**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**

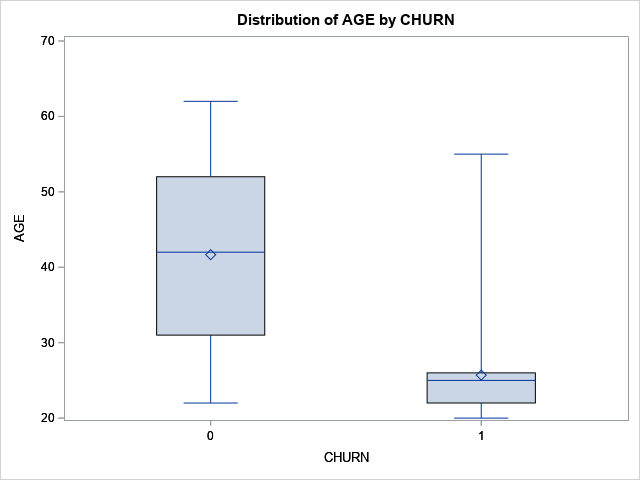
Given the large number of competitors, cell phone carriers are very interested in analyzing and predicting customer retention and churn. The primary goal of churn analysis is to identify those customers that are most likely to discontinue using your service or product. The dataset churn\_train.csv contains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

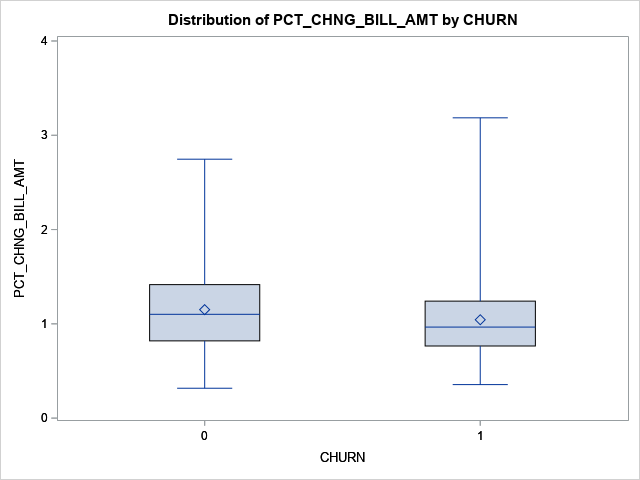
The company is interested in a churn predictive model that identifies the most important predictors affecting probability of switching to a different mobile phone company (churn = 1). Answer the following questions:

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 boxplot**:



**PCT\_CHNG\_BILL\_AMT and Churn boxplot:**



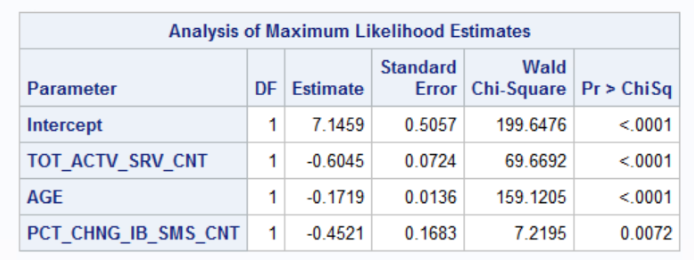
**My findings on the boxplots:**

**Age and Churn boxplot**: The age boxplot shows that those people whose CHURN was 0 meaning they did not switch provider that boxplot’s maximum was higher than someone CHURN is 1 who did switch. The maximum or oldest age that someone did not change was about ~63 compared to someone who did who was ~57. The minimum for someone who did not switch their service was ~23 but someone who did switch was younger than someone who did at ~20. The Q3 and Q1 for someone who did not switch was much higher in both compared to someone who did switch their service Q3 was ~54 and Q1 was ~32 compared to someone who did switch which was much younger Q3 was ~25 and Q1 was ~22. Why I believe the Q1 and Q3 both are older meaning a 0 compared to someone who switch which is 1 because those who are in their 30,40,50’s did not grow up in an era where phones were meant to be this daily used software so when it comes to the CHURN 1 the 20’s group who grew up on this technology will want to switch more. They get better deal for the new phones that come out every year. Advertisements will target younger generations to switch phone plans because they have unlimited data, phone call numbers, and texting enticing the younger generations to make the switch for cheaper prices while trading in their current phone. The median and mean of CHURN 0 is median ~43 and the mean is ~42. The CHURN 1 boxplot is much lower like everything else in comparison the median is ~24 but the mean is higher at ~26.

