Telecom Churn Casestudy

GROUP MEMBERS :

1. AKSHITA GOEL

Problem Statement

The problem at hand is the high churn rate in the telecom industry, where customers have the flexibility to switch between service providers. With an average annual churn rate of 15-25%, the telecommunications sector faces a significant challenge. Considering the substantial cost disparity between acquiring new customers and retaining existing ones (5-10 times more), customer retention has become a critical focus over customer acquisition. Retaining high-profit customers is particularly crucial for incumbent operators. To tackle this issue, telecom companies require an effective solution to predict customers who are most likely to churn. This prediction capability will enable targeted strategies to reduce churn and retain valuable customers.

Business Objective

The business objective is to analyze customer-level data from a leading telecom firm and develop predictive models that can accurately identify customers at high risk of churn. By leveraging the available data, the aim is to uncover the key indicators or factors that contribute to churn. This analysis will provide valuable insights for the telecom firm to implement targeted strategies and interventions aimed at reducing customer churn. Ultimately, the objective is to improve customer retention rates, enhance profitability, and maintain a competitive edge in the highly dynamic telecommunications industry.

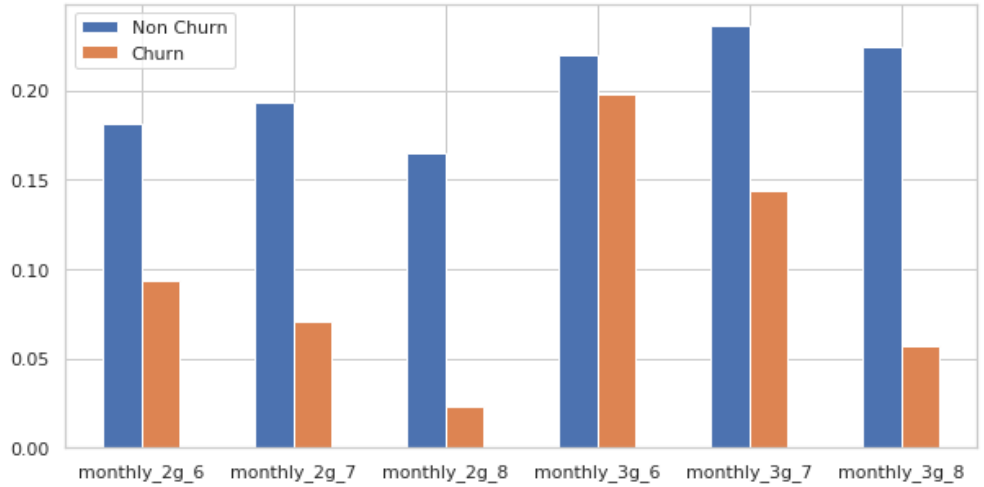
Solution Methodology

- Data cleaning and data manipulation:
    1. Handling and resolving duplicate data.
    2. Removing the columns with single value.
    3. Changing datatype of date columns to Date Time.
    2. Managing NA values and missing values.
    3. Dropping columns with a large number of missing values and no relevance for analysis.
    4. Imputing values as needed.
    5. Addressing outliers in the data.
    6. Drop highly correlated variables.
- Exploratory Data Analysis (EDA):
    - Univariate data analysis:
        - Analyzing value count and variable distribution.
    - Bivariate data analysis:
        - Examining correlation coefficients and patterns between variables.
- Feature Scaling & Dummy Variables and encoding of the data.
- Classification technique:
    - Utilizing logistic regression for model creation and prediction.
- Validation of the model.
- Model presentation.
- Conclusions and recommendations.

