Telecom-Churn Case Study

Rakesh Khanna

Telecom DataSet

This assignment aims to give us an idea of applying prediction model in a real business scenario for Telecom. In this assignment, we need to develop an understanding of the data provided for the telecom users and need to come up with a model to predict whether the user will churn or not. We need to determine what factors are the indicators of the churn behavior and advise the telecom operator to look into those factors to reduce the probability of user churn

We will look into the data using Exploratory Data Analysis Techniques and create a Logistic Regression model.

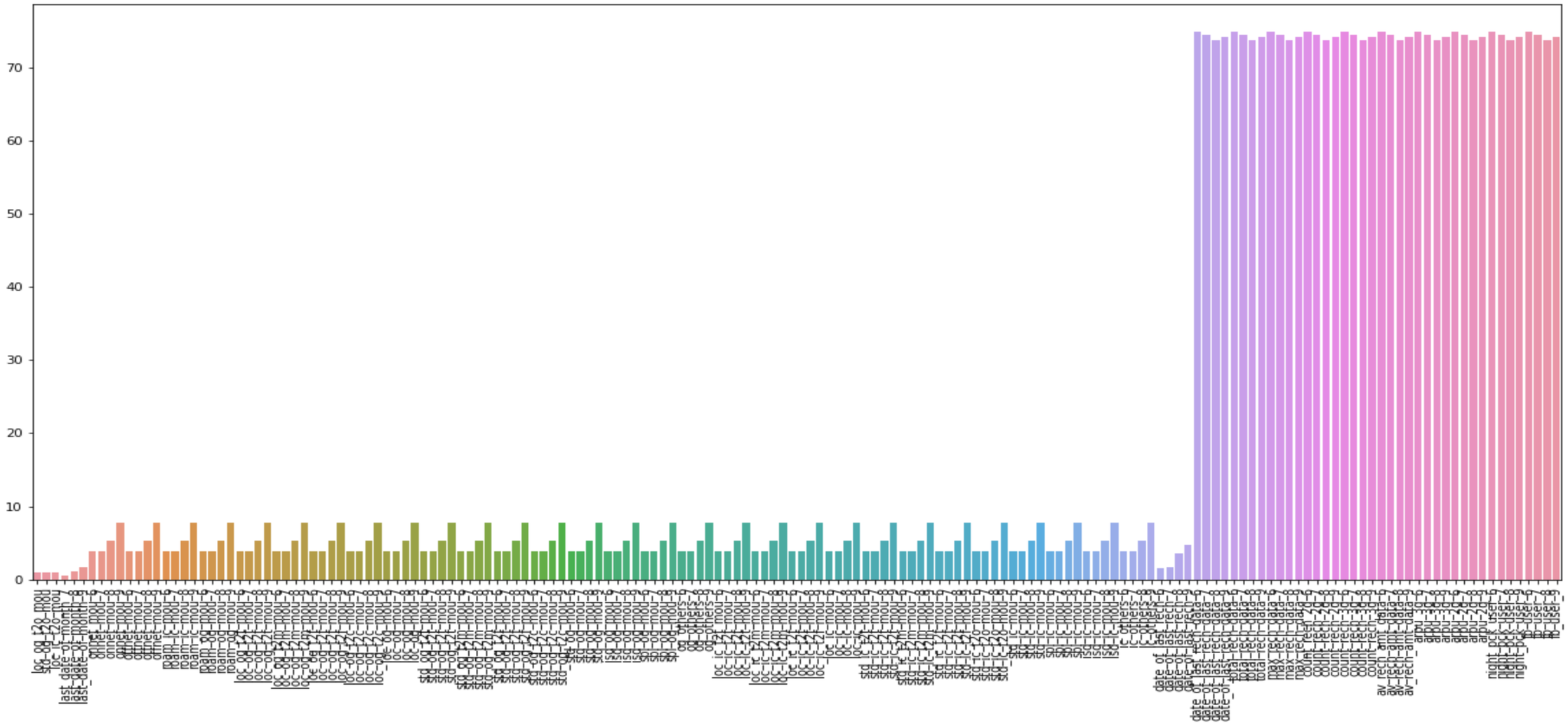
This dataset provided has 2 files as explained below:

1. 'telecom_churn_data.csv' contains all the information of the telecom users at the time of application.
The data tells about different services a user avails and payment details for the services**.**

2. Data+Dictionary-+Telecom+Churn+Case+Study.xlsx' is data dictionary which describes the meaning of the variables.

Null Value Analysis

Following Plot shows the percentage of null values present in each column where null values are existent. As most of the data are numeric and absence of data my mean 0 For sake of importance of the data we will impute relevant columns like recharge columns to 0



High Value Customers

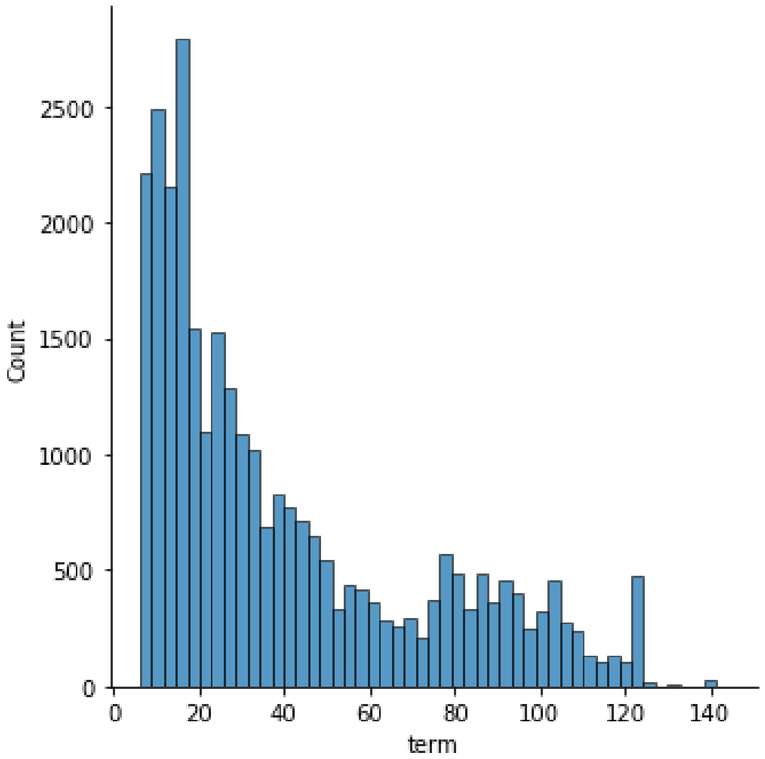
Since we are concerned about the high value customers, we will consider top 70 Percentile data based upon the recharge amount for month 6 and 7.

