Netflix User Subscription Retention Prediction using Generalized Linear Models

Shinjon Ghosh

Department of Mathematics, Illinois State University

Abstract:

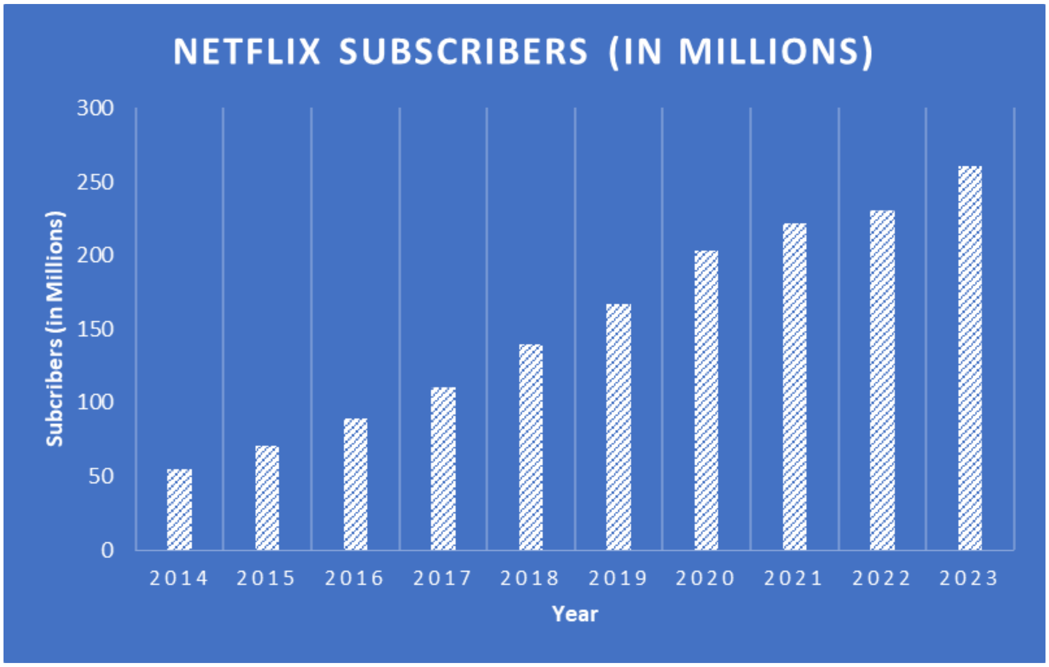
Subscription retention is an important factor in the success and growth of streaming platforms like Netflix. This study proposed the application of Generalized Linear Models (GLMs) to predict and identify factors contributing to user continuation. Developing a robust model that can anticipate churn and guide retention strategies. The analysis involves building GLMs with a binomial probit link function to predict binary outcomes—whether a user remains active or becomes inactive. By examining a range of features, the aim to uncover patterns and relationships that influence user behavior and retention. The results highlight significant predictors (subscription, country, gender) of user retention and offer actionable insights for developing targeted continuation strategies. Moreover, the model diagnostics suggest that the proposed GLMs fitted well and adequate according to Netflix userbase dataset. This project provides the value of predictive modeling in driving business decisions and optimizing customer retention of Netflix.

Keywords: generalized linear models, Netflix subscription, subscription retention, binomial glms, probit link function, interaction variable, odds ratio, quantile residuals, score test, likelihood ratio test, multicollinearity check.

Introduction:

Netflix is one of the world's leading entertainment services, with millions of subscribers all over the world. Almost 15% of Internet downlink traffic was attributed to Netflix. To be specific, it controlled 26.58% of the global market for video streaming services [6]. In the statistical field, predicting the number of future users is worth studying. As the market demand develops and varies frequently, Netflix regularly adjusts its prices to adapt to the market and understand more about consumer needs in order to serve its members better and attract more subscribers [7]. The OTT platforms growing rapidly and being more competitive day by day. The GLMs provides insights into factors that influence subscription retention and guides strategies to reduce churn. Generalized Linear Models (GLMs) provide a suitable approach to predicting binary outcomes, such as whether a user will renew or cancel their subscription. By using GLMs with a probit link function, this study aims to identify key factors that influence user retention on Netflix. The purpose of this study is to develop a predictive model to estimate the probability of user retention, enabling Netflix to understand strategies for customer retention, and optimize marketing efforts.

The dataset used for this predictor analysis based on demographic variables, revenue, subscription plan. The results of this project indicate predictor variables (subscription, country, gender) are highly significant. According to the model diagnostics, the proposed GLM is fitted well. The odds ratio suggests the outcome of response variables with comparison between categorized predictor variables.



The subscription rate of Netflix is gradually growing up. (**Source: Netflix, Statista)**

Background & Literature Review:

Netflix's growth is driven by lower-priced overseas subscriptions, but US subscribers are more profitable. To increase users, Netflix could introduce a 5to7 monthly subscription with ads. Unlike Disney, Netflix releases entire series at once, allowing binge-watching and easy switching between streaming providers [4]. Improve perception of anytime usability, collaborate with cellular data providers for special packages, offer diverse payment methods, provide trial periods to attract more subscribers. Enhance belief in ease of access and benefits to increase subscription intentions. Consider partnerships with cellular data providers for special packages and diverse payment methods [5]. COVID-19 significantly impacted Netflix subscribers from 2020 to 2022, with income, production, competition, and copyright being key factors [1]. Adjusted quantile residual's distribution approximates standard normal distribution better than other residuals, especially with larger sample sizes and higher variance. Applications showed adjusted quantile residual's effectiveness in detecting model misspecification and outliers, outperforming other residuals.

