CAPSTONE PROJECT

FINAL REPORT

MODELLING AND CLTV ANALYSIS ON CUSTOMER CHURN BEHAVIOR IN TELECOM INDUSTRY

CHENNAI
FEBRUARY 2022 BATCH

GROUP 7

Submitted On:

27 July 2022

By
Balakumaran K
Hemantha Kumar K S
Martin Aaron
Pavithra Rushendiran
Praveen Kumar M

Mentored by
Vikash Chandra
Senior Data Scientist
Fractal

TABLE OF CONTENTS

S.No.	TITLE	PAGE No.
1.	Project Details	1
	1.1. Overview	1
	1.2. Business Problem Statement (GOALS)	1
2.	Critical Assessment of Topic Survey	2
3.	Data Understanding	3
	3.1. Categorical Features	3
	3.2. Numerical Features	4
4.	Exploratory Data Analysis	6
	4.1. Imbalance in the Target Variable	6
	4.2. Null Values Analysis	7
	4.3. Analysis of Various Categorical Features	8
	4.3.1. Service Area	8
	4.3.2. Children in Household	8
	4.3.3. Handset Refurbished	9
	4.3.4. Handset Webcapable	9

S.No.	TITLE	PAGE No.
	4.3.5. Truck Owner, RV Owner and Motorcycle Owner	10
	4.3.6. New Cellphone User	11
	4.3.7. Handset Price	11
	4.3.8. Non-US Travel	12
	4.3.9. Mailings	12
	4.3.10. Owns a Computer	12
	4.3.11. Home Ownership	13
	4.3.12. Made Retention Call to Team	13
	4.3.13. Credit Rating	13
	4.3.14. Marital Status	13
	4.3.15. Occupation	14
	4.4. Analysis of Various Numerical Features	14
	4.4.1. Monthly Revenue, Monthly Recurring Charge and Monthly Usage	14
	4.4.2. Percentage Change in Revenues and Minutes	14
	4.4.3. Months in Service	15

S.No.	TITLE	PAGE No.
	4.4.4. Current Equipment Days	15
	4.4.5. Value Added Services	15
	4.4.6. Retention Offers Accepted	16
	4.4.7. Correlation Matrix	17
5.	Model Building	18
6.	Variance Inflation Factor	21
7.	Logistic Regression	22
	7.1. Classification Report	24
	7.2. Confusion Matrix	24
	7.3. ROC Curve	25
8.	Decision Tree Algorithms	25
	8.1. Classification Report	26
	8.2. Confusion Matrix	27
	8.3. ROC Curve	27
9.	Random Forest Classifier	27
	9.1. Classification Report	29

S.No.	TITLE	PAGE No.
	9.2. Confusion Matrix	29
	9.3. ROC Curve	30
10.	XG Boost	30
	10.1. Classification Report	31
	10.2. Confusion Matrix	32
	10.3 ROC Curve	32
11.	K Nearest Neighbours	33
	11.1. Classification Report	33
	11.2. Confusion Matrix	34
	113. ROC Curve	35
12.	Naïve Bayes	35
	12.1. Classification Report	36
	12.2. Confusion Matrix	37
	12.3. ROC Curve	38
13.	Summary of the Findings	38
	13.1. Using VIF techniques derived features	38

S.No.	TITLE	PAGE No.
	13.2. Using Principal Component Analysis (13 Components)	39
14.	Suggestion for Reducing Churn	40
15.	References	41

1. PROJECT DETAILS

1.1. OVERVIEW:

With the enormous increase in the number of customers using telephone services, the marketing division for a telecom company wants to attract more new customers and avoid contract termination from existing customers (churn rate - the annual percentage rate at which customers stop subscribing to a service or employees leave a job) Some of the factors that caused existing customers to leave their telecom companies are better price offers, faster internet services, and a more secure online experience from other companies.

A high churn rate will adversely affect a company's profits and impede growth. Our churn prediction would be able to provide clarity to the telecom company on how well it is retaining its existing customers and understand what are the underlying reasons that are causing existing customers to terminate their contract (high churn rate). Since the cost of acquiring new customers is much higher than retaining its existing customers, the company can use the churn rate analysis to provide discounts, special offers, and superior products to keep current customers.

1.2. BUSINESS PROBLEM STATEMENT (GOALS)

1. **Business Problem Understanding:**

   In this highly competitive market, the telecommunications industry experiences an average of 15-25% annual churn rate. For many incumbent operators, retaining high profitable customers is the number one business goal. To reduce customer churn, telecom companies need to predict which customers are at high risk of churn. In this project, you will analyze customer-level data of a leading telecom firm, build predictive models to identify customers at high risk of churn.

2. Business Objective:

Given the fact that it costs 5-10 times more to acquire a new customer than to retain an existing one, the main objective is **customer retention** has now become even more important than customer acquisition.

3. Approach:

CLTV method has been chosen to approach the problem statement. Customer Life time Value (CLTV) is a quantitative analysis and one of the most important metric to modern customer centric business scenario. It has been a mainstay concept in direct response marketing for many years, and has been increasingly considered in the field of marketing. Here, we will experiment with a implementable CLTV model that is useful for market segmentation and the allocation of marketing resources for acquisition, retention, and cross-selling. Analyzing the current trends and combinations of customer behavior gives the business owner various insights for demand-supply planning.

4. Conclusions

The primary objective of this project is to help the business owner increase sales and in order to do the same we aim to provide actionable insights.

2. CRITICAL ASSESSMENT OF TOPIC SURVEY:

1. Finding the key area, gaps identified in the topic survey where the project can add value to the customers and business
    - Identify CLTV based on customers
    - Identify Churn and Customer Lifetime Value

- CLTV analysis helps to classify customers who are important to the business and who are not adding much value. By evaluating the gap analysis for the customer segments through CLTV and figure out the best strategies to bridge that gap, thus increasing sales for the store.

2. What key gaps are we trying to solve?

- Analyze user behavior
- Uncover hidden correlations
- Customer retention rate
- Customer Life time value
- Reason behind churn rate.

3. DATA UNDERSTANDING

The dataset contains a total of 49752 rows (records) and 58 columns (Features) including, the target variable (Churn).

3.1. CATEGORICAL FEATURES

S.No.	Categorical Features	Column Description
1	Churn	The column describes whether the customer is going to switch to another network or not.(Yes or No)
2	ServiceArea*	Describes the region (coded) where the service is being provided.
3	ChildrenInHH	Is the Handset User a Child or not? (Yes or No)
4	HandsetRefurbished	Is the Handset being used whether a new one or a refurbished one.(Yes or No)
	HandsetWebCapable	Is the Handset Web Capable? (Yes or No)

6	TruckOwner	Is the customer a Truck Owner? (Yes or No)
7	RVOwner	Is the customer a RV Owner? (Yes or No)
8	Homeownership	Does the customer owns a Home? (Known or Unkown)
9	BuysViaMailOrder	Does the customer shops through Mail-order? (Yes or No)
11	OptOutMailings	Has the customer opted out of Mailing Advertisements? (Yes or No)
12	NonUSTravel	Has the customer ever been out of United States? (Yes or No)
13	OwnsComputer	Does the customer own a Computer? (Yes or No)
14	HasCreditCard	Does the Customer have a Credit Card? (Yes or No)
15	NewCellphoneUser	Is this the first time the customer using a mobile phone for the first time? (Yes or No)
16	NotNewCellphoneUser	Is this the first time the customer using a mobile phone for the first time? (Yes or No)
17	OwnsMotorcycle	Does the customer own a Motorcycle? (Yes or No)
18	MadeCallToRetentionTeam	Has the customer made call regarding switching to another network. (Yes or No)
19	CreditRating	Describes the credit worthiness of the the customer. (1 - Highest, 2 - High, 3 - Good, 4 - Medium, 5 - Low, 6 - Very Low, 7 - Lowest)
20	PrizmCode	Gives the location code (Town, Rural, Suburban and Others)
21	Occupation	The colum provides the nature of job the customer is into. (Professional, Crafts, Clerical, Self, Retired, Student, Homemaker and others)
22	MaritalStatus	Is the Customer Married or Not? (Yes, No and Unknown)
23	HandsetModels	Different categories of Handset Models.

24	IncomeGroup	Categories of different Income Group the customer falls under.
***Denotes Columns having null values**		

3.2. NUMERICAL FEATURES

S.No.	Numerical Features	Column Description
1	CustomerID	Provides identification number for every Customer.
2	MonthlyRevenue*	Monthly Revenue the network receives from each customer.
3	MonthlyMinutes*	Total Number of minutes the customer spends on calls.
4	TotalRecurringCharge*	Base plan subscribed by the customer.
5	DirectorAssistedCalls*	Directory Assistance is a phone Service used to find out specific telephone number and/or address of a residence, business, or government entity. The column describes the number of minutes the customer spends on Directory assisted calls.
6	Overageminutes*	When a user goes over the minutes allowed under the particular post-paid plan, they are charged separately for the extra minutes. This is called Overage minutes. The column describes time(in minutes) the customer spends on overage.
7	RoamingCalls*	Minutes spent on call while the customer is away from the home state.
8	PercChangeMinutes*	Percentage change in total minutes on call
9	PercChangeRevenues*	Percentage change in revenues from the customer
10	DroppedCalls	Minutes spent while the Calls dropped.