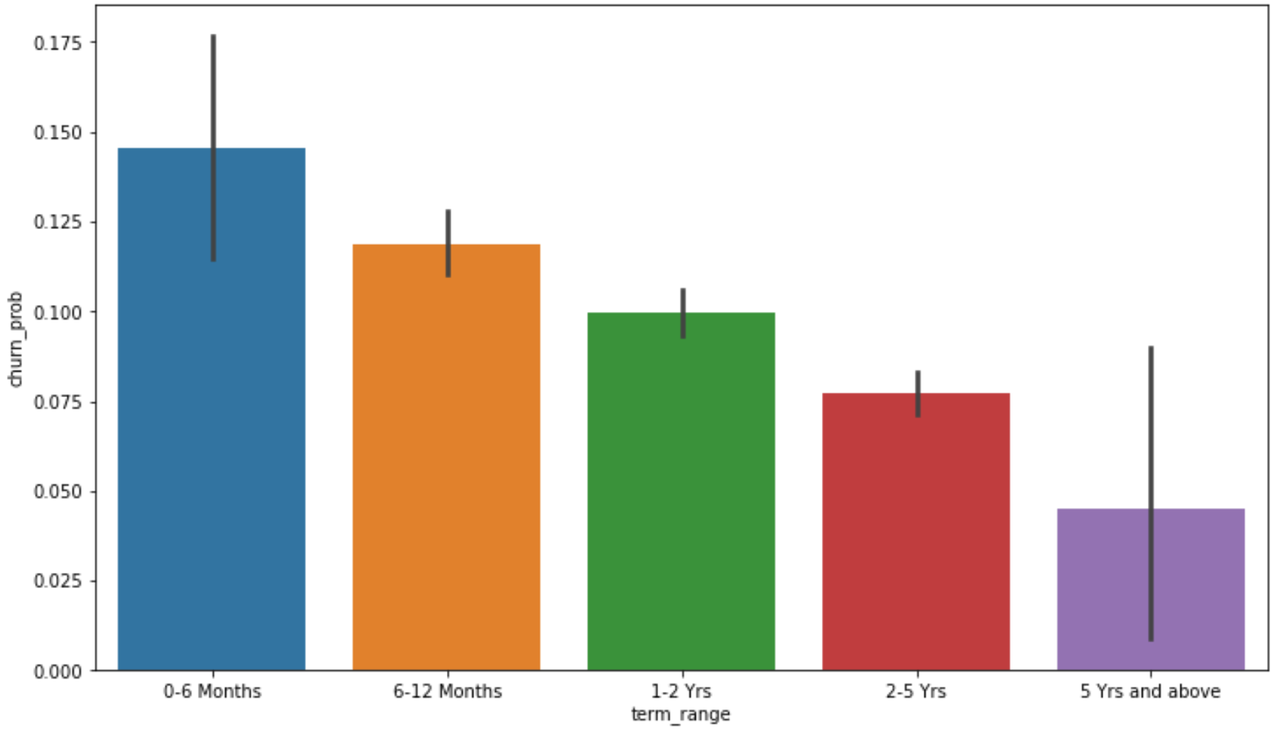
We will use the Following formula to find the average recharge amount for voice and data services and take an average of month 6 and 7. This will reduce the initial 100,000 records to 30,000 records as the initial dataset for consideration.

```
(((total_rech_data_6 * av_rech_amt_data_6)+ total_rech_amt_6) +
((total_rech_data_7 * av_rech_amt_data_7)+ total_rech_amt_7)))/2
```

AON Analysis

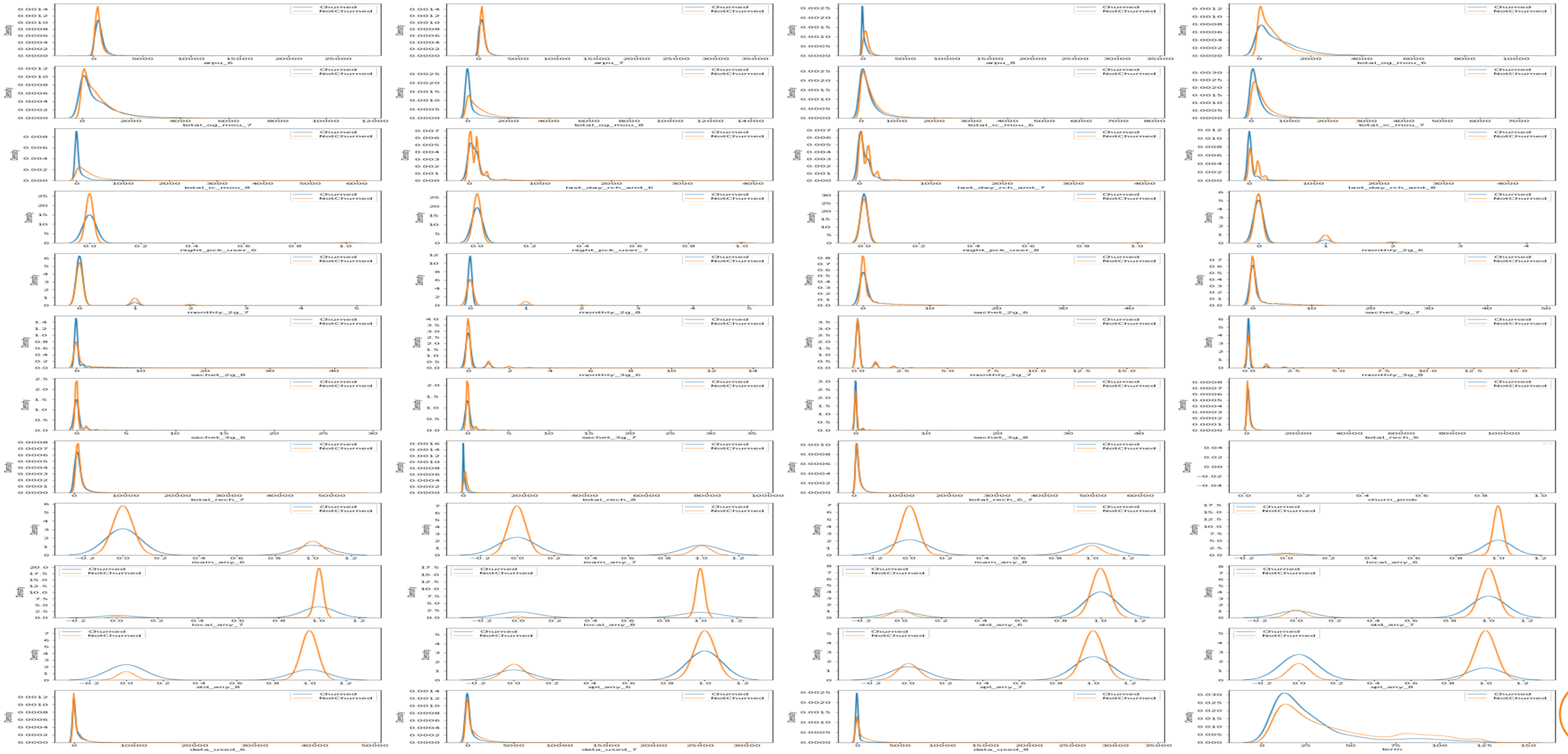
AON (Age on Network) typically defines how long the customer is associated with the Telecom Operator. The analysis here shows that probability of user churning is greatest in first 6 months and it gradually decreases with the increase of tenure with the operator.





