Telecom churn

Presented By : Madhumita Roy

Introduction

- 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 more important than customer acquisition.

- For many incumbent operators, retaining highly profitable customers is the number one business goal.

- To reduce customer churn, telecom companies need to predict which highly profitable customers are at risk of churn.

Business problem

- Retaining high-value customers is critical to reducing revenue leakage.
- Predicting churn in high-value customers can help in taking proactive actions.
- Use customer data from a leading telecom firm to predict churn and identify key indicators.

Understanding Churn Models & Definition of Churn

Churn Models in Telecom:

- Postpaid customers: They terminate services directly, sas a result are easy to track.
- Prepaid customers: They may stop using services without notice.
- Importance: Prepaid churn prediction is more critical in India and Southeast Asia, where prepaid is common.

Definition of Churn :

- Revenue-Based Churn: No usage of revenue-generating services (e.g., calls, SMS) for a given period.
- Usage-Based Churn: No usage (incoming/outgoing calls, internet, etc.) for a specific time.
- Churn Definition Used: Usage-based churn, focusing on a lack of calls or data usage.

Focus on High-Value Churn

- High-value customers are those in the top 30% of recharge amounts in the first two months.
- Reduce churn in high-value customers to prevent significant revenue loss.
- 80% of revenue comes from 20% of high-value customers.

Understanding the Dataset

- Customer-level data obtained was from June to September (4 months).
- We need to predict churn in September using data from June, July, and August.

Phases:

Good Phase: First two months (June & July).

Action Phase: Third month (August).

Churn Phase: Fourth month (September).

Data Preparation

Filtering High-Value Customer:

- Customers whose average recharge in June & July is above the 70th percentile were selected.
- This reduces the dataset to about 30,000 rows.

Tagging Churners:

- Customers who didn’t make any calls or use data in the churn phase (September) are tagged as churners.

Data Cleanup:

- All features related to the churn phase are removed (September) for prediction purposes.

EDA (Exploratory Data Analysis)

- Data Visualization using seaborn and matplotlib.
- Exploratory Data Analysis is an approach to analyse dataset and to summarize their main characteristics, often with visual methods.
- EDA is for seeing what the data can tell us beyond formal modelling or hypothesis

Box plot (Finding Outlier)

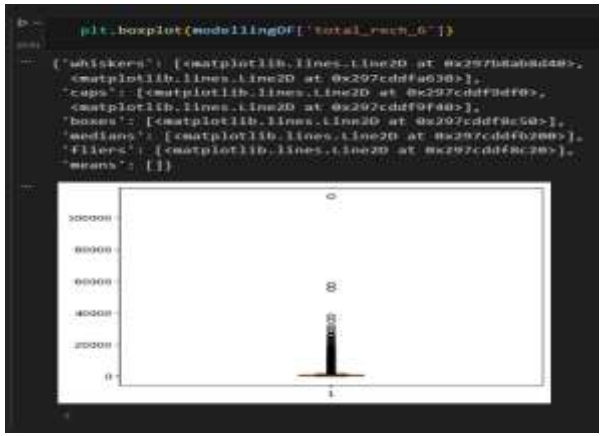


Fig 1 : 1st Box Plot

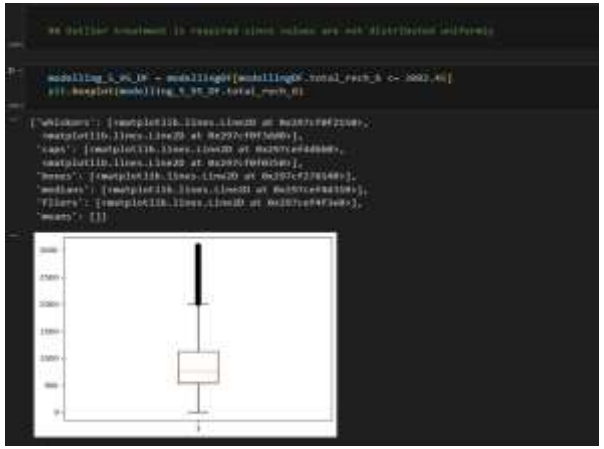
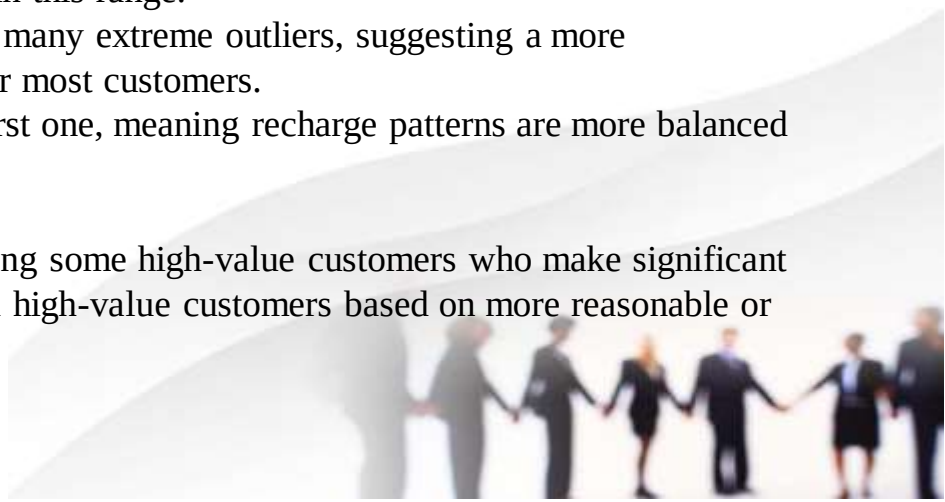


