Taylor Lucero

Predictive Classification Analysis

Telco Customer Churn - Decision Tree

Packages used:

readr
lm.beta
rpart
rpart.plot
e1071
caret
yardstick

Introduction

Telco is a Telecommunication company that provides a wide array of services to customers based on contracts. In this industry churn is the number of customers that leave the service of the company ,cancelling their subscription. By using predictive analytics Telco will be able to determine possible customers that may churn based on customer attributes. This is beneficial because it allows them to selectively target customers predicted to leave with promotions or focus on steps to better improve the customer experience.

Introducing the Dataset

In the provided there are 7043 instances with 21 variables.

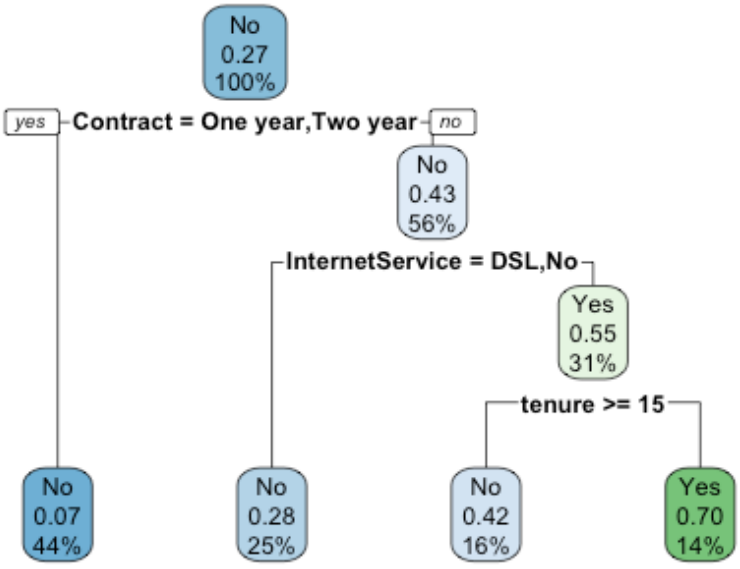
Variable	Data Type	Importance
CustomerID	String	Unique Customer Identifier
Gender	String (2 levels)	Determiner of male or female (Male , Female)
Senior Citizen	Integer (2 levels)	Determiner of senior citizen status (0 , 1)
Partner	String (2 levels)	Determiner of marriage status (No, Yes)
Dependents	String (2 levels)	Determiner of Dependents (No, Yes)
Tenure	Integer	Number of years with Telco
Phone Service	String (2 levels)	Determiner if they have phone service (No, Yes)
Multiple Lines	String (3 levels)	Determiner if the customer has multiple lines setup (No, Yes, No Phone Service)

Internet Service	String (3 levels)	Determiner if the customer has internet service provided (No, DSL, Fiber Optic)
Online Security	String (3 levels)	Determiner of if the customer has online security provided (No, Yes ,No internet Service)
Online Backup	String (3 levels)	Determiner of if data is stored for backup (No, Yes ,No internet Service)
Device Protection	String (3 levels)	Determiner of if devices are protected in the contract (No, Yes ,No internet Service)
Tech Support	String (3 levels)	Determiner if support is provided for issues the customer may be experiencing with internet services (No, Yes ,No internet Service)
Streaming Tv	String (3 levels)	Determiner if the customer uses internet service to stream TV (No, Yes ,No internet Service)
Streaming Movies	String (3 levels)	Determiner if the customer is streaming movies with internet service.
Contract	String (3 levels)	The type of contract customers are under (One Year, Month-to-Month, Two Year)
Paperless Billing	String (2 Levels)	How the customer receives bills.(No , Yes)
Payment Method	String (3 levels)	How the customer is paying for the services. (Bank Transfer, Mailed Check, Credit Card)
Monthly Charges	Numeric	The amount charged per month.

