Logistic Regression

Leticia Lugo

Western Governor University

Case

Hospital readmissions from Medicaid patients alone cost the government $26 billion annually according to Robert Wood Johnson Foundation. According to Tom Prose, CEO at General Medicine, P.C., in his article “Top Reasons for Hospital Readmissions” some reasons to get re-admitted to a hospital are:

* The patients are not well informed about their care, which means patients do not know the extent of their conditions and the expectations of future treatment processes.
* Patients leave their hospital care too soon because of the hospital charges causing a potential extension of their illness.
* Patients naturally require more medical care with age.

(Prose, n.d.)

Research Question

Do Doctor Visits, Age, Income, Gender, Area, Total Charge, Additional Charges, and days spent in a hospital have an influence on hospital readmissions?

Target Variable

The target variable is the binary variable ‘ReAdmis’. ReAdmis is a binary variable that answers whether the patient was readmitted within a month of release or not.

Objective

The analyst should detect the reasons for hospital readmissions. Based on the article of Prose, the following columns were selected for the analysis: doctor visits, age, gender, area, total charge, additional charges, readmissions, and days spent in the hospital. The readmission variable is binary, and the independent variables are a mix of categorical and numerical variables. Therefore, logistic regression is the best fit for the statistical analysis.

Tools

This report will be performed in Python language using jupiter notebook environment due to several reasons compiled on table 1.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Comparisons | Python | R | SAS | Option/Reason |
| Uses | Used for software development, web development, and data science. | Used for data mining and machine learning. | Used for business intelligence, data management, and predictive analytics. | Python/Widely usage. |
| Cost | Open-source programming and free language | Open-source programming and free language | Proprietary software, customers pay to use it. | Python/No cost |
| Speed | A high-level programming language. It is faster for building large applications and web development. | A low-level programming language. It manages extended codes for simple procedures, which results in low speed. | A high-level programming. It uses SAS SQL and automatic code generation with reusable code snippets. | Python/High-level language. |
| Accessibility | It is easy to learn. It uses friendly libraries, and it has become a phenomenon in data analytics. | It has a steep learning curve because it needs a working knowledge of coding. | It is easy to learn, and it has a simple GUI. It does not required knowledge on a prior programming language. | Python/Analyst more familiar with it. |
| Data Handling | It is easy to handle data with popular libraries such as Numpy and Pandas. | It is easy to handle data with packages like plyr, dplyr, and tdyr. However, small tasks will take time to run because it runs on RAM memory. | It is efficient to perform data handling and manipulation using the DATA step, which compiles and runs faster. | Python/Analyst familiar with libraries. |
| Data Visualization | Have packages such as matplotlib, seaborn, and vispy among others that makes it robust in graphical analysis. | Makes visualizations and analysis with packages like ggplot, lattice, ggvis, rgis, among others. | Work still in progress to improve its visualization and graphical capabilities, but it does not yet match Python and R's standards. | Python/Analyst familiar with libraries. |
| Customer and Community Support | It does not have customer support. However, it provides online community support. | It does not have customer support. However, it provides online community support. | It has customer support, and it has an online community to help solve customer’s questions. | Python/Large community support. |
| Popularity | 41% | 30% | 29% | Python/More popular. |

(Comparison of Python, R, and SAS languages)