Fig 1 : 2nd Box Plot

1st Box Plot: (With Outlier)
There are very large recharge amounts that extend beyond 40,000 units, and one extreme outlier goes above 100,000. This indicates that a few customers have made significantly higher recharges than the majority.
The bulk of the data is concentrated near the bottom of the box plot, suggesting that most customers have made much smaller recharges.
This distribution appears highly right-skewed, meaning a few high-recharge values are pulling the average up, but most of the customers are recharging smaller amounts.

2nd Box Plot: (After outlier treatment)
The orange line inside the box represents the median, which is roughly between 1000 to 1500 units. This indicates that 50% of customers recharge an amount around this value.
The data shows a more typical distribution for recharges, with the whiskers extending up to around 3000 units, indicating that most recharges fall within this range.
Unlike the first box plot, this one does not show as many extreme outliers, suggesting a more concentrated and representative recharge pattern for most customers.
This plot looks more symmetric compared to the first one, meaning recharge patterns are more balanced in this subset of data.

The first plot might represent all customers, including some high-value customers who make significant recharges, while the second could represent filtered high-value customers based on more reasonable or realistic recharge amounts.



Heat Map (Correlation matrix)

Interpretation of the matrix :

1. No Significant Correlation: Since most correlation values are very close to zero, this suggests that there is no significant linear relationship between the variables in this dataset. The variables are largely independent of each other.

2. Potential Data Issues: Given the extremely small values (e.g., 10^-16), indicates numerical instability or data-related issues, such as scaling problems, near-zero variance in the variables, or other anomalies in the data.

3. No Multicollinearity: A lack of strong correlations between variables suggests there is no multicollinearity.

In summary, this matrix indicates that the variables in this dataset don't have meaningful linear relationships, and the data might need to be examined for potential issues like scaling or variance problems.

Modelling Approach

Churn in high-value customers needs to be predicted.

Identifying important predictors of churn to inform business decisions is an important step.

Modeling Techniques:

Logistic Regression was used to handle the class imbalance.

Multicollinearity needs to be handled to identify significant features.

Class Imbalance

Challenge: The churn rate is low (5-10%).

Model Evaluation: Metrics such as accuracy, precision, recall, and F1-score will be used.

Key Predictors of Churn :

Identifying Important Variables:

Logistic regression reveals which variables are most strongly associated with churn.

Example of important features: call volumes, internet usage patterns, recharge amounts.

Model Performance

```
# Let's check the overall accuracy.
round(metrics.accuracy_score(y_pred_final.Churn, y_pred_final.Predicted), 4) * 100
```

```
91.86999999999999
```

```
TP = confusion[1,1] # true positive
TN = confusion[0,0] # true negatives
FP = confusion[0,1] # false positives
FN = confusion[1,0] # false negatives
```

```
# Let's see the sensitivity of our logistic regression model
round(TP / float(TP+FN), 4) * 100
```

```
np.float64(6.61)
```

```
# Let us calculate specificity
round(TN / float(TN+FP), 4) * 100
```

```
np.float64(99.57000000000001)
```

```
# Calculate false postive rate - predicting churn when customer does not have churned
round(FP/ float(TN+FP), 4) * 100
```

```
np.float64(0.43)
```

```
# positive predictive value
round(TP / float(TP+FP), 4)* 100
```

```
np.float64(58.08)
```

```
# Negative predictive value
round(TN / float(TN+ FN), 4) * 100
```

```
np.float64(92.19000000000001)
```

```
draw_roc(y_pred_final.Churn, y_pred_final.Predicted)
```

