Business Intelligence and Data Mining

MIST.6060 - 201

PROJECT TEAM:

Matthew Davis.

Stephen Hartigan.

Sridhar Rangan

University of Massachusetts – Lowell

December 3, 2018

Prof. Xiaobai (Bob) Li, PhD



Predicting Customer Churn at Telco Communications

# Abstract

Customer loyalty is often a leading indicator of corporate financial success. As such, customer churn (unexpected termination of service) would intuitively negatively impact a company’s financial performance. In this study, we will look at a large dataset of customer attributes in the hopes of predicting what factors contribute most to customer churn. The resulting model(s) will then be used to create strategies that would minimize churn going forward.

Table of Contents

[Abstract 1](#_Toc531547161)

[Background 1](#_Toc531547162)

[Customer Churn 1](#_Toc531547163)

[Telco Communications 2](#_Toc531547164)

[Objective 2](#_Toc531547165)

[Technique 2](#_Toc531547166)

[Cross-Industry Standard Process for Data Mining (CRISP-DM) 2](#_Toc531547167)

[Business Understanding 3](#_Toc531547168)

[Data Understanding 3](#_Toc531547169)

[Model Building 7](#_Toc531547170)

[Testing and Evaluation 9](#_Toc531547171)

[Deployment 10](#_Toc531547172)

[Practical Implications and Recommendations 10](#_Toc531547173)

[Appendix A – Results from (J48) Decision Tree Model 12](#_Toc531547174)

[Appendix B – Results from Naïve Bayes Model 14](#_Toc531547175)

[Appendix C – Results from Support Vector Machine (SMO) Model 16](#_Toc531547176)

[Appendix D – R Code for Descriptive Statistics and Conversions 18](#_Toc531547177)

[References 23](#_Toc531547178)

# Background

## Customer Churn

As mentioned in the abstract, the goal of this study was to understand the leading indicators of customer retention and to discuss methods of retaining customers that may inevitably move to competitor firms. In the business world, the term customer churn is used to indicate when a firm’s customers halt doing business. Customer churn is an important aspect for any business-to-business (B2B) or business-to-consumer (B2C) company as it is a key factor in determining growth. In fact, customer retention is known to be less expensive than acquiring new customers due to the lengthy process involved to gain new customers.[[1]](#endnote-2) Customer churn plays a vital role in any B2B or B2C company and must be monitored carefully by ensuring the satisfaction of customers.

In a customer experience article, published by Oracle in 2011, several factors were highlighted for reducing customer churn. The main idea put forth by the report indicates firms should focus on the customer by providing great customer service. In the survey’s conducted, Oracle discovered that 86% of consumers will pay for a better customer experience.[[2]](#endnote-3) The takeaway from the article was a recommendation to all firms either B2B or B2C to make a concentrated effort in delivering quality service for customers in order to reduce customer churn.

## Telco Communications

The data set that will be used for this project is from Kaggle. Our data comes from an IBM sample dataset. Therefore, the company has been generalized as “Telco Communications” which is said to provide phone, internet and TV services to customers. Since, Telco Communications is a B2C firm, customer churn is an important factor due to the number of consumers they serve. As noted previously, it is less expensive to retain customers rather than pursue new ones due to the high cost. For Telco Communications, customer churn is important in maintaining constant growth and increasing market share in a competitive industry.

The data obtained in this data set was collected via customer sign-ups and contracts. From the contract, data such as services, account and demographic information can be collected. In general, for telecommunications companies, much of the data comes from contracts signed by the customer. Additionally, companies can verify certain information through external third-party services as well as US Census data.

# Objective

For all of the reasons outlined above, the overall objective of this study is to identify the key contributing factors that drive customer churn. Once identified, we believe that we can develop strategies to mitigate these factors and improve revenues for the organization.

# Technique

## Cross-Industry Standard Process for Data Mining (CRISP-DM)

One of the more popular methods used in approach to Data Mining projects is known as the cross-industry standard process for data mining (CRISP-DM)[[3]](#endnote-4). There is a natural sequence to the process that promotes more meaningful analyses. In this section we will describe our process using header terminology specific to the CRISP-DM process as shown in Figure 1 below.

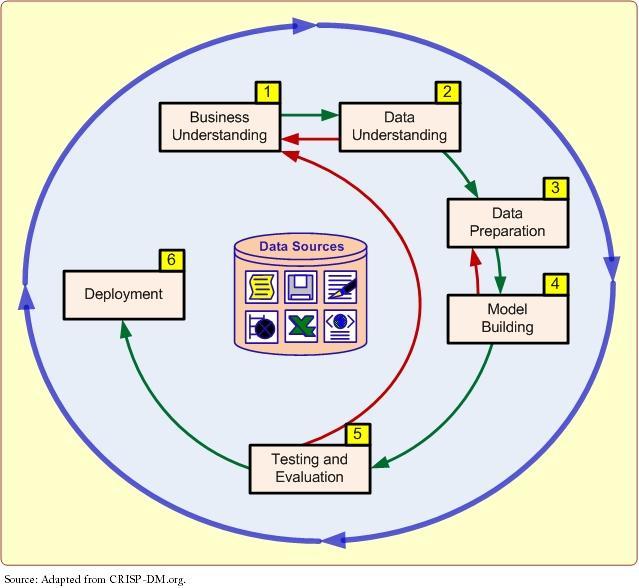


Figure 1 - CRISP-DM Flow Chart

### Business Understanding

In the case of customer churn, we believe that the business process is well defined within the Background and Objective sections of this document.

