BBA Semester – VI
Capstone Project

Name	**Avikumar Talaviya**
Project	**Customer Churn Prediction Project**
Mentor	**Mr. Milind Desai**
Date of Submission	**9th June, 2024**

A study on "Customer Churn Prediction Using Machine Learning Modeling for Business Recommendations"

Capstone Project submitted to Jain Online (Deemed-to-be University)

In partial fulfillment of the requirements for the award of:

Bachelor of Business Administration

Submitted by:

Avikumar Talaviya Kiritbhai

USN:

211VBBR03319

Under the guidance of:

Mr. Milind Desai

(Faculty-JAIN Online)

Jain Online (Deemed-to-be University)

Bangalore

2023-24

DECLARATION

I, ***Avikumar Talaviya*** hereby declare that the Research Project Report titled ***"Customer Churn Prediction Using Machine Learning for Business Recommendations"*** *has been* prepared by me under the guidance of ***Mr. Milind Desai.*** I declare that this Project work is towards the partial fulfillment of the University Regulations for the award of the degree of Bachelor of Business Administration by Jain University, Bengaluru. I have undergone a project for a period of Eight Weeks. I further declare that this Project is based on the original study undertaken by me and has not been submitted for the award of any degree/diploma from any other University / Institution.

Place: Surat, GJ, IN

Date: 20/04/2024

Talaviya Avikumar Kiritbhai

USN: **211VBBR03319**

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to all those who have helped me complete this research project. I am particularly grateful to my advisor, Mr. Milind Desai, for his invaluable guidance, support, and encouragement throughout the entire process. His expertise and constructive feedback were instrumental in shaping the direction and quality of this research.

I would also like to thank our program managers, Sagari Ma'am and Miss Maria Paul, for their time and insightful suggestions as well as helping throughout the research project with guidelines and support.

Finally, I am grateful to my friends and family for their unwavering support and encouragement throughout this journey.

EXECUTIVE SUMMARY

With an increasing demand for Direct-to-Home (DTH) connections across the rural and urban parts of the country, the market has experienced increasing competition to acquire new customers by providing lucrative offers to DTH customers. To survive in the market the company wants to acquire new customers while retaining existing customers which helps to increase the lifetime revenue of the user. To increase customer retention and reduce churn we want to analyze customer data and purchase history to identify the most probable churners through Exploratory data analysis, Churn modeling, and Evaluation of such models. Customer churn modeling helps to identify a segment of customers who are most likely to churn so that businesses can offer customizable pricing and personalized offers to such users leading to reduced churn and increased retention, We use Machine learning modeling techniques such as Logistics regression, Random Forest, Gradient boosting, and K-Nearest neighbors to model the prediction model and do comparative analysis to identify most accurate model. We aim to provide specific business recommendations for the revenue and marketing teams for personalized marketing campaigns targeting potential churners so that the revenue of the company can be increased while mitigating risk for the business. We also aim to analyze data using data pre-processing, univariate, and bivariate analysis to further understand customer preferences and behaviors that drive customer churn.

Direct-to-Home (DTH) television is a method of receiving satellite television by means of signals transmitted from direct broadcast satellites. India is one of the largest markets of DTH customers around the world. with a large market size, an increased number of competitors would strive to gain maximum market share. As per the industry reports published in 2022, DTH and cable sector together have a subscriber base of 125 million across the country. Out of which 55 million subscribers are DTH connections. Customers

always look for better price of subscription, quality customer service support, variety of TV channels availability while companies strive to retain their existing customers and add new customers to increase revenue as well as profits. This project aims to address customer churn problems by analyzing customers data and developing predictive models for the business and suggest suitable recommendations to reduce the churn for the business.

In this project, I derived attributes like Tenure of customers, cashback offered and Complaint raised against DTH service providers, an important factor for customer churn. Proposed machine learning model developed after extensive descriptive and exploratory analysis archive accuracy of 99% percent on test dataset with random forest model. Research project also explores feature importance for the explanation of machine learning models for interpretability.

TABLE OF CONTENTS

List of Graphs		
Graph No.	**Graph Title**	**Page No.**
2.1	Pie chart of churn ratio	26
2.2	Bar chart of account user count	26
2.3	Heatmap of numerical column correlations with each other	27
3.1	Histogram of Tenure	36
3.2	Histogram of revenue per month	36
3.3	Histogram of coupon used for payment	37
3.4	Bar plot of Payment methods	38
3.5	Bar plot of marital status	39
3.6	Relationship between tenure and churn columns	39
3.7	Relationship between coupon used for payment and churn columns	40
3.8	Distribution of customer care contact by churn	41
3.9	Relationship between customer churn by complains	41
3.10	Relationship between customer churn and marital status	42
3.11	Relationship of customer churn by city tier	43
3.12	Analysis of marital status, tenure and churn	46
3.13	Analysis of CC_contacted_LY, Payment and Churn	49
3.14	Customer churn proportions by cluster labels	50
3.15	Confusion matrix of decision tree classifier	52
3.16	Feature importance of the decision tree classifier	52
3.17	Feature importance of fine tuned decision tree classifier	53
3.18	Confusion matrix of SVM	55
3.19	Confusion matrix of Gaussian naive bayes	56

CHAPTER 1

INTRODUCTION AND BACKGROUND

Introduction and Background

In 1996, DTH services were initially proposed in India, but concerns over national security and cultural influence led to the proposal being rejected. The government banned DTH services in 1997 when Indian Sky Broadcasting (ISkyB), owned by Rupert Murdoch, was on the verge of launching. However, after deliberations among government ministers, DTH services were finally permitted in November 2000 by the NDA government. Four key recommendations were made: preventing a monopoly in DTH services, monitoring vertical integration to prevent TV distribution monopolies, avoiding integration between DTH operators and TV channels for fair competition.

The first DTH service in India was launched on 2nd October 2003 by Dish TV, focusing on rural and underserved areas rather than competing directly with urban cable operators. Dish TV quickly gained 350,000 subscribers within two years. Prasar Bharati introduced DD Free Dish in December 2004, offering free-to-air channels. Tata Play (formerly Tata Sky), a joint venture between the Tata Group and Star India's parent company, was established in 2004 and launched DTH services in April 2007.

The Indian DTH industry has undergone significant evolution, marked by technological advancements, regulatory changes, and shifting consumer preferences. With the advent of high-definition (HD) and ultra-high-definition (UHD) content, DTH operators have continuously upgraded their services to cater to the growing demand for enhanced viewing experiences.

Today, the Indian DTH industry is one of the largest and most competitive in the world with millions of subscribers across tier 1,2, and 3 towns as well as rural areas. Its contribution to the media and entertainment sector has been instrumental in democratizing access to diverse content and driving digital inclusion across the country.

1. Problem Statement

The DTH provider is currently encountering significant competition in the market, posing challenges in retaining existing customers. Retention has become a pressing issue for the company. Consequently, there is a need to develop a model capable of predicting churn among customer accounts and offering segmented incentives to potential churners. By building a machine learning model to predict churn our aim is to identify customers who are most likely to churn. By offering segmented offerings we can enhance customer satisfaction, improvement in customer loyalty and increase retention rates.

2. Objectives of Study

1. To describe the dataset and generate a data report for further analysis.
2. To perform data pre-processing and data cleaning of customer churn dataset.
3. To analyze data using exploratory data analysis (EDA) and data visualization techniques for business insights generation to identify most likely churners for the business and potential inference from the data to improve retention rates.
4. To develop Machine learning models for customer churn prediction for future values for the business that identify and segment potential churners for the business.
5. To evaluate, tune, and interpret machine learning models for explainability for the business stakeholders for sound decision making.
6. To generate business recommendations from churn prediction modeling to optimize marketing campaigns that improves customer retention and customer satisfaction.
7. Using machine learning modeling to reduce customer churn and maximize the retention rates to sustain revenue for the business.

3. Industry overview

The Direct-to-Home industry has many competitors who hold significant market share and strive to increase their customer base as well as revenue through innovation in products and services. With such fierce competition in the market it is important to retain

customer base and reduce churn. Below table gives overview of DTH operators available in the market.

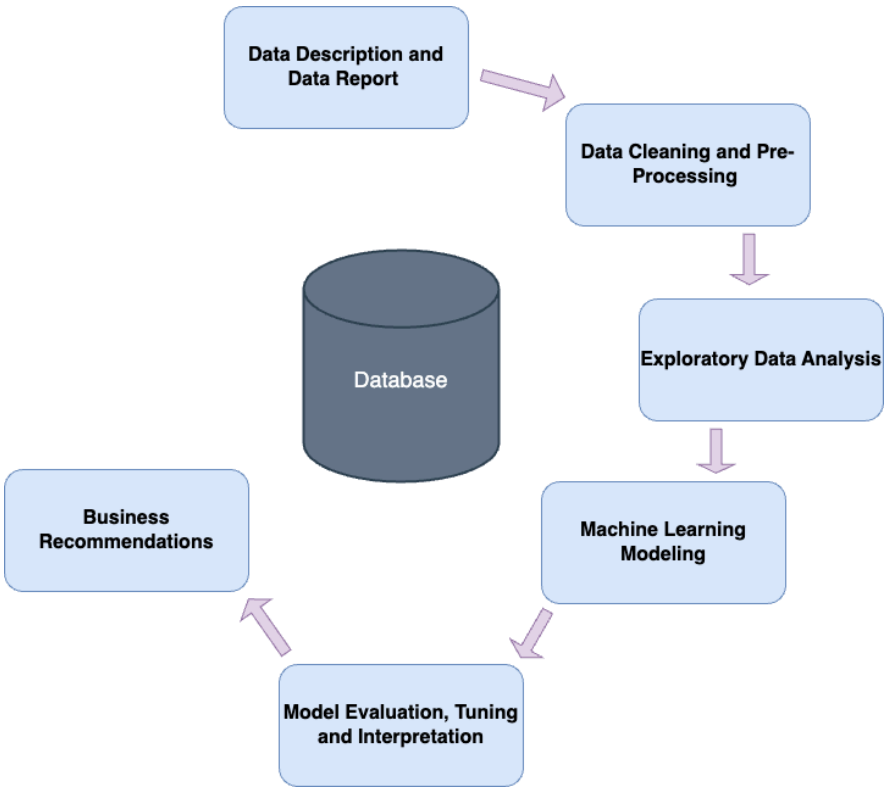
Table No. 1.1: List of DTH service providers in India

S. No.	Service Provider	Launch date	Subscribers	Ownership
1	DD Free Dish	December 2004	40.0 million	Prasar Bharati
2	Tata Play	August 2006	23.44 million	Tata Sons (70%) The Walt Disney Company India (30%)
3	Dish TV d2h Zing Digital	October 2003	18.06 million	Yes Bank (25.63%) Deutsche Bank (6.2%) Jawahar Goel (Promoter & Managing Director) family (5.93%) Housing Development Finance Corporation(4.7 %) IndusInd Bank Ltd.(3.8%)
4	Airtel digital TV	October 2008	17.86 million	Bharti Airtel
5	Sun Direct	December 2007	11.60 million	SUN Group (80%), Astro Group (20%)

Among all market players, DD Free Dish leads with 40.0 million subscribers in India, followed by Tata Play, Dish TV, Airtel digital and Sun direct. The Department of Space (DoS) mandates that all Direct-to-Home (DTH) operators in India utilize satellites authorized by the Indian Space Research Organisation (ISRO). In cases where ISRO satellite capacity is insufficient, DTH operators are permitted to utilize capacity leased by ISRO from foreign satellites.

