SA4 Group Assignment: **Group26**

Yash Srivastava
PGID - 12010060

Charika Bhatia
PGID - 12010036

Jaideep Saraswat
PGID - 12010051

Q1.) **Objective:** Understanding the factors which are contributing to customer churn in order to predict the customers which will potentially churn.

Approach: The analysis was carried out on R and cross comparison of various models was performed on excel. The various steps followed to arrive at the appropriate model are as follows:

1. Perform EDA on the dataset which includes:
    a. Checking for null values
    b. Outlier detection
    c. Contingency tables for checking distribution of variables
    d. Checking for collinearity between variables
    e. Graphical approach to check for distribution of variables – using package ggplot
2. Understanding the dataset from a business standpoint and identifying the variables that may/may not be useful for predicting customer churn.
3. Splitting the Dataset into Test and Train
4. Model Iterations and Evaluation:
    a. LDA
    b. Logistic Regression
5. Model validation and inference

EDA

10Null values were found for TotalCharges, those observations were later dropped.

Distribution of Churn: From the data we can see that ~26% customers are churning, while not churning customers are ~74%.
Churn: 1329
Not Churn: 3672

Variables with no significance: “X”, “customerID”.

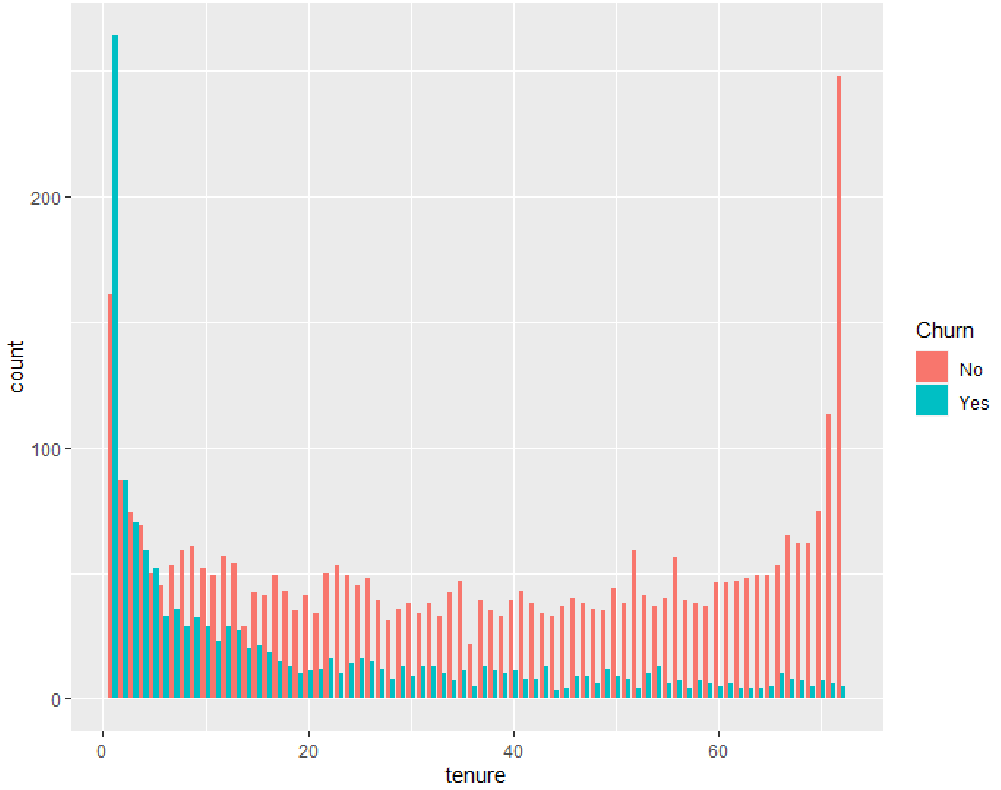
Distribution of some Explanatory variables with respect to Churn:

Gender

	Female	Male
No	1810	1862
Yes	680	649

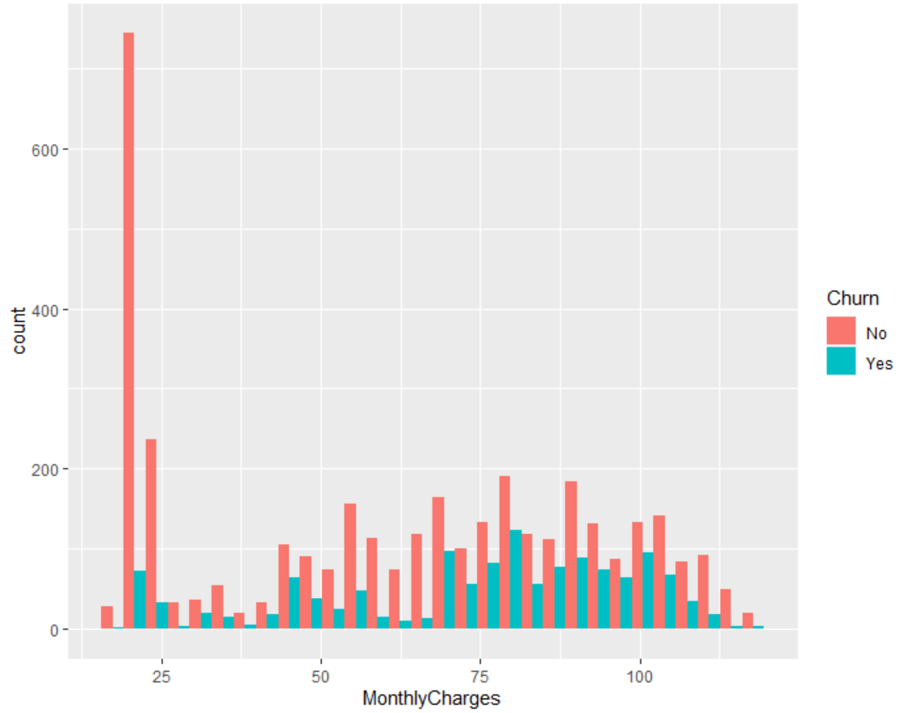
As we can see that the proportion of both male and female churn & not churn is almost the same. Hence, we drop Gender from the model.

Tenure:



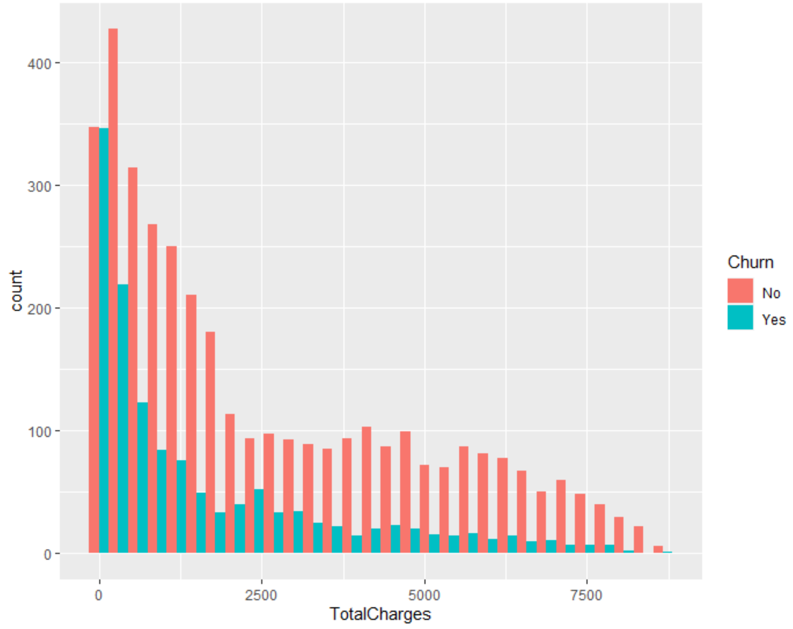
As tenure is increasing the number of not churning customers are also increasing.

MonthlyCharges:



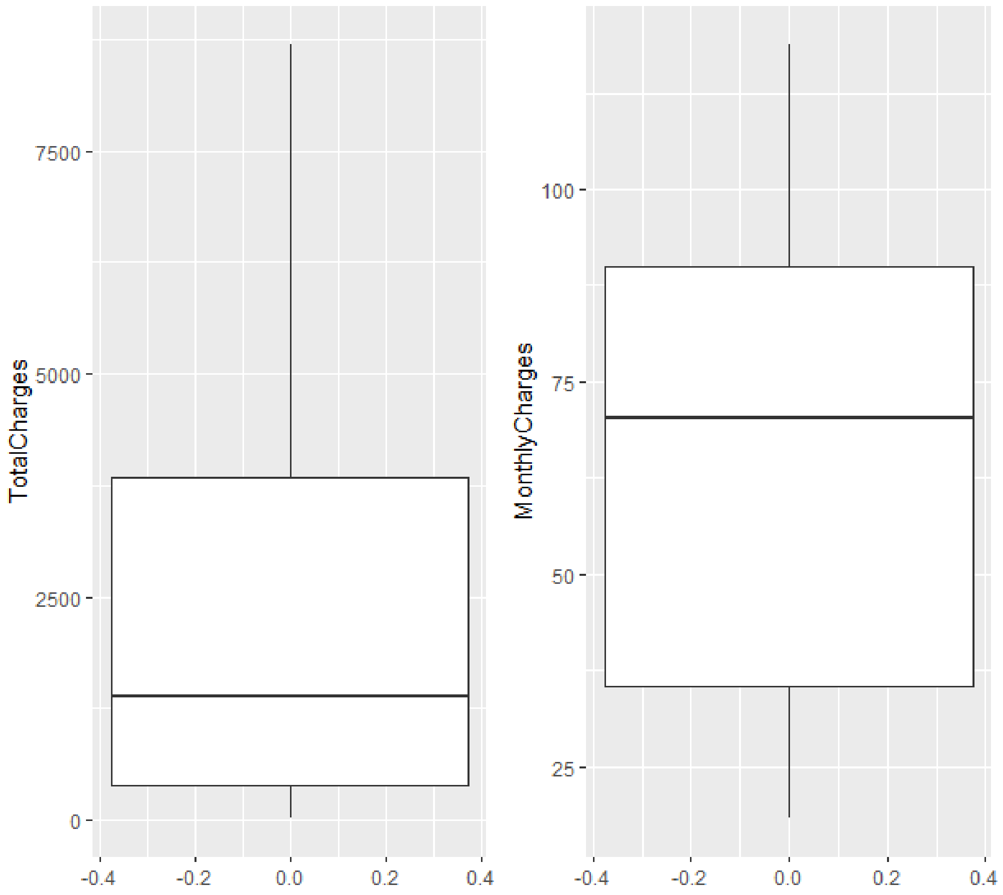
Churn is increasing as monthly charge increases.

Total Charges:



Churn and not churn are decreasing as total charges increase.

No outliers were found for continuous variables as seen in below box plots.

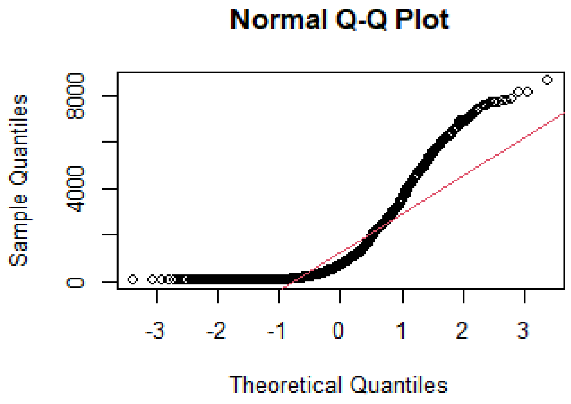


LDA

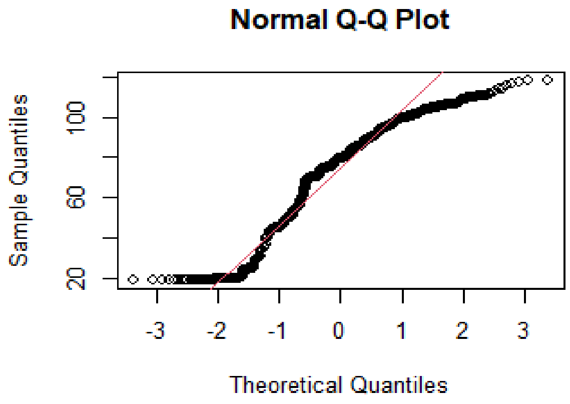
Checking LDA assumptions:

The given data does not follow the assumptions of linear discriminant analysis which are Normality and equal variances between groups. This is verified by the below tests and plots:

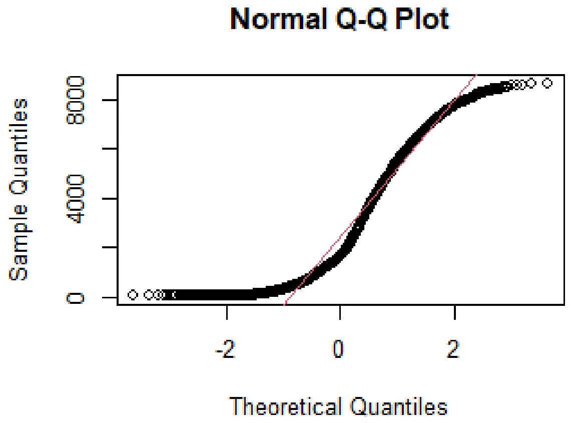
1. **Normality:** Using Shapiro-Wilk normality test we found that none of the continuous variables have a P value which is insignificant enough to accept the Null Hypothesis which states the data is Normally Distributed.



