PREDICTIVE ANALYSIS – D209

Gooden, Nina S. Gooden [ID #: 009823504]

Dr. Eric Straw

PREDICTIVE ANALYSIS – D209

This analysis will explore the medical data of a theoretical real-world organization. By creating this logistic regression, I will address the business’ concern about readmission in a data-first evaluation. I will also provide visualizations to support my assessments, in tandem with code for my models. As previously requested, I will also discuss the limitations and potential course of action this data supports.

## **A. The Research Question:**

1. Research Question
   1. Using Ridge regression, can we identify patients likely to contribute to readmission based on factors pertaining to age, how long they were initially at the hospital, and how much they were billed overall?
2. Data Analysis Goals and Objective
   1. I will be reassessing some of the data I gathered from course D206 with tools from this course. This evaluation is to further explore variables that were identified as having correlation with readmission. In this evaluation, we will look at variables that have been identified as having correlation previously—age, VitD\_levels, Initial\_days, and TotalCharge—through the model tuning method.

## **B. Justification for Method**

1. How Ridge Regression analyses the selected data set and hypothesized outcome:
   * 1. Ridge Regression is a method of estimating the multicollinearity in data that is multi-regressive. It helps to avoid overfitting data within a model.
     2. It is my expectation that this analysis will result in a **low mean squared error (MSE) and a high r2 score**, due to the fact that we’ve evaluated these variables in the past.My goal for this task is to glean insights as to which variables have value in my continued assessment of the medical data. Ridge regression is a good method to use for this analysis because of the multicollinearity of the data.
2. Ridge Regression assumption summary *(Great Learning Team, 2022b)*:
   1. Linearity: the assumption that there is a linear relationship between the independent variable and the dependent variable.
   2. Constant Variance: the assumption that residuals have constant variance on the predictor variables.
   3. Independence: there is no relation between the residuals and the variable itself.
3. Packages and libraries:
   1. I’ve opted to use Python for my analysis as this is the language, I am most comfortable with. Python offers packages and libraries that make visualizations and analysis easy and straightforward. I imported pandas, Numpy, Matplotlib, and Seaborn in this evaluation, though the latter two were unnecessary. In addition, I will also be using sklearn, for Rodge, r2\_score, mean\_squared\_error, and train\_test\_split. The sklearn package is uniquely suited for predicting and classifying data in measurable models.

## **C. Data Preperation Process**

1. Data Preparation Goals and Manipulations
   1. My first data processing goal is to ensure my data is as clean as possible. I do this by first importing my packages and converting the ReAdmis column to numerical values.
   2. I then dropping all columns I will not be working with.

A screenshot of a computer

Description automatically generated with medium confidence

1. Doing this leaves me with a dataframe with only the following columns, which I run two quick analyses on before exporting for submission:
   * 1. ReAdmis, my categorical primary variable.
     2. Age, VitD\_levels, Initial\_days, and TotalCharge are all continuous variables.

A screenshot of a computer

Description automatically generated with medium confidence

1. Data and steps have been provided via screenshots.
2. Cleaned data set has been attached to this submission.

## **D. ANALYSIS**

1. Split data and provide files:
   1. I first created the X and y values by removing and creating exclusive frames with ReAdmis.