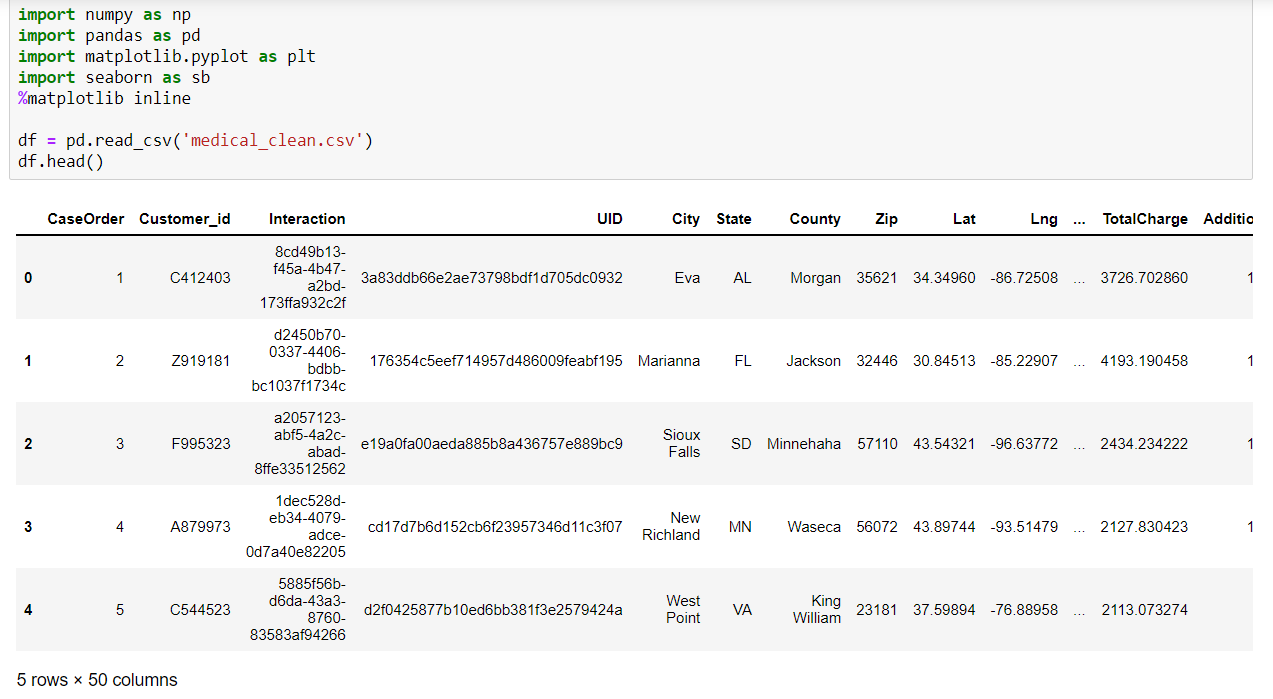
Table 1: Comparison of Python, R, and SAS languages.

Data

The data consist of 10,000 entries of patients readmitted to the hospital with 50 columns of patients’ medical conditions, demographic information, and patient service in the hospital.

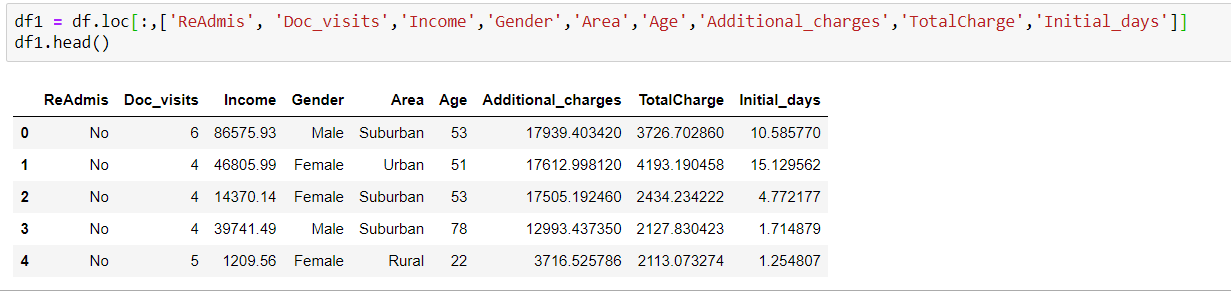
Import the data

The dataset is imported using pandas library to the Python environment.

