CHURN PREDICTION FOR TELECOM COMPANIES

Team Members:

NAME	CONTRIBUTION
Vijay Kumar Bollina vbollina@kent.edu	Model Building and Model Performance
Santhosh Reddy Mallikireddy smalliki@kent.edu	Predictions and Results
Mallikarjun Sasnur msasnur@kent.edu	Data Cleaning and Data Exploration
Vamsi Krishna Darapaneni vdarapan@kent.edu	Documentation and presentation

Index

Introduction

In modern Days most telecom companies suffer from Customer churn. It has become extremely important because of increase in competition among companies. It is very expensive for the companies to acquire new customers than to retain existing customers as the customers can easily choose an alternate service provider.

Telecom companies have used two approaches to address churn so far. They are

a) un-targeted approach (Without analysing the customer behaviour/profile)

b) targeted approach (Analysing customers behaviour and offering custom promotions)

Now, ABC wireless telecom company is suffering from churn rate. They have gathered information of their customers behaviour. By analysing the available data, the company wants to identify the customer churn and intervene them to retain the service.

Objective

The main objective of this project is to build a model to predict the customer churn based on the historical data provided from ABC wireless company. Based on the outcome of the model we do targeted marketing on the customers who are likely to churn.

Overview of data

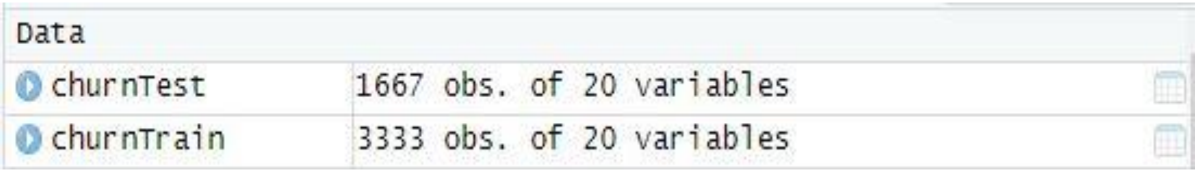
ABC wireless company has provided the following data from which we can infer:

- ➢ Demographics
  - State
  - Account length
  - Area code
  - International plan
  - Voice-mail plan

- ➢ Calling Behaviour
  - Number of messages
  - Total day minutes, Total day calls, Total day charge
  - Total evening minutes, Total evening calls and Total evening charges
  - Total night minutes, Total night calls and Total night charges
  - Total International minutes, Total International calls and Total International charges
  - Number of calls to customer service

Data Exploration:

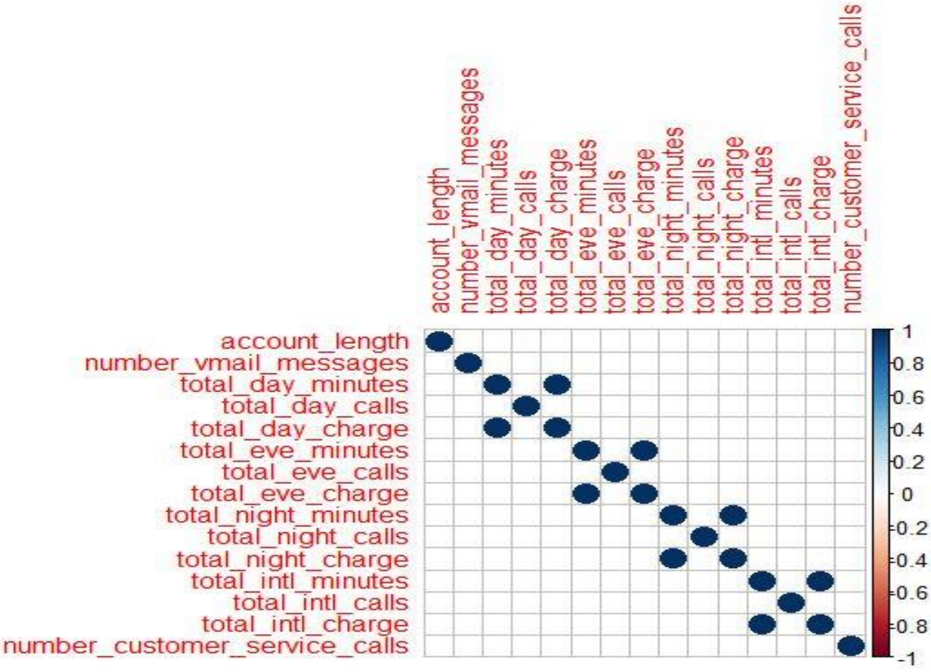
Initially we have 20 variables in our data.



