**DSC423: Data Analysis And Regression / DSC 324: Data Analysis & Statistical Software II**

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

**Due Date: 05/20/2019 by 11:59 pm**

Note:

* All assignments should be submitted in a **single MS WORD format**, no PDFs or any other file types will be accepted. If you submit any other file type, it will not be graded.
* No extensions will be given unless for a documented reason specified in the syllabus, no late assignments past the due date even a couple of minutes late will be accepted as you have an extra day (8-days) to submit your assignments.
* Submitting work that is not yours is grounds for an automatic ‘F’ for the entire course – this includes taking content and ideas from others or consulting others to complete your deliverables other than your instructor.
* SAS software and virtual server stalls, gets slow and crashes; so start early and keep multiple backups in multiple places/mediums. Late submission or inability to do the assignment due to server and/or software issues will not be accepted. Any issues relating with SAS, contact IS using the phone number provided in the syllabus, I won’t be able to help you with DePaul software related issues.

***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.***

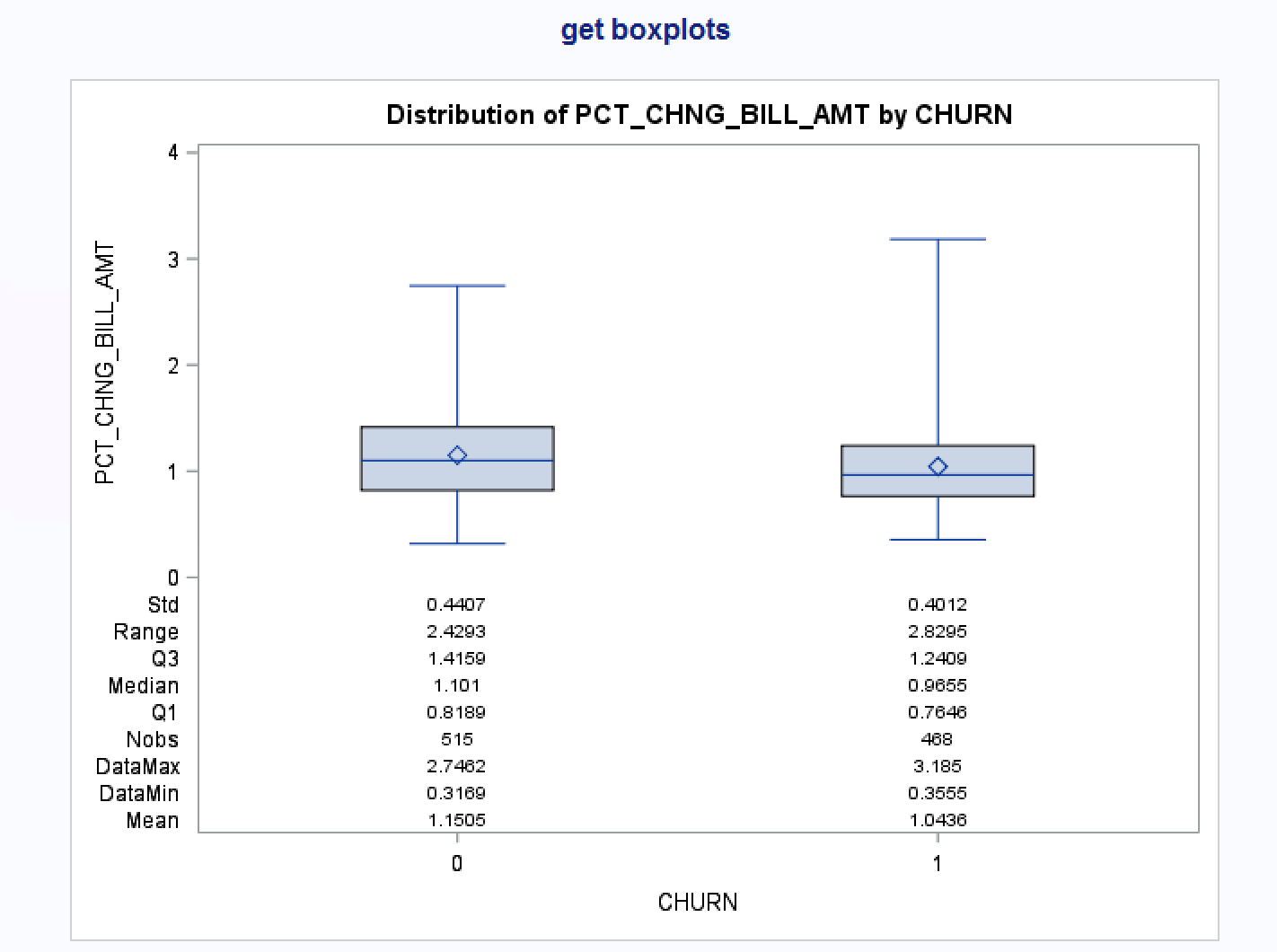
**Problem 1 [10 pts] Churn analysis – to be answered by everyone**