Distribution plot for columns

Univariate analysis of both the data frames put into comparison is inconclusive



Final LR Model Summary

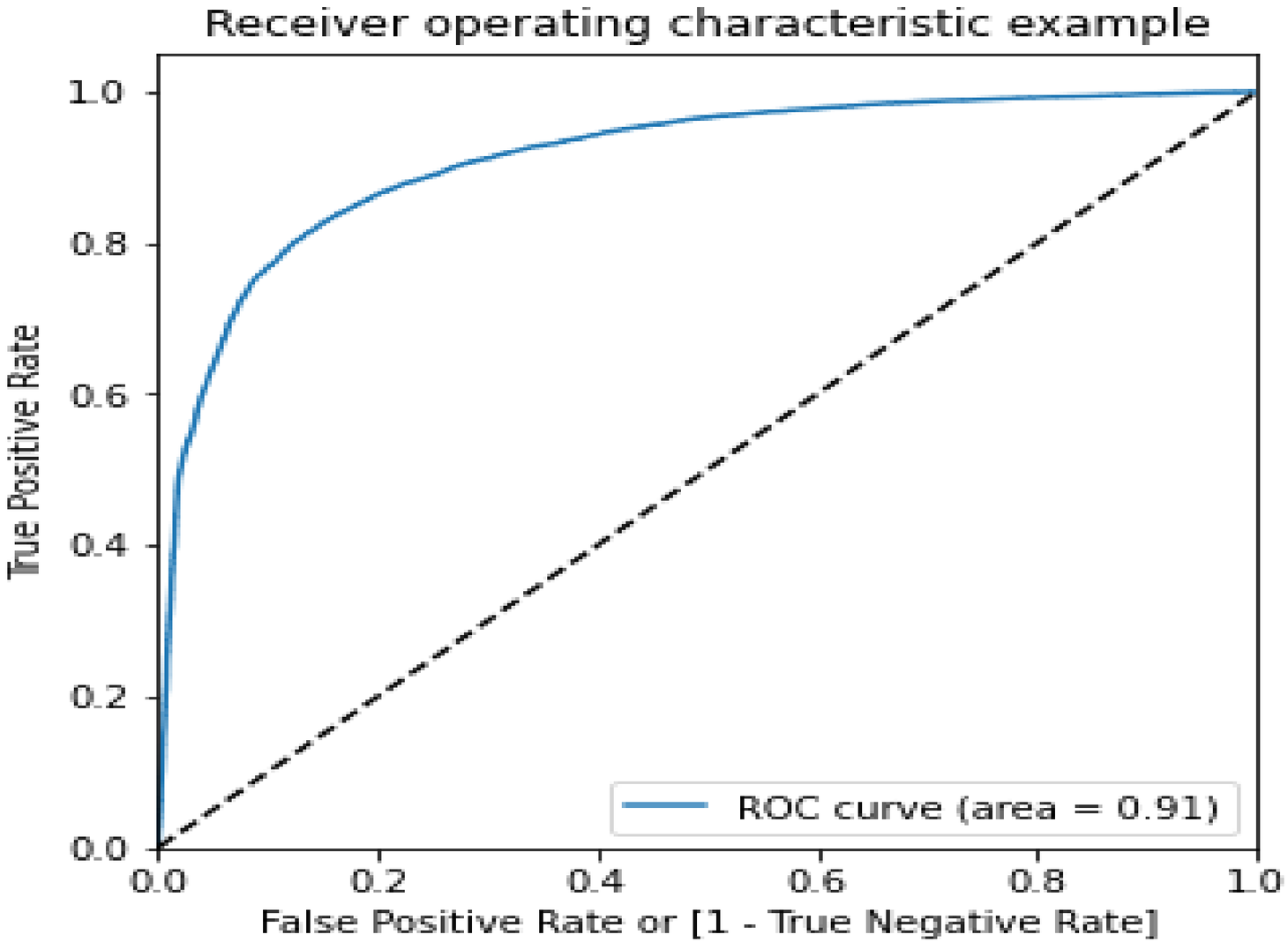
Generalized Linear Model Regression Results