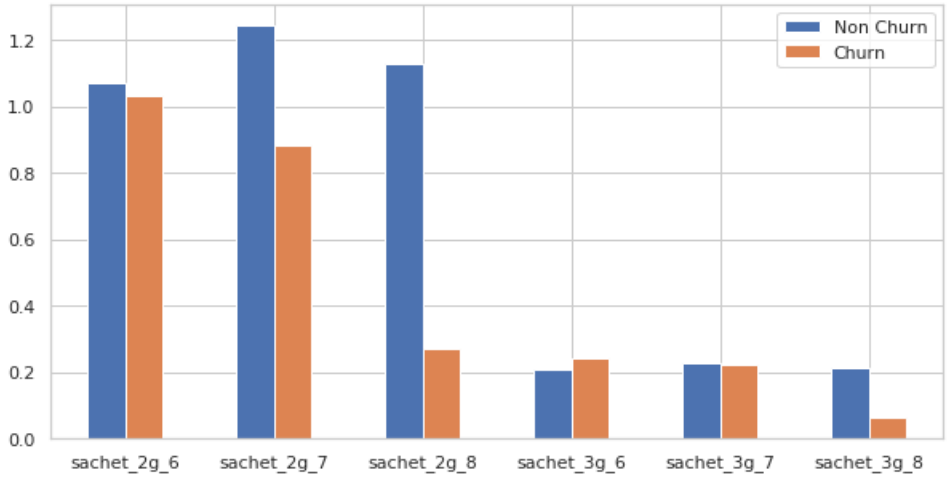
Data Manipulation

- The dataset consists of 9999 rows and 226 columns.
- Features such as 'circle_id', 'loc_og_t2o_mou', 'std_og_t2o_mou', 'loc_ic_t2o_mou', 'last_date_of_month_6', 'last_date_of_month_7', etc have been excluded as they are single-value features.
- The date columns have been changed to datetime datatype.
- Handling missing values of data recharge.
- Checking the correlation between the arpu, recharge amount for month 6,7,8,9 then removing the columns with high correlation.
- If recharge data and amount are null then replace it with 0.
- Columns with more than 75% are being removed.
- Other columns missing values are handled using imputation.
- **Analysis:** We see that the minimum value is 1 while the max is 1555 across months, which indicate the missing values are where no recharges happened for the data, Filling the missing values by 0 , means no recharge.

Drops in 8th Month



	monthly_2g_6	monthly_2g_7	monthly_2g_8	monthly_3g_6	monthly_3g_7	monthly_3g_8
Non Churn	0.18	0.19	0.17	0.22	0.24	0.22
Churn	0.09	0.07	0.02	0.20	0.14	0.06



	sachet_2g_6	sachet_2g_7	sachet_2g_8	sachet_3g_6	sachet_3g_7	sachet_3g_8
Non Churn	1.07	1.25	1.13	0.21	0.23	0.21
Churn	1.03	0.88	0.27	0.24	0.22	0.07

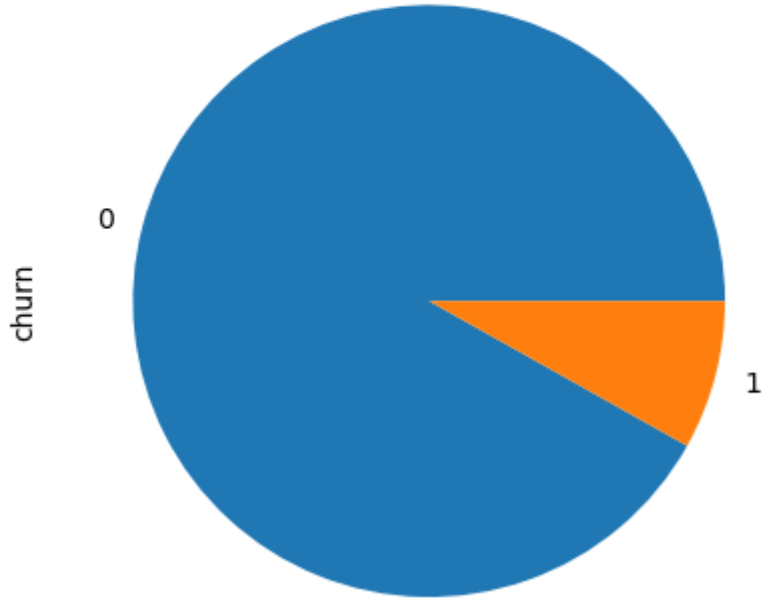
Derive Churn/ Non-Churn %

91 % seems not churned.

```
0    91.863605
1    8.136395
Name: churn, dtype: float64
```

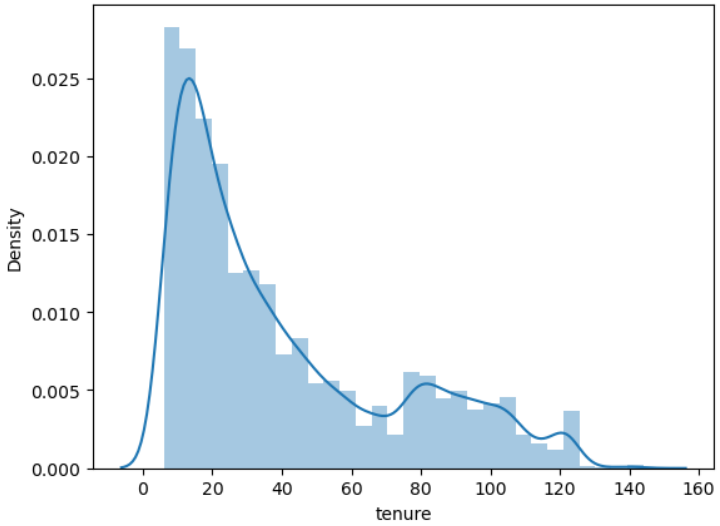
Dropping all churn phase columns.

