MCA Semester – IV
Capstone Project

Name	I C Preetham Jayachandra Sama Jasthi Naveen Jawahar S R Nanda E	212VMTR00735 212VMTR00741 212VMTR00739 212VMTR00740 212VMTR00788
Project	**Customer Churn Final Project**	
Group	**Group – 2**	
Date of Submission	18-11-2023	

A study on "Customer Churn"

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

In partial fulfillment of the requirements for the award of:

Master of Computer Application

Submitted by:

Name	USN
I C Preetham	212VMTR00735
Jayachandra Sama	212VMTR00741
Jasthi Naveen	212VMTR00739
Jawahar S R	212VMTR00740
Nanda E	212VMTR00788

Under the guidance of:

Mr. Nimesh Marfatia

(Faculty-JAIN Online)

Jain Online (Deemed-to-be University)

Bangalore

2022-23

DECLARATION

We, I C Preetham, Jayachandra Sama, Jasthi Naveen, Jawahar S R and Nanda E, hereby declare that the Research Project Report titled "Customer Churn" *has been* prepared by us under the guidance of Mr. Nimesh Marfatia.We declare that this Project work is towards the partial fulfillment of the University Regulations for the award of the degree of Master of Computer Application by Jain University, Bengaluru. We have undergone a project for a period of Eight Weeks. We further declare that this Project is based on the original study undertaken by us and has not been submitted for the award of any degree/diploma from any other University / Institution.

Place :

Date : 17-11-2023

Name	USN
I C Preetham	212VMTR00735
Jayachandra Sama	212VMTR00741
Jasthi Naveen	212VMTR00739
Jawahar S R	212VMTR00740
Nanda E	212VMTR00788

ACKNOWLEDGEMENT

I (Jayachandra Sama) would like to express my sincere gratitude to my supervisor (Mr. Nimesh Marfatia) for his enthusiasm, patience, insightful comments, helpful information and practical advice that have helped me tremendously throughout the Capstone Project. His immense knowledge, profound experience, and professional expertise in performing EDA, Building Model and Evaluation the Model performance helped me to complete this capstone project successfully. Without his support and guidance, this project would not have been possible. I would also like to take this opportunity to thank my other team members who helped me in completing the Capstone Project on time. I would also like to thank Jain Online University and all the professors who taught me various concepts in stream of Data Science and giving me a chance prove my skill in performing EDA, Building ML Model and evaluating various models through this Capstone Project.

I (Jawahar S R) would like to extend my sincere gratitude to all those who have contributed to the completion of this Capstone project. First and foremost, I express my deepest appreciation to my faculty guide, Mr. Nimesh Marfatia, for his invaluable guidance, support, and expertise throughout this research journey. His insightful feedback and constructive criticism have immensely shaped the direction and quality of this paper. I would also like to extend my appreciation to my teammates who provided valuable discussions and insights during the course of the capstone project. Their contributions and support have been invaluable. Each of these individuals has played a significant role in the successful completion of this capstone project, and their contributions are deeply appreciated. I would also like to thank Jain Online University and all the professors who taught me various concepts in stream of Data Science and giving me a chance prove my skill in performing EDA, Building ML Model and evaluating various models through this Capstone Project.

I (I C Preetham) here by acknowledge all the support I got from Team members who helped to complete the project with equal amount of effort from all of them. And I extend my gratitude to my Capstone project Mentor (Mr. Nimesh Marfatia) for expressing either our work or changes in the project and helped to accomplish better project.

I (Nanda E) would like to express my gratitude to Mr. Nimesh Marfatia, my esteemed advisor, for all the guidance, support, and instruction he provided me throughout the Capstone project. I would like to thank the Faculty of Data Science at Jain Online University for providing me with the resources to complete the Capstone Project.

I (Jasthi Naveen) would like to express my gratitude to Mr. Nimesh Marfatia, my esteemed advisor, for all the guidance, support, and instruction he provided me throughout the Capstone project. I would like to thank the Faculty of Data Science at Jain Online University for providing me with the resources to complete the Capstone Project.

CERTIFICATE

This is to certify that the Capstone Project report submitted by Mr./Ms.

Name	USN
I C Preetham	212VMTR00735
Jayachandra Sama	212VMTR00741
Jasthi Naveen	212VMTR00739
Jawahar S R	212VMTR00740
Nanda E	212VMTR00788

bearing on the title *"*Customer Churn*"* is a record of project work done by him/ her during the academic year 2022-23 under my guidance and supervision in partial fulfillment of Master of Computer Application.

Place: Mr. Nimesh Marfatia

Date : 17-11-2023 Faculty Guide

TABLE OF CONTENTS

EXECUTIVE SUMMARY

An E Commerce company is facing a lot of competition in the current market, and it has become a challenge to retain the existing customers in the current situation. Hence, the company wants to develop a machine learning model through which they can do churn prediction of the accounts and provide segmented offers to the potential churners.

In this company, account churn is a major thing because 1 account can have multiple customers. Hence by losing one account the company might be losing more than one customer.

The company is having data about its customers details, using this data the company has decided to do churn prediction by comparing with the common features of the customer who has churned previously.

The company has about 11260 customer accounts out of which 1896 accounts have already churned. So, the company has a target to retain as many customers as possible from the remaining 9364 customers. The details of the customer like the tenure, gender, marital status, city_tier etc. are available. For each customer there are 19 features available.

Data analysis and interpretation were performed using an Excel sheet with 11,260 rows and 19 columns. The table contains various variables related to customer accounts, such as churn flag, tenure, city tier, and customer care contact frequency. The data underwent a cleaning process to address outliers, missing values, and irrelevant data. Data transformation was then performed to assign correct data types to each column.

Five machine learning models were built for the Customer Churn project, with the Random Forest Classifier model performing the best. It achieved 100% accuracy, precision, recall, and F1 score on the training data, and showed good performance on the testing data.

Based on EDA analysis and Churn prediction, the campaign suggestions must be provided to customers to retain them. The campaign suggestion should be unique and be very clear on the campaign offer because the recommendation will go through the revenue assurance team. If they find that you are giving a lot of free (or subsidized) stuff thereby making a loss to the company; they are not going to approve the recommendation.

Thus, the overall objective of the project is to perform data cleaning, EDA analysis, building the classification model, evaluating the classification models and figure out a best suited one, identify the potentially churnable customer and provide appropriate campaign suggestions to those customers and retain them.

CHAPTER 1

INTRODUCTION AND BACKGROUND

1.1 Executive Summary

As the E-Commerce company is facing a lot of challenge to retain the customer and decided to build a ML model to predict the potentially churnable customers, as a team of data scientists we are building a classification model like Logistic regression using the account details provided by the company.

1.2 Introduction and Background

An E-Commerce company is facing a lot of competition in the current market, and it has become a challenge to retain the existing customers in current situations.

The company has decided to develop a Machine learning model through which they can perform churn prediction on the accounts and provide segmented offers to the potential churners.

