TELECOM CHURN - CASE STUDY

Amit Jadhav

TABLE OF CONTENTS

- Overview
- Understanding & Defining Churn
- Understanding Customer Behavior During Churn
- Data preparation
- Exploratory data analysis
- Model Summary conclusion with PCA for Logistic regression, Support
- Vector Machine(SVM) with PCA, Decision Tree, Random Forest
- Logistic regression with No PCA
- Final model summary with top predictors and business recommendation

OVERVIEW

The Competitive Landscape of Telecommunications

- **High Customer Churn:** The telecommunications industry is characterized by a dynamic market with multiple service providers, leading to an average annual customer churn rate of 15-25%.
- **Cost of Customer Acquisition:** Acquiring new customers is significantly more expensive (5-10 times) than retaining existing ones. This economic reality necessitates prioritizing customer retention strategies.

Importance of Customer Retention

- In this highly competitive environment, retaining high-value customers is paramount for telecommunication companies. To achieve this goal, it is crucial to proactively identify customers at risk of churning.

Project Objectives

- This project addresses this challenge by:
- Analyzing customer-level data from a leading telecommunications firm.
- Developing predictive models to identify customers with a high propensity to churn.
- Unveiling the key factors that influence customer churn.

By achieving these objectives, we can empower telecommunication companies to implement targeted retention strategies and mitigate customer churn.

UNDERSTANDING & DEFINING CHURN

The telecommunications industry utilizes two primary payment structures: Postpaid and Prepaid.

- **Postpaid Model:** Customers settle their bills for services used at the end of a billing cycle (monthly/annually). In this scenario, customer churn is readily identifiable when they explicitly notify the operator to terminate services.
- **Prepaid Model:** Customers prepay for a specific service amount beforehand. Identifying churn within the prepaid model presents a challenge. Customers intending to switch providers can simply cease using the service without notification. This inactivity may be due to genuine churn or temporary non-usage, such as traveling abroad.

Churn Measurement Approaches

To address the complexities of churn identification, various metrics can be employed:

- **Revenue-Based Churn:** This metric focuses on the loss of customer-generated revenue within a defined timeframe.
- **Usage-Based Churn:** This metric identifies churn based on a customer's service usage patterns exceeding a predefined period of inactivity.

CUSTOMER BEHAVIOR DURING CHURN

Churn prediction leverages a three-phase customer lifecycle model:

- **Good Phase:** Customers are satisfied with the service and exhibit consistent usage patterns.
- **Action Phase:** Customer experience deteriorates due to factors like competitive offers, service quality issues, or billing problems. This phase is crucial for identifying churn risk, as timely interventions (e.g., matching competitor offers, service improvements) can prevent churn.
- **Churn Phase:** Customers discontinue service. Importantly, churn data is unavailable for prediction purposes as it occurs after the timeframe analyzed. Therefore, data from the churn phase is excluded after labeling churn status (1/0) based on subsequent periods.

Windowing for Churn Prediction

In this four-month window:

- **Months 1 & 2:** Represent the Good Phase.
- **Month 3:** Represents the Action Phase, where churn prediction models are applied.
- **Month 4:** Represents the Churn Phase, excluded from the prediction dataset due to its future nature.

DATA PREPARATION

The following steps are essential for preparing the customer data for churn prediction modeling:

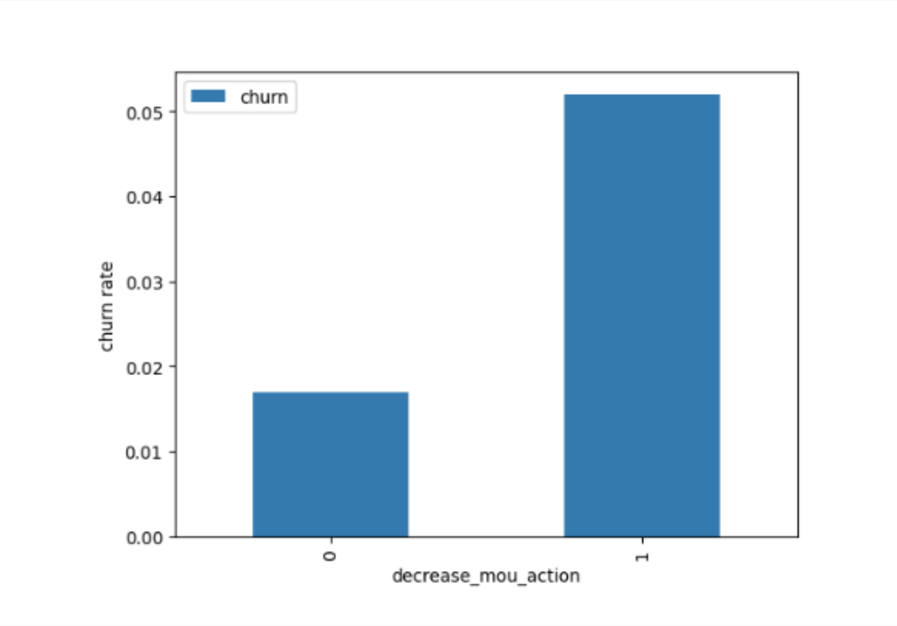
High-Value Customer Segmentation:

