PROJECT I 2022-2023:

Predict Churn Behavior of Customers from a telecommunication company

Panagiotis G. Vaidomarkakis (p2822203)

Using the provided excel file churn.xls

Tutor: Dimitris Karlis

19/02/2023

Table of Contents

Abstract..2

Introduction..3

Descriptive analysis and exploratory data analysis..4

Creating a predictive model..7

Model Assumptions..10

Interpretation of the model..14

Abstract

In our case, data were collected through a telecommunication company. The whole dataset contains approximately 3333 rows, both with the usage characteristics as well as demographic data for customers. We have 15 numerical variables for call duration, call charge, etc. as far as 6 categorical (factorial) variables about churn status (which is the response for our case), gender, state, etc. Moreover, we have none missing attribute values. After pre-processing the training data, we tested several different prediction models to arrive at the final prediction model which is the following:

Logit(churn) = -8.049 + 0.00717* Eve Mins + 0.513* CustServ Calls +2.042* Int’l Plan (if customer has Int’l Plan) - 0.938* VMail Plan (if customer has VMail Plan) + 0.0765* Day Charge + 0.0815* Night Charge -0.0914* Intl Calls + 0.324* Intl Charge

In the end of this assignment, we made some test in order to evaluate it.

Introduction

In our dataset, we have churn variable which corresponds to if a customer has churned or not churned (values are 0 or 1) and it is the response variable of our model that we will try to present later on.

All the other variables are the following:

- Account Length
- VMail Message
- Day Mins
- Eve Mins
- Night Mins
- Intl Mins
- CustServ Calls
- Int’l Plan
- Vmail Plan
- Day Calls
- Day Charge
- Eve Calls
- Eve Charge
- Night Calls
- Night Charge
- Intl Calls
- Intl Charge
- State
- Area Code
- Gender

Descriptive analysis and exploratory data analysis

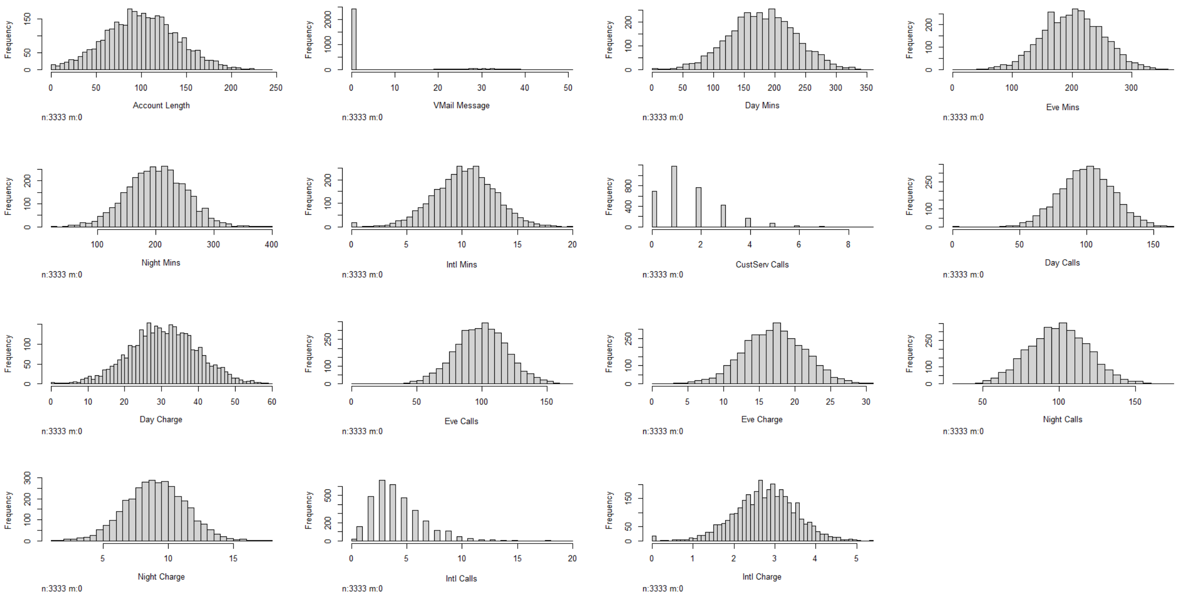
In order to select which variables are the most significant for a customer to be churner, we need a preliminary descriptive analysis to gain some initial insights.