11	BlockedCalls	Average Blocked Calls that tried to connect.
12	UnansweredCalls	Number of unanwered calls
13	CustomerCareCalls	Total number of calls made by the customer to Customer Care.
14	ThreewayCalls	Total Number of calls with two more recipients.
15	ReceivedCalls	Total Calls attended by the customer.
16	OutboundCalls	Total Calls Made by the Customer.
17	InboundCalls	Total Incoming Calls to the Customer
18	PeakCallsInOut	Minutes spent on call during Peak Hours
19	OfPeakCallsInOut	Minutes spent on call during Non-Peak Hours.
20	DroppedBlockedCalls	Blocked Numbers that were dropped.
21	CallForwardingCalls	Minutes spent on call that were forwarded to another number.
22	CallWaitingCalls	Minutes spend by inbound callers on Waiting.
23	MonthsInServiceCalls	Total number of months the customer has been using the service from the network provider.
24	UniqueSubs	Unique subscriptions bought by the customer from the service provider.
25	AcitveSubs	Active subscriptions in use by the customer.
26	Handsets	Total Handsets owned by the customer.
27	CurrentEquipmentDays	Total number of days the handset is in use.
28	AgeHH1*	Age of the Primary User.
29	AgeHH2*	Age of the Secondary User.
30	RetentionCalls	Number of Calls made by the network to persuade the customer to not switch to another network.

31	RetentionOffersAccepted	Total number of calls it took for the customer to accept the retention offer and not to switch network.
32	ReferralsMadeBySubscriber	Total number of referrals made by the customer to acquire new customer.
33	AdjustmentsToCreditRating	Adjustments made to change the credit rating of the customer.
34	HandsetPrice	Cost of the Handset.
***Columns having Null Values**		

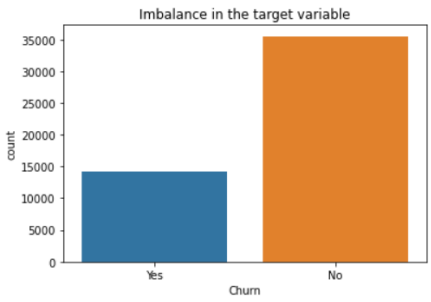
4. EXPLORATORY DATA ANALYSIS

4.1 IMBALANCE IN THE TARGET VARIABLE

We had checked the imbalance in the target variable (Churn) using countplot from seaborn.

From the figure, the Feature Churn is imbalanced. There are 35507 records with 'No' and 14245 with 'Yes' as Churn.

It was decided to not resample the target variable until base model building.



	Churn	Percentage
No	35507	71.37
Yes	14245	28.63

4.2. NULL VALUES ANALYSIS

	Null Value Percentage
CustomerID	0.000
Churn	0.000
MonthlyRevenue	0.306
MonthlyMinutes	0.306
TotalRecurringCharge	0.306
DirectorAssistedCalls	0.306
OverageMinutes	0.306
RoamingCalls	0.306
PercChangeMinutes	0.719
PercChangeRevenues	0.719
DroppedCalls	0.000
BlockedCalls	0.000
UnansweredCalls	0.000
CustomerCareCalls	0.000
ThreewayCalls	0.000
ReceivedCalls	0.000
OutboundCalls	0.000
InboundCalls	0.000
PeakCallsInOut	0.000
OffPeakCallsInOut	0.000
DroppedBlockedCalls	0.000
CallForwardingCalls	0.000
CallWaitingCalls	0.000
MonthsInService	0.000
UniqueSubs	0.000
ActiveSubs	0.000
ServiceArea	0.047
Handsets	0.002
HandsetModels	0.002
CurrentEquipmentDays	0.002
AgeHH1	1.781
AgeHH2	1.781
ChildrenInHH	0.000
HandsetRefurbished	0.000
HandsetWebCapable	0.000
TruckOwner	0.000
RVOwner	0.000
Homeownership	0.000
BuysViaMailOrder	0.000
RespondsToMailOffers	0.000
OptOutMailings	0.000
NonUSTravel	0.000
OwnsComputer	0.000
HasCreditCard	0.000
RetentionCalls	0.000

RetentionOffersAccepted	0.000
NewCellphoneUser	0.000
NotNewCellphoneUser	0.000
ReferralsMadeBySubscriber	0.000
IncomeGroup	0.000
OwnsMotorcycle	0.000
AdjustmentsToCreditRating	0.000
HandsetPrice	0.000
MadeCallToRetentionTeam	0.000
CreditRating	0.000
PrizmCode	0.000
Occupation	0.000
MaritalStatus	0.000

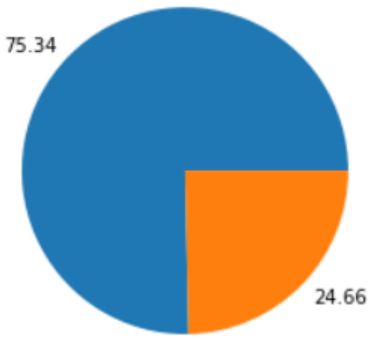
Only 5 features had null values that too very negligible. It was decided to remove the records containing null values.

4.3. ANALYSIS OF VARIOUS CATEGORICAL FEATURES

4.3.1. SERVICE AREA

There is a total of 743 unique area in which the telecom provides its services. It was decided to drop the service area column before model building.

4.3.2. CHILDREN IN HOUSEHOLD



	ChildrenInHH	Percentage
No	37483	75.34
Yes	12269	24.66

Nearly two-thirds of the household didn't have children in the household.

4.3.3. HANDSET REFURBISHED

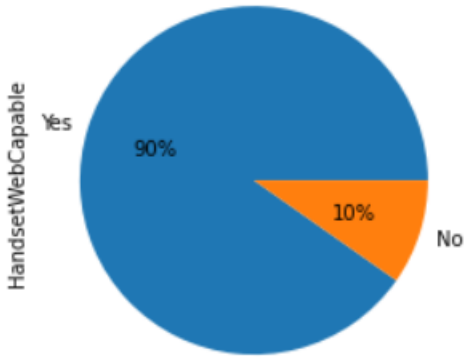
Handset refurbished refers to the customer using second-hand equipment. From the Churn response two things can be inferred. One, the customer falls under low-income category and hence we have to curate special packages exclusively for this income group or two, the customer is planning to switch to a brand-new handset and the possibility of churn is huge.

	HandsetRefurbished	Percentage
No	42852	86.13
Yes	6900	13.87

```
Total response where both Churn and Handset Refurbished was Yes :
 2209
In Percentage Terms out of Total Churn :
 15.51 %
In Percentage Terms out of Total HandsetRefurbished Users :
 32.01 %
```

4.3.4. HANDSET WEB CAPABLE

In the era of Industry 4.0, internet is vital for everyone. As anticipated the number of users with web-capable handset was higher but, it is necessary to look into the customers using handsets that aren't web capable. They may be senior-citizens, users with second line or they may be planning to switch to a newer handset in the future. It is important to anticipate and address such customer needs.




```
Total response where Handset Web Capable was No and Churn was Yes :
 1804
In Percentage Terms out of Total Churn (Yes):
 12.66 %
In Percentage Terms out of Total non-WebCapable Users :
 37.22 %
```

4.3.5. TRUCK OWNER, RV OWNER AND MOTORCYCLE OWNER

Considering that ownership of automobiles like Truck and implies that the customer may move from one place to another and may greatly rely upon RoamingCalls facility. In addition, motorcycle ownership in the United States implies that the customer may come under high income group given that motorcycle is costly in the United States. These three features were kept during base model building and it was decided to drop the columns one by one and iterate the process (trial and error) to understand and improve the model better.

	TruckOwner	Percentage
No	40280	80.96
Yes	9472	19.04

	RVOwner	Percentage
No	45619	91.69
Yes	4133	8.31

```
Total response where Truck Owner was Yes and Churn was Yes :
 2669
In Percentage Terms :
 18.74 %
In Percentage Terms out of Total Truck Owners :
 28.18 %
```

```
Total response where Truck Owner was Yes and Churn was Yes :
 1163
In Percentage Terms out of Total Churn (Yes):
 8.16 %
In Percentage Terms out of Total RVOwners :
 28.14 %
```

	OwnsMotorcycle	Percentage
No	49078	98.65
Yes	674	1.35

```
Total response where New Cellphone users who have responded Yes to Churn :
 213
In Percentage Terms out of Total Churn (Yes):
 1.5 %
In Percentage Terms out of those who own a Motorcycle :
 31.6 %
```

From the above observations, it can be concluded that nearly 28% of the total Truck and RV Owners had responded ‘Yes’ to Churn. Hence, the company has to focus more on curating packages for Roaming Calls.

