­

**Data Mining I: Churn Data**

**PA Task 2**

Stephen E. Porter

College of Information Technology, Western Governors University

Dr. Festus Elleh

March 22, 2022

**Data Mining: Churn Data**

In this paper, I will use a data set containing cleaned customer data from a fictional telecommunications company. The primary purpose is to perform statistical analysis on the cleaned data set to determine which factor (or factors) are the greatest indicator of customer turnover.

# Part I: Research Question

## A1. Proposal of Question

Can we build a model to accurately identify which customers have churned and predict which customers may be at risk of churning based on customers with similar characteristics?

## A2. Defined Goal

The primary goal of this analysis is to use *randomForest* to create a model that will provide insight into customer churn rate. This will enable company leadership to take more targeted action to address issues around customer churn. “Churn rates do correlate with lost revenue and increased acquisition spend.” (Altexsoft, 2020) By taking an active role in preventing customer churn, the company can have greater customer satisfaction/retention and greater revenue by retaining current customers while adding new ones.

# Part II: Method Justification

## B1. Explanation of Classification Method

Using a Random Forest is generally superior to a Decision Tree; as the name implies, a Random Forest is a collection of many Decision Trees offering a more global view of what the optimal outcome is for a question. When using a Decision Tree, an algorithm will make the “best” choice at each level but only from the perspective of that single choice. “Decision Trees are prone to overfitting, especially when the tree is particularly deep.” (Liberman, 2020) The data set used in this analysis has 65 predictor variables which, while not an extreme value, will likely lead to overfitting.

To avoid these disadvantages, a Random Forest analysis will be used. The standard *randomForest* function in *R* is 500 and the number of variables tried at each split is 6. By comparing 500 Decision Trees with fewer levels and taking the simple majority, the algorithm creates a better global view of the “best” outcome (in this case, correctly choosing the “Churn” value as “Yes” or “No”).

## B2. Summary of Method Assumption

There are not major assumptions for data distribution when using Random Forests. (Richmond, 2016) Basic assumptions are that the “best” choice will be made at each level and we will use a bootstrap sample instead of the entire data set. (Laptev, 2013)

## B3. Packages or Libraries List

The following libraries are used in this analysis:

|  |  |
| --- | --- |
| **Library** | **Justification** |
| **tidyverse** | Includes several useful tools, including the piping function %>% |
| **caret** | Contains the *createDataPartition* function to split the data into training and testing data sets and the *confusionMatrix* function to create a confusion matrix for the *randomForest* model |
| **randomForest** | Contains the *randomForest* function which is the main function of this analysis |

# Part III: Data Preparation

## C1. Data Processing

To answer the research question, the data must be tidied by removing irrelevant columns (such as ID numbers) and columns with too many unique entries (such as City and County). Character columns will need to be reassigned as factors in order for *randomForest* to process them. The data set will then be split into a training set and a testing set (70/30 split, respectively).

Once the data set has been tidied, split, and normalized, it is ready for analysis.

## C2. Data Set Variables

The table below outlines the types of variables used in this data set.

|  |  |  |
| --- | --- | --- |
| **Quantitative** | | **Qualitative** |
| **Discrete** | **Continuous** | **Categorical** |
| Population | Income | State |
| Children | Outage\_sec\_perweek | Area |
| Age | MonthlyCharge | Marital |
| Email | Bandwidth\_GB\_year | Gender |
| Contacts |  | Churn |
| Yearly\_equip\_failure |  | Techie |
| Tenure |  | Contract |
| Response |  | Port\_modem |
| Fix |  | Tablet |
| Replacement |  | InternetService |
| Reliability |  | Phone |
| Options |  | Multiple |
| Respectful |  | OnlineSecurity |
| Courteous |  | OnlineBackup |
| Listening |  | DeviceProtection |
|  |  | TechSupport |
|  |  | StreamingTV |
|  |  | StreamingMovies |
|  |  | PaymentMethod |

## C3. Steps for Analysis

To prepare the data, data frame will be reviewed for nulls and the column names will be renamed as needed.

Table

Description automatically generated

Columns named Item1 through Item8 will be given more meaningful names.

Table

Description automatically generated

Table

Description automatically generated

Other columns will be dropped as they will not provide useful information (such as CaseOrder and the various ID codes).

Text

Description automatically generated with low confidence

Character columns need to be converted to factors.



The data set will then be ready to be split into testing and training sets containing 70% and 30% of the data, respectively.

Text

Description automatically generated with low confidence

The data set is now ready for Random Forest analysis.

## C4. Cleaned Data Set

Please find copies of the prepared training and testing data sets included in the submission.

# Part IV: Analysis

## D1. Splitting the Data

The data set is split 70/30 as describe above to ensure there are plenty of data points for the model to learn from while there is still a sufficiently large set to test with.

## D2. Output and Intermediate Calculations

The *randomForest* function created 500 decision trees with 6 variables tried at each split. This resulted in an out-of-the-box estimate of error rate at 10.9%, meaning the model should correctly predict churn 89.1% of the time.

## D3. Code Execution

Please see the code file included with this submission.

# Part V: Data Summary and Implications

## E1. Accuracy and MSE

After applying *predict* on the training data set using the *rf* (random forest) model, the accuracy of *pred* given by the *confusionMatrix* function of the **caret** package is 0.8837 (88.37%) and can be calculated by taking the sum of correct predictions (True Positives + True Negatives) divided by the total number of rows. In this case, the calculation would be .

The mean square error between the actual value and predicted value is which implies the model is a good fit.

## E2. Results and Implications

This model can accurately determine whether or not a customer has churned 88.37% of the time. It can be used to analyze churn potential of current users and adapted to predict new users.

## E3. Limitation

While using a Random Forest over a Decision Tree reduces overfitting, it does not guarantee overfitting does not occur – especially if the data is noisy. “Random forests are biased in favor of those predictors with more levels.” (Richmond, 2016) There are several columns with many levels (such as “State” with 52 levels). It was left in to allow some consideration for geographic location having an effect on Churn.

## E4. Course of Action

The False Positives customers seem of particular interest as the model believes they have churned when they have not yet. These users should be identified and contacted to determine their satisfaction with their current services. This model can be used as new users are enrolled in services to periodically check for their risk of churning.

# Part VI: Demonstration

## F. Panapto Demonstration

Please view video included in the submission.

## G. Sources of Third-Party Code

N/A

## I. Sources

**References**

AltexSoft. (2020, February 27). Customer churn prediction for subscription businesses using Machine Learning: Main Approaches and Models. AltexSoft. Retrieved November 11, 2021, from <https://www.altexsoft.com/blog/business/customer-churn-prediction-for-subscription-businesses-using-machine-learning-main-approaches-and-models/>

Laptev, D. (2013, May 16). Random Forest Assumptions. Cross Validated. Retrieved March 22, 2022, from <https://stats.stackexchange.com/questions/59124/random-forest-assumptions>

Liberman, N. (2020, May 21). Decision trees and random forests. Medium. Retrieved March 22, 2022, from <https://towardsdatascience.com/decision-trees-and-random-forests-df0c3123f991>

Richmond, S. (2016, March 21) Algorithms exposed: Random Forest. BCCVL. Retrieved March 22, 2022, from <https://bccvl.org.au/algorithms-exposed-random-forest/#:~:text=ASSUMPTIONS,are%20ordinal%20or%20non%2Dordinal>.

Sharma, A. (2020, May 12). Decision Tree vs. Random Forest - which algorithm should you use? Analytics Vidhya. Retrieved March 22, 2022, from <https://www.analyticsvidhya.com/blog/2020/05/decision-tree-vs-random-forest-algorithm/>