**Project Report**

**Group 3 --- (Members of team)**

Elmy, Shivani, Vashini, Tongxiang Lu

**Project Goal**

In this project, we use historical data from ABC Wireless Inc. to build a model which can help predict their customers who are likely to churn.

**Overview of Data**

This project uses ABC Wireless Inc. data of customers’ churn across US states. Data includes 20 variables with 5 categorical variables, 7 integer variables and 8 numerical variables. Data consists of 3333 observations. The main variable of interest is the customers’ churn which is categorical variable with “yes” if a customer switches to a different provider from the current one and “no” if she stays with the current provider. Out of 333 customers, 483 churn and 2850 do not. Table 1 presents the summary of data.

**Table 1: Summary Statistics**

## state account\_length area\_code international\_plan  
## Length:3333 Min. :-209.00 Length:3333 Length:3333   
## Class :character 1st Qu.: 72.00 Class :character Class :character   
## Mode :character Median : 100.00 Mode :character Mode :character   
## Mean : 97.32   
## 3rd Qu.: 127.00   
## Max. : 243.00   
## NA's :501   
## voice\_mail\_plan number\_vmail\_messages total\_day\_minutes total\_day\_calls  
## Length:3333 Min. :-10.000 Min. : 0.0 Min. : 0.0   
## Class :character 1st Qu.: 0.000 1st Qu.: 149.3 1st Qu.: 87.0   
## Mode :character Median : 0.000 Median : 190.5 Median :101.0   
## Mean : 7.333 Mean : 418.9 Mean :100.3   
## 3rd Qu.: 16.000 3rd Qu.: 237.8 3rd Qu.:114.0   
## Max. : 51.000 Max. :2185.1 Max. :165.0   
## NA's :200 NA's :200 NA's :200   
## total\_day\_charge total\_eve\_minutes total\_eve\_calls total\_eve\_charge  
## Min. : 0.00 Min. : 0.0 Min. : 0.0 Min. : 0.00   
## 1st Qu.:24.45 1st Qu.: 170.5 1st Qu.: 87.0 1st Qu.:14.14   
## Median :30.65 Median : 209.9 Median :100.0 Median :17.09   
## Mean :30.63 Mean : 324.3 Mean :100.1 Mean :17.08   
## 3rd Qu.:36.84 3rd Qu.: 257.6 3rd Qu.:114.0 3rd Qu.:20.00   
## Max. :59.64 Max. :1244.2 Max. :170.0 Max. :30.91   
## NA's :200 NA's :301 NA's :200 NA's :200   
## total\_night\_minutes total\_night\_calls total\_night\_charge total\_intl\_minutes  
## Min. : 23.2 Min. : 33.0 Min. : 1.040 Min. : 0.00   
## 1st Qu.:167.3 1st Qu.: 87.0 1st Qu.: 7.530 1st Qu.: 8.50   
## Median :201.4 Median :100.0 Median : 9.060 Median :10.30   
## Mean :201.2 Mean :100.1 Mean : 9.054 Mean :10.23   
## 3rd Qu.:235.3 3rd Qu.:113.0 3rd Qu.:10.590 3rd Qu.:12.10   
## Max. :395.0 Max. :175.0 Max. :17.770 Max. :20.00   
## NA's :200 NA's :200 NA's :200   
## total\_intl\_calls total\_intl\_charge number\_customer\_service\_calls  
## Min. : 0.00 Min. :0.000 Min. :0.000   
## 1st Qu.: 3.00 1st Qu.:2.300 1st Qu.:1.000   
## Median : 4.00 Median :2.780 Median :1.000   
## Mean : 4.47 Mean :2.762 Mean :1.561   
## 3rd Qu.: 6.00 3rd Qu.:3.270 3rd Qu.:2.000   
## Max. :20.00 Max. :5.400 Max. :9.000   
## NA's :301 NA's :200 NA's :200   
## churn   
## Length:3333   
## Class :character   
## Mode :character

1. **Data Exploration**

This section involves summarizing the main characteristics of the dataset. Understanding the characteristics of the data that we have plays a critically important role in the success of customer churn analysis. So, we start our data exploration by transforming categorical variables into numeric ones. Then we examine the skewness and distribution of each variable in the dataset.

