Telco Customer Churn

BUAN 6356.003
Business Analytics with R S22
Analytics

Group 1 - Aakash Singh, Ashna Bhardwaj, Gunjan Mishra, Kruthika Anil More, Raquel Guerra Galdamez

Executive Summary

- This research focuses on the fact that huge industries have a high rate of customer attrition:
  - Because the market is extremely competitive, companies must keep clients engaged in its offerings, and businesses end up spending millions in branding

- The data obtained depicts a customer's plan of numerous enrolment elements.

- We decided to visualize the data in order to better understand the influence of each predictor in the determining of the churn rate.

- We used machine learning techniques to analyze the data and discover important indicators of customer attrition.

Background

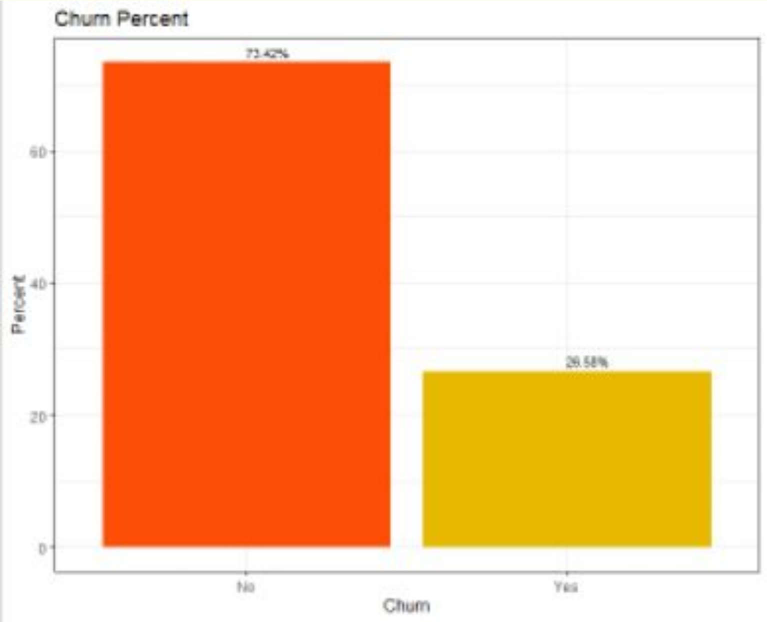
- The rate at which consumers leave doing business with a company is known as the churn rate

- Customer churn is one of the most significant sources of revenue loss in the telecom business

- A reason why a customer might decide to leave is due to the cognitive dissonance phenomenon:
  - When a corporation fails to match a customer's expectations, the disappointment causes customer's post-purchase perplexity

- In this data collection, we're looking for folks who are leaving the AT&T operator.

Objective



⦿Through this project, we aim to identify the major predictors and try to find out the root cause for the voluntary churn.

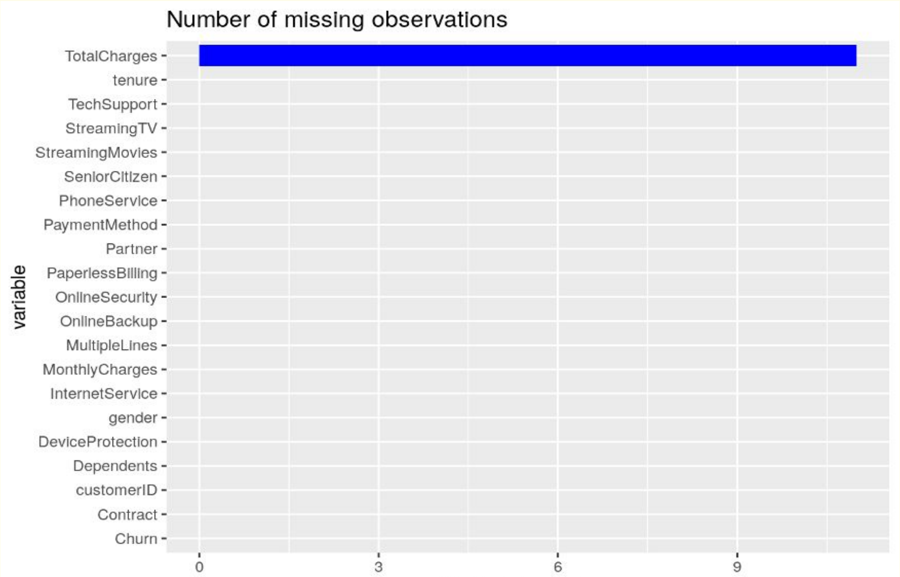
⦿This activity will help the company to increase the customer attrition rate by changing policies and offers to help increase the clientele retention.