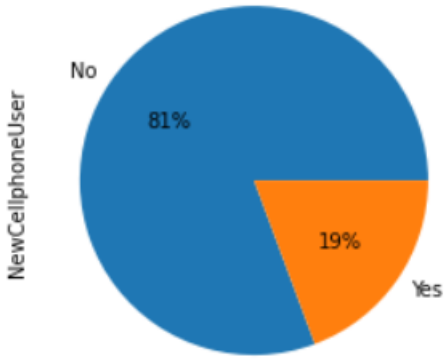
Also, nearly 31.6% of the total motorcycle owners intend to leave the telecom service provider. This means the company might lose high income group customers. It is important to develop strategies to retain the customers.

After model building, we may even drop these columns given that 98% of the customers do not own any motorcycle and only 10% and 20% of the customers use RV and Truck respectively.

4.3.6. NEW CELLPHONE USER

New cell phone users have the potential to bring additional income to the network. It is essential to retain them. Also, there was another feature named, 'NotNewCellPhoneUser',

which is opposite of the column. Hence that column was decided to be dropped before base model building.




```
Total response where New Cellphone users who have responded Yes to Churn :
 2684
In Percentage Terms out of Total Churn (Yes):
 18.84 %
In Percentage Terms out of those who are New Cellphone User :
 27.93 %
```

4.3.7. HANDSET PRICE

Nearly 28263 records had 'Unkown' as the HandsetPrice. This comprises 56% of the total records. Hence it was decided to drop the column before base model building.

4.3.8. NON-US TRAVEL

	NonUSTravel	Percentage
No	46896	94.26
Yes	2856	5.74

Given that 94.26% of the customers had never travelled outside the United States, the column is highly imbalanced and hence can be dropped before model building.

4.3.9. MAILINGS

Three Features namely, 'BuysViaMailOrder', 'RespondsToMailOffers', 'OptOutMailings' provide information about preferences of the customer to Mails. 98.5% of the customers had not opted out of mailings. This shows that the customers are greatly inclined to shop through mail. Hence, 'OptOutMailings' column could be dropped. But, the other two columns are vital given that there is fair distribution between preferences.

	BuysViaMailOrder	Percentage
No	31432	63.18
Yes	18320	36.82

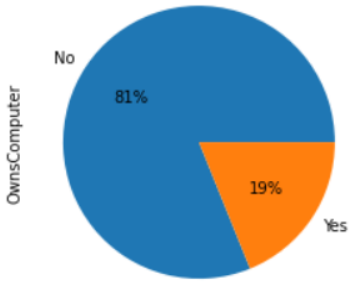
	OptOutMailings	Percentage
No	49006	98.5
Yes	746	1.5

	OptOutMailings	Percentage
No	49006	98.5
Yes	746	1.5

4.3.10. OWNS A COMPUTER

It was a strange observation to note that 81.14% of the customers did not own a computer. We can infer that nearly 80% of the users may rely on wireless service for accessing the internet. Hence, suggestions to provide broadband along with mobile phone services and curated packages for internet can be offered by the company to retain and widen the customer base.

```
Total response where Owns Compter was No and Churn was Yes :
 11569
In Percentage Terms out of Total Churn (Yes):
 81.21 %
In Percentage Terms out of those who do not own a computer :
 36.81 %
```



4.3.11. HOME OWNERSHIP

32.2 % of the Home ownership status came out to be 'Unkown'. If customers do not own a home, the possibility for Churn is high as they may be moving to different location. It is important to focus on this feature.

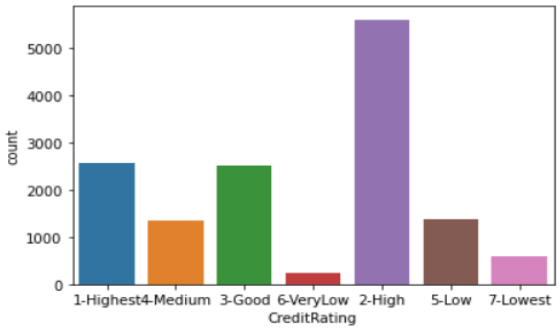
	Homeownership	Percentage
Known	33725	67.79
Unknown	16027	32.21

```
Total response where Home Ownership was Unknown and Churn was Yes :
 4733
In Percentage Terms out of Total Churn (Yes):
 33.23 %
In Percentage Terms out of Uknown Homewonership :
 29.53 %
```

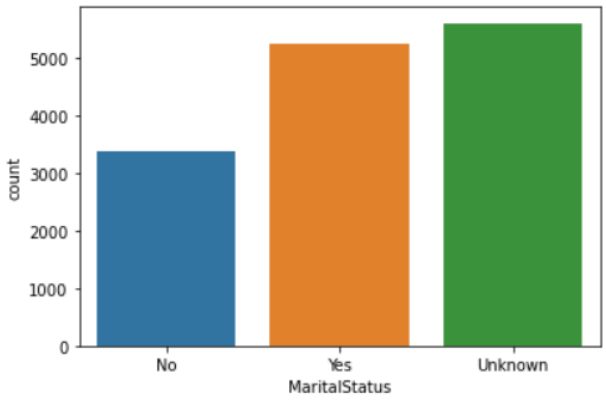
4.3.12. MADE RETENTION CALL TO TEAM

Only 4.96% of the customers had made a retention call to the team out of the total churners. Hence, the customer care team has to contact the customers to understand the reason for their decision to churn.

4.3.13. CREDIT RATING



4.3.14. MARITAL STATUS



4.3.15. OCCUPATION

The occupation for nearly 73% of the customers is categorized under ‘other’.

	Occupation	Percentage
Other	36453	73.27
Professional	8681	17.45
Crafts	1507	3.03
Clerical	979	1.97
Self	872	1.75
Retired	726	1.46
Student	377	0.76
Homemaker	157	0.32

4.4. ANALYSIS OF VARIOUS NUMERICAL FEATURES

4.4.1. MONTHLY REVENUE, MONTHLY RECURRING CHARGE AND MONTHLY MINUTES

After Feature Engineering the following DataFrame was obtained. It can be observed that the company earns a total Monthly Revenue of 2.92 Million USD. If those who responded 'Yes' to Churn leave the network then it stands to lose 28.22 % of the monthly revenue.

	Total	Potential Loss from Churners	Potential Loss Percentage
Monthly Revenue	2921306.78	824404.41	28.22
Monthly Minutes	26120868.00	6882672.00	26.35
Monthly Recurring Charge	2330133.00	635667.00	27.28

Similarly, the company stands to lose 27.28% of the monthly recurring charge if the customers leave.

4.4.2. PERCENTAGE CHANGE IN REVENUES AND MINUTES

It has been observed that both Revenues and Minutes spent on call has decreased significantly. It is vital to identify the reasons and come up with solutions to reduce the rate of decrease.

4.4.3. MONTHS IN SERVICE

It was identified that 61 months (5 years) was the maximum time period a user has subscribed to the network. Hence a DataFrame was created to identify the number of customers who plan to leave the network and their corresponding time period using the service.

	Time	Number of Churners
0	Less than 1 year	3316
1	1 to 2 years	7122
2	2 to 3 years	2934
3	3 to 4 years	704
4	4 to 5 years	166
5	Greater than 5 years	3

4.4.4. CURRENT EQUIPMENT DAYS

The mean number of days for which a mobile handset was handled comes out to be 380.

Neary 46% of the Churners' had mobile equipment for greater than 380 days. This implies that they maybe planning to get a new mobile handset. Hence, tie-ups with mobile phone manufacturers could be planned to reduce the Churn rate.