Despite standardized deviance residual being commonly used, adjusted quantile residual is simpler, easier to calculate, and more effective in diagnostic analysis [3].

Objectives:

- Build an optimum GLM model to forecast user subscription continuation based on user characteristics.
- Identify key factors influencing subscription continuation.
- Evaluate model performance to improve user retention strategies and enhance customer satisfaction.

Methods & Materials:

Data Source & Overview

The Netflix Userbase dataset collects from Kaggle (secondary data resources) which provides demographics and subscription details of Netflix users. The dataset contains 2500 observations and 10 features. Among 10 variables subscription (Basic, Standard, Premium), device (Smart TV, Laptop, Smartphone, Tablet), country (Brazil, USA, UK, Italy, France, Spain, Canada, Mexico, Germany, Australia), gender (Male & Female) are categorical features and user id, revenue, age are numerical variables. Moreover, the dataset has customer join date, last payment date and subscription duration columns. Each row corresponds to a specific user, identifiable by their unique user id. This dataset allows us to explore and analyze Netflix's user base in terms of their subscription choices and demographics measurements. It is a useful resource for statistical modelling, hypothesis testing and developing strategies for marketing, user retention, and mitigate churn risk.

Data Preparation

The aim of our project is to develop Generalized Linear Models to predict subscription retention. In the dataset 'last payment date' column format was several date formats. I transformed this column to single date structure which is applicable for analysis. Then I calculate active & non-active account according to subscription duration. If the last payment date is below 30 days, then it will be active account otherwise non-active. After that, the categorical variables are factorized by label 0 & 1 where subscription (Basic = 0 & others = 1), gender (Male = 0 & Female = 1), country (Brazil, Canada, Italy contains 0 & others countries 1), device (Laptop = 0 & others 1). Later, I define account status (active & non-active account) as a response variable. On the other hand, I set categorical and numerical variable as predictor or explanatory variables according to dataset.

Result & Discussion:

Exploratory Data Analysis

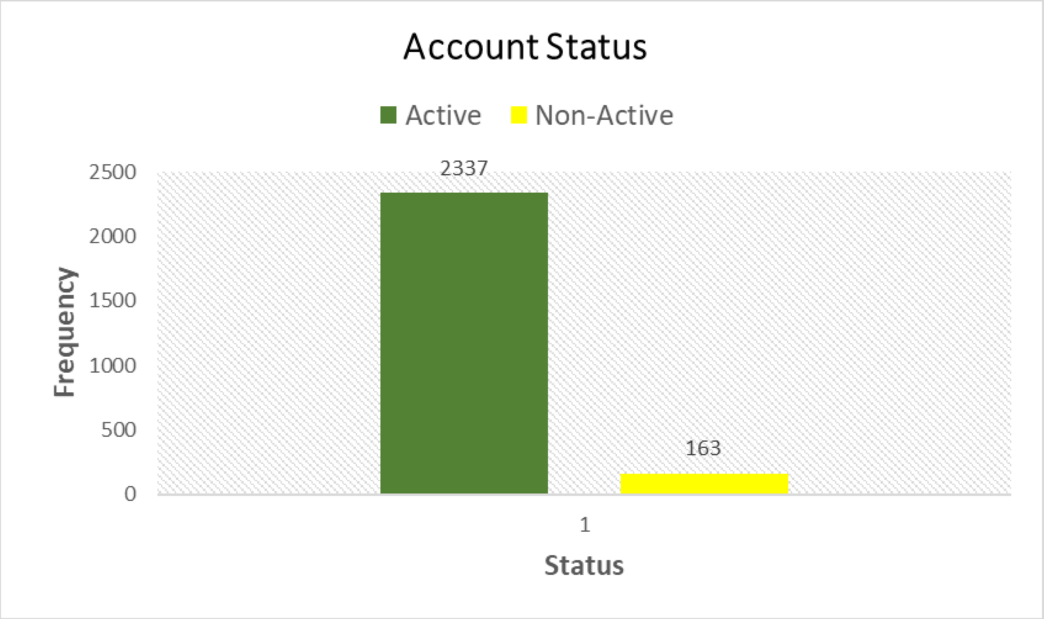


Fig 1: Account Status

Information in Fig 1 illustrates the comparison of Netflix account status values which indicates active account subscribers and non-active subscribers. From the bar diagram, we can say that active account customers numbers (2337) outperformed non-active (163) subscribers. That means the response variable data is highly imbalanced.

