**Causes of Internet Churn PA2**

**D209 Exploratory Data Analysis:**

**By Josue Gonzalez**

Table of Contents

[Part I: Research Question 3](#_Toc174998302)

[A1 State your research question. 3](#_Toc174998303)

[A2 State Objectives and Goals for Analysis. 3](#_Toc174998304)

[Part II: Method Justification 4](#_Toc174998305)

[B1 Explanation of the Chosen Prediction Method: Random Forest 4](#_Toc174998306)

[B2. Assumption of the Chosen Prediction Method 4](#_Toc174998307)

[B3. Programming Language and benefits. 4](#_Toc174998308)

[Part III: Data Preparation and Manipulation (Cleaning → Exploration → Wrangling) 5](#_Toc174998309)

[C1. Describe One Data Preprocessing Goal Relevant to the Prediction Method 5](#_Toc174998310)

[C2. Identify the Initial Data Set Variables for Analysis 5](#_Toc174998311)

[The analysis utilized a combination of numerical and categorical variables to predict customer. 5](#_Toc174998312)

[C3. Explain the Steps Used to Prepare the Data for the Analysis 6](#_Toc174998313)

[C4. Provide the prepared data set as a CSV file. 8](#_Toc174998314)

[Part IV: Model Comparison and Analysis 9](#_Toc174998315)

[Part V: Data Summary and Implications 11](#_Toc174998316)

[E1. Explain the Accuracy and Mean Squared Error (MSE) of Your Prediction Model 11](#_Toc174998317)

[E2. Discuss the results and implications of your prediction analysis 12](#_Toc174998318)

[**https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=e2f3ed8d-a23a-4ca7-ad8b-b1cf00839c79** 13](#_Toc174998319)

[Sources 13](#_Toc174998320)

[References 14](#_Toc174998321)

# 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? using a Random Forest model?

Random Forests are a powerful ensemble learning method that combines multiple decision trees to improve predictive accuracy and reduce the risk of overfitting. By leveraging this method, we aim to create a robust model that can handle the complexity and variability of customer data, ultimately providing the organization with reliable predictions of customer churn.

A2 State Objectives and Goals for Analysis. The objective of this data analysis is to develop a predictive model using the Random Forest technique that accurately identifies customers at high risk of churning. The primary goal is to enable the organization to proactively implement targeted retention strategies based on the insights derived from this model, thereby reducing customer turnover and increasing customer lifetime value.

This analysis leverages a comprehensive set of customer data, including service usage patterns such as monthly charges and bandwidth consumption, as well as customer interaction history (e.g., the number of responses and solutions provided). Additionally, key customer satisfaction metrics (e.g., reliability, respectfulness) and demographic information (e.g., age, income, education, employment) are incorporated to enhance the predictive power of the model. The end goal is to create a robust and interpretable model that provides actionable insights for improving customer retention.

## Part II: Method Justification

### B1 Explanation of the Chosen Prediction Method: Random Forest

The Random Forest method is a powerful ensemble learning technique that enhances prediction accuracy and stability by constructing multiple decision trees and aggregating their outputs. In this analysis, the dataset is first split into multiple subsets using bootstrapping, a technique where random samples of the data are drawn with replacement. Each subset is used to grow a separate decision tree, which allows the model to capture various patterns within the data. During the construction of each tree, the algorithm randomly selects a subset of features at each split to determine the best division, ensuring diversity among the trees and reducing the risk of overfitting. Once all the trees are built, they are used to predict whether each customer will churn. The final prediction for each customer is determined by taking a majority vote among all the trees, which typically results in higher accuracy and robustness compared to a single decision tree.

The expected outcome of using the Random Forest model is a highly accurate classification of customers who are at risk of churning. Additionally, the model will generate a feature importance ranking, highlighting the most influential factors in predicting customer churn. These insights will be invaluable for the organization, enabling it to implement targeted retention strategies based on the identified key drivers of churn.

### B2. Assumption of the Chosen Prediction Method

**Assumption:** One key assumption of the Random Forest method is that the individual decision trees within the forest are uncorrelated. By selecting random subsets of data and features for each tree, the method assumes that the trees will capture different aspects of the data, leading to a more generalized and accurate ensemble model. This assumption helps prevent overfitting, as no single tree dominates the prediction process.