### Data Understanding

#### Data Description

As previously mentioned, the dataset used in this study was obtained from ([www.kaggle.com/blastchar/telco-customer-churn](http://www.kaggle.com/blastchar/telco-customer-churn)).  Within the dataset, there are 7,043 observations, 20 variables and one target variable: “Churn”.  Upon receiving the data, there were actually 22 columns with one column representing a unique customer ID.  This column was deleted as it does not contain any useful information necessary for the analysis of customer churn.  More important are the remaining variables and their definitions which will be provided below in Table 1.

|  |  |  |
| --- | --- | --- |
| **Name of Attribute** | **Data Type** | **Description** |
| Customer ID | Label | Unique identifier for customer |
| Gender | Categorical | Customer gender (Male, Female) |
| Senior Citizen | Numeric | Whether customer is a senior citizen (1) or not (0) |
| Partner | Categorical | Whether customer has a partner or not (Yes, No) |
| Dependents | Categorical | Whether customer has dependents or not (Yes, No) |
| Tenure | Numeric | Number of months the customer has stayed with the company. |
| PhoneService | Categorical | Whether customer has a phone service or not (Yes, No) |
| MultipleLines | Categorical | Whether customer has multiple lines or not (Yes, No, No phone service) |
| InternetService | Categorical | Customer’s internet service provider (DSL, Fiber optic, No) |
| OnlineSecurity | Categorical | Whether customer has online security or not (Yes, No, No internet service) |
| OnlineBackup | Categorical | Whether customer has online backup or not (Yes, No, No internet service) |
| DeviceProtection | Categorical | Whether customer has device protection or not (Yes, No, No internet service) |
| TechSupport | Categorical | Whether customer has tech support or not (Yes, No, No internet service) |
| StreamingTV | Categorical | Whether customer has streaming TV or not (Yes, No, No internet service) |
| StreamingMovies | Categorical | Whether customer has streaming movies or not (Yes, No, No internet service) |
| Contract | Categorical | The contract term of the customer (Month-to-month, One year, Two year) |
| PaperlessBilling | Categorical | Whether customer has paperless billing or not (Yes, No) |
| PaymentMethod | Categorical | The customer’s payment method (Electronic check, Mailed check, Bank transfer (automatic), Credit card (automatic)) |
| MonthlyCharges | Numeric | The amount charged to the customer monthly |
| TotalCharges | Numeric | The total amount charged to the customer |
| TotalDiff | Numeric | Calculation added to dataset: [(Tenure x MonthlyCharges) - TotalCharges] |
| Churn | Categorical | Whether customer churned or not (Yes or No) |

Table 1 - List of Attributes in the dataset

#### Descriptive Statistics

This dataset presented several important relationships that are important to note before a discussion on model performance can be introduced.  Of the 7,032 observations, 3,483 are females, and 3,549 are males (Figure 2).  Figure 3displays the contrast in customer churn (73.4% - No, 26.6% - Yes), all further percentages are relative to this 73.5:26.5 percentage split.  Therefore, an investigation into the difference in churn between the genders (Figure 4), Of the 7,032 observations, 13.4% of churn were females while 13.2% were males.  As is apparent in the chart, the difference is trivial.  This indicates little to no difference in gender and their respective decisions to stay or leave.

|  |  |  |
| --- | --- | --- |
| Figure 2 - Comparison of gender | Figure 3 - Comparison of churn | Figure 4 - Churn between gender |

One of the more important comparisons has to do with the length of a customer’s contract and whether they churn or not.  Figure 5 indicates a substantial difference in longer contracts and the churn rate of these customers.  It appears that customers that have month-to-month contracts are more unpredictable, and therefore should be more of a focus once we begin to investigate the implications of our study.  Approximately 31.6% of loyal customers - that is did not churn - were “month-to-month”, while 23.5% of month-to-month customers decided to leave to find a better deal with another company.  Of the 26.6% of customers whom left the company, month-to-month customers were 70% of the churn customers that decided to leave to find a better deal!  As for 2-year contracted customers, 23.3% remain loyal while only 6.8% brought their business elsewhere.  Part of this difference may be newer customers trying out the services before committing to a long-term contract, however it is an important note as we investigate further trends using data mining techniques.  Whereas, customers committed to a longer contract seem happy with their services.  Similarly, Figure 6displays the length of tenure (in months) as a leading indicator in determining whether a customer will stay or leave the company.  Clearly, a large majority of customers that stay with Telco after 12 months will most likely stay with the company even longer.  In the last bar (customers with Telco between 6 and 7 years), 18.7% decided to stay with the company longer than 60 months, while only 1.3% will move on.  On the opposite side of the spectrum, 16.2% will stay while 14.8% will leave to find a different company, proving to be a bit more unpredictable with their customer loyalty.

|  |  |
| --- | --- |
| Figure 5 - Contract and Churn | Figure 6 - Tenure and churn |

Finally, we address the topics discussed in the preliminary research regarding main indicators of customer churn, it was found that customer service ranks as one of the most important characteristics.  Here we display the statistical comparisons of “InternetService,” “TechSupport” in Figure 7 and Figure 8respectively.  The highest rate of churn with respect to internet service is displayed through the Fiber Optic (FO) option.  Of the nearly 26.6% of churn customers, 18.4% of customers who left had Fiber Optic services.  This indicates that Telco’s Fiber Optic service may have lacked the quality that other companies may have been able to provide (i.e. Verizon Fios).  Furthermore, customers without Internet Service (20% - No churn; 1.6% Yes churn) may be an older demographic, therefore are not looking for all of the “bells and whistles”.

As previously stated, Tech Support and customer service have proven to be important factors in customers decision to stay with a company or leave.  Displayed in **figure 8** is yet another justification to this statement.  Of the 73.5% of customers who stayed with Telco, only 28.8% did not have tech support with their subscription.  Likewise, 20.5% of the 26.6% of churned customers did not have any tech support with their subscription.  This indicates that customer service was important to leaving customers.  Further proof that future prevention would require stellar customer service.

|  |  |
| --- | --- |
| Figure 7 - Internet Services and customer churn | Figure 8 - Tech Support and customer churn |

#### Data Preparation

Below is a sample of the data for context.



Table 2 - Sample from the original dataset

As mentioned, our first step in manipulating the data was to eliminate the “CustomerID” attribute, as it is a unique identifier (label attribute). In addition, it was discovered that there were 11 missing values within “TotalCharges”.  To ensure that the models were not skewed by these omissions, we deleted the entire instance (row) for all occurrences of missing data.

