**An Analysis of Telecommunication Churn**

James P. Glenister

Master of Science Data Analytics, Western Governor’s University

C744: Data Mining and Analytics II

Prof. David Gagner

July 10, 2021

**An Analysis of Telecommunication Churn**

**Tool Selection**

**Analysis Tool**

The author decided to use Python for this data analysis. Python is currently the world’s fourth most popular programming language (Stack Overflow, 2020), in large part to its strong libraries for data science, pseudo-language like syntax, and its ease of deploying final models online. Since Python is free, it makes it a very popular choice among businesses that would like to have the strength of a dedicated data analysis tool like R with the functionality of application languages like Java or C++. As an added bonus, Python has very easy, built-in methods for reading and writing CSV data, which is important for this project.

**Goals of the Analysis**

There are two main goals for this analysis: generate a logistic model to predict customer churn and compile a collection of the important factors that lead to churn. By narrowing in on those factors, a company can focus upon the customers with a greater likelihood of churning to keep them loyal.

**Analysis Methods**

There are three main steps in our data analysis each with different methods to extract meaning. During data cleaning and EDA, we are primarily focused on the general shape of the data and whether the predictor variables are linearly independent or not. We therefore utilize methods such as utilizing the variable inflation factor, or VIF, to detect multicollinearity between the continuous variables. We can use (*insert chi-squared test that tests for independence here)* to detect if the categorical variables are independent.

Correspondence Analysis (CA) is utilized as the evaluative method. CA is used on categorical data to as an extension of Principle Component Analysis to analyze relationships between factors. While it cannot be used directly to make predictions, CA can be used to get a better sense of what behavior similar customers have to then direct marketing in that direction.

Logistic regression is used as the predictive method. Logistic regression is a simple but powerful model that can generate a model for a binary response variable. It works well for categorical and normalized continuous data.

**Data Exploration and Preparation**

**Target Variable**

The target variable for this analysis is binary value representing the customers that churn or not. The boxplot below shows the frequency of customers that don’t churn, represented by 0, to customers that do churn, represented by 1. For this data set, 26.54% of customers will churn.

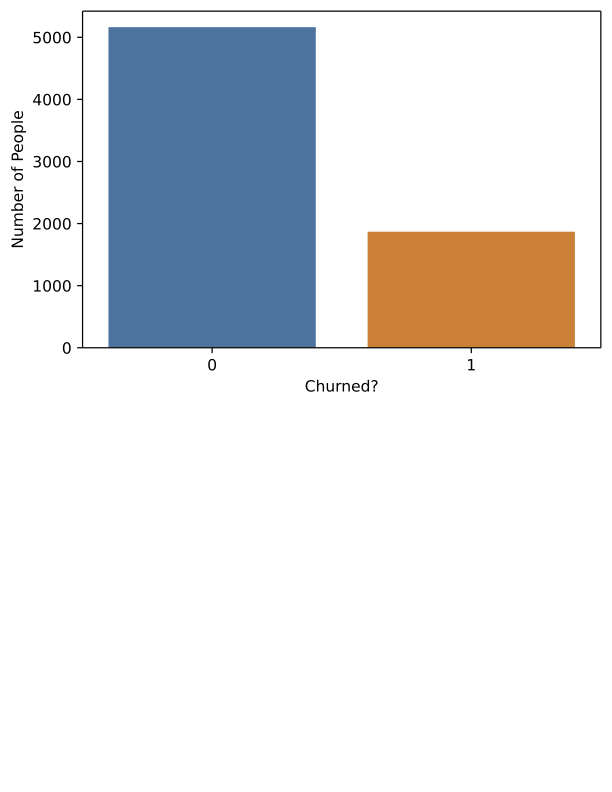


Figure 1: A frequency chart of customers that do not churn to customers that do.

**Predictor Variables**

While there will be further discussion on how these predictor variables were chosen in the sections on data manipulation goals and data cleaning, for now, it is enough to say that all the predictor variables are binary categorical variables representing if a customer does or does not have a particular characteristic. The predictor variables are:

* 'SeniorCitizen'
* 'gender\_Female’
* 'Partner\_Yes'
* 'Dependents\_Yes'
* 'PhoneService\_Yes'
* 'MultipleLines\_Yes'
* 'InternetService\_DSL'
* 'InternetService\_Fiber optic'
* 'InternetService\_No'
* 'OnlineSecurity\_Yes'
* 'OnlineBackup\_Yes'
* 'DeviceProtection\_Yes'
* 'TechSupport\_Yes'
* 'StreamingTV\_Yes'
* 'StreamingMovies\_Yes'
* 'Contract\_Month-to-month'
* 'Contract\_One year'
* 'Contract\_Two year'
* 'PaperlessBilling\_Yes'
* 'PaymentMethod\_Bank transfer (automatic)'
* 'PaymentMethod\_Credit card (automatic)’
* 'PaymentMethod\_Electronic check'
* 'PaymentMethod\_Mailed check'
* 'BinnedTenure\_(0-18]'
* 'BinnedTenure\_(18-36]'
* 'BinnedTenure\_(36-54]'
* 'BinnedTenure\_(54-72]'
* 'BinnedMonthlyCharges\_(0-23.75]'
* 'BinnedMonthlyCharges\_(23.75-47.50]'
* 'BinnedMonthlyCharges\_(47.50-71.25]'
* 'BinnedMonthlyCharges\_(71.25-95.00]'
* 'BinnedMonthlyCharges\_(95-118.75]'

**Data Manipulation Goals**

There are a few goals when manipulating this data. First, for the sake of this analysis, all variables need to be categorical data for the logistic regression model. All continuous variables (Tenure, Monthly Charges, and Total Charges) need to be binned to turn into categorical variables.

Most data was

**Statistical Identity**

**Chart, box and whisker chart

Description automatically generated**

**Data Cleaning**

**A screenshot of a computer

Description automatically generated with medium confidence**

**Data Analysis**

**Univariate Analysis**

**A picture containing PowerPoint

Description automatically generated**

**Bivariate Analysis**

**Chart, bar chart

Description automatically generated**

**Analytic Methods**

Analytic methods are employed to generate predictive models on the data.

Graphical user interface, text, application

Description automatically generated

Chart, bar chart, histogram

Description automatically generated

Chart, waterfall chart

Description automatically generated

**Evaluative Methods**

**Chart, scatter chart

Description automatically generated**

**Data Summary**

# References

Bendixen, M. (2003). *A Practical Guide to the Use of Correspondence Analysis in Marketing Research.* Retrieved from Marketing Bulletin 14: http://marketing-bulletin.massey.ac.nz/V14/MB\_V14\_T2\_Bendixen.pdf

Stack Overflow. (2020). *Stack Overflow Developer Survey 2020*. Retrieved from Stack Overflow: https://insights.stackoverflow.com/survey/2020#technology-programming-scripting-and-markup-languages-professional-developers