Factors Affecting Churn

John Wensink

MIS445 - Statistics in Business Analytics

Colorado State University-Global Campus

Dr. Alin Tomoiaga

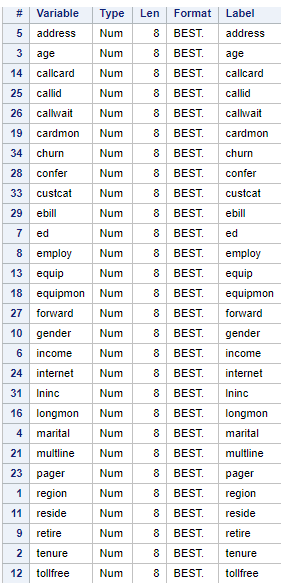
November 11, 2020

Factors Affecting Churn

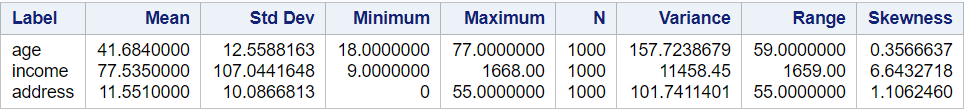
A random sampling of 1000 Telco customers was taken by a third party, and our firm purchased the data set for the purpose of providing insight into the phenomenon of churn. In this sample, approximately 50% of the population of Telco customers are represented with personally identifiable information omitted. Each observation in the data set corresponds to and represents a unique Telco customer. The research was intended to provide insight into the attrition rate of Telco customers over a given period of time. Churn, a statistic typically expressed as a percentage of customers who quit doing business with an organization within a set period of time (Arifin & Samopa, 2018), was examined alongside eight other variables to provide the client insight into its customer attrition. We examined the relationship between the eight of 32 available variables to investigate how they seem to affect churn. These same variables were again examined to investigate the relationship between the test variables and their effect on income. A hypothesis was created and analyzed, based on the apparent relationship between the dependent variables as they relate to the two examined dependent variables, with the results interpreted and predictions made about the odds of a 27-year-old to churn.

**Sample Characteristics**

The variables we are interested in are region, age, marital, address, income, ed, gender, custcat, and churn. Our analysis depended on the type of data examined, and exploration was broken down into three types of approaches. For the exploration of continuous data such as age, years at address, and income, we examined the mean and standard deviation of the datasets. For the exploration of categorical data we differentiated between data with 2 categories (boolean) such as marital status, gender, and churn; as opposed to data with greater than two categories such as region, education, and customer category. For these categorical variables, we performed frequency analyses for each variable, multiple regression analysis on income, and logistic analysis on churn. When examining income and churn, our study excluded each as the independent variable on the other dependent variable’s analysis. Before researching the dataset could begin, a bit of preprocessing was required to set character datatypes to numerical for the variables of region, gender, and marital. We used a technique to retype categorical data types with continuous inputs. With our dataset in numeric form, the data was imported into SAS University Edition (SUE) and the variables were explored.



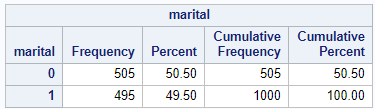
**Summary Statistics for Age, Years at Address, and Income**



As we intended to perform a multivariate linear regression analysis of age and years of continuous residence at one’s address, we examined these variables in terms of their mean (μ) and their standard deviation (σ). This sample’s age has a mean age of μ = 41.7 years old, spread across σ = 12.56. We see that the youngest customer is 18 and the oldest is 77, a range of 59 years. This range proved to have unique implications in the likelihood of churn described later in the study. We see the skewness is in between 0.5 and -0.5 with a skewness of 0.36 this can be interpreted as Telco having slightly more older customers than younger ones.

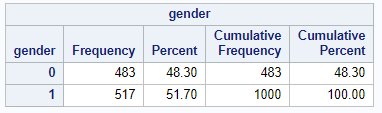
Examining the statistics for years of residence we find a different interpretation. Mean years living continuously at an address was μ = 11.6 years spread across σ = 10.09 years. This is a huge degree of variability. The skewness is highly positive at 1.106, we can interpret this as there are many more long term homeowners as short term homeowners in this sample. The min was 0 years and the max was 55 years, with a range of 55 years.

**Marital Frequency Analysis**



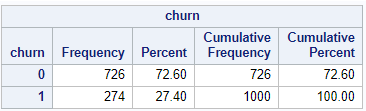
The split between married and single customers was very close with unmarried comprising of 50.50% of the sample and 49.50& of the respondents being married. Our multiple regression analysis described later in the study did not find a statistically significant correlation between marital status and income.

**Gender Frequency Analysis**



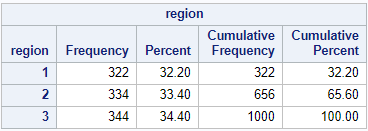
The split between male and female respondents came in as 48.30% being male, and 51.70% being female. Our multivariate regression analysis did not find a statistically significant correlation between gender and income.