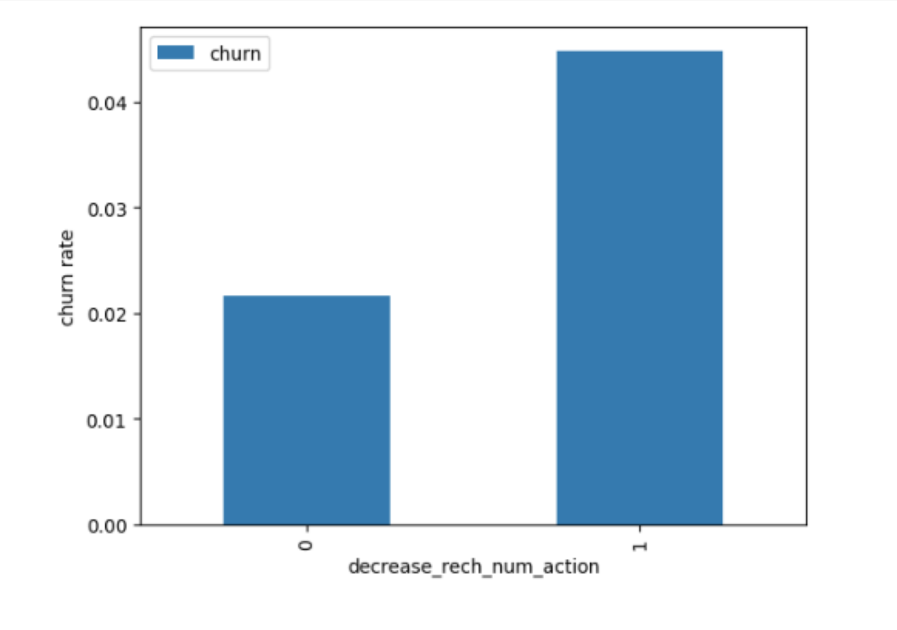
- Identify high-value customers based on their recharge behavior within a defined timeframe (e.g., first two months).
- In this case, high-value customers are those who recharge with an amount greater than or equal to the 70th percentile of the average recharge amount during the good phase (months 1 & 2). This threshold is set at X = 369.5.

Churn Labeling and Feature Selection:

- Churn is identified for customers who exhibit no calling activity (incoming or outgoing) and no mobile internet usage in the designated churn phase (e.g., fourth month). Customer records are labeled accordingly (churn = 1) or assigned a non-churn label (churn = 0) based on this criteria.
- Attributes corresponding to the churn phase (typically denoted by "_9" in their names) are excluded from the analysis since they represent future information not available for prediction at the time.

