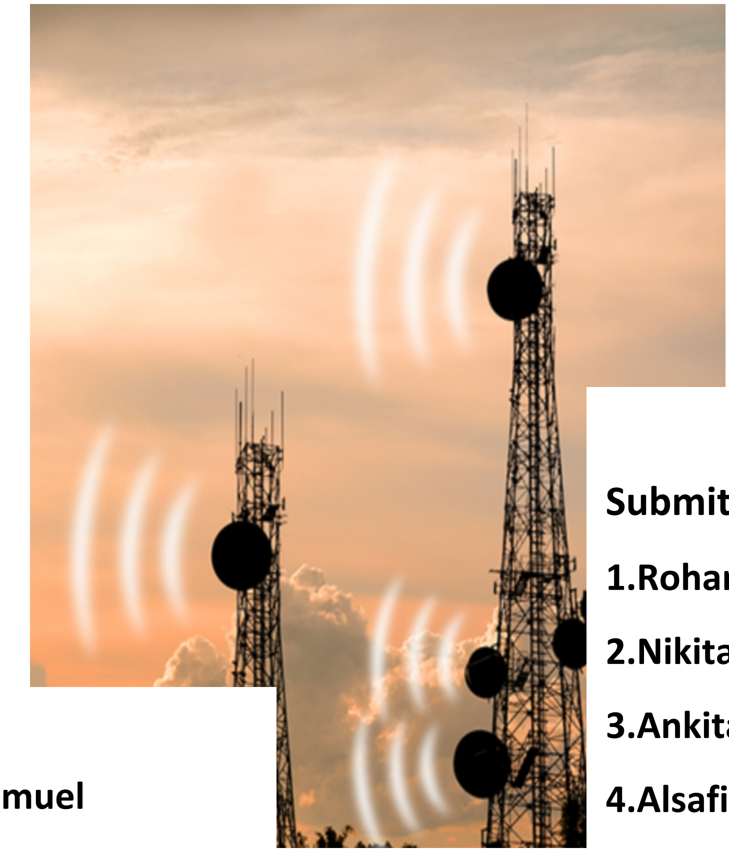
ABSTRACT

As part of this analysis, we plan to identify churners so that we can take measures to avoid churn and retain the customers. Further predictive analysis and identifying the data around



Submitted By:

1.Rohan Kumariya (1002007078)

2.Nikita Mhatre (1001967565)

3.Ankita Phadke (1001963837)

4.Alsafin Samnani (1001979027)

Professor:

Jayarajan Samuel

Table of Contents

a. Executive summary and purpose of report:

1. Executive summary:

The Tele-communications industry is of crucial importance to the United States, service providers are inclined more towards expansion of the subscriber base. The business revenue is directly proportional to customer base and it's far more feasible to retain existing customer base than acquiring new customers. To meet the need of surviving in the competitive environment, the retention of existing customers has become a huge challenge. It is stated that the cost of acquiring a new customer is far more than that for retaining the existing one. Therefore, it is imperative for the telecom industries to use advanced analytics to understand consumer behavior and in-turn predict the association of the customers as to whether they will leave the company. This analysis and findings will in turn help the telecom providers to device strategies and retain the customers.

2. Key takeaway:

Below are the key takeaways from this report:

1. The customer churn findings and analysis would help telecom company to track the important metrics of 'churn' and take necessary actions in time.
2. Incentives suggested at the end of this analysis would help outline factors around which new product and service schemes can be devised.

b. Motivation, business problem and overview of data:

1. Motivation:

Customer churn rate happens to be one of the top-rated metrics when it comes to evaluating the success of a telecom company. Lost customers translate into lost revenue and if that continues, it may cost the company serious consequences on its main business baseline. Although customer churn is inevitable, tracking and improving the churn rate would help bring down the churn and help the company survive in the competitive age. The customer life with a telecom service provider would mainly depend on the below factors:

- Service quality
- Availability of features
- Competitive advantage given by other companies
- Customer service model and turnaround time

2. Business problem:

It is imperative for the telecom industries to use advanced analytics to understand consumer behavior and in-turn predict the association of customers as to whether they will leave the company in future. As part of this analysis, we plan to identify churners so that we can take measures to avoid churn and retain the customers. Further predictive analysis and identifying the data around such customers will let us decide which customers might churn in future. This analysis and findings will in turn help the company to device strategies and retain the customers. We, as a telecom company, would like to build analysis around retaining customers, identify the threshold values for variables like the calls rates, monthly usage amounts, bill charges, international charges, above and below which we would try to provide incentive schemes and offers that will help to retain the identified customers who might leave the company services in future. Below are four main areas of the problem statements:

- Analyze behavior to retain customers
- Predict customer churn trends
- Analyze all relevant customer data and develop focused customer retention programs
- Suggest actions to be taken to stop customers from porting

3. Overview of data:

This dataset consists of state-wise data of the United States based telecom company's customers information, including customer's telecom service details, usage, subscriptions, and churn data. Below data dictionary provides detailed information on the dataset, with details about each column.

