XIMAN LIU

DATA ANALYSIS AND REGRESSION

Assignment-6 | Total Points: 20 pts for DSC 423 and 10 pts for DSC 323

Due Date: 5/17/2021 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.**

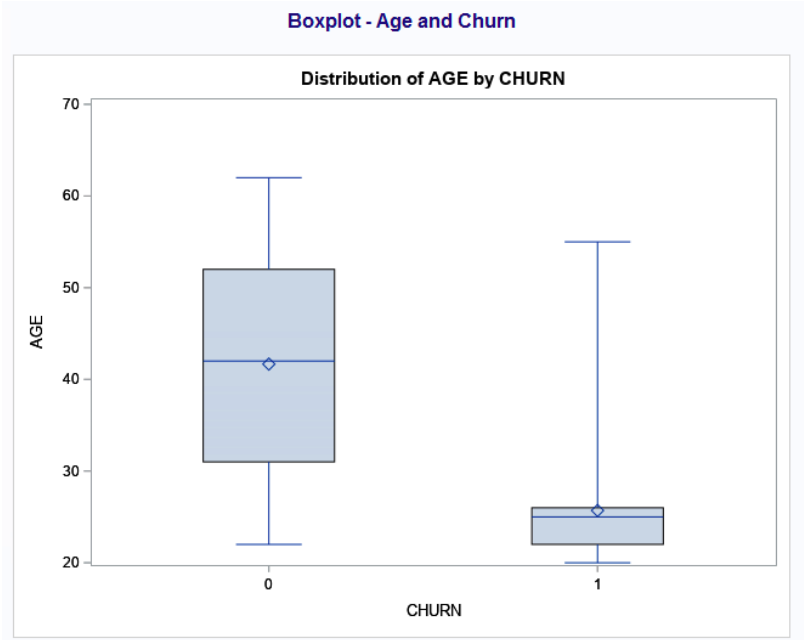
Problem 1 [10 pts] Churn analysis – to be answered by everyone

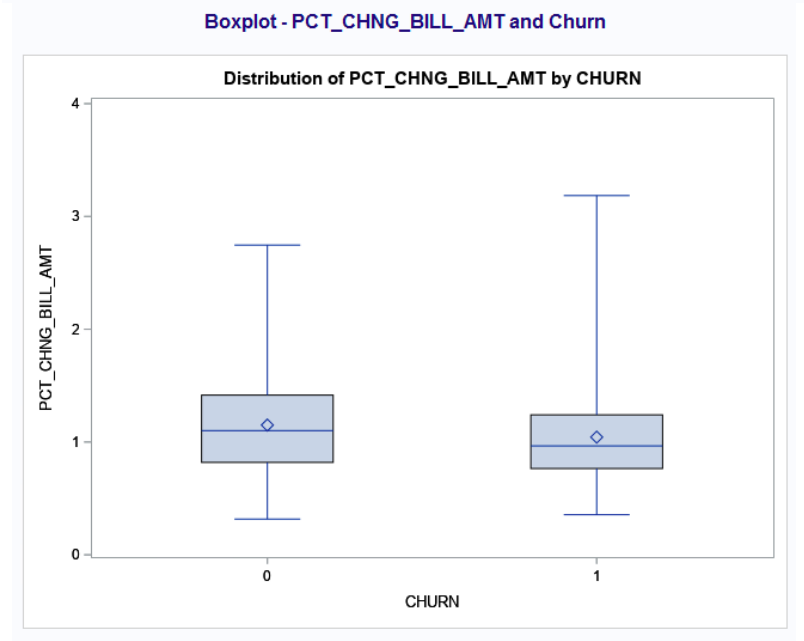
Given the large number of competitors, cell phone carriers are very interested in analyzing and predicting customer retention and churn. The primary goal of churn analysis is to identify those customers that are most likely to discontinue using your service or product. The dataset churn_train.csv contains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

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.





