Customer Churn

New Users

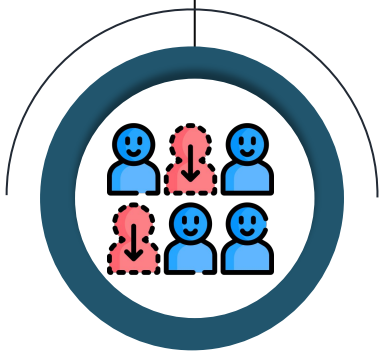
Active Users

Churned Users

Prepared by : Dhruv Dosad

Mentor : Sharath Srivatsa

Agenda



01

Business Problem

- Problem Understanding
- Objective
- Brief overview of Data

02

Modelling Approach

- Modeling Process Flow
- Comparison Table
- Best Model

03

Insights

- Insights from Analysis

04

Recommendations

- Recommendations for Business

Business Problem Understanding

Problem Statement

- ➔ DTH provider facing retention challenges amidst market competition
- ➔ Need of segmented marketing strategies

Objective

- ➔ Form a machine learning model with the smallest false negative (Recall).
- ➔ Predict customers who have the potential to churn with the best prediction model.
- ➔ Providing insights & recommendations to identify factors that influence the churn rate

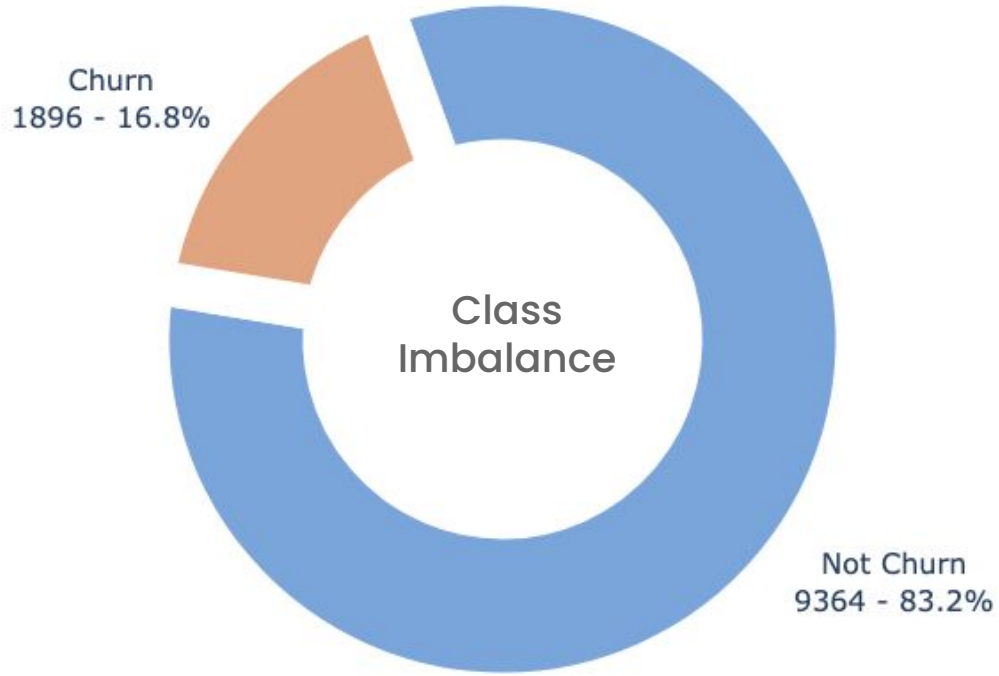
Goal

- ➔ Enhance customer retention
- ➔ Improve market competitiveness
- ➔ Explore revenue growth opportunities

Overview of Data

Data Size	Rows 11260	Features 19
Data Type	Numeric 7	Categorical 12
Data Issues	Yes	
Duplicates	None	
Missing	4361 Data Pts	2.04% of Total
Outliers	Yes	

Churn Overview



** More Details in Appendix for overview of data*

Modeling Approach

Data Cleaning

· Corrected data types
· Fixed Irrelevant Data

Fixed Data Issues

· Treated Missing values*
· No Duplicates
· Treated Outlier*

Feature Encoding

· One Hot Encoder for categorical
· Standard Scaler for Numerical data

Transforming

· 80 : 20 split on Train : Test

Build Different Models

· Decision Tree
· Random Forest
· Linear Discriminant Analysis (LDA)
· k-Nearest Neighbours(KNN)
· Naive Bayes
· Neural Network (ANN)
· SVM

Regularize & Ensemble (Hypertune)

· Decision Tree
· LDA
· Random Forest
· Bagging
· Boosting : Adaboost, XGB, CatBoost
· Neural Network (ANN)
· SVM

Evaluate & Compare Models

· Recall as Evaluation metric
· Other metrics precision, F1 score, and AUC-ROC in combination with recall

Best Model

· SVM was found to be best model with Recall **95.20%**
· CatBoost as Balanced Model with f1-score **95.87%**

** Missing Data : Categorical treated using Mode and Numerical using KNN imputer*
** Outliers : Lower limit capped at 1st and upper limit at 99th quartile*
** Refer Appendix for more details*

Modeling Approach : Model Comparison Table (Test)