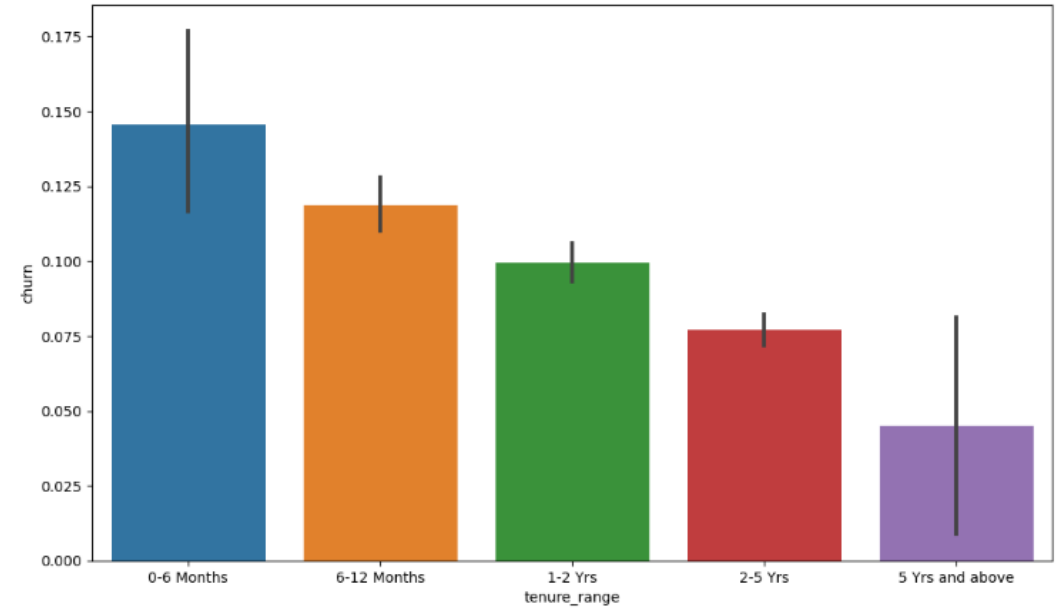
Now we have 30001 rows and 141 columns



Tenure Bucketing

It can be seen that the maximum churn rate happens within 0-6 month, but it gradually decreases as the customer retains in the network.