EXPLORATORY DATA ANALYSIS

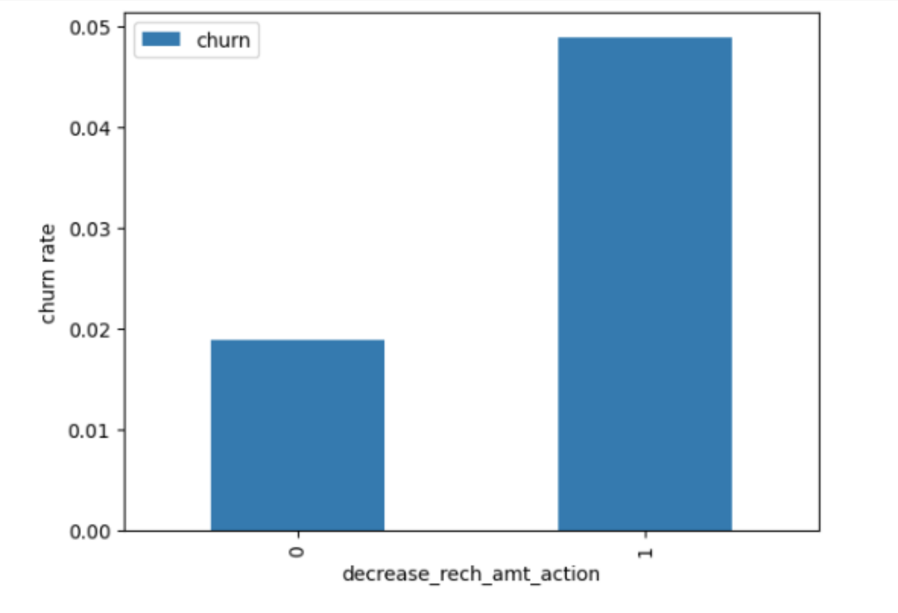


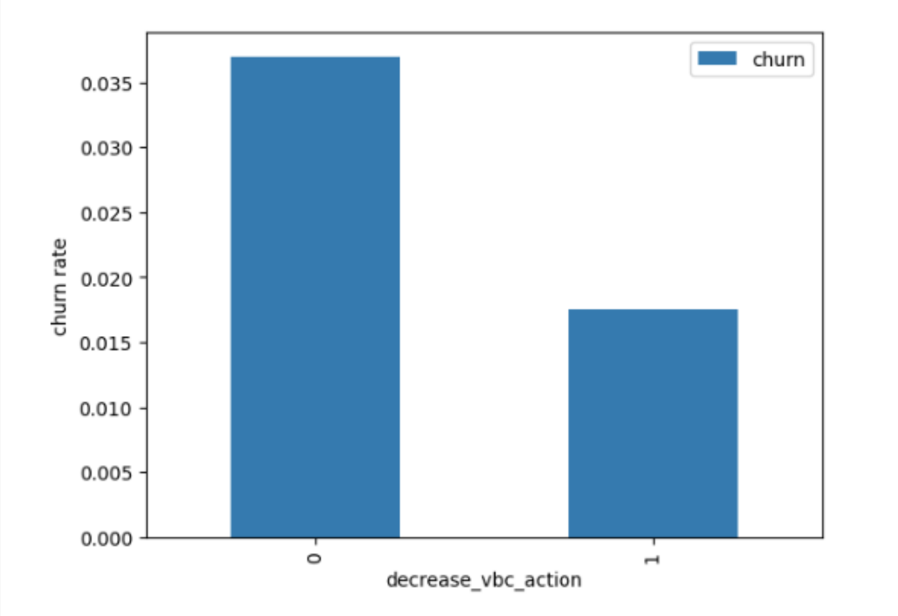


Our analysis indicates a positive correlation between a decrease in customer MoU (minutes of usage) during the action phase, compared to the good phase, and an increased likelihood of churn. In other words, customers who significantly reduce their call activity in the month preceding potential churn are at higher risk of customer churn.

Analysis of customer recharge behavior reveals a correlation between churn risk and a decrease in recharge frequency during the action phase (month 3) compared to the good phase (months 1 & 2). This suggests that customers who recharge less frequently during the action phase may be more likely to churn.