Account churn is a major thing in this company as one account can have multiple customers. Hence by losing one account the company might be losing more than one customer.

1.3 Problem Statement

The problem statement revolves around an E-commerce company facing intense competition in the current market, making customer retention a significant challenge. The company aims to develop a model capable of predicting account churn and offering segmented offers to potential churners. Given that one account can have multiple customers, the loss of a single account affects more than one customer. Therefore, the problem is to devise a churn prediction system that can identify potential churners accurately and provide tailored offers to retain them, ultimately minimizing customer churn and maximizing customer loyalty.

1.4 Objective of Study

The objective of this study is to develop a churn prediction model (Classification Model) for the E-commerce company that accurately identifies potential churners and provide solutions to retain them. The goal is to reduce churn and increase customer satisfaction and loyalty by predicting which accounts are at risk of churning and offering them personalized incentives to encourage them to use the company's services. This model should effectively predict churn and provide effective solutions to decrease churn that are cost-effective for the company and attractive to the customers. This will help in retaining existing customers and minimizing the loss of revenue due to churn.

1.5 Company and Industry Overview

The B2C e-commerce industry is experiencing significant growth and is expected to continue expanding in the coming years. According to recent reports, the global B2C e-commerce market size was valued at USD 3.67 trillion in 2020 and is projected to reach USD 7.0 trillion by 2028. This growth can be attributed to various factors, including rising disposable income, increasing internet penetration, and the convenience of online shopping.

The B2C e-commerce market is driven by the dominance of key players such as Amazon, Alibaba, JD.com, and Flipkart. These companies have capitalized on the convenience factor and the lack of an organized offline retail sector to attract customers. The COVID-19 pandemic also had both positive and negative impacts on the market, with a surge in demand for essential goods but disruptions to supply chains.

The market is segmented based on type, with B2C retailers and classifieds being the major segments. In terms of revenue share, the clothing and footwear segment is the largest, followed by the consumer electronics segment. The report suggests that the consumer electronics segment is expected to grow further due to increasing internet usage and consumer interest in new products and trends.

To stay competitive in the evolving e-commerce industry, businesses need to prioritize customer satisfaction, embrace digital transformation, and leverage technology. Personalized experiences, technological advancements, and the increasing penetration of smartphones are key drivers for the growth of the market.

CHAPTER 2

RESEARCH METHODOLOGY

2.1 Scope of the Study

The scope of the study is to develop a churn prediction model for an E-commerce company. The study aims to identify potential churners accurately and provide tailored solutions to retain them. The objectives include reducing churn, increasing customer satisfaction and loyalty, and minimizing revenue loss.

2.2 Methodology

2.2.1 Data Collection

Secondary data issued by Jain University

2.2.2 Data Analysis Tools

Tableau

Microsoft Excel

Jupyter Notebook

CHAPTER 3

DATA ANALYSIS AND INTERPRETATION

DATA ANALYSIS AND INERPRERTION

The excel Customer Churn contains 11260 rows and 19 columns. The meta data of the table with column names are as mentioned below.

Sl. No	Variable	Description
1	AccountID	account unique identifier
2	Churn	account churn flag (Target)
3	Tenure	Tenure of account
4	City_Tier	Tier of primary customer's city
5	CC_Contacted_L12m	How many times all the customers of the account has contacted customer care in last 12months
6	Payment	Preferred Payment mode of the customers in the account
7	Gender	Gender of the primary customer of the account
8	Service_Score	Satisfaction score given by customers of the account on service provided by company
9	Account_user_count	Number of customers tagged with this account
10	account_segment	Account segmentation on the basis of spend
11	CC_Agent_Score	Satisfaction score given by customers of the account on customer care service provided by company
12	Marital_Status	Marital status of the primary customer of the account
13	rev_per_month	Monthly average revenue generated by account in last 12 months
14	Complain_l12m	Any complaints has been raised by account in last 12 months
15	rev_growth_yoy	revenue growth percentage of the account (last 12 months vs last 24 to 13 month)
16	coupon_used_l12m	How many times customers have used coupons to do the payment in last 12 months
17	Day_Since_CC_connect	Number of days since no customers in the account has contacted the customer care
18	cashback_l12m	Monthly average cashback generated by account in last 12 months
19	Login_device	Preferred login device of the customers in the account

Table 1: Meta Data

3.1 Data Cleaning

Now that basic structure of the table is clear, the table needs to be cleaned and all the irrelevant data and the outliers should be taken care of. To do this each column is analyzed separately and replaced with respective solutions to the outliers and missing data.

a) Churn

There are 9364 records negative to churn and 1896 records positive to churn. Since there is no irrelevant data observed, no action is required here.

The percentage of missing values in the churn features is: 0.0%. There are no missing values in this feature. Hence there is no further missing value treatment required.

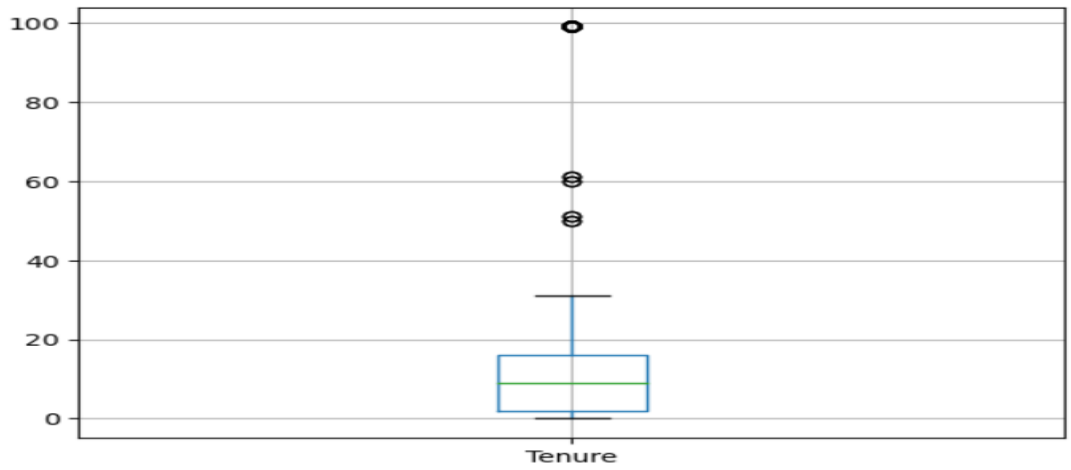
b) Tenure

The Tenure Feature contains discrete numerical data.
It is observed that there are 116 non-numeric values (#) and 102 Null values. The non-numeric values must be replaced with Null values.

After the non-numeric values (#) have been replaced with Null values, there are 218 Null values in the Tenure feature. These Null values will be replaced with median of Tenure. The Median of the Tenure feature is 9. All the Null values have been replaced by 9 making the total count of 9 from 496 to 714. This indicates that all the null values have been replaced. After handling the missing values in Tenure feature, the data type of tenure feature changes from object type to int64 type.