### B3. Programming Language and benefits. In this analysis, several Python libraries will utilized to perform data preprocessing, modeling, and evaluation. The Pandas library was fundamental in handling and manipulating the dataset, providing powerful tools for cleaning and preparing the data for analysis. NumPy complemented Pandas by offering efficient array operations, which were particularly useful in numerical data processing. For model building, Scikit-Learn was the primary library, chosen for its comprehensive suite of machine learning algorithms, including the Random Forest classifier used in this project. Scikit-Learn also provided essential tools for splitting the data, scaling features, and evaluating model performance through metrics such as accuracy, ROC-AUC, and Log Loss. Matplotlib was employed for data visualization, enabling the clear presentation of model results and feature importance, which helped in interpreting the model’s predictions. Together, these libraries formed a robust toolkit that facilitated a smooth and effective analysis, from data preparation to model evaluation and interpretation.

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

### C1. Describe One Data Preprocessing Goal Relevant to the Prediction Method

The primary data preprocessing goal in this analysis was to ensure that the dataset was in a suitable format for training a Random Forest model, a robust classification technique. Given the diverse nature of the data, which included both categorical and numerical variables, preprocessing focused on handling potential inconsistencies such as missing values, converting categorical variables into a format that could be utilized by the model, and standardizing numerical features. This comprehensive approach was essential for improving the model’s accuracy and ensuring that it could effectively learn from the data to predict customer churn.

### C2. Identify the Initial Data Set Variables for Analysis

### The analysis utilized a combination of numerical and categorical variables to predict customer churn. The numerical variables included features such as MonthlyCharge, Bandwidth\_GB\_Year, Outage\_sec\_perweek, Yearly\_equip\_failure, Tenure, Responses, Solutions, Replacements, Reliability, Options, Respectfulness, Courteous, and Listening, which provided detailed insights into customer behavior and service interactions. Categorical variables like Education, Employment, Marital Status, Gender, Contract, Techie, InternetService, Phone, Multiple Services, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, and StreamingMovies were also included to capture demographic information and service preferences. Together, these variables provided a comprehensive view of each customer, contributing to the model’s ability to accurately predict churn.

### C3. Explain the Steps Used to Prepare the Data for the Analysis

The data preparation process began by selecting the relevant features that were most likely to impact customer churn.

A screenshot of a computer

Description automatically generated

After identifying these variables, the first step was to handle any missing values in the dataset to ensure that the analysis would not be skewed by incomplete data.

A white card with black text

Description automatically generated

Next, we addressed the binary categorical variables, which included columns such as Techie, Phone, Multiple, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies, and Churn. These columns originally contained the string values 'Yes' and 'No'. Similar to the Churn column, these values were mapped to 1 and 0, respectively. This conversion was necessary to prepare the data for machine learning, as models like Random Forests cannot directly handle categorical string data.

After converting these binary variables to numerical values, One-Hot Encoding was applied with the drop\_first=True parameter. This process creates dummy variables for each binary category but drops the first category to avoid the dummy variable trap (multicollinearity). The result was a cleaner, more efficient dataset with appropriate binary representations.



The dataset also contained several columns with more than two categories, such as Contract, InternetService, Education, Employment, Marital, and Gender. For these columns, One-Hot Encoding was applied without using the drop\_first option. This approach ensures that all categories are represented by separate binary columns, allowing the Random Forest model to distinguish between all levels of these categorical variables.These steps were crucial in preparing the data, ensuring that the Random Forest model could leverage the full range of information available to predict customer churn accurately.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

### 

### C4. Provide the prepared data set as a CSV file.

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



A white rectangular object with a white background

Description automatically generated

# Part IV: Model Comparison and Analysis

D1. Split the data

A screen shot of a computer code

Description automatically generated

D2.Describe the analysis technique used

A screenshot of a computer program

Description automatically generated

The analysis was conducted using a Random Forest classifier, a robust ensemble learning method that constructs multiple decision trees and aggregates their outputs to enhance predictive accuracy. This technique is well-suited for the complexity and variability of customer data, providing both high accuracy and interpretability through feature importance analysis.

