Performance Assessment: Logistic Regression Modeling

Gabriela Howell

Master of Science Data Analytics, Western Governors University

D208 – Logistic Regression Modeling

Professor Choudhury

May 5, 2024

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.

The dataset I am treating is the Medical data, and the main question I am asking is, "Which factors contribute to the likelihood of patient readmission?". This matters in real life because hospitals can use this information to predict and prevent patients from having to come back to the hospital too soon.

2. Define the goals of the data analysis.

Straight with the research question "Which factors contribute to the likelihood of patient readmission?", logistic regression's purpose is to identify and analyze the several factors that influence the probability of patients being readmitted to a hospital. Through this analysis, the goal is to discover significant predictors and understand their impact on the likelihood of readmission, providing valuable insights for healthcare professionals to improve patient care and potentially reduce readmission rates.

Part II: Method Justification

B. Describe logistic regression methods by doing the following:

1. Summarize four assumptions of a logistic regression model.

The assumptions of logistic regression encompass various key points: Firstly, the response variable is binary, representing outcomes such as “yes” or “no,” or “true” and “false.” Secondly, observations are assumed to be independent of each other. Thirdly, there should be no multicollinearity among the explanatory variables, typically assessed using techniques like Variance Inflation Factor (VIF). Lastly, a large sample size is preferable for reliable estimation and inference (Bobbitt, 2020).

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

I will be using Python for logistic regression analysis as there are two foremost benefits. Firstly, its user-friendly nature and flexibility, owing to libraries like scikit-learn and statsmodels, simplify model setup and evaluation, making the process more accessible. Secondly, Python possesses an excess of data analysis tools such as matplotlib and seaborn, which facilitate thorough exploration and comprehension of datasets. Moreover, the interactive capabilities of Jupyter notebooks enhance the analysis process, enabling dynamic examination of data and seamless report generation. Python elevates logistic regression analysis, offering a smoother and more insightful journey from model creation to result interpretation.

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

Logistic regression is the right choice for my research question because it is tailored for situations trying to predict a “yes” or “no” results based on different factors. Since I want to understand what makes it more likely for patients to be readmitted to the hospital, logistic regression can help me with that. It looks at things like my age, medical history, and treatments to see how they affect my chances of being readmitted. This method gives me clear insights into what’s influencing readmission rates, making it a useful tool for my research.

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.

The first step in the cleaning process is to ensure there are no duplicates, outliers, and missing values. All of which were done and came back with nothing. As outliers seemed reasonable to keep. These steps align with the research question to allow data to follow the Income to the various data points, I included all these steps in my Python code which will be attached to this document.

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.

For this research, the dependent variable is "Readmission," while the independent variables include 'Age', 'Gender', 'Income', 'Marital', 'Children', 'VitD_levels', 'Doc_visits', 'Asthma','Arthritis', 'TotalCharge', 'Soft_drink', 'Initial_admin', 'HighBlood', 'Stroke', 'Complication_risk', and 'Services'. It is notable that not all variables are numeric; some are categorical. The dependent variable, 'Readmission', which is binary, and several of the independent variables are categorical as well. Below, you'll find a summary of the numerical variables and frequency counts for the categorical variables.

Summary statistics for X (numerical features only):

	Age	Income	Children	VitD_levels	Doc_visits
count	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000
mean	53.511700	40490.495160	2.097200	17.964262	5.012200
std	20.638538	28521.153293	2.163659	2.017231	1.045734
min	18.000000	154.080000	0.000000	9.806483	1.000000
25%	36.000000	19598.775000	0.000000	16.626439	4.000000
50%	53.000000	33768.420000	1.000000	17.951122	5.000000
75%	71.000000	54296.402500	3.000000	19.347963	6.000000
max	89.000000	207249.100000	10.000000	26.394449	9.000000