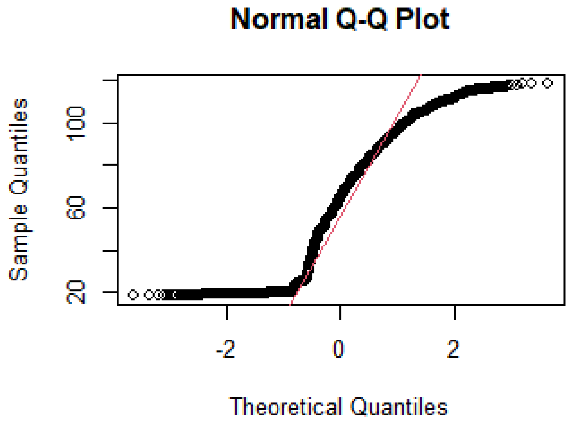
TotalCharges – Churn-Yes



MonthlyCharges-Churn-Yes

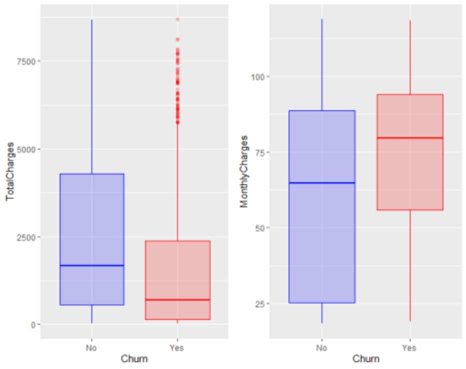


TotalCharges – Churn-No

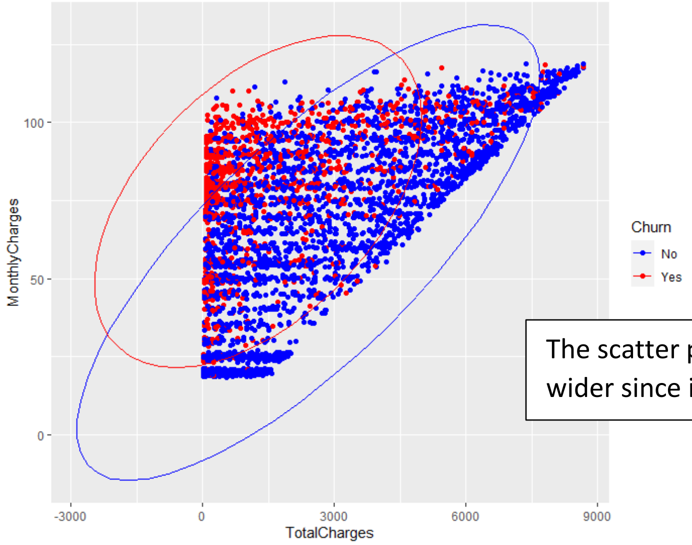


MonthlyCharges-Churn-No

2. **Equal Variances:** The given Box plots show that they have different lengths which is an indication of non-equal variances.



3. **Equal Covariance:** We have used scatter plots and Box M test to check the homogeneity of variance-covariance metrices.



The scatter plots show that Not Churn group is wider since it has wider spread of data points.

BOX M test: The P value for continuous variables is not insignificant enough to accept the Null Hypothesis stating that covariance matrices for Churn has equal variances across all groups.

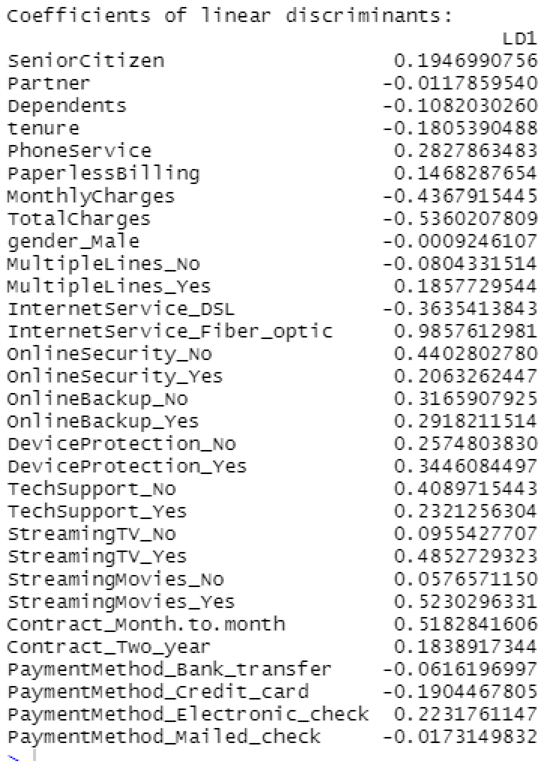
1. ITERATION 1

Converting categorical variables with 2 classes into 0,1 and other categorical variables with more than 2 classes as dummy variables with n-1 classes.

“Partner”, “Dependents”, “PhoneService”, “PaperlessBilling”,

“MultipleLines" "InternetService" "OnlineSecurity" "OnlineBackup"
"DeviceProtection" "TechSupport" "StreamingTV" "StreamingMovies"
“PaymentMethod”

Standardizing Tenure, TotalCharges and Monthly Charges.
Splitting dataset in test and train
Running the model with all variables on train:



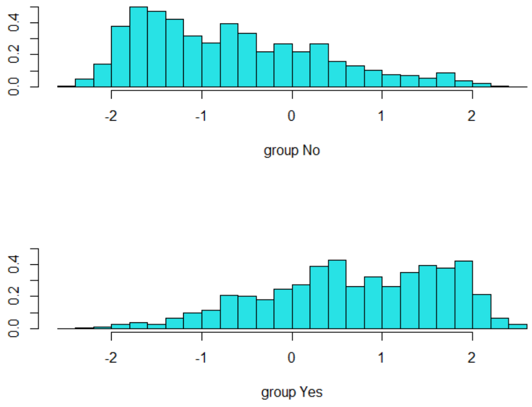
```
Coefficients of linear discriminants:
                                         LD1
SeniorCitizen                   0.1946990756
Partner                        -0.0117859540
Dependents                     -0.1082030260
tenure                         -0.1805390488
PhoneService                    0.2827863483
PaperlessBilling                0.1468287654
MonthlyCharges                 -0.4367915445
TotalCharges                   -0.5360207809
gender_Male                    -0.0009246107
MultipleLines_No               -0.0804331514
MultipleLines_Yes               0.1857729544
InternetService_DSL            -0.3635413843
InternetService_Fiber_optic     0.9857612981
OnlineSecurity_No               0.4402802780
OnlineSecurity_Yes              0.2063262447
OnlineBackup_No                 0.3165907925
OnlineBackup_Yes                0.2918211514
DeviceProtection_No             0.2574803830
DeviceProtection_Yes            0.3446084497
TechSupport_No                  0.4089715443
TechSupport_Yes                 0.2321256304
StreamingTV_No                  0.0955427707
StreamingTV_Yes                 0.4852729323
StreamingMovies_No              0.0576571150
StreamingMovies_Yes             0.5230296331
Contract_Month.to.month         0.5182841606
Contract_Two_year               0.1838917344
PaymentMethod_Bank_transfer    -0.0616196997
PaymentMethod_Credit_card      -0.1904467805
PaymentMethod_Electronic_check  0.2231761147
PaymentMethod_Mailed_check     -0.0173149832
```

Prior probabilities of groups:
No Yes
0.7271519 0.2728481

This indicates that 73% of training observations are customers who did not churn while 27% represent those who churned.

The coefficients of linear discriminant provide the linear combination of predictor variables that are used to form LDA decision rule.

We can use plot to produce plots of the linear discriminants, obtained by computing these coefficients for each of the training observations.



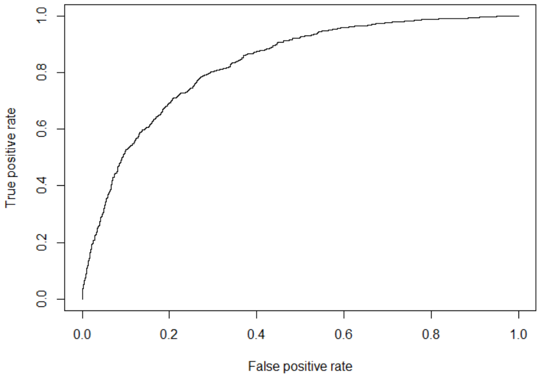
From the plots we can see that when the sum of estimates imputed < 0, the probability increases for Not churning and when the sum of estimates imputed > 0 the probability increases for Churning.

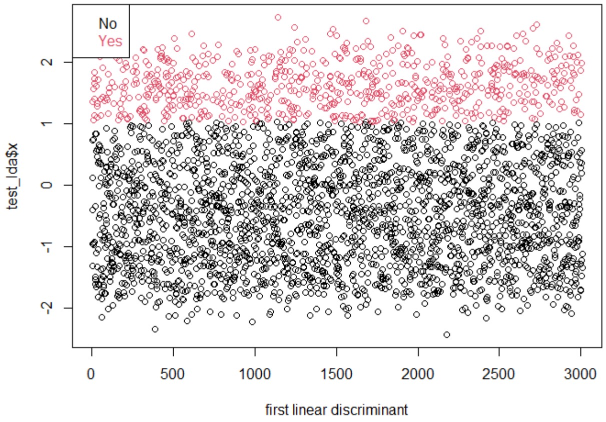
We use the Predict function to get the confusion matrix. The predict function gives us the posterior probability that is the probability weather the customer will churn or not. By default, the threshold is set at 50%. So, if the prob is greater than 50% the LDA classifier will predict that observation as Yes (Churning)

Confusion Matrix:

	No	Yes
No	65%	7.5%
Yes	13%	14%

Model Accuracy: **0.7926221**
Error Rate: **0.2073779**
AUC: **0.8312123**





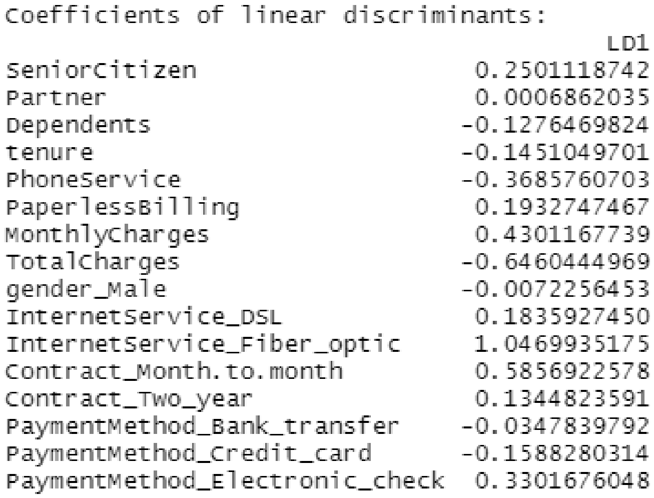
This model is not optimum since we have included all the variables and there is multi-collinearity.

2. ITERATION 2

Removing correlated variables using Chi-square test.

Variables dropped:
MultipleLines, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies

Running the model again on the dataset.



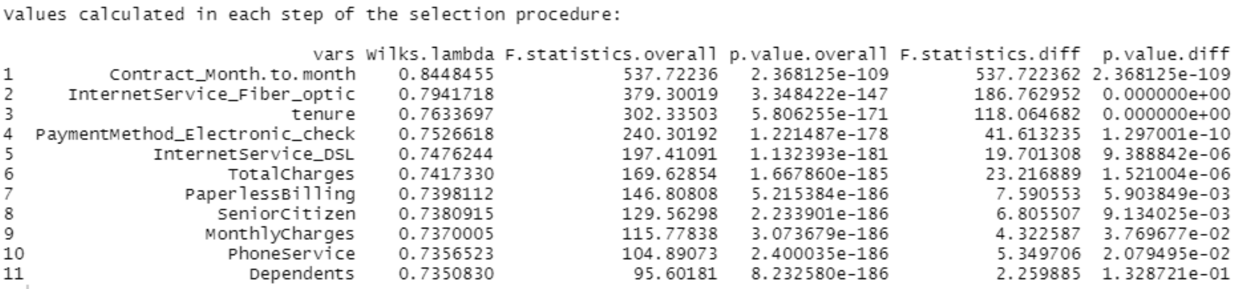
```
Coefficients of linear discriminants:
                                        LD1
SeniorCitizen                  0.2501118742
Partner                        0.0006862035
Dependents                    -0.1276469824
tenure                        -0.1451049701
PhoneService                  -0.3685760703
PaperlessBilling               0.1932747467
MonthlyCharges                 0.4301167739
TotalCharges                  -0.6460444969
gender_Male                   -0.0072256453
InternetService_DSL            0.1835927450
InternetService_Fiber_optic    1.0469935175
Contract_Month.to.month        0.5856922578
Contract_Two_year              0.1344823591
PaymentMethod_Bank_transfer   -0.0347839792
PaymentMethod_Credit_card     -0.1588280314
PaymentMethod_Electronic_check  0.3301676048
```

	No	Yes
No	65%	7%
Yes	14%	13%

Model Accuracy: **0.7879694**
Error Rate: **0.2120306**
AUC: **0.8280333**

Feature selection using Greedy Wilks Lambda

After running greedy.wilks for the variables selected in model 2, we get a list of significant variables which will minimize the overall lambda for the model.

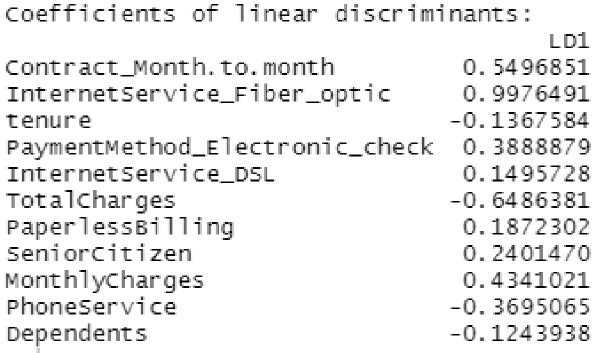


```
Values calculated in each step of the selection procedure:

                            vars Wilks.lambda F.statistics.overall p.value.overall F.statistics.diff  p.value.diff
1        Contract_Month.to.month    0.8448455            537.72236    2.368125e-109        537.722362 2.368125e-109
2     InternetService_Fiber_optic    0.7941718            379.30019    3.348422e-147        186.762952  0.000000e+00
3                         tenure    0.7633697            302.33503    5.806255e-171        118.064682  0.000000e+00
4  PaymentMethod_Electronic_check    0.7526618            240.30192    1.221487e-178         41.613235  1.297001e-10
5             InternetService_DSL    0.7476244            197.41091    1.132393e-181         19.701308  9.388842e-06
6                   TotalCharges    0.7417330            169.62854    1.667860e-185         23.216889  1.521004e-06
7                PaperlessBilling    0.7398112            146.80808    5.215384e-186          7.590553  5.903849e-03
8                  SeniorCitizen    0.7380915            129.56298    2.233901e-186          6.805507  9.134025e-03
9                 MonthlyCharges    0.7370005            115.77838    3.073679e-186          4.322587  3.769677e-02
10                  PhoneService    0.7356523            104.89073    2.400035e-186          5.349706  2.079495e-02
11                    Dependents    0.7350830             95.60181    8.232580e-186          2.259885  1.328721e-01
```

Let us run our Model again using these variables.