Next is to find the outliers of the Tenure feature and this is done using Box Plot

.



Graph1: Tenure Box Plot

Based on the box plot, it is observed that there are outliers in the table. The outliers are 99, 50, 60, 51 and 61 and the maximum cap value from the box plot is 31. Thus, all the outlier values are replaced with 31.

c) City_Tier

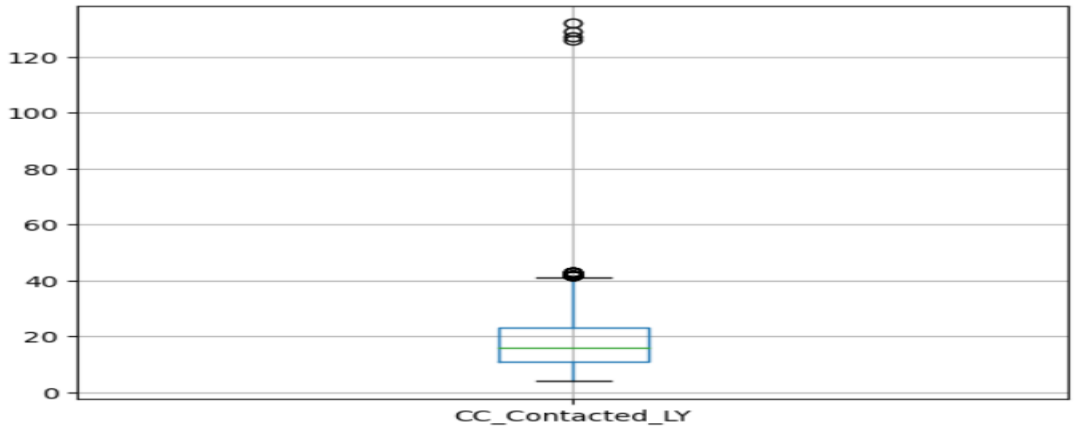
In the City_Tier it can be observed that the count of Tier 1 is 7263, Tier 2 is 480, Tier 3 is 3405 and Null values is 112. There is no irrelevant data in Cit_Tier feature.

The next step is to clear the null values. As per previous data the total records of null value amounts to 112 which is 0.9948% of the total records. Since the variables in City_Tier are categorical variables, the null values can be replaced with either median or mode.
The mode of is 1. Thus, all the null values are replaced with 1 which increased the count of Tier1 to 7375.

d) CC_Contacted_LY

There are no irrelevant data in CC_Contacted_LY feature. Next is to check for the existence of Null value and its treatment.
There are a total of 102 Null values in CC_Contacted_LY feature which is 0.9059% of the total records. Since the variables of CC_Contacted_LY are categorical variable the Null values are replaced with the median of CC_Contacted_LY.

The median of CC_Contacted_LY is 16. After replacing the Null values with median, the total frequency of 16 is increased to 796 from 663.
The next step is outlier treatment which can be viewed through Box Plot



Graph 2: CC_Contacted_LY Box Plot

The Box Plot shows that there are outliers in CC_Contacted_LY. The maximum cap value of the CC_Contacted_LY is 41. Thus, all the outliers are replaced with maximum cap value which is 41.

By replacing the outlier, the total frequency of 41 has increased to 71.

e) Payment

The Payment feature is categorical variable. There are no irrelevant data observed in Payment feature.

Cash on Delivery	1014
Credit Card	3511
Debit Card	4587
E wallet	1217
UPI	822
NaN	109

Table 2: Payment

The total frequency of Null values in Payment feature are 109. The null values are replaced with mode of Payment feature.

Mode of the Payment is Debit card thus replacing the Null values to Debit Card making the total frequency of Debit card 4969.
There are no outliers in this feature.

f) Gender

Gender is a categorical variable.
The frequency of gender is as shown below.

Male	6328
Female	4178
M	376
F	270
NaN	108

Table 3: Gender

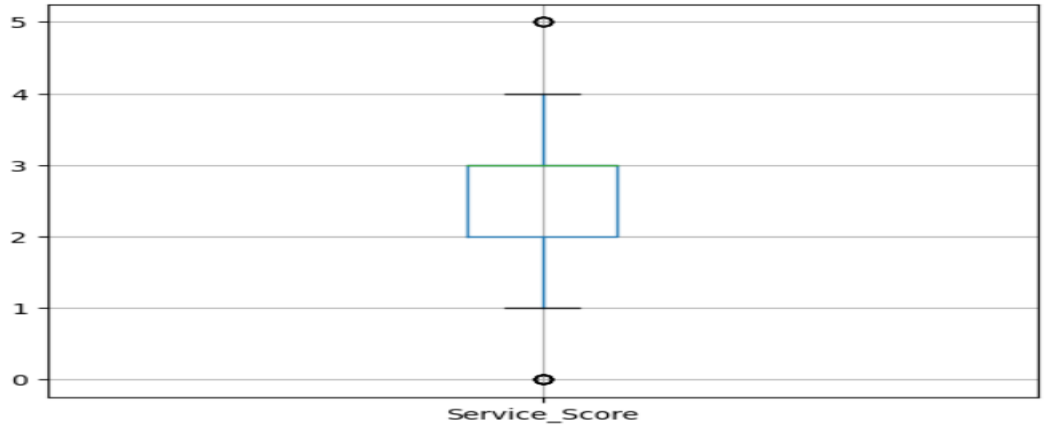
In Gender feature there are few entries with F and few with Female. It is the same for Male and M.

As both F and Female conveys the same information and M and Male conveys the same information, F is replaced with Female and M is replaced with Male. Thus, the irrelevant data has been handled.
The total number of missing values (NULL) are 108. Since this is a categorical data, the missing values can be replaced with mode.
The mode of Gender feature is Male. Thus, all the missing values are replaced with Male increasing the frequency of male to 6812. There are no outliers in Gender feature.

g) Service_Score

Service_Score is categorical data.
There is no irrelevant data based in this feature. There are 98 missing values. Since this data is categorical data, it can be replaced with mode.

The mode of Service_Score is 3.0. Thus, all the missing values are replaced with 3.0 increasing the frequency of 3.0 to 5588.