Model	Majority Class(Not Churn)			Minority Class(Churn)			accuracy	Overfit
	Precision	Recall	f1-score	Precision	Recall	f1-score		
SVM with Tuning	98.97%	97.90%	98.43%	90.63%	95.20%	92.86%	97.42%	No
CatBoost with Tuning	98.67%	99.62%	99.14%	98.15%	93.69%	95.87%	98.58%	No
Random Forest	97.61%	99.14%	98.37%	95.64%	88.64%	92.01%	97.29%	Slightly Overfit
KNN	97.62%	99.25%	98.42%	96.16%	88.64%	92.25%	97.38%	Slightly Overfit
XGBoost with Tuning	97.36%	99.19%	98.27%	95.84%	87.37%	91.41%	97.11%	Slightly Overfit
Adaboost	97.21%	99.62%	98.40%	98.00%	86.62%	91.96%	97.34%	Slightly Overfit
Tuned Random Forest	96.41%	89.60%	92.88%	63.38%	84.34%	72.37%	88.68%	Slightly Overfit
Decision Tree	96.50%	96.55%	96.53%	83.80%	83.59%	83.69%	94.27%	Yes
Bagging Classifier	96.47%	98.71%	97.58%	93.20%	83.08%	87.85%	95.96%	Yes
Regularized Decision Tree	95.44%	96.93%	96.18%	84.47%	78.28%	81.26%	93.65%	Yes
Neural Network	94.63%	98.76%	96.65%	92.70%	73.74%	82.14%	94.36%	Yes
LDA	87.22%	97.84%	92.23%	76.47%	32.83%	45.94%	86.41%	Yes
Tuned LDA	87.22%	97.84%	92.23%	76.47%	32.83%	45.94%	86.41%	Yes
Naive bayes	85.26%	97.84%	91.12%	67.21%	20.71%	31.66%	84.28%	Yes

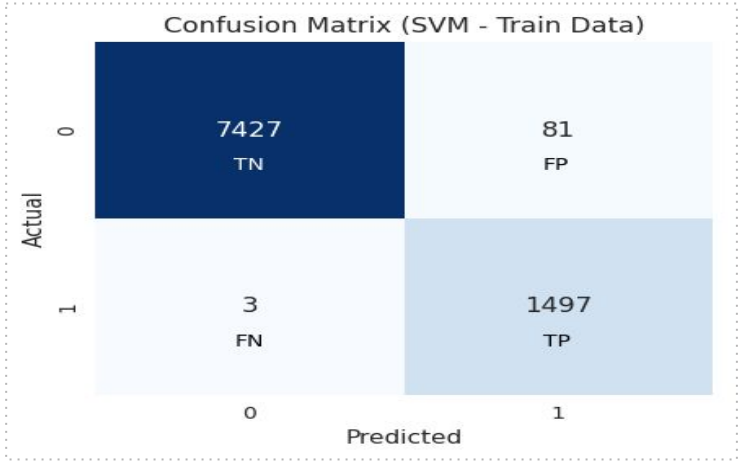
Best Models

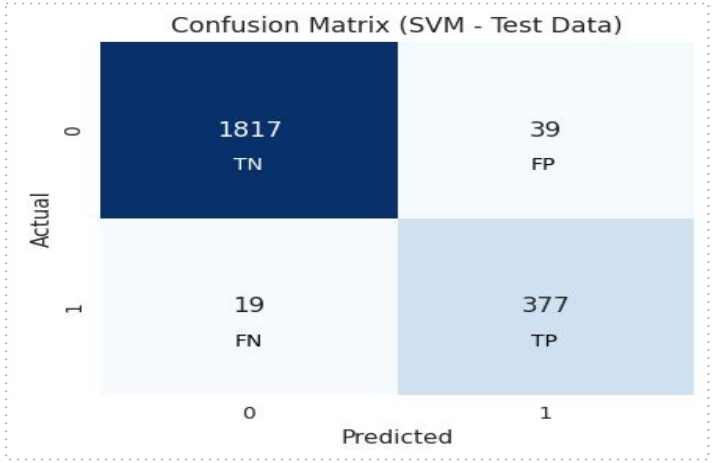
Improved recall but not able to generalize well on new data as compared to train

