**Option #1 – Portfolio Project**

Josiah Williams

Colorado State University Global

MIS445: Statistics in Business Analytics

Alin Tomoiaga

July 4, 2021

**Option #1 – Portfolio Project**

Data analysis is essential for businesses looking to make informed decisions. A critical type of data analysis is regressions. Regression analysis helps sort out which variables in a dataset have an impact (Gallo, 2015). There are two critical variables in regression analysis: dependent variable, a factor someone is attempting to predict, and independent variables postulated to impact the dependent variable. A case study in churn and income predictions from a telecommunication customer dataset demonstrates two regression models. A logistic regression model is used to determine the odds and probability that a customer will churn given their age. Multiple linear regression predicts a customer’s income given their gender and years living at their current address.

**Sample Characteristics**

Figure 1 - Summary Statistics for Telco Data

Table

Description automatically generated

There are three continuous variables in this analysis: age, years at current address, and income. Examining the summary statistic table (figure 1), we can analyze these variables and interpret critical information regarding the data. The mean age is 41.684. Thus, the average age of customers is about 41 years old. The standard deviation for age is 12.556, which means, on average, a customer’s age deviates from the mean by approximately 12.5 years. The mean for years at the current address is 11.551. On average, a customer has lived at their current address for about 11.5 years. The standard deviation for that variable is 10.087. On average, the years a customer has lived at their address deviates from the mean by about ten years. Lastly, the mean for income is 77.535, and the standard deviation is 107.044. This statistic means that, on average, a customer’s annual income is about $77,535 and deviates from the mean by about $107,044.

Figure 2 - Frequency chart for Marital variable.

| **Marital Status** | | | | |
| --- | --- | --- | --- | --- |
| **Status** | **Frequency** | **Percent** | **Cumulative Frequency** | **Cumulative Percent** |
| **Unmarried** | 505 | 50.50 | 505 | 50.50 |
| **Married** | 495 | 49.50 | 1000 | 100.00 |

50.5% of the customers are married, while 49.5% of customers are unmarried. There is a slight percentage difference between unmarried customers versus married customers.

Figure 3 - Frequency chart for Gender variable.

| **Gender** | | | | |
| --- | --- | --- | --- | --- |
| **Gender** | **Frequency** | **Percent** | **Cumulative Frequency** | **Cumulative Percent** |
| **Male** | 483 | 48.30 | 483 | 48.30 |
| **Female** | 517 | 51.70 | 1000 | 100.00 |

48.3% of the customers are male, while 51.7% of customers are female. There are slightly more female than male customers.

Figure 4 - Frequency chart for Churn variable.

| **Churn** | | | | |
| --- | --- | --- | --- | --- |
| **Churn** | **Frequency** | **Percent** | **Cumulative Frequency** | **Cumulative Percent** |
| **Not going to churn** | 726 | 72.60 | 726 | 72.60 |
| **Going to churn** | 274 | 27.40 | 1000 | 100.00 |

Based on the frequency analysis, 72.6% of customers will not churn, while 27.4% of customers will churn. More customers in the sample data are not going to churn than those that will.

Figure 5 - Frequency chart for Region variable.

| **Region** | | | | |
| --- | --- | --- | --- | --- |
| **Zone** | **Frequency** | **Percent** | **Cumulative Frequency** | **Cumulative Percent** |
| **1** | 322 | 32.20 | 322 | 32.20 |
| **2** | 334 | 33.40 | 656 | 65.60 |
| **3** | 344 | 34.40 | 1000 | 100.00 |

32.2% of customers are in region zone 1, 33.4% of customers are in zone 2, and the remaining 34.4% are in zone 3.

Figure 6 - Frequency chart for Education variable.

| **Education** | | | | |
| --- | --- | --- | --- | --- |
| **Level** | **Frequency** | **Percent** | **Cumulative Frequency** | **Cumulative Percent** |
| **Did not complete high school** | 204 | 20.40 | 204 | 20.40 |
| **High school degree** | 287 | 28.70 | 491 | 49.10 |
| **Some college** | 209 | 20.90 | 700 | 70.00 |
| **College degree** | 234 | 23.40 | 934 | 93.40 |
| **Post-undergraduate degree** | 66 | 6.60 | 1000 | 100.00 |

20.4% of customers did not complete high school, while 28.7% of customers obtained their high school degree. Regarding college, 20.9% of customers had some college and 23.4% obtained a college degree. 6.6% of customers received their post-undergraduate degree.