4.4.5. VALUE ADDED SERVICES

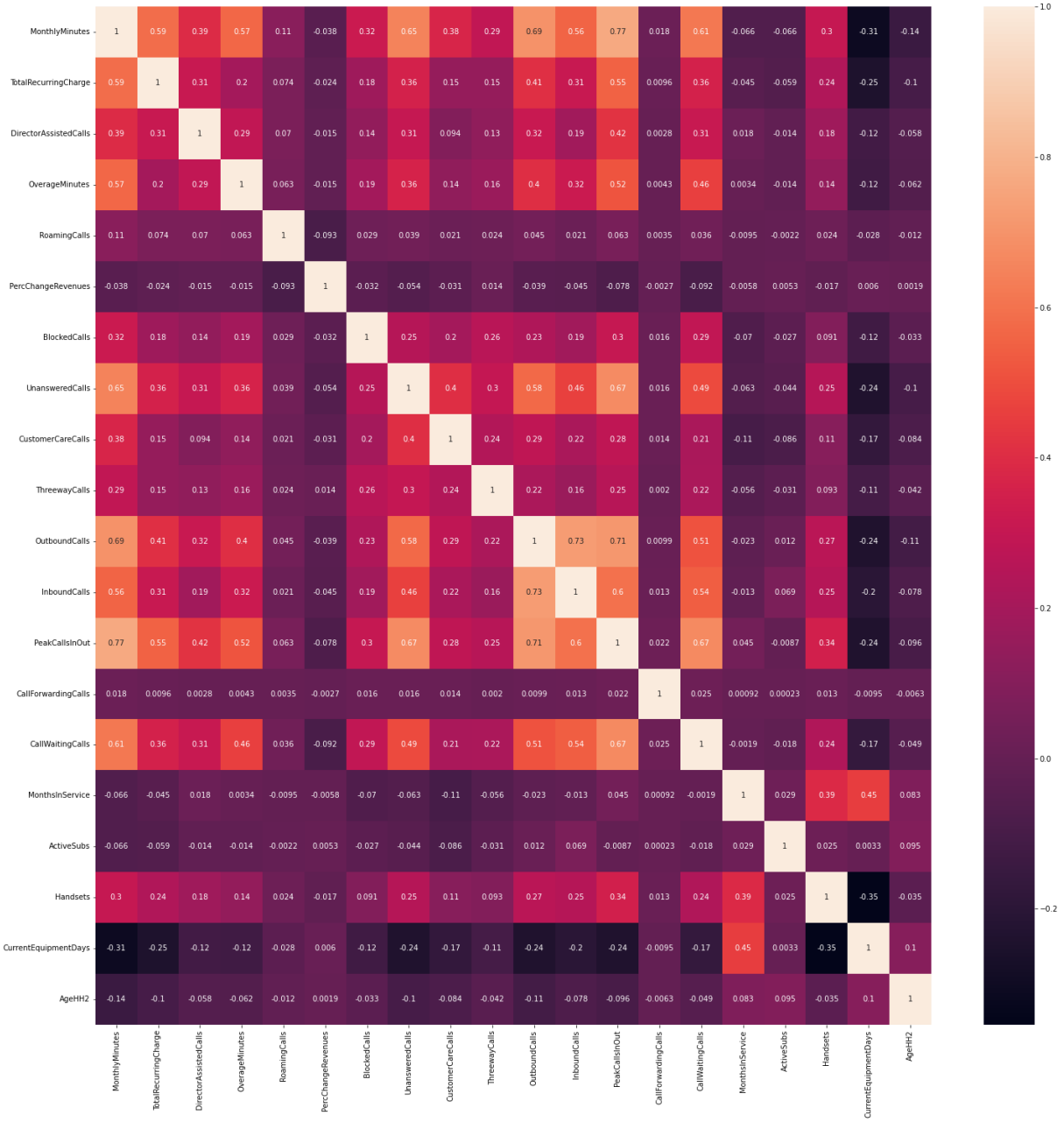
Value Added Services like DirectorAssistedCalls, OverageMinutes, Roaming Calls, ThreewayCalls and CallForwarding were analysed with respect to the number of customers who had responded 'Yes' to Churn. Also, we had decided to convert these value added services into binary before model building.

```
DirectorAssistedCalls

Number of users who hasn't used  DirectorAssistedCalls but still opting out :
7089
Percentage out of those who said Yes :  49.76
Percentage out of total users :  14.25

OverageMinutes

Number of users who hasn't used  OverageMinutes but still opting out : 6417
Percentage out of those who said Yes :  45.05
Percentage out of total users :  12.9

RoamingCalls

Number of users who hasn't used  RoamingCalls but still opting out : 9825
Percentage out of those who said Yes :  68.97
Percentage out of total users :  19.75

ThreewayCalls

Number of users who hasn't used  ThreewayCalls but still opting out : 10643
Percentage out of those who said Yes :  74.71
Percentage out of total users :  21.39

CallForwardingCalls

Number of users who hasn't used  CallForwardingCalls but still opting out : 1
4187
Percentage out of those who said Yes :  99.59
Percentage out of total users :  28.52
```

4.4.6. RETENTION OFFERS ACCEPTED

Only 2.32 % of the total Churners had accepted to stay back. It is vital to understand the demands of every customer and implement necessary changes to retain a significant portion of them.

4.4.7. CORRELATION MATRIX



The correlation matrix has been generated after iterating through columns and reducing the multicollinearity between them by dropping different columns.

5. MODEL BUILDING

Various features were dropped as discussed in the EDA. The final features after encoding came out to be 49 and a Logistic Regression Model was built as part of interim report preparation. The train and Test data performance is shown below.

Classification Report for Train Data :

	precision	recall	f1-score	support
0	0.72	0.99	0.83	24865
1	0.44	0.01	0.02	9961
accuracy			0.71	34826
macro avg	0.58	0.50	0.43	34826
weighted avg	0.64	0.71	0.60	34826

Classification Report for Test Data :

	precision	recall	f1-score	support
0	0.71	0.99	0.83	10642
1	0.41	0.01	0.02	4284
accuracy			0.71	14926
macro avg	0.56	0.50	0.42	14926
weighted avg	0.63	0.71	0.60	14926

With an accuracy of 71% for both train and test data, further analysis using other algorithms like RandomForest, KNeighborsClassifier, NaiveBayes, etc and feature reshaping tools like LDA and PCA are yet to be done to arrive at the best possible working machine learning model and suggest solutions to retain customers. However, the F1 Score seems to be poor.

One of the reasons for poor F1 Score could be because of imbalance in the target variable where ‘Yes’ is about 30% while ‘No’ is about 70%. So we intend to SMOTE the data in furter analysis.

6. VARIANCE INFLATION FACTOR

Variance Inflation Factor gives the multicollinearity among the variables. High multicollinearity is considered bad for model building. Through multiple iterations, some of the columns were dropped and the following features were chosen as final independent features.

	feature	VIF
0	MonthlyMinutes	3.749730
1	TotalRecurringCharge	1.731368
2	PercChangeRevenues	1.009128
3	BlockedCalls	1.167577
4	UnansweredCalls	2.221752
5	CustomerCareCalls	1.332433
6	OutboundCalls	3.155269
7	InboundCalls	2.238011
8	PeakCallsInOut	3.743935
9	MonthsInService	2.337491
10	ActiveSubs	1.075503
11	Handsets	2.214138
12	CurrentEquipmentDays	2.332223
13	AgeHH2	1.688184
14	DirectorAssistedCalls	2.364013
15	OverageMinutes	2.731527
16	RoamingCalls	1.512643
17	ThreewayCalls	1.591826
18	CallForwardingCalls	1.012891
19	CallWaitingCalls	2.709275
20	ChildrenInHH	1.748560
21	HandsetRefurbished	1.248317
22	HandsetWebCapable	8.221456
23	TruckOwner	2.218421
24	RVOwner	1.771550
25	Homeownership	9.116523
26	BuysViaMailOrder	2.556108
27	OptOutMailings	1.028935
28	OwnsComputer	1.533304
29	HasCreditCard	8.935971
30	NewCellphoneUser	1.246390
31	CreditRating	9.383524

7. LOGISTIC REGRESSION

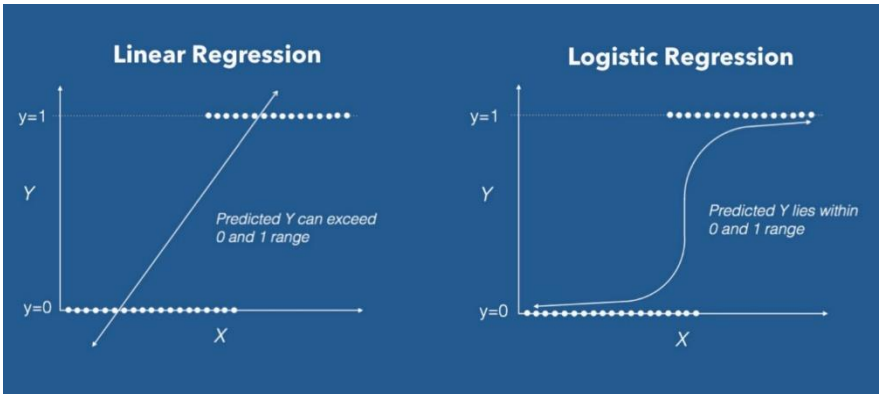
Logistic regression is a classification algorithm used to assign observations to a discrete set of classes. Some of the examples of classification problems are Email spam or not spam, Online transactions Fraud or not Fraud, Tumor Malignant or Benign. Logistic regression transforms its output using the logistic sigmoid function to return a probability value.

What are the types of logistic regression

1. Binary (eg. Tumor Malignant or Benign)

2. Multi-linear functions failsClass (eg. Cats, dogs or Sheep's)

Logistic Regression is a Machine Learning algorithm which is used for the classification problems, it is a predictive analysis algorithm and based on the concept of probability.



Linear Regression VS Logistic Regression Graph| Image: Data Camp

We can call a Logistic Regression a Linear Regression model but the Logistic Regression uses a more complex cost function, this cost function can be defined as the '**Sigmoid function**' or also known as the 'logistic function' instead of a linear function.

