

Group 1

TELCO

Customer Churn Regression Model

Govindu Sathruwan

Abdullah Sheriffdeen

Shahik Shiyam

Iffath Saleem

Aaisha Aamina

Customer Churn experienced by Telco

Telco is a telecommunication and mobile service provider. Our group aims to focus and identify reasons for the companies loss of customers by observing factors such as customer demographics, location and service among several others.

In the telecommunications industry, losing customers is extremely costly as a lost customer is unlikely to return and would prefer to stick with their new service provider thus making this a key factor that needs to be evaluated. In the following slides will be highlighting our work and the outcomes in identifying solutions to overcome said issue.

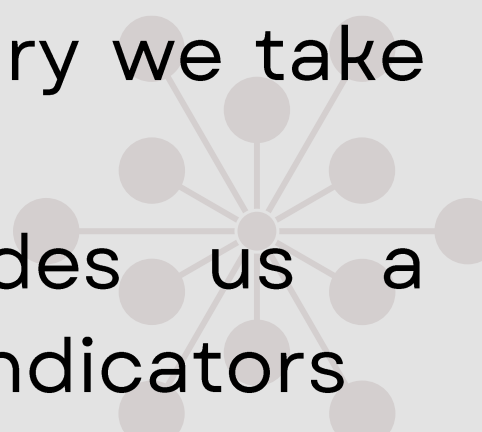


Data Structure and Summary

Data Structure

- We begin by making use of the `str()` function
- This is then used to identify the structure of the data
- This can be used to analyse the data type and similarly the rows and columns as well
- In this scenario it can be noted that there were 7043 rows and 33 columns.
- Similarly, a detailed graphic has been provided showing the relevant data types

Data Summary

- 
- To find the descriptive summary we take advantage of the `summary()`
 - The output provided provides us a statistical summary of several indicators
 - This can then be applied to our quantitative data to identify patterns
 - Some statistics that can be observed include the minimum and maximum value, median, mean and finally the interquartile ranges

```
> str(data)
'data.frame': 7043 obs. of 33 variables:
```

```
Tenure.Months
Min. : 0.00
1st Qu.: 9.00
Median :29.00
Mean :32.37
3rd Qu.:55.00
Max. :72.00
```

Monthly.Charges	Total.Charges
Min. : 18.25	Min. : 0.0
1st Qu.: 35.50	1st Qu.: 398.6
Median : 70.35	Median :1394.5
Mean : 64.76	Mean :2279.7
3rd Qu.: 89.85	3rd Qu.:3786.6
Max. :118.75	Max. :8684.8

Churn.Score	CLTV
Min. : 5.0	Min. :2003
1st Qu.: 40.0	1st Qu.:3469
Median : 61.0	Median :4527
Mean : 58.7	Mean :4400
3rd Qu.: 75.0	3rd Qu.:5380
Max. :100.0	Max. :6500

```
'data.frame': 7043 obs.
 $ CustomerID : chr
 $ Count : int
 $ Country : chr
 $ State : chr
 $ City : chr
 $ Zip.Code : int
 $ Lat.Long : chr
 8.315709" ...
 $ Latitude : num
 $ Longitude : num
 $ Gender : chr
 $ Senior.Citizen : chr
 $ Partner : chr
 $ Dependents : chr
 $ Tenure.Months : int
 $ Phone.Service : chr
 $ Multiple.Lines : chr
 $ Internet.Service : chr
 $ Online.Security : chr
 $ Online.Backup : chr
 $ Device.Protection: chr
 $ Tech.Support : chr
 $ Streaming.TV : chr
 $ Streaming.Movies : chr
 $ Contract : chr
 $ Paperless.Billing: chr
 $ Payment.Method : chr
 $ Monthly.Charges : num
 $ Total.Charges : num
 $ Churn.Label : chr
 $ Churn.Value : int
 $ Churn.Score : int
 $ CLTV : int
 $ Churn.Reason : chr
```

Converting categorical variables into factors

- We assign each categorical variable to a factor form so that a multiple linear regression model can be used.
- By assigning them as factors, it allows the `lm()` function to automatically create dummy variables for them.
- It is a crucial tool for regression modelling.

```
# Setting the variables with more than two categories to factor  
# lm() in regression automatically creates the dummy variables  
  
data$Internet.Service <- factor(data$Internet.Service)  
data$Payment.Method <- factor(data$Payment.Method)  
data$Contract <- factor(data$Contract)  
data$Gender <- factor(data$Gender)  
data$Senior.Citizen <- factor(data$Senior.Citizen)  
data$Partner <- factor(data$Partner)  
data$Dependents <- factor(data$Dependents)  
data$Phone.Service <- factor(data$Phone.Service)  
data$Multiple.Lines <- factor(data$Multiple.Lines)  
data$Online.Security <- factor(data$Online.Security)  
data$Online.Backup <- factor(data$Online.Backup)  
data$Device.Protection <- factor(data$Device.Protection)  
data$Tech.Support <- factor(data$Tech.Support)  
data$Streaming.TV <- factor(data$Streaming.TV)  
data$Streaming.Movies <- factor(data$Streaming.Movies)  
data$Paperless.Billing <- factor(data$Paperless.Billing)  
data$Churn.Label <- factor(data$Churn.Label)
```