3. ITERATION 3



```
Coefficients of linear discriminants:
                                      LD1
Contract_Month.to.month         0.5496851
InternetService_Fiber_optic     0.9976491
tenure                         -0.1367584
PaymentMethod_Electronic_check  0.3888879
InternetService_DSL             0.1495728
TotalCharges                   -0.6486381
PaperlessBilling                0.1872302
SeniorCitizen                   0.2401470
MonthlyCharges                  0.4341021
PhoneService                   -0.3695065
Dependents                     -0.1243938
```

Confusion Matrix

	No	Yes
No	67%	8%
Yes	12%	13%

Model Accuracy: **0.7935154**
Error Rate: **0.2064846**
AUC: **0.829271**

Model Diagnostics:

Model2

Hotelling's T square :
Test stat: 65.91
Numerator df: 16
Denominator df: 2913
P-value: 0
Wilks Lambda
0.7342

Model3

Hotelling's T square :
Test stat: 95.602
Numerator df: 11
Denominator df: 2918
P-value: 0
Wilks Lambda
0.73508

Based on our diagnostic tests we have seen that both Models have significant P values for Hotelling's T square and Wilks Lambda.
However, the Hotelling test statistic is greater for Model 3 and Wilks Lambda value is approximately the same.
Based on AUC values we can observe that Model 3 and Model 2 have almost the same values

Based on above assumptions we have selected Model 3 as the optimum model.

Model Validation:

Evaluating Model3 on Test Data.

	No	Yes
No	1331	170
Yes	284	276

Model Accuracy: **0.7797186**
Error Rate: **0.2202814**

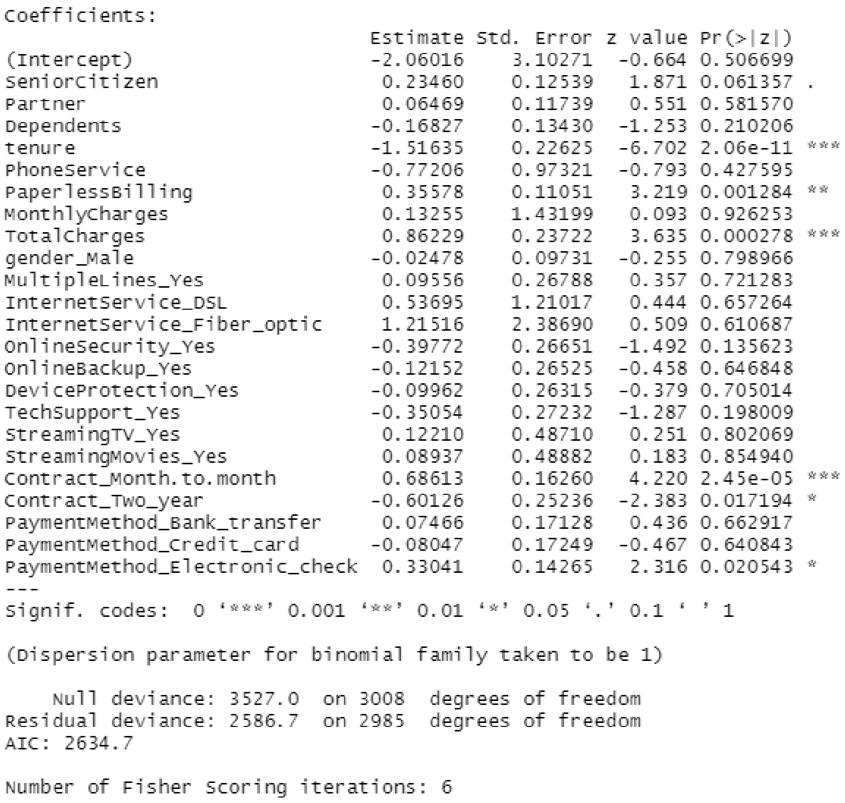
Logistic Regression -------------------------

Iteration 1
Converting Churn into binary response variable 0,1
Converting categorical variables with 2 classes into 0,1 and other categorical variables with more than 2 classes as dummy variables with n-1 classes.

“Partner”, “Dependents”, “PhoneService”, “PaperlessBilling”,

“MultipleLines" "InternetService" "OnlineSecurity" "OnlineBackup" "DeviceProtection" "TechSupport" "StreamingTV" "StreamingMovies" “PaymentMethod”

Standardizing Tenure, TotalCharges and Monthly Charges.
Splitting dataset in test and train
Running the model with all variables on train:



```
Coefficients:
                              Estimate Std. Error z value Pr(>|z|)    
(Intercept)                   -2.06016    3.10271  -0.664 0.506699    
SeniorCitizen                  0.23460    0.12539   1.871 0.061357 .  
Partner                        0.06469    0.11739   0.551 0.581570    
Dependents                    -0.16827    0.13430  -1.253 0.210206    
tenure                        -1.51635    0.22625  -6.702 2.06e-11 ***
PhoneService                  -0.77206    0.97321  -0.793 0.427595    
PaperlessBilling               0.35578    0.11051   3.219 0.001284 ** 
MonthlyCharges                 0.13255    1.43199   0.093 0.926253    
TotalCharges                   0.86229    0.23722   3.635 0.000278 ***
gender_Male                   -0.02478    0.09731  -0.255 0.798966    
MultipleLines_Yes              0.09556    0.26788   0.357 0.721283    
InternetService_DSL            0.53695    1.21017   0.444 0.657264    
InternetService_Fiber_optic    1.21516    2.38690   0.509 0.610687    
OnlineSecurity_Yes            -0.39772    0.26651  -1.492 0.135623    
OnlineBackup_Yes              -0.12152    0.26525  -0.458 0.646848    
DeviceProtection_Yes          -0.09962    0.26315  -0.379 0.705014    
TechSupport_Yes               -0.35054    0.27232  -1.287 0.198009    
StreamingTV_Yes                0.12210    0.48710   0.251 0.802069    
StreamingMovies_Yes            0.08937    0.48882   0.183 0.854940    
Contract_Month.to.month        0.68613    0.16260   4.220 2.45e-05 ***
Contract_Two_year             -0.60126    0.25236  -2.383 0.017194 *  
PaymentMethod_Bank_transfer    0.07466    0.17128   0.436 0.662917    
PaymentMethod_Credit_card     -0.08047    0.17249  -0.467 0.640843    
PaymentMethod_Electronic_check  0.33041    0.14265   2.316 0.020543 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3527.0  on 3008  degrees of freedom
Residual deviance: 2586.7  on 2985  degrees of freedom
AIC: 2634.7

Number of Fisher Scoring iterations: 6
```

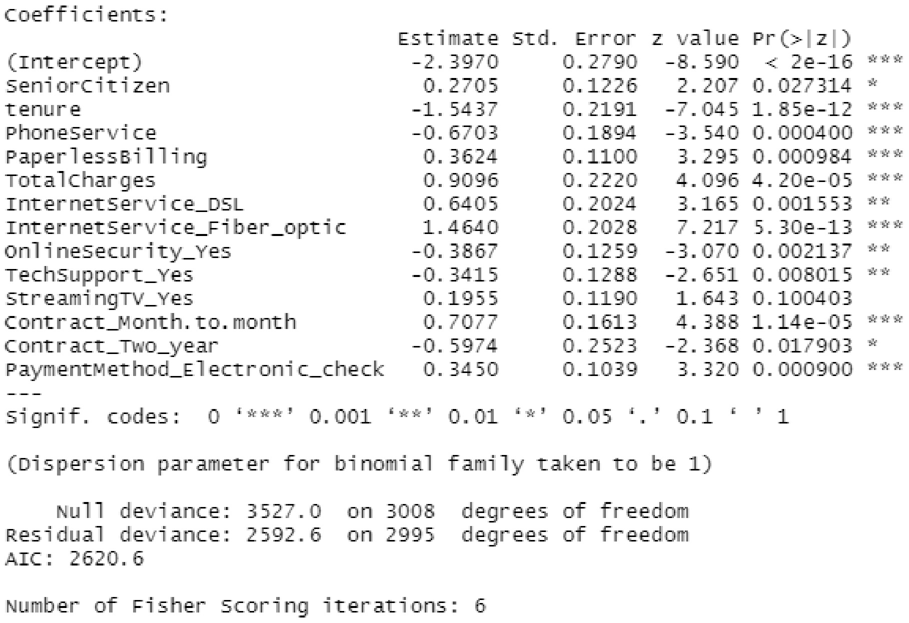
AIC: **2641.8**

From the model above we can see that there are few significant variables as Tenure, PaperlessBilling etc. Although the models seem good in terms of deviances but there is high collinearity among variables.

Dropped these variables as they have strong collinearity verified by Chi-square test MultipleLines_No, OnlineSecurity_No, OnlineBackup_No, DeviceProtection_No, TechSupport_No StreamingTV_No, StreamingMovies_No

Running the model against all remaining variables using Stepwise regression which will provide optimal set of features.

Iteration 2



```
Coefficients:
                              Estimate Std. Error z value Pr(>|z|)    
(Intercept)                    -2.3970     0.2790  -8.590  < 2e-16 ***
SeniorCitizen                   0.2705     0.1226   2.207 0.027314 *  
tenure                         -1.5437     0.2191  -7.045 1.85e-12 ***
PhoneService                   -0.6703     0.1894  -3.540 0.000400 ***
PaperlessBilling                0.3624     0.1100   3.295 0.000984 ***
TotalCharges                    0.9096     0.2220   4.096 4.20e-05 ***
InternetService_DSL             0.6405     0.2024   3.165 0.001553 ** 
InternetService_Fiber_optic     1.4640     0.2028   7.217 5.30e-13 ***
OnlineSecurity_Yes             -0.3867     0.1259  -3.070 0.002137 ** 
TechSupport_Yes                -0.3415     0.1288  -2.651 0.008015 ** 
StreamingTV_Yes                 0.1955     0.1190   1.643 0.100403    
Contract_Month.to.month         0.7077     0.1613   4.388 1.14e-05 ***
Contract_Two_year              -0.5974     0.2523  -2.368 0.017903 *  
PaymentMethod_Electronic_check  0.3450     0.1039   3.320 0.000900 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3527.0  on 3008  degrees of freedom
Residual deviance: 2592.6  on 2995  degrees of freedom
AIC: 2620.6

Number of Fisher Scoring iterations: 6
```

AIC: **2620**

Confusion Matrix:

	FALSE	TRUE
No	65%	8%
Yes	12%	15%

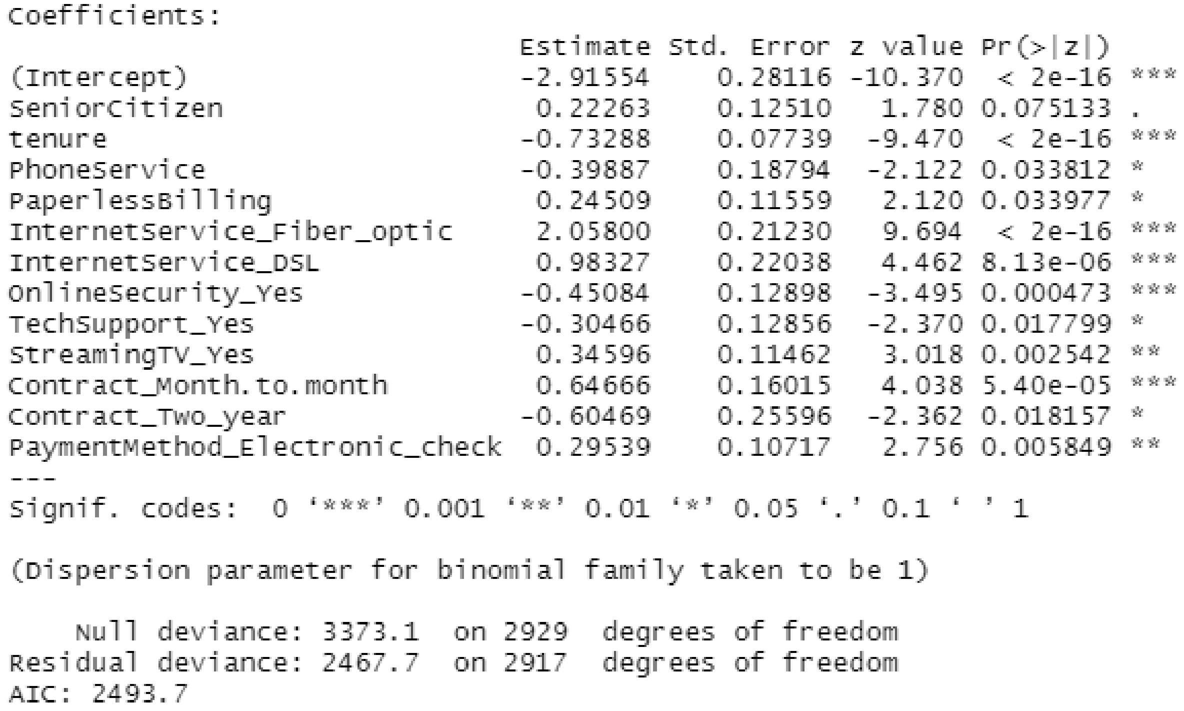
Accuracy**: 0.7942838**
Sensitivity**: 65%**
Specificity**: 84%**

From the above model we can see that Residual deviance has improved which suggests that the model fits the data better than model 1.
The AIC has reduced

All the explanatory variables are now significant except Contract_Month.to.month. The Sensitivity and Specificity have an imbalance.

Checked the VIFs:
TotalCharges and tenure have the highest VIF value.
Dropped TotalCharges and ran the model again.

