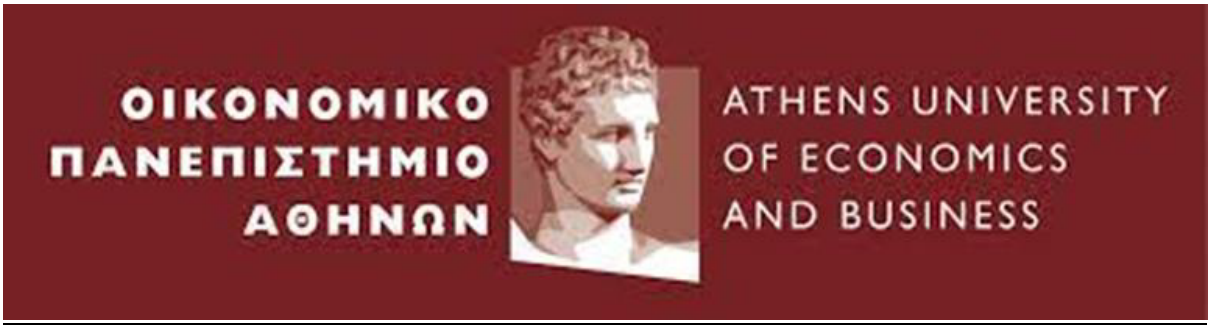


ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
DEPARTMENT OF MANAGEMENT, SCIENCE & TECHNOLOGY
MSc BUSINESS ANALYTICS

Statistics for Business Analytics II

"Churn of customers from a telecommunications company"

Full Name: ATHANASIOS ALEXANDRIS
Register Number:p2822202

ATHENS, 2023

TABLE OF CONTENTS

1.INTRODUCTION

1.1 Description of the problem

The data of this assignment refer to the characteristics of customers of a telecommunications company, regarding their use of services of the telecommunication company for the previous period and their demographic information. The main variable of the study is the churn, and more particularly we are interested to identify the ingredients that lead customers of the telecommunications company to churn.

1.2 DATASET

The dataset we used for our study contained 3.333 records and 21 variables. The dependent variable of the study is if the customer leaves the company or not (‘churn’), and there are 20 independent variables.

2. EXPLORATORY/DESCRIPTIVE ANALYSIS

2.1 Data cleaning

Before we start our analysis we had to check if our dataset need to be cleaned and modified, in a way that would help us to our next steps to make an effective and sufficient analysis. Below are listed all actions we did regarding the data cleaning :

1.We checked for blanks, NAs, NaN or infinite values, but we did not find any of these.
2.We set the categorical columns, and the numerical columns that were nominal as factors.
By the end of the cleaning 21 variables had been remained to our dataset, 15 numerical variables and 6 factors. Thus we decided to create two new data frames, one for the numerical variables and one for the factors.