Data Partition

- The purpose of this code is to split the dataset into training and test sets to prepare for machine learning and statistical modelling.
- 65% for Train
- 35% for Test
- Overall it aims to prevent biasness, which enhances our models reliability.

```
# Data Partition  
  
RanNum <- runif(7043)  
length(RanNum)  
  
Index <- order(RanNum)  
Index  
  
Train <- data[Index[1:4543],]  
Test <- data[Index[4543:7043],]
```

Full Model

- **Dependent Variable: Churn Score**

Represents the likelihood of a customer leaving.

- **Independent Variables:**

- All predictors from the dataset were included to build the full model, ensuring a comprehensive analysis.
- These include both continuous (e.g., Monthly Charges, Total Charges) and categorical variables (e.g., Internet Service, Contract, Payment Method, and service add-ons like Tech Support and Online Security).

As churn score (Dependent Variable) is a continuous variable, we utilized Multiple Linear Regression for modeling.

Call:

```
lm(formula = Churn.Score ~ Total.Charges + Monthly.Charges +  
  Zip.Code + Gender + Senior.Citizen + Partner + Dependents +  
  Tenure.Months + Phone.Service + Multiple.Lines + Internet.Service +  
  Online.Security + Online.Backup + Device.Protection + Tech.Support +  
  Streaming.TV + Streaming.Movies + Contract + Paperless.Billing +  
  Payment.Method + Churn.Label + CLTV, data = Train)
```

Residuals:

Min	1Q	Median	3Q	Max
-45.334	-12.722	0.039	12.832	32.299

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	4.572e+01	1.336e+01	3.421	0.000628	***
Total.Charges	-4.126e-04	3.473e-04	-1.188	0.234771	
Monthly.Charges	1.451e-01	2.332e-01	0.622	0.533734	
Zip.Code	-1.989e-05	1.275e-04	-0.156	0.876054	
GenderMale	8.418e-01	4.760e-01	1.769	0.077043	.
Senior.CitizenYes	-2.272e-01	6.852e-01	-0.332	0.740254	
PartnerYes	-2.327e-01	5.481e-01	-0.425	0.671202	
DependentsYes	-7.588e-01	6.310e-01	-1.203	0.229219	
Tenure.Months	3.156e-02	2.692e-02	1.173	0.241054	
Phone.ServiceYes	-1.469e+00	4.771e+00	-0.308	0.758151	
Multiple.LinesYes	-1.900e-01	1.301e+00	-0.146	0.883910	
Internet.ServiceFiber optic	-3.597e+00	5.866e+00	-0.613	0.539765	
Internet.ServiceNo	3.477e+00	5.920e+00	0.587	0.557016	
Online.SecurityYes	-8.977e-01	1.337e+00	-0.672	0.501862	
Online.BackupYes	7.437e-01	1.313e+00	0.566	0.571091	
Device.ProtectionYes	-2.050e+00	1.316e+00	-1.557	0.119451	
Tech.SupportYes	-7.897e-01	1.343e+00	-0.588	0.556592	
Streaming.TVYes	-1.535e+00	2.404e+00	-0.639	0.523022	
Streaming.MoviesYes	-9.024e-01	2.399e+00	-0.376	0.706756	
ContractOne year	5.610e-02	7.556e-01	0.074	0.940815	
ContractTwo year	4.340e-01	9.055e-01	0.479	0.631740	
Paperless.BillingYes	1.160e-01	5.315e-01	0.218	0.827191	
Payment.MethodCredit card (automatic)	6.167e-02	7.259e-01	0.085	0.932305	
Payment.MethodElectronic check	-8.792e-02	7.072e-01	-0.124	0.901057	
Payment.MethodMailed check	-4.167e-01	7.789e-01	-0.535	0.592628	
Churn.LabelYes	3.256e+01	6.409e-01	50.805	< 2e-16	***
CLTV	1.386e-04	2.202e-04	0.629	0.529193	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 16.01 on 4516 degrees of freedom