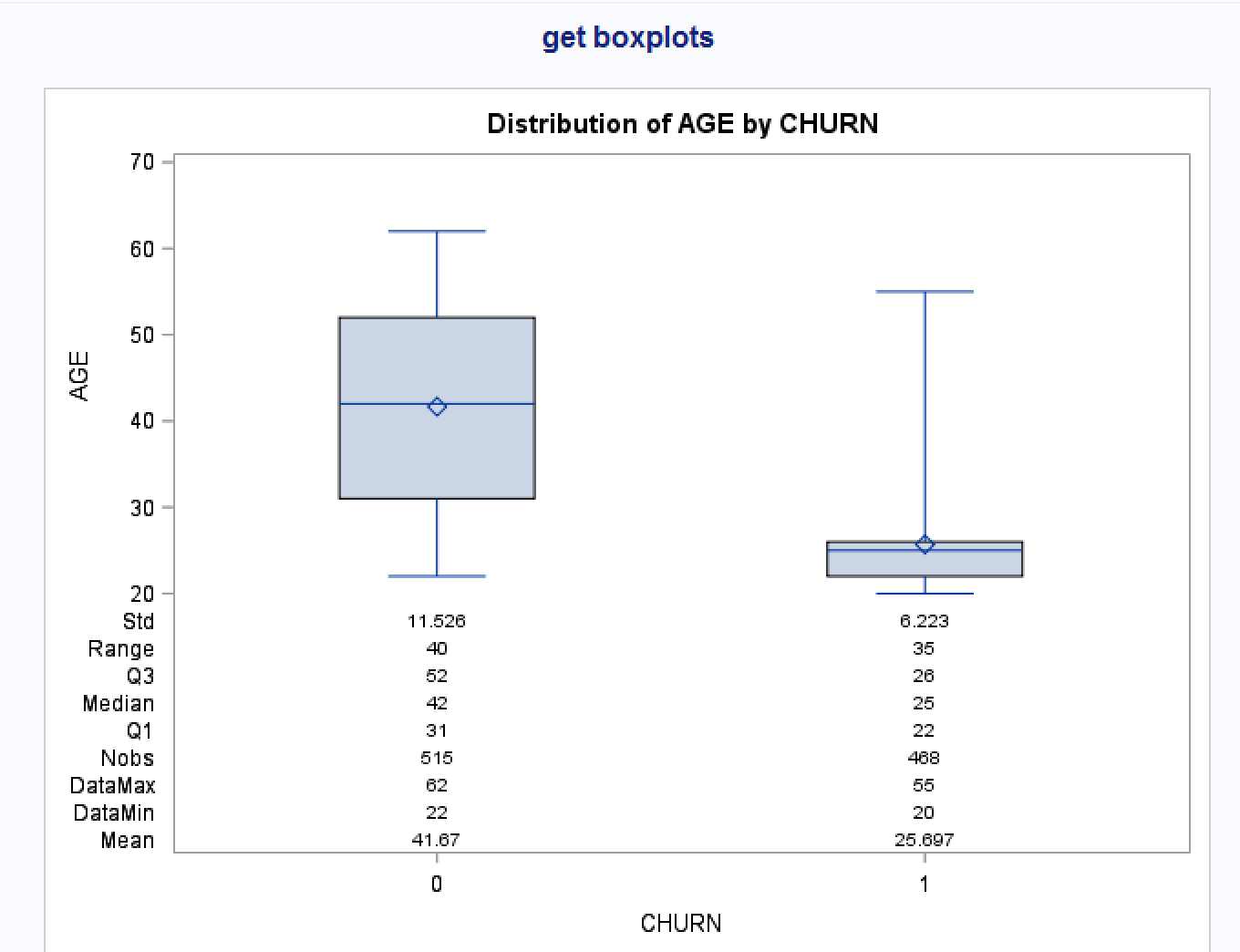
Given the large number of competitors, cell phone carriers are very interested in analyzing and predicting customer retention and churn. The primary goal of churn analysis is to identify those customers that are most likely to discontinue using your service or product. The dataset churn\_train.csv contains 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:

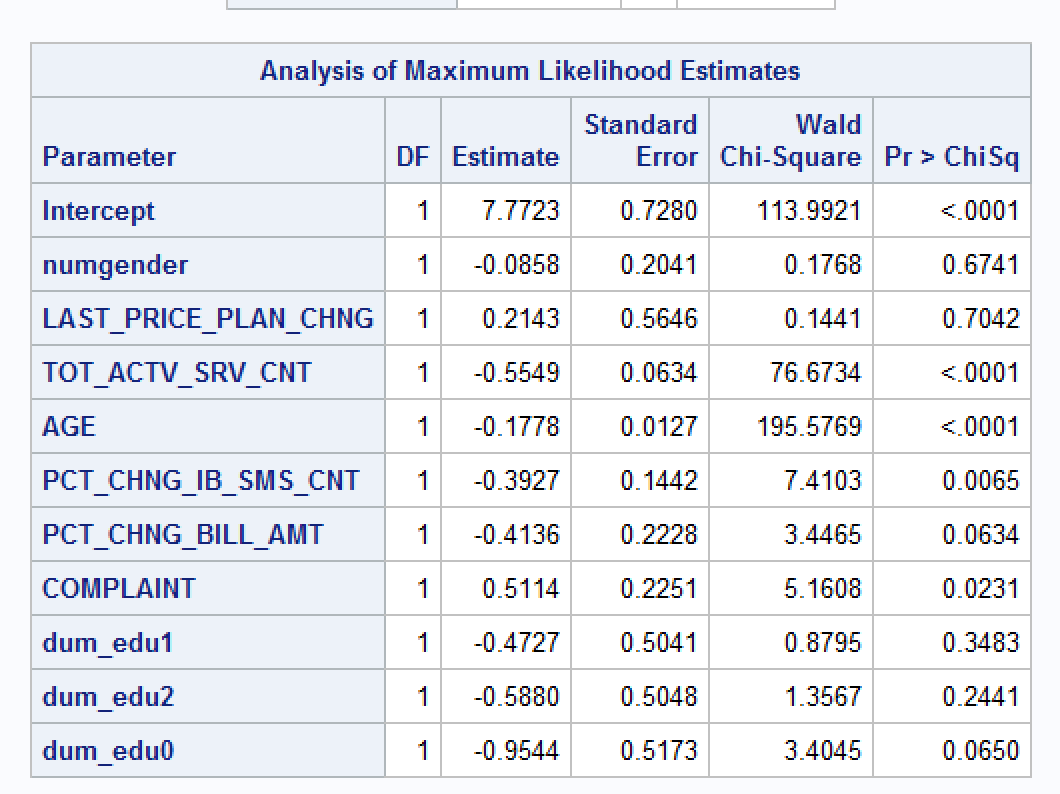
1. Create two boxplots to analyze the observed values of age and PCT\_CHNG\_BILL\_AMT by churn value. Analyze the boxplots and discuss how customer age and changes in bill amount affect churn probabilities. Include the boxplots.