EXPLORATORY DATA ANALYSIS

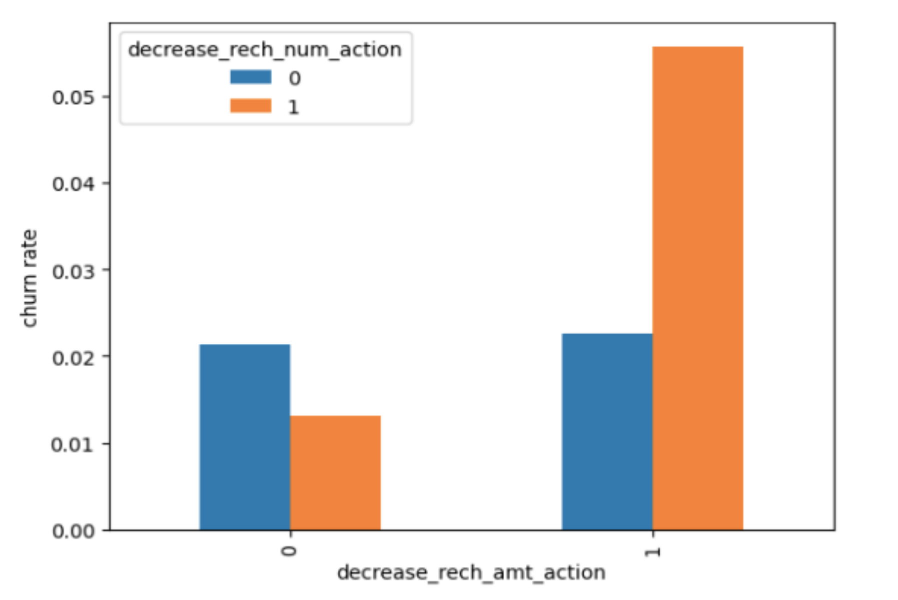


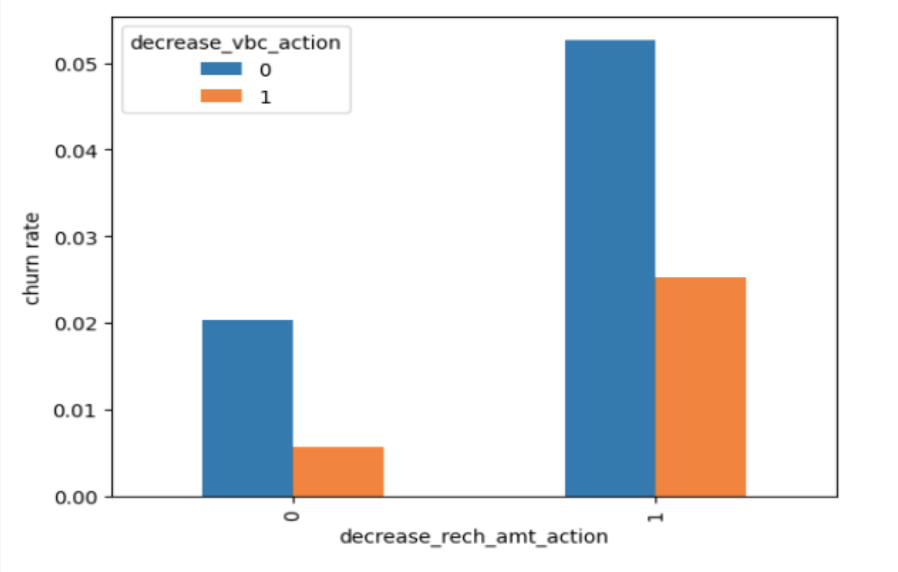


Consistent with previous observations, customer churn exhibits a correlation with recharge behavior. Customers who recharge with a lower amount during the Action phase tend to churn at a higher rate compared to those who maintain or increase their recharge amount relative to the Good phase. This suggests that a decline in recharge amount during this critical period can be a valuable indicator of potential churn.

Our analysis reveals a positive correlation between increased volume-based cost in the action month and customer churn. This suggests that customers with higher usage during the at-risk phase may be less likely to maintain their regular recharge patterns, potentially indicating dissatisfaction or a propensity to switch providers.

EXPLORATORY DATA ANALYSIS





Our analysis reveals a positive correlation between decreased recharge amounts and recharge frequency during the action phase and customer churn. This suggests that customers who reduce both their recharge value and number of recharges compared to the good phase are more likely to churn.

Analysis highlights a potential churn indicator in the action phase: customers who decrease their recharge amount while incurring higher volume-based costs. This combined trend suggests dissatisfaction with current plans and a higher risk of churn.

LOGISTIC REGRESSION

Model summary

Train set

- *Accuracy:- 0.87*
- *Sensitivity:- 0.89*
- *Specificity:- 0.84*

Test set

- *Accuracy:- 0.84*
- *Sensitivity:- 0.82*
- *Specificity:- 0.84*

Overall, the model is performing well in the test set, what it had learnt from the train set.

SUPPORT VECTOR MACHINE(SVM) WITH PCA

Model summary

Train set

- *Accuracy = 0.89*
- *Sensitivity = 0.91*
- *Specificity = 0.87*

Test set

- *Accuracy = 0.86*
- *Sensitivity = 0.78*
- *Specificity = 0.86*

DECISION TREE WITH PCA

Model summary

Train set

- *Accuracy = 0.91*
- *Sensitivity = 0.92*
- *Specificity = 0.89*

Test set

- *Accuracy = 0.86*
- *Sensitivity = 0.72*
- *Specificity = 0.86*

We can see from the model performance that the Sensitivity has been decreased while evaluating the model on the test set. However, the accuracy and specificity is quite good in the test set.

RANDOM FOREST WITH PCA

Model summary

Train set

- *Accuracy = 0.88*
- *Sensitivity = 0.89*
- *Specificity = 0.86*

Test set

- *Accuracy = 0.84*
- *Sensitivity = 0.73*
- *Specificity = 0.85*

We can see from the model performance that the Sensitivity has been decreased while evaluating the model on the test set. However, the accuracy and specificity is quite good in the test set.

CONCLUSION WITH PCA

Our evaluation focused on achieving high sensitivity, prioritizing the model's ability to correctly identify churned customers. Among the evaluated models, Logistic Regression and Support Vector Machines (SVM) demonstrated the strongest performance in this regard. Both models achieved a sensitivity of approximately 90%, indicating a strong ability to capture churn events. Additionally, they delivered good overall accuracy of around 88%. Based on this analysis, Logistic Regression or SVM would be suitable choices for churn prediction in this context.