* 1. **Skewness and distribution of the variables**

We start our data exploration with the distribution of the variables. Figure 1 depicts the skewness of variables.

The graphs in figure 1 below show that the majority of the variables are of bell-shaped confirming the normal distribution. However, some of the variables do not have normal distribution. For example, “total day minutes” and “total evening minutes” have a tiny percentage or sizeable quantity spread to the right with positively skewed tail. Similarly, total nigh calls also show somewhat different than bell-shaped. It seems it has some outliers. The “Customer Service calls” has an irregular skewness.

Figure 1: Skewness of variables

Diagram, shape, polygon

Description automatically generated

* 1. **Breaking down the customers churn and no churn**

Figure 2 splits the number of customers into two categories: those who churn (yes) and those who don’t (no). It can be determined from the above graph that among the customers, 483 customers have switched to other providers while the remaining 2850 of them have decided to stay.

Figure 2: Churn and not churn number of customers

Chart, bar chart

Description automatically generated

* 1. **Breaking down the churn rate across states**

We also break down the churn rate across states. Figure 3 portrays the state-wise churn rate. We put state on the horizontal axis and the churn rate on the vertical axis. The taller the bar, the larger the churn rate. For example, Maryland, New Jersey, Michigan, and Texas are the states with high churn rates.

Figure 3: Churn rate for each state

Chart, bar chart, line chart, histogram

Description automatically generated

* 1. **Customer churn for total day charge**

Next, we decompose the churn data by the total day charges by using the ggplot in R. Box plot in figure 4 shows such decomposition. Measuring churn (yes or no) in the x-axis and total day charges in the y-axis, box plot reveals that customers having the day charge between 30-40 are more inclined towards cancelling their services with the current providers and shift to a different provider.

Figure 4: Box plot of churn data for total day charge

Chart, box and whisker chart

Description automatically generated

* 1. **Churn count with and without international plan**

We also split the churn data to determine the customers who have the international package and shifted to another provider based on the dataset. Figure 5 does this thing. The results depict the percentage of customers who are a part of the international plan and have moved to another provider i.e. 28% of the customers are likely to churn.

Figure 5: Churn count with and without international plan

Chart, bar chart

Description automatically generated

* 1. **Churn data for number of customer service calls**

One more split of the churn data we are doing is based on the number of customer service calls. Figure 6 is for this purpose. The box plot obviously depicts that the customers who have reached out to the customer services more than 2-4 times are

likely to move to other providers. We can interpret that the customers who have churned are approximately 64% and the reason being, they reach out to the customer service 1-4 times.

Figure 6: Churn data for number of customer service calls

Chart, box and whisker chart

Description automatically generated

* 1. **Data cleaning**

Data cleaning is the process of detecting and correcting or removing corrupt or inaccurate records from a dataset and refers to identifying incomplete or inaccurate

parts of the data and then replacing, modifying, or deleting the them. In this presentation we will focus on methods that used to handle missing data and outliers.

Here, what we do is sort and impute the missing values using mice package in R. Then we generate the complete data using mice imputation in random forests. Table 2 summarizes this work.

**Table 2: Summary of the cleaned data**

churn.train\_data\_imputed <- mutate(mice\_output,churn=churn.train\_data$churn)  
summary(churn.train\_data)

## state account\_length area\_code international\_plan  
## WV : 106 Min. :-209.00 area\_code\_408: 838 no :3010   
## MN : 84 1st Qu.: 72.00 area\_code\_415:1655 yes: 323   
## NY : 83 Median : 100.00 area\_code\_510: 840   
## AL : 80 Mean : 97.32   
## OH : 78 3rd Qu.: 127.00   
## OR : 78 Max. : 243.00   
## (Other):2824 NA's :501   
## voice\_mail\_plan number\_vmail\_messages total\_day\_minutes total\_day\_calls