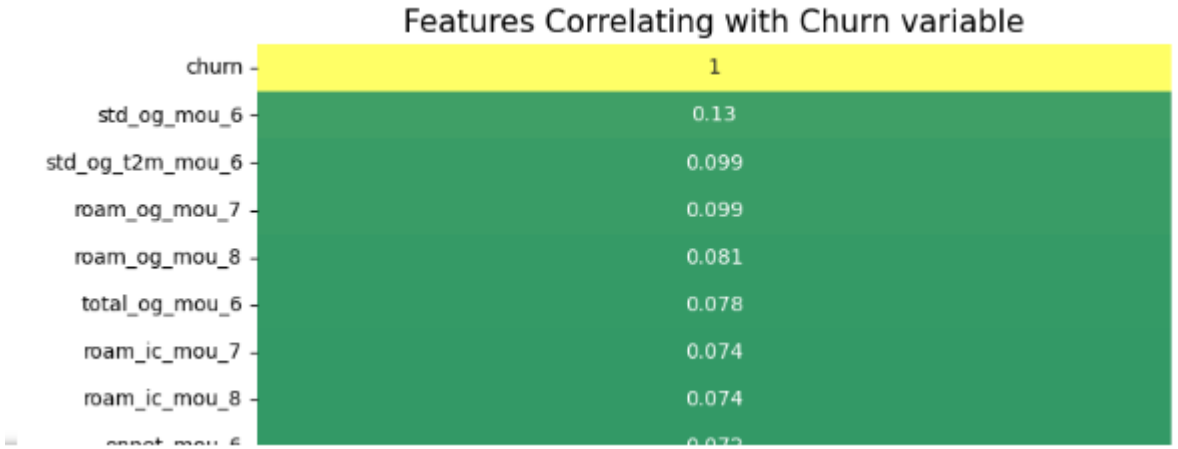


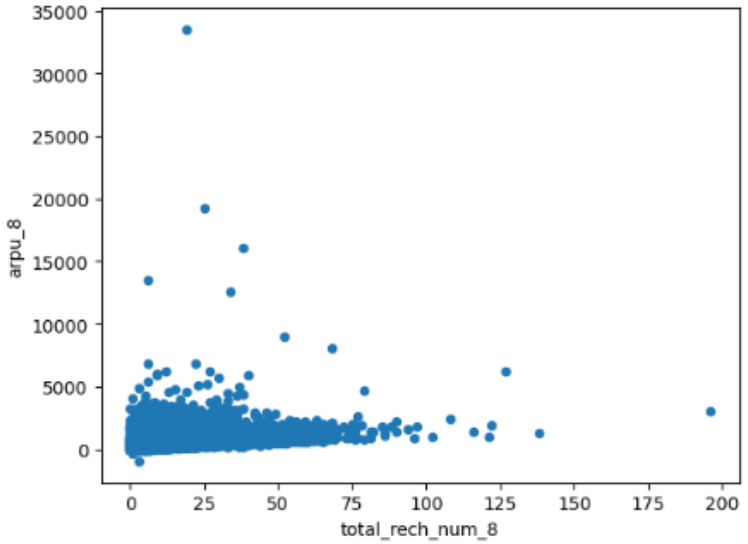


Derive correlation of variables with churn variable.

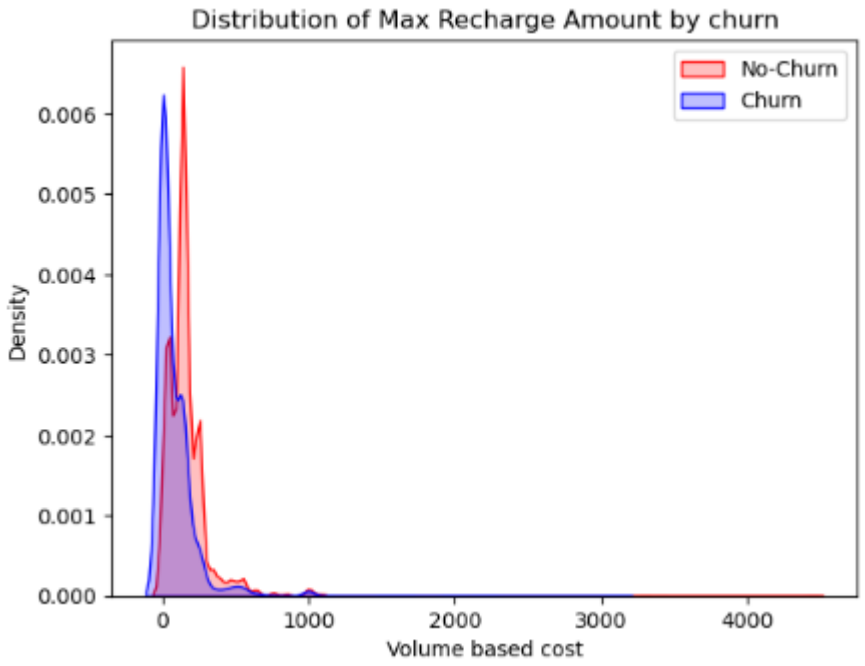
Avg Outgoing Calls & calls on romaning for 6 & 7th months are positively correlated with churn.

