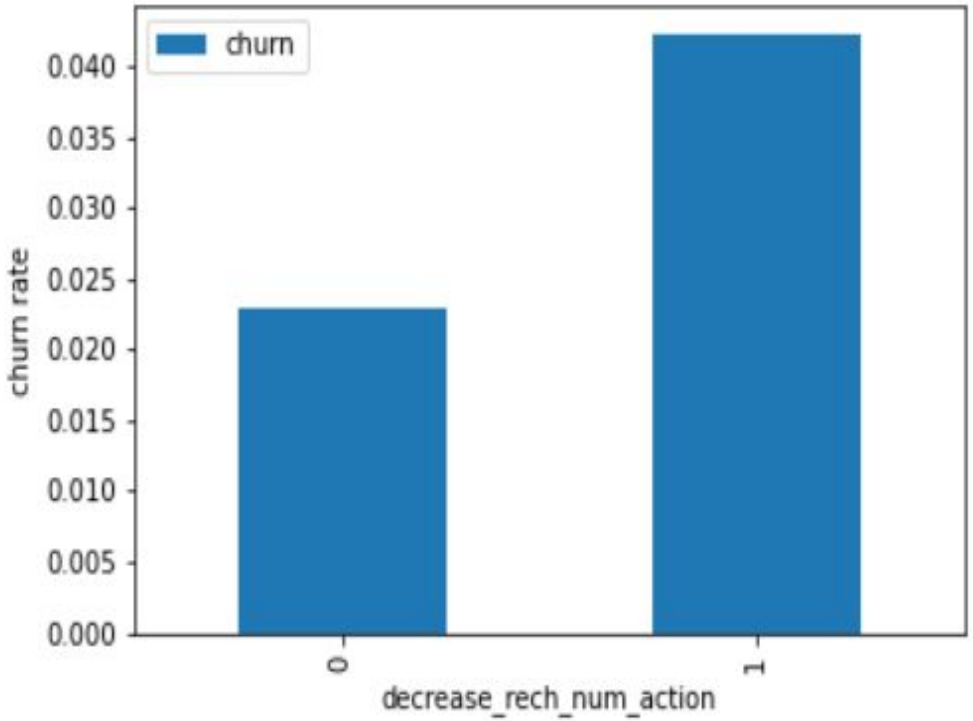
We can see that the churn rate is more for the customers, whose minutes of usage(mou) decreased in the action phase than the good phase.

```
data.pivot_table(values='churn', index='decrease_mou_action', aggfunc='mean').
plt.ylabel('churn rate')
plt.show()
```



- As expected, the churn rate is more for the customers, whose number of recharge in the action phase is lesser than the number in good phase.

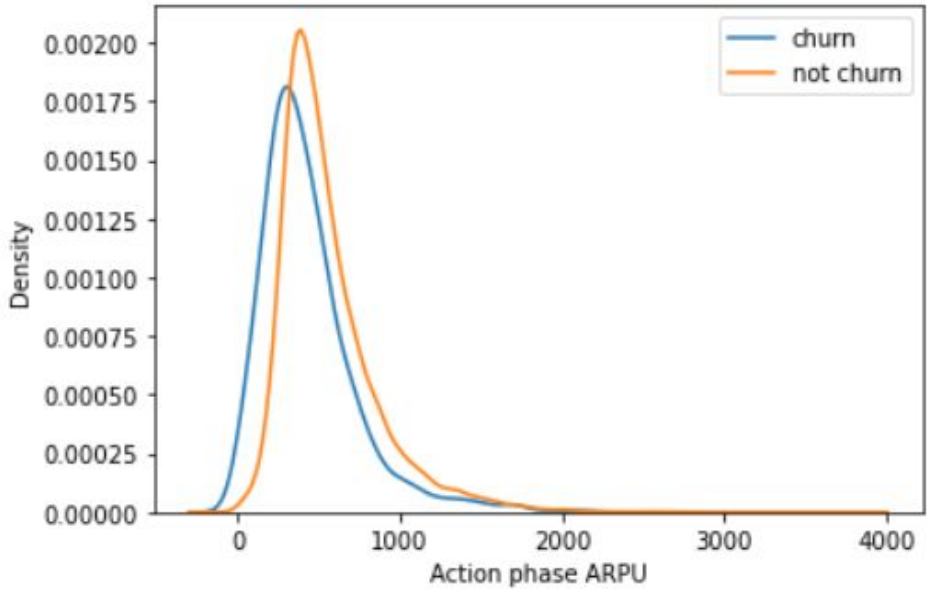
```
data.pivot_table(values='churn', index='decrease_rech_num_action', aggfunc='me
plt.ylabel('churn rate')
plt.show()
```