LOGISTIC REGRESSION WITH NO PCA

Model Analysis

- **Coefficient Analysis:** The model coefficients indicate a mix of positive and negative influences on the churn prediction. Positive coefficients suggest features that increase the likelihood of churn, while negative coefficients indicate features that mitigate churn risk.
- **Feature Significance Evaluation:** The analysis revealed several features with high p-values, indicating a lack of statistical significance for predicting churn within the model. These features will be considered for removal during the coarse-tuning process.

Coarse Tuning (Automated and Manual Approach)

We will employ a two-step approach for feature selection:

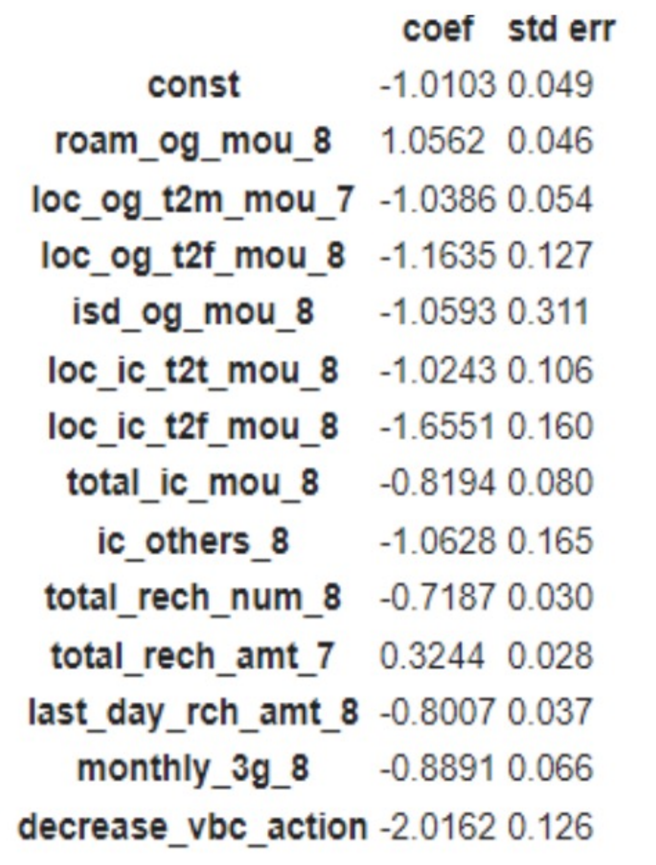
- **Recursive Feature Elimination (RFE):** This automated technique will iteratively remove the least informative features based on a pre-defined metric (e.g., feature importance). This will result in a reduced set of candidate features.
- **Manual Feature Selection:** Following RFE, we will perform a manual evaluation of the remaining features. This will involve examining p-values and Variance Inflation Factors (VIFs) to identify features with low predictive power or redundancy within the model. Features deemed insignificant or highly correlated with others will be removed.

This combined approach aims to optimize the model's performance by retaining the most relevant and informative features for churn prediction.

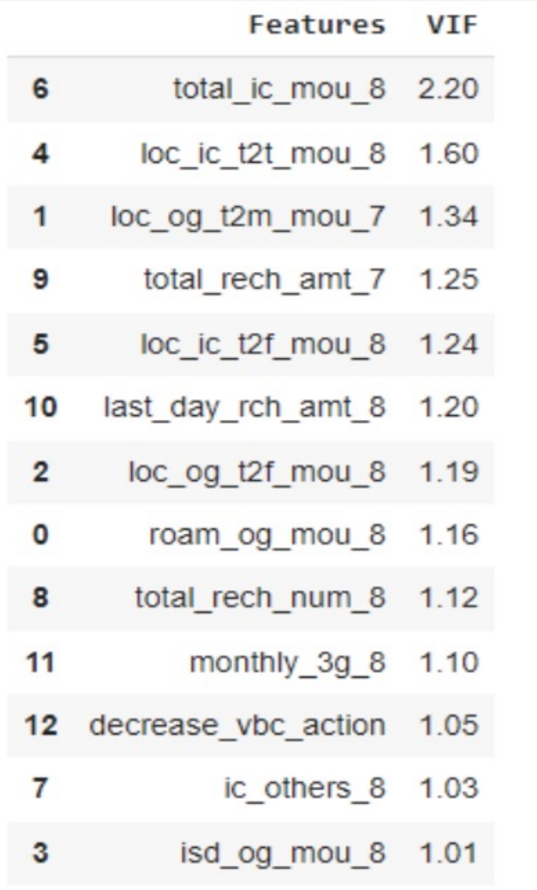
FINAL MODEL III

Based on the analysis of model summaries and Variance Inflation Factors (VIFs), we can confirm that:

- All the independent variables included in Model-3 (log_no_pca_3) are statistically significant for churn prediction (i.e., their p-values are below a predefined threshold).
- The absence of high VIF values indicates a lack of multicollinearity among the variables. This ensures that the features are not redundant and contribute uniquely to the model's predictive power.
- Therefore, considering both statistical significance and the absence of multicollinearity, Model-3 (log_no_pca_3) can be identified as the final model for churn prediction.



	coef	std err
const	-1.0103	0.049
roam_og_mou_8	1.0562	0.046
loc_og_t2m_mou_7	-1.0386	0.054
loc_og_t2f_mou_8	-1.1635	0.127
isd_og_mou_8	-1.0593	0.311
loc_ic_t2t_mou_8	-1.0243	0.106
loc_ic_t2f_mou_8	-1.6551	0.160
total_ic_mou_8	-0.8194	0.080
ic_others_8	-1.0628	0.165
total_rech_num_8	-0.7187	0.030
total_rech_amt_7	0.3244	0.028
last_day_rch_amt_8	-0.8007	0.037
monthly_3g_8	-0.8891	0.066
decrease_vbc_action	-2.0162	0.126



	Features	VIF
6	total_ic_mou_8	2.20
4	loc_ic_t2t_mou_8	1.60
1	loc_og_t2m_mou_7	1.34
9	total_rech_amt_7	1.25
5	loc_ic_t2f_mou_8	1.24
10	last_day_rch_amt_8	1.20
2	loc_og_t2f_mou_8	1.19
0	roam_og_mou_8	1.16
8	total_rech_num_8	1.12
11	monthly_3g_8	1.10
12	decrease_vbc_action	1.05
7	ic_others_8	1.03
3	isd_og_mou_8	1.01

OPTIMAL PROBABILITY CUTOFF POINT

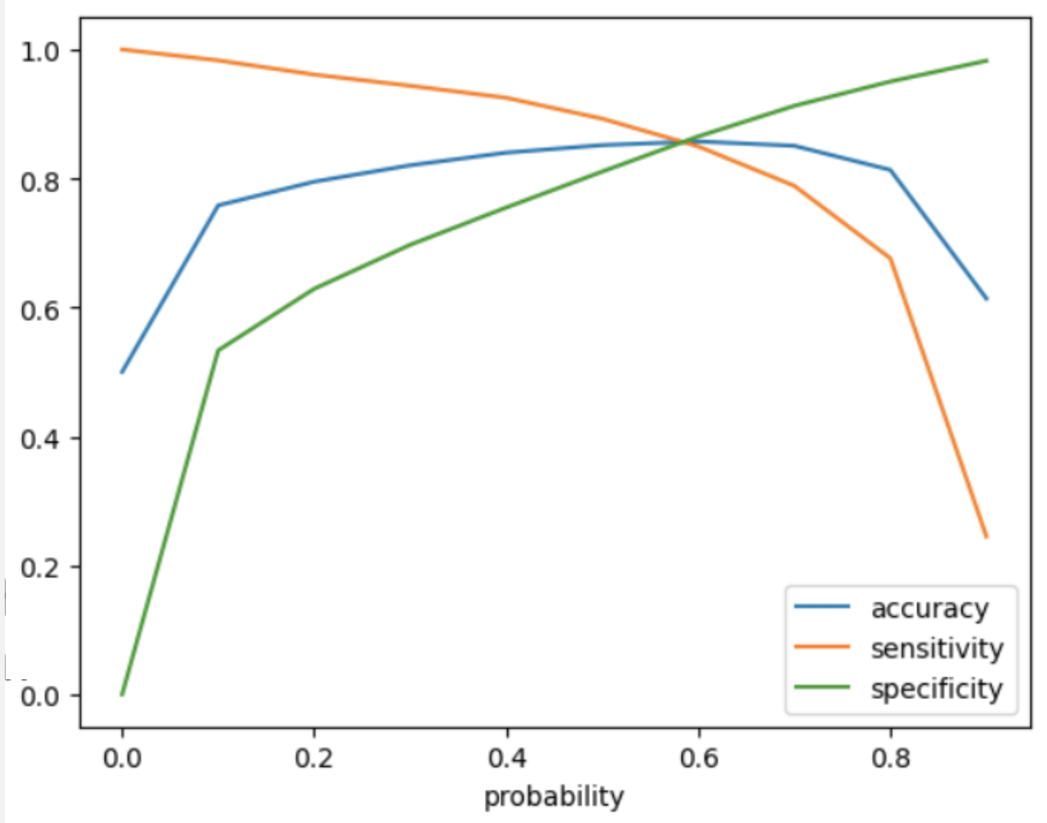
The evaluation examined three key performance metrics: accuracy, sensitivity, and specificity.