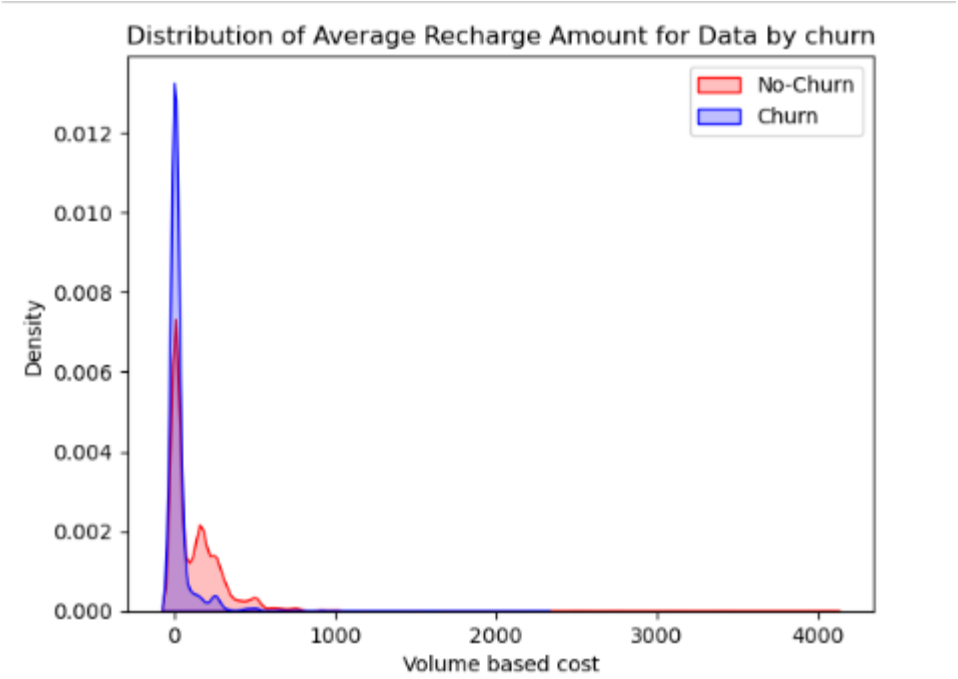
Avg Revenue, No. Of Recharge for 8th month has negative correlation with churn.





Distribution of Average and Max Recharge Amount by Churn



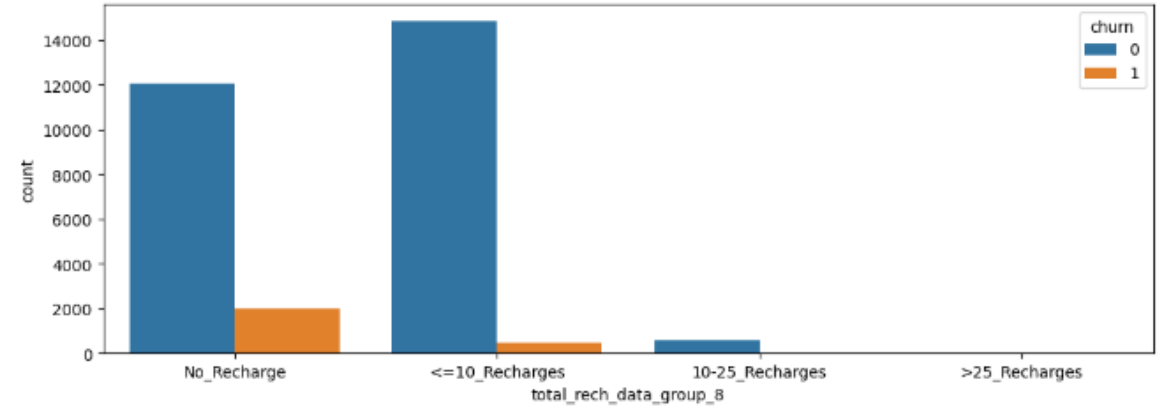


Recharge Rate vs Churn Analysis

As the number of recharge rate increases, the churn rate decreases clearly.

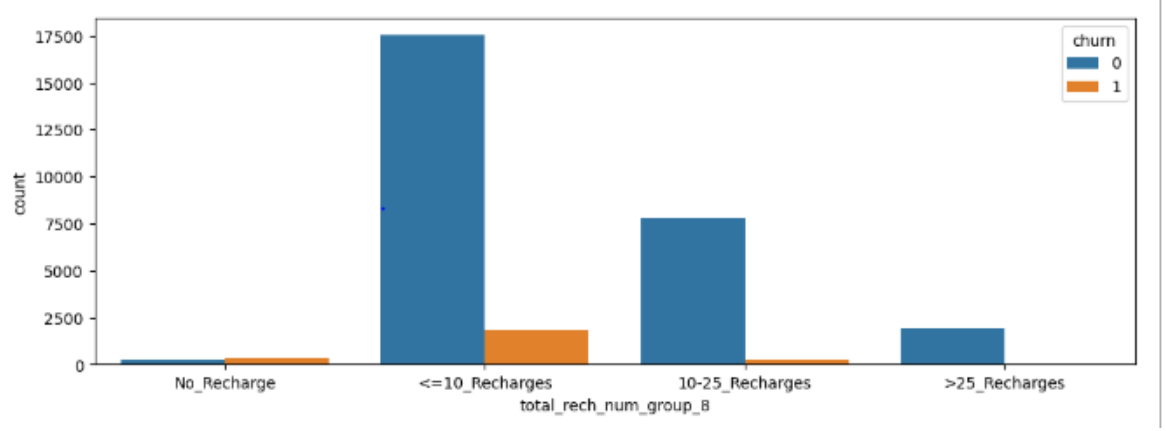
Distribution of total_rech_data_8 variable

```
<=10_Recharges     15307
No_Recharge        14048
10-25_Recharges      608
>25_Recharges         38
Name: total_rech_data_group_8, dtype: int64
```