After importing the dataset in RStudio, we make a summary in order to see the nature of the dataset. Below, you can see the result:

```
Account Length  VMail Message     Day Mins       Eve Mins       Night Mins      Intl Mins     CustServ Calls
Min.   :  1.0   Min.   : 0.000   Min.   :  0.0   Min.   :  0.0   Min.   : 23.2   Min.   : 0.00   Min.   :0.000
1st Qu.: 74.0   1st Qu.: 0.000   1st Qu.:143.7   1st Qu.:166.6   1st Qu.:167.0   1st Qu.: 8.50   1st Qu.:1.000
Median :101.0   Median : 0.000   Median :179.4   Median :201.4   Median :201.2   Median :10.30   Median :1.000
Mean   :101.1   Mean   : 8.099   Mean   :179.8   Mean   :201.0   Mean   :200.9   Mean   :10.24   Mean   :1.563
3rd Qu.:127.0   3rd Qu.:20.000   3rd Qu.:216.4   3rd Qu.:235.3   3rd Qu.:235.3   3rd Qu.:12.10   3rd Qu.:2.000
Max.   :243.0   Max.   :51.000   Max.   :350.8   Max.   :363.7   Max.   :395.0   Max.   :20.00   Max.   :9.000

Churn      Int'l Plan VMail Plan   Day Calls      Day Charge      Eve Calls      Eve Charge     Night Calls
No :2850   No :3010   No :2411   Min.   :  0.0   Min.   : 0.00   Min.   :  0.0   Min.   : 0.00   Min.   : 33.0
Yes: 483   Yes: 323   Yes: 922   1st Qu.: 87.0   1st Qu.:24.43   1st Qu.: 87.0   1st Qu.:14.16   1st Qu.: 87.0
                                 Median :101.0   Median :30.50   Median :100.0   Median :17.12   Median :100.0
                                 Mean   :100.4   Mean   :30.56   Mean   :100.1   Mean   :17.08   Mean   :100.1
                                 3rd Qu.:114.0   3rd Qu.:36.79   3rd Qu.:114.0   3rd Qu.:20.00   3rd Qu.:113.0
                                 Max.   :165.0   Max.   :59.64   Max.   :170.0   Max.   :30.91   Max.   :175.0

 Night Charge      Intl Calls      Intl Charge        State        Area Code     Gender
Min.   : 1.040   Min.   : 0.000   Min.   :0.000   WV      : 106   408: 838   Female:1662
1st Qu.: 7.520   1st Qu.: 3.000   1st Qu.:2.300   MN      :  84   415:1655   Male  :1671
Median : 9.050   Median : 4.000   Median :2.780   NY      :  83   510: 840
Mean   : 9.039   Mean   : 4.479   Mean   :2.765   AL      :  80
3rd Qu.:10.590   3rd Qu.: 6.000   3rd Qu.:3.270   OH      :  78
Max.   :17.770   Max.   :20.000   Max.   :5.400   OR      :  78
                                                 (Other):2824
```

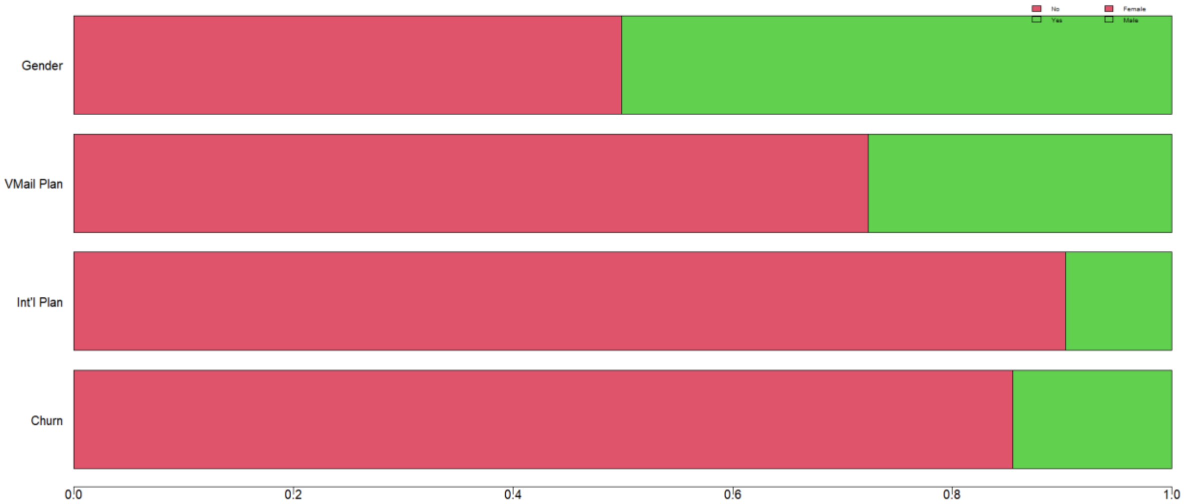
As we can see, from 3333 customers, only 483 have churned, 323 have Int’l Plan and 922 have Vmail Plan. Males and Females are equally distributed in our case, so gender doesn’t seem to be significant variable for our churn model.