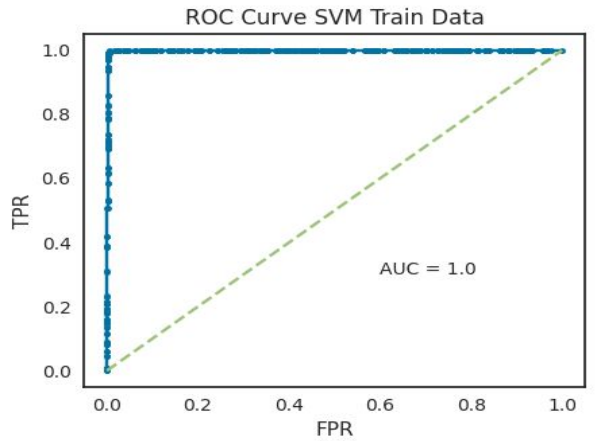
Poor Recall

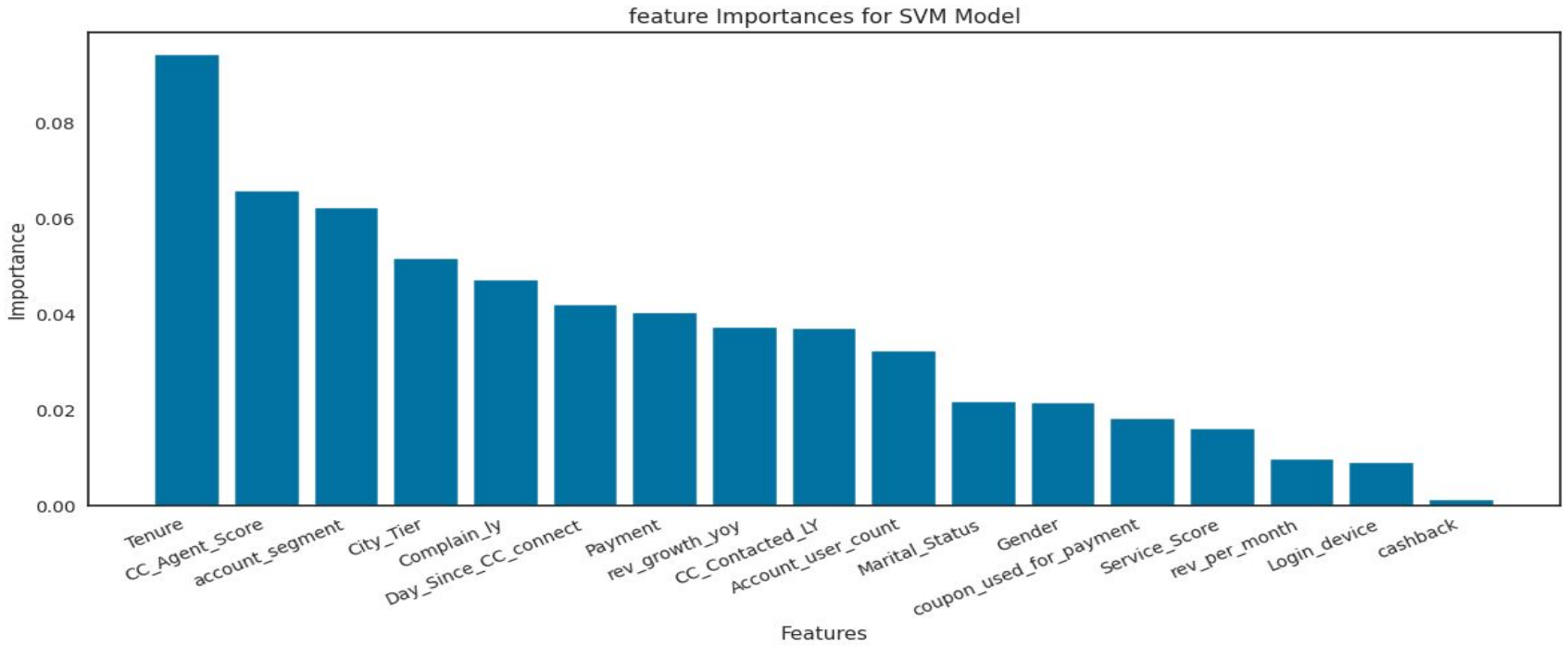
* Train results in Appendix

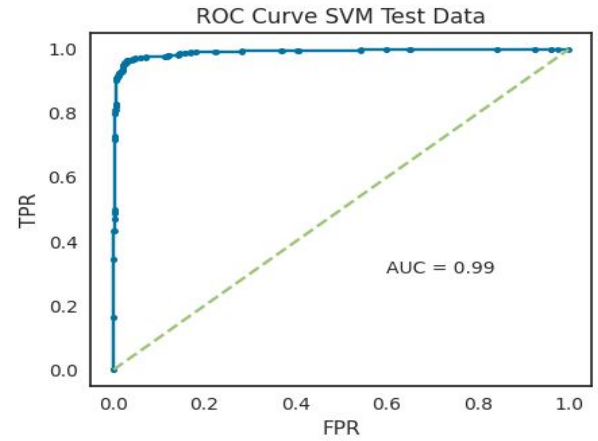
Modeling Approach : Best Model









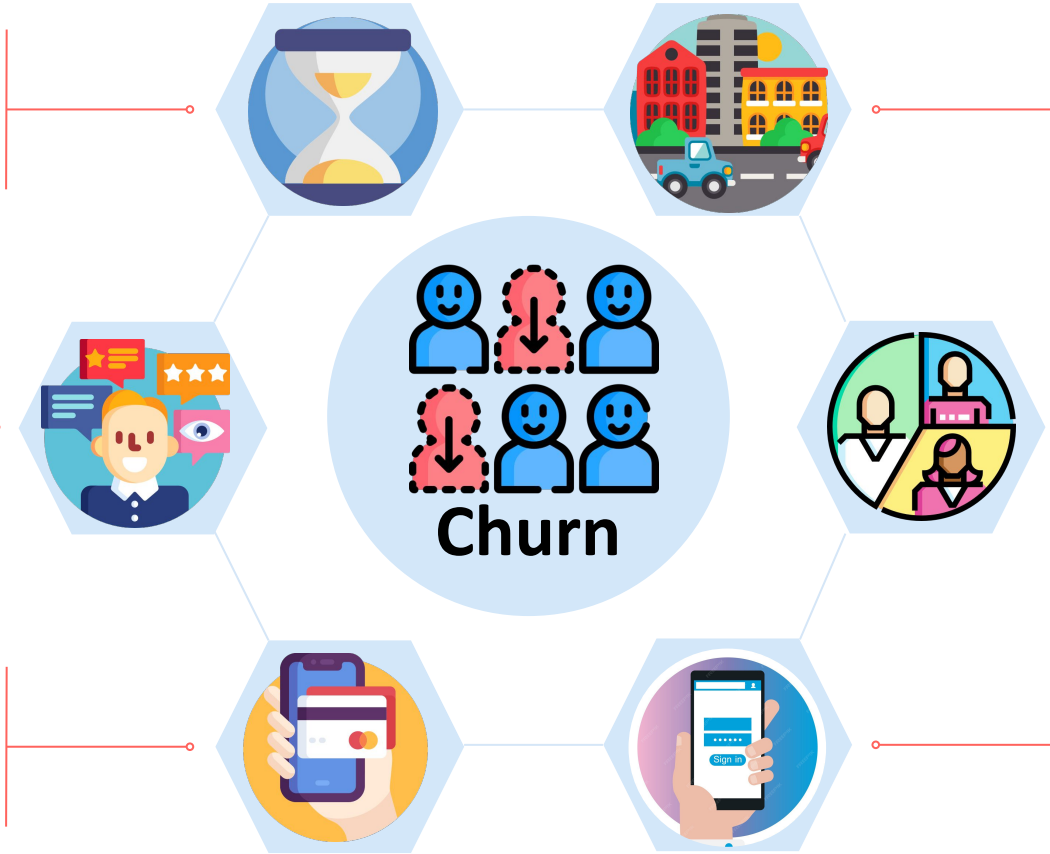


Insights

Tenure
Lower Tenure, Higher Churn (Median: Churn: 1 Mo, Non-Churn - 10 mo)

City Tier
- City_Tier 3 have a higher churn rate (21%)

