|  |
| --- |
| **Western Governors University** |
| M.S. Data Analytics |

|  |
| --- |
| C744 | By: Jeevan G Joseph |



|  |
| --- |
| Analyzing Customer Churn In The Telecom Industry |
| MSDA C744 Performance Assessment Scenario The scenario presented puts us in the position of a data analyst for a telecommunications company that is concerned about the number of customers leaving their landline business for cable competitors. The company needs to know which customers are leaving and attempt to mitigate continued customer loss. Understanding customer behavior like affinity or attrition directly impacts a company’s profitability and growth prospects. Therefore, analyzing data using descriptive methods to understand patterns and causality as well as predictive methods to understand and evaluate corrective actions becomes a central function for the business.    Tool Selection I have chosen to use the R data mining software to analyze the data for this exercise. The choice to use R was driven by the following factors.  Easy Availability & Light Footprint  R and R studio are easily available and can be installed on a wide variety of operating systems. The easy of setup and light footprint meant that I could run this datamining software without specialized hardware designed for larger workloads.  Plugin Ecosystem & Graphics Support  R also has a wide range of plugins available, some of which were used especially in the data exploration phase. The plugin ecosystem is easy to use as well, with the R studio having the ability to search for, and install plugins directly from within the editing environment. Moreover, the R Studio environment supports very rich graphics capabilities to create visualizations, making R a very versatile platform for performing data analysis on small to medium data volumes.  Data Volume  The data volume is a characteristic that can dictate the choice of the tool. Most commercial tools can handle large volumes of data. The data set being analyzed here is relatively small with under 10,000 observations. This data set can easily be analyzed with R, whose primary drawback with large data sets come from the fact that R can only operate on in-memory data. Our data set does not really qualify as a large one, and can be easily analyzed using R on commodity consumer grade computers.  Ease of Data Extraction  R makes it very simple to work with data sets of limited sizes, especially when they are present locally. Large datasets often need to be stored remotely in larger databases or storage solutions, and operated up on remotely. The dataset here is presented as a CSV file, which is less than a megabyte in size. This makes it ideal for R to store the file locally and load the entire dataset in to memory. R’s read.csv() function offers a simple way to load the data, and convert it in to a Data Frame structure that makes it easy to work with in R. The command to load the data file from the working directory is shows below.  phone\_data<- read.csv(file="WA\_Fn-UseC\_-Telco-Customer-Churn.csv") Goals The primary goal of this exercise is to use descriptive and predictive statistical tools to understand customer churn and to take corrective actions. The business is concerned about the customer churn in the landline business, where current customers are moving to cable competitors. To achieve this,   * We will focus on the data set that pertains to landline customers * Use descriptive and exploration techniques like association rules and mosaic plots to discover patterns in the dataset * We will then use predictive methods such as classification by logistic regression to create model that can classify the customer base in to populations with various attrition risks.   The data set contains the several data points or independent variables that we can potentially use to explain the effect on the churn, as well as build a model for classification. In fact, there are 21 variables, including the dependent variable in the data set.  Most of the variables in the dataset are qualitative. With this sort of a dataset, we can use descriptive methods such as mosaic plots or association analysis to determine the factors contributing to churn. Association analysis and mosaic plots are helpful for analyzing the data and discovering patterns. Since the dataset primarily consists of qualitative variables, mosaic plots are helpful in visually identifying the most discriminating variables. They can display several independent variables at once and give a good visual indication to patterns in the data. Association analysis is helpful in this case, because we are trying to find the qualities that occur most frequently with the individuals that have left the carrier’s service.  Descriptive analysis however, only gives us insights into the current dataset and help us identify patterns.  To predict the attrition risk, we need to apply a predictive method to classify the population with high attrition risk. Logistic regression seems like a good choice here to predict churn, since the target variable we are predicting, churn, is a categorical variable and not a continuous one, which would have yielded to scoring. However, we will need to ensure that there are no links between the independent variables, since logistic regression would be sensitive to it. The descriptive analysis of the data should yield evidence to support or reject the choice of logistic regression.  