# Home-Work-2

VigneshwarPesaru

3/19/2021

# R Reading Libraries

#Check the number of missing values in each column

```
##       customerID           gender    SeniorCitizen          Partner
##                0                0                0                0
##       Dependents           tenure     PhoneService    MultipleLines
##                0                0                0                0
##  InternetService   OnlineSecurity     OnlineBackup DeviceProtection
##                0                0                0                0
##      TechSupport      StreamingTV  StreamingMovies         Contract
##                0                0                0                0
## PaperlessBilling    PaymentMethod   MonthlyCharges     TotalCharges
##                0                0                0               11
##            Churn
##                0
```

Description:

############################OUTCOME VARIABLE##################################

The potential outcome variable in the main churn data set is the “churn.” If the customer is already left from the service, then he/she is given by churn “YES” else “NO”

# Summary Statistics

Datatype : Binary (0-1) churn : Yes(Customer already left the service because of several reasons) churn : No(Customer is in the service program)

########The variables invloved in the main_churn_data process are:#######

1. CustomerID: This is the unique(primary key) is given to all the customers who are currently in the service along with the customers who already left the service. This column is mainly used for east identification of any given customer.

   Datatype: character Length:7032

2. Gender: This column is just know the gender who are enrolled and left the telecom services

   Datatype: character Gender : Male Gender : Female

3. SeniorCitizen: This is the measurement taken to know whether the given citizen is a senior citizen or not.

   Datatype: Binary (0-1) SeniorCitizen: 1 (if the given customer is senior citizen) SeniorCitizen: 0 (if the given customer is not a senior citizen)

4. Partner: Whether the customer has a partner or not (Yes, No)

   Datatype: Binary (0-1) Partner: 1 (if the given customer have a partner) Partner: 0 (if the given customer dont have a partner)

5. Dependent: Whether the customer has a dependents or not (Yes, No)

   Datatype: Binary (0-1) Partner: 1 (if the given customer have a dependent) Partner: 0 (if the given customer dont have a dependent)

6. Tenure:Number of months the customer has stayed with the company

   Datatype: Integer (Non-negative) -Min. :1.00
   -1st Qu.:9.00
   -Median :29.00
   -Mean :32.42
   -3rd Qu.:55.00
   -Max. :73.00
   -Range :[1, 72]

7. PhoneService: Whether the customer has a phone service or not (Yes, No)

   Datatype: Binary (0-1) PhoneService: 1 (if the given customer have a PhoneService) PhoneService: 0 (if the given customer dont have a PhoneService)

8. MultipleLines : Whether the customer has multiple lines or not (Yes, No, No phone service)

   Datatype: Binary (0-1) MultipleLines: 1 (if the given customer have a MultipleLines) MultipleLines: 0 (if the given customer dont have a MultipleLines)

9. InternetService : Customer’s internet service provider (DSL, Fiber optic, No)

   Datatype: character InternetService : DSL(if the customer have DSL) InternetService : Fiber optic(if the customer have a Fiber optic) InternetService : No(if the customer dont have any)

10. OnlineSecurity : Whether the customer has online security or not (Yes, No, No internet service)

    Datatype: Binary (0-1) OnlineSecurity : Yes (if the customer have online security) OnlineSecurity : No (if the customer dont have an online security).

11. OnlineBackup : Whether the customer has online backup or not (Yes, No, No internet service)

    Datatype: Binary (0-1) OnlineBackup : Yes (if the customer have OnlineBackup ) OnlineBackup : No (if the customer dont have an OnlineBackup).

12. DeviceProtection : Whether the customer has device protection or not (Yes, No, No internet service)

    Datatype: Binary (0-1) DeviceProtection : Yes (if the customer have DeviceProtection ) DeviceProtection : No (if the customer dont have an DeviceProtection).

13. Techsupport :Whether the customer has tech support or not (Yes, No, No internet service)

    Datatype: Binary (0-1) Techsupport : Yes (if the customer have Techsupport ) Techsupport : No (if the customer dont have an Techsupport).

14. Streaming Tv : Whether the customer has streaming TV or not (Yes, No, No internet service)

Datatype: Binary (0-1) streamingtv : Yes (if the customer have streamingtv ) streamingtv : No (if the customer dont have an streamingtv).

15. Streaming Movies : Whether the customer has streaming movie or not (Yes, No, No internet service)

    Datatype: Binary (0-1) streamingMovie : Yes (if the customer have streamingmovie ) streamingMovie : No (if the customer dont have an streaming movie).

16. Contract : The contract term of the customer (Month-to-month, One year, Two year)

    Datatype: character Contract : Month-to-month Contract : One Year Contract : Two Year

17. PaperlessBilling : Whether the customer has paperless billing or not (Yes, No)

    Datatype: Binary (0-1) PaperlessBilling : Yes (if the customer have paperless billing ) PaperlessBilling : No (if the customer dont have paperless billing )

18. MonthlyCharges:Number of months the customer has stayed with the company

    Datatype: Real (Non-negative) -Min. :18.25
    -1st Qu.:35.59
    -Median :70.35
    -Mean :64.80
    -3rd Qu.:89.86
    -Max. :118.75 -Range :[18.25, 118.75]

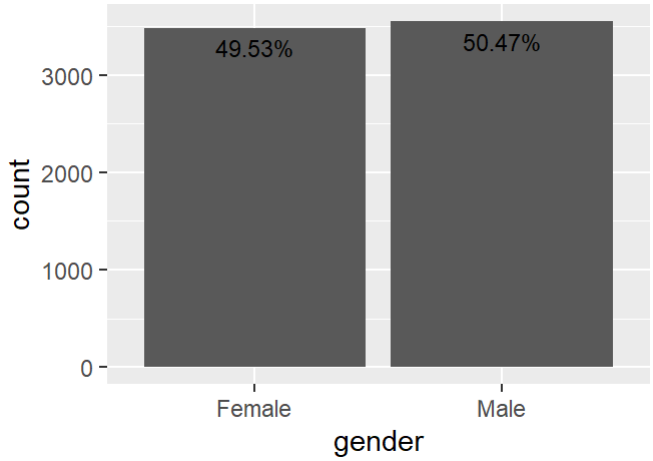
19. TotalCharges:The total amount charged to the customer

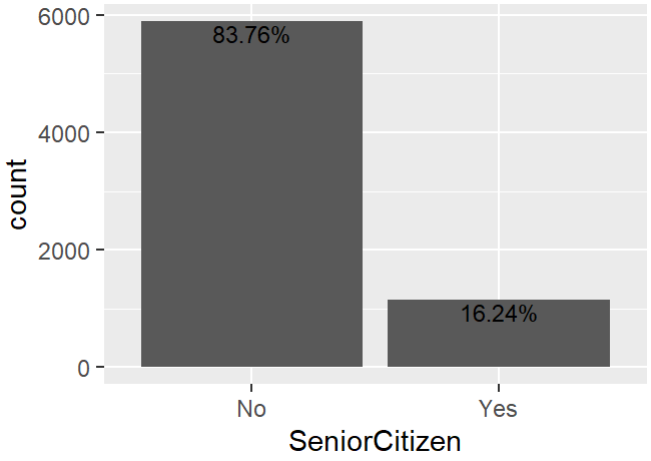
    Datatype: Real (Non-negative) -Min. :18.8
    -1st Qu.:401.4
    -Median :1397.5
    -Mean :2283.5
    -3rd Qu.:3794.7
    -Max. :8684.4 -Range :[18.8, 8684.4]

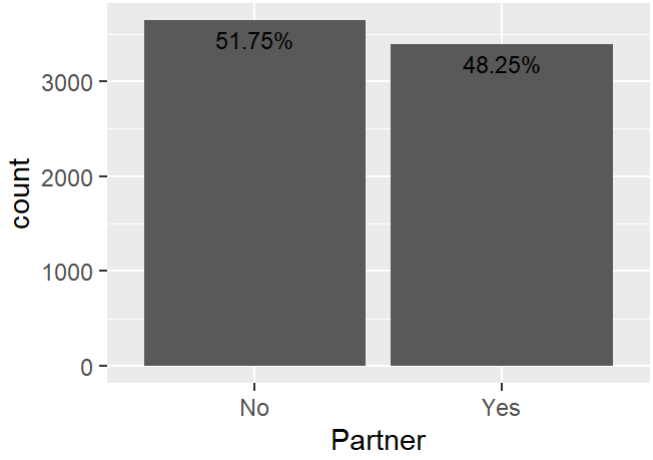
20. Churn: Whether the customer churned or not (Yes or No)

