greatlearning
Power Ahead

Capstone Presentation - Customer Churn

By ISHA SHUKLA

Date - 07 Sep 2024

Business Problem Understanding

The DTH provider is struggling to retain customers due to rising competition. Losing even one account can have a significant impact because each account may include multiple customers. To address this, the company needs to develop a churn prediction model that identifies at-risk accounts and helps create targeted, cost-effective campaigns to retain them.

As data analysts, we are working with the firm to develop a churn prediction model that will enable them to make informed adjustments to their business strategies. This model will help the company customize their marketing and advertising campaigns to retain existing customers and attract new ones.

Exploratory Data Analysis (EDA)

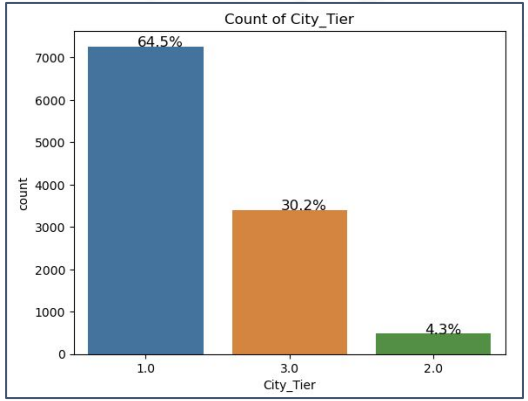
- The dataset consists of 11,260 rows and 19 columns(including Target variable - Churn).
- No duplicate records are found in the dataset.
- Null values are present in some columns.
- The dataset includes 5 columns with float data types, 2 integer data types, and 12 with object data types.
- Outliers have been identified within the dataset.
- Churn shows data is imbalance in nature.
- Numerical variables are not normally distributed and exhibit skewness.

```
(11260, 19)
```

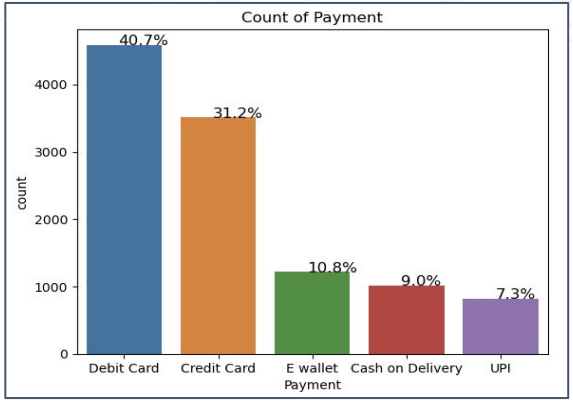
Shape of the data

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 11260 entries, 0 to 11259
Data columns (total 19 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   AccountID               11260 non-null  int64
 1   Churn                   11260 non-null  int64
 2   Tenure                  11158 non-null  object
 3   City_Tier               11148 non-null  float64
 4   CC_Contacted_LY         11158 non-null  float64
 5   Payment                 11151 non-null  object
 6   Gender                  11152 non-null  object
 7   Service_Score           11162 non-null  float64
 8   Account_user_count      11148 non-null  object
 9   account_segment         11163 non-null  object
 10  CC_Agent_Score          11144 non-null  float64
 11  Marital_Status          11048 non-null  object
 12  rev_per_month           11158 non-null  object
 13  Complain_ly             10903 non-null  float64
 14  rev_growth_yoy          11260 non-null  object
 15  coupon_used_for_payment 11260 non-null  object
 16  Day_Since_CC_connect    10903 non-null  object
 17  cashback                10789 non-null  object
 18  Login_device            11039 non-null  object
dtypes: float64(5), int64(2), object(12)
memory usage: 1.6+ MB
```

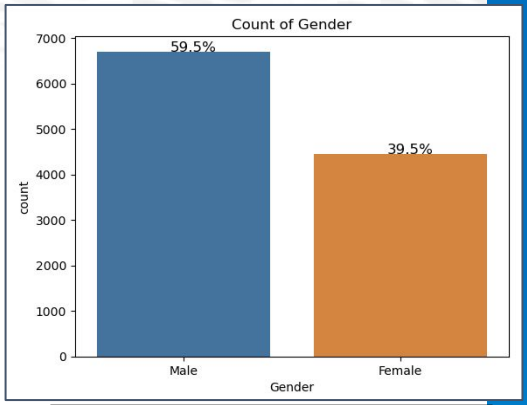
Univariate Analysis



Most customers are from Tier 1 cities, followed by Tier 3 , with the fewest in Tier 2 .