Customer complaint
Higher Churn when:
- Recent Complaints(Avg 2 days)
- Complaint in last year (32% Churn)



Account Segment
Regular Plus customers, have a higher churn rate (27%)

Payment mode
"Debit Card" and "Credit Card" are popular (72.89% users)

Login Device
Mobile login, lower churn (16% as compared to computers 20%)

- Tenure in months
- If no Complaint last year- 15% Churn rate | Median Days since last connect for churn: 4, Non Churn- 2
- Churn rate 23% for ewallet, 25% for COD, UPI, 17% . DC/CC ~15%

- City Tier 1 & 2 has Churn rate as 15% and 20%, however, city 2 has very low subs count.
- Other Account segments has less Churn rate
- Mobile login churn 16% , Computer 20%

Recommendations

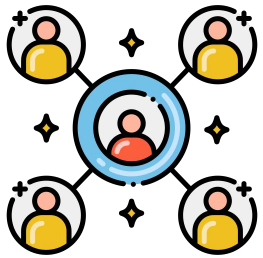
In order to Increase Customers Loyalty and Tenure

Create Customer Experience

- ➔ **Deep dive** in **customer complaints** to identify gaps.
- ➔ **promptly address customer complaints**
- ➔ **Ask for customer feedback** at regular intervals

Strategic Retention Approach

- ➔ Offering exclusive upgrades **to higher account segment**
- ➔ Increases loyalty and reduces **reliance on cashbacks**
- ➔ **Premium services** to **Higher Account Segments**. **Pay attention** to **regular plus** customers



Introduce Referral Drive

- ➔ Helps acquire **new customers**
- ➔ Against that **offer Existing users** exclusive benefits of **cashback or upgrade** to next tier

Other Key points

- ➔ Encourage customers to set up payments either through **debit card or credit card**
- ➔ Introduce **curated contents for Tier 2 & 3 cities** to increase visibility.
- ➔ **Promote app/mobile** based **logins**
- ➔ Competitive benchmarking

Appendix

Appendix | Business Problem (Detailed)

PROBLEM STATEMENT:

The problem at hand is developing a churn prediction model for a DTH service provider. The primary challenge faced by the company is retaining existing customers in a highly competitive market. Account churn, where one account can encompass multiple customers, poses a significant threat, as the loss of a single account can result in the attrition of several customers. The problem is to create a predictive model to identify potential churners and design segmented offers to retain them.

WHY IS IT NEEDED?

The need for this study or project arises from the company's requirement to address the critical issue of customer churn. Retaining customers is pivotal, and the company aims to achieve this by creating a model that can predict churn and provide tailored offers to potential churners. In this context, the project is essential for enhancing customer retention, staying competitive in the market, exploring revenue growth opportunities, and elevating the overall customer experience.

UNDERSTANDING BUSINESS OPPORTUNITY

The business opportunity lies in devising a sophisticated churn prediction model and formulating distinct campaign recommendations to combat churn effectively. The potential benefits include strengthening customer loyalty, improving market competitiveness, unlocking revenue growth, and delivering an exceptional customer experience. The project represents a unique business opportunity to align recommendations with strategic objectives, ensuring they don't inadvertently lead to losses for the company. It underscores the importance of creating offers that are both enticing to customers and financially viable for the business, presenting a balance between customer satisfaction and revenue assurance.

Appendix | Data Description

Variable	Description
AccountID	Account unique identifier
Churn	Account churn flag (Target)
Tenure	Tenure of account
City_Tier	Tier of the primary customer's city
CC_Contacted_L1	How many times all the customers of the account have contacted customer care in the last 12 months
Payment	Preferred payment mode of the customers in the account
Gender	Gender of the primary customer of the account
Service_Score	Satisfaction score given by customers of the account on the service provided by the company
Account_user_count	Number of customers tagged with this account
account_segment	Account segmentation based on spending
CC_Agent_Score	Satisfaction score given by customers of the account on customer care service provided by the company
Marital_Status	Marital status of the primary customer of the account
rev_per_month	Monthly average revenue generated by the account in the last 12 months
Complain_l12m	Any complaints raised by the account in the last 12 months
rev_growth_yoy	Revenue growth percentage of the account (last 12 months vs. last 24 to 13 months)
coupon_used_l12m	How many times customers have used coupons to make payments in the last 12 months
Day_Since_CC_connect	Number of days since no customers in the account have contacted customer care
cashback_l12m	Monthly average cashback generated by the account in the last 12 months
Login_device	Preferred login device

Appendix | Data Types

Original

```
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)
```

Corrected

```
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   Tenure                   11260 non-null  float64
 1   CC_Contacted_LY          11260 non-null  float64
 2   rev_per_month            11260 non-null  float64
 3   rev_growth_yoy           11260 non-null  float64
 4   coupon_used_for_payment  11260 non-null  float64
 5   Day_Since_CC_connect     11260 non-null  float64
 6   cashback                 11260 non-null  float64
 7   City_Tier                11260 non-null  object
 8   Payment                  11260 non-null  object
 9   Gender                   11260 non-null  object
 10  Service_Score            11260 non-null  object
 11  Account_user_count       11260 non-null  object
 12  account_segment          11260 non-null  object
 13  CC_Agent_Score           11260 non-null  object
 14  Marital_Status           11260 non-null  object
 15  Complain_ly              11260 non-null  object
 16  Login_device             11260 non-null  object
 17  Churn                    11260 non-null  int64
dtypes: float64(7), int64(1), object(10)
```

