**Data mining causes of internet churn**

**D209 Data mining 1**

**By Josue Gonzalez**

Table of Contents

[Part I: Research Question 9](#_Toc172837914)

[A1 State your research question 9](#_Toc172837915)

[A2 State Objectives and Goals for Analysis. 9](#_Toc172837916)

[Part II: Method Justification 10](#_Toc172837917)

[B1 Assumptions 10](#_Toc172837918)

[B2 Programming Language and benefits 10](#_Toc172837919)

[B3 Justification of using Regression 10](#_Toc172837920)

[Part III: Data Preparation and Manipulation (Cleaning → Exploration → Wrangling) 11](#_Toc172837921)

[C1 Data Cleaning 11](#_Toc172837922)

[C2. Data Exploration (EDA). 11](#_Toc172837923)

[C3. Visualizations. 13](#_Toc172837924)

[C4. Data Transformation(Data Wrangling) 16](#_Toc172837929)

[C5. Provide the prepared data set as a CSV file. 19](#_Toc172837930)

[Part IV: Model Comparison and Analysis 20](#_Toc172837932)

[D1.Initial model 20](#_Toc172837933)

[D2.Justification for Model Reduction 21](#_Toc172837934)

[D3. Reduced model 24](#_Toc172837938)

[E1.Model Comparison 25](#_Toc172837940)

[E2.Reduced model 26](#_Toc172837941)

[E2.Residuals 27](#_Toc172837942)

[Part V: Data Summary and Implications 28](#_Toc172837943)

[F1. Regression Equation for the Reduced Model 28](#_Toc172837944)

[c) Discussion on Statistical and Practical Significance 29](#_Toc172837945)

[Disadvantages and Implications of the Methods Used 30](#_Toc172837946)

[Part VI: Demonstration 31](#_Toc172837947)

[Sources 31](#_Toc172837948)

[References 31](#_Toc172837949)

# Part I: Research Question

A1 State your research question. How can we predict whether a customer will churn based on their usage patterns and demographic information. We will be using Naïve Bayes for this assignment.

A2 State Objectives and Goals for Analysis. The goal of this data analysis is to develop a predictive model that accurately identifies customers at high risk of churning. This model will help the organization implement targeted retention strategies, thereby reducing customer turnover and increasing customer lifetime value. The analysis will leverage customer usage data, such as call duration, internet usage, and service complaints, along with demographic information like age, location, and income level.

## Part II: Method Justification

### B1. Explanation of the Classification Method

Naive Bayes is chosen for this analysis due to its efficiency and suitability for large datasets with both continuous and categorical features. Naive Bayes classifiers work by applying Bayes' theorem with strong (naive) independence assumptions between the features. The model calculates the probability of a customer churning based on the likelihood of each feature given the class (churn or not churn). For each customer, the model computes the posterior probability for each class and assigns the class with the highest probability. Expected outcomes include a clear probabilistic model that can be used to predict churn with reasonable accuracy. This method is particularly effective for initial screening and identifying high-risk customers quickly (Zhang, 2004).

### B2. Assumption of Naive Bayes

One key assumption of Naive Bayes is the conditional independence assumption, which posits that the features are independent given the class label. This means that the presence or absence of a particular feature is assumed to be unrelated to the presence or absence of any other feature, given the class. While this assumption is rarely true in real-world scenarios, Naive Bayes tends to perform well even when the independence assumption is violated to some extent (Rish, 2001). This robustness makes it a practical choice for many applications, including customer churn prediction.

### B3. Packages or Libraries for Python

We will be using, scikit-learn: This library provides efficient implementations of various machine learning algorithms, including Naive Bayes classifiers. It is well-documented and widely used, making it a reliable choice for building and evaluating the model. Scikit-learn's implementation of Naive Bayes is straightforward to use and integrates well with other preprocessing and evaluation tools within the library (Pedregosa et al., 2011). Pandas:This library is essential for data manipulation and analysis. It allows for easy handling of dataframes, which is useful for cleaning and preparing the dataset. Pandas supports efficient data manipulation operations, which are crucial for preparing the dataset before applying the Naive Bayesclassifier (McKinney, 2010). numpy: This library provides support for large, multi-dimensional arrays and matrices, along with a collection of mathematical functions to operate on these arrays.Numpy is used for numerical computations, which are necessary for data processing and feature extraction (Van Der Walt, Colbert, & Varoquaux, 2011). matplotlib and seaborn: These libraries are used for data visualization, helping to explore the dataset and understand the distributions and relationships between features.Visualization is key to understanding the data and presenting the results of the analysis effectively (Hunter, 2007).

By leveraging these libraries, we can efficiently preprocess the data, build the Naive Bayes model, and visualize the results, ensuring a comprehensive approach to predicting customer churn.

# Part III: Data Preparation and Manipulation (Cleaning → Exploration → Wrangling)

### C1. Data Preprocessing Goal

One data preprocessing goal relevant to the Naive Bayes classification method is to ensure that all features are in an appropriate format for the model. This involves handling missing values and encoding categorical variables properly. Unlike some machine learning algorithms, Naive Bayes does not require scaling or normalizing numerical variables, as the model is based on the probability distribution of each feature rather than its magnitude. The preprocessing goal is to transform the raw data into a clean and structured format that the Naive Bayes algorithm can effectively process, ensuring optimal performance and accuracy in the churn prediction model.