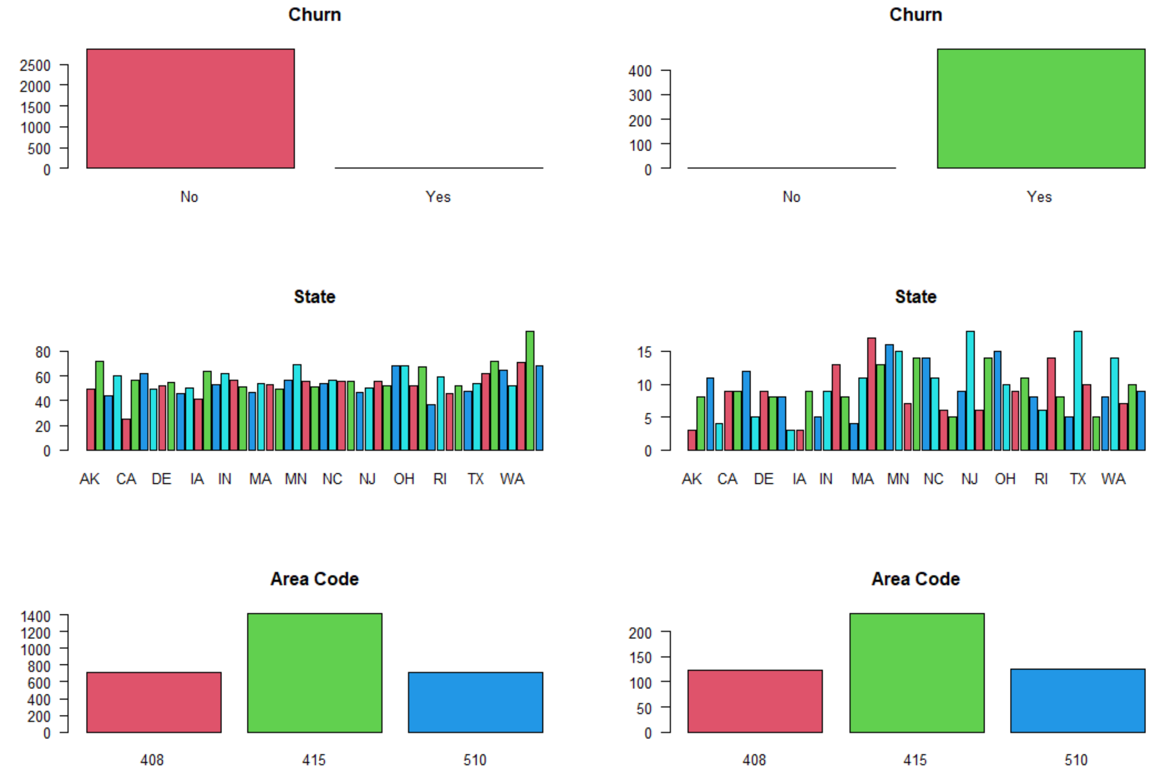
We need to visualize both our numeric and categorical variables in order to have more insights of which are significant and which aren’t. Below, you can see RStudio’s results:



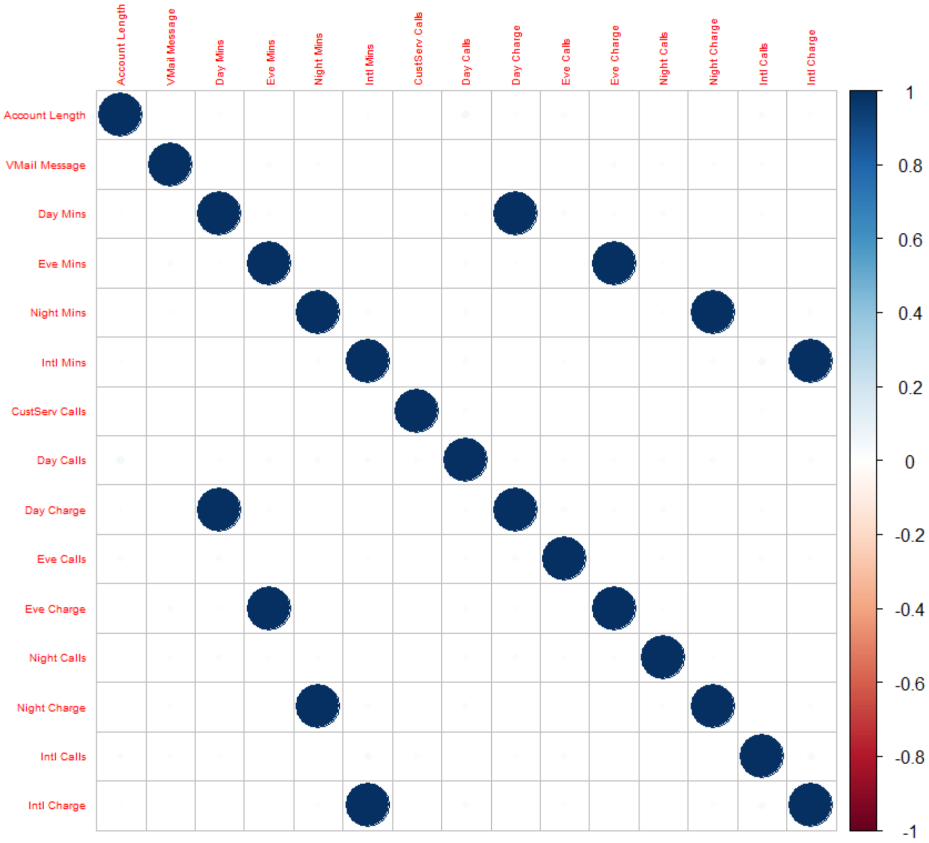
Most of our numerical variables seems to follow normal distribution. Vmail Message seems to have most zeros and a few numbers in range (18-38). CustServ Calls seems to have discrete values from 0 to 8 with most calls to be 1.

