|  |
| --- |
| **Analyzing Customer Churn in the Telecom Industry** |
| Submitted By: Jeevan Joseph |

|  |
| --- |
| Data Mining & Analytics II - C744 Performance Assessment  Western Governor’s University M.S. Data Analytics |



|  |
| --- |
| Table of Contents  [Scenario 2](#_Toc526249190)  [Tool Selection 2](#_Toc526249191)  [Goals 3](#_Toc526249192)  [Data Exploration and Preparation 3](#_Toc526249193)  [Data Exploration 3](#_Toc526249194)  [Data Preparation 5](#_Toc526249195)  [Irrelevant Data 5](#_Toc526249196)  [Values / Levels meaning “Not Applicable” 5](#_Toc526249197)  [Missing Values 6](#_Toc526249198)  [Discretizing Values (Binning) 7](#_Toc526249199)  [Monthly and Total Charges. 9](#_Toc526249200)  [Senior Citizens 9](#_Toc526249201)  [Data Analysis 9](#_Toc526249202)  [Univariate Statistics 9](#_Toc526249203)  [Bivariate Statistics 14](#_Toc526249204)  [Mosaic Plots & Multivariate Statistics 22](#_Toc526249205)  [Predictive methods: Logistic regression vs. Decision Trees 25](#_Toc526249206)  [Logistic Regression 25](#_Toc526249207)  [Decision Tree Model 28](#_Toc526249208)  [Comparing the Logistic Regression Model with RPART model 29](#_Toc526249209)  [Predicting the churn and cross validating the results 30](#_Toc526249210)  [Data Summary 31](#_Toc526249211)  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. A model based on Logistic regression or Decision trees seem 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. To decide what model works best, we can build both models and compare the ROC curves to identify which one performs better and make a choice. 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. Data Exploration We start by examining the data, the various variables that are present, its encoding and so on. We then start preparing the data for processing. 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.   Data Preparation Now that we have examined the raw dataset as it was presented to us, we can start preparing the dataset for processing. This step includes identifying irrelevant data, encoding variables, identifying and missing values and possibly imputing he values and so on. Irrelevant Data The goal for this analysis is to identify and make recommendations to mitigate the customer churn in the company’s landline business. In keeping with this goal our first step is to eliminate observations that do not pertain to the landline business. We have the “Phone Service” variable that indicates if the observation is for a landline customer or not. Counting the observations for customers who are not subscribers to the phone service, we see that we have 682 observations that do not pertain to the population we are interested in (we can easily see this from the summary for the loaded data). We can also check if these affected the churn, and we see that they did not, in other words we do not have observations where a non phone customer left the company. We can remove these observations to narrow the dataset to only the population that are phone service subscribers. The screenshot below shows how we counted the customers in the dataset that were not phone service subscribers and how they were removed from the dataset.   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 data frame.   Missing Values From the summary of the data frame, we can see that there are some missing values (NA) in the “Total Charges” variable. We can quickly filter the NAs in R using the complete.cases() function. There are only 9 observations that have missing values among 6361 total observations, or about 0.141% of the total dataset. Incidentally these 9 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.    The “Total Charges” field represents the amount of money the customer has spent with the carrier throughout his tenure. We can confirm this by examining a few observations where the Total Charges are available.    We can also calculate the error ratios and compute the percentiles for it. When plotted, this looks like the following    The 95th percentile for the error ratio is .102, it effectively tells us that that the there is a 5% chance to have an error exceeding 10.2%, if we impute the value of the total charges. Since we are imputing the value of total charges for accounts less than a month old, even a 10% error rate will not greatly affect the total univariate statistics for the Total Charges variable. The imputation and and the univariate statistics for the Total charges before and after imputation are shown beow in the screenshot.   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 calculate the bin sizes 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 * 41-61 months * 62+ months   Although this seems reasonable at first glance, given the contract based business model for the telecom industry, we can assume that the tenure is related to the contract type, and this may be a better way to discretize and segment the population[[1]](#footnote-1). There are month-to-month, 1-year and 2-years contracts available. We use the univariate statistics to understand the distribution of these observations based on contract type to get better bin sizes for the tenure to find its discriminating power to describe and predict churn. The univariate statistics used are explained in a separate section below. We can arrive at the following bin sizes.   * < 3 months * 3 - 6 months * 6 - 11 months * 1 – 1.5 years * 1.5 - 2 years * 2 – 5 years * 5+ years   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 9 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” from the dataset, since we just discretized it. Monthly and Total Charges. Similar to Tenure, we need to discretize the monthly and total charges as well. For the monthly and Total charges, we can simply use the quantiles to split the data in to appropriately sized bins. The screenshot below shows both variables being discretized.   Senior Citizens The variable Senior Citizen appears to be a factor with possible values of 0 and 1, where 0 represents the fact that the customer is not a senior citizen and a 1 represents that he/she is. However, R sees this as numeric data and we need to convert the data type to be a factor to be treated as such in analyses. To do this, we can replace all (0)s with (No)s and all (1)s with (Yes)s. Then we can convert the independent variable with the factor() function. The screenshot below shows the functions being applied, followed by a summary to show the levels in the factor.   Data Analysis In the data analysis step, we take a closer look at the data and make our inferences. We start by looking at each of the independent variables, and their distributions using univariate statistics, and then use bivariate and multi-variate statistics to get a more complete picture of the data and the interactions between variables. We use visual tools like scatter plots and mosaic plots to better understand the data. Then we build and compare our prediction models and try to predict the target variable using our trained model. Univariate Statistics The first step in any investigation of data is to examine the univariate statistics of the variables[[2]](#footnote-2). We do this in order to:   * detect any anomalies in their distribution (especially outliers or missing values) * get an idea of some orders of magnitude (such as the average ages and income of the  population) which will be useful in the subsequent analysis * see how to discretize the continuous variables if this is necessary.  **Identify the distribution of variables using univariate statistics from your cleaned and prepared data. Represent your findings visually as part of your submission.** We have our date discretized and cleaned up now. We also used and relied upon some univariate statistics earlier to discretize some of the continuous variables. We can now examine each variable to see its distribution and identify any patters, outliers or any apparent anomalies. The most basic way to see univariate statistics in R is to use the summary function.    This shows us the univariate statistics for all the variable sin our data frame. A glance at this can indicate if we have the right categories or if we need to adjust the levels of the factors, since this data set contains a lot of categorical variables.  Next, we visualize each of the variables’ distribution by plotting them. Since this is repetitive work, we can create a function in R to automate the repetitive tasks. The function definition is shown below.    The basic frequency plots for the factor variables are below.  Analyzing the categorical variables can be done easily by plotting their frequencies.    The Contract type and Senior citizen factors variables show that there are more customers on month-to-month contracts than 1-year and 2-year contracts combined. The majority of the customers are not senior citizens either.    We observe that roughly equal number of customer have (or is) a partner account. We can also see that the ratio of customers with dependents to the ones that don’t is roughly 2:1. We can also make a note here to combine the dependents with churn to see if there is any co-relation between the customers without dependents to churn.    We can also bserve from the plots for multipe lines that there are roughly the same number of people who subscribe to more than one lines as the one who do not. As for customers’ preference for internet, only about a quarter of the customers do not use Internet services, and the fiber optioc servive is more popular than DSL.    From the bar plots for the Paperless billing and payment method, we see that the customers who prefer paperless billing marginally outpace the ones that receive a paper bill. Among payment methods, Electronic checks are the clear winner, and all the other payment methods have relatively the same popularity.    The tenure distribution is really interesting in that we can easily see that this carrier has a lot of tenured users, who are with the company for more than 25 months, which is ineffect more than the longest term contract offered. This sems to indicate that once a customer gets used to the service, they seem to appreciate it. While there could be many factors to this, some of which not represented in the dataset we have, we can identify two inferences here. If the tenured customer base contributes to the churn more, then we are losing very valable customers. On the otherhand, if the churn is more for recent customers, then we can infer that if we can keep the customer for a longer term there is a good chance that they will stay with the company. We can identify these patterns in detail using bivariate statistics to plot the realtionship between tenure and churn.  