Upon investigating various methods for modeling, it was decided that several variables should have shorter titles to ensure simplicity when viewing the models created through Weka.  Therefore, variables were altered using Excel’s Find / Replace tool.

|  |  |
| --- | --- |
| After replacement, each variable has the new levels below. | |
| MultipleLines – “No”, “Nps”, “Yes” | InternetService – “DSL”, “FO”, “No” |
| OnlineSecurity – “No”, “Nis”, “Yes” | OnlineBackup – “No”, “Nis”, “Yes” |
| DeviceProtection – “No”, “Nis”, “Yes” | TechSupport – “No”, “Nis”, “Yes” |
| StreamingTV – “No”, “Nis”, “Yes” | StreamingMovies – “No”, “Nis”, “Yes” |
| Contract – “M-T-M”, “1yr”, “2yr” | PaymentMethod – “BT”, “CC”, “EC”, “MC” |

Table 3 - Abbreviated attribute classes

Further analysis of the dataset intimated that there was an expected value for the TotalCharges based on the monthly charges and tenure.  In order to vet whether that was a more accurate predictor of churn, we created a calculated column (TotalDiff) to demonstrate the difference between this expected value and the total charges incurred (more on that in the Testing and Evaluation section below).  Note that our models will use either TotalCharges, or TotalDiff, since the latter is derived from the former. The calculation was as follows:

Finally, as a means of displaying descriptive statistics, as well as creating more efficient model performance (where warranted), it was decided that we would “bin” the attributes titled “MonthlyCharges” and “tenure”.  With “MonthlyCharges”, the data was cut into increments of 24.  This created 5 levels between the minimum charge ($0) and the maximum charge ($118.75).   In order to “bin” these attributes the following code was created using R, all lines are available for viewing in Appendix D – R Code for Descriptive Statistics and Conversions.

* *CleanProj1$MonthlyCharges1=cut(x=CleanProj1$MonthlyCharges, breaks=c(0, 24, 48, 72, 96, 120))*

With respect to “tenure”, buckets were created in yearly increments (0 to 12, 12-24, etc.), and since the minimum and maximum length of tenure for customers within this dataset were 0 and 72 respectively, 6 buckets were created using the following line of code.

* *CleanProj1$tenure1=cut(x=CleanProj1$tenure, breaks=c(min(CleanProj1$tenure)-0.001, 12, 24, 36, 48, 60, 72))*

### Model Building

In this section we will discuss the models that we have decided to use and the justification for why we have chosen such models. It is important to note that modeling provides a method of prediction, allowing its director to both infer about the data as well as predict future behavior. In our study, we chose to investigate the data using three models: (J48) Decision Trees, Naïve Bayes, and (SMO) Support Vector Machines all through the use of Weka. All methods provide an investigation into the analytical understanding of the data. We shall begin our investigation by defining the parameters for Decision Trees.

**(J48) Decision Trees**

Descision trees provide analyst with an opportunity to visualize the data and its most important predictors. Although we have the opportunity to determine the coefficient correlation in Weka, it does not visually appeal to the user and their audience. Furthermore, the critical coefficient to modify within a decision tree model is the number of instances in each leaf. Through trial and error, we found that 70 gave a reasonably sized tree of 20 leaves. After adjusting various parameters, we successfully ran a (J48) Decision Tree and identified that the root node was the customers length of contract. The output (including visualization of the tree) can be found in Appendix A – Results from (J48) Decision Tree Model.

**Naïve Bayes**

Naive Bayes is a probabilistic classifier that assumes that all events are conditionally independent. Furthermore, Naïve Bayes is typically used for categorical data. However, we decided to leave these numerical attributes in the data to remain consistent among all models. In the following tables we will discuss the results of the model, however it is important to note the error rate and the ROC value. Naïve Bayes manages the highest false positive to true positive rate (ROC value), indicating that Naïve Bayes has some predictive advantages to the other models run. However, because the visualization of the model is limited, we found this model to be the least descriptive of future datasets. The results of the model can be found in Appendix B – Results from Naïve Bayes Model.

**Support Vector Machine (SMO)**

Support Vector Machines allow the user to create a hyperplane that separates our dataset into two classes (class attribute of Churn: “No” or “Yes”). The usefulness of the SVM model is in the attribute weights which are in the classifier output window. Similar to Decision Trees, we see a correlation between the variables with the highest predictive power within their model. Again, tenure, InternetService (FO) and TotalCharges have the highest impact in the overall selection of the class attribute. This impact falls in line with the descriptive statistics as well (refer to figures 5 through 8). In implementation of the Support Vector Machine model, we left all default values in Weka as found. The results of the model can be found in Appendix C – Results from Support Vector Machine (SMO) Model.

**Unused Methods**

Given that our objective in this analysis was to predict customer churn given an array of variables, it was unnecessary to investigate unsupervised datamining techniques such as Clustering and Weka’s Apriori algorithm. These techniques are useful for identifying unknown relationships between both categorical and numeric datasets, but as mentioned, they have little use in our current study.

### Testing and Evaluation

Before we address the output of our individual models, we would like to address some of the findings discovered while testing our models. As mentioned in the Data Preparation section above, we created several new attributes for analysis. It was no surprise that our continuous attributes (Tenure and MonthlyCharges) were more accurate predictors of churn than their categorized counterparts (Tenure1 and MonthlyCharges1). While the categorized attributes proved useful from a descriptive analytics perspective (visualization of the distribution for each attribute), the importance of their original counterparts provided more variation in the final model results. For example, using decision trees with the continuous data for “tenure”, it was clear that 15 months was the decisive boundary where customers would either remain with the company or “churn” and find another company for their services.

Additionally, we created a calculated attribute that attempted to discern whether a customer was paying more than they were anticipating. By multiplying the tenure (in months) by the monthly charges, and subtracting the actual total charges, we assumed that a positive number (predicted > actual spend) would suggest there were some savings, and cause customers to stay. Similarly, a negative value (actual value > predicted) would signify exorbitant charges and lead customers to leave.

Not only did the data refute this notion, but it did so emphatically. The only attribute with a lower correlation to churn than TotalDiff was gender. As such, none of the models in our evaluation included the TotalDiff attribute.

