

Telecom Churn Analysis Case Study

Team – Suresh, Shashi & Jita

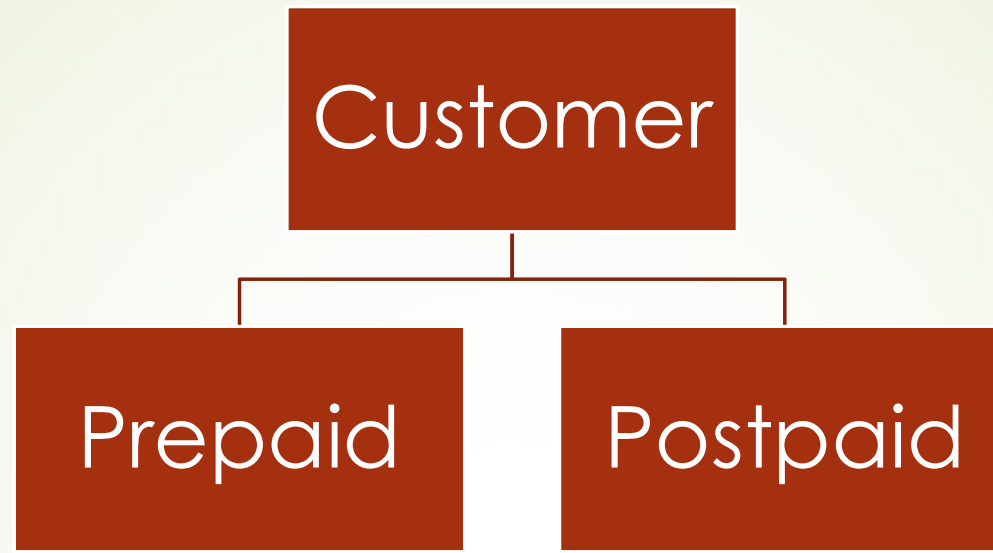
Business problem overview

In the telecom industry, customers are able to choose from multiple service providers and actively switch from one operator to another. In this highly competitive market, the telecommunications industry experiences an average of 15-25% annual churn rate. Given the fact that it costs 5-10 times more to acquire a new customer than to retain an existing one, **customer retention** has now become even more important than customer acquisition. For many incumbent operators, retaining high profitable customers is the number one business goal.

To reduce customer churn, telecom companies need to **predict which customers are at high risk of churn**.

In this project, you will analyse customer-level data of a leading telecom firm, build predictive models to identify customers at high risk of churn and identify the main indicators of churn

Understanding and defining churn



(customers pay/recharge with a certain amount in advance and then use the services).

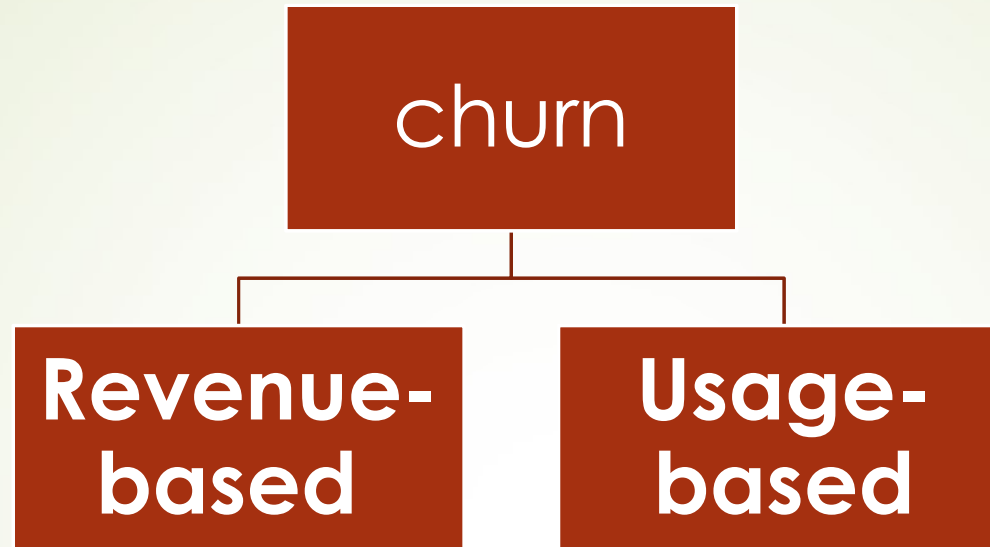
In the prepaid model, customers who want to switch to another network can simply stop using the services without notice, and it is hard to know whether someone has churned or is simply not using the services temporarily (e.g. someone may be on a trip abroad for a few days or two and then intend to resume using the services).