Next we look at some of the continuous variables we had, to see their distribution and to find if they axhibit any significant patterns. Scatter Plots for the continuous variables For the continuous variables that we have in the dataset, we can use scatter plots to visually inspect patterns in the value distribution.    Looking at the Tenure, we see that there is more or less an even distribution of the data, the scatterplot shows a random distribution of values within the bounds. Monthly charges on the other hand presents an interesting pattern. We can clearly see that there is a pattern here, with values concentrated around the $15 – $30 mark. We can subset this dataset to analyze this closely and we see that the distribution is concentrated around $21 mark, with a median of $20.15 and a mean of $21.08. There are 1526 customers in this subset, which tells us that this is likely a popular plan/contract.      The scatter plot for the total charges above shows that we have most customers concentrated at around the bottom of the plot. This can potentially be interpreted in many ways, one being that we are losing customers at around the $2000 mark, which could be their contracts expiring or some other reason. The top portion of the scatterplot represents the most valuable/loyal customers, who have spent the most on the services across a Bivariate Statistics Using bivariate statistics, we can start to determine or visualize the empirical relation between variables. Since we are using categorical variables in our dataset, we can use contingency table analysis using the Chi-Square statistic. As our goal is to understand the dependence between the categorical variables, and specifically their influence on the target variable “Churn”, we can use the Chi-Square test of independence to evaluate the contingency tables we build. We can also visualize out analysis using stacked bar plots and then follow on with some inferential analysis using some mosaic plots to visualize how much or how little the variables affect the Churn. Building contingency matrices and Performing the Chi-Square test of independence Chi Squared test seems appropriate here and more applicable than the Fischer’s test since we have a large sample size. To build the contingency matrix in R, we just need to use the table() function with the two vectors we need in the matrix. The vectors we need are already available in our data frame. The screen shots below show the Chi-Squared test and the corresponding plots for the contingency tables. Please note that for 2x2 matrices the Yates continuity correction is applied in all cases. We will also assume a significance level of 0.01. A low value for the Chi Squared statistic indicates less independence between the categorical variables. In other words, a low value for the Chi-Squared value means that the categorical variables in the contingency matrix are highly correlated. However, we need to put the chi-squared statistic in perspective using the p-value to measure if the difference between the expected values and actual observations is statistically significant.  Gender x Churn  Here we see the contingency matrix for Gender and Churn. These seem to be fairly independent. Although we have a low value for the statistic, we can see that the p-value for the chi-squared statistic is well above the significance level. So this means that the value is not statistically significant.    Senior Citizen x Churn  Here we see the contingency matrix for Senior Citizen and Churn. These seem to be more dependent as we have a large Chi-Squared statistic value with a p-value for the chi-squared statistic below the threshold value. In practice, however all this says is that these two variables are not completely independent. As we shall see below, there are far more discriminating values with much larger chi-squared statistic values below the threshold, which are far more discriminating.    Partner x Churn  Here we see the contingency matrix for Partner and Churn. As before, we see that the two variables are not statistically independent. The moderate value for the statistic combined with the low P-value indicates that we can reject the null hypothesis of the variables being independent.    Dependents x Churn  Here we see the contingency matrix for Dependents and Churn. As before, we see that the two variables are not statistically independent. The moderate value for the statistic combined with the low P-value indicates that we can reject the null hypothesis of the variables being independent.    Tenure x Churn  Here we see the contingency matrix for Tenure and Churn. As before, we see that the two variables are not statistically independent. We have a high value for the chi-squared statistic and combined with the low P-value, indicates that we can reject the null hypothesis of the variables being independent. The Yates continuity correction is not applied here since this is not a 2 x 2 contingency matrix.    Multiple Lines x Churn  Here we see the contingency matrix for Multiple Lines and Churn. We have a fairly low chi-squared value here indicating that the difference between the observations and the expectations were smaller, but the p-value is still lower than the significance level, therefore we conclude that the variables are not independent.    