Predictive Analytics Using the Telco Extra Dataset

Steven Duchene

MIS 445 – Justin Bateh

Colorado State University – Global Campus

June 5, 2020

Predictive Analytics Using the Telco Extra Dataset

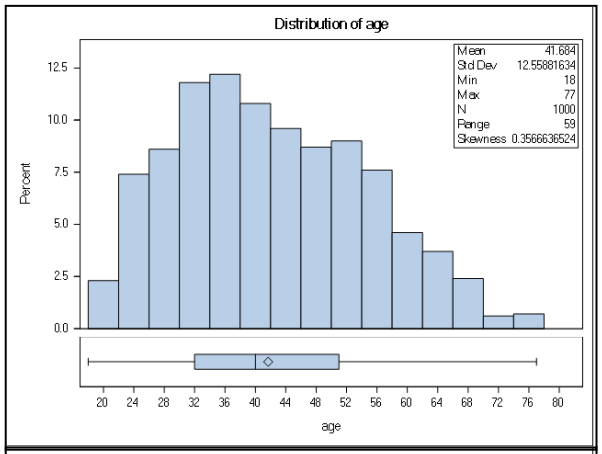
The Telco Extra dataset was developed from the Telco Extra customer database and is comprised of a sample of 1000 customers; roughly half of the organizations customer base. The main purpose for this report is to determine whether or not a selection of independent variables can accurately predict the likelihood of customer churn rate as well as customer income. In order to achieve this, multiple linear regression will be conducted. Using multiple linear regression for prediction purpose, one must be careful to ensure that they utilize an adequate sample size (Knofcyznski and Mundfrom, 2008). As we are using a sample size of 1000 customers, the analysis has the possibility of making accurate predictions. The following sections will detail the summary statistics as well as the statistics gained from running multiple linear regression analyses. At the conclusion of the report, surprises or unexpected findings will be discussed.

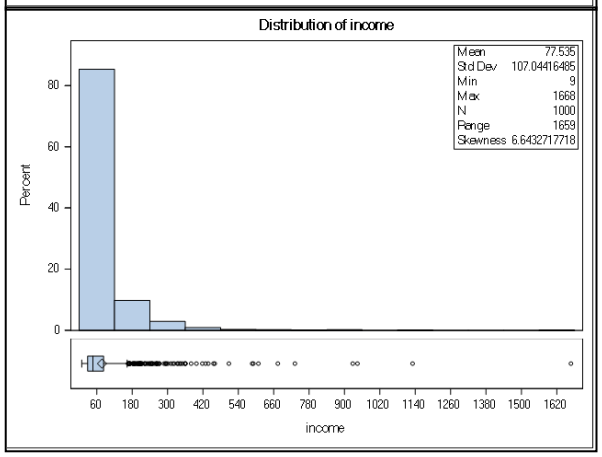
Sample Characteristics and Descriptive Statistics for the Telco Extra Dataset

The Telco dataset is comprised of 34 different variables that include characteristics such as age and gender as well as characteristics specific to each customer's account. The sample being analyzed in this report was selected, at random, from the Telco Extra customer database. It is comprised of 1000 observations. A preliminary review of the dataset indicates that there are no missing values within the dataset; thus, no further action is required in this area. In order to conduct the predictive analytics discussed in the proceeding sections, a selection of variables was selected from the dataset. Theses variables include customer age, gender, marital status, income, level of education, region, category, years at current address, and churn.