4. Overview of Theoretical Concepts

This research project aims to conduct in-depth analysis using data analytics methodologies to describe, predict and inference insights and recommendations that can help mitigate customer churn for the DTH companies. Below are some of the techniques we will be using in this project for our analysis and generation of business reports.



1. **Data description and data report:**

   Methods such as a data description, metadata, shape of data and data summarization to prepare a comprehensive data report

2. **Data cleaning methods:**

   Data cleaning methods and techniques include missing value treatment, outlier detection, and removal, and data quality checks as well as removing unwanted variables.

3. **Exploratory data analysis and Data visualization:**

   We will perform univariate, bivariate, and multivariate data analysis to derive insights and recommendations for DTH business to reduce customer churn.

4. **Data pre-processing techniques:**

   Data pre-processing techniques such as data standardization, data normalization, Imbalance data treatment and data encoding would be used to pre-process data for machine learning modeling.

5. **Machine learning modeling:**

   Machine learning modeling techniques such as Logistics regression, Decision Tree, Random Forest, Gradient boosting, and K-nearest neighbors would be used for customer churn modeling.

6. **Model evaluation, tuning, and interpretation:**

   Machine learning modeling evaluation metrics such as an accuracy score, precision, recall, and confusion matrix would be used. Moreover, model tuning and interpretation techniques like feature importance, and hyperparameter tuning would be used.

7. **Business recommendations:**

   Finally, we will derive insights and business recommendations from our modeling to develop personalized marketing campaigns for the businesses.

CHAPTER 2

RESEARCH METHODOLOGY

RESEARCH METHODOLOGY

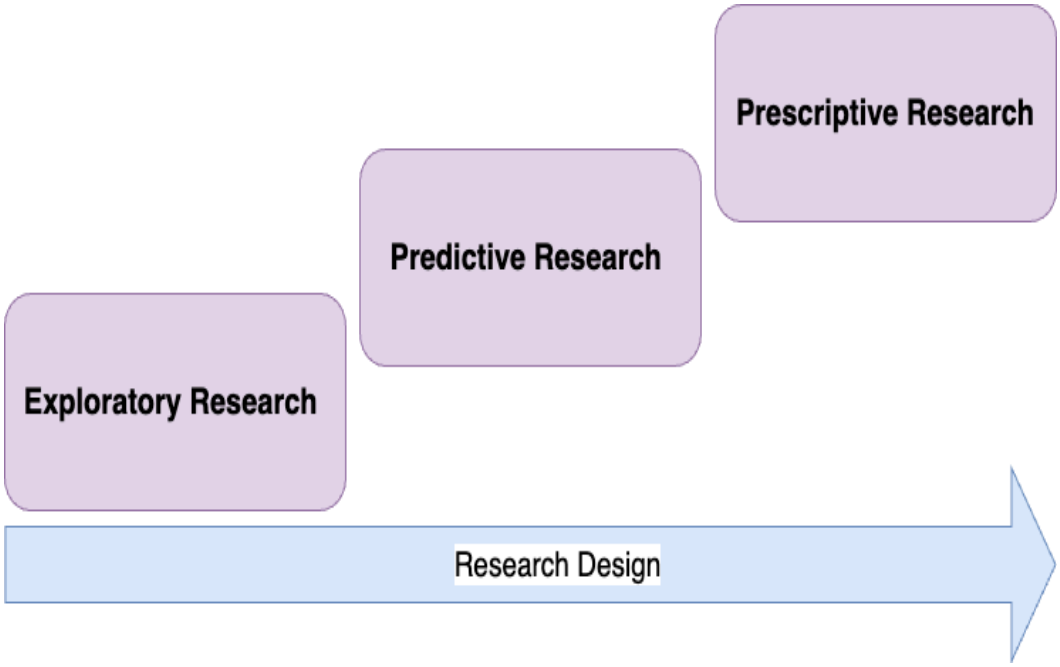
2.1 Scope of the Study

The research project will focus on developing a churn prediction model for the DTH services provider. The business recommendations will be provided on implementing targeted marketing campaigns based on the model's predictions to improve retention rates for the DTH service provider.

2.2 Methodology

1. **Research Design**

   We will be using mainly 3 types of research methodologies for this project. Exploratory research, Predictive research, Prescriptive research for detailed insights and inference generation from dataset. These methodologies will also help to develop machine learning modeling methods to predict potential churners to segment it priorly for customized targeted marketing campaigns with better offers and improvement in services that will eventually lead to increase in revenue for the business. let's look at the all 3 methods in detail:



- **Exploratory research:**

  In this research methodology, we will transform and analyze the dataset using exploratory data analysis techniques as well as data visualizations to understand how various factors impact customer churn for DTH businesses that can help mitigate potential customer loss to competition. thus, enabling an increase in revenue for the business.

- **Predictive research:**

  In this research methodology, we will use data transformation methods to preprocess and prepare dataset for machine learning models to build predictive models for customer churn prediction.

- **Prescriptive research:**

  This research methodology aims to derive business recommendations to reduce customer churn by launching personalized marketing campaigns for potential churners.

2. Data Collection and Data Report

We have data collected from a known DTH company with anonymized information of customers' information. Dataset was collected over a period of 12 months in a DTH company. We will look at the brief data report generated using data analytics tools and techniques. First let's look at the variable descriptions for our business report.

Table No. 2.1: Variable description report

Variable	Description
AccountID	account unique identifier
Churn	account churn flag (Target)
Tenure	Tenure of account
City_Tier	Tier of primary customer's city
CC_Contacted_L12m	How many times all the customers of the account has contacted customer care in 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 service provided by company
Account_user_count	Number of customers tagged with this account
account_segment	Account segmentation on the basis of spend
CC_Agent_Score	Satisfaction score given by customers of the account on customer care service provided by company
Marital_Status	Marital status of the primary customer of the account
rev_per_month	Monthly average revenue generated by account in last 12 months
Complain_l12m	Any complaints has been raised by account in last 12 months
rev_growth_yoy	revenue growth percentage of the account (last 12 months vs last 24 to 13 month)
coupon_used_l12m	How many times customers have used coupons to do the payment in last 12 months
Day_Since_CC_connect	Number of days since no customers in the account has contacted the customer care
cashback_l12m	Monthly average cashback generated by account in last 12 months
Login_device	Preferred login device of the customers in the account

The customer churn dataset has ***11260 samples*** and ***19 features*** including the target feature of the dataset. Let's look at the metadata of the dataset to identify missing values present in the dataset as well as data types of each feature.

Table No. 2.2: Metadata of customer churn dataset

Sr. No.	Feature Name	Non-Null Count	Dtype
1	AccountID	11260	int64
2	Churn	11260	int64
3	Tenure	11158	object
4	City_Tier	11148	float64
5	CC_Contacted_L12m	11158	float64
6	Payment	11151	object
7	Gender	11152	object
8	Service_Score	11162	float64
9	Account_user_count	11148	object
10	account_segment	11163	object
11	CC_Agent_Score	11144	float64
12	Marital_Status	11048	object
13	rev_per_month	11158	object
14	Complain_l12m	10903	float64
15	rev_growth_yoy	11260	object
16	coupon_used_l12m	11260	object
17	Day_Since_CC_connect	10903	object
18	cashback_l12m	10789	object
19	Login_device	11039	object

3. Data Analysis Tools

We will be using a wide range of data analytics tools for analysis and machine learning modeling, evaluation purposes. The following are some of the tools used for the research project.

- **Python** - An object oriented programming language used widely for data analysis, data visualization and machine learning tasks.
- **Numpy** - A powerful library for numerical computing in Python, providing support for large, multi-dimensional arrays and matrices, along with a collection of mathematical functions to operate on these arrays efficiently.
- **Pandas** - A versatile data manipulation and analysis library for Python, offering data structures like DataFrames and Series, along with tools for reading and writing data from various file formats, data alignment, reshaping, and handling missing data.
- **Matplotlib** - A comprehensive plotting library for Python, enabling the creation of a wide variety of static, interactive, and publication-quality visualizations, including line plots, scatter plots, bar charts, histograms, and more.
- **Seaborn** - A statistical data visualization library for Python, built on top of Matplotlib, providing a high-level interface for creating attractive and informative statistical graphics, including complex multi-plot grids, categorical plots, and specialized visualizations for exploring relationships in datasets.
- **Scikit-Learn** - A powerful machine learning library for Python, providing tools for data mining and data analysis, including classification, regression, clustering, dimensionality reduction, model selection, and preprocessing, built on NumPy, SciPy, and Matplotlib.
- **Imblearn** - A Python library for tackling the problem of imbalanced datasets in machine learning, offering a variety of techniques for

resampling, including over-sampling, under-sampling, and a combination of both, to improve the performance of classifiers on minority classes.

- **Jupyter Notebook** - An interactive computing environment for Python, offering a web-based interface for creating and sharing documents containing live code, equations, visualizations, and narrative text, enabling users to explore data, prototype algorithms, and communicate their findings in a seamless manner.

2.3 Data Summary and Primary Findings:

Using analytics tools, we have summarized the dataset to identify a five point summary of each feature as well as other potential characteristics of the dataset that affect the customer churn of a DTH business.

Below table describes the descriptive statistics of numerical data types columns.

Table No. 2.3: Descriptive statistics of numerical columns

	Count	Mean	Std	Min	25%	50%	75%	Max
AccountID	11260.0	25629.500000	3250.626350	20000.0	22814.75	25629.5	28444.25	31259.0
Churn	11260.0	0.168384	0.374223	0.0	0.00	0.0	0.00	1.0
City_Tier	11148.0	1.653929	0.915015	1.0	1.00	1.0	3.00	3.0
CC_Contacted_LY	11158.0	17.867091	8.853269	4.0	11.00	16.0	23.00	132.0

	Count	Mean	Std	Min	25%	50%	75%	Max
Service_Score	11162.0	2.902526	0.725584	0.0	2.00	3.0	3.00	5.0
CC_Agent_Score	11144.0	3.066493	1.379772	1.0	2.00	3.0	4.00	5.0
Complain_ly	10903.0	0.285334	0.451594	0.0	0.00	0.0	1.00	1.0

Now, we will look at the descriptive statistics of objective or categorical columns of the customer churn dataset for our data report for further analysis.