Graph 3: Service_Score Box Plot

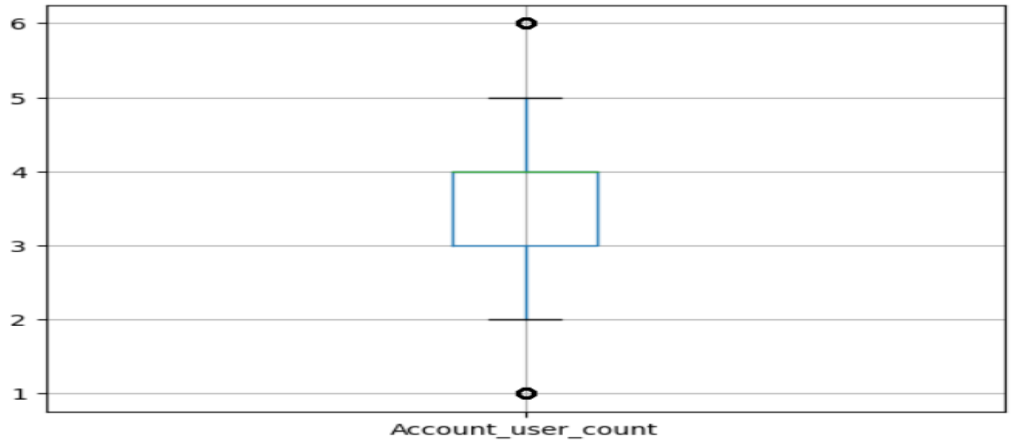
There are few outliers above the upper cap and few outliers below the lower cap. The lower cap is 1.0 and upper cap is 4.0.

All the values less than 1.0 that is 0.0 are replaced with 1.0 and all the values above 4.0 that is 5.0 are replaced with 4.0.

h) Account_user_count

There are irreleven data that is '@'. The irrelevant data has to be replaced with none.
There are 444 missing values after replacing '@' to none. These missing values are to be replaced with mode of Account_user_count.

The mode of the Account_user_count feature is 4. After replacing the null values in Account_user_count with mode, the frequency of 4.0 is increased to 5013 from 4569. The null value frequency is reduced to zero. Thus, all the null values are replaced with Mode successfully.



Graph 4: Account_user_count Box Plot

There are a few outliers above the upper cap and few outliers below the lower cap.

1 is an outlier below the lower cap which is 2. Hence replaced 1 with 2.6 is an outlier above the upper cap which is 5. Hence replaced 6 with 5.Thus outlier treatment performed successfully on Account_user_count feature.

i) Account_segment

There are few irrelevant data such as ‘Regular plus’ and “Regular +’. Thus we will replace these files

Observed that 'Super +' and 'Super Plus' convey the same information. Hence repalced 'Super +' with 'Super Plus'. Observed that 'Regular +' and 'Regular Plus' conveys the same information. Hence repalced 'Regular +' with 'Regular Plus’. Thus, irrelevant data in the account_segment feature is handled.

There are 97 missing values n account_segment. These null values are replaced by mode. The mode of account_segment is Regular Plus. Hence replaced null values with mode After replacing null values with mode, the frequency of Regular plus is increased to 4221 from 4124. The null values frequency is reduced to zero. Thus, all the null values are replaced with mode successfully. There are no outliers in the data.

j) CC_Agent_Score

No irrelevant data is observed in CC_Agent_Score feature. Hence no further action required on irrelevant data is required.
There are 116 missing values in CC_Agent_Score. These missing values are replaced with mode.

The mode of CC_Agent_Score is 3.0. After replacing the null values with mode, the frequency of 3.0 is increased from 5490 to 3476, and the null values frequency decreased to zero. Thus, all the null values are replaced with mode.

There are no outliers in CC_Agent_Score

k) Marial_Status

No Irrelevant Data is found in this feature.
There are 212 missing values in Martial_Status feature. The Null values will be replaced by mode of Martial_Status feature.

The mode is ‘Married’. After replacing the null values with mode the frequency of 'Married' is increased from 5860 to 6072, and the null values frequency decreased to zero.
Thus all the null values are replaced with mode in Marital_Status feature successfully. There are no outliers detected.

l) Rev_per_month

Observed that there is an irrelevant value '+' in the rev_per_month feature. Hence replaced the irrelevant value '+' with None.Thus irrelevant data is handled in rev_per_month feature.

There are 791 missing values in rev_per_month features. These missing values will be replaced by median.
The median is 5. Thus, all the null values are replaced with 5.

The maximum cap value of rev_per_month is 13.0. Since all the outliers are upper outliers, they are replaced with 13.0.

m) Complain_ly

No irrelevant data observed from Complain_ly feature. Hence no further action required on irrelevant data is required.
There are 357 missing values in Copmlain_ly features. These null value will be replaced with mode.

The mode is 0.0. Thus, all the null values are replaced with 0.0 increasing its frequency to 8149. There are no outliers in Copmalin_lyfactors

n) Rev_growth_yoy

Observed that there is an irrelevant data '$' in the rev_growth_yoy feature. Hence replaced the irrelevant value with None. Thus irrelevant data is handled in the rev_growth_yoy feature.

There are 3 Null values in rev_growth_yoy. All the null values of rev_growth_yoy will be replaced with its median
The median of rev_growth_yoyfeture is 15. Thus all the null values are replaced with 15.
There are no more null values in rev_growth_yoy feature.
There are no outliers in Rev_growth_yoy

As observed, there are outliers in Cash back which are both above upper cap and below lower cap. These outliers will be replaced by their respective caps.

The maximum cap value of the cashback feature is 271.36. The minimum cap value of the cashback feature is 81.0. The outliers above upper cap value is replaced with 271.36 and the outliers below lower cap value is replaced with 81.0.

o) Login_device_feature

There are 539 values of irrelevant data (&&&&) in Login_device feature. This irrelevant data is replaced with null values.

There are 760 null values in Login_device feature. These null values are to be replaced by mode because the Login_device feature is a categorical variable.
There are no outliers in Login_devicefeature

3.2 Data Transformation

Now that all the data cleaning has been done with all the columns, the next step is data transformation. The datatype for each column is as shown below.