- **Accuracy:** Stabilizes around 0.6, indicating the overall ability of the model to correctly classify churn and non-churn customers.
- **Sensitivity:** Exhibits an inverse relationship with probability threshold. In other words, as the probability threshold for churn prediction increases, the sensitivity decreases. Sensitivity reflects the model's capacity to correctly identify true churners.
- **Specificity:** Demonstrates a positive correlation with probability threshold. Specificity measures the model's ability to correctly classify non-churning customers.

The performance curves depict a trade-off between these metrics. While the accuracy reaches a peak at 0.6, where all three curves intersect, our primary objective prioritizes achieving a high sensitivity for churn prediction.

Prioritizing Sensitivity for Churn Prediction

- Given the importance of identifying churners, we strategically select a probability threshold of 0.5. This choice prioritizes capturing a larger proportion of true churn events, even if it comes at a slight cost to overall accuracy. This approach ensures that we can proactively target at-risk customers with retention efforts and minimize customer churn.



MODEL SUMMARY

Train Set

- Accuracy = 0.85
- Sensitivity = 0.89
- Specificity = 0.81

Test Set

- Accuracy = 0.82
- Sensitivity = 0.83
- Specificity = 0.82

Overall, the model is performing well in the test set, what it had learnt from the train set.

Our analysis revealed that the logistic regression model without PCA (Model-3) achieved comparable sensitivity and accuracy to models that incorporated PCA dimensionality reduction.

Benefits of Model-3 (Logistic Regression without PCA):

- **Interpretability:** Model-3 offers a clear advantage in interpretability. It directly uses the original features, allowing for a straightforward explanation of the important predictor variables and their individual significance in churn prediction. This feature-centric approach aligns well with business needs, as it facilitates the identification of specific factors that influence churn risk.
- **Simplicity:** Model-3 avoids the complexity introduced by PCA. This simpler model is easier to implement and understand, making it a more practical choice for business stakeholders.

Conclusion:

While PCA-based models can be effective, Model-3 (logistic regression without PCA) emerges as the preferable choice for this scenario. It delivers competitive performance while offering superior interpretability and a simpler structure, which resonates with business requirements for understanding and addressing churn.

TOP PREDICTORS

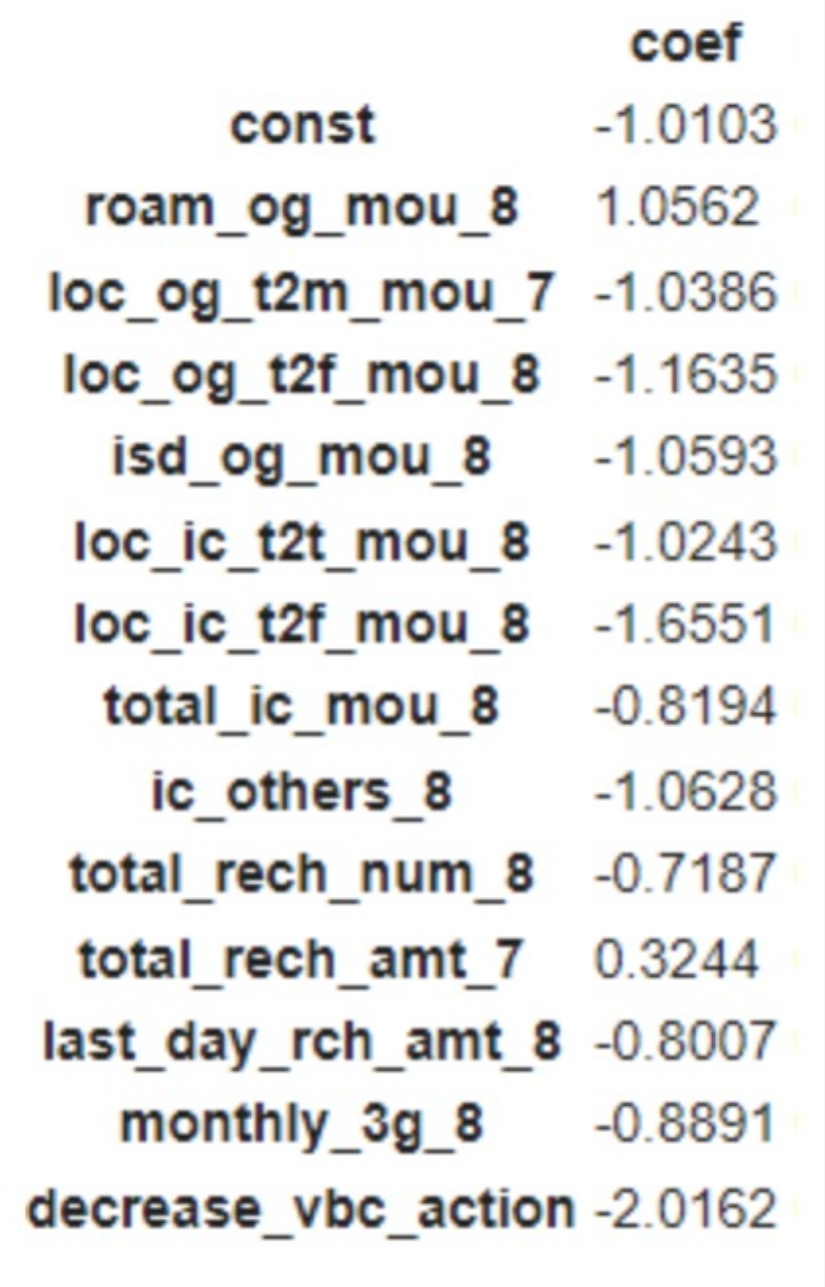
The logistic regression model identified several significant features that influence customer churn propensity. Notably, most of these features exhibit negative coefficients, indicating an inverse correlation with churn probability. In other words, higher values of these features are associated with a lower likelihood of churn.