Table No. 2.4: Descriptive statistics of categorical columns

	Count	Unique	Top	Freq
Tenure	11158	38	1	1351
Payment	11151	5	Debit Card	4587
Gender	11152	4	Male	6328
Account_user_count	11148	7	4	4569

account_segment	11163	7	Super	4062
Marital_Status	11048	3	Married	5860
rev_per_month	11158	59	3	1746
rev_growth_yoy	11260	20	14	1524
coupon_used_for_payment	11260	20	1	4373
Day_Since_CC_connect	10903	24	3	1816
cashback	10789.0	5693.0	155.62	10.0
Login_device	11039	3	Mobile	7482

Following analysis based on descriptive statistics of the dataset we can notice some initial findings that can be crucial for the further exploratory and predictive research of the customer churn.

1. Only 4 features do not have any missing values, the rest of the features have many samples with missing values. We need to tackle this during data cleaning.
2. Most used payment method is debit card, while most accounts are people with married status.
3. Over 83% of samples are the ones that didn't churn while around 16% of samples are cases of customer churn, so this is a case of imbalance class classification. Project research also aims to reduce such churn to retain maximum customers which ultimately leads to increased revenue per user.
4. On average, customers contacted the customer service center 17 times in a year with a maximum number of customer service calls reaching 132 in a year.
5. The dataset contains customers from tier 1, tier 2, and tier 3 cities.
6. There are five unique types of payment methods used by the customers.
7. The average customer service satisfaction score is higher than the company service satisfaction score.
8. Most used login device is mobile phone and the most recurring cashback amount is 155.62 rupees.
9. It is also observed that many categorical columns are having numerical values which need to be corrected during data cleaning.
10. Most customers are from tier 1 city followed by tier 3 and 2 respectively with as much as 7263 customers from tier 1 white 3405 from tier 3 and 480 from tier 2 city.
11. Customers account segments based on their spending on various package of subscriptions before cleaning of the dataset are as follows:

Table No. 2.5: Count of customers based on account segment

Account segment	

Super	4062
Regular Plus	3862
HNI	1639
Super Plus	771
Regular	520
Regular +	262
Super +	47

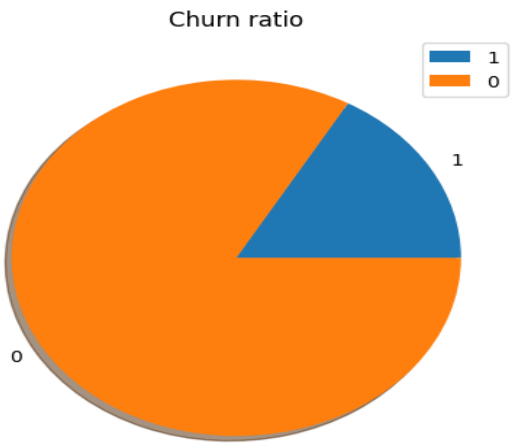
12. Analyzing data summary it is noticed that columns like Login_device, Account_user_count, Tenure, Rev_per_month, cashback, etc have many missing values as well as corrupted data values that will need to be replaced with correct values during data cleaning tasks.
13. It is found that customer service agent scores were a bit higher in tier 1 cities than tier 2 or 3 cities. Below are the feature interaction columns with interaction between marital status, city tire and customer service agent score.

Table No. 2.6: Feature interaction between marital status, city tier and CC agent score

City_Tier	**1.0**	**2.0**	**3.0**
Marital_Status			
Divorced	3.228324	3.742424	3.236190
Married	3.051053	3.278146	2.960414
Single	3.035230	2.755102	3.038767

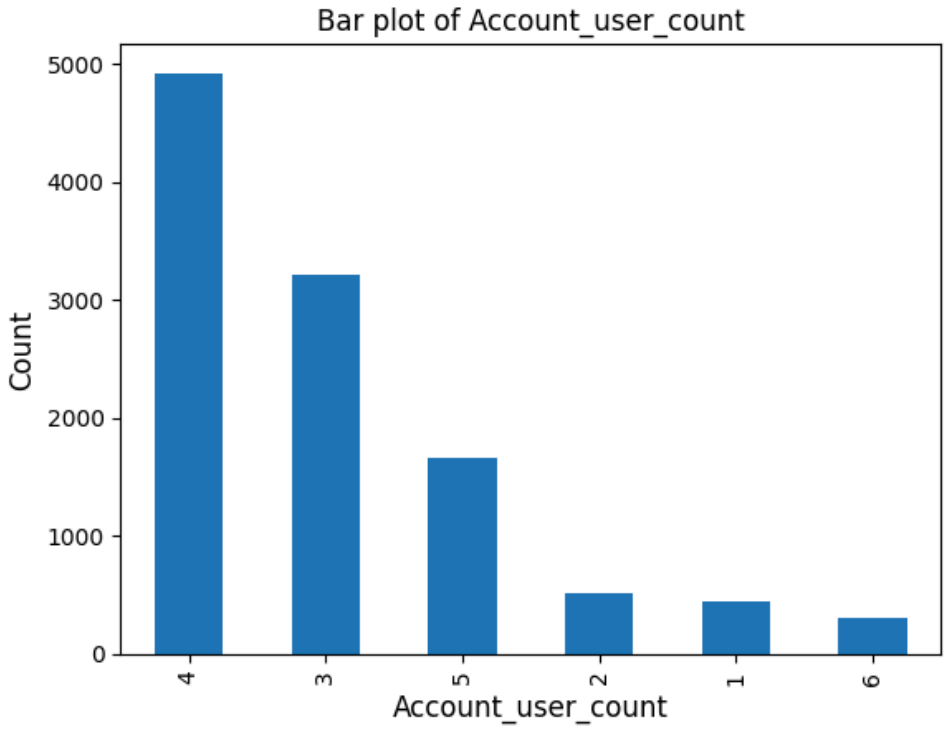
14. We noticed approx 83% of customers are not from customer churn while only 16.8% samples are that of customer churn. Below is the visual representation of the ratio of churn or not churn customers with **'1'** representing **customer churn** while **'0'** representing **customer not churning**.

Graph No. 2.1: Pie chart of churn ratio



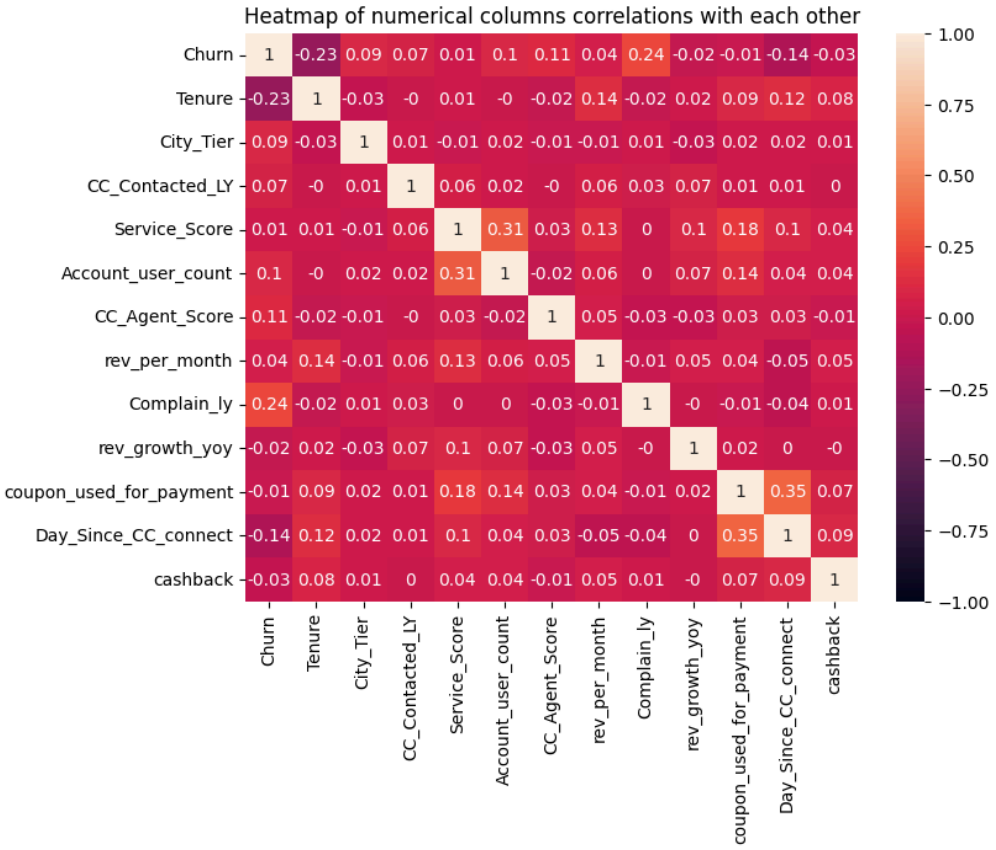
15. We also noticed that most accounts contain 4 users followed by 3 and 5 users per account. Below is the graph that represents how many users are there per account in our customer churn dataset.

Graph No. 2.2: Bar chart of account user count



16. Another dataset insight to notice using numeric columns in our featureset. Using seaborn's heatmap we can measure correlations between various numerical columns to identify most correlated as well as effective features. Below chart shows **'Compalin_ly'** feature is most correlated with target feature among all variables. Such analysis gives us many business insights that can be used to mitigate customer churn.

Graph No. 2.3: Heatmap of numerical column correlations with each other



17. Next, the chart below tells the Marital status proportion in the DTH customer dataset.

Table No. 2.7: Count of customers based on their Marital Status

Marital status	
Married	5860
Single	3520
Divorced	1668

2.4 Approach to EDA and Pre-Processing

With the relevant insights from descriptive statistics of the dataset and data report, we will now proceed for further data cleaning and preprocessing followed by exploratory data analysis to identify potential features and interactions that affect the customer churn for the DTH business. By identifying such factors we can further transform and develop machine learning models to flag potential churners for targeted marketing campaigns that reduce churn and increase revenue for the business. Below are the steps taken to pre-process the dataset and carry out an exploratory data analysis.

1. We will need to treat missing values from the dataset using the pandas library of python.
2. Once the missing values are treated, we need to replace unwanted corrupted values from features with median or mode value of respective features.
3. There are many features that contain numeric values but have **'Object'** data type. We will change categorical data types with numerical data types.
4. Remove redundant features from the dataset to prepare the dataset for machine learning modeling.
5. Outlier treatment for **'rev_per_month'** feature creates a robust dataset.
6. Once the dataset is clean, we will explore the dataset to find business insights using univariate, bivariate and multivariate data analysis.
7. In EDA, we will use data visualizations techniques using pandas, matplotlib and seaborn library to find out in-depth insights from the dataset

2.5 Approach to data transformations and machine learning modeling

With EDA we will get business insights and dataset characteristics that can be used to transform the dataset into a final training dataset. The training dataset will be used for machine learning modeling and tuning to develop accurate models that can be leveraged

for business improvement by reducing customer churn. Below are the steps taken for data transformations and machine learning modeling.

