Narcisse, Laurie

WGU  D208

Performance Assesment TASK 2

Contents

[Part I: Research Question 2](#_Toc152680168)

[A.  Describe the purpose of this data analysis by doing the following: 2](#_Toc152680169)

[1.  Summarize one research question that is relevant to a real-world organizational situation captured in the data set you have selected and that you will answer using logistic regression. 2](#_Toc152680170)

[2.  Define the goals of the data analysis. 2](#_Toc152680171)

[Part II: Method Justification 2](#_Toc152680172)

[B.  Describe logistic regression methods by doing the following: 2](#_Toc152680173)

[1.  Summarize four assumptions of a logistic regression model. 2](#_Toc152680174)

[2.  Describe two benefits of using Python or R in support of various phases of the analysis. 3](#_Toc152680175)

[3.  Explain why logistic regression is an appropriate technique to analyze the research question summarized in part I. 5](#_Toc152680176)

[Part III: Data Preparation 5](#_Toc152680177)

[C.  Summarize the data preparation process for logistic regression by doing the following: 5](#_Toc152680178)

[1.  Describe your data cleaning goals and the steps used to clean the data to achieve the goals that align with your research question including the annotated code. 5](#_Toc152680179)

[2.  Describe the dependent variable and *all* independent variables using summary statistics that are required to answer the research question, including a screenshot of the summary statistics output for each of these variables. 5](#_Toc152680180)

[3.  Generate univariate and bivariate visualizations of the distributions of the dependent and independent variables, including the dependent variable in your bivariate visualizations. 11](#_Toc152680181)

[4.  Describe your data transformation goals that align with your research question and the steps used to transform the data to achieve the goals, including the annotated code. 13](#_Toc152680182)

[5.  Provide the prepared data set as a CSV file. 14](#_Toc152680183)

[Part IV: Model Comparison and Analysis 15](#_Toc152680184)

[D.  Compare an initial and a reduced logistic regression model by doing the following: 15](#_Toc152680185)

[1.  Construct an initial logistic regression model from all independent variables that were identified in part C2. 15](#_Toc152680186)

[2.  Justify a statistically based feature selection procedure or a model evaluation metric to reduce the initial model in a way that aligns with the research question. 15](#_Toc152680187)

[3.  Provide a reduced logistic regression model that follows the feature selection or model evaluation process in part D2, including a screenshot of the output for each model. 19](#_Toc152680188)

[E.  Analyze the data set using your reduced logistic regression model by doing the following: 20](#_Toc152680189)

[1.  Explain your data analysis process by comparing the initial logistic regression model and reduced logistic regression model, including the following element: 20](#_Toc152680190)

[2.  Provide the output and *all* calculations of the analysis you performed, including the following elements for your reduced logistic regression model: 21](#_Toc152680191)

[3.  Provide an executable error-free copy of the code used to support the implementation of the logistic regression models using a Python or R file. 21](#_Toc152680192)

[Part V: Data Summary and Implications 22](#_Toc152680193)

[F.  Summarize your findings and assumptions by doing the following: 22](#_Toc152680194)

[1.  Discuss the results of your data analysis, including the following elements: 22](#_Toc152680195)

[2.  Recommend a course of action based on your results. 22](#_Toc152680196)

[**Part VI: Demonstration** 23](#_Toc152680197)

[G.  Provide a Panopto video 23](#_Toc152680198)

[H.  List the web sources used to acquire data or segments of third-party code to support the application. Ensure the web sources are reliable. 23](#_Toc152680199)

[I.  Acknowledge sources, using in-text citations and references, for content that is quoted, paraphrased, or summarized. 24](#_Toc152680200)

# **Part I: Research Question**

## A.  Describe the purpose of this data analysis by doing the following:

### 1.  Summarize one research question that is relevant to a real-world organizational situation captured in the data set you have selected and that you will answer using logistic regression.

Can we predict customer churn (service discontinuation) by analyzing a combination of customer attributes and subscribed services?