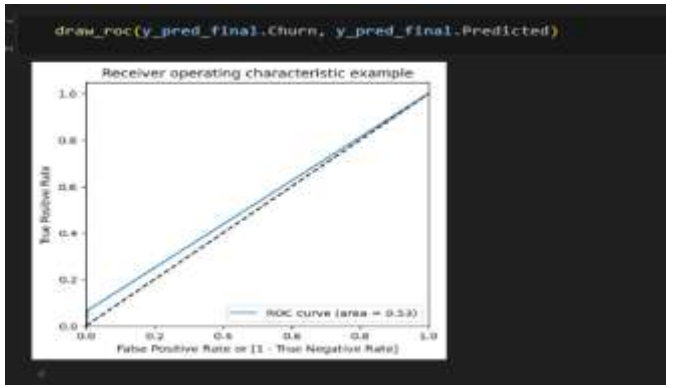


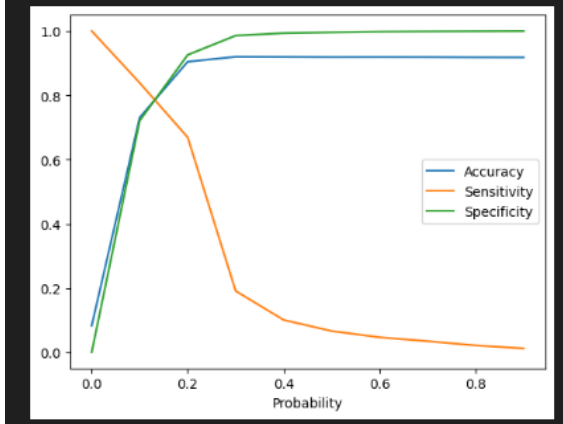
Fig3 : ROC Curve

- The ROC curve has an AUC (Area Under the Curve) value of 0.53, it indicates that the model's ability to distinguish between the two classes (e.g., "churn" vs. "not churn") is only slightly better than random guessing.
- Further hyperparameter tuning is done based on the current ROC curve.
- Threshold Selection : The ROC curve can help in selecting an appropriate threshold for decision-making. By observing different thresholds, we can choose one that offers an optimal balance between sensitivity and specificity for the particular business use case.

Threshold value

	Probability	Accuracy	Sensitivity	Specificity	Precision	Recall
0.0	0.0	0.082857	1.000000	0.000000	0.082857	1.000000
0.1	0.1	0.730571	0.838506	0.720820	0.213429	0.838506
0.2	0.2	0.904238	0.668966	0.925493	0.447865	0.668966
0.3	0.3	0.919667	0.190805	0.985514	0.543372	0.190805
0.4	0.4	0.919238	0.100575	0.993198	0.571895	0.100575
0.5	0.5	0.918667	0.066092	0.995691	0.580808	0.066092
0.6	0.6	0.918857	0.046552	0.997664	0.642857	0.046552
0.7	0.7	0.918619	0.034483	0.998494	0.674157	0.034483
0.8	0.8	0.918048	0.021264	0.999065	0.672727	0.021264
0.9	0.9	0.917810	0.012069	0.999637	0.750000	0.012069

```
# Let's plot accuracy sensitivity and specificity for various probabilities.
cutoffmatrix_df.plot.line(x='Probability', y=['Accuracy','Sensitivity','Specificity'])
plt.show()
```




```
## 0.10 is optimal value
y_pred_final['Predicted'] = y_pred_final.Churn_Prob.map( lambda x: 1 if x > 0.1 else 0)
y_pred_final.head()
```

	I	Churn	Churn_Prob	Predicted	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0	65002	1	0.251386	1	1	1	1	0	0	0	0	0	0	0
1	67088	0	0.016534	0	1	0	0	0	0	0	0	0	0	0
2	36410	0	0.053804	0	1	0	0	0	0	0	0	0	0	0
3	90870	0	0.195898	1	1	1	0	0	0	0	0	0	0	0
4	50581	0	0.055124	0	1	0	0	0	0	0	0	0	0	0

```
# Confusion matrix
confusion = metrics.confusion_matrix(y_pred_final.Churn, y_pred_final.Predicted)
confusion
TP = confusion[1,1] # true positive
TN = confusion[0,0] # true negatives
FP = confusion[0,1] # false positives
FN = confusion[1,0] # false negatives
```

```
# Let's see the sensitivity of our logistic regression model
round(TP / float(TP+FN), 4) * 100
```

