

Telecom Industry Customer Churn Prediction

BUSINESS ANALYTICS WITH R
PROJECT
AKSHAY PATHAK

1 TABLE OF CONTENTS

2	Executive Summary	4
3	Background	5
4	Objective	6
5	Data dictionary	7
6	Data Summary:	8
7	Exploratory Data Analysis	11
7.1	Gender Distribution	11
7.2	Senior Citizen Distribution	11
7.3	Partner Distribution	12
7.4	Dependents Distribution	12
7.5	Phone Service Distribution	13
7.6	Multiple Lines Distribution	13
7.7	Internet Service Distribution	14
7.8	Online Security Distribution	14
7.9	Online Backup Distribution	15
7.10	Device Protection Distribution	15
7.11	Tech Support Distribution	16
7.12	Streaming Movies Distribution	16
7.13	Streaming TV Distribution	17
7.14	Contract Distribution	17
7.15	Paperless Billing Distribution	18
7.16	Payment Method Distribution	18
7.17	Tenure Group Distribution	19
7.18	Scatter Plot Between Monthly Charge And Tenure	19

7.19	Scatter Plot Between Total Charges And Tenure	20
7.20	Scatter Plot Between Monthly Charge And Total Charges	20
7.21	Scatter Plot Between Monthly Charge And Total Charges	21
7.22	Churned Customer Radar Chart	22
7.23	Non-Churn Customer Radar Chart	23
8	Logistic Regression.....	24
8.1	Logistic Regression Output:.....	25
8.2	Model performances:.....	27
8.2.1	Decile-wise Lift Chart.....	27
8.2.2	Lift Chart.....	27
8.2.3	Confusion Matrix	27
9	K Nearest Neighbors.....	28
9.1	Cross – Validation Curve For Knn.....	29
10	Classification And Regression Tree.....	30
10.1	Decision Tree.....	31
10.2	Interpretation of rules	32
10.2.1	Confusion Matrix	32
11	Linear Discriminant Analysis	33
11.1	Model Interpretation:.....	34
11.2	Model performance.....	35
11.2.1	Confusion Matrix	35
11.2.2	Lift and Decile Charts	35
12	Model Comparison.....	36
12.1	Strengths and Weaknesses of the model :	36
12.1.1	KNN:.....	36

12.1.2	Logistic Regression:.....	36
12.1.3	Decision Tree:	36
12.2	Parameter Based Comparison.....	37
13	Insights and Recommendation.....	38
14	Table of Figures	40
15	Appendix.....	41
16	References.....	41

2 EXECUTIVE SUMMARY

Service industries invest millions of dollars to ensure that customers feel welcomed and become brand loyal. The market is very competitive, and the business needs to keep their customers interested in their services. This report basically focuses on the fact that big industries have high rate of customer attrition and through this project, we are trying to find out ways to control it. The data under study is picked up from Kaggle. It talks about the various enrollment features of a customer's plans. Different visualization provides a brief understanding about the impact of each predictors in determining the churn rate, in order to identify key features heuristically. We then move to various machine learning algorithms like Logistic Regression, K Nearest Neighbors, Decision Tree and Linear Discriminant Analysis, for analyzing these features and identifying major predictors responsible for customer attrition. These models are compared and evaluated based on accuracy. We have provided recommendations to prevent customer attrition for different domains by leveraging customer behavior and domain expertise.

3 BACKGROUND

Customer Churn Analysis, also known as Customer Attrition Analysis, is the analysis of customers who cease using the company products. Telecom business treat a customer as churned when the customer stops using telecom services. It has become a critical problem lately since it has become more challenging and expensive to acquire a new customer than it is to retain a current customer especially when there are multiple options for the customer. Customer churn rate is one of the key business metrics because reclaimed long-term customers can be worth much more to the business than newly earned customers.

We can clearly classify churn customers into two distinct types of churn namely

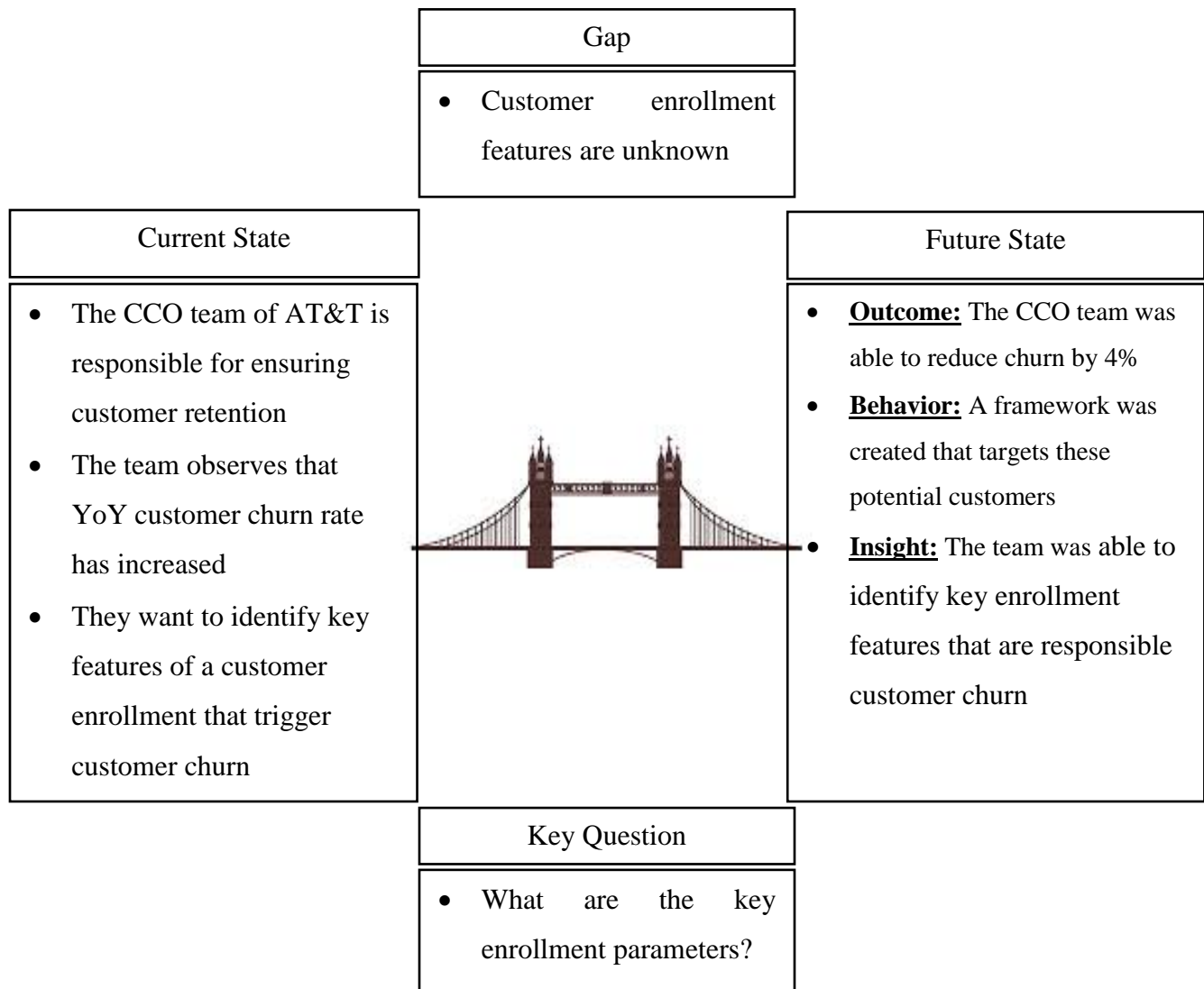
- **Voluntary:** The voluntary churners are those who leave by either cancelling or porting out to another carrier. This can be characterized by Cognitive Dissonance.
- **Involuntary:** Involuntary churners are terminated by us due to unpaid bills, fraudulent activity, etc. It can also be due to variety of other factors like relocation to a distant place, death etc.

The involuntary churn is beyond the control of the service provider and hence our main target is to find out the critical reasons for churn and retain customers who churn voluntarily. Accordingly, a single model would have a hard time capturing such complex patterns and having a separate model for each churn type is preferred.

Cognitive dissonance is the post purchase conundrum of customer about the purchase decision. When the company is not able to meet the customers expectation, the disappointment leads to cognitive dissonance hence they are likely to churn to the competitors. In this data set we are trying to find out the people churning out from the AT&T operator.

In order to succeed at retaining customers who would otherwise abandon the business, marketers and retention experts must be able to predict in advance which customers are going to churn through churn analysis. Armed with this knowledge, a large proportion of customer churn can be eliminated.

4 OBJECTIVE



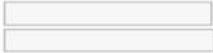

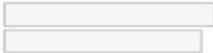


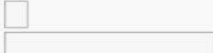
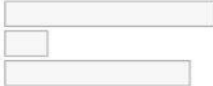






AT&T is observing a huge loss in customers over the years, the customer care team is working diligently on customer retention as retaining a current customer is very vital to an organization's survival. Identification of factors which govern the churn rate is important as they will be the predictors in defining the decision of churn by the customer. Through this project, we aim to identify the major predictors and try to find out the root cause for the voluntary churn. This activity will help AT&T in increasing the customer attrition rate by changing policies and offers to help increase the clientele retention.

5 DATA DICTIONARY

The data for our analysis is Customer Churn data of AT&T. We have taken this data from Kaggle.com (Kaggle, n.d.).

<i>Metric</i>	<i>Metric Definition</i>
<i>customerID</i>	Customer ID
<i>gender</i>	Customer Gender (Female=0, Male=1)
<i>SeniorCitizen</i>	Whether the customer is senior citizen or not (1,0)
<i>Partner</i>	Whether the customer has a Partner or not (1,0)
<i>Dependents</i>	Whether the customer has dependents or not (1,0)
<i>tenure</i>	Number of months the customer has stayed with the company
<i>PhoneService</i>	Whether the customer has phone service or not (1,0)
<i>MultipleLines</i>	Whether the customer has multiple lines or not (1,0)
<i>InternetService</i>	Customer's internet service provider (DSL, Fiber optic, No)
<i>OnlineSecurity</i>	Whether the customer has online security or not
<i>OnlineBackup</i>	Whether the customer has online backup or not
<i>DeviceProtection</i>	Whether the customer has device protection or not
<i>TechSupport</i>	Whether the customer has tech support or not
<i>StreamingTV</i>	Whether the customer has streaming TV or not
<i>StreamingMovies</i>	Whether the customer has streaming movies or not
<i>Contract</i>	The contract term of the customer
<i>PaperlessBilling</i>	Whether the customer has paperless billing or not
<i>PaymentMethod</i>	The customer's payment method
<i>MonthlyCharges</i>	The amount charged to the customer monthly
<i>TotalCharges</i>	The total amount charged to the customer
<i>Churn</i>	Whether the customer churned or not

