­

**Data Mining I: Churn Data**

Stephen E. Porter

College of Information Technology, Western Governors University

Dr. Festus Elleh

March 20, 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 *k-*nearest neighbor (kNN) analysis 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

The core mechanism of kNN is “the value of a data point is determined by the data points around it.” (Yildirim, 2021) The *k-*value represents the number of closest neighboring points to check (with “closeness” determined by Euclidian distance).

***Euclidian Distance Definition:*** For two points , the distance between the points is determined by the square root of the sum of the squares of the change in coordinates and the change in coordinates, or

.

From the perspective of customer churn, if , then the closest single point is checked. The new point will be assigned a churn value that matches this closest point. If , then the five closest points are checked, and the value of the new point is assigned by the majority (ex: three have a churn value of “Yes” and two “No” then the new point will be assigned “Yes”).

To determine the accuracy of the model, a data set will be split into training and testing subsets

## B2. Summary of Method Assumption

A kNN analysis assumes that similar points with similar characteristics will be close together (Grant, 2019) and that the entries are independent.

## B3. Packages or Libraries List

The following libraries are used in this analysis:

|  |  |
| --- | --- |
| **Library** | **Justification** |
| **tidyverse** | Includes ggplot2 and tidyr for data visualizations and data tidying |
| **fastDummies** | Used to automatically create dummy columns for categorical variables |
| **class** | Contains the *knn* function which is at the core of this analysis |
| **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 kNN model |
| **pROC** | Contains the *auc* function which is used to determine the accuracy of the model |

# 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 to create dummy variables (such as City and County). Dummy variables will be created for the remaining categorical variables (such as Gender, Contract, and InternetService) and the new column headers will need to be tidied as well so they can be selected by name as part of the linear regression process. The data set will then be split into a training set and a testing set (70/30 split, respectively) and normalized to ensure large values (like Income) do not obscure the comparatively smaller values (like number of Children).

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

|  |  |
| --- | --- |
| **Quantitative** | |
| **Discrete** | **Continuous** |
| Population | Income |
| Children | Outage\_sec\_perweek |
| Age | Tenure |
| Email | MonthlyCharge |
| Contacts |  |
| Yearly\_equip\_failure |  |
| Bandwidth\_GB\_Year |  |
| Response |  |
| Fix |  |
| Replacement |  |
| Reliability |  |
| Options |  |
| Respectful |  |
| Courteous |  |
| Listening |  |

## C2. Data Set Variables

The target variable for this model is Churn (or Churn\_Yes once the dummy variable is created) and this will be compared against all remaining variables. Because dummy variables were created for categorical variables, the initial model has too many variables (118) to list in a reasonable manner, so instead the quantitative variables are listed specifically below. All other variables are considered categorical. See screenshots below for code output.

**Table

Description automatically generatedTable

Description automatically generatedTable

Description automatically generated with medium confidence**

## 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

Dummy variables will need to be created for categorical data, and those columns will need to be renamed to remove characters that would interfere with the code running successfully (like spaces, hyphens, and parenthesis).

Text, letter

Description automatically generated

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

A picture containing chart

Description automatically generated

The data set is now ready for kNN 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

To determine the first *k*-value to try for kNN, I found the square root of the number of rows in the training data set which returned . I then ran a kNN function on the training set for and . These returned nearly identical accuracy values

## D3. Code Execution

Please see the code file included with this submission.

# Part V: Data Summary and Implications

## E1. Accuracy and AUC

The accuracy of *pred43* given by the *confusionMatrix* function of the **caret** package is 0.9527 (95.27%) 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 AUC curve and corresponding AUC value of 0.905 (90.5%) confirm that the *pred43* model is performing very well.

## E2. Results and Implications

This model can accurately determine whether or not a customer has churned 95.27% of the time. It can be used to analyze churn potential of current users and adapted to predict new users.

## E3. Limitation

One limitation of my method is that I was not able to check every *k* value to determine which is truly the best. The process determined a model that performs well with a high level of accuracy, but there could be a *k* value that has a higher level of accuracy without over-fitting. A drawback of kNN is that the function uses the entire data set each time it is run so it can be very slow when running for many *k* values.

## 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

Learn by marketing. Learn by Marketing | Data Mining + Marketing in Plain English. (n.d.). Retrieved March 20, 2022, from <https://www.learnbymarketing.com/tutorials/k-nearest-neighbors-in-r-example/#:~:text=K-Nearest-Neighbors%20in%20R%20Example%20KNN%20calculates%20the%20distance,class%20%7D%20library%20and%20uses%20the%20knn%20function>

## 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/>

Grant, P. (2019, July 21). Introducing K-nearest neighbors. Medium. Retrieved March 20, 2022, from <https://towardsdatascience.com/introducing-k-nearest-neighbors-7bcd10f938c5>

Yıldırım, S. (2021, December 13). *K-Nearest Neighbors (kNN) — Explained - Towards Data Science*. Medium. Retrieved March 20, 2022, from <https://towardsdatascience.com/k-nearest-neighbors-knn-explained-cbc31849a7e3>