1. Encode categorical features with numerical encodings using the pandas library of python.
2. Standardize the dataset using the **'StandardScaler()'** method.
3. Using K-means clustering to segregate the customer churn dataset to find out the most affected segment for customer churn.
4. Create a new feature of average customer care score using customer service score and customer care service score.
5. Balance the dataset using **'SMOTE'** technique to reduce bias during model development.
6. Once the pre-processed dataset is prepared, we can use sci-kit learn's train and test split method to split the dataset for machine learning modeling.
7. Develop the baseline DecisionTree and k-NN models using train dataset followed by evaluation of accuracy on test dataset.
8. Once the baseline models are built, we will build Random forest classifier, Decision tree classifier, Gaussian Naive Bayes, Support vector machine, Gradient Boosting Classifier, and XGBoost classifier improvements in model accuracy and performance of machine learning models.
9. Compare models and tune the hyper parameters to improve accuracy of the models. Use feature importance methods to identify most impactful features that will determine customer churn possibility and help DTH business to reduce the customer churn.

2.6 Utility of Research outcomes

This research project aims to generate potential outcomes that result in a wide range of business implications. The following are some of the outcomes that this project will generate.

1. Enhanced customer retention rates for the DTH business.
2. Potential insights into targeted marketing campaigns for potential churners.
3. Improved customer satisfaction and loyalty towards the company.
4. Increased profitability through reduced churn and targeted marketing efforts.

CHAPTER 3

DATA ANALYSIS AND INTERPRETATION

DATA ANALYSIS AND INTERPRETATION

3.1 Exploratory Data Analysis and Business Insights from EDA

Using python's pandas library methods such as a **'fillna()'**, **'replace()'**, **'astype()'**, and **'quantile()'** to clean and remove outliers from the dataset we have dataset with **11075 samples** and **18 columns** including target column. Detailed code for data cleaning methods are available in annexure listed at bottom of this report. Below are the descriptive statistics of the clean dataset.

Table No. 3.1 Descriptive statistics of numerical columns of clean dataset

	count	mean	std	min	25%	50%	75%	max
Churn	11075.0	0.167223	0.373192	0.0	0.00	0.00	0.00	1.0
Tenure	11075.0	10.897336	12.835648	0.0	2.00	8.00	16.00	99.0
City_Tier	11075.0	1.657517	0.916308	1.0	1.00	1.00	3.00	3.0
CC_Contacted_LY	11075.0	17.839729	8.809774	4.0	11.00	16.00	23.00	132.0
Service_Score	11075.0	2.901761	0.722276	0.0	2.00	3.00	3.00	5.0
Account_user_co	11075.0	3.70320	1.0051	1.0	3.00	4.00	4.00	6.0

unt		5	09					
CC_Agent_Score	11075.0	3.063386	1.373249	1.0	2.00	3.00	4.00	5.0
rev_per_month	11075.0	4.987720	2.771900	1.0	3.00	4.00	7.00	13.0
Complain_ly	11075.0	0.275576	0.446824	0.0	0.00	0.00	1.00	1.0
rev_growth_yoy	11075.0	16.180497	3.753587	4.0	13.00	15.00	19.00	28.0
coupon_used_for_payment	11075.0	1.785914	1.970878	0.0	1.00	1.00	2.00	16.0
Day_Since_CC_connect	11075.0	4.579233	3.648118	0.0	2.00	3.00	7.00	47.0
cashback	11075.0	194.960662	175.583900	0.0	147.85	165.24	197.25	1997.0

Above table shows we now have 11075 samples as compared to previous 12060 samples. Above table shows we have an average tenure of 10 months for our customers while

average cashback stands at 195 Rupees. In that manner, We can observe the five point summary of each feature from the above table.

Now, let's look at the summary statistics of categorical columns.

Table No. 3.2: Descriptive statistics of categorical columns of clean dataset

	Count	Unique	Top	freq
Payment	11075	5	Debit Card	4618
Gender	11075	2	Male	6667
account_segment	11075	5	Super	4090
Marital_Status	11075	3	Married	5885
Login_device	11075	3	Mobile	7568

Once we have information about clean dataset statistics, we now want to learn more about categorical values each feature contains. This helps to know how many values are there per category in our dataset. Below are the values per category in each column.

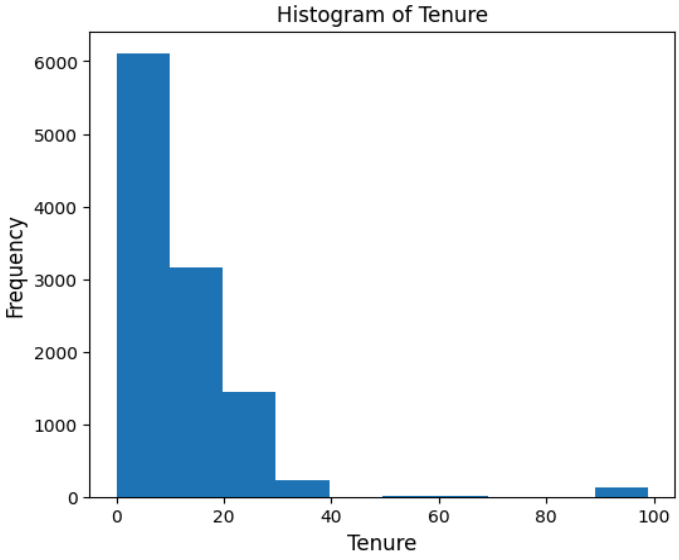
- Value counts for each categories of *Gender*: {'Male': 6667, 'Female': 4408}
- Value counts for each categories of *Payment*: {'Debit Card': 4618, 'Credit Card': 3446, 'E wallet': 1210, 'Cash on Delivery': 997, 'UPI': 804}

- Value counts for each categories of *account_segment*: {'Super': 4159, 'Regular Plus': 4142, 'HNI': 1639, 'Super Plus': 818, 'Regular': 520}
- Value counts for each categories of *Marital_Status*: {'Married': 5885, 'Single': 3523, 'Divorced': 1667}
- Value counts for each categories of *Login_device*: {'Mobile': 7568, 'Computer': 2968, 'Mobile & Computer': 539}

Univariate Analysis

1. Histogram of Tenure:

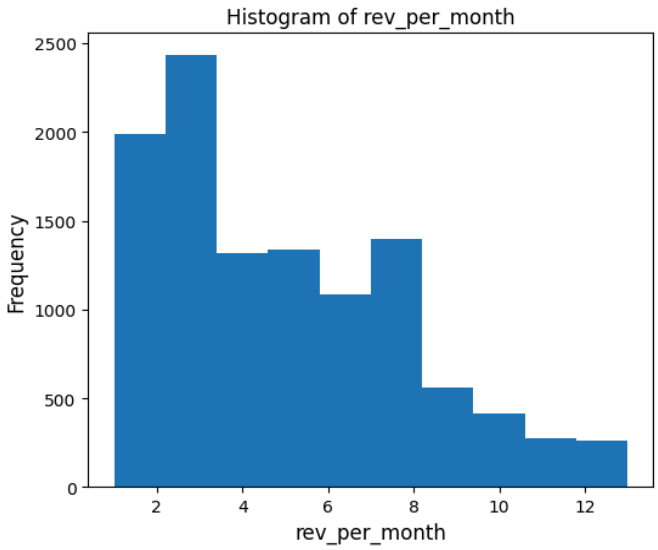
Graph No. 3.1: Histogram of Tenure



The histogram of tenure illustrates the frequency distribution of tenure lengths. It reveals that a significant number of observations have shorter tenures (between 0 and 10), with frequencies decreasing as tenure increases. The inference is that customer retention strategies should focus on early tenure stages to prevent churn and enhance loyalty.

2. Histogram of rev_per_month:

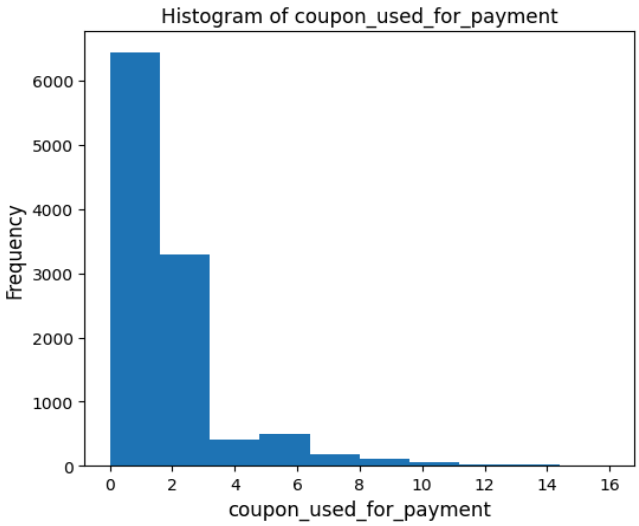
Graph No. 3.2: Histogram of revenue per month



The histogram of rev_per_month displays the frequency distribution of a metric called "rev_per_month." The highest frequency occurs in the range of 0 to 2, suggesting that this range is the most common for the given metric. Businesses can infer that a significant proportion of observations have relatively low revenue per month, which may impact strategic decisions related to pricing, customer acquisition, or retention efforts.

3. Histogram of coupon_used_for_payment:

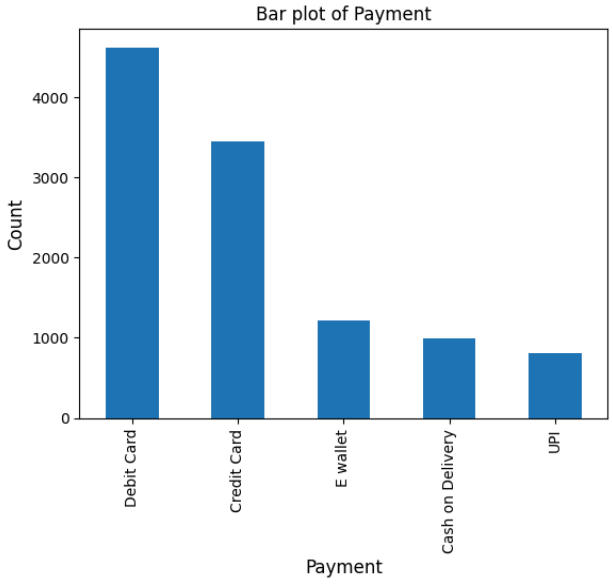
Graph No. 3.3: Histogram of coupon used for payment



The histogram of coupon_used_for_payment displays the frequency distribution of a metric called number of coupons used for payment. The highest frequency occurs in the range of 0 to 2, suggesting that this range is the most common for the given metric.

4. Bar plot of Payment column:

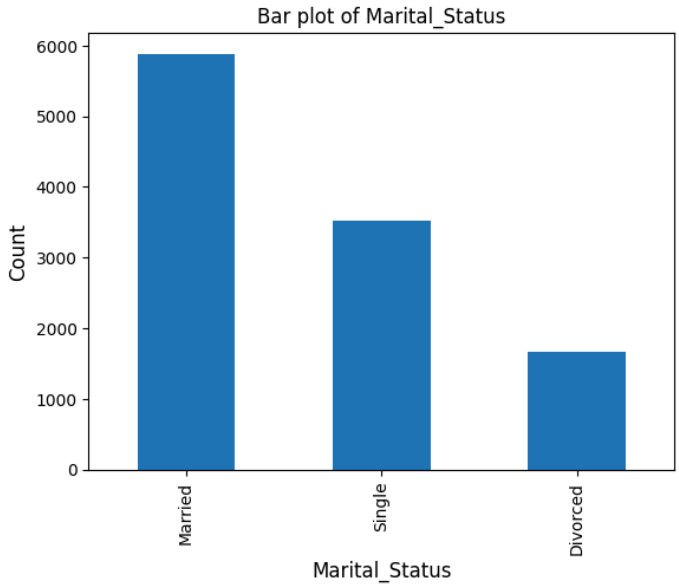
Graph No. 3.4: Bar plot of Payment methods



Debit Card is the most frequently used payment method, with a significantly higher count compared to others. Credit Card follows as the second most used method. E-Wallet, Cash on Delivery, and UPI have much lower counts in comparison.