Postpaid (customers pay a monthly/annual bill after using the services)

In the postpaid model, when customers want to switch to another operator, they usually inform the existing operator to terminate the services, and you directly know that this is an instance of churn.

Churn prediction is usually more critical (and non-trivial) for prepaid customers, and the term should be defined carefully. Also, prepaid is the most common model in India and Southeast Asia, while postpaid is more common in Europe and North America.

Understanding and defining churn




have generated less than INR 4
month in total/average/median
ue

Customers who have not done any usage,
incoming or outgoing - in terms of calls, int
over a period of time.

project, we will use the **usage-based definition** to define churn.

High Value Churn

In the Indian and Southeast Asian markets, approximately 80% of revenue comes from the top 20% of customers (called high-value customers). Thus, if we can reduce the churn of high-value customers, we will be able to reduce significant revenue leakage.



In this project, you will define high-value customers based on a certain metric (mentioned later below) and predict churn only on high-value customers.

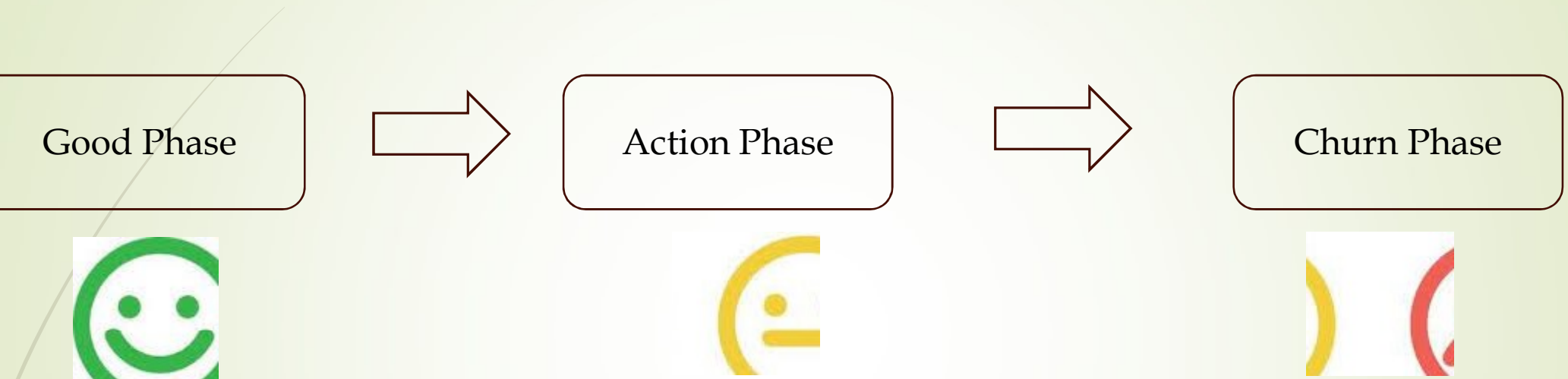
Understanding Data

Understanding the business objective and the data

The dataset contains customer-level information for a span of four consecutive months - June, July, August and September. The months are encoded as 6, 7, 8 and 9, respectively.