Appendix : Data Cleaning

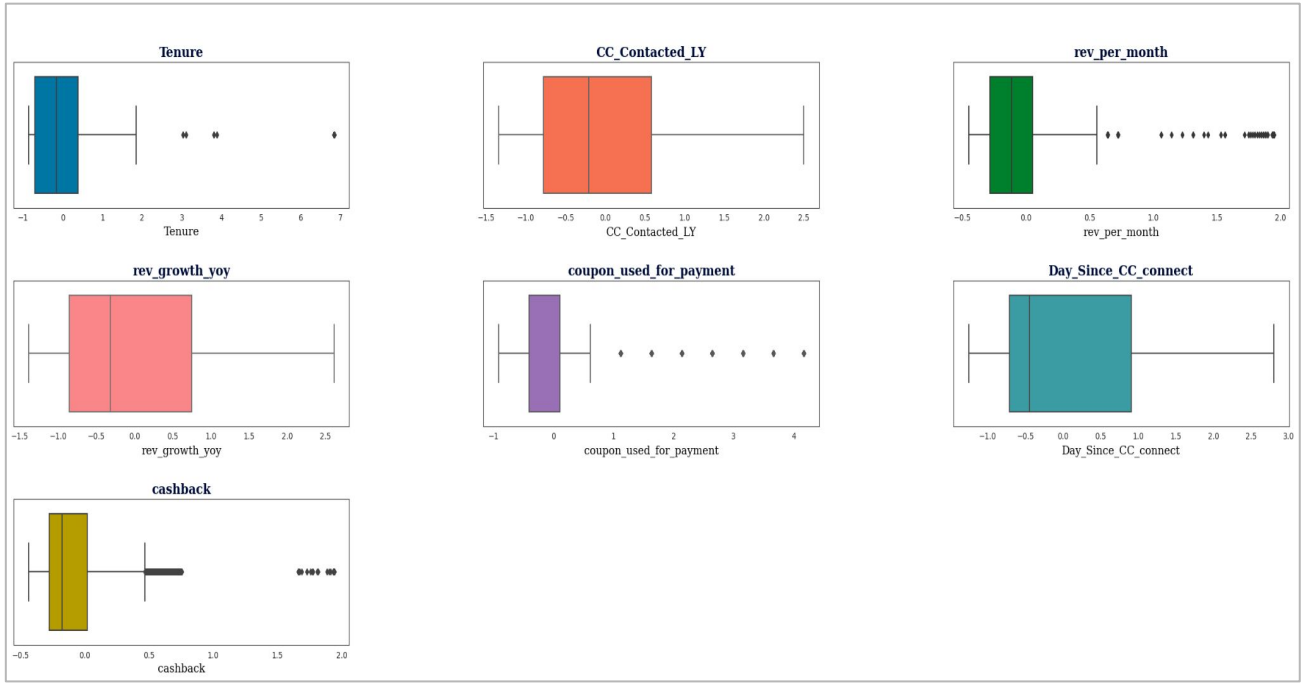
- Tenure (Unexpected values like '#', high numerical values)
- Account_user_count (Unexpected value '@' alongside numerical counts)
- Gender (Inconsistent formatting with 'F' and 'M' alongside 'Female' and 'Male')
- rev_per_month (Unexpected values like ' + ', high numerical values)
- rev_growth_yoy (Unexpected values like ' $ ' mixed with numerical values)
- coupon_used_for_payment (Unexpected values like '#', '$', and '*' mixed with numerical values)
- Day_Since_CC_connect (Unexpected values like '$' mixed with numerical values)
- Login_device (Unexpected value '&&&&' alongside 'Mobile' and 'Computer')

All unexpected values were **replaced by "NAN"** (missing values) & later treated using missing value Treatment technique.

Appendix | Missing Values & Outliers

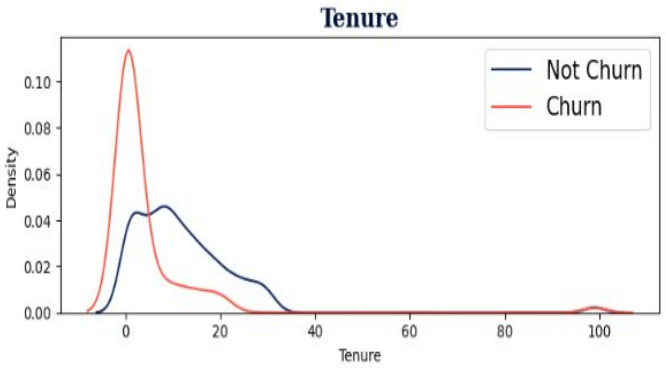
Column	Missing Count	Missing %
Rev_per_month	791	7.02%
Login_device	760	6.75%
Cashback	473	4.20%
Account_user_count	444	3.94%
Day_Since_CC_connect	358	3.18%
Complain_ly	357	3.17%
Tenure	218	1.94%
Marital_Status	212	1.88%
CC_Agent_Score	116	1.03%
City_Tier	112	0.99%
Payment	109	0.97%
Gender	108	0.96%
CC_Contacted_LY	102	0.91%
Service_Score	98	0.87%
Account_segment	97	0.86%
Rev_growth_yoy	3	0.03%
Coupon_used_for_payment	3	0.03%

Column	Outliers
CC_Contacted_LY	80
rev_per_month	113
rev_growth_yoy	52
coupon_used_for_payment	98
Day_Since_CC_connect	94
cashback	226

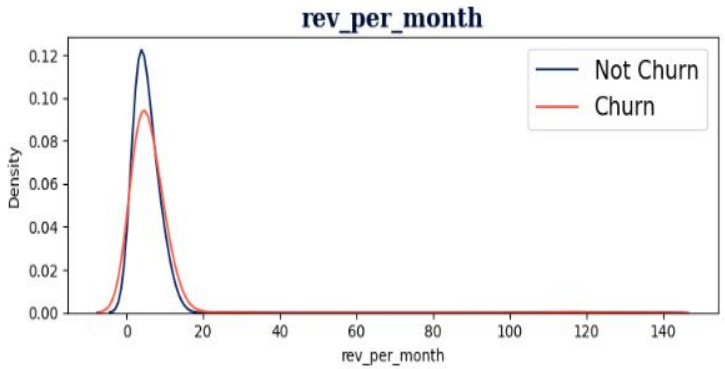


** Missing Data : Categorical treated using Mode and Numerical using KNN imputer*
** Outliers : Lower limit capped at 1st and upper limit at 99th quartile*

