**Performance Assessment**

**NBM3 TASK 1: LINEAR REGRESSION MODELING**

Bader Ale

Department of Information Technology, Western Governors University

D208 Exploratory Data Analysis

July 2023

Table of Contents

[Part I: Research Question 3](#_Toc140945832)

[Part II: Method Justification 3](#_Toc140945833)

[Part III: Data Preparation 4](#_Toc140945834)

[Univariate Visualizations 6](#_Toc140945835)

[Bivariate Visualizations 8](#_Toc140945836)

[Part IV: Model Comparison and Analysis 13](#_Toc140945837)

[Initial Regression Model 13](#_Toc140945838)

[Reduced Regression Model 16](#_Toc140945839)

[Model Comparison 17](#_Toc140945840)

[Works Cited 19](#_Toc140945841)

# Part I: Research Question

For this performance assessment, my research question is: **What patient factors contributed to the highest total charge billed to the patient during their hospital stay.** This will help determine which of these factors are the biggest influence of a patient’s total charge during their stay. The goal is to help patients determine what has the biggest influence on their hospital charges.

# Part II: Method Justification

For this performance assessment, we will be using multivariate linear regression to determine the relationship between our explanatory variables and our target variable. In order to do this, we must first take into account certain assumptions. First of all, there has to exist a linear relationship between each of the explanatory variables (independent variables, X) and the target variable (dependent variable, Y). Secondly, we must determine if no multicollinearity exists between the explanatory variables. This is because the linear model will not be great if more than one variable affects the independent variable the same way. Thirdly, all explanatory variables must be independent of each other. Lastly, the residuals of the model should have a constant variance, or deviation from the true value, at every point in the model (Zach, 2021).

Python was chosen over R for two reasons. The first reason is the existing familiarity with Python; the only thing to learn was which libraries needed to be imported and the parameters. Even though R was written specifically for statistical analysis and Python as more of a general programming language, many libraries have been created for Python to perform like R. The second reason is its flexibility and ease of reading/structuring. Certain Python packages were used in order to facilitate statistical calculations and visualizations. Some of these packages were Seaborn (visualization), SKLearn (data splitting and regression analysis), Statsmodel (perform linear regression) and others.

Multiple linear regression (MLR) is the appropriate technique for analyzing the research question because we effectively want to predict a total charged to the patient depending on several other, independent factors. Multiple linear regression, or multivariate linear regression was used in lieu of regular, univariate linear regression because of the multiple explanatory variables considered.

# Part III: Data Preparation

In order to begin running the regression model, the data had to be cleaned and pre-processed. This reduces the model’s error and allows for seamless splitting of the data for both x-and y- values. The data preparation included replacing missing values, if any, Furthermore, checking for any duplicated values that could interfere with our model. Lastly, outliers were dealt with accordingly, ensuring that the values used for the model were accurate and within an acceptable range.

The following table shows the independent and dependent variables used in regression model as well as the output of the summary statistics for each. Visualizations have also been included for both univariate and bivariate.

The variables used for the initial regression model include *Area, Age, Income, Marital, Gender, VitD\_levels, Doc\_visits, Initial\_admin,Complication\_risk, Overweight, Arthritis, Diabetes, Hyperlipidemia, Asthma, Services, Initial\_days and TotalCharge*. The dependent variable is *TotalCharge.* The data cleaning steps included the detection and treatment of null values, duplicates, and outliers.

The figure below shows the summary statistics for all the quantitative variables only, as these can only be performed on numeric quantities.

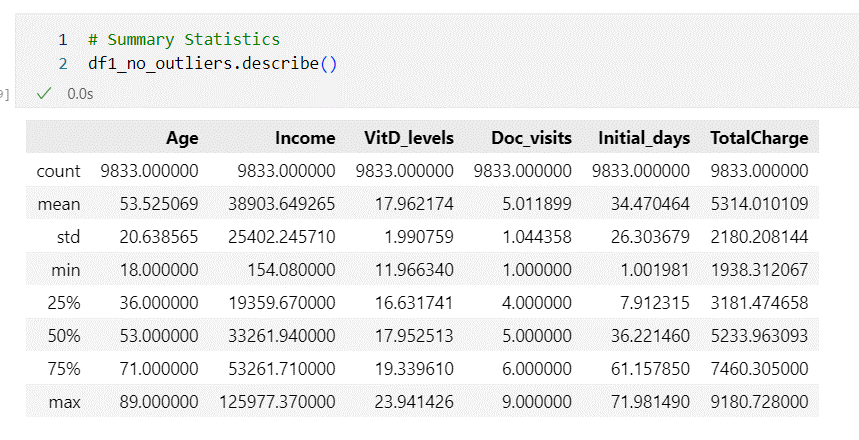


Figure 1:Variable Summary Statistics

The variables shown in Figure 1 represent only numeric quantities. The describe function runs a basic statistical overview of the variables and therefore can only run operations such as mean, and standard deviation on number values.