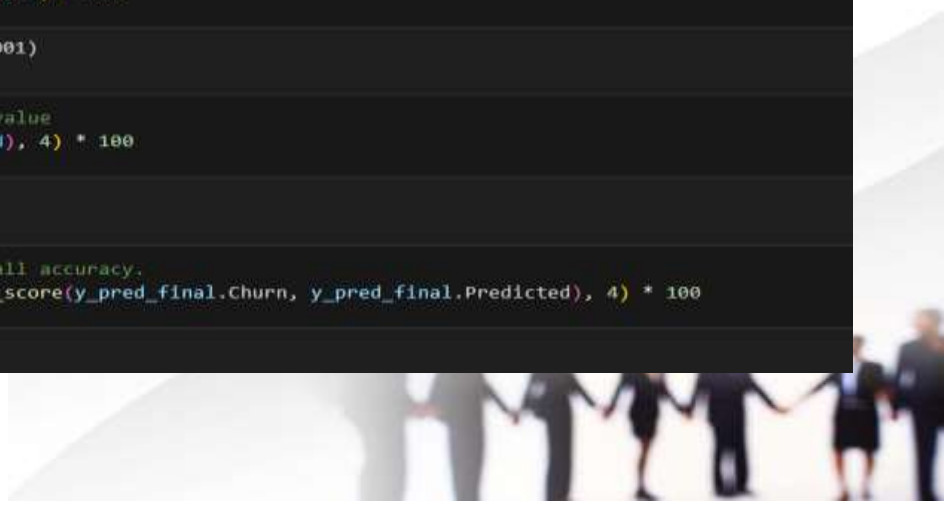
```
np.float64(83.85000000000001)
```

```
# Negative predictive value
round(TN / float(TN+ FN), 4) * 100
```

```
np.float64(98.02)
```

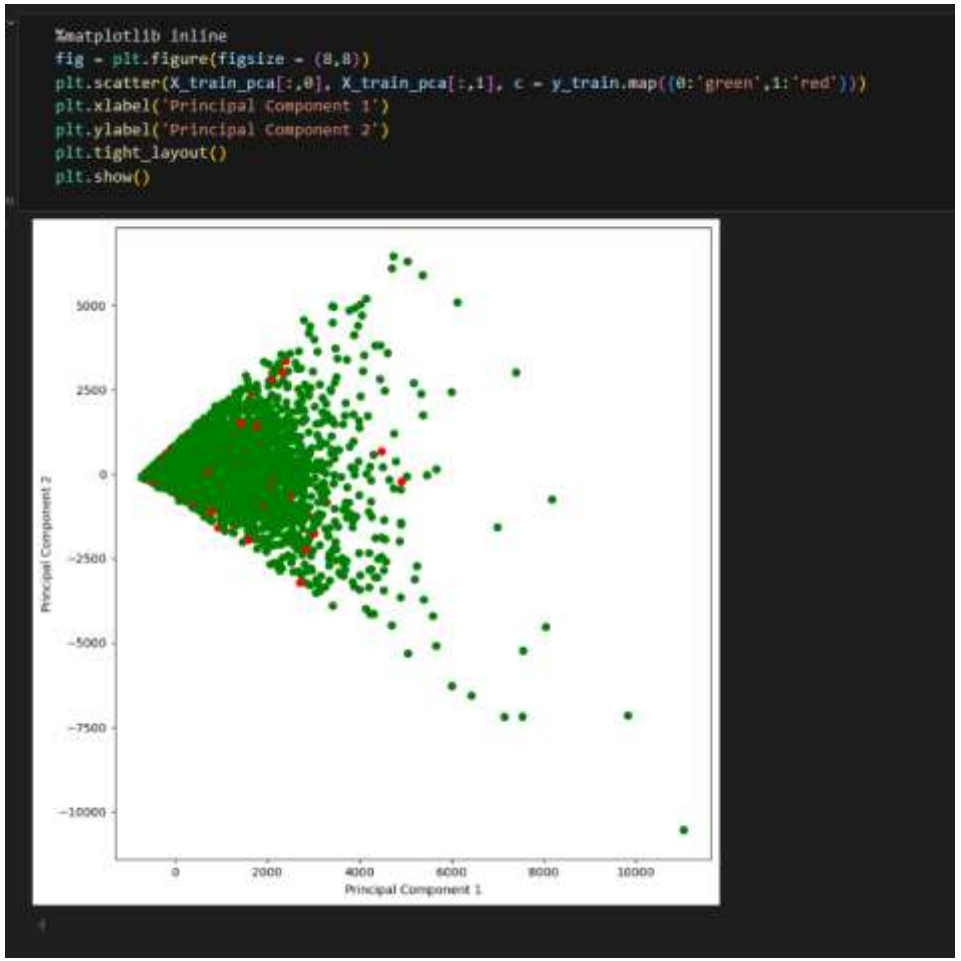
```
##Let's check the overall accuracy.
round(metrics.accuracy_score(y_pred_final.Churn, y_pred_final.Predicted), 4) * 100
```

```
73.06
```