6 DATA SUMMARY:

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Valid	Missing
1	gender [character]	1. Female 2. Male	3488 (49.5%) 3555 (50.5%)		7043 (100%)	0 (0%)
2	SeniorCitizen [integer]	mean (sd) : 0.16 (0.37) min < med < max : 0 < 0 < 1 IQR (CV) : 0 (2.27)	0 : 5901 (83.8%) 1 : 1142 (16.2%)		7043 (100%)	0 (0%)
3	Partner [character]	1. No 2. Yes	3641 (51.7%) 3402 (48.3%)		7043 (100%)	0 (0%)
4	Dependents [character]	1. No 2. Yes	4933 (70.0%) 2110 (30.0%)		7043 (100%)	0 (0%)
5	tenure [integer]	mean (sd) : 32.37 (24.56) min < med < max : 0 < 29 < 72 IQR (CV) : 46 (0.76)	73 distinct values		7043 (100%)	0 (0%)
6	PhoneService [character]	1. No 2. Yes	682 (9.7%) 6361 (90.3%)		7043 (100%)	0 (0%)
7	MultipleLines [character]	1. No 2. No phone service 3. Yes	3390 (48.1%) 682 (9.7%) 2971 (42.2%)		7043 (100%)	0 (0%)
8	InternetService [character]	1. DSL 2. Fiber optic 3. No	2421 (34.4%) 3096 (44.0%) 1526 (21.7%)		7043 (100%)	0 (0%)
9	OnlineSecurity [character]	1. No 2. No internet service 3. Yes	3498 (49.7%) 1526 (21.7%) 2019 (28.7%)		7043 (100%)	0 (0%)
10	OnlineBackup [character]	1. No 2. No internet service 3. Yes	3088 (43.8%) 1526 (21.7%) 2429 (34.5%)		7043 (100%)	0 (0%)
11	DeviceProtection [character]	1. No 2. No internet service 3. Yes	3095 (43.9%) 1526 (21.7%) 2422 (34.4%)		7043 (100%)	0 (0%)
12	TechSupport [character]	1. No 2. No internet service 3. Yes	3473 (49.3%) 1526 (21.7%) 2044 (29.0%)		7043 (100%)	0 (0%)
13	StreamingTV [character]	1. No 2. No internet service 3. Yes	2810 (39.9%) 1526 (21.7%) 2707 (38.4%)		7043 (100%)	0 (0%)

14	StreamingMovies [character]	1. No 2. No internet service 3. Yes	2785 (39.5%) 1526 (21.7%) 2732 (38.8%)		7043 (100%)	0 (0%)
15	Contract [character]	1. Month-to-month 2. One year 3. Two year	3875 (55.0%) 1473 (20.9%) 1695 (24.1%)		7043 (100%)	0 (0%)
16	PaperlessBilling [character]	1. No 2. Yes	2872 (40.8%) 4171 (59.2%)		7043 (100%)	0 (0%)
17	PaymentMethod [character]	1. Bank transfer (automatic) 2. Credit card (automatic) 3. Electronic check 4. Mailed check	1544 (21.9%) 1522 (21.6%) 2365 (33.6%) 1612 (22.9%)		7043 (100%)	0 (0%)
18	MonthlyCharges [numeric]	mean (sd) : 64.76 (30.09) min < med < max : 18.25 < 70.35 < 118.75 IQR (CV) : 54.35 (0.46)	1585 distinct values		7043 (100%)	0 (0%)
19	TotalCharges [numeric]	mean (sd) : 2283.3 (2266.77) min < med < max : 18.8 < 1397.47 < 8684.8 IQR (CV) : 3393.29 (0.99)	6530 distinct values		7032 (99.84%)	11 (0.16%)
20	Churn [character]	1. No 2. Yes	5174 (73.5%) 1869 (26.5%)		7043 (100%)	0 (0%)

The data collected has 20 predictors namely gender, senior citizen, partner, tenure etc. The total number of records are 7043 out of which only TotalCharges has 11 missing values which we have removed from the data. Out of the 20 predictors, we have 3 numeric columns i.e

Tenure

- mean (sd): 32.37 (24.56)
- min < med < max: 0 < 29 < 72
- IQR (CV): 46 (0.76)

TotalCharges

- mean (sd): 2283.3 (2266.77)
- min < med < max: 18.8 < 1397.47 < 8684.8
- IQR (CV): 3393.29 (0.99)

MonthlyCharges

- mean (sd): 64.76 (30.09)
- min < med < max: 18.25 < 70.35 < 118.75
- IQR (CV): 54.35 (0.46)

The remaining 17 predictors are categorical. For OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies we have 3 categories:

- Yes
- No
- No internet service

As per business logic “**No internet service**” is similar to “**No**” for these variables. So, all “Yes” and “No” have been converted to “1” and “0” for analysis purposes. Other categorical variables have been converted to dummy variables.

We have 17 categorical variables, hence the need for dimension reduction is not required.

7 EXPLORATORY DATA ANALYSIS

7.1 GENDER DISTRIBUTION

Gender distribution shows that 49.8% were male out of churned customer and 50.2% were found out to be female. Similarly, for non churned customers, 49.3% of women were and 50.7% were male. Hence, gender is a poor indicator to identify likely to churn.

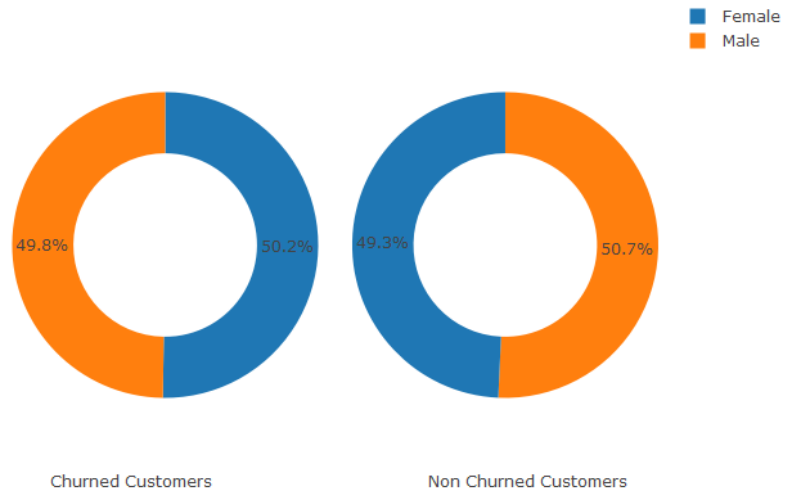


Figure 1: Gender Distribution

7.2 SENIOR CITIZEN DISTRIBUTION

Non-senior citizens constitute of 74.5% of the total churned population. While if we observe the same pattern for non-churned customers, the population of non-churned customers is again dominated by the younger people as they constitute 87.1% of the total population as compared to just 12.9% senior citizens who did not churn. This can be a good identifier.

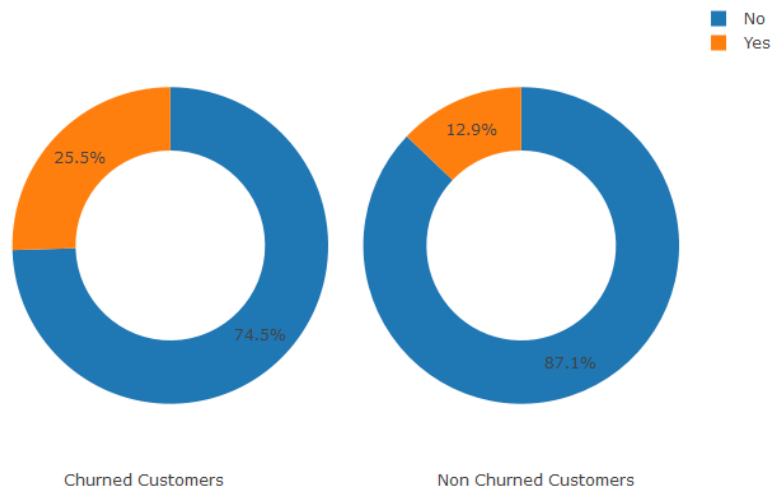


Figure 2: Senior Citizen Distribution

7.3 PARTNER DISTRIBUTION

We see that 64.2% of the churned customers do not have a partner enrolled whereas the distribution in non-churned is almost similar. This means that partner can be a good identifier for predicting churn.

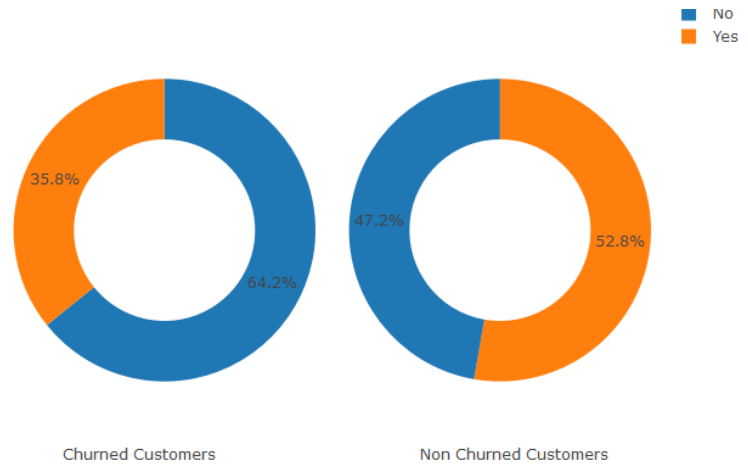


Figure 3: Partner Distribution

7.4 DEPENDENTS DISTRIBUTION

17.4% customers who churned out had dependents whereas 82.6% of the customers who churned out did not have any dependents. Similarly, for non-churned customers 34.3% had dependents while 65.7% did not have any dependents. Hence, dependents can be a good identifier.

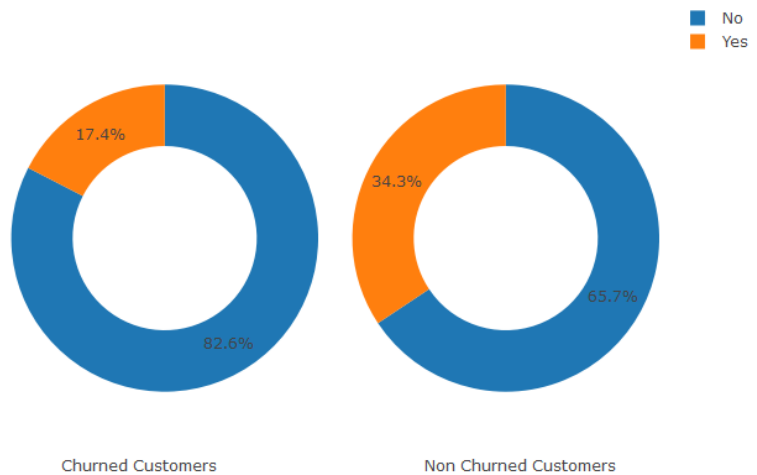


Figure 4: Dependents Distribution

7.5 PHONE SERVICE DISTRIBUTION

Here we can see that, 90.9% of the customers who churned has phone service and other 9.1% churned customers do not have Phone service and in the Non-churned customers, 90.1% of the non-churned customer has phone service and rest 9.88% of the total non-churned customers do not have phone service. We can identify that most of the churned and non-churned customers has phone service.