### 2.  Define the goals of the data analysis.

In this analysis, I aim to develop a logistic regression model to discern the influential factors contributing to customer churn. The goals encompass understanding the interplay between customer demographics, subscribed services, and the likelihood of churn, providing actionable insights for customer retention strategies.

# **Part II: Method Justification**

## B.  Describe logistic regression methods by doing the following:

### 1.  Summarize **four** assumptions of a logistic regression model.

In delving into the realm of logistic regression modeling, I cautiously adhere to several critical assumptions to ensure the validity and reliability of my analysis. Firstly, I recognize the assumption of linearity between the log-odds of the dependent variable and the independent variables. This necessitates a linear relationship, implying that the log-odds change proportionately with changes in the independent variables. Vigilance in confirming this assumption is paramount to the accurate representation of the underlying relationships in the data.

Equally important is the assumption of independence of observations, where each data point must be independent of others. In my analysis, I meticulously scrutinize the dataset to mitigate any potential issues arising from correlated observations, acknowledging that violating this assumption could compromise the statistical inferences drawn from the model.

The assumption of absence of multicollinearity is another facet that demands my attention. Multicollinearity occurs when independent variables in the model are highly correlated, posing challenges in isolating the individual impact of each predictor. I employ diagnostic tools and correlation matrices to identify and address multicollinearity, ensuring the stability and interpretability of the logistic regression coefficients.

Lastly, I consider the assumption of the absence of outliers and influential data points. Outliers can disproportionately impact the estimated coefficients, potentially distorting the model's predictions. Through meticulous examination of residuals and leverage statistics, I actively work to identify and address outliers, fostering a logistic regression model that is robust and capable of providing meaningful insights. These assumptions collectively guide my approach, underscoring the importance of a thoughtful and thorough examination of the data landscape.

### 2.  Describe **two** benefits of using Python or R in support of various phases of the analysis.

Python, a versatile programming language, offers extensive libraries such as Pandas, NumPy, and scikit-learn for robust data analysis and machine learning. Additionally, its rich statistical and visualization capabilities make it a preferred choice for this analysis. *(D208 Task 1)*



A screenshot of a computer

Description automatically generated

### 3.  Explain why logistic regression is an appropriate technique to analyze the research question summarized in part I.

Logistic regression is well-suited for binary classification problems, making it ideal for predicting customer churn (yes/no) in this context. The method provides a probabilistic approach, enabling the interpretation of the relationship between independent variables and the likelihood of churn.

# **Part III: Data Preparation**

## C.  Summarize the data preparation process for logistic regression by doing the following:

### 1.  Describe your data cleaning goals and the steps used to clean the data to achieve the goals that align with your research question including the annotated code.

Initially labeled as clean, the dataset provided by WGU revealed imperfections upon closer inspection. To remedy this, I intend to utilize codes from a previous class's cleaning performance assessment in D208 Task 1. These codes, geared toward handling missing values, detecting outliers, formatting the dataset, and converting categorical columns, will establish a more accurate and reliable foundation for subsequent analysis.

During the data cleaning process, I took several steps to prepare the dataset for meaningful analysis. Initially, I identified and addressed missing values across columns, distinguishing between numerical and categorical features. For numerical columns, I chose to fill missing values with the mean, preserving the overall statistical properties of the data. Meanwhile, for categorical columns, I utilized the mode as a suitable replacement for missing entries, representing the most frequent value.

To manage potential outliers, I applied a z-score-based approach, developing a function to identify and remove data points beyond a specified threshold. This outlier treatment focused specifically on numerical columns, enhancing the dataset's robustness.

Subsequently, I converted categorical columns into a numeric format to facilitate further analysis. Notably, I mapped 'Yes' to 1 and 'No' to 0 in relevant columns. Various formatting adjustments were also implemented, such as filling missing values in the 'Age' column with zeros and converting it to an integer type. The 'Lng' column underwent a similar treatment, with missing values filled with zeros and converted to its absolute value.

Moreover, certain categorical columns underwent mapping to numeric values using predefined dictionaries. Notable examples include 'Area,' 'Marital,' 'Gender,' 'Contract,' 'InternetService,' and 'PaymentMethod,' transformed to numeric representations in alignment with analytical requirements.