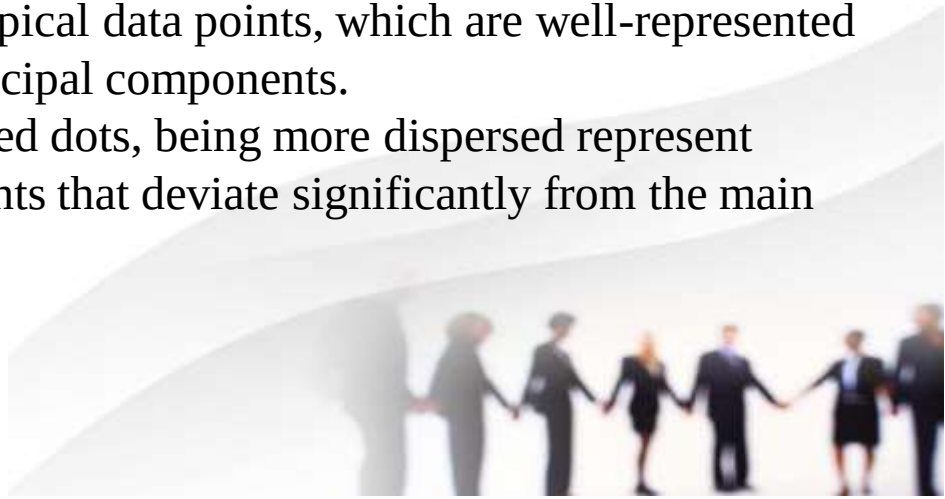


- The optimum threshold value was seen to be 0.1
- After setting the threshold value to 0.1. The model acuracy is reduced to 73 % , from this we can say that the accuracy score obtained earlier was misleading
- Initially, the model might be using a default threshold (usually 0.5) to classify predictions. Given the class imbalance, this threshold favors the majority class, resulting in high accuracy but poor detection of the minority class (churners).
- By selecting an optimal threshold using the ROC curve, we’re adjusting the model to be better at detecting the minority class (churners), which improves the True Positive Rate (sensitivity).
- However, this can reduce overall accuracy because the model now misclassifies more majority class instances (non-churners).
- So, accuracy drops to 73%, but the model's ability to correctly identify churners has improved.
- The goal of predictive modeling, especially in tasks like customer churn prediction, is often to balance sensitivity and specificity, not just maximize accuracy.
- The lower accuracy (73%) after threshold adjustment likely means the model is making better, more meaningful predictions, particularly for the class you're most interested in (churners), even though it sacrifices some accuracy on the majority class.

Scatter Plot (PCA)



- The majority of points are concentrated within a triangular or cone-like shape, which indicates that most of the variance in the data lies within a particular range i.e.,0 to 10,000 in x axis and -10,000 to 5,000 in y axis along the two principal components.
- The presence of green and red dots suggests two different groups or categories. The red points could be representing anomalies, outliers, or a special subset within the data, while the green points represent the general data distribution.
- The densely packed green dots likely represent a core cluster of normal or typical data points, which are well-represented by the two principal components.
- The scattered red dots, being more dispersed represent outliers, or points that deviate significantly from the main data cluster.



Assesing with statsmodel

```
X_train_sm = sm.add_constant(X_train[col])
logm2 = sm.GLM(y_train,X_train_sm, family = sm.families.Binomial())
res = logm2.fit()
res.summary()
```