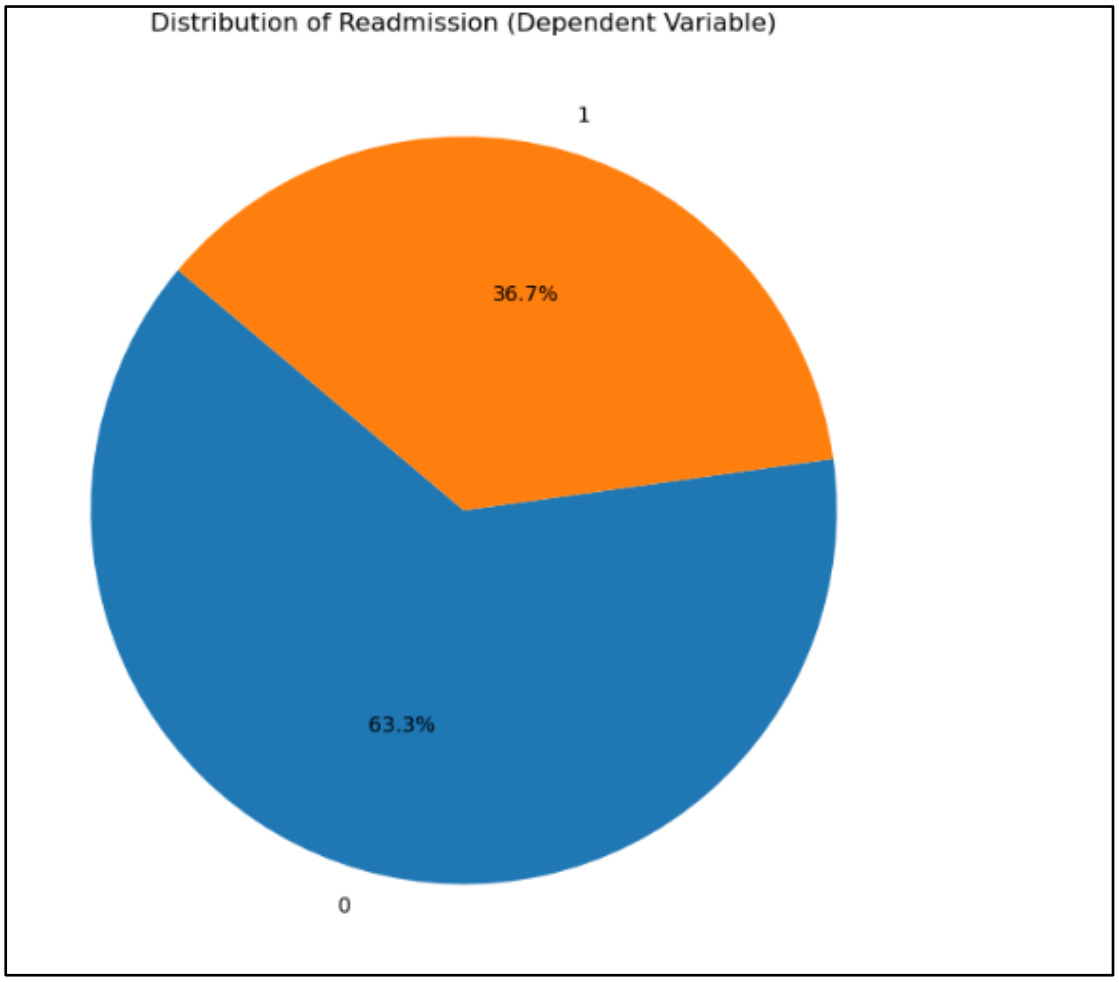
	TotalCharge
count	10000.000000
mean	5312.172769
std	2180.393838
min	1938.312067
25%	3179.374015
50%	5213.952000
75%	7459.699750
max	9180.728000

```
Summary statistics for Y:
ReAdmis
No     6331
Yes    3669
Name: count, dtype: int64

Summary statistics for X (categorical features only):

Gender:
Gender
Female       5018
Male         4768
Nonbinary     214
Name: count, dtype: int64

Marital:
Marital
Widowed          2045
Married          2023
Separated        1987
Never Married    1984
Divorced         1961
Name: count, dtype: int64

Asthma:
Asthma
No     7107
Yes    2893
Name: count, dtype: int64
```

```
Arthritis:
Arthritis
No     6426
Yes    3574
Name: count, dtype: int64

Soft_drink:
Soft_drink
No     7425
Yes    2575
Name: count, dtype: int64

Initial_admin:
Initial_admin
Emergency Admission      5060
Elective Admission       2504
Observation Admission    2436
Name: count, dtype: int64

HighBlood:
HighBlood
No     5910
Yes    4090
Name: count, dtype: int64

Stroke:
Stroke
No     8007
Yes    1993
Name: count, dtype: int64

Complication_risk:
Complication_risk
Medium    4517
High      3358
Low       2125
Name: count, dtype: int64

Services:
Services
Blood Work     5265
Intravenous    3130
CT Scan        1225
MRI             380
Name: count, dtype: int64
```