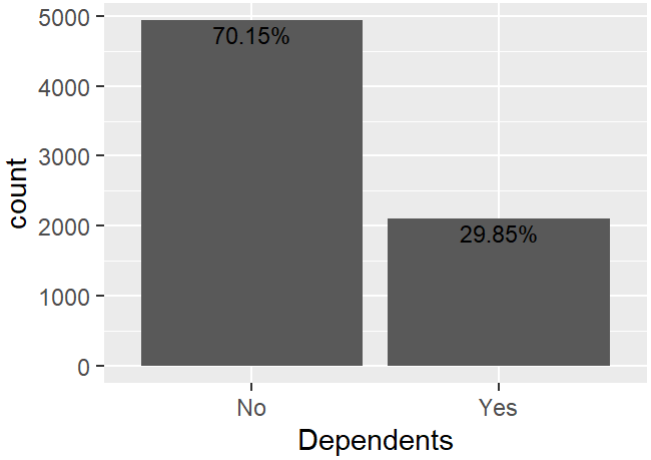
    Datatype: Binary (0-1) Churn : Yes (if the customer churned ) Churn : No (if the customer not churned )

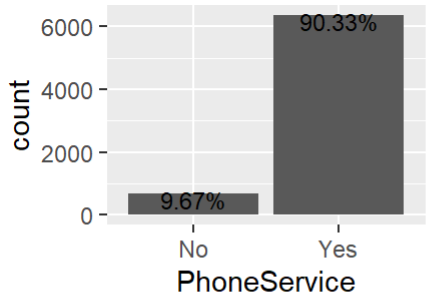
# Data Pre-processing

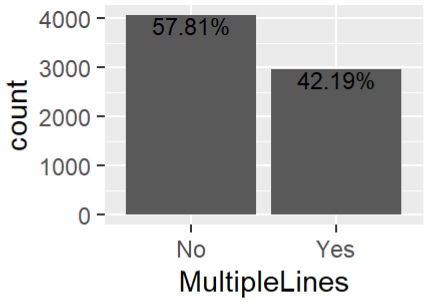

count
3000
2000
1000
0
49.53%
50.47%
Female
Male
gender

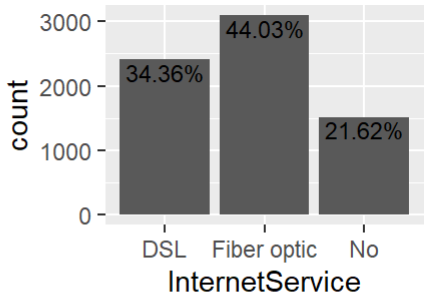

count
6000
4000
2000
0
83.76%
16.24%
No
Yes
SeniorCitizen

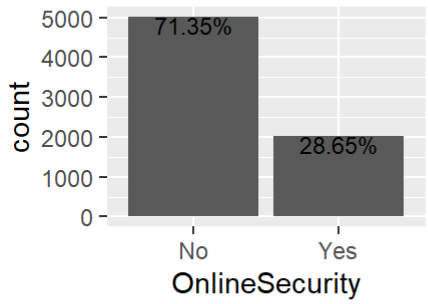

count
3000
2000
1000
0
51.75%
48.25%
No
Yes
Partner

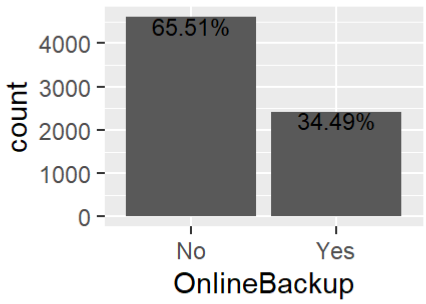

count
5000
4000
3000
2000
1000
0
70.15%
29.85%
No
Yes
Dependents

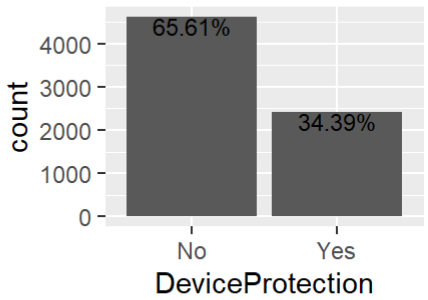

count
6000
4000
2000
0
9.67%
90.33%
No
Yes
PhoneService

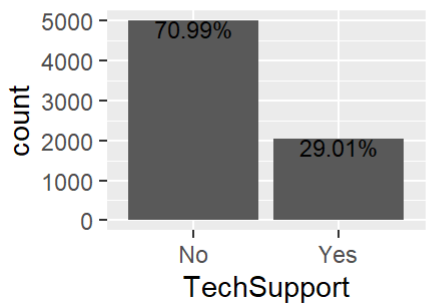

count
4000
3000
2000
1000
0
57.81%
42.19%
No
Yes
MultipleLines