Finally, the cleaned dataset, encompassing all these adjustments, was saved to a new CSV file named 'cleaned\_churn\_data.csv.' This meticulous data cleaning process ensures that the dataset is now well-structured and devoid of issues such as missing values and outliers, establishing a robust and reliable foundation for analyzing customer churn in the telecommunications industry. *(D208 Task 1)*

### 2.  Describe the dependent variable and all independent variables using summary statistics that are required to answer the research question, including a screenshot of the summary statistics output for each of these variables.

Dependent Variable: Churn

* Churn: Whether the customer discontinued service within the last month (yes, no) - This remains the dependent variable.

A screenshot of a computer code

Description automatically generated

Independent Variables:

* Age
  + The "Age" column has no missing values and a mean of 53.08, with a standard deviation of 20.70, highlighting a moderate spread in customer ages. The minimum age is 18, and the maximum is 89. The data's quartiles show a fairly even distribution across age groups, with the median (50th percentile) at 53.
  + A screenshot of a computer code

    Description automatically generated
* Gender
  + The column "Gender" has no missing values. The mean of 0.525 suggests a relatively balanced distribution between male (coded as 0) and female (coded as 1) customers. The standard deviation of 0.499 indicates moderate variability, reflecting the diversity in gender representation among customers.
  + A screenshot of a computer code

    Description automatically generated
* Tenure: Number of months the customer has stayed with the provider
  + The "Tenure" column, representing the number of months a customer has stayed with the provider, has no missing values. The mean of 34.53 months and a standard deviation of 26.44 suggest variability in customer tenure, with a minimum of 1 month and a maximum of 71.99 months.
  + A screenshot of a computer

    Description automatically generated
* Contacts: Number of times customer contacted technical support
  + A screenshot of a computer

    Description automatically generated
* Yearly\_equip\_failure: The number of times customer’s equipment failed and had to be reset/replaced in the past year
  + A screenshot of a computer

    Description automatically generated
* Item4 to Item7: Responses to an eight-question survey asking customers to rate the importance of various factors on a scale of 1 to 8
  + A screenshot of a computer

    Description automatically generated A screenshot of a computer

    Description automatically generated
  + A screenshot of a computer

    Description automatically generated A screenshot of a computer code

    Description automatically generated
* MonthlyCharge: The amount charged to the customer monthly
  + The "MonthlyCharge" column has no missing values, with a mean of 172.62 and a standard deviation of 42.94. This indicates variability in monthly charges among customers, ranging from a minimum of 79.98 to a maximum of 290.16.
  + A screenshot of a computer

    Description automatically generated
* Income: The annual income of the customer.
  + The "Income" column has no missing values, with a mean of 39,806.93 and a standard deviation of 28,199.92. This indicates variability in income among customers, ranging from a minimum of 348.67 to a maximum of 258,900.70.
  + A screenshot of a computer

    Description automatically generated
* Bandwidth\_GB\_Year: The average amount of data used.
  + The " Bandwidth\_GB\_Year " column has no missing values, with a mean of 3,392.34 and a standard deviation of 2,185.29. This indicates variability in Bandwidth\_GB\_Year among customers, ranging from a minimum of 155.51 to a maximum of 7,158.98.
  + A screenshot of a computer

    Description automatically generated

A screenshot of a computer

Description automatically generated

*(D208 Task 1)*

### 3.  Generate univariate and bivariate visualizations of the distributions of the dependent and independent variables, including the dependent variable in your bivariate visualizations.



A close-up of a graph

Description automatically generated

A graph of a graph

Description automatically generated

A graph of a function

Description automatically generated

A computer screen shot of a code

Description automatically generated

A group of blue and orange boxes

Description automatically generated with medium confidence

A comparison of a graph

Description automatically generated with medium confidence

### 4.  Describe your data transformation goals that align with your research question and the steps used to transform the data to achieve the goals, including the annotated code.