Generalized Linear Model Regression Results

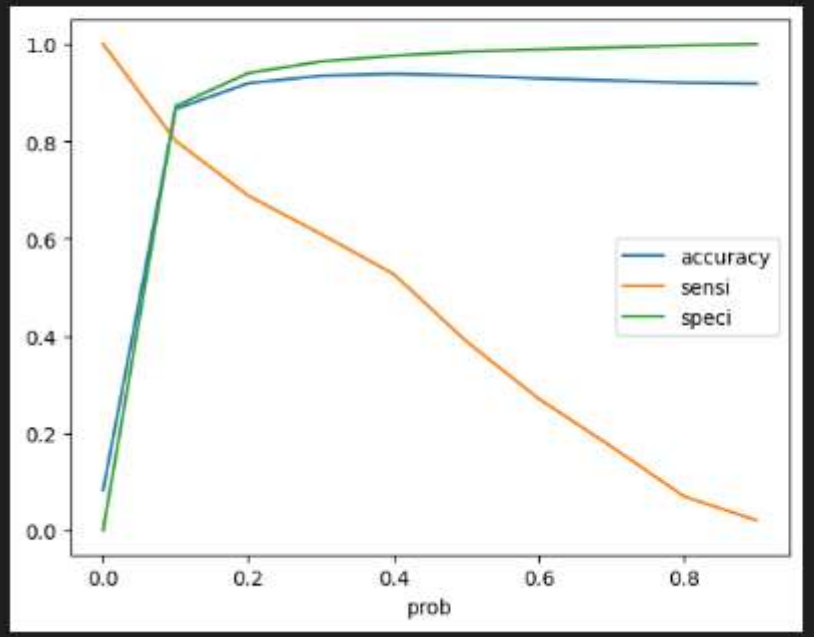
Dep. Variable:	Churn	No. Observations:	21000
Model:	GLM	Df Residuals:	20979
Model Family:	Binomial	Df Model:	20
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-3713.5
Date:	Tue, 17 Sep 2024	Deviance:	7427.0
Time:	18:47:43	Pearson chi2:	2.18e+05
No. Iterations:	9	Pseudo R-squ. (CS):	0.1956
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-3.9198	0.065	-60.134	0.000	-4.048	-3.792
arpu_6	0.3036	0.038	8.033	0.000	0.230	0.378
arpu_7	0.3629	0.055	6.569	0.000	0.255	0.471
arpu_8	-0.7515	0.087	-8.665	0.000	-0.922	-0.582
total_ic_mou_7	0.5089	0.063	8.097	0.000	0.386	0.632
total_ic_mou_8	-1.6105	0.122	-13.230	0.000	-1.849	-1.372
last_day_rch_amt_8	-0.3127	0.055	-5.734	0.000	-0.420	-0.206
arpu_2g_7	0.2013	0.069	2.927	0.003	0.067	0.336
monthly_2g_7	-0.1888	0.048	-3.935	0.000	-0.283	-0.095
monthly_2g_8	-0.5051	0.075	-6.706	0.000	-0.653	-0.357
sachet_2g_8	-0.8466	0.084	-10.036	0.000	-1.012	-0.681
monthly_3g_7	-0.2354	0.077	-3.075	0.002	-0.385	-0.085
monthly_3g_8	-0.4044	0.091	-4.453	0.000	-0.582	-0.226
sachet_3g_8	-0.2811	0.088	-3.201	0.001	-0.453	-0.109
aon	-0.2693	0.040	-6.791	0.000	-0.347	-0.192
total_rech_8	0.3099	0.170	1.822	0.068	-0.023	0.643
roam_any_7	-0.2001	0.034	-5.968	0.000	-0.266	-0.134
roam_any_8	0.5203	0.031	16.619	0.000	0.459	0.582
std_any_8	-0.5971	0.030	-19.831	0.000	-0.656	-0.538
spl_any_8	-0.2460	0.033	-7.538	0.000	-0.310	-0.182
data_used_8	-0.3101	0.097	-3.201	0.001	-0.500	-0.120

- Coefficients reflect the log odds of a customer churning based on each feature. Positive values increase the likelihood of churn, while negative values reduce it.

- arpu_6 (Avg Revenue per User in month 6): Coefficient = 0.3036 (significant with $P < 0.05$)
  Higher revenue in month 6 slightly increases churn probability, but the effect size is modest.
- total_ic_mou_7 (Total Incoming Minutes of Use in month 7): Coefficient = 0.0589 (not significant with $P > 0.05$)
  This variable is not statistically significant, so we can't make strong conclusions about its effect on churn.
- last_day_rch_amt_8 (Last Recharge Amount in month 8): Coefficient = -0.3127 (significant)
  A higher last recharge amount decreases the likelihood of churn, indicating that customers who recently spent more are less likely to churn.
- monthly_2g_7 (2G usage in month 7): Coefficient = 0.1888 (significant)
  Higher 2G usage in month 7 increases the probability of churn, suggesting that customers using more 2G data are more likely to churn, potentially due to dissatisfaction.
- monthly_3g_8 (3G usage in month 8): Coefficient = -0.2354 (significant)
  Increased 3G usage in month 8 decreases the likelihood of churn, possibly indicating satisfied customers with better connectivity.
- roam_any_8 (Roaming activity in month 8): Coefficient = 0.2513 (significant)
  Positive coefficient suggests that customers using roaming in month 8 are more likely to churn.
- data_used_8 (Total data used in month 8): Coefficient = -0.3101 (significant)
  Higher data usage decreases churn, indicating that active data users are less likely to leave.

```
cutoff_df.plot.line(x='prob', y=['accuracy','sensi','speci'])
plt.show()
```