## no :2411 Min. :-10.000 Min. : 0.0 Min. : 0.0   
## yes: 922 1st Qu.: 0.000 1st Qu.: 149.3 1st Qu.: 87.0   
## Median : 0.000 Median : 190.5 Median :101.0   
## Mean : 7.333 Mean : 418.9 Mean :100.3   
## 3rd Qu.: 16.000 3rd Qu.: 237.8 3rd Qu.:114.0   
## Max. : 51.000 Max. :2185.1 Max. :165.0   
## NA's :200 NA's :200 NA's :200   
## total\_day\_charge total\_eve\_minutes total\_eve\_calls total\_eve\_charge  
## Min. : 0.00 Min. : 0.0 Min. : 0.0 Min. : 0.00   
## 1st Qu.:24.45 1st Qu.: 170.5 1st Qu.: 87.0 1st Qu.:14.14   
## Median :30.65 Median : 209.9 Median :100.0 Median :17.09   
## Mean :30.63 Mean : 324.3 Mean :100.1 Mean :17.08   
## 3rd Qu.:36.84 3rd Qu.: 257.6 3rd Qu.:114.0 3rd Qu.:20.00   
## Max. :59.64 Max. :1244.2 Max. :170.0 Max. :30.91   
## NA's :200 NA's :301 NA's :200 NA's :200   
## total\_night\_minutes total\_night\_calls total\_night\_charge total\_intl\_minutes  
## Min. : 23.2 Min. : 33.0 Min. : 1.040 Min. : 0.00   
## 1st Qu.:167.3 1st Qu.: 87.0 1st Qu.: 7.530 1st Qu.: 8.50   
## Median :201.4 Median :100.0 Median : 9.060 Median :10.30   
## Mean :201.2 Mean :100.1 Mean : 9.051 Mean :10.23   
## 3rd Qu.:235.3 3rd Qu.:113.0 3rd Qu.:10.590 3rd Qu.:12.10   
## Max. :395.0 Max. :175.0 Max. :17.770 Max. :20.00

## NA's :200 NA's :200 NA's :200   
## total\_intl\_calls total\_intl\_charge number\_customer\_service\_calls churn   
## Min. : 0.00 Min. :0.000 Min. :0.000 no :2850   
## 1st Qu.: 3.00 1st Qu.:2.300 1st Qu.:1.000 yes: 483   
## Median : 4.00 Median :2.780 Median :1.000   
## Mean : 4.47 Mean :2.761 Mean :1.561   
## 3rd Qu.: 6.00 3rd Qu.:3.270 3rd Qu.:2.000   
## Max. :20.00 Max. :5.400 Max. :9.000   
## NA's :301 NA's :200 NA's :200

str(churn.train\_data)

## 'data.frame': 3333 obs. of 20 variables:  
## $ state : Factor w/ 51 levels "AK","AL","AR",..: 34 12 8 12 36 25 28 39 13 16 ...  
## $ account\_length : int 125 108 82 NA 83 89 135 28 86 65 ...  
## $ area\_code : Factor w/ 3 levels "area\_code\_408",..: 3 2 2 1 2 2 2 2 1 2 ...  
## $ international\_plan : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...  
## $ voice\_mail\_plan : Factor w/ 2 levels "no","yes": 1 1 1 2 1 1 1 1 1 1 ...  
## $ number\_vmail\_messages : int 0 0 0 30 0 0 0 0 0 0 ...  
## $ total\_day\_minutes : num 2013 292 300 110 337 ...  
## $ total\_day\_calls : int 99 99 109 71 120 81 81 87 115 137 ...  
## $ total\_day\_charge : num 28.7 49.6 51 18.8 57.4 ...  
## $ total\_eve\_minutes : num 1108 221 181 182 227 ...