Multiple R-squared: 0.4496, Adjusted R-squared: 0.4464

F-statistic: 141.9 on 26 and 4516 DF, p-value: < 2.2e-16

Techniques Used To Obtain Optimum Model

1. Forward Selection Model

- Starts with a null model (No independent variables).
- Adds a variable each iteration.
- The significant variables are identified based on the AIC value.
- Iteration stops when the addition of new variables does not improve the model (eg: r^2 value and AIC value)

```
> # Summarize the selected model
> summary(forward_model)

Call:
lm(formula = Churn.Score ~ Churn.Label + Online.Backup + Device.Protection +
    Gender + Phone.Service, data = Train)

Residuals:
    Min       1Q   Median       3Q      Max
-44.821 -12.720  -0.052  12.948  32.130

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    48.3211     0.8719   55.419 < 2e-16 ***
Churn.LabelYes  32.5064     0.5377   60.453 < 2e-16 ***
Online.BackupYes  1.4191     0.5239    2.709  0.00677 **
Device.ProtectionYes -1.2979     0.5257   -2.469  0.01360 *
GenderMale      0.8466     0.4746    1.784  0.07454 .
Phone.ServiceYes  1.3782     0.8142    1.693  0.09058 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.99 on 4537 degrees of freedom
Multiple R-squared:  0.4485,    Adjusted R-squared:  0.4479
F-statistic: 738 on 5 and 4537 DF,  p-value: < 2.2e-16
```

Techniques Used To Obtain Optimum Model

2. Backward Elimination Model

- Starts with a full model with all the selected variables.
- Removes a variable at each iteration.
- The significant variables are identified based on the AIC value.
- Iteration stops when the removal of new variables does not improve the model (eg: r^2 value and AIC value)

```
> summary(backward_model)
```

```
Call:
```

```
lm(formula = Churn.Score ~ Gender + Phone.Service + Online.Backup +  
    Device.Protection + Churn.Label, data = Train)
```

```
Residuals:
```

```
      Min       1Q   Median       3Q      Max  
-44.821 -12.720  -0.052  12.948  32.130
```

```
Coefficients:
```

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	48.3211	0.8719	55.419	< 2e-16	***
GenderMale	0.8466	0.4746	1.784	0.07454	.
Phone.ServiceYes	1.3782	0.8142	1.693	0.09058	.
Online.BackupYes	1.4191	0.5239	2.709	0.00677	**
Device.ProtectionYes	-1.2979	0.5257	-2.469	0.01360	*
Churn.LabelYes	32.5064	0.5377	60.453	< 2e-16	***

```
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 15.99 on 4537 degrees of freedom
```

```
Multiple R-squared:  0.4485,    Adjusted R-squared:  0.4479
```

```
F-statistic:   738 on 5 and 4537 DF,  p-value: < 2.2e-16
```

Techniques Used To Obtain Optimum Model

3. Stepwise selection Model

- Combines the forward and backward selection model techniques.
- Starts with either a full model or null model.
- Removes or adds a variable at each iteration depending on the AIC and r^2 values.
- Iteration stops when the removal or addition of new variables does not improve the model (eg: r^2 value and AIC value)

```
> summary(stepwise_model)

Call:
lm(formula = Churn.Score ~ Churn.Label + Online.Backup + Device.
    Protection +
    Gender + Phone.Service, data = Train)

Residuals:
    Min       1Q   Median       3Q      Max
-44.821 -12.720  -0.052  12.948  32.130

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      48.3211     0.8719   55.419 < 2e-16 ***
Churn.LabelYes    32.5064     0.5377   60.453 < 2e-16 ***
Online.BackupYes   1.4191     0.5239    2.709  0.00677 **
Device.ProtectionYes -1.2979     0.5257   -2.469  0.01360 *
GenderMale         0.8466     0.4746    1.784  0.07454 .
Phone.ServiceYes   1.3782     0.8142    1.693  0.09058 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.99 on 4537 degrees of freedom
Multiple R-squared:  0.4485,    Adjusted R-squared:  0.4479
F-statistic: 738 on 5 and 4537 DF,  p-value: < 2.2e-16
```

Final Model

- Dependent Variable - Churn score
 - Independent Variables - Churn Label
Online Backup
Gender
Phone Service
Device Protection
-
- Accept Alternative hypothesis at 5% significance level for 3 variables.
 - Accept Alternative hypothesis at 10% significance level for 2 variables.

```
> summary(CustChurnRegression)
```

```
Call:
```

```
lm(formula = Churn.Score ~ Churn.Label + Online.Backup + Gender +  
    Phone.Service + Device.Protection, data = Train)
```

```
Residuals:
```

Min	1Q	Median	3Q	Max
-44.821	-12.720	-0.052	12.948	32.130

```
Coefficients:
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	48.3211	0.8719	55.419	< 2e-16 ***
Churn.LabelYes	32.5064	0.5377	60.453	< 2e-16 ***
Online.BackupYes	1.4191	0.5239	2.709	0.00677 **
GenderMale	0.8466	0.4746	1.784	0.07454 .
Phone.ServiceYes	1.3782	0.8142	1.693	0.09058 .
Device.ProtectionYes	-1.2979	0.5257	-2.469	0.01360 *