2.2 Visualizations

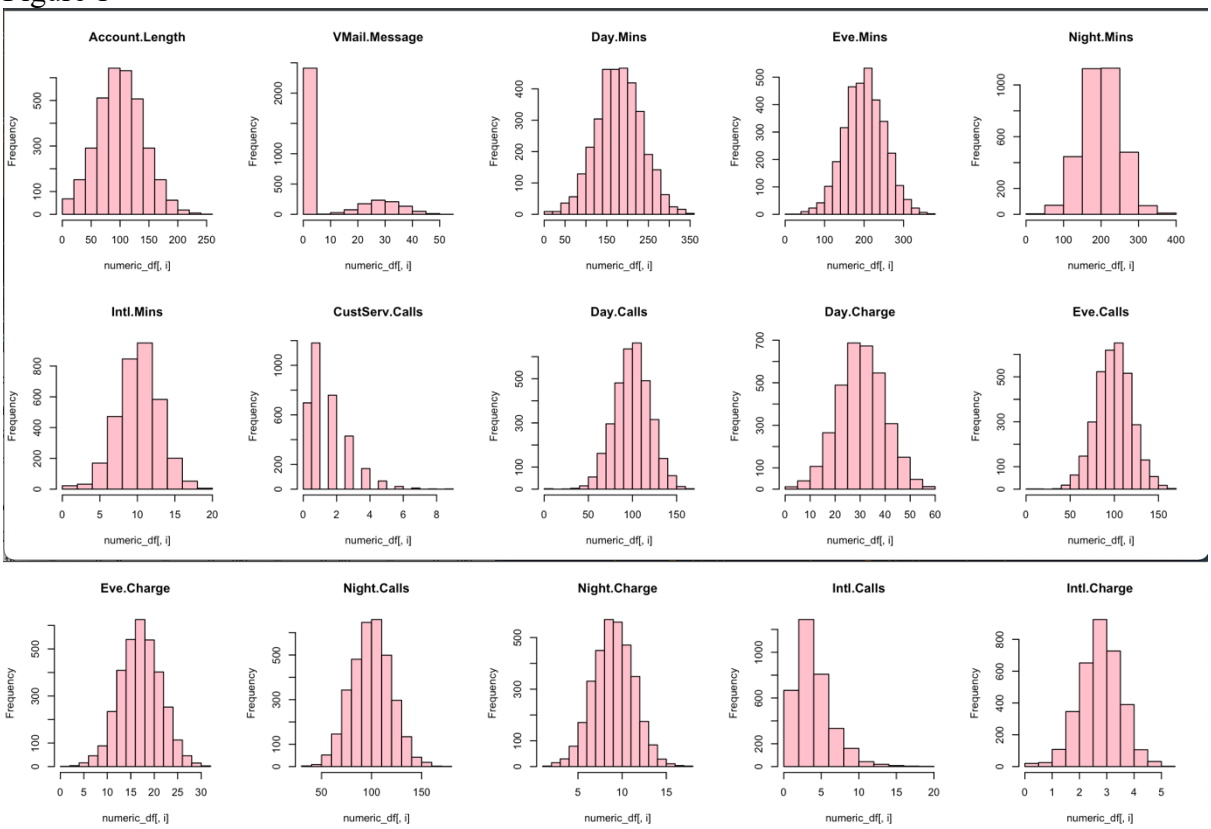
2.2.1 Numerical variables

Regarding the numerical variables, we made histograms to understand the type of distribution and the symmetry of each variable. By observing Figure 1 we can assume the following:

1. Most variables seem to follow normal distribution. More particularly the variables that are related to everyday use services like the minutes of day-evening-night calls, the charge of them or their number of calls. Normal distribution also seem to be followed by the variable that indicates the duration of the account for each customer.

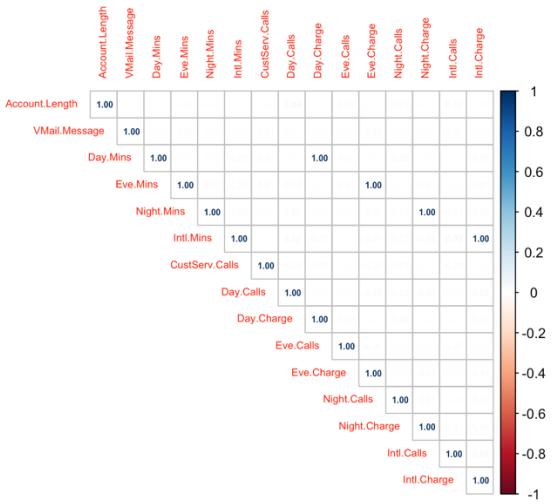
2.On the other hand rest numerical variables seem to have asymmetry and significant positive/negative skewness. Those variables are associated with services that customers use to use more rarely, like the minutes of international calls, the number of voice mail messages or the calls that the customers made to the customer service of the company.

Figure 1



Next we designed a correlation plot for the numeric variables, to identify if there is any high association between our variables. As we can observe from Figure 2, there is high absolute correlation only between the minutes and the charge for the all types of those variables (day, eve, night, international).

Figure 2



2.2.2 Factor variables

In Figure 3 and Figure 4, we have designed barplots for the factor variables with two levels and the factors that have more than two levels, respectively. From these barplots, we can understand the approximate number the customers that have the respective attributes, and whether the customers have churn or not.

Figure 3.