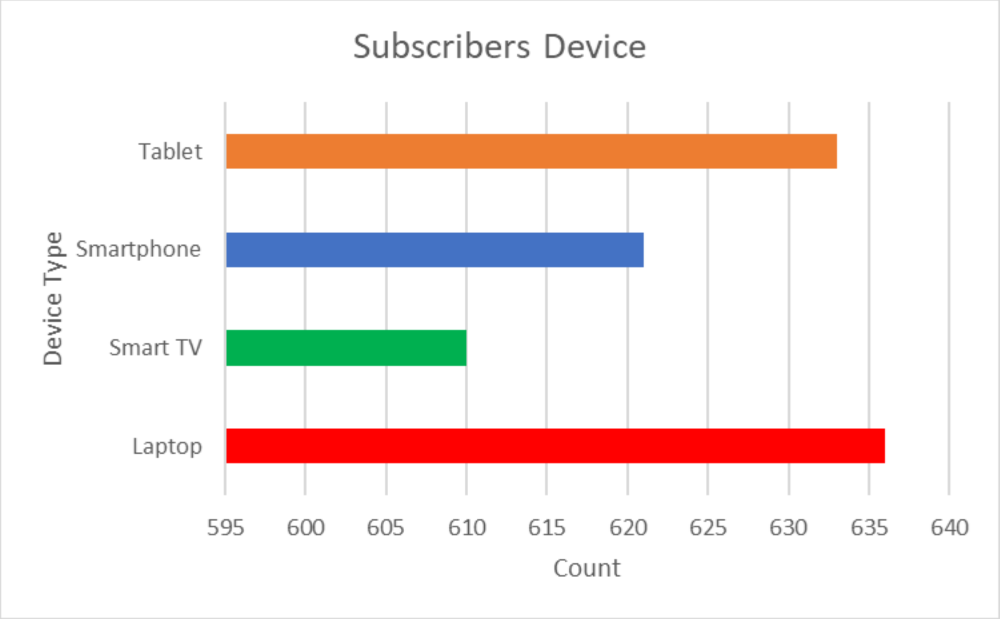


Fig 2: Subscribers device variation

Fig 2 shows the comparison between customers Netflix account log in device. From the horizontal bar plot, we find maximum number of subscribers use laptop and the number is more than 635. On the other hand, 610 users Netflix account logged-in in smart tv which is the minimum numbers. Moreover, more than 620 people use smartphone and almost 633 people use tablet for watching Netflix.

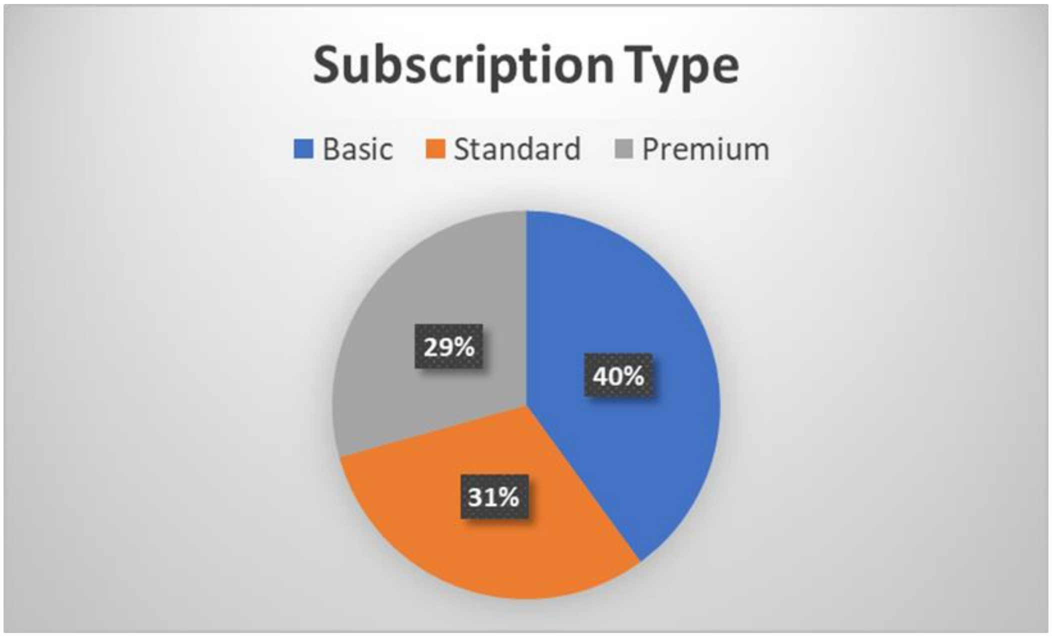


Fig 3: Customers Subscription Type

This pie chart depicts the overall subscriber subscription plan variation. From Fig 3 we can say that, 40 percent account holders buy basic plan and 31% subscriber holds standard plan. In addition, only 29 percent customers use premium subscription plan.

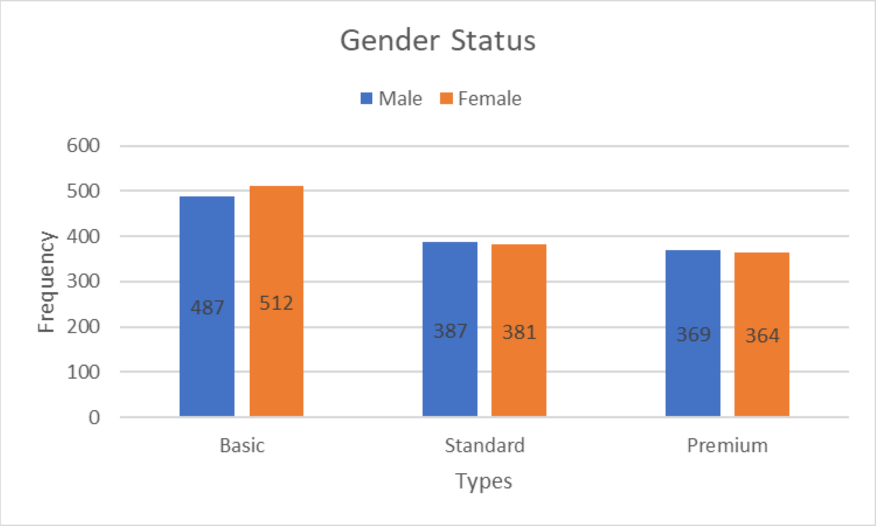


Fig 4: Gender wise subscription plan

Fig 4 compares subscription plan customers according to gender. The bar plot shows that female users are higher than male users in basic subscription package. The basic plan female users value is 512. Moreover, 487 male account holders choose basic plan. On contrary, male subscriber number is higher than females for standard & premium plan. The plot depicts 387 male account holders buy standard and 369 uses premium. Moreover, the female premium customers value is 364 and 381 buy standard plan.