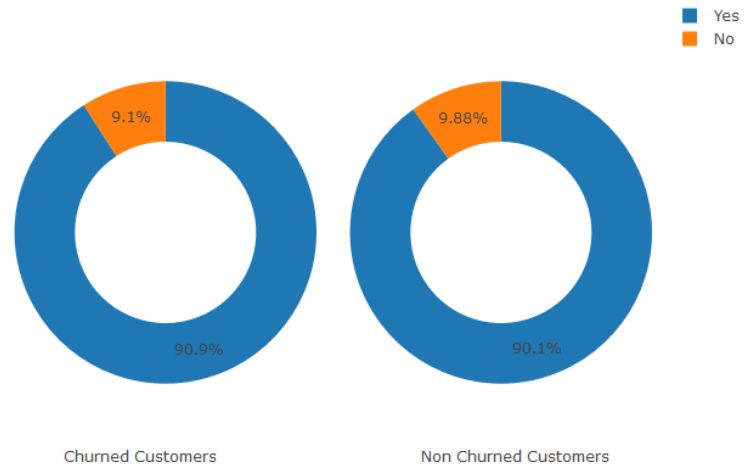


Figure 5: Phone Service Distribution

7.6 MULTIPLE LINES DISTRIBUTION

We infer that from churned customers 45.4% didn't not have multiple line distribution and 45.5% customer had multiple line distribution whereas 9.1% were not having phone service. Similarly, from non-churned customers 41% customers have had

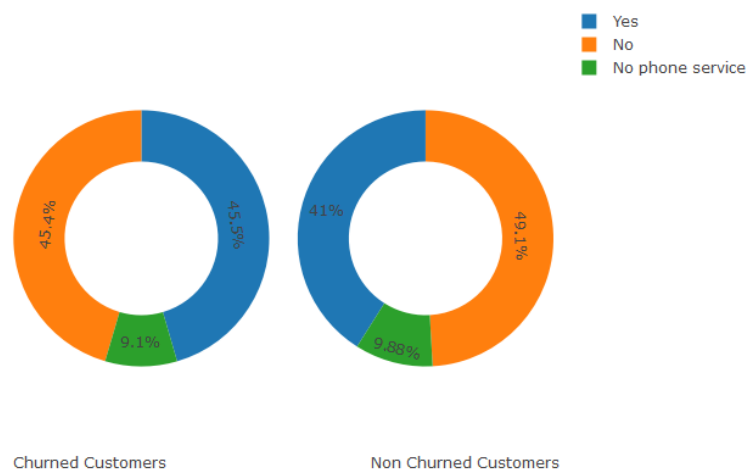


Figure 6: Multiple Lines Distribution

multiple line distributions and 49.1% customers did not where as 9.88% were not having phone service. Hence, it is might not be a good predictor.

7.7 INTERNET SERVICE DISTRIBUTION

For churned and non-churned customer's the distribution of different types of service plan shows that 6.05% of the customers who churned did not have any internet service whereas 69.4% used fiber optic and 24.6% used DSL. A similar pattern is

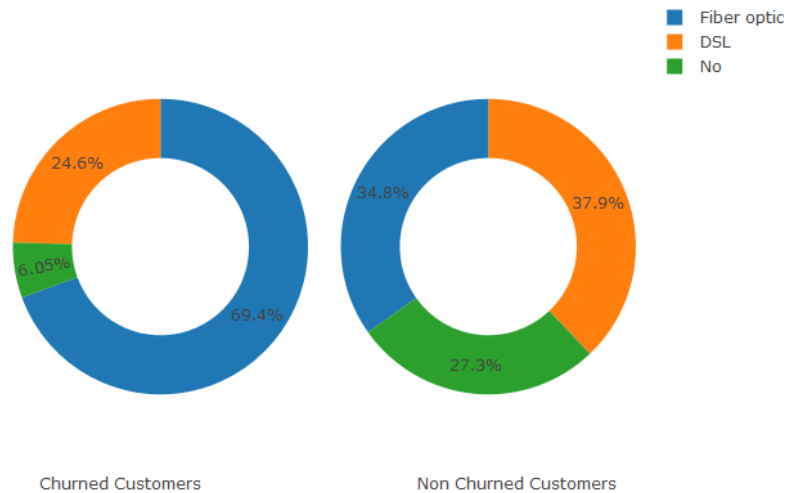


Figure 7: Internet Service Distribution

observed in the non-churned customers as well which is again dominated by the fiber optic users which constitute the major chunk of 37.9% in comparison to DSL users which comprise of 34.8% and 27.3% did not have any internet service. Hence, Internet service seems to be strong predictor.

7.8 ONLINE SECURITY DISTRIBUTION

For churned customers we see that 84.2% of customers didn't have online security distribution whereas remaining 15.8% had. Similarly, from non-churned customers donut chart we infer 66.7% customers didn't have online security

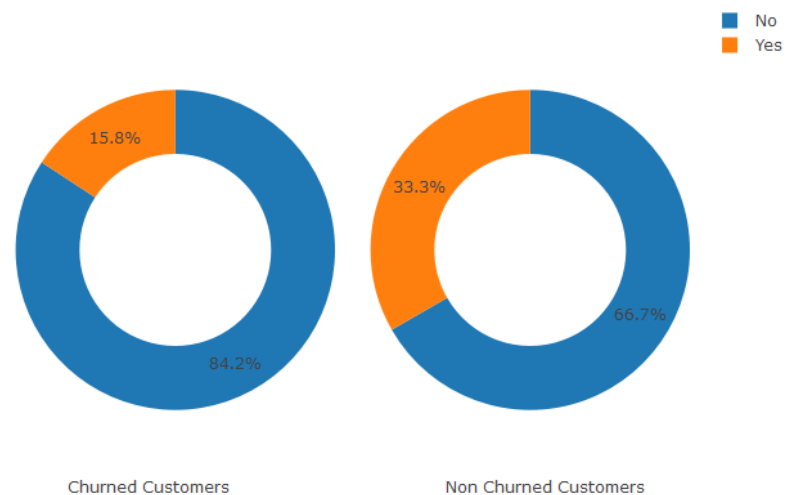


Figure 8: Online Security Distribution

distribution where as 33.3% had. This might be a good predictor.

7.9 ONLINE BACKUP DISTRIBUTION

28% of the customers who churned has Online backup distribution and other 72% customers who churned who do not have Online Backup Distribution. For Non-churned customers, 36.8% of the non-churned customer Online Backup Distribution and rest 63.2%

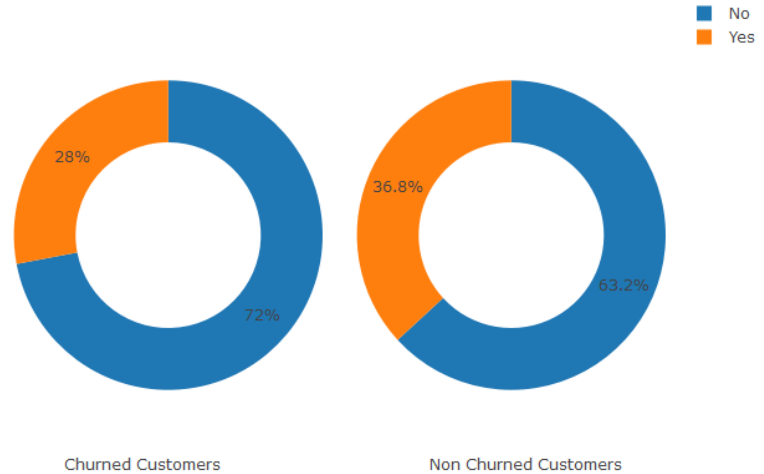


Figure 9: Online Backup Distribution

of the total non-churned customers do not have phone service. We can identify that the percentage of customers who has online service distribution in churned and non-churned customers are almost equal.

7.10 DEVICE PROTECTION DISTRIBUTION

For churned customers we see that 70.8% of customers didn't have device protection distribution whereas remaining 29.2% had. Similarly, from non-churned customers donut chart we infer 63.7%

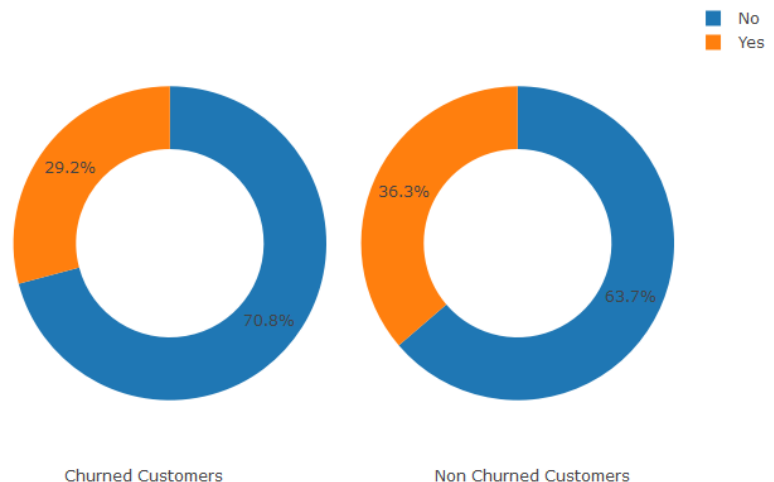


Figure 10: Device Protection Distribution

customers didn't have device protection distribution where as 36.3% had. Hence, this is a poor predictor.

7.11 TECH SUPPORT DISTRIBUTION

Here we can see from the donut chart that, 16.6% of the customers who churned has Tech Support Distribution and other 83.4% churned customers do not have Tech Support Distribution. For Non-churned customers, 33.5% of the non-churned customer Tech Support Distribution and rest 66.5% of the total non-churned

customers do not have Tech Support Distribution. We can identify that the percentage of non-churned customers who has Tech Support Distribution is almost double than the percentage of churned customer who has Tech Support Distribution.

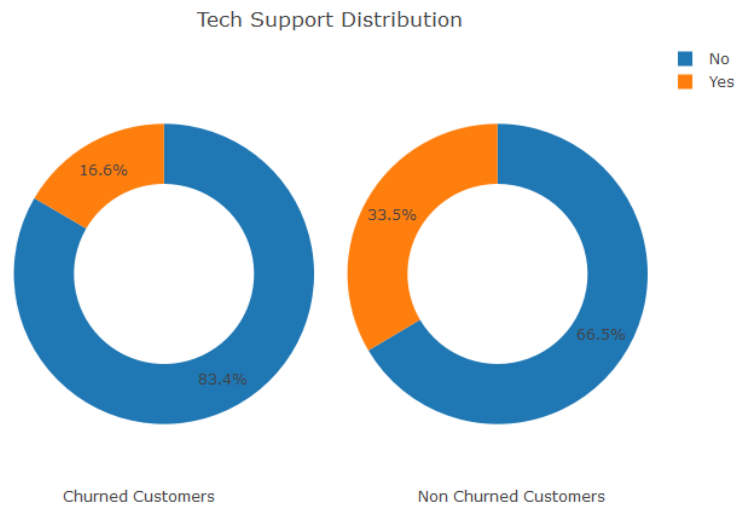


Figure 11: Tech Support Distribution

7.12 STREAMING MOVIES DISTRIBUTION

It can be inferred that 43.8% customers who churned out had streaming movies distribution whereas 56.2% of the customers who churned out did not have the streaming movies facility in the AT&T plan. Similarly, for non-churned customers 37.1% had streaming movie distribution

while 62.9% did not have that facility. Hence, this is a poor predictor.

