D209 Data Mining I: Performance Assessment

Task 2: Predictive Analysis

Jill Rivera

Western Governors University

Table of Contents

[I. Overview 3](#_Toc106955918)

[II. Method Justification 3](#_Toc106955919)

[III. Data Preparation 4](#_Toc106955920)

[IV. Analysis 11](#_Toc106955921)

[V. Data Summary and Implications 14](#_Toc106955922)

[VI. Scholarly References 16](#_Toc106955923)

[VII. Third-Party Code References 16](#_Toc106955924)

[VIII. Table of Figures 17](#_Toc106955925)

# **Overview**

1. A telecommunications company has gathered a data set on 10,000 customers to analyze the churn rate they were experiencing. This analysis will attempt to answer a business-related question using the data gathered and make recommendations based on the results.
   1. **Research Question**: The provider must determine the factors that are contributing to the customers churning form the company. This will answer the question of if the churn variable can be accurately predicted using limited predictor variables.
   2. **Objectives**: This analysis will use a predictive analysis method called a decision classification tree to accurately identify the limited features that are affecting churn.

# **Method Justification**

1. The predictive analysis is performed using a decision tree classification method. The assumptions and tools used for this analysis are the following:
   1. **Method**: A decision classification tree is a predictive modeling technique that graphically plots nodes as branches to represent outcomes in the shape of a tree. The decision or target node is at the top and all subsequent levels branch out with a set of attributes, or classification rules, associated with a class label. (Magee, 2022) These branches keep splitting of until it reaches an endpoint where no more outcomes are possible.
   2. **Assumption**: A decision classification tree analysis assumes that the target variable is categorical. This is necessary so that the predictive features can be classified into one of the values associated with the target. (Geeks, 2022) The target variable chosen for this analysis fits this assumption since churn is a categorical variable with yes/no values.
   3. **Tools**: This analysis is conducted using the open-source Jupyter Notebook platform with the Python3 programming language. This was chosen for the powerful computations, ease of use and the ability to markup and visualize the data in the same platform. Python has several libraries that will be used for this analysis. Pandas will be used to read, manipulate, and write new data frames. Numpy is a useful library for data frame computations. Matplotlib and Seaborn are helpful libraries that graph and chart the data into various forms. Scikit-learn has several libraries that will be used to split up the data into training and testing groups and perform the decision tree classification. Finally, PydotPlus is used to interface with the Graphviz language, and will be used to see the final decision tree.

# **Data Preparation**

1. The data set is cleaned, explored, and wrangled before the analysis can be performed.
   1. **Objective**: The data is prepared so that there are no missing or irrelevant data in the data set, so that the data is not distorted. The categorical features are converted so that the algorithm can perform numerical calculations to determine probability.
   2. **Variable Statistics**: The variables used for this analysis are displayed in the tables shown. The categorical features with their values are displayed on the first table, and the numerical features table shows the mean and range for each. The target variable for this analysis is the churn feature, which is categorical with yes and no values.

|  |  |
| --- | --- |
| **Categorical Variables** | |
| **Label** | **Value** |
| Area | Urban, Suburban, Rural |
| Marital | Widowed, Married, Separated, Never Married, Divorced |
| Gender | Male, Female, Nonbinary |
| Churn | No, Yes |
| Techie | No, Yes |
| Contract | One year, Month-to-Month, Two Year |
| Port\_modem | No, Yes |
| Tablet | No, Yes |
| InternetService | Fiber Optic, DSL, None |
| Phone | No, Yes |
| Multiple | No, Yes |
| OnlineSecurity | No, Yes |
| OnlineBackup | No, Yes |
| DeviceProtection | No, Yes |
| TechSupport | No, Yes |
| StreamingTV | No, Yes |
| StreamingMovies | No, Yes |
| PaperlessBilling | No, Yes |
| PaymentMethod | Credit Card (automatic), Bank Transfer (automatic), Mailed Check, Electronic Check |