With our newly created attributes completely stripped from the dataset, we began running the dataset through various models. As expected, the output of the various models produced mixed performance results. Below, you can see the confusion matrices for each model.

|  |  |  |
| --- | --- | --- |
| Decision Tree | Naïve Bayes | Support Vector Machine |
|  |  |  |

Table 4 - Confusion matrices

Note that the positive result was assigned to “No” since that is the more prevalent value. While counter-intuitive (one would typically associate “Yes” with the positive outcome), the manipulation required to remediate the outcome to make it more intuitive (transpose the results to a separate confusion matrix) were not pursued in the interest of time.

The performance metrics for each model were tabulated for comparison’s sake, as shown in Table 5 below. Note that maximum and minimum values are highlighted. We could certainly choose and deploy the model with the highest overall accuracy (Support Vector Machine). However, since churn is the target attribute, and “yes” is the only outcome we care to take action on, we might choose the model with the greatest specificity (Naïve Bayes). That model has most accurately classified this dataset provided in regards to those that churn.



Table 5 - Performance Metrics

Depending upon how the models are deployed, there are certainly other metrics to be considered, and we have included an extensive list of those available from the Weka model for that reason.

### Deployment

The models produced adequate information to begin strategizing methods to address churn. In fact, even though the decision tree “performed” worse than our other models, the visualization of the tree proved valuable in itself. We will discuss further in the next section. For the purposes of this study, we believe that any future data obtained would be best analyzed using Naive Bayes based on the specificity therein.

# Practical Implications and Recommendations

The models and metrics described above are mathematical representations of the dataset, that provide little value without proper context. We know that we want to prevent churn, and therefore hope to identify any insights that might predict future churn. According to all the models, there are four major contributors to churn: tenure, length of contract, internet service and total charges.

Intuitively, tenure makes sense. People that churn would, by definition, have less tenure than their more loyal counterparts. It is the analysis of the various inputs in regards to tenure that we find value. Our decision tree suggest that tenure less than 15 months is where we see the most churn. This provides context to the customers that should be targeted from a marketing standpoint to ensure that they remain on-board. However to focus on tenure as an actionable indication of churn would be misleading. Reducing churn by definition will increase tenure.

Length of contract is similarly intuitive in its prognosticative value. Those with a longer contract have more to lose in defecting. Likewise, those that choose a month to month contract are more likely to be considering an early departure in the first place. The data bears that out. Whether or not the company can provide any more value to long term contracts is open for debate. Regardless, the effects of the contract toward mitigating churn are obvious from the dataset provided.

Total charges also proved to be a misleading indicator (similar to tenure). Those that have stayed longer, have accumulated more total monthly costs as a result. The data suggests that those with a GREATER overall total cost are more likely STAY than churn. We believe that this is an effect of customer loyalty as opposed to a cause of customer churn.

What proved most valuable in terms of actionable data was the internet service attribute. According to the data, 1,869 of the 7,032 customers “churned”, yielding a churn rate of 26.6%. Using the output from the naïve bayes model, we can see that Fiber optic internet users churned at a rate of 41.9%. Although DSL subscribers churn at a lower than average rate of 19%, we can tell from the decision tree that the greatest predictor of churn there (after negating total charges, as specified above) is indicated by who opted out of both Online Security and Tech Support (43% churn rate). For fiber optic subscribers (this time negating tenure), we can see that those with multiple phone lines that also pay by electronic check have a 38% churn rate.

Armed with this new perspective, we can market to our DSL customers with more emphasis on the added services, and work to dissuade fiber optic customers (or investigate further the actual root cause) that pay by electronic check.

In conclusion, we believe that internet subscribers need to be better understood, and the Telco marketing group should be focusing their efforts in customer engagement there. A more comprehensive survey of WHY customers defer technical support or opt for electronic check payment could go a long way toward improving churn rate, and inevitably revenues / market share.

# Appendix A – Results from (J48) Decision Tree Model

=== Run information ===

Scheme: weka.classifiers.trees.J48 -C 0.25 -M 70

Relation: CleanProj1-weka.filters.unsupervised.attribute.Remove-R20,22-23

Instances: 7032

Attributes: 20

gender

SeniorCitizen

Partner

Dependents

tenure

PhoneService

MultipleLines

InternetService

OnlineSecurity

OnlineBackup

DeviceProtection

TechSupport

StreamingTV

StreamingMovies

Contract

PaperlessBilling

PaymentMethod

MonthlyCharges

TotalCharges

Churn

Test mode: 10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree

------------------

Contract = M-T-M

| InternetService = DSL

| | TotalCharges <= 310.6

| | | OnlineSecurity = No

| | | | TechSupport = No: Yes (335.0/144.0)

| | | | TechSupport = Yes: No (71.0/27.0)

| | | | TechSupport = Nis: Yes (0.0)

| | | OnlineSecurity = Yes: No (92.0/34.0)

| | | OnlineSecurity = Nis: Yes (0.0)

| | TotalCharges > 310.6: No (725.0/142.0)

| InternetService = FO

| | tenure <= 15: Yes (1036.0/319.0)

| | tenure > 15

| | | tenure <= 51

| | | | PaperlessBilling = Yes

| | | | | TechSupport = No

| | | | | | MultipleLines = Nps: Yes (0.0)

| | | | | | MultipleLines = No: No (164.0/67.0)

| | | | | | MultipleLines = Yes

| | | | | | | PaymentMethod = EC: Yes (244.0/93.0)

| | | | | | | PaymentMethod = MC: No (17.0/5.0)

| | | | | | | PaymentMethod = BT: No (72.0/34.0)

| | | | | | | PaymentMethod = CC: No (60.0/25.0)

| | | | | TechSupport = Yes: No (142.0/49.0)

| | | | | TechSupport = Nis: No (0.0)

| | | | PaperlessBilling = No: No (178.0/59.0)

| | | tenure > 51: No (215.0/55.0)

| InternetService = No: No (524.0/99.0)

Contract = 1yr: No (1472.0/166.0)

Contract = 2yr: No (1685.0/48.0)

Number of Leaves : 20

Size of the tree : 31

Time taken to build model: 0.05 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances 5552 78.9534 %