## $ total\_eve\_calls : int 107 93 100 108 116 74 114 92 112 83 ...  
## $ total\_eve\_charge : num 14.9 18.8 15.4 15.5 19.3 ...  
## $ total\_night\_minutes : num 243 229 270 184 154 ...  
## $ total\_night\_calls : int 92 110 73 88 114 120 82 112 95 111 ...  
## $ total\_night\_charge : num 2 10.31 12.15 8.27 6.93 ...  
## $ total\_intl\_minutes : num 10.9 14 11.7 11 15.8 9.1 10.3 10.1 9.8 12.7 ...  
## $ total\_intl\_calls : int 7 9 4 8 7 4 6 3 7 6 ...  
## $ total\_intl\_charge : num 0.5 3.78 3.16 2.97 4.27 2.46 2.78 2.73 2.65 3.43 ...  
## $ number\_customer\_service\_calls: int 0 2 0 2 0 1 1 3 2 4 ...  
## $ churn : Factor w/ 2 levels "no","yes": 1 2 2 1 2 1 1 1 1 2 ...

* 1. **Correlation between variables**

We use ggplot to represent the correlation between the variables where churn is equal to yes. As can be seen in the correlation plot, it can be depicted that for the people who have churned, there lies a significant negative correlation between total day charge and the number of customer service calls and also total international charges and total evening charges. The statistics show that customer service calls have a greater churn rate than other calls since the charges are higher.

Chart

Description automatically generated

1. **Modelling Strategy**

This section describes the empirical framework used to analyze the customer’ churn. Using a predictive model such as regression and decision tree, it is possible to demonstrate the influence of various variables on the churn and the importance of each in foreseeing the outcome of the dependent variable.

A logistic regression model is preferred to others since the dependent variable (target variable) in this data is categorical and classification being our prime objective. While in a linear regression model, performance probability may be negative or more than 1, making it ineffective for predicting a binomial feature. The best result for this model is a likelihood of possibilities that fall between 0 and 1 i.e. logistic regression.

For our analysis we will be using both the models and selecting the best among the two to be the final model. Using Logistic Regression and Decision Tree Models to determine Predictive Ability: Before choosing a model, the following procedures are followed:

* The dataset has been divided into training and validation sets to prevent overfitting the model.
* Constructing a logistic regression model and forecasting the outcomes from the validation set.
* Using a confusion matrix to confirm the validity of the model.
* Making a decision tree model and predicting the results of the validation set.
* Validating the model’s performance with a confusion matrix.
* Considering the results of both models and selecting the best one.

**4.1. Data Partitioning**

set.seed(111)  
index<- createDataPartition(churn.train\_data\_imputed $churn,p=0.8,list=FALSE)  
train\_data<-churn.train\_data\_imputed [index,]  
valid\_data <- churn.train\_data\_imputed [-index,]

**4.2. Building a Logistic Regression model**

Logistic regression is a statistical analytic approach for predicting a binary outcome, such as yes or no.

set.seed(222)  
log\_model <- glm(churn~.,data=train\_data ,family = "binomial" ) *#summary(Logistic Model)*  
predict\_valid<-predict(log\_model,valid\_data,type="response")  
head(predict\_valid)

## 10 21 28 34 39 40   
## 0.16572725 0.07577522 0.05251576 0.02260023 0.22631119 0.02431119

result\_check<-ifelse(predict\_valid > 0.5,'yes','no')  
*#Accuracy Check*  
error<-mean(result\_check!=valid\_data$churn)  
accuracy <-1- error  
print(accuracy)

## [1] 0.8468468

plot.roc(valid\_data$churn,predict\_valid)

## Setting levels: control = no, case = yes

## Setting direction: controls < cases

Figure 7: Specificity and sensitivity

Chart, line chart

Description automatically generated

#Using confusion matrix for the logistic regression model.

set.seed(333)  
log\_confusion\_matrix <- confusionMatrix(as.factor(result\_check),as.factor(valid\_data$churn))  
log\_confusion\_matrix