⦿Based on the analysis we know that:

•CHURN columns tells us about the number of Customers who left within the last month.

•Around 26% of customers left the platform within the last month.

Data Summary



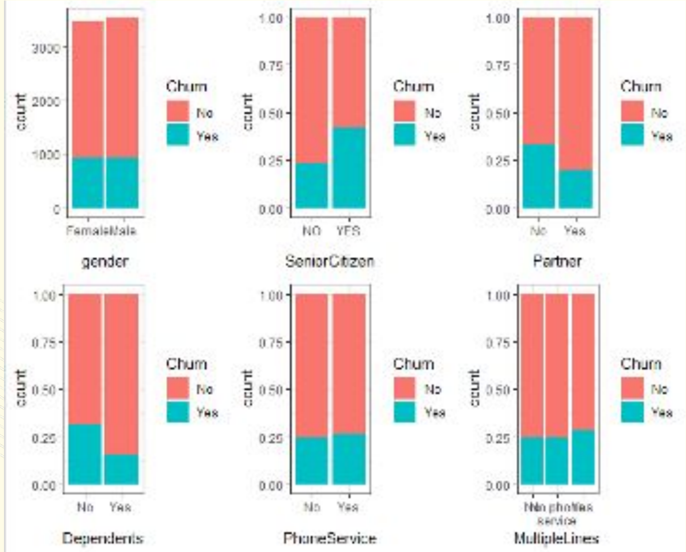
- The data collected has 20 predictors namely gender, senior citizen, partner, tenure etc.
- The total number of records are 7043 out of which only TotalCharges has 11 missing values which we have substituted with the mean of the remaining data
- Senior Citizen is in 'int' form, that can be changed to categorical.
- There are three continuous variables
- Out of 20 predictors, we have 3 numeric columns :-

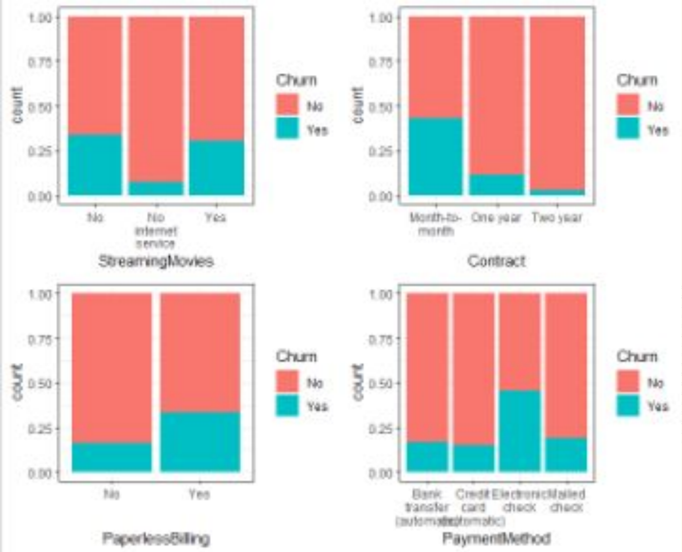
TotalCharges
·mean (sd): 2283.3 (2266.77)
·min < med < max: 18.8 < 1397.47 <8684.8
·IQR (CV): 3393.29 (0.99)

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

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

Churn Based on Categorization





⦿Gender - The churn percent is almost equal in case of Male and Females

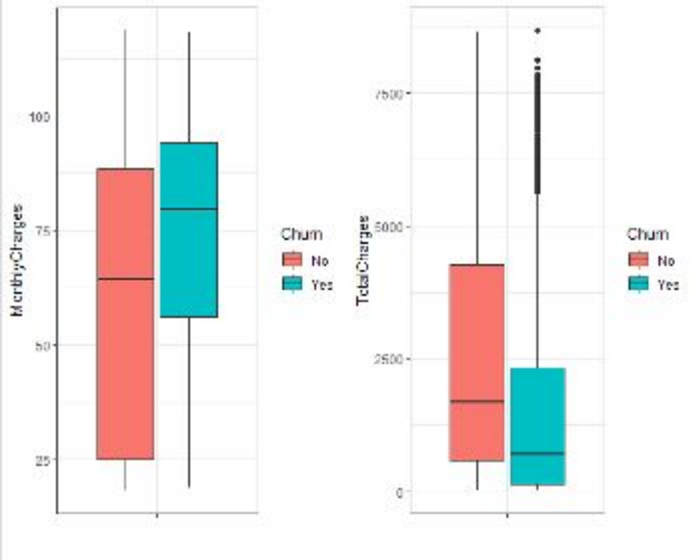
⦿The percent of churn is higher in case of senior citizens

