BUAN 6356.003
Business Analytics with R
S22 Analytics

Telco Customer Churn

Group-1

Aakash Singh (AXS200397)

Ashna Bhardwaj (AXB210124)

Gunjan Mishra (GXM210015)

Kruthika Anil More (KXA210028)

Raquel Guerra Galdamez (RXG210050)

Executive Summary

Service businesses spend millions of dollars to make clients feel welcome and devoted to their brand. The market is extremely competitive; thus, the company must keep its clients engaged in its offerings. This research focuses on the fact that huge industries have a high rate of customer attrition, and we are attempting to figure out how to reduce it through this project. The data for this study was obtained from Kaggle. It discusses a customer's plan's numerous enrolment elements. In order to select significant features heuristically, several visualizations give a concise comprehension of the influence of each predictor in determining the churn rate. We next use several machine learning techniques to analyse these data and discover important indicators of customer attrition, such as Logistic Regression, K Nearest Neighbours, Decision Tree, and Linear Discriminant Analysis. The accuracy of these models is compared and rated. By using consumer behaviour and domain expertise, we have developed ideas for preventing client attrition in several domains.

Background

Customer Attrition Analysis, also known as Customer Churn Analysis, is the study of customers who stop using a company's goods. When a consumer stop utilizing telecom services, the company considers them to have churned. It has recently become a crucial issue since acquiring a new client is more difficult and expensive than keeping a current customer, especially when the customer has several alternatives. Because recaptured long-term clients are worth significantly more to a corporation than freshly acquired customers, customer churn rate is one of the most important business indicators.

As involuntary churn is outside the service provider's control, our primary goal is to identify the major reasons for churn and retain customers who willingly quit. As a result, a single model would struggle to capture such complicated patterns, and it is preferable to have a different model for each churn type. Customers' post-purchase perplexity over their purchasing decision is known as cognitive dissonance. When a corporation fails to match a customer's expectations, the disappointment causes cognitive dissonance, and the consumer is more likely to switch to a rival. In this data collection, we're looking for folks who are leaving the AT&T operator. Marketers and retention experts must be able to predict which customers will churn in advance using churn analysis in order to keep customers who would otherwise abandon the business. With this insight, a significant part of client attrition may be avoided.

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.

Current State:

- The CCO team of AT&T is responsible for ensuring customer retention
- The team observes that YoY customer churn rate has increased
- They want to identify key features of a customer enrolment that trigger customer churn

Gaps:

- Customer enrolment features are unknown.
- What are the key enrolment parameters?

Future State:

- Outcome: The CCO team was able to reduce churn by 4%
- Behaviour: A framework was created that targets these potential customers
- Insight: The team was able to identify key enrolment features that are responsible customer churn

Challenges faced

- NA in total charges
- Multi collinearity in all no internet services variable
- While generating the Confusion matrix we were getting an error of data and reference not being on the same level
- Accuracy score of decision tree was high but AUC score was low so there was a possibility of choosing the incorrect model

Data Summary

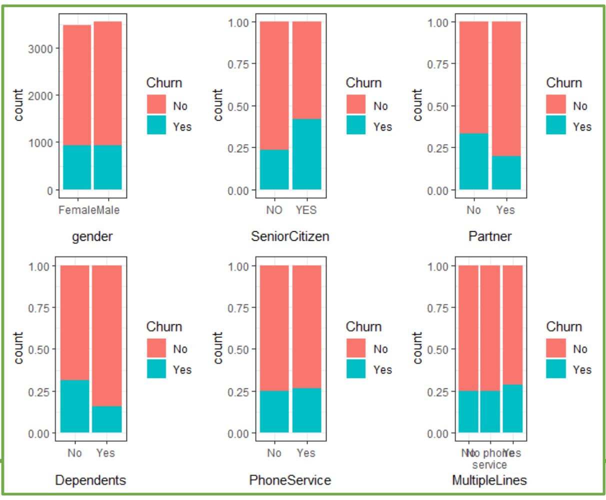
The data used for the project is Customer Churn Data of AT&T, the source of the data from Kaggle.com - *https://www.kaggle.com/code/bandiatindra/telecom-churn-prediction/data.*

Column Name	Column Description
customerID	Customer ID
Gender	Whether the customer is a male or a female
SeniorCitizen	Whether the customer is a senior citizen or not (1, 0)
Partner	Whether the customer has a partner or not (Yes, No)
Dependents	Whether the customer has dependents or not (Yes, No)
Tenure	Number of months the customer has stayed with the company
PhoneService	Whether the customer has a phone service or not (Yes, No)