Data transformation: We added new columns for per minute call rates in day, evening, night, which are equal to USD 0.2, 0.1, 0.05 respectively for each customer. These columns can be useful in designing incentive ideas to be offered to customers.

4. Data dictionary:

Sr. No.	Variable	Variable	Description
1	State	Text	Name of the State
2	Account length	Number	Number of weeks customer has had active account
3	Area code	Number	Coded area for each state
4	Phone number	Number	Contact number
5	International plan D1	Boolean	'Yes', if customer is subscribed to International plan, else 'No' – Dummy variable
6	Voicemail plan D2	Boolean	'Yes', if customer is subscribed to voicemail plan, else 'No' - Dummy variable
7	Number vmail messages	Number	Number of voice-mail messages sent
8	Total Day Minutes	Number	Call minutes used in daytime
9	Total Day Calls	Number	No. of calls made in day
10	Total Day Charge	Number	Charges for day calls
11	Rate per minute for day	Number	Rate charged by telecom company for day call per minute – Added new using rows 8 and 10 above
12	Total Evening Minutes	Number	Call minutes used in evening time
13	Total Evening Calls	Number	No. of calls made in evening

14	Total Evening Charge	Number	Charges for evening calls
15	Rate Per Minute for Evening	Number	Rate charged by telecom company for day call per minute – Added new using rows 12 and 14 above
16	Total Night Minutes	Number	Call minutes used in nighttime
17	Total Night Calls	Number	No. of calls made in night
18	Total Night Charge	Number	Charges for night calls
19	Rate per minute for Night	Number	Rate charged by telecom company for day call per minute – Added new using rows 16 and 18 above
20	Total domestic charges	Number	Total domestic call charges paid by customer
21	Total International Minutes	Number	No. of minutes spent on international calls
22	Total International Calls	Number	No. of international calls made
23	Total International Charge	Number	Total amount of money spent on international calls.
24	Customer Service Calls	Number	No. of calls made to customer service by customer
25	Churn D3	Boolean	True if customer cancelled service, False if not

Table. 1: Data dictionary

c. Methodology:

As the dependent variable is qualitative, we will use Logistic Regression and Classification tree. We are also planning to use these techniques to predict the customer traits which contribute to potential churners to avoid churn. As a part of pre-processing, we didn't find any missing values to deal with. We will make use of Histogram, Bar chart, Scatter plot for presenting the data analysis, trends, outcomes, and outliers if any. We combined the 3 separate daytime call charges i.e., day + evening + night call charges columns into one column named total domestic charges.

1. Data analysis using visualization and findings:

a) USA state-wise distribution of churners percentage:

Top 3 States with percentage of churners more than 25% of the total no. of users:
CA: 26.47%
NJ: 26.47%
TX: 25.00%

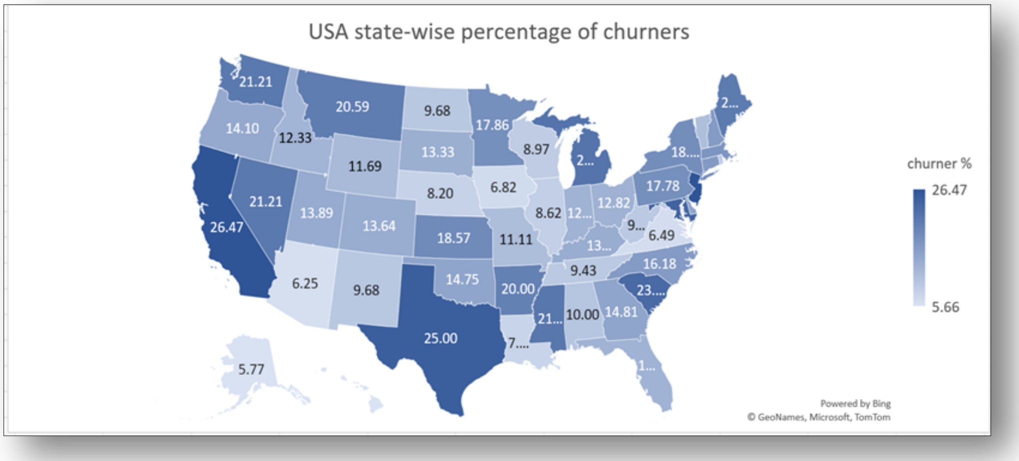


Figure 1:

USA state-wide churners %

b) Churners statistics as per subscription to international plan:

It can be concluded that, only 9% customers are subscribed to an international plan. Out of the customers who have subscribed to international plan, around 42% customers are currently churners.