```
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 15.99 on 4537 degrees of freedom
```

```
Multiple R-squared:  0.4485,    Adjusted R-squared:  0.4479
```


```
F-statistic:   738 on 5 and 4537 DF,  p-value: < 2.2e-16
```


Final Model

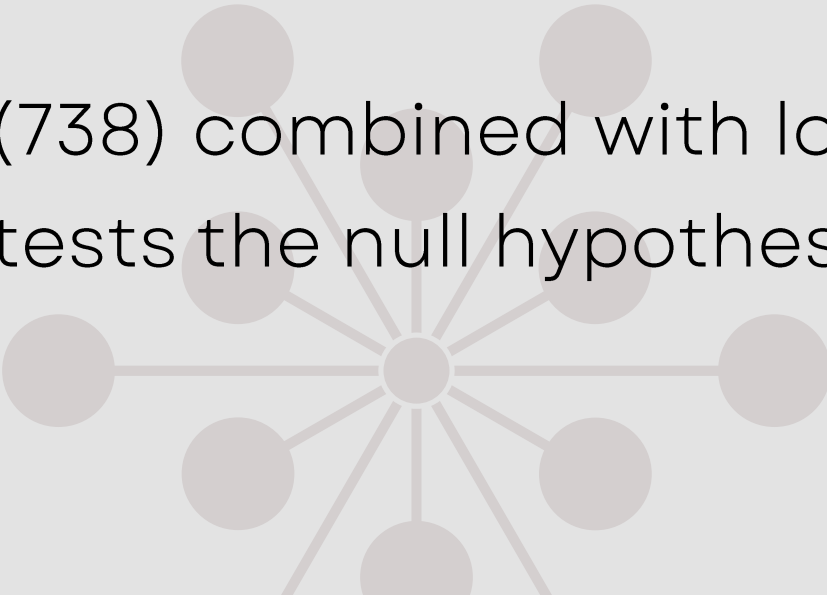
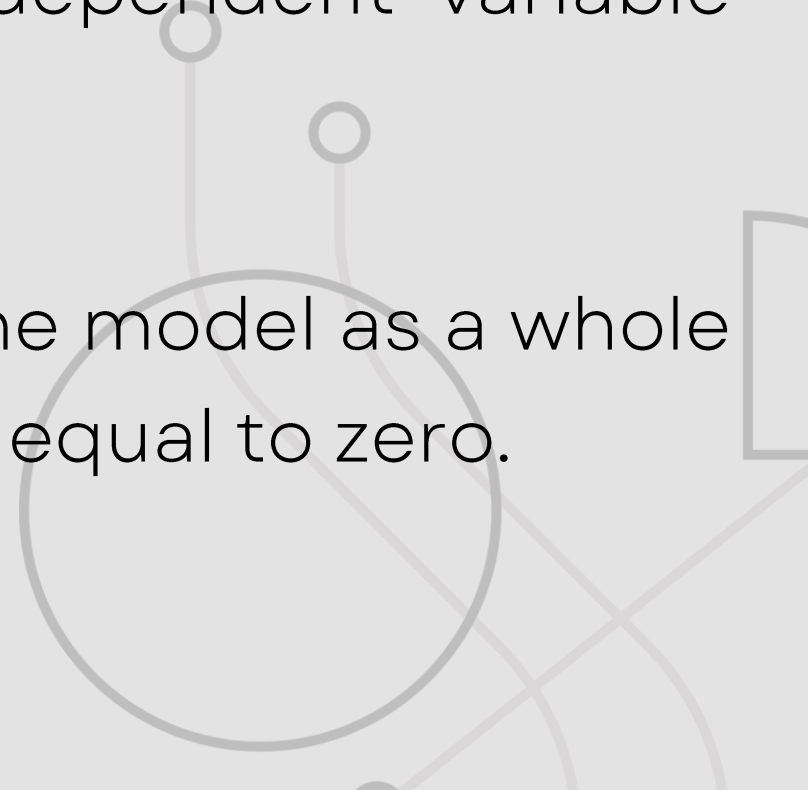
$$\text{Churn.Score} = 48.3211 + 32.5064 \times (\text{Churn.LabelYes}) + 1.4191 \times (\text{Online.BackupYes}) + 0.8466 \times (\text{GenderMale}) + 1.3782 \times (\text{Phone.ServiceYes}) - 1.2979 \times (\text{Device.ProtectionYes})$$

- Intercept (48.3211): The predicted Churn.Score when all predictors are at their baseline.