Figure 7 - Frequency chart for Customer Cat variable.

| **Customer Category** | | | | |
| --- | --- | --- | --- | --- |
| **Category** | **Frequency** | **Percent** | **Cumulative Frequency** | **Cumulative Percent** |
| **Basic service** | 266 | 26.60 | 266 | 26.60 |
| **E-service** | 217 | 21.70 | 483 | 48.30 |
| **Plus service** | 281 | 28.10 | 764 | 76.40 |
| **Total service** | 236 | 23.60 | 1000 | 100.00 |

26.6% of customers receive basic service, 21.7% subscribe to the e-service, 28.1 have the plus service, and the remaining 23.6% of customers receive the total service.

Figure 8 - Histogram chart for Age, Address, and Income variables

|  |  |  |
| --- | --- | --- |
| **Chart, line chart, histogram  Description automatically generated** | Chart, line chart, histogram  Description automatically generated | Chart, box and whisker chart  Description automatically generated |

Figure 9 - Pie charts for Marital, Gender, and Churn variables

|  |  |  |
| --- | --- | --- |
| Chart, pie chart  Description automatically generated | Chart, pie chart  Description automatically generated | Chart, pie chart  Description automatically generated |

Figure 10 - Bar charts for Education, Region, and Customer Category variables.

|  |  |  |
| --- | --- | --- |
| Chart, bar chart  Description automatically generated |  |  |

**Logistic Regression – Churn**

According to Jain et al. (2020), churn in telecommunication services is very frequent because of a competitive market. Thus, it is essential for companies to proactively analyze customer behavior or characteristics to tailor the right retention campaign.Logistic regression is used to find the odds ratio in light of more than one variable (Sperandei, 2014). In this case study, a logistic regression model estimates the odds and probability that a customer will churn based on their age.

Figure - Logistic regression results for churn

| **Model Information** | | |
| --- | --- | --- |
| **Data Set** | WORK.IMPORT |  |
| **Response Variable** | churn | churn |
| **Number of Response Levels** | 2 |  |
| **Model** | binary logit |  |
| **Optimization Technique** | Fisher’s scoring |  |

|  |  |
| --- | --- |
| **Number of Observations Read** | 1000 |
| **Number of Observations Used** | 1000 |

| **Response Profile** | | |
| --- | --- | --- |
| **Ordered Value** | **churn** | **Total Frequency** |
| **1** | 1 | 274 |
| **2** | 0 | 726 |

|  |
| --- |
| ***Probability modeled is churn='1'.*** |

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

| **Model Fit Statistics** | | |
| --- | --- | --- |
| **Criterion** | **Intercept Only** | **Intercept and Covariates** |
| **AIC** | 1176.394 | 1110.296 |
| **SC** | 1181.301 | 1120.111 |
| **-2 Log L** | 1174.394 | 1106.296 |

| **Testing Global Null Hypothesis: BETA=0** | | | |
| --- | --- | --- | --- |
| **Test** | **Chi-Square** | **DF** | **Pr > ChiSq** |
| **Likelihood Ratio** | 68.0981 | 1 | <.0001 |
| **Score** | 63.9173 | 1 | <.0001 |
| **Wald** | 60.0945 | 1 | <.0001 |

| **Analysis of Maximum Likelihood Estimates** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Parameter** | **DF** | **Estimate** | **Standard Error** | **Wald Chi-Square** | **Pr > ChiSq** |
| **Intercept** | 1 | 1.0424 | 0.2599 | 16.0912 | <.0001 |
| **age** | 1 | -0.0505 | 0.00652 | 60.0945 | <.0001 |

| **Odds Ratio Estimates** | | | |
| --- | --- | --- | --- |
| **Effect** | **Point Estimate** | **95% Wald Confidence Limits** | |
| **age** | 0.951 | 0.939 | 0.963 |

| **Association of Predicted Probabilities and Observed Responses** | | | |
| --- | --- | --- | --- |
| **Percent Concordant** | 65.3 | **Somers’ D** | 0.328 |
| **Percent Discordant** | 32.5 | **Gamma** | 0.336 |
| **Percent Tied** | 2.2 | **Tau-a** | 0.131 |
| **Pairs** | 198924 | **c** | 0.664 |

The regression equation is as follows:

First, we must perform a hypothesis test on the significance of the predictor Age:

H0: =0; Age is not a significant predictor.

H1: ; Age is a significant predictor.