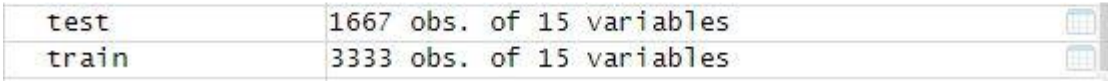
Data	
churnTest	1667 obs. of 20 variables
churnTrain	3333 obs. of 20 variables

We want to reduce the complexity of the model by reducing the number of variables without losing the accuracy. By observing the data, we excluded the "state" variable which is not significant to the model.

Similarly using "corrplot()" function we have found correlation between the continuous variables.



From the above plot we observe that 8 variables are highly correlated with each other. So, 4 variables can be omitted among them and now, we have 15 variables.



test	1667 obs. of 15 variables
train	3333 obs. of 15 variables

Model Construction

We have used Logistic regression technique because it is the appropriate regression analysis to conduct when the dependent variable is Multinomial(binary). Like all regression analysis, the logistic regression is a predictive analysis. It is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

In short, we use logistic regression when:

1.there is a binary variable.

2.we have explanatory variables that can be related to binary variable.

3.It is reasonable to say that the probability of getting a value of binary variable as yes/no depends on explanatory variable.

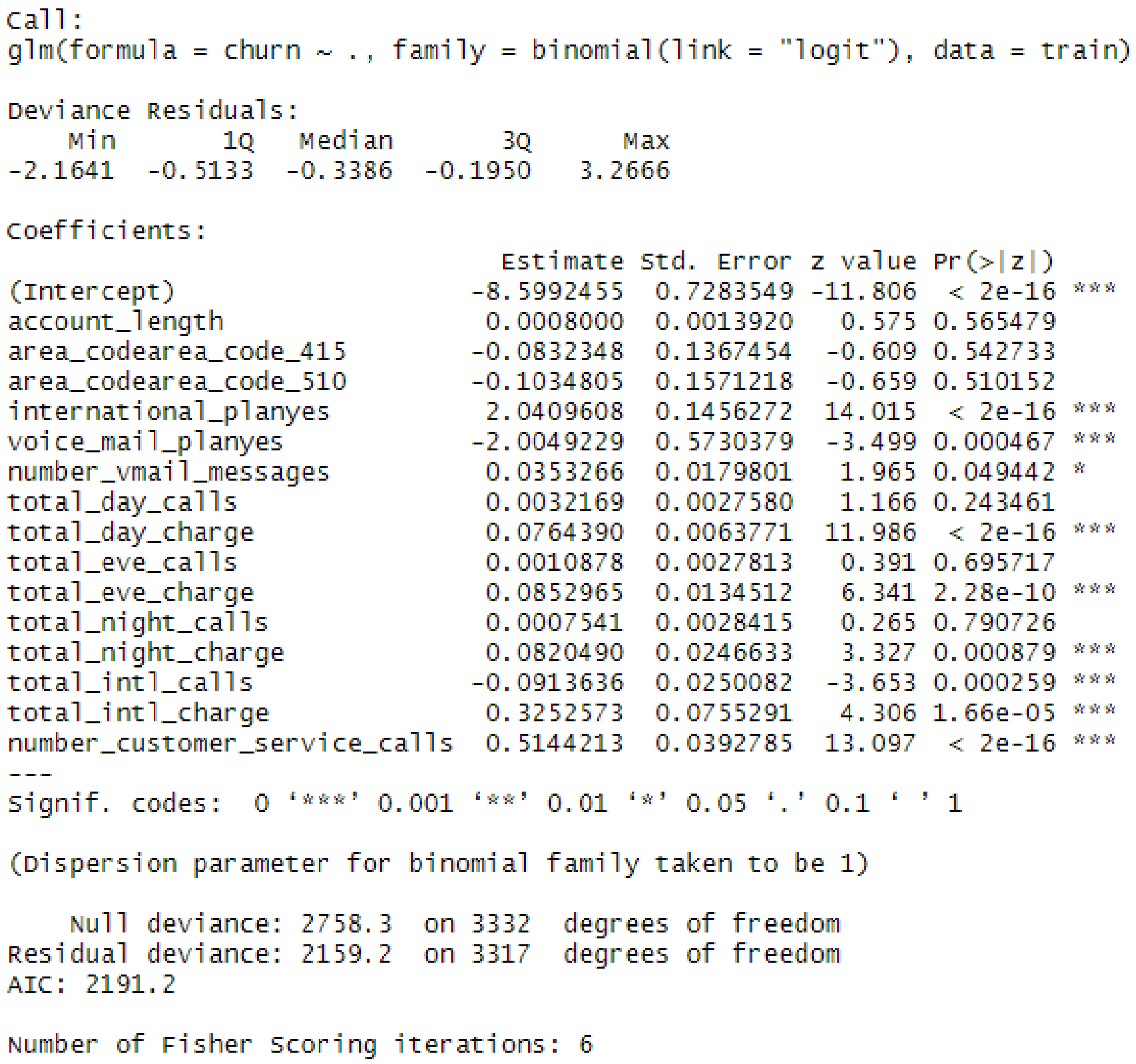
From the given data we infer that our dependent variable describes churn which is binary variable (yes or no) and the remaining variables are independent variables.