- Average revenue per user (ARPU) for the churned customers is mostly densed on the 0 to 900. The higher ARPU customers are less likely to be churned.
- ARPU for the not churned customers is mostly densed on the 0 to 1000.

```
# Distribution plot
ax = sns.distplot(data_churn['avg_arpu_action'],label='churn',hist=False)
ax = sns.distplot(data_non_churn['avg_arpu_action'],label='not churn',hist=Fal
ax.legend(labels=["churn","not churn"])
ax.set(xlabel='Action phase ARPU')
```

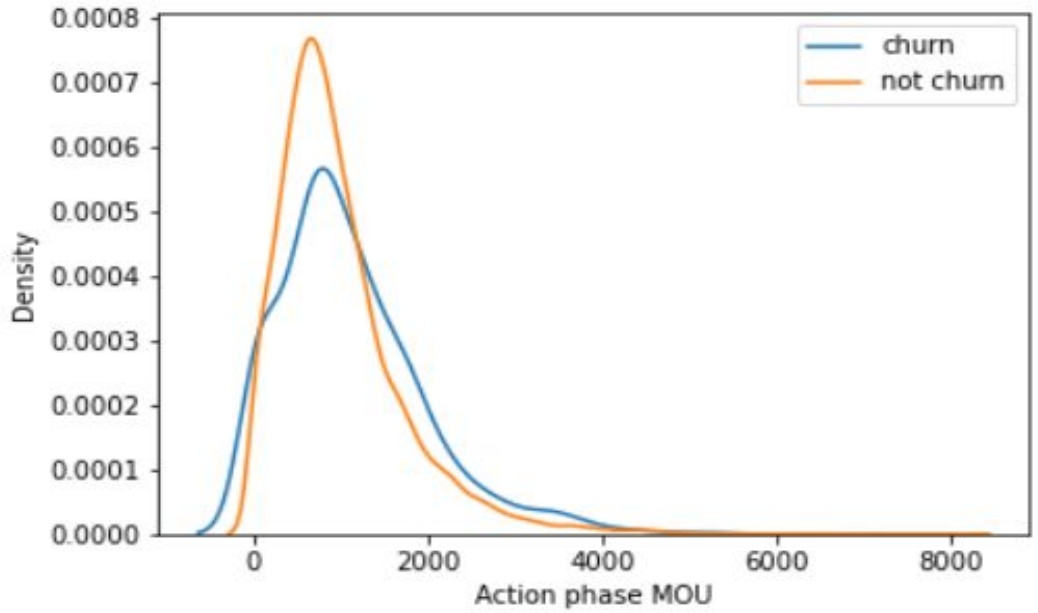
```
[Text(0.5, 0, 'Action phase ARPU')]
```



- Minutes of usage(MOU) of the churn customers is mostly populated on the 0 to 2500 range. Higher the MOU, lesser the churn probability..

```
# Distribution plot
ax = sns.distplot(data_churn['total_mou_good'],label='churn',hist=False)
ax = sns.distplot(data_non_churn['total_mou_good'],label='non churn',hist=Fals
ax.legend(labels=["churn","not churn"])
ax.set(xlabel='Action phase MOU')
```