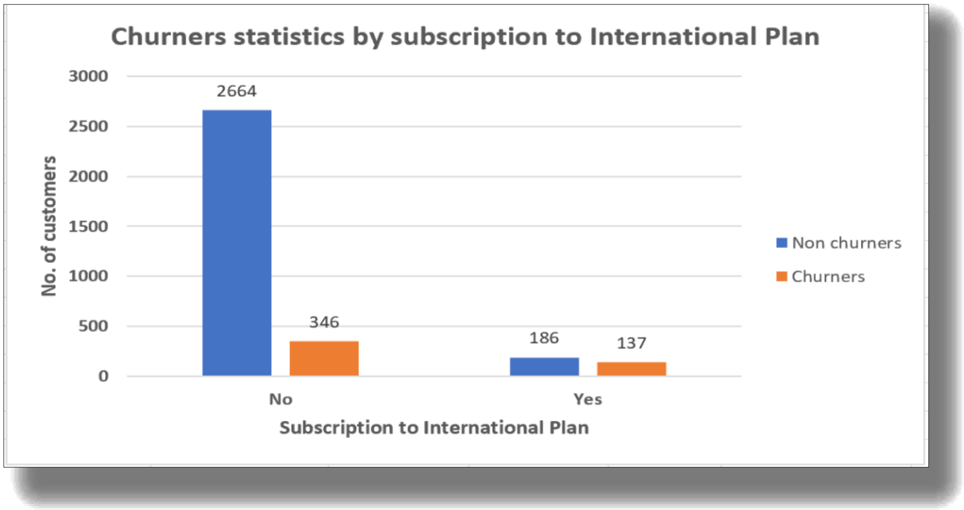


Figure 2: Churners statistics by international plan subscription

c) Churners statistics as per subscription to voicemail plan:

It can be concluded that, around 15% customers are subscribed to voicemail plan. Out of the customers who have subscribed to voicemail plan, around 16% customers are currently churners. Hence, we can say that the voicemail plan subscription is not related to churners trend by a great extent.

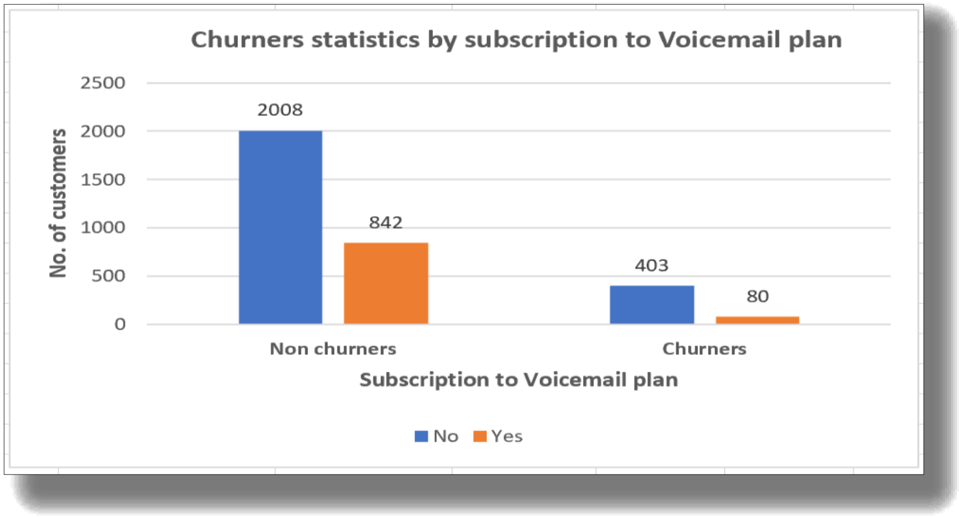


Figure 3: Churners statistics by voicemail plan subscription

2. Techniques used:

a) Logistic Regression analysis:

Logistic regression helps model probability of certain class or event happening. It gives relationship between the dependent variable and one or more independent variables by estimating probabilities using a logistic regression equation. The dependent variable being predicted here is the Churn – which has 2 possible values. Churn = 0 would mean that the customer would not churn and churn = 1 would mean that the customer would churn and leave the telecom service provider. Since logistic regression plots the output in form of S-curve, the model output values range between 0 and 1.
Tool used: SAS on demand

b) Classification tree analysis:

Classification tree is a classification and prediction technique, which constructs rules that classify data into churners (churn = 0) and non-churners (churn = 1) cases. Further, using the rules given by classification tree, it helps to predict any new incoming customers as churners or non-churners.
Tool used: Minitab

d. Results:

1. Initial Analysis:

Upon running initial tests on the data, we observed the below –

- Approximate churn rate is 14.5%.
- Upon running the data visualization techniques, we got the following observations:
  - Texas is one of the states having highest percentage of churn rate.
  - Area code 415 has highest percentage of churn rate among the states.
  - People who are not subscribed to international plan and voice mail plan are more likely to churn.