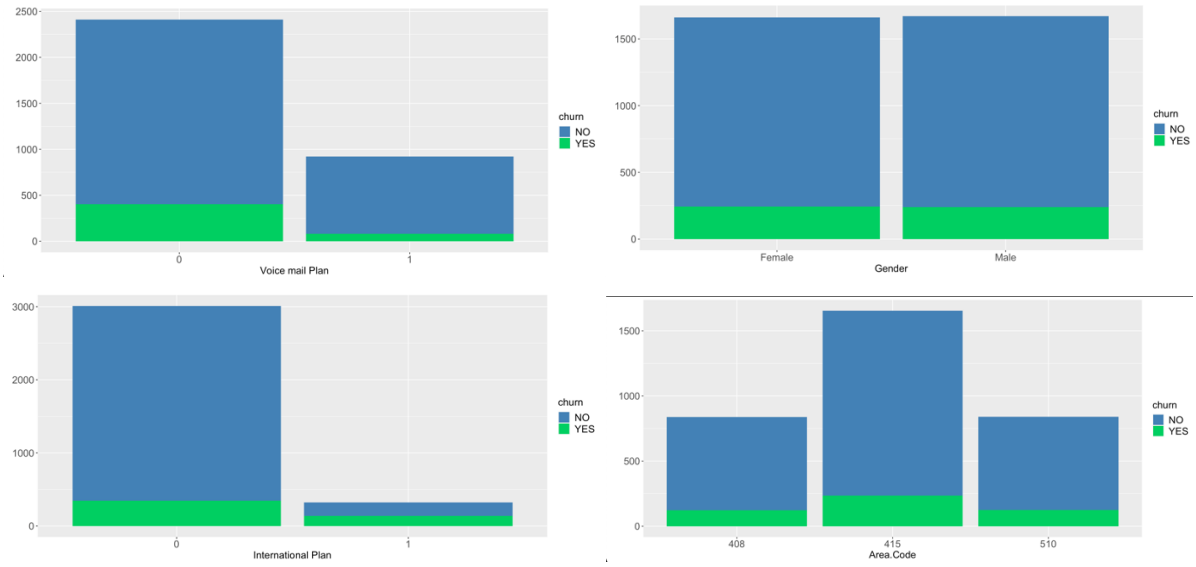
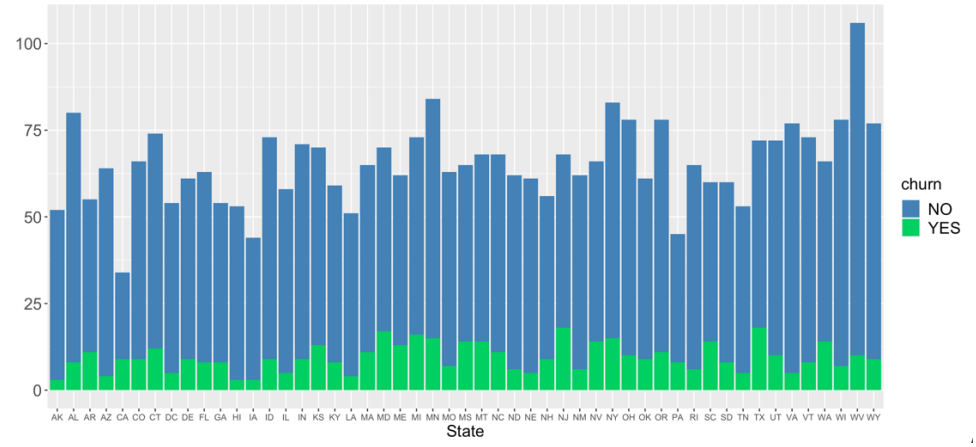


Figure 4.



Regarding the voice mail plan, it seems that the number of customers who had this program is very low. We also observe that for both categories the number of the customers that have left the company are is very low in comparison to those who have not.

The gender of the customer does not seem to play a significant role to our analysis, as the number of males and females seem to be equal. Also, we observe that for both genders, the

number of customers is much lower than those who have not and the proportion for those genders seem to be the same.

Regarding the International calls plan, it seems that the number of customers who had this program is approximately 10% of our sample. We observe that the number for the category of customers that they do not participate to the program, the proportion of churn is very low. However, for the other category the proportion of the customers who leaved the company in comparison to those who have not, seem to be equal.

Customers from the area with code '415' seem to be approximately the 50 % of our observations, while the two other areas seem to have equal number of customers. For all three areas, we observe that the proportion of the customers who leaved the company in comparison to those who have not is very low, approximately at 10-15%.