Iteration 3



```
Coefficients:
                              Estimate Std. Error z value Pr(>|z|)    
(Intercept)                   -2.91554    0.28116 -10.370  < 2e-16 ***
SeniorCitizen                  0.22263    0.12510   1.780 0.075133 .  
tenure                        -0.73288    0.07739  -9.470  < 2e-16 ***
PhoneService                  -0.39887    0.18794  -2.122 0.033812 *  
PaperlessBilling               0.24509    0.11559   2.120 0.033977 *  
InternetService_Fiber_optic    2.05800    0.21230   9.694  < 2e-16 ***
InternetService_DSL            0.98327    0.22038   4.462 8.13e-06 ***
OnlineSecurity_Yes            -0.45084    0.12898  -3.495 0.000473 ***
TechSupport_Yes               -0.30466    0.12856  -2.370 0.017799 *  
StreamingTV_Yes                0.34596    0.11462   3.018 0.002542 ** 
Contract_Month.to.month        0.64666    0.16015   4.038 5.40e-05 ***
Contract_Two_year             -0.60469    0.25596  -2.362 0.018157 *  
PaymentMethod_Electronic_check  0.29539    0.10717   2.756 0.005849 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3373.1  on 2929  degrees of freedom
Residual deviance: 2467.7  on 2917  degrees of freedom
AIC: 2493.7
```

AIC: **2493.7**

Confusion Matrix:

	FALSE	TRUE
No	66%	7%
Yes	13%	14%

Accuracy**: 80%**
Sensitivity**: 65%**
Specificity**: 84%**

From the above model we can see that after dropping TotalCharges the overall accuracy has increased as well as the model deviances have a significant change.

Checked the VIFs again:
The VIFs good look good.

MODEL DIAGNOSTICS

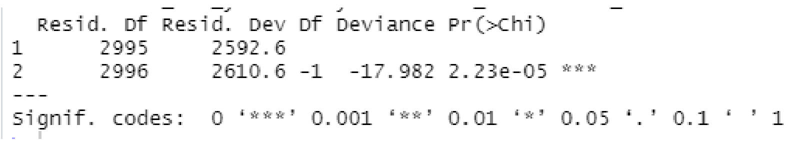
Model2

1. **Chi-square test for residual deviance**: P value: 0.99.
   Failed to reject Null Hypothesis which states that the model fits the data.
2. **Hosmer and Lemeshow goodness of fit (GOF) test:** P value: 0.1572
   Failed to reject Null Hypothesis which states that the model fits the data.
3. **PseudoR2**: McFadden: 0.2649119

Model3

4. **Chi-square test for residual deviance**: P value: 1
   Failed to reject Null Hypothesis which states that the model fits the data.
5. **Hosmer and Lemeshow goodness of fit (GOF) test:** P value: 0.6329
   Failed to reject Null Hypothesis which states that the model fits the data.
6. **PseudoR2**: McFadden: 0.2684265

Selecting Optimum model by using ANOVA.



```
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1      2995     2592.6
2      2996     2610.6 -1  -17.982 2.23e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Based on the output above, selecting Model 3 as statistically significant at p = 0.001. This suggests that model3 does provide an improved model fit.

Model Validation:

Evaluating Model3 on Test Data.

Confusion Matrix:

	FALSE	TRUE
No	66%	7%
Yes	14%	13%

Accuracy: **79%**
Sensitivity**: 66%**
Specificity**: 83%**

Q2. **Predictive Accuracy:**

LDA: Confusion Matrix for the Optimum model

	No	Yes
No	67%	8%
Yes	12%	13%

From the Confusion matrix we can see that 67% of predicted observations are true negatives and 13% are true positives. The type 2 error is 13% where the model predicts customer will not churn but they did. Type 1 error is 8% where model predicts that customer will churn but, they did not.

The predictive accuracy is **79%** which means that we successfully able to classify 79% of the predicted values. The best accuracy is at 50% threshold only because as we increase the probability the accuracy decreases.

LR: Confusion Matrix for the Optimum model

Confusion Matrix

	FALSE	TRUE
No	65%	8%
Yes	13%	14%

From the Confusion matrix we can see that 65% of predicted observations are true negatives and 14% are true positives. The type 2 error is 13% where the model predicts customer will not churn but they did. Type 1 error is 8% where model predicts that customer will churn but, they did not.

The predictive accuracy is ~**80%** which means that we successfully able to classify 80% of the predicted values. The best accuracy is at 50% threshold only because as we increase the probability the accuracy decreases.

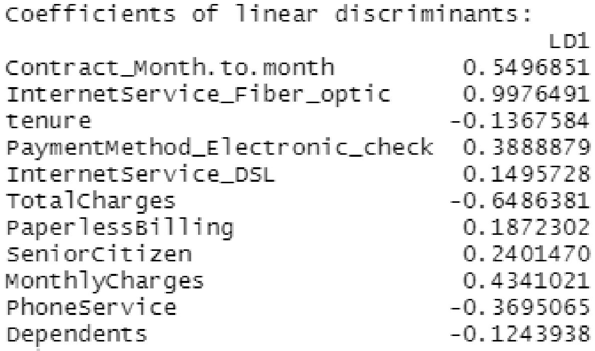
For both models we have observed that there is a class imbalance, due to this the models are biased toward negatives. This is because of the distribution of Churn in the dataset. The number of observations for Not Churn is way higher than Churn as result the model is able predict correctly for Not churn, but for Churn it is under predicting.

This problem can be solved by either getting more data points such that Churn and Not Churn is balanced or by using sampling techniques such as:

I. Oversampling minority class
II. Under sampling majority class.

Impact of Variables:

LDA:



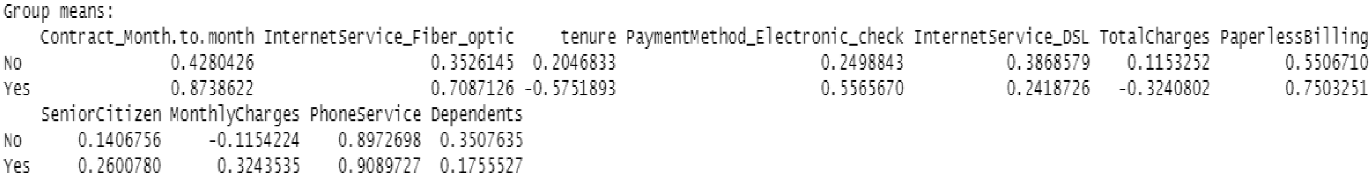
```
Coefficients of linear discriminants:
                                      LD1
Contract_Month.to.month         0.5496851
InternetService_Fiber_optic     0.9976491
tenure                         -0.1367584
PaymentMethod_Electronic_check  0.3888879
InternetService_DSL             0.1495728
TotalCharges                   -0.6486381
PaperlessBilling                0.1872302
SeniorCitizen                   0.2401470
MonthlyCharges                  0.4341021
PhoneService                   -0.3695065
Dependents                     -0.1243938
```

From the coefficients above we make some interpretations:

Order of important variables in terms of Magnitude in decreasing order:
Internet Service – Fiber Optic
Total Charges
Contract – Month-month
Monthly Charges
Payment Method – Electronic Check
Phone Service
Senior Citizen
Paperless Billing
Internet Service – DSL
Tenure
Dependents

Based on this we can see that Internet service – Fiber Optic has the largest impact on Customer Churn.
In terms of sign tenure, Phone Service, Total Charges and Dependents have a negative impact on Churn.

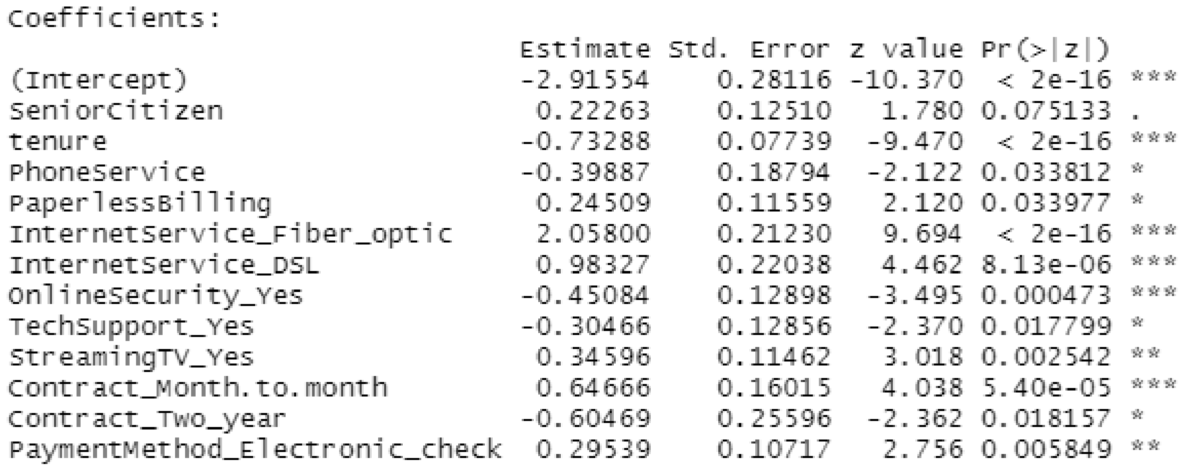
Group Means:



```
Group means:
    Contract_Month.to.month InternetService_Fiber_optic     tenure PaymentMethod_Electronic_check InternetService_DSL TotalCharges PaperlessBilling
No                0.4280426                   0.3526145  0.2046833                      0.2498843           0.3868579    0.1153252        0.5506710
Yes               0.8738622                   0.7087126 -0.5751893                      0.5565670           0.2418726   -0.3240802        0.7503251
    SeniorCitizen MonthlyCharges PhoneService Dependents
No      0.1406756     -0.1154224    0.8972698  0.3507635
Yes     0.2600780      0.3243535    0.9089727  0.1755527
```

In terms of group means we can see that there is a significant difference in groups for all variables hence each variable is able to classify well between churn and not churn.

LR:



```
Coefficients:
                              Estimate Std. Error z value Pr(>|z|)    
(Intercept)                   -2.91554    0.28116 -10.370  < 2e-16 ***
SeniorCitizen                  0.22263    0.12510   1.780 0.075133 .  
tenure                        -0.73288    0.07739  -9.470  < 2e-16 ***
PhoneService                  -0.39887    0.18794  -2.122 0.033812 *  
PaperlessBilling               0.24509    0.11559   2.120 0.033977 *  
InternetService_Fiber_optic    2.05800    0.21230   9.694  < 2e-16 ***
InternetService_DSL            0.98327    0.22038   4.462 8.13e-06 ***
OnlineSecurity_Yes            -0.45084    0.12898  -3.495 0.000473 ***
TechSupport_Yes               -0.30466    0.12856  -2.370 0.017799 *  
StreamingTV_Yes                0.34596    0.11462   3.018 0.002542 ** 
Contract_Month.to.month        0.64666    0.16015   4.038 5.40e-05 ***
Contract_Two_year             -0.60469    0.25596  -2.362 0.018157 *  
PaymentMethod_Electronic_check  0.29539    0.10717   2.756 0.005849 ** 
```

SeniorCitizen - If the customer is a senior citizen then the odds ratio or the likelihood of customer churning increases exponentially by the factor 0.22.

Tenure - Increase in tenure will decrease the odds ratio exponentially by the factor 0.73. So as tenure increases the change of churn will decrease.

PhoneService - If the customer has phone service then the odds ratio or the likelihood of customer churning decrease exponentially by the factor 0.39.

PaperlessBilling - If the customer has phone service then the odds ratio or the likelihood of customer churning increase exponentially by the factor 0.24.

InternetService_Fiber_Optic - If the customer fiber optic as the internet service then the odds ratio or the likelihood of customer churning increase exponentially by the factor 2.05. Such a high possibility of churn could be because fire optic service is very expensive hence customer usually moves to cheap internet service provider.

InternetService_DSL - If the customer fiber optic as the internet service then the odds ratio or the likelihood of customer churning increase exponentially by the factor 0.98, less than that of fiber optics.

OnlineSecurity_Yes - If the customer has online security then the odds ratio or the likelihood of customer churning will decrease exponentially by the factor 0.45.

TechSupport_Yes - If the customer has tech support then the odds ratio or the likelihood of customer churning will decrease exponentially by the factor 0.3.

StreamingTv_Yes - If the customer has streaming TV then the odds ratio or the likelihood of customer churning increase exponentially by the factor 0.34.

Contract_Month_to_month - If the customer has opted for monthly contract then the odds ratio or the likelihood of customer churning increase exponentially by the factor 0.64.
Contract_2_Year - If the customer has opted for a 2 year contract then the odds ratio or the likelihood of customer churning decrease exponentially by the factor 0.60.

PaymentMethod_Electronic_check - If the customer has opted for a electronic check mode of payment then the odds ratio or the likelihood of customer churning increase exponentially by the factor 0.29.

Q3. Created new variable Customer_value_segment based on Total Charges.

This variable has 3 classes:
Low - Total Charges < 398.900

Medium - 398.900 < Total Charges < 3844

High - Total Charges > 3844

As the dependent variable is not a binary response variable hence we cannot use Logistic regression therefore we use LDA for classification.

Converting Churn into binary response variable 0,1
Converting categorical variables with 2 classes into 0,1 and other categorical variables with more than 2 classes as dummy variables with n-1 classes.

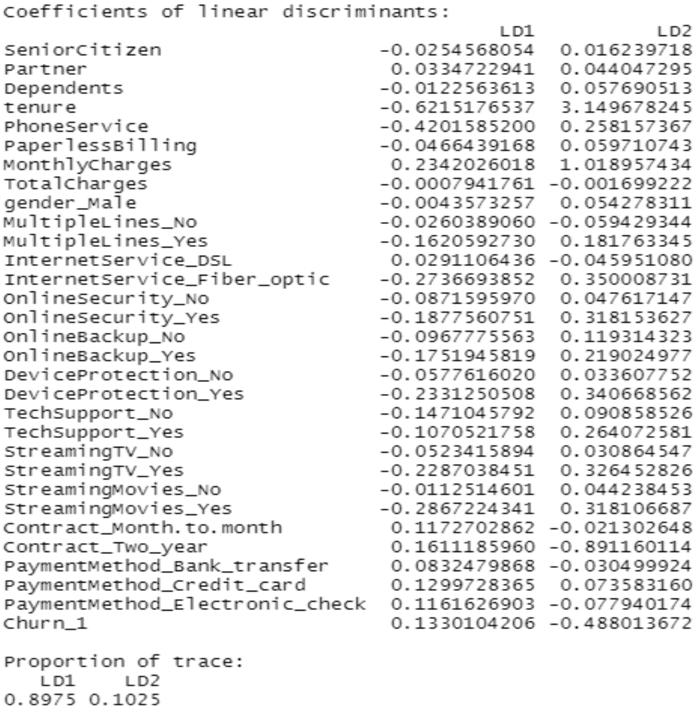
“Partner”, “Dependents”, “PhoneService”, “PaperlessBilling”,

“MultipleLines" "InternetService" "OnlineSecurity" "OnlineBackup" "DeviceProtection" "TechSupport" "StreamingTV" "StreamingMovies" “PaymentMethod”

Standardizing Tenure and Monthly Charges.
Splitting dataset in test and train

Running the model with all variables on train:

Iteration 1