Number of Observations	3333
Number of binary/ categorical variables	3
Number of continuous variables	21
Outcome/target variable	Churn (currently at 14.49%)
Percentage of binary/categorical variable	1. Area Code: a. 415-> 49.65% b. 408-> 25.14% c. 510-> 25.20% 2. Churned customers: a. True-> 14.49% b. False-> 85.50%

Prediction of the target variable	Upon visualizing data, we noticed total domestic charges, voice mail subscription, international plan subscription and customer service calls have significant impact on churners. Here Churn is the target variable, where Churn rate for customers should be reduced to ~0% from 14.49%.
Significant predictor variables	1. State (distributed across 50 states) 2. Account length (1 to 243 months) 3. Total domestic charges (USD 20 to 92.56) 4. Total international charges (USD 0 to 5.4) 5. International plan subscription (9.6%) 6. Voicemail plan subscription (27.67%) 7. Customer service calls (0 to 9 calls per day)

Table 2: Initial analysis findings

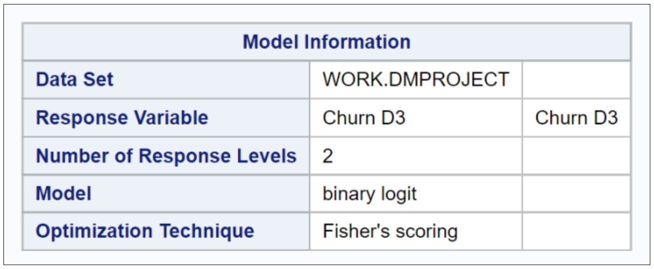
2. Prediction models and findings

a) Logistic regression Analysis:

We started with using eight independent variables as input to logistic regression analysis i.e., (1) Total domestic charges, (2) Total international charges, (3) Customer service calls, (4) Voicemail plan, (5) International plan, (6) State, (7) Account length, (8) No. of voicemail messages.

However, when we ran the Logistic regression in SAS studio, we found below 3 variables to be not significant, based on their p-value.
(6) State, (7) Account length, (8) No. of voicemail messages

Hence using backward elimination method, we removed the above 3 independent variables and finally retained below variables in the model.
(1) Total domestic charges, (2) Total international charges, (3) Customer service calls, (4) Voicemail plan, (5) International plan



Model Information		
Data Set	WORK.DMPROJECT	
Response Variable	Churn D3	Churn D3
Number of Response Levels	2	
Model	binary logit	
Optimization Technique	Fisher's scoring	

Figure 4: Logistic regression model details using SAS on demand

Based on the accuracy criteria, we are 95% confident that above mentioned 5 variables are significant with their p-value < 0.01. Even with this p-value, we are 99% confident that these are highly significant variables.

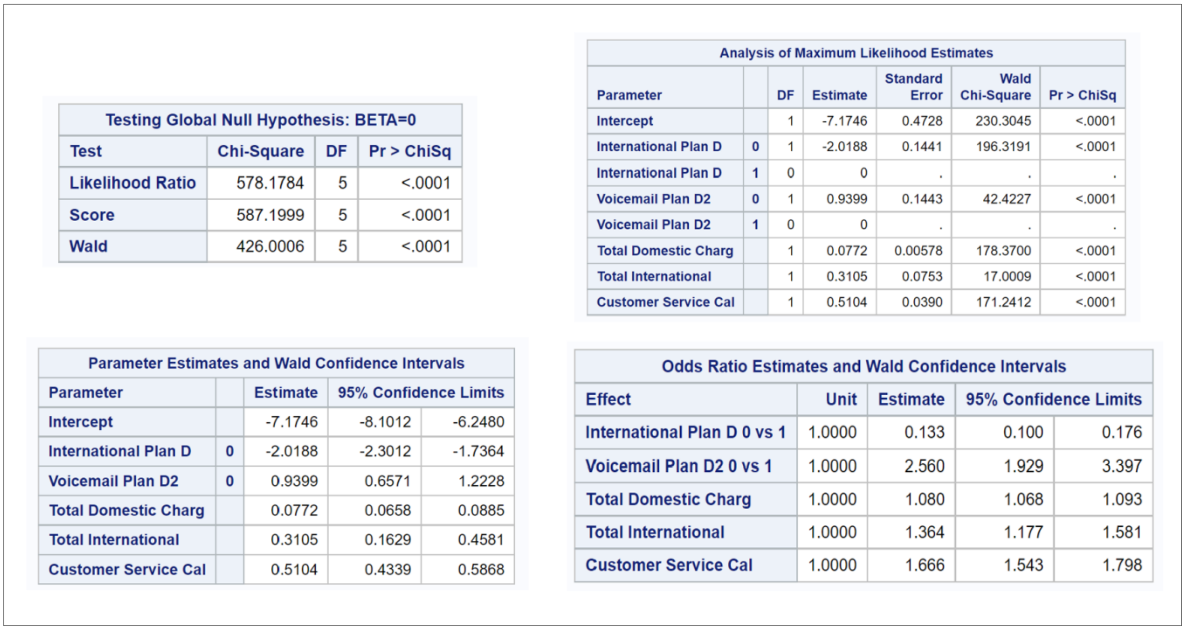
The beta coefficients can be described as: directly proportional if the sign is +ve, else inversely proportional.