Distribution of total_rech_num_8 variable

```
<=10_Recharges     19349
10-25_Recharges     8073
>25_Recharges       1996
No_Recharge          583
Name: total_rech_num_group_8, dtype: int64
```

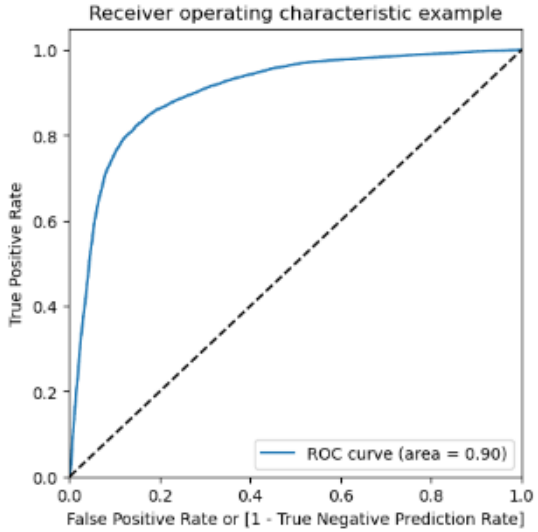


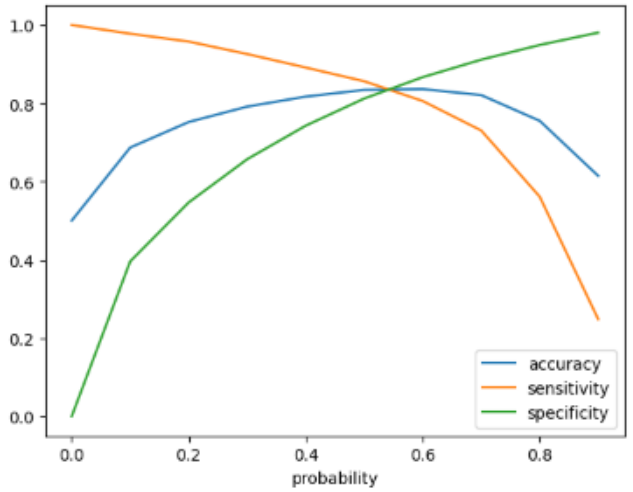
Model Building

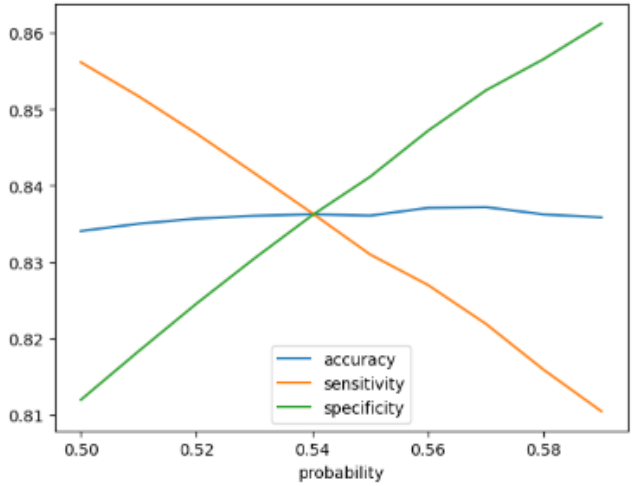
- The data has been divided into training and testing sets using a 70:30 ratio, which is a common practice in logistic regression.
- Scale the dataset using MinMaxScaler()
- SMOTE to manage the imbalance in dataset.
- Recursive Feature Elimination (RFE) has been employed for feature selection. This technique selects the top 20 variables as the output.
- A model has been built by removing variables with a p-value greater than 0.8 and a VIF (Variance Inflation Factor) value greater than 5.
- Predictions have been made on the test dataset.
- The overall accuracy of the Logistic Regression model is 83.4% and precision is 80%