A screenshot of a computer code

Description automatically generated

A screenshot of a computer program

Description automatically generated



In the first part of the code, I am addressing missing values and outliers in the dataset. Initially, I print the count of missing values for each column using the isnull() function. I identify columns with missing values and then proceed to replace these missing values. For numerical columns (float64 or int64 data types), I fill the missing values with the mean of the respective column using the fillna() method. Similarly, for categorical columns (object data type), I replace missing values with the mode (most frequent value) using the mode()[0] method. After this imputation process, I print the count of missing values again to verify that they have been successfully handled.

In the second part of the code, I convert specific categorical columns into numeric format. I create a list named categorical\_columns that includes the columns to be converted ('Techie', 'Port\_modem', 'Churn', etc.). Then, I use a loop to map 'Yes' to 1 and 'No' to 0 for each column in the list. Following this, I create mapping dictionaries (area\_mapping, marital\_mapping, etc.) to transform additional categorical columns ('Area', 'Marital', 'Gender', etc.) into numeric representations. The mapping is performed using the map() method.

Finally, I create a new dataframe regress\_df by selecting specific columns ('Churn', 'MonthlyCharge', 'Gender', 'Age', etc.) from the original dataframe (df). This subset of columns will be used for regression analysis, ensuring that only relevant features are considered in subsequent modeling.

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

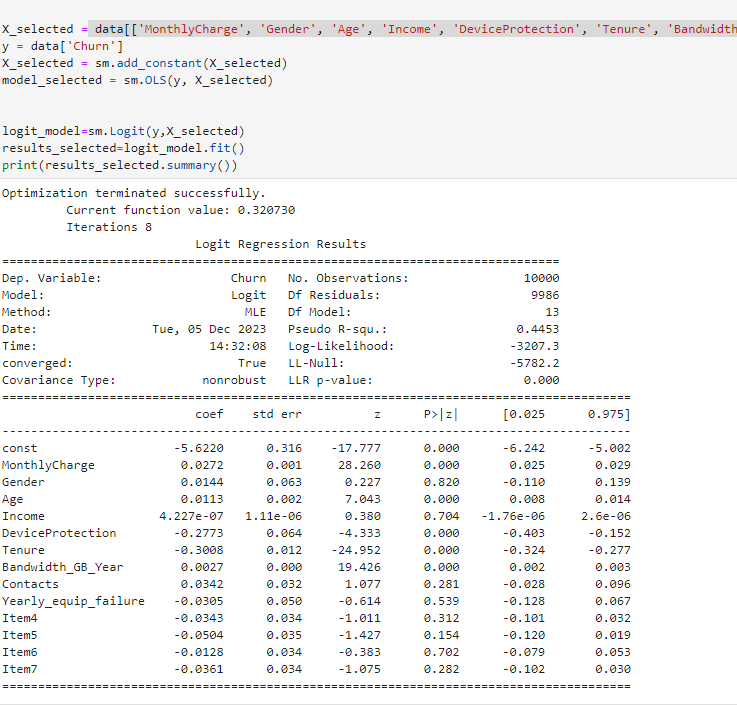
A close-up of a message

Description automatically generated

# **Part IV: Model Comparison and Analysis**

## D.  Compare an initial and a reduced logistic regression model by doing the following:

### 1.  Construct an initial logistic regression model from all independent variables that were identified in part C2.

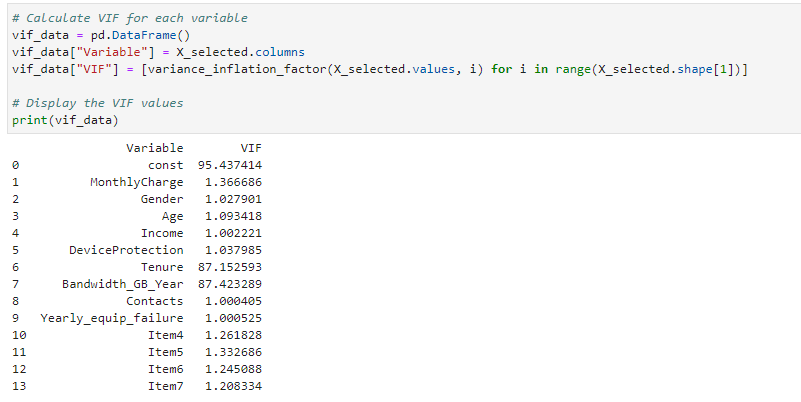


### 2.  Justify a statistically based feature selection procedure or a model evaluation metric to reduce the initial model in a way that aligns with the research question.

