**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? This analysis will help the company be more proactive by identifying customers who are likely to churn and reaching out to them with retention strategies. I plan on using the Native Bayes method to answer the question.

A2. The goal of this analysis is to identify key factors that affect customers' churn rates. These factors can be used to predict likelihood of churn.

## 

## Part II: Method Justification

B1. The Naive Bayes classifier will predict whether a customer is likely to churn or not based on the provided data. It should output the probabilities for each customer, indicating the likelihood of a churn. By analyzing the probabilities and predicted labels against the actual churn status, the model's accuracy in predicting customer churn can be evaluated. The model will provide insights into which variables are most influential in predicting churn.

B2. One assumption of Naive Bayes is that all variables are independent of each other. The churn data meets this assumption (Gandhi, 2018).

B3. For the data analysis, I will use Python within Jupyter Notebook, leveraging several essential libraries and packages. Pandas will facilitate data manipulation tasks. NumPy will assist in handling numerical operations. For the classifier and scaling, I'll primarily utilize scikit. Visualizations will be created using Matplotlib to gain insights from the data and model performance.

## Part III: Data Preparation

C1. One data preprocessing goal that is important to Naive Bayes is scaling. In this step, I will apply scaling to numerical variables. This makes the data more usable.

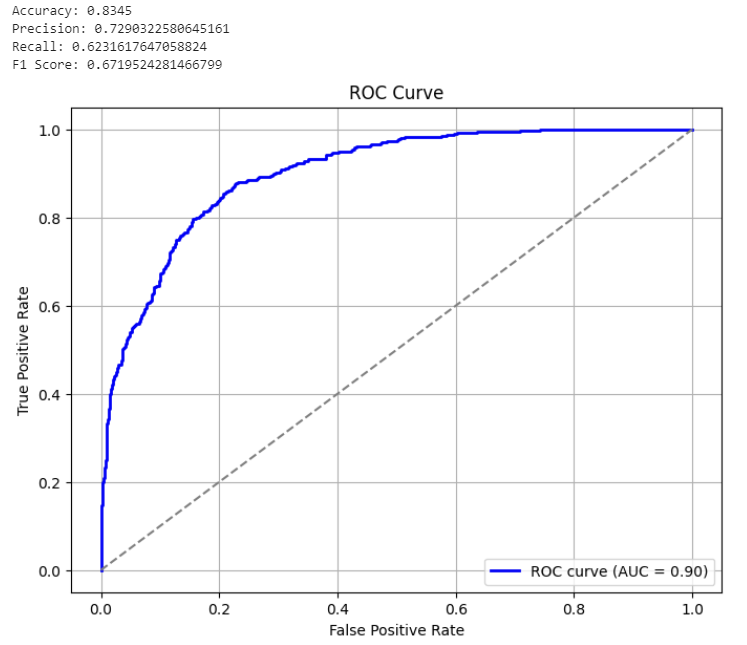
C2. My numeric variables will be ‘Tenure’ and ‘MonthlyCharge’. My categorical variable will be ‘Churn’.

C3. To prepare the data for the analysis, I will use a label encoder and a standard scaler. Churn will be transformed from yes or no to 1 or 0. The numerical variables of tenure and 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. The numerical variables were scaled, and categorical variables were encoded. The data was then split into training and test data. Then the Naive Bayes classifier was then used for classification. 

D3. The file is attached.

## Part V: Data Summary and Implications

E1. The model's accuracy was .8345. This suggests that 83.45% of the predictions made by the model were correct. A precision of .7290 suggests churn was predicted 72.90% of the time. A recall of .6232 suggests that the model identified 62.32% of churn occurrences. An F1 score of .6720 suggests that the model had a good balance of precision and recall. The AUC was .90. This suggests that the model performed well in distinguishing between churn and non-churn. Overall, the model performs well in predicting churn based on tenure and monthly charge (Deepchecks, 2023).

E2. Based on the model’s accuracy and AUC, it can be concluded that tenure and monthly charges have a significant impact on churn. Furthermore, the model's good performance means that it would be useful for predicting churn in a business environment and decision-making.

E3. The limitations of the data analysis include the simplicity of the model, which only takes tenure and monthly charges into account as predictive factors. It’s possible that there are other factors that influence customer behavior and lead to churn. The predictive power may be limited by the exclusion of these factors.

E4. Utilizing this predictive model enables businesses to identify customers at risk of churn accurately. The company can implement targeted retention strategies aimed at reducing churn rates. These strategies may involve proactive initiatives such as personalized offers, incentives, or service quality check-ins tailored to meet the specific needs of at-risk customers. Furthermore, continuous monitoring and evaluation of the model's accuracy are essential. Customer behaviors and preferences evolve over time, and staying updated on these changes ensures that retention strategies remain effective and aligned with shifting customer dynamics.

## Part VI: Demonstration

F. https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=005d9f1b-3f14-44cb-9e7f-b0c1007671b2

G.

Scikit. (n.d.). *sklearn.metrics.roc\_curve — scikit-learn 1.3.2 documentation*. Scikit-learn. Retrieved November 20, 2023, from https://scikit-learn.org/stable/modules/generated/sklearn.metrics.roc\_curve.html

Scikit. (n.d.). *sklearn.model\_selection.train\_test\_split — scikit-learn 1.3.2 documentation*. Scikit-learn. Retrieved November 20, 2023, from https://scikit-learn.org/stable/modules/generated/sklearn.model\_selection.train\_test\_split.html

H.

Deepchecks. (2023, April 4). *Understanding F1 Score, Accuracy, ROC-AUC, and PR-AUC Metrics for Models*. Deepchecks. Retrieved November 20, 2023, from https://deepchecks.com/f1-score-accuracy-roc-auc-and-pr-auc-metrics-for-models/

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