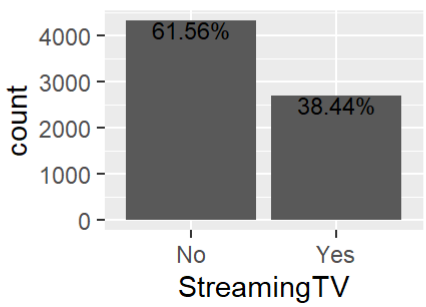

count
3000
2000
1000
0
34.36%
44.03%
21.62%
DSL
Fiber optic
No
InternetService

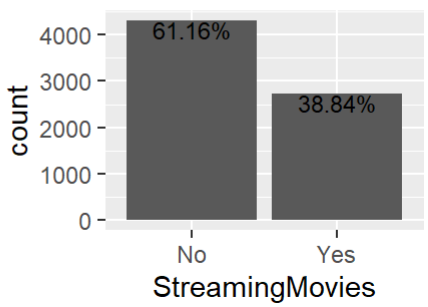

count
5000
4000
3000
2000
1000
0
71.35%
28.65%
No
Yes
OnlineSecurity

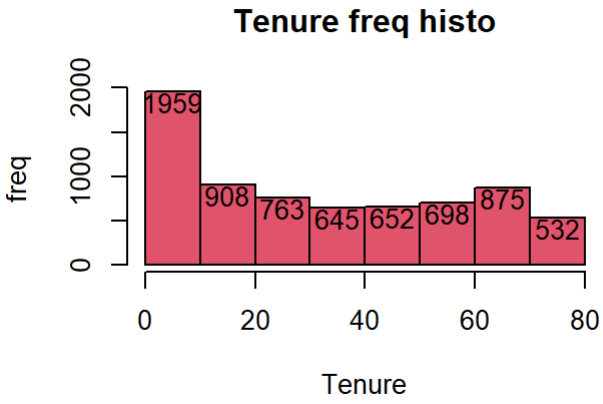

count
4000
3000
2000
1000
0
65.51%
34.49%
No
Yes
OnlineBackup

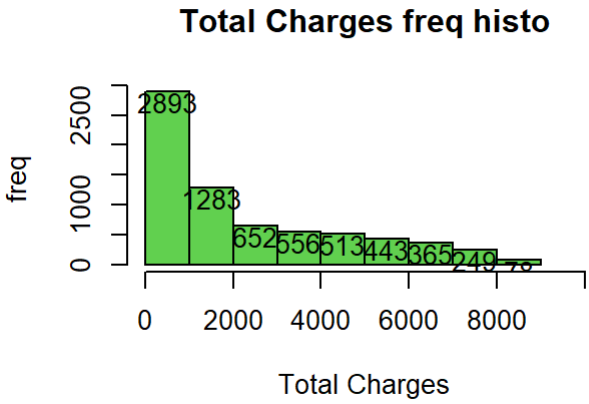

count
4000
3000
2000
1000
0
65.61%
34.39%
No
Yes
DeviceProtection

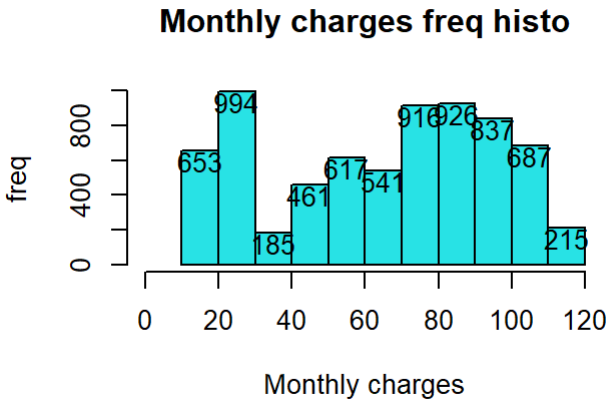

count
5000
4000
3000
2000
1000
0
70.99%
29.01%
No
Yes
TechSupport

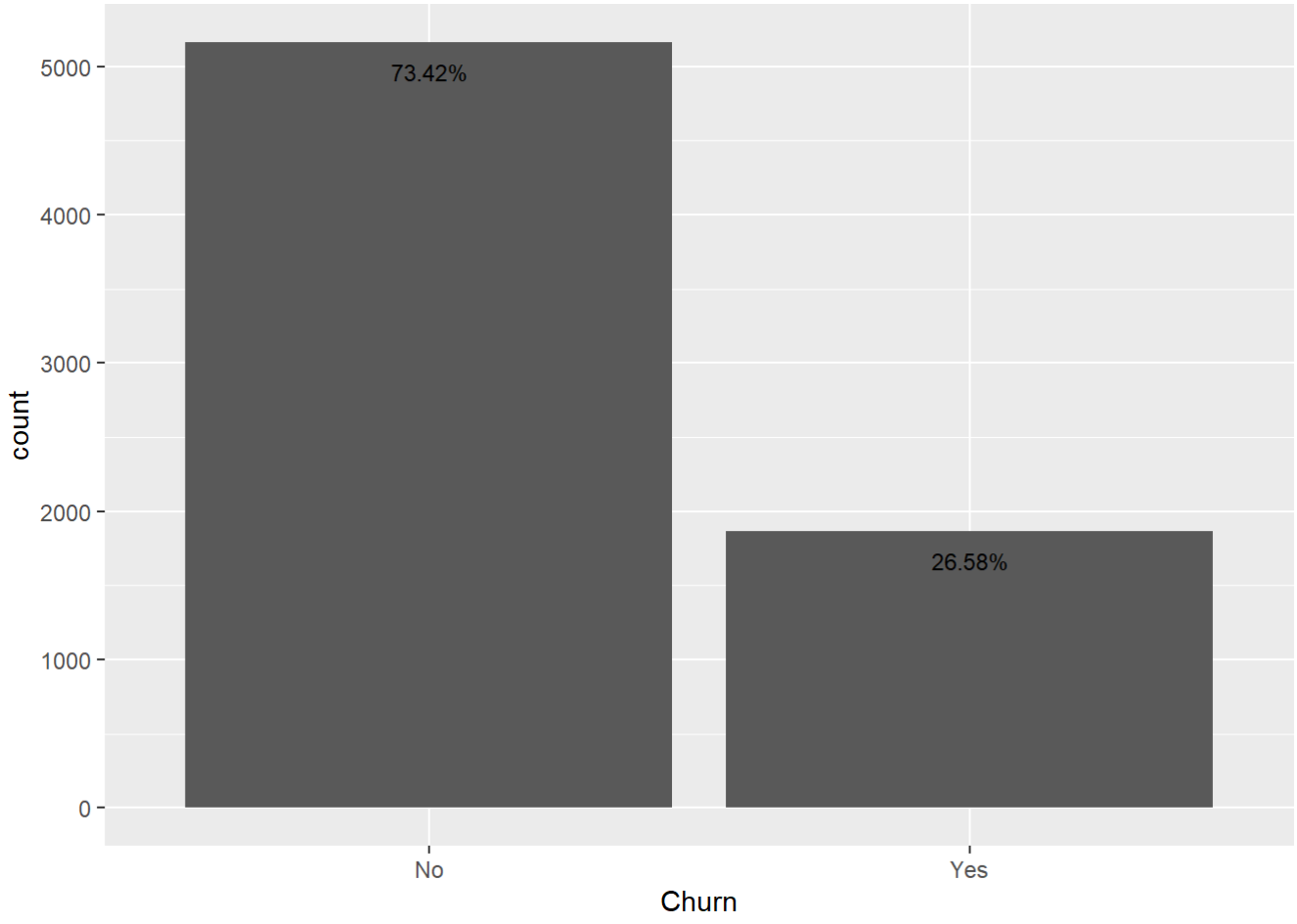

count
4000
3000
2000
1000
0
61.56%
38.44%
No
Yes
StreamingTV


count
4000
3000
2000
1000
0
61.16%
38.84%
No
Yes
StreamingMovies


Tenure freq histo
freq
0
1000
2000
1959
908
763
645
652
698
875
532
0
20
40
60
80
Tenure