Here, the probability of churn is given by below equation:

$$P(churn=1) = \frac{e^{(-7.1746 - 2.0188*\boldsymbol{IP} + 0.9399*\boldsymbol{VP} + 0.0772*\boldsymbol{TDC} + 0.5104*\boldsymbol{SC} + 0.3105*\boldsymbol{TIC})}}{1 + e^{(-7.1746 - 2.0188*\boldsymbol{IP} + 0.9399*\boldsymbol{VP} + 0.0772*\boldsymbol{TDC} + 0.5104*\boldsymbol{SC} + 0.3105*\boldsymbol{TIC})}}$$

We can conclude that, 1 unit change in the service calls, international and domestic charges, will cause increase in odds of churning by factor of approximately 2.

Since we are using SAS on demand tool for the analysis and modelling the regression equation, the coefficients of the dummy variables, international plan and voicemail plan are shown in reverse way, i.e., when we select a customer who has subscribed to international plan, we will use a value of 0 instead of 1 and vice versa in IP. The same applies for the voicemail plan too i.e., variable VP. 1 unit change in international plan and voicemail plan subscription would change the odds of churn by a factor of 1 to 2.



Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	578.1784	5	<.0001
Score	587.1999	5	<.0001
Wald	426.0006	5	<.0001

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-7.1746	0.4728	230.3045	<.0001
International Plan D	0	1	-2.0188	0.1441	196.3191	<.0001
International Plan D	1	0	0	.	.	.
Voicemail Plan D2	0	1	0.9399	0.1443	42.4227	<.0001
Voicemail Plan D2	1	0	0	.	.	.
Total Domestic Charg		1	0.0772	0.00578	178.3700	<.0001
Total International		1	0.3105	0.0753	17.0009	<.0001
Customer Service Cal		1	0.5104	0.0390	171.2412	<.0001

Parameter Estimates and Wald Confidence Intervals

Parameter		Estimate	95% Confidence Limits	
Intercept		-7.1746	-8.1012	-6.2480
International Plan D	0	-2.0188	-2.3012	-1.7364
Voicemail Plan D2	0	0.9399	0.6571	1.2228
Total Domestic Charg		0.0772	0.0658	0.0885
Total International		0.3105	0.1629	0.4581
Customer Service Cal		0.5104	0.4339	0.5868

Odds Ratio Estimates and Wald Confidence Intervals

Effect	Unit	Estimate	95% Confidence Limits	
International Plan D 0 vs 1	1.0000	0.133	0.100	0.176
Voicemail Plan D2 0 vs 1	1.0000	2.560	1.929	3.397
Total Domestic Charg	1.0000	1.080	1.068	1.093
Total International	1.0000	1.364	1.177	1.581
Customer Service Cal	1.0000	1.666	1.543	1.798

Figure 5: Logistic regression output using SAS on demand

b) Classification tree analysis:

According to initial analysis, we started with multiple factors including below as well as area code, no. of voicemail messages, account length etc. We found that not all these factors are significant. Factors that can affect churn are the input variables provided to the final classification tree:

1) Total domestic charges
2) International plan
3) Customer service calls
4) Voicemail plan