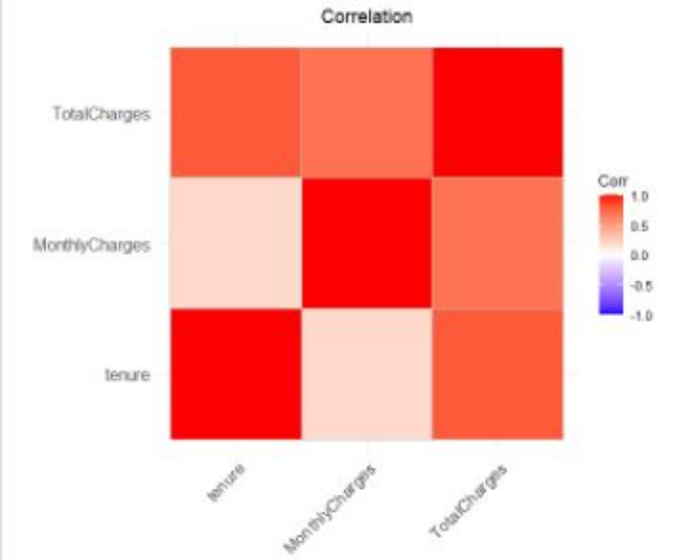
⦿Customers with Partners and Dependents have lower churn rate as compared to those who don't have partners & Dependents.

⦿A larger percent of Customers with monthly subscription have left when compared to Customers with one or two year contract.

⦿Churn percent is higher in case of customers having paperless billing option.

⦿Customers who have Electronic Check Payment Method tend to leave the platform more when compared to other options.





·The median tenure for customers who have left is around 10 months.
·Customers who have churned, have high monthly charges. The median is above 75
·The median Total charges of customers who have churned is low.

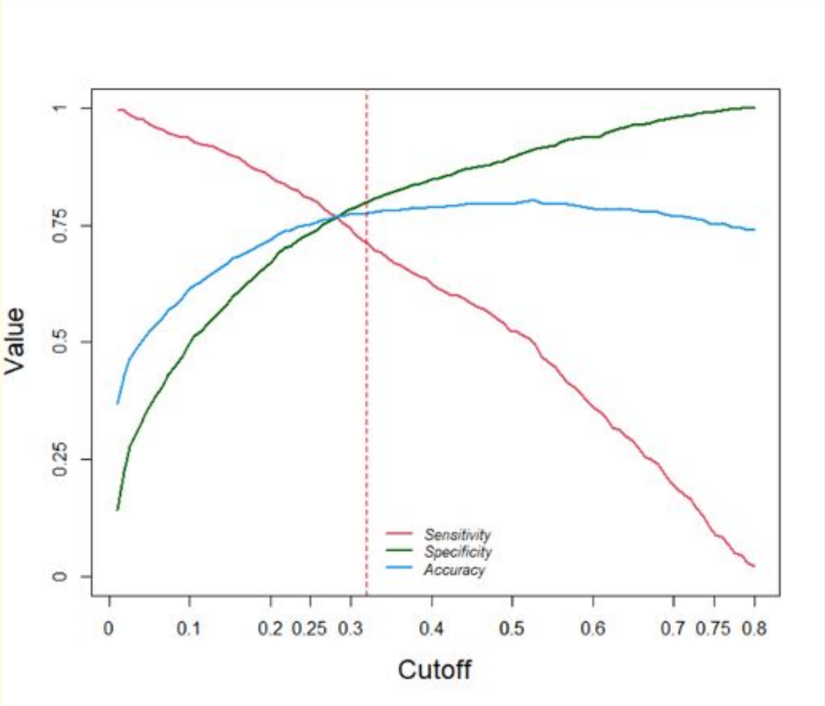
The correlation between continuous variables.
·Total Charges has positive correlation with MonthlyCharges and tenure