## Confusion Matrix and Statistics  
##

## Reference  
## Prediction no yes  
## no 549 81  
## yes 21 15   
## Accuracy : 0.8468   
## 95% CI : (0.8172, 0.8734)  
## No Information Rate : 0.8559   
## P-Value [Acc > NIR] : 0.7653   
## Kappa : 0.1613   
## Mcnemar's Test P-Value : 5.162e-09   
## Sensitivity : 0.9632   
## Specificity : 0.1562   
## Pos Pred Value : 0.8714   
## Neg Pred Value : 0.4167

## Prevalence : 0.8559   
## Detection Rate : 0.8243   
## Detection Prevalence : 0.9459   
## Balanced Accuracy : 0.5597   
## 'Positive' Class : no

Results produced from the confusion matrix:

* Accuracy: 84.68%
* Sensitivity: 96.32%
* Specificity: 15.62%

**4.3. Building a Decision Tree Model**

Decision tree analysis is basically producing a tree-shaped diagram to chart out a course of action or a statistical probability analysis.

set.seed(444)  
decisiontree\_model<- rpart(churn ~ .,data=train\_data,method = 'class')  
*# Show the variable importance*  
*#DT\_model$variable.importance*

*# Show the split for variable*  
head(decisiontree\_model$splits)

## count ncat improve index adj  
## total\_day\_charge 2667 -1 78.51181 44.975 0  
## number\_customer\_service\_calls 2667 -1 57.34523 3.500 0  
## international\_plan 2667 2 37.82693 1.000 0  
## total\_day\_minutes 2667 -1 22.37794 263.600 0  
## state 2667 51 15.22592 2.000 0  
## number\_customer\_service\_calls 2485 -1 59.34436 3.500 0

*#Predicting the probability*  
prob\_decisiontree <- predict(decisiontree\_model, newdata = valid\_data, type = "prob")  
*#determining AUC Value*  
roc(valid\_data$churn,prob\_decisiontree[,2])

## Setting levels: control = no, case = yes

## Setting direction: controls < cases

##   
## Call:  
## roc.default(response = valid\_data$churn, predictor = prob\_decisiontree[, 2])  
##   
## Data: prob\_decisiontree[, 2] in 570 controls (valid\_data$churn no) < 96 cases

(valid\_data$churn yes).  
## Area under the curve: 0.8234

Using a Confusion Matrix for the Decision Tree Model.

set.seed(555)  
decisiontree\_class<- predict(decisiontree\_model, newdata = valid\_data, type = "class")  
confusionMatrix(as.factor(decisiontree\_class),as.factor(valid\_data$churn))

Confusion Matrix and Statistics  
## Reference  
## Prediction no yes  
## no 555 42  
## yes 15 54   
## Accuracy : 0.9144   
## 95% CI : (0.8905, 0.9345)  
## No Information Rate : 0.8559   
## P-Value [Acc > NIR] : 3.035e-06   
## Kappa : 0.6072   
##   
## Mcnemar's Test P-Value : 0.0005736   
## Sensitivity : 0.9737   
## Specificity : 0.5625   
## Pos Pred Value : 0.9296

## Neg Pred Value : 0.7826   
## Prevalence : 0.8559   
## Detection Rate : 0.8333   
## Detection Prevalence : 0.8964   
## Balanced Accuracy : 0.7681   
## 'Positive' Class : no

From the Confusion Matrix, the following conclusions have been made :-

* Accuracy: 91.44%
* Sensitivity: 97.37%
* Specificity: 56.25%

**4.4. Choosing the optimal model**

On the comparison of the two models, Decision Tree Model is interpreted the best model to put in use as it has higher accuracy than the logistical regression model.

Though the Sensitivities of both the models are almost equal, Decision Tree has a higher specificity. Therefore, Decision Tree Model is the right and optimal model to use.

Predicting the churn using the test data and the decision tree algorithm for the final model analysis.

