NBM2 Task 2: Logistic Regression for Predictive Modeling

RathiPriyanka Srinivasan

Department of Information Technology, Western Governors University

D208: Predictive Modeling

Dr. Daniel Smith

May 8, 2021

A. Research Question

1. Do various demographic factors impact customer churn?
2. Analysis on which demographic factors like income, age, occupation status, number of children and marital status effect churn, will be crucial data for the telecommunications company to tailor and center their marketing on customers that are predicted to have lower churn rates. With this information, the telecommunications company should be able to market services on a community and corporate level leading to increased and maintained sales while lowering the current churn rate. Longer implications of customer demographic analysis allow the telecommunications company to gradually replace their current customer base with customers that will be less likely to discontinue services leading to more consistent quarterly and annual sales projections.

B. Method Justification

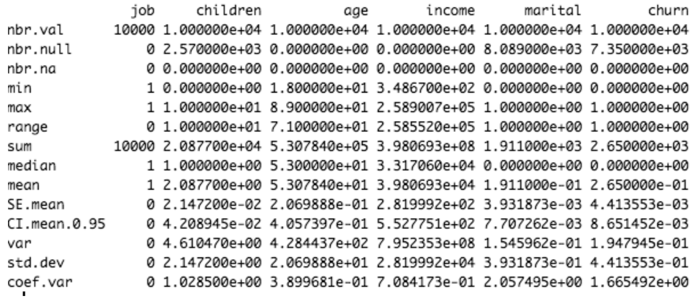
1. Binary logistic regression assumes that the dependent variables are binary while ordinal logistic regression is performed with the assumption that the dependent variables are ordinal and that there are no strongly influential outliers (Stoltzfus, 2011). Logistic regression also presumes that there is little or no multicollinearity among the independent variables while assuming the linearity of independent variables and log odds (Stoltzfus, 2011). Finally, logistic regression is utilized with the understanding that there is a large sample of observations within the data being analyzed (Stoltzfus, 2011).
2. R will be the chosen tool for data preparation, analysis and manipulation. R is flexible to use as there is a large user community that provides packages for

various data preparations, data manipulations, and generally has higher time performance for smaller datasets (Brittain et al., 2018). R can also be used for modeling boxplots, histograms, and visualizing the distribution of data (Kabacoff, 2017).

3. Logistic regression is an appropriate analysis technique for this data set as it is linear, simple to understand, and there is an option to have the output as a classification (Massaron & Boschetti, 2016). Logistic regression is also easy to train allowing the model to be stored easily (Massaron & Boschetti, 2016).. Finally, logistic regression is less prone to overfitting and has an extension for “multiclass classification” (Massaron & Boschetti, 2016).

C. Data Preparation

1. The data was scrubbed of all information except for the following variables: churn, income, age, job, number of children, and marital status (Scenario: Telecommunications Churn, 2021). The three categorical variables: churn, job, and marital status will be converted to numerical by marking each relevant or ‘Yes’ entry as ‘1’ and non-relevant or ‘No’ entry as ‘0’.



```
               job     children          age       income      marital        churn
nbr.val      10000 1.000000e+04 1.000000e+04 1.000000e+04 1.000000e+04 1.000000e+04
nbr.null         0 2.570000e+03 0.000000e+00 0.000000e+00 8.089000e+03 7.350000e+03
nbr.na           0 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00
min              1 0.000000e+00 1.800000e+01 3.486700e+02 0.000000e+00 0.000000e+00
max              1 1.000000e+01 8.900000e+01 2.589007e+05 1.000000e+00 1.000000e+00
range            0 1.000000e+01 7.100000e+01 2.585520e+05 1.000000e+00 1.000000e+00
sum          10000 2.087700e+04 5.307840e+05 3.980693e+08 1.911000e+03 2.650000e+03
median           1 1.000000e+00 5.300000e+01 3.317060e+04 0.000000e+00 0.000000e+00
mean             1 2.087700e+00 5.307840e+01 3.980693e+04 1.911000e-01 2.650000e-01
SE.mean          0 2.147200e-02 2.069888e-01 2.819992e+02 3.931873e-03 4.413553e-03
CI.mean.0.95     0 4.208945e-02 4.057397e-01 5.527751e+02 7.707262e-03 8.651452e-03
var              0 4.610470e+00 4.284437e+02 7.952353e+08 1.545962e-01 1.947945e-01
std.dev          0 2.147200e+00 2.069888e+01 2.819992e+04 3.931873e-01 4.413553e-01
coef.var         0 1.028500e+00 3.899681e-01 7.084173e-01 2.057495e+00 1.665492e+00
```