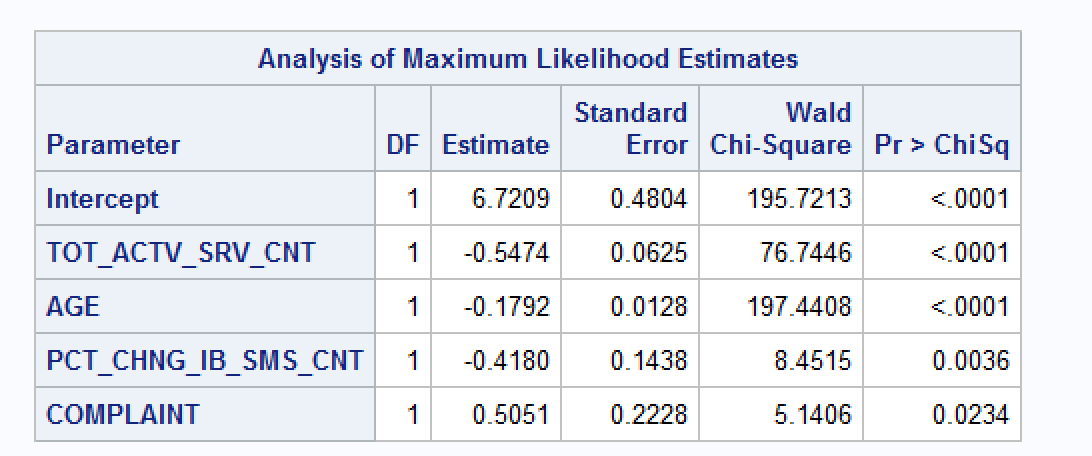




A 1 is given if the customer switched providers whereas a 0 is given if the customer did not switch.   
We can see that pct\_chng\_bill\_amt shows a little bit influence on churn decision. People who switched providers saw a 1% bill\_amt change whereas people who did not switch saw a 1.15% change. The only obvious difference is people who switched provider have a bigger pct\_chng\_bill\_amt. The range for this is 0.36-3.19 % compared to people who didn’t switch from 0.32% - 2.75%.

1. 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.





=

Complaint is going to be 1 when there was a customer complaint in the last two months, and 0 if there were no complaints. These are the only predictors with a value less than .05, all the others were removed form the model.

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

Without changing anything within the other predictors (no change), an additional count of active service decreases the churn odds by 42% and with a 95% CI, the decrease is within 48 and 34% for tot-actv-srv-cnt. Any additional year of age decreases the odds of churn by 16.4% and with a 95% CI that range is between 14-18%. Any additional count of Percent change of latest 2 months incoming SMS wrt previous 4 months incoming SMS decreases the odds of churn by 34.2% and with a 95% CI the decrease is between 12.7 – 50.3%. If there was at least one customer’s complaint in the two months, the odds of churn would increase by 67.1%, and with a 95% CI the increase is between 7.1% - 156.5%. If there was no customer’s complaint, the odds of churn would not change.

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.

Based on my model, for the specific input given in this question, the model gives the churn probability is 0.043 (4.3% probability to switch providers). This is a pretty low probability based on the other results we have seen thus far. With a 95% CI, 95% of the time, the predicted probability will fall within 0.027 and 0.068.

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

title 'import churn dataset';

proc import datafile="s:\assignment\_4\churn\_train.csv" out=churntrain replace;

delimiter=',';

getnames=yes;

run;

proc freq data = churntrain;

tables education;

run;

title 'get dummy variable';

data churntrain;

set churntrain;

numgender = (gender = 'M');

\*if education has missing value, I will make it to 0;

if education = .

then education = 0;

run;

data churntrain;

set churntrain;

dum\_edu1 = (education = 1);

dum\_edu2 = (education = 2);

dum\_edu0 = (education = 0);

run;

proc freq data = churntrain;

tables dum\_edu1 dum\_edu2 dum\_edu0;

run;

proc print data = churn\_train;

run;

proc sort data = churntrain;

by churn;

run;

proc boxplot data = churntrain;

plot (pct\_chng\_bill\_amt age) \* churn;

insetgroup mean min max n

q1 q2 q3 range stddev;

run;

proc logistic data = churntrain;

model churn(event = '1') = numgender LAST\_PRICE\_PLAN\_CHNG\_DAY\_CNT

TOT\_ACTV\_SRV\_CNT age PCT\_CHNG\_IB\_SMS\_CNT PCT\_CHNG\_BILL\_AMT COMPLAINT dum\_edu1

dum\_edu2 dum\_edu0;

run;

proc logistic data = churntrain;

model churn(event = '1') = TOT\_ACTV\_SRV\_CNT age PCT\_CHNG\_IB\_SMS\_CNT COMPLAINT;

run;

data new;

input GENDER $ AGE LAST\_PRICE\_PLAN\_CHNG\_DAY\_CNT TOT\_ACTV\_SRV\_CNT PCT\_CHNG\_IB\_SMS\_CNT PCT\_CHNG\_BILL\_AMT COMPLAINT;

datalines;

M 43 0 4 1.04 1.19 1

;

run;

proc print data = new;

run;

data pred;

set new churntrain;

numgender = (gender = 'M');

run;

proc print data = pred;

run;

proc logistic;

model churn(event = '1') = TOT\_ACTV\_SRV\_CNT age PCT\_CHNG\_IB\_SMS\_CNT COMPLAINT;

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

run;

proc print data = pred;

run;