Finally, by observing the Figure 3 we can say that the state with the highest number of customers by far is 'WV'. For all states the proportion of the customers who leaved the company in comparison to those who have not is very low.

3.MODEL SELECTION & EVALUATION

3.1 Model Construction

The response variable in our study is binary (the customer leaves or not the company), so we decided to implement the logistic regression method. Our model will have the following form:

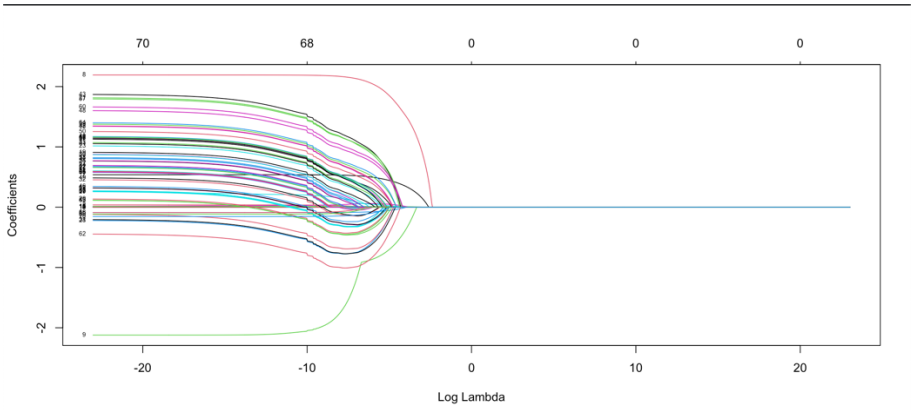
$\text{logit}(p) = b0 + b1X1 + b2X2 + ... + bnXn$

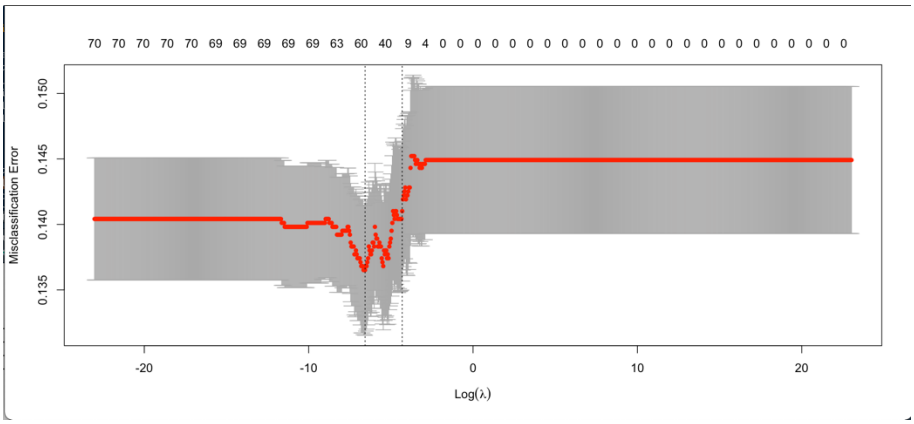
$\text{logit}(p) = \ln(p/(1-p))$

With the implementation of the GLM function in R, we constructed our first model, which involves all variables of our dataset. The summary of the first model (Table 1) indicates a great number of variables that are not statistically significant (Pr > 0.05 for Ho: b1=b2 = ... = bn = 0). The ratio $\frac{Residual\ Deviance}{\text{Degrees of freedom}}$ = 0.6342 and the AIC = 2210.9.

In order to select only the appropriate variables we used the LASSO method, as it performs well in large data sets and removing all the unnecessary variables without important loss of information.

By observing Figure 5 and Figure 6 below, we understand that the optimal Lasso λ parameter decide to keep around 15 variables of the original dataset, as we decided to use lambda.1se (right vertical line), as we can have a more clear model and more regularized model.





Then we used the outcome of the cross validation method, shown in Table 2, as input of the dependent variables of our new logistic model.

As in the case of our previous model, the summary of the lasso model (Table 3) indicates a great number of variables that are not statistically significant (Pr > 0.05 for Ho: b1=b2 = ... = bn = 0).