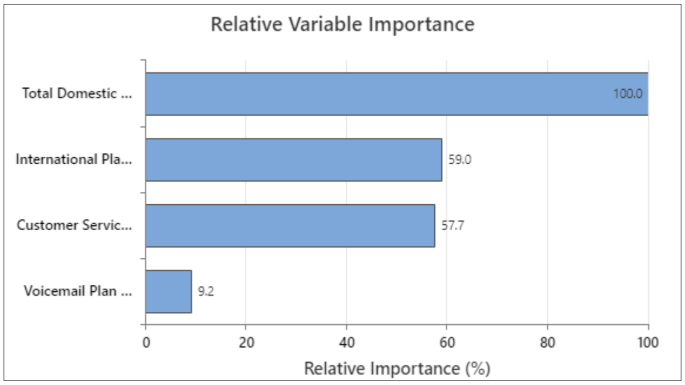


Figure 6: Relative % importance of 4 input variables to classification tree

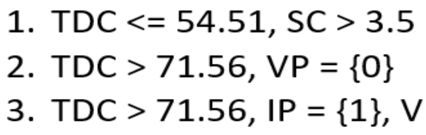
Outcome variable being predicted:

Churn variable, that can have 2 values – True (1) or False (0)

We got a classification tree as output that has 3 split levels for our case and 7 terminal nodes. 1st split level is done on total domestic charges, next on customer service calls and last on voicemail plan and international plan.

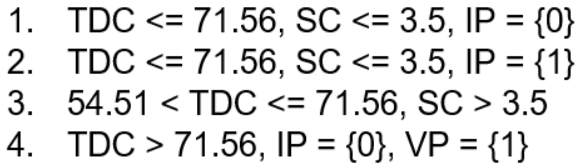
Output leaf rules of the classification tree:

Churn = 1 ->



1. TDC <= 54.51, SC > 3.5
2. TDC > 71.56, VP = {0}
3. TDC > 71.56, IP = {1}, VP = {1}

Churn = 0 ->



1. TDC <= 71.56, SC <= 3.5, IP = {0}
2. TDC <= 71.56, SC <= 3.5, IP = {1}
3. 54.51 < TDC <= 71.56, SC > 3.5
4. TDC > 71.56, IP = {0}, VP = {1}

Here, TDC = total domestic charges, SC = customer service calls, IP = international plan subscription, VP = voicemail plan subscription. Red node => non churner, Blue node => churner.

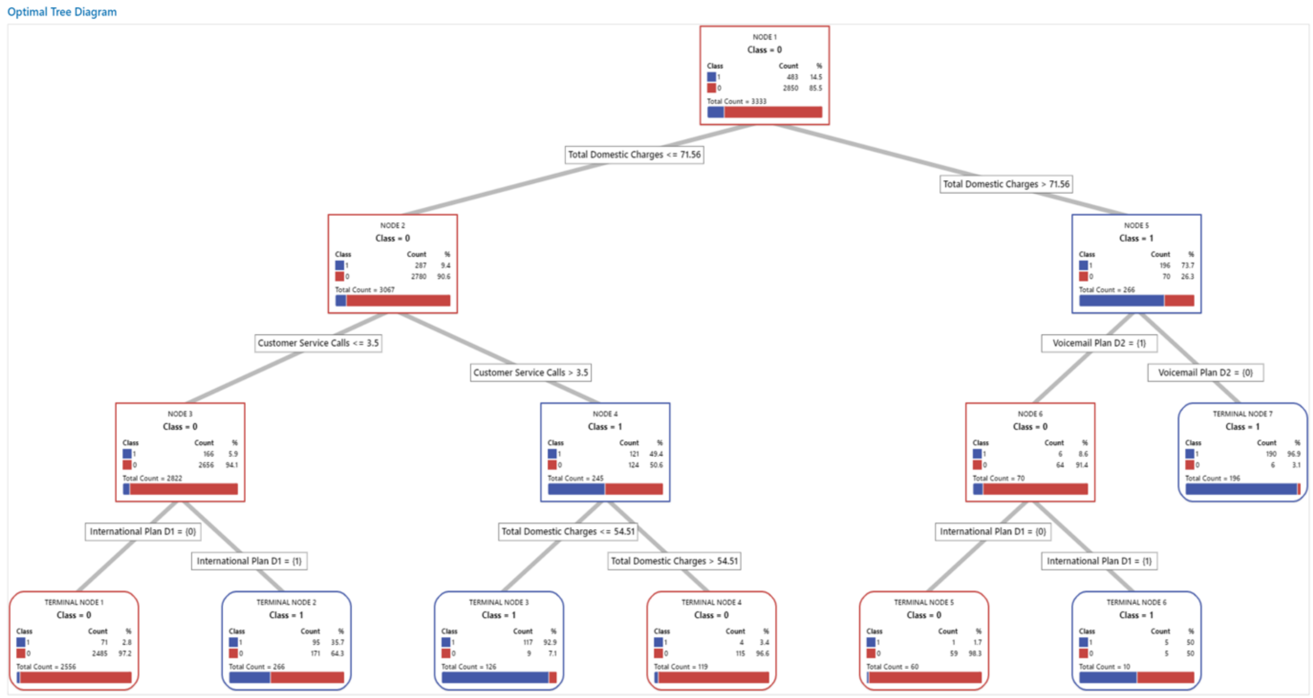


Figure 7: Classification tree

c) Better model out of the 2 techniques used:

The classification tree happens to provide a better model based on the misclassification rates. The comparative confusion matrix and misclassification rates calculations are given below.