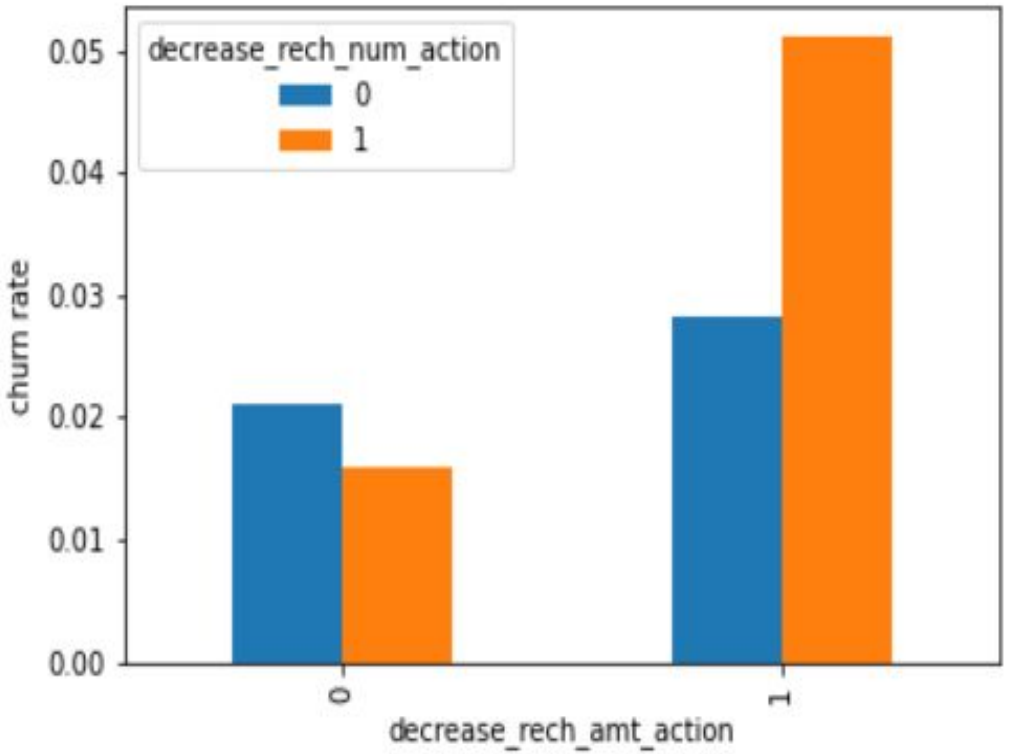
```
[Text(0.5, 0, 'Action phase MOU')]
```