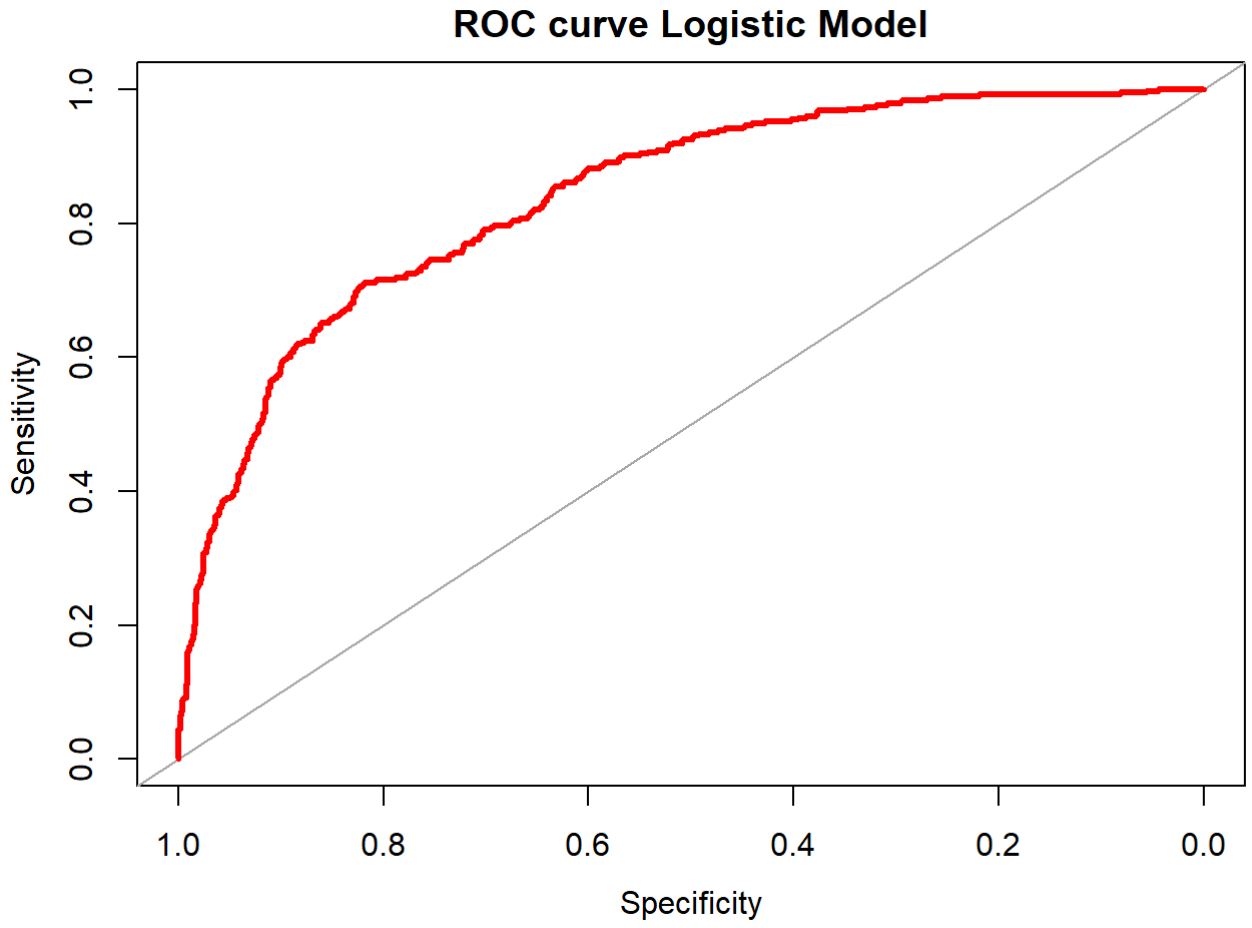

Total Charges freq histo
freq
0
1000
2500
2893
1283
652
556
513
443
365
249
0
2000
4000
6000
8000
Total Charges

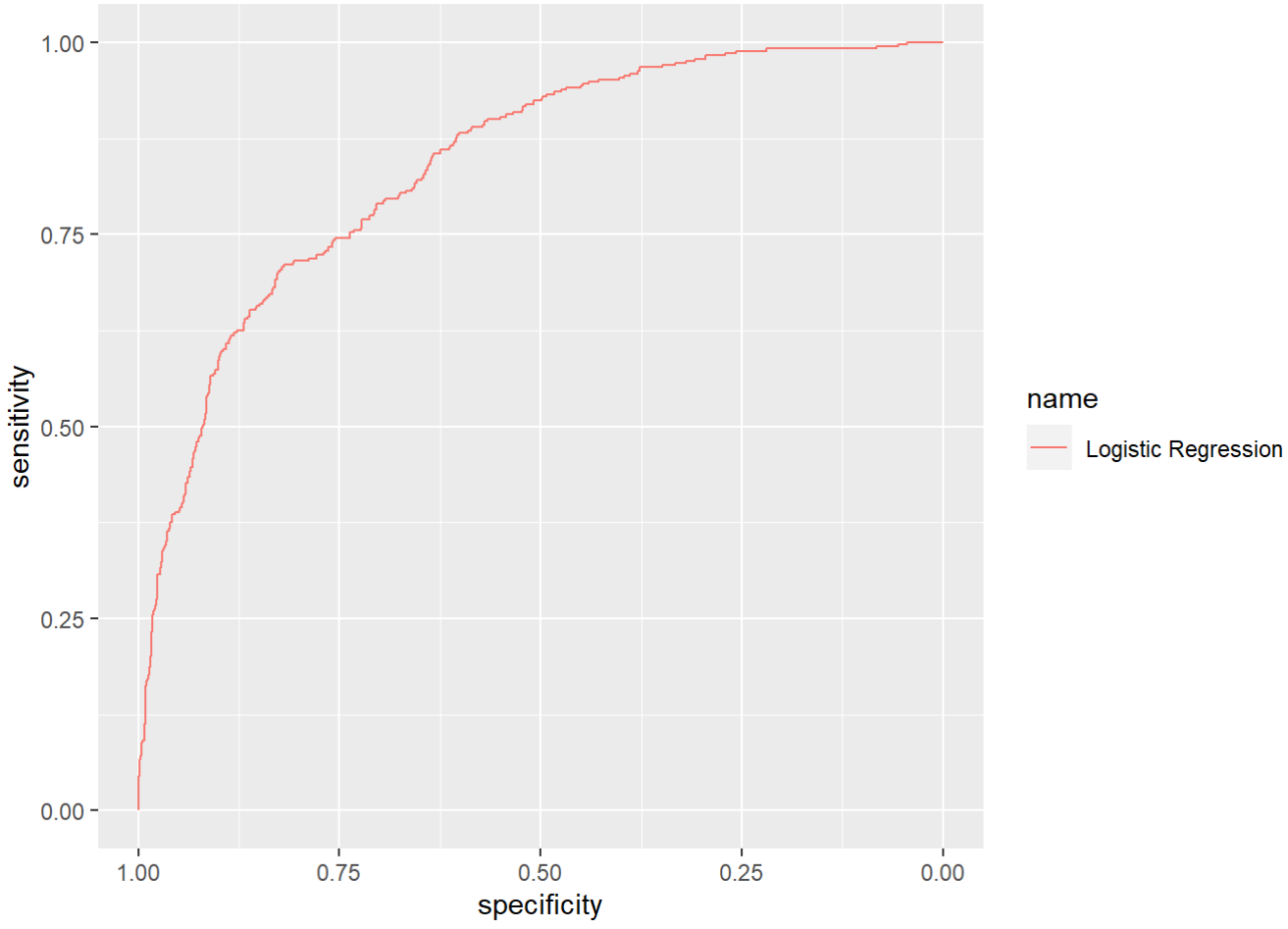

Monthly charges freq histo
freq
0
400
800
653
994
185
461
617
541
916
926
837
687
215
0
20
40
60
80
100
120
Monthly charges