**Churn Frequency Analysis**



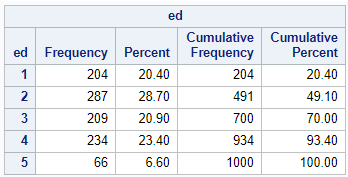
In a frequency analysis of churn across the entire sample, we found that we had 27.4% of customers churn in the period examined. This is a very high attrition rate and more exploration into the makeup of the churning customers is warranted.

**Regional Frequency Analysis**



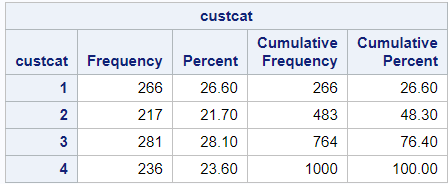
This table shows a nearly perfect three-way split between the three regions in the sample. The difference of 2.2% between the highest and lowest represented regions seems to be insignificant.

**Education Frequency Analysis**



The sample of highest education level is represented by 1 = did not complete high school, 2 = graduated high school, 3 = some college, 4 = college graduate, 5 = postgraduate degree. We can see that high school graduates make up 28% of respondents and are the winning plurality. only 6.6% of respondents hold post-graduate degrees.

**Customer Category Analysis**

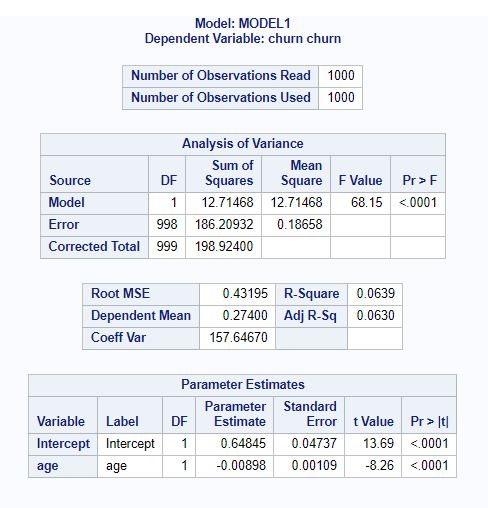
****

The customer category frequency table can be interpreted as having the highest percentage of customers with tier three service. The four tiers represent 1 = basic service, 2 = e-service, 3 = plus service, 4 = total service packages. The packages seem fairly evenly distributed with only a 6.4% range between the distribution of the service tiers.

**Predicting Churn as a Function of Age**

Building the logistic regression model using the dependent variable as churn, and the independent variable as age. We will be examining the relationship between age and churn. We had hoped that the strong correlation between age and income would translate into correlation in logistic regression of age and churn. We were excited to see that again that R2 = 0.0639 Adjusted R2 = 0.0630, representing a significantly positive correlation, 2 and the p-statistic remained constant at <0.0001 with significance level alpha = 0.05:

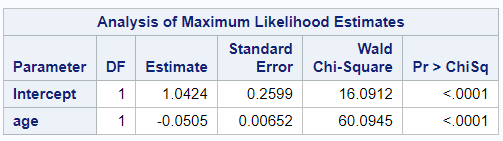
**Logistic Regression of Age and Churn**



With our model validated, we can estimate the probability that a customer will churn based on his or her age. When churn = 1, a customer is guaranteed to churn, we estimated the likelihood of a customer churn based on his or her age. The logistic regression equation is:

log(odds of churn) = β0 + β1 \* age.

**Analysis of Maximum Likelihood Estimates table:**



From this table we performed the following test on the significance of our predictor Age:

H0: β1 = 0 Age is not a significant predictor of churn.

H1: β1 ≠ 0 Age is a significant predictor of churn.

Using the p-value <0.0001, we see this is well below the significance level α = 0.05, we reject the null hypothesis. Age is a significant predictor of a customer’s likelihood to churn. Now that we have validated the model, we can use it to predict the likelihood of churn based on a customer’s age:

log(odds of churn) = b0 + b1 \* age, where b0 and b1 are our estimates of β0 and β1.

We find these statistics in the Analysis of Maximum Likelihood Estimates table as b0 = 1, and b1 = -0.0505. This indicates a strong inverse relationship where as age increases the likelihood to churn decreases:

log(odds of churn) = 1- 0.0505 \* age

When we exponentiate the equation we see that:

odds of churn = e1 - 0.0505 \* age

For a one year increase in age, the expected change in odds is ebi = e -0.0505 = 0.95075. **We can confidently predict this customer will have a 4.925% decrease in odds of churn.**

Making a prediction that we have a 27-year old customer, his odds to churn are

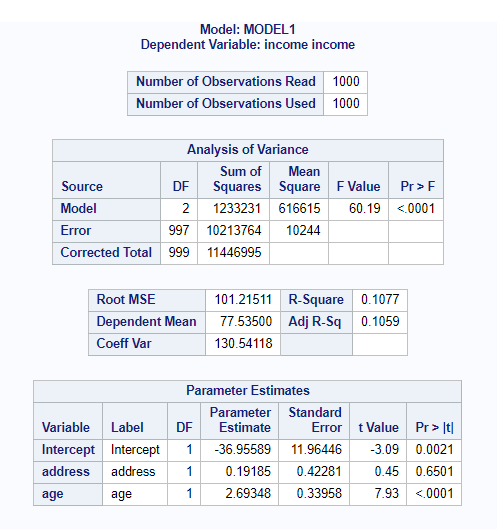
odds of churn = e1-0.0505 \*27 = 0.25576.