- **Total Incoming Minutes of Usage (total_ic_mou_8):** A decrease in incoming call minutes during the action phase compared to previous months suggests a potential churn risk. This could indicate reduced customer engagement with the network.

Explanation of Feature Importance:

The negative coefficients associated with these features imply that they act as protective factors against churn. Customers who exhibit higher values on these features are less likely to churn.

This analysis provides valuable insights into customer behavior and the factors that contribute to churn. By understanding these key drivers, telecommunication companies can develop targeted retention strategies to mitigate customer churn and maintain a healthy subscriber base.



	coef
const	-1.0103
roam_og_mou_8	1.0562
loc_og_t2m_mou_7	-1.0386
loc_og_t2f_mou_8	-1.1635
isd_og_mou_8	-1.0593
loc_ic_t2t_mou_8	-1.0243
loc_ic_t2f_mou_8	-1.6551
total_ic_mou_8	-0.8194
ic_others_8	-1.0628
total_rech_num_8	-0.7187
total_rech_amt_7	0.3244
last_day_rch_amt_8	-0.8007
monthly_3g_8	-0.8891
decrease_vbc_action	-2.0162

RECOMMENDATION

The logistic regression model identified several key customer behaviours associated with an increased likelihood of churn. These insights can be leveraged to develop targeted retention campaigns:

1. **Low Usage of Local and International Calls:** Customers exhibiting a decrease in incoming and outgoing local call minutes, as well as outgoing international call minutes, during the action phase require attention. This decline in usage signifies reduced engagement and potential churn risk.
2. **Decreased Outgoing Other Charges and Incoming Other Charges:** A drop in "outgoing other charges" in July and "incoming other charges" in August might indicate reduced customer satisfaction or a shift towards alternative communication channels. These customers could benefit from targeted promotions or service plan adjustments.
3. **Increase in Value-Based Costs:** Customers experiencing a rise in value-based costs (likely due to increased data usage) during the "action phase" might be dissatisfied with their current plan. Offering targeted promotions or flexible data packages could help retain these customers.
4. **Decreasing Local Fixed-Line Calls:** A decline in local incoming call minutes from a specific operator's fixed-line network (T to fixed lines of T) in August suggests a potential search for alternative service providers. Proactive outreach with competitive offers might be necessary to retain these customers.
5. **Reduced Monthly 3G Usage:** Customers with a significant decrease in monthly 3G data usage during August might be considering switching to a competitor offering better data plans. Tailored data packages or addressing network connectivity issues could help prevent churn.
6. The roam_og_mou_8 variable has a positive coefficient, indicating a positive correlation with churn. Customers with an increase in roaming outgoing minutes (traveling more) are more likely to churn. While this might seem counterintuitive, it could be due to factors like high roaming charges or a lack of suitable roaming packages. Offering competitive roaming plans or targeted promotions during travel periods could be beneficial.

These targeted interventions based on customer behavior patterns can significantly improve customer retention efforts and minimize churn.

THANK YOU