5. Bar plot of Marrital_Status:

Graph No. 3.5: Bar plot of marital status

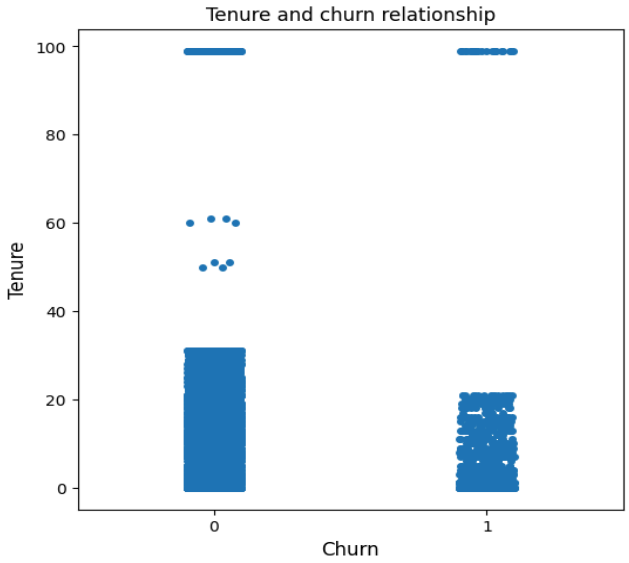


Married is the most common status among the individuals represented, with a significantly higher count. Single follows as the second most common status. Divorce is the least common status in customer churn dataset.

Bivariate Analysis

1. Analysis of Tenure and Churn features

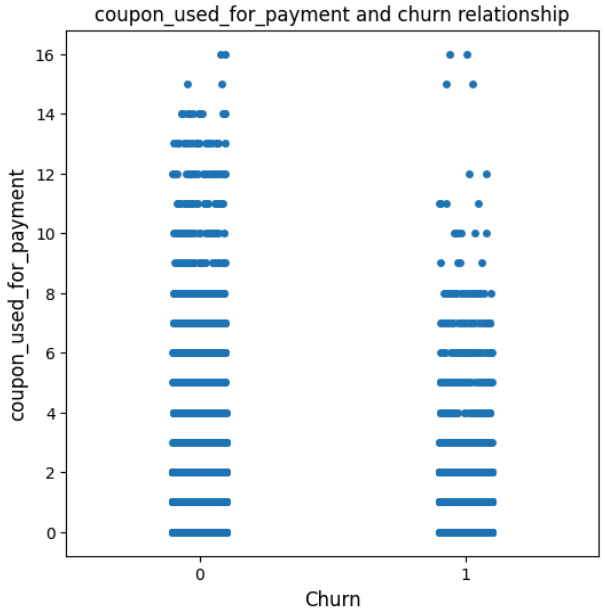
Graph No. 3.6: Relationship between tenure and churn columns



Customers with higher tenure (closer to 100) are less likely to churn (Churn value 0). Lower tenure customers (spread across lower values) have a higher likelihood of churning (Churn value 1).

2. Analysis of coupon used for payment and Churn column

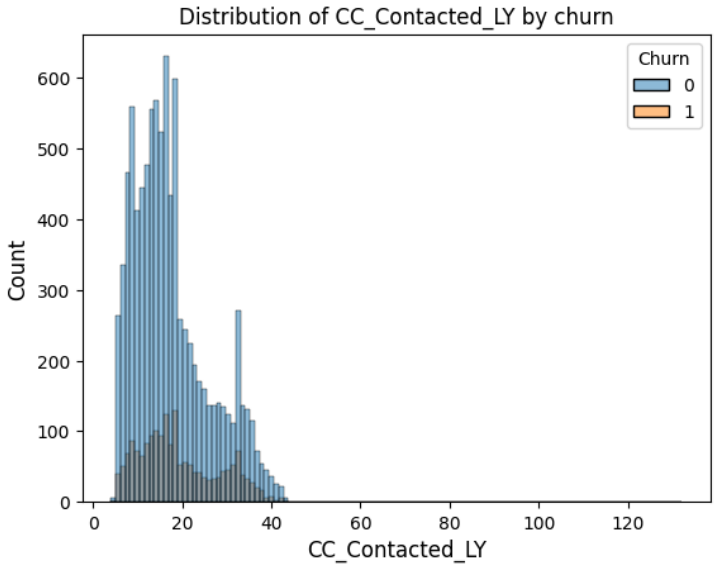
Graph No. 3.7: Relationship between coupon used for payment and churn columns



As the number of coupons used for payment increases, there seems to be no clear trend in the impact on customer churn. but it can be said that there's a slight trend that if less coupons are used than customers are likely to churn. The scatter of points suggests that coupon usage alone may not strongly influence churn behavior.

3. Distribution of customer care contacted in last year by Churn column

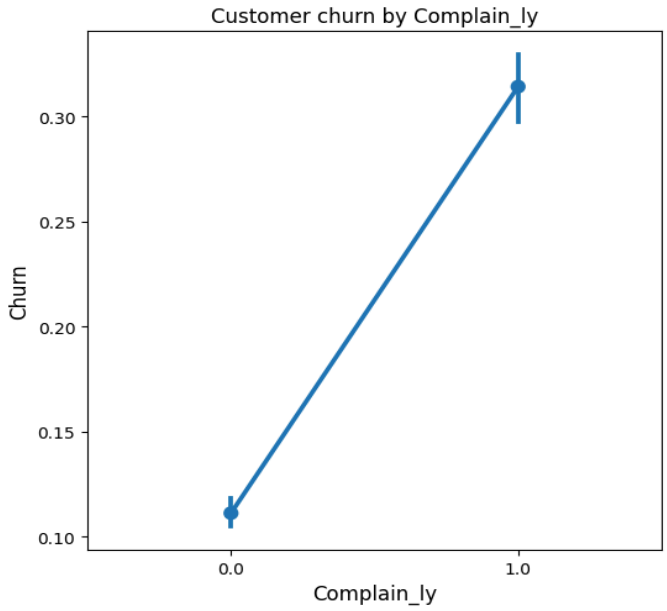
Graph No. 3.8: Distribution of customer care contact by churn



Most data points cluster at the lower end of the x-axis, indicating that many customers had fewer customer care interactions in the last year. The higher frequency of no churn (Churn 0) suggests that customers with more customer care contacts are less likely to churn. However, there doesn't seem to be a clear trend or pattern based solely on the distribution of data points.

4. Analysis of customer complain by Churn column

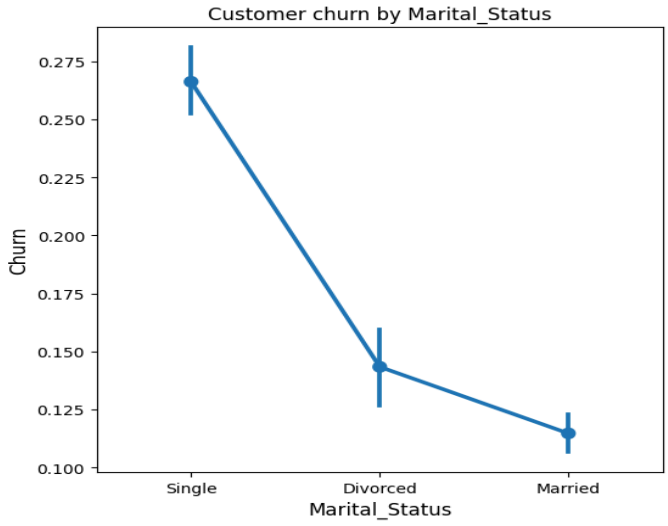
Graph No. 3.9: Relationship between customer churn by complains



The upward trend of the blue line suggests a positive correlation between the variable 'Complain_ly' and customer churn. As the value of 'Complain_ly' increases, the churn rate also tends to increase. In other words, when customers raise more complaints (higher 'Complain_ly'), there is a higher likelihood of them stopping or churning from using the company's services.

5. Analysis of Marital status by Churn column

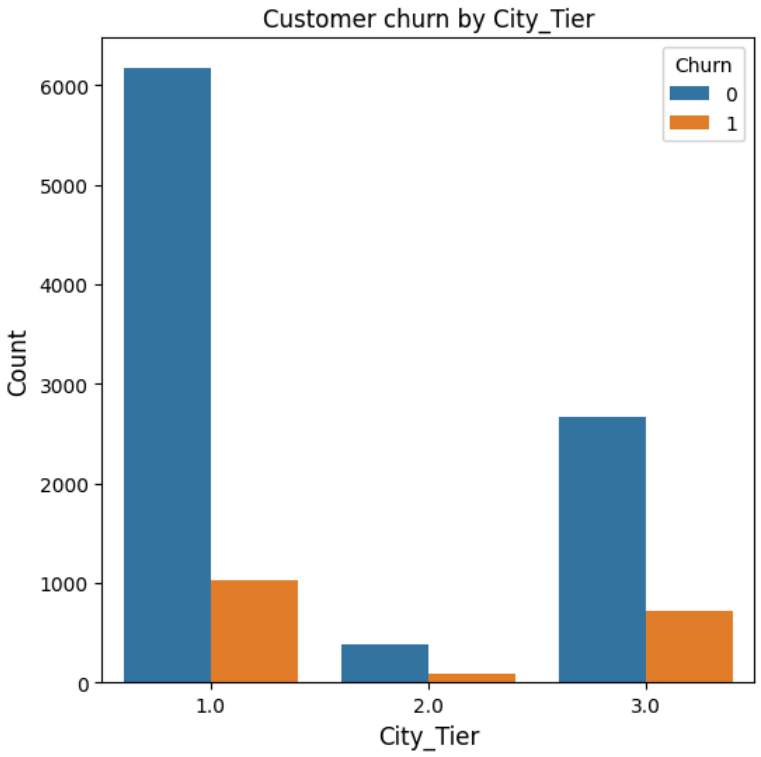
Graph No. 3.10: Relationship between customer churn and marital status



The trend shown by the line indicates that Singles have the highest churn rate (around 0.28). Divorced individuals have an intermediate churn rate (around 0.20). Married individuals have the lowest churn rate (around 0.12). Marital status appears to significantly influence customer churn behavior. Companies should consider tailoring retention strategies based on marital status to reduce churn. For example, personalized communication or loyalty programs might be more effective for singles compared to married customers.