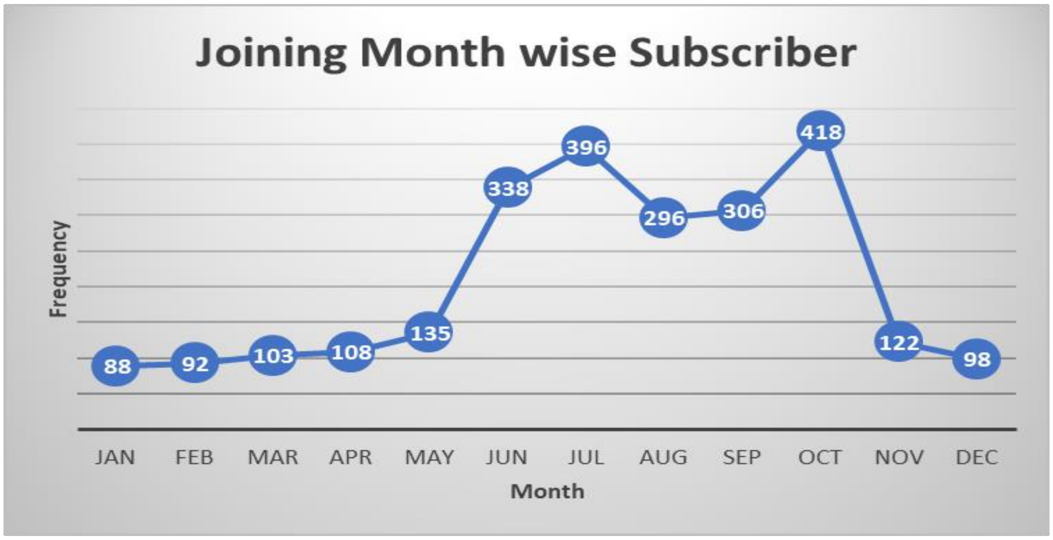


Fig 5: Subscriber distribution by joining month

In Fig 5 illustrates the subscriber joining numbers according to several month. From the line chart, we say that the highest numbers subscriber joins Netflix in October which is more than 400 and the lowest numbers customer start using Netflix in January which is 88. From January to May, the subscriber increasing rate is very low but May to July it's growing rapidly. Moreover, in September the joining rate decrease and it is peak high again in October. Furthermore, the subscriber subscription rate is rapidly decreasing in November and it's continuing till December.

Generalized Linear Models

From processed Netflix userbase dataset, I use account status as a response variable which is binary data (0 for active account and 1 for non-active). That's why I set and run several binomial generalized linear models according to different predictor variables and link functions. In Table 1, nf1, nf2, nf3 model predictor variables are Gender, Age*Revenue, Country*Subscription and I use logit, probit, complementary log-log link functions respectively. Moreover, nf6, nf7, nf8 are forward regression, backward elimination, stepwise regression models respectively. Furthermore, model nf4 and nf5 based on without any interaction predictor variables. In nf4 model, the explanatory variables are Age, Gender, Subscription, Revenue and Country. On the other hand, I use only categorical variables (Gender, Subscription, Country) in model nf5.

Models	AIC
nf1	1201.4
nf2	1201.5
nf3	1201.4
nf4	1211.6
nf5	1207.7
nf6	1205.7
nf7	1205.7
nf8	1205.7

Table 1: GLMs with AIC

We know lower AIC model performs better than higher AIC model. From Table 1, we can say that nf1(logit link) and nf3(complementary loglog link) models AIC is similar and lower than other models.

```
Call:
glm(formula = active ~ Gender + Age * Revenue + Country * Subscription, 
    family = binomial("logit"), data = nf)

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)   
(Intercept)            -9.955287   3.388056  -2.938  0.00330 **
GenderMale             -0.222160   0.163680  -1.357  0.17469   
Age                     0.201765   0.084612   2.385  0.01710 * 
Revenue                 0.611495   0.265504   2.303  0.02127 * 
Country1               -0.875235   0.269991  -3.242  0.00119 **
Subscription1          -0.636307   0.348518  -1.826  0.06789 . 
Age:Revenue            -0.015884   0.006706  -2.369  0.01785 * 
Country1:Subscription1  1.149585   0.426385   2.696  0.00702 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1205.2  on 2499  degrees of freedom
Residual deviance: 1185.4  on 2492  degrees of freedom
AIC: 1201.4

Number of Fisher Scoring iterations: 5
```

```
Call:
glm(formula = active ~ Gender + Age * Revenue + Country * Subscription, 
    family = binomial("cloglog"), data = nf)

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)   
(Intercept)            -9.743272   3.259189  -2.989  0.00279 **
GenderMale             -0.213297   0.157772  -1.352  0.17640   
Age                     0.194709   0.081282   2.395  0.01660 * 
Revenue                 0.590470   0.255194   2.314  0.02068 * 
Country1               -0.843498   0.261577  -3.225  0.00126 **
Subscription1          -0.611490   0.337752  -1.810  0.07022 . 
Age:Revenue            -0.015330   0.006442  -2.380  0.01732 * 
Country1:Subscription1  1.110516   0.414099   2.682  0.00732 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1205.2  on 2499  degrees of freedom
Residual deviance: 1185.4  on 2492  degrees of freedom
AIC: 1201.4

Number of Fisher Scoring iterations: 6
```

Both of this model coefficients are significant by p-value without GenderMale. Then I check multicollinearity by Variance Inflation Factor (VIF) and found Age and Revenue variables are highly corelated in this model.