Confusion matrix:

Logistic regression:

	Actual class			
Predicted class		**0**	**1**	**Total**
	0	2387	463	2850
	1	296	187	483
	Total	2683	650	3333

Table 3: Logistic regression confusion matrix

Classification tree:

	Actual class			
		0	**1**	**Total**
Predicted class	**0**	2659	191	2850
	1	76	407	483
	Total	2735	598	3333

Table 4: Classification tree confusion matrix

Misclassification rates comparison:

Since the classification tree provides a better overall misclassification %, we go for classification tree as our final model. Also, since our interest event is customer churning, the misclassified as 0 rate is more important than misclassified as 1 rate. Since the misclassified as 0 rate of classification tree is low at 6.70%, this model passes as a better model in our event of interest as well.

Sr. No.	Classification type	Logistic regression	Classification tree
1	Misclassified as 0	16.25%	6.70%
2	Misclassified as 1	61.28%	15.73%
3	**Overall misclassification rate**	**22.77%**	**8.01%**

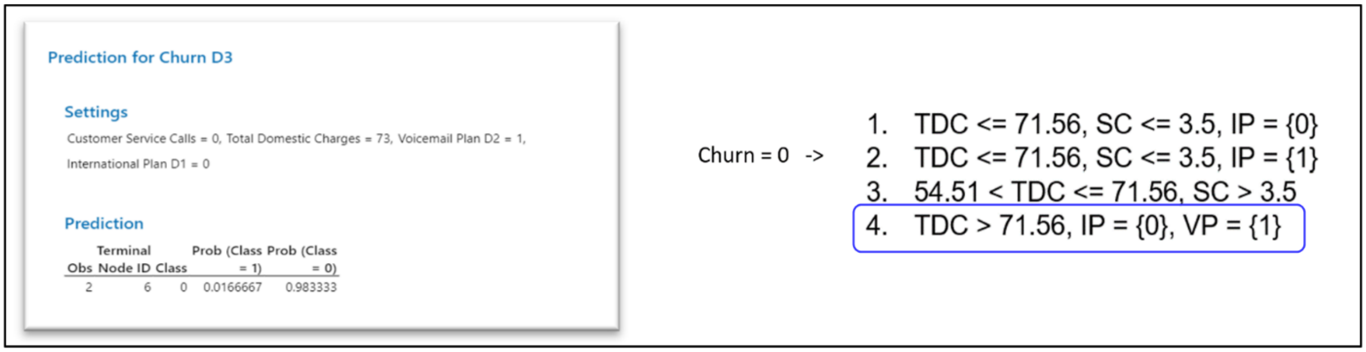
Table 5: Comparison of misclassification rates

e. Managerial implications and Conclusions:

1. Prediction of future customers churn using classification tree:

Customer 3334: classified as a non-churner: P(churn=0) = 0.9833

Explanation: Since the TDC > 71.56, IP = 0, VP = 1, the prob (churn = 0) =1, Rule no. 4 holds true here and customer is classified as a non-churner.



Prediction for Churn D3

Settings

Customer Service Calls = 0, Total Domestic Charges = 73, Voicemail Plan D2 = 1, International Plan D1 = 0

Prediction

```
        Terminal          Prob (Class Prob (Class
Obs  Node ID Class              = 1)        = 0)
  2        6     0        0.0166667    0.983333
```