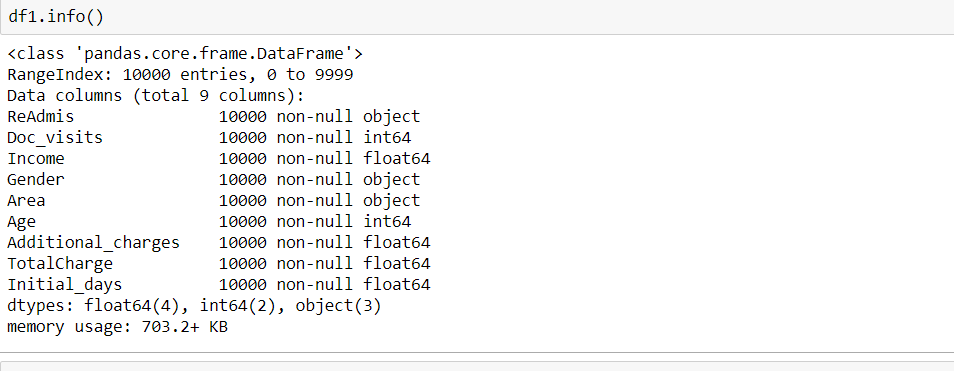


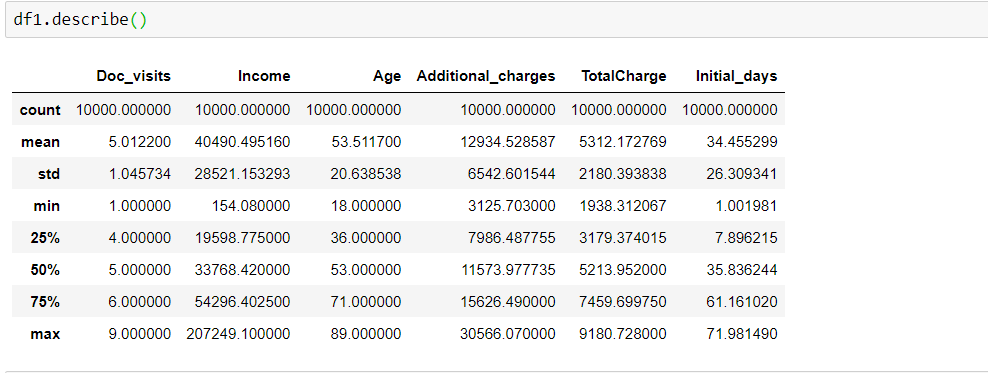
Select the variables and clean data

Based on the research question, the analyst's variables were taken from the dataset to a new dataset name ‘df1’.



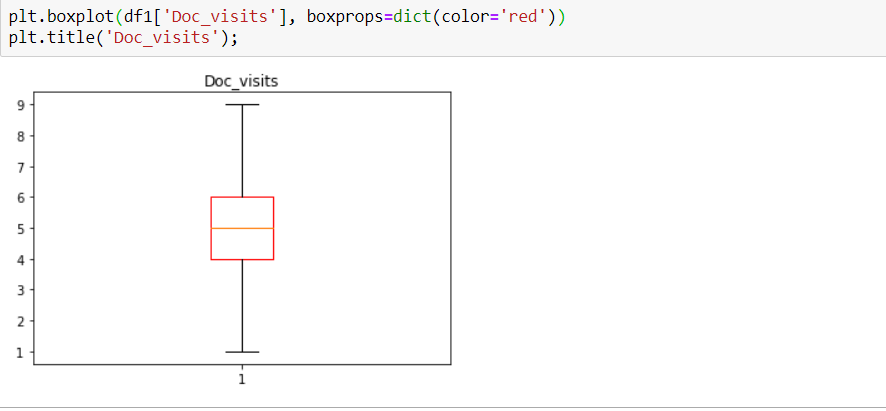
Starting the data cleaning process using info() method of pandas to detect missing values, based on the result table, there are not detected missing values.



After detecting no missing values, the described() method is run to observe the statistical values of the data. This method will only apply to numerical variables. For example, with the table of statistical analysis below, we can see how the mean of days in the hospital is 34, which is about a month. Another insight from the statistics table gives us the average age of 53. The average income is $40,490. The addition of total charges and additional charges is about $19,000, about half of the average income. 