- Churn.LabelYes (32.5064): If Churn.Label is "Yes," the Churn.Score increases by 32.5064, holding all else constant.
- Online.BackupYes (1.4191): If Online.Backup is "Yes," the Churn.Score increases by 1.4191, holding all else constant.
- GenderMale (0.8466): If Gender is "Male," the Churn.Score increases by 0.8466, holding all else constant.
- Phone.ServiceYes (1.3782): If Phone.Service is "Yes," the Churn.Score increases by 1.3782, holding all else constant.
- Device.ProtectionYes (-1.2979): If Device.Protection is "Yes," the Churn.Score decreases by 1.2979, holding all else constant.

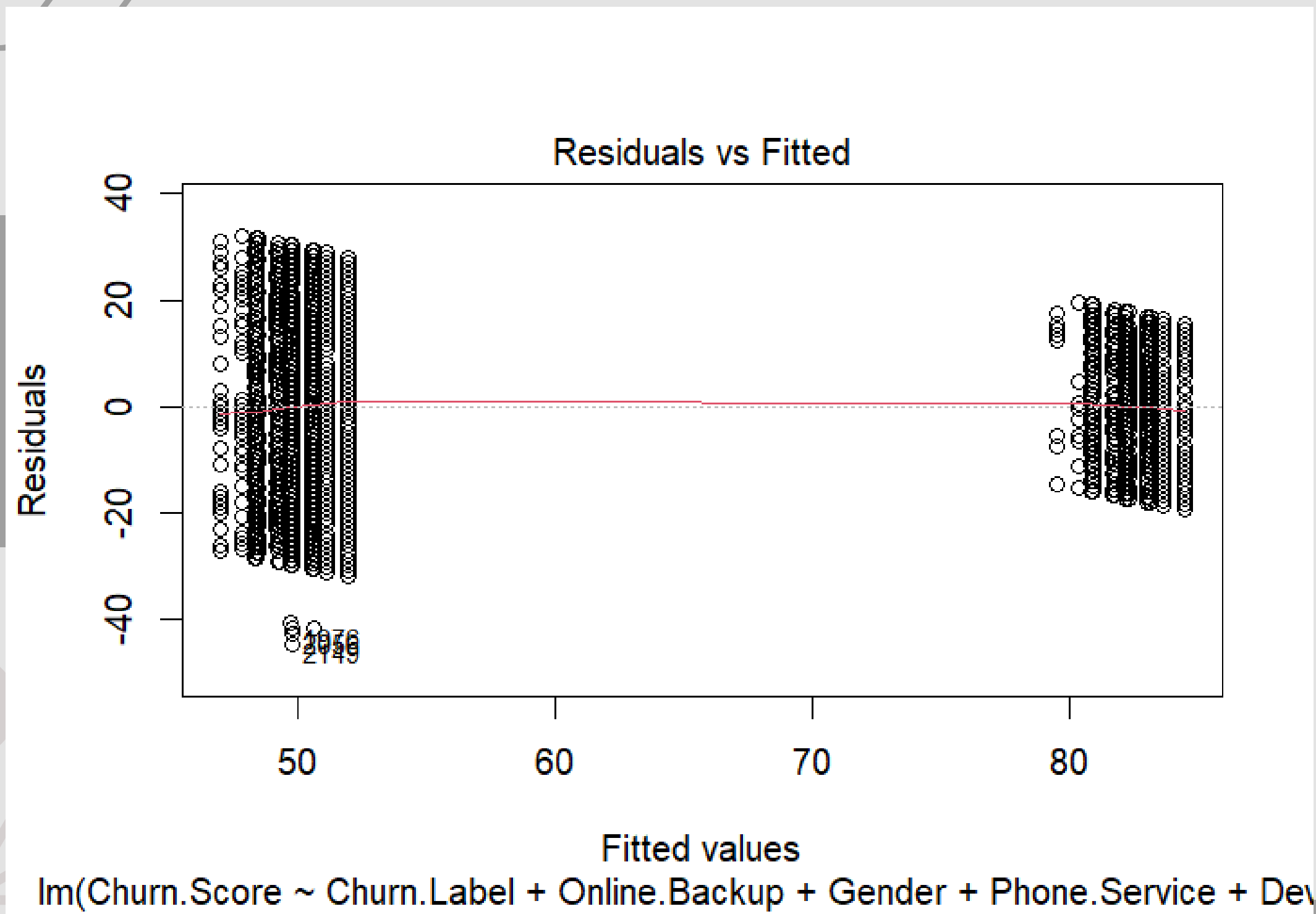
Final Model