**PCT\_CHNG\_BILL\_AMT and Churn boxplot:** When looking at the boxplots it does not surprise me when people change service their bill went up. The maximum for CHURN 0 not changing their phone service the maximum the bill went up was ~2.8% compared to when CHURN 1 when someone changes their service was higher and was ~3.3% a higher ½ a percent higher.The minimum for both boxplots was about the same CHRUN 0 boxplot was ~.3% and CHURN 1 was slightly higher at ~.4% but not much of a difference that I see when looking at price for phone bill change before and after changing or staying with a service. The Q3 for CHURN 0 was higher than CHURN 1 the Q3 for CHURN 0 was ~1.4% while CHURN 1 was ~1.2%. Q1 for CHURN 0 was ~.8% while CHURN 1 was ~.7%. My thinking for why CHURN 1 Q3 and Q1 is lower than CHURN 0 for Percent Bill Change because when signing up for a new service they tend to offer cheaper prices than the person’s original phoner service and when they switch will keep it relative to the price they advertised. Contracts with phone services tend to be a couple years and that is when over time the price for the phone bill will go up when they try to hit people with more fees. So when they have the new phone service it will be lower but as time goes on them being in CHURN 0 will turn into CHURN 1 and their percent their phone bill could go up is those rates CHURN 0 has.

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.

**Final Model Forward Selection**:



**The final expression of the model is**:

Log(CHURN=1/CHURN=0)=7.1459 – 0.6045 TOT\_ACTV\_SRV\_CNT – 0.1719 AGE – 0.4521 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**: The effect Total number of active services has a negative effect if someone will not change their phone service being a -60.45% the less likely someone be willing to switch their phone service.

**AGE**: The effect of someone’s Age has a -17.19% for someone to not switch their phone service.

**PCT\_CHNG\_IB\_SMS\_CNT**: The effect Percent change of latest 2 months incoming SMS wrt previous 4 months incoming SMS has a -45.21% chance of making a person not switch their phone service.

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.

**Values obtained:**

Phat-0.04184. It is saying that the predicted probability is 0.04184.

Lcl-0.02591 95% of the time, the lower bound of the predicted probability is 0.025

Ucl-0.06688 95% of the time, the higher bound of the predicted probability is 0.066

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 replace;

delimiter=',';

getnames=yes;

datarow=**2**;

**run**;

TITLE"Churn";

**PROC** **PRINT**;

**RUN**;

\*Boxplot of age by Churn;

TITLE"AGE By Churn";

**PROC** **BOXPLOT**;

plot AGE \* CHURN;

**RUN**;

\*Boxplot of PCT\_CHNG\_BILL\_AMT by Churn;

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

**PROC** **BOXPLOT**;

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

**RUN**;

\*Dummy variable creation for Gender;

**data** Churn;

set Churn;

Male=(GENDER="M");\*Dummy variable for Male;

**RUN**;

\*Fit logistic model, and run forward selection procedure (USING THIS TEST);

**PROC** **LOGISTIC** data=Churn;

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

**RUN**;

\*Fit logistic model, and run backward selection procedure (DID THIS TO TEST HOW IT WORKS);

**PROC** **LOGISTIC** data=Churn;

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

**RUN**;

\*Fit logistic model, and run stepwise selection procedure (DID THIS TO TEST HOW IT WORKS);

**PROC** **LOGISTIC** data=Churn;

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

**RUN**;

\*Predicted Values;

**data** new;

input LAST\_PRICE\_PLAN\_CHNG\_DAY\_CNT 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**;

\*Merge prediction dataset with original dataset;

**data** prediction;

set new Churn;

**RUN**;

**PROC** **PRINT**;

**RUN**;

\*Run prediction;

**PROC** **LOGISTIC**;

TITLE"Predictions";

MODEL CHURN (event='1')=LAST\_PRICE\_PLAN\_CHNG\_DAY\_CNT 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 predicted probabilities and confidence intervals;

**PROC** **PRINT** data=prediction;

**RUN**;