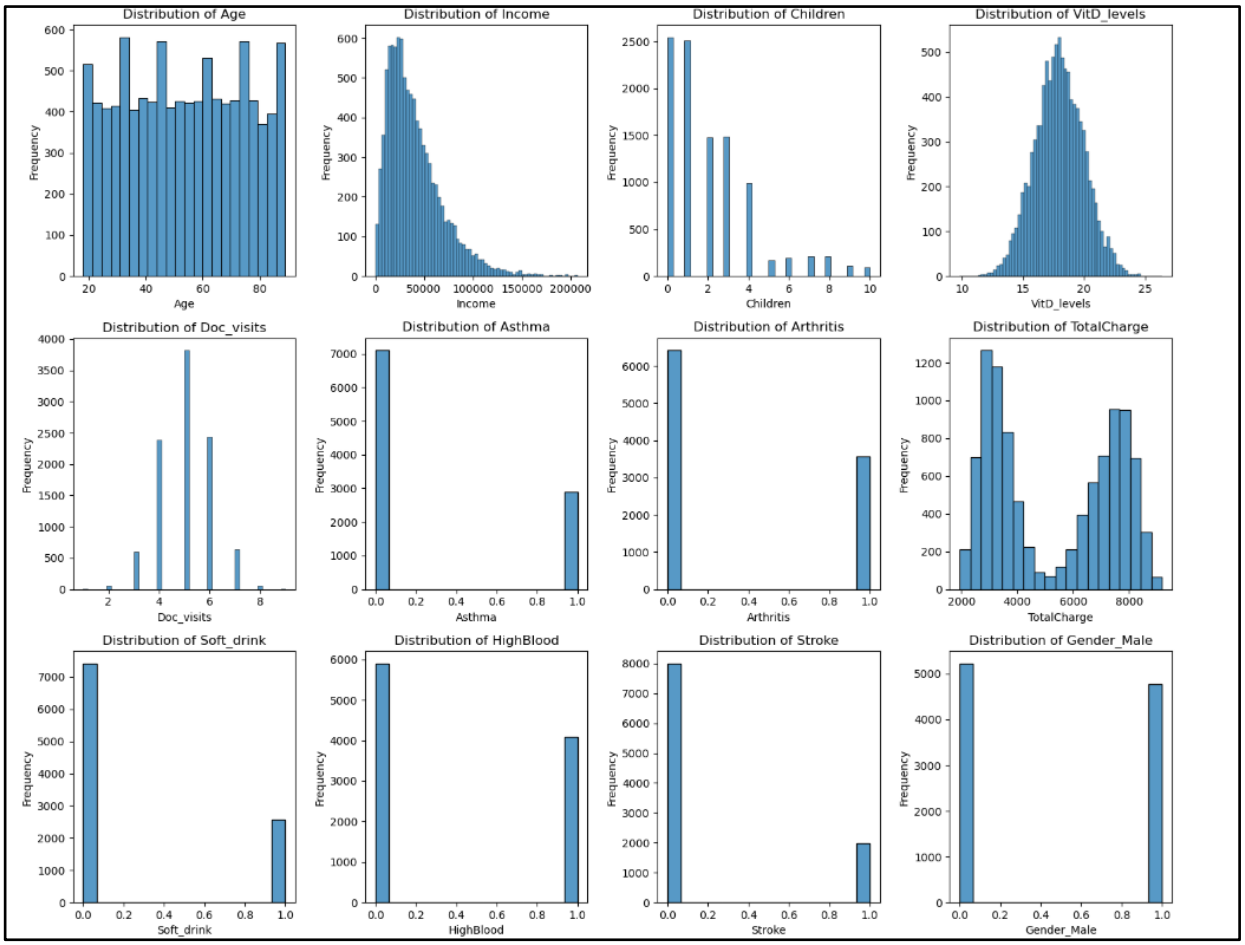
3. Generate univariate and bivariate visualizations of the distributions of the dependent and independent variables, including the dependent variable in your bivariate visualizations.