Constructing the model

- we have constructed the model using following libraries
  - **C50**: To import the churn dataset
  - **CARET** (Classification and Regression Training): To run most of the supervised machine learning algorithms.
  - **Gmodels**: To perform cross validation across the data.
  - **Ggplot2**: To visualize the data.
  - **Corrplot**: To visualize correlation plot
  - **pROC**: To measure the performance of the model using ROC plot.
- Imported training data from the library and selected variables to run the algorithm.
- We apply logistic regression algorithm on train data using “glm()” function.
- Using the above model, we predict the dependent variable of test data using “predict()” function.
- Now we set the cut off value to greater than 60% for probability of prediction.
- To check the accuracy of the prediction we use “CrossTable()” function.
- To measure the performance metric of the model we use “roc()” function.

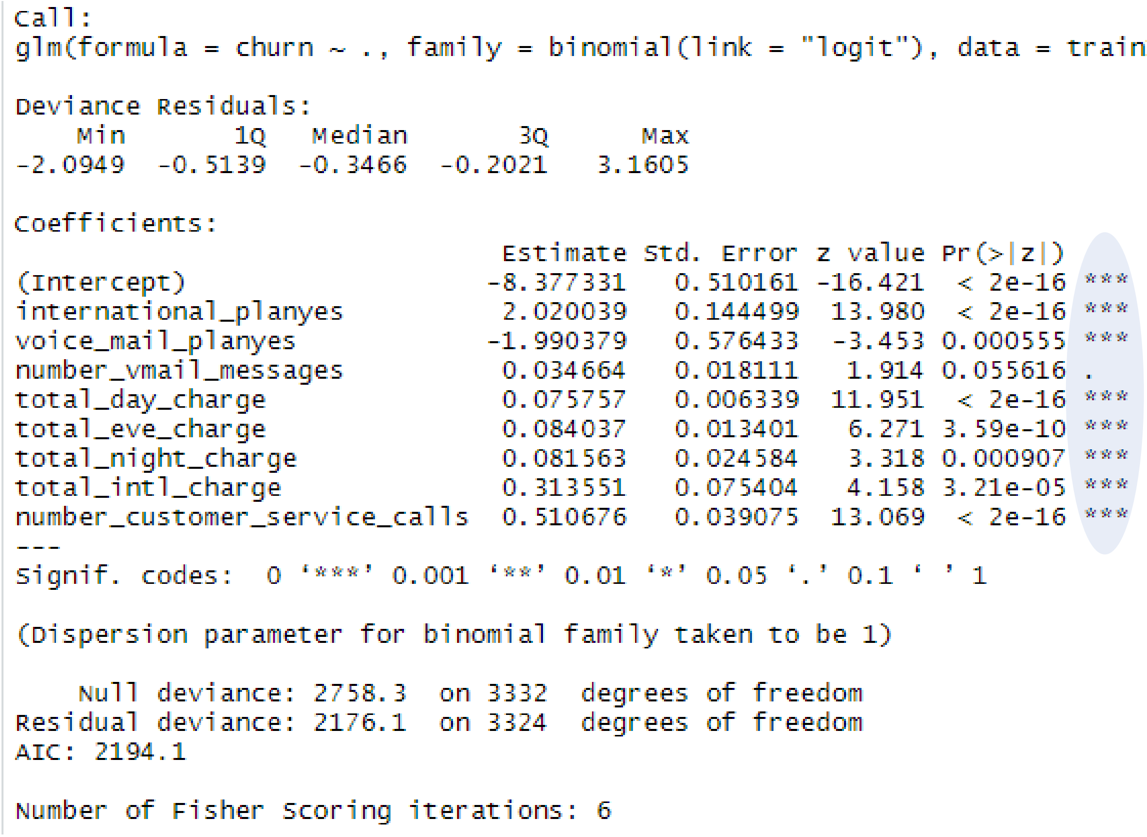
Model Performance

- we built the first model with 19 variables.
  - we got the prediction accuracy as 87.70%.
  - The performance metric of the model is Area Under Curve (AUC) as 0.8392.
- In the second model we used 15 variables.
  - Using corrplot we have omitted 4 variables from 19 variables since they are highly correlated.
  - Prediction accuracy of this model is 87.64%.
  - The performance metric of the model AUC is 0.84.
  - Summary of the model II.



```
Call:
glm(formula = churn ~ ., family = binomial(link = "logit"), data = train)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-2.1641  -0.5133  -0.3386  -0.1950   3.2666  

Coefficients:
                               Estimate Std. Error z value Pr(>|z|)    
(Intercept)                  -8.5992455  0.7283549 -11.806  < 2e-16 ***
account_length                0.0008000  0.0013920   0.575 0.565479    
area_codearea_code_415       -0.0832348  0.1367454  -0.609 0.542733    
area_codearea_code_510       -0.1034805  0.1571218  -0.659 0.510152    
international_planyes         2.0409608  0.1456272  14.015  < 2e-16 ***
voice_mail_planyes           -2.0049229  0.5730379  -3.499 0.000467 ***
number_vmail_messages         0.0353266  0.0179801   1.965 0.049442 *  
total_day_calls               0.0032169  0.0027580   1.166 0.243461    
total_day_charge              0.0764390  0.0063771  11.986  < 2e-16 ***
total_eve_calls               0.0010878  0.0027813   0.391 0.695717    
total_eve_charge              0.0852965  0.0134512   6.341 2.28e-10 ***
total_night_calls             0.0007541  0.0028415   0.265 0.790726    
total_night_charge            0.0820490  0.0246633   3.327 0.000879 ***
total_intl_calls             -0.0913636  0.0250082  -3.653 0.000259 ***
total_intl_charge             0.3252573  0.0755291   4.306 1.66e-05 ***
number_customer_service_calls 0.5144213  0.0392785  13.097  < 2e-16 ***
---
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: 2159.2  on 3317  degrees of freedom
AIC: 2191.2

Number of Fisher Scoring iterations: 6
```

