**Aamir Patel**

**Data Analysis And Regression**

**Assignment-6** | **Total Points: 10**

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.***

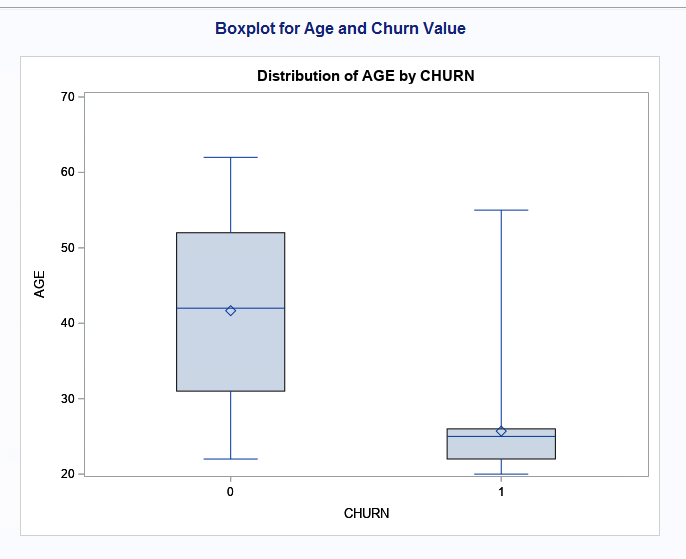
**Problem 1 [10 pts] Churn analysis**

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:

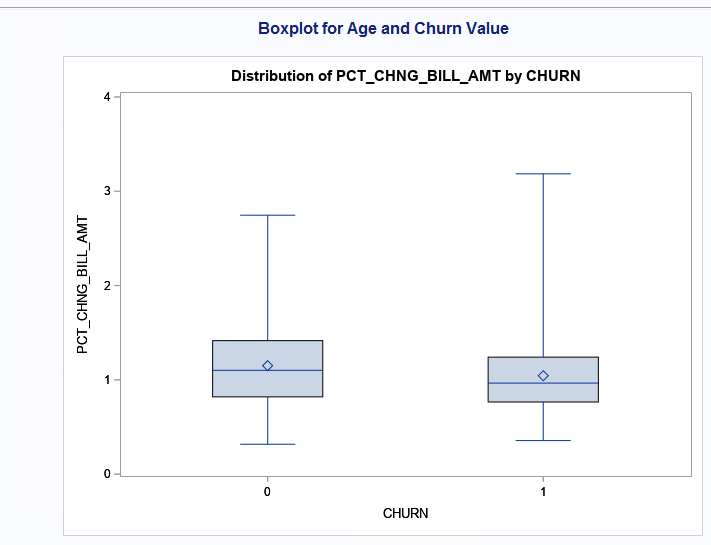
1. CHURN: 1 if customer switched provider, 0 if customer did not switch
2. GENDER: M, F
3. EDUCATION (categorical): code 1 to 6 depending on education levels
4. LAST\_PRICE\_PLAN\_CHNG\_DAY\_CNT: No. of days since last price plan change
5. TOT\_ACTV\_SRV\_CNT: Total no. of active services
6. AGE: customer 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.

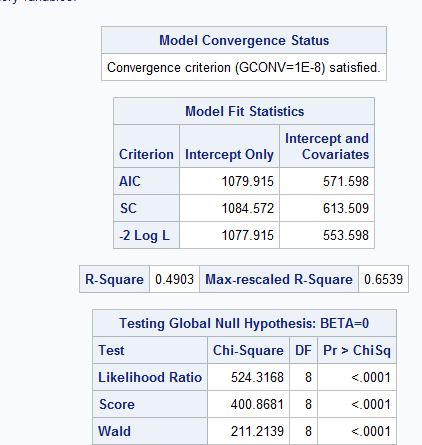
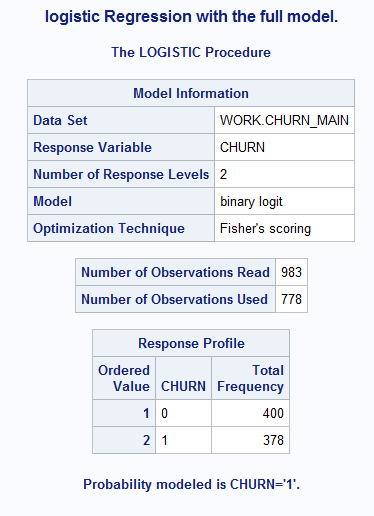