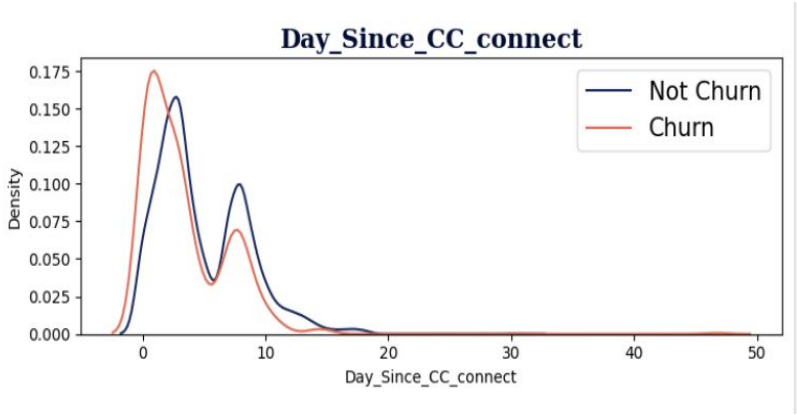
Appendix | EDA Insights



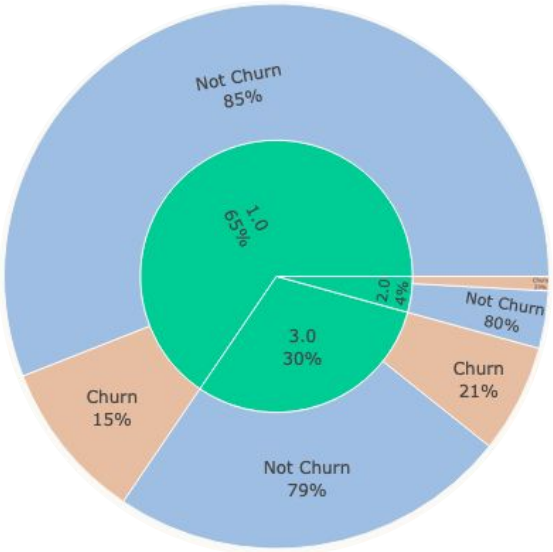
Lower Tenure, Higher Churn



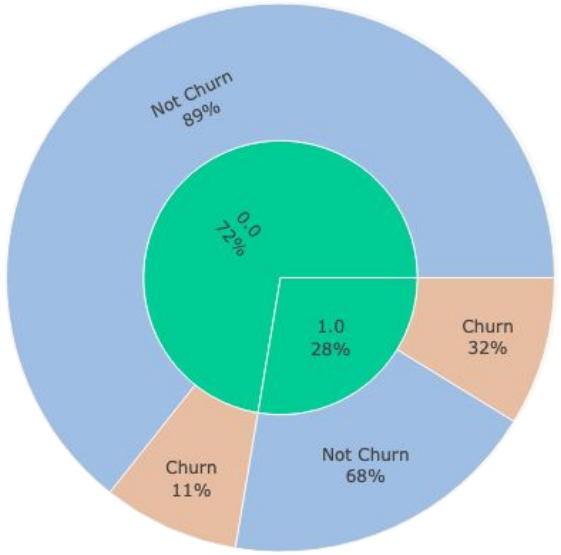
Higher Revenue for non-Churn subscribers



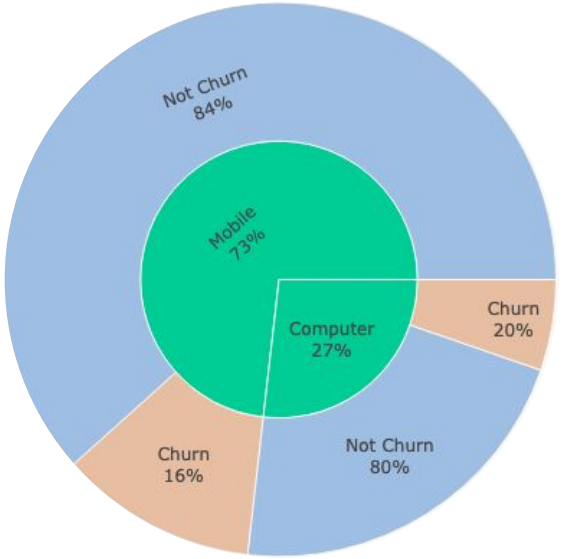
Recent Complaints, higher Churn



City_Tier 3 have a higher churn rate



Complaint in last year has higher churn rate



Computer login has higher churn rate

Appendix | EDA Insights

1. Churn Analysis:

➔ The class imbalance, with only 16.8% churned customers, indicates the need to focus on retaining existing customers to maintain business stability.

2. Demographics and Churn:

➔ Customers from City_Tier 3 have a higher churn rate, suggesting that tailored offerings or incentives may be needed to retain these customers.

➔ Churn rates differ based on payment methods and gender. Efforts should be made to engage male customers more effectively.

3. Service and Customer Care:

➔ The "Service_Score" analysis shows that higher scores don't necessarily lead to lower churn. Service quality may need improvement.

➔ Customers who complained in the last year had a significantly higher churn rate, indicating a need to address their issues promptly.

4. Engagement and Loyalty:

➔ Increasing the number of account users leads to a higher churn rate. Strategies for retaining multi-user accounts should be explored.

➔ Encouraging customers to use mobile devices for logins may reduce churn, as mobile users exhibit a lower attrition rate

Appendix | Modeling Approach(Detailed)

Step 1: Segregate dependent variable "Churn" from the independent variables.

We assigned all predictors to X & response on y variable before split.