```
Residual standard error: 15.99 on 4537 degrees of freedom  
Multiple R-squared: 0.4485, Adjusted R-squared: 0.4479  
F-statistic: 738 on 5 and 4537 DF, p-value: < 2.2e-16
```

- RSE - measures the average deviation (16) of the observed Churn Score values from the values predicted by the model.
 - Adjusted R^2 - indicates the proportion of the variance (45%) in the dependent variable (Churn Score) explained by the independent variables.
 - High F statistic (738) combined with low P value ($< 2.2 \times 10^{-16}$) means that the model as a whole is significant. It tests the null hypothesis that all regression coefficients are equal to zero.
- 
- 

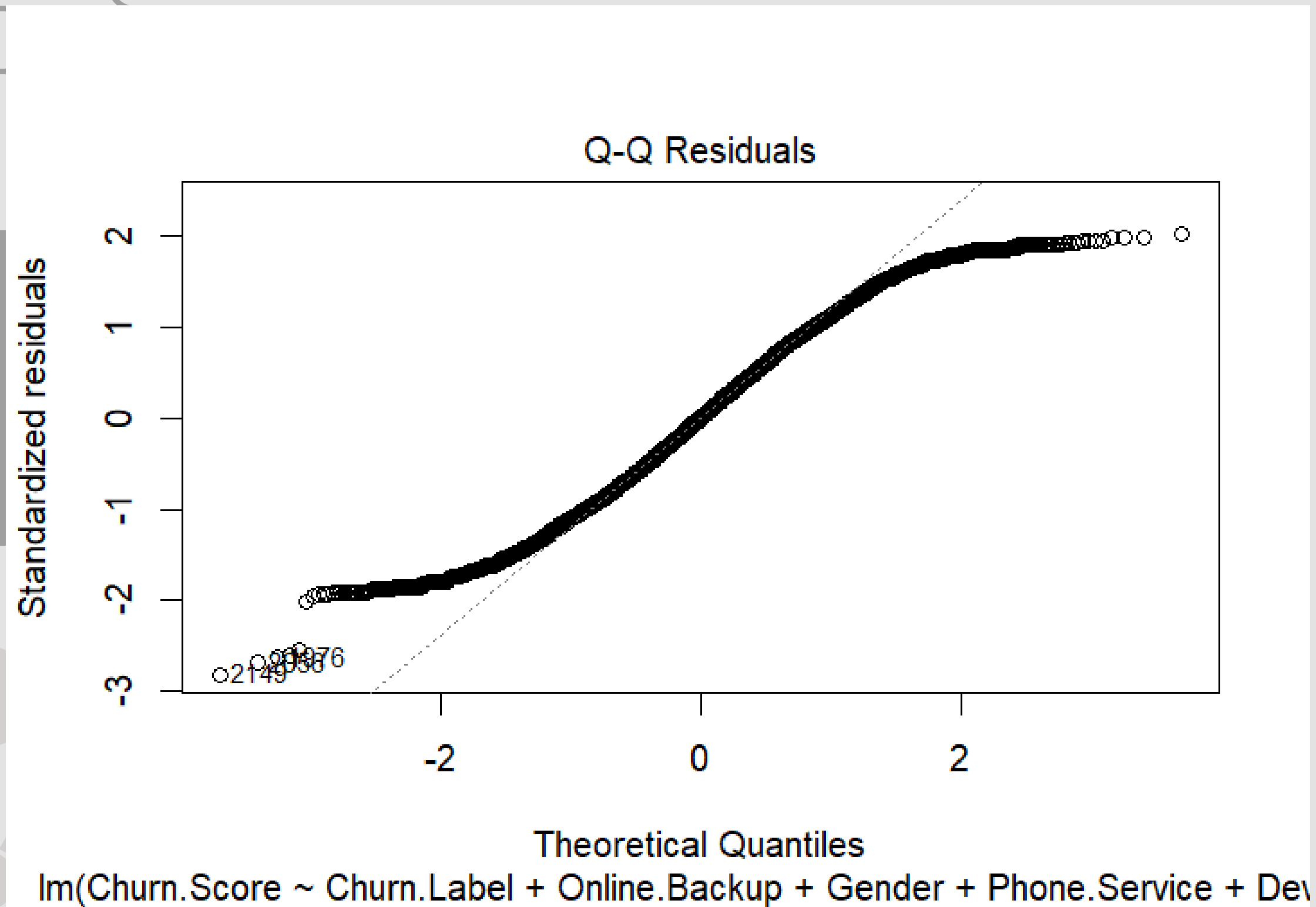
Model Diagnostics



The Residuals vs Fitted plot helps identify non-linearity, unequal error variance, and outliers. Ideally, the points should be randomly distributed around the horizontal line without any distinct pattern.