6. Analysis of customer churn by city tier

Graph No. 3.11: Relationship of customer churn by city tier



City Tier 1 has the highest overall customer count (over 6000). Despite having the most customers, City Tier 1 also has a higher retention rate (more no churn) compared to Tiers 2 and 3. City Tier 2 and City Tier 3 have lower customer counts, but the proportion of churn to no churn is more balanced. Companies operating in City Tier 1 may be doing well in retaining customers, while those in

Tiers 2 and 3 need to focus on churn reduction strategies. Thus, DTH businesses operating Tier 2 and 3 need more focus on retaining customers.

7. Analysis of churn by Login device

Table No. 3.3: Churn proportion by login device

	0 (Not churn)	1 (Churn)
Login Device		
Mobile	0.84342	0.15658
Computer	0.803908	0.196092
Mobile & Computer	0.842301	0.157699

Login device category ***'Computer'*** shows Higher proportion of customer churn than other device categories. To reduce churn DTH business can consider marketing on Mobile devices rather than computers.

8. Analysis of churn by Account Segment

Table No. 3.4: Churn proportion by account segment

	0 (Not churn)	1 (Churn)
Account segment		
Super	0.894866	0.105134
Regular Plus	0.728493	0.271507

Regular	0.922179	0.077821
HNI	0.847785	0.152215
Super Plus	0.949868	0.050132

Account segment **'Regular Plus'** has a higher customer churn proportion than any other account segment for a DTH company. Due to high churn in such account segments it would be expected that companies reconsider such packages or entirely remove segments to offer more profitable offers to the customers.

9. Analysis of churn by Payment methods used

Table No. 3.5: Churn proportion by Payment mode

	0 (Not churn)	1 (Churn)
Payment		
Debit card	0.84712	0.15288
UPI	0.825871	0.174129
Credit card	0.858677	0.141323
Cash on delivery	0.755266	0.244734
E Wallet	0.772727	0.227273

Payment modes like **'Cash on delivery'** and **'E Wallet'** are having higher churn proportions compared to other payment modes. DTH companies should look at the potential causes of such customer churn. while it doesn't directly signify any correlation but it should be further investigated.

10. Analysis of churn by Gender

Table No. 3.6: Churn proportion by Gender

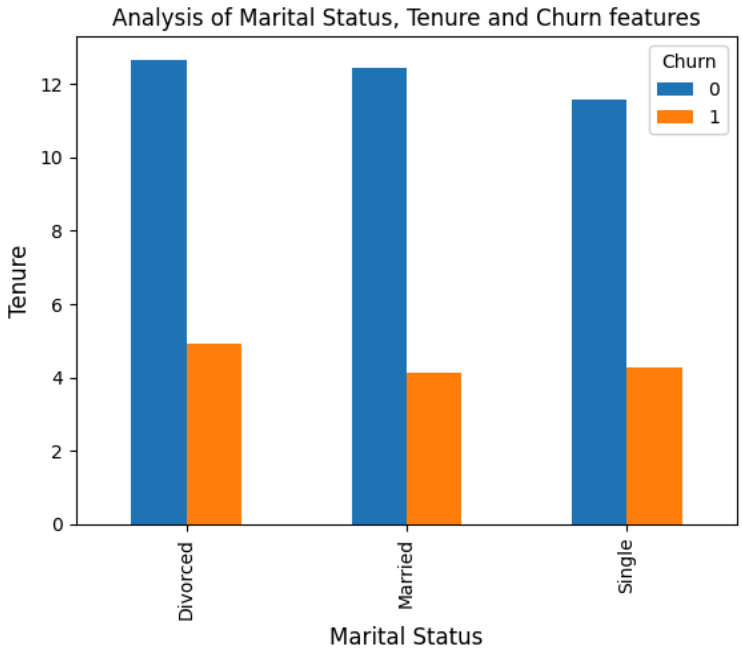
	0 (Not churn)	1 (Churn)
Gender		
Male	0.845508	0.154492
Female	0.824359	0.175641

Female customers have a higher churn proportion than Male customers. companies should provide higher discounts and special offers to Female customers to reduce churn.

Multivariate Analysis

1. Analysis of Marital Status, Tenure, and Churn features

Graph No. 3.12: Analysis of marital status, tenure and churn



Single individuals have the highest churn rate, as indicated by the shorter orange bar. Divorced individuals have an intermediate churn rate. Married individuals have the lowest churn rate and longer tenure. Marital status appears to significantly influence customer churn behavior while lower Tenure of customers significantly contributes to churn. such feature interactions should be considered for machine learning modeling and customer segmentation.

2. Analysis of cashback, marital status, and churn

Table No. 3.7: Analysis of cashback, marital status and churn

	0 (Not churn)	1 (Churn)
Marital Status		
Divorced	204.169188	184.889707
Married	198.786077	187.312770

Single	190.770151	179.312015

It appears that cashback values differ based on both marital status and churn status. Married individuals tend to have slightly higher cashback values, especially among those who have churned. Single individuals have the lowest cashback values, both for churned and non-churned cases. Overall those who churn have lower cashback than those who do not churn.

3. Analysis of Tenure, Login device, and churn

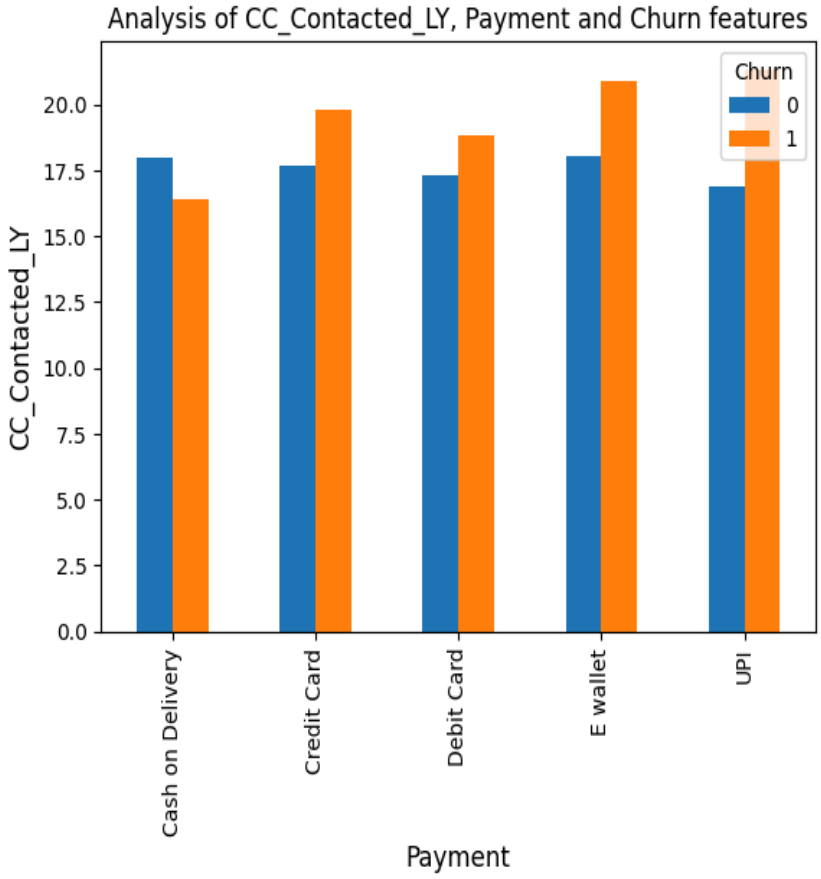
Table No. 3.8: Analysis of tenure, login device and churn

	0 (Not churn)	**1 (Churn)**
Login_device		
Computer	12.208718	3.864261
Mobile	12.243929	4.578059
Mobile & Computer	11.951542	3.588235

Above table clearly shows Tenure of churners is very low compared to those who do not churn across the Login devices of the customers. It's especially low where customers use computers to sign up for the DTH services. companies should focus more on these segments to improve retention rates and reduce churn.

4. Analysis of CC_Contacted_LY, Payment and Churn features

Graph No 3.13: Analysis of CC_contacted_LY, Payment and Churn



The chart illustrates the relationship between the number of times customers contacted the call center last year (CC_Contacted_LY), different payment methods, and their churn status (whether they stopped using the service).

For all payment methods, churned customers (orange bars) tend to have a higher number of call center contacts compared to non-churned customers (blue bars).

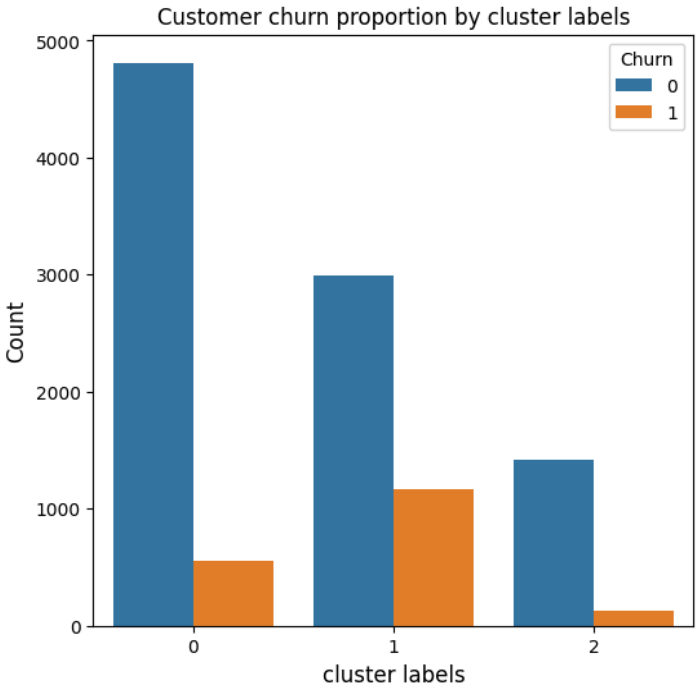
The difference in contact frequency between churned and non-churned customers is most noticeable for Credit Card and E-wallet users.

3.2 Model Development, Evaluation, Tuning and Interpretation.

Given the insights we gathered from the exploratory data analysis and data description analysis, next we need to transform and engineer the customer churn dataset to develop machine learning modeling for our business use case. Machine learning modeling helps predicting target variables using algorithms without explicit coding. In our case, we want to predict customer churn for the DTH business. let's look at the data transformation steps:

1. Encode categorical features with numerical encodings using the pandas library of python.
2. Standardize the dataset using the **'StandardScaler()'** method from sci-kit learn library of python.
3. Using K-means clustering to segregate the customer churn dataset to find out the most affected segment for customer churn. Below is the clustering segmentation with respect to customer churn proportions.

Graph No. 3.14: Customer churn proportions by cluster labels



- **Cluster Label 0**:
    - Has the **highest count** of non-churning customers.
    - The count of churning customers is **very low**.
- **Cluster Label 1**:
    - Has a **moderate count** of both non-churning and churning customers.
    - Churning customers are **slightly higher** in number relative to other clustering labels 0 and 2.
- **Cluster Label 2**:
    - Has for both churning and non-churning customers.
    - Non-churning customers are **more** than churning customers.

