Contents:

1. EXECUTIVE SUMMARY

Telecom industry is one of the industries which is cursed by cut throat competition and customers usually navigate through companies in search of better services, value for money, products and overall experience. Given the constant investment in equipment, modern technology and Research and development, telecom companies often try to poach customers subscribed to rival companies. Hence, it is important for companies to protect their customers by getting feedback and understanding the needs and problems of the customers. This feedback can be used along with other customer behaviors to predict whether a customer is likely to leave (churn) the company or not. This project serves the problem by comparing machine learning models trained on consumer data. The dataset leveraged for the scope of this project is from IBM. The features include data on consumer subscriptions, charges, tenure with the company, demographic information and products used. The target variable is the label for each customer as churned or not churned. The dataset is randomly divided into training, validation and testing datasets in the ratio- 70:15:15 respectively. Different machine learning models are trained and evaluated to compare key statistics using a confusion matrix. The scope of the project can be expanded to include more features and data points for use in industrial applications.

1.1 PROJECT MOTIVATION

To effectively retain consumers, telecommunications businesses must be able to predict customer attrition. Getting new clients is more expensive than keeping your current ones. Large telecoms companies are attempting to create models to identify which consumers are more likely to change and take appropriate action as a result.
In this article, we develop a model to forecast a customer's likelihood of leaving by examining its characteristics: (1) demographic data, (2) account data, and (3) services data. Our goal is to find a data-driven solution that would enable us to lower churn rates, which will boost customer happiness and business income.

2. DATA SOURCE

The data is an open-sourced dataset first published in 2020 on the IBM platform. The dataset consists of 7043 data points/rows and 33 features/columns. Dimension reduction techniques are implemented which is explained in detail in this report.

LINK TO THE DATASET

3. DATA DESCRIPTION

Demographics

CustomerID: A unique ID that identifies each customer.

Count: A value used in reporting/dashboarding to sum up the number of customers in a filtered set.

Gender: The customer's gender: Male, Female

Age: The customer's current age, in years, at the time the fiscal quarter ended.

Senior Citizen: Indicates if the customer is 65 or older: Yes, No

Married: Indicates if the customer is married: Yes, No

Dependents: Indicates if the customer lives with any dependents: Yes, No. Dependents could be children, parents, grandparents, etc.

Number of Dependents: Indicates the number of dependents that live with the customer.

Location

CustomerID: A unique ID that identifies each customer.

Count: A value used in reporting/dashboarding to sum up the number of customers in a filtered set.

Country: The country of the customer's primary residence.

State: The state of the customer's primary residence.

City: The city of the customer's primary residence.

Zip Code: The zip code of the customer's primary residence.

Lat Long: The combined latitude and longitude of the customer's primary residence.

Latitude: The latitude of the customer's primary residence.

Longitude: The longitude of the customer's primary residence.

Population

ID: A unique ID that identifies each row.

Zip Code: The zip code of the customer's primary residence.

Population: A current population estimate for the entire Zip Code area.

Services

CustomerID: A unique ID that identifies each customer.

Count: A value used in reporting/dashboarding to sum up the number of customers in a filtered set.

Quarter: The fiscal quarter that the data has been derived from .

Referred a Friend: Indicates if the customer has ever referred a friend or family member to this company: Yes, No

Number of Referrals: Indicates the number of referrals to date that the customer has made.

Tenure in Months: Indicates the total amount of months that the customer has been with the company by the end of the quarter specified above.

Offer: Identifies the last marketing offer that the customer accepted, if applicable. Values include None, Offer A, Offer B, Offer C, Offer D, and Offer E.

Phone Service: Indicates if the customer subscribes to home phone service with the company: Yes, No

Avg Monthly Long Distance Charges: Indicates the customer's average long distance charges, calculated to the end of the quarter specified above.

Multiple Lines: Indicates if the customer subscribes to multiple telephone lines with the company: Yes, No

Internet Service: Indicates if the customer subscribes to Internet service with the company: No, DSL, Fiber Optic, Cable.

Avg Monthly GB Download: Indicates the customer's average download volume in gigabytes, calculated to the end of the quarter specified above.

Online Security: Indicates if the customer subscribes to an additional online security service provided by the company: Yes, No

Online Backup: Indicates if the customer subscribes to an additional online backup service provided by the company: Yes, No

Device Protection Plan: Indicates if the customer subscribes to an additional device protection plan for their Internet equipment provided by the company: Yes, No

Premium Tech Support: Indicates if the customer subscribes to an additional technical support plan from the company with reduced wait times: Yes, No

Streaming TV: Indicates if the customer uses their Internet service to stream television programing from a third party provider: Yes, No. The company does not charge an additional fee for this service.

Streaming Movies: Indicates if the customer uses their Internet service to stream movies from a third party provider: Yes, No. The company does not charge an additional fee for this service.

Streaming Music: Indicates if the customer uses their Internet service to stream music from a third party provider: Yes, No. The company does not charge an additional fee for this service.

Unlimited Data: Indicates if the customer has paid an additional monthly fee to have unlimited data downloads/uploads: Yes, No

Contract: Indicates the customer's current contract type: Month-to-Month, One Year, Two Year.

Paperless Billing: Indicates if the customer has chosen paperless billing: Yes, No

Payment Method: Indicates how the customer pays their bill: Bank Withdrawal, Credit Card, Mailed Check

Monthly Charge: Indicates the customer's current total monthly charge for all their services from the company.

Total Charges: Indicates the customer's total charges, calculated to the end of the quarter specified above.

Total Refunds: Indicates the customer's total refunds, calculated to the end of the quarter specified above.

Total Extra Data Charges: Indicates the customer's total charges for extra data downloads above those specified in their plan, by the end of the quarter specified above.

Total Long Distance Charges: Indicates the customer's total charges for long distance above those specified in their plan, by the end of the quarter specified above.

Status

CustomerID: A unique ID that identifies each customer.

Count: A value used in reporting/dashboarding to sum up the number of customers in a filtered set.

Quarter: The fiscal quarter that the data has been derived from .

Satisfaction Score: A customer's overall satisfaction rating of the company from 1 (Very Unsatisfied) to 5 (Very Satisfied).

Satisfaction Score Label: Indicates the text version of the score (1-5) as a text string.

Customer Status: Indicates the status of the customer at the end of the quarter: Churned, Stayed, or Joined

Churn Label: Yes = the customer left the company this quarter. No = the customer remained with the company. Directly related to Churn Value.

Churn Value: 1 = the customer left the company this quarter. 0 = the customer remained with the company. Directly related to Churn Label.

Churn Score: A value from 0-100 that is calculated using the predictive tool IBM SPSS Modeler. The model incorporates multiple factors known to cause churn. The higher the score, the more likely the customer will churn.

Churn Score Category: A calculation that assigns a Churn Score to one of the following categories: 0-10, 11-20, 21-30, 31-40, 41-50, 51-60, 61-70, 71-80, 81-90, and 91-100

CLTV: Customer Lifetime Value. A predicted CLTV is calculated using corporate formulas and existing data. The higher the value, the more valuable the customer. High value customers should be monitored for churn.

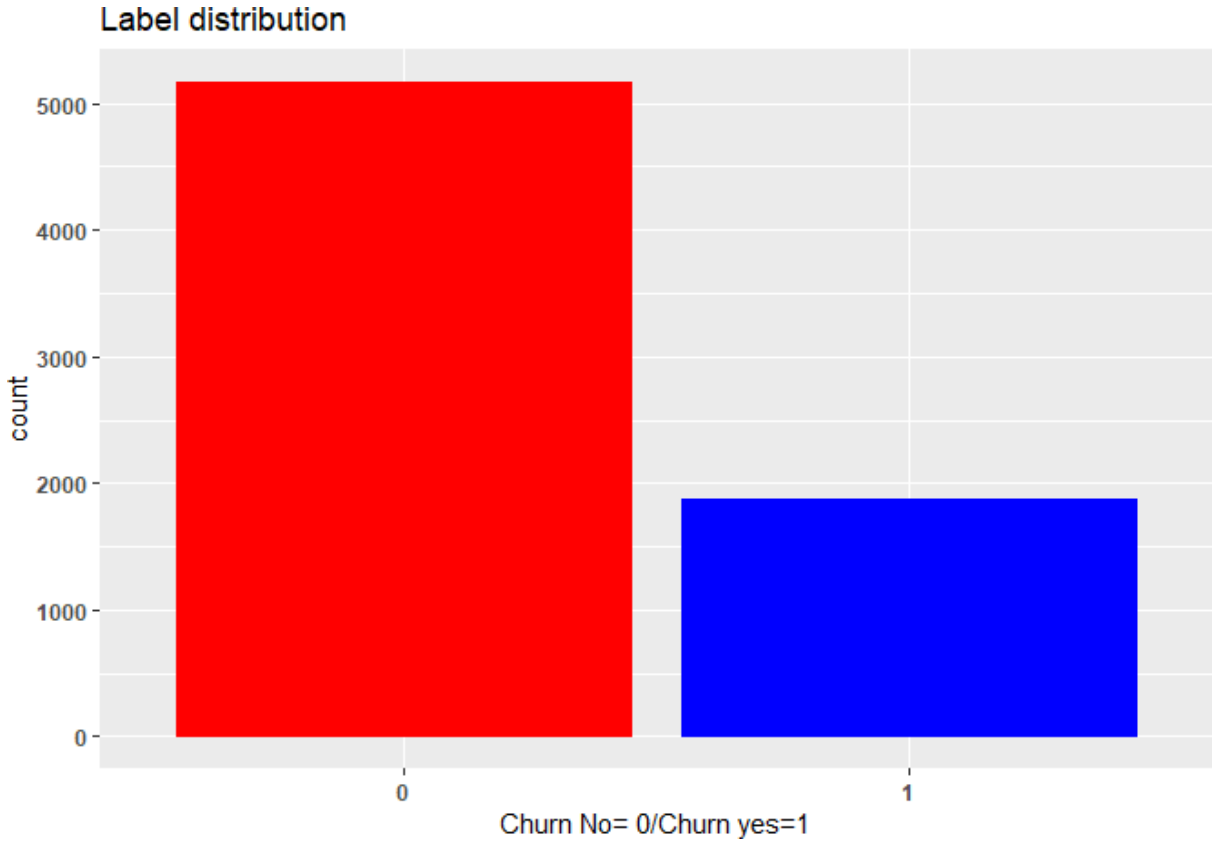
CLTV Category: A calculation that assigns a CLTV value to one of the following categories: 2000-2500, 2501-3000, 3001-3500, 3501-4000, 4001-4500, 4501-5000, 5001-5500, 5501-6000, 6001-6500, and 6501-7000.

Churn Category: A high-level category for the customer's reason for churning: Attitude, Competitor, Dissatisfaction, Other, Price. When they leave the company, all customers are asked about their reasons for leaving. Directly related to Churn Reason.

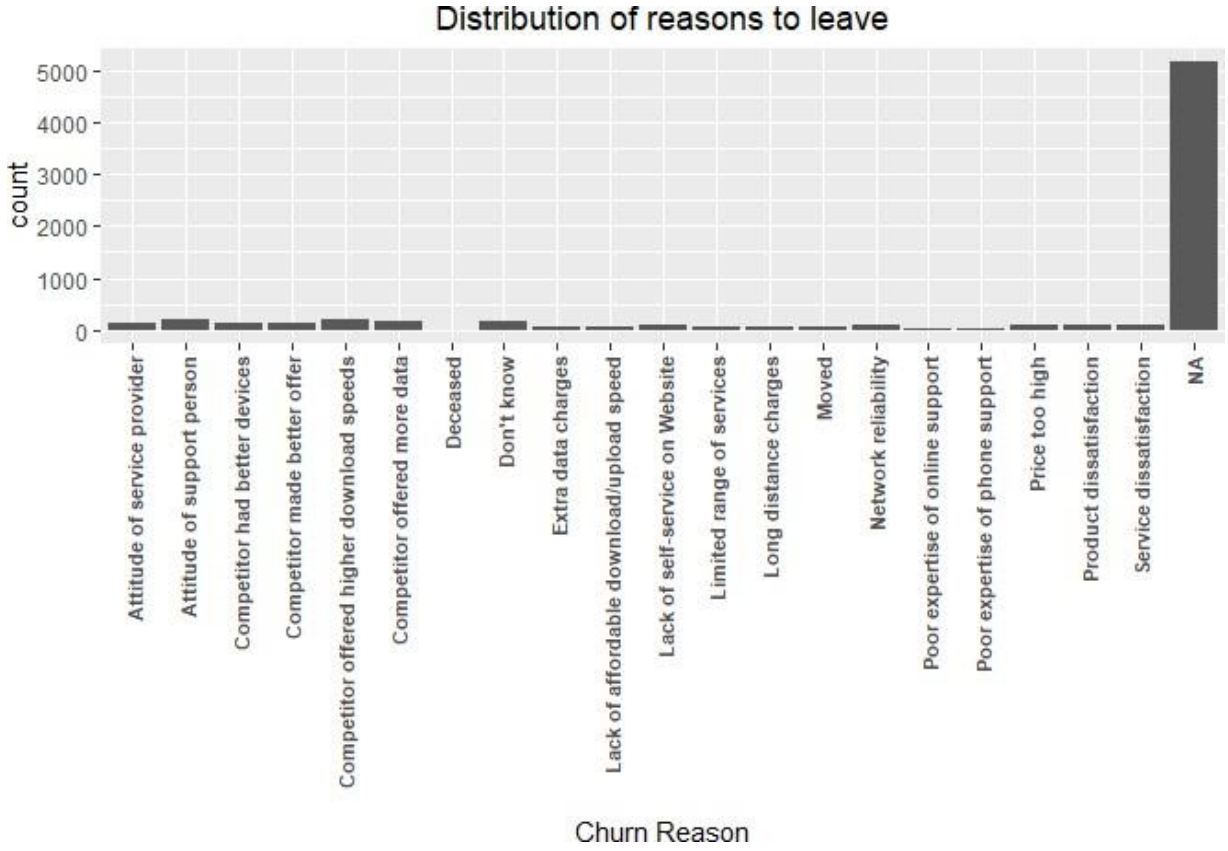
Churn Reason: A customer's specific reason for leaving the company. Directly related to Churn Category.