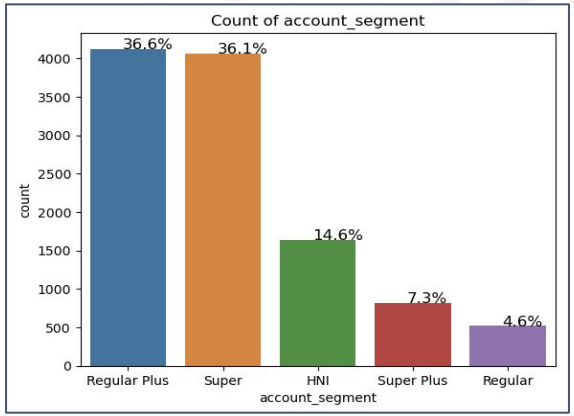


Most of customers prefer Debit Card payments, 31.2% use Credit Cards, 10.8% choose E-wallets, 9% prefer Cash on Delivery, and UPI is the least preferred payment method.

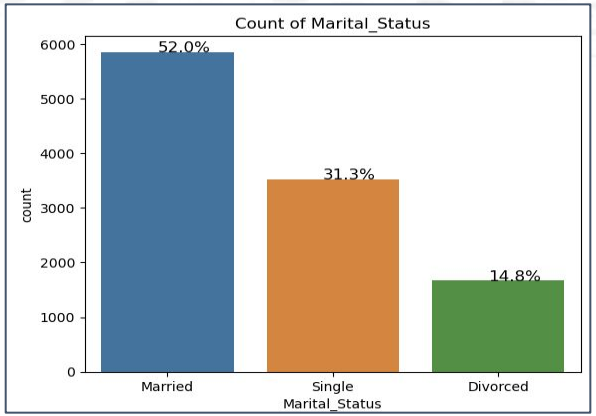


Most customers are male (59.5%), while 39.5% are female.

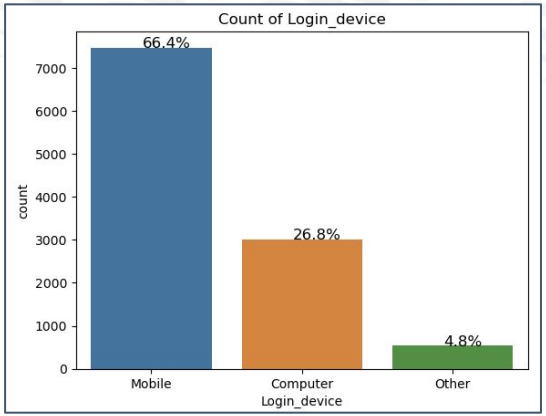
Univariate Analysis



'Regular Plus' and 'Super' have the most customers (over 36%), 'Regular' the least, with High Net Income (HNI) at 14.6% and Super Plus at 7.3%.

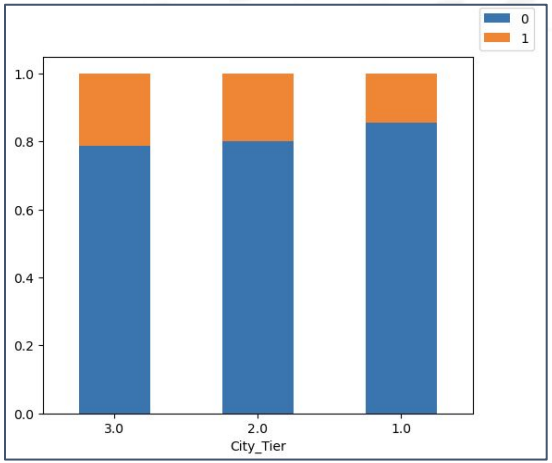


The majority of primary account holders are married, followed by singles, and then divorced individuals.