Variables	VIF
Gender	1.002878
Age	56.663211
Revenue	30.997494
Country	2.441696
Subscription	4.413007
Age:Revenue	71.053446
Country:Subscription	6.846309

Table 2 : Multicollinearity check by VIF

After detecting multicollinearity issue, I remove Age and revenue variable from this model and run again this corresponding model with different link function.

Models	AIC
nf9	1201
nf10	1201.1
nf11	1201.1

Table 3: Modified GLMs Model

In the models nf9, nf10, nf11 predictor variables are Gender, Country*Subscription and link functions are probit, logit, complementary log-log respectively. From Table 3, model nf9 AIC is lower than others. In this case, I choose model nf9 for final model.

```
glm(formula = active ~ Gender + Country * Subscription, family = binomial(lin
k = "probit"),
    data = nf)

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)    
(Intercept)            -1.26060    0.08869 -14.213  < 2e-16 ***
GenderMale             -0.11013    0.07846  -1.404 0.160442    
Country1               -0.42515    0.12656  -3.359 0.000781 ***
Subscription1          -0.32387    0.16415  -1.973 0.048484 *  
Country1:Subscription1  0.56346    0.19847   2.839 0.004525 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1205.2  on 2499  degrees of freedom
Residual deviance: 1191.0  on 2495  degrees of freedom
AIC: 1201

Number of Fisher Scoring iterations: 5
```

From this model summary we can say that, all of the predictor variables are significant without GenderMale according to p-values. Moreover, GenderMale (-0.11013) indicates that every unit increase of GenderMale, the response variable slightly decreases the log-odds of 0.11013 unit compared to GenderFemale. Furthermore, Country1 (-0.42515) suggests that every unit increase of Country1 (UK, USA, Spain, France, Germany, Mexico) the log-odds of the active outcome decreases by 0.42515 units compared to Country0 (Brazil, Canada, Italy). Subscription1 (-0.32387) suggests every unit increase of Subscription1 (standard, premium) the log-odds of the active outcome decreases by 0.32387 units compared to basic subscription. In addition, Country1 :Subscription1 (0.56346) suggesting that the combination of Country1 and Subscription1 increases the log-odds of the response variables compared to Country0:Subscription0.

Our response variable data is highly imbalanced. In this case, Wald test is not trustworthy. Then I draw Score test for checking significancy of explanatory variables.

Score Test:

z-score	p-value
48.80721	0.00000

From this Score test, we get p-value is less than 0.05. That means the explanatory variables are significant.

Likelihood Ratio Tests to Compare Nested Models:

```
Model 1: active ~ Gender + Country * Subscription
Model 2: active ~ Gender + Country * Subscription + Age + Revenue

  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1      2495     1191.0                     
2      2493     1190.9  2 0.094866   0.9537
```

The residual deviance is almost similar for both models. Here, p-value is 0.9537 which is much greater than 0.05. So, we can say the additional variable is not significant. That means the model 1(nf9) is sufficient for the data.

After that, I check influential observation according to cook's distance, DFFITS and covariance ratio. I don't find any influential measures in my dataset. I illustrate first 25 observations influential measures here.

```
     dfb.1 dfb.GndM  dfb.Cnt1  dfb.Sbs1 dfb.C1.S   dffit cov.r   cook.d     hat
1  -0.00475  0.01131  0.000281  1.36e-04 -0.00473 -0.0199     1 1.97e-05 0.00123
2   0.00312 -0.00744 -0.012810 -8.94e-05  0.00824 -0.0180     1 1.57e-05 0.00208
3   0.00442 -0.01053 -0.000261 -1.27e-04 -0.00346 -0.0173     1 1.47e-05 0.00116
4  -0.00475  0.01131  0.000281  1.36e-04 -0.00473 -0.0199     1 1.97e-05 0.00123
5   0.00442 -0.01053 -0.000261 -1.27e-04 -0.00346 -0.0173     1 1.47e-05 0.00116
6  -0.00293  0.00698 -0.015239  8.39e-05  0.00965 -0.0212     1 2.19e-05 0.00227
7   0.00442 -0.01053 -0.000261 -1.27e-04 -0.00346 -0.0173     1 1.47e-05 0.00116
8  -0.00475  0.01131  0.000281  1.36e-04 -0.00473 -0.0199     1 1.97e-05 0.00123
9   0.00442 -0.01053 -0.000261 -1.27e-04 -0.00346 -0.0173     1 1.47e-05 0.00116
10  0.00312 -0.00744 -0.012810 -8.94e-05  0.00824 -0.0180     1 1.57e-05 0.00208
11 -0.00475  0.01131  0.000281  1.36e-04 -0.00473 -0.0199     1 1.97e-05 0.00123
12 -0.00475  0.01131  0.000281  1.36e-04 -0.00473 -0.0199     1 1.97e-05 0.00123
13  0.00312 -0.00744 -0.012810 -8.94e-05  0.00824 -0.0180     1 1.57e-05 0.00208
14  0.00442 -0.01053 -0.000261 -1.27e-04 -0.00346 -0.0173     1 1.47e-05 0.00116
15  0.00442 -0.01053 -0.000261 -1.27e-04 -0.00346 -0.0173     1 1.47e-05 0.00116
```

```
16  -0.00475  0.01131  0.000281  1.36e-04 -0.00473 -0.0199     1 1.97e-05 0.00123
17   0.00442 -0.01053 -0.000261 -1.27e-04 -0.00346 -0.0173     1 1.47e-05 0.00116
18  -0.00475  0.01131  0.000281  1.36e-04 -0.00473 -0.0199     1 1.97e-05 0.00123
19  -0.00475  0.01131  0.000281  1.36e-04 -0.00473 -0.0199     1 1.97e-05 0.00123
20  -0.00293  0.00698 -0.015239  8.39e-05  0.00965 -0.0212     1 2.19e-05 0.00227
21  -0.00475  0.01131  0.000281  1.36e-04 -0.00473 -0.0199     1 1.97e-05 0.00123
22  -0.00475  0.01131  0.000281  1.36e-04 -0.00473 -0.0199     1 1.97e-05 0.00123
23   0.00442 -0.01053 -0.000261 -1.27e-04 -0.00346 -0.0173     1 1.47e-05 0.00116
24   0.00442 -0.01053 -0.000261 -1.27e-04 -0.00346 -0.0173     1 1.47e-05 0.00116
25  -0.00475  0.01131  0.000281  1.36e-04 -0.00473 -0.0199     1 1.97e-05 0.00123
```