Dep. Variable:	churn_prob	No. Observations:	38520
Model:	GLM	Df Residuals:	38502
Model Family:	Binomial	Df Model:	17
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-14807.
Date:	Tue, 14 Mar 2023	Deviance:	29615.
Time:	12:03:42	Pearson chi2:	2.12e+09
No. Iterations:	7	Pseudo R-squ. (CS):	0.4607
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	1.8045	0.133	13.554	0.000	1.544	2.065
arpu_6	10.2428	1.389	7.376	0.000	7.521	12.964
total_og_mou_7	9.3176	0.426	21.878	0.000	8.483	10.152
total_og_mou_8	-18.4410	0.695	-26.535	0.000	-19.803	-17.079
total_ic_mou_6	1.1911	0.522	2.282	0.022	0.168	2.214
total_ic_mou_7	13.1324	0.869	15.120	0.000	11.430	14.835
total_ic_mou_8	-30.2675	0.886	-34.166	0.000	-32.004	-28.531
last_day_rch_amt_6	2.0887	0.471	4.436	0.000	1.166	3.011
last_day_rch_amt_8	-13.6145	0.703	-19.378	0.000	-14.992	-12.237
monthly_2g_6	-2.0857	0.205	-10.177	0.000	-2.487	-1.684
monthly_2g_8	-7.6344	0.406	-18.791	0.000	-8.431	-6.838
sachet_2g_8	-14.1413	0.571	-24.764	0.000	-15.260	-13.022
monthly_3g_8	-16.3197	0.956	-17.066	0.000	-18.194	-14.445
sachet_3g_8	-16.2047	1.367	-11.858	0.000	-18.883	-13.526
total_rech_7	9.8552	0.827	11.920	0.000	8.235	11.476
total_rech_8	14.8310	3.143	4.718	0.000	8.670	20.992
local_any_8	-2.3873	0.054	-44.542	0.000	-2.492	-2.282
data_used_8	-6.1450	1.028	-5.979	0.000	-8.159	-4.130