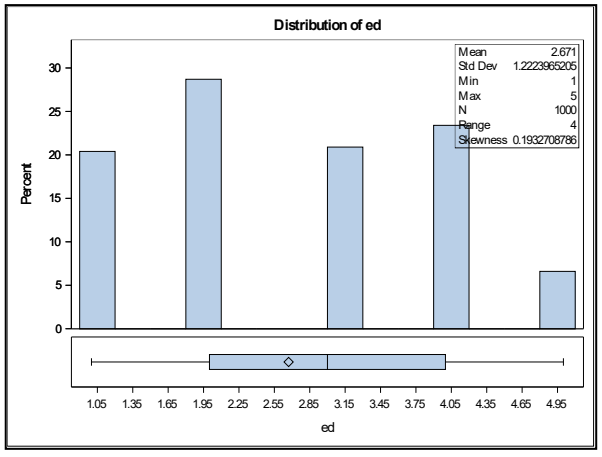
Variable	Label	Mean	Std Dev	Minimum	Maximum	N	Range	Skewness
age	age	41.6840000	12.5588163	18.0000000	77.0000000	1000	59.0000000	0.3566637
income	income	77.5350000	107.0441648	9.0000000	1668.00	1000	1659.00	6.6432718
ed	ed	2.6710000	1.2223965	1.0000000	5.0000000	1000	4.0000000	0.1932709
reside	reside	2.3310000	1.4357926	1.0000000	8.0000000	1000	7.0000000	1.0334458
custcat	custcat	2.4870000	1.1203062	1.0000000	4.0000000	1000	3.0000000	-0.0316353
churn	churn	0.2740000	0.4462321	0	1.0000000	1000	1.0000000	1.0149556

Below, the descriptive statistics, also known as sample statistics, will be presented for each numerical-based variable. In the case of the three categorical variables, pie charts will be presented to indicate their proportions. Table 1, below, presents that sample statistics for the numerical variables age, income, years at address, education level, customer category, and churn. Included in table 1 is a measure of skewness. This measure will assist in the analysis for each variable.



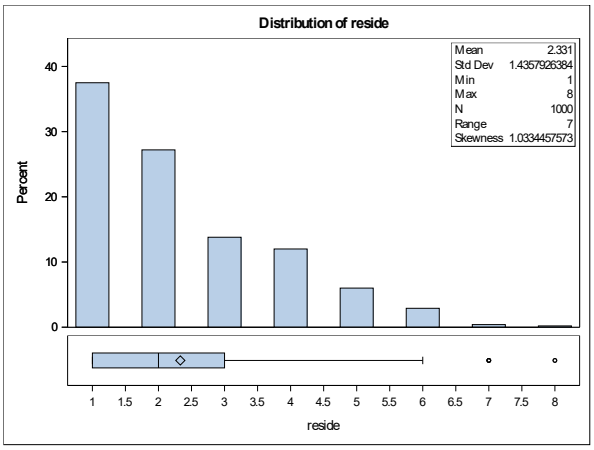


1. Age Variable: As indicated by table 1 above, the age variable has a mean value of 41.68 years with a standard deviation of 12.56 years. Customer age ranges from 18 to 77 years corresponding to a range of 59 years. It is weakly skewed to the right.
2. Income Variable: As indicated by table 1 above, the income variable has a mean value of $77.54 with a standard deviation of $107.04. The range for income is $1659.00. It is strongly skewed to the right.

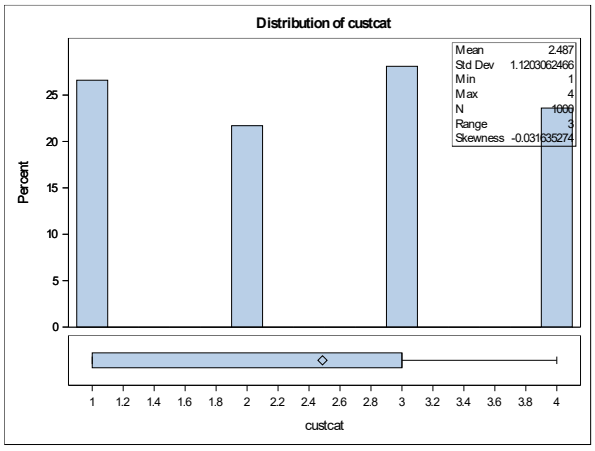


3. Level of Education: As indicated by the table above, the mean value for level of education is 2.67 years with a standard

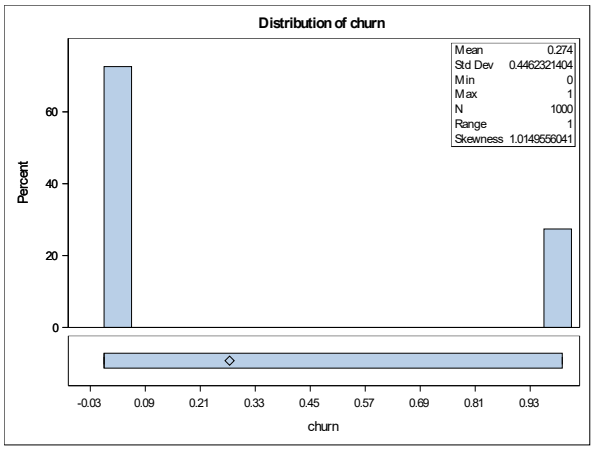
deviation of 1.22 years. The range indicated for years of education is 4 years. It does not appear to exhibit any relative skewness.