Step 2: Split the data into train & Test set

We split the data with 20% of the records going to the test set & 80% to train, random state of 42 was selected.
Shape of Train dataset: (9008, 17)
Shape of Train dataset: (2252, 17)

Step 3: Model(s) Building

Model Selection : In total 14 models were built and their performance was evaluated to find the best fitting model. We built the following models on the dataset & evaluate the performance:

1. Decision Tree
2. Random Forest
3. Linear Discriminant Analysis (LDA)
4. k-Nearest Neighbours(KNN)
5. Naive Bayes

Of these models, the models listed below were ensembled and hyperparameter tuned as well to obtain the optimum model.

1. Decision Tree
2. Linear Discriminant Analysis (LDA)
3. Random Forest
4. Bagging (base- Decision Tree)
5. Adaboost (base- Decision Tree)
6. Extreme Gradient Boost (XGB)
7. Categorical Boost(CatBoost)
8. Neural Network (ANN)
9. SVM

Appendix | Modeling Approach(Detailed)

Step 4: Selection of appropriate Evaluation Matrix

In a churn prediction scenarios wherein the dataset is also imbalanced with 17% Churn cases, the primary goal is to identify potential churners (Churn cases).

Refer to Evaluation Matrix in slide 21

Step 4: Model Tuning

A combination of ensemble methods and hyperparameter tuning can lead to highly accurate and robust machine learning models.

Ensemble Methods:

Ensemble methods involve combining multiple machine learning models to improve overall predictive performance. The key idea is to leverage the strengths of different models and overcome their individual weaknesses

Hyperparameter Tuning:

Grid Search, also known as parameter sweeping, is one of the most basic and traditional methods of hyperparametric optimization. Hyperparameters are parameters that are not learned from the data but need to be set before training a machine learning model. Hyperparameter tuning is the process of finding the best combination of hyperparameters to optimize a model's performance.

Data Imbalance

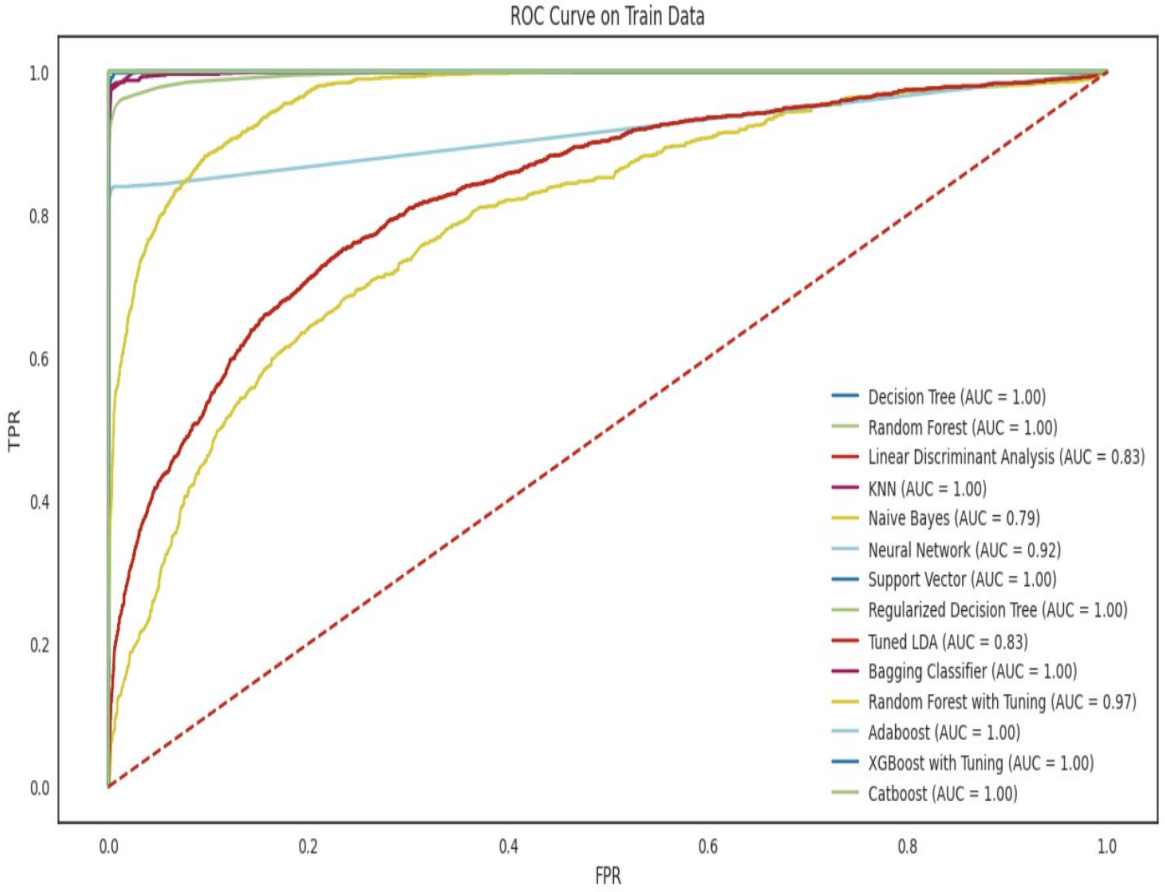
While the dataset is imbalanced we have not oversampled the minority class to avoid synthetic data, since the models were able to perform fairly good without need to introducing oversampled data. However, this is also one way to tune the model for better results when data is imbalanced.

