Problem 1 [10 pts] – to be answered by everyone Given the large number of competitors, cell phone carriers are very interested in analyzing and predicting customer retention and churn. The primary goal of churn analysis is to identify those customers that are most likely to discontinue using your service or product. The dataset churn.csv contains information about a random sample of customers of a cell phone company. For each customer, the company recorded the following variables:

1. CHURN: 1 if customer switched provider, 0 if customer did not switch

2. GENDER: M, F

3. EDUCATION (categorical): code 1 to 6 depending on education levels

4. PRICE\_PLAN\_CHNG: 1 if price plan was changed, 0 otherwise

5. TOT\_ACTV\_SRV\_CNT: Total no. of active services

6. AGE: customer age

7. PCT\_CHNG\_IB\_SMS\_CNT: Percent change of latest 2 months incoming SMS wrt previous 4 months incoming SMS

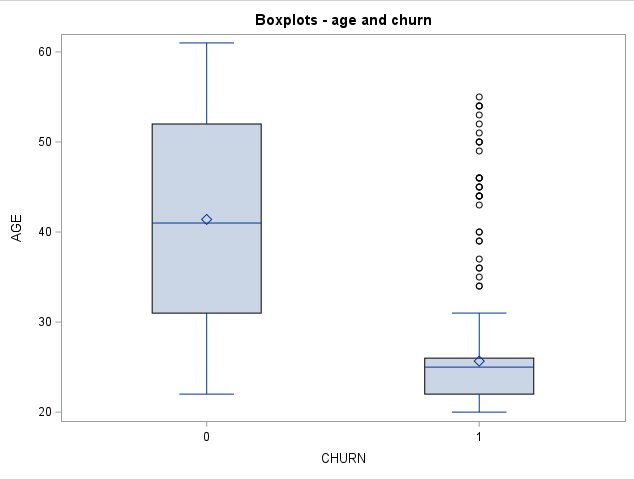
8. PCT\_CHNG\_BILL\_AMT: Percent change of latest 2 months bill amount wrt previous 4 months bill amount

9. COMPLAINT: 1 if there was at least a customer’s complaint in the two months, 0 no complaints

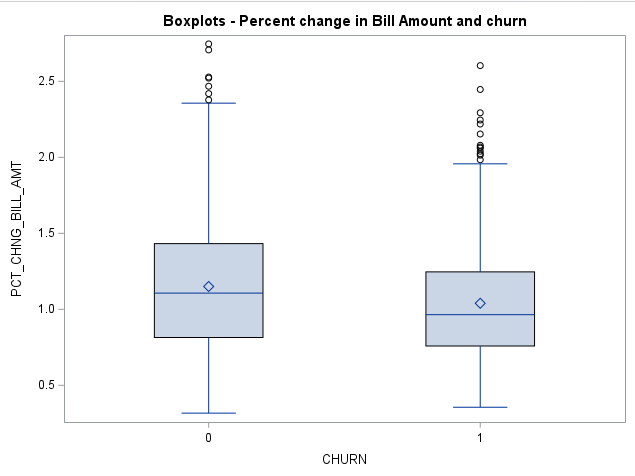
The company is interested in a churn predictive model that identifies the most important predictors affecting the probability of switching to a different mobile phone company (churn= 1). Answer the following questions:

a) Create two boxplots to analyze the observed values of age and PCT\_CHNG\_BILL\_AMT by churn value. Analyze the boxplots and discuss how customer age and changes in bill amount affect churn probabilities. Include the boxplots.

* **Boxplots:**
  + Age & Churn:



* + Percent Change in Bill Amount & Churn



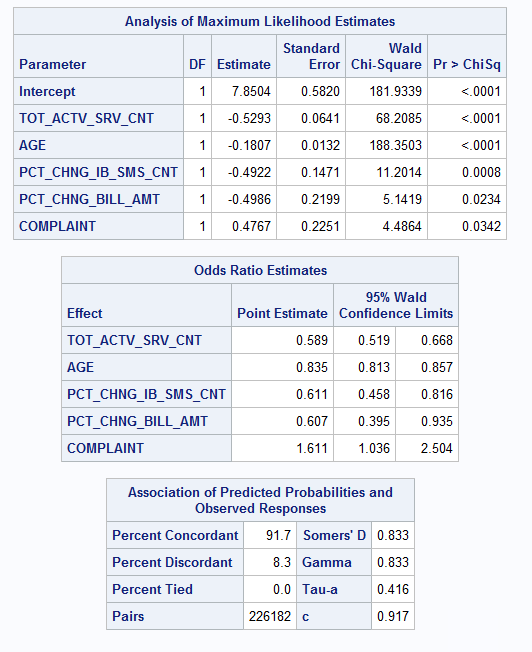
**Discussion: Note: Churn at 1 is if a customer switched providers, Churn at 0 is if customer did not switch providers.**

By regarding the boxplots starting at the relationship between AGE and CHURN, we can notice that customers that did not change providers were on average approximately 42 years old with an interquartile range of approximately 32-52 years old. By comparison, individuals who did change their providers were on average approximately 25 years old, with an interquartile range of approximately 22-26 years old. This may mean that there is a higher churn over rate when concerning younger individuals than older individuals.