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

Table 4: Columns data type before Data Transformation

All the object data types are transformed to int64

Variable Transformation for Payment Feature

0	1014
1	3511
2	4696
3	1217
4	822

Table 5:Payment after data transformation

Variable Transformation for Gender Feature

0	4448
1	6812

Table 6: Gender after data transformation

Variable Transformation for Account_segment

0	1639
1	520
2	4221
3	4062
4	818

Table 7: Account_segment after data transformation

Account_segment Feature

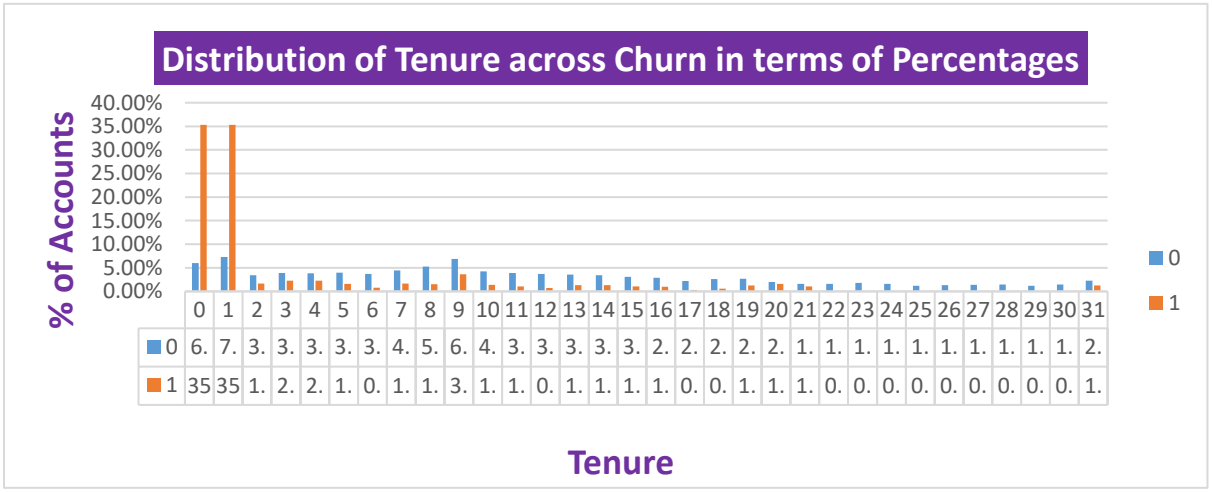
After transforming all the variables, the data type of each variable is as below.

#	Column	Non-Null Count	Dtype
0	AccountID	11260 non-null	int64
1	Churn	11260 non-null	int64
2	Tenure	11260 non-null	int64
3	City_Tier	11260 non-null	float64
4	CC_Contacted_LY	11260 non-null	float64
5	Payment	11260 non-null	int8
6	Gender	11260 non-null	int8
7	Service_Score	11260 non-null	float64
8	Account_user_count	11260 non-null	int64
9	account_segment	11260 non-null	int8
10	CC_Agent_Score	11260 non-null	float64
11	Marital_Status	11260 non-null	int8
12	rev_per_month	11260 non-null	float64
13	Complain_ly	11260 non-null	float64
14	rev_growth_yoy	11260 non-null	int64
15	coupon_used_for_payment	11260 non-null	float64
16	Day_Since_CC_connect	11260 non-null	float64
17	cashback	11260 non-null	float64
18	Login_device	11260 non-null	int8

Table 8: Columns data type after data transformation

3.2 Data Analysis and Business Insights

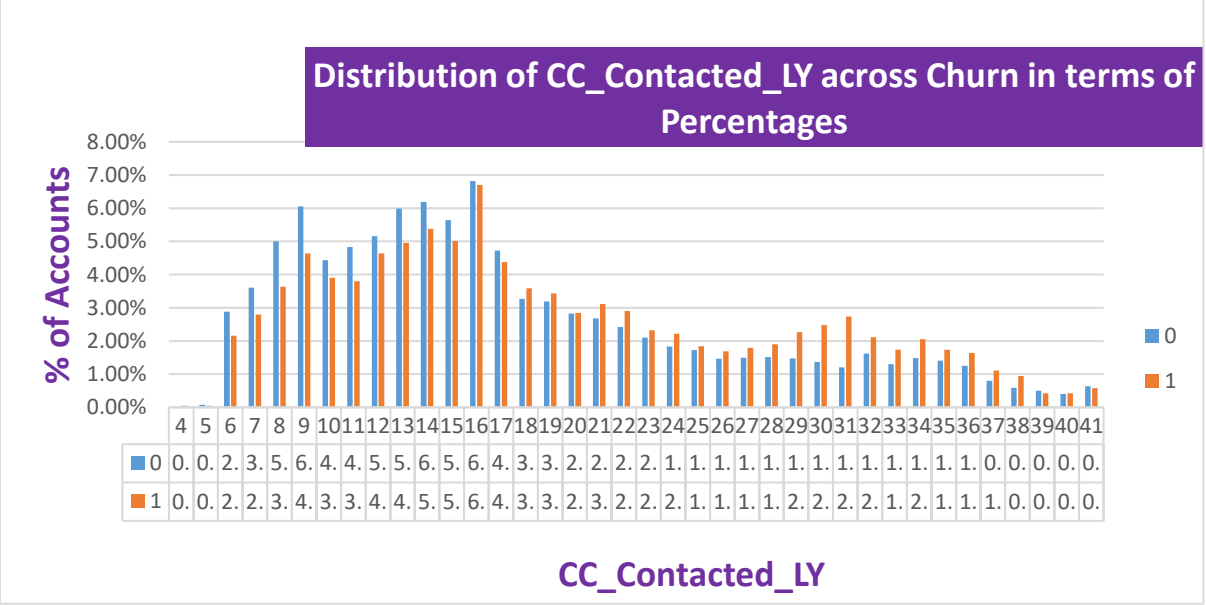
Tenure



Graph 5: Distribution of Tenure across Churn in terms of percentages

- If the Tenure is less than 1, the probability of churn is increased by 70%.
- If the Tenure is greater than 21, there is very less risk of churn.

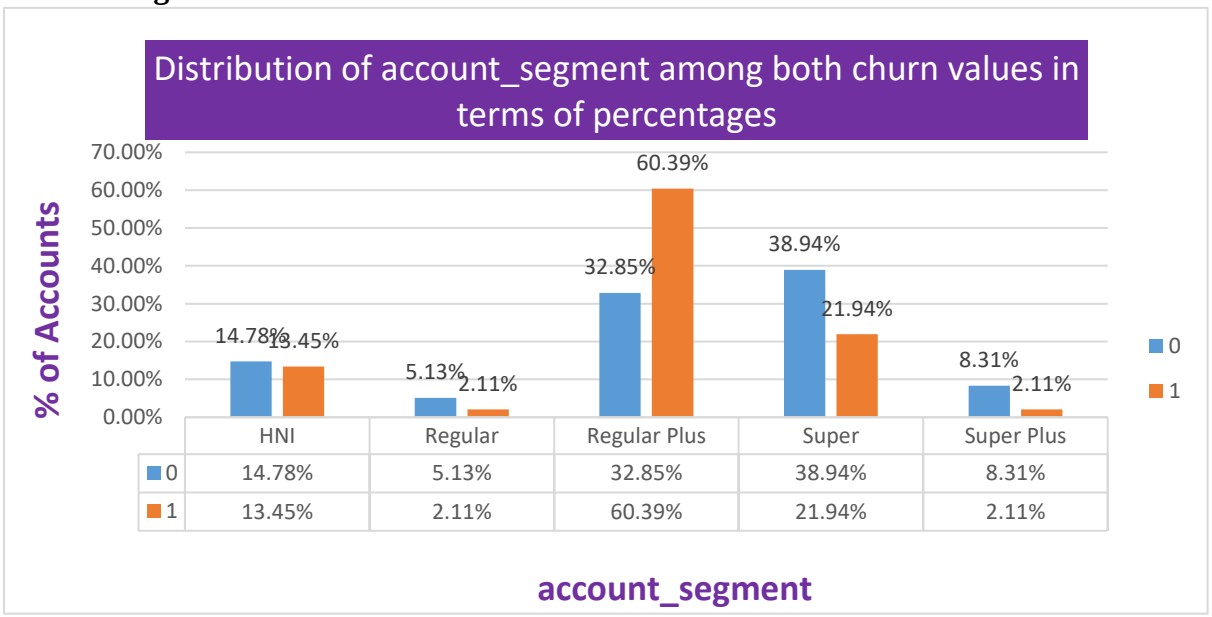
CC_Contacted_LY



Graph 6: Distribution of CC_Contacted_Ly across Churn in terms of percentages.