The ratio $\frac{Residual\ Deviance}{\text{Degrees of freedom}}$ = 0.6355 and the AIC = 2204.2 .

Then we implemented the BIC method, which is more appropriate for interpretation, in order to end up to the most effective model. We ended up with a model with 8 variables.

The summary of our last model (Table 4) has ratio $\frac{Residual\ Deviance}{\text{Degrees of freedom}}$ = 0.6514 and the AIC = 2183.5 .

Finally, we used VIF to understand if we have multicollinearity in our model. However, all coefficients of our model are below 10, so we accept them all (Table 5).

3.2 Model Evaluation

At this point we need to compare all models that we have constructed in the previous section, in order to state which model perform better to our dataset and gives the more clear view of the variables that play instrumental role for a customer to leave the company.

The criteria that should take into consideration for the model selection are the following:

AIC - The AIC (Akaike Information Criterion) provides an approximation of a constant value, as well as the difference between the unknown true likelihood function of the data and the probability density function of the model that has been fitted to the data. Therefore, a lower AIC value indicates that the model is more likely to accurately represent the true relationship between the variables, while a higher AIC value suggests that the model may not fit the data as well.

The ratio $\frac{Residual\ Deviance}{\text{Degrees of freedom}}$ - The closest to 1 the better model.

McFadden's pseudo R2 - Defined as 1−[LLk/LLnull)] where LLnull is the log-likelihood when only constant is present and LLk is the log-likelihood with k predictors. Values of pseudo R2 ranging from 0.2 to 0.4 indicates very good model fit.

Regarding AIC, the model that has the lowest value is the third model (2,183.5). For the initial model the AIC was 2,210.9 and for the second model 2204.2. Thus, from AIC scope the best model is the third one.

Regarding the ratio $\frac{Residual\ Deviance}{\text{Degrees of freedom}}$, the model that its value is closest to 1 is again the third one (0.6514). The ratio's values for the initial and the second model are 0.6342 and 0.6355 respectively.

The values of the McFadden's pseudo R2 (Table 6) are 0.2499 for the initial model, 0.2465 for the second model and 0.2149 for the third model. All values are between 0.2 and 0.4, which indicates good model fit for all models.

Based on all the above, and taking into consideration that for the third model, there is not multicollinearity problem, and all variables are statistically significant, so we believe that this is the model that describes better the factors that make a customer leave the telecommunication company.

4. MODEL INTERPRETATION

The model we have selected in the previous section has the following format:

logit(p) = -8.048873 + 0.007167*Eve.Mins + 0.512584*CustServ.Calls + 2.041929*Int.l.Plan1 -0.938145*VMail.Plan1 + 0.0765*DayCharge + 0.081465*Night.Charge -0.091430*Intl.Calls +0.32399*Intl.Charge

In logistic regression, the coefficients for the independent variables reflect the estimated change in the log odds of the outcome for a one unit increase in the corresponding independent variable, holding all other independent variables constant. The odds indicate the estimated probability of the customer to leave the telecommunications company. The positive and negative signs preceding each covariate in the model indicate the direction and magnitude of the estimated effect of a one-unit increase in the corresponding independent variable on the probability of churn.

The interpretation of our model is the following:

- For every one unit increase in Eve.Mins (number of minutes of evening calls), the log odds of churn increase by 0.007167, having all other covariates constant.
- For every one unit increase in CustServ.Calls (number of calls in customer services), the log odds of churn increase by 0.512584, having all other covariates constant.
- If the customer has an international plan the log odds of churn decrease by 6.006944 (-8.048873+2.041929), and all the other variables are equal to zero.
- If the customer has a voice mail plan the log odds of churn decrease by -8.987018 (-8.048873-0.938145), and all the other variables are equal to zero.
- For every one unit increase in DayCharge (charge for services during the day hours), the log odds of churn increase by 0.0765, having all other covariates constant.
- For every one unit increase in Night.Charge (charge for services during the night hours), the log odds of churn increase by 0.081465, having all other covariates constant.
- For every one unit increase in Intl.Calls (number of international calls), the log odds of churn decrease by 0.091430, having all other covariates constant.
- For every one unit increase in Intl.Charge (charge for international services), the log odds of churn increase by 0.32399, having all other covariates constant.