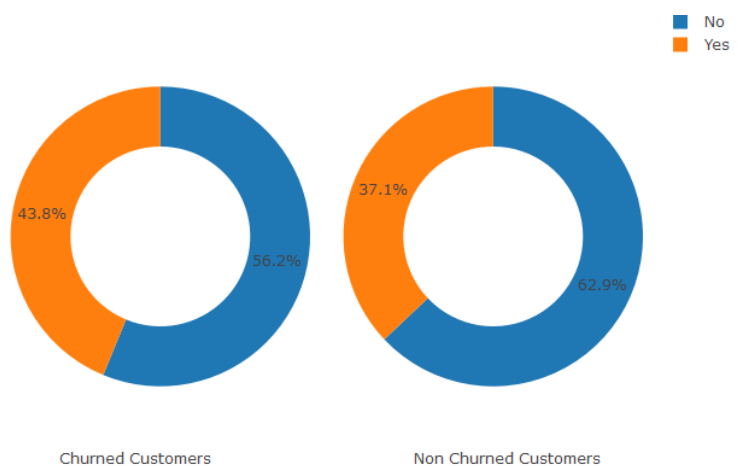


Figure 12: Streaming Movies Distribution

7.13 STREAMING TV DISTRIBUTION

For churned customers we see that 56.4% of customers didn't have Streaming TV Distribution whereas 43.6% had. Similarly, from non-churned customers donut chart we infer 63.4% customers didn't have

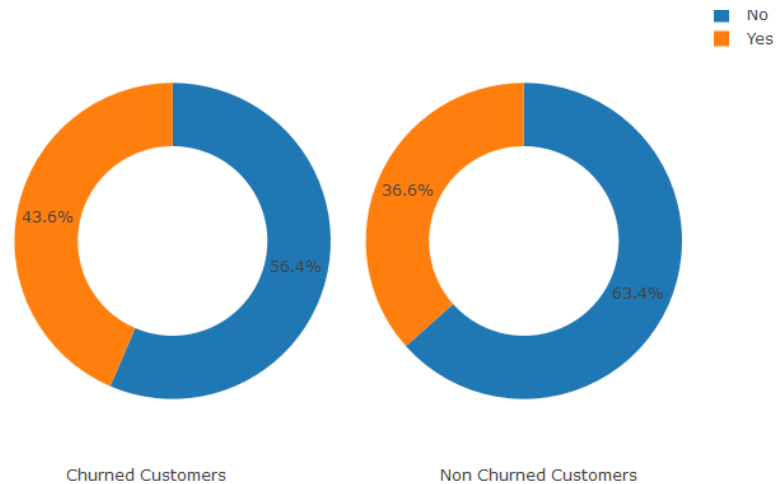


Figure 13: Streaming TV Distribution

Streaming TV Distribution where as 36.3% had. Hence, this is a poor predictor.

7.14 CONTRACT DISTRIBUTION

2.57% of the customers who churned out had two-year contract whereas 8.88% had one-year contract and 88.6% had month to month contract whereas in the non-churned customers segment, 43% people had month to month contract,

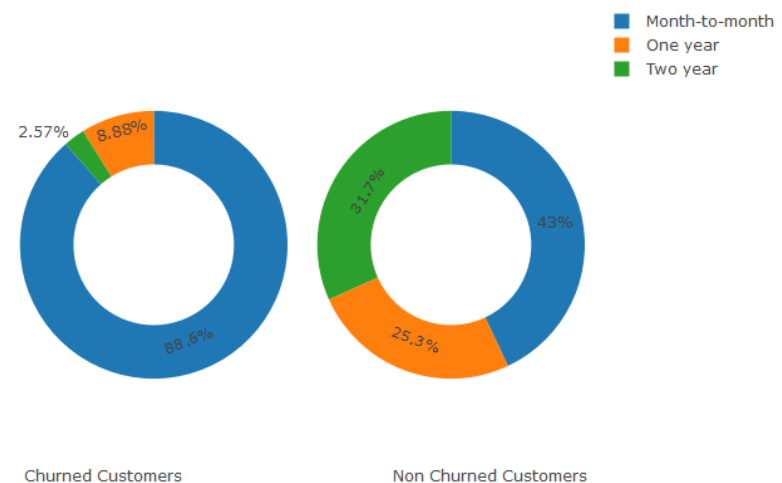


Figure 14: Contract Distribution

25.3% had yearly contract and 31.7% people had a contract of two years. Hence, this is a good predictor.

7.15 PAPERLESS BILLING DISTRIBUTION

For churned customers we see that 74.9% of customers had paperless billing distribution whereas remaining 25.1% didn't have. Similarly, from non-churned customers donut chart we infer 53.64% customers have had paperless billing distribution whereas 46.4% did not. Hence, this is a good predictor.

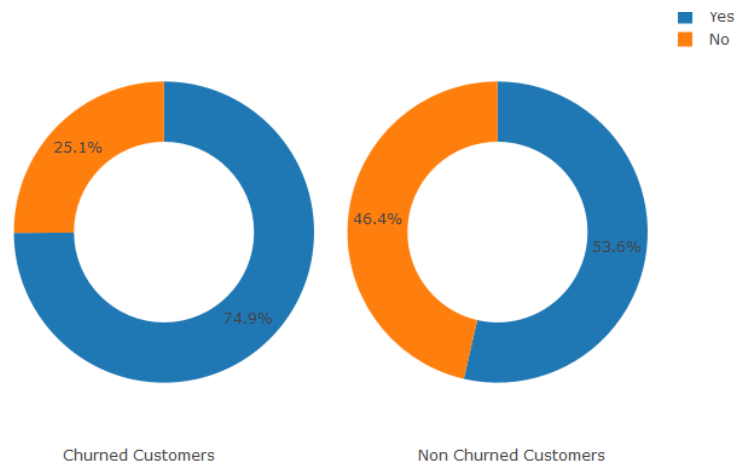


Figure 15: Paperless Billing Distribution

7.16 PAYMENT METHOD DISTRIBUTION

16.5% of the churned customers are paying by Mailed Check, 13.8% churned customers are paying by Bank Transfer, 12.4% of the churned customers are paying by Credit card and rest 57.3% of the churned customers are paying by Electronic Check. In the Non-churned customers, 25.1% of

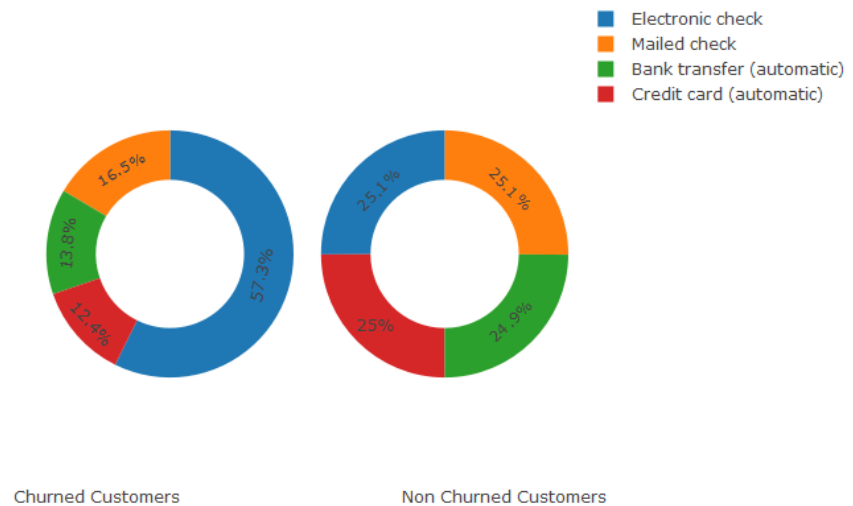


Figure 16: Payment Method Distribution

the non-churned customers are paying by Mailed Check, 24.9% non-churned are paying by Bank Transfer, 25% of the non-churned customers are paying by Credit card and rest 25.1% of the non-churned customers are paying by Electronic Check. Hence, this is a good predictor.

7.17 TENURE GROUP DISTRIBUTION

55.5% of the churned customers have a tenure between 0-12 months, 17.4% of the churned customers have a tenure between 24 to 48 months, 15.7% of the customers have tenure between 12 and 24 months, 6.42% between 48 and 60 months and 4.98% have tenure of

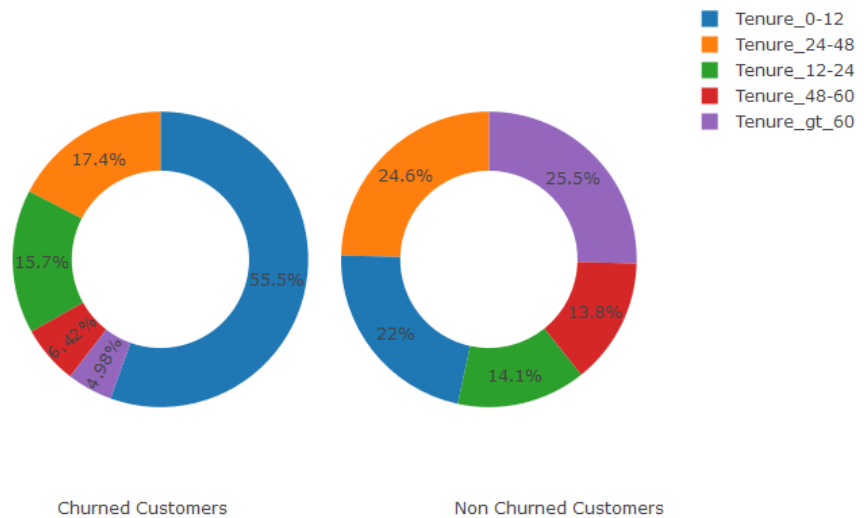


Figure 17: Tenure Group Distribution

greater than 60 months. In the non-churned customers, 25.5% have tenure greater than 60 months, 24.6% have a tenure between 24 and 48 months, 22% have a tenure between 0 to 12 months, 14.1% have a tenure between 12 and 24 months and 13.8% have a tenure between 48 and 60 months.

7.18 SCATTER PLOT BETWEEN MONTHLY CHARGE AND TENURE

It can be concluded that when monthly charges fall over \$60 the customer is more likely to churn irrespective of the tenure. So, monthly charges are a good classifier for predicting customer churn rate.



Figure 18: Monthly Charges VS Tenure based on Churn

7.19 SCATTER PLOT BETWEEN TOTAL CHARGES AND TENURE

All customers follow the same pattern irrespective of whether they churn or not. So, we can conclude that tenure with respect to total charges are not a good determinant of the customer attrition rate.

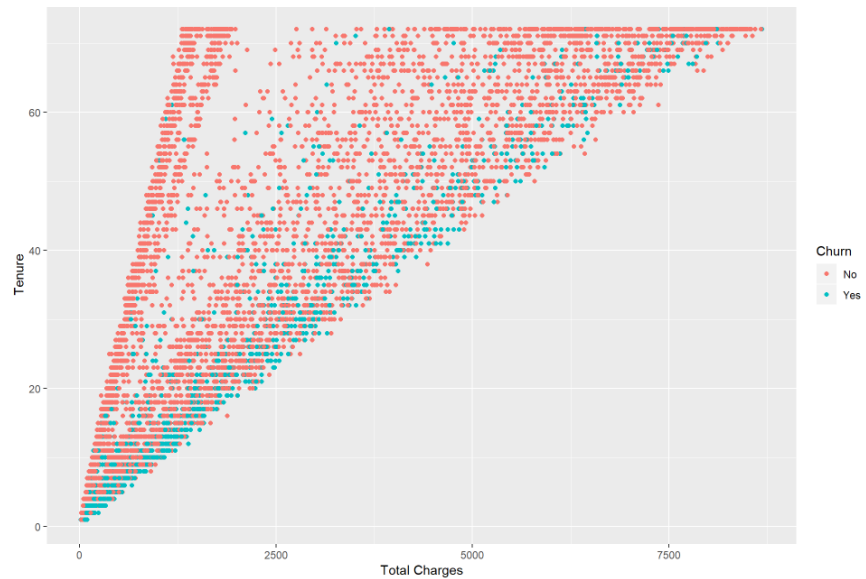


Figure 19: Total Charges VS Tenure based on Churn

7.20 SCATTER PLOT BETWEEN MONTHLY CHARGE AND TOTAL CHARGES

We can see in the scatter plot that in the quadrant where monthly charges are high and total charges are low, the customer attrition rate is the maximum. So, monthly charges with respect to total charges is a good predictor and determinant whether the customer will churn or not.



Figure 20: Total Charges VS Monthly Charges based on Churn

7.21 SCATTER PLOT BETWEEN MONTHLY CHARGE AND TOTAL CHARGES



Figure 21: Monthly Charges VS Total Charges based on Tenure Group

The plot shows that customers within a tenure group have similar monthly charges. This shows that over a period, customers do not change their plans. This can be because there are no lucrative deals for customer to change plans.