count
0
1000
2000
3000
4000
5000
73.42%
26.58%
No
Yes
Churn

# Logistic regression

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  No Yes
##        No  939 166
##        Yes  93 207
##
##                Accuracy : 0.8157
##                  95% CI : (0.7944, 0.8356)
##     No Information Rate : 0.7345
##     P-Value [Acc > NIR] : 5.469e-13
##
##                   Kappa : 0.4958
##
##  Mcnemar's Test P-Value : 7.682e-06
##
##             Sensitivity : 0.9099
##             Specificity : 0.5550
##          Pos Pred Value : 0.8498
##          Neg Pred Value : 0.6900
##              Prevalence : 0.7345
##          Detection Rate : 0.6683
##    Detection Prevalence : 0.7865
##       Balanced Accuracy : 0.7324
##
##        'Positive' Class : No
##
```

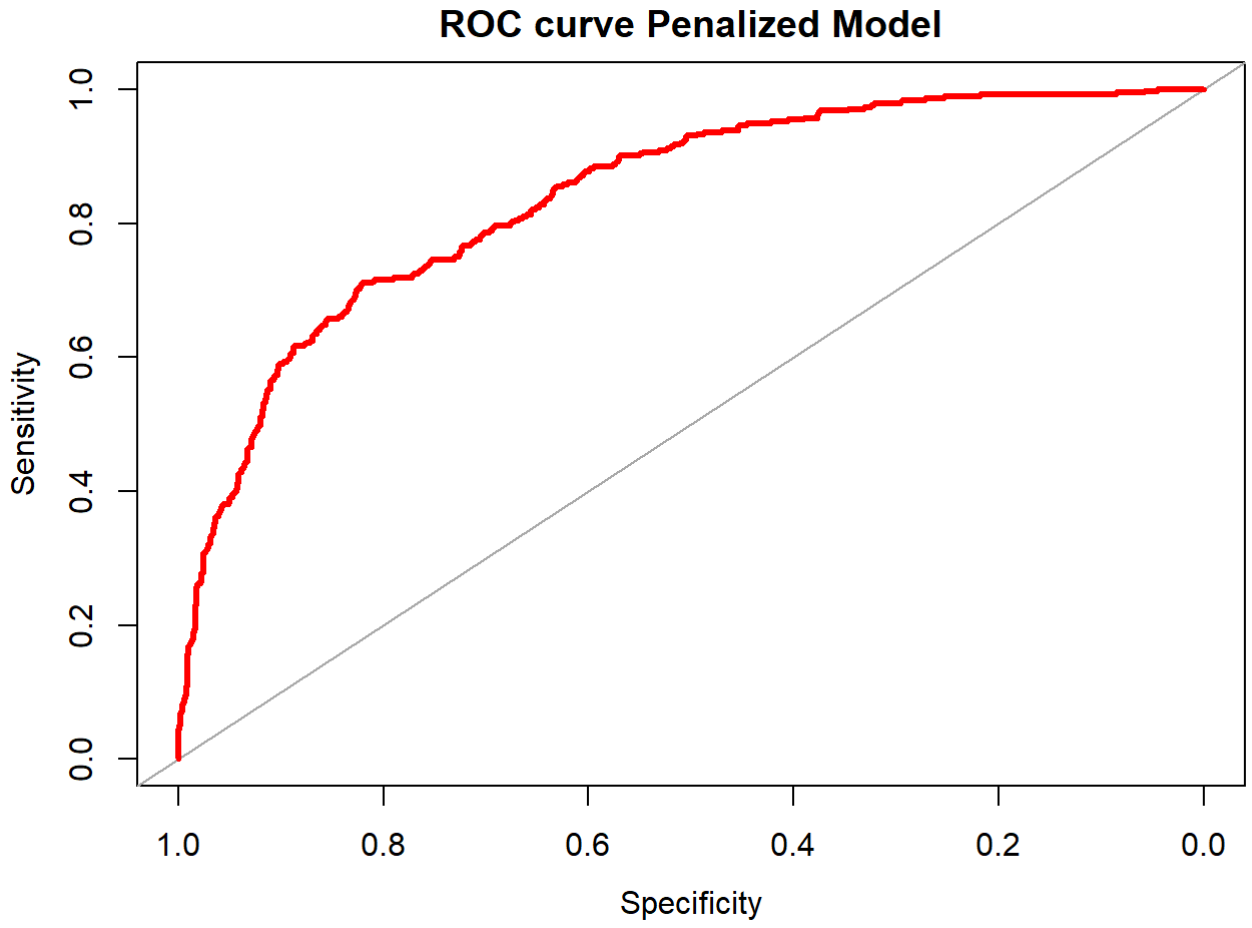

ROC curve Logistic Model
Sensitivity
1.0
0.8
0.6
0.4
0.2
0.0
Specificity
1.0
0.8
0.6
0.4
0.2
0.0

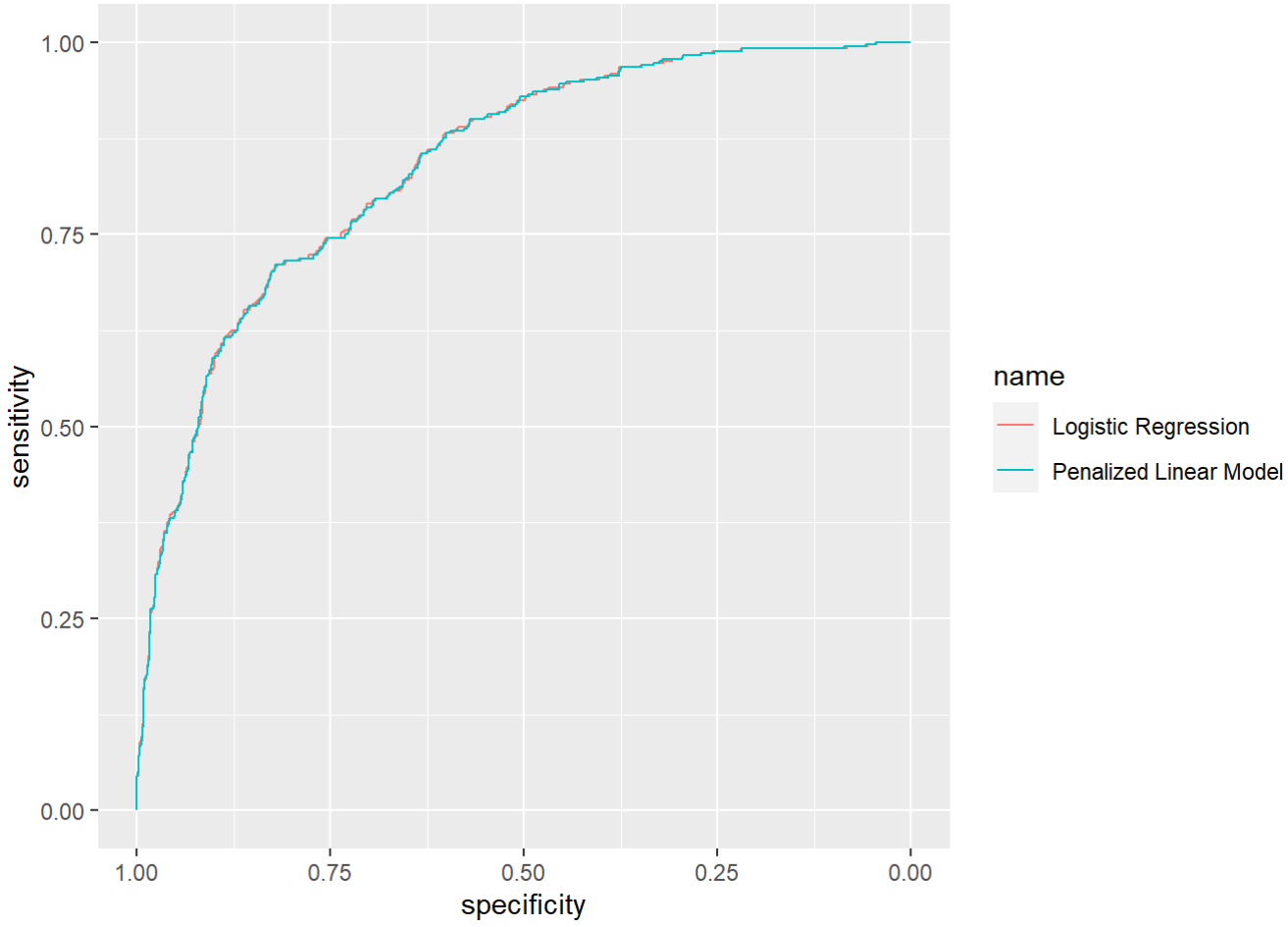

sensitivity
1.00
0.75
0.50
0.25
0.00
specificity
1.00
0.75
0.50
0.25
0.00
name
Logistic Regression

# Penalized Model