EXPLORATORY DATA ANALYSIS

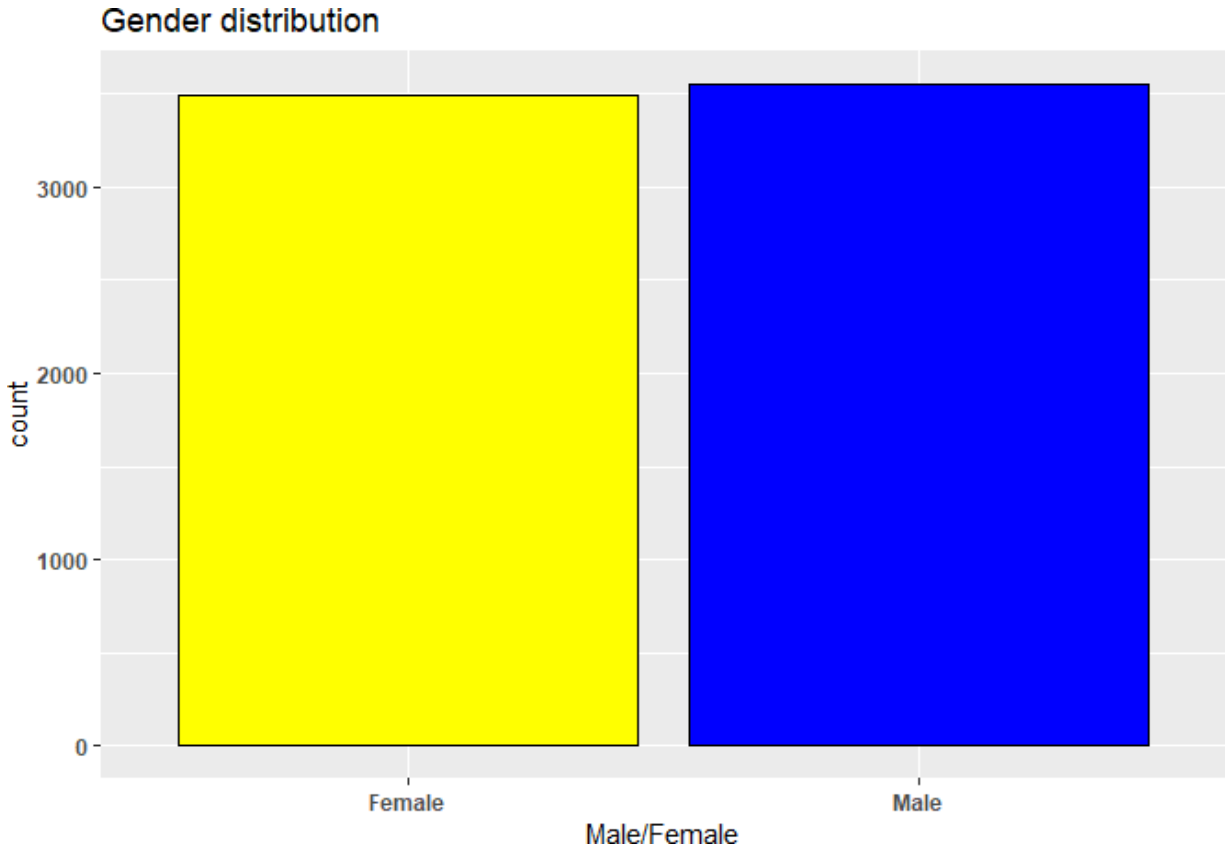
Label Distribution:



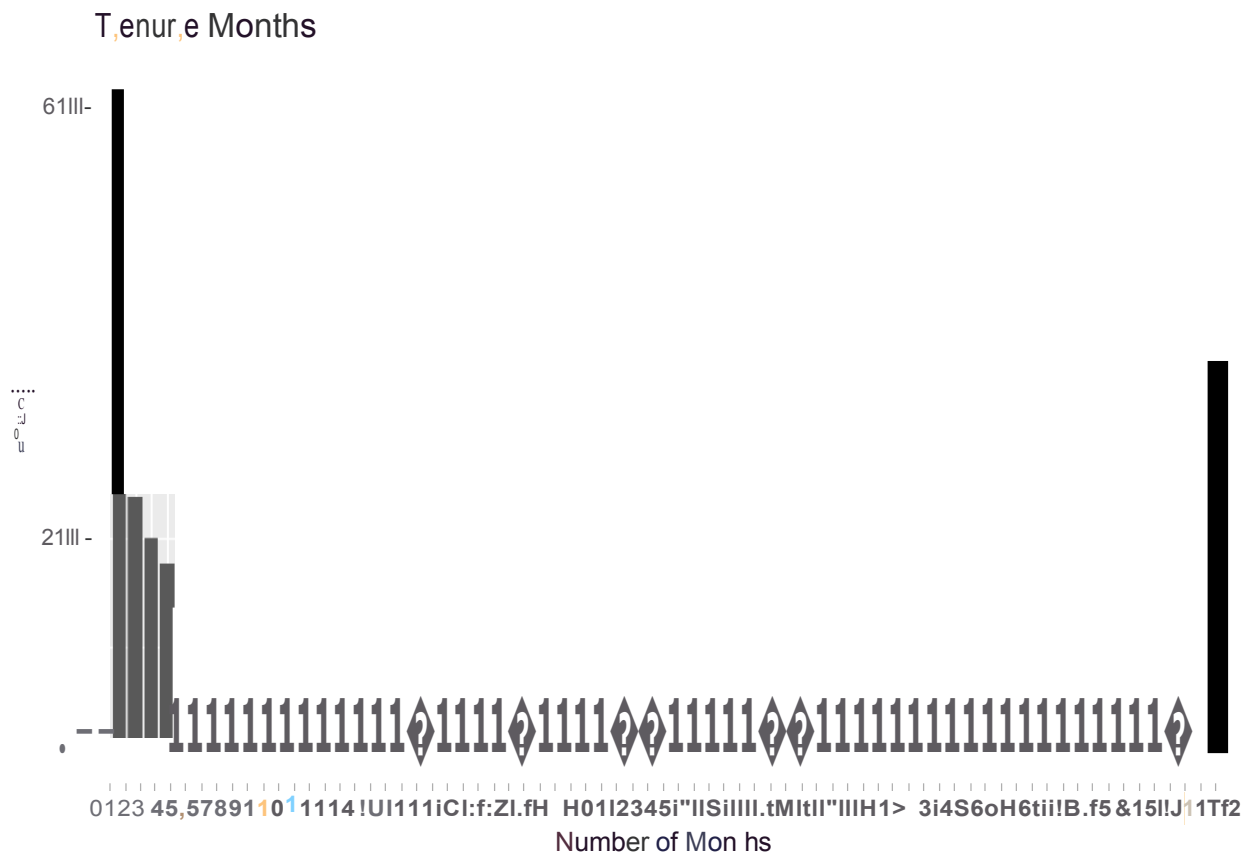
Distribution Of Reason To Leave:



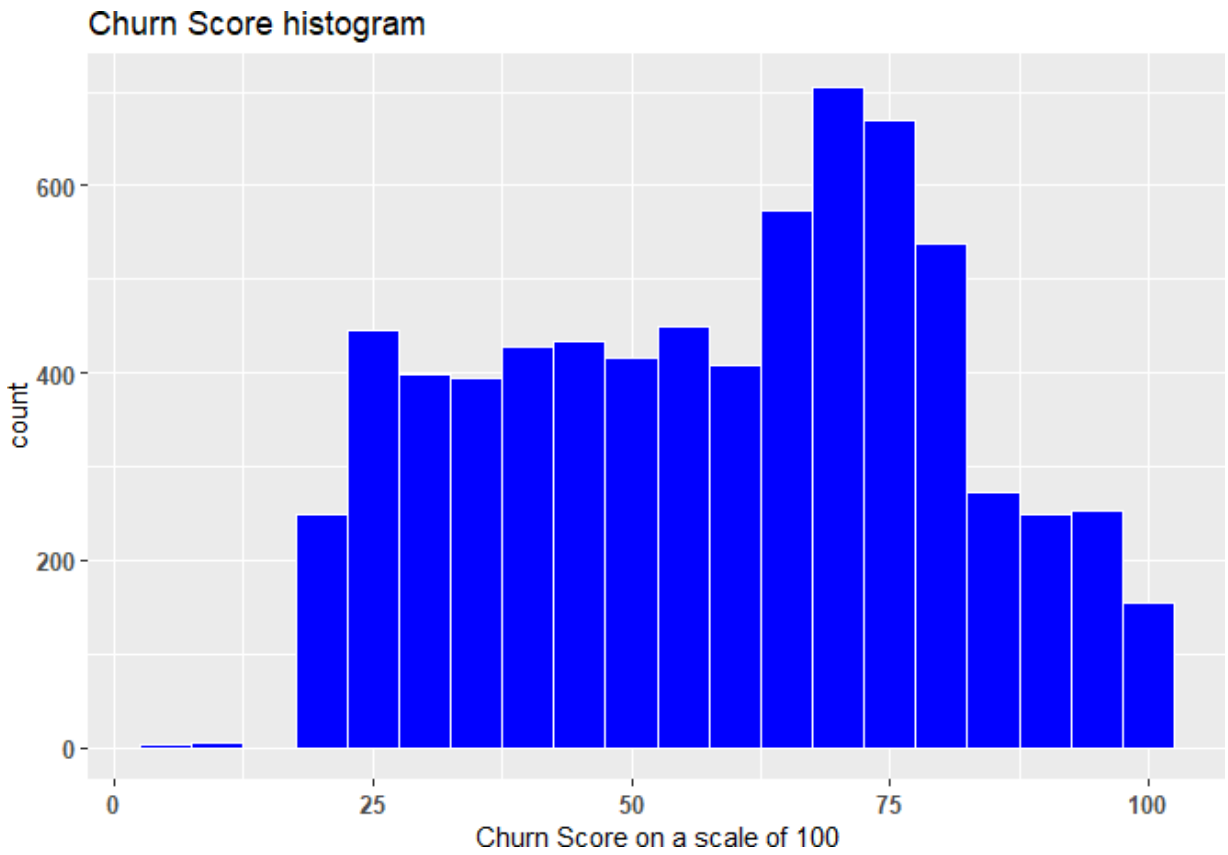
Gender Distribution:



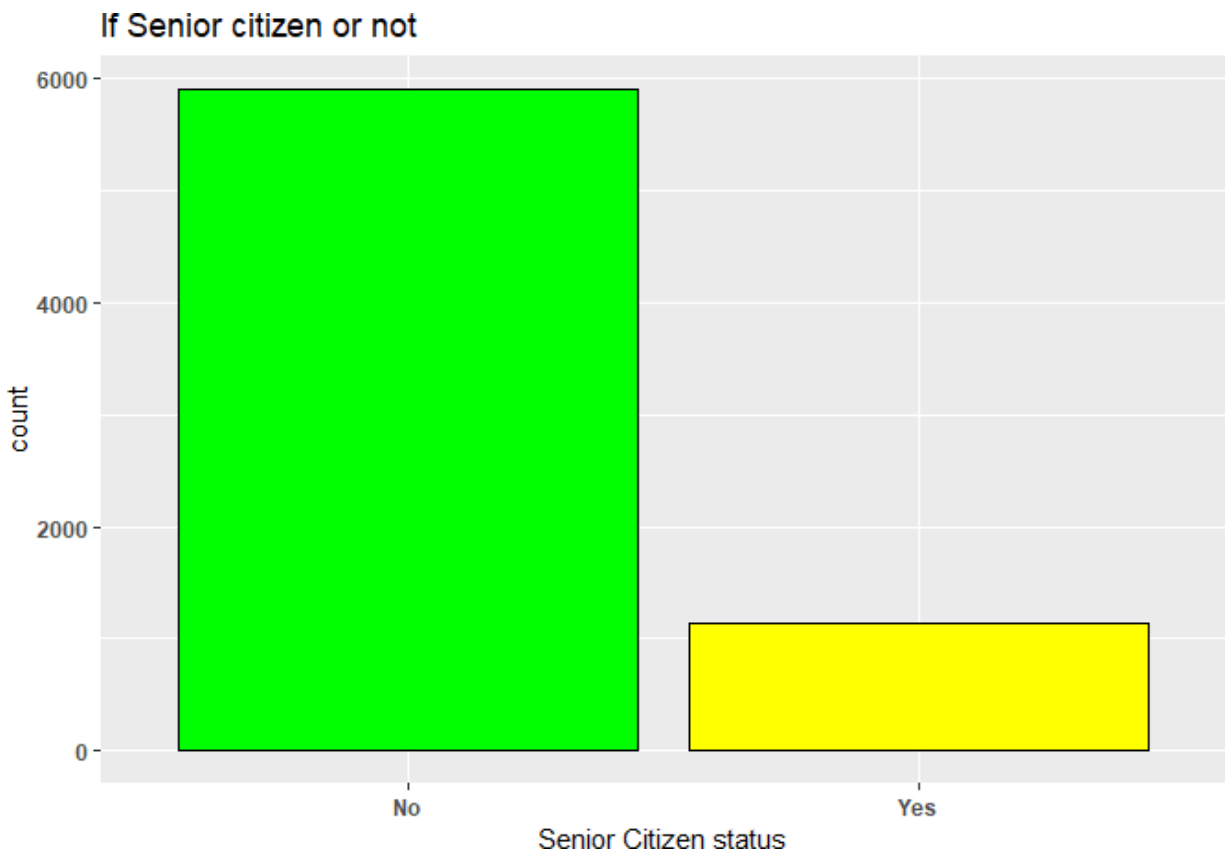
Tenure Months:


T,enur,e Months
61lll-
21lll -
0123 45,57891101 1114 !Ul111iCI:f:ZI.fH H01I2345i"llSiIlll.tMItll"lllH1> 3i4S6oH6tii!B.f5 &15l!J11Tf2
Number of Mon hs

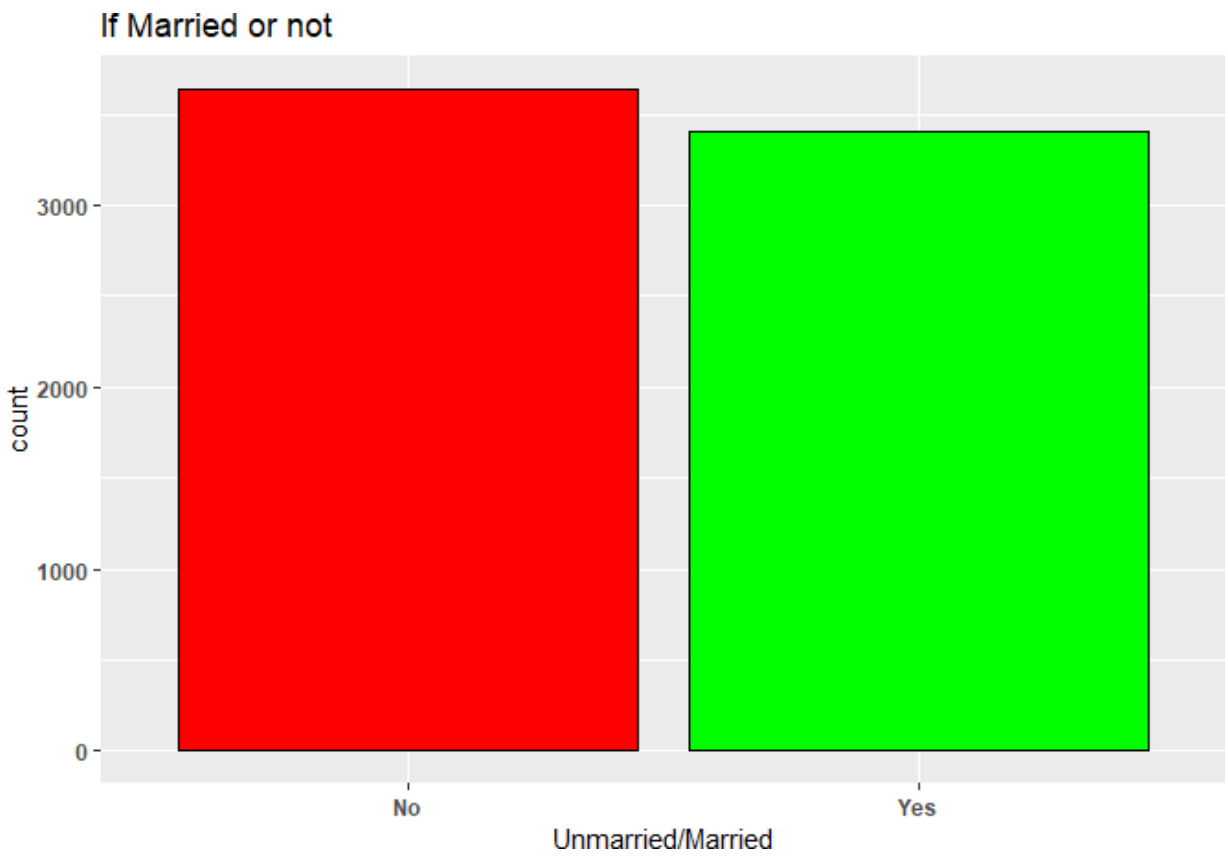
Churn Score Histogram:



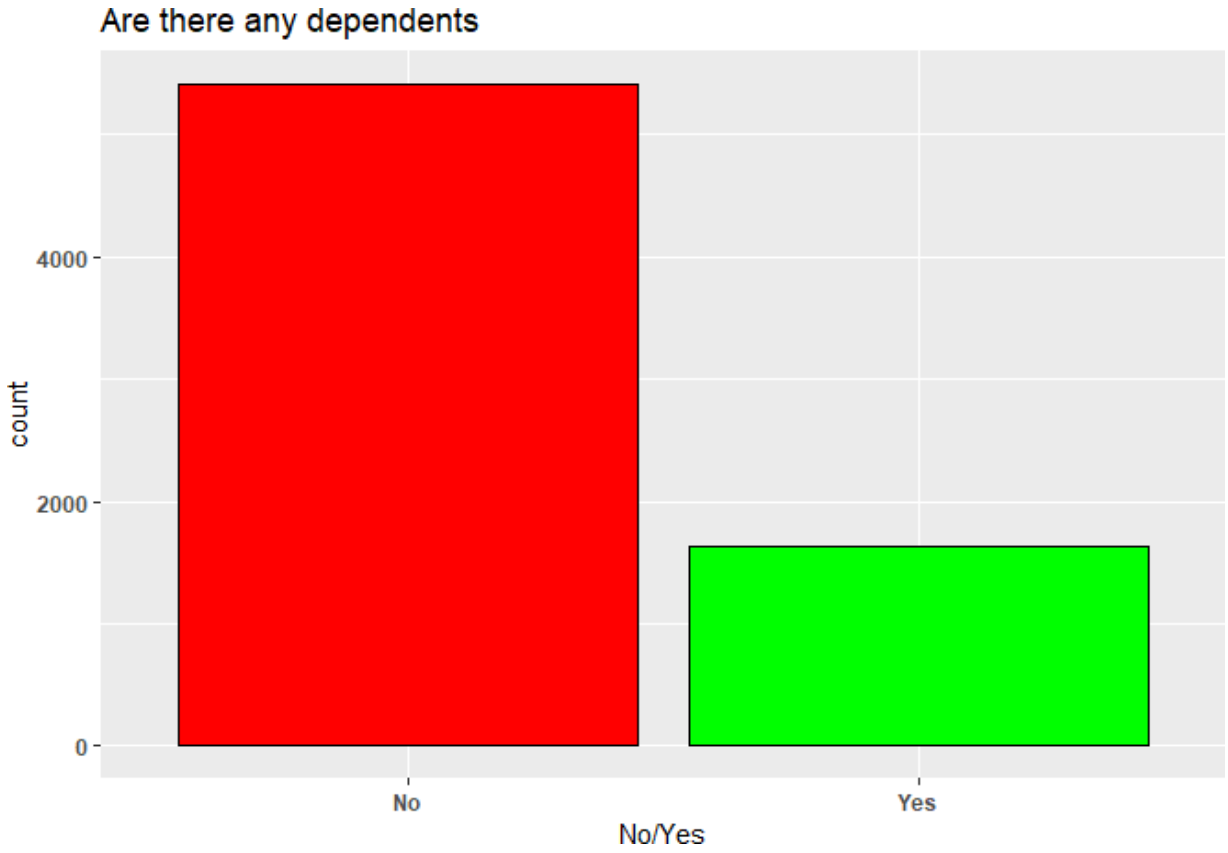
If Senior Citizen or Not:



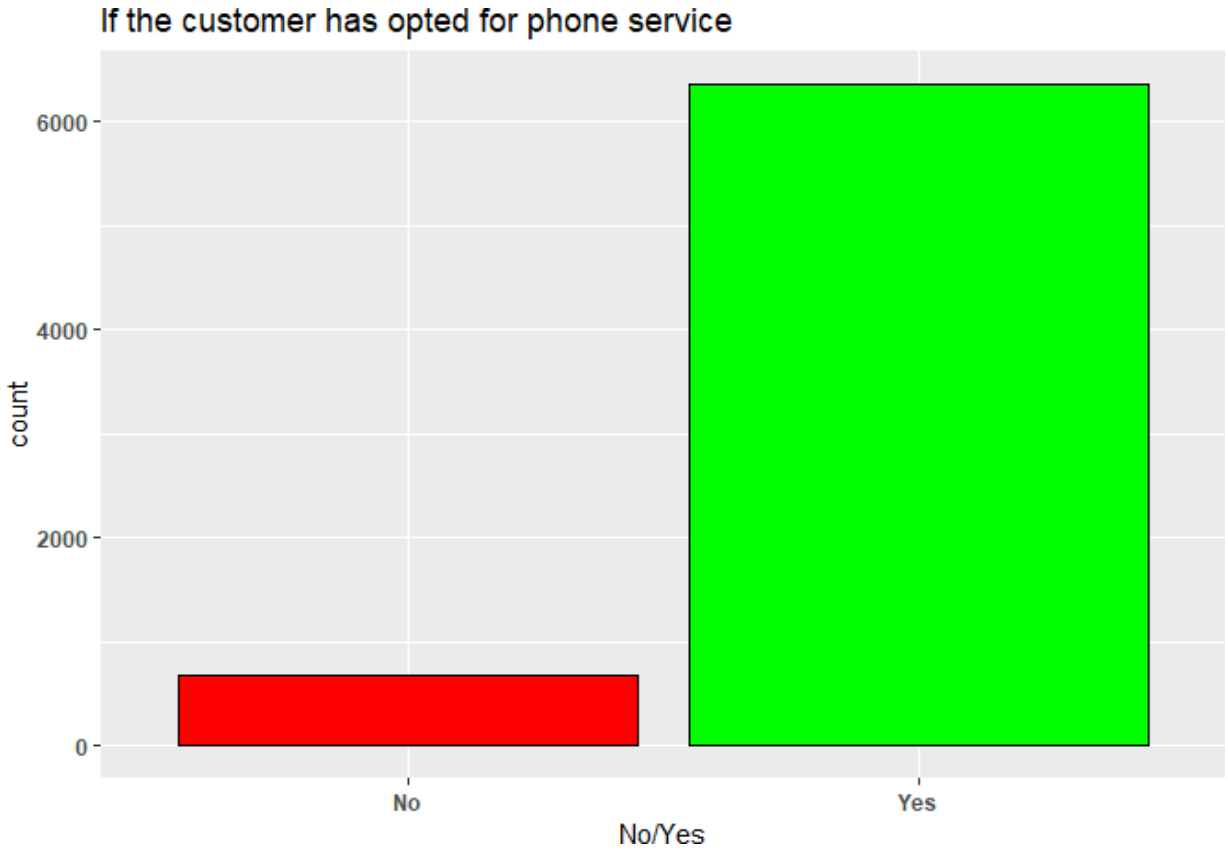
If Married Or Not:



Are There Any Dependents:

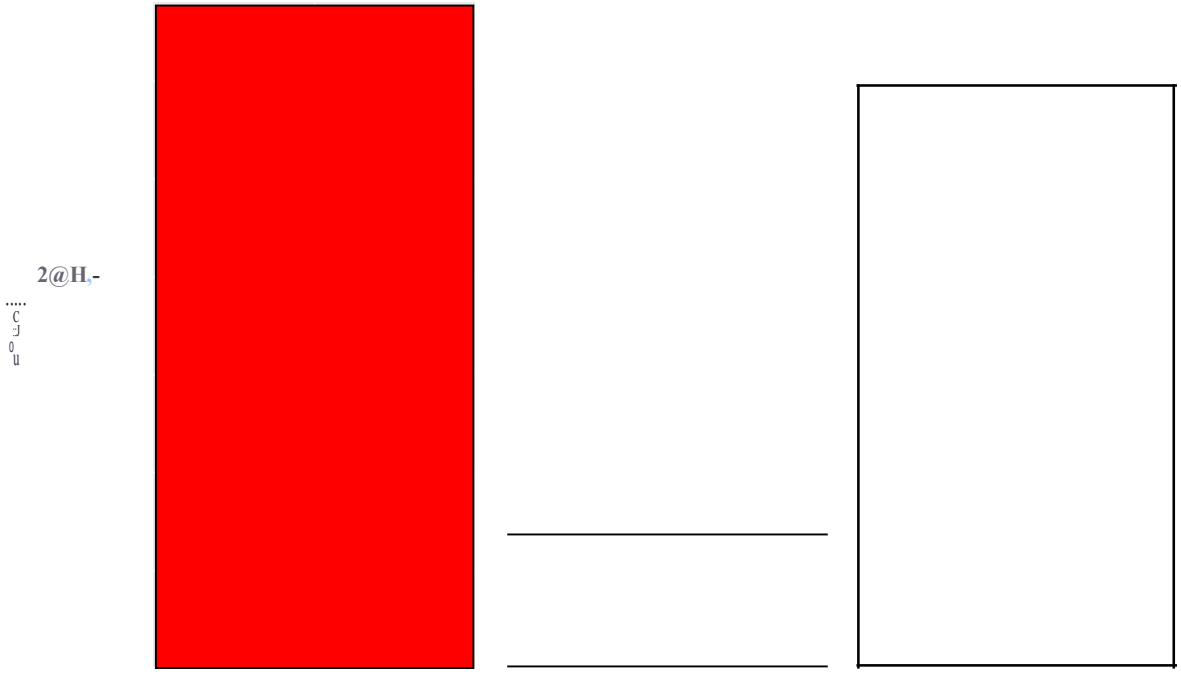


If the customer has opted for phone service:

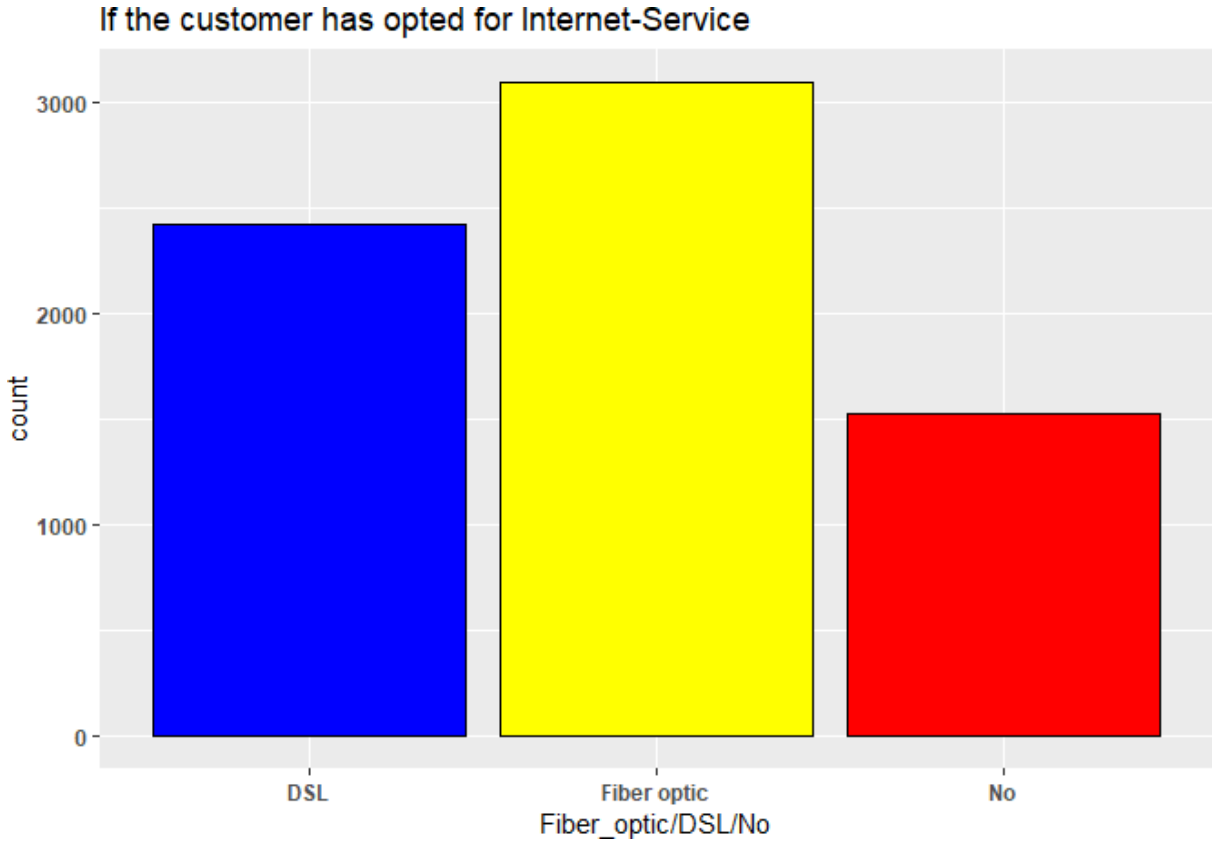


If the customer has opted for multiple-lines:

Im the customer has opted for multiple-lines

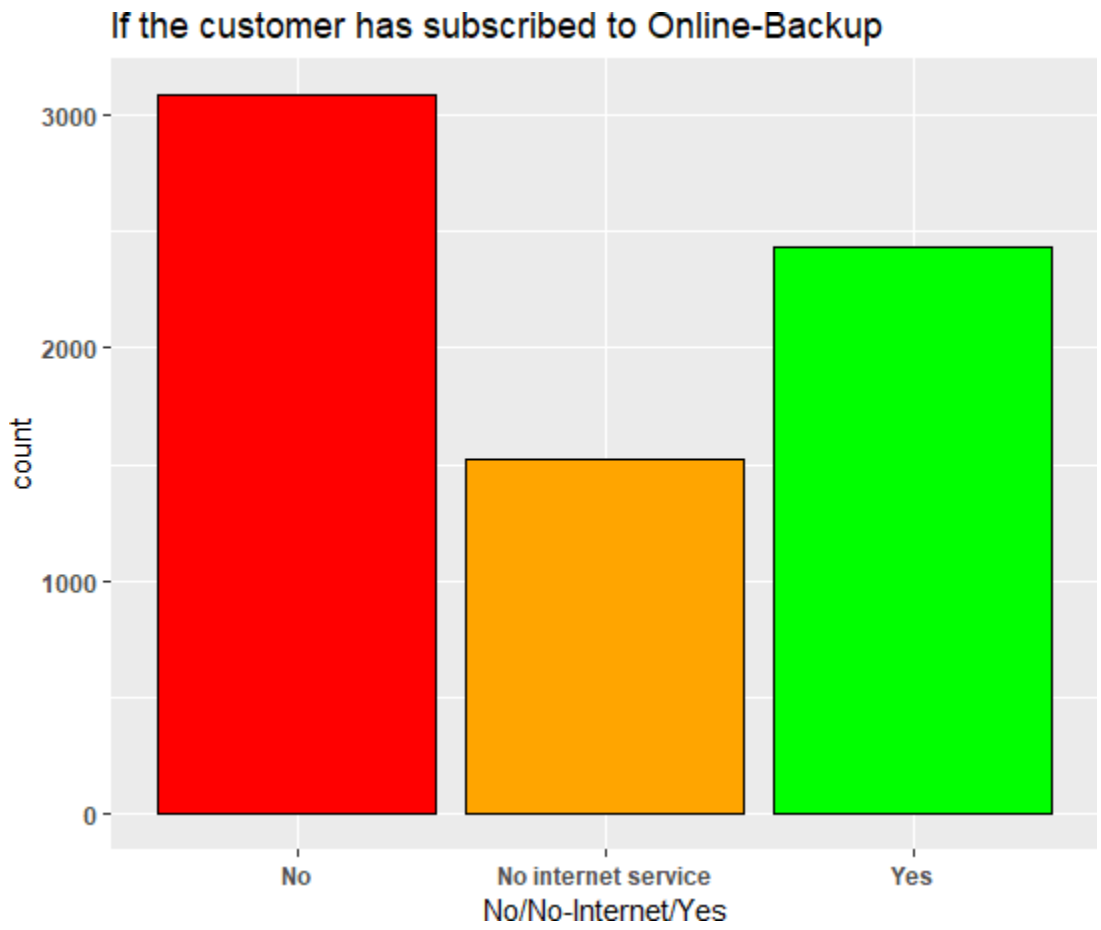


If the customer has opted for Internet-Service:

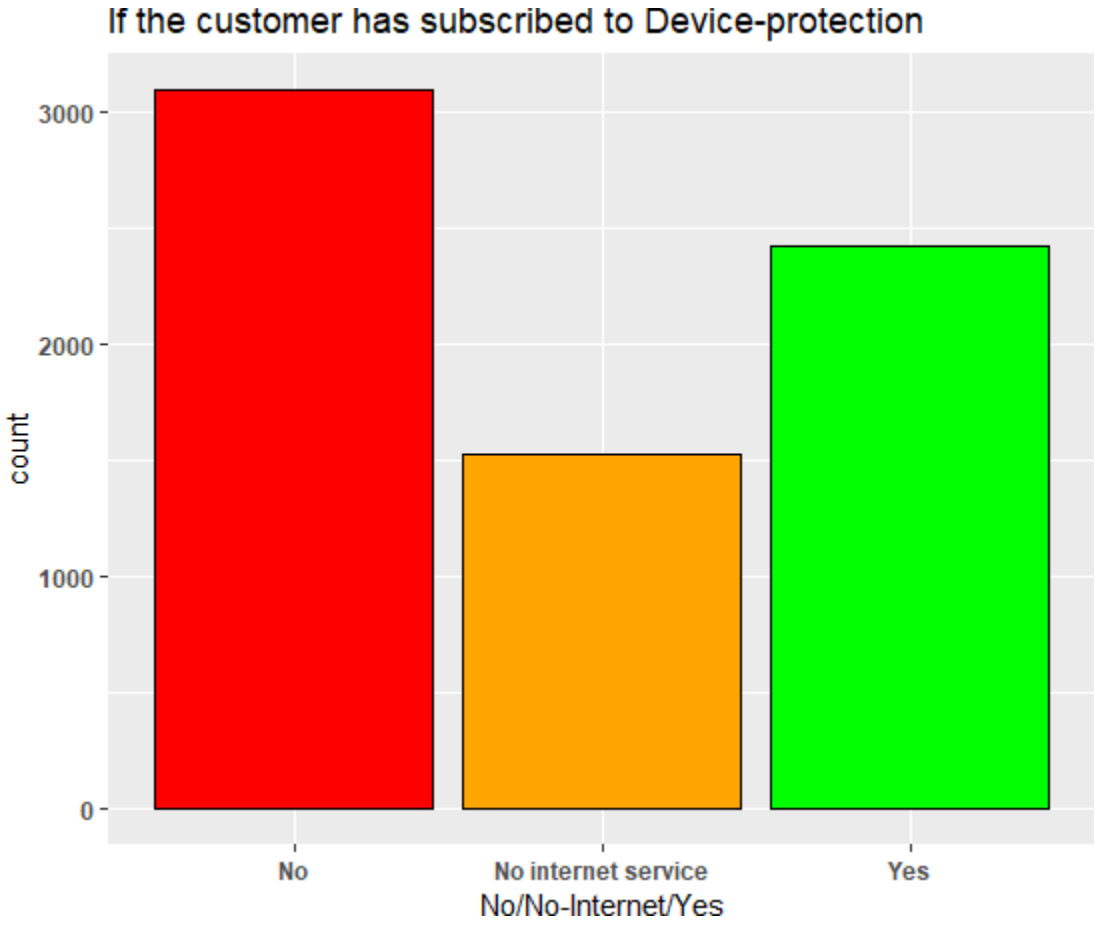


If the customer has subscribed to Online Security:

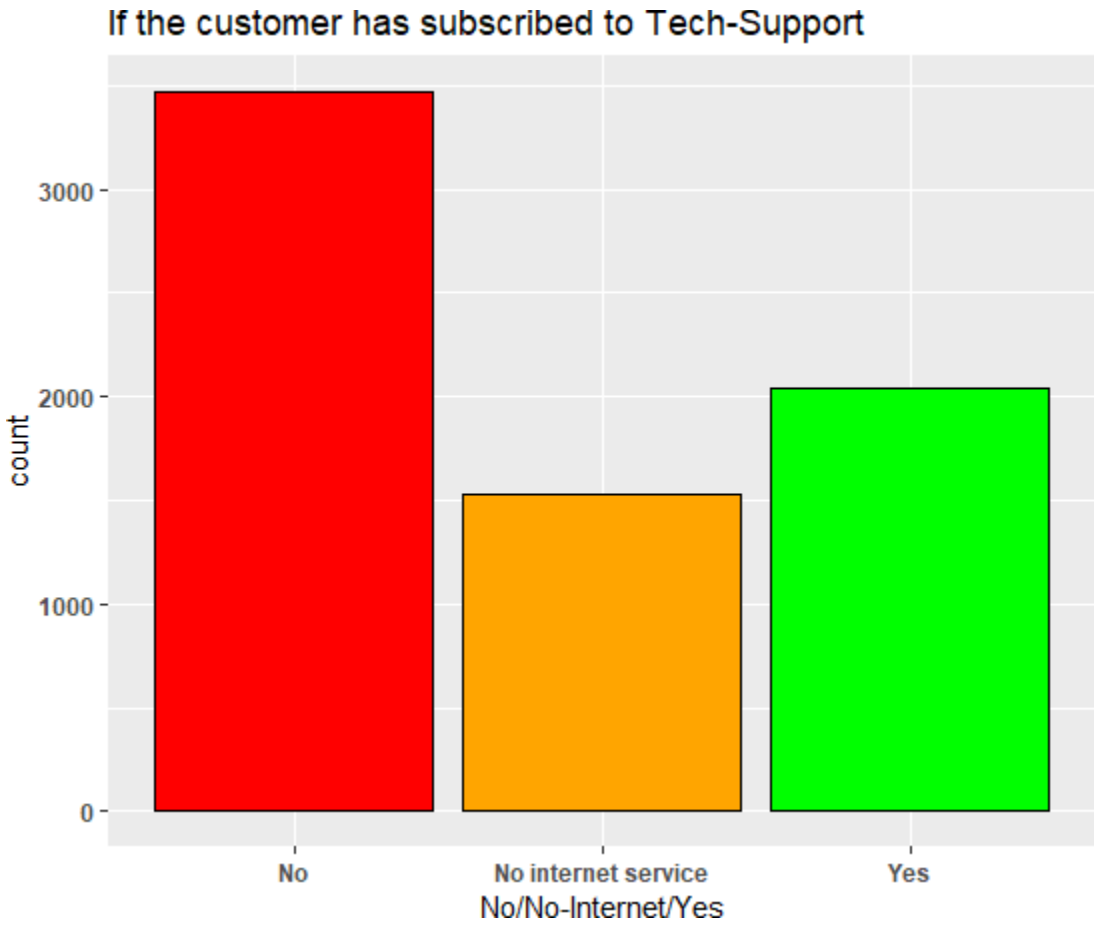
If the customer has subscribed to Online-Backup:



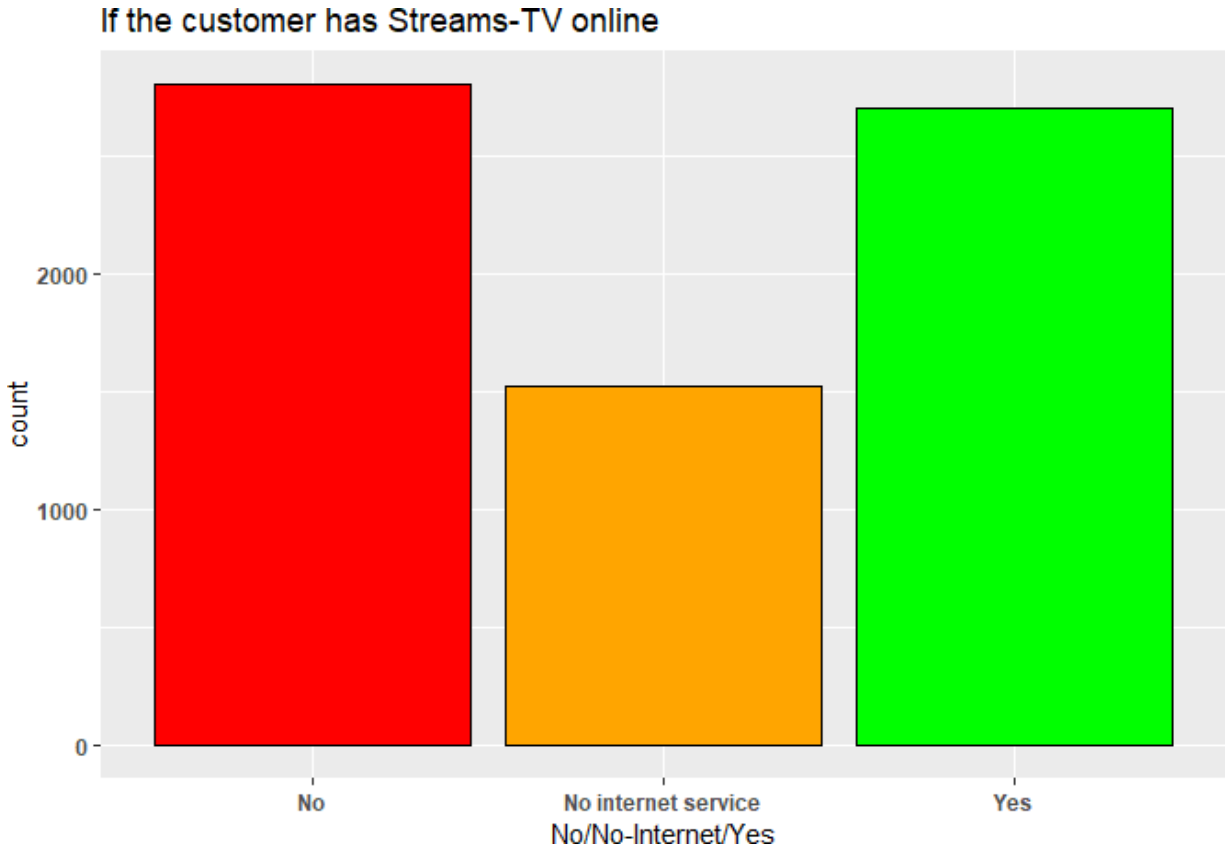
If the customer has subscribed to Device-protection:



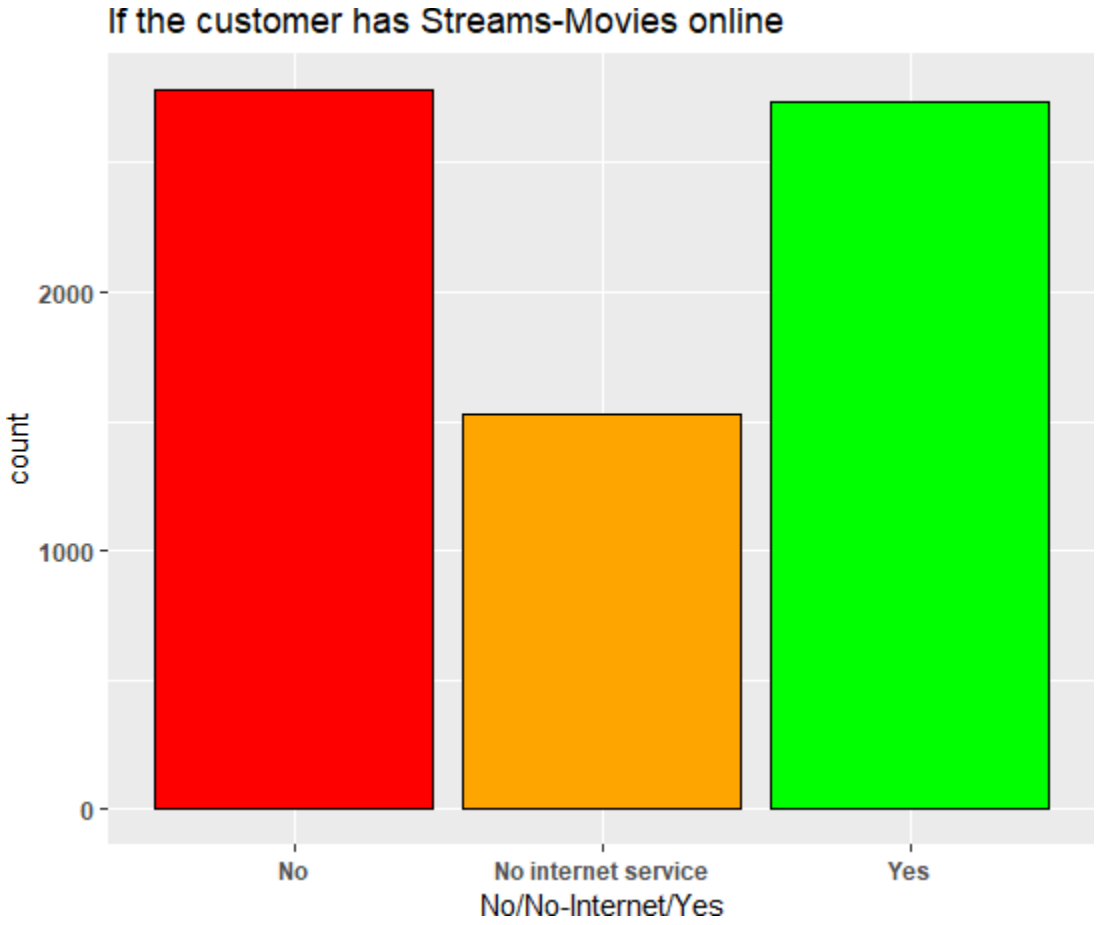
If the customer has subscribed to Tech-Support:



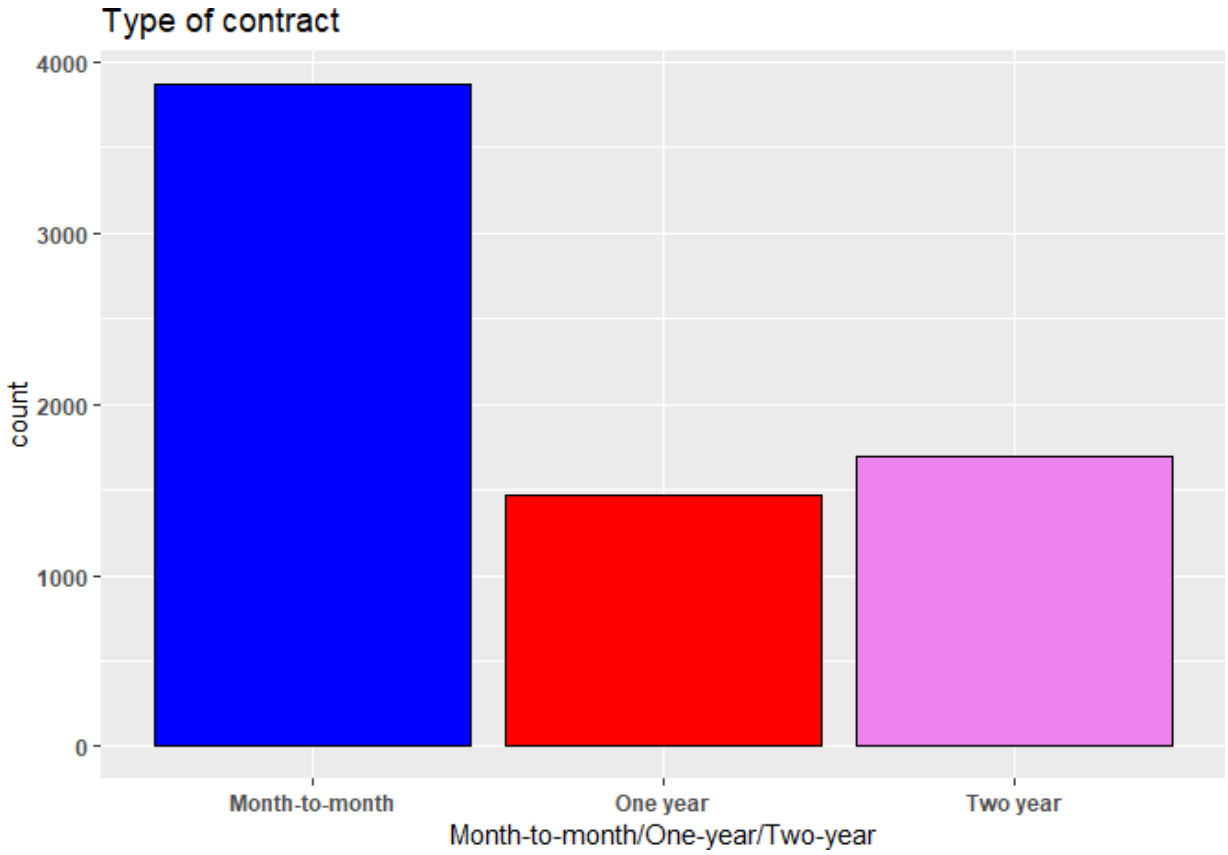
If the customer has Streams-TV online:



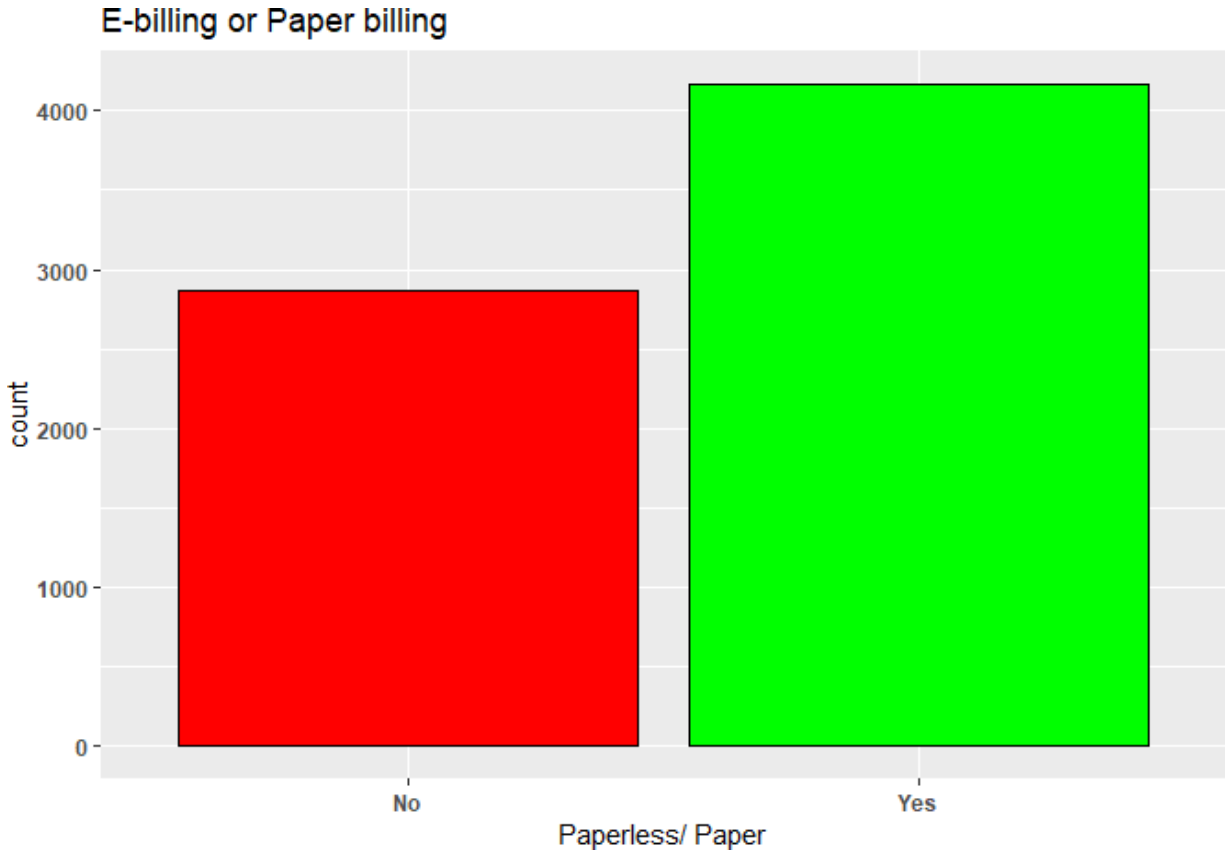
If the customer has Streams-Movies online:



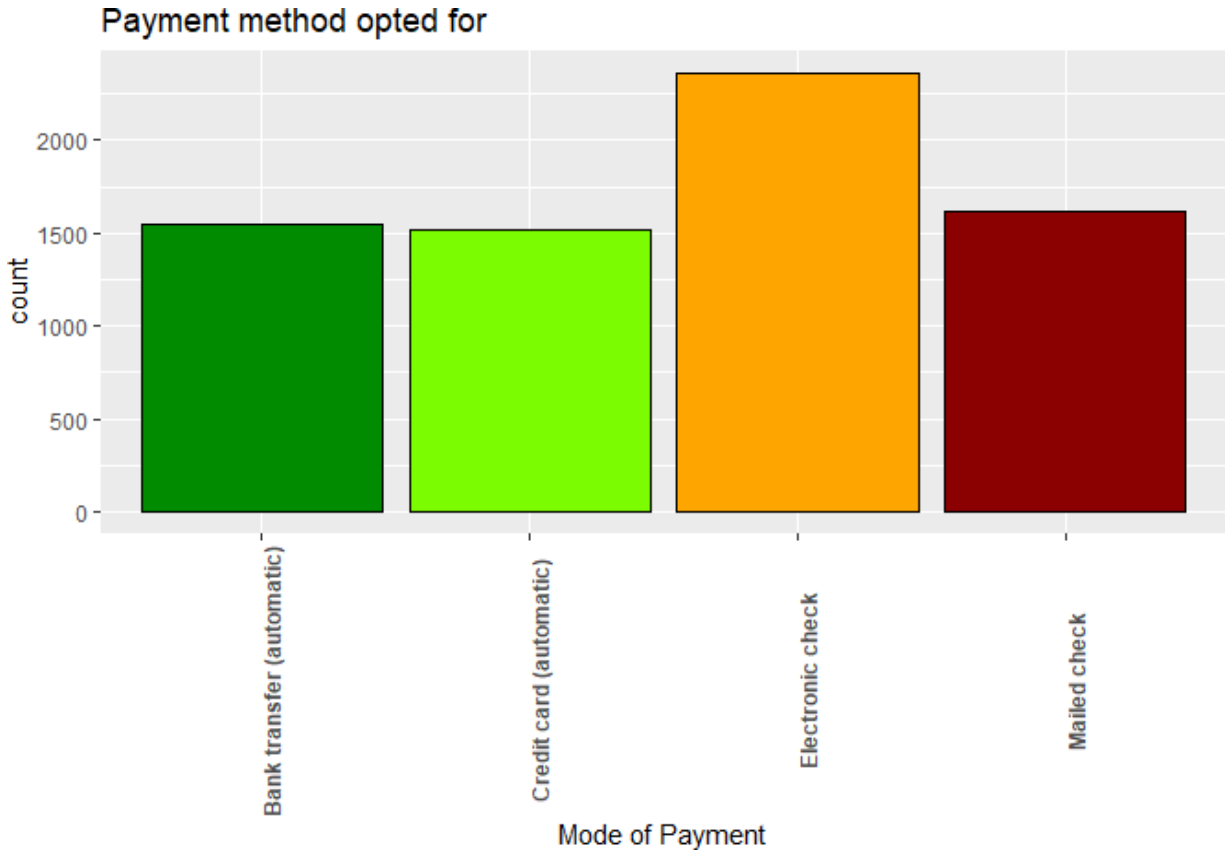
Type of contract:



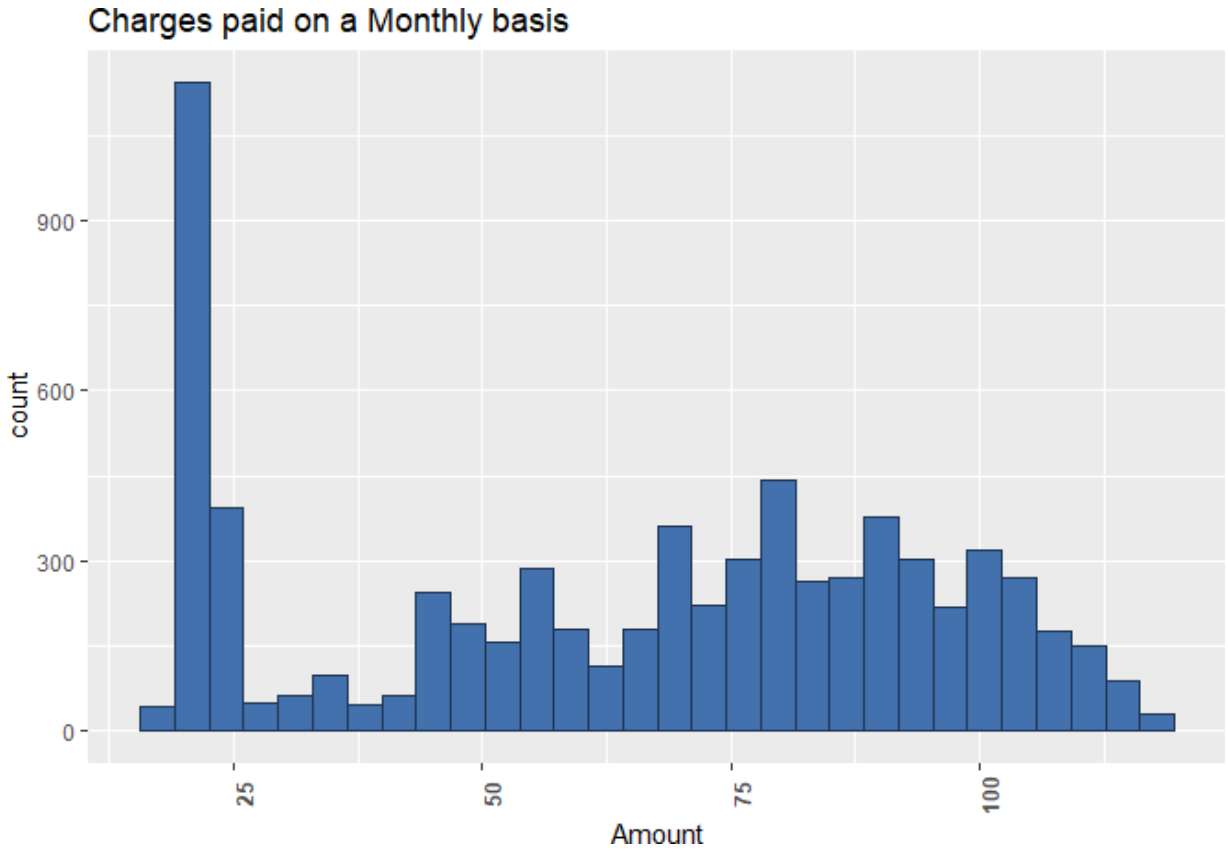
E-billing or Paper billing:



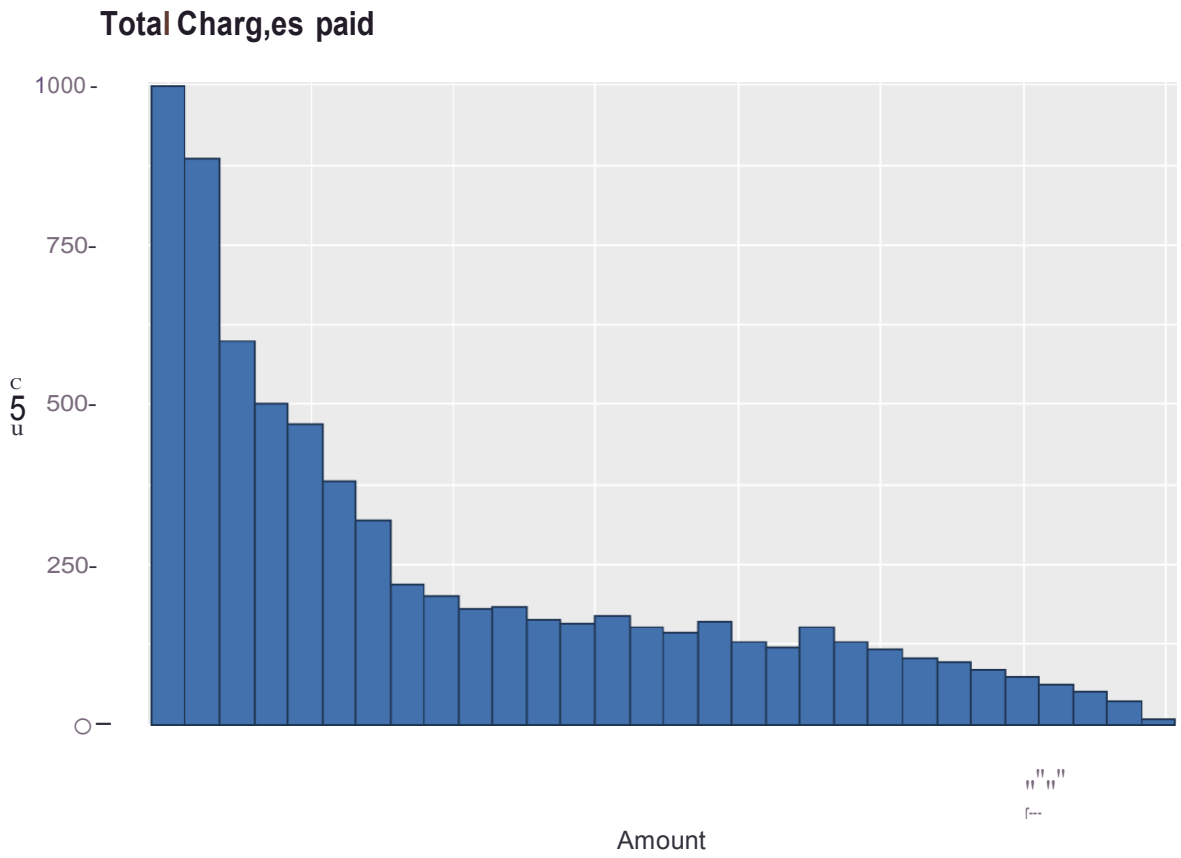
Payment method opted for:



Charges paid on a Monthly basis:



Total Charges paid:


Total Charg,es paid
1000
750
500
250
0
Amount

5. DIMENSION REDUCTION/FEATURE ENGINEERING

Dimension reduction or feature engineering is a set of techniques practiced to reduce the number of features without losing important information. Furthermore, categorical features are encoded using one-hot encoding technique while reducing the number of categories.

5.1 ONE-HOT ENCODING

1. Churn Reason feature is concatenated to less number of categories- 5 categories. The categories chosen are:

Reason	Category
Attitude of support person	Customer service
Attitude of service provider	Customer service
Service dissatisfaction	Customer service
Lack of self-service on Website	Customer service
Poor expertise of phone support	Customer service
Poor expertise of online support	Customer service
Competitor offered higher download speeds	Competitor
Competitor offered more data	Competitor
Competitor had better devices	Competitor
Competitor made better offer	Competitor
Network reliability	Product
Product dissatisfaction	Product
Limited range of services	Product
Price too high	Price

Extra data charges	Price
Long distance charges	Price
Lack of affordable download/upload speed	Price
Moved	Other
Deceased	Other

2. Gender variable is converted to numerical using one-hot encoding where Male is assigned 0 and Female is assigned 1.
3. Senior.Citizen variable is converted to numerical using one-hot encoding where No is assigned 0 and Yes is assigned 1.
4. Partner variable is converted to numerical using one-hot encoding where No is assigned 0 and Yes is assigned 1.
5. Dependent variable is converted to numerical using one-hot encoding where No is assigned 0 and Yes is assigned 1.
6. Phone.Service variable is converted to numerical using one-hot encoding where No is assigned 0 and Yes is assigned 1.
7. Multiple.Lines variable is converted to numerical using one-hot encoding where No is assigned 0 and Yes is assigned 1.
8. Online.Security variable is converted to numerical using one-hot encoding where No is assigned 0 and Yes is assigned 1.
9. Online.Backup variable is converted to numerical using one-hot encoding where No is assigned 0 and Yes is assigned 1.
10. Device.Protection variable is converted to numerical using one-hot encoding where No is assigned 0 and Yes is assigned 1.
11. Tech.Support variable is converted to numerical using one-hot encoding where No is assigned 0 and Yes is assigned 1.
12. Streaming.TV variable is converted to numerical using one-hot encoding where No is assigned 0 and Yes is assigned 1.
13. Streaming.Movies variable is converted to numerical using one-hot encoding where No is assigned 0 and Yes is assigned 1.
14. Paperless.Billing variable is converted to numerical using one-hot encoding where No is assigned 0 and Yes is assigned 1.

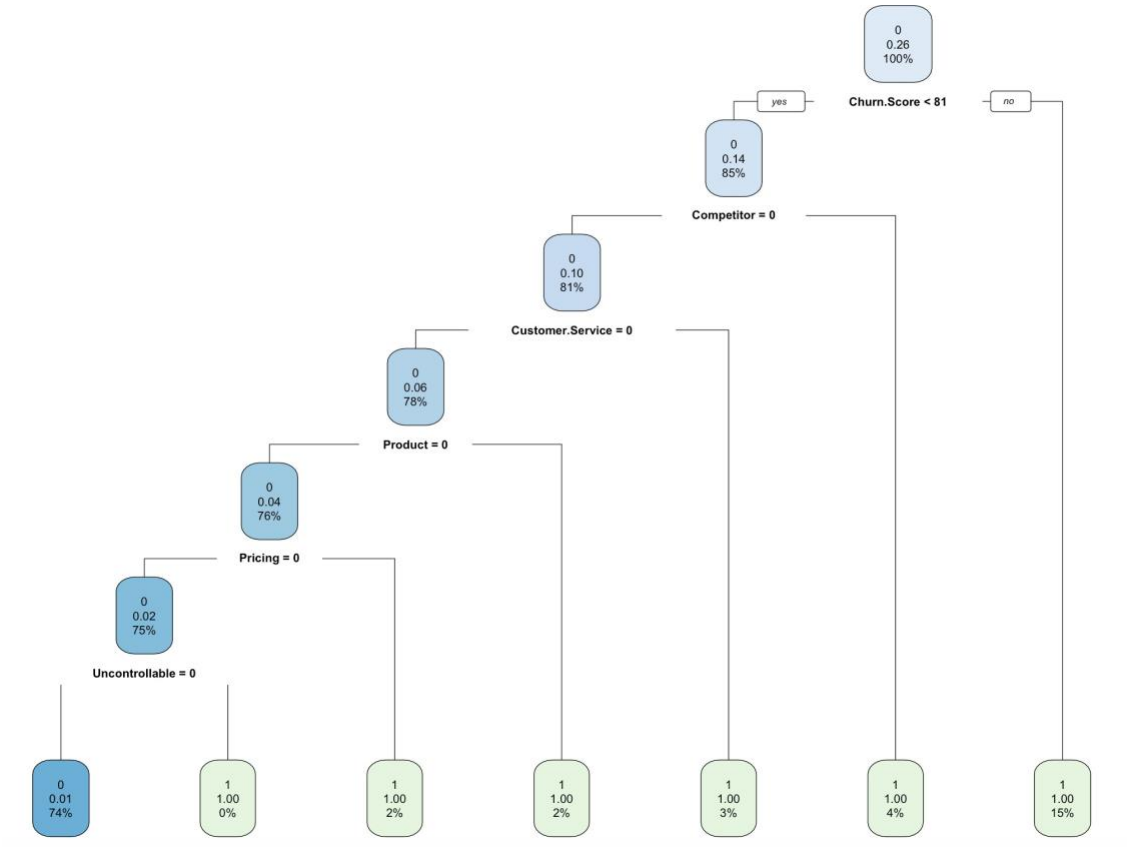
15. Churn.Label variable is converted to numerical using one-hot encoding where No is assigned 0 and Yes is assigned 1.
16. Internet.Service variable is converted to numerical using one-hot encoding where No is assigned 0, DSL assigned is assigned 1 and No DSL is assigned 2.
17. Contract variable is converted to numerical using one-hot encoding where ‘One Year’ is assigned 0 and ‘Two Year’ is assigned 1 and ‘Three Year’ is assigned 2.
18. Payment.method variable is converted to numerical using one-hot encoding where ‘Credit card (automatic) ' is assigned 0 and ‘Bank Transfer (automatic)’ is assigned 1, ‘Electronic check’ is assigned 2 and ‘Cash’ is assigned 3.
19. The zipcode has been replaced by

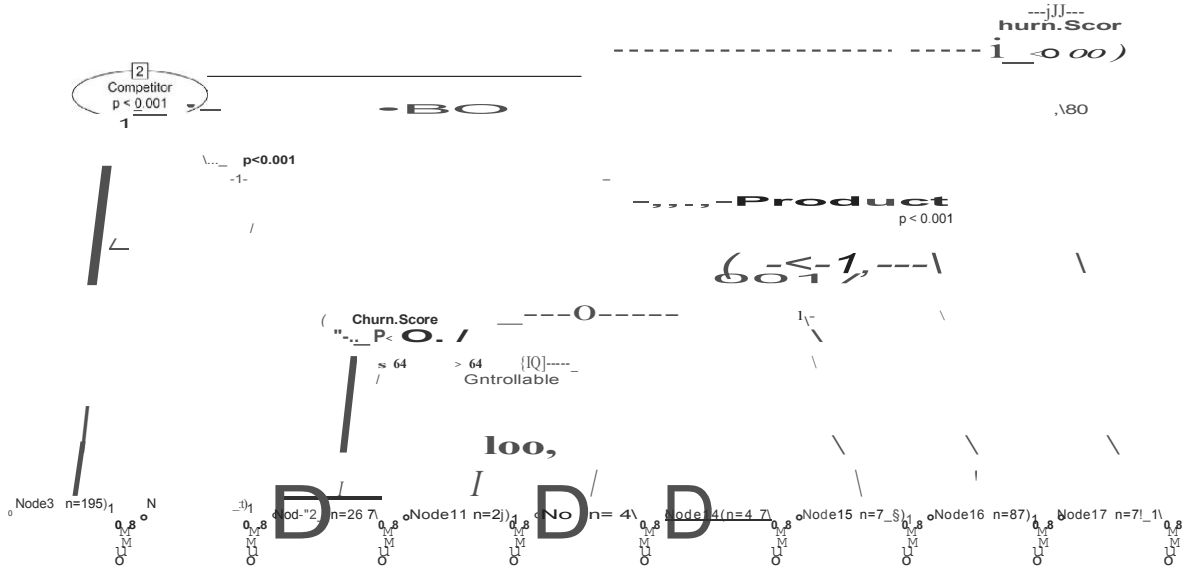
1. DATA MODELLING:

1.1. Model 1 – Decision Tree:

The Decision Tree is a powerful and widely used tool for classification and prediction. A flowchart is similar to a tree structure, with each internal node representing a test on an attribute, each branch representing a test outcome, and each leaf node (terminal node) holding a class label.