After the accuracy has been tested for the validation and training data, we can use the entire data to build the final model. Actual dataset can be used to predict the churn only after testing for accuracy.

set.seed(666)  
ABC\_model<- rpart(churn ~ .,data= churn.train\_data\_imputed,method = 'class')

*#Model Splits.*  
head(ABC\_model$splits)

## count ncat improve index adj  
## total\_day\_charge 3333 -1 88.13813 44.975 0

## number\_customer\_service\_calls 3333 -1 68.18448 3.500 0  
## international\_plan 3333 2 55.77483 1.000 0  
## total\_day\_minutes 3333 -1 26.21947 223.250 0  
## state 3333 51 14.95004 2.000 0  
## number\_customer\_service\_calls 3116 -1 71.09667 3.500 0

*#Plotting Decision Tree*  
fancyRpartPlot(ABC\_model)

Figure 8: Plot of Decision TreeTimeline

Description automatically generated

Figure 9: Probability Prediction (decision tree)

Diagram

Description automatically generated

*Probability Prediction(decision tree)*  
decisiontree\_prob <- predict(ABC\_model, newdata = churn.train\_data\_imputed, type = "prob")  
*#Determining the AUC Value*  
roc(churn.train\_data\_imputed$churn,decisiontree\_prob[,2])

## Setting levels: control = no, case = yes

## Setting direction: controls < cases  
## Call:  
## roc.default(response = churn.train\_data\_imputed$churn, predictor = decisiontree\_prob[, 2])

##   
## Data: decisiontree\_prob[, 2] in 2850 controls (churn.train\_data\_imputed$churn no) < 483 cases (churn.train\_data\_imputed$churn yes).  
## Area under the curve: 0.8879

set.seed(777)  
load("~/Desktop/MS BA/Business Analytics/Assignment-4/Customers\_To\_Predict.RData")  
  
count(Customers\_To\_Predict)

## # A tibble: 1 × 1  
## n  
## <int>  
## 1 1600

**Summary(Customers\_To\_Predict)**

## state account\_length area\_code international\_plan  
## Length:1600 Min. : 1.00 Length:1600 Length:1600   
## Class :character 1st Qu.: 71.00 Class :character Class :character   
## Mode :character Median : 98.00 Mode :character Mode :character   
## Mean : 98.52   
## 3rd Qu.:126.00   
## Max. :238.00   
## voice\_mail\_plan number\_vmail\_messages total\_day\_minutes total\_day\_calls

## Length:1600 Min. : 0.000 Min. : 6.6 Min. : 34.00   
## Class :character 1st Qu.: 0.000 1st Qu.:143.8 1st Qu.: 86.00

## Mode :character Median : 0.000 Median :180.9 Median : 99.00   
## Mean : 7.043 Mean :181.6 Mean : 99.06   
## 3rd Qu.: 0.000 3rd Qu.:215.9 3rd Qu.:112.00   
## Max. :52.000 Max. :351.5 Max. :160.00   
## total\_day\_charge total\_eve\_minutes total\_eve\_calls total\_eve\_charge  
## Min. : 1.12 Min. : 22.3 Min. : 38.0 Min. : 1.90   
## 1st Qu.:24.45 1st Qu.:165.8 1st Qu.: 88.0 1st Qu.:14.10   
## Median :30.76 Median :199.9 Median :101.0 Median :17.00   
## Mean :30.87 Mean :199.6 Mean :100.6 Mean :16.96   
## 3rd Qu.:36.70 3rd Qu.:231.8 3rd Qu.:114.0 3rd Qu.:19.70   
## Max. :59.76 Max. :359.3 Max. :169.0 Max. :30.54   
## total\_night\_minutes total\_night\_calls total\_night\_charge total\_intl\_minutes  
## Min. : 0.0 Min. : 0.00 Min. : 0.000 Min. : 0.00   
## 1st Qu.:166.6 1st Qu.: 86.00 1st Qu.: 7.500 1st Qu.: 8.60   
## Median :199.2 Median : 99.00 Median : 8.960 Median :10.40   
## Mean :199.2 Mean : 99.45 Mean : 8.963 Mean :10.32   
## 3rd Qu.:232.4 3rd Qu.:113.00 3rd Qu.:10.463 3rd Qu.:12.00   
## Max. :381.6 Max. :170.00 Max. :17.170 Max. :19.70   
## total\_intl\_calls total\_intl\_charge number\_customer\_service\_calls  
## Min. : 0.000 Min. :0.000 Min. :0.000   
## 1st Qu.: 3.000 1st Qu.:2.320 1st Qu.:1.000