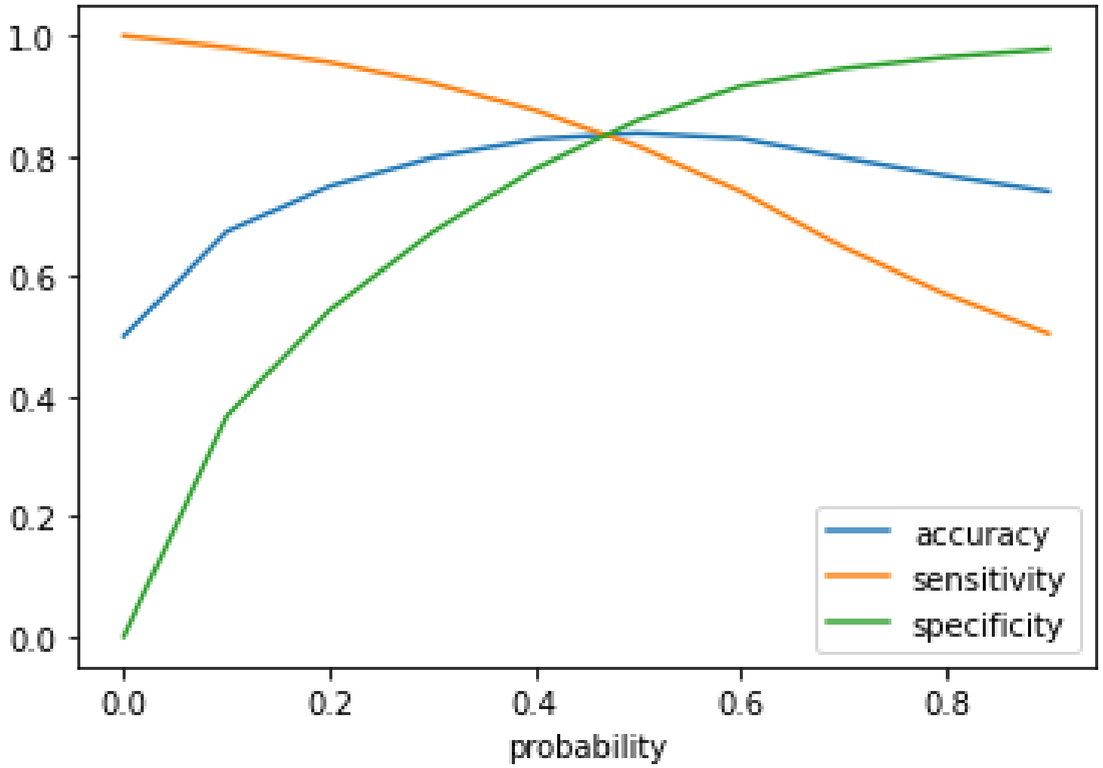
ROC-AUC Curve

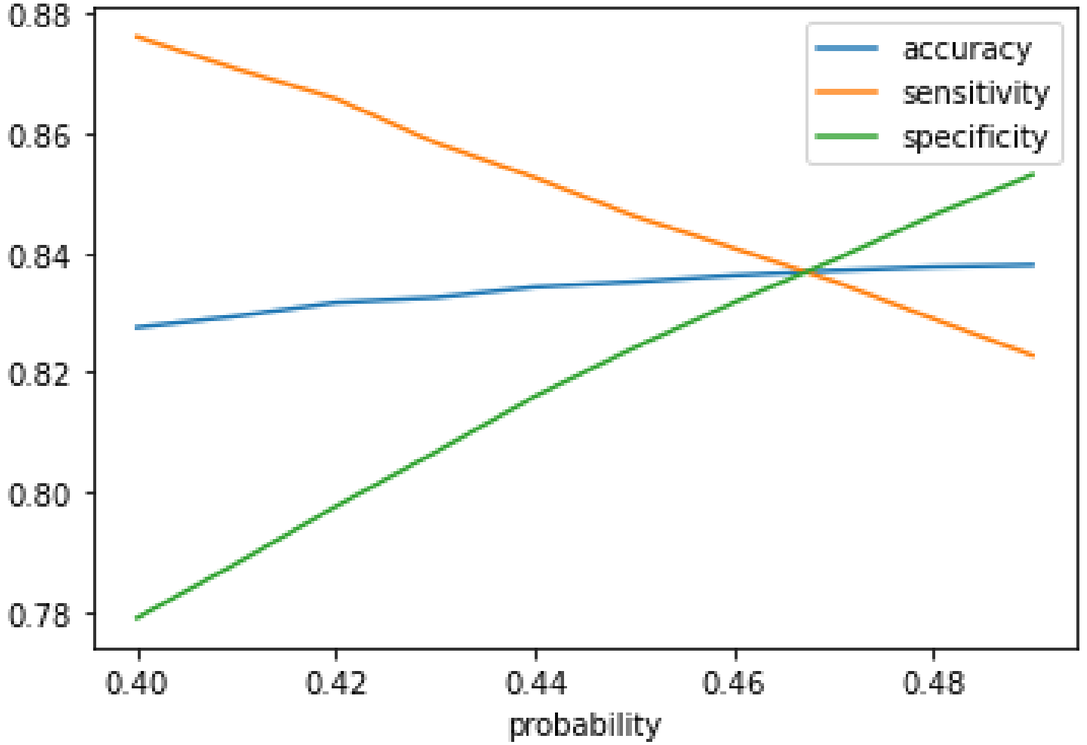
Area under the ROC-AUC Curve is 0.91



Cutoff Calculation

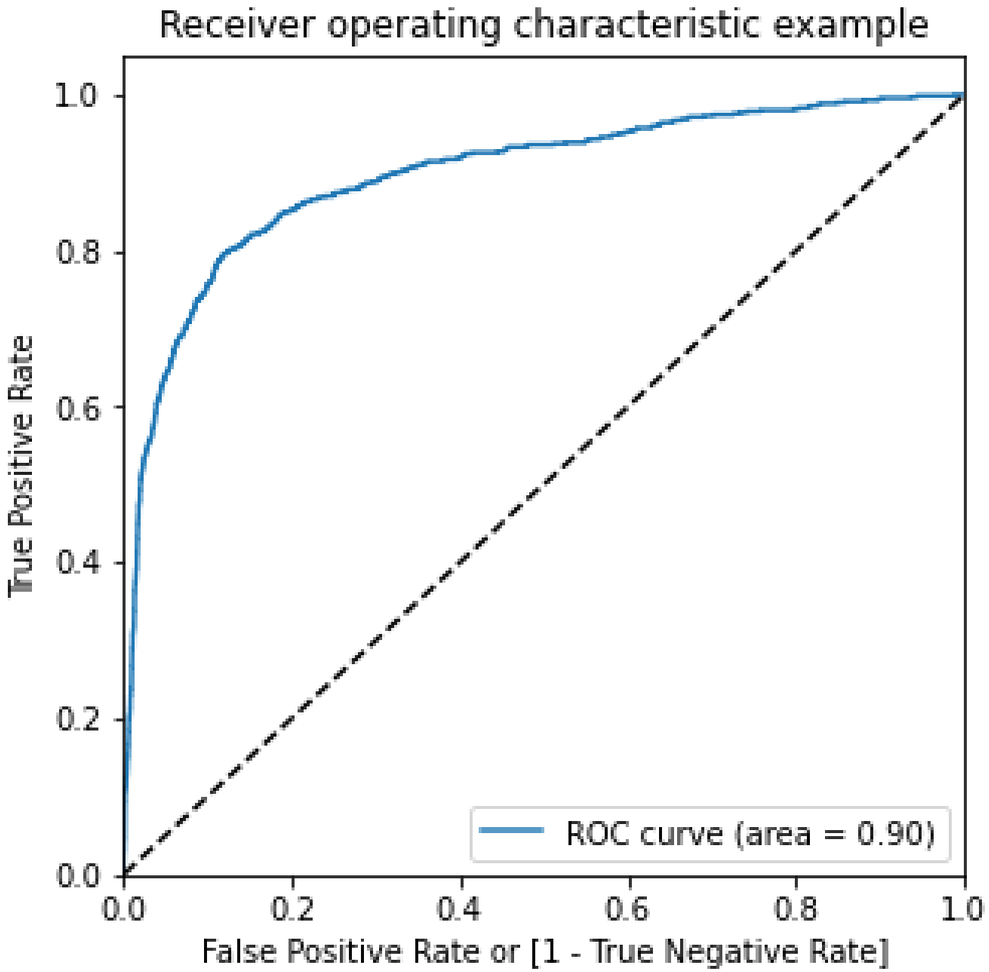
Initial cutoff calculations for Accuracy Sensitivity and Specificity was calculated between 0.4 and 0.5. Further on calculation of more refined calculation exact cutoff is calculated to be **0.47**



