**D209 PA**

Jon Isak

D209: Data Mining

Contents

[**D209 3**](#_3xni6yjcaxld)

[**Part I: Research Question 3**](#_v3v5jdbuxjo5)

[**Part II: Method Justification 3**](#_1ty9oqwbr9xq)

[**Part III: Data Preparation 4**](#_edq5yp60pzx9)

[**Part IV: Analysis 4**](#_hu3fuma5y7h4)

[**Part V: Data Summary and Implications 5**](#_184xb8dznw1a)

[**Part VI: Demonstration 6**](#_9d0i2u1cemko)

## 

## D209

## Part I: Research Question

A1. This data analysis will look into the question, can we predict if a customer will churn based on several factors? Through this analysis, the company will identify customers who are likely to churn and contact them with retention strategies. I plan to use the random forests model to answer the question.

A2. The goal of this analysis is to identify the key factors that influence customer churn. The likelihood of churning can be predicted by these factors.

## 

## Part II: Method Justification

B1. The random forest model should construct multiple decision trees as part of the training phase and output the mode of the classes or mean prediction of the individual trees. Each tree should make an independent churn prediction, and the outcomes are aggregated through a voting or averaging process. The expected outcome is correct predictions of churn. It is also supposed to identify the most important factors that contribute to churn (Guide, 2023).

B2. One assumption of random forest is that random decision trees are independent of each other (Guide, 2023).

B3. The data analysis will be done using Python within Jupyter Notebook. I will use Pandas for data manipulation, NumPy for numerical operations, Scikit-learn for the random forest operations.

## Part III: Data Preparation

C1. One data preprocessing goal is label encoding. In this step, I will apply a number to each category. This makes the data compatible with random forests.

C2. My numeric variables will be 'Outage\_sec\_perweek,' 'Email,' 'Contacts,' 'Yearly\_equip\_failure,' 'Tenure,' 'MonthlyCharge,' 'Bandwidth\_GB\_Year,' 'Population,' 'Age,' 'Income'. My categorical variables will be 'Marital,' 'Gender,' 'Churn' 'Port\_modem,' 'Tablet,' 'InternetService,' 'Phone,' 'Multiple,' 'OnlineSecurity,' 'OnlineBackup,' 'DeviceProtection,' 'TechSupport,' 'StreamingTV,' 'StreamingMovies,' 'PaperlessBilling,' 'PaymentMethod,' and 'Techie’.

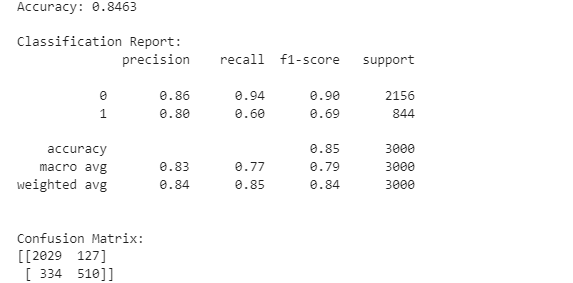
C3. To prepare the data for the analysis, I will use a label encoder, scale numeric variables, and split data into training and test. Churn and other categorical variables will be transformed from yes or no to 1 or 0. The numerical variables like monthly charge will be standardized to be on a similar scale.

C4. The file is attached.

## Part IV: Analysis

D1. The files are attached.

D2. Scaling was applied to numerical data, and categorical variables were encoded. The data was then divided into training and testing sets. The random forest technique was then used to predict churn.



D3. The code is included with the attached file.

## Part V: Data Summary and Implications

E1. The MSE was .1537, indicating the average squared difference between the predicted probabilities and the actual values.

E2. The accuracy of the model was .8463, indicating that the model makes correct predictions 84.63% of the time. The high accuracy indicates that the model was a good fit. For no churn, the precision indicated that 86% of the predictions were correct. For churn, percussion was 80%. For no churn, recall indicates that model identified 94% of no churn instances. For churn, recall was 60%. For no churn, the F1-score of .90 indices a balance between percussion and recall. For churn the F1-score was .69 indicating a lower performance compared to no churn. In the confusion matrix, there were 510 true positives for churn, 127 false positives, 2029 true negatives, and 334 false negatives.

E3. The limitations of the data analysis include the data imbalance in the data. No churn is much more prevalent than churn. This limits the performance of the model, as indicated by the lower F1 and recall score for churn compared to no churn.

E4. Based on the results, there are many actions the company can take. The company can address the imbalance in the data by continuing to collect more data. The company can use the model to predict churn with high accuracy. The company can also use the model to identify customers who are likely to churn and reach out to them with personalized retention offers.

## Part VI: Demonstration

F. https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=011159ef-af78-4812-82c9-b0c4014b9c96

G.

Scikit. (n.d.). *sklearn.ensemble.RandomForestClassifier — scikit-learn 1.3.2 documentation*. Scikit-learn. Retrieved November 24, 2023, from https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

Scikit. (n.d.). *sklearn.ensemble.RandomForestClassifier — scikit-learn 1.3.2 documentation*. Scikit-learn. Retrieved November 24, 2023, from https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

H.

Gandhi, R. (2018, May 5). *Naive Bayes Classifier. What is a classifier? | by Rohith Gandhi*. Towards Data Science. Retrieved November 19, 2023, from https://towardsdatascience.com/naive-bayes-classifier-81d512f50a7c