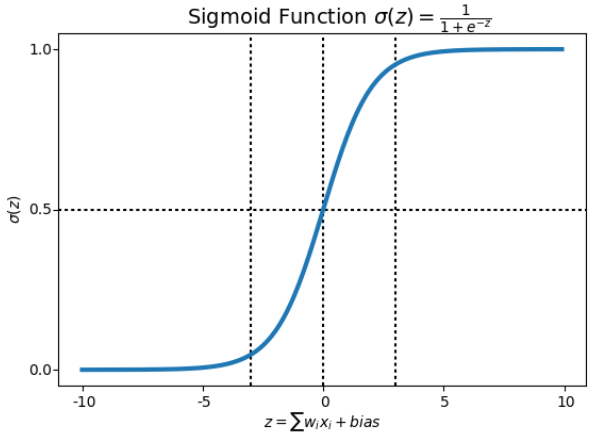
The hypothesis of logistic regression tends it to limit the cost function between 0 and 1. Therefore linear functions fail to represent it as it can have a value greater than 1 or less than 0 which is not possible as per the hypothesis of logistic regression.

$$0 \leq h_\theta(x) \leq 1$$

Logistic regression hypothesis expectation

What is the Sigmoid Function?

In order to map predicted values to probabilities, we use the Sigmoid function. The function maps any real value into another value between 0 and 1. In machine learning, we use sigmoid to map predictions to probabilities.



Sigmoid Function Graph

$$f(x) = \frac{1}{1+e^{-(x)}}$$

7.1. CLASSIFICATION REPORT

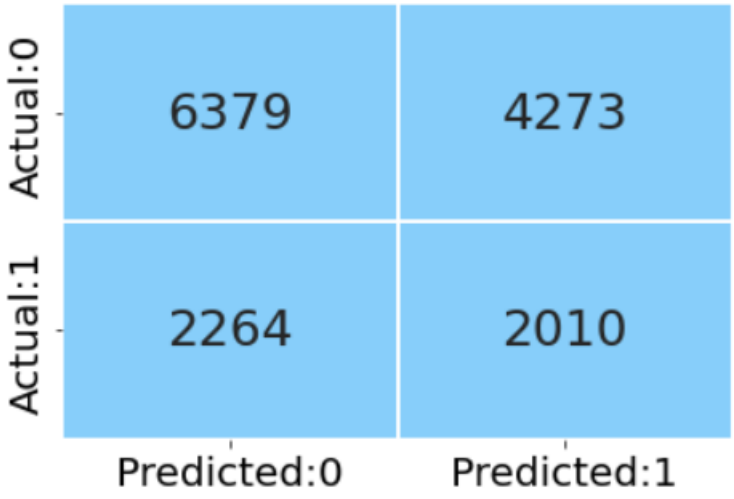
Classification Report for Train Data :

	precision	recall	f1-score	support
0	0.62	0.60	0.61	24855
1	0.62	0.64	0.63	24855
accuracy			0.62	49710
macro avg	0.62	0.62	0.62	49710
weighted avg	0.62	0.62	0.62	49710

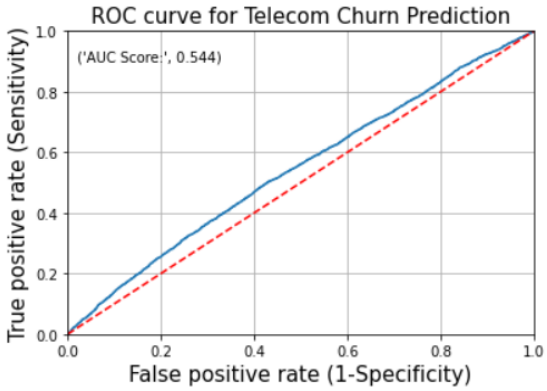
Classification Report for Test Data :

	precision	recall	f1-score	support
0	0.74	0.60	0.66	10652
1	0.32	0.47	0.38	4274
accuracy			0.56	14926
macro avg	0.53	0.53	0.52	14926
weighted avg	0.62	0.56	0.58	14926

7.2. CONFUSION MATRIX

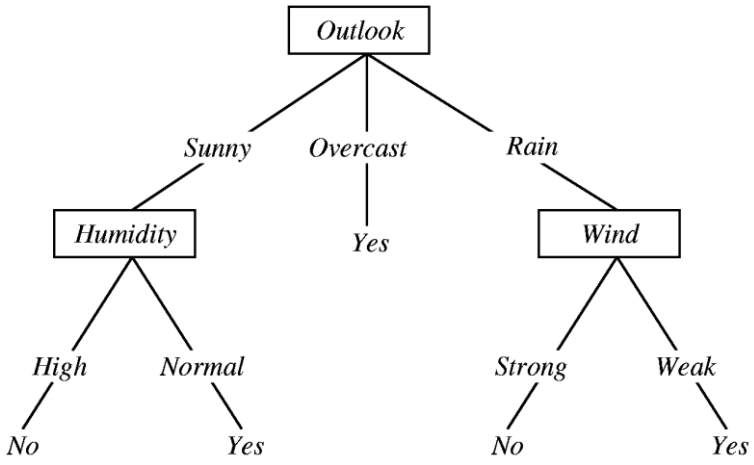


7.3. ROC CURVE



8. DECISION TREE ALGORITHM

A decision tree is a flowchart-like structure in which each internal node represents a test on a feature (e.g. whether a coin flip comes up heads or tails) , each leaf node represents a class label (decision taken after computing all features) and branches represent conjunctions of features that lead to those class labels. The paths from root to leaf represent classification rules. Below diagram illustrate the basic flow of decision tree for decision making with labels (Rain(Yes), No Rain(No)).



Decision tree is one of the predictive modelling approaches used in statistics, data mining and machine learning.

Decision trees are constructed via an algorithmic approach that identifies ways to split a data set based on different conditions. It is one of the most widely used and practical methods for supervised learning. Decision Trees are a non-parametric **supervised learning** method used for both **classification** and **regression** tasks.

Tree models where the target variable can take a discrete set of values are called **classification trees**. Decision trees where the target variable can take continuous values (typically real numbers) are called **regression trees**. Classification And Regression Tree (CART) is general term for this.

8.1. CLASSIFICATION REPORT

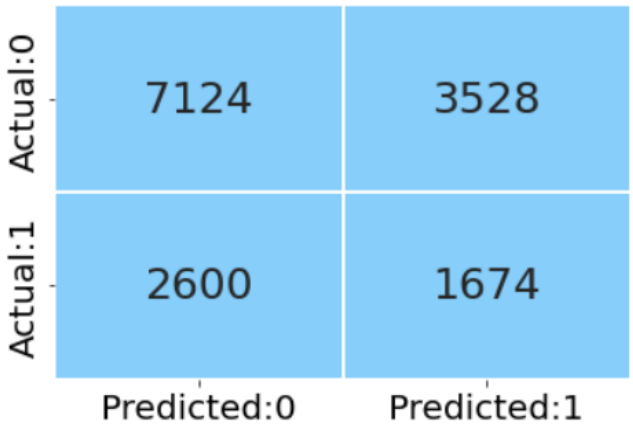
Classification Report for Train Data :

	precision	recall	f1-score	support
0	1.00	1.00	1.00	24855
1	1.00	1.00	1.00	24855
accuracy			1.00	49710
macro avg	1.00	1.00	1.00	49710
weighted avg	1.00	1.00	1.00	49710

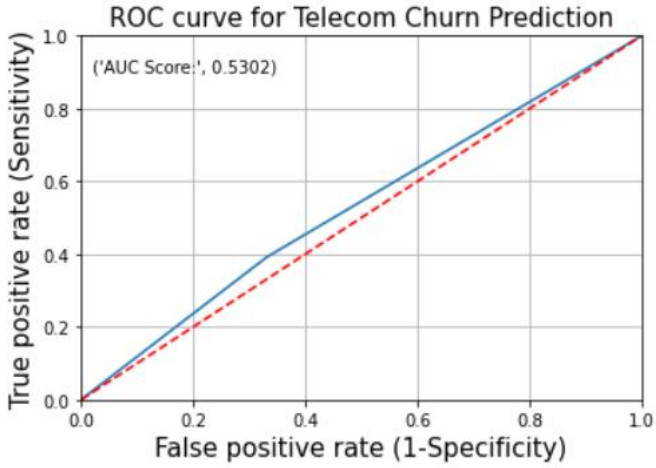
Classification Report for Test Data :

	precision	recall	f1-score	support
0	0.73	0.67	0.70	10652
1	0.32	0.39	0.35	4274
accuracy			0.59	14926
macro avg	0.53	0.53	0.53	14926
weighted avg	0.61	0.59	0.60	14926