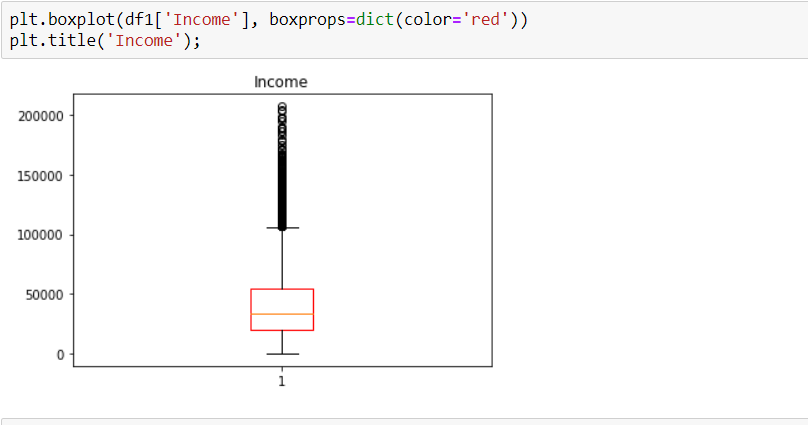
Univariate and Bivariate plots

The first univariate plot indicates the average of doctor visits is about 5. It also shows a maximum of 6 doctor visits and a minimum of 4 doctor visits.



(Matplotlib, n.d.)

The maximum income is about $200,000, and the average income is about $40,000. The plot shows several outliers.



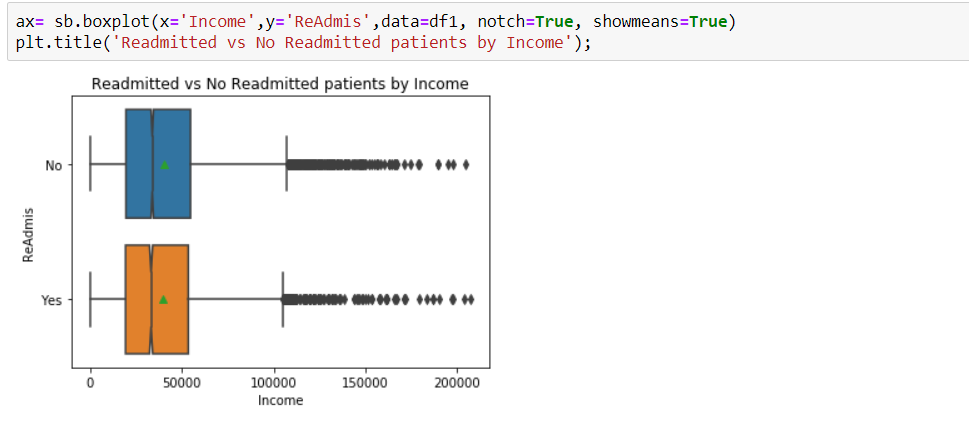
(Matplotlib, n.d.)

The next plot indicates the number of patients readmitted and not readmitted to the hospital. There are more than 6,000 patients not readmitted, and about 3,800 patients readmitted to the hospital.



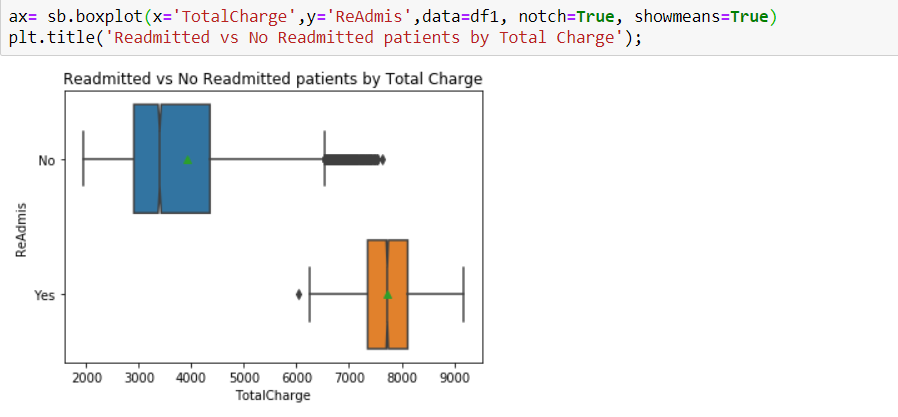
(Seaborn, n.d.)

The bivariate plot below shows the readmitted patients by income. The plot presents several outliers on patient’s higher income for both readmitted and not readmitted patients.