Total Charges	Numeric	The total amount charged so far.
Churn (Target Variable)	**String (2 Levels)**	**If a customer leaves or not.**

Once the background and information provided for each variable has been understood, the data was then cleaned to correct any erroneous instances that could provide issues during the analysis. With the small number of missing values in the Total Charges column, the data could have been omitted without having a large impact on the total number of customers and the statistics with the training and test data sets. After the training and test data sets have been assigned by randomly using 80% of the population in the Training and 20% in the Test data sets, the model was constructed. The training data is normally larger then the test data as it is used to help the algorithm learn determinant and factors that are significant in predicting the Churn of customers.

The Initial Constructed Model



The above Decision tree was constructed with the rpart package to fit the model using the test against the training data, then using rpart.plot to model a tree using this fitted information. This model shows a binary tree that splits based on three chosen variables, Contract, InternetService, and Tenure. Starting at the root node or the starting node and branching out towards the leaf nodes and terminal nodes based on “Yes” or “No” conditionals, this tree provided information stating, with the root node stating a probability of 27% of Churning in the entire population ,and then asking a conditional, if the Contract was one to two years, if “Yes” it

moves left, if “No” it moves right. Moving to the right at a decision node, this states that 56% don't have one or two year contracts with a probability of 43% Churning. This breakdown occurs until it reaches a terminal node, which is at the end of the Decision tree.

Confusion Matrix

	Predicted: No	*Predicted: Yes*
Actual: No	*1001*	*63*
Actual: Yes	*212*	*133*

The table provided shows the predicted against the actual instances of the Churn variable. Starting at the top left values in this cell are identified as True Positives (1001) as these are the correctly predicted values of those customers who actually don't churn, moving to the right this cell is identified as False Positive (63) as these are the customers who were predicted to Churn but actually did not. The bottom cells starting at the left, these are the False Negatives (212) meaning these are the customers that were predicted to not Churn but actually did. The last cell on the right, are identified as the True Negatives (133), meaning these are the predicted customers who are going to churn and did in fact Churn.

Understanding Parameters

The initial model and confusion matrix put through the confusionMatrix function from the caret and yardstick package provided a few important statistics on determining if the presented model is the best. Focusing on four of the most important, Precision, Recall, Accuracy, and the F1-measure. Precision is the ratio of correctly predicted positive observations to the total predicted positive observations, this makes the precision for this model .94. A high precision relates to how close estimates are to each other. The Accuracy is the most intuitive, it is a ratio of correctly predicted observation over the total, making the Accuracy for this model .80. This means that the model is approximately 80% accurate. The Recall statistic is the ratio of predicted positive observations to all the observations in the actual population. This is called sensitivity, and for this model it is .83. This shows out of the customers that truly Churned how many were correctly predicted. Taking both the Precision and Recall and finding their weighted average is called the F1-measure. The measure in this case would be .88. This takes into account both False Positive and False Negatives giving a better understanding when the cost of False positives and False Negatives are very different.

In simple terms, the initial model is approximately 80% accurate, with 94% percent of customers predicted to Churn actually Churned and 83% of customers that actually Churned that were, the model shows a decently constructed model. Looking at the implications that a False Positive and False Negative may have, a False Positive is more costly as customers that weren’t expected to Churn ended up Churning, this means customers that were not identified to be targeted to receive promotions or increased focus for marketing teams and account

managers left. If the False Positives had a larger degree of being correctly predicted as those in True Negatives, less may leave by being given special attention like those correctly predicted to Churn in the True Negative portion provided by the model. Customers placed in the False Negative portion don't have a negative impact towards Telco as they were incorrectly predicted to Churn but really did not. Meaning these customers stayed, the largest cost that may be caused by this may be labor efforts and a reduction of employee efficiency if they had to divide their focus further to include these customers in campaigns.