8.2. CONFUSION MATRIX



8.3. ROC CURVE

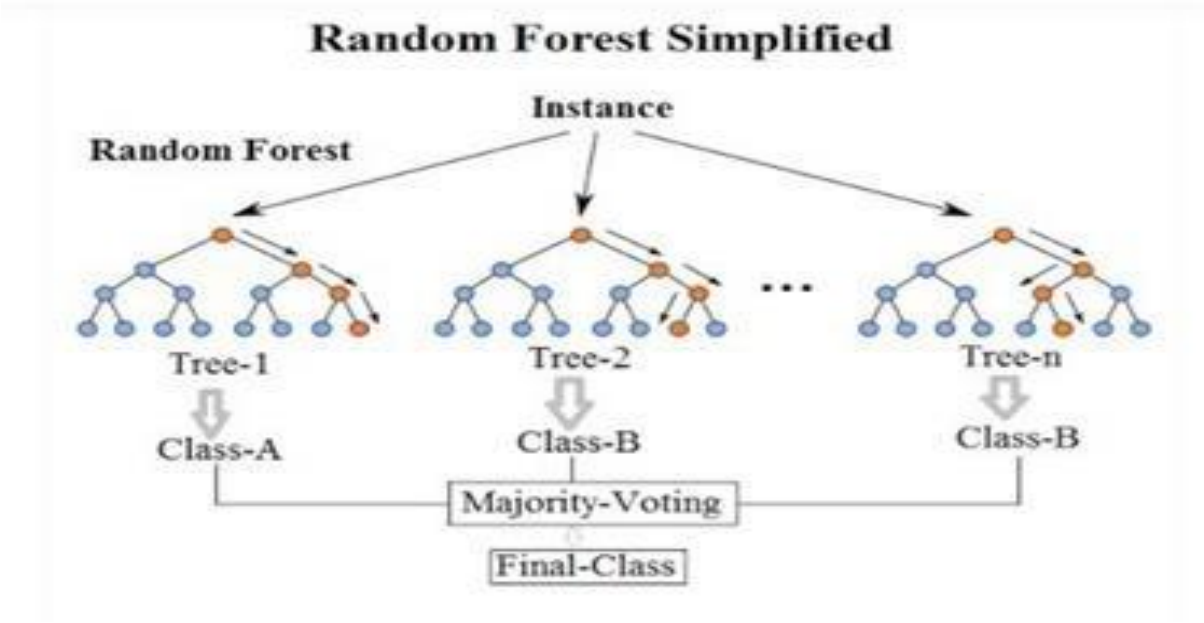


9. RANDOM FOREST CLASSIFIER

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.

To say it in simple words: Random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction.

One big advantage of random forest is, that it can be used for both classification and regression problems.



Random Forest has nearly the same hyper parameters as a decision tree or a bagging classifier.

Fortunately, we don't have to combine a decision tree with a bagging classifier and can just easily use the classifier-class of Random Forest. Like I already said, with Random Forest, you can also deal with Regression tasks by using the Random Forest repressor.

Random Forest adds additional randomness to the model, while growing the trees. Instead of searching for the most important feature while splitting a node, it searches for the best feature among a random subset of features. This results in a wide diversity that generally results in a better model.

Advantages of Random Forest:

- There is no need for feature normalization
- Individual decision trees can be trained in parallel
- Reduced overfitting
- Require almost no input preparation
- Performs implicit feature selection

- It’s very quick to train

Modeling and Predicting Online Purchasing Intention of Shopper

Disadvantages of Random Forest:

- No interpretability

9.1. CLASSIFICATION REPORT

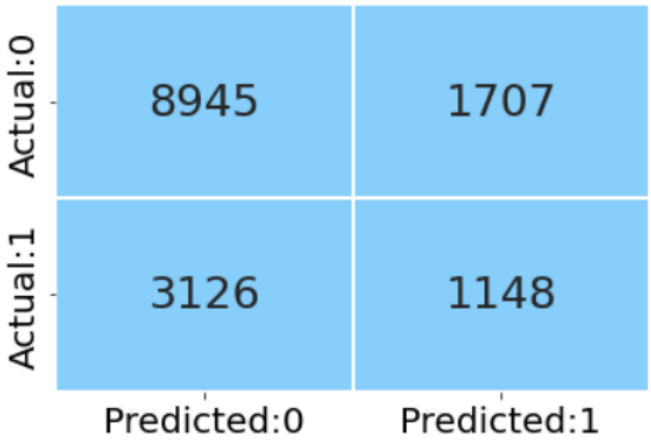
Classification Report for Train Data :

	precision	recall	f1-score	support
0	1.00	1.00	1.00	24855
1	1.00	1.00	1.00	24855
accuracy			1.00	49710
macro avg	1.00	1.00	1.00	49710
weighted avg	1.00	1.00	1.00	49710

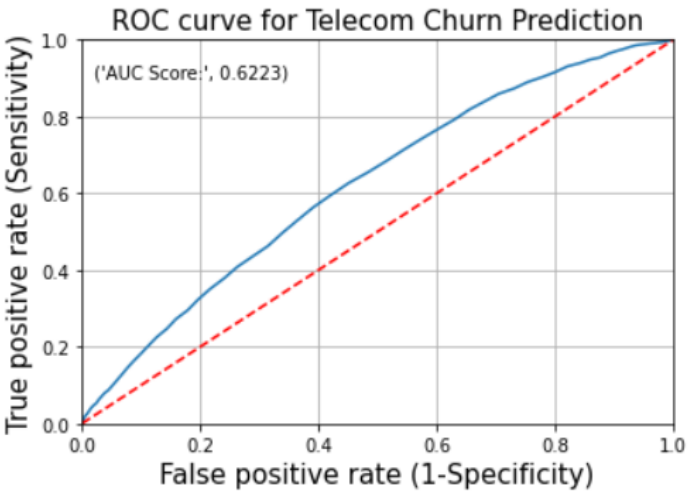
Classification Report for Test Data :

	precision	recall	f1-score	support
0	0.74	0.84	0.79	10652
1	0.40	0.27	0.32	4274
accuracy			0.68	14926
macro avg	0.57	0.55	0.55	14926
weighted avg	0.64	0.68	0.65	14926

9.2. CONFUSION MATRIX



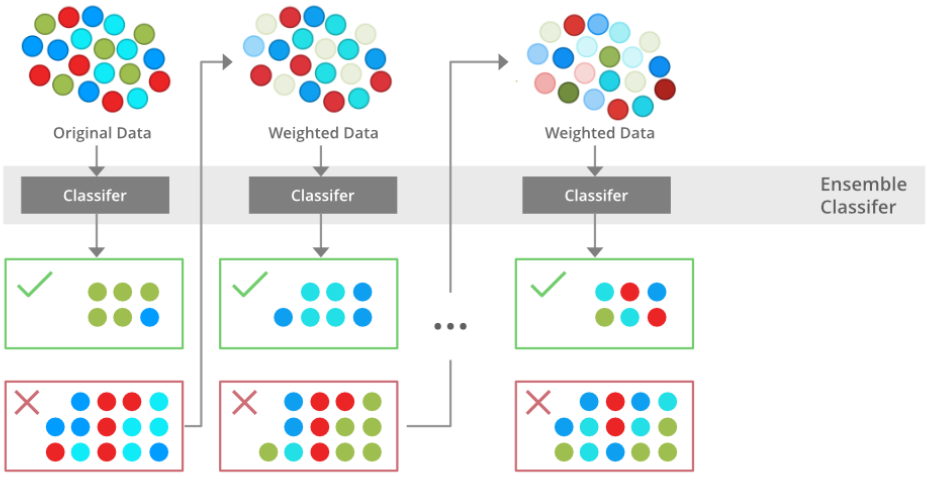
9.3. ROC CURVE



10. XG BOOST

XGBoost stands for Extreme Gradient Boosting, which was proposed by the researchers at the University of Washington. It is a library written in C++ which optimizes the training for Gradient Boosting.

Boosting is an ensemble modelling, technique that attempts to build a strong classifier from the number of weak classifiers. It is done by building a model by using weak models in series. Firstly, a model is built from the training data. Then the second model is built which tries to correct the errors present in the first model. This procedure is continued and models are added



until either the complete training data set is predicted correctly or the maximum number of models are added.

In XGBoost, decision trees are created in sequential form. Weights play an important role in XGBoost. Weights are assigned to all the independent variables which are then fed into the decision tree which predicts results. The weight of variables predicted wrong by the tree is increased and these variables are then fed to the second decision tree. These individual classifiers/predictors then ensemble to give a strong and more precise model. It can work on regression, classification, ranking, and user-defined prediction problems.