The residuals appear scattered without a clear pattern, which suggests that the assumption of linearity is reasonably met.

Model Diagnostics

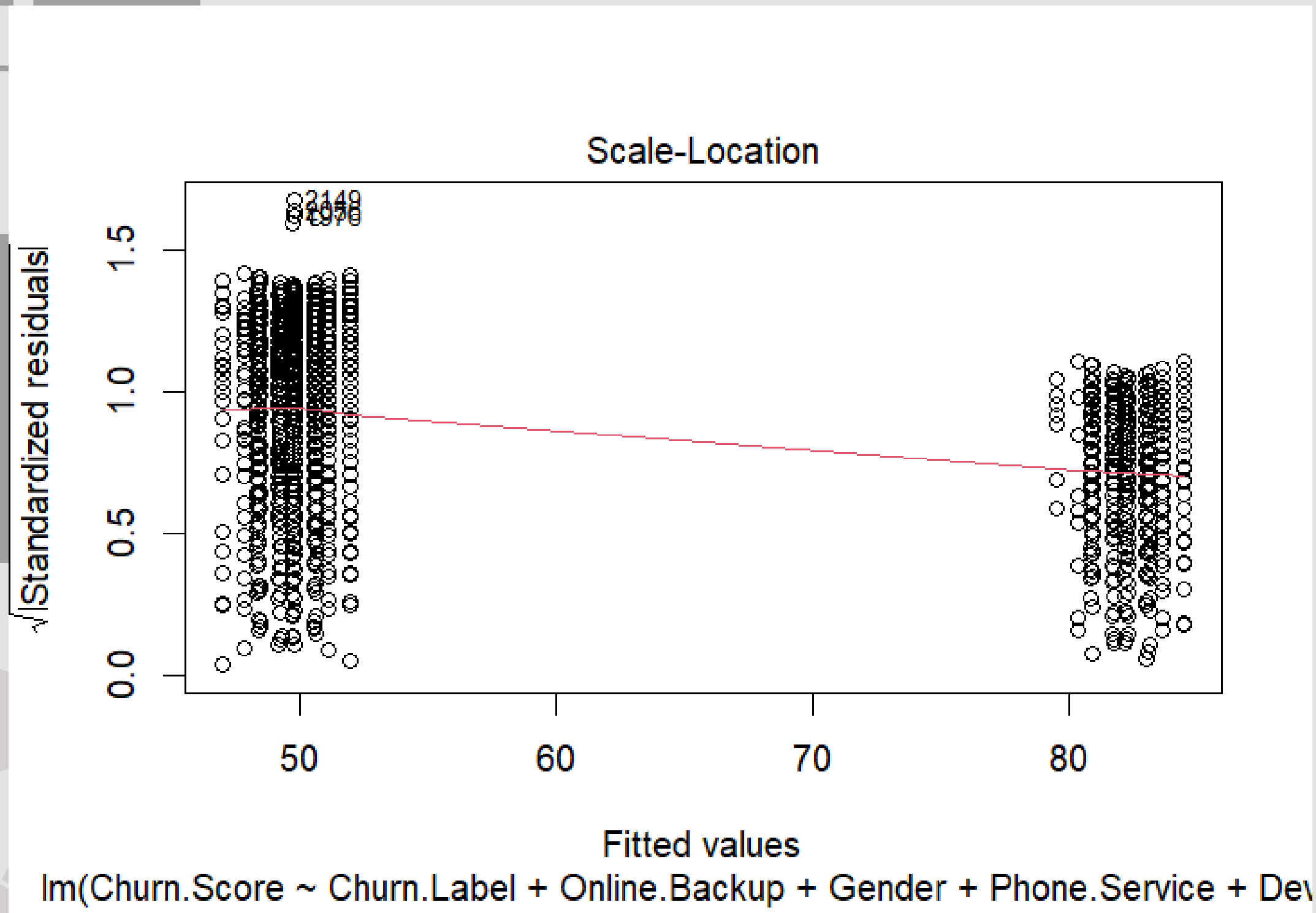


A Q-Q plot is a tool used to check if data, follows a specific distribution, usually the normal distribution.

Left Tail (bottom-left): The points curve below the line, indicating heavy tails (Negative residuals - predicted value > actual value).

Right Tail (top-right): The points curve above the line, indicating heavy positive tails or high outliers. (Positive residuals - predicted value < actual value)

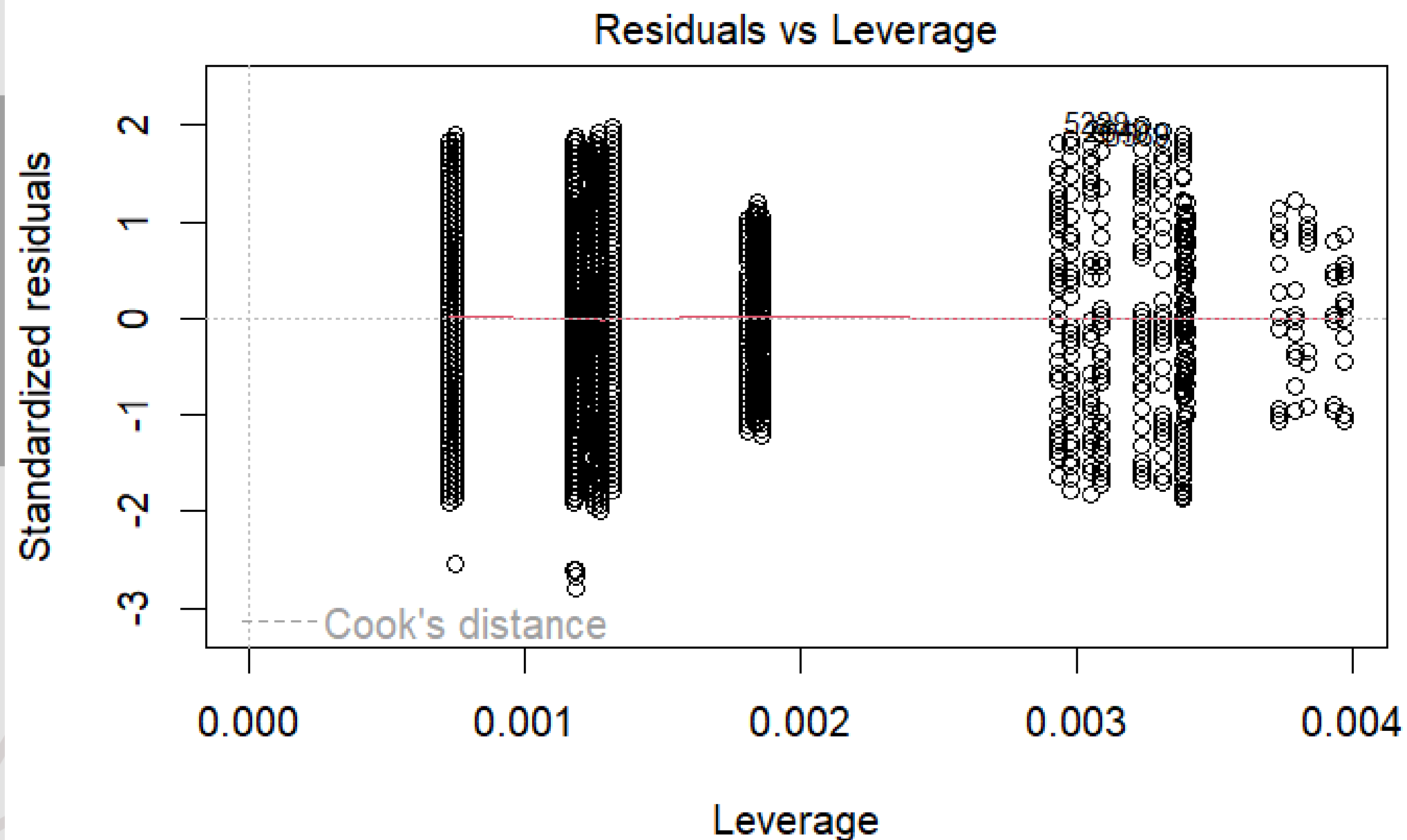
Model Diagnostics



This plot tests for homoscedasticity and ensures that the variance of residuals is consistent across fitted values. For this plot to be effective, the plots should be randomly spread.

The spread of points appears relatively consistent

Model Diagnostics



lm(Churn.Score ~ Churn.Label + Online.Backup + Gender + Phone.Service + Dev

This plot identifies the influential data points that have a significant impact on the regression model. Impactful observations would play a key role in influencing the model

Prediction Test

A new column was created in test data frame with the predicted churn scores from the test data and our final model.

```
predict = predict(CustChurnRegression, newdata =Test)
predict
Test$PREDICTED.Churn.Score=predict
```

Churn.Score	CLTV	Churn.Reason	PREDICTED.Churn.Score
78	4049		49.28894
79	5431		50.54592
58	5254		49.28894
25	5473		49.69933
67	6203		48.40144
32	5038		50.66716
50	4432		49.69933
25	4210		50.66716
62	6069		49.24804
80	5198		51.96505
64	3222		49.28894
78	4067		49.69933
20	4439		50.54592
79	4218		49.69933
53	3217		50.54592
26	5453		49.74023
27	4759		48.40144
73	3555		48.40144
65	4034		49.74023
23	5917		49.69933
59	3200		50.54592
45	3733		49.69933
48	4408		49.69933
27	3068		50.54592
62	4191		49.24804

THANK YOU