7.22 CHURNED CUSTOMER RADAR CHART

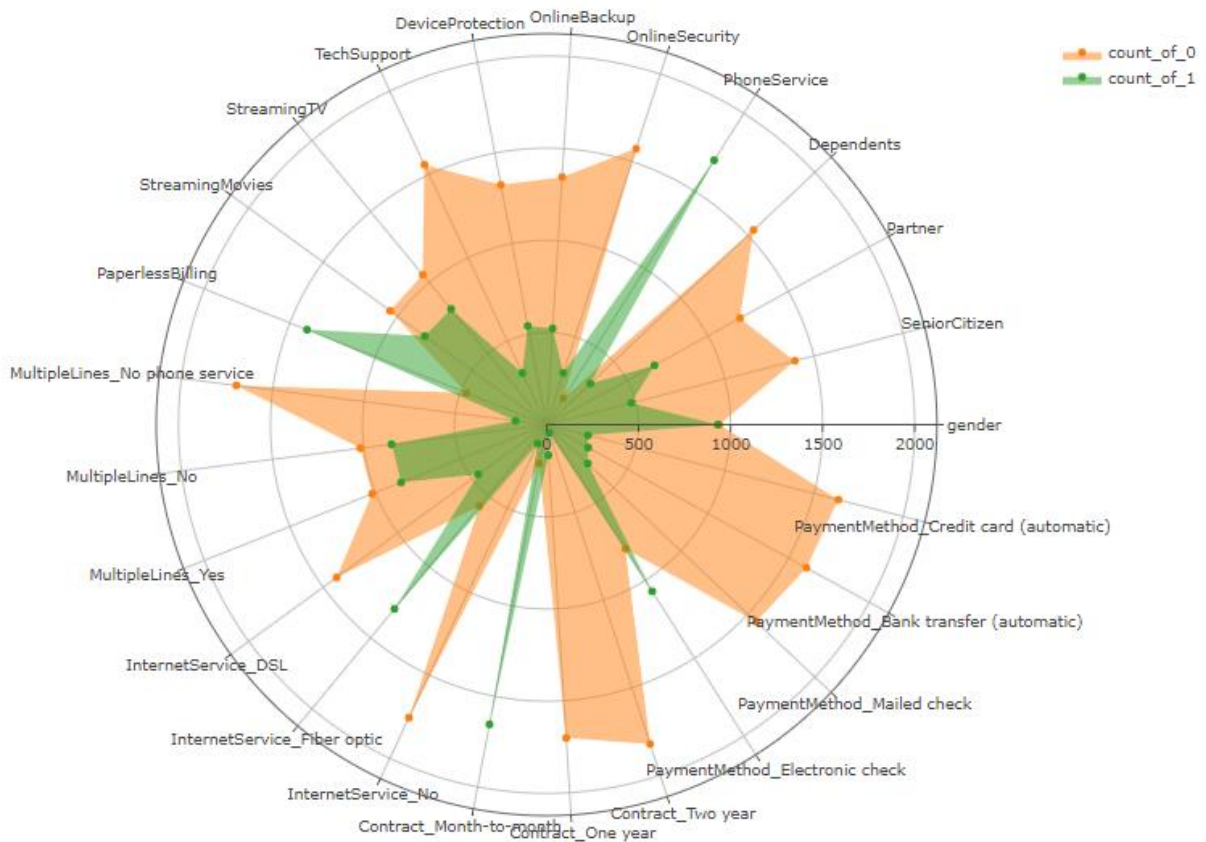


Figure 22: Churned Customer Radar Chart

From this Radar Chart we can identify the parameters which are significant in the churned customers. Here the customers with Phone service, Paperless Billing, Fiber Optic Internet Service, Electronic Check payment method and month to month Contract are prominent. We can sum up from the chart that the customers with the above parameters churned the most.

7.23 NON-CHURN CUSTOMER RADAR CHART

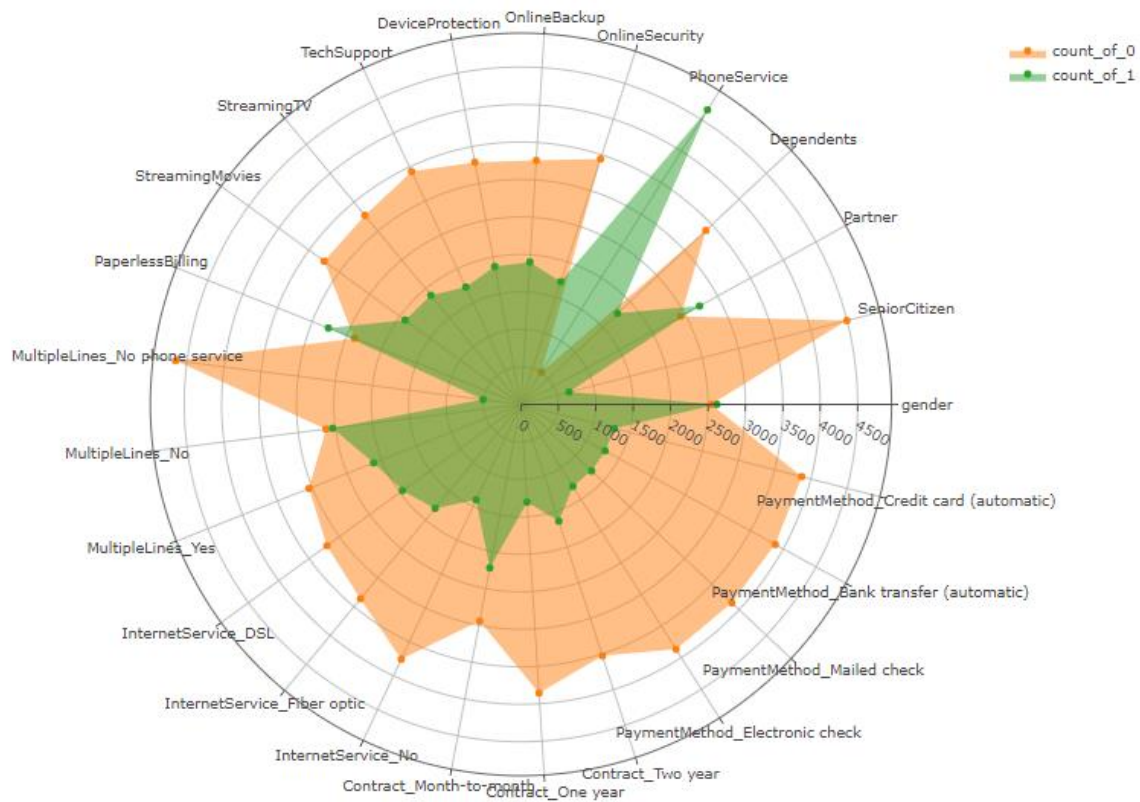


Figure 23: Non-Churned Customer Radar Chart

From this Radar Chart we can identify the parameters which are significant in the non-churned customers. Here the customers with Phone service, Paperless Billing and Partners are prominent. We can sum up from the chart that the customers with the above parameters are less likely to churn.

8 LOGISTIC REGRESSION

Logistic regression, A predictive analysis is used for analyzing a dataset in which there are one or more independent variables determining an outcome. It is a method for fitting a regression curve, $y = f(x)$, when y is a categorical variable. The typical use of this model is for predicting y given a set of predictors x . The predictors can be continuous, categorical or a mix of both.

Logistic regression can be used for classifying a new record, where its class is unknown, into one of the classes, based on the values of its predictor variables (called classification). It can also be used in data where the class is known, to find factors distinguishing between records in different classes in terms of their predictor variables. A logistic response function can be defined as:

$$p = 1/(1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_q x_q)})$$

The odds of belonging to class 1 are defined as the ratio of the probability of belonging to class 1 to the probability of belonging to class 0:

$$\text{Odds}(Y = 1) = p / (1 - p)$$

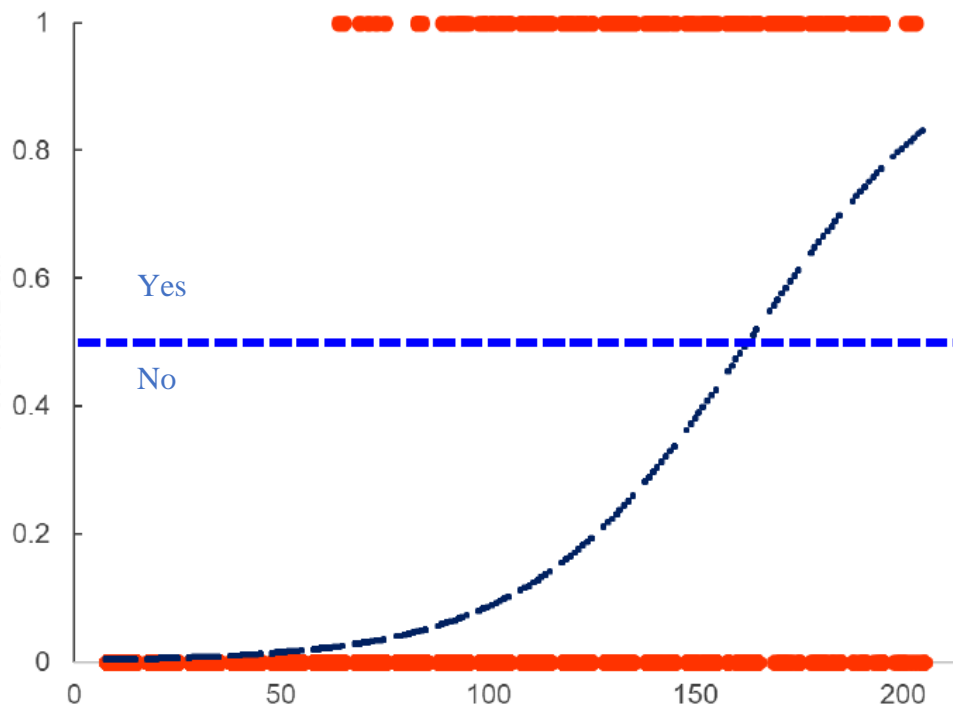


Figure 24: Sigmoid Function

8.1 LOGISTIC REGRESSION OUTPUT:

```
call:
glm(formula = churn ~ ., family = "binomial", data = telecom.train.norm[,
  c(-1)])

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8908 -0.6713 -0.2955  0.7102  3.3694

Coefficients: (5 not defined because of singularities)
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   -6.288267    2.403732  -2.616  0.008896 **
SeniorCitizen    0.262442    0.097048   2.704  0.006846 **
Partner         0.008353    0.089992   0.093  0.926043
Dependents     -0.228338    0.104212  -2.191  0.028446 *
tenure        -1.414263    0.175115  -8.076  6.68e-16 ***
PhoneService    0.893690    0.936760   0.954  0.340072
OnlineSecurity  -0.138056    0.206670  -0.668  0.504134
OnlineBackup    0.011271    0.203769   0.055  0.955891
DeviceProtection 0.124969    0.202852   0.616  0.537856
TechSupport    -0.106599    0.209866  -0.508  0.611496
StreamingTV     0.717541    0.377488   1.901  0.057324 .
StreamingMovies 0.653318    0.379653   1.721  0.085281 .
PaperlessBilling 0.329459    0.085734   3.843  0.000122 ***
MonthlyCharges -1.397024    1.104791  -1.265  0.206046
TotalCharges    0.635076    0.184262   3.447  0.000568 ***
`MultipleLines_No phone service`
MultipleLines_No -0.517724    0.205874  -2.515  0.011911 *
MultipleLines_Yes
MultipleLines_Yes NA          NA          NA          NA
InternetService_DSL
InternetService_DSL 2.066939    0.935936   2.208  0.027215 *
`InternetService_Fiber optic`
InternetService_Fiber optic 3.995940    1.848001   2.162  0.030595 *
InternetService_No
InternetService_No NA          NA          NA          NA
`Contract_Month-to-month`
Contract_Month-to-month 1.209910    0.200064   6.048  1.47e-09 ***
`Contract_One year`
Contract_One year 0.623142    0.198835   3.134  0.001725 **
`Contract_Two year`
Contract_Two year NA          NA          NA          NA
`PaymentMethod_Electronic check`
PaymentMethod_Electronic check 0.247922    0.112625   2.201  0.027715 *
`PaymentMethod_Mailed check`
PaymentMethod_Mailed check -0.024200    0.134531  -0.180  0.857242
`PaymentMethod_Bank transfer (automatic)`
PaymentMethod_Bank transfer (automatic) -0.056739    0.131323  -0.432  0.665698
`PaymentMethod_Credit card (automatic)`
PaymentMethod_Credit card (automatic) NA          NA          NA          NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 25: Logistic Regression Output

Logistic model summary for predicting churn clearly shows that not all variables are significant while explaining the model. The following variables are show a significant contribution in predicting the model:

1. Senior Citizen: If the customer is a Senior citizen then the odds of a customer churning is 1.3 times to that of a non-Senior Citizen
2. Dependent: If the customer has a dependent, the odds of the customer churning 0.8 times to that of a customer with no Dependent

3. Tenure: As the tenure of a customer's enrollment increases by 1 unit, the odds of the customer churning is 0.24 times
4. Paperless Billing: If the customer opts for paperless billing then the odds of a customer churning is 1.4 times
5. Multiple Lines: If the customer does not opt for Multiple lines then the odds of a customer churning is 0.6 times to that of customers with multiple lines
6. Total Charges: If the Total Charges of the enrollment increase by 1 unit then the odds of a customer churning are 1.9 times more
7. Internet Services:
 - a. DSL: If the customer has opted for a DSL line then the odds of a customer churning is 7.9 times more than customers with No Internet Service
 - b. Optic Fiber: If the customer has opted for an Optic Fiber the odds of a customer churning is 54 times more than customers with No Internet Service
8. Contract:
 - a. Month-to-month: If the customer has opted for Month-to-month contract then the odds of a customer churning is 3.3 times more than customers with a Two-year contract
 - b. One Year: If the customer has opted for One-Year contract then the odds of a customer churning is 1.8 times more than customers with a Two-year contract
9. Payment Method:
 - a. Electronic Check: If the customer has opted for Electronic Check payment method then the odds of a customer churning is 1.28 times to that customers paying through credit card
 - b. Mailed Check: If the customer has opted for Mailed Check payment method then the odds of a customer churning is 0.97 times to that customers paying through credit card
 - c. Automatic bank transfer: If the customer has opted for Automatic bank transfer then the odds of a customer churning is 0.94 times to that customers paying through credit card

8.2 MODEL PERFORMANCES:

8.2.1 Decile-wise Lift Chart

We observe that identify that mean of the Classification accuracy for the top 20 Percentile of the data is more than 2.5. Even the mean of the Classification accuracy for the 20-40 Percentile is 1.4.

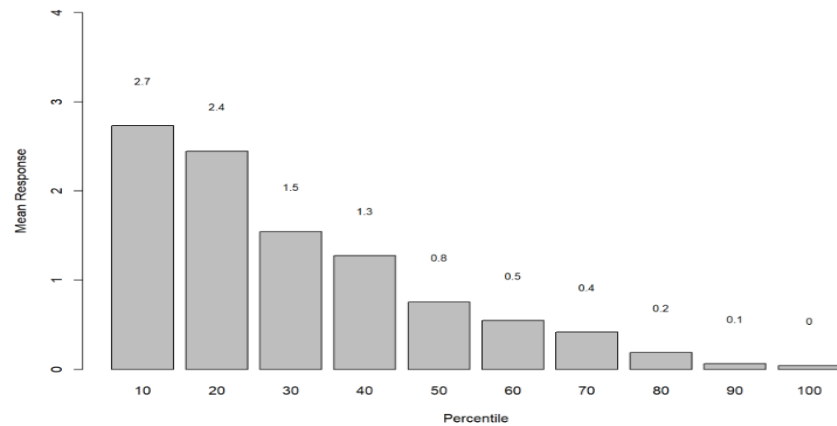


Figure 26: Decile Wise Lift Chart

8.2.2 Lift Chart

The chart shows that the first 400 cases were correctly predicted by the model. The performance of the model is better than the naïve Bayes model of 50% accuracy.

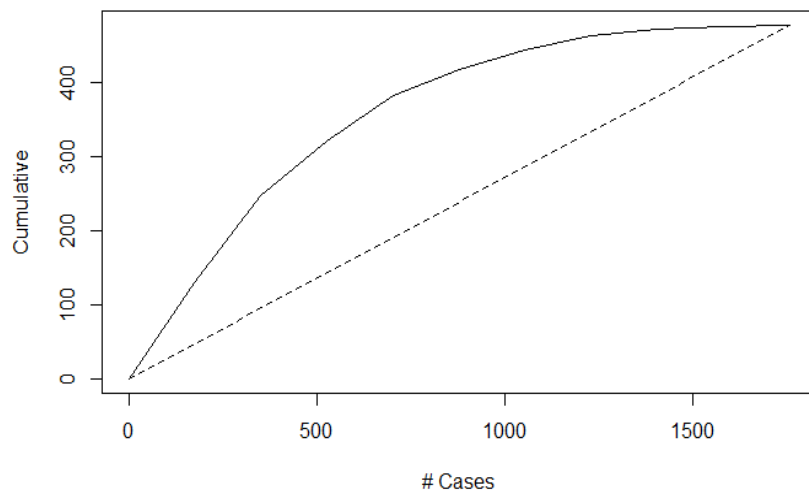


Figure 27: Lift Chart

8.2.3 Confusion Matrix

The model able to predict the possibility of a customer to churn with an accuracy of 80.5%.

					Kappa : 0.4767
	Prediction	0	1		McNemar's Test P-Value : 2.019e-06
	0	1153	216		
	1	127	262		
					Sensitivity : 0.9008
					Specificity : 0.5481
					Pos Pred Value : 0.8422
					Neg Pred Value : 0.6735
					Prevalence : 0.7281
					Detection Rate : 0.6559
					Detection Prevalence : 0.7787
					Balanced Accuracy : 0.7244
	Accuracy :	0.8049			
	95% CI :	(0.7856, 0.8232)			
	No Information Rate :	0.7281			
	P-value [Acc > NIR] :	4.457e-14			
					'Positive' Class : 0

9 K NEAREST NEIGHBORS

K nearest neighbors is a classification/supervised learning algorithm. To classify or predict a new record, this method finds similar records in the training data. These neighbors are then used to derive a classification or prediction for the new record by voting (for classification) or averaging (for prediction).

KNN is a non-parametric algorithm which means it does not need any training data points for model generation. All training data used is generally used in the testing phase. This makes training faster and testing phase slower and costlier.

In KNN, K is the number of nearest neighbors. The number of neighbors is the core deciding factor. The idea in k-nearest-neighbor's methods is to identify k records in the training dataset that are similar to a new record that we wish to classify. We then use these similar records to classify the new record into a class, assigning the new record to the predominant class among these neighbors.

Now, the main agenda is to measure the distance between records based on their predictor values. Euclidean distance is the most popular method of finding distance of the new member from the existing ones. The Euclidean distance between two records can be calculated as follows.

$$\sqrt{(x_1 - u_1)^2 + (x_2 - u_2)^2 + \dots + (x_p - u_p)^2}.$$

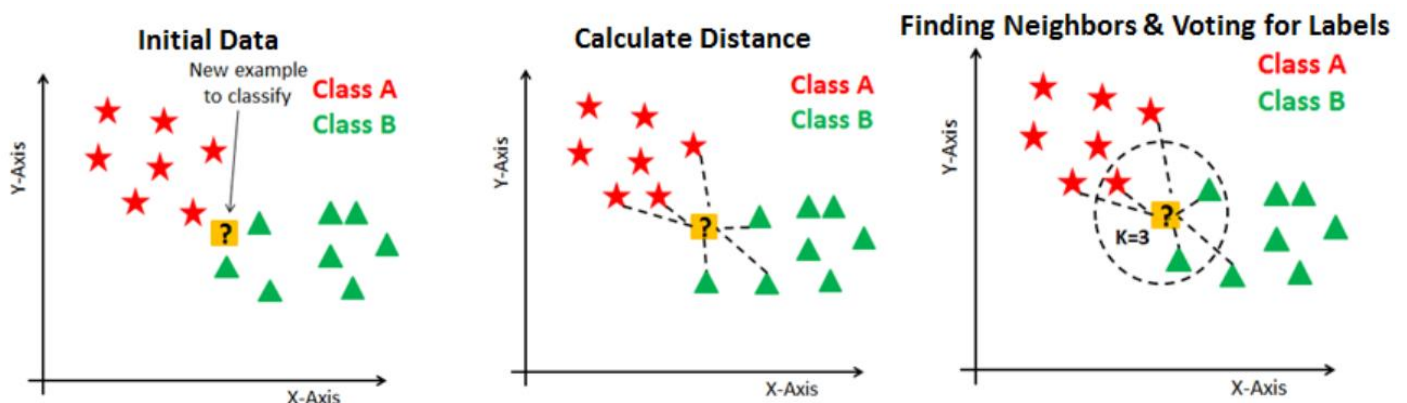


Figure 28: KNN Explanation (Datacamp, n.d.)

9.1 CROSS – VALIDATION CURVE FOR KNN

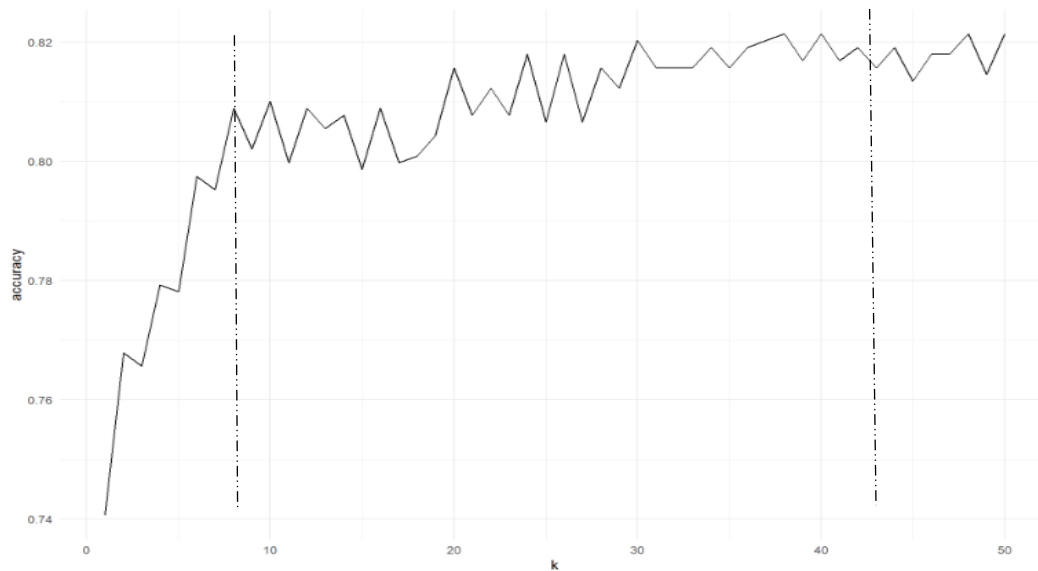


Figure 29: Cross-Validation Curve

Validation Dataset:

From the cross-validation curve, we can identify that as we increase the number of K in KNN model, accuracy of the model will quickly increase till K=9 and then the increase in accuracy is minimal. We can get maximum accuracy with K=38, i.e. 79% accuracy is achieved and with K=9, 76.7% accuracy is achieved. At K=38 the model is computationally, so one can choose K=9 at the expense of accuracy to cut down on cost.

Confusion Matrix and Statistics

