**Homework #11 (Due Nov 12 11:59 PM)**

IST 3420 - Fall 2017, Chen

**Name**: \_\_\_\_\_\_Adam Forestier\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Customer Churn Analysis (20 points)**

1. Download data file “Telco-Customer-Churn.csv”. Read the dataset into R.
2. Transform variable “Churn” as numeric type. Use 1 to represent “Yes” and 0 to represent “No”. (1 point)
3. Show summary statistics of the dataset. Are there missing values in the dataset? If yes, explain the detail of missing data. (2 points)

|  |
| --- |
| Yes, many of the columns in the data frame have values such as 0 to represent “No.” |

1. Calculate the percentage of customers who churned. Fill the result in the following box. (1 point)

|  |
| --- |
| % of churn = 26.5% |

1. Use multiple linear regression model (also known as OLS, ordinary least squares) to fit the data. In the linear model, regress Churn on other variables. Use summary() function to show the result of multiple regression model. Paste the result in the following box. (1 point)

|  |
| --- |
|  |

1. Explain the effect of SeniorCitizen on customer churn in the following box. (2 points)

|  |
| --- |
| 1. There is significance because p-value < .001. It has positive effect on churn with a coefficient .0443 which is not very significant. |

1. Use logistic regression model to fit the data. In the logistic model, regress Churn on other variables. Use summary() function to show the result of regression analysis. Paste the result in the following box. (1 point)

|  |
| --- |
|  |

1. Explain the effect of SeniorCitizen on customer churn in the following box. Comparing with multiple linear regression model, do you get similar conclusion from the logistic model? (2 points)

|  |
| --- |
| There is significance because p-value < .05. It has a positive effect on churn with a coefficient of .2153. Both show a positive coefficient, and a significant p-value. The multilinear model has a more significant small p-value, but the logistic model is much more statistically significant with a much higher coefficient. |

1. Use stargazer() function (in stargazer package) to show the results of both multiple linear regression and logistic regression analyses. Paste the result in the following box. (2 points)

|  |
| --- |
| Multiple Linear vs. Logistic Regression  ============================================================================  Dependent variable:  ---------------------------------------  Churn  OLS logistic  (1) (2)  ----------------------------------------------------------------------------  SeniorCitizen 0.0443\*\*\* 0.2153\*  (0.0129) (0.0840)    DependentsYes -0.0207\* -0.1485  (0.0103) (0.0814)    Tenure -0.0020\*\*\* -0.0599\*\*\*  (0.0005) (0.0062)    MultipleLinesNo phone service 0.0318 0.6317\*\*\*  (0.0180) (0.1328)    MultipleLinesYes 0.0521\*\*\* 0.2484\*\*  (0.0109) (0.0796)    InternetServiceFiber optic 0.1778\*\*\* 0.7439\*\*\*  (0.0139) (0.0983)    InternetServiceNo -0.1459\*\*\* -0.7668\*\*\*  (0.0178) (0.1374)    OnlineSecurityNo internet service      OnlineSecurityYes -0.0490\*\*\* -0.4031\*\*\*  (0.0120) (0.0848)    OnlineBackupNo internet service      OnlineBackupYes -0.0177 -0.1732\*  (0.0114) (0.0775)    TechSupportNo internet service      TechSupportYes -0.0505\*\*\* -0.3849\*\*\*  (0.0122) (0.0858)    StreamingTVNo internet service      StreamingTVYes 0.0505\*\*\* 0.1852\*  (0.0120) (0.0813)    StreamingMoviesNo internet service      StreamingMoviesYes 0.0525\*\*\* 0.1938\*  (0.0120) (0.0813)    ContractOne year -0.1059\*\*\* -0.6676\*\*\*  (0.0140) (0.1073)    ContractTwo year -0.0703\*\*\* -1.3622\*\*\*  (0.0170) (0.1762)    PaperlessBillingYes 0.0450\*\*\* 0.3454\*\*\*  (0.0100) (0.0744)    PaymentMethodCredit card (automatic) -0.0061 -0.0889  (0.0135) (0.1140)    PaymentMethodElectronic check 0.0675\*\*\* 0.3057\*\*  (0.0133) (0.0944)    PaymentMethodMailed check -0.0067 -0.0564  (0.0145) (0.1147)    TotalCharges -0.00004\*\*\* 0.0003\*\*\*  (0.00001) (0.0001)    Constant 0.3459\*\*\* -0.5046\*\*\*  (0.0181) (0.1325)    ----------------------------------------------------------------------------  Observations 7,032 7,032  R2 0.2840  Adjusted R2 0.2821  Log Likelihood -2,914.2290  Akaike Inf. Crit. 5,868.4590  Residual Std. Error 0.3743 (df = 7012)  F Statistic 146.4055\*\*\* (df = 19; 7012)  ============================================================================  Note: \*p<0.05; \*\*p<0.01; \*\*\*p<0.001 |

1. Compare the results of multiple regression and logistic regression. Which model is more appropriate to explain the effects of explanatory variables on customer churn? Why? Explain in the following box. (3 points)

|  |
| --- |
| The results are fairly close. Logistic regression is more appropriate to explain the effects of the explanatory effects variables on the dependent variable customer churn, because churn is categorical. The number 1 stands for yes, and the number 0 stands for no; the data is binary |

1. Calculate the McFadden R2 for the logistic regression model. What is the value of the McFadden R2? Is it similar to the R2 of the multiple linear regression model? Explain in the following box. (2 points)

|  |
| --- |
| McFadden pseudo R2 = 0.2848645  Adjusted R2 = 0.2821  R2 = 0.2840  The pseudo R2 in the linear model is fairly close to Adjusted R and R2 in the multiple linear regression model. The pseudo R2 is just barely larger than Adjusted R2 and almost identical to R2. |

1. How many observations are in the dataset? How many observations are actually used to fit the logistic regression model? Are these two numbers the same? Why? Explain in the following box. (2 points)

|  |
| --- |
| There are 7043 observations in the dataset. 7,032 observations were used to fit the logistic regression model. There are 11 NA values; these numbers are not the same, this is because NA values were omitted when fitting the model. |

***Homework Submission***

1. Upload this document with your answers to “Homework 11” on Canvas.
2. Upload your R Markdown file to “Homework 11” on Canvas. (1 point)