**D208 PA**

Jon Isak

D208: Predictive Modeling

Eric Straw

Contents

[D208 PA 2](#_Toc151119339)

[Part I: Research Question 2](#_Toc151119340)

[Part II: Method Justification 2](#_Toc151119341)

[Part III: Data Preparation 3](#_Toc151119342)

[Part IV: Model Comparison and Analysis 4](#_Toc151119343)

[Part V: Data Summary and Implications 6](#_Toc151119344)

## D208 PA

## Part I: Research Question

A1. The data analysis will explore the question: Can we predict the likelihood of customer churn based on various factors? This question seeks to find what factors affect churn. Furthermore, it allows the company to create strategies to tackle the customer churn and increase revenue.

A2. One of the goals of the data analysis is to use logistic regression to predict and understand the chance of customer churn based on various factors. Another goal is to use the significant factors of churn to create a strategy to reduce churn rates.

## Part II: Method Justification

B1. Four assumptions need to be met when fitting a logistic regression model. First, the response variable should be binary. Second, the observations need to be independent. Third, there should be no multicollinearity among explanatory variables. Finally, extreme outliers should be absent (Zach, 2020).

B2. I plan to perform the analysis using Python within a Jupyter notebook environment. Python stands out as an excellent option because of its abundant resources, including thorough documentation, tutorials, and strong community support, which are invaluable for every stage of the analysis. Furthermore, Python boasts a multitude of powerful libraries such as Pandas, NumPy, and Matplotlib, which greatly enhance its suitability for a wide range of data analysis tasks. Pandas provides efficient capabilities for manipulating data, allowing for seamless handling and transformation of datasets. NumPy offers crucial numerical operations and efficient array handling, enabling swift and optimized mathematical computations. Additionally, Matplotlib facilitates the creation of insightful visualizations, aiding in thorough exploration and presentation of findings. The adaptability, reliability, and extensive range of Python libraries make it an optimal choice for this analysis. Python offers the essential tools needed to effectively carry out data processing, statistical modeling, visualization, and interpretation of results (Terra, 2019).

B3. Logistic regression is an appropriate statistical technique for analyzing my research question for several reasons. First, logistic regression is suitable for predicting binary outcomes, which makes it good for my ‘churn’ data since it consists of yes or no values. Second, my variables are independent, which makes it suitable. Furthermore, it can estimate the probability of something occurring based on independent variables.

## Part III: Data Preparation

C1. The data was verified to be clean in task 1. The data was checked for data type consistency and missing values. As a result, no additional data cleaning needs to be performed and the data is ready for analysis.

C2. The independent variables will be 'Email', 'Contacts', 'Yearly\_equip\_failure', 'Tenure', and 'MonthlyCharge'. The dependent variable will be ‘Churn’. Looking into the independent variables, the average email count is 12, and ranges from 1-23. Contacts had an average of 1 and ranged from 0-7. Equipment failure averaged .4 and ranged from 06. Tenure has an average of 34.5 and ranges from 1-72. Monthly charges averaged 172.62 and ranged from 8-290.

C3. Examining the univariate visualizations, the distribution of emails exhibits a bell curve, suggesting a normal distribution. On the other hand, the distributions of contacts and yearly\_equip\_failure appear to be left-skewed. Tenure demonstrates a bimodal distribution, signifying two distinct modes within the dataset. Additionally, the distribution of monthly charges forms a bell curve with a slight left skew. Regarding the bivariate visualizations, the means and quartiles for Email, Contacts, and Yearly\_equip\_failure remain relatively consistent between churn and non-churn customers. However, a notable disparity emerges in Tenure, non-churn customers tend to exhibit significantly longer tenures compared to churn customers. Moreover, in the case of MonthlyCharge, churn customers appear to have higher average charges than non-churn customers, evident from the quartile values and mean. The visualizations are included with the code.

C4. To prepare the data, first I will scale the numerical values to be on the same scale. Second, I will encode ‘Churn’ from Yes/No to 0/1. This data will be saved to a new file.

C5. The file is attached.

## Part IV: Model Comparison and Analysis

D1. The initial results show that Email, Contacts, and Yearly\_equip\_failure have coefficients near zero and p-values higher than the significant level of .05. This suggests that these values are insignificant. Tenure has a coefficient of -1.95, which suggests that as tenure increases, the chance of churn decreases. It’s also significant according to p-value. MonthlyCharge has a coefficient of 1.4. This suggests that churn increases as monthly charge increases. It’s also significant according to the p-value.

A screenshot of a computer

Description automatically generated

D2. My question seeks to identify variables that affect churn, so my goal is to only retain significant variables. I will utilize p-value to reduce the variables.

D3. Using p-values, tenure and monthly charge were the only variables with p-values less than .05. This indicates that they are significant in predicting churn.

A screenshot of a computer

Description automatically generated

E1. The initial model included multiple variables that were identified from the data that could affect churn. Then, the variables were reduced according to their significance using p-values. The initial model had a pseudo-R-squared value of 0.4085, which suggests that the 5 variables accounted for approximately 41% of the variability in churn. The reduced model had a pseudo-R-squared value of 0.4084, which suggests that the 2 variables can explain the same amount of variability in churn as the model with 5 variables. The reduced model is more efficient because it has the same explanatory power as the initial model.

E2. According to the confusion matrix, there were 6738 true negatives, 1050 false negatives, 612 false positives, 1600 true positives. The accuracy of the model was 83.38%, which suggests that the model predicts churn and non-churn correctly 83.38% of the time.

A close-up of a number

Description automatically generated

E3. The code is attached.

## Part V: Data Summary and Implications

F1. The reduced logistic regression model can be represented with this equation, Churn=−1.9754−1.9526×Tenure+1.4245×MonthlyCharge. According to the equation, for every unit in tenure the log odds of churn decrease by 1.9526 and for every unit increase in monthly charge there is a 1.4245 increase in churn. Only tenure and monthly charge were found to be statistically significant. The limitations include the oversimplification of the model by reducing it to two variables. Furthermore, the analysis assumes linearity, absence of multicollinearity, and no interaction effects between variables. Violations of these assumptions could impact the model's accuracy.

F2. Based on the results, the company can pursue multiple strategies to reduce churn. Predictive models can be used to identify high risk customers to send personalized retention offers. In addition, they can create offers to reward and retain high value and long-term customers. With an average churn on the 13th month of service, the company can investigate this issue further. The company can also have regular check-ins to gather feedback and ensure satisfaction within the first year of the service.

## Part VI: Demonstration

G. https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=87a88aec-a686-49bd-8b0f-b0bd01716d69

H.

Seaborn. (n.d.). Seaborn.pairplot. Retrieved November 17, 2023, from https://seaborn.pydata.org/generated/seaborn.pairplot.html

Investigate. (n.d.). Picking what matters and what doesn’t in a regression. Investigate.Ai: Data Science for Journalists. Retrieved November 17, 2023, from https://investigate.ai/apm-reports-jury-bias/in-the-dark-feature-selection-with-p-values/

I.

Javapoint. (n.d.). Accuracy\_Score in Sklearn. Retrieved November 17, 2023, from https://www.javatpoint.com/accuracy\_score-in-sklearn

Terra, J. (2019, March 5). *Python for Data Science and Data Analysis*. Simplilearn.Com. https://www.simplilearn.com/why-python-is-essential-for-data-analysis-article

Zach. (2020, October 13). The 6 Assumptions of Logistic Regression. *Statology*. https://www.statology.org/assumptions-of-logistic-regression/