Logistic Regression

- ⦿ A predictive analysis is used for analyzing a dataset; this method is 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 variable of interest, i.e. the target variable here is ‘Churn’ which will tell us whether or not a particular customer has churned.
- ⦿ It is a binary variable - 1 means that the customer has churned and 0 means the customer has not churned.

```
Call:
glm(formula = Churn ~ ., family = "binomial", data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9016  -0.6712  -0.2720   0.7125   3.4526

Coefficients: (7 not defined because of singularities)
                                      Estimate Std. Error z value Pr(>|z|)
(Intercept)                         -3.7828384  1.5166146  -2.494 0.012622 *
tenure                              -1.5279395  0.1877298  -8.139 3.98e-16 ***
MonthlyCharges                      -1.9456247  1.1341813  -1.715 0.086264 .
TotalCharges                         0.7888622  0.1957142   4.031 5.56e-05 ***
gender                               0.0005592  0.0782388   0.007 0.994297
Partner                              0.0652922  0.0931267   0.701 0.483234
Dependents                          -0.2303858  0.1065373  -2.162 0.030580 *
PhoneService                         0.8156778  0.7712637   1.058 0.290244
MultipleLines.xNo.phone.service             NA         NA      NA       NA
MultipleLines.xYes                   0.6255750  0.2107995   2.968 0.003001 **
InternetService.xFiber.optic         2.3061941  0.9475150   2.434 0.014935 *
InternetService.xNo                 -2.4458233  0.9579362  -2.553 0.010673 *
OnlineSecurity.xNo.internet.service         NA         NA      NA       NA
OnlineSecurity.xYes                 -0.0728116  0.2144186  -0.340 0.734175
OnlineBackup.xNo.internet.service           NA         NA      NA       NA
OnlineBackup.xYes                    0.1422095  0.2093413   0.679 0.496936
DeviceProtection.xNo.internet.service       NA         NA      NA       NA
DeviceProtection.xYes                0.2161347  0.2119372   1.020 0.307821
TechSupport.xNo.internet.service            NA         NA      NA       NA
TechSupport.xYes                    -0.1348444  0.2169579  -0.622 0.534255
StreamingTV.xNo.internet.service            NA         NA      NA       NA
StreamingTV.xYes                     0.8680622  0.3894207   2.229 0.025806 *
StreamingMovies.xNo.internet.service        NA         NA      NA       NA
StreamingMovies.xYes                 0.8451914  0.3893666   2.171 0.029955 *
Contract.xOne.year                  -0.7281318  0.1324802  -5.496 3.88e-08 ***
Contract.xTwo.year                  -1.3856856  0.2122134  -6.530 6.59e-11 ***
PaperlessBilling                     0.3522914  0.0896886   3.928 8.57e-05 ***
PaymentMethod.xCredit.card..automatic. 0.0063052 0.1381872  0.046 0.963607
PaymentMethod.xElectronic.check      0.3941347  0.1150158   3.427 0.000611 ***
PaymentMethod.xMailed.check         -0.0273945  0.1398321  -0.196 0.844680
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5700.7  on 4923  degrees of freedom
Residual deviance: 4012.1  on 4901  degrees of freedom
  (6 observations deleted due to missingness)
AIC: 4058.1

Number of Fisher Scoring iterations: 6
```

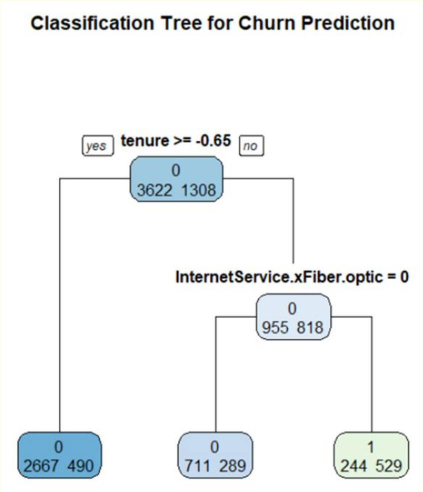
Cutoff & Confusion Matrix




```
>
> confusionMatrix(validation$Churn, DTPred)
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1448  104
         1  350  211

               Accuracy : 0.7851
                 95% CI : (0.767, 0.8025)
    No Information Rate : 0.8509
    P-Value [Acc > NIR] : 1

                  Kappa : 0.3594

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.8053
            Specificity : 0.6698
         Pos Pred Value : 0.9330
         Neg Pred Value : 0.3761
             Prevalence : 0.8509
         Detection Rate : 0.6853
   Detection Prevalence : 0.7345
      Balanced Accuracy : 0.7376

       'Positive' Class : 0
```