```

Reference
Prediction 0 1
0 1233 233
1 167 300

Accuracy : 0.7931
95% CI : (0.7743, 0.8109)
No Information Rate : 0.7243
P-Value [Acc > NIR] : 1.953e-12

Kappa : 0.4613
McNemar's Test P-Value : 0.001154

Sensitivity : 0.8807
Specificity : 0.5629
Pos Pred Value : 0.8411
Neg Pred Value : 0.6424
Prevalence : 0.7243
Detection Rate : 0.6379
Detection Prevalence : 0.7584
Balanced Accuracy : 0.7218

```

Figure 31: For K = 38

Confusion Matrix and Statistics

```

Reference
Prediction 0 1
0 1203 252
1 197 281

Accuracy : 0.7677
95% CI : (0.7482, 0.7864)
No Information Rate : 0.7243
P-Value [Acc > NIR] : 7.761e-06

Kappa : 0.3992
McNemar's Test P-Value : 0.01082

Sensitivity : 0.8593
Specificity : 0.5272
Pos Pred Value : 0.8268
Neg Pred Value : 0.5879
Prevalence : 0.7243
Detection Rate : 0.6223
Detection Prevalence : 0.7527
Balanced Accuracy : 0.6932

```

Figure 30: For K = 9

10 CLASSIFICATION AND REGRESSION TREE

The CART model consists of decision tree. It's an algorithm which uses machine learning to perform both regression and classification. Basically, Decision tree makes it easy to represent the complex data and makes the process easier to showcase. There are two kinds of nodes in decision tree – Decision node and terminal node. The node with successors is called decision node whereas nodes with no successors are called terminal node. The Variable used for splitting is mentioned above the decision node, the number mentioned below the decision node are number of records in the node with value that is larger or equal the right side and less than the value at the left side. Yes or no at top of nodes tells on which side the lesser value is found. The two key ideas of CART are Recursive partitioning and Pruning.

The recursive partitioning is done of the space of predictor variables and pruning is done using validation data.

The Recursive partitioning – The classification is done in a manner to separate the input variable by using predictor variables such that the output is categorical variable. Suppose it X is the input variable; the predictor variable will be if $X \leq T$ or $X > T$. These conditions help in converting the variables into non-overlapping multidimensional pure data, so that they belong to just one class. The process is called recursive because it is continued till the complex data is sorted and all variables are adjusted into a class.

Pruning – The pruning tree works on the platform of CART. The validation data is used to prune back the tree that is emerged from training data (it is used to prune as well as grow the tree). It is mostly used to overcome the overfitting problem. The overfitting problem arises because of large grown trees and pruning removes all weak branches which eventually results in negligible changes for error rate.

10.1 DECISION TREE

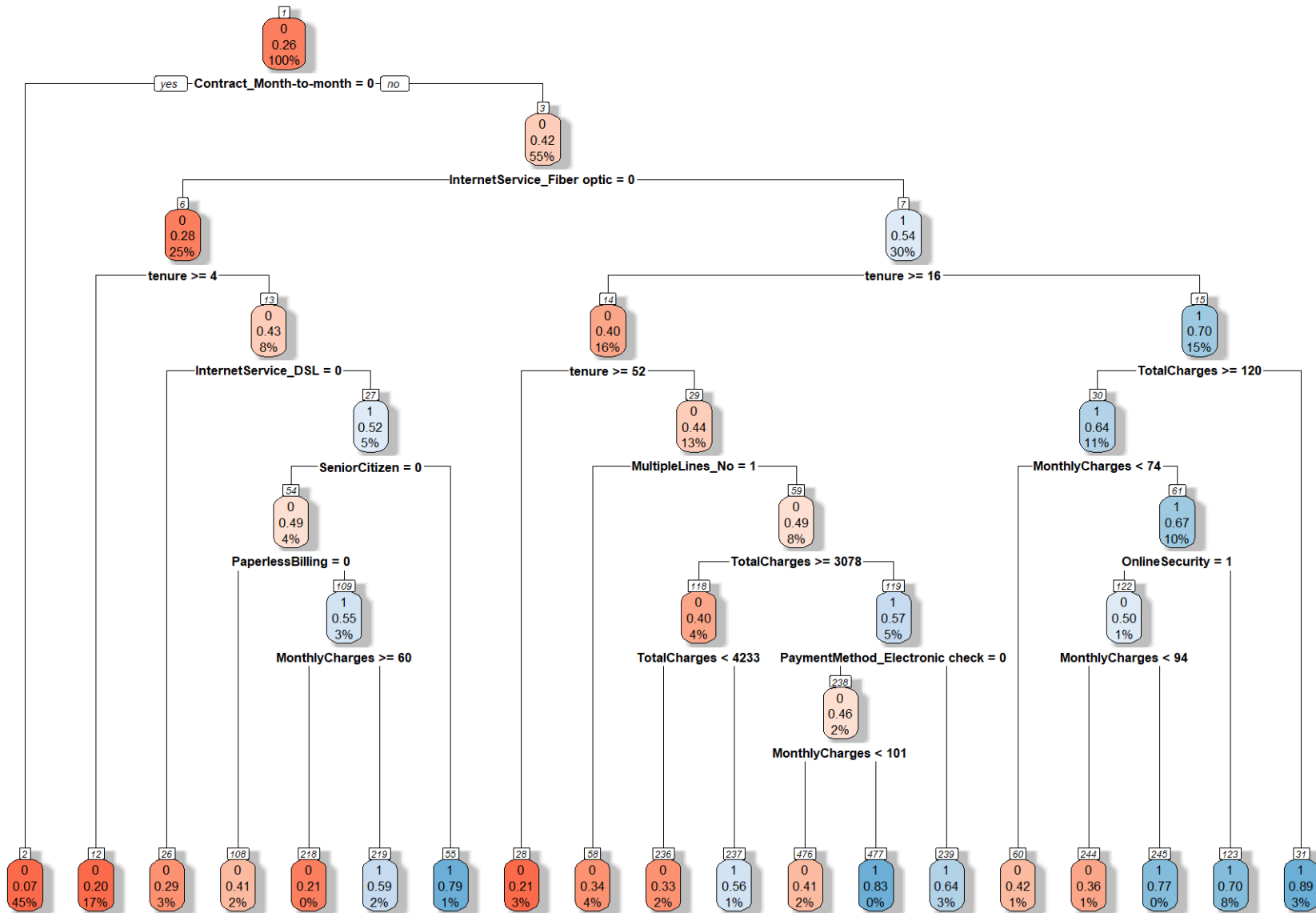


Figure 32: Decision Tree

10.2 INTERPRETATION OF RULES

Customers with month to month contract, Fiber optic internet service, with tenure less than 16 months and total charges lower than 120\$ are more likely to churn.

Customer with Month-to-month contract, Fiber optic internet service, with tenure is between 16 months to 52 months, total charges less than 3078\$, monthly charges greater than or equal to 101\$, without multiple lines and payment method is not Electronic check are second prominent group who are most likely to churn.

Customers who are senior citizens with Month-to-month contract, tenure less than 4 months, with DSL internet service and without Fiber optic Internet Service are third most prominent group who are most likely to churn.

Customers with month to month contract, Fiber optic internet service, with tenure less than 16 months, total charges not less than 120\$, monthly charges not less than 94\$ and with online security added to their account are fourth prominent group who are most likely to churn.

Customers who do not have month-to-month contract are less likely to churn.

Customers with month-to-month contract and tenure not less than 4 months without fiber optic service are less likely to churn.

Customers with Month-to-month contract and tenure not less than 52 months with fiber optic Internet service are less likely to churn.

Customers who are not senior citizens with month-to-month contract, tenure less than 4 months without fiber optic service but with paperless billing and DSL internet service subscribed and monthly charges not less than 60\$ are less likely to churn.

10.2.1 Confusion Matrix

```

      Reference
Prediction  0    1
0    1184  276
1      96  202

      Accuracy : 0.7884
      95% CI : (0.7685, 0.8073)
No Information Rate : 0.7281
P-Value [Acc > NIR] : 3.332e-09

      Kappa : 0.3941
McNemar's Test P-Value : < 2.2e-16
```

```

Sensitivity : 0.9250
Specificity : 0.4226
Pos Pred Value : 0.8110
Neg Pred Value : 0.6779
Prevalence : 0.7281
Detection Rate : 0.6735
Detection Prevalence : 0.8305
Balanced Accuracy : 0.6738

'Positive' Class : 0
```

11 LINEAR DISCRIMINANT ANALYSIS

Discriminant analysis is a classification method. It is a classical statistical technique that can be used for classification and profiling. It uses sets of measurements on different classes of records to classify new records into one of those classes (classification).

Linear Discriminant Analysis (LDA) is most commonly used as dimensionality reduction technique in the pre-processing step for pattern-classification and machine learning applications. The goal is to project a dataset onto a lower-dimensional space with good class-separability in order avoid overfitting (“curse of dimensionality”) and reduce computational costs.

It segments groups in a way as to achieve maximum separation between them. Below is our formula:

$$D = b_0 + b_1X_1 + b_2X_2 + ..b_nX_n$$

Here, D is the discriminant score, b is the discriminant coefficient, and X1 and X2 are independent variables.

The discriminant coefficient is estimated by maximizing the ratio of the variation between the classes of customers and the variation within the classes. In other words, points belonging to the same class should be close together, while also being far away from the other clusters.

11.1 MODEL INTERPRETATION:

```
Call:
lda(Churn ~ ., data = telecom.train.df)

Prior probabilities of groups:
  0          1 
0.7362533 0.2637467 

Group means:
      gender SeniorCitizen   Partner Dependents   tenure PhoneService OnlineSecurity OnlineBackup DeviceProtection TechSupport StreamingTV StreamingMovies
0 0.5075972   0.1300541 0.5274272 0.3486995 37.79217   0.9000773   0.3332475   0.3711048   0.3636364   0.3332475   0.3649240   0.3739377
1 0.4982027   0.2652768 0.3551402 0.1689432 18.01438   0.9137311   0.1631919   0.2731848   0.2875629   0.1718188   0.4370956   0.4428469
  PaperlessBilling MonthlyCharges TotalCharges
0      0.5323204      61.22998      2566.447
1      0.7440690      74.72926      1529.746