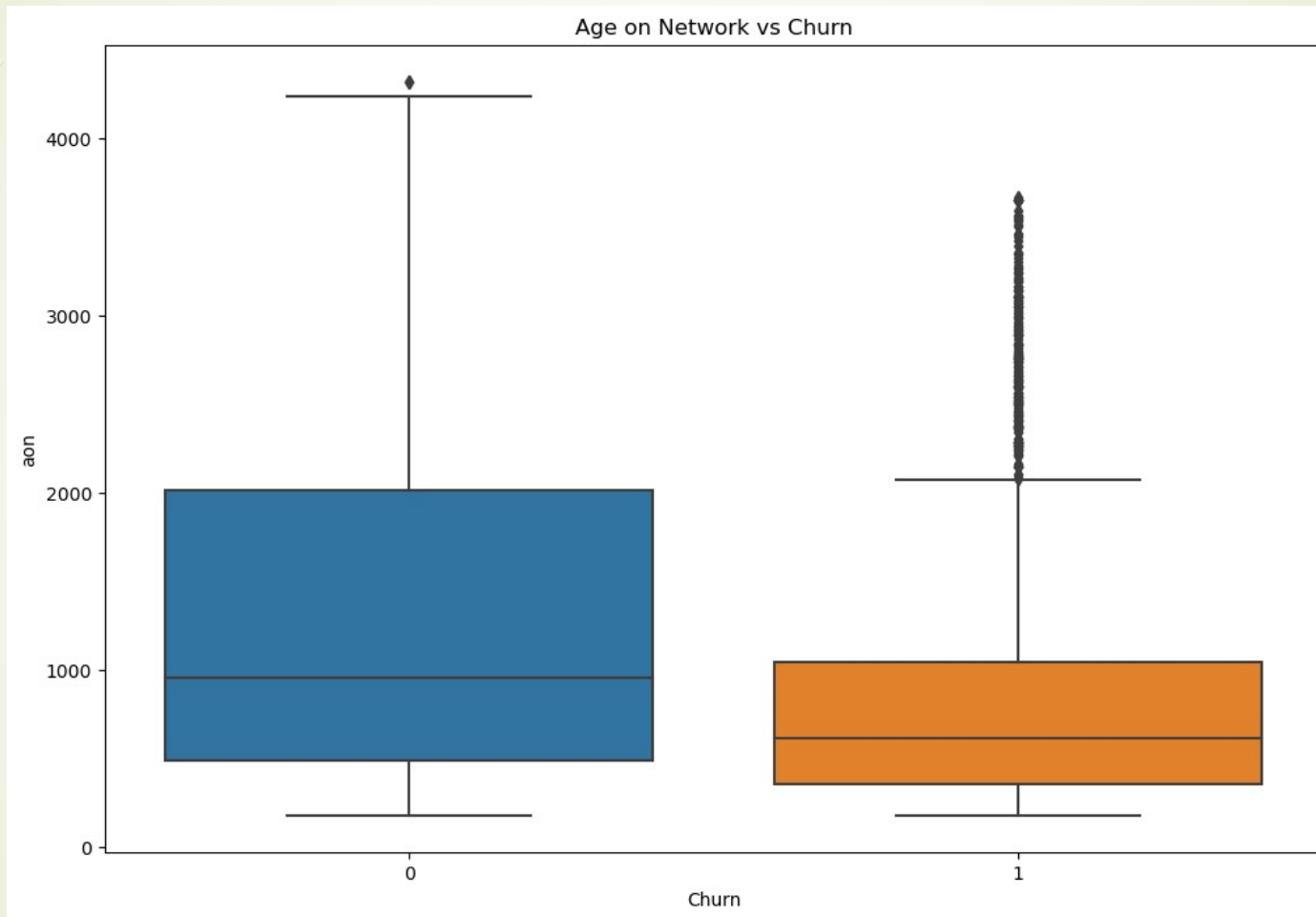
The **business objective** is to predict the churn in the last (i.e. the ninth) month using the data (features) from the first three months. To do this task well, understanding the typical customer behaviour during churn will be helpful