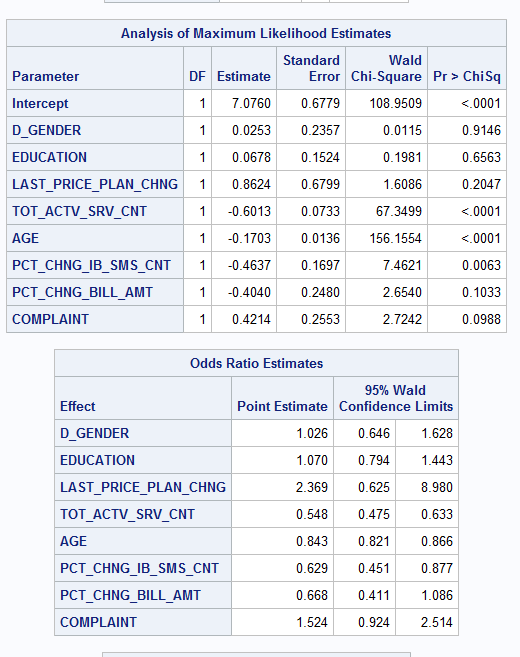
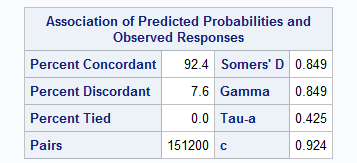
* Age and Churn: - Upon looking at the box plot for Age and Churn we can see that ages between 22 to 25 tend to switch providers often, and people between the age of 30 to 52 tend to not change their providers as much. In other words, we can say that younger people tend to not be more open to different providers than older people.



* CHURN and PCT\_CHNG\_BIL\_AMT: - Upon looking at the boxplot, we can see that the difference between changing providers and percent of bill change is very minimal. We can also see that people who did not churn have a minor higher percentage difference compared to people who did churn. In other words, changing providers does not affect the change in bill amount as much.

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.

* **Log(churn = 1 / churn = 0 ) = 7.0760 – 0.6013 (TOT\_ACTV\_SRV\_CNT) – 0.1703(AGE) – 0.4637 (PCT\_CHNG\_IB\_SMS\_CNT)**

**Log(churn = 1 / churn = 0 ) = 7.0760 – 0.6013 (TOT\_ACTV\_SRV\_CNT) – 0.1703(AGE) – 0.4637 (PCT\_CHNG\_IB\_SMS\_CNT)**

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

* **TOT\_ACTV\_SRV\_CNT: -** Looking at the odds ratio, we can see that everytime the TOT\_ACTV\_SRV\_CNT increases the chances of the customer churning decreases by 45.2%. and with 95% confidence interval we can say that the change will be between -52.5% to -36.7%.
* Calculation: - (0.548 – 1) \* 100 = -45.2%

(0.475 - 1) \* 100 = -52.5%

(0.633 - 1) \* 100 = -36.7%

* **AGE: -** Looking at the odds ratio for Age,we can see that everytime the AGE increases the chances of the customer churning decreases by 15.3%, and with the confidence interval of 95%, we can say that the change will be between -17.9% and -13.4%.
* Calculation: - (0.843 – 1) \* 100 = -15.7%

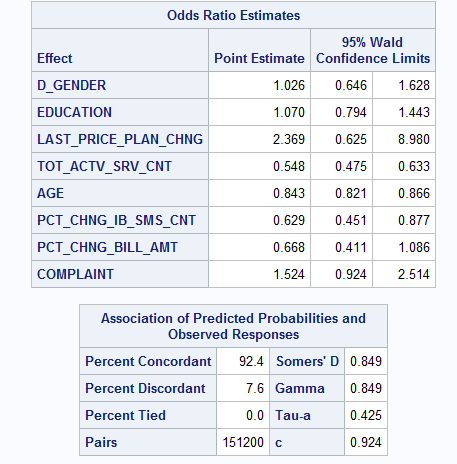
(0.821 - 1) \* 100 = -17.9%

* 1. - 1) \* 100 = -13.4%
* **PCT\_CHNG\_IB\_SMS\_CNT: -** Looking at the odds ratio for PCT\_CHNG\_IB\_SMS\_CNT, we can see that everytime the PCT\_CHNG\_IB\_SMS\_CNT increases the chances of the customer churning decreases by -37.1%. We can also say with 95% confidence interval that the change will be between -54.9% to -12.3%.