```
Coefficients of linear discriminants:
                                      LD1          LD2
SeniorCitizen                 -0.0254568054  0.016239718
Partner                        0.0334722941  0.044047295
Dependents                    -0.0122563613  0.057690513
tenure                        -0.6215176537  3.149678245
PhoneService                  -0.4201585200  0.258157367
PaperlessBilling              -0.0466439168  0.059710743
MonthlyCharges                 0.2342026018  1.018957434
TotalCharges                  -0.0007941761 -0.001699222
gender_Male                   -0.0043573257  0.054278311
MultipleLines_No              -0.0260389060 -0.059429344
MultipleLines_Yes             -0.1620592730  0.181763345
InternetService_DSL            0.0291106436 -0.045951080
InternetService_Fiber_optic   -0.2736693852  0.350008731
OnlineSecurity_No             -0.0871595970  0.047617147
OnlineSecurity_Yes            -0.1877560751  0.318153627
OnlineBackup_No               -0.0967775563  0.119314323
OnlineBackup_Yes              -0.1751945819  0.219024977
DeviceProtection_No           -0.0577616020  0.033607752
DeviceProtection_Yes          -0.2331250508  0.340668562
TechSupport_No                -0.1471045792  0.090858526
TechSupport_Yes               -0.1070521758  0.264072581
StreamingTV_No                -0.0523415894  0.030864547
StreamingTV_Yes               -0.2287038451  0.326452826
StreamingMovies_No            -0.0112514601  0.044238453
StreamingMovies_Yes           -0.2867224341  0.318106687
Contract_Month.to.month        0.1172702862 -0.021302648
Contract_Two_year              0.1611185960 -0.891160114
PaymentMethod_Bank_transfer    0.0832479868 -0.030499924
PaymentMethod_Credit_card      0.1299728365  0.073583160
PaymentMethod_Electronic_check 0.1161626903 -0.077940174
Churn_1                        0.1330104206 -0.488013672

Proportion of trace:
   LD1    LD2
0.8975 0.1025
```

Prior probabilities of groups:

high	low	medium
0.2524950	0.2498337	0.4976713

From the prior probabilities we can see that 49% of training observation fall under medium segment, the rest of the observations have almost the same probabilities.

Confusion Matrix:

	high	low	medium
high	25%	0%	0%
low	0%	22%	3%
medium	0%	5%	44%

From the matrix we can see 25% of predicted values are true positives for high, 22% are for low and 44% are of medium.

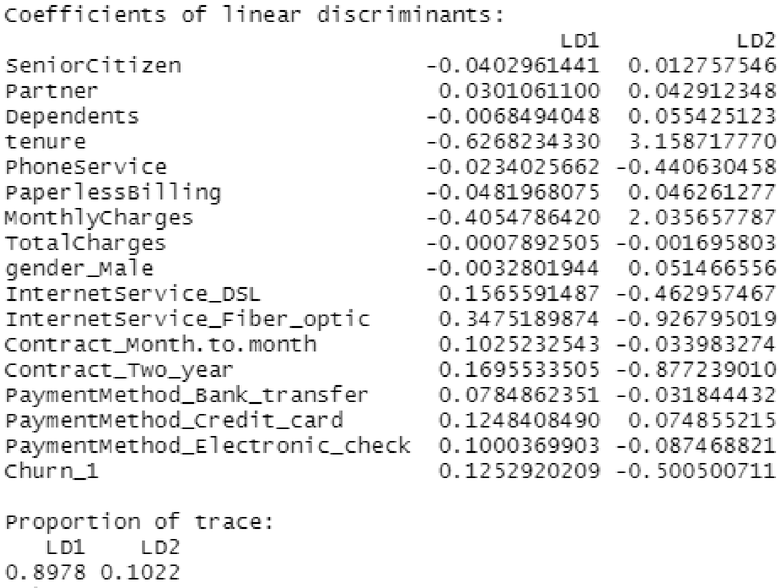
Model Accuracy: **0.9184963**
Error Rate: **0.08150366**

This model is not optimum since there is collinearity among variables.

Removing correlated variables using Chi-square test.

Variables dropped:
MultipleLines, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies

Running model again
Iteration2



```
Coefficients of linear discriminants:
                                      LD1          LD2
SeniorCitizen                -0.0402961441  0.012757546
Partner                       0.0301061100  0.042912348
Dependents                   -0.0068494048  0.055425123
tenure                       -0.6268234330  3.158717770
PhoneService                 -0.0234025662 -0.440630458
PaperlessBilling             -0.0481968075  0.046261277
MonthlyCharges               -0.4054786420  2.035657787
TotalCharges                 -0.0007892505 -0.001695803
gender_Male                  -0.0032801944  0.051466556
InternetService_DSL           0.1565591487 -0.462957467
InternetService_Fiber_optic   0.3475189874 -0.926795019
Contract_Month.to.month       0.1025232543 -0.033983274
Contract_Two_year             0.1695533505 -0.877239010
PaymentMethod_Bank_transfer   0.0784862351 -0.031844432
PaymentMethod_Credit_card     0.1248408490  0.074855215
PaymentMethod_Electronic_check 0.1000369903 -0.087468821
Churn_1                       0.1252920209 -0.500500711

Proportion of trace:
   LD1    LD2
0.8978 0.1022
```

Confusion Matrix:

	high	low	medium
high	25%	0%	0%
low	0%	22%	3%
medium	0%	5%	44%

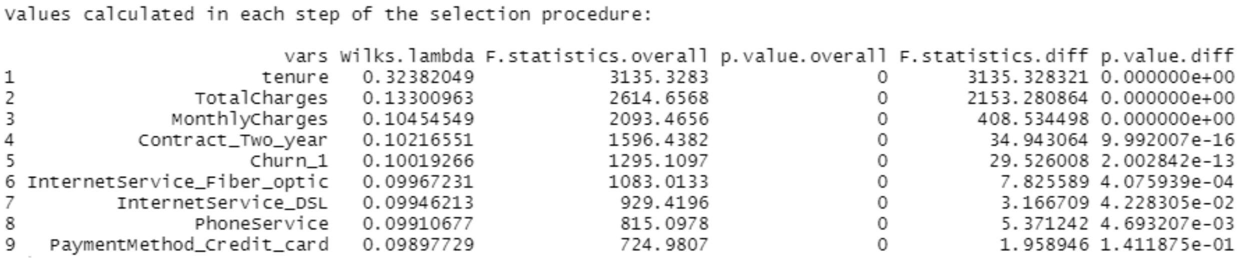
From the matrix we can see 25% of predicted values are true positives for high, 22% are for low and 44% are of medium.

Model Accuracy: **0.917831**
Error Rate: **0.082169**

The model is almost the same as model 1 in terms of accuracy but the collinearity has been removed.

Feature selection using Greedy Wilks Lambda

After running greedy.wilks for the variables selected in model 2, we get a list of significant variables which will minimize the overall lambda for the model.

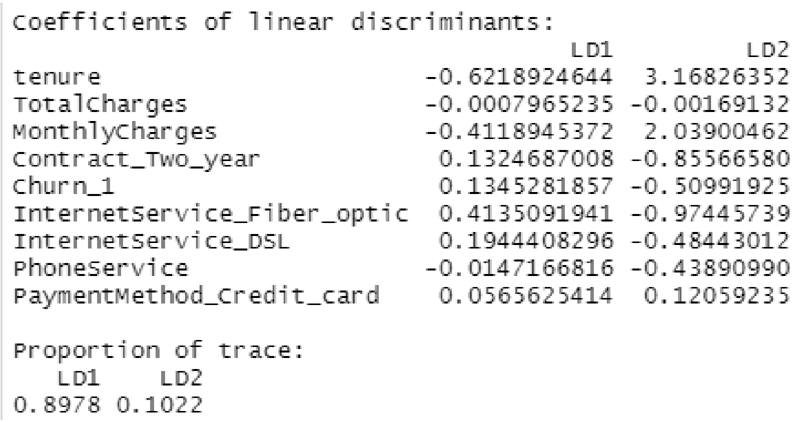


```
Values calculated in each step of the selection procedure:

                       vars Wilks.lambda F.statistics.overall p.value.overall F.statistics.diff p.value.diff
1                    tenure   0.32382049            3135.3283               0       3135.328321 0.000000e+00
2              TotalCharges   0.13300963            2614.6568               0       2153.280864 0.000000e+00
3            MonthlyCharges   0.10454549            2093.4656               0        408.534498 0.000000e+00
4         Contract_Two_year   0.10216551            1596.4382               0         34.943064 9.992007e-16
5                   Churn_1   0.10019266            1295.1097               0         29.526008 2.002842e-13
6 InternetService_Fiber_optic   0.09967231            1083.0133               0          7.825589 4.075939e-04
7        InternetService_DSL   0.09946213             929.4196               0          3.166709 4.228305e-02
8              PhoneService   0.09910677             815.0978               0          5.371242 4.693207e-03
9   PaymentMethod_Credit_card   0.09897729             724.9807               0          1.958946 1.411875e-01
```

Running the model against these variables

Iteration3



```
Coefficients of linear discriminants:
                                      LD1         LD2
tenure                      -0.6218924644  3.16826352
TotalCharges                -0.0007965235 -0.00169132
MonthlyCharges              -0.4118945372  2.03900462
Contract_Two_year            0.1324687008 -0.85566580
Churn_1                      0.1345281857 -0.50991925
InternetService_Fiber_optic  0.4135091941 -0.97445739
InternetService_DSL          0.1944408296 -0.48443012
PhoneService                -0.0147166816 -0.43890990
PaymentMethod_Credit_card    0.0565625414  0.12059235

Proportion of trace:
   LD1    LD2
0.8978 0.1022
```

Confusion Matrix:

	high	low	medium
high	25%	0%	0%
low	0%	22%	3%
medium	0%	5%	45%

From the matrix we can see 25% of predicted values are true positives for high, 22% are for low and 45% are of medium.

Model Accuracy: **0.919827**
Error Rate: **0.082169**

As we can see Accuracy has improved slightly so we can stop here.

Model Diagnostics:

Model2

Hotelling's T square :
Test stat: 1930.5
Numerator df: 17
Denominator df: 1492
P-value: 0
Wilks Lambda
0.098611

Model3

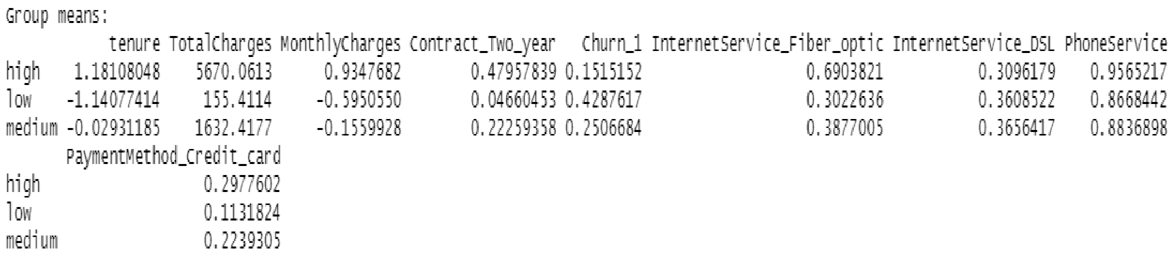
Hotelling's T square :
Test stat: 3642
Numerator df: 9
Denominator df: 1500
P-value: 0
Wilks Lambda
0.098977

Based on our diagnostic tests we have seen that both Models have significant P values for Hotelling's T square and Wilks Lambda.
However, the Hotelling test statistic is greater for Model 3 and Wilks Lambda value is approximately the same.

Based on above assumptions we have selected Model 3 as the optimum model.

Profiling of Customer Segments category:

Group means:



```
Group means:
            tenure TotalCharges MonthlyCharges Contract_Two_year   Churn_1 InternetService_Fiber_optic InternetService_DSL PhoneService
high    1.18108048    5670.0613      0.9347682        0.47957839 0.1515152                   0.6903821          0.3096179    0.9565217
low    -1.14077414     155.4114     -0.5950550        0.04660453 0.4287617                   0.3022636          0.3608522    0.8668442
medium -0.02931185    1632.4177     -0.1559928        0.22259358 0.2506684                   0.3877005          0.3656417    0.8836898
       PaymentMethod_Credit_card
high                   0.2977602
low                    0.1131824
medium                 0.2239305
```

Customer Segment HIGH
Tenure, TotalCharges, MonthlyCharges, Contract_two_years, Internet_service_fiber optic, phone service and PaymentMethod_Credit_Card have the highest average values for this category.

Customer Segment MEDIUM
InternetService_DSL has the highest average value for this segment.

Customer Segment LOW
Churn has highest average value for this segment.

As we can see that customer with high tenure are mostly classified in High segments. In order to shift customers from low to high categories some benefits need to be provided like discount and other offers to keep them engaged such that they stay longer with the service.

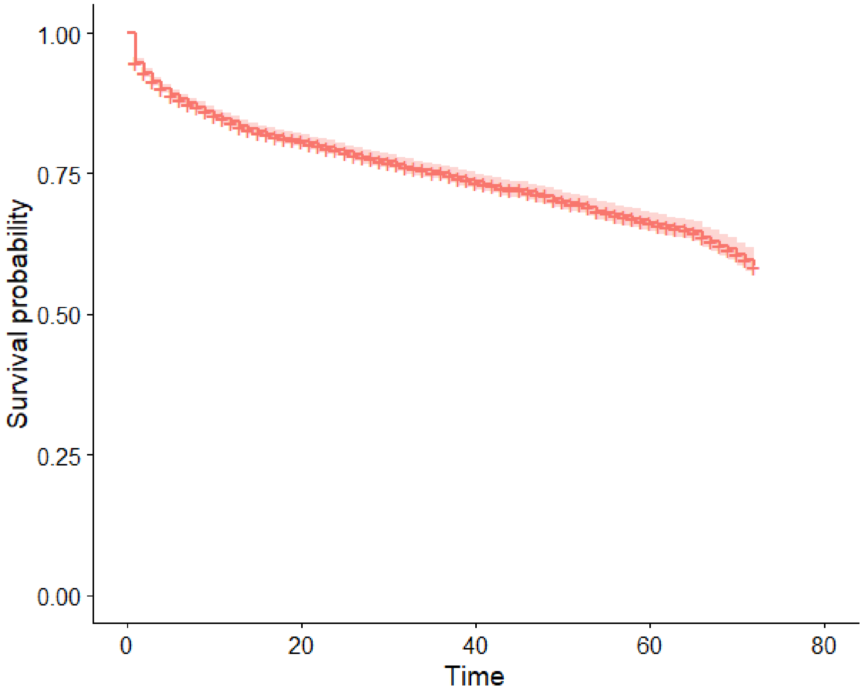
These discounts could be for internet services such as Fiber Optics. Since this service is expensive than DSL therefore most customers stay with medium category as HIGH segment customers are usually better in terms of finances.

In terms of contract we need to shift customers from Monthly to Annual contract this will directly help our tenure issue. The reason why HIGH customer have a high tenure is because they usually have annual contracts.

For phone services offers should be provided such as free calling for few months or internet add-ons.

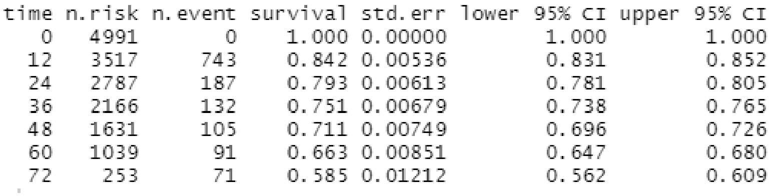
Customers of LOW category usually have higher Churn rate compared to HIGH and Medium. Availing above offers might help reduce the overall churn rate as well as shifting customers from one segment to other.

Q4. Survival Curve



From the curve we can see the survival percentage is 100% at tenure T = 0.
After 30 months about 25% of customers have churned.

The half life of this curve occurs somewhere at 70months which means that we are losing 50% of customers at time T = 70 months.

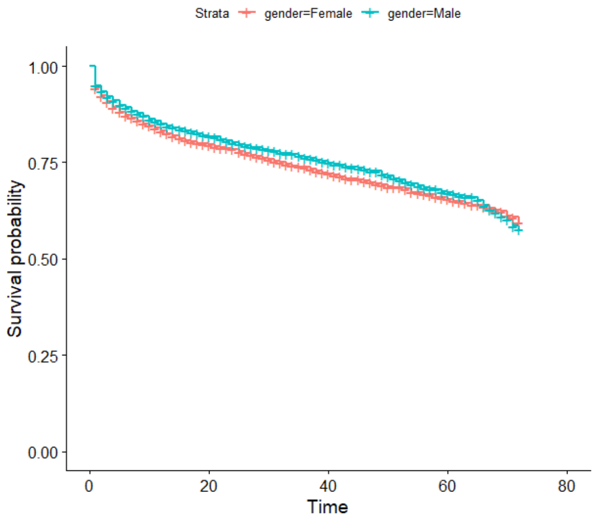