## Univariate Visualizations

A graph with several blue rectangular bars

Description automatically generated A graph of blue bars

Description automatically generatedA graph of a number of blue bars

Description automatically generated A graph of a number of people

Description automatically generated

A graph of a number of people

Description automatically generated A graph of a number of levels

Description automatically generated

A graph with numbers and lines

Description automatically generated A graph of a number of blue squares

Description automatically generated

A graph of a high complicatization

Description automatically generated A blue rectangular graph with white text

Description automatically generated

A graph with blue squares

Description automatically generated A blue rectangular graph with white bars

Description automatically generated

A graph with blue squares

Description automatically generated A graph with blue squares

Description automatically generated

A graph of a bar graph

Description automatically generated A graph of a number of blue squares

Description automatically generated

## Bivariate Visualizations

A graph of different colored bars

Description automatically generated A graph of different colored rectangles

Description automatically generated

A graph of different colored bars

Description automatically generated A graph of a chart

Description automatically generated

A graph of different colored squares

Description automatically generated A graph of overweight vs total charge

Description automatically generated

A graph of a comparison of a number of people

Description automatically generated A graph of a comparison of the same type of charge

Description automatically generated

A graph of a comparison of the average charge

Description automatically generated A graph of a comparison of the same color

Description automatically generated

A graph of a medical service

Description automatically generated

A graph of blue and white lines

Description automatically generated

A blue and white dotted diagram

Description automatically generated

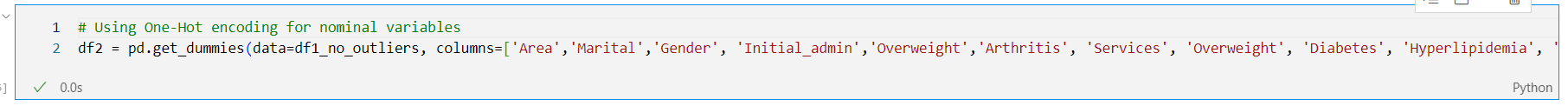
A graph with blue lines

Description automatically generated

A graph of a bar graph

Description automatically generated

The goal for transforming the data is to prepare it in a way that is easy for the model to use and that minimizes error. This included re-expressing categorical variables accordingly so the code could run efficiently. The variables needed to be re-expressed were *Area, Marital, Gender, Initial\_admin, Overweight, Arthritis, Services, Overweight, Diabetes, Hyperlipidemia, and Asthma.* This was accomplished by using the Pandas functionality *pd.get\_dummies –* the code is shown below. No missing values or duplicates were noted in the dataset.



A copy of the final code is attached.

# Part IV: Model Comparison and Analysis

## Initial Regression Model

An initial multivariate regression model was created using the prepared dataframe containing all the independent variables(shown inTable 1: Variables used in Model.**).** We first had to define the X (independent variables) and Y (TotalCharge). The code snippet is shown below for this process:

A screenshot of a computer program

Description automatically generated

Figure 2: X and Y variable assignments

. After adding a constant using .*add\_constant()* attribute of the OLS module, the linear regression object was created with the additional *.fit()* command, as shown below.

A screenshot of a computer code

Description automatically generated

Figure 4: Regression Model parameters

The following shows the summary display of the initial regression model.

A screenshot of a computer

Description automatically generated

Figure 5: Summary of Initial Regression Model

The X-data was then used to pass to the prediction module of the regression model and the results were graphed to compare it to the original y-values.

A graph of value and comparison

Description automatically generated

An initial regression model equation was used using the constant and coefficients summarized in the summary report.

***TotalCharge*** *= 2358.675577 + 0.060576(Age) +  0.000005(Income) + 0.220942(VitD\_levels) + 1.222980(Doc\_visits) - 200.577435(Complication\_risk) + 81.929811(Initial\_days) + 3.975710(Area\_Suburban) + 4.727445(Area\_Urban) + 4.514919(Marital\_Married) - 3.355894(Marital\_Never Married) - 4.737488(Marital\_Separated) - 5.767059(Marital\_Widowed) - 0.520453(Gender\_Male) + 13.478394(Gender\_Nonbinary) + 511.564613(Initial\_admin\_Emergency Admission) - 1.771107(Initial\_admin\_Observation Admission) + 2.337127(Overweight\_Yes) + 72.877156(Arthritis\_Yes) + 8.360456(Services\_CT Scan) - 2.740121(Services\_Intravenous) - 1.166686(Services\_MRI) + 2.337127(Overweight\_Yes) + 71.513515(Diabetes\_Yes) + 93.168755(Hyperlipidemia\_Yes) + 2.945512(Asthma\_Yes)*

The equation includes all 17 explanatory variables in the initial regression model and their respective coefficients. Moreover, the constant was also included. Our next task is to reduce the number of features and run a reduced regression model.