For the age graph, it's clear that the middle 50% of consumers in the survey who didn't switch providers (churn=0) and those who did (churn=1) have clear differences. As we all know, the boxes located in the middle of higher and lower extremes, the consumers who did not switch providers seem to have an average age. 75% of the chosen sample (Q3) is under the age of 52, while the remaining 25% is between the ages of 52 and 64. Vice versa, for the customers who switched providers, they have box which way lower than 25% (Q1) of the customers who did not switch. 75% (Q3) of consumers who switched are under 25 years old, with the remaining 25% being between the ages of 25 and 55. We can deduce from the longer whisker that, for consumers who did not switch providers, they have a broader range of retention from age of 22 to 64. The medians of both genres are quite close to their means. It is quite obviously that for customers who did not switch providers, their median and mean are overlapped. It represents

that the distribution of age by churn = 0 is symmetric and normal. On the other side, for customers who has switched provider, the median is lower than the mean indicates that the distribution of age by churn = 1 is light skewed right.

For the second graph, in terms of percent difference in the last two months' bill number, the middle 50% of customers in the study who did not move provider (churn=0) and customers who did (churn=1) are reasonably close. Since these boxes are closest to the lower extreme, all groups of consumers seem to see a lower percentage shift. 75% (Q3) of consumers who did not switch providers, the number is less than 1.6%, and the remaining 25% is up to 2.8%. Vice versa, for the customers who have switched provider, 75% (Q3) of the selected sample is under 1.5% and with the remaining 25% up to 3.2%. We may deduce that consumers who has changed providers have a greater range of retention, ranging from 0.5% to 3.2%. The medians for both genres are slightly lower than the means. It indicates that the distributions of percent change of bill amount by churn are slightly skewed right.

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.

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	1.6238	0.3185	25.9869	<.0001
PCT_CHNG_IB_SMS_CNT	1	-0.4156	0.1186	12.2768	0.0005
PCT_CHNG_BILL_AMT	1	-0.5088	0.1735	8.5968	0.0034
TOT_ACTV_SRV_CNT	1	-0.6103	0.0505	145.8899	<.0001
dGEN	1	0.0286	0.1551	0.0340	0.8537
dEDU1	1	-0.0286	0.1529	0.0351	0.8515
dEDU2	1	0.7325	0.4521	2.6249	0.1052
dEDU3	1	0.7903	0.6503	1.4768	0.2243
dEDU4	1	12.8939	507.2	0.0006	0.9797
dEDU5	1	0.5573	1.1614	0.2302	0.6313
dLPPC	1	0.0815	0.4514	0.0326	0.8566
dCOMP	1	0.5575	0.1751	10.1320	0.0015

Hypotheses

H0: $\beta1=\beta2=0$ (hypothesis of all parameters=0 corresponds to an "empty" model or M0: logit(p)= $\beta0$)

Ha: all $\beta i \neq 0$ (hypothesis corresponds to model with some covariates or Model M1: logit(p) = $\beta0+ \beta1x1+\beta2x2$)

Based on t-test, ChiSq with p-value less than 0.05. It provides strong support that the null hypothesis is rejected. In that case, the corresponding x-variables is valid in the model. Thus, PCT_CHNG_IB_SMS_CNT, PCT_CHNG_BILL_AMT, TOT_ACTV_SRV_CNT and dCOMP (dummy variable) are important variables as all the p-values for ChiSq are less than 0.05. They should appear in the model.

The next step is to use stepwise method and backward method to identify the final model.

Stepwise Method

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	1.6918	0.2906	33.9006	<.0001
PCT_CHNG_IB_SMS_CNT	1	-0.4290	0.1182	13.1618	0.0003
PCT_CHNG_BILL_AMT	1	-0.5129	0.1717	8.9237	0.0028
TOT_ACTV_SRV_CNT	1	-0.6030	0.0501	144.8608	<.0001
dCOMP	1	0.5407	0.1725	9.8222	0.0017

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	1362.479	1144.529
SC	1367.370	1168.982
-2 Log L	1360.479	1134.529

R-Square	0.2054	Max-rescaled R-Square	0.2740

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	225.9499	4	<.0001
Score	203.6428	4	<.0001
Wald	170.2585	4	<.0001