2. The mean is greater than the median for the target and all predictor variables, suggesting that the data is not symmetric and skewed to the right. The observations for the children, age, and income predictor variables are spread within approximately two standard deviations on either side of the mean. For the marital predictor variable and churn target

variable, the data is spread within approximately four standard deviations on either of the mean. A higher standard deviation indicates a greater spread within the data. The minimum, maximum, and range will be the same for job, marital and churn variables as they were converted from qualitative to quantitative data.

3. The following steps were used to prepare the medical dataset for analysis:

    i. Access the installed packages including tidyverse, openxlsx, plotly and zoo

    ii. Load the data from the provided CSV file (Scenario: Telecommunications Churn, 2021)

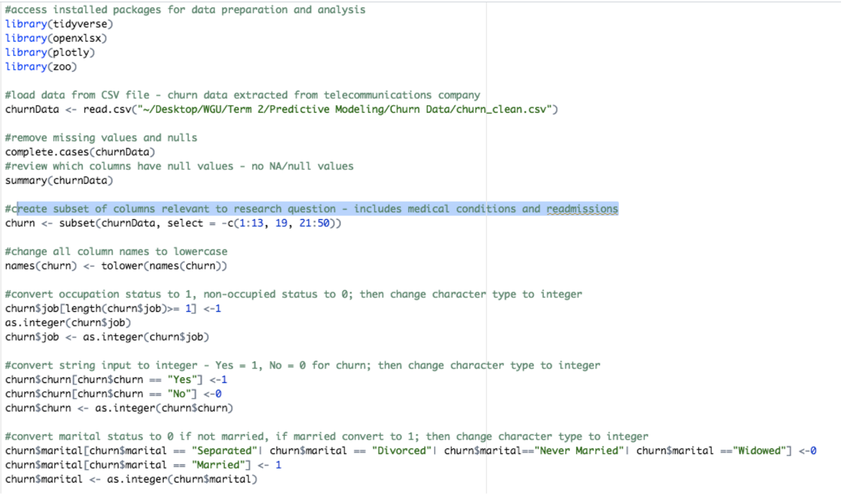
    iii. Scrub data of missing values and nulls

    iv. Create subset 'churn' that includes columns relevant to the research question specified in Part A. Columns include: job, children, age, income, marital, and churn

    v. Change all columns names to lowercase

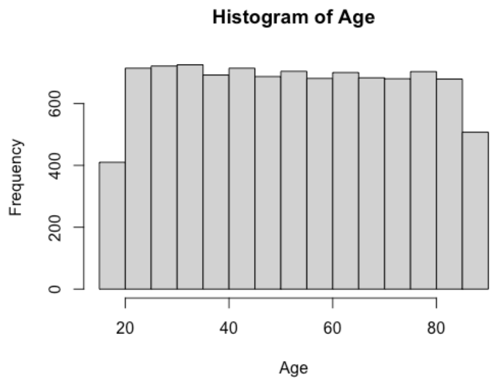
    vi. In 'churn' convert all character variables to integers by changing relevant or 'Yes' entries to 1 and non-relevant or 'No' entries to 0

The screenshot of the annotated code in RStudio has been provided below:

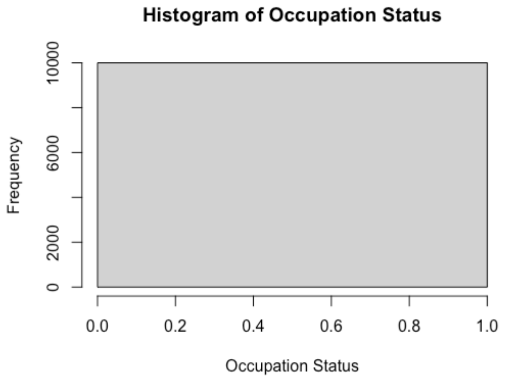


```
#access installed packages for data preparation and analysis
library(tidyverse)
library(openxlsx)
library(plotly)
library(zoo)

#load data from CSV file - churn data extracted from telecommunications company
churnData <- read.csv("~/Desktop/WGU/Term 2/Predictive Modeling/Churn Data/churn_clean.csv")

#remove missing values and nulls
complete.cases(churnData)
#review which columns have null values - no NA/null values
summary(churnData)

#create subset of columns relevant to research question - includes medical conditions and readmissions
churn <- subset(churnData, select = -c(1:13, 19, 21:50))

#change all column names to lowercase
names(churn) <- tolower(names(churn))

#convert occupation status to 1, non-occupied status to 0; then change character type to integer
churn$job[length(churn$job)>= 1] <-1
as.integer(churn$job)
churn$job <- as.integer(churn$job)

#convert string input to integer - Yes = 1, No = 0 for churn; then change character type to integer
churn$churn[churn$churn == "Yes"] <-1
churn$churn[churn$churn == "No"] <-0
churn$churn <- as.integer(churn$churn)

#convert marital status to 0 if not married, if married convert to 1; then change character type to integer
churn$marital[churn$marital == "Separated"| churn$marital == "Divorced"| churn$marital=="Never Married"| churn$marital =="Widowed"] <-0
churn$marital[churn$marital == "Married"] <- 1
churn$marital <- as.integer(churn$marital)
```

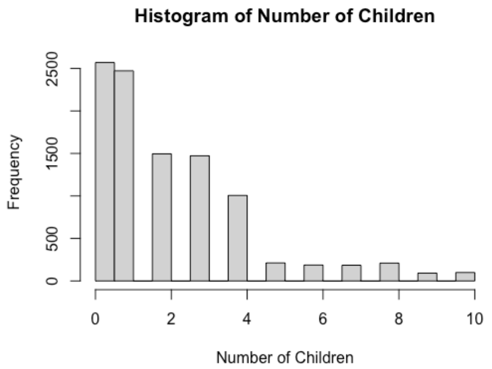
4. Univariate visualization of all predictor variables are provided below:



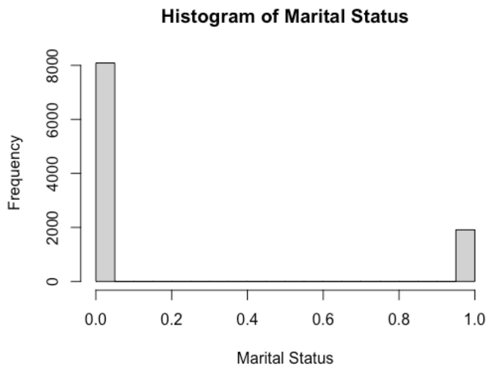
Histogram of Age



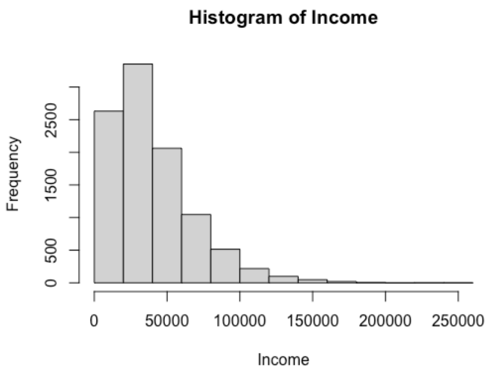
Histogram of Occupation Status



Histogram of Number of Children

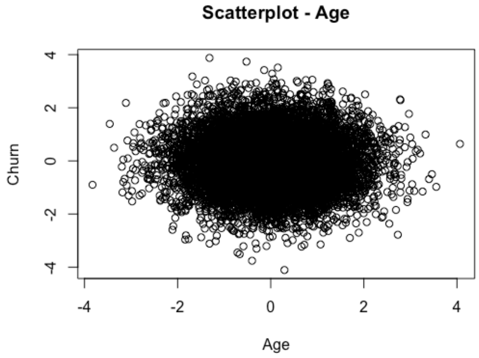


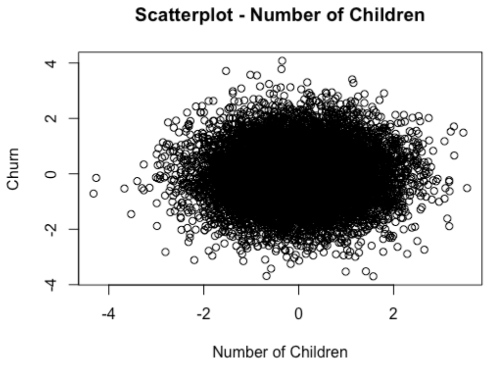
Histogram of Marital Status

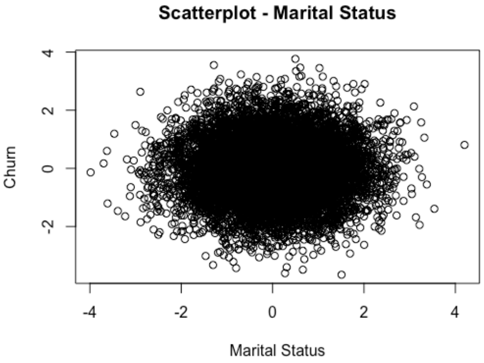


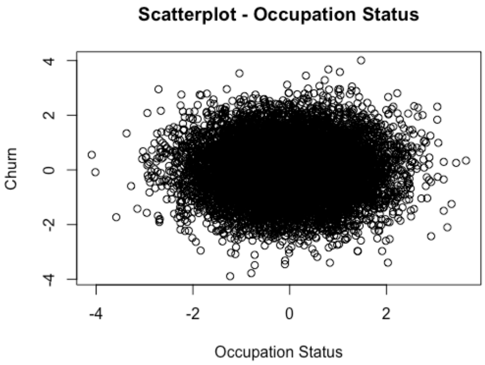
Histogram of Income

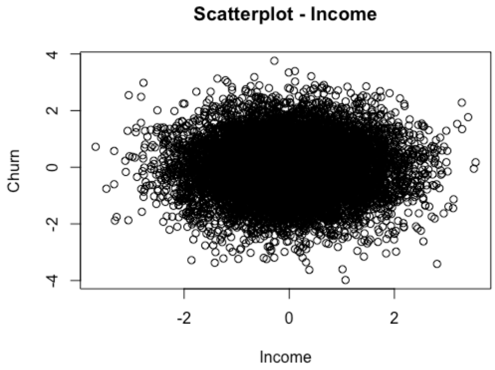
Bivariate visualizations with the target variable and each predictor variable is included below:


Scatterplot - Age


Scatterplot - Number of Children

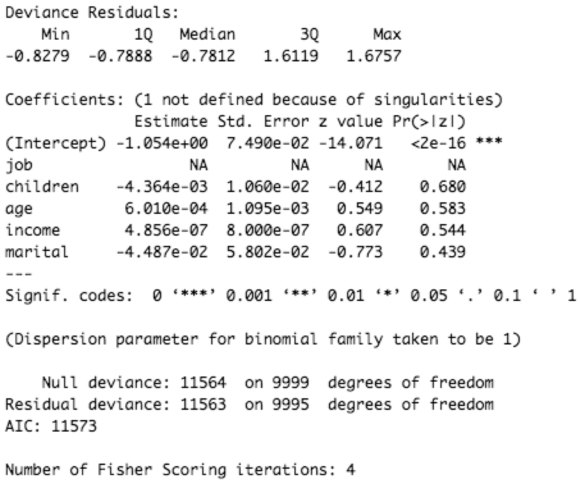

Scatterplot - Marital Status


Scatterplot - Occupation Status


Scatterplot - Income

D. Model Comparison and Analysis

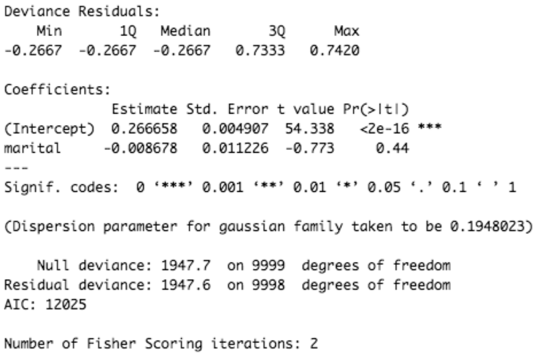
1. Initial logistic regression model with all predictor variables (Occupation Status is not included because of singularities): Churn = .004364*Children + .000601*Age + .0000004856*Income - .04487*MaritalStatus – 1.054. Occupation status was excluded from the initial regression model due to singularities, or a high level of collinearity.



```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.8279  -0.7888  -0.7812   1.6119   1.6757

Coefficients: (1 not defined because of singularities)
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.054e+00  7.490e-02 -14.071   <2e-16 ***
job                 NA         NA      NA       NA
children    -4.364e-03  1.060e-02  -0.412    0.680
age          6.010e-04  1.095e-03   0.549    0.583
income       4.856e-07  8.000e-07   0.607    0.544
marital     -4.487e-02  5.802e-02  -0.773    0.439
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 11564  on 9999  degrees of freedom
Residual deviance: 11563  on 9995  degrees of freedom
AIC: 11573

Number of Fisher Scoring iterations: 4
```

2. Forward step-wise regression will be the variable selection technique used for analyzing churn data. Advantages to using forward step-wise regression include processing predictor variables and adjusting the logistic regression model to choose the best predictor variables from the available options (NCSS Statistical Software, 2021). Forward step-wise regression also allows for step by step analysis of the model by providing information about the quality of predictor variables and how they influence the logistic regression model as they are added in the variable selection technique (NCSS Statistical Software, 2021). Observing which demographic factors have greater influence on customer churn will be crucial for marketing purposes and reducing customer churn over the long run.

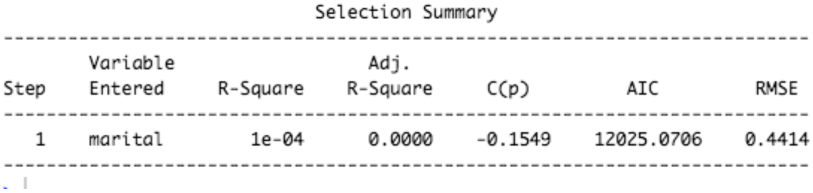
3. Reduced logistic regression model: Churn = -0.008678*MaritalStatus + 0.266658



```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.2667  -0.2667  -0.2667   0.7333   0.7420

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.266658   0.004907  54.338   <2e-16 ***
marital     -0.008678   0.011226  -0.773     0.44
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.1948023)

    Null deviance: 1947.7  on 9999  degrees of freedom
Residual deviance: 1947.6  on 9998  degrees of freedom
AIC: 12025

Number of Fisher Scoring iterations: 2
```

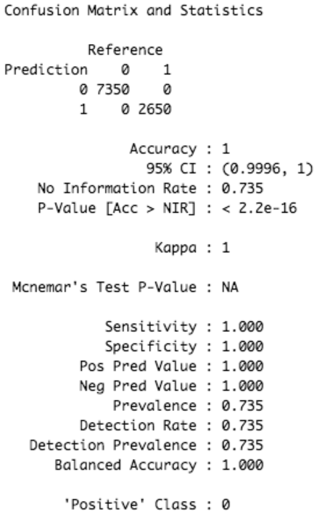
4. Forward stepwise regression starts with no predictor variables in the model, then selects the predictor variable with the highest R-squared. At each step, predictor variables are then selected by observing the greatest influence on R-squared as it increases. The analysis is finished when none of the predictor variables are significant. In this case, significance is measure by the model evaluation metric, p-value = 0.05.

5. Output for forward stepwise regression:



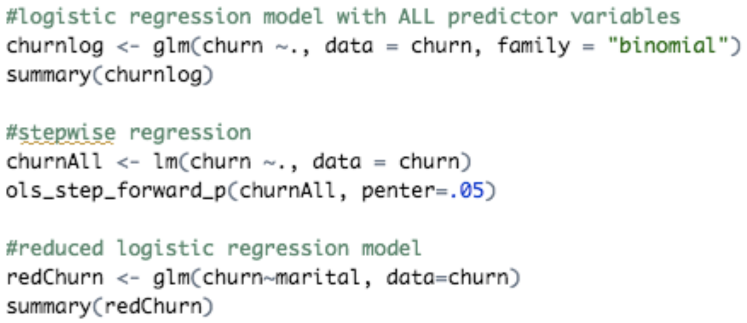
```
                              Selection Summary
---------------------------------------------------------------------------
       Variable                   Adj.
Step   Entered      R-Square    R-Square     C(p)         AIC         RMSE
---------------------------------------------------------------------------
   1   marital        1e-04      0.0000     -0.1549    12025.0706    0.4414
---------------------------------------------------------------------------
```