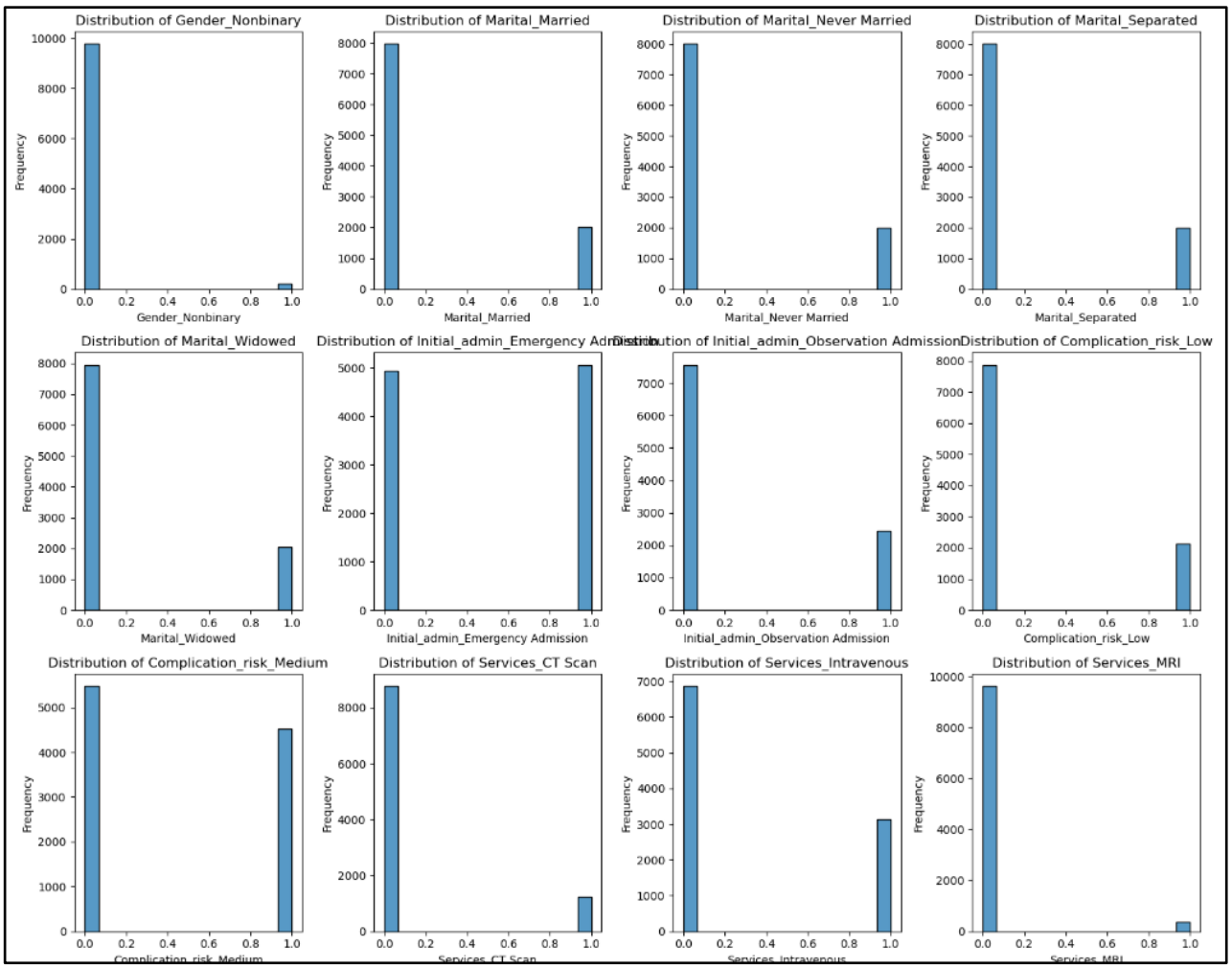
The pie chart depicting the dependent variable “Readmission” illustrates two distinct categories: 0 and 1. Here, 0 signifies no readmissions, while 1 indicates the presence of readmissions.



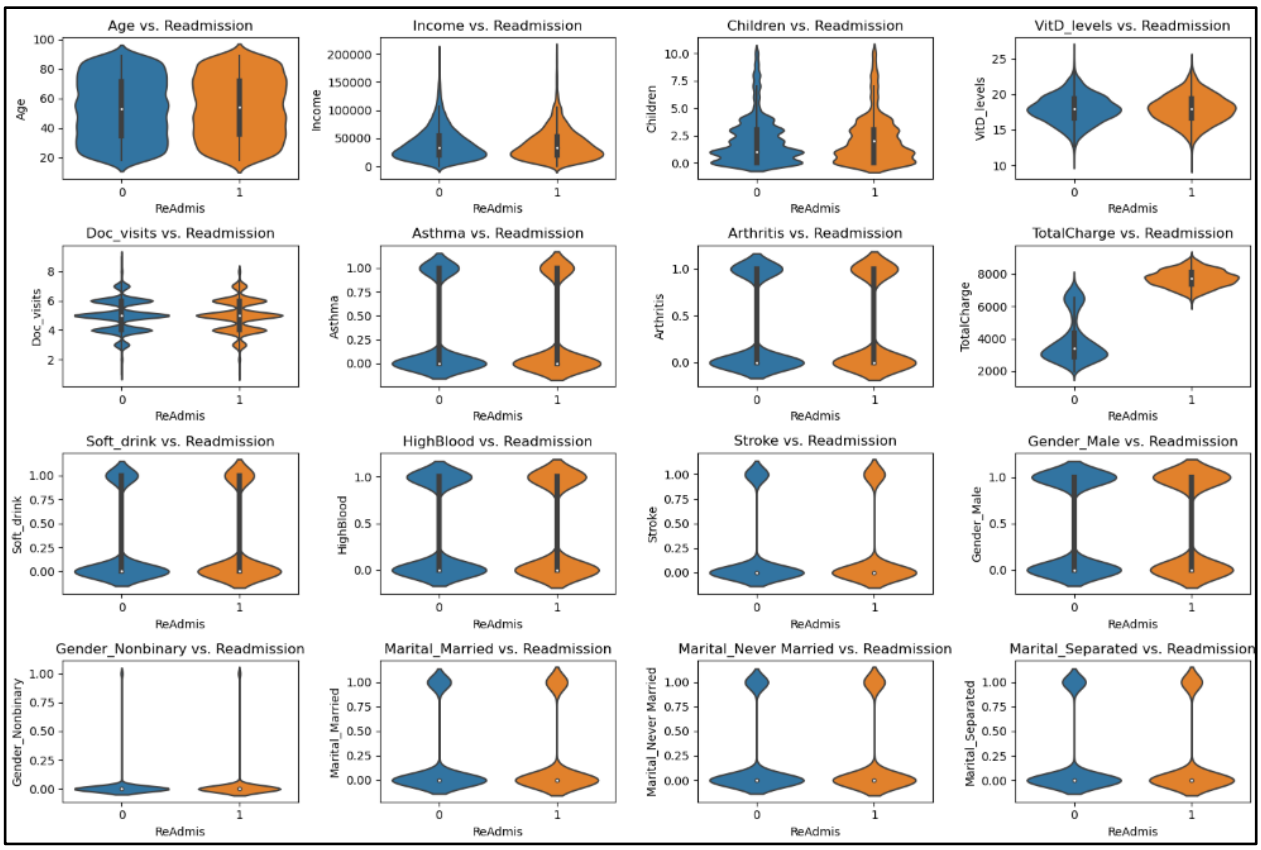
Additionally, a total of twenty-four independent variables underwent individual examination through univariate analysis. These variables include ‘Age’, ‘Income’, ‘Children’, ‘VitD_levels’, ‘Doc_visits’, ‘Asthma’, ‘Arthritis’, ‘TotalCharge’, ‘Soft_drink’, ‘HighBlood’, ‘Stroke’,