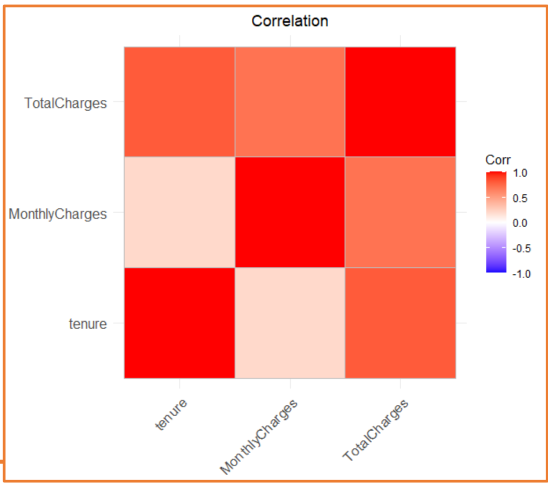
MultipleLines	Whether the customer has multiple lines or not (Yes, No, No phone service)
InternetService	Customer's internet service provider (DSL, Fiber optic, No)
OnlineSecurity	Whether the customer has online security or not (Yes, No, No internet service)
OnlineBackup	Whether the customer has online backup or not (Yes, No, No internet service)
DeviceProtection	Whether the customer has device protection or not (Yes, No, No internet service)
TechSupport	Whether the customer has tech support or not (Yes, No, No internet service)
StreamingTV	Whether the customer has streaming TV or not (Yes, No, No internet service)
StreamingMovies	Whether the customer has streaming movies or not (Yes, No, No internet service)
Contract	The contract term of the customer (Month-to-month, One year, Two year)
PaperlessBilling	Whether the customer has paperless billing or not (Yes, No)
PaymentMethod	The customer's payment method (Electronic check, Mailed check, Bank transfer (automatic)
MonthlyCharges	The amount charged to the customer monthly
TotalCharges	The total amount charged to the customer
Churn	Whether the customer churned or not (Yes or No)

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. 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.

Exploratory Data Analysis



- 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.



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

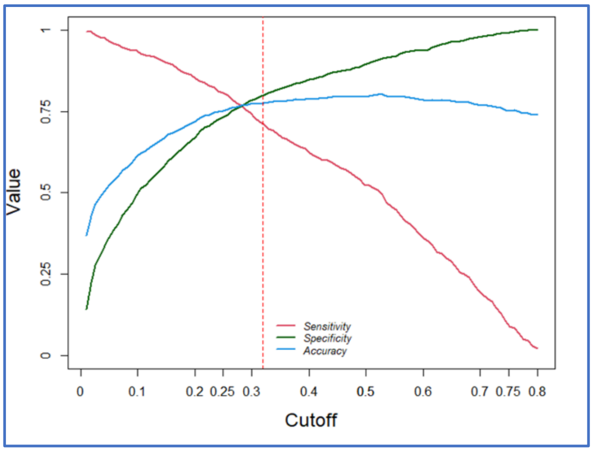
Logistic Regression

Logistic regression, A predictive analysis is used for analysing 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.

Using stepAIC for variable selection, which is a iterative process of adding or removing variables, in order to get a subset of variables that gives the best performing model. Model_3 all has significant variables, so let's just use it for prediction first. model_2<- stepAIC(model, direction="both"). 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.

Let's choose 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 0.32 gives us better values of accuracy, sensitivity, and specificity in the validation data.

```
> model = glm(Churn ~ ., data = train, family = "binomial")
> summary(model)

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_churn <- factor(ifelse(pred >=0.28, "Yes", "No"))
> conf_final <- confusionMatrix(cutoff_churn, actual_churn, positive = "Yes")
> accuracy <- conf_final$overall[1]
> sensitivity <- conf_final$byClass[1]
> specificity <- conf_final$byClass[2]
> accuracy
 Accuracy
0.7667141
> sensitivity
Sensitivity
  0.7682709
> specificity
Specificity
  0.7661499
```

Demographic Information	**Customer Account Information**	**Services Information**
• gender: Whether the client is a female or a male (Female, Male). • SeniorCitizen: Whether the client is a senior citizen or not (0, 1). • Partner: Whether the client has a partner or not (Yes, No). • Dependents: Whether the client has dependents or not (Yes, No)	• tenure: Number of months the customer has stayed with the company (Multiple different numeric values). • Contract: Indicates the customer’s current contract type (Month-to-Month, One year, Two year). • PaperlessBilling: Whether the client has paperless billing or not (Yes, No). • PaymentMethod: The customer’s payment method (Electronic check, mailed check, Bank transfer (automatic), Credit Card (automatic)). • MontlyCharges: The amount charged to the customer monthly (Multiple different numeric values). • TotalCharges: The total amount charged to the customer (Multiple different numeric values).	• PhoneService: Whether the client has a phone service or not (Yes, No). • MultipleLines: Whether the client has multiple lines or not (No phone service, No, Yes). • InternetServices: Whether the client is subscribed to Internet service with the company (DSL, Fiber optic, No) • OnlineSecurity: Whether the client has online security or not (No internet service, No, Yes). • OnlineBackup: Whether the client has online backup or not (No internet service, No, Yes). • DeviceProtection: Whether the client has device protection or not (No internet service, No, Yes). • TechSupport: Whether the client has tech support or not (No internet service, No, Yes). • StreamingTV: Whether the client has streaming TV or not (No internet service, No, Yes) • StreamingMovies: Whether the client has streaming movies or not (No internet service, No, Yes).