Backward Method

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	1.6918	0.2906	33.9006	<.0001
PCT_CHNG_IB_SMS_CNT	1	-0.4290	0.1182	13.1618	0.0003
PCT_CHNG_BILL_AMT	1	-0.5129	0.1717	8.9237	0.0028
TOT_ACTV_SRV_CNT	1	-0.6030	0.0501	144.8608	<.0001
dCOMP	1	0.5407	0.1725	9.8222	0.0017

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	1362.479	1144.529
SC	1367.370	1168.982
-2 Log L	1360.479	1134.529

R-Square	0.2054	Max-rescaled R-Square	0.2740

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	225.9499	4	<.0001
Score	203.6428	4	<.0001
Wald	170.2585	4	<.0001

These variables should appear in the final model according to above methods: PCT_CHNG_IB_SMS_CNT, PCT_CHNG_BILL_AMT, TOT_ACTV_SRV_CNT and dCOMP.

Final model (churn=1) for customers who have switched provider:

Log odds = 1.6918 – 0.429 * PCT_CHNG_IB_SMS_CNT – 0.5129 * PCT_CHNG_BILL_AMT – 0.603 * TOT_ACTV_SRV_CNT + 0.5407 * dCOMP + e

(dummy variable dCOMP = 1 if at least one consumer complaints appears in last two months)

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

For each percent change of latest 2 months incoming SMS wrt previous 4 months incoming SMS, the log odds of churn decrease by 0.429. As exp(0.429) = 1.54, the odds (p/1-p) decreases 54%, for each percent change of latest 2 months incoming SMS wrt previous 4 months incoming SMS.

For each percent change of latest 2 months bill amount wrt previous 4 months bill amount, the log odds of churn decrease by 0.5129. As exp(0.5129) = 1.67, the odds (p/1-p) decreases 67% for each percent change of latest 2 months bill amount wrt previous 4 months bill amount.

For each number increase in total number of active services, the log odds of churn decrease by 0.603. As exp(0.603) = 1.83, the odds (p/1-p) decreases 83% for each number increase in total number of active services.

As if there was at least a customer's complaint appeared in two months, the log odds of churn increase by 0.5407. As exp(0.5407) = 1.717, the odds (p/1-p) increases 71.7% for each number increase in total number of active services as if there was at least a customer's complaint appeared in 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 LAST_PRICE_PLAN_CHNG_DAY_CNT=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.

Prediction - Final Model

Obs	PCT_CHNG_IB_SMS_CNT	PCT_CHNG_BILL_AMT	TOT_ACTV_SRV_CNT	dCOMP	GENDER	EDUCATION	LAST_PRICE_PLAN_CHNG_DAY_CNT	AGE	CHURN	COMPLAINT
1	1.04	1.19	4	1		.	.	.	.	.

LEVEL	phat	lcl	ucl
1	0.22514	0.18135	0.27594

For the final model, age and LAST_PRICE_PLAN_CHNG_DAY_CNT or its dummy variable LAST_PRICE_PLAN_CHNG_DAY_CNT = 1 are not included. Predicted probability is 0.22514 phat and the 95% confidential interval is from 0.18135 to 0.27594. It means 95% of the time, the predicted probability will fall in 0.18135 to 0.27594.

As there is a percent change of latest 2 months incoming SMS wrt previous 4 months incoming SMS, percent change of latest 2 months bill amount wrt previous 4 months bill amount, total no. of active services is 4, and at least a customer's complaint in the two months, the chances of consumers switching providers rise from 19.88% and 31.78%.