‘Gender_Male’, ‘Gender_Nonbinary’, ‘Marital_Married’, ‘Marital_Never Married’, ‘Marital_Separated’, ‘Marital_Widowed’, ‘Initial_admin_Emergency Admission’, ‘Initial_admin_Observation Admission’, ‘Complication_risk_Low’, ‘Complication_risk_Medium’, ‘Services_CT Scan’, ‘Services_Intravenous’, and ‘Services_MRI’. Each of these variables presents distinctive patterns as observed in their respective graphical representations.





For bivariate analysis, violin plots were employed to visualize the relationship between independent variables and the dependent variable, 'Readmission'. This choice was made to effectively represent the distribution of 'Readmission', which is binary (0 or 1), in conjunction with the independent variables.



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.

The initial transformation plan is to include categorical variables in the logistic regression analysis, meaning I must convert them into numerical data.. Categorical variables such as ‘Services’, representing various medical services, and ‘Marital’, indicating marital status, contain non-numeric values unsuitable for mathematical computations. So, converting these categorical variables into numerical ones enables statistical analyses and the training of machine learning models.

This transformation process serves several purposes. Firstly, it ensures uniformity in the dataset's format. Additionally, this step is crucial as it aligns with the assumptions of logistic regression, which inherently necessitate numerical values. Logistic regression relies on a linear relationship between the log odds of the outcome and independent variables, making numerical representations vital for calculating coefficients and predictions in the dataset.

Furthermore, transforming categorical variables into numerical form, using techniques like one-hot encoding (creating dummy variables) or ordinal encoding (mapping categories to numerical values), enables the model to capture inherent relationships between categories. For instance, if a categorical variable denotes different severity levels (e.g., 'Low', 'Medium', 'High'), numerical encoding preserves the ordinal relationship between these levels, crucial for certain analyses.

In the process of transforming categorical variables into numerical data for logistic regression analysis, the aim is to ensure comparability and meet the assumptions of the model. These transformations are implemented using Python libraries like scikit-learn and numpy, enhancing the reliability and interpretability of the analysis for logistic regression.

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

The provided CSV file is called 'Data_using.csv'.

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.