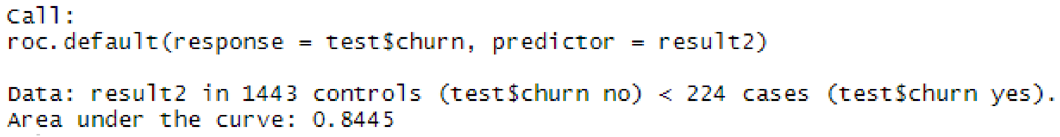
- In the third model
    - From the summary of model 2 we choose the variables whose probability of not rejecting null hypothesis is very low.
    - So, we are considering only 9 variables to construct the model which are statistically significant.
    - Prediction accuracy of this model is 87.34%.
    - The performance metric of the model AUC is 0.8445.
    - Summary of model 3



```
Call:
glm(formula = churn ~ ., family = binomial(link = "logit"), data = train1)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-2.0949  -0.5139  -0.3466  -0.2021   3.1605  

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)    
(Intercept)                  -8.377331   0.510161 -16.421  < 2e-16 ***
international_planyes         2.020039   0.144499  13.980  < 2e-16 ***
voice_mail_planyes           -1.990379   0.576433  -3.453 0.000555 ***
number_vmail_messages         0.034664   0.018111   1.914 0.055616 .  
total_day_charge              0.075757   0.006339  11.951  < 2e-16 ***
total_eve_charge              0.084037   0.013401   6.271 3.59e-10 ***
total_night_charge            0.081563   0.024584   3.318 0.000907 ***
total_intl_charge             0.313551   0.075404   4.158 3.21e-05 ***
number_customer_service_calls 0.510676   0.039075  13.069  < 2e-16 ***
---
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: 2176.1  on 3324  degrees of freedom
AIC: 2194.1

Number of Fisher Scoring iterations: 6
```