- The negative intercept indicates that when all covariates are zero and the categorical variables has the value of the refernce level, log odds is equal to 8.048873.

5. CONCLUSION

The aim of this project was to construct a logistic regression model, that describes well the factors that can lead a customer of an telecommunications company to break his contract with the company. Finally, we ended up to a logistic regression model, with 8 predictors and the constant. The predictors of our model are associated with the international activity of the customer, the charge of various services in day and night hours, the frequency of the calls to customer service.

We conclude saying that while a well-fitting logistic regression model has been constructed, there may be opportunities for improvement by increasing the size of the dataset or by exploring different transformations of the data.

Appendix : Tables and Figures

Table 1

```
> summary(logit1)

Call:
glm(formula = Churn ~ ., family = binomial(link = "logit"), data = df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9295  -0.4975  -0.3101  -0.1657   3.0786

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    -9.612e+00  9.810e-01  -9.798  < 2e-16 ***
Account.Length  9.959e-04  1.436e-03   0.693 0.488070
VMail.Message   3.883e-02  1.866e-02   2.081 0.037414 *
Day.Mins       -4.508e-01  3.383e+00  -0.133 0.893983
Eve.Mins        9.578e-01  1.701e+00   0.563 0.573281
Night.Mins     -2.130e-01  9.049e-01  -0.235 0.813924
Intl.Mins      -4.251e+00  5.495e+00  -0.774 0.439225
CustServ.Calls  5.378e-01  4.102e-02  13.111  < 2e-16 ***
Int.l.Plan1     2.197e+00  1.534e-01  14.324  < 2e-16 ***
VMail.Plan1    -2.147e+00  5.949e-01  -3.609 0.000307 ***
Day.Calls       4.115e-03  2.865e-03   1.436 0.150877
Day.Charge      2.729e+00  1.990e+01   0.137 0.890909
Eve.Calls       8.768e-04  2.894e-03   0.303 0.761935
Eve.Charge     -1.118e+01  2.001e+01  -0.559 0.576400
Night.Calls     1.613e-04  2.930e-03   0.055 0.956088
Night.Charge    4.820e+00  2.011e+01   0.240 0.810566
Intl.Calls     -9.075e-02  2.572e-02  -3.528 0.000419 ***
Intl.Charge     1.605e+01  2.035e+01   0.789 0.430249
StateAL         3.335e-01  7.637e-01   0.437 0.662289
StateAR         9.172e-01  7.527e-01   1.219 0.223007
StateAZ         1.169e-01  8.459e-01   0.138 0.890118
StateCA         1.807e+00  7.829e-01   2.308 0.020982 *
StateCO         6.477e-01  7.643e-01   0.847 0.396727
```

```
StateCO         0.66857    0.75718   0.883 0.377253
StateCT         1.07361    0.72122   1.489 0.136595
StateDC         0.74325    0.80347   0.925 0.354943
StateDE         0.76921    0.74641   1.031 0.302750
StateFL         0.66116    0.75732   0.873 0.382645
StateGA         0.71152    0.77214   0.921 0.356795
StateHI        -0.16695    0.89410  -0.187 0.851881
StateIA         0.23981    0.89851   0.267 0.789548
StateID         0.90342    0.74224   1.217 0.223546
StateIL        -0.16124    0.82736  -0.195 0.845485
StateIN         0.52919    0.74756   0.708 0.479009
StateKS         1.11000    0.72559   1.530 0.126068
StateKY         0.79615    0.76161   1.045 0.295860
StateLA         0.60388    0.83295   0.725 0.468459
StateMA         1.19252    0.73716   1.618 0.105724
StateMD         1.21233    0.71276   1.701 0.088962 .
StateME         1.36362    0.72402   1.883 0.059646 .
StateMI         1.43877    0.70858   2.031 0.042305 *
StateMN         1.18806    0.71191   1.669 0.095150 .
StateMO         0.62422    0.77101   0.810 0.418160
StateMS         1.37913    0.72499   1.902 0.057134 .
StateMT         1.86002    0.71197   2.612 0.008989 **
StateNC         0.69277    0.74876   0.925 0.354849
StateND         0.22298    0.79093   0.282 0.778000
StateNE         0.35396    0.80113   0.442 0.658612
StateNH         1.20372    0.76347   1.577 0.114878
StateNJ         1.63066    0.70549   2.311 0.020811 *
StateNM         0.48174    0.78518   0.614 0.539522
StateNV         1.30094    0.72110   1.804 0.071216 .
StateNY         1.18933    0.71599   1.661 0.096693 .
StateOH         0.73806    0.74097   0.996 0.319214
StateOK         0.92359    0.75191   1.228 0.219326
StateOR         0.76641    0.73267   1.046 0.295539
StatePA         1.17638    0.77426   1.519 0.128671
```

```
StatePA         1.17638    0.77426   1.519 0.128671
StateRI        -0.09400    0.81830  -0.115 0.908546
StateSC         1.82749    0.72960   2.505 0.012253 *
StateSD         0.87544    0.75603   1.158 0.246886
StateTN         0.28775    0.81509   0.353 0.724063
StateTX         1.68905    0.70343   2.401 0.016343 *
StateUT         1.08534    0.74058   1.466 0.142778
StateVA        -0.34787    0.81877  -0.425 0.670929
StateVT         0.16042    0.77267   0.208 0.835532
StateWA         1.47540    0.71921   2.051 0.040226 *
StateWI         0.32254    0.77564   0.416 0.677531
StateWV         0.62728    0.72950   0.860 0.389851
StateWY         0.34163    0.75051   0.455 0.648969
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2758.3  on 3332  degrees of freedom
Residual deviance: 2078.2  on 3270  degrees of freedom
AIC: 2204.2

Number of Fisher Scoring iterations: 6
```