Data Exploration and Preparation  In the data preparation stage, we examine the dataset and make sure the data set is fit for analysis and also identify redundant data and missing values and similar errors. Describe the target variable in the data and indicate the specific type of data the target variable is using, including examples that support your claims. The target variable in the dataset is “Churn”. This variable is “Yes”, when the customer has already left the carrier and “No” when the customer is still a subscriber. This type of data in the telephony industry is categorized as “Line Data”, as it pertains to the status and other characteristics of the line the customer has subscribed to. The variable is categorical nominal variable, and specifically a dichotomous variable since the levels for variable are only one of two possibilities – Yes or No.  In R, we can see the possible levels for each variable in a data-frame using the levels() function.  > levels(phone\_data$Churn)  [1] "No" "Yes" *Describe an independent predictor variable in the data and indicate the specific type of data being described. Use examples from the data set that support your claims.* An example independent variable in the dataset is, tenure. Tenure is expressed in months and indicates the length of time(in months) this customer has been a subscriber to the carrier’s service. This type of data is classified as “Customer Data” in the telephony industry. This is a continuous variable. We can quiclkly summarize this variable’s characteristics using the summary() function in R.  > summary(data$tenure)  Min. 1st Qu. Median Mean 3rd Qu. Max.  0.00 9.00 29.00 32.37 55.00 72.00  This shows the distribution of the values for this variable, which is useful to identify the rage of values present in the dataset and its rough distribution.  The dataset also includes other data types common to the telephony industry as well, some of them are :   |  |  |  | | --- | --- | --- | | Customer Data | Line Data | Billing Data | | Gender | PhoneService | PaperlessBilling | | SeniorCitizen | MultipleLines | PaymentMethod | | Partner | InternetService | MonthlyCharges | | Dependents | OnlineSecurity | TotalCharges | | CustomerID | OnlineBackup |  | | Tenure | DeviceProtection |  | |  | TechSupport |  | |  | StreamingTV |  | |  | StreamingMovies |  | |  | Contract |  |  *Propose the goal in manipulation of the data and define your data preparation aims.* The main goal during the data preparation phase is to ensure that the data is fit for evaluation. This includes imputing any missing values, ensuring that categorical variables all have values that are uniform. To apply categorical analysis, we need to convert some of the continuous variables to ranges or other categories that represent these ranges. The include variables like tenure, and the billing amounts. With a quick analysis of the data, we can also see that some of the categorical variables have values that need to be adjusted to be uniform, like changing some of the line data like Streaming TV, to replace values like “No Internet Service” to “No”, since for our analysis, both are equivalent. Define the statistical identity of the data, including the essential criteria and phenomenon to be predicted. We can quickly summarize the data loaded in to a data frame in R by using the summary.data.frame() function. This gives us an overview of the whole data frame to help us identify areas where we need to apply transformations or impute erroneous values.  From the output we can immediately glean the distribution of the data and summary statistics for each of the variables. Our goal is to use descriptive methods to understand the factors contributing to the Churn, and we have finite categories for most of the variables. We will discretize the continuous variables in the dataset, based on their distributions. The customerID is unique identifier, that can potentially become useful for association analysis, we we consider this similar to transaction data.   Explain the steps used to clean the data and how you addressed any anomalies or missing data. For each of the variables, the following techniques were used to clean or address anomalies, or to discretize a continuous variable in to a categorical one.  Values / Levels meaning “Not Applicable”  From the summary of the data frame, we can plainly see that there are some redundant values for some of the variables. These are duplicate negative levels are provided in some of the variables, especially those that are not applicable given a previous value. For example, the DeviceProtection variable has 3 levels – “Yes”, “No” and “No Internet Service”. The last level is redundant, and for our data analysis goals, equivalent to “No”.  We can convert these levels in to “No”, by selectively updating the data frame. We can do this using the levels() function to combine existing levels. The screenshot below shows the application of the function and the observations in each level before and after combining the levels.    We continue to do this for the other variables that exhibit a similar pattern in redundant levels. These variables are:   * Online Security * Online Backup * Device Protection * Tech Support * Streaming TV * Streaming Movies   Once completes, we can use the summary.data.frame() function again to see the updated levels for all dimensions in the data frame. The screenshot below shows the updated dataframe.    Discretizing Values (Binning)  From the data, we can see that we have three independent variables that are continuous – Tenure, Monthly Charges and Total Charges. We shall discretize (bin) these in to ranges so that we can work with our descriptive analysis methods.  