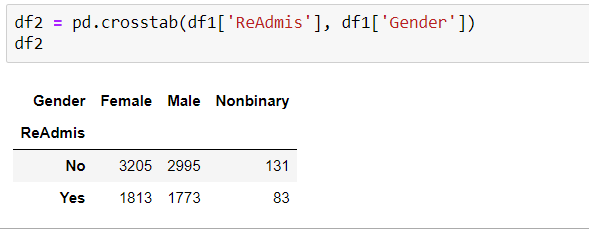
(Seaborn, n.d.)

On the other hand, the patients readmitted to the hospital have a significant difference with their total charges. For patients not readmitted to the hospital the total charge average is about $3,500, and patients readmitted to the hospital have an average of about $7,800.



(Seaborn, n.d.)

The following cross-tabulation table presents the patient readmission by gender. Females and males have a similar total number of patients being readmitted to the hospital.



Statistical Model

Logistic regression is used to model the probability of certain classes of events existing such as yes/no. For the analysis, logistic regression is used because the dependent variable is a binary variable, and the independent variables are a mix of continuous and categorical variables.

(Logistic regression,2021)

Logistic Regression Assumptions

1. Data is free of missing values. Overfitting can occur due to limited data points.
2. The dependent variable is binary (only accepts two values) or ordinal (a categorical variable with ordered values).
3. All predictors are independent of each other. Multicollinearity occurs when high association (correlation) among independent variables.
4. There are at least 50 observations per predictor variable (to ensure reliable results).

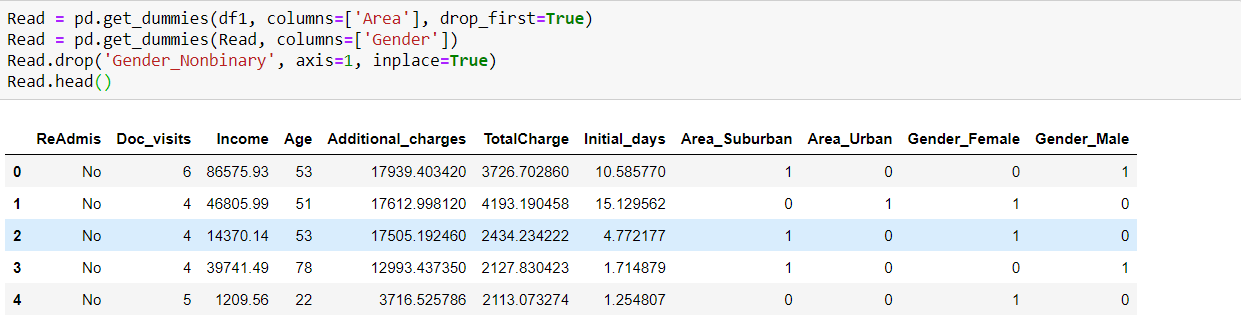
(Kassambara,2018)

Pre-processing the data

Logistic regression analysis requires all independent variables to be numerical. The categorical variables need to be encoded, creating additional dummy variables with all the levels they have. The dummy variables are the conversion from categorical to numerical variables. The dependent variable is encoded separately to convert it to numerical variable.