Incorrectly Classified Instances 1480 21.0466 %

Kappa statistic 0.4312

Mean absolute error 0.2819

Root mean squared error 0.3791

Relative absolute error 72.2195 %

Root relative squared error 85.8126 %

Total Number of Instances 7032

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class

0.886 0.477 0.837 0.886 0.861 0.434 0.818 0.915 No

0.523 0.114 0.624 0.523 0.569 0.434 0.818 0.599 Yes

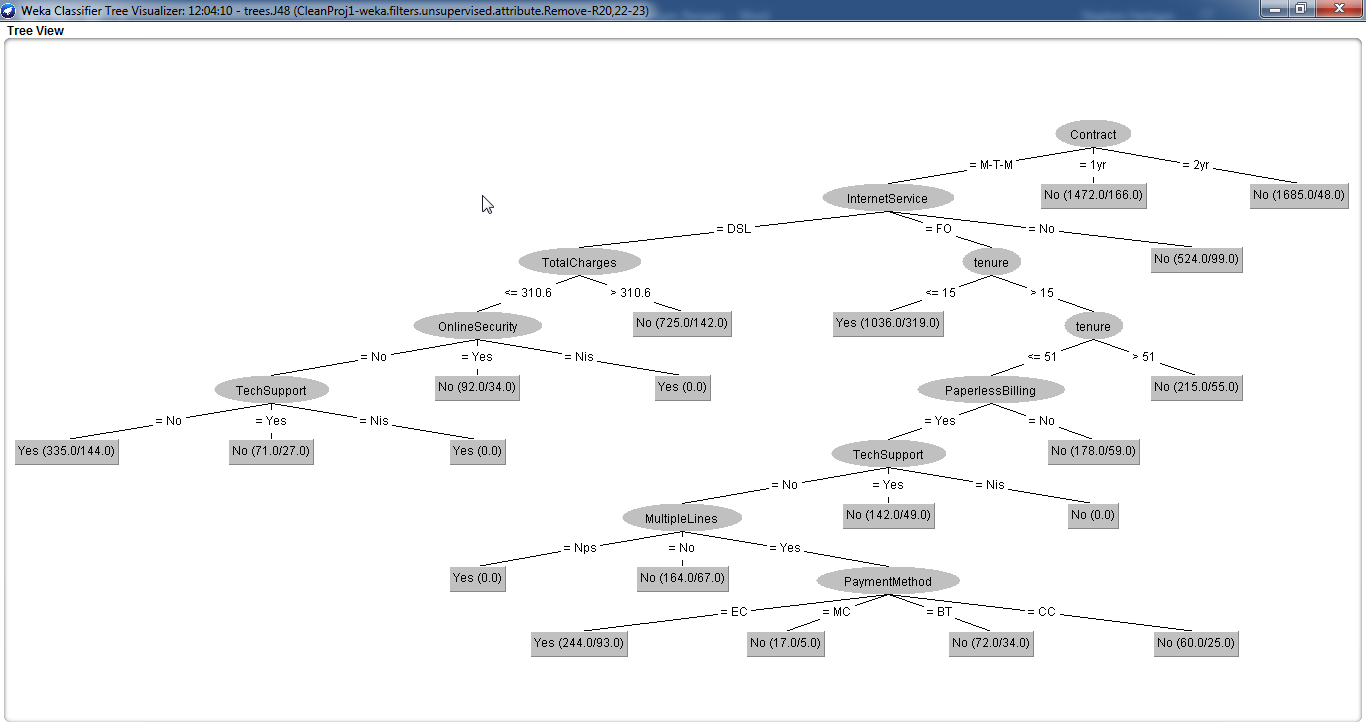
Weighted Avg. 0.790 0.381 0.780 0.790 0.783 0.434 0.818 0.831

=== Confusion Matrix ===

a b <-- classified as

4575 588 | a = No

892 977 | b = Yes



# Appendix B – Results from Naïve Bayes Model

=== Run information ===

Scheme: weka.classifiers.bayes.NaiveBayes

Relation: CleanProj1-weka.filters.unsupervised.attribute.Remove-R20,22-23

Instances: 7032

Attributes: 20

gender

SeniorCitizen

Partner

Dependents

tenure

PhoneService

MultipleLines

InternetService

OnlineSecurity

OnlineBackup

DeviceProtection

TechSupport

StreamingTV

StreamingMovies

Contract

PaperlessBilling

PaymentMethod

MonthlyCharges

TotalCharges

Churn

Test mode: 10-fold cross-validation

=== Classifier model (full training set) ===

Naive Bayes Classifier

Class

Attribute No Yes

(0.73) (0.27)