In order to discretize the values, we need to take the data distribution in to account. We take the simple method of dividing the data in to quintiles and using these as our bins. In R this is simple, using the quantile() function.    We have five bins, for various tenures:   * 0-6 months * 7-20 months * 21-40 months * 41-60 months * 61+ months   In R we can discretize a continuous variable using the cut() function. The screenshot below shows the function to discretize.    We can see that there are 11 observations where the tenure is 0 months. These indicate accounts that are less than a month old, and since the default behavior of the function is to exclude the first break (closed on the left), we include the include.lowest=TRUE parameter to include the observations with tenure=0. We take the discretized variable and add it to our data frame as “Tenure”. The summary shows the distribution of the number of observations in each bin for our new independent variable. We also drop the continuous variable “tenure”  Missing Values  From the summary of the data frame, we can see that there are some missing values in the “Total Charges” variable. There are only 11 observations that have missing values among 7043 total observations, or about 0.156% of the total dataset. Incidentally these 11 observations also have a tenure of 0 months, but none of them have been marked as Churn=Yes. Therefore, these can potentially be valid customer accounts that are less than a month old. Although the Total Charges are missing values and the number of observations is an insignificant number, we do not yet know how discriminating the “Total Charges” variable is to the target variable “Churn”. If there is no significant influence, we can safely drop these observations, or even the variable itself.  We will therefore wait until we can estimate how discriminating this independent variable is to the target variable before we take any action. This is a potential refinement we will need to revisit in our iterative process, after we finish our descriptive analysis of the data.  Data Analysis For each of the following steps, be sure to clearly indicate each step within your data sheet with a screen shot and annotations in your final submission. All algorithms used need to be clearly identified in the screen shot and submission.   1. Identify the distribution of variables using univariate statistics from your cleaned and prepared data. Represent your findings visually as part of your submission. 2. Identify the distribution of variables using bivariate statistics from your cleaned and prepared data. Represent your findings visually as part of your submission. 3. Apply an analytic method and an evaluative method. Annotate the data showing both methods and your findings. 4. Justify the methods you have chosen to analyze your data. Be sure to include details about how the methods you have chosen better represents your findings than other methods. 5. Justify the methods you have chosen to visually present your data. Be sure to include details about how the presentation methods you chose better represents your findings than other presentation methods.   Data Summary Summarize the findings of your data evaluation. Provide the final findings dataset, including evaluation measures.   1. Explain how your data shows that it was discriminating or not and whether the phenomenon you wanted to detect was present in your findings. Provide specific examples from the data to support your claims. 2. Describe the methods you used for detecting interactions and for selecting the most important predictor variables. Include the specific interactions you detected and the most important predictor variables that you found. 3. Acknowledge sources, using in-text citations and references, for content that is quoted, paraphrased, or summarized.  Heading 2 You can easily change the formatting of selected text in the document text by choosing a look for the selected text from the Quick Styles gallery on the Home tab. You can also format text directly by using the other controls on the Home tab. Most controls offer a choice of using the look from the current theme or using a format that you specify directly.  To change the overall look of your document, choose new Theme elements on the Page Layout tab. To change the looks available in the Quick Style gallery, use the Change Current Quick Style Set command. Both the Themes gallery and the Quick Styles gallery provide reset commands so that you can always restore the look of your document to the original contained in your current template. Heading 3 On the Insert tab, the galleries include items that are designed to coordinate with the overall look of your document. You can use these galleries to insert tables, headers, footers, lists, cover pages, and other document building blocks. When you create pictures, charts, or diagrams, they also coordinate with your current document look.  You can easily change the formatting of selected text in the document text by choosing a look for the selected text from the Quick Styles gallery on the Home tab. You can also format text directly by using the other controls on the Home tab. Most controls offer a choice of using the look from the current theme or using a format that you specify directly.  To change the overall look of your document, choose new Theme elements on the Page Layout tab. To change the looks available in the Quick Style gallery, use the Change Current Quick Style Set command. Both the Themes gallery and the Quick Styles gallery provide reset commands so that you can always restore the look of your document to the original contained in your current template. |
|  |