Internet Service x Churn  Here we see the contingency matrix for Internet Service and Churn. As before, we see that the two variables are not statistically independent. The high value for the statistic combined with the low P-value indicates that we can reject the null hypothesis of the variables being independent. The chi-squared statistic is much larger than other comparisons like Dependents, so we could have a more discriminating variable here.    Online Security x Churn  Here we see the contingency matrix for Online Security and Churn. As before, we see that the two variables are not statistically independent. The moderate value for the statistic combined with the low P-value indicates that we can reject the null hypothesis of the variables being independent. The chi-squared statistic is much lower than some of the other comparisons we have done so far, so this variable may not be as discriminating as some of the others.    Online Backup x Churn  Here we see the contingency matrix for Online Backup and Churn. As before, we see that the two variables are not statistically independent. The low value for the statistic combined with the low P-value indicates that we can reject the null hypothesis of the variables being independent. The chi-squared statistic is much lower than some of the other comparisons we have done so far, so this variable may not be as discriminating as some of the others.  Device Protection x Churn  Here we see the contingency matrix for Device Protection and Churn. As before, we see that the two variables are not statistically independent. The low value for the statistic combined with the low P-value indicates that we can reject the null hypothesis of the variables being independent. The chi-squared statistic is much lower than some of the other comparisons we have done so far, so this variable may not be as discriminating as some of the others.    Tech Support x Churn  Here we see the contingency matrix for Tech Support and Churn. As before, we see that the two variables are not statistically independent. The moderate value for the statistic combined with the low P-value indicates that we can reject the null hypothesis of the variables being independent. The chi-squared statistic is closer to some of the other comparisons we have done so far, so this variable may be a good discriminating variable along with others.    Streaming TV x Churn  Here we see the contingency matrix for Streaming TV and Churn. As before, we see that the two variables are not statistically independent. The low value for the statistic combined with the low P-value indicates that we can reject the null hypothesis of the variables being independent. The chi-squared statistic is much lower than some of the other comparisons we have done so far, so this variable may not be as discriminating as some of the others.    Streaming Movies x Churn  Here we see the contingency matrix for Streaming Movies and Churn. As before, we see that the two variables are not statistically independent. The low value for the statistic combined with the low P-value indicates that we can reject the null hypothesis of the variables being independent. The chi-squared statistic is much lower than some of the other comparisons we have done so far, so this variable may not be as discriminating as some of the others.    Contract x Churn  Here we see the contingency matrix for Contract and Churn. As before, we see that the two variables are not statistically independent. The high value for the statistic combined with the low P-value indicates that we can reject the null hypothesis of the variables being independent. The chi-squared statistic is much higher than some of the other comparisons we have done so far, so this variable is one of the most discriminating that we have analyzed.    Paperless Billing x Churn  Here we see the contingency matrix for Paperless Billing and Churn. As before, we see that the two variables are not statistically independent. The moderate value for the statistic combined with the low P-value indicates that we can reject the null hypothesis of the variables being independent. The chi-squared statistic is higher than some of the other comparisons we have done so far, so this variable may be a good discriminating variable.    Payment Method x Churn  Here we see the contingency matrix for Payment Method and Churn. As before, we see that the two variables are not statistically independent. The high value for the statistic combined with the low P-value indicates that we can reject the null hypothesis of the variables being independent. The chi-squared statistic is much higher than some of the other comparisons we have done so far, so this variable may be a good discriminating variable.    Monthly Spend x Churn  Here we see the contingency matrix for Monthly Spend and Churn. As before, we see that the two variables are not statistically independent. The high value for the statistic combined with the low P-value indicates that we can reject the null hypothesis of the variables being independent. The chi-squared statistic is much higher than some of the other comparisons we have done so far, so this variable may be a good discriminating variable.  Total Spend x Churn  Here we see the contingency matrix for Gender and Churn. As before, we see that the two variables are not statistically independent. The moderate value for the statistic combined with the low P-value indicates that we can reject the null hypothesis of the variables being independent. The chi-squared statistic is much higher than some of the other comparisons we have done so far, so this variable may be a good discriminating variable.   