Customer behavior during churn

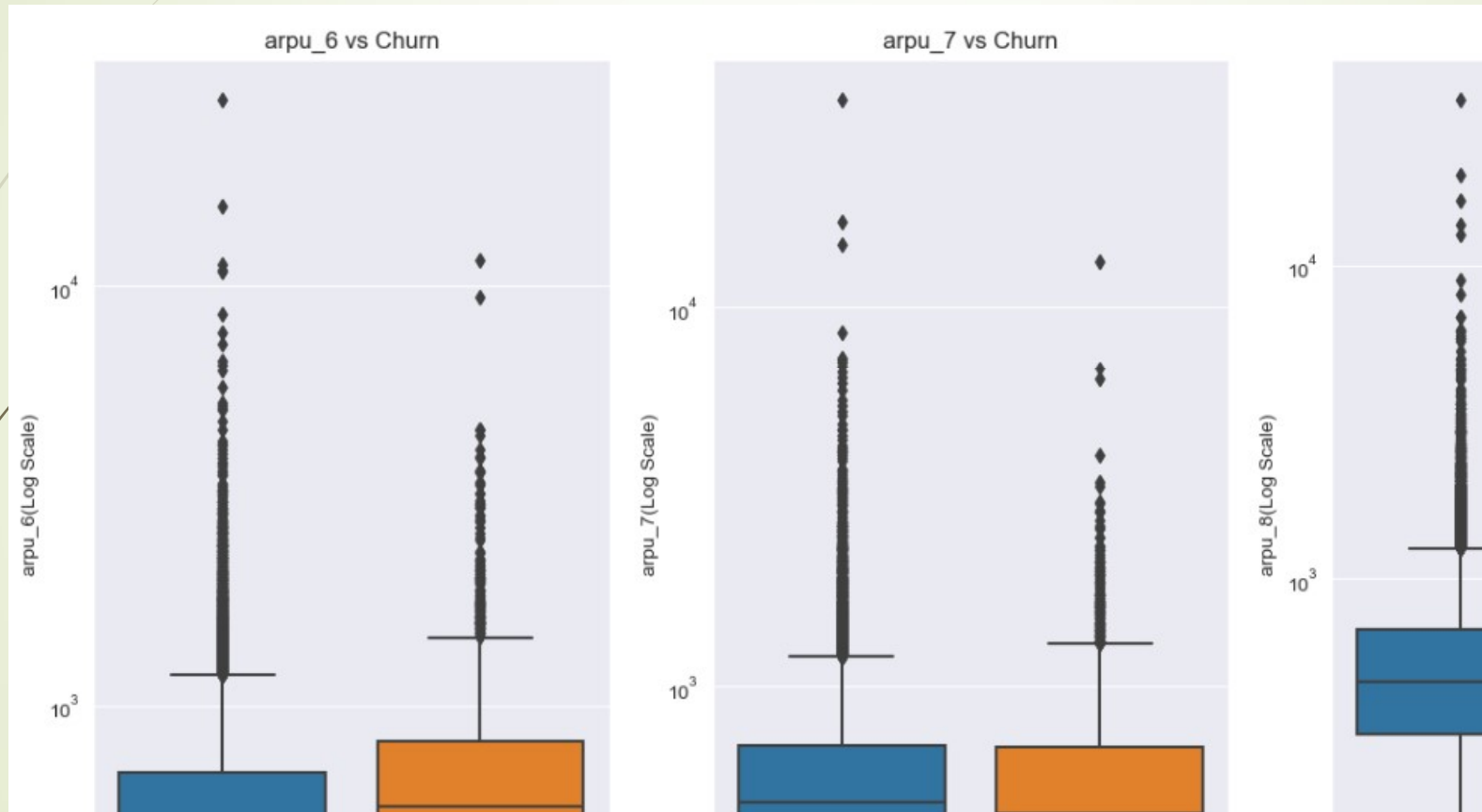


In this case, since we are working over a four-month window, the first two months are the 'good' phase, the third month is the 'action' phase, and the fourth month is the 'churn' phase.