Below, you will see all categorical variables using bar plots:





The first bar plot shows the categorical variables with 2 possible outcomes. The other bar plots show the categorical variables with more than 2 possible outcomes and are divided in 2 pieces. The first "column" shows State and Area Code of customers who haven't churned and the second "column" shows exactly the same variables for people who have churned. We can see that gender is equally distributed in our dataset so it might not be significant in the final model. We can also see that Area Code has the same shape, both in churned and not churned customers. As for State, we have more customers who hasn't churned so it is logical for having higher bar plots than the bar plots next to them.



From the correlation matrix above, we can see that each of Day Charge, Eve Charge, Night Charge and Intl Charge is fully positive correlated with Day Mins, Eve Mins, Night Mins and Intl Mins (which seems reasonable). So, we have a hint that in the Final Model, only one from each pair will be as the share the same information.

Creating a predictive model

In our case, we will use a generalized linear model in order to determine the Churn variable, which is corresponds to if a customer is a churner or not. We need an appropriate distribution in order to model the categorization of a customer. It seems that Binomial distribution is the most appropriate to choose, because it models the possibility of “success” when the outcome is between two distinct categories. It seems like it fits to our case so we will continue with that.

After we execute the full GLM model for a binomial distribution, we have a first guess of which variables are statistically important for the churn variable, the response. With the term statistically important, we mean that with the increase or decrease of one unit in a value of one of the variables in the model, it can modify the possibility of a customer to be churner or not.

These variables which our model indicated as significant are below:

- VMail Message
- CustServ Calls
- Int’l Plan
- VMail Plan
- Intl Calls
- State

Because we run the full model, we took into account all the variables of our dataset and not all of them seem to be significant, we need to make a variable “screening” method for choosing only the most valuable variables and the execute a “leaner” model.

Lasso is the variable screening method with relied on λ, a specific parameter which shows how strict the method is going to be. In a few words, the value of λ indicates the number of significant parameters our model is going to have. The goal is to minimize the λ value but when you minimize the λ, you still have many parameters and our task is to keep only the significant ones. We will choose the λ based on the λ_{min} but we will keep the λ with one standard error away from the minimum, the λ_{lse}. We will use glmnet package for lasso.

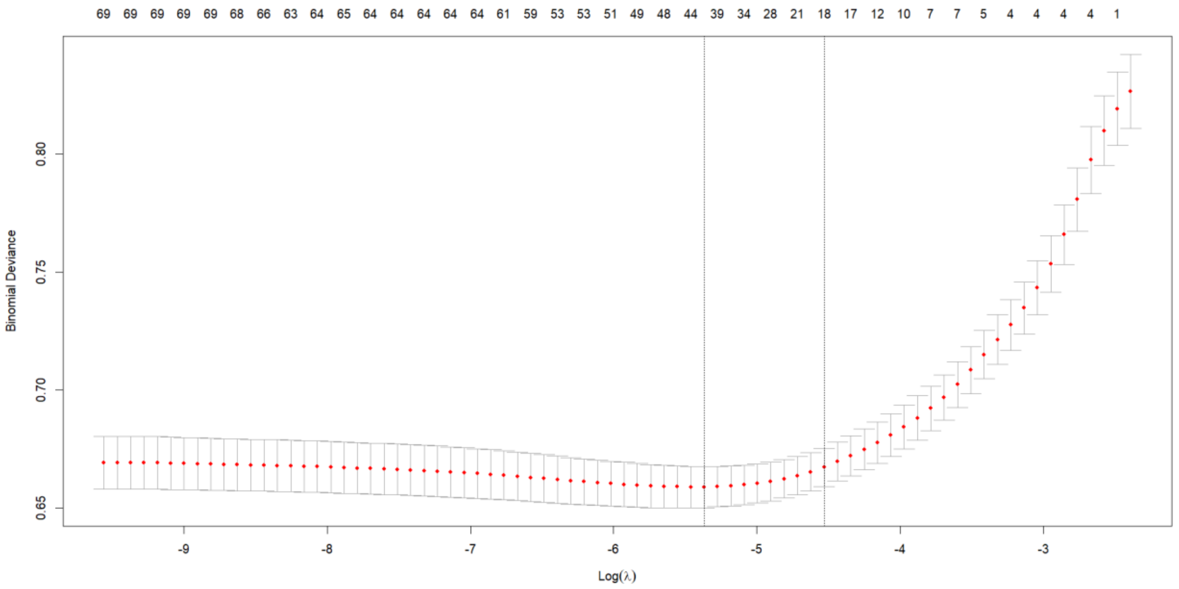
These are the variables that Lasso has selected from the screening:

- Day Mins
- Eve Mins
- Night Mins
- Intl Mins
- CustServ Calls
- Int’l Plan

- VMail Plan
- Day Charge
- Eve Charge
- Night Charge
- Intl Calls
- Intl Charge
- State