- When we are using a cutoff of 0.50 we are getting a good accuracy and specificity, but the sensitivity is very less. Hence, we need to find the optimal probability cutoff which will give maximum accuracy, sensitivity, and specificity hence we used a cutoff value of 0.28 for final model, where the three curves for accuracy, specificity and sensitivity meet.Logistic Regression with a cutoff probability value of 0.28 gives us better values of accuracy, sensitivity , and specificity in the validation data.

Decision Tree




```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1448  104
         1  350  211

               Accuracy : 0.7851
                 95% CI : (0.767, 0.8025)
    No Information Rate : 0.8509
    P-Value [Acc > NIR] : 1

                  Kappa : 0.3594

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.8053
            Specificity : 0.6698
         Pos Pred Value : 0.9330
         Neg Pred Value : 0.3761
             Prevalence : 0.8509
         Detection Rate : 0.6853
   Detection Prevalence : 0.7345
      Balanced Accuracy : 0.7376

       'Positive' Class : 0
```

- Type of supervised Learning algorithm
- Classifier that mostly predicts mostly categorical variables
- Each branch is a node that quantifies the churn request that needs to be predicted against the features it satisfies

NAIVE BAYES' CLASSIFIER

Principle: -

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

A, B = events

$P(A|B)$ = probability of A given B is true

$P(B|A)$ = probability of B given A is true

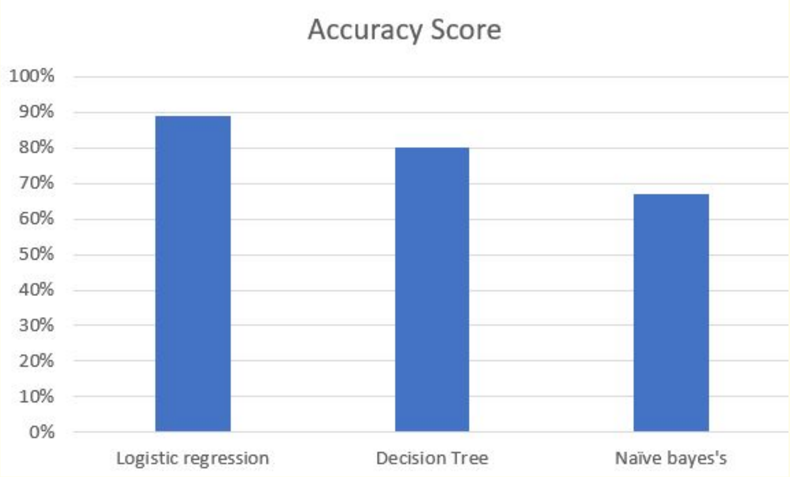
$P(A), P(B)$ = the independent probabilities of A and B