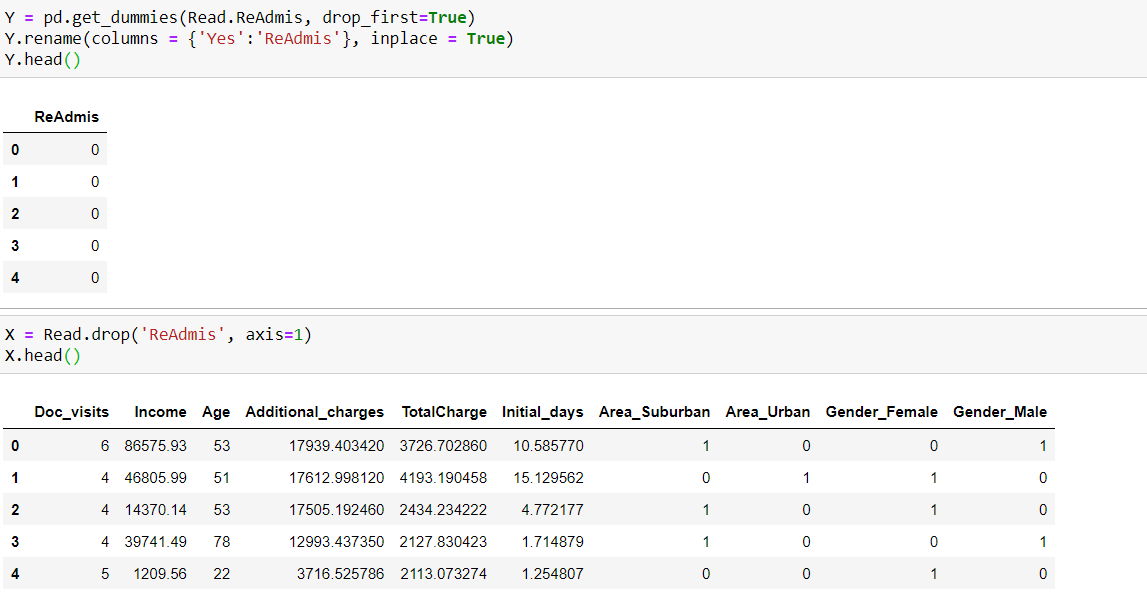
The process of creating dummy variables results in the problem of multicolinearity. The dummy variables tend to be highly correlated because they came from the same categorical variable. To resolve multicolinearity, one of the levels of the categorical variables is dropped.

(Babu, 2020)



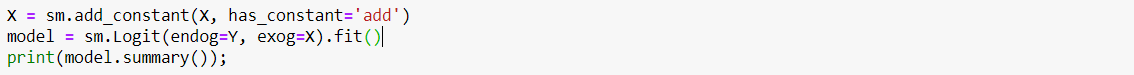
Define X and Y

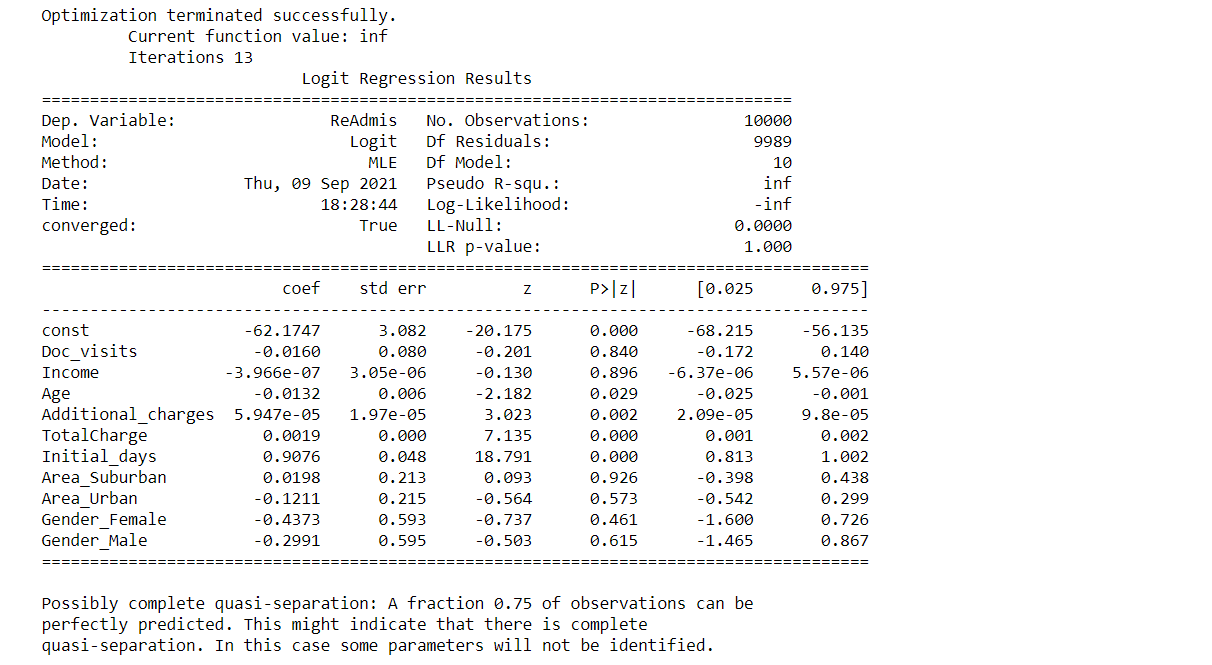
The data features ‘ReAdmis’ as a target variable, which is a categorical variable. The binary target variable needs to be numeric, so it is separated and converted to a numerical variable. The process of conversion starts by transforming the variable into two numerical variables according to its level (Yes/No). Then, the ‘X’ and ‘Y’ datasets are defined from the dataset. The ‘ReAdmis’ values are presented as Yes = 1, and No = 0.