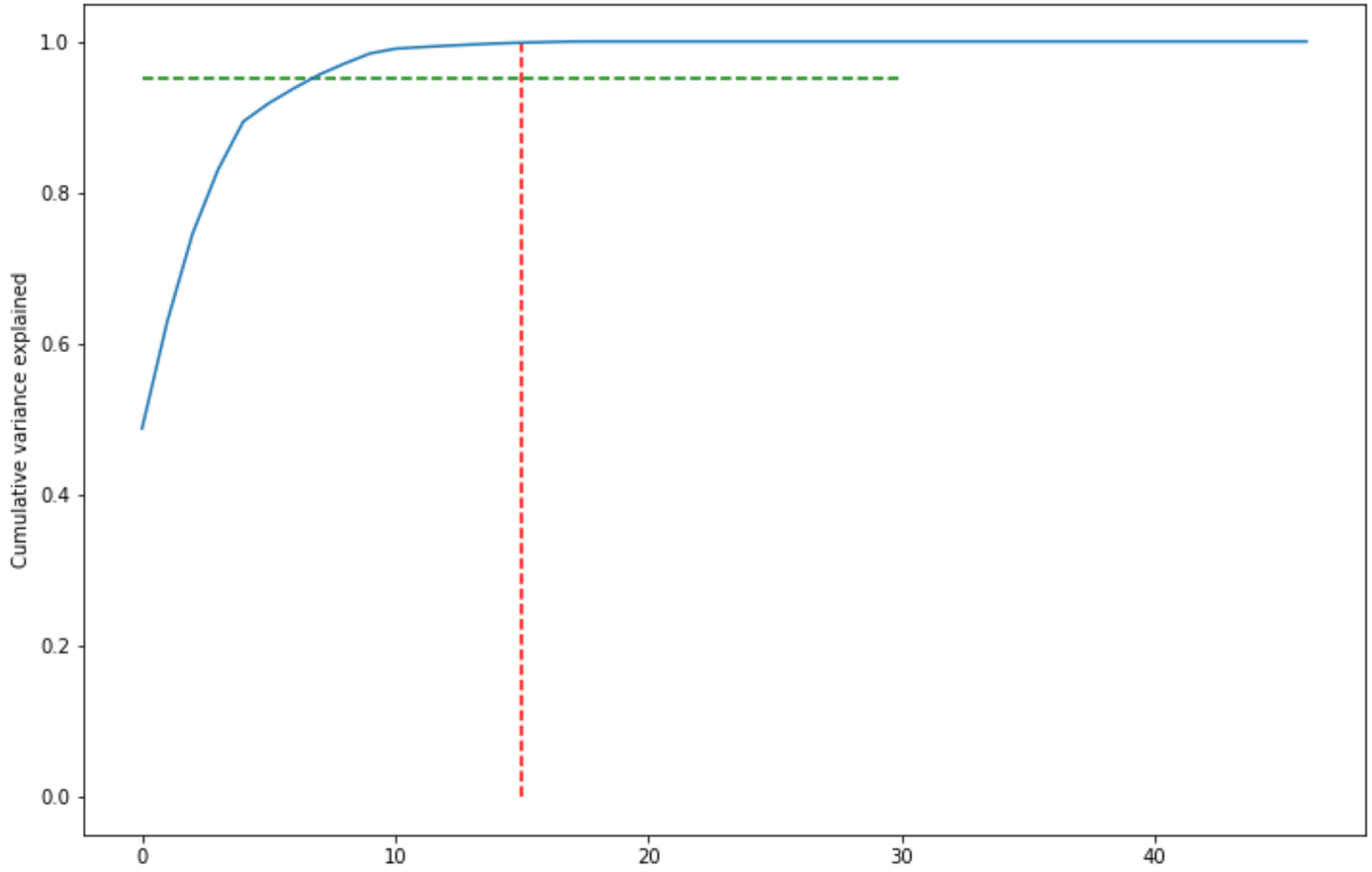


ROC-AUC Curve for Test Data

Area under the ROC-AUC Curve is 0.90



PCA Cumulative variance



Summary

We have analysed the two model creation techniques Logistic Regression with RFE Logistic regression with PCA.
Following is the summary of results

Logistic Regression
- Train Accuracy : ~91% . Test Accuracy : ~90%

Logistic regression with PCA
- Train Accuracy : ~92% . Test Accuracy : ~92%

Suggestions

The incoming calls of all categories are important factor in understanding the possibility of churn. Hence, the telecom operator should focus on incoming calls data should provide some kind of special offers to the customers whose incoming calls are turning lower.