DSC 423 - Data Analysis and Regression
Assignment 4 - Due on Friday November 10th, 2023
Total Points: 27 pts

Problem 1 Churn analysis [16 pts]

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
   LAST_PRICE_PLAN_CHNG_DAY_CNT: No. of days since last price plan change
4. TOT_ACTV_SRV_CNT: Total no. of active services
5. AGE: customer age
6. PCT_CHNG_IB_SMS_CNT: Percent change of latest 2 months incoming SMS wrt previous 4 months incoming SMS
7. PCT_CHNG_BILL_AMT: Percent change of latest 2 months bill amount wrt previous 4 months bill amount
8. 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. **[1 pt for R code for boxplots, 1 pt for analysis = 2 pts]**

b) Fit a 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). Remove x-variables that are not significant using alpha=0.05. Write down the expression of the fitted model. (HINT: probability of interest is p = pr(churn = 1) **[1 pt R code for model, 1 pt non-significant x-variables, 1 pt expression = 3 pts]**

c) Analyze the final logistic regression model and discuss the effect of each variable on the churn probability. Discuss results in terms of odds ratios. **[1 pt residual plot, 1 pt odds ratio, 1 pt discussion = 3 pts]**

d) Compute the predicted churn probability and the prediction 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. **[1 pt computing predicted probability, 1 pt prediction interval = 2 pts]**

e) The dataset *churn_test.csv* contains a new set of customers, and can be used to test the validity of the churn predictive model. Apply the methods discussed **in week 8 lecture** to identify a threshold T for the predicted churn probability in order to define a classification rule for customers, so that

- predicted probability p(churn) > = T, then customer is a "likely churn", and
- predicted probability p(churn) < T, then customer is a "unlikely churn".

Compute the optimal T value, and create the classification matrix summarizing classification results. Hint: You can use the Classify_functions.R in your solution. **[1 pt for R code for getting range of thresholds to choose optimal T, 1 pt for computing predicted churn outcomes corresponding to predicted probabilities, 1 pt for computing confusion matrix and metrics, 1 pt for selecting the P* that optimizes a certain metric, 1 pt for computing predicted churn outcomes for a separate testing set (churn_test.csv), 1 pt for computing the confusion matrix that summarizes classification results and metrics = 6 pts]**

Problem 2 [11 pts]

A researcher is interested in evaluating the relationship between energy consumption by the homeowner and the difference between the internal and external temperatures. A sample of 30 homes was used in the study. During an extended period of time, the average temperature difference (in ^{o}F) (TEMPD) inside and outside the homes was recorded. The average energy consumption (ENERGY) was also recorded for each home. The data are stored in the *energytemp.txt* data file.

a) Create a scatterplot of ENERGY (y) versus TEMPD (x) to visualize the association between the two variables. Analyze the association displayed by the scatterplot. **[1 pt for scatterplot, 1 pt for analysis = 2 pts]**

b) Fit a cubic model $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + e$ (*HINT: create two new variables TEMP2 and TEMP3:*

In R use the code:
tempd2 = tempd^2;
tempd3 = tempd^3;

Include the new variables in the regression model) **[1 pt]**

c) Are all variables in the model significant? **[1 pt]**

d) Create the residual plots (residuals vs predicted; residuals vs x variable; and normal plot of residuals). Analyze residual plots to evaluate the normality and constant variance assumptions. Discuss your findings. **[3 pts for residual plots, 1 pt for analysis = 4 pts]**

e) If you are satisfied with the fitted regression model, write down its expression. **[1 pt]**

f) Use the fitted regression model to predict the average energy consumption for an average difference in temperature equal to TEMPD=10.
(HINT:

In R use the following code:

```
new <- data.frame(tempd=c(10), tempd2=c(100), tempd3=c(1000))
```

then use the predict() function with the fitted regression model as explained in the document under week 5. **[1 pt for R code, 1 pt for answer = 2 pts]**

Submission instructions:

Submit the homework at the Course Web page at https:\\d2l.depaul.edu

- Keep a copy of all your submissions!
- If you have questions about the homework, email me BEFORE the deadline.
- The assignment will lose 20% of the points if submission is late.
- Assignments submitted four days after the due date will not be accepted.