Bivariate Analysis

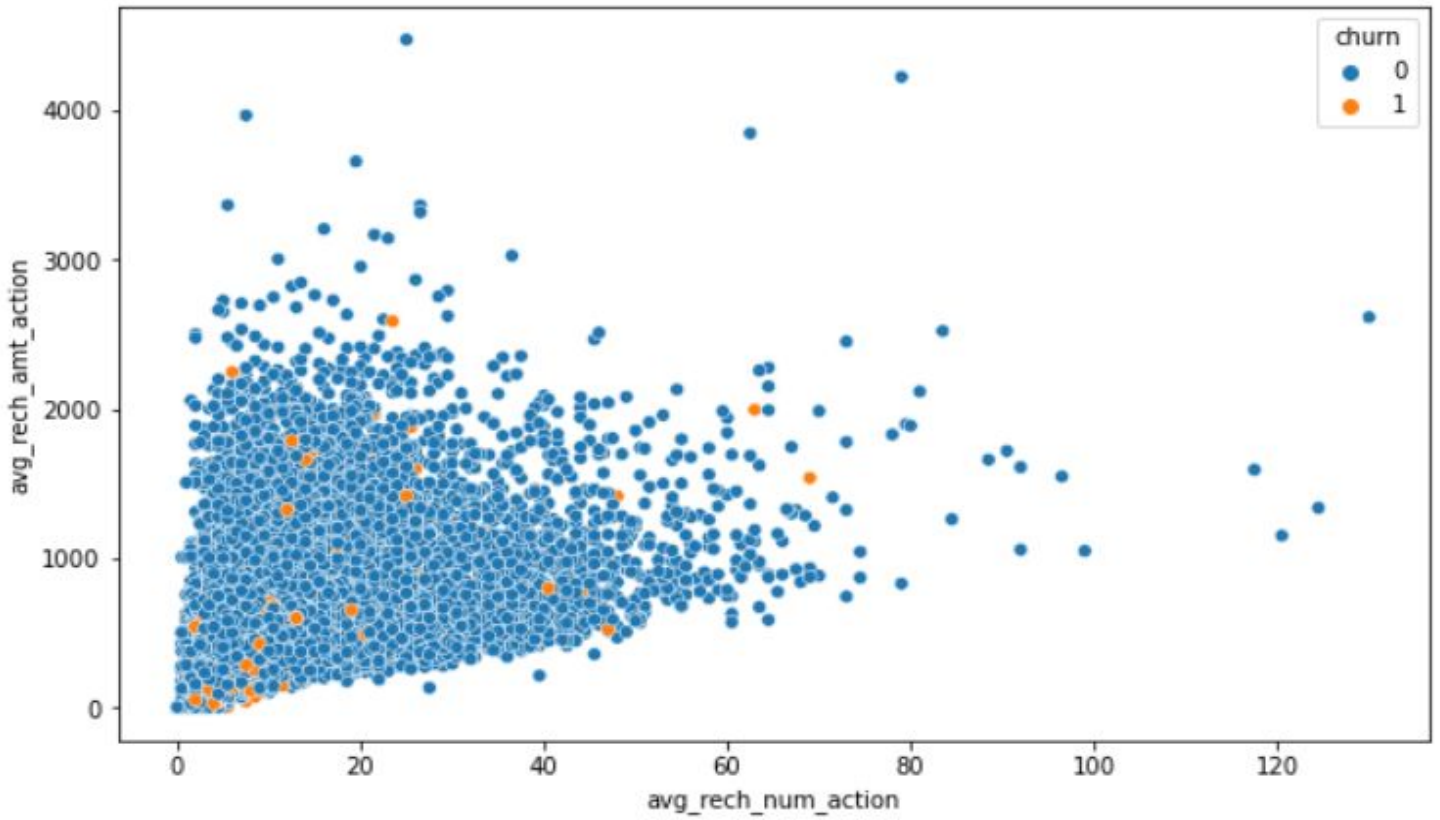
- We can see from the above plot, that the churn rate is more for the customers, whose recharge amount as well as number of recharge have decreased in the action phase than the good phase.

```
data.pivot_table(values='churn', index='decrease_rech_amt_action', columns='de
plt.ylabel('churn rate')
plt.show()
```



- We can see from the above pattern that the recharge number and the recharge amount are mostly propotional. More the number of recharge, more the amount of the recharge.

```
plt.figure(figsize=(10,6))
ax = sns.scatterplot('avg_rech_num_action','avg_rech_amt_action', hue='churn',
```



Model Building

- ***Model analysis***

1. We can see that there are few features have positive coefficients and few have negative.
2. Many features have higher p-values and hence became insignificant in the model.

- ***Coarse tuning (Auto+Manual)***
- We'll first eliminate a few features using Recursive Feature Elimination (RFE), and once we have reached a small set of variables to work with, we can then use manual feature elimination (i.e. manually eliminating features based on observing the p-values and VIFs).