10.1. CLASSIFICATION REPORT

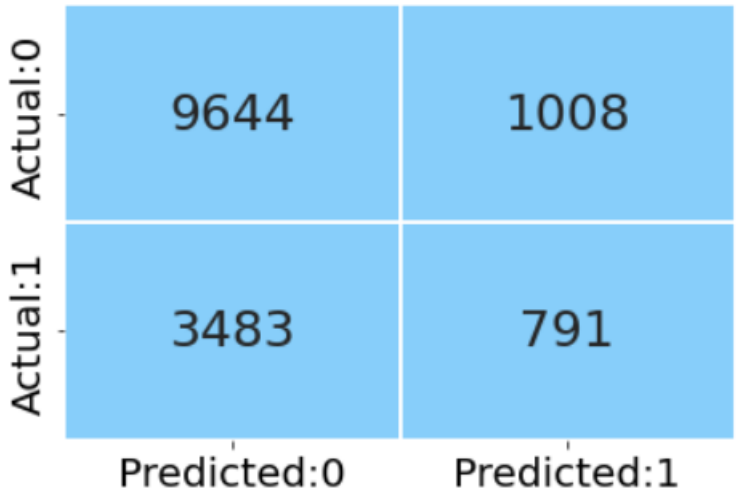
Classification Report for Train Data :

	precision	recall	f1-score	support
0	0.79	0.96	0.87	24855
1	0.95	0.75	0.83	24855
accuracy			0.85	49710
macro avg	0.87	0.85	0.85	49710
weighted avg	0.87	0.85	0.85	49710

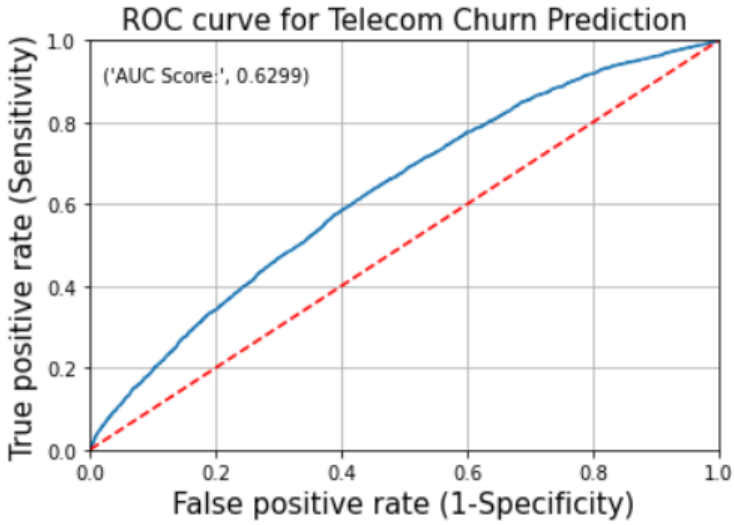
Classification Report for Test Data :

	precision	recall	f1-score	support
0	0.73	0.91	0.81	10652
1	0.44	0.19	0.26	4274
accuracy			0.70	14926
macro avg	0.59	0.55	0.54	14926
weighted avg	0.65	0.70	0.65	14926

10.2. CLASSIFICATION REPORT



10.3. ROC CURVE



11. K-NEAREST NEIGHBOURS

The KNN algorithm assumes that similar things exist in close proximity. In other words, similar things are near to each other.

"Birds of a feather flock together.

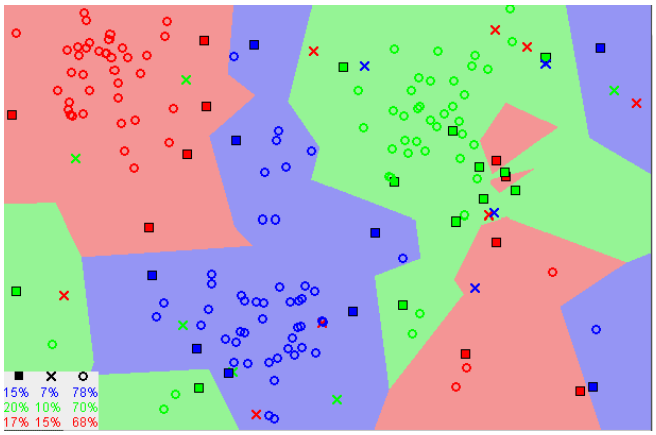


Image showing how similar data points typically exist close to each other

Notice in the image above that most of the time, similar data points are close to each other. The KNN algorithm hinges on this assumption being true enough for the algorithm to be useful. KNN captures the idea of similarity (sometimes called distance, proximity, or closeness) with some mathematics we might have learned in our childhood— calculating the distance between points on a graph.

KNN's main disadvantage of becoming significantly slower as the volume of data increases makes it an impractical choice in environments where predictions need to be made rapidly. Moreover, there are faster algorithms that can produce more accurate classification and regression results.

11.1. CLASSIFICATION REPORT

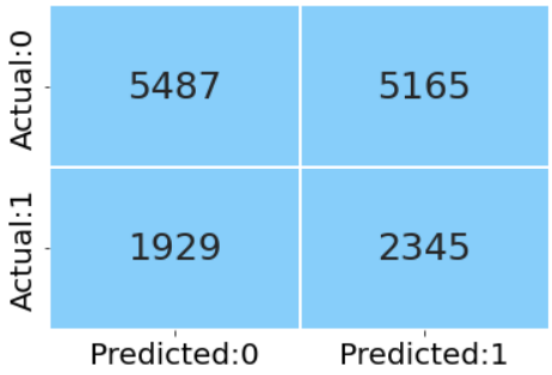
Classification Report for Train Data :

	precision	recall	f1-score	support
0	0.94	0.66	0.77	24855
1	0.74	0.96	0.83	24855
accuracy			0.81	49710
macro avg	0.84	0.81	0.80	49710
weighted avg	0.84	0.81	0.80	49710

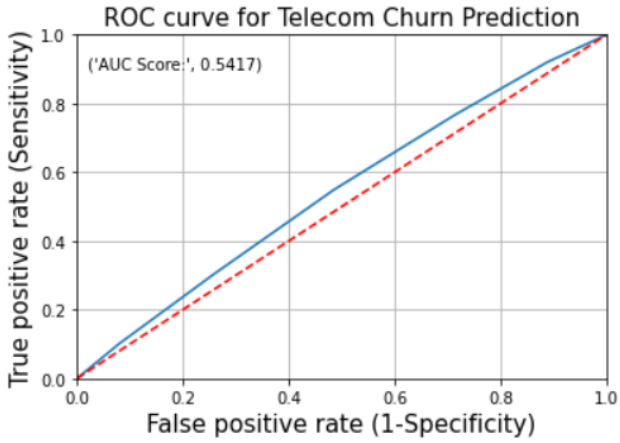
Classification Report for Test Data :

	precision	recall	f1-score	support
0	0.74	0.52	0.61	10652
1	0.31	0.55	0.40	4274
accuracy			0.52	14926
macro avg	0.53	0.53	0.50	14926
weighted avg	0.62	0.52	0.55	14926

11.2. CONFUSION MATRIX



11.3. ROC CURVE



12. NAÏVE BAYES

Naive Bayes classifiers are a collection of classification algorithms based on **Bayes' Theorem**. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e. every pair of features being classified is independent of each other. The fundamental Naive Bayes assumption is that each feature makes an:

- independent
- equal

contribution to the outcome.

Bayes' Theorem

Bayes' Theorem finds the probability of an event occurring given the probability of another event that has already occurred. Bayes' theorem is stated mathematically as the following equation:

$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)}$$

A, B = events
$P(A|B)$ = probability of A given B is true
$P(B|A)$ = probability of B given A is true
$P(A), P(B)$ = the independent probabilities of A and B

where A and B are events and P(B) ≠ 0.

- Basically, we are trying to find probability of event A, given the event B is true. Event B is also termed as **evidence**.
- P(A) is the **priori** of A (the prior probability, i.e. Probability of event before evidence is seen). The evidence is an attribute value of an unknown instance(here, it is event B).
- P(A|B) is a posteriori probability of B, i.e. probability of event after evidence is seen.

12.1. CLASSIFICATION REPORT

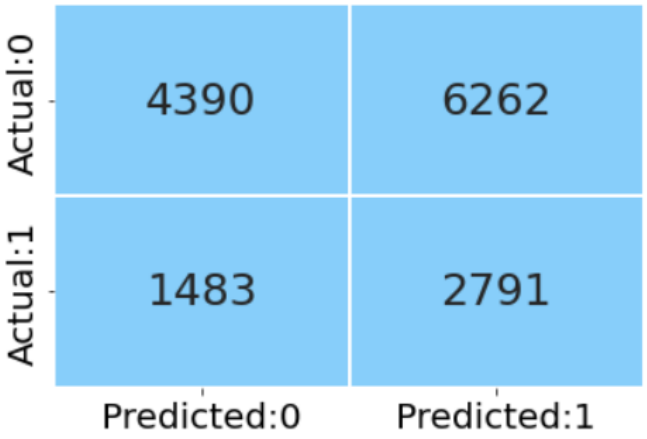
Classification Report for Train Data :

	precision	recall	f1-score	support
0	0.64	0.41	0.50	24855
1	0.57	0.77	0.65	24855
accuracy			0.59	49710
macro avg	0.61	0.59	0.58	49710
weighted avg	0.61	0.59	0.58	49710