Model

The logistic regression model results are shown below.





The results identified the target variable as ‘ReAdmis’. The model used to fit the data is Logit. The method used is MLE (Maximum Likelihood Estimation). The date and time were captured when the model was created. The convergence of the model is true because the model was able to be executed successfully. The number of observations used is 10,000. Degrees of freedom of the model and residuals are calculated and presented on the summary table. The pseudo R-square measures the proportion of variance for the dependent variable explained by the independent variables.

The p-value results indicate Area\_Suburban is not a significant variable. The variable ‘Area’ was separated into its three different levels, and one was dropped to fix multicolinearity. The Area\_Urban is also removed because the rest of the variables were not part of the model.

(statsmodels, n.d.)

The logistic regression equation is:

Log Odds = -62.1747 – 0.0160(Doc\_visits) – 3.966e-07(Income) – 0.0132(Age) + 5.947e-05(Additional\_charges) + 0.0019(TotalCharge) + 0.9076(Initial\_days) + 0.0198(Area\_Suburban) – 0.1211(Area\_Urban) – 0.4373(Gender\_Female) – 0.2991(Gender\_male)

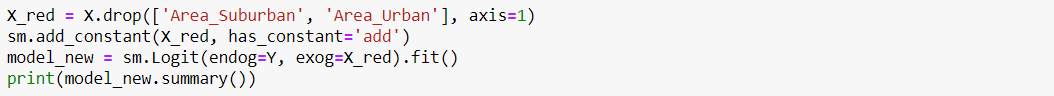
The log odds from the logistic regression equation is log(p/1-p) where ‘p’ is the probability of having the outcome and p/1-p is the odds of the outcome.

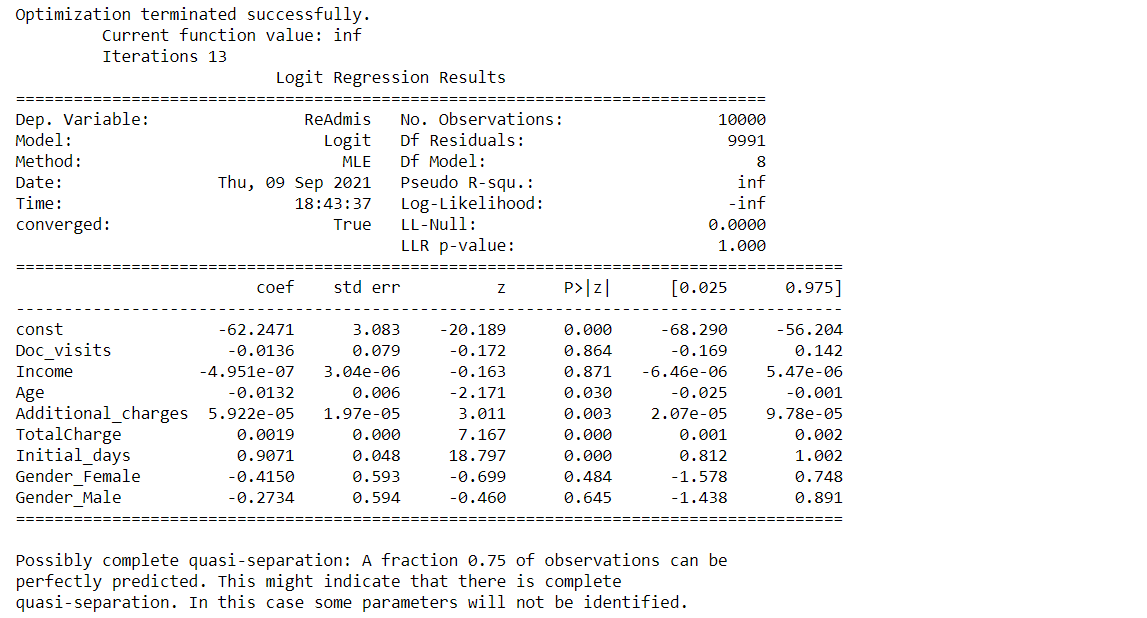
When the independent variables are 0, the log of the odds of having the outcome is the intercept, which is -62.1747. The intercept has a negative sign, meaning the probability of having the outcome would be <0.05. In the analysis the intercept has p-value of 0.00 meaning that it is statistically significant.