Model - 1

Dep. Variable:	churn	**No. Observations:**	42850
Model:	GLM	**Df Residuals:**	42834
Model Family:	Binomial	**Df Model:**	15
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	nan
Date:	Tue, 17 Jan 2023	**Deviance:**	30008.
Time:	18:20:13	**Pearson chi2:**	4.49e+06
No. Iterations:	41		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-53.0128	4235.111	-0.013	0.990	-8353.678	8247.652
offnet_mou_7	0.6096	0.026	23.449	0.000	0.559	0.661
offnet_mou_8	-3.2532	0.106	-30.548	0.000	-3.462	-3.045
roam_og_mou_8	1.2482	0.032	39.496	0.000	1.186	1.310
std_og_t2m_mou_8	2.4408	0.094	26.101	0.000	2.258	2.624
isd_og_mou_8	-1.0212	0.194	-5.271	0.000	-1.401	-0.641
og_others_7	-1.1915	0.862	-1.382	0.167	-2.881	0.498
og_others_8	-3780.7239	3.08e+05	-0.012	0.990	-6.08e+05	6.01e+05
loc_ic_t2f_mou_8	-0.7547	0.072	-10.487	0.000	-0.896	-0.614
loc_ic_mou_8	-1.9744	0.066	-30.078	0.000	-2.103	-1.846
std_ic_t2f_mou_8	-0.7922	0.075	-10.607	0.000	-0.939	-0.646
ic_others_8	-1.4913	0.132	-11.305	0.000	-1.750	-1.233
total_rech_num_8	-0.4840	0.018	-26.977	0.000	-0.519	-0.449
monthly_2g_8	-0.9031	0.043	-20.851	0.000	-0.988	-0.818
monthly_3g_8	-0.9871	0.043	-22.711	0.000	-1.072	-0.902
decrease_vbc_action	-1.3078	0.073	-17.956	0.000	-1.451	-1.165

	Features	VIF
1	offnet_mou_8	7.45
3	std_og_t2m_mou_8	6.27
0	offnet_mou_7	1.92
8	loc_ic_mou_8	1.68
7	loc_ic_t2f_mou_8	1.21
11	total_rech_num_8	1.19
2	roam_og_mou_8	1.16
14	decrease_vbc_action	1.08
13	monthly_3g_8	1.06
6	og_others_8	1.05
12	monthly_2g_8	1.05
5	og_others_7	1.04
9	std_ic_t2f_mou_8	1.02
10	ic_others_8	1.02
4	isd_og_mou_8	1.01

- Removing column og_others_8, which is insignificatnt as it has the highest p-value 0.99

Model - 2

Dep. Variable:	churn	**No. Observations:**	42850
Model:	GLM	**Df Residuals:**	42835
Model Family:	Binomial	**Df Model:**	14
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-15034.
Date:	Tue, 17 Jan 2023	**Deviance:**	30068.
Time:	18:27:47	**Pearson chi2:**	4.51e+06
No. Iterations:	11		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.1052	0.031	-35.342	0.000	-1.167	-1.044
offnet_mou_7	0.6081	0.026	23.427	0.000	0.557	0.659
offnet_mou_8	-3.2557	0.106	-30.603	0.000	-3.464	-3.047
roam_og_mou_8	1.2491	0.031	39.747	0.000	1.188	1.311
std_og_t2m_mou_8	2.4428	0.093	26.146	0.000	2.260	2.626
isd_og_mou_8	-1.0982	0.196	-5.590	0.000	-1.483	-0.713
og_others_7	-1.8793	0.818	-2.299	0.022	-3.482	-0.277
loc_ic_t2f_mou_8	-0.7548	0.072	-10.491	0.000	-0.896	-0.614
loc_ic_mou_8	-1.9714	0.066	-30.058	0.000	-2.100	-1.843
std_ic_t2f_mou_8	-0.8020	0.075	-10.727	0.000	-0.949	-0.655
ic_others_8	-1.4871	0.132	-11.278	0.000	-1.746	-1.229
total_rech_num_8	-0.4864	0.018	-27.146	0.000	-0.522	-0.451
monthly_2g_8	-0.9066	0.043	-20.866	0.000	-0.992	-0.821
monthly_3g_8	-0.9862	0.043	-22.700	0.000	-1.071	-0.901
decrease_vbc_action	-1.3097	0.073	-17.994	0.000	-1.452	-1.167