ROC Curve

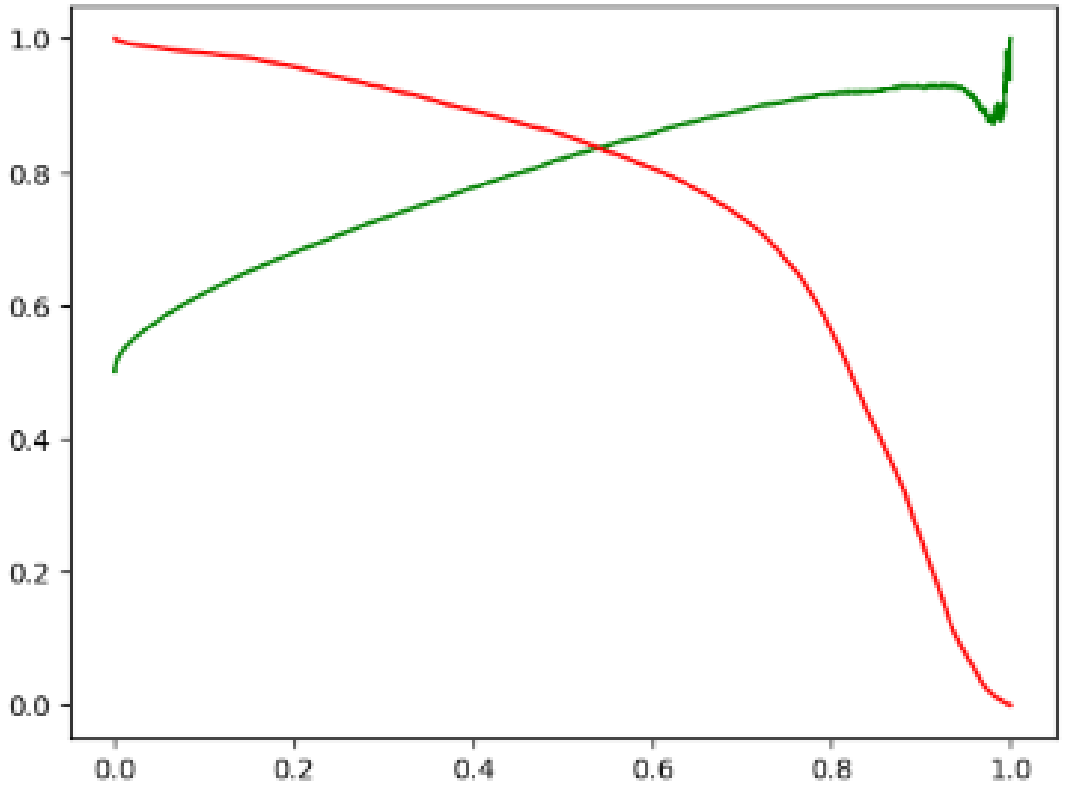
- The optimal cut-off point refers to the probability threshold where there is a balance between sensitivity and specificity.
- Initially we selected the optimum point of classification as 0.5
- From the second graph, we can see the optimum cutoff is slightly higher than 0.5 but lies lower than 0.6. So lets tweak a little more within this range.
- From the third graph we can conclude, the optimal cutoff point in the probability to define the predicted churn variable converges at 0.54.