In terms of the relationship between PCT\_CHNG\_BILL\_AMT and CHURN, we notice that the approximate average percent change in bill amount based on latest 2-month bill amount from 4-month bill amount for those who do not change providers is approx. 1.1%, with an interquartile range of 0.8%-1.45%. By comparison, the approximate average percent change in bill amount based on latest 2-month bill amount from 4-month bill amount for those who do change providers is approx. 1%, with an interquartile range 0.75%-1.2%. This may mean that for individuals who do not change providers, their bills may change more on average than individuals who do change providers.

b) Using a selection method, fit the final logistic regression model to predict the churn probability using the data in the dataset (Churn is the response variable and the remaining variables are the independent x-variables). Include the SAS output. Write down the expression of the fitted model.

* Stepwise:

Graphical user interface

Description automatically generated

Table

Description automatically generated

* Backward:

Table

Description automatically generatedTable

Description automatically generated

Table

Description automatically generated

* Forward:

Table

Description automatically generatedTable

Description automatically generatedTable

Description automatically generated

**Estimated Correlation Matrix:**

Table

Description automatically generated

**Outliers:**

Obs: 342, 390, 421, 457, 473, 526, 623, 680, 784, 801

**Stepwise with Removed Outliers:**

**Table

Description automatically generatedTable

Description automatically generated**

**Table

Description automatically generated with medium confidence**

**Fitted Model (after removing outliers):**

c) Analyze the final logistic regression model and discuss the effect of each variable on the churn probability.

**Note:** The slope parameter measures the degree of association between the probability p=pr(Y=1) and the value of X.

if :

**Discussion:** In consideration to the above formula, we can determine the effect each x-variable has on the probability of churn (success would mean the client churns over).

**TOT\_ACTV\_SRV\_CNT:** The odds of success decreases with an increase in X. If total number of active services increases (is large) the odds of the client churning over decreases.

**AGE:** The odds of success decreases with an increase in X. If the age of the client increases (is older/large) the odds of the client churning over decreases.

**PCT\_CHNG\_IB\_SMS\_CNT:** The odds of success decreases with an increase in X. If the percent change of the latest two months incoming SMS from previous four months incoming SMS increases the odds of the client churning over decreases.

**PCT\_CHNG\_BILL\_AMT:** The odds of success decreases with an increase in X. If the percent change of latest 2 months bill amount from previous 4 months bill amount increases the odds of the client churning over decreases.

**COMPLAINT:** The odds of success increases with an increase in X. If the client makes a complaint the odds of success increases.

d) Using SAS, compute the predicted churn probability and the confidence interval for a male customer who is 43 years old, and has the following information PRICE\_PLAN\_CHNG=0, TOT\_ACTV\_SRV\_CNT=4, PCT\_CHNG\_IB\_SMS\_CNT= 1.04, PCT\_CHNG\_BILL\_AMT= 1.19, and COMPLAINT =1. Include the output, interpret, and explain the 3 values you obtained.

**Computing CHURN:**

**= -3.314883**

**CI:**

****

**Predicted probability: 0.03509**

**95% confidence interval: (0.02019, 0.06031)**

**Discussion:**

In our confidence interval made of the above requisites, we looked for three values: phat, lcl and ucl. The phat value allows us to estimate the predicted probability in which churn is successful based on the qualites given to our x-variables (for example, Complaint = 1). In the case that we may predict the predicted probability (phat) of a customer who is of the above qualities, we receive a phat of 0.03509 such that the predicted probability of that customer churning over is 0.03509 (high chance of failure). The lcl value is the confidence interval’s lower value and ucl is the confidence interval’s upper value. In the case of the same costumer, 95% of the time, the predicted probability will fall within 0.02019 and 0.06031.

e) Copy and paste your FULL SAS code into the word document along with your answers.