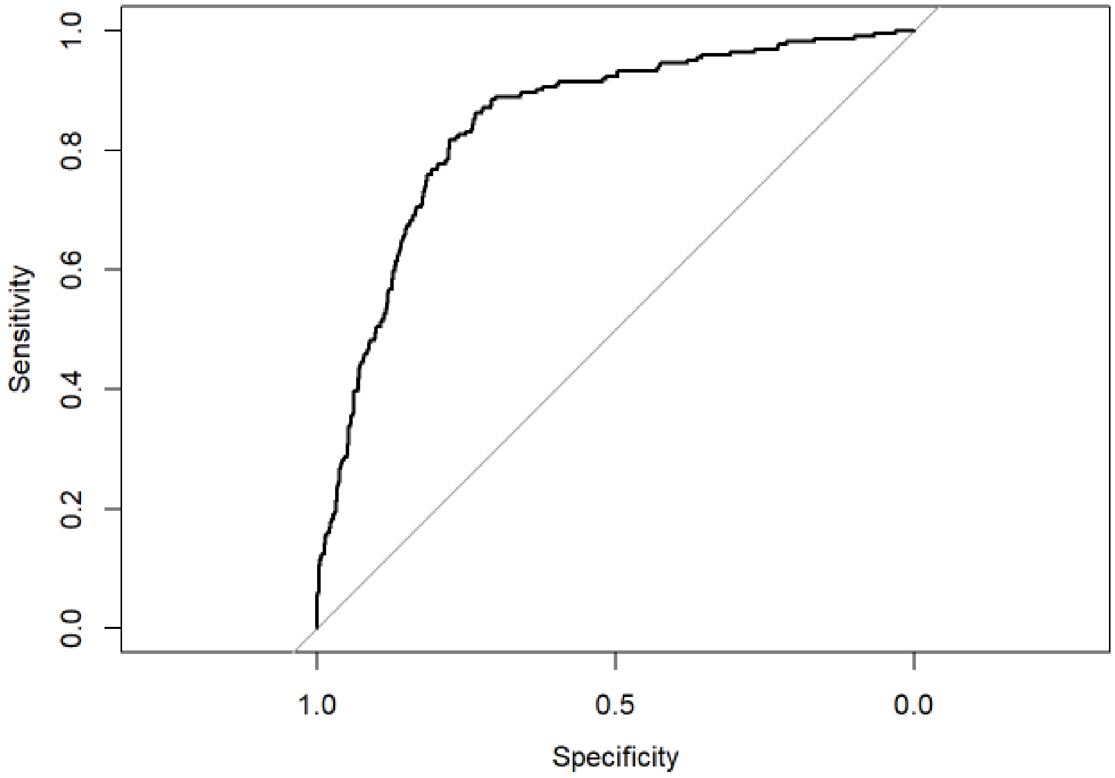
Choosing Best model

From the above observations we choose model 3 as the best one. Since it has only 9 variables which are significant. Compared with other two models, Model 3 has least number of variables and yields higher AUC value with similar prediction accuracy.



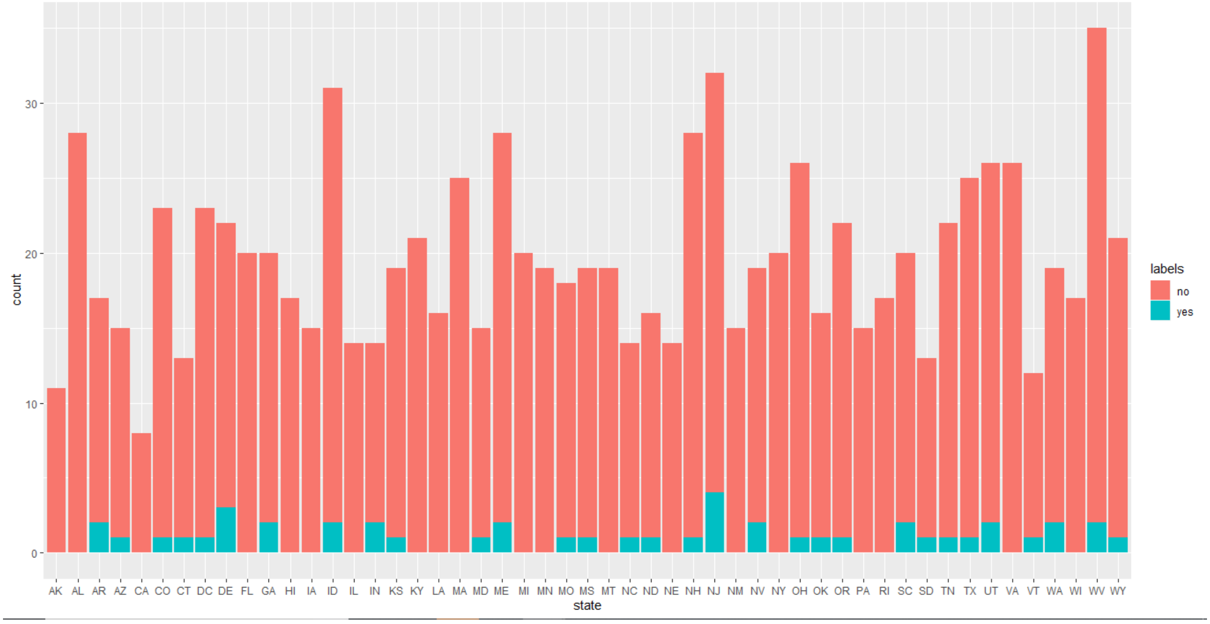
```
Call:
roc.default(response = test$churn, predictor = result2)

Data: result2 in 1443 controls (test$churn no) < 224 cases (test$churn yes).
Area under the curve: 0.8445
```

Plot of ROC Curve



Conclusion

We have a data of 1000 customers whose churn should be predicted. So, by using best model from the above we observed that 954 customers are not likely to churn, and 46 customers are likely to churn. Based on the predictive analysis we can target the customers to retain them who are likely to churn by doing the needful.



From the above plot we can visualize the number of customers state wise who are going to churn or not. Based on analysis we can do target marketing state wise.