=======================================

gender

Female 2545.0 940.0

Male 2620.0 931.0

[total] 5165.0 1871.0

SeniorCitizen

mean 0.129 0.2547

std. dev. 0.3352 0.4357

weight sum 5163 1869

precision 1 1

Partner

Yes 2725.0 670.0

No 2440.0 1201.0

[total] 5165.0 1871.0

Dependents

No 3391.0 1544.0

Yes 1774.0 327.0

[total] 5165.0 1871.0

tenure

mean 37.65 17.9791

std. dev. 24.0746 19.5259

weight sum 5163 1869

precision 1 1

PhoneService

No 511.0 171.0

Yes 4654.0 1700.0

[total] 5165.0 1871.0

MultipleLines

Nps 511.0 171.0

No 2537.0 850.0

Yes 2118.0 851.0

[total] 5166.0 1872.0

InternetService

DSL 1958.0 460.0

FO 1800.0 1298.0

No 1408.0 114.0

[total] 5166.0 1872.0

OnlineSecurity

No 2037.0 1462.0

Yes 1721.0 296.0

Nis 1408.0 114.0

[total] 5166.0 1872.0

OnlineBackup

Yes 1903.0 524.0

No 1855.0 1234.0

Nis 1408.0 114.0

[total] 5166.0 1872.0

DeviceProtection

No 1884.0 1212.0

Yes 1874.0 546.0

Nis 1408.0 114.0

[total] 5166.0 1872.0

TechSupport

No 2027.0 1447.0

Yes 1731.0 311.0

Nis 1408.0 114.0

[total] 5166.0 1872.0

StreamingTV

No 1868.0 943.0

Yes 1890.0 815.0

Nis 1408.0 114.0

[total] 5166.0 1872.0

StreamingMovies

No 1844.0 939.0

Yes 1914.0 819.0

Nis 1408.0 114.0

[total] 5166.0 1872.0

Contract

M-T-M 2221.0 1656.0

1yr 1307.0 167.0

2yr 1638.0 49.0

[total] 5166.0 1872.0

PaperlessBilling

Yes 2769.0 1401.0

No 2396.0 470.0

[total] 5165.0 1871.0

PaymentMethod

EC 1295.0 1072.0

MC 1297.0 309.0

BT 1285.0 259.0

CC 1290.0 233.0

[total] 5167.0 1873.0

MonthlyCharges

mean 61.3073 74.4412

std. dev. 31.0911 24.6588

weight sum 5163 1869

precision 0.0635 0.0635

TotalCharges

mean 2555.3435 1531.7976

std. dev. 2329.237 1890.3342

weight sum 5163 1869

precision 1.3273 1.3273

Time taken to build model: 0.06 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances 5107 72.6251 %

Incorrectly Classified Instances 1925 27.3749 %

Kappa statistic 0.4166

Mean absolute error 0.2753

Root mean squared error 0.4754

Relative absolute error 70.5224 %

Root relative squared error 107.6074 %

Total Number of Instances 7032

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class

0.699 0.198 0.907 0.699 0.789 0.446 0.819 0.920 No

0.802 0.301 0.491 0.802 0.609 0.446 0.819 0.612 Yes

Weighted Avg. 0.726 0.225 0.796 0.726 0.741 0.446 0.819 0.838

=== Confusion Matrix ===

a b <-- classified as

3608 1555 | a = No

370 1499 | b = Yes

# 

# Appendix C – Results from Support Vector Machine (SMO) Model

=== Run information ===

Scheme: weka.classifiers.functions.SMO -C 1.0 -L 0.001 -P 1.0E-12 -N 0 -V -1 -W 1 -K "weka.classifiers.functions.supportVector.PolyKernel -E 1.0 -C 250007" -calibrator "weka.classifiers.functions.Logistic -R 1.0E-8 -M -1 -num-decimal-places 4"

Relation: CleanProj1-weka.filters.unsupervised.attribute.Remove-R20,22-23

Instances: 7032

Attributes: 20

gender

SeniorCitizen

Partner

Dependents

tenure

PhoneService

MultipleLines

InternetService

OnlineSecurity

OnlineBackup

DeviceProtection

TechSupport

StreamingTV

StreamingMovies

Contract

PaperlessBilling

PaymentMethod

MonthlyCharges

TotalCharges

Churn

Test mode: 10-fold cross-validation

=== Classifier model (full training set) ===

SMO

Kernel used:

Linear Kernel: K(x,y) = <x,y>

Classifier for classes: No, Yes

BinarySMOMachine linear: showing attribute weights, not support vectors.

-0.0428 \* (normalized) gender=Male

+ 0.184 \* (normalized) SeniorCitizen

+ -0.0311 \* (normalized) Partner=No

+ -0.0917 \* (normalized) Dependents=Yes

+ -1.3602 \* (normalized) tenure

+ -0.0882 \* (normalized) PhoneService=Yes

+ 0.0882 \* (normalized) MultipleLines=Nps

+ -0.1626 \* (normalized) MultipleLines=No

+ 0.0744 \* (normalized) MultipleLines=Yes

+ -0.5997 \* (normalized) InternetService=DSL

+ 0.6861 \* (normalized) InternetService=FO

+ -0.0865 \* (normalized) InternetService=No

+ 0.1625 \* (normalized) OnlineSecurity=No

+ -0.0761 \* (normalized) OnlineSecurity=Yes

+ -0.0865 \* (normalized) OnlineSecurity=Nis

+ 0.0081 \* (normalized) OnlineBackup=Yes

+ 0.0783 \* (normalized) OnlineBackup=No

+ -0.0865 \* (normalized) OnlineBackup=Nis

+ 0.0316 \* (normalized) DeviceProtection=No

+ 0.0549 \* (normalized) DeviceProtection=Yes

+ -0.0865 \* (normalized) DeviceProtection=Nis

+ 0.1776 \* (normalized) TechSupport=No

+ -0.0911 \* (normalized) TechSupport=Yes

+ -0.0865 \* (normalized) TechSupport=Nis

+ -0.0616 \* (normalized) StreamingTV=No

+ 0.1481 \* (normalized) StreamingTV=Yes

+ -0.0865 \* (normalized) StreamingTV=Nis

+ -0.0799 \* (normalized) StreamingMovies=No

+ 0.1664 \* (normalized) StreamingMovies=Yes

+ -0.0865 \* (normalized) StreamingMovies=Nis

+ 0.1301 \* (normalized) Contract=M-T-M

+ -0.1638 \* (normalized) Contract=1yr

+ 0.0337 \* (normalized) Contract=2yr

+ -0.1879 \* (normalized) PaperlessBilling=No

+ 0.21 \* (normalized) PaymentMethod=EC

+ -0.0355 \* (normalized) PaymentMethod=MC

+ -0.0527 \* (normalized) PaymentMethod=BT

+ -0.1218 \* (normalized) PaymentMethod=CC

+ -0.2967 \* (normalized) MonthlyCharges

+ -1.3484 \* (normalized) TotalCharges

- 0.0177

Number of kernel evaluations: 92177457 (43.544% cached)

Time taken to build model: 26.79 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances 5615 79.8493 %

Incorrectly Classified Instances 1417 20.1507 %

Kappa statistic 0.4544

Mean absolute error 0.2015

Root mean squared error 0.4489

Relative absolute error 51.626 %

Root relative squared error 101.6175 %

Total Number of Instances 7032

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class

0.893 0.463 0.842 0.893 0.867 0.458 0.715 0.830 No

0.537 0.107 0.645 0.537 0.586 0.458 0.715 0.470 Yes

Weighted Avg. 0.798 0.369 0.790 0.798 0.792 0.458 0.715 0.734

=== Confusion Matrix ===

a b <-- classified as

4612 551 | a = No

866 1003 | b = Yes

# Appendix D – R Code for Descriptive Statistics and Conversions

###########################################################################################

