**Kubra iqbal**

**homework 6**

**CSC423: Data Analysis And Regression / CSC 324: Data Analysis & Statistical Software II**

**Assignment-6** | **Total Points: 10 pts for CSC 423/ CSC 324**

**Due Date: 05/15/2018 by 11:59 pm**

***Note: For all questions, immaterial if whether the relevant output is asked to be attached or not, make sure to include it. Also, it is important to include the sign (negative/positive or increase/decrease, and units of measurements e.g. $ or $ 99 million,%, etc.) otherwise points will be deducted.***

**Problem 1 [10 pts] Churn analysis – 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\_train.csv contains information about a random sample of customers of a cell phone company. For each customer, 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. LAST\_PRICE\_PLAN\_CHNG\_DAY\_CNT: No. of days since last price plan change
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 probability of switching to a different mobile phone company (churn = 1). Answer the following questions:

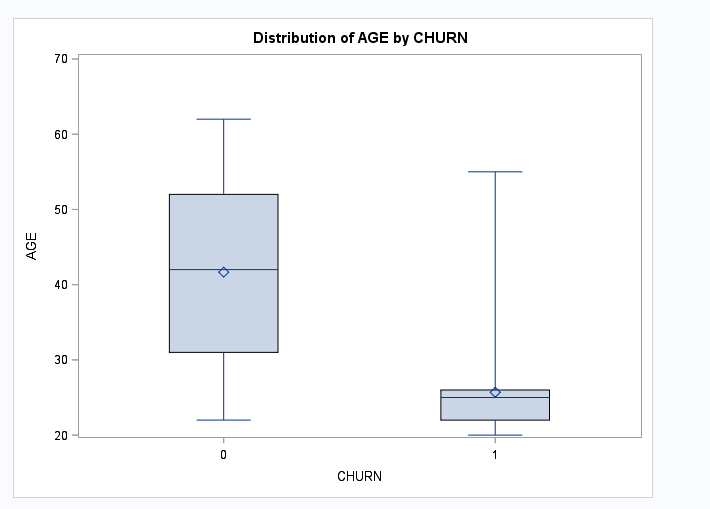
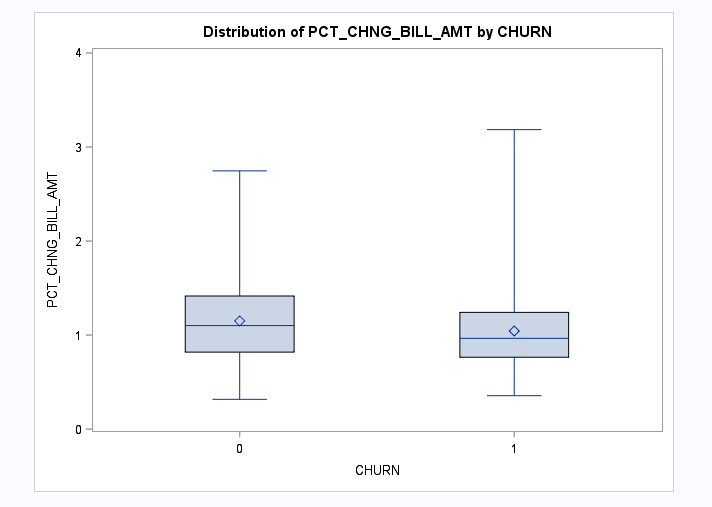
1. Create two boxplots to analyze the observed values of age and PCT\_CHNG\_BILL\_AMT by churn value. Analyze the boxplots and discuss how customer age and changes in bill amount affect churn probabilities. Include the boxplots.

Box plot 1 – The boxplot that shows the relation of Churn and PCT\_CHNG\_BILL\_AMT shows that the median is 1 and 0.75. The median of the first set of value is higher than the other one. The boxplots are comparatively short in the data below. The boxplots that shows the value at (o) is higher than the other one. The upper whiskers are taller and the lower whiskers are shorter if compared to each other. There are no outliers in this boxplot and both of them are almost symmetric.

Lower Quartile Value for both sets: 0.75 and 0.5

Middle quartile value: 1 and 0.75

Upper quartile value: 1.5 and 1.25

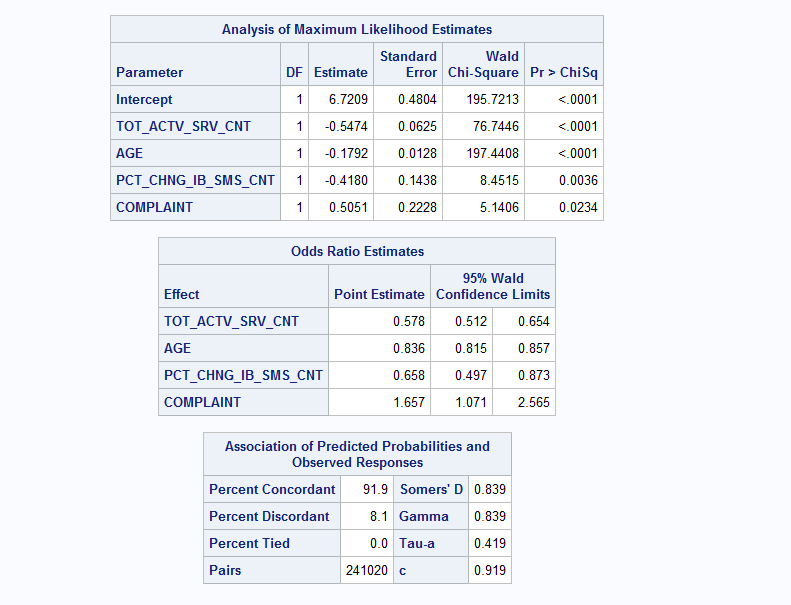


The above boxplot shows the relation between age and churn. The medians are as follows: 42 and 28. If compared the boxplot at (o) us comparatively taller if compared to the other one. For the first one, the whiskers are almost the same but whiskers for the second one have a huge difference. There is an obvious difference between the boxplots. The first one is symmetrically skewed but the left one is skewed right.   
Lower Quartile Value for both sets: 32 and 25