	Features	VIF
1	offnet_mou_8	7.45
3	std_og_t2m_mou_8	6.27
0	offnet_mou_7	1.92
7	loc_ic_mou_8	1.68
6	loc_ic_t2f_mou_8	1.21
10	total_rech_num_8	1.19
2	roam_og_mou_8	1.16
13	decrease_vbc_action	1.08
12	monthly_3g_8	1.06
11	monthly_2g_8	1.05
8	std_ic_t2f_mou_8	1.02
4	isd_og_mou_8	1.01
9	ic_others_8	1.01
5	og_others_7	1.00

As we can see from the model summary that all the variables p-values are significant and offnet_mou_8 column has the highest VIF 7.45. Hence, deleting offnet_mou_8 column

Model - 3

Dep. Variable:	churn	**No. Observations:**	42850
Model:	GLM	**Df Residuals:**	42836
Model Family:	Binomial	**Df Model:**	13
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-15720.
Date:	Tue, 17 Jan 2023	**Deviance:**	31440.
Time:	18:28:24	**Pearson chi2:**	3.92e+06
No. Iterations:	11		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.2058	0.032	-37.536	0.000	-1.269	-1.143
offnet_mou_7	0.3665	0.022	16.456	0.000	0.323	0.410
roam_og_mou_8	0.7135	0.024	29.260	0.000	0.666	0.761
std_og_t2m_mou_8	-0.2474	0.022	-11.238	0.000	-0.291	-0.204
isd_og_mou_8	-1.3811	0.212	-6.511	0.000	-1.797	-0.965
og_others_7	-2.4711	0.872	-2.834	0.005	-4.180	-0.762
loc_ic_t2f_mou_8	-0.7102	0.075	-9.532	0.000	-0.856	-0.564
loc_ic_mou_8	-3.3287	0.057	-58.130	0.000	-3.441	-3.216
std_ic_t2f_mou_8	-0.9503	0.078	-12.181	0.000	-1.103	-0.797
ic_others_8	-1.5131	0.129	-11.771	0.000	-1.765	-1.261
total_rech_num_8	-0.5060	0.018	-28.808	0.000	-0.540	-0.472
monthly_2g_8	-0.9279	0.044	-21.027	0.000	-1.014	-0.841
monthly_3g_8	-1.0943	0.046	-23.615	0.000	-1.185	-1.004
decrease_vbc_action	-1.3293	0.072	-18.478	0.000	-1.470	-1.188

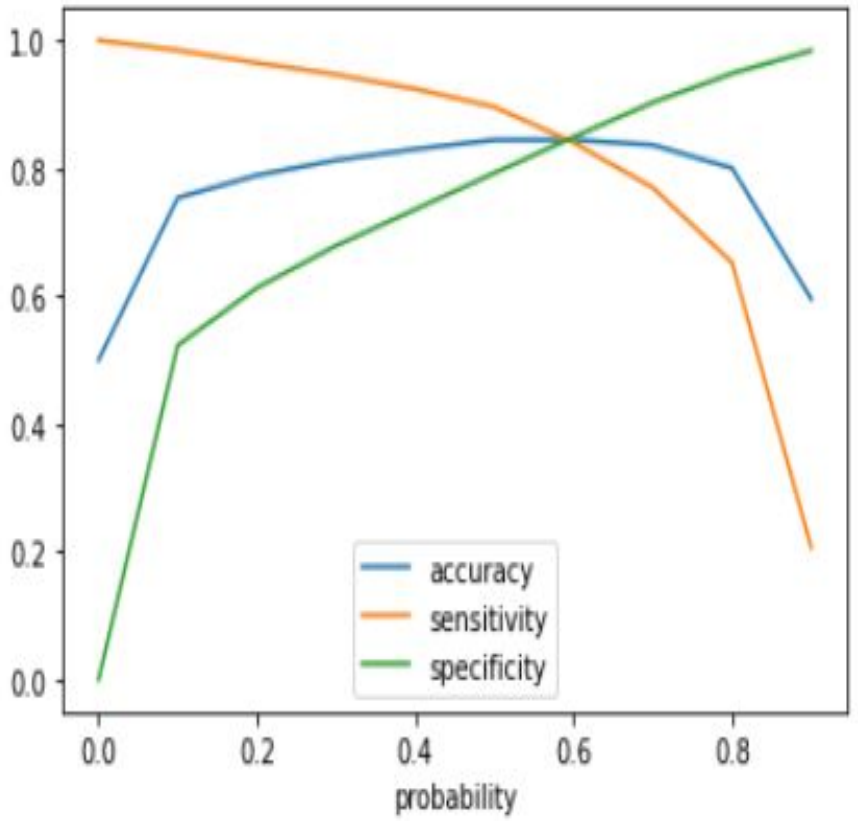
	Features	VIF
2	std_og_t2m_mou_8	1.87
0	offnet_mou_7	1.72
6	loc_ic_mou_8	1.33
5	loc_ic_t2f_mou_8	1.21
9	total_rech_num_8	1.17
12	decrease_vbc_action	1.07
1	roam_og_mou_8	1.06
11	monthly_3g_8	1.06
10	monthly_2g_8	1.05
7	std_ic_t2f_mou_8	1.02
3	isd_og_mou_8	1.01
8	ic_others_8	1.01
4	og_others_7	1.00