- If the number of times customer contacted last year is greater than 17, then the risk of churn increases.

Account Segment



Graph 7: Distribution of Account Segment across Churn in terms of percentages

- If the customer belongs to Regular Plus plan, then the Churn probability is very high at above 60%. Only 40% of customers churned belongs to other 4 segments.

CC_Agent_Score

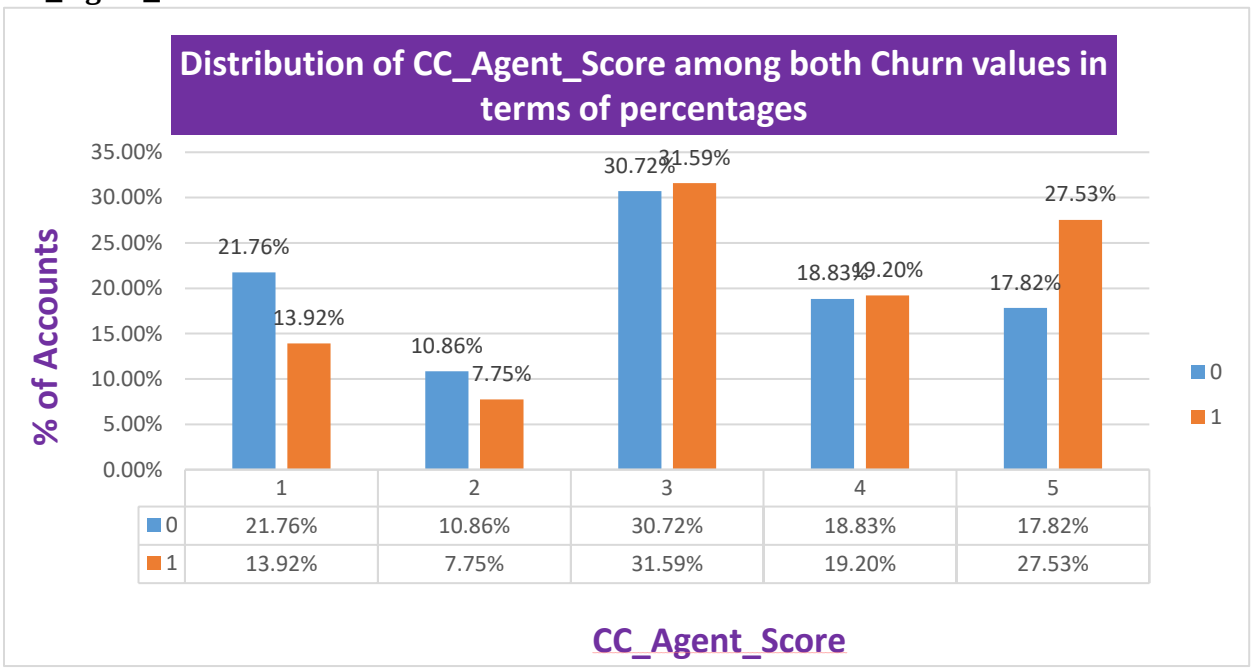


Chart 8: Distribution of CC_Agent_Scoreamong both Churn values in terms of percentages

- If the CC_Agent_Score is 5, there is a high risk of churn. 27.5% of churn is observed at CC_Agent_Score as 5 while the no churn is at 17.8%.

Martial Status

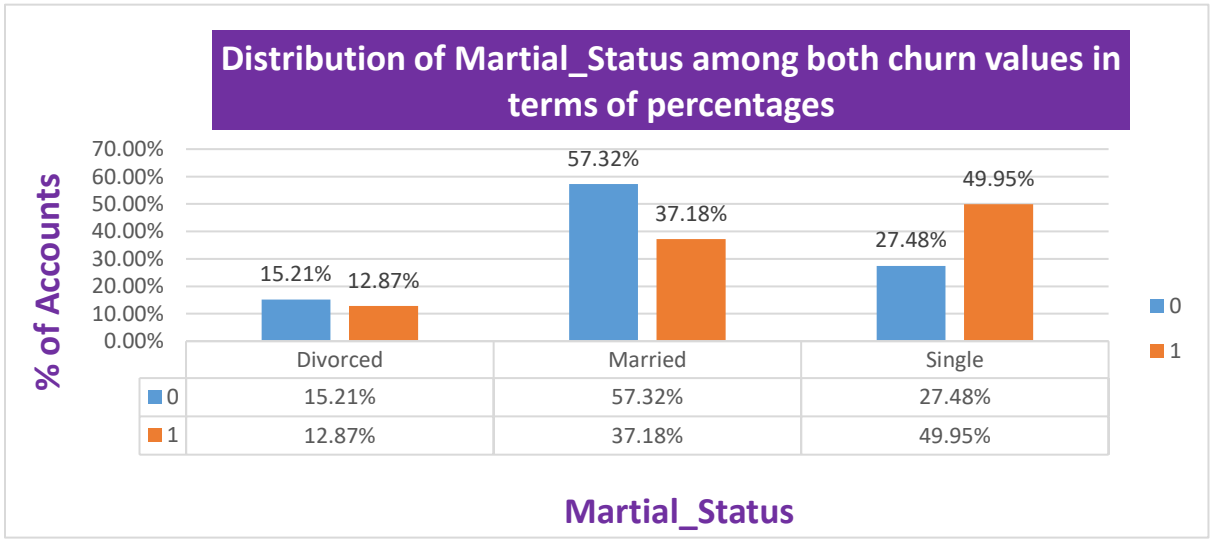


Chart 9: Distribution of Martial Status among both Churn values in terms of percentages

- If the Martial_Status of primary customer is Single, then there is a high probability of churn. Nearly 50% of customers churned are Singles, while the remaining 50% of customers churned belongs to Married and Divorced category.