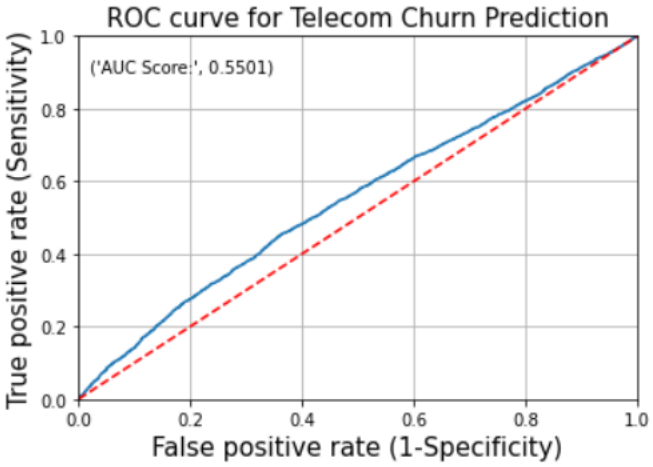
Classification Report for Test Data :

	precision	recall	f1-score	support
0	0.75	0.41	0.53	10652
1	0.31	0.65	0.42	4274
accuracy			0.48	14926
macro avg	0.53	0.53	0.48	14926
weighted avg	0.62	0.48	0.50	14926

12.2. CONFUSION MATRIX



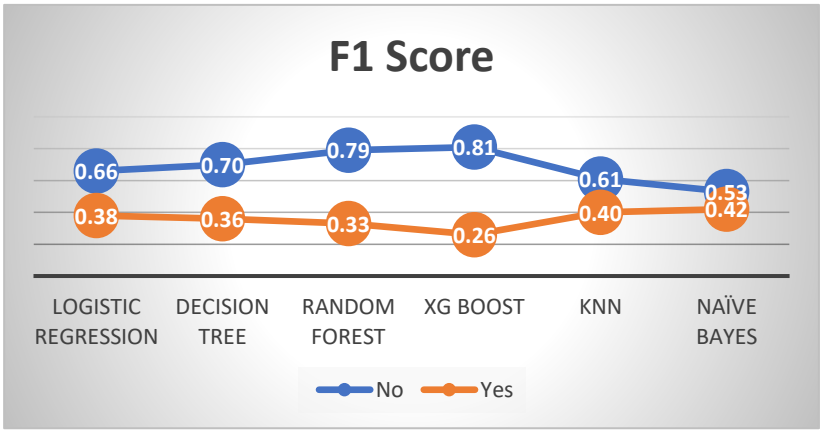
12.3. ROC CURVE

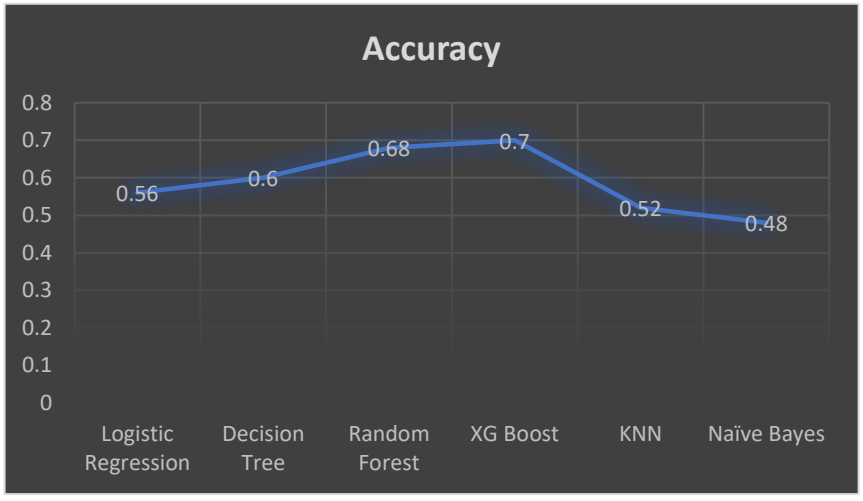


13. SUMMARY OF THE FINDINGS

13.1. USING VIF TECHNIQUE DERIVED FEATURES

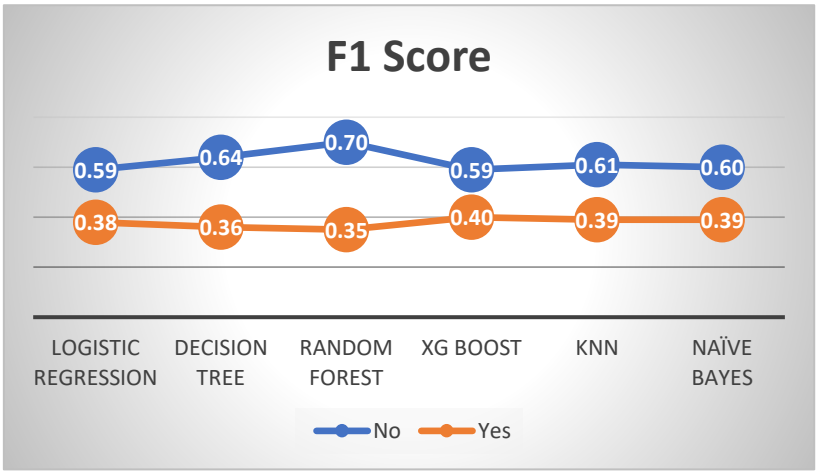
	F1 Score		Accuracy
	No	Yes	
Logistic Regression	0.66	0.38	0.56
Decision Tree	0.70	0.36	0.60
Random Forest	0.79	0.33	0.68
XG Boost	0.81	0.26	0.70
KNN	0.61	0.40	0.52
Naïve Bayes	0.53	0.42	0.48

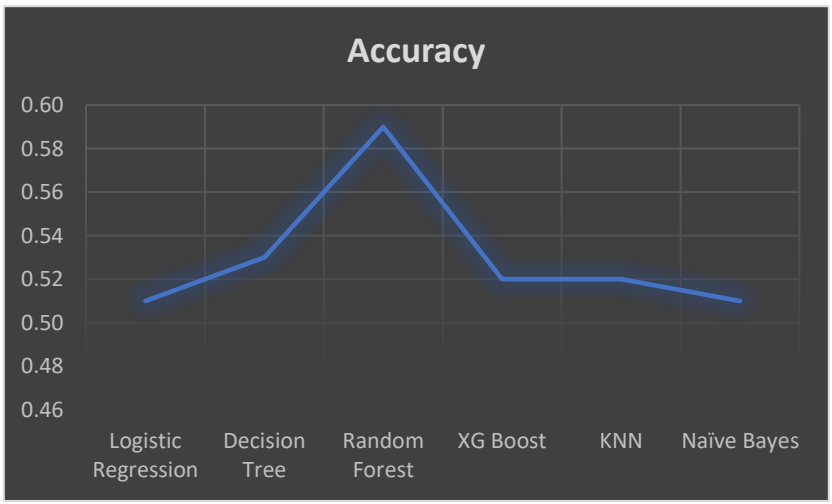




13.2. USING PRINCIPAL COMPONENT ANALYSIS (13 COMPONENTS)

	F1 Score		Accuracy
	No	Yes	
Logistic Regression	0.59	0.38	0.51
Decision Tree	0.64	0.36	0.53
Random Forest	0.70	0.35	0.59
XG Boost	0.59	0.40	0.52
KNN	0.61	0.39	0.52
Naïve Bayes	0.60	0.39	0.51





From the above we can find that DecisionTree and RandomForest are overfitted due to 1.00 in Train Data set. The XGBoost algorithm for VIF selected features seems to be performing good.

14. SUGGESTIONS FOR REDUCING CHURN

Based on EDA observations the following suggestions have been made:

- Targeted customers who may be switching to new phones:
    - 10 % (12.66% Churners) customers not using web capable phones.
    - 13.87% (15.51% Churners) customers using Handset Refurbished.
    - 46% of Churners had mobile equipment greater than 380 days.

  Curated discounts for customers who might be buying a new mobile phone should be given by the company.

- Roaming offers for those who might be travelling frequently.
- Anniversary offers for customers who complete every year with the company.
- Awareness to customers about value added services like Call Forwarding, Threeway Calls, Director Assisted Calls, etc.
- Training of customer care staff to handle retention calls.

15. REFERENCES

The references can be blogs, articles or even social media news relevant to explain the importance of the projects.

- Data Source:
    - https://www.kaggle.com/datasets/jpacse/datasets-for-churn-telecom
- Others:
    - https://towardsdatascience.com/telco-customer-churnrate-analysis-d412f208cbbf
    - https://strategicmarketingpartner.com/recency-frequency-monetary-rfm-marketing-analysis/?gclid=Cj0KCQjw4uaUBhC8ARIsANUuDjVtBG2WIE_3iCxrnd7TPT6si5QXVdCpK4upJl-8H0oCpzsAKBCdHNUaArrpEALw_wcB
    - https://paginas.fe.up.pt/~ec/files_0405/slides/02%20CRISP.pdf
    - https://en.wikipedia.org/wiki/Cross-industry_standard_process_for_data_mining
    - https://towardsdatascience.com/measuring-users-with-customer-lifetime-value-cltv-94fccb4e532e

Notes For Project Team

Sample Reference for Datasets (to be filled by team and mentor)

Original owner of data	Kaggle
Data set information	This dataset is unprocessed and a balanced version provided for analyzing Process. Consists of 71,047 instances and 58 attributes.

Any past relevant articles using the dataset	No past articles found on the dataset
Reference	Kaggle
Link to web page	https://www.kaggle.com/datasets/jpacse/datasets-for-churn-telecom