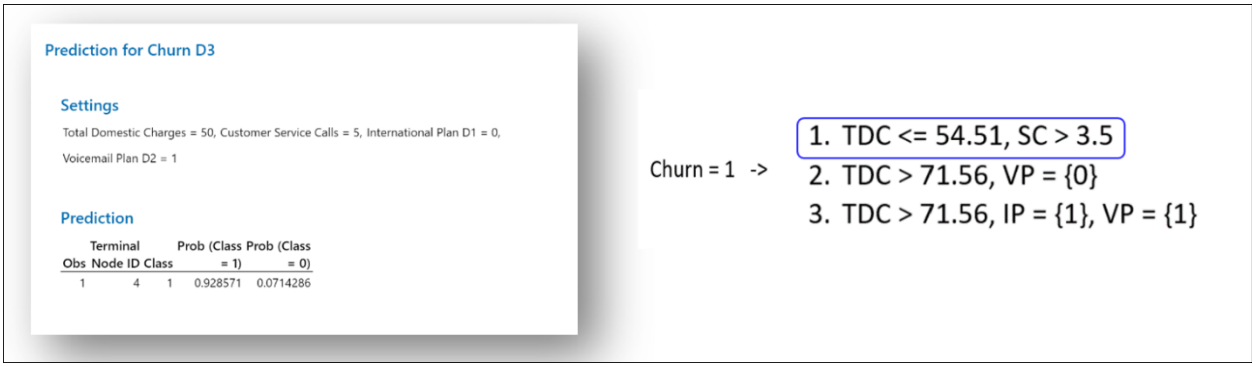
Churn = 0 ->

1. TDC <= 71.56, SC <= 3.5, IP = {0}
2. TDC <= 71.56, SC <= 3.5, IP = {1}
3. 54.51 < TDC <= 71.56, SC > 3.5
4. TDC > 71.56, IP = {0}, VP = {1}

Figure 8: Prediction of future churn for customer 3334 using Minitab

Customer 3335: classified as a churner: P(churn=1) = 0.9286

Explanation: Since the TDC < 54.51, SC > 3.5, the prob (churn = 1) =1, Rule no. 1 holds true here and customer is classified as a churner.



Prediction for Churn D3

Settings

Total Domestic Charges = 50, Customer Service Calls = 5, International Plan D1 = 0, Voicemail Plan D2 = 1

Prediction

Obs	Terminal Node ID	Class	Prob (Class = 1)	Prob (Class = 0)
1	4	1	0.928571	0.0714286

Churn = 1 ->

1. TDC <= 54.51, SC > 3.5
2. TDC > 71.56, VP = {0}
3. TDC > 71.56, IP = {1}, VP = {1}

Figure 9: Prediction of future churn for customer 3335 using Minitab

2. Managerial incentives:

Classification tree results and leaf rules are useful to reduce the customer churn rate, main emphasis is to be put on the most significant predictor fields: TDC (total domestic charges) and SC (service calls).

We put forth a target to bring down the customer churn from 14.49% to less than 5% in 1st year by using below incentives.

Incentives:

a) Using leaf rule: churn=1: TDC > 71.56, IP= {1}, VP= {1}

% Customers targeted: 0.15%

Offered incentive: **Customers Loyalty Program**

For loyal customers (Minimum Account Length > 24 months), all call minutes after 700th minute are free. For customers that fall in this criteria, currently Average daily minutes = 730 and Average daily charge = $ 0.12. Hence, this incentive would cost the company to spend $4 more per month per customer, however, can result in retention of 5 customers and they can be converted to non-churners.

b) Using leaf rule: churn=1: TDC<= 54.51, SC > 3.5

% Customers targeted: 3.51%

Offered incentive: **Customer Satisfaction Program**

Based on the customer service calls, we can identify customer complaints, group similar complaints and design custom solutions for clusters based on below criteria. These would be offered to customers who agree to stay with the company for 36 months.

- Discounted handsets on monthly plan contract – for customers who report an issue with network quality or higher call drop, we can offer them mobile handsets at a lower price.
- Lower call rate for prepaid to postpaid conversion – this would hold good for customers with a concern of higher monthly bill amounts which is the main reason behind churn.
- Creating brand value by offering subscription model – we can offer OTT platform subscriptions to customers for lower cost, who sign a contract for next 36 months.

The benefits to the company are continued customer association and customer retention for 36 months.

c) Using leaf rule: TDC > 71.56, VP = {0}

% Customers targeted: 5.88%

Offered incentive: **No Cost Voice Mail Plan**

Customers having minimum bill amount of $75 per month, will get free upgrade to facility of voice mail plan and enrolment in family plan at no cost. The benefits to the company are continued customer association and customer retention for 12 months. This would also offer addition of new connections due to Family plan.

3. Potential future research:

As part of the future research possibilities, below areas can be analyzed to deeper extent:

a) More detailed analysis can be done by splitting domestic charges based on day, evening and night calls minutes and charges. This analysis would help better incentives to be designed based on time of the day.

b) Survey results can be collected from customers and joined with this database to understand customer complaints. Emails, mobile apps can be used for this purpose.

f. References:

Dataset Source: Customer Churn Analysis:

https://www.kaggle.com/sandipdatta/customer-churn-analysis/data