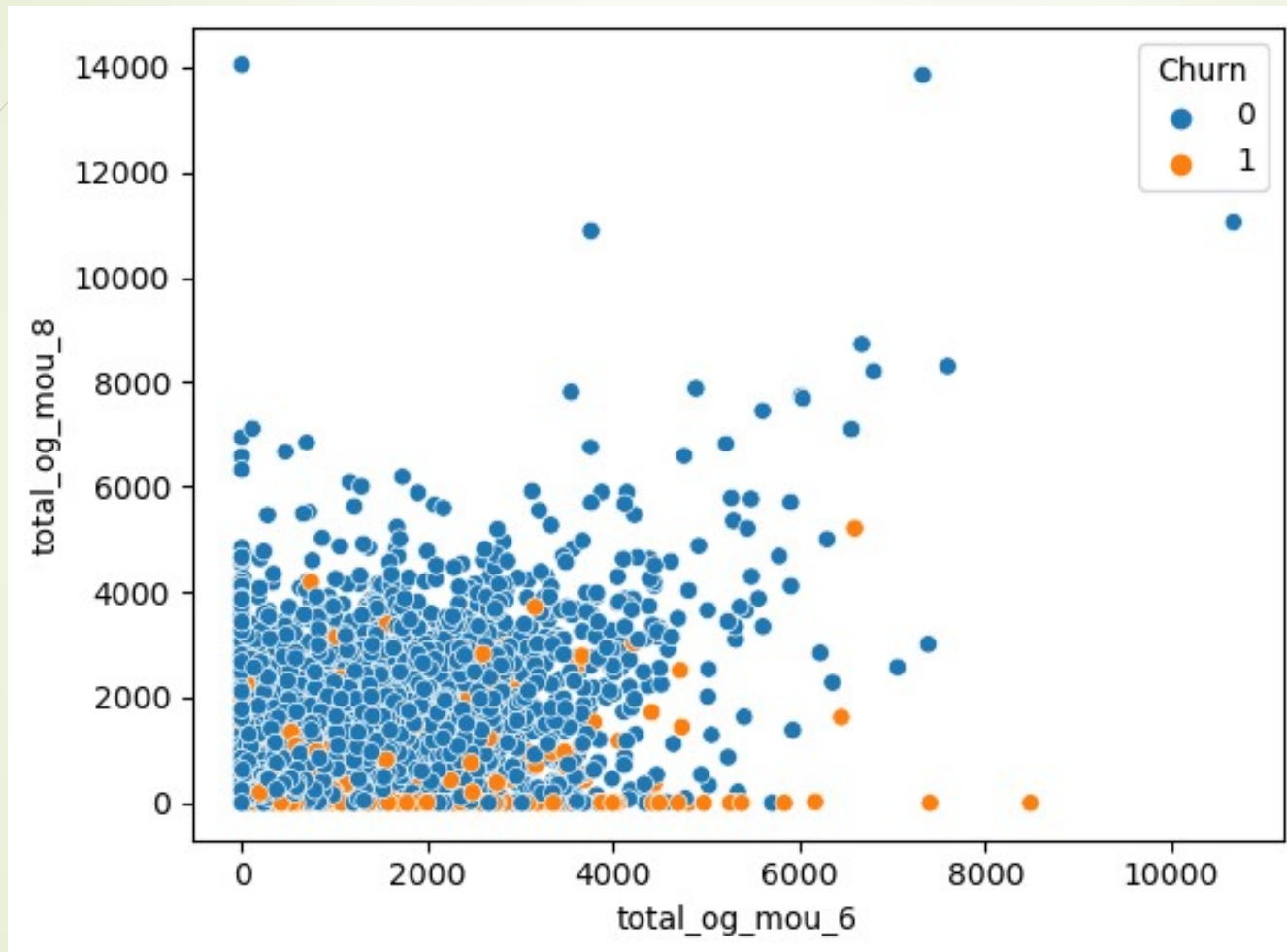
Univariate Analysis