Complain_LY

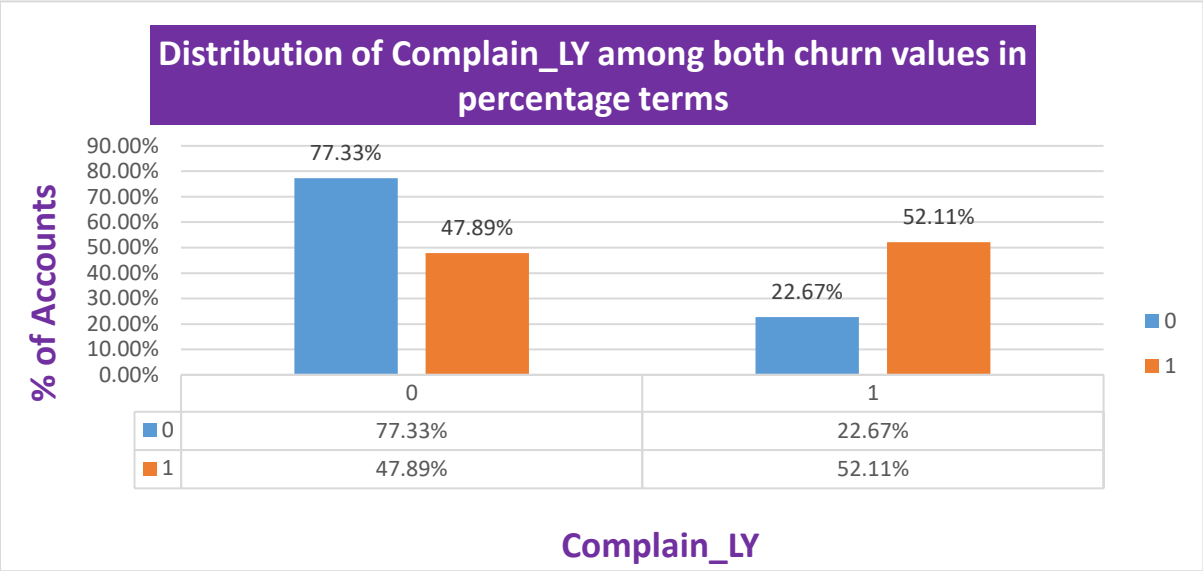
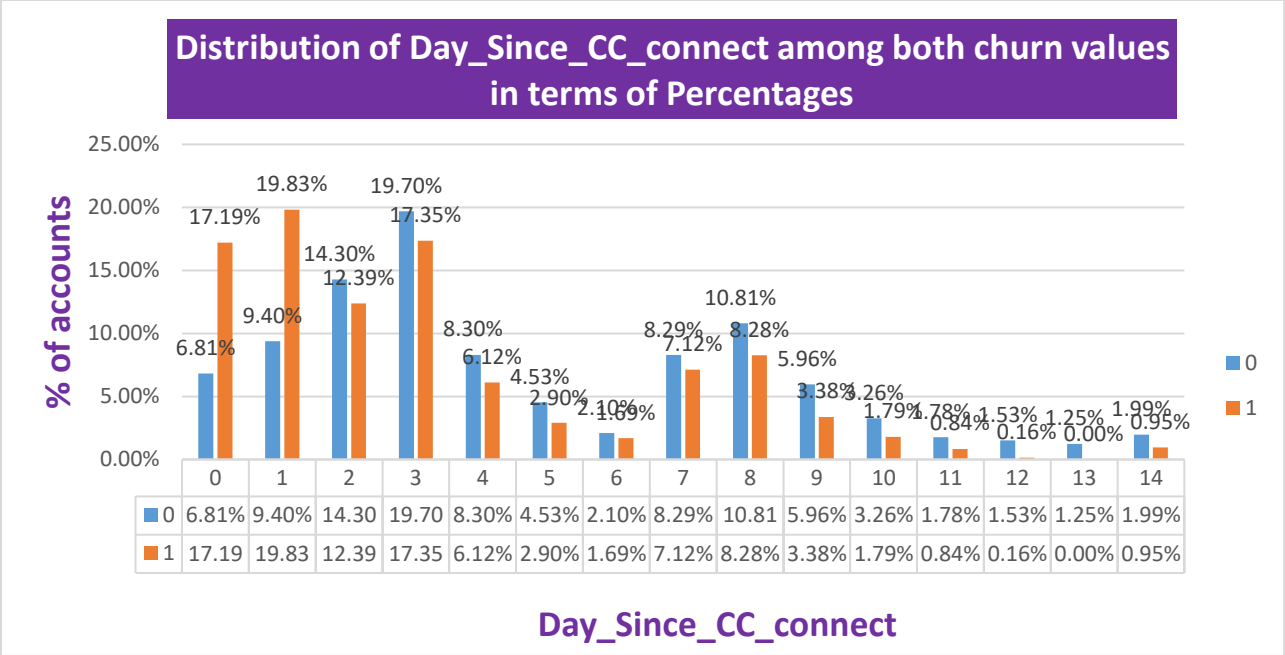


Chart 10: Distribution of Complain_LY amongboth Churn values in terms of percentages.

- If the customer makes a complaint in last one year, he has a high probability to churn. There are about 52.12% of customers who raised complaints last year and churned, while only 22.67% of retained customers have raised complaints.

- **Day_Since_CC_connect**



	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
0	6.81%	9.40%	14.30	19.70	8.30%	4.53%	2.10%	8.29%	10.81	5.96%	3.26%	1.78%	1.53%	1.25%	1.99%
1	17.19	19.83	12.39	17.35	6.12%	2.90%	1.69%	7.12%	8.28%	3.38%	1.79%	0.84%	0.16%	0.00%	0.95%

Chart 11: Distribution of Day_Since_CC_connect amongboth Churn values in terms of percentages

- If a customer contacted CC in last two days, then there is a high risk of churn. Among all customers contacted CC in last 15 days and churned, 30% of customers contacted CC in last 2 days.

Model building and interpretation.

Model building is an essential part of data analytics and is used to extract insights and knowledge from the data to make business decisions and strategies. In this phase of the project, team needs to develop data sets for training and testing. These data sets enable data scientists to develop an analytical method and train it while holding aside some of the data for testing the model.

Model building in data analytics is aimed at achieving not only high accuracy on the training data but also the ability to generalize and perform well on new, unseen data. Therefore, the focus is on creating a model that can capture the underlying patterns and relationships in the data, rather than simply memorizing the training data.

a) Splitting Data

The data provided by Customer is split in 70:30 ratio. 70% data is used to train the models and the remaining 30% of the data is used to evaluate the performance of the model.

b) Balancing Data

- The data provided by the customer is noticed as imbalanced data as it is having 17% of records related to churned customer having value as 1 and the reaming 83% of records is related to customers who have not churned.
- In order to build a model with better performance it is required to use the data with nearly 50% or records with churned customers information and 50% of the customers with non-churned customers.
- Thus, we used SMOTE algorithm to balance the data.
- After balancing the data with SMOTE algorithm, it is confirmed that the data is now having 50% of records with Churned customers and 50% of records with non-churned customers.
- Thus, this data can now be used for building the model.

c) Building Models

. It is required to build classification model as the churn feature in the data is a binary variable.

a. Using the preprocessed, cleaned and balanced data, Five classification namely Logistic Regression, Random Forest Classifier, CART classifier, Bagging classifier and Boosting classifier models were build.

b. All the above stated models were built successfully using the data provided.

d) Evaluating Models

. As the models are built successfully, there are evaluated using the ML performance metrics namely Accuracy, Precision, Recall, F1 Score and Accuracy Score.

a. The above stated metrics were captured for all the built models and kept in tabular format as below.

b. Based on the below table, it is observed that the Random Forest Model is showing 100% performance on the Training data and more than 95% of performance on the testing data**.**

c. Hence the random forest model is chosen as the best model among the other 4 models for the Customer Churn project with the givendata.