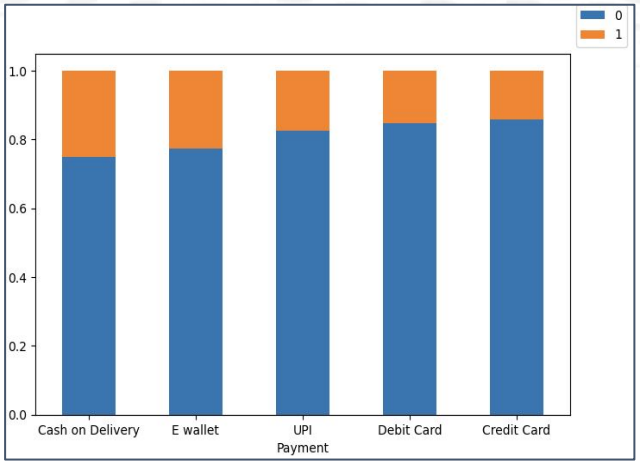


The majority of login devices are mobile, then followed by computers.

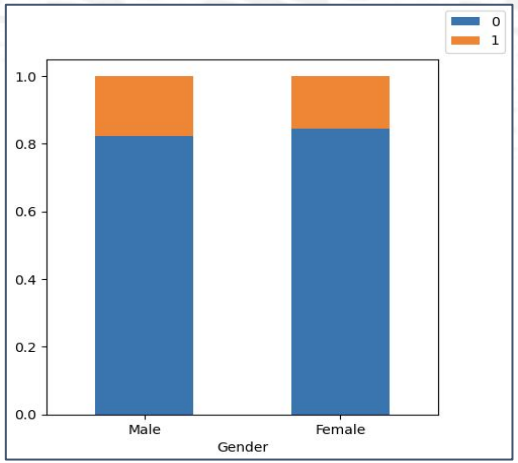
Bivariate Analysis



Tier 1 has the most loyal customers, while Tier 3 poses the greatest churn risk.

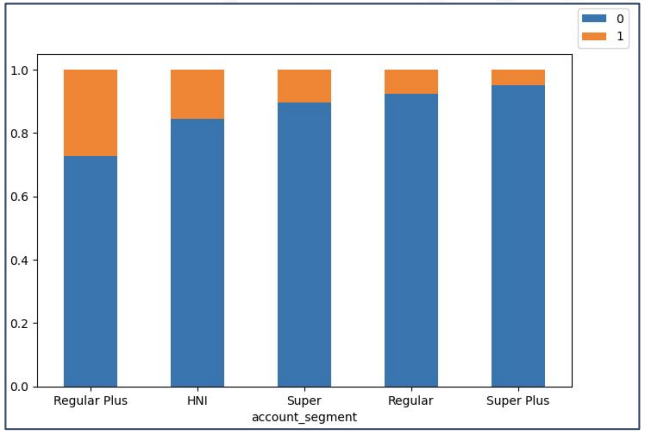


Alternative payment methods (E-wallets, Cash on Delivery) have higher churn rates than traditional methods (Debit/Credit Cards).

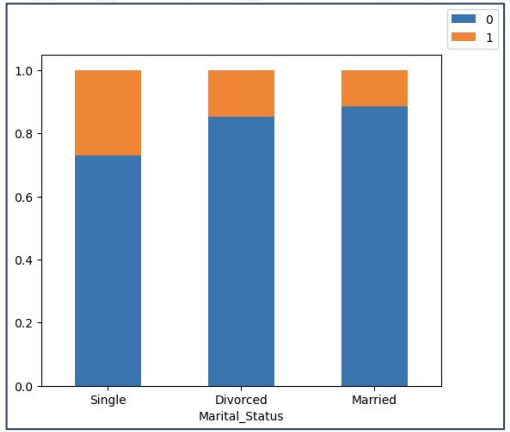


Males tend to churn at a slightly higher rate than females, but both genders have similar churn risk.

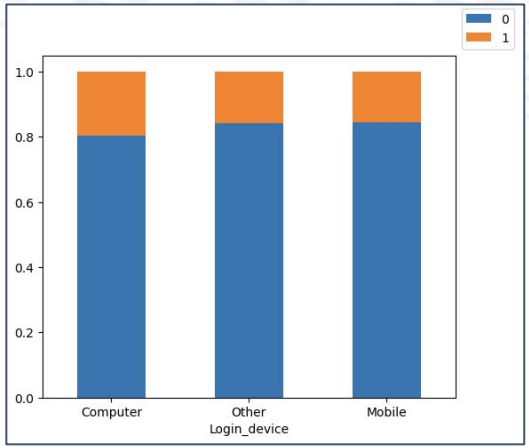
Bivariate Analysis



Regular Plus has the highest churn rate, while Super Plus has the lowest.