The log of the odds of being readmitted into the hospital decreases by 0.0160 per every doctor visit. The log odds of being readmitted into the hospital increases by 0.9076 per day spent in the hospital.

Reduced Model

For the reduced model, the logistic regression is run without the ‘Area’ variable defining a new X dataset named ‘X\_red’.





Log[y/1-y] = -62.2471 – 0.0136(Doc\_visits) – 4.951e-07(Income) – 0.0132(Age) + 5.922e-05 (Additional\_charges) + 0.0019(TotalCharge) + 0.9071(Initial\_days) – 0.4150(Gender\_Female) – 0.2734(Gender\_male)

(statsmodels, n.d.)

Conclusion

As a result of the analysis, we can conclude that the area from where the patient lives has no influence if the patient is readmitted into the hospital. However, the age, gender, doctor visits, income, total charge, additional charges and days spend in the hospital have a significant effect on hospital readmissions.

The results of the two pseudo R-squares and log-likelihood numbers very close to provide the same evaluation of the model. Consequently, the reduced model is selected for its simplicity.

Recommendations for future research

* Use different set of variables to predict the readmission in hospitals.
* Use a different target variable. If it is numerical or categorical, it will need to be process to be use on logistic regression.
* Use different method of regression analysis.
* Split the data in training and test datasets for the model.

Reference

Wikimedia Foundation. (2021, August 23). Logistic regression. Wikipedia. <https://en.wikipedia.org/wiki/Logistic_regression#:~:text=Logistic%20regression%20is%20a%20statistical%20model%20that%20in,a%20logistic%20model%20%28a%20form%20of%20binary%20regression%29>

Kassambara, & U, M. (2018, March 11). Logistic regression assumptions and diagnostics in r. STHDA. <http://www.sthda.com/english/articles/36-classification-methods-essentials/148-logistic-regression-assumptions-and-diagnostics-in-r/#:~:text=Logistic%20regression%20assumptions%20The%20logistic%20regression%20method%20assumes,logit%20of%20the%20outcome%20and%20each%20predictor%20variables>

Visualizing distributions of data¶. Visualizing distributions of data - seaborn 0.11.2 documentation. (n.d.). <https://seaborn.pydata.org/tutorial/distributions.html?highlight=univariate>

statsmodels.discrete.discrete\_model.logit¶. statsmodels. (n.d.). <https://www.statsmodels.org/stable/generated/statsmodels.discrete.discrete_model.Logit.html?highlight=logit#statsmodels.discrete.discrete_model.Logit>

matplotlib.axes.axes.boxplot¶. matplotlib.axes.Axes.boxplot - Matplotlib 3.4.3 documentation. (n.d.). <https://matplotlib.org/stable/api/_as_gen/matplotlib.axes.Axes.boxplot.html?highlight=boxplot#matplotlib.axes.Axes.boxplot>

Babu, A. (2020, September 23). Multicollinearity - a beginner's guide. Medium. Retrieved September 9, 2021, from <https://medium.com/analytics-vidhya/multicollinearity-a-beginners-guide-fea13d01a171#:~:text=Reasons%20for%20multicollinearity.%20It%20can%20be%20caused%20by,same%20kind%20of%20variable.-%20e.g.%20Sex%20and%20Gender>