```
time n.risk n.event survival std.err lower 95% CI upper 95% CI
   0   4991       0    1.000 0.00000        1.000        1.000
  12   3517     743    0.842 0.00536        0.831        0.852
  24   2787     187    0.793 0.00613        0.781        0.805
  36   2166     132    0.751 0.00679        0.738        0.765
  48   1631     105    0.711 0.00749        0.696        0.726
  60   1039      91    0.663 0.00851        0.647        0.680
  72    253      71    0.585 0.01212        0.562        0.609
```

As we can see that survival rate is 84% at T=12 which is the highest(not including T=0)
After 24 months we can see that for every 12-month cycle the survival rate is decreasing further. This could be because after every year after contract ends some of the customers move towards a cheaper service provider.

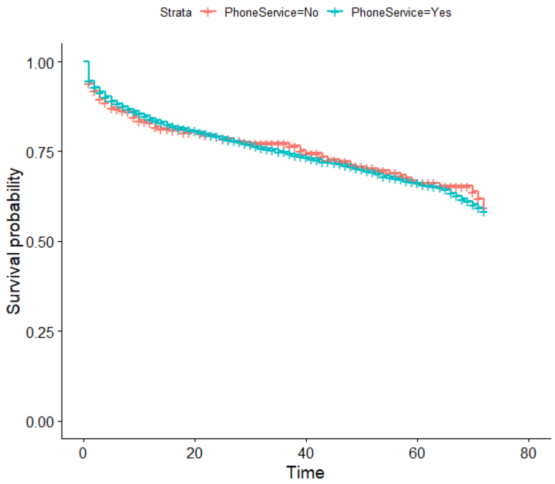
Q5.

Survival Curve for Gender:



There is not much difference between Male and Female. However, Males have a slightly higher survival rate for initial months but after 70months we can see that females take the lead in survival percentage.

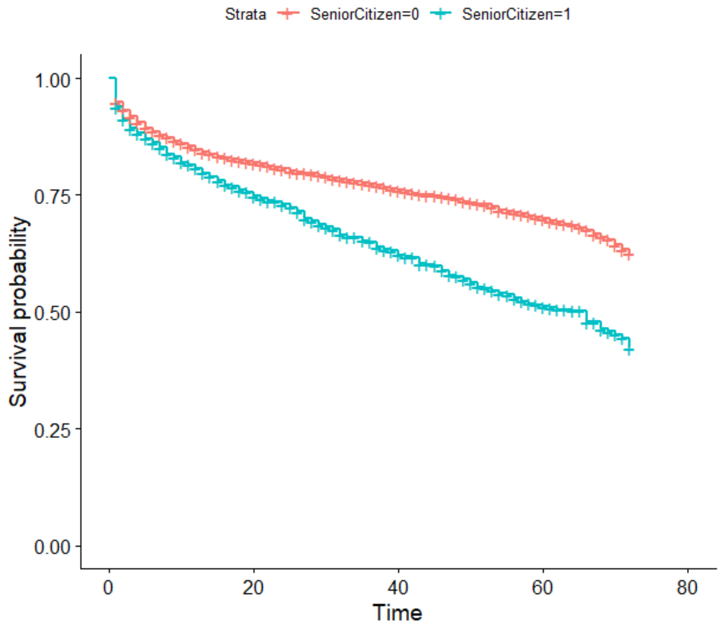
Survival Curve for PhoneService:



Customers with and without phone service don’t have much difference in probabilities. For initial 20 months customers with phone service have slightly higher probability but this gradually declines as we move away from 20 months. It is interesting to see that at tenure 30 and 70 both groups have almost the survival rate.

Phone service customers have gradual decline throughout and after 60months a steep decline. For non-phone service customers, the curve becomes constant for 30-40months and 60-70months.

Survival Curve for SeniorCitizen:

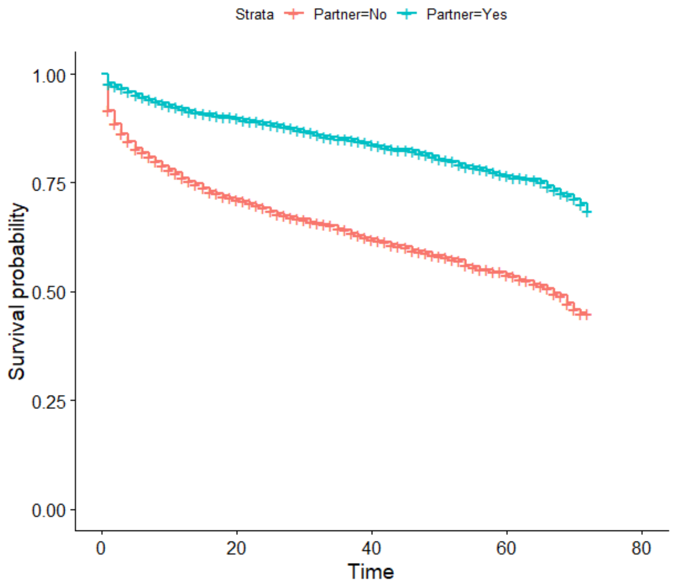


Senior Citizen have a low survival probability compared to non-senior citizen.

75% of customers who were senior citizens survived at tenure = 20 months but for other customers the tenure was about 40months.

One of the reason for low probability of senior citizens who be their health condition since they choose to pay more towards medical facilities rather than telecom providers.

Survival Curve for Partners

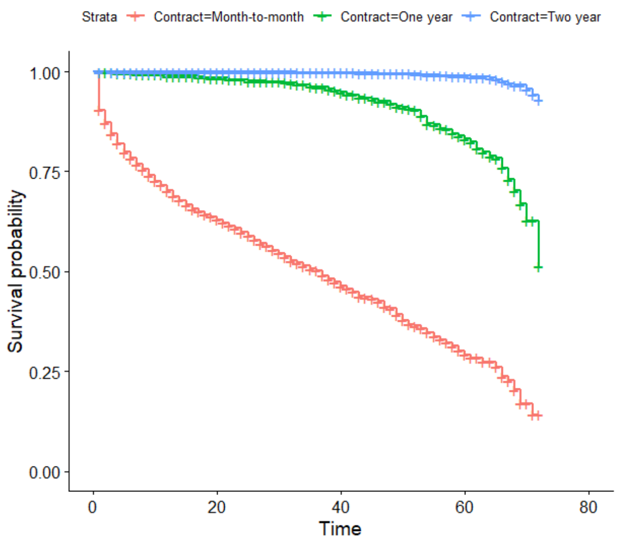


Customers who have partners have a high survival rate compared to customers who do not have partners.

As we can see from the curve 25% of customers without partners are lost in the initial 15months while the 25% of customers with partners are lost after 60months.

This behavior could be because partners may be using the same service providers such as Online Streaming eg. Husband and wife as they will be living together.

Survival Curve for Contract:



There is massive difference between different mode of contracts. As we can see that customers with a 2 year contract have the highest survival rate, followed by 1 year and monthly being the least.

For monthly contract we are losing 50% of customers in a span of 30months.

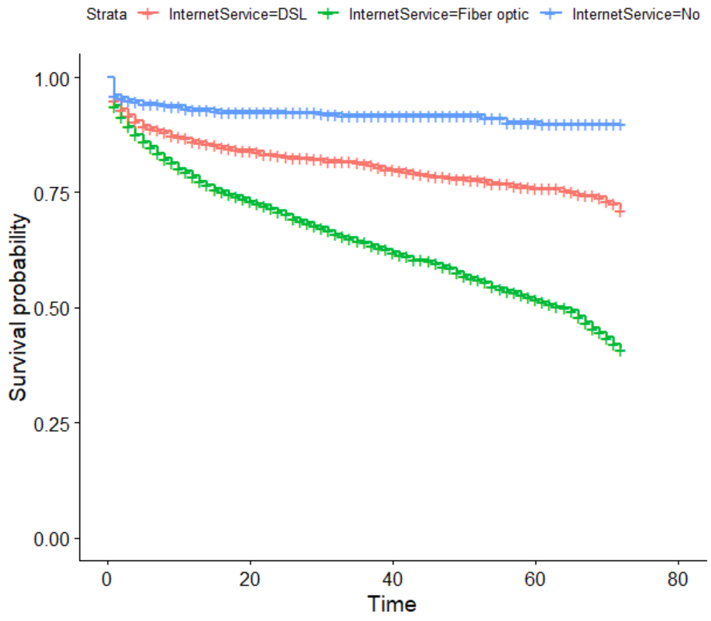
This is because of many reasons:

1.Customer has lost interest in the service and decided to move to a cheaper provider.

2.The customer did not know what he/she was getting into and after initial months decided that the service was suitable for him.

Customers that have 2-year contracts are usually loyal customers who are retained by the company hence have the highest survival probability.

Survival Curve for Internet Service:



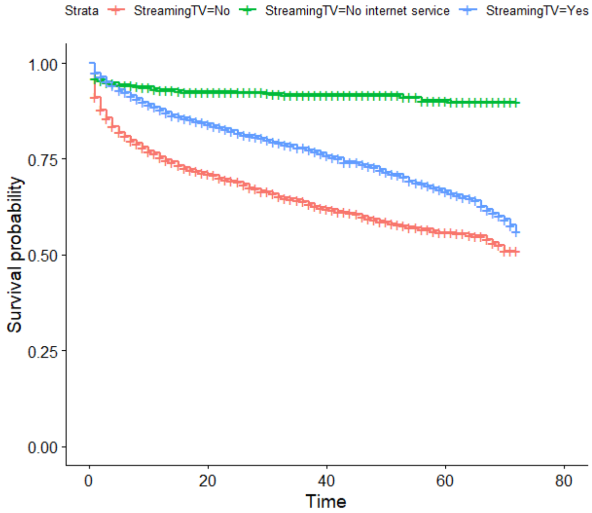
As we can see customers having no internet services have the highest survival rate followed by DSL and fiber optic being the least.

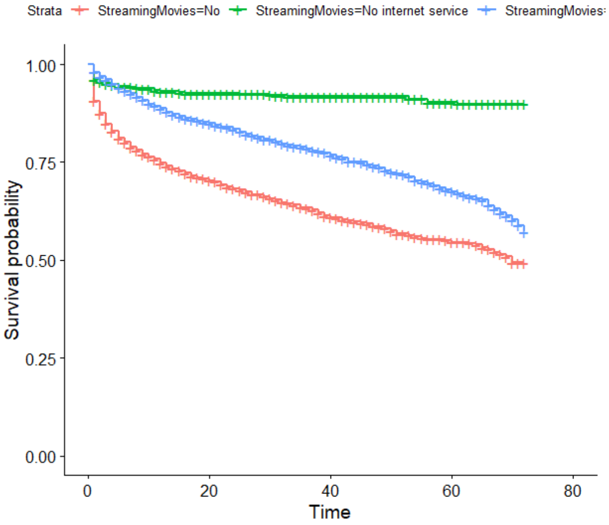
Customers who have fiber optic as a service have the lowest curve because the service is very expensive hence they choose opt out and to some cheaper service available.

Customers having DSL service usually belong to the medium customer segment as the service is not as expensive as fiber optic.

For fiber optic service we can see that half-life is around 60months.

Survival Curve for Streaming TV and Streaming Movies:

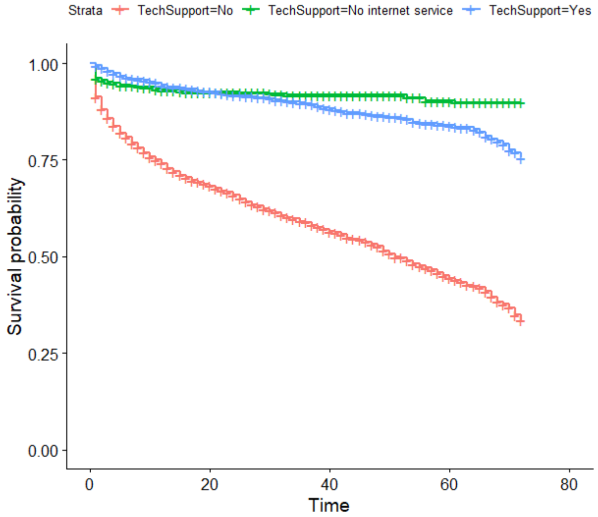


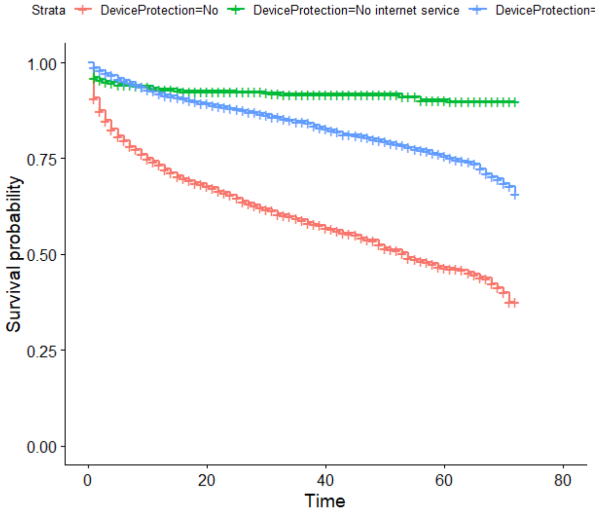


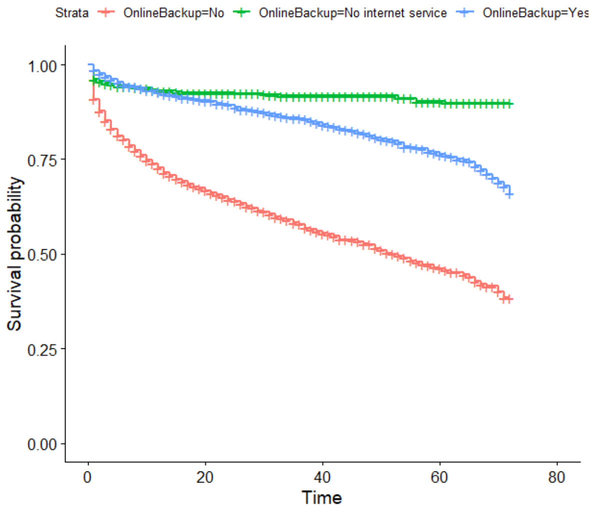
As we can see the survival curves for both of them are more or less the same with customers having streaming services having medium survival rate.

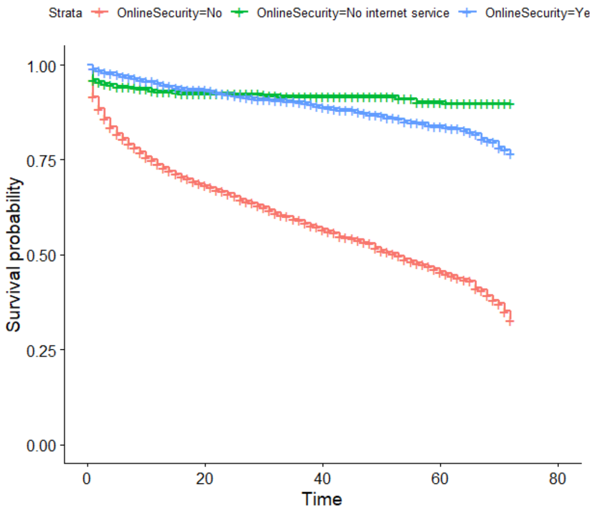
Customers who do not have this services have the lowest survival rate and customers with no internet service have the highest rate.

Survival Curve for Tech Support, Device Protection, Online Backup and Online Security:







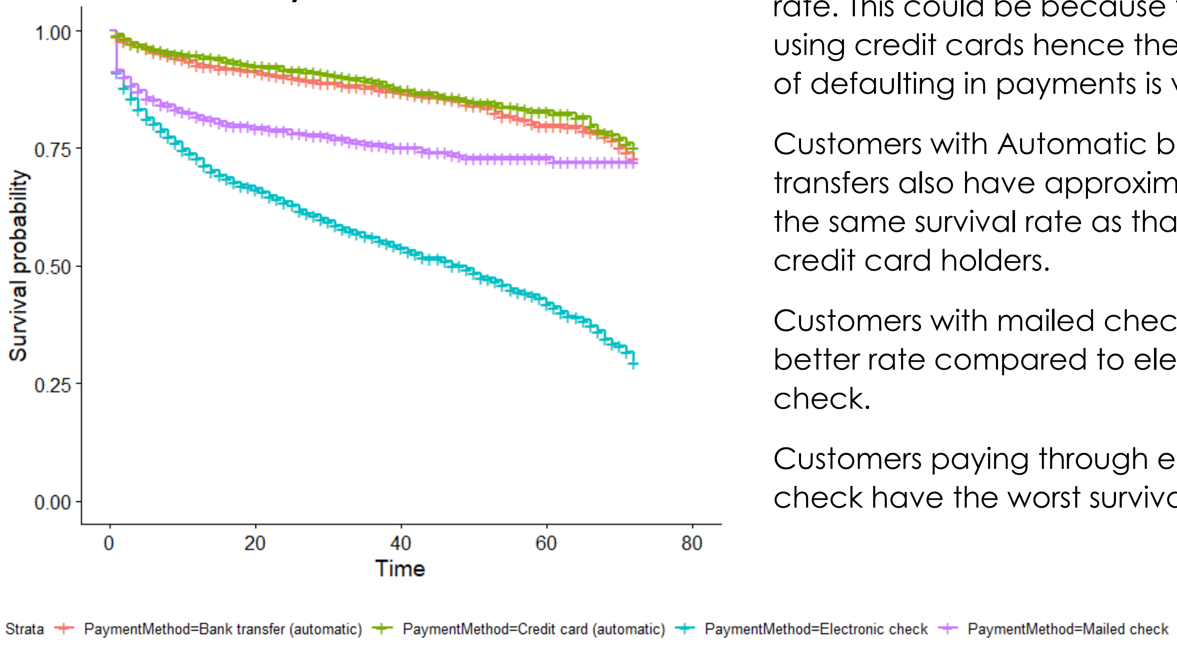


As we can see the survival curves for all of them are the same with customers having online services having medium survival rate.

Customers who do not have this services have the lowest survival rate and customers with no internet service have the highest rate.

Customer with No Internet service have the same survival probability through the variables.

Survival Curve for Payment Method:



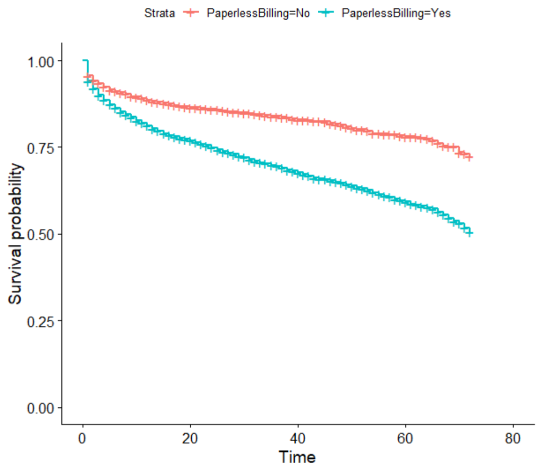
As we can see that customers who choose credit cards as mode of payment have the highest survival rate. This could be because they are using credit cards hence the chance of defaulting in payments is very less.

Customers with Automatic bank transfers also have approximately the same survival rate as that of credit card holders.

Customers with mailed check have a better rate compared to electronic check.

Customers paying through electronic check have the worst survival rate.

Survival Curve for Paperless Billing:

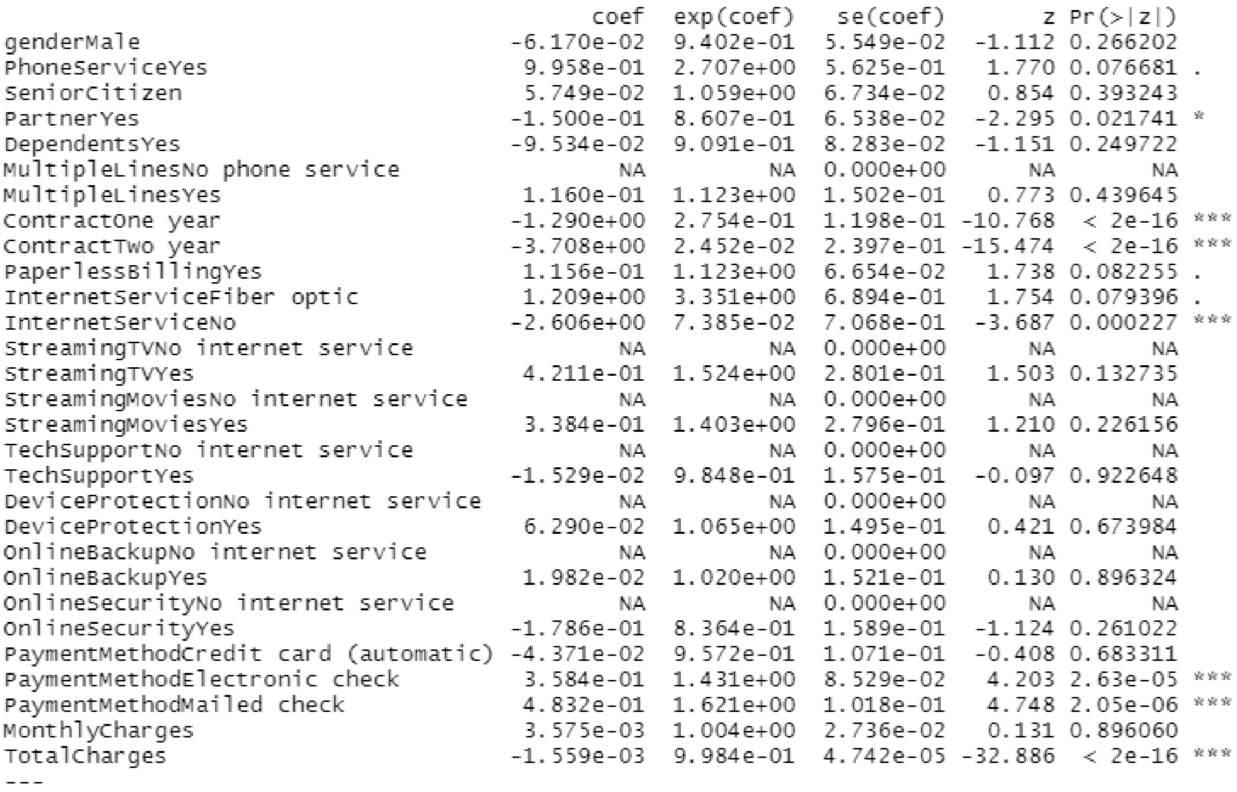


Customers who didn’t have paperless billing have a higher survival rate compared to customers who have Paperless billing.

Q6. COX Hazard Model

Let us fit Cox Proportional Hazard Model using all variables.

Iteration 1