Mosaic Plots & Multivariate Statistics A mosaic plot is a good way to visualize a contingency table, by graphically representing the cell frequencies using boxes whore area are proportional to the cell frequencies. Therefore, in a contingency table where a cell frequency is high will be represented by a box of larger area in the mosaic plot. It is a good tool to quickly glean relationships and their relative strength in a dataset.  We can start off with a simple mosaic plot that simply represents a contingency matrix for Churn and Contract. It clearly and quickly displays the relative impact of Contract on Churn.    From the plot, it’s easy to see that among the customers who moved away from the business (resulted in churn), a large proportion had the month-to-month contract.  Similarly, if we generate the mosaic plot for Churn and gender, we get the following:    Its plainly visible in this plot that the both genders are more or less equally represented in the customers with churn as well as the customers who did not move away from the business. it indicates that gender has a high likelihood of being independent of the churn.  The power of mosaic plots however lies in visualizing larger contingency matrices with more dimensions.  The mosaic plot for Churn plotted against tenure and Monthly Spend is provided below. Here we can see that the among the customers that had churn, the 0-3 month and 2-5 years tenured customers represent a high proportion. This can then be further broken up in to the monthly spend brackets. Here we see that in the short tenured segment, a monthly spend of $45-$80 makes up a higher proportion, while among the customers we lost who had 2-5 years tenure, the it’s the $80+ segment that contributes most to the churn.    Plotting the churn by tenure and monthly spend we see more interesting patterns. In the plot below its very clear that the customers with churn mostly had a month-to-month contract. We also see that in this high risk category of month-to-month subscribers, they had a low churn rate, when their monthly spend was under $45    We shall take a look at one last mosaic plot with one more dimension. Here we plot the churn against payment method, monthly spend and tenure.    Here we see that among the customers who had churn, a large proportion is made up of people who used the Electronic check payment method. Within this segment we can see that people who had a very short tenure of 0-3 months with a monthly spend of $45-$95 make up a lot of the churn. However, we also see that in the same segment when the tenure increased to 2-5 years a higher portion of the churn is produced by customers with $80+ in monthly spend. Predictive methods: Logistic regression vs. Decision Trees To predict a categorical variable like Churn, we can leverage either a Logistic regression or a Decision tree. Both have their advantages and disadvantages. For instance, decision trees are generally better at accuracy and Logistic regression better at probability estimation. To evaluate which of these methods offer a better fit, we can build both and compare their respective ROC curves. Logistic Regression Logistic regression is suited for creating predictive models that involve the prediction of nominal variable like in this case. For logistic regression we divide the dataset into a training set and a test set. To split the dataset, we use the sample.split() function in the caTools library. We also encode the Yes/No values in the Churn variable in to 0 and 1, where 0 means No – No Churn, and 1 means Yes – The customer has left the company.  With this we are able to split the dataset in to a training set and a test set. We will use the training set to train the model and then the test set to evaluate the model we built. To make the training more effective, we will use 90% the data to train and use the remaining 10% data to test the model and match our predictions with the actual churn in the data. The following screenshot shows how the data is split and our training and validation datasets are created.    Now we have taken our dataset and created a subset with 90% random sample that will serve as our training dataset. The remaining samples are saved separately as a test set, to validate our trained model against.  Now for our model, we can use the glm() function in R. We will include all the attributes in our dataset in our first iteration of the model and evaluate the model. Then we can remove the less discriminating variables and create a leaner model.    The summary of the model shows the following:   Finding discriminating values and refining the Model Here we can see the significant codes for the independent variables indicate the P-values, and hence the most discriminating or influential independent variables. For instance, we can see that gender does not have a significant impact on the result of Churn, and we can corroborate that with the bivariate statistics as well. We also see that Contract is highly influential, with a very low P-value.  With this we can refine our model to have only a handful of discriminating independent variables.    