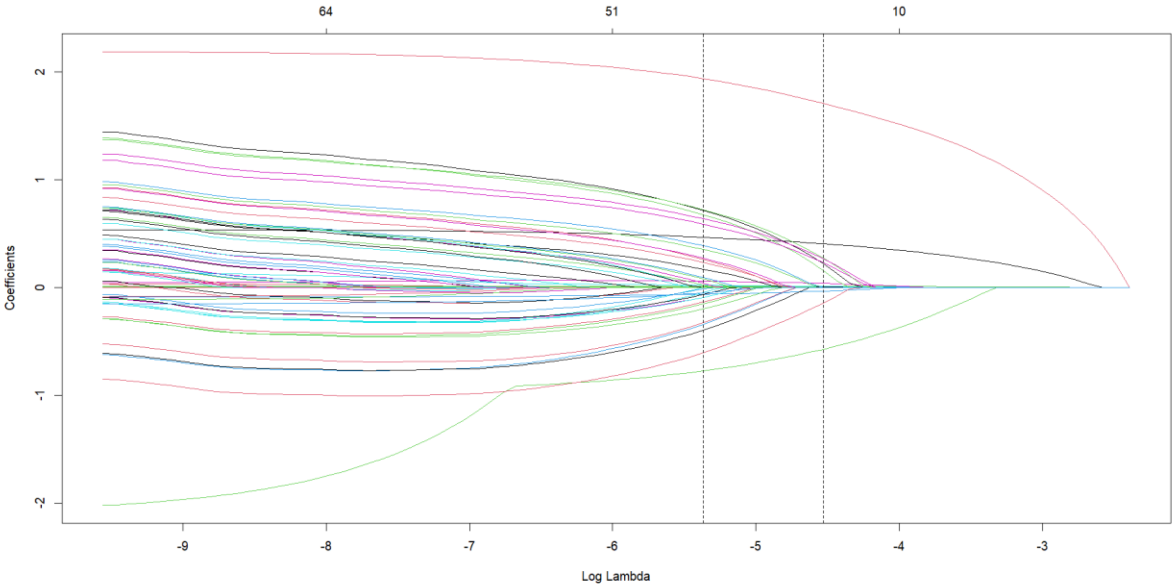
λ_{min} was 0.004659692 but we chose λ_{lse} which was 0.01076449 in order to have less variables.

Below we will see how λ was selected:



The log(λ_{min})= -5.368806(the first line) while the one that we kept in order to continue is log(λ_{lse})= -4.531502(the second line).

Below, you will see another graph which indicates the number of variables the model will have after lasso using different λ values. When a line goes to 0, then this variable will not be in the model.



As we can see, we have significant less variables in the second line compared to the first line.

After Lasso, we have to choose which of the variables that Lasso screened are the most important variables. For this purpose, a stepwise method is needed. AIC will be our final judge. In every step of this procedure, we add or subtract variables from the model in order to minimize AIC score, which is an indicator of keeping the significant variables.

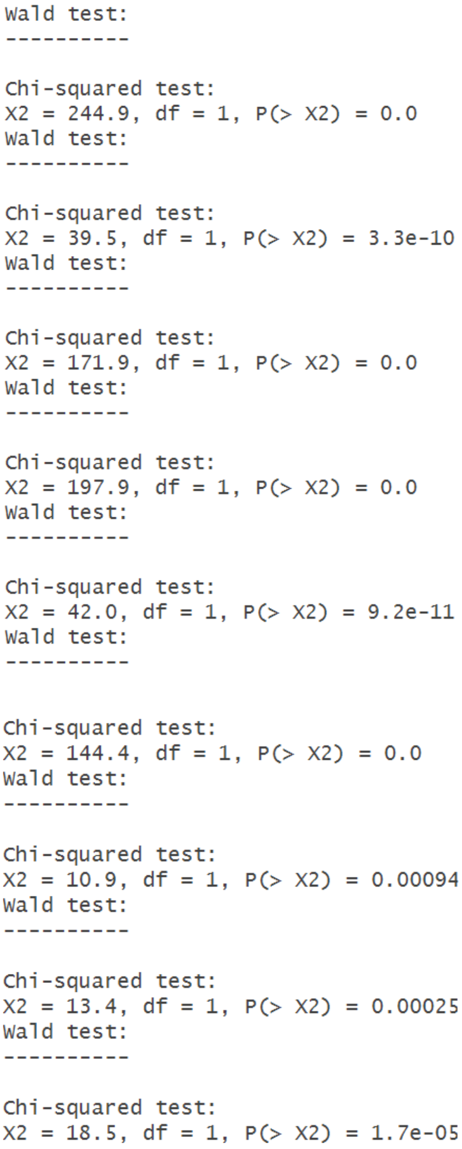
After running the stepwise procedure, the model contains the following variables:

- Eve Mins
- CustServ Calls
- Int'l PlanYes
- VMail PlanYes
- Day Charge
- Night Charge
- Intl Calls
- Intl Charge

Between the variables that have been deemed important for churn, we can point out that the pairs that we saw at the correlation matrix aren't here as we thought that they weren't going to be because they share the same information. Moreover, gender isn't in our model as a parameter as we have pointed out before.