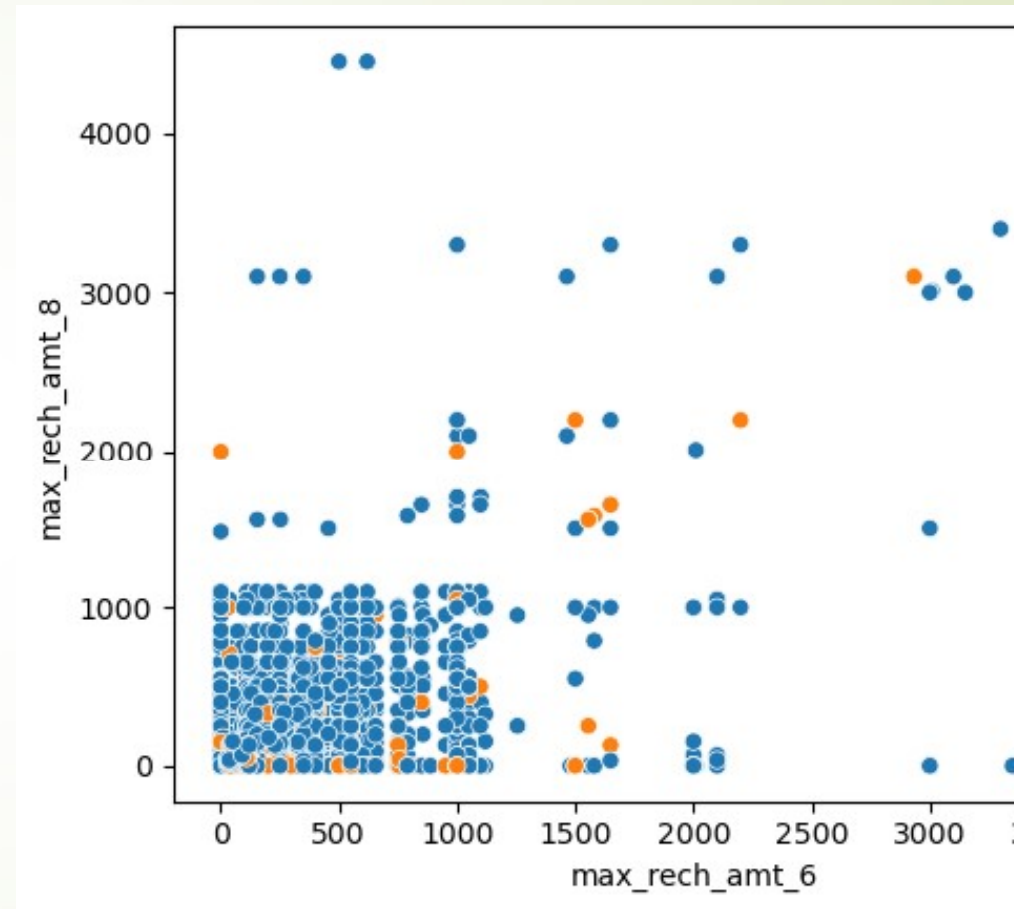
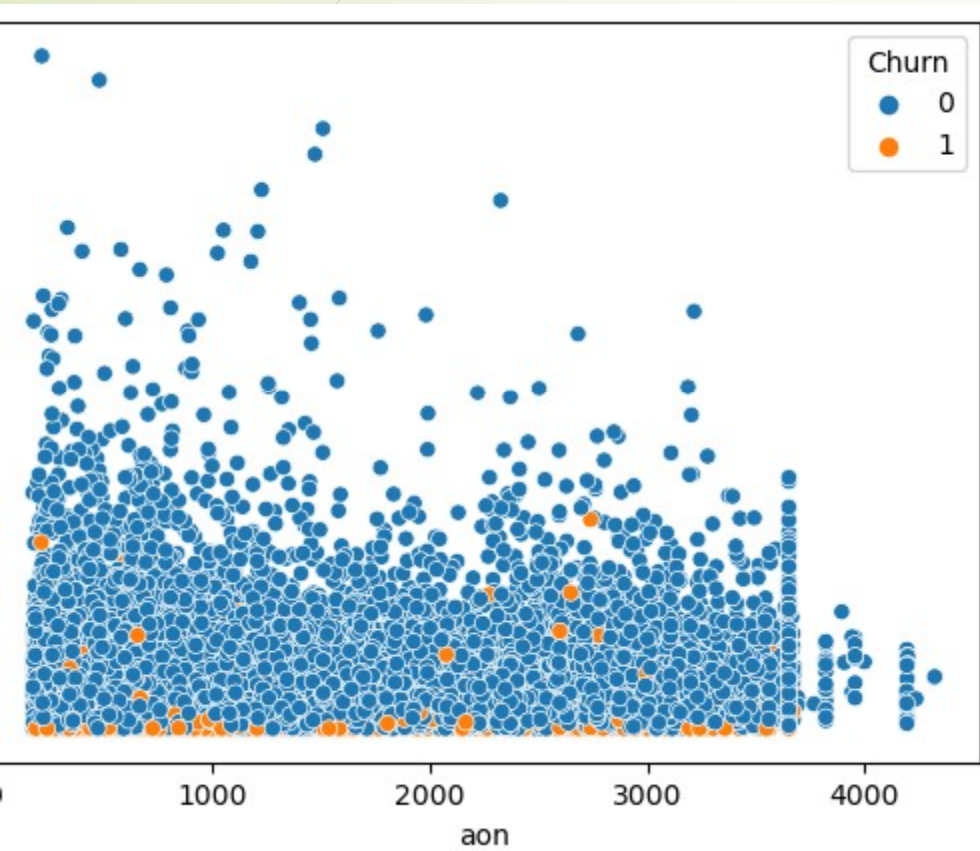
Univariate Analysis