Table 2

```
> rownames(coef(lasso.cv, s = 'lambda.1se'))[coef(lasso.cv, s = 'lambda.1se')[,1]!= 0]
 [1] "(Intercept)"    "Day.Mins"       "Eve.Mins"       "Night.Mins"     "Intl.Mins"
 [6] "CustServ.Calls" "Int.l.Plan1"    "VMail.Plan1"    "Day.Charge"     "Eve.Charge"
[11] "Night.Charge"   "Intl.Calls"     "Intl.Charge"    "StateNJ"        "StateSC"
[16] "StateTX"
```

Table 3

```
> summary(model2)

Call:
glm(formula = Churn ~ 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, family = binomial(link = "logit"),
    data = df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9252  -0.5044  -0.3133  -0.1708   3.1292

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    -9.13787    0.82417 -11.087  < 2e-16 ***
Day.Mins       -0.33733    3.37644  -0.100 0.920417
Eve.Mins        0.69847    1.69288   0.413 0.679905
Night.Mins     -0.21013    0.90226  -0.233 0.815843
Intl.Mins      -4.10771    5.48476  -0.749 0.453899
CustServ.Calls  0.53447    0.04078  13.106  < 2e-16 ***
Int.l.Plan1     2.19245    0.15276  14.352  < 2e-16 ***
VMail.Plan1    -0.97619    0.14935  -6.536  6.3e-11 ***
Day.Charge      2.06127   19.86147   0.104 0.917342
Eve.Charge     -8.12697   19.91606  -0.408 0.683229
Night.Charge    4.75601   20.04935   0.237 0.812490
Intl.Calls     -0.09061    0.02565  -3.532 0.000413 ***
Intl.Charge    15.52340   20.31277   0.764 0.444737
StateAL         0.37171    0.75757   0.491 0.623666
StateAR         0.95259    0.74616   1.277 0.201728
StateAZ         0.14758    0.84160   0.175 0.860802
StateCA         1.85426    0.78099   2.374 0.017585 *
StateCO         0.66857    0.75718   0.883 0.377253
StateCT         1.07361    0.72122   1.489 0.136595
StateDC         0.74325    0.80347   0.925 0.354943
StateDE         0.76921    0.74641   1.031 0.302750
StateFL         0.66116    0.75732   0.873 0.382645
StateGA         0.71152    0.77214   0.921 0.356795
StateHI        -0.16695    0.89410  -0.187 0.851881
StateIA         0.23981    0.89851   0.267 0.789548
StateID         0.90342    0.74224   1.217 0.223546
StateIL        -0.16124    0.82736  -0.195 0.845485
StateIN         0.52919    0.74756   0.708 0.479009
StateKS         1.11000    0.72559   1.530 0.126068
StateKY         0.79615    0.76161   1.045 0.295860
StateLA         0.60388    0.83295   0.725 0.468459
StateMA         1.19252    0.73716   1.618 0.105724
StateMD         1.21233    0.71276   1.701 0.088962 .
StateME         1.36362    0.72402   1.883 0.059646 .
StateMI         1.43877    0.70858   2.031 0.042305 *
StateMN         1.18806    0.71191   1.669 0.095150 .
StateMO         0.62422    0.77101   0.810 0.418160
StateMS         1.37913    0.72499   1.902 0.057134 .
StateMT         1.86002    0.71197   2.612 0.008989 **
StateNC         0.69277    0.74876   0.925 0.354849
StateND         0.22298    0.79093   0.282 0.778000
StateNE         0.35396    0.80113   0.442 0.658612