The probability that 27-year old churns can be described as

p = odds of churn / 1 + odds of churn = 0.25576/1.25576 = 0.2037 .

So, for a 27- year old, the estimated probability that he will churn is 20.37%. One in five 27-year-olds is likely to churn.

**Multiple Regression of Address and Age**



Regression Model:

income = β0 + β1 \* Address + β2 \* Age

**Hypothesis**

H0: β1 = β2 = 0 There is no correlation between income and neither years of residence at the same address nor customer’s age.

H1: β1 = β2 ≠ 0 At a minimum either years of residence or age is correlated with income.

We will compare the p-value given by the Analysis of Variance table, in column Pr<F which is less than .0001. This is much less than the significance level α = 0.05, and as such we reject the null hypothesis. One of our predictors is significant, perhaps both, we will form a third hypothesis to determine whether years of residence or age is a significant predictor of income. We checked the significance of each predictor:

H0: β1 = 0 Years of residence is not a significant predictor of income according to this model.

H1: β1 ≠ 0 Years of residence is a significant predictor of income according to this model.

and:

H0: β2 = 0 Age is not a significant predictor of income according to this model.

H1: β2 ≠ 0 Age is a significant predictor of income according to this model.

The p-values which correspond to these two tests are shown in the table labled parameter estimates in the column P>|t|. For years at residence p= 0.6501 we fail to reject the null hypothesis. Years at residence is not a significant predictor of income. Examining the p-value for age however p > 0.0001, we fail to reject the null hypothesis. Age is a significant predictor of income based on this sample of the population. As we have failed to validate our model, we find the estimate of the regression equation to be:

income = b0 + b1 \* Years at residence + b2 \* Age,

where b0, b1 b2 are the estimates of β1, β2, and β3 whose values are shown in the given column of Parameter Estimate.

We ran three additional multivariate regressions keeping age and substituting gender, address, and marital status to try to obtain a statistically significant result and failed. While Age consistently falls below significance level alpha, no significance was found for Years at Address, Gender, or Marital status, with Gender being the closest

p-values:

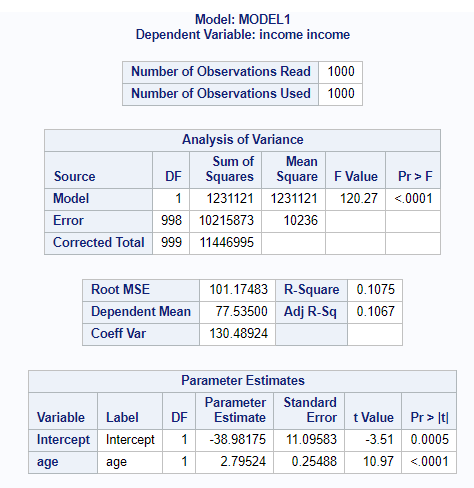
Address .6501

Gender 0.1558

Age <0.0001

Marital 0.1894

As the variable for age seems to be the only variable for which the p-level is < significance level alpha, we modified our multivariate approach to univariate regression analysis of age and its significance as a predictor of income.



This univariate regression seems to more significantly predict income with the single independent variable of age. r2 = 0.1075 the results are too poorly correlated to make a meaningful prediction (Hamilton, Ghert, & Simpson, 2015). However, the p-value <0.0001, and alpha = 0.05. If the r2 statistic of variance had been more strongly correlated, the estimate of the regression equation would have been:

income = b0 + b1 \* age

References

Arifin, S., &amp; Samopa, F. (2018). Analysis of Churn Rate Significantly Factors in Telecommunication Industry Using Support Vector Machines Method. Journal of Physics: Conference Series, 1108, 012018. Retrieved November 1, 2020, from <https://www.researchgate.net/publication/329379548_Analysis_of_Churn_Rate_Significantly_Factors_in_Telecommunication_Industry_Using_Support_Vector_Machines_Method>

Hamilton, D. F., Ghert, M., &amp; Simpson, A. H. (2015). Interpreting regression models in clinical outcome studies. Bone &amp; Joint Research, 4(9), 152-153. Retrieved November 1, 2020, from <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4678365/#:~:text=R2%20is%20a%20measure,perfectly%20fit%20the%20linear%20model>.