ivariate Analysis



ivariate Analysis



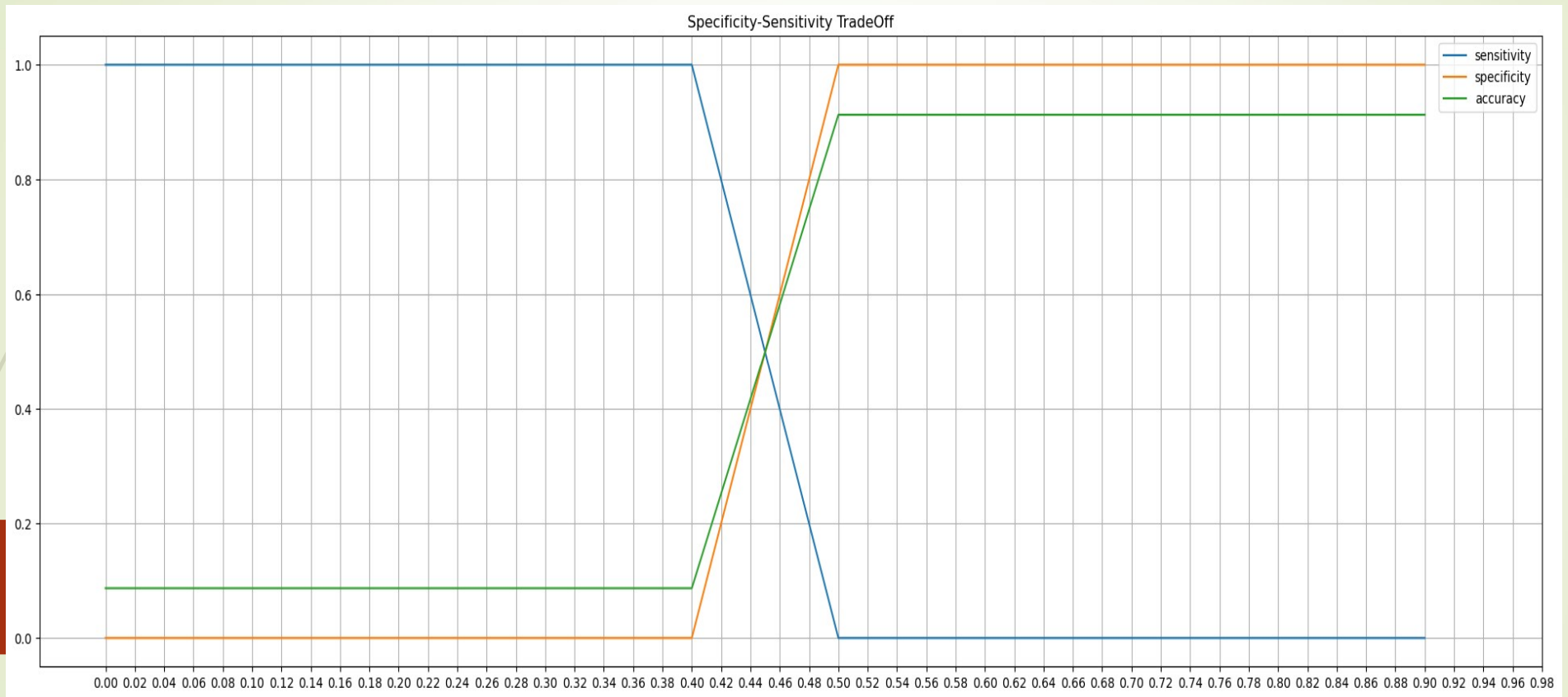
Procedure

- Test Train Split
- Class Imbalance
- Standardization
- Modelling
- Model 1 : Logistic Regression with RFE & Manual Elimination (Interpretable Model)
- Model 2 : PCA + Logistic Regression
- Model 3 : PCA + Random Forest Classifier
- Model 4 : PCA + XGBoos

Modelling

A decorative graphic on the left side of the slide, featuring a red arrow pointing right, a thin grey line, and a curved grey line.

Baseline Performance - Finding Optimum Probabilistic Cutoff



Baseline Performance at Optimum Cutoff

```
print('\n\nTest Performance : \n')  
model_metrics(test_matrix)
```

 Train Performance:

Accuracy : 0.087

Sensitivity / True Positive Rate / Recall

Specificity / True Negative Rate : 0.

Precision / Positive Predictive Value

F1-score : 0.16

Test Performance :



Logistic Regression with RFE Selected Columns

Model -I

```
vif(X_train_resampled, logr_
```




VIF P-

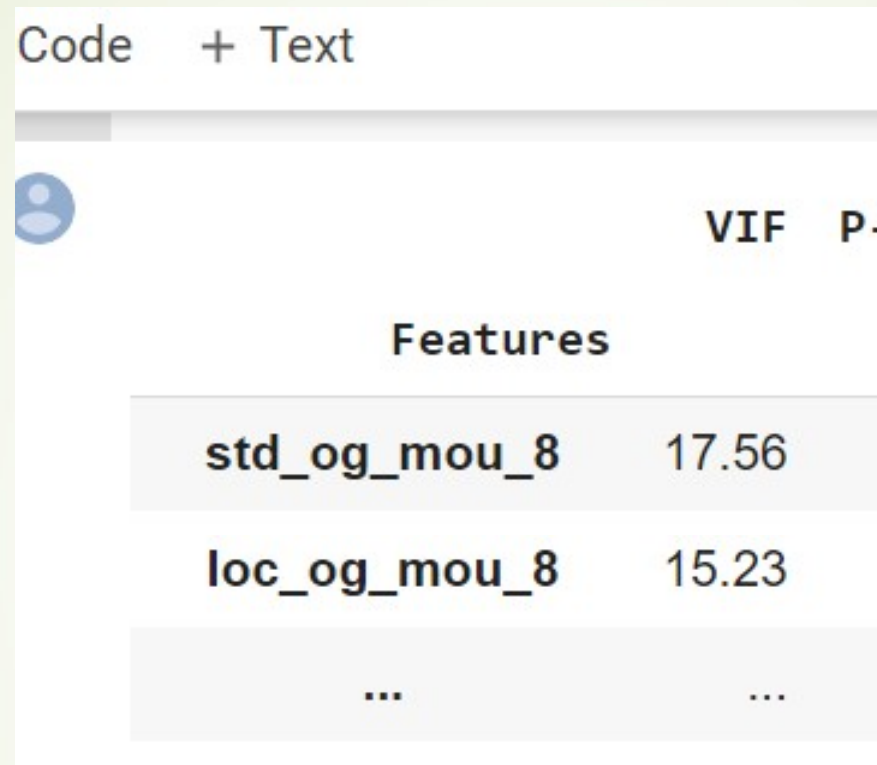
Features

std_og_mou_8	17.56
loc_og_mou_8	15.23
...	...