```
                                           coef exp(coef)  se(coef)       z Pr(>|z|)    
genderMale                           -6.170e-02 9.402e-01 5.549e-02  -1.112 0.266202    
PhoneServiceYes                       9.958e-01 2.707e+00 5.625e-01   1.770 0.076681 .  
SeniorCitizen                         5.749e-02 1.059e+00 6.734e-02   0.854 0.393243    
PartnerYes                           -1.500e-01 8.607e-01 6.538e-02  -2.295 0.021741 *  
DependentsYes                        -9.534e-02 9.091e-01 8.283e-02  -1.151 0.249722    
MultipleLinesNo phone service                NA        NA 0.000e+00      NA       NA    
MultipleLinesYes                      1.160e-01 1.123e+00 1.502e-01   0.773 0.439645    
ContractOne year                     -1.290e+00 2.754e-01 1.198e-01 -10.768  < 2e-16 ***
ContractTwo year                     -3.708e+00 2.452e-02 2.397e-01 -15.474  < 2e-16 ***
PaperlessBillingYes                   1.156e-01 1.123e+00 6.654e-02   1.738 0.082255 .  
InternetServiceFiber optic            1.209e+00 3.351e+00 6.894e-01   1.754 0.079396 .  
InternetServiceNo                    -2.606e+00 7.385e-02 7.068e-01  -3.687 0.000227 ***
StreamingTVNo internet service               NA        NA 0.000e+00      NA       NA    
StreamingTVYes                        4.211e-01 1.524e+00 2.801e-01   1.503 0.132735    
StreamingMoviesNo internet service           NA        NA 0.000e+00      NA       NA    
StreamingMoviesYes                    3.384e-01 1.403e+00 2.796e-01   1.210 0.226156    
TechSupportNo internet service               NA        NA 0.000e+00      NA       NA    
TechSupportYes                       -1.529e-02 9.848e-01 1.575e-01  -0.097 0.922648    
DeviceProtectionNo internet service          NA        NA 0.000e+00      NA       NA    
DeviceProtectionYes                   6.290e-02 1.065e+00 1.495e-01   0.421 0.673984    
OnlineBackupNo internet service              NA        NA 0.000e+00      NA       NA    
OnlineBackupYes                       1.982e-02 1.020e+00 1.521e-01   0.130 0.896324    
OnlineSecurityNo internet service            NA        NA 0.000e+00      NA       NA    
OnlineSecurityYes                    -1.786e-01 8.364e-01 1.589e-01  -1.124 0.261022    
PaymentMethodCredit card (automatic) -4.371e-02 9.572e-01 1.071e-01  -0.408 0.683311    
PaymentMethodElectronic check         3.584e-01 1.431e+00 8.529e-02   4.203 2.63e-05 ***
PaymentMethodMailed check             4.832e-01 1.621e+00 1.018e-01   4.748 2.05e-06 ***
MonthlyCharges                        3.575e-03 1.004e+00 2.736e-02   0.131 0.896060    
TotalCharges                         -1.559e-03 9.984e-01 4.742e-05 -32.886  < 2e-16 ***
---
```

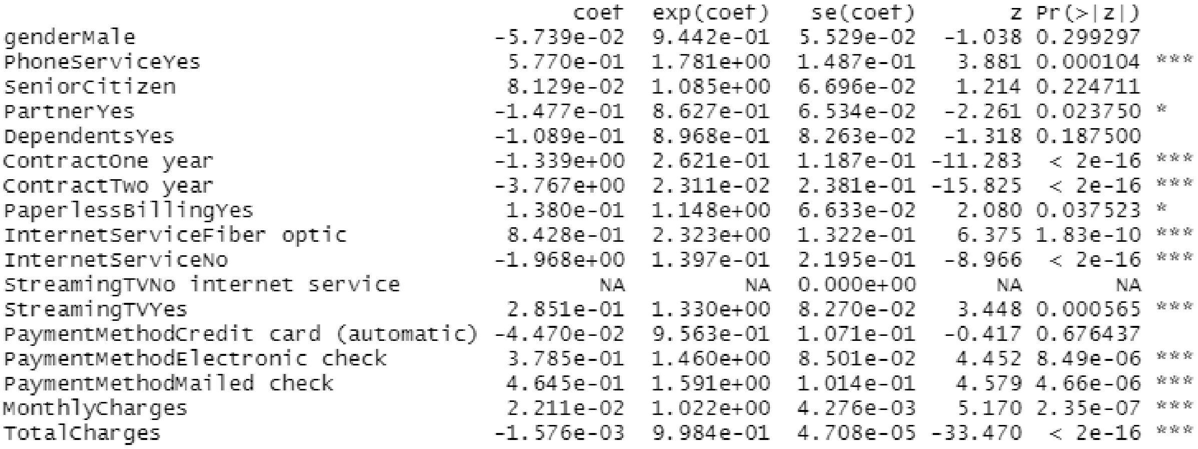
N = 4991
Events (Churn=1) = 1329
Concordance= **0.926** (se = 0.003)
Likelihood ratio test= **4194** on 22 df, p=<2e-16
Wald test = **1826** on 22 df, p=<2e-16
Score (logrank) test = **3242** on 22 df, p=<2e-16

The column exp(coef) gives us the exponentiated coefficient which gives us the hazard ratio.

Internet service: Fiber Optic and Contract Two Year has the highest hazard ratio in terms of magnitude but, hazard ratio is increasing for Fiber Optic and decreasing for Contract Two Year.

From the model above we can see there some insignificant variables few NAs which is mainly because of the collinearity in the data.

Dropping strongly correlated variables:
Iteration 2

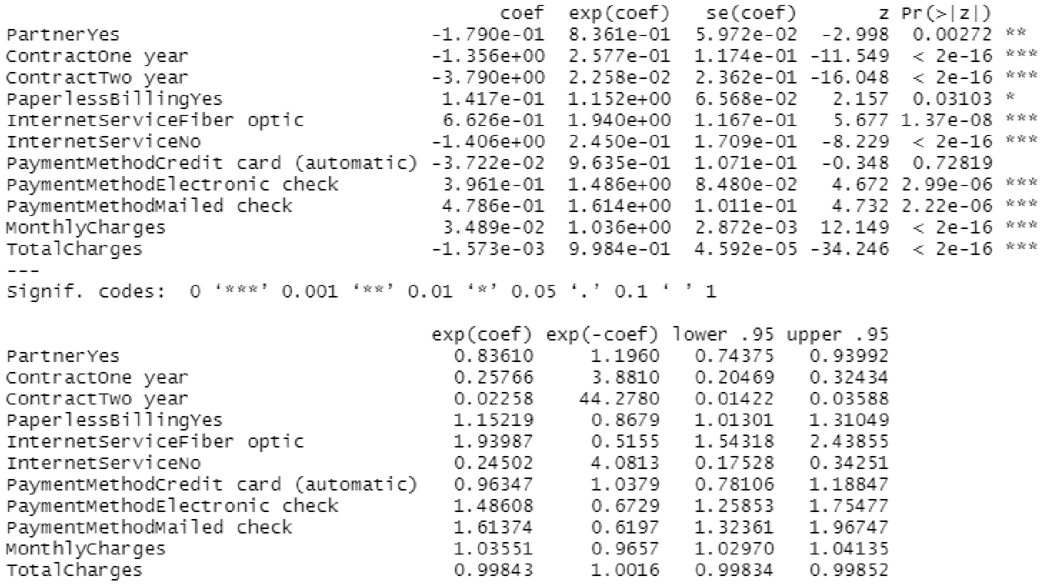


```
                                         coef  exp(coef)  se(coef)       z Pr(>|z|)    
genderMale                         -5.739e-02  9.442e-01  5.529e-02  -1.038 0.299297    
PhoneServiceYes                     5.770e-01  1.781e+00  1.487e-01   3.881 0.000104 ***
SeniorCitizen                       8.129e-02  1.085e+00  6.696e-02   1.214 0.224711    
PartnerYes                         -1.477e-01  8.627e-01  6.534e-02  -2.261 0.023750 *  
DependentsYes                      -1.089e-01  8.968e-01  8.263e-02  -1.318 0.187500    
ContractOne year                   -1.339e+00  2.621e-01  1.187e-01 -11.283  < 2e-16 ***
ContractTwo year                   -3.767e+00  2.311e-02  2.381e-01 -15.825  < 2e-16 ***
PaperlessBillingYes                 1.380e-01  1.148e+00  6.633e-02   2.080 0.037523 *  
InternetServiceFiber optic          8.428e-01  2.323e+00  1.322e-01   6.375 1.83e-10 ***
InternetServiceNo                  -1.968e+00  1.397e-01  2.195e-01  -8.966  < 2e-16 ***
StreamingTVNo internet service             NA         NA  0.000e+00      NA       NA    
StreamingTVYes                      2.851e-01  1.330e+00  8.270e-02   3.448 0.000565 ***
PaymentMethodCredit card (automatic) -4.470e-02  9.563e-01  1.071e-01  -0.417 0.676437    
PaymentMethodElectronic check       3.785e-01  1.460e+00  8.501e-02   4.452 8.49e-06 ***
PaymentMethodMailed check           4.645e-01  1.591e+00  1.014e-01   4.579 4.66e-06 ***
MonthlyCharges                      2.211e-02  1.022e+00  4.276e-03   5.170 2.35e-07 ***
TotalCharges                       -1.576e-03  9.984e-01  4.708e-05 -33.470  < 2e-16 ***
```

Concordance= 0.926 (se = 0.003)
Likelihood ratio test= **4169** on 15 df, p=<2e-16
Wald test = **1808** on 15 df, p=<2e-16
Score (logrank) test = **3220** on 15 df, p=<2e-16

Still we are getting some insignificant variables and NA. Let us select only those variables which had a significant difference between groups as seen in the survival curves.

Iteration3



```
                                   coef  exp(coef)   se(coef)       z Pr(>|z|)    
PartnerYes                   -1.790e-01  8.361e-01  5.972e-02  -2.998  0.00272 ** 
ContractOne year             -1.356e+00  2.577e-01  1.174e-01 -11.549  < 2e-16 ***
ContractTwo year             -3.790e+00  2.258e-02  2.362e-01 -16.048  < 2e-16 ***
PaperlessBillingYes           1.417e-01  1.152e+00  6.568e-02   2.157  0.03103 *  
InternetServiceFiber optic    6.626e-01  1.940e+00  1.167e-01   5.677 1.37e-08 ***
InternetServiceNo            -1.406e+00  2.450e-01  1.709e-01  -8.229  < 2e-16 ***
PaymentMethodCredit card (automatic) -3.722e-02  9.635e-01  1.071e-01  -0.348  0.72819    
PaymentMethodElectronic check   3.961e-01  1.486e+00  8.480e-02   4.672 2.99e-06 ***
PaymentMethodMailed check     4.786e-01  1.614e+00  1.011e-01   4.732 2.22e-06 ***
MonthlyCharges                3.489e-02  1.036e+00  2.872e-03  12.149  < 2e-16 ***
TotalCharges                 -1.573e-03  9.984e-01  4.592e-05 -34.246  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

                                     exp(coef) exp(-coef) lower .95 upper .95
PartnerYes                             0.83610     1.1960   0.74375   0.93992
ContractOne year                       0.25766     3.8810   0.20469   0.32434
ContractTwo year                       0.02258    44.2780   0.01422   0.03588
PaperlessBillingYes                    1.15219     0.8679   1.01301   1.31049
InternetServiceFiber optic             1.93987     0.5155   1.54318   2.43855
InternetServiceNo                      0.24502     4.0813   0.17528   0.34251
PaymentMethodCredit card (automatic)   0.96347     1.0379   0.78106   1.18847
PaymentMethodElectronic check          1.48608     0.6729   1.25853   1.75477
PaymentMethodMailed check              1.61374     0.6197   1.32361   1.96747
MonthlyCharges                         1.03551     0.9657   1.02970   1.04135
TotalCharges                           0.99843     1.0016   0.99834   0.99852
```

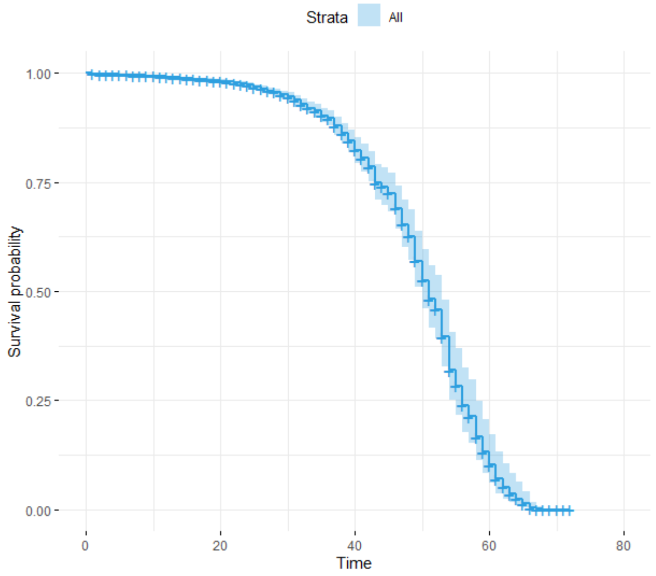
As we can see NAs have been removed and all of the variables are now significant except Payment Method Credit Card.

Testing for best model on the basis of Likelihood Ratio test:

Model 2: loglik: -8575.2
Model3: loglik: -8586.7 p values: 0.0003412 ***

As per the test Model3 is the optimum model.

Survival Curve for Optimum Cox Hazard model:



As we can see this is a smooth curve with half-life coming at 50months.

Maximum survival rate is for initial 15 to 20 months.

Interpretation of Coefficients:

Partner_Yes, exp(coef) = **0.836**
Hazard probability: **-16%**
This means that if a customer has a partner then the likelihood of churning will decrease by 16%.

ContractOne year, exp(coef) = **0.25766**
Hazard probability: **-74%**
For a customer with one-year contract plan, the likelihood of churning will reduce by 74%.

ContractTwo year, exp(coef) = **0.02258**
Hazard probability: **-98%**
For a customer with two-year contract plan, the likelihood of churning will reduce by 98%.

PaperlessBillingYes, exp(coef) = **1.15219**
Hazard probability: **15%**
For a customer with paperless billing plan, the likelihood of churning will increase by 15%.

InternetServiceFiber optic, exp(coef) = **1.93987**
Hazard probability: **94%**
For a customer who has a fiber optic as internet service, the likelihood of churning will increase by 94%.

InternetServiceNo, exp(coef) = **0.24502**
Hazard probability: **-75%**
For a customer who does not have an internet service, the likelihood of churning will decrease by 75%.

PaymentMethodCredit card (automatic)= **0.24502**
Hazard probability: **-75%**
For a customer with a mode of payment as credit card, the likelihood of churning will decrease by 48%.

PaymentMethodElectronic check= **1.48608**
Hazard probability: **48%**
For a customer with a mode of payment as electronic check, the likelihood of churning will increase by 48%.

PaymentMethodMailed check= **1.61374**
Hazard probability: **61%**
For a customer with a mode of payment as mailed check, the likelihood of churning will increase by 61%.

MonthlyCharges= **1.03551**
Hazard probability: **3%**
For a unit increase in monthly charges the likelihood of a customer churning will increase by 3%.

TotalCharges= **0.99843**
Hazard probability: **-0.01%**
For a unit increase in total charges the likelihood of a customer churning will reduce by 0.01%.

From the above interpretation we can see that least impact on the hazard ratio in terms of magnitude is by total charges, highest impact is by contract 2 year and internet service fire optic.