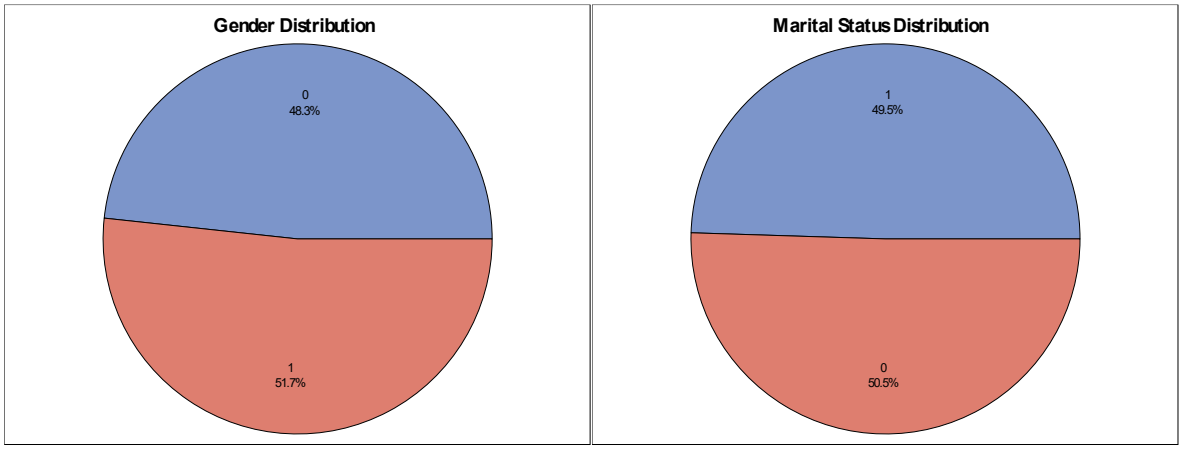
4. Years at Residence: As indicated in the table, the mean value for years at residence is 2.33 years with a standard deviation of 1.44 years. The values range from a minimum value of 1 year to a maximum value of 8 years with a total range of 7 years. The data exhibits a strong skewness to the right.



5. Customer Category: The mean customer category is 2.45 with a standard deviation of 1.12. Due to the nature of this variable, customer category has a minimum value of 1 with a maximum value of 4. Based on a visual analysis of the histogram below, there appears to be relatively equal proportions for each category.
6. Churn: The churn variable has a mean value of .264 and a standard deviation of .446. In the case of churn, minimum and maximum values are not of significance due to the value of 0 indicating no churn and 1 indicating churn.



7. Gender, Region, and Marital Status: The pie charts presented below demonstrate the proportions for gender, region, and marital status. With regards to gender 48.3% of the sample was male and 51.7% was female. With regards to marital status, 49.5% percent of the sample indicated they were married and 50.5% indicated single. With regards to region, the sample comprised of nearly 1/3 of customers from each region: 1, 2, and 3.



Linear Regression Analysis: Dependent Variable – Churn

Linear regression is a linear model, where the model attempts to determine the relationship between two variables by fitting a linear equation (Yale, n.d.). It is a well-known and highly used algorithm in the fields of statistics, data mining, and machine learning. Using a simple, linear equation, linear regression is able to accurately predict the value of one variable

based on the value of another. The linear regression equation typically takes the form of y = B0 + B1x. Using the SAS Studio statistical platform, the dependent variable was set to Churn with the continuous variables set to Age, Education, Category, and Years at Address. The classification variables were set to Gender, Marital Status, and Region. The ANOVA and Parameter Estimates are presented below. Prior to completing the regression analysis, we must develop the hypothesis. In this case, the null hypothesis would be "There is no linear association between the independent variables and churn" with the alternate hypothesis being "There is a linear relationship between the independent variables and churn."