Performance Summary of each classification model											
SNO	Model Name	Accuracy		Precision		Recall		F1 Score		Accuracy Score	
		Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
1	Logistic Regression	81.84%	82.22%	83.82%	83.57%	78.90%	80.20%	81.29%	81.85%	81.84%	82.22%
2	Random Forest Classifier	100.00%	96.99%	100.00%	95.63%	100.00%	98.48%	100.00%	97.03%	100.00%	96.99%
3	CART Classifier	100.00%	93.80%	100.00%	92.10%	100.00%	95.82%	100.00%	93.92%	100.00%	93.80%
4	Bagging Classifier	99.89%	95.71%	99.80%	93.76%	99.97%	97.95%	99.89%	95.81%	99.89%	95.71%
5	Boosting Classifier	92.75%	92.90%	91.90%	92.14%	93.78%	93.80%	92.83%	92.96%	92.75%	92.90%

Table 9: Performance summary of each classification models

Business Insights:

1. If the Tenure is less than 1, the probability of churn is increased by 70%. It is recommended to be more focused on customers who joined in last one year. To retain the new customers, it would be suggested to provide them better offers and provide good customer support service to them on priority.

2. If the customer belongs to Regular Plus plan, then the Churn probability is very high at above 60%. Amont other segments, the Customers in the Regular Plus plan segment are more probable to churn. It is recommended to be more focused on customers in Regular Plus account segment. To retain these customers, it would be suggested to provide them better offers and provide good customer support service to them on priority.

3. If the CC_Agent_Score is 5, there is a high risk of churn. 27.5% of churn is observed at CC_Agent_Score as 5 while the no churn is at 17.8%.

4. If the Martial_Status of primary customer is Single, then there is a high probability of churn. Nearly 50% of customers churned are Singles, while the remaining 50% of customers churned belongs to Married and Divorced category.

5. If the customer makes a complaint in last one year, he has a high probability to churn. There are about 52.12% of customers who raised complaints last year and churned, while only 22.67% of retained customers have raised complaints.

Business Implications

1. Provide attractive offers to customers who tenure is less than 2 years as the customer with less than two years tenure are more prone to churn.
2. Reduce the number of times the Customers reach to Customer Care service by providing better service from Customer Care teams, as the customers who contacted Customer Care team for more than 17 times in last year are more prone to churn.
3. Provide attractive offers to Customers belonging to Regular-Plus account segment, as the customers in this account segment are more prone to churn.
4. Provide attractive offers to Customers who are single, as they are more prone to churn when compared with divorced and married customers.
5. Pay more attention to Customer who are reaching Customer Care in last two days as the Customers reaching to Customer care in last two days are more prone to Churn. Try to get more information about those customers and get feed back form those customers to find out the reasons for high Customer Churn in this area.

CHAPTER 4

FINDINGS, RECOMMENDATIONS AND CONCLUSION

FINDINGS, RECOMMENDATIONS AND CONCLUSION

4.1 Findings Based on Observations

1. Observed null values in all the features except Churn, AccountID, revenue_growth_yoy and coupon_used_for_payment features. All the null values are replaced with mean/median/mode based on the nature of the feature.
2. Observed some special charecters in the features namely Login_device, revenue_per_month, Account_User_Count, Tenure features. All the special charecters and irrelevant data is removed and replace with None.
3. Observed outliers in features like cashback, Coupons_used_for payment and Account_User Count. All the outliers are replaced for the particular feature based on the nature of the feature.

4.2 Findings Based on analysis of data

1. Below is the category of customers who are more probable to churn.
    a. Customer who joined in last one year
    b. Customers who belong to Regular Plus account segment.
    c. Customers who contacted customer support team for more than 17 times in last one year.
    d. Customers who are Singles
    e. Customers who raised complaints in last one year

4.3 General findings

1. The data provided is imbalanced data as there is 17% of records which has the churn value as 1 while the remaining 83% of records has the churn value as 0.
2. The data is balanced by using SMOTE algorithm and thus attained 50% of records having the churn value as 1 and remaining 50% of the records having churn value as 0.

4.4 Recommendation based on findings.

1. Provide better offers, best service, and good Customer Service support on priority to the customer falling in below stated category.
    a. Customer who joined in last one year
    b. Customers who belong to Regular Plus account segment.
    c. Customers who contacted customer support team for more than 17 times in last one year.
    d. Customers who are Singles
    e. Customers who raised complaints in last one year

4.5 Suggestions for areas of improvement

1. The data provided has a lot of noisy data in the form of outliers, irrelevant data, missing data etc. It is suggested to capture the clean data for which steps need to be planned an considered.

2. The clean data can help in building better ML models with high accuracy in prediction.

3. Provide better offers, high quality service and Customer Care support on a priority basis for the for the categorized customer listed in the recommendations based on findings section.

4.6 Scope for future research

1. The nature of business and Customer changes with time. Hence it is suggested to keep on building the ML models periodically with the latest available data to perform churn prediction.

2. This helps in better maintainability of the ML model built.

3. If any needy features are added to the churn data, then is s recommended to perform EDA on the new feature and build the ML models with the new data.

4.7 Conclusion

1. The data is split in 70:30 ratio. 70% portion of data is used for Training the ML models and the other 30% portion of data is used for Testing the ML Models.

2. Five Machine Learning models namely Logistic Regression, CART classifier model, Random Forest Classifier model, Bagging Classifier Model and Boosting Classifier Models are built for the Customer Churn project with the given data.

3. Among the five models, the Random Forest Regression model is providing good performance.

4. 100% Accuracy, Precision, Recall and F1 Score are observed on training data.

5. Very good Accuracy, Precision, Recall and F1 Score are observed on testing data as well with very minimal deviation of less than 5% when compared with training data.

6. Hence Random Forest Classification Model is the suggested model to build ML model for the Customer Churn project for the given data.

REFERENCES

(APA style; below is only a sample)

1. Great Learning Lecture Videos
2. Microsoft Excel
3. https://www.geeksforgeeks.org/understanding-logistic-regression/
4. https://www.geeksforgeeks.org/random-forest-classifier-using-scikit-learn/
5. https://www.geeksforgeeks.org/cart-classification-and-regression-tree-in-machine-learning/
6. https://www.geeksforgeeks.org/ml-bagging-classifier/
7. https://www.geeksforgeeks.org/ml-gradient-boosting/