|  |
| --- |
| D208 |
| Logistic Regression Modeling |
| Task 2 |

|  |
| --- |
| Shantel Johnson  3-30-2023 |

Contents

[Section A 3](#_Toc130665641)

[Part 1: Research Question 3](#_Toc130665642)

[This information could prove practical for the allocation of government funds to areas with higher incidences of factors that increase the additional charges billed. 3](#_Toc130665643)

[Part 2: Goals 3](#_Toc130665644)

[Section B 3](#_Toc130665645)

[Part 1: Assumptions of Linear Regression 3](#_Toc130665646)

[Part 2: Python for Data Analysis 3](#_Toc130665647)

[Part 3: Justification of Methods 3](#_Toc130665648)

[Section C 4](#_Toc130665649)

[Part 1: Data Cleaning 4](#_Toc130665650)

[Duplication 4](#_Toc130665651)

[Missing Values 4](#_Toc130665652)

[Outliers 4](#_Toc130665653)

[Part 2: Data Descriptions 4](#_Toc130665654)

[Part 3: Visualization 4](#_Toc130665655)

[Part 4: Data Transformation 4](#_Toc130665656)

[Ordinal Encoding 4](#_Toc130665657)

[Nominal Encoding 5](#_Toc130665658)

[Part 5: Output 5](#_Toc130665659)

[Section D 5](#_Toc130665660)

[Part 1: Initial Model 5](#_Toc130665661)

[Part 2: Feature Selection 5](#_Toc130665662)

[Part 3: Reduced Model 5](#_Toc130665663)

[Section E 5](#_Toc130665664)

[Part 1: Model Comparison 5](#_Toc130665665)

[Part 2: Output 6](#_Toc130665666)

[Part 3: Code 6](#_Toc130665667)

[Section F 6](#_Toc130665668)

[Part 1: Results 6](#_Toc130665669)

[Regression Equation 6](#_Toc130665670)

[Interpretation 6](#_Toc130665671)

[Statistical Significance 6](#_Toc130665672)

[Practical Significance 7](#_Toc130665673)

[Limitations 7](#_Toc130665674)

[Part 2: Next Steps 7](#_Toc130665675)

[Section G 7](#_Toc130665676)

[Part 1: Data Preparation Code 7](#_Toc130665677)

[Data Cleaning 10](#_Toc130665678)

[Issue Handling 13](#_Toc130665679)

[Data Descriptions 13](#_Toc130665680)

[Visualizations 14](#_Toc130665681)

[Data Transformation 41](#_Toc130665682)

[Part 2: Linear Regression Modeling Code 45](#_Toc130665683)

[Linear Regression Model 1: All Features 45](#_Toc130665684)

[Linear Regression Model 2: Reduced Features 48](#_Toc130665685)

[Section H 53](#_Toc130665686)

[Section I 53](#_Toc130665687)

[Part 1: Web Sources 53](#_Toc130665688)

[Part 2: References 54](#_Toc130665689)

# Section A

## Part 1: Research Question

The dataset for this project comprises observations on patient demographics, medical conditions, and hospital readmission. As the project's data analyst, my responsibility is to answer the research question: What factors cause a patient to be readmitted to the hospital within 30 days of release?

## This information could prove practical because by identifying the factors that drive readmission, hospitals can work towards lowering readmission rates and avoiding penalties from the US Centers for Medicare and Medicaid Services (CMS).

## Part 2: Goals

This project aims to identify factors (i.e., variables) in the dataset that contribute to patient readmission and establish the correlation between them. To achieve these goals, the dataset requires preparation for statistical analysis using logistic regression.

# Section B

## Part 1: Assumptions of Logistic Regression

Four assumptions to be aware of when using a linear regression model are:

* Binary response: Logistic regression assumes that the dependent variable can only take on two possible outcomes (Bobbitt, 2020).
* Independent observations: Observations in the dataset must be independent of one another (Bobbitt, 2020).
* Multicollinearity: Multicollinearity can lead to dubious regression coefficients. As such, predictor variables should be removed from the model until multicollinearity is gone (Bruce, Bruce, & Gedeck, 2019).
* Outliers: Outliers can distort the results of regression modeling. For this reason, predictor variables should have normal distributions (Bruce, Bruce, & Gedeck, 2019).

## Part 2: Python for Data Analysis

For logistic regression in particular, Python is useful because 1) of the availability of packages that are used for data preprocessing and exploratory data analysis (pandas and scipy) and 2) of the availability of packages that are used for building logistic regression models (scikit-learn and statsmodels).

|  |  |
| --- | --- |
| **Package/Library** | **Purpose** |
| pandas | Data manipulation & exploratory data analysis |
| scipy | Exploratory data analysis |
| scikit-learn | Feature engineering & logistic regression modeling |
| statsmodels | Logistic regression modeling |
| matplotlib | Visualization |
| seaborn | Visualization |

## Part 3: Justification of Methods

Like linear regression, logistic regression quantifies the nature of the relationship between one or more predictor variables and the target variable. The key difference between linear and logistic regression, however, is that the outcome for logistic regression is binary (Bruce, Bruce, & Gedeck, 2019). As defined in Section A, Part 1 the research question seeks to understand what factors (the predictor variables) influence whether a patient is readmitted (the target variable). Based on this, we can see that logistic regression is an appropriate technique for analyzing the research question.