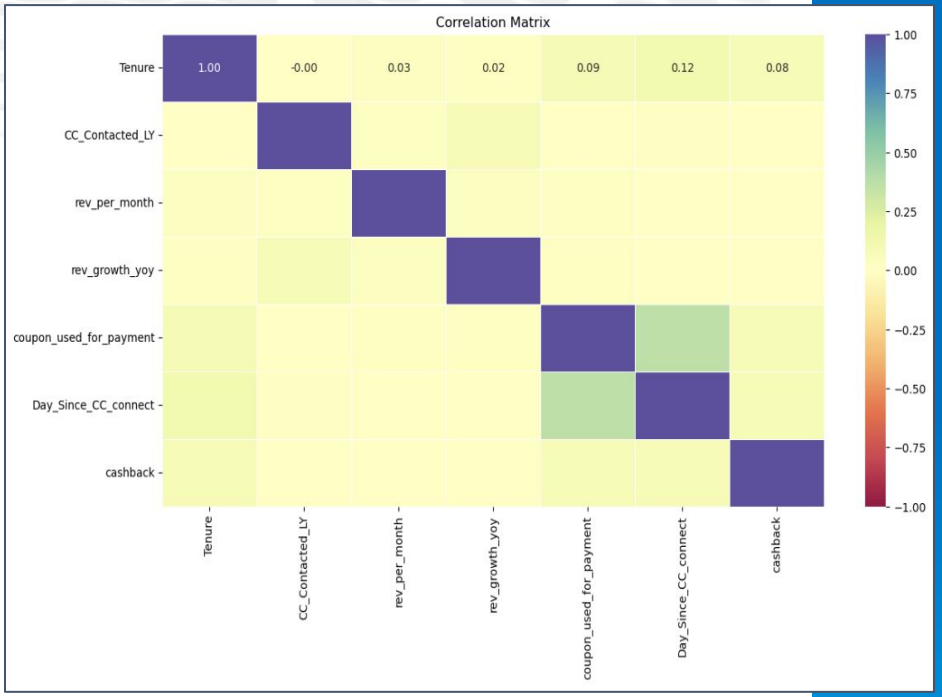
Single customers exhibit the highest churn rate, while married customers are the most stable group.



Mobile users make up the largest segment, but also have a high churn rate, making them a key target for retention efforts.

Multivariate Analysis

- **Churn and Complain_ly:** Positive Correlation (0.251): Customers who have lodged complaints in the previous year are more likely to churn.
- **Churn and Tenure:** Negative Correlation (-0.233): Customers with longer tenure are less likely to churn.
- **Service_Score and Account_user_count:** Strong Positive Correlation (0.323): Higher service scores are associated with accounts that have more users.



	Tenure	CC_Contacted_LY	rev_per_month	rev_growth_yoy	coupon_used_for_payment	Day_Since_CC_connect	cashback
Tenure	1.000000	-0.004261	0.028431	0.018824	0.089171	0.122612	0.078416
CC_Contacted_LY	-0.004261	1.000000	0.015675	0.072913	0.004969	0.012938	0.002679
rev_per_month	0.028431	0.015675	1.000000	0.024114	0.016548	-0.000923	0.002974
rev_growth_yoy	0.018824	0.072913	0.024114	1.000000	0.018341	0.002206	-0.001157
:oupon_used_for_payment	0.089171	0.004969	0.016548	0.018341	1.000000	0.361735	0.072861
Day_Since_CC_connect	0.122612	0.012938	-0.000923	0.002206	0.361735	1.000000	0.084465
cashback	0.078416	0.002679	0.002974	-0.001157	0.072861	0.084465	1.000000

Insights from Analysis

1. City Tier 1 has the highest retention rate, with a majority of customers not churning.
2. Male customers are more likely to churn than female customers.
3. Regular Plus segment has the highest churn rate, while Super Plus segment has the lowest churn rate.
4. Mobile users have a higher churn rate compared to computer users.
5. Customers who have complained in the last year are more likely to churn.
6. Longer-tenured customers are less likely to churn.
7. Customers with more recent customer care connections are less likely to churn.
8. Most customers rated services and customer care interactions as "3", indicating a neutral sentiment.
9. Transaction via UPI and e-wallet is very low, suggesting a lack of adoption of these payment methods.
10. Customers with marital status "single" contribute max towards churn, indicating a higher risk of churn among single customers.