**-** Calculation: - (0.629 – 1) \* 100 = -37.1%

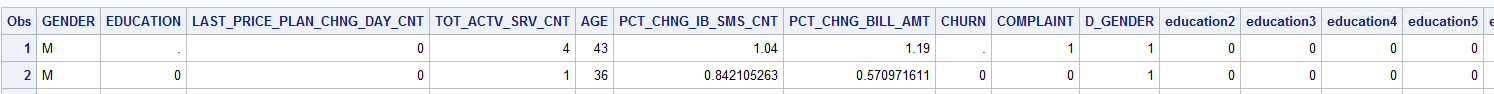
(0.451 - 1) \* 100 = -54.9%

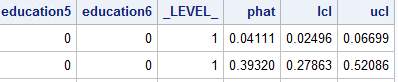
(0.877 - 1) \* 100 = -12.3%



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.

* Looking at the information for 43 year old customer, we can say with 95% confidence that the churn probability of our customer is 0.04%, with the confidence range from 0.02496 to 0.06699.





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

\*import data from file;

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

delimiter=',';

getnames=yes;

**run**;

\*Printing entire Dataset;

title "Printing main Dataset";

**proc** **print** data=churn\_main;

**run**;

\*For Dummy Variables Gender;

title "Dummy Variables for GENDER";

**DATA** churn\_main;

set churn\_main;

if (GENDER = 'M') then D\_GENDER = **1**;

else D\_GENDER = **0**;

education = (education = **2**);

education = (education = **3**);

education = (education = **4**);

education = (education = **5**);

education = (education = **6**);

**run**;

\*Printing entire Dataset;

title "Printing Dummy Dataset";

**proc** **print** data=churn\_main;

**run**;

\*Printing Boxplot for AGE AND CHURN;

title "Boxplot for Age and Churn Value";

**proc** **boxplot**;

plot AGE\*CHURN;

**run**;

\*Printing Boxplot for PCT\_CHNG\_BILL\_AMT and CHURN ;

title "Boxplot for PCT\_CHNG\_BILL\_AMT and Churn Value";

**proc** **boxplot**;

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

**run**;

\*Full Logistic Model;

title "logistic Regression with the full model.";

**proc** **logistic**;

model CHURN (event='1') = D\_GENDER EDUCATION LAST\_PRICE\_PLAN\_CHNG\_DAY\_CNT TOT\_ACTV\_SRV\_CNT AGE PCT\_CHNG\_IB\_SMS\_CNT PCT\_CHNG\_BILL\_AMT COMPLAINT / rsquare;

**run**;

\*Reduced Model;

title "Logistic Regression with Reduced Model.";

**proc** **logistic**;

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

**run**;

\*Inserting data for Question D ;

title "Inserting data for Question D";

**data** new;

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

datalines;

M . 0 4 43 1.04 1.19 . 1

;

**run**;

**PROC** **PRINT**;

**RUN**;

\*Setting up dummy variables;

**data** predict;

set new churn\_main;

if (GENDER='M') then d\_gender=**1**;

else d\_gender=**0**;

education2=(education=**2**);

education3=(education=**3**);

education4=(education=**4**);

education5=(education=**5**);

education6=(education=**6**);

**RUN**;

title "Printing after doing dummy variables";

**proc** **print**;

**run**;

title "Proc Logistic for new model";

**PROC** **LOGISTIC** data=predict;

MODEL CHURN (event='1') = d\_gender LAST\_PRICE\_PLAN\_CHNG\_DAY\_CNT TOT\_ACTV\_SRV\_CNT AGE PCT\_CHNG\_IB\_SMS\_CNT PCT\_CHNG\_BILL\_AMT COMPLAINT ;

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

**RUN**;

title "Printed Data after predictions";

**proc** **print** data=predict;

**run**;