**Jon Doretti**

**DATA ANALYSIS AND REGRESSION**

**Assignment-4** | **Total Points: 10 pts for DSC 323; 16 pts for DSC 423**

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 (7-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.

• **Make sure to double check your submissions. After you submit the assignment, log out of D2L, log back in, and click on your submission to see if you submitted the right file(s) and it is the correct version. Wrong submissions will not be graded.**

***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] – 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, 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

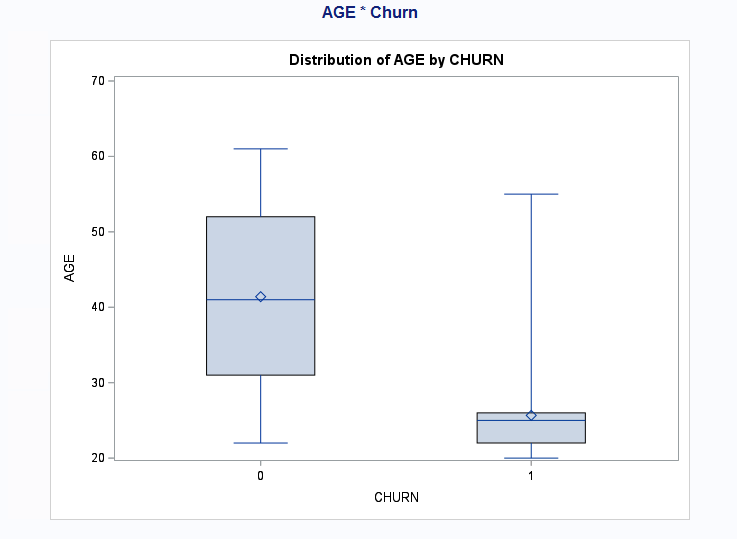
7. PCT\_CHNG\_IB\_SMS\_CNT: Percent change of latest 2 months incoming SMS wrt previous 4 months incoming SMS

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

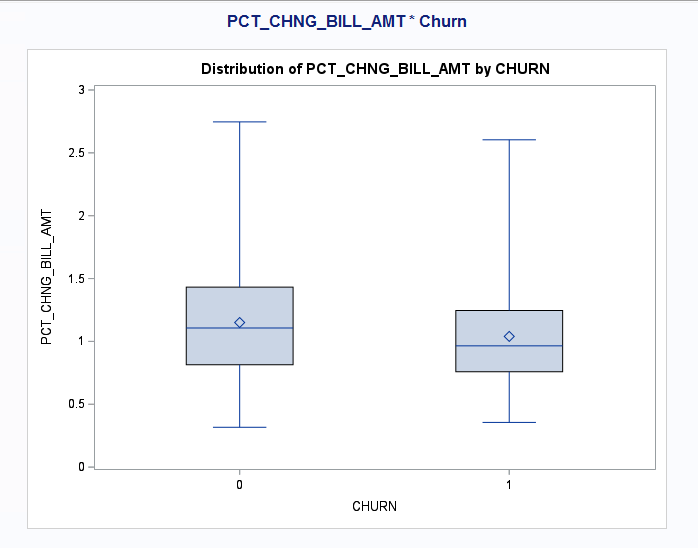
9. COMPLAINT: 1 if there was at least a customer’s complaint in the two months, 0 no complaints

The company is interested in a churn predictive model that identifies the most important predictors affecting 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.

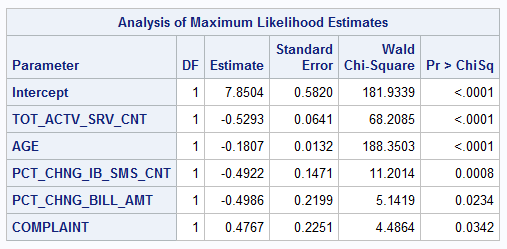


In the above box plot you can tell that younger people (20 through 27) are super likely to change their phone phone provider compared to people older than 30. Although both box plots are around the same min and max range. Overall, it seems more people are not willing to change their phone provider.



In the above graph, both box plots are about the same. This tells us that no matter if you do change your phone bill it will go up as all points whether they changed or not have increased. However those who did not change their phone bill have a wider range in the increase compared to those who changed their phone bill. This makes sense as those who change their phone bill - within the first few years - may receive a special deal from the company to entice them to stay with the company. This results in the tiger spread within those who changed their phone bill.

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.



Log(CHURN(1)/CHURN(0)) = 7.85 - .53(TOT\_ACTV\_SRV\_CNT) - .18(AGE) - .49(PCT\_CHNG\_IB\_SMS\_CNT) - .50 (PCT\_CHNG\_BILL\_AMT) + .48(COMPLAINT)

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

TOT\_ACTV\_SRV\_CNT - There is a negative 53% effect on someone changing their phone plan based on the total number of active services.

AGE - There is a negative 18% effect on someone changing their phone plan based on their age.

PCT\_CHNG\_IB\_SMS\_CNT - There is a negative 49% effect on someone changing their phone plan based on the percent change of the latest 2 months of incoming SMS wrt previous 4 months of incoming SMS.

PCT\_CHNG\_BILL\_AMT - There is a negative 50% effect on someone changing their phone plan based on the percent change of the latest 2 months bill amount wrt previous 4 months bill amount.

COMPLAINT - There is a positive 48% change on someone changing their phone plan based on if there was a customer complaint in the two months.

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.

Phat(0.06095): The chance they change their phone is .06095 or 6.095%.

Lcl(0.03863): The chance they change their phone is .03863 or 3.863%, with 95% certainty.

Ucl(0.09489):The chance they change their phone is .09489 or 9.489%, with 95% certainty.

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

proc import datafile="C:\Users\JDORETTI\Downloads\churn.csv" out=Churn replace;

delimiter=',';

getnames=yes;

datarow=2;

run;

PROC SORT;

BY CHURN;

RUN;

\*Boxplot;

TITLE"AGE \* Churn";

PROC BOXPLOT;

plot AGE \* CHURN;

RUN;

\*Boxplot;

TITLE"PCT\_CHNG\_BILL\_AMT \* Churn";

PROC BOXPLOT;

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

RUN;

\*Dummy variable;

data Churn1;

set Churn;

Male=(GENDER="M");RUN;

\*final log reg model;

PROC LOGISTIC data=Churn1;

model CHURN(event='1')=Male EDUCATION PRICE\_PLAN\_CHNG TOT\_ACTV\_SRV\_CNT AGE PCT\_CHNG\_IB\_SMS\_CNT PCT\_CHNG\_BILL\_AMT COMPLAINT/selection=forward rsquare;

RUN;

\*Predicted Values;

data prediction;

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

DATALINES;

0 4 43 1.04 1.19 1 1

;

RUN;

PROC PRINT;

RUN;

\*combine all data;

data prediction1;

set prediction Churn1;

RUN;

PROC PRINT;

RUN;

\*prediction;

PROC LOGISTIC;

TITLE"Predictions";

MODEL CHURN (event='1')=PRICE\_PLAN\_CHNG TOT\_ACTV\_SRV\_CNT AGE PCT\_CHNG\_IB\_SMS\_CNT PCT\_CHNG\_BILL\_AMT COMPLAINT Male;

OUTPUT out=prediction p=phat lower=lcl upper=ucl predprob=(individual);

RUN;

\*Print;

PROC PRINT data=prediction;

RUN;