Through my logistic regression analysis, the goal is to gain insight into the likelihood of a patient being readmitted to the hospital. Starting with investigating the AIC value, which is at 3668.59, this mathematical method checks how well the model fits the data. Reviewing the pseudo-R-squared value of the model is at 0.7246, which indicates that a meaningful section of the variability in readmission is reported for by the model. When focusing on the p-value, key factors like 'Age', 'Income', 'VitD_levels', 'Doc_visits', 'Asthma', 'Arthritis', 'TotalCharge', 'HighBlood', 'Gender_Male', 'Marital_Married', 'Marital_NeverMarried', 'Marital_Separated', 'Marital_Widowed', 'Initial_admin_Emergency Admission', 'Initial_admin_Observation Admission', 'Complication_risk_Low', 'Complication_risk_Medium' and 'Services_Intravenous' demonstrate considerable impacts on readmission probabilities based on their low p-values. However, certain variables like 'Children', 'Soft_drink', 'Stroke', 'Gender_Nonbinary', 'Services_CT Scan' and 'Services_MRI do not seem to make much difference in predicting readmission, even though they're in the model.

```
Optimization terminated successfully.
         Current function value: 0.181029
         Iterations 9
                           Logit Regression Results
==============================================================================
Dep. Variable:                 ReAdmis   No. Observations:                10000
Model:                           Logit   Df Residuals:                     9976
Method:                            MLE   Df Model:                           23
Date:                Mon, 20 May 2024   Pseudo R-squ.:                  0.7246
Time:                        19:44:39   Log-Likelihood:                -1810.3
converged:                       True   LL-Null:                       -6572.9
Covariance Type:            nonrobust   LLR p-value:                     0.000
=================================================================================================================
                                      coef    std err          z      P>|z|      [0.025      0.975]
-----------------------------------------------------------------------------------------------------------------
Age                                -0.0168      0.002     -8.131      0.000      -0.021      -0.013
Income                          -6.987e-06   1.45e-06     -4.810      0.000   -9.83e-06   -4.14e-06
Children                           -0.0165      0.019     -0.867      0.386      -0.054       0.021
VitD_levels                        -0.5465      0.019    -28.050      0.000      -0.585      -0.508
Doc_visits                         -0.5631      0.040    -14.206      0.000      -0.641      -0.485
Asthma                             -0.4042      0.094     -4.321      0.000      -0.588      -0.221
Arthritis                          -0.4927      0.088     -5.593      0.000      -0.665      -0.320
TotalCharge                         0.0023   6.26e-05     37.462      0.000       0.002       0.002
Soft_drink                         -0.0303      0.097     -0.311      0.756      -0.221       0.160
HighBlood                          -0.4105      0.086     -4.764      0.000      -0.579      -0.242
Stroke                             -0.0154      0.105     -0.146      0.884      -0.221       0.191
Gender_Male                        -0.2297      0.086     -2.686      0.007      -0.397      -0.062
Gender_Nonbinary                   -0.1018      0.294     -0.346      0.730      -0.679       0.475
Marital_Married                    -0.6578      0.142     -4.641      0.000      -0.936      -0.380
Marital_Never Married              -0.8407      0.140     -5.999      0.000      -1.115      -0.566
Marital_Separated                  -0.8373      0.138     -6.053      0.000      -1.108      -0.566
Marital_Widowed                    -0.6875      0.140     -4.914      0.000      -0.962      -0.413
Initial_admin_Emergency Admission  -1.1694      0.109    -10.702      0.000      -1.384      -0.955
Initial_admin_Observation Admission -0.3649     0.117     -3.111      0.002      -0.595      -0.135
Complication_risk_Low               0.2233      0.113      1.972      0.049       0.001       0.445
Complication_risk_Medium            0.5823      0.099      5.884      0.000       0.388       0.776
Services_CT Scan                    0.1189      0.137      0.869      0.385      -0.149       0.387
Services_Intravenous               -0.2390      0.095     -2.525      0.012      -0.425      -0.054
Services_MRI                       -0.0366      0.211     -0.173      0.862      -0.450       0.377
=================================================================================================================
```

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 my analysis, I employed the Variance Inflation Factor (VIF) to assess multicollinearity among predictor variables in the logistic regression model. The VIF values help me understand how close variables are correlated with each other. This allows me to identify if there might be a problem called multicollinearity, where variables are too related. For instance, variables such as

‘VitD_levels’ and ‘Doc_visits’ exhibited high VIF values, indicating strong correlations with other predictors in the model. Through an iterative process of removing variables with the highest VIF, I examined the remaining VIF values to ensure none exceeded the threshold of 5, signifying significant multicollinearity. This is the order of VIF establish a total of 4 rounds: The order of VIF values identified four rounds of iterations:

1. ‘VitD_levels’: Approximately 32.407
2. ‘Doc_visits’: Approximately 13.5156
3. ‘Age’: Approximately 6.001
4. ‘TotalCharge’: Approximately 5.332

After completing these iterations, the model retained 19 variables, each with a VIF under 2.65, demonstrating significant improvement from the initial high VIF of 32 associated with ‘VitD_levels’.

Besides VIF, I used backward elimination to remove variables with high p-values in refining my model. This iterative process allowed me to restructure the model by eliminating predictors that were not statistically significant. By doing this, I ensured that the final model contained only the most influential predictors, enhancing its interpretability and predictive performance. This way of doing things fits my research question because it is all about evaluating and retaining the important factors that affect what I'm studying.