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	8	18.95706	2.36963	13.05	<.0001
Error	991	179.96694	0.18160		
Corrected Total	999	198.92400			

Root MSE	0.42615
Dependent Mean	0.27400
R-Square	0.0953
Adj R-Sq	0.0880
AIC	-694.98211
AICC	-694.75966
SBC	-1652.81231

Parameter Estimates					
Parameter	**DF**	**Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	1	0.486254	0.087830	5.54	<.0001
age	1	-0.008452	0.001152	-7.34	<.0001
ed	1	0.062264	0.011460	5.43	<.0001
reside	1	-0.014367	0.012726	-1.13	0.2592
custcat	1	-0.003062	0.012393	-0.25	0.8049
region 1	1	0.007716	0.033263	0.23	0.8166
region 2	1	0.024810	0.032777	0.76	0.4493
region 3	0	0	.	.	.
marital 0	1	0.011077	0.035323	0.31	0.7539
marital 1	0	0	.	.	.
gender 0	1	-0.003130	0.027010	-0.12	0.9078
gender 1	0	0	.	.	.

In order to determine whether or not the overall model is significant, an examination must be completed of the regression analysis. Similar to simple linear regression, it is the p-value of the F-test that determines significance. With a p-value of less than .0001, the model appears to be statistically significant. However, the R^2 value for the analysis returns as .0953; meaning that 9.53% of the variability in churn is accounted for by the

independent variables. A low R^2 value indicates that the independent variables do not explain much of the variation seen in the churn variable. Further interpretation will be provided in a later section.

Linear Regression Analysis: Dependent Variable – Income

Using the SAS Studio statistical platform, the dependent variable was set to Income with the continuous variables set to Age, Education, Category, and Years at Address. The classification variables were set to Gender, Marital Status, and Region. The ANOVA and Parameter Estimates are presented below. Like the prior section, we must also develop the null and alternate hypothesis. For this analysis, the null hypothesis is "There is no linear relationship between the independent variables and Income" while the alternate hypothesis is "There is a liner relationship between the independent variables and Income.

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	8	1686608	210826	21.41	<.0001
Error	991	9760387	9849.02827		
Corrected Total	999	11446995			

Root MSE	99.24227
Dependent Mean	77.53500
R-Square	0.1473
Adj R-Sq	0.1405
AIC	10206
AICC	10206
SBC	9248.25713