Figure 1: Categorical Variables Table

|  |  |  |
| --- | --- | --- |
| **Numerical Variables** | | |
| **Label** | **Range** | **Mean** |
| Children | 0 - 10 | 2.09 +/- 2.15 |
| Age | 18 - 89 | 53.08 +/- 20.70 |
| Income | 348.67 – 258900.7 | 39806.93 +/- 28199.92 |
| Outage\_sec\_perweek | 0.09974694 – 21.20723 | 10.00 +/- 2.98 |
| Email | 1 - 23 | 12.02 +/- 3.03 |
| Contacts | 0 - 7 | 0.99 +/- 0.99 |
| Yearly\_equip\_failure | 0 - 6 | 0.40 +/- 0.64 |
| Tenure | 1.00025934 – 71.9928 | 34.53 +/- 26.44 |
| MonthlyCharge | 79.97886 – 290.160419 | 172.62 +/- 42.94 |
| Bandwidth\_GB\_Year | 155.5067148 – 7158.98153 | 3392.34 +/- 2185.29 |
| Item1 | 1 - 7 | 3.49 +/- 1.04 |
| Item2 | 1 - 7 | 3.51 +/- 1.03 |
| Item3 | 1 - 7 | 3.49 +/- 1.03 |
| Item4 | 1 - 7 | 3.50 +/- 1.03 |
| Item5 | 1 - 7 | 3.49 +/- 1.02 |
| Item6 | 1 - 7 | 3.50 +/- 1.03 |
| Item7 | 1 - 7 | 3.51 +/- 1.03 |
| Item8 | 1 - 7 | 3.50 +/- 1.03 |

Figure 2: Numerical Variables Table

* 1. **Data Cleaning**: The libraries that will be used are first imported.

Graphical user interface, text

Description automatically generated

Figure 3: Importing Standard Libraries

Text

Description automatically generated

Figure 4: Importing Scikit-Learn Packages

The data set is loaded using Pandas read\_csv method and the first five rows are displayed using the head method.

Graphical user interface, text

Description automatically generated

Figure 5: Loading the Data Frame

The variables and size of the data set is shown using the columns and shape methods. This data set has 50 variables and has 10,000 rows of data.

Text

Description automatically generated

Figure 6: Data Frame Structure

Irrelevant features and those with a high degree of cardinality are removed from the data frame using the drop method. This removes 13 features from the data frame, leaving 37 variables for analysis.



Figure 7: Removing Variables

The data frame is checked for null values using the isnull and sum methods. This data set has no missing values.

Table

Description automatically generated with low confidence

Figure 8: Checking for Null Values

The rows are checked for duplicated data using the duplicated and any methods. There are no duplicated rows.

Graphical user interface, text, application, chat or text message

Description automatically generated

Figure 9: Check for Duplicated Rows

The data is now checked for outlying values that could skew the results. This is done by calculating the z-score for each feature and removing the rows with outlying values. This has removed 709 rows from the data frame, leaving 9291 customer records for analysis.

Graphical user interface, text, application

Description automatically generated

Figure 10: Removing Outlying Values

Finally, the categorical features are converted into individual columns for each value.

Text

Description automatically generated

Figure 11: Categorical Feature Conversion

Text

Description automatically generated

Figure 12: Categorical Features Post Conversion

* 1. **Extract**: A copy of the cleaned data set is extracted using Pandas to\_csv method. This is labeled as 209clean\_task2.csv and attached for submission of this assessment. **A picture containing text

     Description automatically generated**

Figure 13: Exporting Clean Data

# **Analysis**

1. The data frame is first split into training and testing groups of data. Following the split, the classification analysis can be performed and analyzed.
   1. **Train/Test Split**: The data needs to be split into raining and testing groups for the data to be trained on one set of data and then tested on a clean set to ensure accuracy. First, the target variable, along with the predictive variables are defined using the loc method.

Text

Description automatically generated

Figure 14: Define Target and Predictive Features

The data is split into training and testing groups consisting of 30% of the data reserved for training and 70% for testing. Scikit-learn’s train\_test\_split function is passed the decimal value of the groups desired and is easily split for the model.

Text, letter

Description automatically generated

Figure 15: Splitting the Data (Pedregosa, Varoquaux, & Gramfort, sklearn.model\_selection.train\_test\_split, 2022)

These groups of data are exported to CSV files using Pandas to\_csv method for submission to this assessment.

Graphical user interface, text, application

Description automatically generated

Figure 16: Exporting Split Data Frames

* 1. **Analysis Technique**: The DecisionTreeClassifier is a package from Scikit-learn that has all the functions for creating decision trees built into it. The model is created and fit to the training data, while specifying the maximum depth of two to limit the tree’s complexity.

**Graphical user interface, text, application

Description automatically generated**

Figure 17: Model Creation (Pedregosa, Varoquaux, & Gramfort, sklearn.tree.DecisionTreeClassifier, 2022)

Using PyplotPlus and Graphviz, the decision tree is visualized. The features that the tree used to predict churn are monthly charge and tenure.

Diagram

Description automatically generated

Figure 18: Decision Tree

The training data is graphed in a scatter plot for a better visualization of the display pattern for the variables.