Evaluating this model, we get the following:   Decision Tree Model Our next step is to build a decision tree model. In R we can use the ROSE package to build decision trees and plot ROC curves. Once installed, we can use the rpart() function to build a regression tree model. The screenshot below shows the model. We use the same discriminating variables we found with our Logistic regression model here.    Notably the model summary will output the variable importance in order:   Comparing the Logistic Regression Model with RPART model We need to choose between the two models, and to do that we can plot their respective ROC curves and compare the areas under the curve.  We make a prediction using the test data set and then we plot the ROC curve for it.        Comparing the models, we can see that the logistic regression model is better suited here, since the area under the ROC curve is more for this model. Predicting the churn and cross validating the results With our updates to the model, we have fewer independent variables to consider, and the AIC value is slightly lesser indicating a slightly better fit for the model.  Using the refined model, we can now make some predictions on out test set. Once the predictions are made, we need to convert the probability scores to binary. We use the threshold level of 0.5, any probability score above 0.5 is considered true and anything below 0.5 is considered false.  To make comparison with the dataset easier, we create a new data frame to hold just the results and summarize them.    Here we can see that based on the training dataset, our trained model predicted the churn correctly 155 times out of 170 actual churn observations in the validation dataset. This yields us a model that is approximately 91% accurate. Data Summary Based on our analysis so far, we can establish that there is a strong link between the several independent variables and the churn rate.   * Tenure has a strong influence on the churn, we see that the customers in 0-3 month tenure bracket make up a large proportion of the churn. We can also see that within this bracket, Churn is high if the monthly charges are above $45. This effect can be visually detected from the bivariate statistics. We could offer short term discounts to keep customers. * The Contract type is also an influencer on the churn rate. The Month-to-Month subscribers have the highest rate of churn, and hence the highest risk. Here as well we see the influence of pricing though the bivariate statistics. For month-to-month subscribers, we could introduce lower cost plans, or more flexible plans that can lower the monthly spend. Bringing the monthly spend to under $45 seems to help with the curn in this segment. * We see churn around the two-year tenure mark. This is somewhat expected as there would be a good number of subscribers whose 2-year contract period is expiring. We could try to keep these customers by offering updated devices with new long-term contracts. * We also have a good amount of very loyal subscribers who have surpassed the 5 year mark. Although they do not represent a high churn risk, they could be vocal advocated for the services offered, just due to their tenure. They can be kept happy with additional customer service initiatives, and by identifying and prioritizing their issues. * Multiple line subscribers seem to have a lesser churn risk. We could expand on this trend by offering “family” plans and free communications between lines within the same family or shared limits. This could prompt families with multiple providers to converge on to our network. * Payment method seems to have an impact on the churn rate. People who use electronic checks as a payment method seems to have a higher churn risk. This could be due to the fact that banks usually send notifications to the users about payments, and it forms a constant reminder about the amount they spend. We could offer one time offers or other incentives to get customers to switch to another payment method like an automatic debit, so that the bills are not constantly scrutinized by the customers. * Tech support seems to have a positive influence on the churn rate. We see lesser churn on customers who have tech support. This can be interpreted in a couple of ways. One is that we may be selling devices that are hard to operate or un reliable, so that customers who do not have tech support are having ahrd time using our services. We can evaluate the devices we offer and allow to operate on our network and perhaps offer better devices. We can also interpret this as having a highly effective tech support team who can mitigate churn risk. However, exposing the tech support team to the entire customer base might decrease their effectiveness by overburdening them. Instead we can take the high-risk segments like customers on month-to-month contracts with a monthly spend of $45 or more and bundle free tech support for these customers. This gives the customers more value as well as the support of an effective team without over loading the tech support team. |
|  |
|  |

1. “3.10 Choosing Ranges of Values of Binned Variables.” *Data Mining and Statistics for Decision Making*, by Tufféry Stéphane, John Wiley, 2011. [↑](#footnote-ref-1)
2. *“Examining Distribution of Variables.” Data Mining and Statistics for Decision Making, by Tufféry Stéphane, John Wiley, 2011.* [↑](#footnote-ref-2)