DEFAULT TREE:






```
> tree <- rpart(Churn. Labelenco - data trai.n, cp 0.01, maxdepth 22)
> rpart.plot(tree)
> tree <- rpart(Churn. Labelenco - data trai.n, cp 0.01, maxdepth 22)
> rpart.plot(tree)
>rpart.rules(tree)
 Churn. Labelenco
             0.01 when Churn. Score < 81 & Competi. tor i.s 0 & Customer. Servi.ce i.s 0 & Product i.s 0 & Pricing i.s 0 & Uncontrollable i.s 0
             1. 00 when Churn. Score < 81 & Competi. tor i.s 0 & Customer. Service i.s 0 & Product i.s 0 & Pri.ci.ng i.s 0 & Uncontrollable i.s 1
             1.00 when Churn. Score < 81 & Competi. tor i.s 0 & Customer. Service i.s 0 & Product i.s 0 & Pricing i.s 1
             1. 00 when Churn. Score < 81 & Competi. tor i.s 0 & Customer. Service i.s 0 & Product i.s 1
             1. 00 when Churn. Score < 81 & Competi. tor i.s 0 & Customer. Service i.s 1
             1. 00 when Churn. Score < 81 & Competi. tor i.s 1
             1. 00 when Churn. Score >= 81
> ctree <- ctree(Churn.Labelenco ~ . , train)
> plot(ctree_)
> pri.ntcp(tree)

Classi. ficati.on tree:
rpart(formula = Churn.Labelenco - . , data trai.n, cp 0.01,
    maxdepth = 22)

Vari.ables actually used in tree construction:
[1] Churn. Score Competi.tor Customer. Servi.ce Pri.ci.ng Product Uncontrollable

Root node error: 1296/4891 = 0.26498

n= 4891

         CP nspli.t rel error xerror xstd
1 0.548611 0 1. 000000 1. 000000 0. 0238149
2 0.150463 1 0.451389 0.451389 0.0175110
3 0.124228 2 0.300926 0.300926 0.0146178
4 0.067130 3 0.176698 0.176698 0.0113999
5 0.057870 4 0.109568 0.109568 0. 0090603
6 0.018519 5 0. 051698 0. 051698 0. 0062725
7 0.010000 6 0.033179 0.033179 0.0050375
> prp(tree, type= 1, extra= 1, under= FALSE, spli.t.font= 1, varlen -10)
> # Checki.ng the models on val i.dati.on set
> pred_trees <- predi.ct(tree, vali.dati.on, type = "class")
> summary(pred_trees)
  0 1
785 273
```

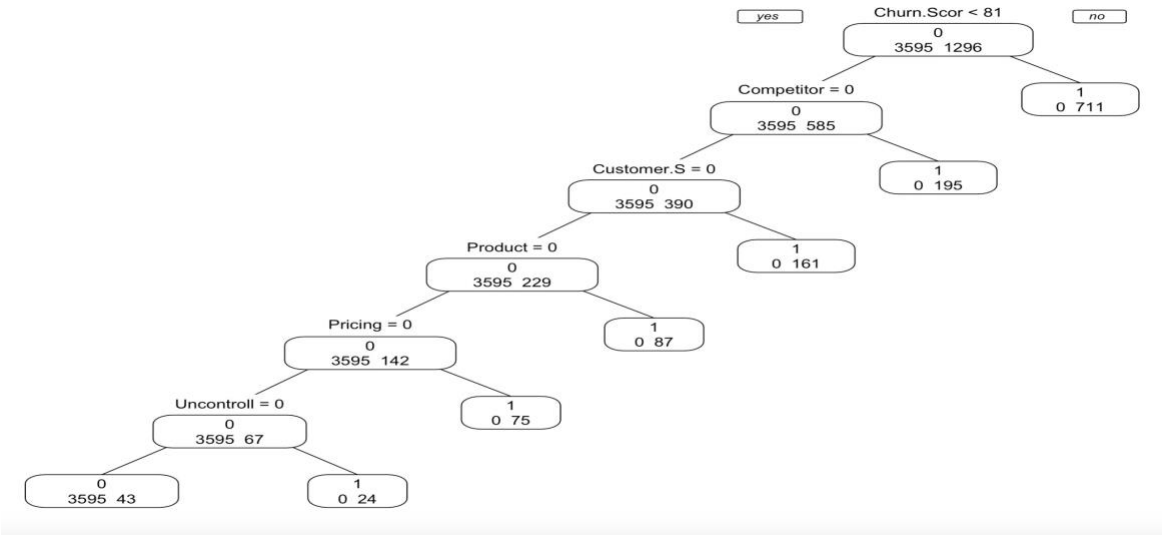
Model Accuracy of Validation Data of the Deepest Tree- 0.9924:

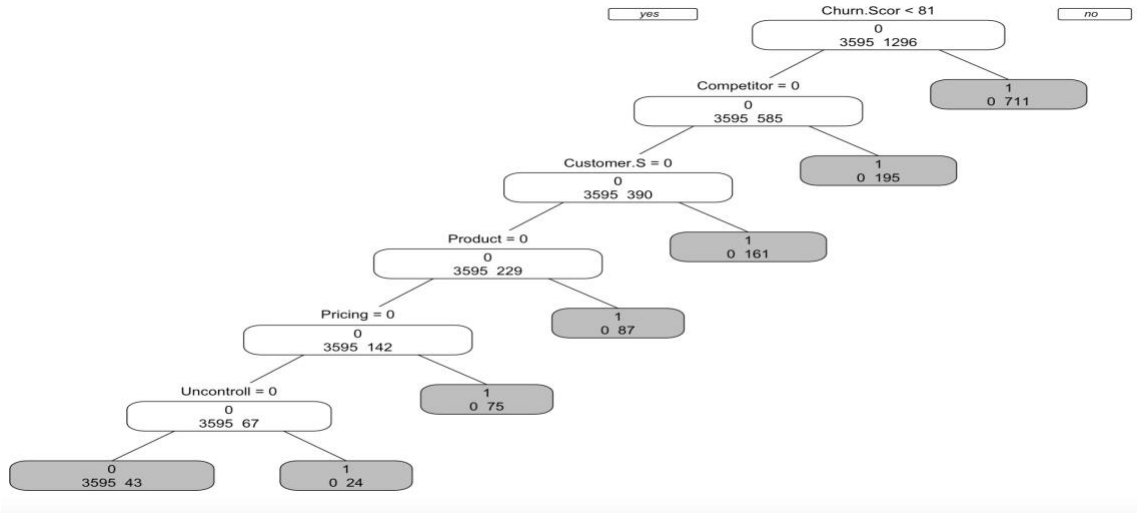
```
> confusionMatrix(pred_trees,validation$Churn.Labelenco )
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 777   8
         1   0 273

               Accuracy : 0.9924
                 95% CI : (0.9852, 0.9967)
    No Information Rate : 0.7344
    P-Value [Acc > NIR] : < 2e-16

                  Kappa : 0.9804

 Mcnemar's Test P-Value : 0.01333

            Sensitivity : 1.0000
            Specificity : 0.9715
         Pos Pred Value : 0.9898
         Neg Pred Value : 1.0000
             Prevalence : 0.7344
         Detection Rate : 0.7344
   Detection Prevalence : 0.7420
      Balanced Accuracy : 0.9858

       'Positive' Class : 0

> # Testing the model on testing set
> pred_trees <- predict(tree, test, type = "class")
> summary(pred_trees)
  0   1
822 272
```

Model Accuracy of Validation Data of the Deepest Tree- 0.9817:

```
> confusionMatrix(pred_trees,test$Churn.Labelenco)
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 802  20
         1   0 272

               Accuracy   0.9817
                 95% CI   (0.9719, 0.9888)
    No Information Rate   0.7331
    P-Value [Ace> NIR]    < 2.2e-16

                  Kappa   0.9522

 Mcnemar's Test P-Value   2.152e-05

            Sensih vity   1.0000
           Specificity    0.9315
         Pos Pred Value   0.9757
         Neg Pred Value   1.0000
            Prevalence    0.7331
        Detection Rate    0.7331
  Detection Prevalence    0.7514
     Balanced Accuracy    0.9658

       'Positive' Class   0

>tree<- rpart(Churn.Labelenco ~   data= train, cp   0.01, maxdepth   22)
> rpart.plot(tree)
```

1.2. PRUNED TREE





Model Accuracy of Validation Data of the Deepest Tree- 0.9924:

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 777   8
         1   0 273

               Accuracy : 0.9924
                 95% CI : (0.9852, 0.9967)
    No Information Rate : 0.7344
    P-Value [Acc > NIR] : < 2e-16

                  Kappa : 0.9804

 Mcnemar's Test P-Value : 0.01333

            Sensitivity : 1.0000
            Specificity : 0.9715
         Pos Pred Value : 0.9898
         Neg Pred Value : 1.0000
             Prevalence : 0.7344
         Detection Rate : 0.7344
   Detection Prevalence : 0.7420
      Balanced Accuracy : 0.9858

       'Positive' Class : 0
```

Model Accuracy of Validation Data of the Deepest Tree- 0.9817:

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 802  20
         1   0 272

               Accuracy : 0.9817
                 95% CI : (0.9719, 0.9888)
    No Information Rate : 0.7331
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.9522

 Mcnemar's Test P-Value : 2.152e-05

            Sensitivity : 1.0000
            Specificity : 0.9315
         Pos Pred Value : 0.9757
         Neg Pred Value : 1.0000
             Prevalence : 0.7331
         Detection Rate : 0.7331
   Detection Prevalence : 0.7514
      Balanced Accuracy : 0.9658

       'Positive' Class : 0
```

1.3. Logistic Regression

```
> options(scipen = 11)
> summary(logit_model)

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

Deviance Residuals:
         Min            1Q        Median            3Q           Max
-0.000002409  -0.000002409  -0.000002409  -0.000002409  -0.000002409

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)           -2.657e+01  5.916e+05       0        1
Longitude              4.658e-15  4.757e+03       0        1
Tenure.Months         -7.166e-16  1.993e+03       0        1
Monthly.Charges       -1.077e-14  1.027e+04       0        1
Total.Charges         -9.645e-18  2.220e+01       0        1
Churn.Score            3.738e-16  1.003e+03       0        1
CLTV                   7.424e-20  8.689e+00       0        1
Customer.Service1      6.586e-14  4.009e+04       0        1
Competitor1           -7.461e-16  3.944e+04       0        1
Product1               3.917e-15  4.489e+04       0        1
Pricing1              -2.268e-16  4.517e+04       0        1
Uncontrollable1        1.128e-14  6.979e+04       0        1
Genderenco1            2.312e-14  2.014e+04       0        1
Senior.Citizenenco1    8.518e-14  2.394e+04       0        1
Partnerenco1          -2.448e-14  2.225e+04       0        1
Dependentsenco1        3.926e-15  4.383e+04       0        1
Phone.Serviceenco1     2.454e-13  2.077e+05       0        1
Multiple.Linesenco1   -7.719e-15  5.702e+04       0        1
Online.Securityenco1   2.815e-14  5.814e+04       0        1
Online.Backupenco1     2.317e-14  5.656e+04       0        1
Device.Protectionenco1 1.281e-13  5.559e+04       0        1
Tech.Supportenco1      2.308e-14  5.839e+04       0        1
Streaming.TVenco1      1.552e-13  1.058e+05       0        1
Streaming.Moviesenco1  1.396e-13  1.044e+05       0        1
Paperless.Billingenco1 1.181e-14  2.430e+04       0        1
Internet.Serviceenco1  2.486e-13  2.616e+05       0        1
Internet.Serviceenco2  5.463e-13  5.159e+05       0        1
Contractenco1          2.595e-14  6.918e+04       0        1
Contractenco2          4.880e-16  4.205e+04       0        1
Payment.Methodenco1   -1.548e-13  3.955e+04       0        1
Payment.Methodenco2   -1.714e-13  3.221e+04       0        1
Payment.Methodenco3   -1.524e-13  3.972e+04       0        1
Avg_Score             -7.042e-16  1.074e+03       0        1

(Dispersion parameter for binomial family taken to be 1)
```

```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 0.0000000000000  on 1295  degrees of freedom
Residual deviance: 0.0000000075188  on 1263  degrees of freedom
  (3635 observations deleted due to missingness)
AIC: 66

Number of Fisher Scoring iterations: 25
```

Confusion Matrix and Statistics

```
          Reference
Prediction  0   1
         0 802  20
         1   0 272
```

Accuracy	0.9817
95% CI	(0.9719, 0.9888)
No Information Rate	0.7331
P-Value [Ace> NIR]	< 2.2e-16
Kappa	0.9522
Mcnemar's Test P-Value	2.152e-05
Sensitivity	1.0000
Specificity	0.9315
Pas Pred Value	0.9757
Neg Pred Value	1.0000
Prevalence	0.7331
Detection Rate	0.7331
Detection Prevalence	0.7514
Balanced Accuracy	0.9658
'Positive' Class	0

6. Conclusions

As per the scope of work performed in this project and comparing the results from various models like Decision trees, pruned trees, random forests, and Logistic Regression, we recommend the use of Random forests as the model was correctly able to predict the customers' churn the best. Moreover, the main disadvantage of Random forest trees i.e extreme sensitivity to change in data is automatically avoided here as the data collected for such a measure is mostly automated for every telecom company and hence no chance of error. A set system of rules are ensured in an industry setting to collect the data.

Presentation Recording:
https://cometmail-my.sharepoint.com/:v:/g/personal/mxp210085_utdallas_edu/Eb-wmsE5G6FKiM9xgazmjjEBBo9EK_XNXJFjV8VE28Wmww