# Section C

## Part 1: Data Cleaning

Data cleaning was performed to identify and remediate the common data quality issues of duplication, missing values, and outliers. Refer to Section G, Part 1 to view the associated data cleaning code.

### Duplication

The pandas.DataFrame.duplicated() function was applied on the dataset to return duplicated rows. No issues were detected after using this method.

### Missing Values

The pandas.DataFrame.isnull.sum() function was applied on the dataset to return the number of missing values found in each variable. No issues were detected after using this method.

### Outliers

The scipy.stats.zscore() function was applied on the dataset to return the number of outliers found in each numeric variable. The following data issues were identified:

|  |  |
| --- | --- |
| **Variable** | **Number of Outliers** |
| Population | 218 |
| Children | 202 |
| Income | 143 |
| VitD\_levels | 24 |
| Doc\_visits | 8 |
| Full\_meals\_eaten | 33 |
| vitD\_supp | 70 |

As noted in Section B, Part 1, outliers impact the results of regression modeling. For this reason, records containing outliers were dropped from the dataset using the pandas.DataFrame.drop() function.

## Part 2: Data Descriptions

Summary statistics were calculated for the dependent and all independent variables. Refer to Section G, Part 1 for more information.

## Part 3: Visualization

Distributions of the dependent and independent variables were depicted using univariate and bivariate plots. To view these visualizations, refer to Section G, Part 1.

## Part 4: Data Transformation

Data transformation was performed on categorical variables in preparation for linear regression modeling (refer to Section G, Part 1).

### Ordinal Encoding

Statistical models should be presented with a numeric encoding of ordered categories that represent linear ordering (Kuhn & Johnson, 2019). If a categorical variable consisted strictly of binary values (e.g., yes/no, true false), then these variables were transformed using the pandas.DataFrame.replace() function. Otherwise, they were transformed into numeric values using the sklearn.preprocessing.OrdinalEncoder() function.

### Nominal Encoding

According to Kuhn and Johnson, the most basic approach for re-expressing categorical variables as numeric data is to create C-1 indicator variables, where C represents the possible values of the predictor. (2019). Categorical variables of the nominal type were transformed into numeric values using the pandas.get\_dummies() function.

## Part 5: Output

To view the output of the data preparation tasks performed in Section C, refer to the attached file, medical\_prepared\_data2.csv.

# Section D

## Part 1: Initial Model

An initial logistic regression model was constructed using the independent variables identified in Section C, Part 2. Refer to Section G, Part 2 to view the initial model.

## Part 2: Feature Selection

Feature selection is important in predictive modeling because the predictive performance of the model often decreases as the number of uninformative predictors increases. In addition, removing predictors can reduce the cost of data acquisition and can increase throughput of the software used to make predictions (Kuhn & Johnson, 2019). For this reason, the number of predictor variables was reduced using feature selection.

The sklearn.feature\_selection.SequentialFeatureSelection() function was used to reduce the number of predictor variables included in the model. The sequential feature selector (SFS) uses forward selection to form a feature subset in a greedy fashion; the features are chosen based on the cross-validation score of an estimator, which in the case of logistic regression, is accuracy (Pedregosa, et. al, 2011). The SFS falls under the wrapper methodology for feature selection. According to Kuhn and Johnson, wrapper methods use iterative search procedures that repeatedly supply predictor subsets to the model (2019).

## Part 3: Reduced Model

A reduced logistic regression model was constructed using a subset of the independent variables. Refer to Section G, Part 2 to view the reduced model.

# Section E

## Part 1: Model Comparison

The following table compares the initial and reduced linear regression models:

|  |  |  |
| --- | --- | --- |
| **Model** | **Predictor Variables** | **Accuracy** |
| Initial | 33 | 0.986 |
| Reduced | 8 | 0.981 |

|  |  |  |
| --- | --- | --- |
| **Confusion Matrix:**  **Initial Model** | | |
|  | **y (0)** | **y (1)** |
| **y\_pred (0)** | 5843 | 67 |
| **y\_pred (1)** | 87 | 3359 |

|  |  |  |
| --- | --- | --- |
| **Confusion Matrix:**  **Reduced Model** | | |
|  | **y (0)** | **y (1)** |
| **y\_pred (0)** | 5824 | 86 |
| **y\_pred (1)** | 71 | 3334 |

While the initial model includes significantly more predictor variables than the reduced model, both models have approximately the same accuracy score. This means that the 33 independent variables in the initial model predict the dependent variable with the same accuracy as the 8 independent variables in the reduced model. In addition, the confusion matrices for the initial and reduced models are near identical for predicted and actual values.