```
##       alpha lambda       ROC      Sens      Spec      ROCSD     SensSD
## 1 0.3333333      0 0.8367118 0.9012358 0.5133915 0.01600036 0.01138995
##       SpecSD
## 1 0.03630297
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  No Yes
##        No  939 167
##        Yes  93 206
##
##                Accuracy : 0.8149
##                  95% CI : (0.7936, 0.8349)
##     No Information Rate : 0.7345
##     P-Value [Acc > NIR] : 8.780e-13
##
##                   Kappa : 0.4934
##
##  Mcnemar's Test P-Value : 5.975e-06
##
##             Sensitivity : 0.9099
##             Specificity : 0.5523
##          Pos Pred Value : 0.8490
##          Neg Pred Value : 0.6890
##              Prevalence : 0.7345
##          Detection Rate : 0.6683
##    Detection Prevalence : 0.7872
##       Balanced Accuracy : 0.7311
##
##        'Positive' Class : No
##
```

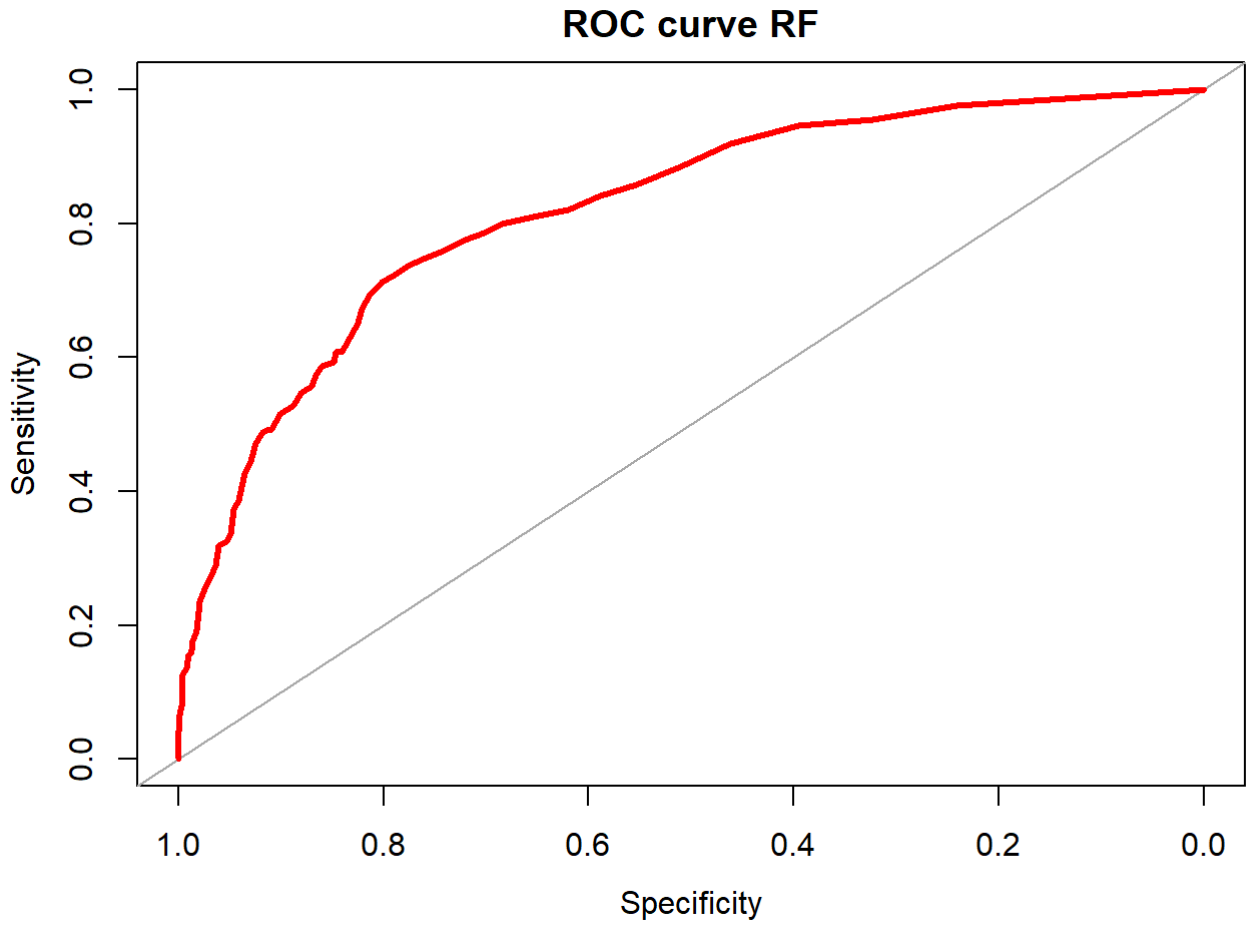

ROC curve Penalized Model
Sensitivity
1.0
0.8
0.6
0.4
0.2
0.0
Specificity
1.0
0.8
0.6
0.4
0.2
0.0

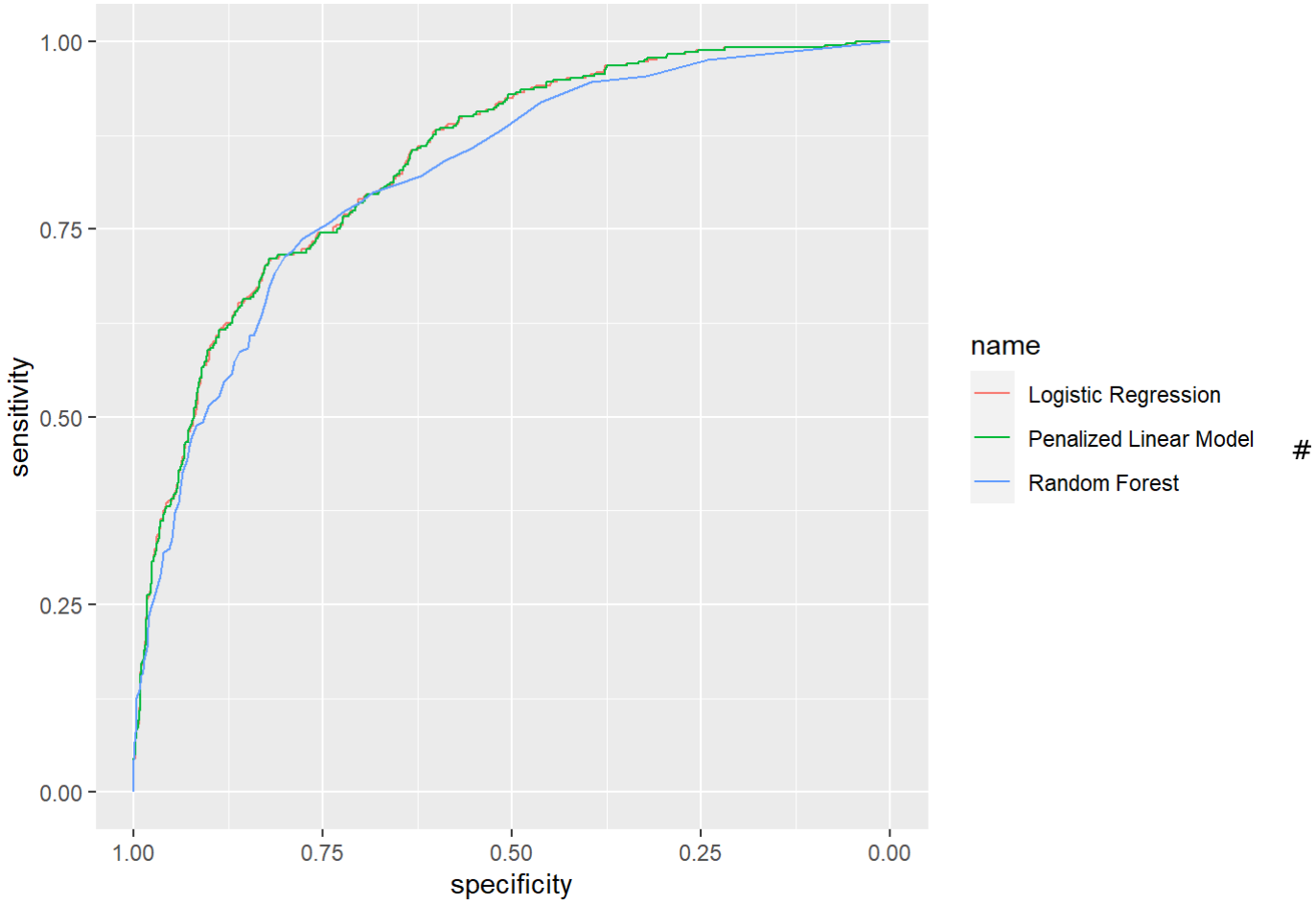

sensitivity
1.00
0.75
0.50
0.25
0.00
specificity
1.00
0.75
0.50
0.25
0.00
name
Logistic Regression
Penalized Linear Model

# Random Forest