StateNH         1.20372    0.76347   1.577 0.114878
StateNJ         1.63066    0.70549   2.311 0.020811 *
StateNM         0.48174    0.78518   0.614 0.539522
StateNV         1.30094    0.72110   1.804 0.071216 .
StateNY         1.18933    0.71599   1.661 0.096693 .
StateOH         0.73806    0.74097   0.996 0.319214
StateOK         0.92359    0.75191   1.228 0.219326
StateOR         0.76641    0.73267   1.046 0.295539
StatePA         1.17638    0.77426   1.519 0.128671
StateRI        -0.09400    0.81830  -0.115 0.908546
StateSC         1.82749    0.72960   2.505 0.012253 *
StateSD         0.87544    0.75603   1.158 0.246886
StateSD         0.87544    0.75603   1.158 0.246886
StateTN         0.28775    0.81509   0.353 0.724063
StateTX         1.68905    0.70343   2.401 0.016343 *
StateUT         1.08534    0.74058   1.466 0.142778
StateVA        -0.34787    0.81877  -0.425 0.670929
StateVT         0.16042    0.77267   0.208 0.835532
StateWA         1.47540    0.71921   2.051 0.040226 *
StateWI         0.32254    0.77564   0.416 0.677531
StateWV         0.62728    0.72950   0.860 0.389851
StateWY         0.34163    0.75051   0.455 0.648969
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2758.3  on 3332  degrees of freedom
Residual deviance: 2078.2  on 3270  degrees of freedom
AIC: 2204.2

Number of Fisher Scoring iterations: 6

>
```

Table 4

```
> summary(BIC_model)

Call:
glm(formula = Churn ~ Eve.Mins + CustServ.Calls + Int.l.Plan +
    VMail.Plan + Day.Charge + Night.Charge + Intl.Calls + Intl.Charge,
    family = binomial(link = "logit"), data = df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1195  -0.5147  -0.3383  -0.2004   3.1063

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    -8.048873   0.514333 -15.649  < 2e-16 ***
Eve.Mins        0.007167   0.001141   6.284 3.31e-10 ***
CustServ.Calls  0.512584   0.039100  13.109  < 2e-16 ***
Int.l.Plan1     2.041929   0.145163  14.066  < 2e-16 ***
VMail.Plan1    -0.938145   0.144801  -6.479 9.24e-11 ***
Day.Charge      0.076500   0.006366  12.018  < 2e-16 ***
Night.Charge    0.081465   0.024621   3.309 0.000937 ***
Intl.Calls     -0.091430   0.024951  -3.664 0.000248 ***
Intl.Charge     0.323990   0.075363   4.299 1.72e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2758.3  on 3332  degrees of freedom
Residual deviance: 2165.5  on 3324  degrees of freedom
AIC: 2183.5

Number of Fisher Scoring iterations: 6
```

Table 5

```
> vif(BIC_model)
      Eve.Mins CustServ.Calls     Int.l.Plan     VMail.Plan      Day.Charge
      1.027812       1.084085       1.067565       1.018520       1.045532
  Night.Charge     Intl.Calls    Intl.Charge
      1.014429       1.008897       1.013228
```

Table 6

```
> PseudoR2(BIC_model, which = "McFadden")
 McFadden
0.2149065
> PseudoR2(model2, which = "McFadden")
 McFadden
0.2465641
> PseudoR2(logit1, which = "McFadden")
 McFadden
0.2499313
```