Coefficients of linear discriminants:
LD1
gender      -0.0117436183
SeniorCitizen 0.3086932747
Partner      0.0111038584
Dependents   -0.1837026831
tenure       -0.0120483489
PhoneService -0.9050413313
OnlineSecurity -0.4855006378
OnlineBackup -0.3534932786
DeviceProtection -0.2922375242
TechSupport  -0.5036030184
StreamingTV  -0.0978071759
StreamingMovies -0.1523559110
PaperlessBilling 0.2654955277
MonthlyCharges  0.0419122868
TotalCharges   -0.0002674877
```

In the given model, The prior probabilities of predicting Churned customer was 0.26 and non-churned customers is 0.73. The model highlights following parameters that are more prominent in identifying churned customers (Identified by comparing the grouped means of each variable):

1. Senior Citizen
2. Streaming TV
3. Streaming Movies
4. Paperless Billing
5. Monthly charges

Here from the above result we can see that the probabilities of predicting churned customers is more in the above listed variables than the probabilities of predicting non-churned customers in the same variables. Thus, we can conclude that the above five parameters are prominent in identifying churned customers.

11.2 MODEL PERFORMANCE

11.2.1 Confusion Matrix

Confusion Matrix and Statistics

	0	1
0	1143	226
1	137	252

Accuracy : 0.7935
95% CI : (0.7738, 0.8122)
No Information Rate : 0.7281
P-Value [Acc > NIR] : 1.394e-10

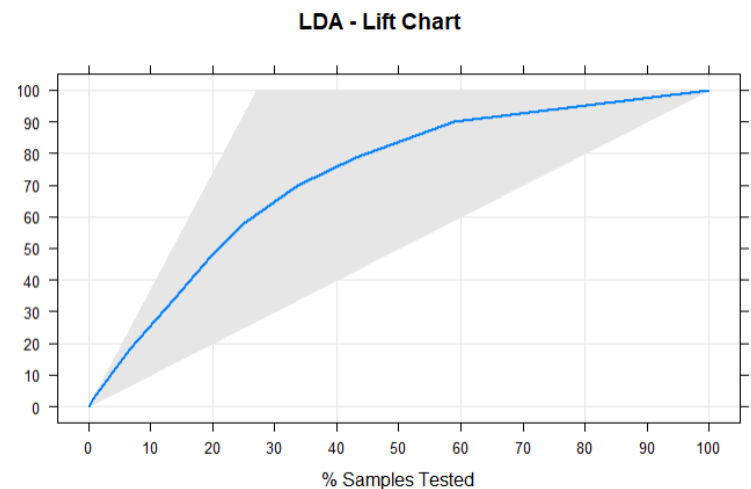
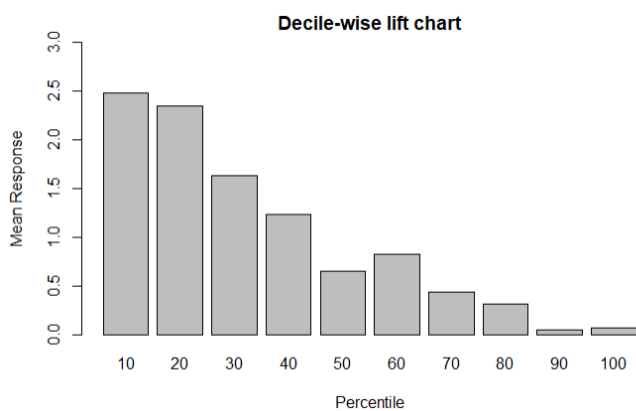
Kappa : 0.4462
McNemar's Test P-Value : 3.860e-06

Sensitivity : 0.8930
Specificity : 0.5272
Pos Pred Value : 0.8349
Neg Pred Value : 0.6478
Prevalence : 0.7281
Detection Rate : 0.6502
Detection Prevalence : 0.7787
Balanced Accuracy : 0.7101

'Positive' Class : 0

LDA is able to effectively identify churned customers with an accuracy of 79%.

11.2.2 Lift and Decile Charts



The Lift curve and Decile charts both show that the model can predict customer attrition with good accuracy. In the decile chart, we can identify that mean of the Classification accuracy for the top 40 Percentile of the data is more than 1.5. Thus, we can say that the model is accurate.

12 MODEL COMPARISON

12.1 STRENGTHS AND WEAKNESSES OF THE MODEL :

12.1.1 KNN:

Strengths:

- It is useful if the number of observations in the training data is huge.
- It is robust to noisy training data especially if we use the inverse square if weighted distance as the “distance”.

Weaknesses:

- KNN is computationally expensive.
- Need to determine value of K.

12.1.2 Logistic Regression:

Strengths:

- Outputs have a nice probabilistic interpretation.
- The algorithm can be regularized to avoid overfitting.
- Logistic models can be updated easily with new data using stochastic gradient descent.

Weaknesses:

- Logistic regression tends to underperform when there are multiple or non-linear decision boundaries.
- They are not flexible enough to naturally capture more complex relationships.

12.1.3 Decision Tree:

Strengths:

- Easy to use and understand.
- Produce rules that are easy to interpret & implement
- Variable selection and reduction is automatic
- Do not require the assumptions of statistical models
- Can work without extensive handling of missing data

Weaknesses:

- Instability and Poor Predictive Performance

12.2 PARAMETER BASED COMPARISON

Sensitivity: If C1 is the important class, sensitivity is defined as the percentage of C1 class correctly classified.

Specificity: If C1 is the important class, specificity is defined as the percentage of C0 class correctly classified.

Precision: Precision describes how precise/ accurate the model is out of those predictive positive, how many of them are actual positive.

It can be described as:

$$\text{True positive} / (\text{True positive} + \text{False positive})$$

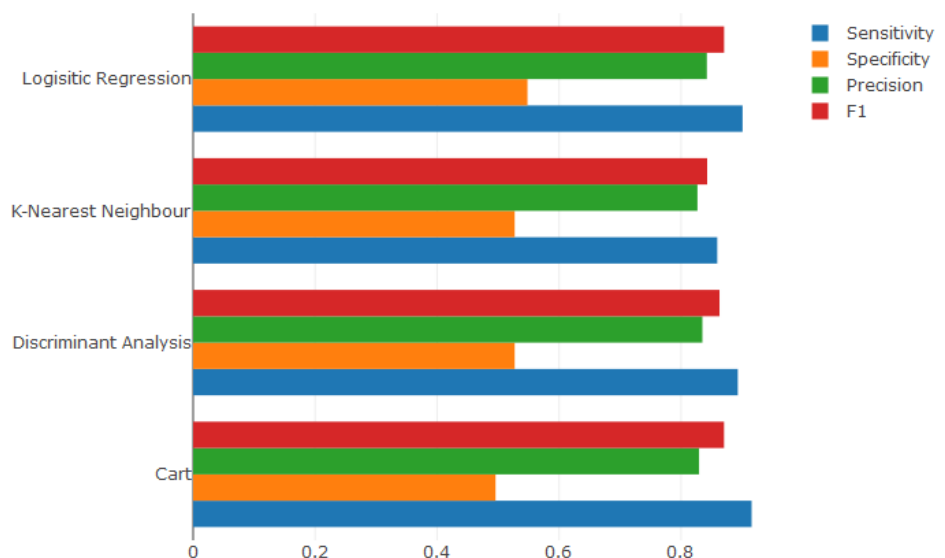
F1 Score: F1 is a function of precision and recall.

It can be described as:

$$F1 = 2 * \{(\text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall})\}$$

F1 score is needed when you need a balance between precision and recall.

For our case, we have considered sensitivity and precision to be the major deciding factors. Hence, **logistic regression** comes out to be the best model out of all.



13 INSIGHTS AND RECOMMENDATION

Based on logistic regression model we have inferred that these parameters are critical and significant in making a prediction.

- **Dependent:** If the customer has a dependent, the odds of the customer churning 0.8 times to that of a customer with no Dependent

Customers who have a dependent, are less likely to churn. Hence, in order to prevent a higher percentage of people from churning the company should come with lucrative family plans, in order to increase the dependents of each customer. This can be done by introducing cheaper data packages in family plans as compared to individual plans.

- **Tenure:** As the tenure of a customer's enrollment increases by 1 unit, the odds of the customer churning is 0.24 times

Customers who have enrolled for a longer period tend to stay. In order to prevent more people from churning out, the company should target to strengthen the customer relationship and increase the longevity of tenure maybe with some personal touch by giving some offers on their special days by understanding customer needs. This in turn creates a loyal customer base.

- **Paperless Billing:** If the customer opts for paperless billing then the odds of a customer churning is 1.4 times

Paperless billing process might be convoluted, and customers might find it difficult to navigate and obtain their monthly bills. We need to re-structure the process for a smooth navigation that would create a seamless experience for the customer.

- **Total Charges:** If the Total Charges of the enrollment increase by 1 unit then the odds of a customer churning are 1.9 times more

The plans are very expensive, and we need to provide reform and rebate, lucrative deals to reduce total cost brand loyal customers.

- **Internet Services:**

- **DSL:** If the customer has opted for a DSL line then the odds of a customer churning is 7.9 times more than customers with No Internet Service

- Optic Fiber: If the customer has opted for an Optic Fiber the odds of a customer churning is 54 times more than customers with No Internet Service

Internet services seems to a pain point for the customer. We need to analyze customer feedbacks and complaints to identify areas of concerns like bandwidth, speed or availability in different areas.

- Contract:

- Month-to-month: If the customer has opted for Month-to-month contract then the odds of a customer churning is 3.3 times more that customers with a Two-year contract
- One Year: If the customer has opted for One-Year contract then the odds of a customer churning is 1.8 times more that customers with a Two-year contract

Customers with month-to- month contract are highly likely to churn as compared to those with One-year contract. Therefore, in order to prevent customer from churning, we need to introduce more membership offers that lowers cost of enrollment, if one takes up 3 months or 6 months contract. This will help to pull customers attention whereas increase the customer base for a longer duration.

- Payment Method: When compared to Payment method by credit card

- Electronic Check: If the customer has opted for Electronic Check payment method then the odds of a customer churning is 1.28 times
- Mailed Check: If the customer has opted for Mailed Check payment method then the odds of a customer churning is 0.97 times
- Automatic bank transfer: If the customer has opted for Automatic bank transfer then the odds of a customer churning is 0.94 times

Different Payment method govern the whether the customer would churn or not. Electronic check payment appears to be less preferred as compared to other payment methods, mailed check and automatic bank transfer. We can promote payment method through mailed check and automatic bank transfer through rebates and cashbacks on payment plans.

14 TABLE OF FIGURES

Figure 1: Gender Distribution.....	11
Figure 2: Senior Citizen Distribution.....	11
Figure 3: Partner Distribution	12
Figure 4: Dependents Distribution.....	12
Figure 5: Phone Service Distribution.....	13
Figure 6: Multiple Lines Distribution	13
Figure 7: Internet Service Distribution	14
Figure 8: Online Security Distribution	14
Figure 9: Online Backup Distribution.....	15
Figure 10: Device Protection Distribution.....	15
Figure 11: Tech Support Distribution	16
Figure 12: Streaming Movies Distribution	16
Figure 13: Streaming TV Distribution.....	17
Figure 14: Contract Distribution.....	17
Figure 15: Paperless Billing Distribution	18
Figure 16: Payment Method Distribution	18
Figure 17: Tenure Group Distribution	19
Figure 18: Monthly Charges VS Tenure based on Churn.....	19
Figure 19: Total Charges VS Tenure based on Churn.....	20
Figure 20: Total Charges VS Monthly Charges based on Churn	20
Figure 21: Monthly Charges VS Total Charges based on Tenure Group.....	21
Figure 22: Churned Customer Radar Chart	22
Figure 23: Non-Churned Customer Radar Chart	23
Figure 24: Sigmoid Function	24
Figure 25: Logistic Regression Output	25
Figure 26: Decile Wise Lift Chart	27
Figure 27: Lift Chart	27
Figure 28: KNN Explanation	28
Figure 29: Cross-Validation Curve.....	29

Figure 30: For $K = 9$	29
Figure 31: For $K = 38$	29
Figure 32: Decision Tree	31

15 APPENDIX

- [R Markdown Code](#) – Please use UTD credentials to access it.

16 REFERENCES

Datacamp. (n.d.). Retrieved from Datacamp.com:
<https://www.datacamp.com/community/tutorials/k-nearest-neighbor-classification-scikit-learn>

Kaggle. (n.d.). Retrieved from Kaggle: https://www.kaggle.com/blatchar/telco-customer-churn/downloads/WA_Fn-UseC_-Telco-Customer-Churn.csv/1