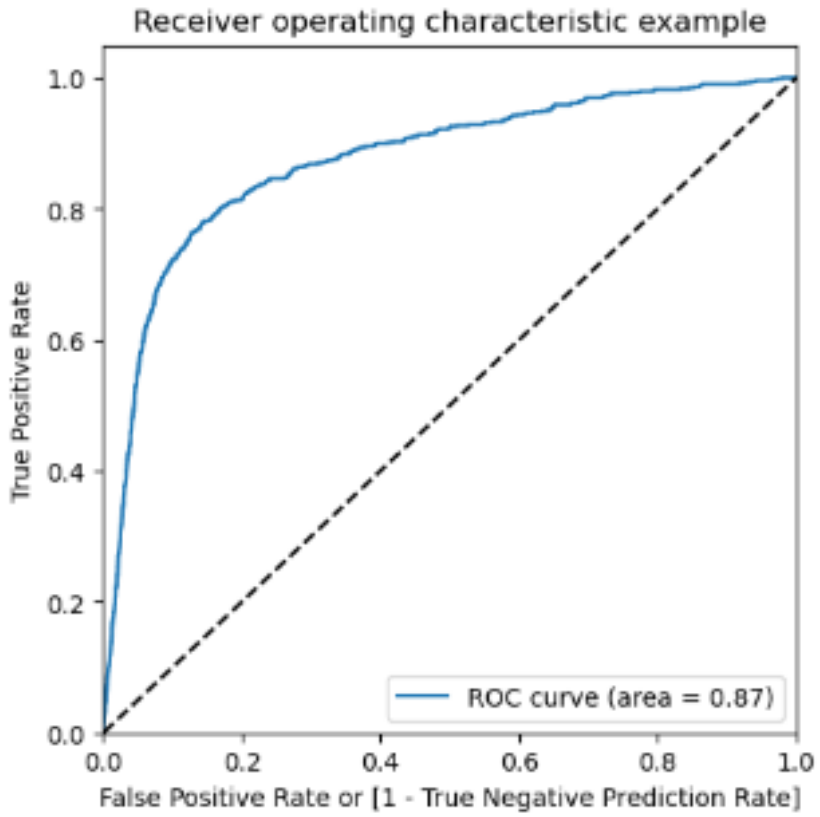


Precision and Recall Trade-off



ROC Curve – Test Dataset

The AUC score for train dataset is 0.90 and the test dataset is 0.87. This model can be considered as a good model.

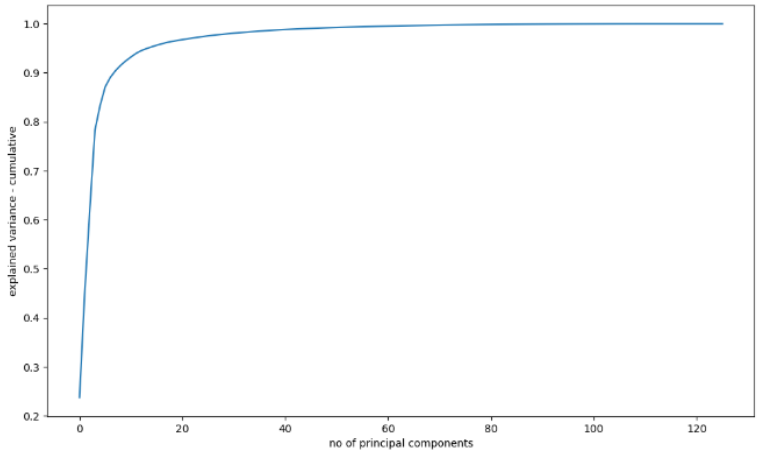


Logistic Regression using PCA

Accuracy of the logistic regression model with PCA: 0.818

90% of the data can be explained with 8 PCA components

Accuracy of the logistic regression model with PCA: 0.754582824130652 1



Model Insights

The dataset analysis revealed that SVM with tuned hyperparameters achieved the highest accuracy of 0.92. Random forest also performed well, with an accuracy of 0.91 (using default hyperparameters, which may indicate overfitting) and 0.90 with tuned hyperparameters. XGBoost yielded a respectable accuracy of 0.86 (using default hyperparameters) and 0.85 with tuned hyperparameters. Based on our findings, SVM and Random forest demonstrated the best accuracy, making them suitable choices for predicting churn data in future datasets or production scenarios.

	Model	Accuracy	Precision	Recall	AUC	F1
0	SVM (Default)-linear	0.83	0.79	0.30	0.81	0.43
1	SVM (Default)-rbf	0.87	0.74	0.36	0.81	0.49
2	SVM(rfb) [Hyper]	0.92	0.48	0.49	0.72	0.49
3	RandomForest (Default)	0.91	0.49	0.45	0.72	0.47
4	RandomForest (Hyper)	0.90	0.66	0.41	0.79	0.51
5	XGBoost (Default)	0.85	0.75	0.33	0.81	0.45
6	XGBoost (Hyper Tuned)	0.84	0.75	0.31	0.80	0.44

Conclusion

1. The churn rate among high-value customers is relatively low. However, it is concerning that no new high-value customers have been on-boarded in the past six months. The company should prioritize addressing this issue.
2. Customers with a tenure of less than four years are more prone to churn. It is recommended that the company focus on this segment by introducing new schemes and offers tailored to their needs.
3. The average revenue per user is a crucial factor in predicting churn. It holds significant importance in determining whether a customer is likely to churn or not.
4. The volume of incoming and outgoing calls while roaming in the eighth month is a strong indicator of churn behavior.
5. Paying attention to this aspect can help in identifying potential churners.
6. The frequency of local outgoing calls made to landline, fixedline, mobile, and call centers serves as a robust indicator of churn behavior. Analyzing this data can provide valuable insights into customer churn tendencies.
7. A strong indicator of churn behavior is the quality of 2G/3G coverage. If the 2G/3G services are inadequate in certain areas, customers are more likely to churn. Enhancing coverage in these areas can help mitigate churn.