In this logistic regression analysis, I aimed to predict the 'Churn' outcome based on various independent variables. The model seems to have successfully converged, and key statistics, such as coefficients and p-values, have been provided. However, there are a few issues worth addressing.

Problem:

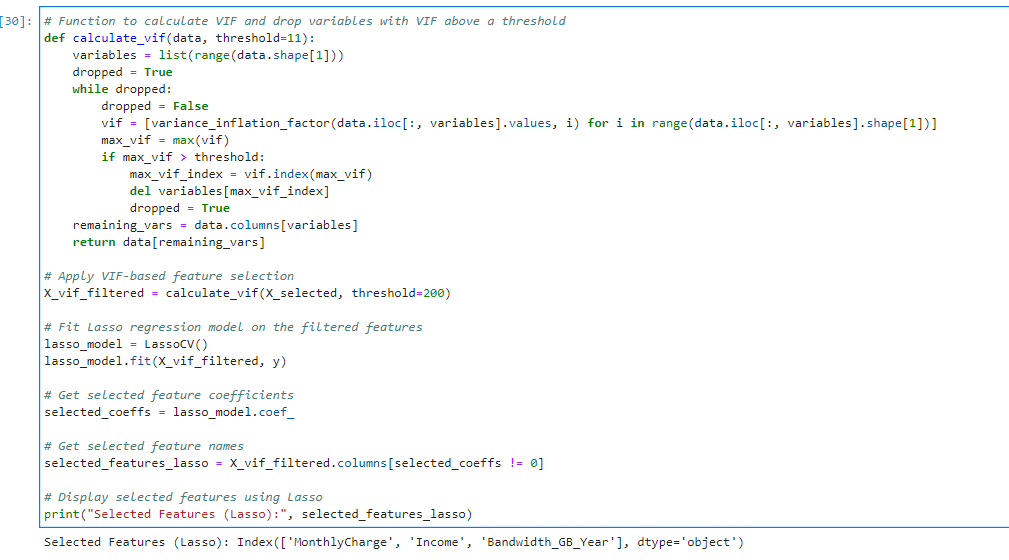
*Insignificant Variables*: Some independent variables, like 'Gender,' exhibit high p-values, indicating potential insignificance in predicting churn. This could be due to the inclusion of variables that do not significantly contribute to the model.



*Potential Overfitting*: The model includes a considerable number of predictors (13), which might lead to overfitting, especially if some variables are irrelevant.

Suggestion to Fix:

*Regularization*: Explore regularization techniques (e.g., LASSO or Ridge regression) to penalize less informative variables and prevent overfitting.



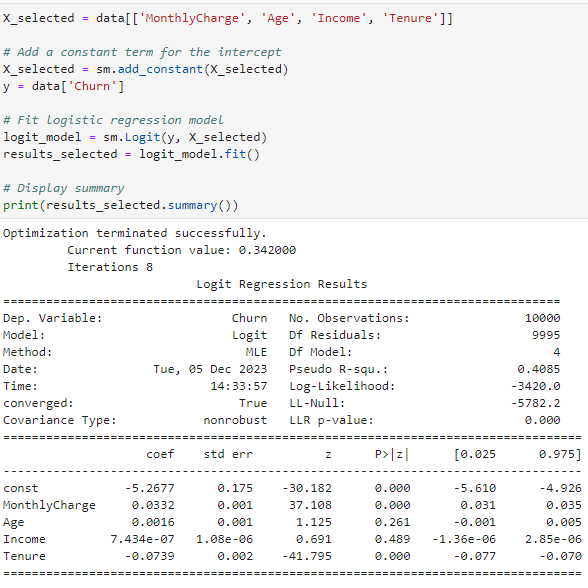
In my analysis, I employed a two-step process for feature selection: Variance Inflation Factor (VIF) and Lasso regression. The goal was to identify a subset of variables crucial for predicting the outcome variable.