# Step 1: Loading & Cleaning the Data (MATT)

###########################################################################################

#Loading the data

proj=read.csv("WA\_Fn-UseC\_-Telco-Customer-Churn.csv")

View(proj)

#There are 11 missing entries based on the information in the next line.

sum(is.na(proj))

names(proj)

dim(proj)

head(proj)

###########################################################################################

#Step 1 (a) - Identifying the Missing Values

###########################################################################################

apply(is.na(proj), 2, any)

#from the above line, there are NA's in "TotalCharges"

names(which(sapply(proj, anyNA)))

# Get the count of columns with NA

sum(is.na(proj$TotalCharges))

###########################################################################################

#Step 1 (b) - Imputing the Missing Values

###########################################################################################

# This method was discussed, but not applied to the final dataset.

###########################################################################################

#Step 1 (c): Renaming Variables for Ease of Viewing (Done through EXCEL by Stephen)

###########################################################################################

str(CleanProj\_DeleteRows)

sum(is.na(CleanProj\_DeleteRows))

# These lines change the data from Longer Phrases, to abbreviations for the following variables

# MultipleLines, InternetService, OnlineSecurity, OnlineBackup, DeviceProtection,

# TechSupport, StreamingTV, StreamingMovies, Contract, PaymentMethod.

# KEY: "Nis"= "No Internet Service"

# "FO" = "Fiber Optic"

# "M-T-M" = "Month-to-Month"

# "OY" = "One year"

# "TY" = "Two year"

# "BT" = "Bank Transfer (automatic)"

# "MC" = "Mailed Check"

# "CC" = "Credit Card (automatic)"

# "EC" = "Electronic Check"

# "Nps" = "No phone service"

#str(CleanProj\_DeleteRows)

#library(plyr)

#levels(CleanProj$MultipleLines)

#levels(CleanProj$MultipleLines)=c("No","Nis","Yes")

#levels(CleanProj$MultipleLines)

#levels(CleanProj$PhoneService)

#levels(CleanProj$PhoneService)=c("No","Nps","Yes")

#levels(CleanProj$PhoneService)

#levels(CleanProj$InternetService)

#levels(CleanProj$InternetService)=c("DSL","FO","No")

#levels(CleanProj$InternetService)

#levels(CleanProj$OnlineSecurity)

#levels(CleanProj$OnlineSecurity)=c("No","Nis","Yes")

#levels(CleanProj$OnlineSecurity)

#levels(CleanProj$OnlineBackup)

#levels(CleanProj$OnlineBackup)=c("No","Nis","Yes")

#levels(CleanProj$OnlineBackup)

#levels(CleanProj$DeviceProtection)

#levels(CleanProj$DeviceProtection)=c("No", "Nis", "Yes")

#levels(CleanProj$DeviceProtection)

#levels(CleanProj$TechSupport)

#levels(CleanProj$TechSupport)=c("No", "Nis", "Yes")

#levels(CleanProj$TechSupport)

#levels(CleanProj$StreamingTV)

#levels(CleanProj$StreamingTV)=c("No", "Nis", "Yes")

#levels(CleanProj$StreamingTV)

#levels(CleanProj$StreamingMovies)

#levels(CleanProj$StreamingMovies)=c("No", "Nis", "Yes")

#levels(CleanProj$StreamingMovies)

#levels(CleanProj$Contract)

#levels(CleanProj$Contract)=c("M-T-M", "1yr", "2yr")

#levels(CleanProj$Contract)

#levels(CleanProj$PaymentMethod)

#levels(CleanProj$PaymentMethod)=c("BT", "CC", "EC", "MC")

#levels(CleanProj$PaymentMethod)

#str(CleanProj\_DeleteRows)

#sum(is.na(CleanProj\_DeleteRows))

#CleanProj1=CleanProj\_DeleteRows

###########################################################################################

#Step 1 (d): Bucketing several Variables.

###########################################################################################

# "Tenure" & "MonthlyCharges" are bucketed into "tenure1" & "MonthlyCharges1"

CleanProj1$tenure1=cut(x=CleanProj1$tenure, breaks=c(min(CleanProj1$tenure)-0.001,12,24,36,48,60,72))

levels(CleanProj1$tenure1)

levels(CleanProj1$tenure1)=c("0-12", "12-24", "24-36", "36-48", "48-60", "60-72")

levels(CleanProj1$tenure1)

# Checking to see if there are any missing values.

sum(is.na(CleanProj1))

apply(is.na(CleanProj1), 2, any)

levels(CleanProj1$MonthlyCharges)

CleanProj1$MonthlyCharges1=cut(x=CleanProj1$MonthlyCharges, breaks=c(0, 24, 48, 72, 96, 120))

levels(CleanProj1$MonthlyCharges1) = c("$0-$24", "$24-$48", "$48-$72", "$72-$96", "$96-$120")

levels(CleanProj1$MonthlyCharges1)

sum(is.na(CleanProj1))

###########################################################################################

#Step 2: Now for some visuals

CleanProj1=read.csv("CleanProj1.csv")

#If necessary: install.packages("ggplot2")

sum(is.na(CleanProj1))

apply(is.na(CleanProj1), 2, any)

library(ggplot2)