Now, I recheck multicollinearity by VIF. From Table 4, we can illustrate every explanatory variables VIF value is below the threshold 10. So, there is not any multicollinearity effect.

Variables	VIF
Gender	1.002878
Country	2.441696
Subscription	4.413007
Country:Subscription	6.846309

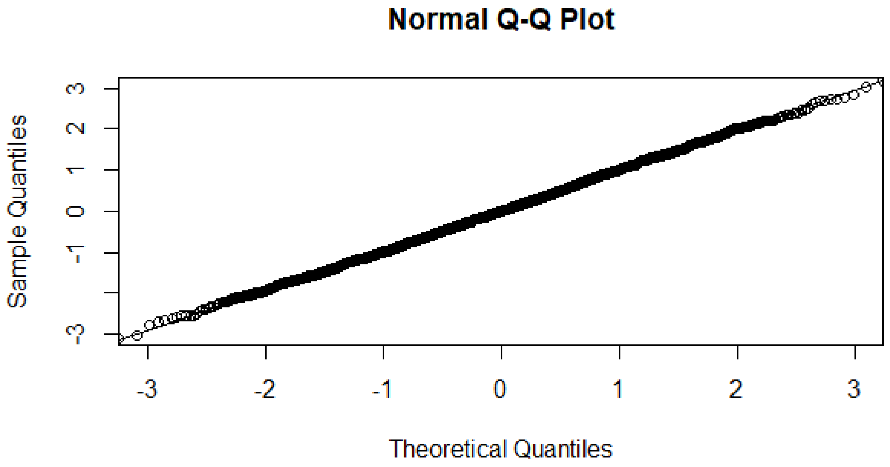
Table 4: Multicollinearity check by VIF

For checking random component, I use quantile residuals to draw Q-Q plot. For binomial families, quantile residual performance better than deviance residual. At first, I find out quantile residuals for model nf9. I illustrate first 81 observations quantile residuals here.

```
[1]  0.311338378 -0.429405175  0.728268918 -0.005781057 -1.672425551

   [6] -2.065600851 -0.250508726  0.864079042 -2.424531748 -1.115321781

  [11] -0.026859752 -1.553901615  0.067153171  0.267422430  0.453636024

  [16]  0.428266455  0.883469551  0.741530694 -2.289940503 -1.334473735

  [21] -1.666607914 -0.438535009 -1.341531792 -0.678832924 -0.305487255

  [26] -1.908715463 -0.324750743 -0.231149836  1.148039525 -0.897967242

  [31]  0.634730360  0.113187641 -1.728191835  0.521488623 -0.919225097

  [36] -0.258557679  0.689118465  0.014060425 -1.009945592 -0.185076912

  [41]  0.687414261  0.629750910 -1.054777048 -0.039602878 -0.861806531
```

```
 [46]  1.591802669 -0.045992466  1.314461567 -1.581539114  1.383716452
 [51] -0.300151253 -1.947485650  0.015748950 -0.896591123  0.572587113
[56] -0.761829343  0.277897995 -0.234295145 -0.469579186  0.930504352
 [61] -0.177709000 -0.231633797 -1.658800527  0.255551926  0.179681315
 [66]  0.056259189 -0.126551508 -1.100087796  0.891765835  0.013240446
 [71] -0.050357081  1.091608419  0.362325880  1.314065680  1.155788462
 [76]  0.194917442 -1.190423773  0.270040448 -1.437155015  0.601249004
 [81] -2.043044760 -0.973281791  0.200982330  1.150165202 -0.076285430
```

Then I draw the Q-Q plot based on quantile residuals. From the Normal Q-Q plot, we show quantile residuals have an exact normal distribution.