**proc** **import** file = "C:\Users\DVALDE12\Desktop\A6\churn.csv"

out = churn

dbms = csv

replace;

**run**;

**proc** **print** data = churn;

**run**;

\*creating dummy variables;

**data** churn;

set churn;

d\_gender = (GENDER = 'M');

**run**;

\*boxplot churn/age & churn/Bill;

TITLE "Boxplots - age and churn";

**proc** **sgplot** data = churn;

vbox AGE / category = CHURN;

**run**;

TITLE "Boxplots - Percent change in Bill Amount and churn";

**proc** **sgplot** data = churn;

vbox PCT\_CHNG\_BILL\_AMT / category = CHURN;

**run**;

\*full model estimation procedure;

title 'full model';

**proc** **logistic** data = churn;

model CHURN(event ='1') = d\_gender EDUCATION PRICE\_PLAN\_CHNG TOT\_ACTV\_SRV\_CNT AGE PCT\_CHNG\_IB\_SMS\_CNT PCT\_CHNG\_BILL\_AMT COMPLAINT;

**run**;

\*creating fitted model based on stepwise procedure;

title 'fitted model: stepwise';

**proc** **logistic** data = churn;

model CHURN(event ='1') = d\_gender EDUCATION PRICE\_PLAN\_CHNG TOT\_ACTV\_SRV\_CNT AGE PCT\_CHNG\_IB\_SMS\_CNT PCT\_CHNG\_BILL\_AMT COMPLAINT

/ selection = stepwise rsquare;

**run**;

\*backward;

title 'fitted model: backward';

**proc** **logistic** data = churn;

model CHURN(event ='1') = d\_gender EDUCATION PRICE\_PLAN\_CHNG TOT\_ACTV\_SRV\_CNT AGE PCT\_CHNG\_IB\_SMS\_CNT PCT\_CHNG\_BILL\_AMT COMPLAINT

/ selection = backward rsquare;

**run**;

\*forward method;

title 'fitted model: forward';

**proc** **logistic** data = churn;

model CHURN(event ='1') = d\_gender EDUCATION PRICE\_PLAN\_CHNG TOT\_ACTV\_SRV\_CNT AGE PCT\_CHNG\_IB\_SMS\_CNT PCT\_CHNG\_BILL\_AMT COMPLAINT

/ selection = forward rsquare;

**run**;

\*checking correlation;

title 'checking correlatin';

**proc** **logistic** data = churn;

model CHURN(event = '1') = TOT\_ACTV\_SRV\_CNT AGE PCT\_CHNG\_IB\_SMS\_CNT PCT\_CHNG\_BILL\_AMT COMPLAINT/ corrb;

**run**;

\*finding outliers;

ods graphics on;

title 'outliers';

**proc** **logistic** data = churn;

model CHURN(event = '1') = TOT\_ACTV\_SRV\_CNT AGE PCT\_CHNG\_IB\_SMS\_CNT PCT\_CHNG\_BILL\_AMT COMPLAINT/ influence iplots corrb stb;

**run**;

ods graphics off;

\*remove outliers Obs: 342, 390, 421, 457, 473, 526, 623, 680, 784, 801;

**data** new\_churn;

set churn;

if \_n\_ in (**342**, **390**, **421**, **457**, **473**, **526**, **623**, **680**, **784**, **801**) then delete;

**run**;

\*fitted model with no outliers;

title 'Fitted Model without Outliers';

**proc** **logistic** data = new\_churn;

model CHURN(event ='1') = d\_gender EDUCATION PRICE\_PLAN\_CHNG TOT\_ACTV\_SRV\_CNT AGE PCT\_CHNG\_IB\_SMS\_CNT PCT\_CHNG\_BILL\_AMT COMPLAINT

/ selection = stepwise rsquare;

**run**;

\*CI;

**data** new;

input d\_gender PRICE\_PLAN\_CHNG TOT\_ACTV\_SRV\_CNT AGE PCT\_CHNG\_IB\_SMS\_CNT PCT\_CHNG\_BILL\_AMT COMPLAINT;

datalines;

1 0 4 43 1.04 1.19 1

;

**proc** **print**;

**run**;

**data** pred;

set new new\_churn;

**run**;

**proc** **print**;

**run**;

**proc** **logistic** data = pred;

title 'CI';

model CHURN(event ='1')

=TOT\_ACTV\_SRV\_CNT AGE PCT\_CHNG\_IB\_SMS\_CNT PCT\_CHNG\_BILL\_AMT COMPLAINT;

output out = prediction p=phat lower=lcl upper=ucl;

**run**;

**proc** **print** data = prediction;

**run**;