## Median : 4.000 Median :2.810 Median :1.000   
## Mean : 4.356 Mean :2.786 Mean :1.583   
## 3rd Qu.: 5.000 3rd Qu.:3.240 3rd Qu.:2.000   
## Max. :19.000 Max. :5.320 Max. :7.000

*#Checking NA Values*  
colMeans(is.na(Customers\_To\_Predict))

## state account\_length   
## 0 0   
## area\_code international\_plan   
## 0 0   
## voice\_mail\_plan number\_vmail\_messages   
## 0 0   
## total\_day\_minutes total\_day\_calls   
## 0 0   
## total\_day\_charge total\_eve\_minutes   
## 0 0   
## total\_eve\_calls total\_eve\_charge   
## 0 0   
## total\_night\_minutes total\_night\_calls   
## 0 0   
## total\_night\_charge total\_intl\_minutes   
## 0 0   
## total\_intl\_calls total\_intl\_charge   
## 0 0

## number\_customer\_service\_calls   
## 0

prob\_churn <- predict(ABC\_model,Customers\_To\_Predict,type = "prob")  
head(prob\_churn)

## no yes  
## 1 0.9683973 0.03160271  
## 2 0.9683973 0.03160271  
## 3 0.9683973 0.03160271  
## 4 0.9289941 0.07100592  
## 5 0.9683973 0.03160271  
## 6 0.6756757 0.32432432

predict\_churn <- predict(ABC\_model,Customers\_To\_Predict,type = "class")  
head(predict\_churn)

## 1 2 3 4 5 6   
## no no no no no no   
## Levels: no yes

predict\_churn<- as.data.frame(predict\_churn)  
summary(predict\_churn)

## predict\_churn  
## no :1460   
## yes: 140

ggplot(predict\_churn) +  
 aes(x = predict\_churn) +  
 geom\_bar(fill = "red")+  
 labs(x = "Customers Not Churning/Churning",  
 y = "Number of Customers", title = "Number of Customers likely to Churn") +  
 theme\_minimal() +  
 theme(plot.title = element\_text(size = 14L,  
 face = "bold", hjust = 0.5), axis.title.y = element\_text(size = 14L, face = "bold"), axis.title.x = element\_text(size = 14L,face = "bold"))

Figure 10: Number of customers likely to churn

Chart, bar chart

Description automatically generated

From figure 10, we can predict that the customers who churn are 140 and who do not churn are 1460.

**Details of our Modeling Strategy**

We particularly utilized various tools and techniques of analytics, including skewness, box plot, correlation, regression, and decision tree to see how variables of ABC company are related to each other and to what factors the customer’s churn is related.

A clear understanding of the factors that influence churn may assist companies to craft an appropriate policy environment that enhances sales and profits of the company.

**Insights and Conclusions**

By data analysis, our insights are stated as below:

* First, customers are more inclined to switch to another provider if they have paid more than $30 in daily fees.
* Second, customers will undoubtedly go to another supplier if they have to pay international day charges, telling by 28% of clients left the company.
* Third, the results show that the company has dissatisfactory customers, because those customers who have called customer service 2-4 times have left the company.
* Finally, states such as Maryland, New Jersey, Michigan, and Texas have a higher rate of churn.

To help the company reduce customer Churn rate, we recommend following methods:

* First of all, enhance client satisfaction through action.
* Secondly, use a competitive pricing strategy.
* Last, conduct a thorough market analysis in the states with a higher churn rate.