Looking at the graph we can see that both specifitivity and sensitivity are high when the cut off is 0.1, hence we will change it accordingly

Final Model

```
confusion2 = metrics.confusion_matrix(y_train_pred_final.Converted, y_train_pred_final.final_predicted )
confusion2
```

✓ 0.0s

```
array([[16800,  2460],
       [  347,  1393]])
```

```
TP = confusion2[1,1] # true positive
TN = confusion2[0,0] # true negatives
FP = confusion2[0,1] # false positives
FN = confusion2[1,0] # false negatives
```

✓ 0.0s

Precision of the model

```
TP/float(TP+FP)
```

✓ 0.0s

```
np.float64(0.361536465092136)
```

Specificity

```
TN / float(TN+FP)
```

✓ 0.0s

```
np.float64(0.8722741433021807)
```

Sensitivity

```
TP / float(TP+FN)
```

✓ 0.0s

```
np.float64(0.8005747126436782)
```

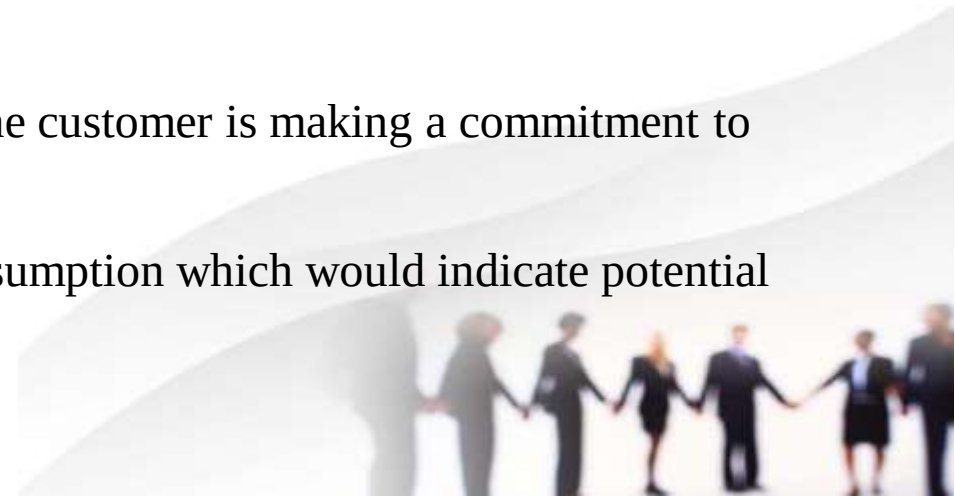
The final model performance are as such after applying the necessary rfe and vif.

Factors responsible & Recommendations

We notice that the following 5 factors affect the churn rate considerably -

- Total Incoming Minutes of usage in the August
- Total Incoming Minutes of usage in the July
- 2G data pack
- Roaming
- Sachet 2g Also these metrics are inversely proportion to churn which means that we need to come up with campaigns that would keep people engaged either via calls (incoming) or on internet. One interesting thing to note here is that we see that a lot of people are hooked on 2G and hence are not churning. It presents us with a great opportunity that if shift these people from 2g to 3g then we have a greater chance of these people not churning. Hence discounts on 3G pack can be one of the popular maarketing campaigns

- The number of recharges, amount of recharge and the last day of recharge are important indicators of customer usage. If any of them decrease, it could be a sign that the customer is trying to wait the remaining period of validity and then try to switch providers. Recharge packs with discount can be offered during this period to retain customer.
- Internet usage is also an important variable as it indicates if the customer is actively using the mobile for operations such as social networking, mobile banking, bill payments etc., and any reduction/non-usage indicates higher possibility of churn. Data Packs can be offered as a bundle/reduced prices/free for a month to attract customer to be in the network.
- Incoming and Outgoing calls in the last good month 7 and action phase month 8 also need to be taken into consideration
- STD, ISD & Special incoming calls for month 8 need to monitored for usage consistently and see if we are able to identify any drop in the usage pattern
- Depending on the volume based cost for 8th month, we can observe if the customer is making a commitment to stay further on the network through month 9
- Average revenue per user should also be monitored for reduction in consumption which would indicate potential customer churn