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

```
TITLE "Analysis & Prediction - churn_train";

PROC IMPORT datafile = "C:\Users\XLIU115\Desktop\Assignment6\churn_train.csv"
out = phone replace;
getnames=yes;
delimiter=',';
RUN;

DATA phone_new;
set phone;
dGEN=(GENDER='M');
dEDU1=(EDUCATION=2);
dEDU2=(EDUCATION=3);
dEDU3=(EDUCATION=4);
dEDU4=(EDUCATION=5);
dEDU5=(EDUCATION=6);
dLPPC=(LAST_PRICE_PLAN_CHNG_DAY_CNT=1);
dCOMP=(COMPLAINT=1);
RUN;

PROC PRINT;
RUN;


/*1a*/
TITLE "Boxplot - Age and Churn";
PROC SORT;
by Churn;
RUN;
```

```
PROC BOXPLOT;
plot Age*Churn;
RUN;
TITLE "Boxplot - PCT_CHNG_BILL_AMT and Churn";
PROC SORT;
by Churn;
RUN;
PROC BOXPLOT;
plot PCT_CHNG_BILL_AMT*Churn;
RUN;


/*1b*/
TITLE "Logistic Regression - Full Model";
PROC LOGISTIC;
model Churn (event='1') = PCT_CHNG_IB_SMS_CNT PCT_CHNG_BILL_AMT TOT_ACTV_SRV_CNT
dGEN dEDU1 dEDU2 dEDU3 dEDU4 dEDU5 dLPPC dCOMP;
RUN;
/*Remove unimportant variables*/
PROC LOGISTIC;
model Churn (event='1') = PCT_CHNG_IB_SMS_CNT PCT_CHNG_BILL_AMT TOT_ACTV_SRV_CNT
dCOMP
                                          /selection=stepwise rsquare;
RUN;
PROC LOGISTIC;
model Churn (event='1') = PCT_CHNG_IB_SMS_CNT PCT_CHNG_BILL_AMT TOT_ACTV_SRV_CNT
dCOMP
                                          /selection=backward rsquare;
RUN;
TITLE "Logistic regression - Final Model";
PROC LOGISTIC;
model Churn (event='1') = PCT_CHNG_IB_SMS_CNT PCT_CHNG_BILL_AMT TOT_ACTV_SRV_CNT
dCOMP;
RUN;


/*1c*/
TITLE "Prediction - Final Model";
DATA phone_pred;
input PCT_CHNG_IB_SMS_CNT PCT_CHNG_BILL_AMT TOT_ACTV_SRV_CNT dCOMP;
datalines;
1.04 1.19 4 1
;
DATA phone_comb;
set phone_pred phone_new;
RUN;
PROC PRINT;
RUN;
PROC LOGISTIC data=phone_comb;
model Churn (event='1') = PCT_CHNG_IB_SMS_CNT PCT_CHNG_BILL_AMT TOT_ACTV_SRV_CNT
dCOMP;
output out=finalPrediction p=phat upper=ucl lower=lcl;
RUN;
PROC PRINT data=finalPrediction;
RUN;
```

Problem 2 [10 pts] – ONLY for GRADUATES

Based on the model built in Problem 1 above, answer the following questions:

1. Compute the R2 value using SAS, provide the relevant output and
    a. Explain the R2 value for your final model.

R-Square	0.2054	Max-rescaled R-Square	0.2740

R2 (20.54%) indicate the log odds of churn explained by the final logistic regression model. However, R2 has a relatively low rate which indicated that this is not a fairly ideally model.

    b. What percentage is unexplained by the model?

The model R-square value is 0.2054 therefore, 79.46% (100-20.54) is unexplained by the model.

    c. What actions can you take to further improve the model? List and explain 5 actions.

1. Binary logistic regression requires the DV to be binary (1,0)

2. Since logistic regression assumes that P(Y=1) is the probability of the event occurring, it is necessary that the DV is coded accordingly. For a binary regression, the factor level 1 of the dependent variable should represent the desired outcome

3. The model requires quite large sample sizes. Because maximum likelihood estimates are less powerful than ordinary least squares.

• At least 10 cases per independent variable; In certain situation we need at least 30 cases for each parameter to be estimated

• Make sure enough observations for each case (1 & 0)

4. Model should have little or no multicollinearity. If multicollinearity is present centering the variables might resolve the issue. If this does not lower the multicollinearity, a factor analysis with orthogonally rotated factors should be done before the logistic regression is estimated

5. Model should have no outliers or significant influential points.

• Outliers: Use Pearson or Deviance residual close to or exceeding ± 3

• Influential Points: Use Dfbetas

2. Explain why you didn't check the model assumptions for this model?

First, logistic regression does not require a linear relationship between the dependent and independent variables. Second, residuals do not need to be normally distributed. Third, homoscedasticity is not required. Finally, the dependent variable in logistic regression is not measured on an interval or ratio scale.