Model -II

Code	+ Text
	VIF P-
Features	
std_og_mou_8	17.56
loc_og_mou_8	15.23
...	...
delta total ic mou	1.32

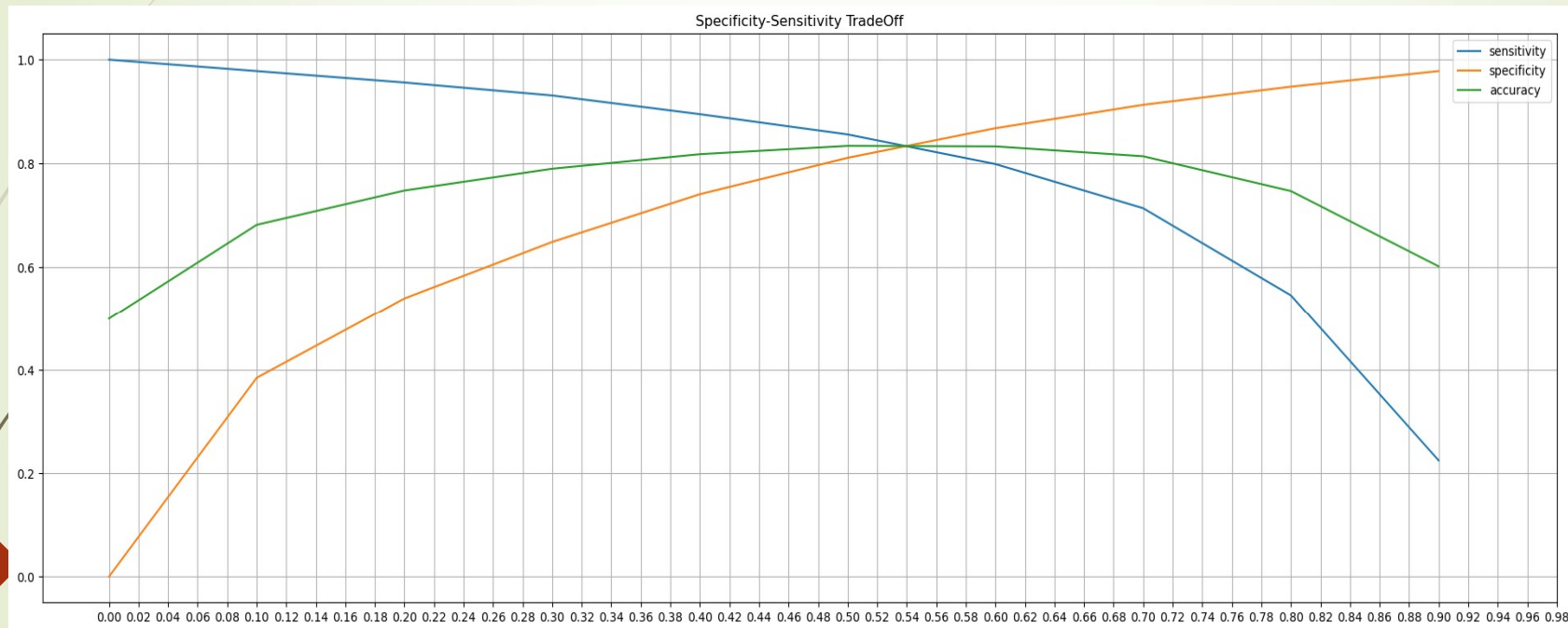
Model III



Code		+ Text	
		VIF	P-
Features			
std_og_mou_8		17.56	
loc_og_mou_8		15.23	
...		...	

All features have low p-values(<0.05) and VIF (<5)
This model could be used as the interpretable logistic regression model

Performance Finding Optimum Probability Cutoff



Strongest indicators of churn

- ❖ Customers who churn show lower average monthly local incoming calls from fixed line in the action period by 1.27 standard deviations, compared to users who don't churn, when all other factors are held constant. This is the strongest indicator of churn.
- ❖ Customers who churn show lower number of recharges done in action period by 1.20 standard deviations, when all other factors are held constant. This is the second strongest indicator of churn.
- ❖ Further customers who churn have done 0.6 standard deviations higher recharge than non-churn customers. This factor when coupled with above factors is a good indicator of churn.
- ❖ Customers who churn are more likely to be users of 'monthly 2g package-0 / monthly 3g package-0' in action period (approximately 0.3 std deviations higher than other packages), when all other factors are held constant.

Recommendations

- ❖ Concentrate on users with 1.27 std deviations lower than average incoming calls from fixed line. They are most likely to churn.
- ❖ Concentrate on users who recharge less number of times (less than 1.2 std deviations compared to avg) in the 8th month. They are second most likely to churn.
- ❖ Models with high sensitivity are the best for predicting churn. Use the PCA + Logistic Regression model to predict churn. It has an ROC score of 0.87, test sensitivity of 100%