## Part 2: Output

To view the output of the analysis, refer to Part 1 of this section.

## Part 3: Code

To review the executable code used to build the initial and reduced linear regression models, refer to the attached file, ‘D208 Task2 Model.csv’.

# Section F

## Part 1: Results

### Regression Equation

The regression equation for the reduced logistic regression model is given below:

### Interpretation

For each predictor variable, the coefficient sign indicates whether the probability of ReAdmis=1 increases or decreases as the predictor variable increases. Stroke, Complication\_risk, Initial\_days, and Emergency\_Admission all have positive coefficients, meaning that as these variables increase, the probability that ReAdmis=1 increases. The coefficient values represent the estimated change in the log-odds of the target variable given a one-unit increase in the predictor variable (Larose & Larose, 2019). The change in log-odds is specified for each variable below:

|  |  |  |
| --- | --- | --- |
| **Variable** | **Odds** | **Meaning** |
| Stroke | exp(1.3984) = 4.05 | A patient is 4.05 times more likely to be readmitted if they have had a stroke. |
| Complication\_risk | exp(0.6286) = 1.87 | As their complication risk increases, a patient is 1.87 times more likely to be readmitted. |
| Initial\_days | exp(1.126) = 3.08 | For every day of their initial stay in the hospital, a patient is 3.08 times more likely to be readmitted. |
| Emergency\_Admission | exp(1.6313) = 5.11 | A patient is 5.11 times more likely to be readmitted if they had an emergency admission. |

For example, to predict the probability that a patient who 1) did not have a stroke, 2) with a medium complication risk, 3) that spent three days in the hospital initially, and 4) had an emergency admission was readmitted, plug the variables into the above regression equation:

Based on this result, the patient will not be readmitted to the hospital.

### Statistical Significance

According to the summary output (refer to Section G, Part 2), the predictor variables included in the reduced model are significant as they all have p-values less than 0.05.

### Practical Significance

The practical significance of this model lies in its ability to identify and quantify the factors that drive hospital readmission (stroke, complication risk, initial length of stay, and emergency admission). By using this model, healthcare providers can proactively identify patients at risk of readmission and implement targeted intervention to prevent it.

### Limitations

The following are limitations associated with the analysis:

* The dataset does not include the feature “reason for hospital admission”. Including this factor as a predictor variable in the model may improve its accuracy score, leading to better predictions regarding a patient's likelihood of readmission.
* The feature "Complication\_risk" is assessed by the primary patient assessment and has values of "low", "medium", and "high". However, it is unclear whether this assessment is administered by a medical professional or self-administered. As a result, it is possible that the values of "Complication\_risk" are not standardized, which could potentially skew the results of the model.

## Part 2: Next Steps

I would first recommend adding “reason for hospital admission” as a feature in the dataset. From there, SFS can be re-performed to determine whether “reason for hospital admission” improves the model. Next, I recommend investigating the administration of the primary patient assessment to better understand the quality of data included in the “Complication\_risk” factor. Once these steps are complete, the data engineering team can work on scaling the model for commercial use.

# Section G

# Section H

To view a walkthrough demonstration of the code referenced in Section G, refer to the following Panopto link:

# Section I

## Part 1: Web Sources

Caswell, et al. (2023, March). doi:10.5281/zenodo.7697899

Gommers, et al. (2023). doi:10.5281/zenodo.7655153

Grisel, et al. (2022). doi:10.5281/zenodo.6543413

Seabold, et al. (2017). doi:10.5281/zenodo.275519

The pandas development team. (2023). doi:10.5281/zenodo.7741580

Waskom, M. (2021). doi:10.5281/zenodo.4645478

## Part 2: References

Bobbitt, Z. (2020, October). *The 6 Assumptions of Logistic Regression (with Examples)*. Retrieved March 2023, from Statology: https://www.statology.org/assumptions-of-logistic-regression/

Bruce, P., Bruce, A., & Gedeck, P. (2019). *Practical Statistics for Data Scientists : 50+ Essential Concepts Using R and Python* (2 ed.). O'Reilly Media, Incorporated. Retrieved March 2023

Kuhn, M., & Johnson, K. (2019). *Feature Engineering and Selection : A Practical Approach for Predictive Models.* CRC Press LLC. Retrieved March 2023

Pedregosa, e. (2011). *scikit-learn*. Retrieved March 2023, from sklearn.feature\_selection.SequentialFeatureSelector: https://scikit-learn.org/stable/modules/generated/sklearn.feature\_selection.SequentialFeatureSelector.html