#Figure 2

dfgender=as.data.frame(table(CleanProj1$gender))

colnames(dfgender) = c("gender", "freq")

dfgender

dfgender$Percent = (dfgender$freq)/7032

dfgender

genderpie=ggplot(dfgender, aes(x="", y=freq, fill = factor(gender)))+

geom\_bar(width=1, stat="identity", colour ="gray45")+

theme(axis.text.x = element\_blank(),

axis.text.y = element\_blank(),

axis.line=element\_blank(),

plot.title=element\_text(hjust=0.5),text = element\_text(size=15))+

labs(fill="gender",

x=NULL,

y=NULL,

title="Pie Chart of Gender",

caption = "Source: www.kaggle.com - Telcom") +

theme(axis.text = element\_blank(),

axis.ticks = element\_blank(),

panel.grid = element\_blank()) +

coord\_polar(theta="y", start=0)

genderpie

#Figure 3

dfchurn=as.data.frame(table(CleanProj1$Churn))

colnames(dfchurn) = c("Churn", "freq")

dfchurn=cbind(dfchurn, (dfchurn$freq/7032))

colnames(dfchurn)=c("Churn", "Count", "Percent")

dfchurn

churnpie=ggplot(dfchurn, aes(x="", y=Count, fill = Churn))+

geom\_bar(width=.5, stat="identity", colour ="gray45")+

theme(axis.text.x = element\_blank(),

axis.text.y = element\_blank(),

axis.line=element\_blank(),text = element\_text(size=15),

plot.title=element\_text(hjust=0.5))+

labs(title="Pie Chart of Customer Churn",

subtitle="Customer Churn",

caption="Source: www.kaggle.com - Telcom Churn") +

theme(axis.text = element\_blank(),

axis.ticks = element\_blank(),

panel.grid = element\_blank()) +

coord\_polar(theta="y", start=0)

churnpie

#Figure 4

dfgenderchurn=as.data.frame(table(CleanProj1$Churn, CleanProj1$gender))

dfgenderchurn

colnames(dfchurn) = c("Churn", "gender", "freq")

dfgenderchurn

dfchurn=cbind(dfchurn, (dfchurn$freq/7032))

colnames(dfgenderchurn)=c("Churn", "Count", "freq")

dfgenderchurn

dfgenderchurn$Percent = (dfgenderchurn$freq/7032)

dfgenderchurn

genderChurn = ggplot(data=CleanProj1,aes(x=gender, y=Percentage,fill=Churn)) +

geom\_bar(mapping= aes(x=gender, fill=Churn, y=(..count..)/sum(..count..)),position="dodge",

colour ="gray45", width = 0.5) + scale\_y\_continuous (labels = scales::percent) +

theme(axis.text.x = element\_text(angle=0, vjust=0.6),text = element\_text(size=15)) +

labs(title="Categorywise Bar Chart",

subtitle="Gender Vs. Churn",

caption="Source: www.kaggle.com - Telcom Churn")

genderChurn

#Figure 5

dfcontractchurn=as.data.frame(table(CleanProj1$Churn, CleanProj1$Contract))

dfcontractchurn

colnames(dfcontractchurn) = c("Churn", "Contract", "freq")

dfcontractchurn

dfcontractchurn=cbind(dfcontractchurn, (dfcontractchurn$freq/7032))

colnames(dfcontractchurn)=c("Churn", "Count", "freq", "Percent")

dfcontractchurn

ContractChurn = ggplot(data=CleanProj1,aes(x=Contract, y=Percentage,fill=Churn)) +

geom\_bar(mapping= aes(x=Contract, fill=Churn, y=(..count..)/sum(..count..)),position="dodge",

colour ="gray45", width=0.5) +

theme(axis.text.x = element\_text(angle=0, vjust=0.6),text = element\_text(size=15)) +

scale\_y\_continuous (labels = scales::percent) +

labs(title="Categorywise Bar Chart",

subtitle="Contract Length Vs. Tenure",

caption="Source: www.kaggle.com - Telcom Churn")

ContractChurn

#Figure 6

dftenurechurn=as.data.frame(table(CleanProj1$Churn, CleanProj1$tenure1))

dftenurechurn

colnames(dftenurechurn) = c("Churn", "Tenure", "freq")

dftenurechurn

dftenurechurn=cbind(dftenurechurn, (dftenurechurn$freq/7032))

colnames(dftenurechurn)=c("Churn", "Tenure", "freq", "Percent")

dftenurechurn

tenureChurn = ggplot(CleanProj1, aes(tenure1)) +

geom\_bar(aes(fill=Churn), width = .5, position="dodge",colour ="gray45") +

theme(axis.text.x = element\_text(angle=0, vjust=0.6), text = element\_text(size = 15)) +

labs(title="Categorywise Bar Chart",

subtitle="Length of Tenure (in months)",

caption="Source: www.kaggle.com - Telcom Churn")

#Figure 7

dfischurn=as.data.frame(table(CleanProj1$Churn, CleanProj1$InternetService))

dfischurn

colnames(dfischurn) = c("Churn", "Int. Ser.", "freq")

dfischurn

dfischurn=cbind(dfischurn, (dfischurn$freq/7032))

colnames(dfischurn)=c("Churn", "Int. Ser.", "freq", "Percent")

dfischurn

Internet=ggplot(CleanProj1, aes(InternetService)) +

geom\_bar(aes(fill=Churn), width = 0.5, position="dodge", colour ="gray45") +

theme(axis.text.x = element\_text(angle=0, vjust=0.6), text = element\_text(size = 15)) +

labs(title="Categorywise Bar Chart",

subtitle="Internet Service",

caption="Source: www.kaggle.com - Telcom Churn")

#Figurre 8

dfTechSupport=as.data.frame(table(CleanProj1$TechSupport, CleanProj1$Churn))

dfTechSupport

colnames(dfTechSupport)=c("TechSupport", "Churn", "Freq")

Percent=(dfTechSupport$Freq/7043)

dfTechSupport=cbind(dfTechSupport, Percent)

dfTechSupport

TechSupport=ggplot(CleanProj1, aes(TechSupport)) +

geom\_bar(aes(fill=Churn), width = 0.5, position="dodge",colour ="gray45") +

theme(axis.text.x = element\_text(angle=0, vjust=0.6), text = element\_text(size = 15)) +

labs(title="Categorywise Bar Chart",

subtitle="Tech Support",

caption="Source: www.kaggle.com - Telcom Churn")

# Writing a PDF that has all of the plots used in the paper/presentation.

pdf(file = "Presentation\_Plots.pdf")

genderpie

churnpie

genderChurn

ContractChurn

tenureChurn

Internet

TechSupport

dev.off()

# References

1. <https://www.ngdata.com/what-is-customer-churn/> [↑](#endnote-ref-2)
2. <http://www.oracle.com/us/products/applications/cust-exp-impact-report-epss-1560493.pdf> [↑](#endnote-ref-3)
3. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.198.5133&rep=rep1&type=pdf [↑](#endnote-ref-4)