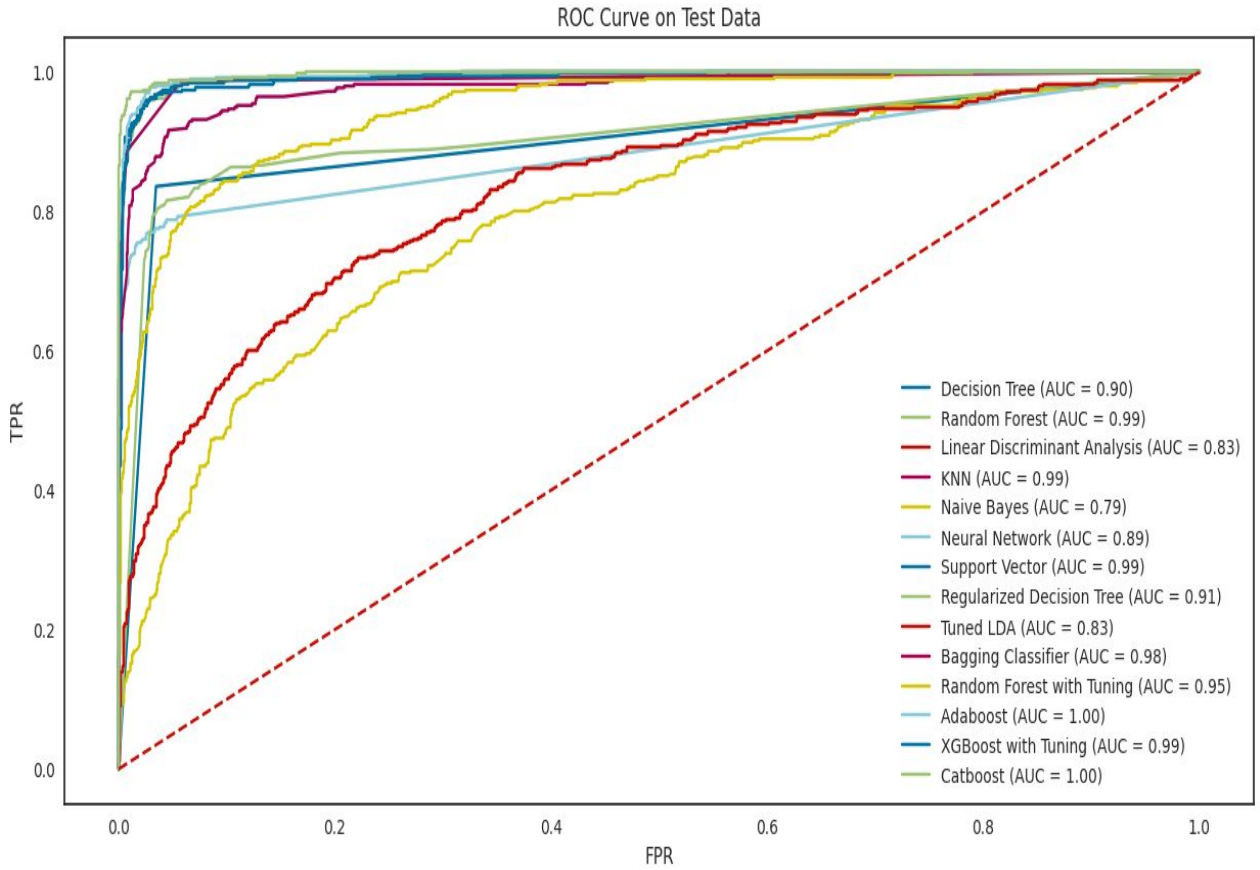
Step 5: Model(s) Comparison & Optimum Model.

Appendix | Model Comparison (Train)

Model	Majority Class(Not Churn)			Minority Class(Churn)			accuracy
	Precision	Recall	f1-score	Precision	Recall	f1-score	
Decision Tree	100%	100%	100%	100%	100%	100%	100%
Random Forest	100%	100%	100%	100%	100%	100%	100%
Adaboost	100%	100%	100%	100%	100%	100%	100%
XGBoost with Tuning	100%	100%	100%	100%	100%	100%	100%
Catboost	100%	100%	100%	100%	100%	100%	100%
Support Vector	99.96%	98.92%	99.44%	94.87%	99.80%	97.27%	99.07%
KNN	99.50%	99.73%	99.61%	98.65%	97.47%	98.05%	99.36%
Bagging Classifier	99.26%	99.92%	99.59%	99.59%	96.27%	97.90%	99.31%
Regularized Decision Tree	98.85%	99.52%	99.18%	97.52%	94.20%	95.83%	98.63%
Random Forest with Tuning	97.45%	90.21%	93.69%	64.29%	88.20%	74.37%	89.88%
Neural Network	96.73%	99.80%	98.24%	98.81%	83.13%	90.30%	97.02%
Linear Discriminant Analysis	87.56%	97.82%	92.41%	73.59%	30.47%	43.09%	86.60%
Tuned LDA	87.56%	97.82%	92.41%	73.59%	30.47%	43.09%	86.60%
Naive Bayes	85.80%	97.54%	91.29%	60.89%	19.20%	29.19%	84.49%

Appendix | Model Comparison ROC –AUC Curve





Appendix | Performance Evaluation Metrics

- **Accuracy**: This term tells us how many right classifications were made out of all the classifications.

- **Precision**: Out of all that were marked as positive, how many are actually truly positive.

- **Recall or Sensitivity**: Out of all the actual real positive cases, how many were identified as positive.

- **F1-Score:** F1 score is a weighted average of Precision and Recall, which means there is equal importance given to FP and FN

  F1 score = 2* (Precision * Recall) / (Precision + Recall)

- **Area Under Curve (AUC) and ROC Curve:** A ROC Curve is drawn by plotting TPR or True Positive Rate or Recall or Sensitivity (which we saw above) in the y-axis against FPR or False Positive Rate in the x-axis. FPR = 1- Specificity
  TPR = TP/ (TP + FN)
  FPR = 1 – TN/ (TN+FP) = FP/ (TN + FP)

	Predicted Positive	Predicted Negative	
Actual Positive	TP *True Positive*	FN *False Negative*	Sensitivity $\frac{TP}{(TP+FN)}$
Actual Negative	FP *False Positive*	TN *True Negative*	Specificity $\frac{TN}{(TN+FP)}$
	Precision $\frac{TP}{(TP+FP)}$	Negative Predictive Value $\frac{TN}{(TN+FN)}$	Accuracy $\frac{TP+TN}{(TP+TN+FP+FN)}$

Key metric for our model