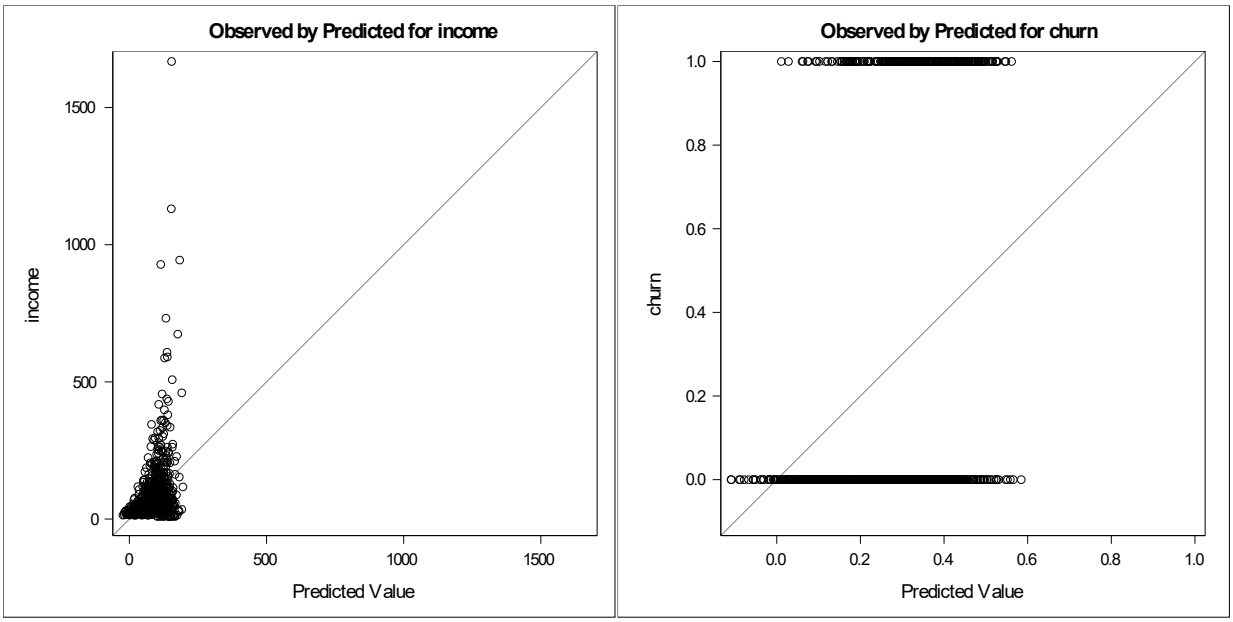
Parameter Estimates					
Parameter	**DF**	**Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	1	-106.412735	20.454112	-5.20	<.0001
age	1	2.982964	0.268257	11.12	<.0001
ed	1	13.388016	2.668841	5.02	<.0001
reside	1	1.217760	2.963671	0.41	0.6812
custcat	1	8.503836	2.886038	2.95	0.0033
region 1	1	-6.404312	7.746466	-0.83	0.4086
region 2	1	0.900497	7.633285	0.12	0.9061
region 3	0	0	.	.	.
marital 0	1	11.644590	8.226100	1.42	0.1572
marital 1	0	0	.	.	.
gender 0	1	-8.820440	6.290185	-1.40	0.1612
gender 1	0	0	.	.	.

Like the above analysis, the p-value of the F-test was less than .0001; indicating that the model was statistically significant. However, the R^2 value returned as .1405, indicating that 14.05% of the variability seen in the Income variable can be explained by the independent variables. A low R^2 value indicates that the independent variables do not explain much of

the variation seen in the churn variable. Further interpretation will be provided in a later section.

Interpretation of the Outcome for the Churn and Income Dependent Variables

As discussed above, while the p-values indicated that each statistical model was indeed significant, the R^2 values, .0953 and .1405 respectively, indicate that a relatively low percentage of variation in churn and income can be explained by the independent variables. In the case of both Churn and Income, we have failed to reject the null hypotheses that "There is no linear relationship between the independent and dependent variables." With respect to the strength and direction of the relationship between the dependent and independent variables, there appears to be a weak positive relationship between the independent and dependent variables. This can be observed in the intercept values of both the parameter estimate charts above. The same is also observed in the charts detailed below. If the analyst were to calculate the correlation coefficients for each variable against Income and Churn, it is likely they would reach the same conclusions.



Summary of Expected Findings

At the beginning of the analysis, it was expected that the independent variables: age, years at address, marital status, level of education, region, category, and gender would have been

a good model for prediction of churn and income; the dependent variables. It was surprising to find that there was a very weak linear relationship between the independent and dependent variables. That being said, when looking at the rest of the variables in the dataset, it becomes obvious that there were far better options to chose from to predict churn. Take for example the following variables: tenure, wireless, multiline, voice, internet, and ebill. Using these variables to conduct the linear regression provides a much greater R^2 value compared to the analyses conducted in the previous sections. It is likely that, if the analyst were to use variables that corresponded more with services received by Telco, that a higher linear relationship would be observed.

While the statistical results failed to reject the null hypothesis, it does not necessarily indicate that there is a lack of correlation between the independent and dependent variables. When taken individually, the analyst can determine the likelihood of churn, based on the variable This is especially true of factors such as employment status, retired status, and income status due to their likelihood of determining the customer's chances of staying or leaving. In order to gain a greater understanding of the variables affecting churn rate, further study should be conducted with respect to the other variables in the dataset.

References

Knofczynski, G. T., & Mundfrom, D. (2008). Sample sizes when using multiple linear regression for prediction. *Educational and psychological measurement*, *68*(3), 431-442.

Yale University. (n.d.). Linear Regression. Retrieved from http://www.stat.yale.edu/Courses/1997-98/101/linreg.htm