Model Assumptions

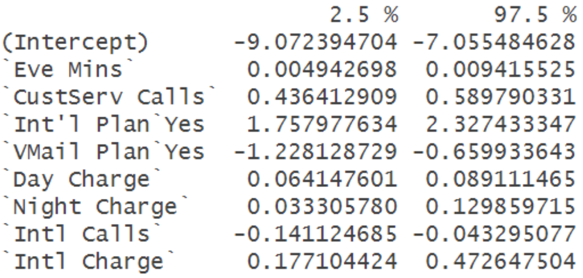
To validate the model and the selected important variables for churn, specific conditions must be met. The first condition involves disproving the null hypothesis that assumes all coefficients are equal to zero. This is evaluated using the Wald test, where the null hypothesis is that the coefficients (β0, β1, β2, ..., βn) for each explanatory variable are zero, implying that these variables have no impact on whether a customer churns or not. The test results showed that the p-value for each variable in the model was below 0.05, indicating that we can reject the null hypothesis at a 95% confidence level. This suggests that a unit increase or decrease in the respective variable will impact (either increase or decrease) the logit probability of a customer churning or not.



```
Wald test:
----------

Chi-squared test:
X2 = 244.9, df = 1, P(> X2) = 0.0
Wald test:
----------

Chi-squared test:
X2 = 39.5, df = 1, P(> X2) = 3.3e-10
Wald test:
----------

Chi-squared test:
X2 = 171.9, df = 1, P(> X2) = 0.0
Wald test:
----------

Chi-squared test:
X2 = 197.9, df = 1, P(> X2) = 0.0
Wald test:
----------

Chi-squared test:
X2 = 42.0, df = 1, P(> X2) = 9.2e-11
Wald test:
----------


Chi-squared test:
X2 = 144.4, df = 1, P(> X2) = 0.0
Wald test:
----------

Chi-squared test:
X2 = 10.9, df = 1, P(> X2) = 0.00094
Wald test:
----------

Chi-squared test:
X2 = 13.4, df = 1, P(> X2) = 0.00025
Wald test:
----------

Chi-squared test:
X2 = 18.5, df = 1, P(> X2) = 1.7e-05
```

After selecting the model, it is necessary to assess its goodness of fit. This is accomplished by employing a chi-square-based goodness of fit test to determine how well the model aligns with the data. In this case, the test returned a value of 1, which indicates that the null hypothesis proposed by the model regarding customer churn is not rejected. As a result, it can be concluded that the selected model fits the data well and effectively explains churn behavior.

Below we see how the coefficients are changing compared to significant level. As we can see, we don’t have major changes in the numbers.



```
                        2.5 %       97.5 %
(Intercept)      -9.072394704 -7.055484628
`Eve Mins`        0.004942698  0.009415525
`CustServ Calls`  0.436412909  0.589790331
`Int'l Plan`Yes   1.757977634  2.327433347
`VMail Plan`Yes  -1.228128729 -0.659933643
`Day Charge`      0.064147601  0.089111465
`Night Charge`    0.033305780  0.129859715
`Intl Calls`     -0.141124685 -0.043295077
`Intl Charge`     0.177104424  0.472647504
```

- Deviance of the null model: 2758.3
- Deviance of the selected model: 2165.5

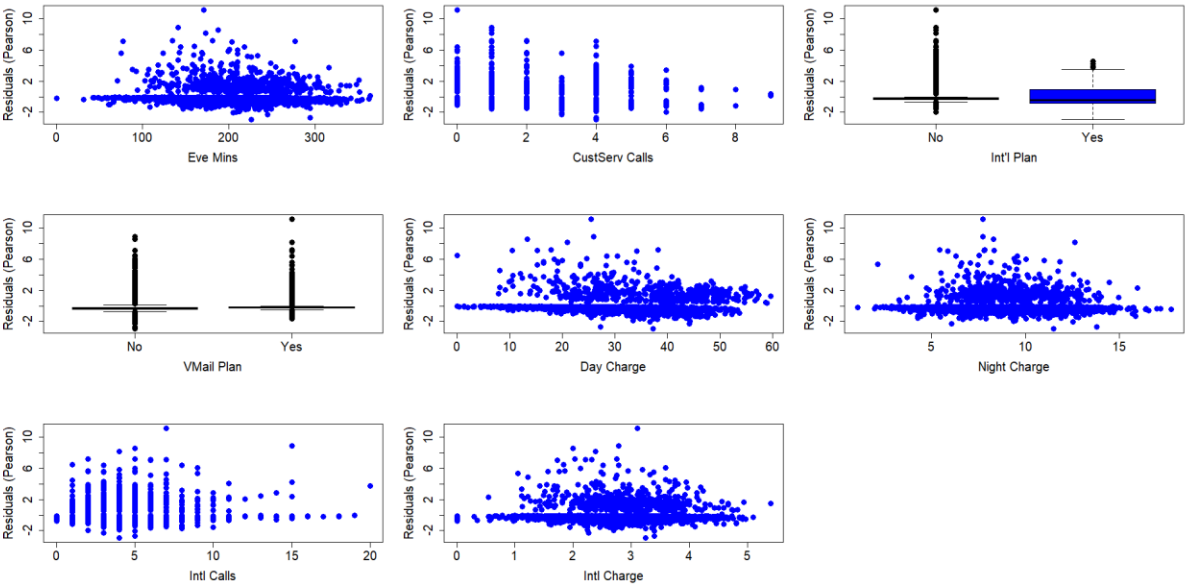
We have succeeded to reduce deviance by 592.7751.