## Reduced Regression Model

In order to reduce the explanatory variables, Recursive Feature Elimination (RFE) was used. For this, we used the SKlearn library and the *RandomForestRegressor* as the estimator parameter. RFE was chosen due to its reduced complexity and selection of estimator algorithm – RFE works by first running the all the explanatory variables first and assigning an coefficient or “value of importance” to each one. In this specific case by specifying the parameters *step=1* , one by one the features were dropped from least importance and the top 3 (explicitly given in *n\_features\_to\_select=3*) (Brownlee, 2020). A screenshot of the code is shown below.

A screenshot of a computer program

Description automatically generated

Here, we specified the algorithm to remove one feature at a time and to reach a maximum of three total variables to use in the reduced model.

A screenshot of a computer program

Description automatically generated

The RFE mechanism chose *Complication\_risk, Initial\_days, and Initial\_admin\_Emergency Admission.* We will now use this as input into our reduced model.

A screenshot of a computer code

Description automatically generated

As before, the new parameters were passed to the model object and the regression was executed. Below is a screenshot of the summary output.

A screenshot of a computer

Description automatically generated

## Model Comparison

From the summary outputs of both models, we can see the R2 value changed from 0.997 to 0.996 when executing the initial and reduced models respectively. Moreover, a histogram of the predicted values vs the actual values for the target variable was plotted and are shown below.

A graph of value and comparison

Description automatically generated A graph of value and value

Description automatically generated

As shown above, the reduced model shows greater predictive value overall but is most evident after around $5,000. From $0 - $5,000, both models showed similar accuracy.

A reduced regression model equation is the following:

*TotalCharge = 2451.189539 - 199.260061(Complication\_risk) + 81.945552(Initial\_days) + 513.432540(Initial\_admin\_Emergency Admission)*

A simple interpretation of the above equation is: With all else being zero, the average starting Total Charge to a patient is $2451.19. This is considered the y-intercept in the equation. For the three variables chosen as parameters in the reduced model, *Complication\_risk* and *Initial\_admin\_Emergency Admission* are dummy variables and so need to be input correctly – for *Complication\_risk,* the appropriate values are 0,1,2 if they are considered Low, Medium or High risk respectively. For *Initial\_admin\_Emergency Admission,* a value of 1 would represent the patient was originally admitted via the Emergency Department and a value of 0 would represent otherwise.

Also, QQ-Plots were also created to show how well the models performed.

A graph with a red line

Description automatically generated A graph with a red line

Description automatically generated

The reduced model shows a slightly better fit than the initial model – a slightly better fit is more evident towards the ends of the plots. Moreover, the residual standard error was calculated and is shown below.

A screenshot of a computer error

Description automatically generated

A similar performance was obtained with the reduced model than with the initial – in this case, a significant number of explanatory variables were removed, and the model still performed well.

## Discussion

The reduced linear regression model showed a 0.996 R-squared – this means that the 99.6% variation in Total Charge can be explained by the *Complication\_risk*, *Initial\_days* and *Initial\_admin\_Emergency Admission* variables. The statistical significance of the results shown for the reduced model are that these three variables can more or less predict the amount the patient would be charged (according to the R-squared value).

A screenshot of a computer screen

Description automatically generated

Out of the three coefficients, *Initial\_admin\_Emergency Admission* had the highest value, meaning it had the most positive effect on the target variable *TotalCharge.* On the other hand, *Complication\_risk* showed a very negative correlation with *TotalCharge.*

Although the model showed high predictive value, the practical use and real-world significance can be much different. Other variables affect the *TotalCharge* such as insurance allowance that are not included in the model. Also, the model does not take into account if the patient had complication while hospitalized – this would lead to more treatments and procedures which would elevate the *TotalCharge.* This makes the model statistically significant but not practically significant – in reality, there are more nuances to a patient’s total charge than just the three variables chosen.

There are disadvantages to the reduced model. For one, the data preparation phase removed some outliers even though they were chosen based on z-scores. Also, the feature selection method used in this case ,RFE, used the RandomForest Regressor as an estimator – other estimators exists that could have output a different set of predictive variables completely different from the ones chosen here.

## Recommendations

Based on the finding of the reduced linear model, admission through the Emergency Room accounted for the highest correlation to the total charge a patient was billed. This information might be beneficial for a patient because depending on the severity of the issue, they might be better off (being billed a lower amount) being admitted through a different department or seeing an urgent care (if the condition/issue is not life threatening). In this way, a patient might not only obtain medical help but also be billed the smallest amount.

# 

Works Cited

Brownlee, J. (2020, August). *Recursive Feature Elimination (RFE) for Feature Selection in Python*. Retrieved from Machine Learning Mastery: https://machinelearningmastery.com/rfe-feature-selection-in-python/

Zach. (2021, November 16). *The Five Assumptions of Multiple Linear Regression*. Retrieved July 3, 2023, from Statology: https://www.statology.org/multiple-linear-regression-assumptions/