4. Create a new feature of '**Average Customer Care Score'** using customer service score and customer care service score. Once new features are created drop old features for machine learning modeling.
5. Our customer churn dataset is an imbalance class classification problem. For the ML modeling we have to balance the dataset using **'SMOTE'** technique to reduce

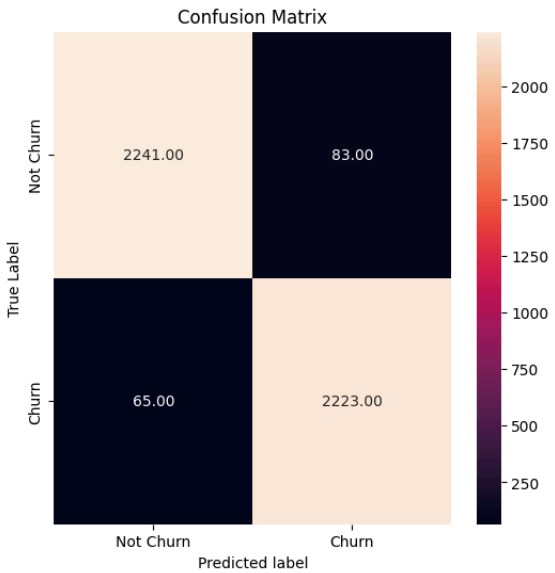
bias during model development. once the data balance is performed, our dataset has now **9223 samples of each target categories** (i.e. '0' (Not churn) and '1' (Churn))

Machine Learning Model Development

1. Decision Tree Classifier:

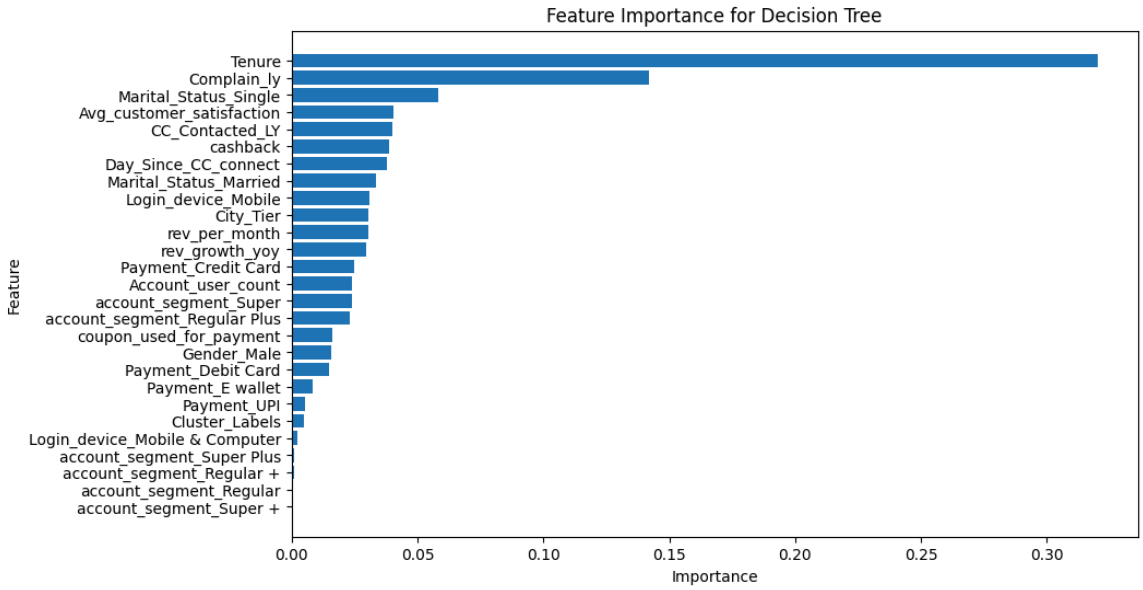
   Decision tree classifier with hyper parameters ***criterion='entropy', random_state=42,max_depth=20*** gives **accuracy of 96.7%** along with **precision and recall of 96% and 97%** respectively on the test dataset. Below is the plot of the confusion matrix of the decision tree classifier.

Graph No. 3.15: Confusion matrix of decision tree classifier



Feature Importance for Model Interpretation

Graph No. 3.16: Feature importance of the decision tree classifier

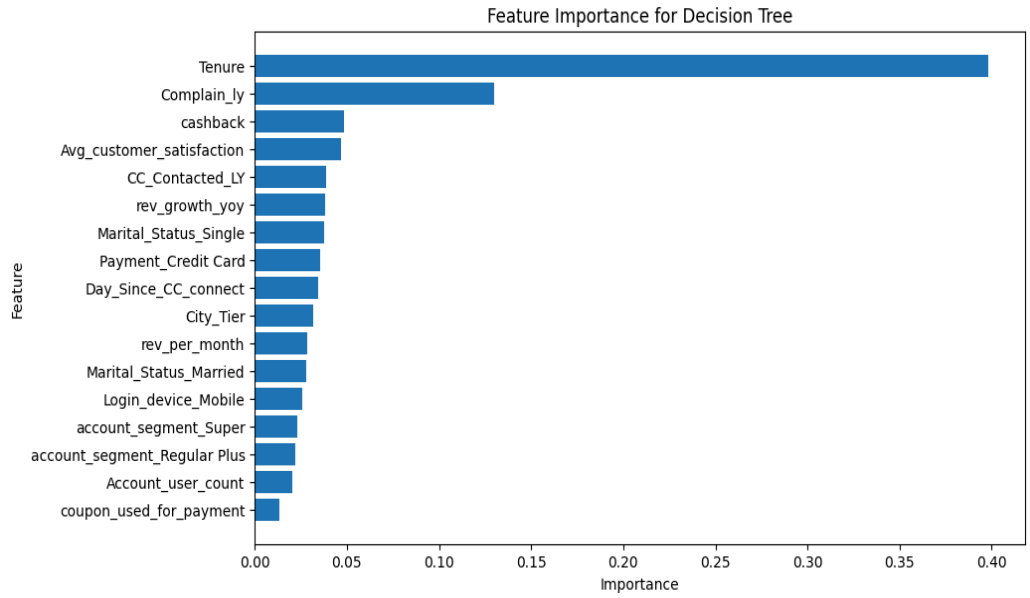


Above chart indicates highly influential features that the model utilizes to predict the target variable. Features like Tenure, Complain_ly. Marital_status_Single, Avg_customer_satisfaction, CC_Contacted_LY, cashback, etc are some of the features that contribute most to the prediction of customer churn. Above graph also helps to remove less important features from machine learning modeling to build less complex models. Feature importance can help identify most influential features in the featureset that impacts the target variable which is Churn in our case.

2. Fined tuned Decision Tree Classifier:

Decision tree classifier with tuned hyper parameters **`criterion='gini', random_state=42, max_depth=15, min_samples_split=2, max_features=15`** with less features gives **accuracy of 96%** while **precision and recall of 95.5% and 96.4%** respectively on test dataset. We can plot the feature importance chart to find most useful features to predict churn predictions.

Graph No. 3.17: Feature importance of Fine-tuned Decision Tree Model



Tenure is the most significant feature, with an importance score approaching 0.40. This suggests that the duration a customer has been with the service is a critical factor in the model's predictions. *Complain_ly* (likely representing the number of complaints last year) is the second most important feature, indicating customer dissatisfaction plays a significant role. *Cashback* and *Avg_customer_satisfaction* are also important, though to a lesser extent, suggesting financial incentives and overall satisfaction are relevant but not as dominant as Tenure and complaints.

3. k-NN Modeling:

   k-nearest neighbors classifier with hyper parameter **`n_neighbors=3`** gives **accuracy of 98%** while **precision and recall of 96% and 99%** respectively on the test dataset. K-NN model gives best accuracy among decision trees and k-nn modeling.

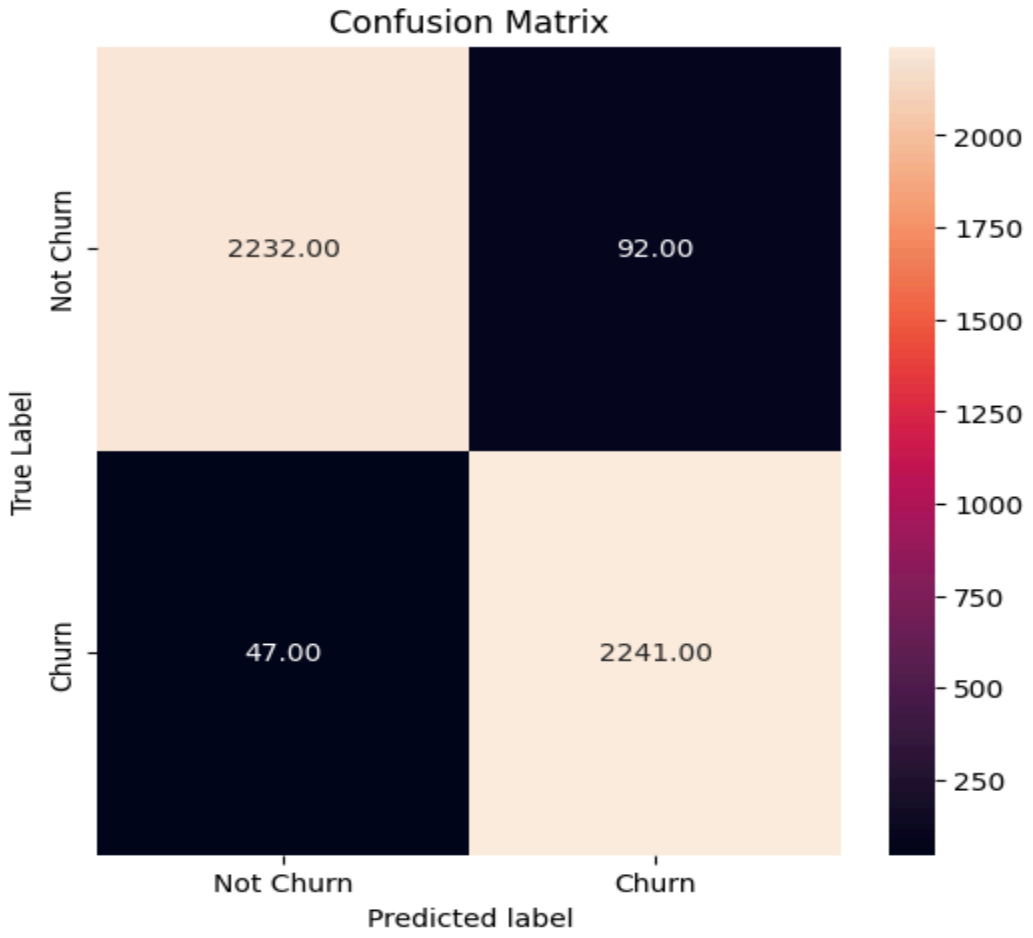
4. Random Forest Classifier:

   Random forest classifier with hyper parameter **`n_esimators=500, Max_depth=15, and random_state=42`** gives **accuracy of 99%** while **precision and recall of 99% and 99%** respectively on the test dataset.