Constructing a Secondary Model

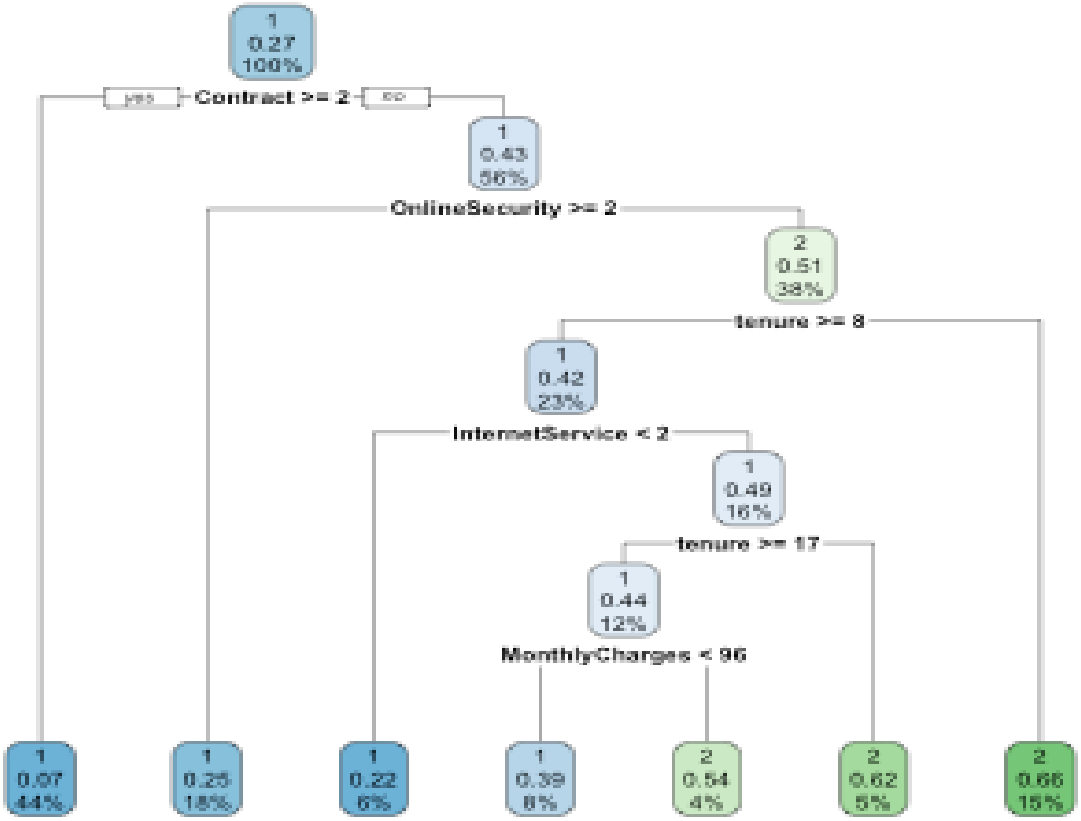
Sifting through the information in the first model, multiple points of improvement began to arise. This led to some skewed data that could have impacted the accuracy and predictive capabilities in the machine learning algorithm. To mitigate these issues, the variables were converted into the numeric factor forms, this allows the algorithm to easily utilize the numeric values representing non-numerical in the analysis. Then, I identified statistically significant variables in relation to the target variable Churn, by using a multiple linear regression model that provided a few parameters identifying significant attributes. However, since these parameters are not standardized these values could vary drastically in number and impact despite having similar significant figures. Correcting for this, I sought to standardize the statistics and focus on the beta coefficient and the t-value. Using the standardized statistics I was then able to determine variables that were not statistically significant to the target variable and then removed them from the data frame being analyzed. The removed variables are Gender, Streaming TV , Streaming Movies, and Payment Method. Of the data, this leaves 15 variables that are noted to have a significant impact on Churn. Once the information was groomed the data followed a similar procedure to the first model in producing a predictive model.