### C2. Initial Data Set Variables

The initial dataset variables to be used for the analysis our categorical variables are Gender, Marital, Children, Phone, Multiple, InternetService, OnlineSecurity, OnlineBackup, Device Protection, TechSupport, StreamingTV, Contract, PaperlessBilling, PaymentMethod, Our target variable Churn. Our Numerical variables are: Monthly Charges, Tenure, Age

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

Next will check for null and duplicates, here we see we have no nulls or duplicates if we did then we would impute the values and remove any duplicates if that was the case.

A screenshot of a computer

Description automatically generated

Next will get the count values of our categorical variables and numerical before encoding the data to set it up to be used in the model.

A screenshot of a computer

Description automatically generated

A screenshot of a computer program

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

### C3. Provide the prepared data set as a CSV file.

See submission attached files final out into the CSV will as follows

# Part IV: Analysis

### D1.Split data A screenshot of a computer Description automatically generated

### D2.Analysis technique

The Naive Bayes classification algorithm was chosen for this analysis due to its simplicity, efficiency, and effectiveness with large datasets. It is particularly well-suited for binary classification problems like predicting customer churn, where the goal is to classify customers into one of two categories: "Churn" or "Not Churn." Naive Bayes works well with categorical data, which makes it a good fit given that many of the features in the dataset (e.g., gender, internet service type, contract type) are categorical.

A screenshot of a computer

Description automatically generated A screenshot of a computer

Description automatically generated

Naive Bayes is a probabilistic classifier based on Bayes' theorem, which describes the probability of an event occurring given prior knowledge of conditions related to the event. The "naive" part of Naive Bayes comes from the assumption that all features (predictors) are independent of each other given the class label, which simplifies the computation.

A screenshot of a computer program

Description automatically generated

A lot of the prep work was done in part III

### D3. Code

### See code in the attached files

# Part V: Data Summary and Implications

### E1. Accuracy and AUC curve A screenshot of a computer program Description automatically generated

### A screen shot of a graph Description automatically generated

### E2. Results and Implications of Your Classification Analysis

* **Results:** Based on the accuracy and AUC, you can determine how well the Naive Bayes model performed. For instance, if the accuracy is 0.85 and the AUC is 0.92, it suggests that the model correctly predicts 85% of the cases, and the AUC indicates a high level of discrimination between customers who churn and those who do not.
* **Implications:** These results imply that the Naive Bayes model is a reliable tool for predicting customer churn. The organization can use this model to identify at-risk customers and implement retention strategies. The high AUC suggests that the model is particularly effective in distinguishing between customers who are likely to churn and those who are not, which can be valuable for targeted marketing and personalized interventions.

### E3. Discuss One Limitation of Your Data Analysis

* **Limitation:** One limitation of the Naive Bayes model is its assumption of feature independence, which may not hold true in real-world data. This can lead to suboptimal performance if the features are highly correlated. Additionally, the model's performance may be sensitive to the quality and completeness of the data. If the dataset contains noise or missing values that are not handled properly, the model's accuracy and AUC could be negatively impacted.

### E4. Recommend a Course of Action for the Real-World Organizational Situation

* **Recommendation:** Based on the analysis, I recommend that the organization use the Naive Bayes model as part of a broader customer retention strategy. The model can be integrated into the organization's CRM system to flag customers who are at high risk of churning. By identifying these customers early, the organization can deploy targeted retention campaigns, such as personalized offers, discounts, or enhanced customer service. Additionally, the organization should continuously monitor the model's performance and update it with new data to maintain its accuracy and effectiveness.

# Part VI: Demonstration

**See Code in video provided**

# **<https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=b006c262-4e49-44f8-abc9-b1cf00856d59>**

# Sources

*1.9. naive Bayes. scikit. (n.d.-a). https://scikit-learn.org/stable/modules/naive\_bayes.html*

*Labelencoder. scikit. (n.d.-c). https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.LabelEncoder.html*

Moffitt, C. (2017, February 6). *Guide to encoding categorical values in python*. Practical Business Python Atom. https://pbpython.com/categorical-encoding.html

*Roc\_auc\_score. scikit. (n.d.-d). https://scikit-learn.org/stable/modules/generated/sklearn.metrics.roc\_auc\_score.html*

# References

Bhandari, A. (2024, August 6). *Guide to AUC ROC curve in machine learning : What is specificity?*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2020/06/auc-roc-curve-machine-learning/

Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. Computing in Science & Engineering, 9(3), 90-95.

McKinney, W. (2010). Data Structures for Statistical Computing in Python. Proceedings of the 9th Python in Science Conference, 51-56.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825-2830.

Rish, I. (2001). An Empirical Study of the Naive Bayes Classifier. IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence.

Van Der Walt, S., Colbert, S. C., & Varoquaux, G. (2011). The NumPy Array: A Structure for Efficient Numerical Computation. Computing in Science & Engineering, 13(2), 22-30.

Zhang, H. (2004). The Optimality of Naive Bayes. FLAIRS Conference.