During the analysis, the dataset was split into training and testing sets, with 70% of the data used to train the model and 30% reserved for testing.

The model was then trained on the features derived from customer usage patterns and demographic information to predict the likelihood of churn. The Random Forest model was chosen due to its ability to handle a mix of categorical and numerical data, reduce the risk of overfitting, and offer insights into which features most influence customer behavior.

The model's performance was evaluated using several metrics, including accuracy, precision, recall, and the ROC-AUC score.

A screenshot of a computer

Description automatically generated

Additionally, Log Loss was calculated to assess the confidence of the probability predictions. These metrics provided a comprehensive evaluation of the model’s effectiveness in predicting customer churn, helping to identify areas where further improvements could be made.

This approach ensures that the model is both accurate and practical for informing business strategies aimed at reducing customer churn.

A screenshot of a computer program

Description automatically generated

D3. Code will be provided in the uploaded assignment.

# Part V: Data Summary and Implications

### E1. Explain the Accuracy and Mean Squared Error (MSE) of Your Prediction Model

A screenshot of a computer program

Description automatically generated

The Random Forest model achieved an accuracy of **88.43%,** indicating that it correctly predicted customer churn in a significant majority of cases. While Mean Squared Error (MSE) is 0.0868 This suggests that the model's probability estimates are relatively well-calibrated, with a lower Log Loss indicating better performance (Zhang et al., 2020)..

### E2. Discuss the results and implications of your prediction analysis

The high accuracy of **88.43%** suggests that the Random Forest model is effective in distinguishing between customers who are likely to churn and those who are not. The model's performance is strong, but the **precision** and **recall** for the churned customers indicate that there is room for improvement, particularly in capturing the positive class (customers who churn). This aligns with findings from prior research, where Random Forest models have been shown to perform well in similar churn prediction tasks but may require additional tuning to improve the balance between precision and recall (Verbeke et al., 2012). The results imply that the organization can confidently use this model to target retention efforts, potentially reducing churn rates. However, continuous monitoring and model refinement are recommended to maintain and improve performance.

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

One limitation of this analysis is the potential issue of **class imbalance**, which may have contributed to the lower recall for the churned customers. Class imbalance is a common challenge in churn prediction models, as highlighted by Burez and Van den Poel (2009), and can lead to biased predictions if not properly addressed. Techniques such as oversampling the minority class, adjusting class weights, or using synthetic data generation methods could be employed in future iterations of the model to mitigate this issue and improve the model's ability to identify customers at risk of churning.

**E4. Recommend a Course of Action Based on Your Results and Implications**

Given the model's results, it is recommended that the organization implement targeted retention strategies for customers identified as likely to churn. Regularly monitoring and retraining the model will ensure that it adapts to changes in customer behavior, maintaining its predictive accuracy over time. Additionally, addressing class imbalance through techniques like oversampling or synthetic data generation, as suggested by research in predictive modeling (Burez & Van den Poel, 2009), could further enhance the model’s effectiveness. This proactive approach will help the organization make data-driven decisions to improve customer retention.Part VI: Demonstration

**See Code in video provided**

# **<https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=550a7b07-392a-4e35-b32f-b1d4013f3ef7>**

# Sources

*Pandas.get\_dummies#*. pandas.get\_dummies - pandas 2.2.2 documentation. (n.d.). https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.get\_dummies.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

* Burez, J., & Van den Poel, D. (2009). Handling class imbalance in customer churn prediction. *Expert Systems with Applications, 36*(3), 4626-4636. https://doi.org/10.1016/j.eswa.2008.05.027
* Verbeke, W., Martens, D., Mues, C., & Baesens, B. (2012). Building comprehensible customer churn prediction models with advanced rule induction techniques. *Expert Systems with Applications, 39*(17), 1166-1176. https://doi.org/10.1016/j.eswa.2012.09.017
* Zhang, C., Ling, C. X., & Li, J. (2020). An improved algorithm for predicting customer churn using feature selection and Random Forest. *Journal of Business Research, 118*, 265-275. https://doi.org/10.1016/j.jbusres.2020.06.041