Moreover, using X^2 as an indicator in order to see if our model is better from null model, we subtracted null deviance and our deviance as far as null residuals and our residuals. The result was 8.362939e-123 which shows significant difference between our model and null model.

After analyzing the selected model in comparison to the null model, it is evident that the deviance of the selected model is lower than that of the null model. In addition to this, a direct comparison between the two models was performed to assess whether the additional variables in the selected model significantly contribute to explaining churn, as compared to a model with no variables. The results of this test indicated that the p-value was less than 0.05, indicating that at a 95% confidence level, the selected model is different from the null model. This implies that the coefficients of all covariates in the selected model are not equal to zero, which confirms the findings of previous tests.

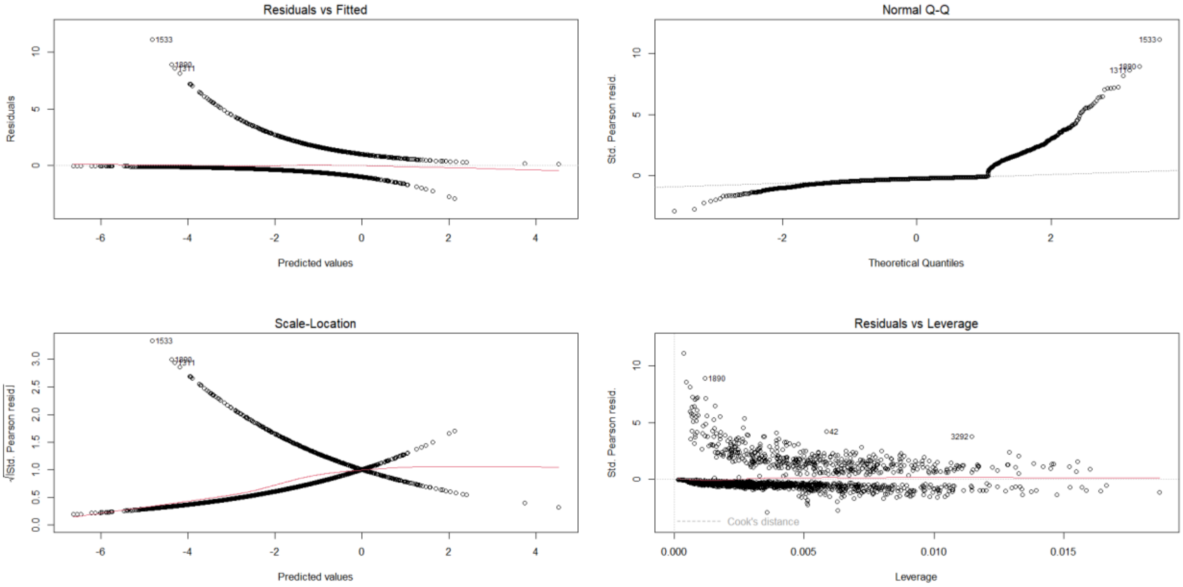
To evaluate the appropriateness of the selected model, it is essential to examine the assumptions for residuals, which are measured by Pearson and Deviance residuals. Both types of residuals represent the difference between the observed response variable and the corresponding predicted value from the model.
Pearson residuals are computed based on the variance function, whereas Deviance residuals are calculated based on the deviance function, which measures the difference between the maximum possible value of the log-likelihood function of a perfectly fitted model and the log-likelihood under the selected model.
To assess the overall goodness of fit of the model, it is crucial to scrutinize the assumptions for both types of residuals. Firstly, it is essential to check for any patterns between the residuals and each explanatory variable in the selected model. Additionally, both types of residuals must be normally distributed, have a mean of zero, and maintain constant variance.

Pearson Graphs:



From the plots presented above, it can be observed that the residuals do not follow any discernible pattern, suggesting that they are randomly distributed and uniformly spread around zero.

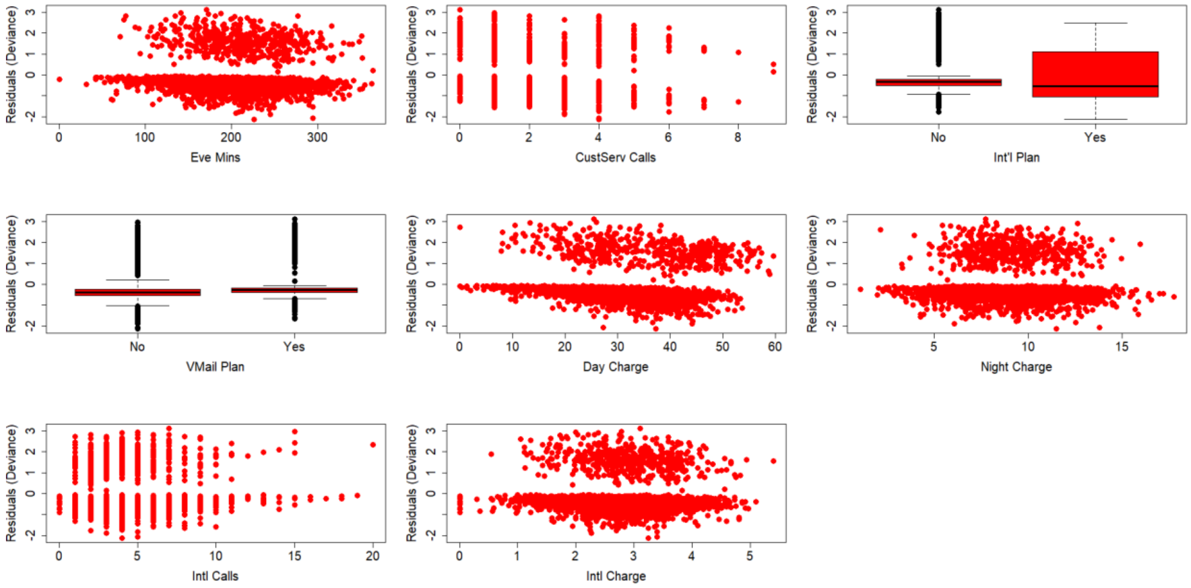
To verify the normality assumption of the residuals, a QQ plot was constructed, which is displayed below(2nd graph):