The first round of iterations started with p values up to 0.887 Which belongs to ‘Gender_Nonbinary’ which was then removed and remodeled to look over for other high p-values. Subsequent iterations involved examining and eliminating variables with progressively

lower but still significant p-values. After 13 rounds of iteration, all remaining predictors had p-values below 0.05, indicating their statistical significance in the model. Below are the variables remaining after each iteration along with their respective p-values:

- Income -0.000004
- Asthma -0.162037
- Initial_admin_Emergency Admission -0.155247
- Initial_admin_Observation Admission -0.227880
- Complication_risk_Low -0.151808
- Complication_risk_Medium -0.152608
- Services_Intravenous -0.176650

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 reduced logistic regression model indicates a successful optimization process, with the model's current function value being 0.659460 and convergence achieved in just 4 iterations. This model was built with 10,000 observations and includes 5 predictor variables. However, the model's pseudo-R-squared value is negative (-0.003309), which suggests that the model descriptions for only a small portion of the variability in the dependent variable, ReAdmis.

Looking at the coefficients of the predictor variables, we observe that 'Income', 'Asthma', 'Initial_admin_Emergency Admission', 'Initial_admin_Observation Admission', 'Complication_risk_Low', 'Complication_risk_Medium ' and

'Services_Intravenous' are retained in the model. These coefficients represent the estimated effect of every predictor on the log odds of the probability of readmission. For example, a one-unit increase in 'Income' is associated with a decrease of approximately 3.796e-06 in the log odds of readmission, holding all other predictors constant.

Overall, the reduced model provides a simplified yet statistically sound representation of the relationship between predictor variables and the likelihood of readmission. The significance of the retained predictors aligns with the research question, indicating that factors such as income, complication risk, gender, marital status, initial administration type, and asthma diagnosis play meaningful roles in predicting readmission probabilities.

Reduced model:

Logit Regression Results

Dep. Variable:	ReAdmis	No. Observations:	10000
Model:	Logit	Df Residuals:	9993
Method:	MLE	Df Model:	6
Date:	Wed, 29 May 2024	Pseudo R-squ.:	-0.004416
Time:	23:05:13	Log-Likelihood:	-6601.9
converged:	True	LL-Null:	-6572.9
Covariance Type:	nonrobust	LLR p-value:	1.000

	coef	std err	z	P>|z|	[0.025	0.975]
Income	-3.796e-06	6.47e-07	-5.866	0.000	-5.06e-06	-2.53e-06
Asthma	-0.1620	0.045	-3.606	0.000	-0.250	-0.074
Initial_admin_Emergency Admission	-0.1552	0.042	-3.709	0.000	-0.237	-0.073
Initial_admin_Observation Admission	-0.2279	0.053	-4.333	0.000	-0.331	-0.125
Complication_risk_Low	-0.1518	0.054	-2.827	0.005	-0.257	-0.047
Complication_risk_Medium	-0.1526	0.042	-3.605	0.000	-0.236	-0.070
Services_Intravenous	-0.1767	0.044	-4.032	0.000	-0.263	-0.091

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:

During my logistic regression analysis, I initially applied a logistic regression model to a dataset comprising 10,000 observations which contained 24 predictor variables along with the target variable for readmission. Upon fitting, the model exhibited a pseudo-R-squared value of 0.7246 and an AIC value of 3668.59. The pseudo-R-squared value indicates of the model's explanatory power, showing how well the predictor variables explain the variability in the target variable. The AIC (Akaike Information Criterion) value measures the model's goodness of fit, with lower values indicating a better fit relative to the complexity of the model.

After completing feature selection using backward elimination which is based on p-values, the reduced logistic regression model preserved 6 significant predictor variables. This reduction aimed to simplify the model while maintaining predictive accuracy. The reduced model displayed a pseudo-R-squared value of -0.004084 and an AIC value of 13227.85, indicating a decline in explanatory power however, hypothetically improved efficiency and interpretability. Regardless of the higher AIC and lower pseudo-R-squared values, the reduced model's simplicity and focus on statistically significant predictors may enhance its practical applicability and ease of understanding.

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

```
Accuracy of logistic regression classifier on test set: 0.64

Confusion matrix:
[[1934    0]
 [1066    0]]
```

The reduced logistic regression model portrayed various key outputs which will be discussed. First, by examining the confusion matrix, which looks at how well the model predicted readmissions, I found that out of 3000 cases, 1934 were correctly predicted as not being readmitted, while 1066 cases were mistakenly labeled as not readmitted when they were. These results indicate the model did well in identifying non-readmissions but struggled with predicting cases of readmission. Overall, the model's accuracy was 64%, meaning that the reduced model performed better than half.

3.Provide an executable error-free copy of the code used to support the implementation of the logistic regression models.

The Python code used is attached to this file called, "Gab -D208_Part2.ipynb".