Middle quartile value: 42 and 28

Upper quartile value: 52 and 28

1. Fit a 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). Remove x-variables that are not significant using alpha=0.05. Include the SAS output. Write down the expression of the fitted model. (HINT: probability of interest is p = pr(churn = 1)



The full logistic regression model to predict probability :

Log(p(1-p)=6.7209-0.5474 tot\_actv\_srv\_cnt -0.1792 age -0.4180 pct\_chng\_ib\_sms\_cnt +0.5051 complaint

1. Analyze the final logistic regression model and discuss the effect of each variable on the churn probability. Discuss results in terms of odds ratios.

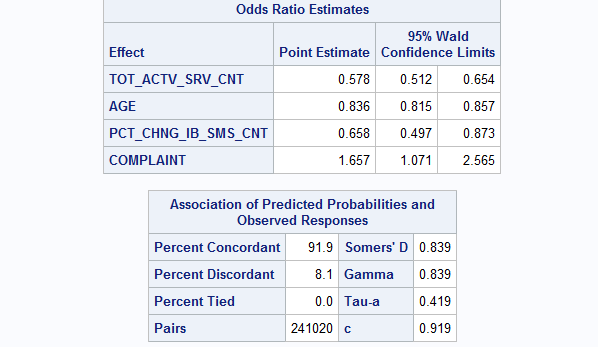
All the variables that include : tot\_actv\_srv\_cnt, age, pct\_chng\_ib\_sms\_cnt and complaint have a positive and negative parameters individually that shows that they both have a positive and negative associations with p.

tot\_actv\_srv\_cnt – ((0.57-1)\*100)) = -43%

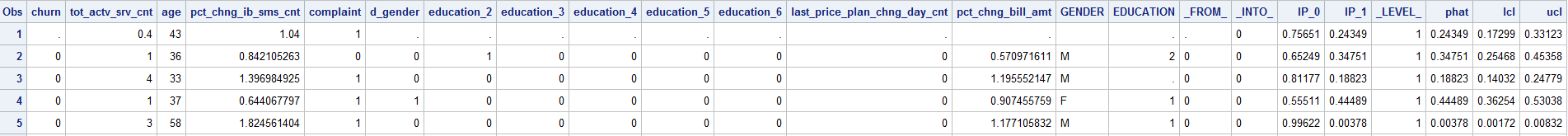
age - – ((0.836-1)\*100))= -16.4%

pct\_chng\_ib\_sms\_cnt = – ((0.658-1)\*100))= -34.2%

complaint = – ((1.657-1)\*100)) =67%



1. 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 LAST\_PRICE\_PLAN\_CHNG\_DAY\_CNT=0, TOT\_ACTV\_SRV\_CN=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.



The data above computes the predicted churn probability and the confidence interval for a male customer who is 43 years old and has the information listed below.

The pedicited value will be : p=0.24 with a 95% prediction interval of (0.173,0.331)

The odds will increase between 24% and 75%



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

/\*importing file\*/

**proc** **import** datafile ="churn\_train.csv" out=myd replace;

delimiter = ',';

getnames = yes;

**run**;

**proc** **print**;

**run**;

**data** churn;

set myd;

d\_gender=(gender="F");

education\_2=(education=**2**);

education\_3=(education=**3**);

education\_4=(education=**4**);

education\_5=(education=**5**);

education\_6=(education=**6**);

**run**;

**proc** **print**;

**run**;

**proc** **sort**;

by churn;

**run**;

**proc** **boxplot**;

plot PCT\_CHNG\_BILL\_AMT\*churn;

**run**;

**proc** **sort**;

by churn;

**run**;

**proc** **boxplot**;

plot age\*churn;

**run**;

**proc** **logistics** data=churn;

model churn (event='1')= d\_gender education\_2 education\_3 education\_4 education\_5 education\_6 last\_price\_plan\_chng\_day\_cnt tot\_actv\_srv\_cnt age pot\_chng\_ib\_sms\_cnt pct\_chng\_bill\_amt complaint/rsquare;

**run**;

**proc** **logisstics** data=churn;

title "final model";

model churn(event='1')= tot\_actv\_srv\_cnt age pct\_chng\_ib\_sms\_cnt complaint/rsquare;

**run**;

**proc** **logistics** data=churn;

title "fitted model";

model churn(event ='1') = tot\_actv\_srv\_cnt age pct\_chng\_ib\_sms\_cnt complaint/rsquare influecnce corrb stb;

**run**;

/\*churn probability and the confidence interval for a male customer\*/

**data** new3;

input churn

tot\_actv\_srv\_cnt

age

pct\_chng\_ib\_sms\_cnt

complaint

d\_gender

education\_2

education\_3

education\_4

education\_5

education\_6

last\_price\_plan\_chng\_day\_cnt

pct\_chng\_bill\_amt

;

datalines;

.

.4

43

1.04

1

.

.

.

.

.

.

.

.

**run**;

**data** predss;

set new3 churn;

**run**;

**proc** **print** data=predess;

**run**;

**proc** **logistic** data=predss;

model churn (event='1')= tot\_actv\_srv\_cnt age pct\_chng\_ib\_sms\_cnt complaint;

output out=press p=phat lower=lcl upper=ucl predprobs=(individual);

**run**;

**proc** **print** data=press;

**run**;