The QQ plot indicates that the residuals do not conform to a normal distribution, which suggests that the model's goodness of fit is less than satisfactory.

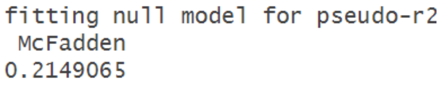
As far as the deviance of the model, we can see below the graphs in order to have some assumptions.

Deviance Graphs:



From the graphs presented above, it is clear that the residuals exhibit no specific pattern, indicating that they are randomly and evenly dispersed around zero.

In the end of our analysis, we calculate a pheudo-R^2 indicator for our model comparing it with the null model using McFadden's R^2 and the result is below:



```
fitting null model for pseudo-r2
 McFadden
0.2149065
```

As we can see, we have explained only 21,5% of the deviance compared to the null model and that our model doesn't fit the data well and doesn't have high predictive power.

Interpretation of the model

The model below represents the most important variables associated with customer churn as determined by the methodology used previously:
Logit(churn) = -8.049 + 0.00717* Eve Mins + 0.513* CustServ Calls +2.042* Int’l Plan (if customer has Int’l Plan) - 0.938* VMail Plan (if customer has VMail Plan) + 0.0765* Day Charge + 0.0815* Night Charge -0.0914* Intl Calls + 0.324* Intl Charge

Comments about the model above at its initial stage:

Given the previous tests conducted, it can be inferred that none of the coefficients for the covariates are equal to 0. Thus, if there is a unit increase or decrease (in the case of numeric variables) or the presence or absence (in the case of categorical variables) of any of the covariates, it is expected to affect the log odds of a customer to churn.

Interpretation of the numerical variables in the model is as follows:

- Based on the absolute value of their coefficients, the most significant numeric variables for predicting churn are CustServ Calls and Intl Charge.
- A unit increase in CustServ Calls, assuming all other variables are constant, is expected to increase the log odds of churn by 0.513.
- A unit increase in the Intl Charge, assuming all other variables are constant, is expected to increase the log odds of churn by 0.324.

Although the absolute values of the coefficients for the other numeric variables are smaller, they are still important predictors of churn. Their interpretations are:

- A unit increase in Eve Mins, assuming all other variables are constant, is expected to increase the log odds of churn by 0.00717.
- A unit increase in Day Charge, assuming all other variables are constant, is expected to increase the log odds of churn by 0.0765.
- A unit increase in Night Charge, assuming all other variables are constant, is expected to increase the log odds of churn by 0.0815.
- A unit increase in Intl Calls, assuming all other variables are constant, is expected to decrease the log odds of churn by 0.0914.

Interpretation of the categorical variables in the model is as follows:

- Choosing an Int’l Plan is expected to increase the log odds of a customer to churn by 2.042.
- Choosing a VMail Plan is expected to decrease the log odds of a customer to churn by 0.938.

Further Comments:

Upon examining the final model proposed before, it is apparent that the variables deemed significant in the preliminary tests are included in the model. Nevertheless, the magnitude of their effect on the log odds of churning varies. For example, although variables related to charges, such as Day and Night Charge, are important, they do not have as much impact as the CustServ Calls variable, which substantially increases the log odds of a customer churning. Similarly, the variable Intl Charge has the second-largest coefficient in absolute value, and it is reasonable to assume that a high initial charge may prompt a customer to switch providers.

Remarkably, the two categorical variables in the model have a significant impact on the log odds of churning. Opting for an initial line plan raises the log odds of churning by the most substantial amount. This suggests that customers who are locked into an unsatisfactory initial plan or lack the means to switch providers are more likely to churn. In contrast, choosing a VMail Plan reduces the log odds of churning, indicating that customers who receive regular communication and engagement through email campaigns are less likely to churn than those who do not opt for the plan.

It is crucial to note, however, that the model with the selected covariates does not yield significantly better results than the null model (without covariates). This implies that either more data is needed to re-fit the model or, more importantly, a dataset with a broader range of variables is required to determine all the crucial factors influencing customer churn.