Part V: Data Summary and Implications

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

1.Discuss the results of your data analysis:

According to the article, "20 Questions to Test Your Skills on Logistic Regression" (Analytics Vidhya, 2021), the regression equation for the reduced model delivers insights into the relationships between predictor variables and the likelihood of readmission. For example, a one-unit increase in Income is associated with a decrease in the log-odds of readmission by 0.000004, individuals diagnosed with Asthma exhibit a larger decrease in log-odds (-0.14959). These coefficients assist in understanding the relative importance of predictors in predicting readmission likelihood. Despite the reduction in explanatory power compared to the initial

model, the reduced model maintains statistical significance, as indicated by significant coefficients and a reasonable accuracy of 64%. However, it is crucial to acknowledge potential limitations such as omitted variable bias and model assumptions.

In the reduced logistic regression model, the equation is constructed as follows:

Logistic Regression Equation: ReAdmis = -0.000004 * Income - 0.162037 * Asthma - 0.155247 * Initial_admin_Emergency_Admission - 0.227880 * Initial_admin_Observation_Admission - 0.151808 * Complication_risk_Low - 0.152608 * Complication_risk_Medium - 0.176650 * Services_Intravenous

Now, let's interpret the coefficients of the reduced model:

- Holding all other variables constant, for one unit increase in patient income, the odds of a patient having readmission decrease by 0.0004%.
- Holding all other variables constant, a patient with asthma has a decrease in their odds of readmission by 14.9590%.
- Holding all other variables constant, a patient with initial_admin_emergency admission has a decrease in their odds of readmission by 14.3797%.
- Holding all other variables constant, a patient with initial_admin_observation admission has a decrease in their odds of readmission by 20.3780%.

- Holding all other variables constant, a patient with complication_risk_low has a decrease in their odds of readmission by 14.0846%.
- Holding all other variables constant, a patient with complication_risk_medium has a decrease in their odds of readmission by 14.1534%.

- Holding all other variables constant, a patient with services_intravenous has a decrease in their odds of readmission by 16.1927%.

Reviewing these interpretations will provide insights into the impact of each variable on the likelihood of readmission in this dataset. Adjustments made from the initial model to the reduced logistic regression model, will emphasize a model that is defined and follows logistic regression assumptions. The significance of the reduced model has all variables with a p-value less than 0.05, making them statistically significant. The model now converges faster, taking just four rounds. However, the pseudo-R-squared value, now at -0.003309, suggests it is not performing better than a model with no predictors. In essence, log-likelihood is at -6594.6, which means the model fits the data, the best it can with these features.

The AIC value, at 13227.85, allows easy comparison of models. The lower the values mean the better models. It balances how well the model fits the data and its complexity. My reduced model, despite its significant variables and fast convergence, might not be the best. However, it is crucial to see how well it predicts real outcomes beyond these numbers.

According to a Geeks for Geeks article, there are certain limitations associated with logistic regression. If the number of observations is fewer than the number of features, logistic regression may not be suitable as it could result in overfitting. Logistic regression is only applicable for predicting discrete outcomes, as its dependent variable is restricted to a discrete number set. Additionally, logistic regression cannot manage non-linear problems due to its linear decision surface. However, non-linearly separable data is common in real-world situations (GeeksforGeeks, 2023).

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

Based on the outcomes from the reduced logistic regression model, unfortunately, I conclude that this model lacks practical significance. However, I can obtain some insightful recommendations for effectively reducing readmissions. To start, fundamentally understanding how each variable affects the likelihood of patient readmission is essential. For example, while changes in income do not seem to affect readmission chances (implied by having a coefficient of 1.0), variables like services_intravenous and initial_admin_observation exhibit hopeful reductions in readmission odds with each additional case. Furthermore, the other variables exhibit similar trends, displaying reductions in readmission odds.

While the initial research question expected to identify variables associated with the likelihood of patient readmission, the results took a surprising turn. Opposite to expectations, a discovery of certain variables is reducing readmission rates. For example, conditions like asthma and initial admin observation, along with intravenous services, appear to play a significant role in lowering readmission odds. This presents an opportunity for learning from our results and understanding the practices or interventions associated with these variables. My recommendation is to integrate these effective strategies to all sectors of the hospital to maintain low readmission rates.

Part VI: Demonstration

G. Panopto video recording

The Panopto video can be found using this link: Gabriela Howell D208 Part 2 Video

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

Goyal, C. (2022, June 24). *20+ questions to test your skills on logistic regression*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/05/20-questions-to-test-your-skills-on-logistic-regression/

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

Scikit. (n.d.). https://scikit-learn.org/stable/auto_examples/index.html#examples

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

Bobbitt, Z. (2020, October 13). *The 6 assumptions of logistic regression (with examples)*. Statology. https://www.statology.org/assumptions-of-logistic-regression/

GeeksforGeeks. (2023, January 10). *Advantages and disadvantages of logistic regression*. https://www.geeksforgeeks.org/advantages-and-disadvantages-of-logistic-regression/

Goyal, C. (2022, June 24). 20+ questions to test your skills on logistic regression. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/05/20-questions-to-test-your-skills-on-logistic-regression/