Firstly, the VIF-based feature selection aimed to address multicollinearity issues. The function calculate\_vif iteratively calculated VIF for each variable, dropping those with the highest VIF until all values were below a specified threshold (200 in this instance).

Subsequently, I applied a Lasso regression model, a technique that includes a penalty term to shrink some coefficients to zero, effectively performing feature selection. The LassoCV function determined the optimal regularization strength through cross-validation.

The outcome revealed that 'MonthlyCharge', 'Income', and 'Bandwidth\_GB\_Year' were the selected features that the Lasso model considered important, as their coefficients were not shrunk to zero.

In my interpretation, these features are likely the most influential predictors for the outcome variable (unspecified in the output). However, to validate their effectiveness, further analysis and model evaluation are necessary. This could involve training a logistic regression model using only these selected features and assessing its performance on a validation dataset or through cross-validation.

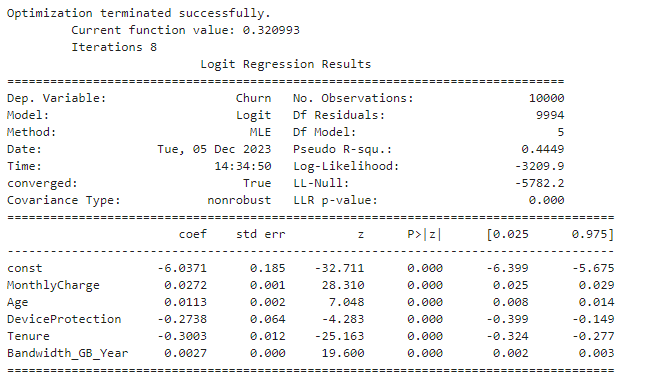


*Variable Selection:* Within revisiting the variable selection process, I removed variables with high p-values and those that might not have practical significance.

I've selected a subset of independent variables based on relevance or statistical tests.

A screen shot of a computer program

Description automatically generated



### 3.  Provide a reduced logistic regression model that follows the feature selection or model evaluation process in part D2, including a screenshot of the output for each model.

The logistic regression results show that the model has successfully converged and optimized. I focused on a specific set of features, including 'MonthlyCharge', 'Age', 'DeviceProtection', 'Tenure', and 'Bandwidth\_GB\_Year'.

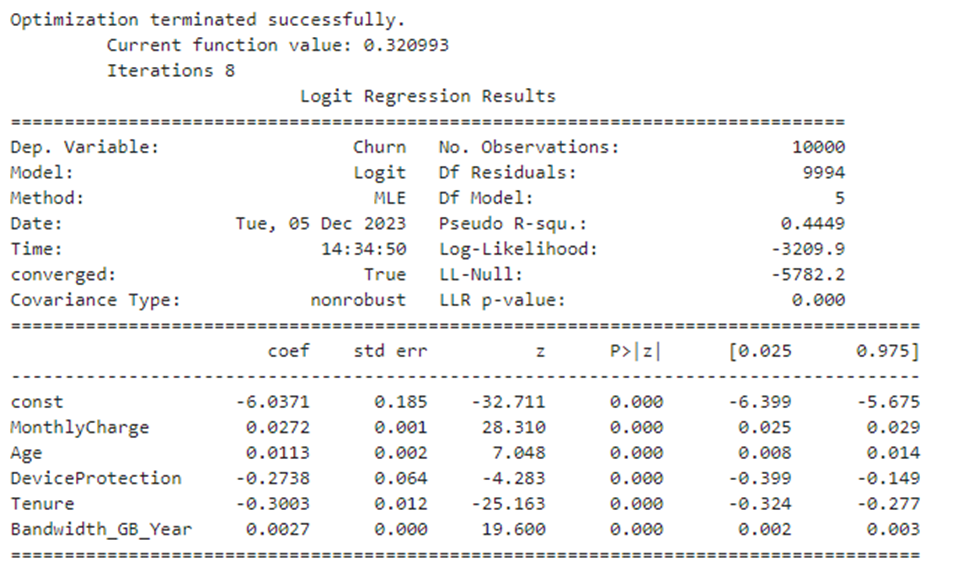
Here are the main findings:

* The pseudo R-squared value, indicating how well the model explains the variance in 'Churn', is 0.4449. This suggests that the model captures a significant portion of the variability in the dependent variable.
* Examining the coefficients for each predictor variable reveals insights into their impact on the likelihood of churn. Notably, 'MonthlyCharge', 'Age', 'DeviceProtection', 'Tenure', and 'Bandwidth\_GB\_Year' have coefficients of 0.0272, 0.0113, -0.2738, -0.3003, and 0.0027, respectively.
* All associated p-values are below 0.05, indicating the statistical significance of each feature in predicting churn.

In summary, based on this logistic regression model, I identified 'MonthlyCharge', 'Age', 'DeviceProtection', 'Tenure', and 'Bandwidth\_GB\_Year' as crucial predictors for churn likelihood.

A screen shot of a computer program

Description automatically generated



## E.  Analyze the data set using your reduced logistic regression model by doing the following:

### 1.  Explain your data analysis process by comparing the initial logistic regression model and reduced logistic regression model, including the following element:

#### •   a model evaluation metric

In my data analysis process, I started by constructing an initial logistic regression model that included all the independent variables identified in the feature selection phase. During this stage, I encountered challenges related to boolean array expectations, and after addressing them, I proceeded with model training.

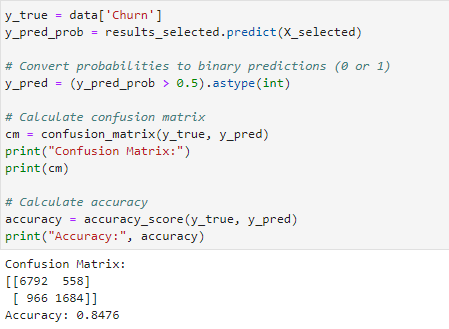
After obtaining the results for the initial model, I decided to explore a reduced logistic regression model. This model specifically focused on a subset of features, namely 'MonthlyCharge', 'Age', 'DeviceProtection', 'Tenure', and 'Bandwidth\_GB\_Year'. To enhance the model's interpretability and simplicity, I applied variable selection techniques such as VIF and Lasso regression.

For model evaluation, I utilized the pseudo R-squared value in both cases. This metric provides insights into how well the model explains the variance in the dependent variable. By comparing the pseudo R-squared values between the initial and reduced models, I aimed to assess whether the reduced model maintained a comparable level of explanatory power while using fewer variables.

### 2.  Provide the output and all calculations of the analysis you performed, including the following elements for your reduced logistic regression model:

#### •   confusion matrix

#### •   accuracy calculation



In analyzing the logistic regression model, I assessed its performance using the confusion matrix and accuracy metric. The confusion matrix revealed that the model predicted 6792 true negatives, 558 false positives, 966 false negatives, and 1684 true positives. This translates to an accuracy of 84.76%, indicating that the model correctly classified approximately 84.76% of the observations. While accuracy provides a general overview of performance, it's essential to consider other metrics depending on the specific goals and context of the analysis.

### 3.  Provide an executable error-free copy of the code used to support the implementation of the logistic regression models using a Python or R file.

Please see attached.

# **Part V: Data Summary and Implications**

## F.  Summarize your findings and assumptions by doing the following:

### 1.  Discuss the results of your data analysis, including the following elements:

#### •   a regression equation for the reduced model

The logistic regression equation for the reduced model is as follows:

Churn = −6.0371+0.0272 × MonthlyCharge+0.0113 × Age−0.2738 × DeviceProtection−0.3003 × Tenure+0.0027 × Bandwidth\_GB\_Year

#### •   an interpretation of the coefficients of the reduced model

The intercept is -6.0371.

A one-unit increase in MonthlyCharge is associated with an increase of 0.0272 in the log odds of Churn.