Confusion Matrix:



```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 7350    0
         1    0 2650

               Accuracy : 1
                 95% CI : (0.9996, 1)
    No Information Rate : 0.735
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 1

 Mcnemar's Test P-Value : NA

            Sensitivity : 1.000
            Specificity : 1.000
         Pos Pred Value : 1.000
         Neg Pred Value : 1.000
             Prevalence : 0.735
         Detection Rate : 0.735
   Detection Prevalence : 0.735
      Balanced Accuracy : 1.000

       'Positive' Class : 0
```

6. Annotated code for implementation of logistic regression models:



```
#logistic regression model with ALL predictor variables
churnlog <- glm(churn ~., data = churn, family = "binomial")
summary(churnlog)

#stepwise regression
churnAll <- lm(churn ~., data = churn)
ols_step_forward_p(churnAll, penter=.05)

#reduced logistic regression model
redChurn <- glm(churn~marital, data=churn)
summary(redChurn)
```

E. Data Summary and Implications

1. The reduced logistic regression equation Churn = -0.008678*MaritalStatus + 0.266658 shows that the average customer churn is influenced by -.008678 unit of change in the customer's marital status. Focusing marketing on customers that were either: divorced, separated, never married, or widowed may prove to increase sales for the long run as customers that do not have a partner at the time of purchasing services from the telecommunications company will tend to stay with the same telecommunications company for the month. These results may lead to a better understanding of customer trends, but is not a guarantee for increased sales. There are certain limitations within the data that was used for this analysis. For example, user-level data is 'fundamentally biased' (Yamaguchi, 2015). The fact that many customers now use multiple devices within one company can lead to questionable user data as the sources they are derived from are fragmented (Yamaguchi, 2015). Another issue is the large level of noise that is associated with the given user data (Yamaguchi, 2015). Having outliers can lead to different interpretations of the data. For instance, divorced parents that purchased services separately may have created double the number of outliers as

both of the separated partners may have reported the number of children they share custody of.

2. The given analysis suggests that developing marketing solutions that target specific populations that are known for not having lifetime or household partners may lead to better customer retention. For example, having advertisements and offering student discounts on college campuses. Another possible opportunity would be negotiating special service rates with local nonprofits that have large member communities of single parents such as: Pennsylvania Women Work, College Area Pregnancy Services Inc., Aggieland Pregnancy Outreach, and GATE Pregnancy Resource Center (Single Parent Agencies, 2021).

References

Brittain, J., Cendon, M., Nizzi, J., & Pleis, J. (2018). Data Scientist's Analysis Toolbox: Comparison of Python, R, and SAS Performance. *SMU Data Science Review*, 1(2).

Kabacoff, R. (2017). Scatterplots. *Quick-R*. https://www.statmethods.net/graphs/scatterplot.html

Massaron, L. & Boschetti, A. (2016). *Regression analysis with Python*. Packt Publishing.

NCSS Statistical Software. *Stepwise Regression*. (2021). https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Stepwise_Regression.pdf

Scenario: Telecommunications Churn. (2021). *Churn Data and Dictionary Files*. Western Governors University. https://access.wgu.edu/ASP3/aap/content/g9rke9s0rlc9ejd92md0.html

Single Parent Agencies. (2021). Find single parent agencies nonprofits and charities. *Great Nonprofits*. https://greatnonprofits.org/categories/view/single-parent-agencies

Stoltzfus, J. (2011). Logistic regression: a brief primer. *Society for Academic Emergency Medicine,* 18(10). 1099-104. doi: 10.1111/j.1553-2712.2011.*01185*.x

Yamaguchi, K. (2015). *7 limitations of big data in marketing analytics.* Martech. https://martech.org/7-limitations-big-data-marketing-analytics/