Business Implications:

The business implications of these factors impacting churn and customer behavior are crucial for designing targeted retention strategies. Here's how each finding can influence business decisions:

- Incoming Minutes Usage (July & August)

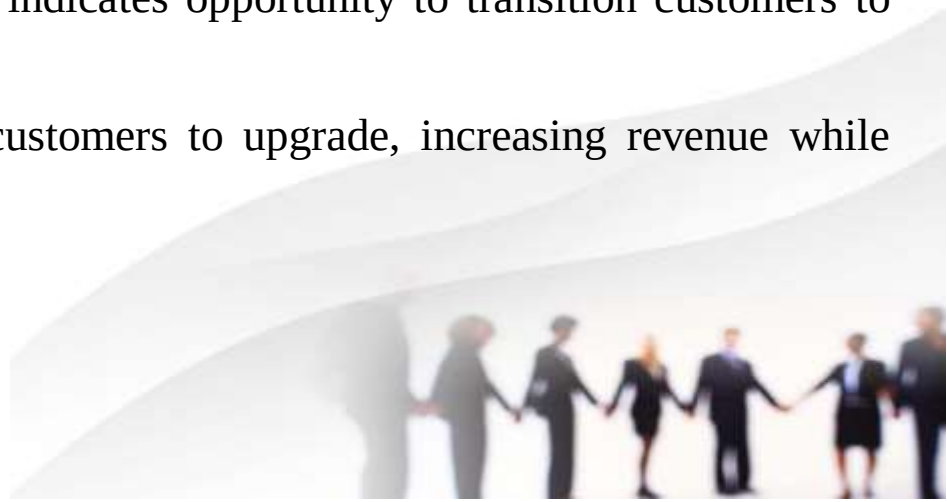
Since incoming minutes are inversely related to churn, maintaining customer engagement through calls is critical.

Campaigns that incentivize more incoming calls (e.g., free incoming during roaming or family call plans) could help prevent churn.

- 2G Data Pack Usage

A large number of customers are still using 2G and are less likely to churn. This indicates opportunity to transition customers to higher-value 3G/4G plans.

Offering discounts on 3G data packs, or bundling 2G-3G plans can encourage customers to upgrade, increasing revenue while reducing churn.



- Roaming Usage

Customers using roaming services are likely to remain engaged.

Promoting affordable roaming packs or offering discounts during national holidays/travel seasons can help keep customers from switching, especially for high-usage travelers.

- Sachet 2G Usage

Sachet 2G packs cater to cost-conscious users who might not be heavy data consumers but are consistent. Micro-plans for higher-speed data (e.g., 3G/4G sachet packs) can attract these customers, offering better service without significantly increasing their costs.

- Recharge Behavior (Number, Amount, and Last Day)

Decreases in recharge frequency, value, or timing can signal that a customer is preparing to leave the network. Proactive offers, such as discounted recharge packs or bonus data for timely recharges, could retain these customers and encourage longer-term commitment.

- Internet Usage & Social Engagement

Decline in internet usage is an early indicator of churn, especially as customers disengage from digital services. Offering free or discounted data packs, especially for popular activities like social media or mobile banking, can encourage sustained usage and loyalty.

- Call Behavior (Incoming & Outgoing in Last Good Month and Action Phase)

Monitoring call activity in key periods like the last "good" month (July) and the action phase (August) can help identify early signs of disengagement. Special calling plans targeting inactive users, like bonus call minutes or loyalty rewards, can re-engage them before they churn.

- STD, ISD, & Special Incoming Calls (August)

Drops in special call usage could indicate an impending churn. Offering discounted STD/ISD rates or promoting exclusive international plans during holidays could help keep these customers loyal.

- Volume-Based Cost in Action Phase (August)

If customers reduce their usage in August, it may indicate that they are testing another network.
Customized offers (like unlimited usage for a fixed price) can encourage customers to stay on the network and avoid switching.

- Average Revenue Per User (ARPU)

Monitoring ARPU is critical as a drop could signal customer disengagement.
Offering personalized retention deals for high-value customers based on their usage patterns can help maintain engagement and prevent churn.

Conclusion

Churn prediction helps retain high-value customers and reduce revenue leakage.

The predictive model can be used to take proactive actions for customer retention.

Identifying key indicators of churn is crucial for developing targeted strategies.