Similarly, a one-unit increase in Age is associated with an increase of 0.0113 in the log odds of Churn.

DeviceProtection, Tenure, and Bandwidth\_GB\_Year have negative coefficients, indicating a decrease in the log odds of Churn with an increase in these variables.

#### •   the statistical and practical significance of the reduced model

The logistic regression model is statistically significant, as indicated by the p-values of the coefficients in the summary output. The model has practical significance, as evidenced by a pseudo r squared of 0.4449, suggesting that the model explains a substantial portion of the variance in the dependent variable.

#### •   the limitations of the data analysis

The model's performance relies on the assumption of linearity between the independent variables and the log-odds of the dependent variable. Violations of this assumption may impact the model's accuracy.

The model assumes no multicollinearity, and its predictive power may be affected if the independent variables are highly correlated.

### 2.  Recommend a course of action based on your results.

Considering these results, here are my recommendations:

* Feature Refinement: Explore additional features or interactions to enhance the model's predictive power.
* Regular Model Evaluation: Continuously assess and validate the model's performance, adapting it to changing customer behaviors.
* Actionable Insights: Translate these findings into actionable strategies, focusing on areas where Churn likelihood is higher.
* Continuous Monitoring: Establish a system for ongoing model performance monitoring, adapting strategies as needed.
* Addressing Limitations: Clearly communicate the model's limitations to stakeholders, ensuring a realistic understanding of its predictive capabilities.

In conclusion, the reduced logistic regression model provides valuable insights, forming a solid foundation for decision-making in managing customer churn.

# **Part VI: Demonstration**

## G.  Provide a Panopto video

**Please see attached video.**

## H.  List the web sources used to acquire data or segments of third-party code to support the application. Ensure the web sources are reliable.

Avcontentteam. (2023, March 30). *Simple guide to logistic regression in R and Python*. Analytics Vidhya. <https://www.analyticsvidhya.com/blog/2015/11/beginners-guide-on-logistic-regression-in-r/>

*Bivariate and multivariate analysis*. (n.d.). Cross Validated. <https://stats.stackexchange.com/questions/603377/bivariate-and-multivariate-analysis>

Galarnyk, M. (2022a, April 27). Logistic Regression using Python (scikit-learn) - Towards Data Science. *Medium*. <https://towardsdatascience.com/logistic-regression-using-python-sklearn-numpy-mnist-handwriting-recognition-matplotlib-a6b31e2b166a>

*How to calculate logistic regression accuracy*. (n.d.). Stack Overflow. <https://stackoverflow.com/questions/47437893/how-to-calculate-logistic-regression-accuracy>

*Is reporting univariate regression next to multivariate regression acceptable methodology?* (n.d.). Cross Validated. <https://stats.stackexchange.com/questions/576661/is-reporting-univariate-regression-next-to-multivariate-regression-acceptable-me>

Li, S. (2019a, February 27). Building a logistic regression in Python, step by step. *Medium*. <https://towardsdatascience.com/building-a-logistic-regression-in-python-step-by-step-becd4d56c9c8>

Navlani, A. (2019, December 16). *Understanding logistic Regression in Python Tutorial*. <https://www.datacamp.com/tutorial/understanding-logistic-regression-python>

*Python Machine Learning - Logistic Regression*. (n.d.). <https://www.w3schools.com/python/python_ml_logistic_regression.asp>

Python, R. (2023, June 26). *Logistic regression in Python*. <https://realpython.com/logistic-regression-python/>

Simplilearn. (2023, February 16). *What is a Confusion Matrix in Machine Learning?* Simplilearn.com. <https://www.simplilearn.com/tutorials/machine-learning-tutorial/confusion-matrix-machine-learning>

Zach. (2021, September 1). *How to create a confusion matrix in Python*. Statology. <https://www.statology.org/confusion-matrix-python/>

## I.  Acknowledge sources, using in-text citations and references, for content that is quoted, paraphrased, or summarized.

D208 Task 1 – Laurie Narcisse (Dec 4, 2023) – I utilized a lot of code and paraphrased.