```
## Random Forest
##
## 5627 samples
##   17 predictor
##    2 classes: 'No', 'Yes'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 5064, 5064, 5065, 5064, 5064, 5064, ...
## Resampling results across tuning parameters:
##
##   mtry  ROC        Sens       Spec
##    2    0.8231442  0.9298037  0.4191007
##    5    0.8136244  0.8973611  0.4832662
##    9    0.8062671  0.8893749  0.4919955
##   13    0.8041182  0.8828368  0.4839508
##   17    0.8039368  0.8830766  0.4845593
##
## ROC was used to select the optimal model using the largest value.
## The final value used for the model was mtry = 2.
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  No Yes
##        No  968 221
##        Yes  64 152
##
##                Accuracy : 0.7972
##                  95% CI : (0.7752, 0.8179)
##     No Information Rate : 0.7345
##     P-Value [Acc > NIR] : 2.778e-08
##
##                   Kappa : 0.3991
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.9380
##             Specificity : 0.4075
##          Pos Pred Value : 0.8141
##          Neg Pred Value : 0.7037
##              Prevalence : 0.7345
##          Detection Rate : 0.6890
##    Detection Prevalence : 0.8463
##       Balanced Accuracy : 0.6727
##
##        'Positive' Class : No
##
```


ROC curve RF
Sensitivity
1.0
0.8
0.6
0.4
0.2
0.0
Specificity
1.0
0.8
0.6
0.4
0.2
0.0

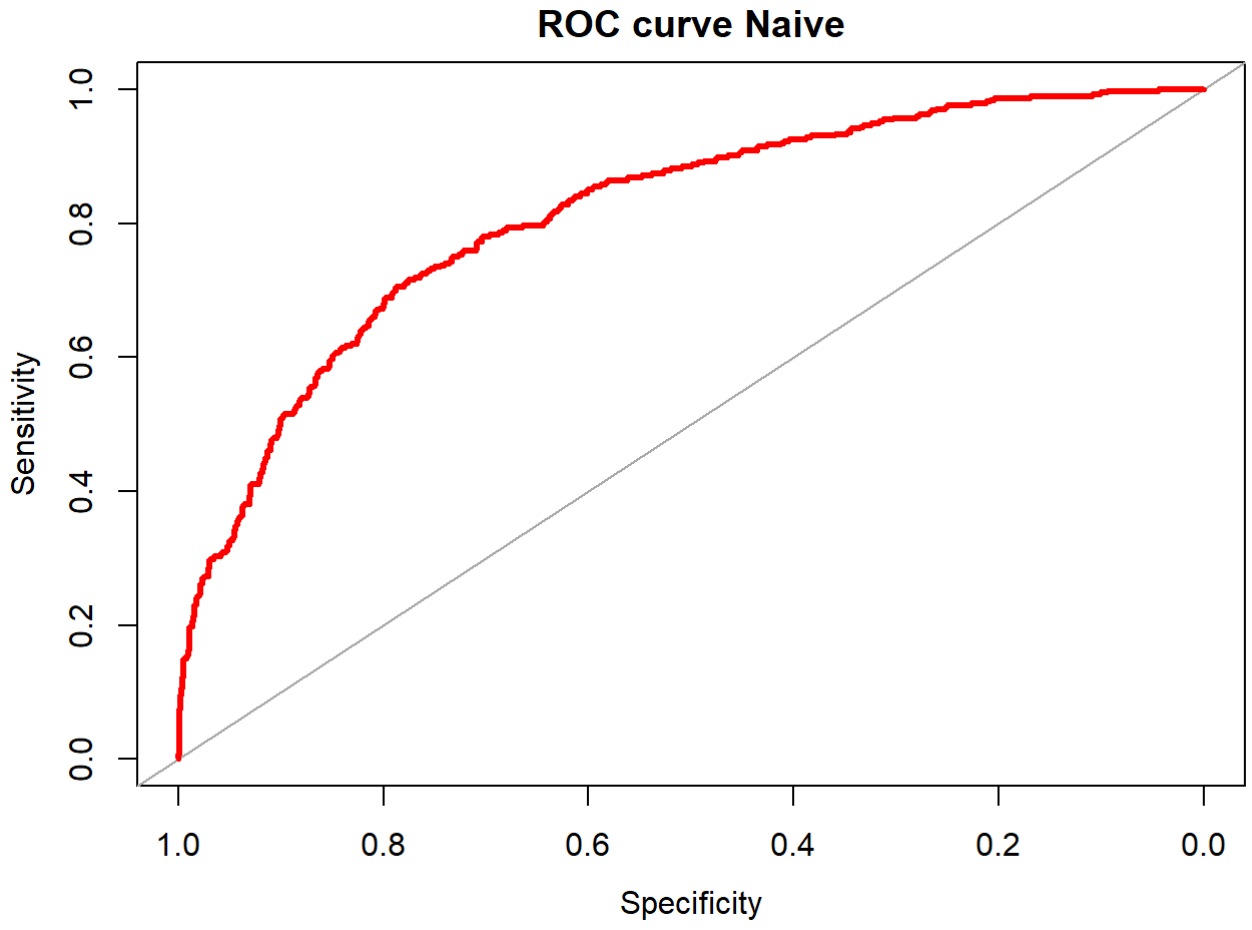

sensitivity
1.00
0.75
0.50
0.25
0.00
specificity
1.00
0.75
0.50
0.25
0.00
name
Logistic Regression
Penalized Linear Model
Random Forest

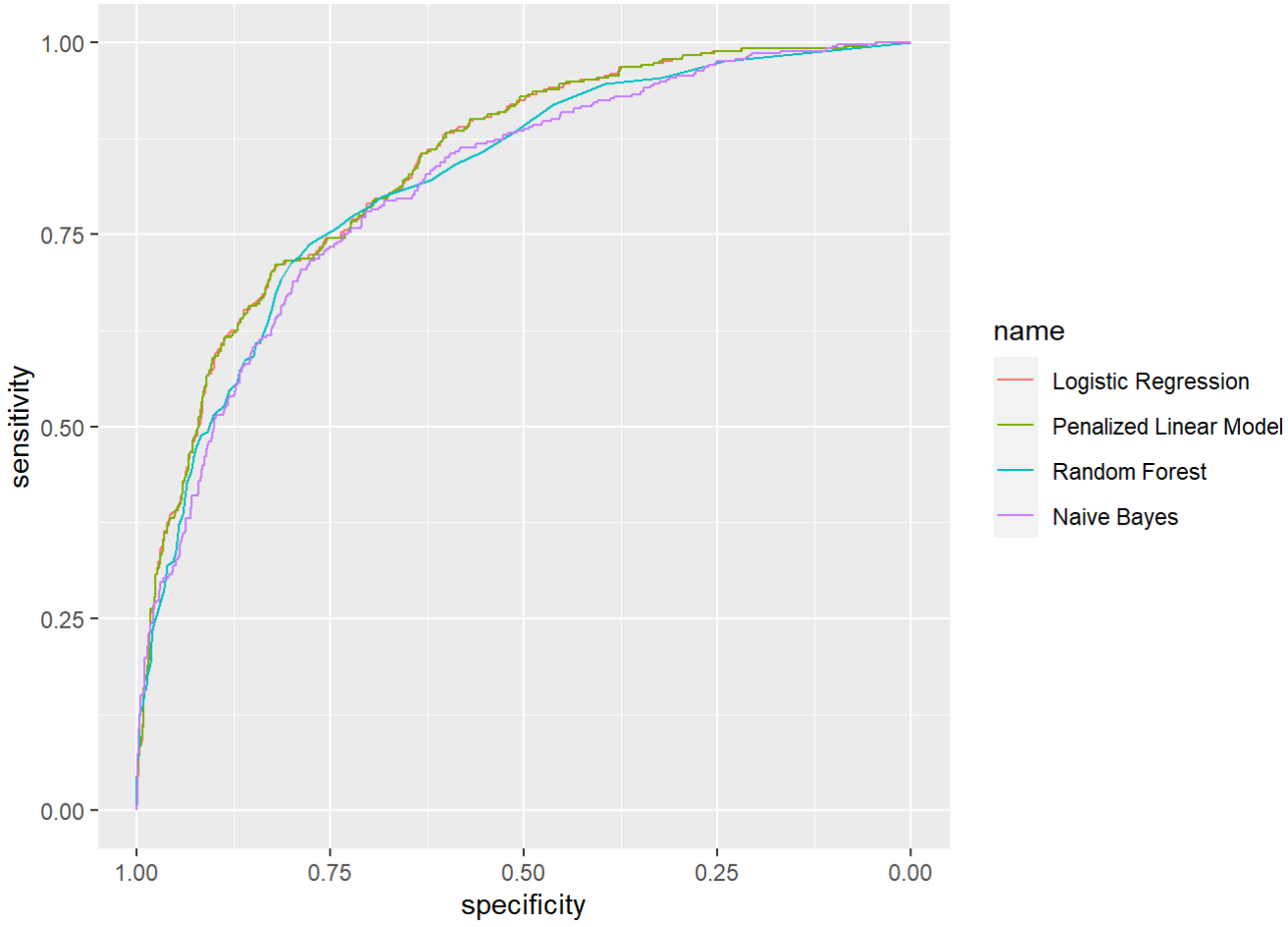
#

Naive Bayes