- Now from the model summary and the VIF list we can see that all the variables are significant and there is no multicollinearity among the variables.
  Hence, we can conclused that ***Model-3 log_no_pca_3 will be the final model***.

Plotting accuracy, sensitivity and specificity

At point 0.8 where the three parameters cut each other, we can see that there is a balance between sensitivity and specificity with a good accuracy.
Here we are intended to acheive better sensitivity than accuracy and specificity.
Though as per the above curve, we should take 0.8 as the optimum probability cutoff, we are taking ***0.6*** for acheiving higher sensitivity, which is our main goal.

```
# Plotting accuracy, sensitivity and specificity for different probabilities.
cutoff_df.plot('probability', ['accuracy','sensitivity','specificity'])
plt.show()
```



Metrics – Observation’s

- Train set
  - Accuracy:- 84%
  - Sensitivity:- 83%
  - Specificity:- 84%
- Test set
  - Accuracy:- 83%
  - Sensitivity:- 77%
  - Specificity:- 84%
- Overall, the model is performing well in the test set, what it had learnt from the train set.

Predictors

- Below are few top variables selected in the logistic regression model.
- loc_ic_mou_8 = -3.3287
- og_others_7 = -2.4711
- ic_others_8 = -1.5131
- isd_og_mou_8 = -1.3811
- decrease_vbc_action = -1.3293
- monthly_3g_8 = -1.0943
- std_ic_t2f_mou_8 = -0.9503
- monthly_2g_8 = -0.9279
- loc_ic_t2f_mou_8 = -0.7102
- roam_og_mou_8 = 0.7135
- We can see most of the top variables have negative coefficients. That means, the variables are inversely correlated with the churn probability.
- E.g.:- If the local incoming minutes of usage (loc_ic_mou_8) is lesser in the month of August than any other month, then there is a higher chance that the customer is likely to churn.

Business recommendation

- Target the customers, whose minutes of usage of the incoming local calls and outgoing ISD calls are less in the action phase (mostly in the month of August).
- Target the customers, whose outgoing others charge in July and incoming others on August are less.
- Also, the customers having value based cost in the action phase increased are more likely to churn than the other customers. Hence, these customers may be a good target to provide offer.
- Customers, whose monthly 3G recharge in August is more, are likely to be churned.
- Customers having decreasing STD incoming minutes of usage for operators T to fixed lines of T for the month of August are more likely to churn.
- Customers decreasing monthly 2g usage for August are most probable to churn.
- Customers having decreasing incoming minutes of usage for operators T to fixed lines of T for August are more likely to churn.
- roam_og_mou_8 variables have positive coefficients (0.7135). That means for the customers, whose roaming outgoing minutes of usage is increasing are more likely to churn.