Factor Level Representation

Variable	Codex
Partner	1 = No, 2 = Yes
Dependents	1 = No, 2 = Yes

Phone Service	1 = No, 2 = Yes
Multiple lines	1 = No, 2 = No Phone Service, 3 = Yes
Internet Service	1= DSL,2 = FiberOptics, 3= No
Online Security	1 = No, 2 = No Internet Service, 3 = Yes
Online Backup	1 = No, 2 = No Internet Service, 3 = Yes
Device Protection	1 = No, 2 = No Internet Service, 3 = Yes
Tech Support	1 = No, 2 = No Internet Service, 3 = Yes
Contract	1 = Month to Month, 2 = One month , 3 = Two months
Paperless Billing	1 = No, 2 = Yes
Monthly Charges	This is an amount over a period of time, making it a numerical value at the start, no change was needed.
Total Charges	This is an amount over a period of time, making it a numerical value at the start, no change was needed.
Senior Citizen	This variable is unique as it utilizes a feature similar to the factor level, but in fact is relying on a boolean identifier. Meaning 0 = False, 1 =True. A value of 1 would identify if the customer was a Senior Citizen.
Customer ID	This is a Unique Identifier meaning for every customer this would have a different value making the factor level for this variable 7043.

The Secondary Model



Based on the newly presented predictive model, it shows a slightly more complex Decision Tree structure. This tree has a depth of 6 as that's the number of movements it would take to get to the farthest terminal node from the root node. The conditionals of the tree also changed to fit the new layout of the data, while still presenting some similarities to the original model. For example the condition set when departing from the root node is the same condition set in the original model. The reason behind its unique appearance is that the string values within each instance were changed to be represented by its specific numerical factor level. So if a contract is greater than or equal to the value of 2, that means based on the conversion that if a customer has either a month-to-month or one month contract the condition is True, moving the cursor to the left. This numerical representation is utilized throughout this entire model, allowing for more statistical and quantitative conditions to be set than the categorical or quantitative values.

The Second Confusion Matrix

Second Model	*Prediction No (1):*	*Prediction Yes (2):*
Actual No (1):	935	129
Actual Yes (2) :	160	185

Initial Model	*Predicted: No*	*Predicted: Yes*
Actual: No	1001	63
Actual: Yes	212	133

Second Model

Precision	.88
Accuracy	.80
Recall	.85
F-measure	.87

Initial Model

Precision	.94
Accuracy	.80
Recall	.83
F-measure	.88

Based on the presented confusion matrix and evaluating the same parameters from the initial matrix we can determine if the model is an improvement. Looking over the values of the True Positive values, there is a noticeable decrease (-66) of correctly predicted customers that didn't churn, while True Negative shows an increase (+52) in correctly predicted instances when the customer might churn and actually did. Next, looking at the False Positives there is an increase (+66) in the incorrectly predicted values, of those who were predicted to Churn but did not in fact leave. While False Negative, shows a decrease (-52) in the number of incorrectly predicted instances of those who were expected not to Churn but actually did.

Looking at the Parameters and comparing them to the previous matrix, there are a few differences. First, Precision has shown a slight decrease in the correctly predicted instances, it went from .94 to .88. This slight .06 decrease may not be that impactful in the overall scheme, looking at the Accuracy there is another slight decrease, from .80 to .795, but this isn't overly impactful because it is still rounded to .80. With Recall, we identify what would be known as the

sensitivity of the model. The Recall increased from .83 to .85. The F-measure shows a decrease from .88 to .87 which is not very impactful to the data.

With the largest change being Precision with a .06 decrease, we can see the data hasn't shown a drastic change in statistical information despite the removal of specific attributes. The data needs to be looked at holistically, compared to the original model we may have less True Positives in matrix but we also show an increase in True Negatives, which is more important because these are the customers that are going to churn. This new model allows a better True Negative group for individuals that are churning, allowing marketers and campaigns to better target these predicted customers before they leave, in turn giving a better probability of reducing the actual number of those who will leave after the end of the campaign and promotion. False Positives do show an increase from the initial to the second model but the actual risk derived from this group of incorrectly predicted instances is small. These are the customers that were predicted to Churn but did not. The only loss here may be employee efficiency if they are focusing on customers that were never going to leave regardless. False Negatives in this case are the most risky and impactful group, the second model in this regard shows an improvement with a decrease from 212 to 160. This change shows a major impact in the predictive model as less customers who were predicted to stay didn't. This group has the largest impact per customer because they weren't expected to leave so they wouldn't fall into the groups incorporated into the campaigns and promotions. To better understand, these False Negative customers are similar to a Sucker punch, with no warning or defense on the part of the recipient they leave, while the False Positives, are a group similar to that of a avoided punch because there is no major impact received by the recipient. Of these two models, the second model shows the best mitigative and predictive measures while presenting similar predictive prowess as the original model.