Looking at the case study of Telco data the, p-value for age is <0.0001, which is less than the significance level of 0.05. Thus we reject the null hypothesis, meaning the age is a significant predictor for churn, and the model is validated. Using the logistic regression equation, we can plug in the values of the maximum likelihood estimates. *B0* is the intercept parameter which 1.0424, while *B*1 is -0.0505. Thus, the estimate of the logistics regression equation looks like the following:

We also need to exponentiate this equation to obtain the odds to churn:

The expected change in odds is . These odds mean that for about a one-unit increase in age, we can expect about 9.14% (1.0424 – 0.951) decrease in the odds of churn. The odds interpret to as a customer gets older, the odds of churn decrease.

**Churn Prediction**

Using this information, we can make a case study prediction for churn. Assuming that we want to estimate the odds and probability that a 32-year old customer will churn, we can apply the following equation:

Then the probability that a 32-year old customer is going to churn is:

The estimated probability that a 32-year customer is going to churn is 36%. We can prove the expected odds change by looking at different ages. The odds of churn for a 35-year old is 0.484, and the probability for churn is about 32.61%. Whereas a 30-year old’s odd of churn is 0.623, with a probability of 38.39%.

**Multiple Linear Regression - Income**

Unlike simple linear regression, multiple linear regression encompasses multiple predictors. The r-squared (R2) value for the Telco data for income given address and gender is 0.0528. Thus 5.28% of the variability of the dependent variable (income) is explained by this regression model. However, since there are predictors, hypothesis tests should be conducted on the significance of the coefficients.

Figure 12 - Regression analysis for income

|  |  |
| --- | --- |
| **Number of Observations Read** | 1000 |
| **Number of Observations Used** | 1000 |

| **Analysis of Variance** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
| **Model** | 2 | 604226 | 302113 | 27.78 | <.0001 |
| **Error** | 997 | 10842769 | 10875 |  |  |
| **Corrected Total** | 999 | 11446995 |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
| **Root MSE** | 104.28516 | **R-Square** | 0.0528 |
| **Dependent Mean** | 77.53500 | **Adj R-Sq** | 0.0509 |
| **Coeff Var** | 134.50076 |  |  |

| **Parameter Estimates** | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
| **Variable** | **Label** | **DF** | **Parameter Estimate** | **Standard Error** | **t Value** | **Pr > |t|** |
| **Intercept** | Intercept | 1 | 45.69816 | 6.04806 | 7.56 | <.0001 |
| **address** | address | 1 | 2.40343 | 0.32712 | 7.35 | <.0001 |
| **gender** | gender | 1 | 7.88168 | 6.59962 | 1.19 | 0.2327 |

Graphical user interface, diagram

Description automatically generated

Chart, scatter chart

Description automatically generated

For the entire model, the significant test is:

* H0: There is no significant correlation between years at address and gender to income.
* H1: at least one of the predictors, years at address or gender, is significant;

The correlation between the independent variables and the dependent variable of income is significant since it has a p-value (<0.0001) which is less than the significance level a = 0.05. Thus, we reject the null hypothesis and proceed to check the significance of each of the predictors.

The hypothesis test for the significance of address is as follows:

* + H0: ; Address is not a significant predictor for the model.
  + H1: ; Address is a significant predictor for the model.

The hypothesis test for the significance of gender is as follows:

* + H0: ; Gender is not a significant predictor for the model.
  + H1: ; Gender is a significant predictor for the model.

The p-value for address is <0.0001, and the p-value for gender is 0.2327. Since the p-value is greater than the significance level we conclude that gender is a not significant factor for churn. Therefore, we must run a linear regression model on address.

**Conclusion**

In conclusion, by analyzing the correlations of various independent variables to dependent variables, we can determine the impacts of customer churn factors. First, using logistic regression, data analysts can evaluate both the odds of churn and the probability that a customer at a certain age is going to churn. Second, with a valid linear regression model, we can predict the income of a customer based years they have lived at their current address. Multiple linear regression and logistic regressions help make actionable predictions that inform business decision-making.

**References**

Gallo, A. (2015, November 4). *A Refresher on Regression Analysis*. Harvard Business Review. <https://hbr.org/2015/11/a-refresher-on-regression-analysis>

Jain, H., Khunteta, A., & Srivastava, S. (2020). Churn Prediction in Telecommunication using Logistic Regression and Logit Boost. *Procedia Computer Science*, *167*, 101–112. <https://doi.org/10.1016/j.procs.2020.03.187>

Sperandei, S. (2014). Understanding logistic regression analysis. *Biochemia Medica*, 12–18. <https://doi.org/10.11613/bm.2014.003>

SAS Institute Inc. (2020). *SAS Studio OnDemand.* (Version 3.8 Enterprise Edition). [Online Software]

‌

‌