11. Any complaints raised in last 12 months doesn't show any impact toward churn, suggesting that complaints may not be a key driver of churn.
12. Tenure and cashback are directly proportional to each other, indicating that longer-tenured customers receive more cashback.
13. Computer usage is more in Tier 1 city followed by Tier 3 and Tier 2 city, highlighting differences in device usage across cities.

Modelling Approach Used & Why

Data Preprocessing

- **Handling Missing Values:** Imputed missing values with KNN imputer
- **Outlier Detection and Treatment:** Identified and addressed outliers to maintain data integrity
- **Anomaly Detection:** Detected and treated anomalies to prevent data contamination
- **Data Scaling:** Scaled data using StandardScaler to prevent feature dominance.
- **Categorical Encoding:** Replaced object-type categorical variables with integer types for efficient modeling

Modelling Approach

- **Customer Behaviour Analysis:** Analyzed customer behaviour to identify patterns and alert potential churners
- **Pain Point Identification:** Identified customer pain points to inform effective retention strategies
- **Classification Modeling:** Built a classification model to identify shared behavioural patterns of customers who have abandoned purchases, enabling targeted interventions

Models Used

1. Built a classification model to categorize customers into churner/non-churner classes
2. Utilized the following models to achieve this:
    a. **Logistic Regression**
    b. **Linear Discriminant Analysis (LDA)**
    c. **K-Nearest Neighbors (KNN)**
    d. **Gaussian Naive Bayes**
    e. **Random Forest**
    f. **Bagging - Random Forest**
    g. **Gradient Boost**
    h. **XGBoost**
    i. **Ada-Boosting**
    j. **Support Vector Machine (SVM)**
3. **Hyperparameter Tuning:** Performed hyperparameter tuning to optimize model accuracy and achieve better results

Model's Comparison Across Parameters

Models	Training Dataset							Testing Dataset						
	Accuracy	Precision-0	Recall - 0	F1-Score - 0	Precision - 1	Recall - 1	F1-Score - 1	Accuracy	Precision-0	Recall - 0	F1-Score - 0	Precision - 1	Recall - 1	F1-Score - 1
Logistic Regression	0.89	0.9	0.97	0.93	0.77	0.47	0.59	0.89	0.91	0.97	0.94	0.79	0.5	0.61
Logistic Regression - GridSearchCV	0.89	0.9	0.97	0.93	0.77	0.47	0.58	0.89	0.9	0.97	0.94	0.77	0.48	0.59
Logistic regression - SMOTE	0.81	0.82	0.79	0.81	0.8	0.83	0.81	0.79	0.95	0.78	0.86	0.43	0.82	0.56
Linear Discriminant Analysis Model	0.88	0.89	0.97	0.93	0.77	0.42	0.54	0.89	0.9	0.98	0.93	0.79	0.44	0.56
LDA model - GridSearchCV	0.88	0.89	0.97	0.93	0.77	0.42	0.54	0.89	0.9	0.98	0.93	0.79	0.44	0.56
LDA model - SMOTE	0.81	0.83	0.77	0.8	0.79	0.84	0.81	0.77	0.96	0.76	0.85	0.41	0.83	0.55
K-Nearest Neighbors (KNN) Model	0.93	0.94	0.98	0.96	0.86	0.68	0.76	0.88	0.91	0.96	0.93	0.7	0.51	0.59
KNN with n_neighbors = 3	0.95	0.97	0.98	0.97	0.89	0.83	0.86	0.89	0.92	0.94	0.93	0.68	0.61	0.64
KNN - GridSearchCV	1	1	1	1	1	1	1	0.94	0.96	0.98	0.97	0.88	0.78	0.82
KNN - SMOTE	0.94	1	0.88	0.93	0.89	1	0.94	0.82	0.97	0.81	0.88	0.48	0.86	0.62
Gaussian Naive Bayes	0.86	0.92	0.91	0.92	0.58	0.59	0.59	0.85	0.92	0.9	0.91	0.55	0.6	0.58
Gaussian Naive Bayes - SMOTE	0.74	0.77	0.68	0.72	0.71	0.8	0.76	0.69	0.94	0.67	0.78	0.32	0.78	0.46
Random Forest	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**0.96**	**0.97**	**0.99**	**0.98**	**0.95**	**0.83**	**0.89**
Random Forest - SMOTE	1	1	1	1	1	1	1	0.97	0.98	0.99	0.98	0.94	0.89	0.91
Bagging - Random Forest	0.99	0.99	1	1	1	0.96	0.98	0.88	0.91	0.96	0.93	0.7	0.51	0.59
Bagging - Random Forest - SMOTE	1	1	1	1	1	1	1	0.96	0.97	0.98	0.97	0.88	0.85	0.86