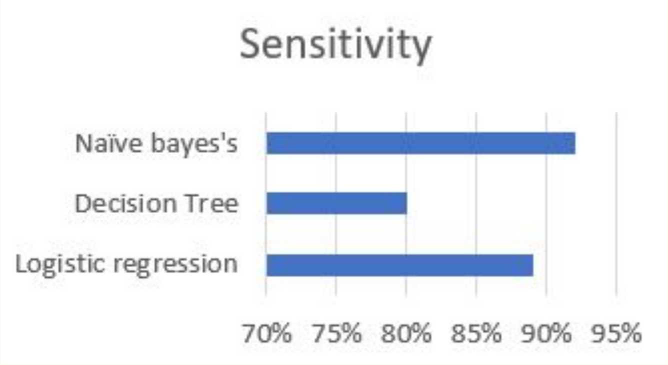
Steps: -

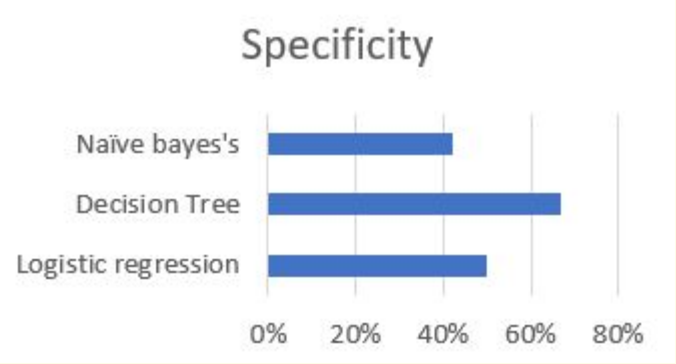
- ⦿ Select the target variable (A) and dependent variable (B)
- ⦿ Calculate all the possible probabilities for the 2 o/p (Yes or No given dependent variables)
- ⦿ Calculate the individual probabilities
- ⦿ Solution is the higher probability

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 2156 1466
         1  166 1142

               Accuracy : 0.669
                 95% CI : (0.6556, 0.6821)
    No Information Rate : 0.529
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.3555

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9285
            Specificity : 0.4379
         Pos Pred Value : 0.5953
         Neg Pred Value : 0.8731
             Prevalence : 0.4710
         Detection Rate : 0.4373
   Detection Prevalence : 0.7347
      Balanced Accuracy : 0.6832

       'Positive' Class : 0
```

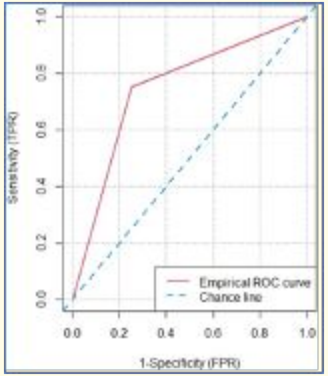
COMPARISON OF DIFFERENT MODELS



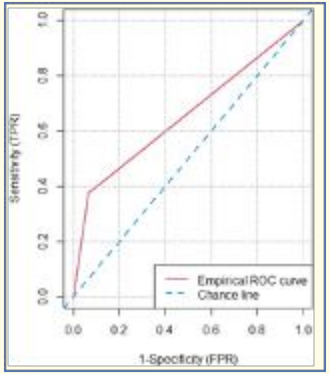




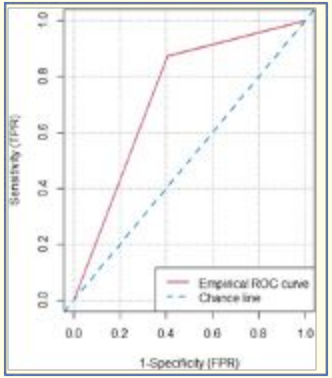
ROC Curve



Logistic Regression



Decision Tree



Naïve Bayes' Classifier

```
> roc_NBC$AUC
[1] 0.73417
> roc_DT$AUC
[1] 0.6545519
> roc_logistic$AUC
[1] 0.7504698
```

·Therefore, from this we can say that the best model in predicting the Churn variable is Logistic Regression that shows an accuracy score of 76% and an AUC score of 75%.

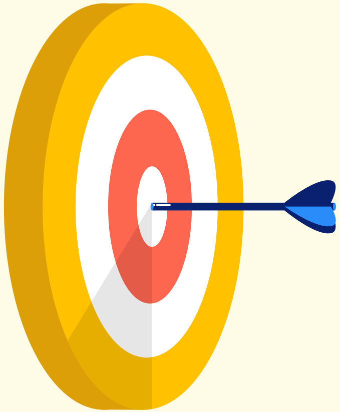
⦿ROC curve is the best indicator to judge a model's efficiency, and in selecting the right model.

⦿From the ROC curve(receiver operating curve), we get the AUC score of each model which tells us about the efficiency of the model in distinguishing between positive and negative classes.

⦿It displays the model which shows the least Type - II error that is represented from the AUC score.

⦿We get the best AUC score from Logistic regression at 75%, while Decision tree gives the worst at 65%.

Conclusion



⦿Attributes and features such as tenure group, Contract,Paperless Billing, Monthly Charges and Internet Service appear to play a role in customer churn.

⦿There seems to be no relationship between the gender and the churn rate.

⦿Customers having a service plan of month-to month contract,with Paperless Billing and are within 12 months tenure, are more likely to churn.

⦿On the other hand, customers with one- or two-year contract, with longer than 12 months tenure, that are not using Paperless Billing, are less likely to churn.

Thanks !