2. 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
```

Decision trees can be applied to both classification and regression problems. It is used to predict a qualitative response rather than a quantitative response. We predict that each of the observations belongs to the most commonly occurring class. It is a type of supervised learning algorithm with a predefined target variable. While mostly used in classification tasks, it can handle numeric data as well. This algorithm splits a data sample into two or more homogeneous sets based on the most significant differentiator in input variables to make a prediction.

Splits the data into multiple sets and each set is further split into subsets to arrive at a tree like structure and make a decision. Homogeneity is the basic concept that helps to determine the attribute on which a split should be made.

A split that results into the most homogenous subset is often considered better and step by step each attribute is chosen that maximizes the homogeneity of each subset.

3. Naive Bayee's Classification

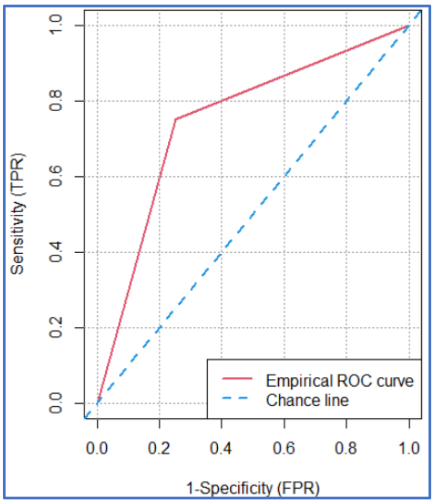
The naive bayee's classification is a family of simple probabilistic classifiers based on applying Bayee's theorem with strong independence assumption between the features or variable. It is a super vised nonlinear algorithm with a pre-defined target variable (Churn).

```
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
```

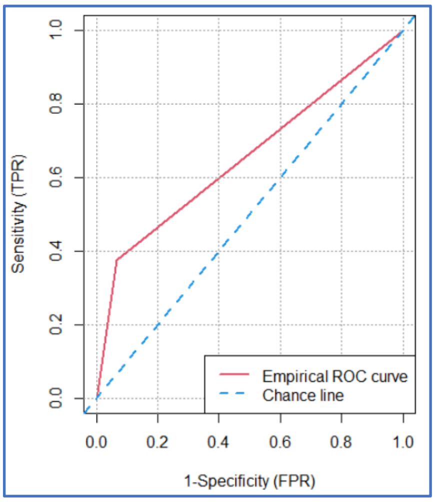
$$Bayees\ Theorem\ P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Confusion Matrix generated from the Regression model calculated for the test data set. The decision tree model (accuracy - 78.5%) gives slightly better accuracy with respect to the logistic regression model (accuracy 76.6%) and Naive Bayee's Class(67%). The sensitivity is also better in case of Decision tree which is 80.53%. However, the specificity has decreased to 66.98% in case of Decision Tree as compared to logistic regression model. When we look at the factors that represent which model must be selected. There are various factors to choose from such as Accuracy score, Sensitivity and Specificity. But we shall plot the roc curves to decide which model gives the best result.

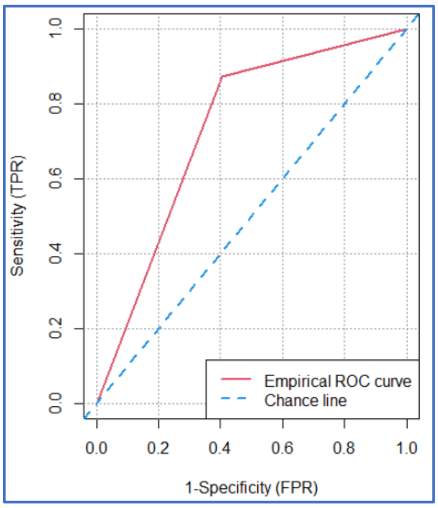
4. ROC Curve



3Logistics Regression



2Decision Tree



1Naive Bayee's Classifier

From the ROC 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. We get the best AUC score from Logistic regression at 75%, while Decision tree gives the worst at 65%. ROC curve is the best indicator to judge a model's efficiency, and in selecting the right model. It displays the model which shows the least Type - II error that is represented from the AUC score. 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_NBC$AUC
[1] 0.73417
> roc_DT$AUC
[1] 0.6545519
> roc_logistic$AUC
[1] 0.7504698
```