Gradient Boost	0.92	0.93	0.98	0.95	0.85	0.64	0.73	0.91	0.92	0.97	0.95	0.81	0.58	0.68
Gradient Boost - SMOTE	0.93	0.92	0.94	0.93	0.94	0.92	0.93	0.9	0.94	0.94	0.94	0.7	0.73	0.71
XGBoost	0.98	0.98	1	0.99	0.98	0.91	0.94	0.95	0.96	0.98	0.97	0.9	0.77	0.83
XGBoost - SMOTE	0.98	0.97	0.99	0.98	0.99	0.97	0.98	0.94	0.96	0.97	0.96	0.85	0.78	0.82
Ada-Boost	0.89	0.9	0.97	0.94	0.78	0.46	0.58	0.89	0.9	0.97	0.94	0.77	0.48	0.59
Ada-Boost-SMOTE	0.86	0.86	0.86	0.86	0.86	0.86	0.86	0.85	0.95	0.87	0.91	0.54	0.78	0.64
SVM	0.89	0.9	0.98	0.93	0.79	0.44	0.57	0.89	0.9	0.98	0.94	0.8	0.46	0.58
SVM - GridSearchCV	0.89	0.9	0.97	0.93	0.77	0.46	0.57	0.89	0.9	0.97	0.93	0.75	0.48	0.59
SVM - SMOTE	0.82	0.83	0.79	0.81	0.8	0.84	0.82	0.79	0.96	0.79	0.86	0.44	0.82	0.57

Insights from Model Building

- **Random Forest with default parameters** is the top-performing model, but **XGBoost and Gradient Boost** are close contenders.
- **XGBoost, Random Forest, and Gradient Boosting** should be preferred for deployment as they handle class imbalance effectively while maintaining high accuracy.
- **SMOTE** is an effective technique to improve churn detection in models that initially struggle with class imbalance, though it may slightly impact overall accuracy.
- **Gradient Boosting** and **KNN** (with tuning) also perform well, especially for identifying churn cases.
- **Logistic Regression, LDA, and Naive Bayes** struggle with churn prediction, even with SMOTE, making them less ideal for churn-focused tasks.

Insights from Model Building

- Using techniques like GridSearchCV improves model performance, especially for **KNN** and **SVM**, by finding the optimal hyperparameters. However, excessive tuning can lead to **overfitting**, as seen with KNN.
- **Bagging techniques, such as Random Forest**, help reduce variance and provide stable, reliable predictions. They are particularly effective in balancing overall accuracy and identifying non-churn customers (class 0). With the addition of SMOTE, Random Forest's ability to detect churn cases (class 1) improves, making it a solid choice when stability is crucial.
- **Boosting techniques, like Gradient Boosting and XGBoost**: These models perform well in identifying churn cases, delivering high accuracy and recall even without oversampling methods. XGBoost, in particular, is highly effective for imbalanced datasets, consistently delivering superior performance across all metrics.

Recommendations

- Implement a customer retention program targeting high-risk customers.
- Improve customer service and support to reduce churn.
- Provide training and incentives to customer-facing staff to improve customer service and retention.
- Offer personalized promotions and offers to high-value customers.
- Conduct regular customer surveys to gather feedback and identify areas for improvement.
- Develop a loyalty program to reward loyal customers and encourage retention.
- Analyze customer behavior and transactional data to identify early warning signs of churn.
- Offer personalized product recommendations to customers based on their purchase history and preferences.

Recommendations

- Improve the user experience of the company's website and mobile app to reduce friction and improve customer retention.
- Enhance customer engagement through regular communication and feedback mechanisms.
- Increase visibility in Tier-2 and Tier-3 cities to improve customer acquisition by partnering with local businesses, sponsoring local events, developing targeted marketing campaigns, and establishing a local presence through physical offices
- Introduce a referral drive for existing customers to acquire new customers.

Thank You!