```
## Naive Bayes
##
## 5627 samples
##   17 predictor
##    2 classes: 'No', 'Yes'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 5065, 5064, 5064, 5064, 5064, 5064, ...
## Resampling results across tuning parameters:
##
##   usekernel  ROC        Sens       Spec
##   FALSE      0.7869084  0.8218368  0.5929441
##    TRUE      0.8057422  0.9786972  0.1817852
##
## Tuning parameter 'fL' was held constant at a value of 0
## Tuning
##  parameter 'adjust' was held constant at a value of 1
## ROC was used to select the optimal model using the largest value.
## The final values used for the model were fL = 0, usekernel = TRUE and adjust
##  = 1.
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  No Yes
##        No  977 251
##        Yes  55 122
##
##                Accuracy : 0.7822
##                  95% CI : (0.7597, 0.8035)
##     No Information Rate : 0.7345
##     P-Value [Acc > NIR] : 2.093e-05
##
##                   Kappa : 0.329
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.9467
##             Specificity : 0.3271
##          Pos Pred Value : 0.7956
##          Neg Pred Value : 0.6893
##              Prevalence : 0.7345
##          Detection Rate : 0.6954
##    Detection Prevalence : 0.8740
##       Balanced Accuracy : 0.6369
##
##        'Positive' Class : No
##
```


ROC curve Naive
Sensitivity
1.0
0.8
0.6
0.4
0.2
0.0
Specificity
1.0
0.8
0.6
0.4
0.2
0.0


sensitivity
1.00
0.75
0.50
0.25
0.00
specificity
1.00
0.75
0.50
0.25
0.00
name
Logistic Regression
Penalized Linear Model
Random Forest
Naive Bayes

# Comparing the models with repect to their Performances.

1. Linear Model:

   In the case of Linear Models, logistic regression and Penalized model performed same.

   1.1 Logistic regression:

   ```
   Accuracy    : 0.8157
   Sensitivity : 0.9099
   Specificity : 0.5550
   ```

   1.2 Penalized Model:

   ```
   Accuracy    : 0.8149
   Sensitivity : 0.9099
   Specificity : 0.5523
   ```

   - Logistic Regression performed better when compared to the penalized model.

2. Tree:

   The following are the performance parameters of the performed Random Forest Model

   2.1: Random Forest:

   ```
   Accuracy    : 0.7972
   Sensitivity : 0.9389
   Specificity : 0.4075


    *Although this model has performed better in terms of sensitivity
    but it performed bad in terms of accuracy.

    Hence we can drop the idea of implementing the Tree model
    for this Data set.
   ```

3. Non-Linear Model:

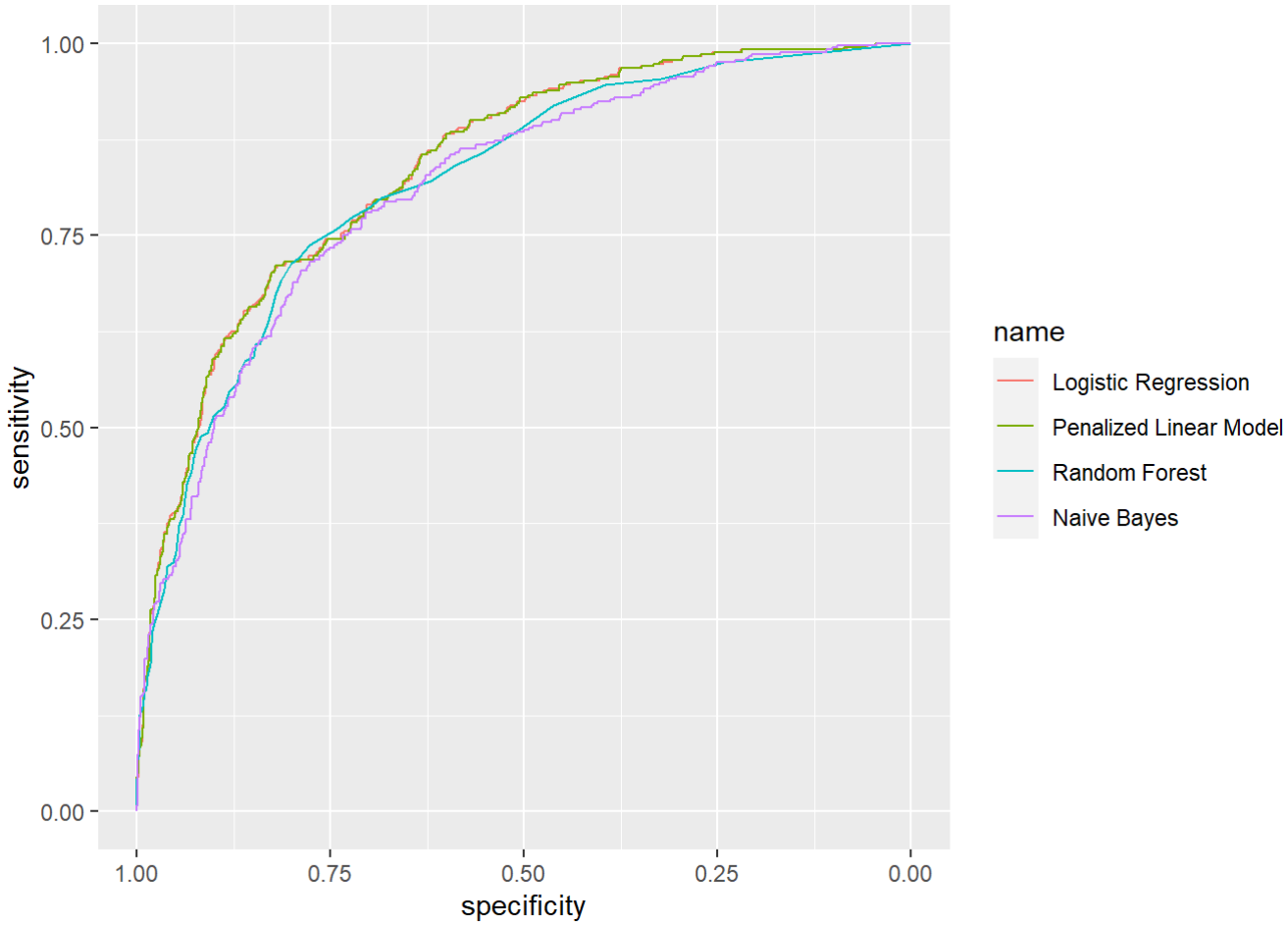
   3.1 : Naive Bayes:

   ```
   Accuracy    : 0.7822
   Sensitivity : 0.9467
   Specificity : 0.3271


    *Although the Non-linear model-Naive Bayes has performed better in terms of sensi
  tivity
    but it performed bad in terms of accuracy and specificity.

    Hence we can drop the idea of implementing the Non-linear models
    for this Data set.
   ```

# Final model selection:




```
* Out of all the models, based on the performance indicators, we can
now conclude that Linear models i.e. Logistic regression performned better than
the rest of the models

I would strongly recommend to use the Linear models for this Churn-telecom
data set.Infact, the logistic regression and penalized models performed same.
Hence further working on the linear models would be a good idea. Because the
accuracy, sensitivity and the Specificity are


Also, clearly from the ROC curve summarizes the performance of the model at different
threshold values by combining confusion matrices at all threshold  values.
X axis of ROC curve is the true positive rate (sensitivity) and y axis of
the ROC curve is the false positive rate (1- specificity) for all the models

From the above chart, we see that the above comaparsion chart of ROC  the
logistic regression and  penalized model(Linear model) have almost
have same area, which is maximum compared to rest of the models.

*Therefore, I conclude that implementing the Linear Models perform better for this dat
a.
```