Text

Description automatically generatedChart, scatter chart

Description automatically generated

Figure 19: Training Data Scatter Plot (Waskom, 2021)

# **Data Summary and Implications**

1. Results of the predictive classification analysis, accuracy of the model, limitations to the analysis method and recommendations for the business are as follows:
   1. **Accuracy**: The mean squared error (MSE), or the average squared difference between the estimated and actual values, is used to test the accuracy of the model. A confusion matrix is created for the training and testing data sets to obtain this calculation. The training data set is testing at an accuracy rate of 83.5%. This is calculated by dividing the total number of true positives and true negatives by the total number of predictions in the matrix. The confusion matrix shows the number of correctly and incorrectly predicted data points in the model. This training set included 5431 accurate predictions and 1072 inaccurate predictions. This can be calculated by adding up the diagonals of the matrix.

Graphical user interface, text, application

Description automatically generated

Figure 20: Confusion Matrix on Training Data

The confusion matrix on the testing data reflects a similar accuracy score of 82.5%. This data set resulted in 2300 accurate predictions and 488 inaccurate predictions.

Text

Description automatically generated with low confidence

Figure 21: Confusion Matrix for Test Data

* 1. **Results**: The decision tree resulted in the two features of monthly charge and tenure able to predict churn with an accuracy of 83%. The tree is easy to read and can be expanded for further analysis.
  2. **Limitations**: If a categorical variable is not available for the target variable, this method will not work. The analysis must classify the predictor nodes or features into a value for the tree to be created. Incorrect variable selection or skipping the data wrangling step will result in incorrect analysis.
  3. **Recommendations**: It is recommended that the company use the model created to periodically review customers and proactively offer additional services or discounts if deemed to be churn positive. This can be done on a quarterly basis and can reduce the churn rate on a proactive basis instead of a reactive one.

# **Scholarly References**

Geeks, G. f. (2022, June 17). *Decision Tree*. Retrieved from Geeks for Geeks: https://www.geeksforgeeks.org/decision-tree/

Magee, J. F. (2022, June 22). *Decision Trees for Decision Making*. Retrieved from Harvard Business Review: https://hbr.org/1964/07/decision-trees-for-decision-making

# **Third-Party Code References**

Pedregosa, F., Varoquaux, G., & Gramfort, A. (2022, June 22). *sklearn.model\_selection.train\_test\_split*. Retrieved from Scikit Learn: https://scikit-learn.org/stable/modules/generated/sklearn.model\_selection.train\_test\_split.html

Pedregosa, F., Varoquaux, G., & Gramfort, A. (2022, June 22). *sklearn.tree.DecisionTreeClassifier*. Retrieved from Scikit Learn: https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html

Waskom, M. L. (2021). *seaborn.scatterplot*. Retrieved from Seaborn: https://seaborn.pydata.org/generated/seaborn.scatterplot.html

# **Table of Figures**

[Figure 1: Categorical Variables Table 5](#_Toc106955897)

[Figure 2: Numerical Variables Table 6](#_Toc106955898)

[Figure 3: Importing Standard Libraries 6](#_Toc106955899)

[Figure 4: Importing Scikit-Learn Packages 6](#_Toc106955900)

[Figure 5: Loading the Data Frame 7](#_Toc106955901)

[Figure 6: Data Frame Structure 7](#_Toc106955902)

[Figure 7: Removing Variables 8](#_Toc106955903)

[Figure 8: Checking for Null Values 9](#_Toc106955904)

[Figure 9: Check for Duplicated Rows 9](#_Toc106955905)

[Figure 10: Removing Outlying Values 10](#_Toc106955906)

[Figure 11: Categorical Feature Conversion 10](#_Toc106955907)

[Figure 12: Categorical Features Post Conversion 11](#_Toc106955908)

[Figure 13: Exporting Clean Data 11](#_Toc106955909)

[Figure 14: Define Target and Predictive Features 12](#_Toc106955910)

[Figure 15: Splitting the Data (Pedregosa, Varoquaux, & Gramfort, sklearn.model\_selection.train\_test\_split, 2022) 12](#_Toc106955911)

[Figure 16: Exporting Split Data Frames 13](#_Toc106955912)

[Figure 17: Model Creation (Pedregosa, Varoquaux, & Gramfort, sklearn.tree.DecisionTreeClassifier, 2022) 13](#_Toc106955913)

[Figure 18: Decision Tree 14](#_Toc106955914)

[Figure 19: Training Data Scatter Plot (Waskom, 2021) 15](#_Toc106955915)

[Figure 20: Confusion Matrix on Training Data 16](#_Toc106955916)

[Figure 21: Confusion Matrix for Test Data 16](#_Toc106955917)