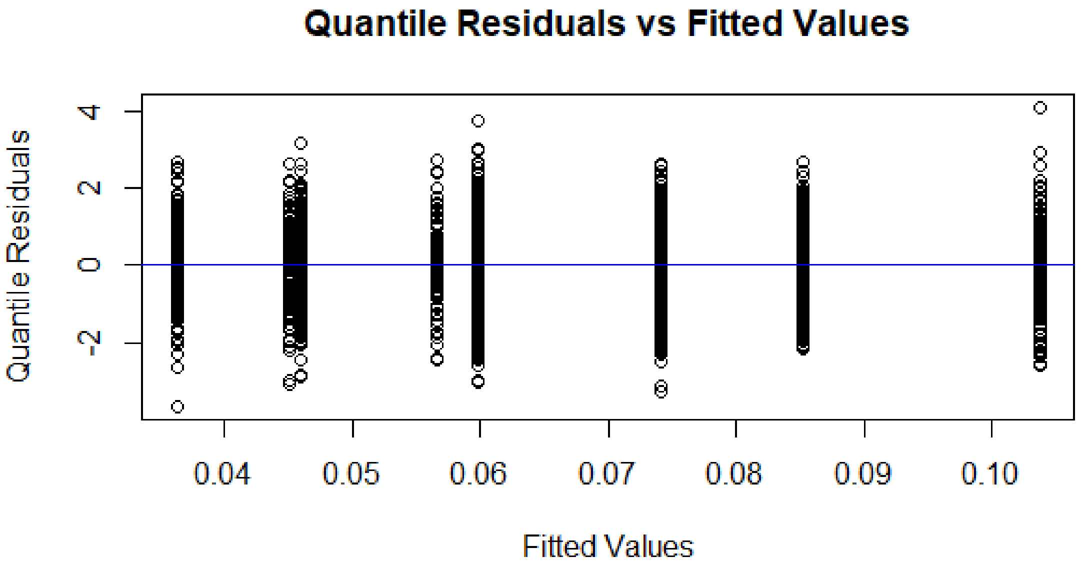


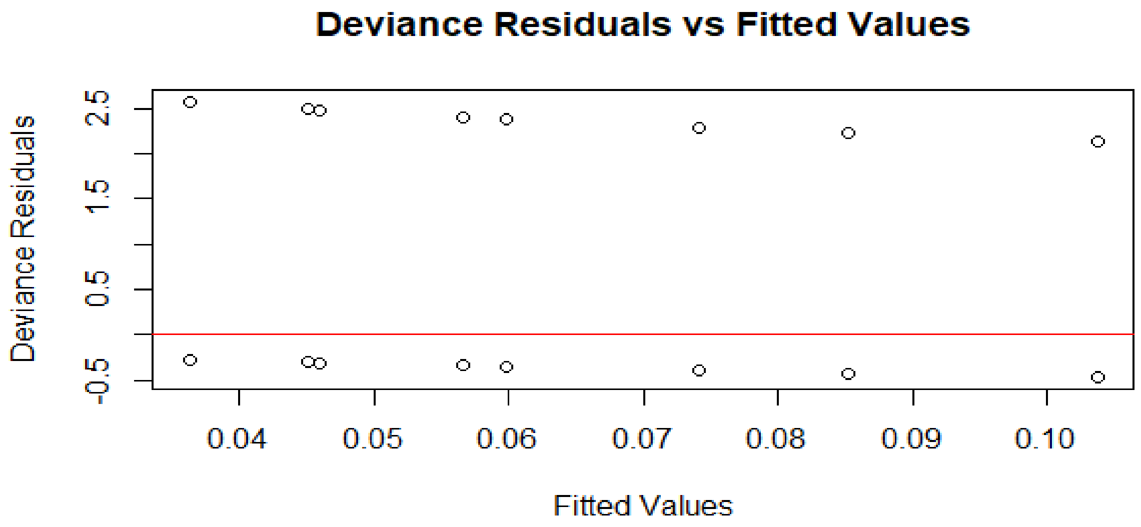
Now, I draw the plot of fitted values against quantile residuals to check constant variance. In this case, I calculate the fitted values of model nf9. I illustrate first 126 observations fitted values here.

```
         1          2          3          4          5          6          7 
0.07406658 0.03625748 0.05982026 0.07406658 0.05982026 0.04592253 0.05982026 
         8          9         10         11         12         13         14 
0.07406658 0.05982026 0.03625748 0.07406658 0.07406658 0.03625748 0.05982026 
        15         16         17         18         19         20         21 
0.05982026 0.07406658 0.05982026 0.07406658 0.07406658 0.04592253 0.07406658 
        22         23         24         25         26         27         28 
0.07406658 0.05982026 0.05982026 0.07406658 0.05982026 0.07406658 0.05982026 
        29         30         31         32         33         34         35 
0.04592253 0.03625748 0.07406658 0.05982026 0.04592253 0.07406658 0.07406658 
        36         37         38         39         40         41         42 
0.07406658 0.05982026 0.05982026 0.05982026 0.04592253 0.05982026 0.07406658 
        43         44         45         46         47         48         49 
0.05982026 0.05982026 0.05982026 0.04592253 0.07406658 0.05982026 0.07406658 
        50         51         52         53         54         55         56 
```

```
0.07406658 0.07406658 0.03625748 0.04592253 0.07406658 0.05982026 0.05982026 
        57         58         59         60         61         62         63 
0.05982026 0.05982026 0.07406658 0.05982026 0.04592253 0.04592253 0.05982026 
        64         65         66         67         68         69         70 
0.07406658 0.07406658 0.07406658 0.07406658 0.03625748 0.07406658 0.05982026 
        71         72         73         74         75         76         77 
0.05982026 0.05982026 0.07406658 0.07406658 0.05982026 0.03625748 0.05982026 
        78         79         80         81         82         83         84 
0.07406658 0.03625748 0.07406658 0.03625748 0.05982026 0.07406658 0.04592253 
        85         86         87         88         89         90         91 
0.07406658 0.07406658 0.05982026 0.05982026 0.07406658 0.05982026 0.07406658 
        92         93         94         95         96         97         98 
0.07406658 0.05982026 0.05982026 0.04592253 0.07406658 0.03625748 0.05982026 
        99        100        101        102        103        104        105 
0.07406658 0.07406658 0.05982026 0.05982026 0.04592253 0.05982026 0.07406658 
       106        107        108        109        110        111        112 
0.07406658 0.03625748 0.05982026 0.07406658 0.07406658 0.05982026 0.05982026 
       113        114        115        116        117        118        119 
0.05982026 0.05982026 0.05982026 0.05982026 0.07406658 0.07406658 0.07406658 
       120        121        122        123        124        125        126 
0.03625748 0.05982026 0.05982026 0.05982026 0.05982026 0.05982026 0.05982026 
```