Tuning Hyper-Parameters

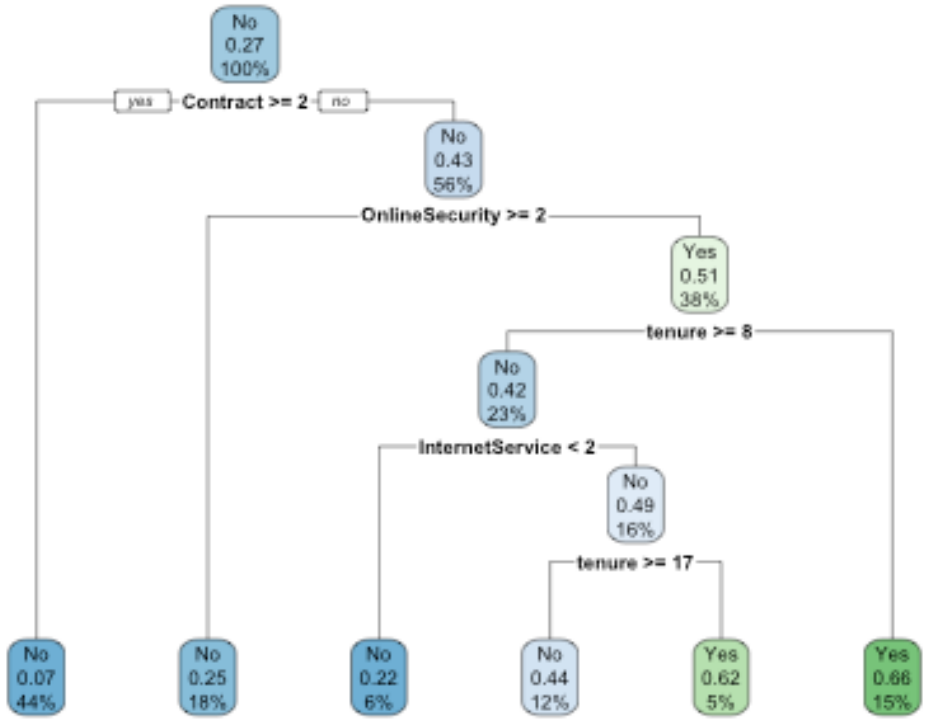
With the variations the Decision Tree algorithm provides, the last step is making sure the Hyper-parameters are tuned so they can be identified for creating the best predictive model. This focuses on, questions "Why should the depth of this Tree be 6? Would a Depth of 5 or 7 be better for the model?" Tuning these Hyper parameters help to improve the strength of the model, however, it is important to keep in mind Underfitting and Overfitting the model. In this analysis I used the prune function to help ease the process of refining the model. The overall shape of the model stayed consistent but it did alter the confusion matrix.

The Final Model

Final Model	*Prediction No (1):*	*Prediction Yes (2):*
Actual No (1):	964	100
Actual Yes (2) :	178	167

Second Model	*Prediction No (1):*	*Prediction Yes (2):*
Actual No (1):	935	129
Actual Yes (2) :	160	185

As you can see the change in number was not drastic but it still reduced the predicted impact from the False Negatives.



Overall

The refined predictive model informs users of information to make actionable decisions, in this case to help limit and mitigate customers that may leave. Based on probability adn conditionals deploying this model will help Telco make informed decisions.