5. Support Vector Machine:

   SVM classifier with hyper parameter **`C=2.0`** gives **accuracy of 97%** while **precision and recall of 97% and 97%** respectively on the test dataset. K-NN model gives best accuracy among decision trees and k-nn modeling. Confusion Matrix of the SVM is displayed below:

Graph No. 3.18: Confusion Matrix of SVM model

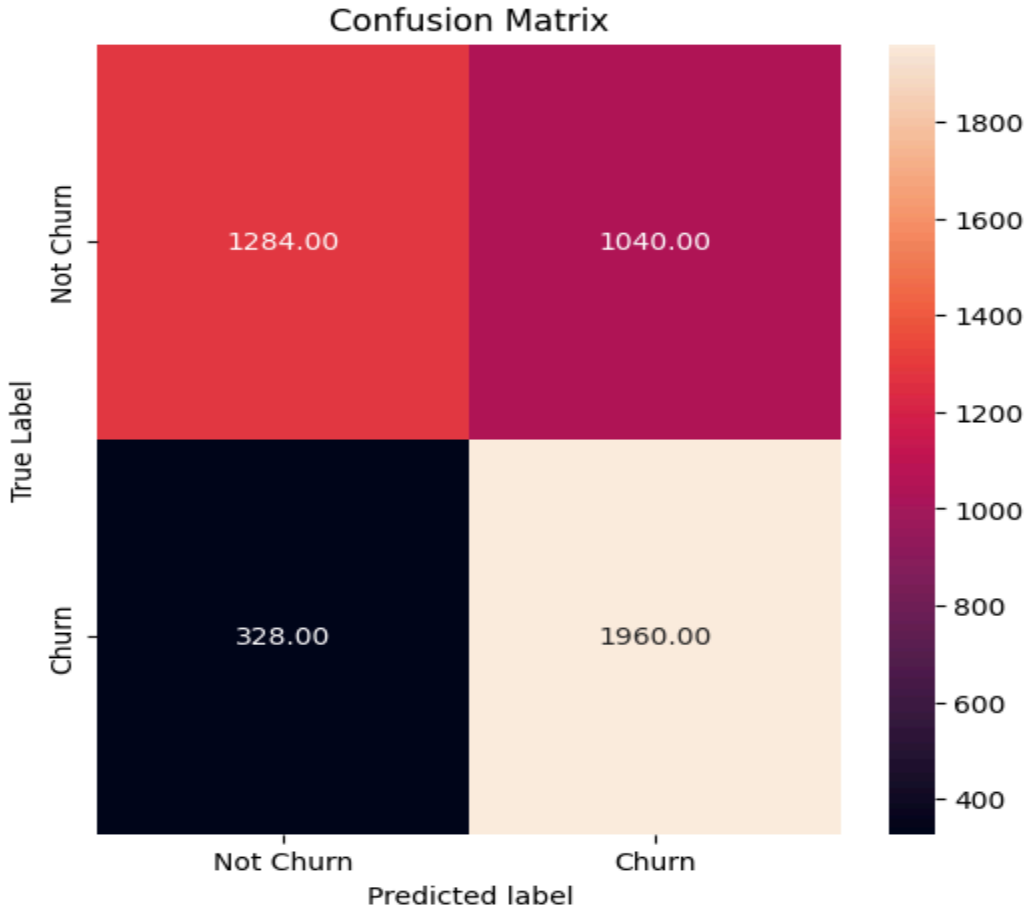


6. Gaussian Naive Bayes:

   Gaussian Naive Bayes classifier with hyper parameter gives **accuracy of 70%** while **precision and recall of 72% and 70%** respectively on the test dataset.

K-NN model gives best accuracy among decision trees and k-nn modeling. Confusion Matrix of the GNB is displayed below:

Graph No. 3.19: Confusion Matrix of Gaussian Naive Bayes



Model Comparison to Identify Best Model

Table No. 3.9: Model comparison

	Accuracy	**Recall**	**Precision**
Decision tree classifier	96.7%	97%	96%
Tuned decision	96%	96.4%	95.5%

tree classifier			
k-NN modeling	98%	99%	96%
Random Forest Classifier	99%	99%	99%
SVM	97%	97%	97%
Gaussian Naive Bayes	70%	70%	72%

In the above model comparison, one can clearly learn that the ensemble learning method ***Random Forest Model*** is clearly giving the highest accuracy of 99% along with recall and precision of 99%.

CHAPTER 4

FINDINGS, RECOMMENDATIONS AND CONCLUSION

FINDINGS, RECOMMENDATIONS AND CONCLUSION

4.1 Findings Based on Observations

1. We found that DTH businesses face high customer churn with a wide range of factors affecting churn.
2. DTH services began in the early 2000s with DD Dish TV which today has grown to millions of subscribers across rural and urban areas of the country.
3. As per the problem statement, we have to find out the reason for high customer churn and segment the potential churners with the use of a set of characteristics.
4. Dataset has mixed features with numerical and categorical columns.

4.2 Findings Based on analysis of Data

1. On average, customers contacted the customer service center 17 times in a year with a maximum number of customer service calls reaching 132 in a year.
2. The average customer service satisfaction score is higher than the company service satisfaction score.
3. Analyzing data summary it is noticed that columns like Login_device, Account_user_count, Tenure, Rev_per_month, cashback, etc have many missing values as well as corrupted data values that will need to be replaced with correct values during data cleaning tasks.
4. Based on analysis of correlations between each features, it is found that no features show significant correlation with each other
5. Using K-means clustering, we found that ***cluster 1*** has a higher churn rate than other clusters. That indicates significant features like Marital status, Customer agent score, Tenure, and cashback within a certain limit are crucial to address for reduction in customer churn.
6. Customers who have a regular plus account type are more likely to churn than any other account type.

7. If the number of users increases per account then such customers are more likely to churn than any other account type.

4.3 General findings

1. It is found that many features in the dataset have missing values, incorrect values and incorrect data types which needed to be corrected.
2. The customer churn dataset has 11260 samples and 19 features including target variables.
3. On average, customers contacted the customer service center 17 times in a year with a maximum number of customer service calls reaching 132 in a year.
4. Lower tenure of customers, low cashback, and complaints against customer service done in the last year of subscription are some of the major reasons for customer churn.
5. If customer Marital status is single then they are more likely to churn compared to married customers.
6. Machine learning models built on training datasets using random forest and decision tree models gave the highest accuracy of 99% and 97% respectively.
7. Random forest models performed well with most accuracy, precision and recall values.
8. Using feature importance it has become known that Tenure, Compain_LY, Marital_Status_Single, and Cashback are some of the most important features for customer churn.

4.4 Recommendation based on findings

1. It is recommended that DTH businesses deploy the ***Random forest model*** to predict customer churn because it has high accuracy which is reliable to predict potential churners.
2. DTH companies should adopt machine learning models that can help mitigate their customer churn.

3. Companies can launch campaigns to offer discounts on periodic subscriptions that increase the tenure of the customers resulting in higher retention rates. DTH businesses can use direct and personalized email marketing to target such customers who are most likely to churn.
4. Increase in cashbacks to those who are new customers can help reduce customer churn.
5. Customers with marital status as a single should be offered with special offers and packages to increase retention rates.
6. Companies should employ quality customer services and best practices to reduce the number of complaints for reduction in customer churn.
7. Those companies which operate in tier 3 cities should provide specialized discounts to the customer in order to improve retention in such towns and cities.
8. Customers with ***'Regular Plus'*** are more likely to churn so it is recommended that such account segments be removed or modified as per the needs of customers.
9. Businesses can ensure a more robust payment mechanism to reduce cash on delivery options as such customers are more likely to churn than other payment options.
10. DTH companies need to address low customer service scores and improve process and service efficiency to reduce customer churn for the business.

4.5 Suggestions for areas of improvement

1. DTH Companies can improve in areas of customer service and efficiency.
2. Customers with low tenure can be targeted using email marketing.

4.6 Scope for future research

In this research project, one can notice comprehensive review and analysis of secondary data sources with objective analysis of a given dataset with CRISP-DM methodology and Exploratory, predictive and prescriptive research methods. It resulted in derivation of a wide range of insights and machine learning modeling to predict customer churn.

The future scope of research for this project would be to enhance the quality and quantity of data with inputs from other sources to find much detailed insights and develop models based on those data. Furthermore, model interpretation and explanation can be crucial topics to research on using various methods to identify specific attributes for customer churn.

4.7 Conclusion

In conclusion, this research project on customer churn prediction has provided a comprehensive approach to understanding and mitigating customer churn through detailed exploratory data analysis (EDA), machine learning (ML) modeling, and actionable recommendations. The EDA illuminated key factors influencing churn, such as customer tenure, service usage, and customer service interactions. By leveraging various ML models, including support vector machines, decision trees, and ensemble methods, we identified patterns and predictors of churn with significant accuracy. The insights derived from these models informed targeted strategies for customer retention, such as personalized marketing campaigns, improved customer service protocols, and tailored service offerings. Implementing these recommendations can help businesses proactively address churn, enhance customer satisfaction, and ultimately, drive long-term profitability. This project underscores the importance of data-driven decision-making in fostering customer loyalty and sustaining competitive advantage in the marketplace.

REFERENCES

1. Prabhadevi, B., Shalini, R., & Kavitha, B. (2023a). Customer churning analysis using machine learning algorithms. *International Journal of Intelligent Networks*, *4*, 145–154. https://doi.org/10.1016/j.ijin.2023.05.005
2. Gandhi, D. (2023, October 25). *How to Improve Retention with Churn Prediction Analytics*. Amplitude. https://amplitude.com/blog/churn-prediction
3. *Customer churn analysis in the telecom industry*. (2015, September 1). IEEE Conference Publication | IEEE Xplore. https://ieeexplore.ieee.org/document/7359318
4. *Direct-to-home television in India*. (2024, April 10). Wikipedia. https://en.wikipedia.org/wiki/Direct-to-home_television_in_India#:~:text=The%20DTH%20and%20the%20cable,quarter%20ending%2030%20September%202022.
5. Singh, A. G. C. C. G. (2019). Predicting Customer Churn for DTH: Building Churn Score Card for DTH. *ideas.repec.org*. https://ideas.repec.org/h/spr/prbchp/978-981-13-1208-3_9.html
6. https://www.researchgate.net/publication/327536278_Predicting_Customer_Churn_for_DTH_Building_Churn_Score_Card_for_DTH#:~:text=The%20primary%20drivers%20of%20churn,recharged%20their%20set%20up%20box.
7. https://www.sciencedirect.com/science/article/pii/S221256711400197X/pdf?md5=62994277acfce8414191cf46e77c3781&pid=1-s2.0-S221256711400197X-main.pdf

ANNEXURE

Sr. No.	Title	Type of file	Remarks
1.	Customer Churn Prediction Modeling	.ipynb	Code file of data analysis and modeling performed for this project.
2.	Research Project Interim Report - Customer Churn Prediction Project	.docx	Interim report submitted to LMS
3.	Research Project Synopsis	.docx	Project synopsis submitted on LMS
3.	Customer churn prediction slides	.pptx	Final report slides