Later, I draw a plot of fitted values vs deviance residuals and quantile residuals. We know any trends appearing in these plots indicate that the systematic component can be improved. On the other hand, if the random component is correct, the variance of the points is approximately constant. From these plots, we do not denote any trend. In this reason, the model (nf9) is adequate.





Goodness of Fit test

For binomial generalized linear model, I use Hosmer-Lemeshow test for checking goodness-of-fit of the model. It compares the observed number of events with the expected number of events based on the model's predictions. The p-value is used to test the null hypothesis that there is no significant difference between observed and expected outcomes. A low p-value (typically < 0.05) suggests rejection of the null hypothesis, indicating poor fit. A high p-value suggests that the model fits the data well.

p-value	0.5597

We get higher p-value (>0.05) from the Hosmer-Lemeshow tests. Now, we can say the model's predictions align with the observed outcomes. This indicates that the model fits the data well and that there's no significant deviation.

Odds Ratio

Odds ratios to convey the relationship between predictors and the binary outcome. An odds ratio (OR) quantifies the odds of an event occurring in one group relative to another group. Odds Ratio (>1) indicates an increased likelihood of the event occurring in one group compared to the reference group. Moreover, Odds Ratio (< 1) describes a decreased likelihood of the event occurring in one group compared to the reference group. Furthermore, Odds Ratio (= 1) denotes no difference in the odds between groups.

Variables	OR	Reference Level
GenderMale	0.90	GenderFemale
Subscription1	0.72	Subscription0
Country1	0.65	Country0
Country1:Subscription1	1.76	Country0:Subscription0

Table 5: Odds Ratio

From Table 5, we describe the odds that male subscriber subscription retention rate is 10% lower than the female subscriber subscription continuation rate. Moreover, the odds that subscription continuation rate is 28% lower for standard and premium subscriber rather than basic user. Subsequently, the odds that subscription continuation rate is 35% lower for Country1 (US, UK, France, Spain, Mexico, Germany, Australia) location subscriber rather than Country0 (Canada, Brazil, Italy) Netflix subscriber. In addition, the odds that subscription continuation rate is 76% higher for standard and premium subscriber who are lives in Country 1 rather than basic user who are staying Country0.

Limitations:

The Netflix userbase dataset based on only demographic and subscription variables. There is not any content related data in this dataset. If the dataset have content data (Genre, content duration), then it may be impact on model significancy and provide more clear image of consumer subscription retention. In this dataset have only 2500 observation and the response variable is highly imbalanced.

Conclusion:

This project aims to leverage predictive analytics to forecast Netflix subscription continuation and provide actionable insights for optimizing user retention strategies. Specifically, constructing an optimum generalized linear model with the help of R language to predict Netflix subscription retention. According to the EDA analysis, the basic subscription plan consumer percentage is higher than standard & premium customers. Moreover, I found better generalized linear model by AIC when the response variable is active (account status) and the predictor variables are gender, country*subscription. According to score test, the explanatory variables of the models are significant. In addition, the diagnostics (Q-Q plot, Residuals vs Fitted values plot, Goodness-of-fit test) indicates the model performance is well. Overall, these results offer a guideline for predicting subscriber retention of Netflix with influential predictor variables and enhance user-centric strategies.

Reference:

[1] Wang, G., Wang, Z., & Xie, Y. (2022). Subscribers Forecasting of Netflix Based on Multiple Linear Models. BCP Business & Management, 34, 229-236. https://doi.org/10.54691/bcpbm.v34i.3018.

[2] https://www.kaggle.com/datasets/arnavsmayan/netflix-userbase-dataset.

[3] cudilio, J., Pereira, G.H.A. (2020). Adjusted quantile residual for generalized linear models. Comput Stat 35, 399–421. https://doi.org/10.1007/s00180-019-00896-w

[4] Paramasivan, K. A., Ying, J. K. K., Abidin,I. N. S. B. Z., Wenji, J., KG, A. (2023).Subscription and Customer Loyalty: A Study of Netflix Before and After Covid-19 Pandemic. Asia Pasific Journal of Management and Education, 6(3), 129-138. https://doi.org/10.32535/apjme.v6i3.2674.

[5] Gusti Nyoman Wiradarma, P.S. Piartrini (2022). Understanding netflix subscription intentions using technology acceptance model: a study in denpasar. Russian Journal of Agricultural and Socio-Economic Sciences, 121(1):37-44. 10.18551/rjoas.2022-01.05.

[6] Kumar J., Gupta A., Dixit, S. (2020). Netflix: SVoD entertainment of next gen. Emerald Emerging Markets Case Studies, Vol. 10 No. 3.

[7] Natalie Sherman, Streaming Netflix's subscribers and profits plummet for the first time in a decade Five reasons behind. (April 21, 2022). Retrieved from: https://www.bbc.com/zhongwen/simp/business-61161567