A screenshot of a computer

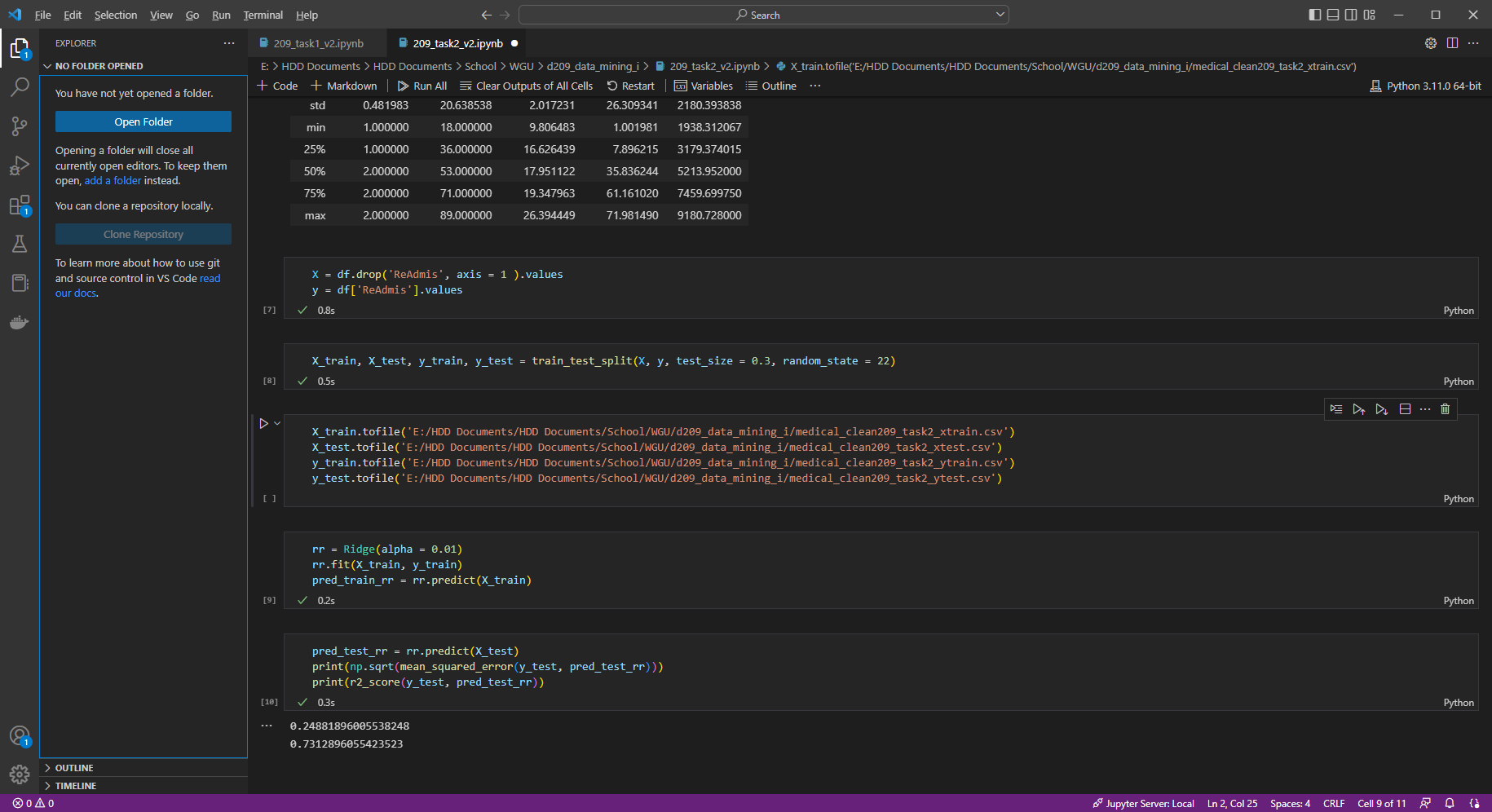
Description automatically generated

* 1. Next, I created a training and test set of the data using a 3:7 ratio for test and training respectively. I also set a seed value so I could replicate this model.

A screenshot of a computer

Description automatically generated

* 1. I exported the files, per the assignment.



1. I used the Ridge() class and set the lambda (alpha) argument to .01 for a 1% penalty. I then defined the model evaluation with .fit() and .predict() just as we’ve done in the past.  
   Finally, I printed the mean square error (MSE) and r2 score. *(Rowe, n.d.)*

A screenshot of a computer

Description automatically generated

1. All code has been provided with detailed steps and screenshots.

## **E. Data SUMMARY AND IMPLICATIONS**

1. Explaining the accuracy and mean square error (MSE) of the model:
   1. The mean square error (MSE) value for my model is 0.25, which falls into the ideal range of 0.2-0.5 for showing that the model can relatively predict accurately. When paired with the r2 score of .73, or 73%, I can say with a degree of certainty that the variables are correlated.
2. Discuss the results and implications of the analysis:
   1. The model has an r2 score of 73%, which is the measurement of whether or not a model is sound or “good.” While that number could be pushed higher by working with the available variables, the results from this assessment are more than enough for stakeholders—particularly those in the data management sections of this healthcare system—to further explore.
3. Data analysis limitations:
   1. Ridge regression includes all the predictors in the final model and is incapable of performing specific feature selection. This means that if a future data analysist wanted to tighten up the model, the only way she could do so would be to change the model numerous times and rerun it manually.
4. Recommendations:
   1. This analysis should be rerun with changes to the explored variables in order to boost the r2 score.
   2. This medical center should spend resources to house data in a more way that better designates actionable information for evaluation.
   3. In order to maintain the health of this analysis, the medical center should launch campaigns to gather as much additional data as possible. I would also recommend keeping track of the variation KPIs in order to segment response efforts.
   4. This analysis should be run quarterly in order to ensure the value of this assessment.

## **F. Panopto**

[Data Mining I – NVM2| D209](https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=46122512-aee7-4b23-940a-af5d01590ad2)

## **G. Reference web sources**

Great Learning Team. (2022, November 16). *Ridge Regression Definition & Examples | What is Ridge Regression?* Great Learning Blog: Free Resources What Matters to Shape Your Career! https://www.mygreatlearning.com/blog/what-is-ridge-regression/

sklearn.linear\_model.Ridge. (n.d.). Scikit-learn. https://scikit-learn.org/stable/modules/generated/sklearn.linear\_model.Ridge.html

## **H. Acknowledge sources**

Rowe, W. (n.d.)*. Mean Square Error & R2 Score Clearly Explained*. BMC Blogs. https://www.bmc.com/blogs/mean-squared-error-r2-and-variance-in-regression-analysis/

Slegh, S. (2021, August 10). *What is Mean Squared Error, Mean Absolute Error, Root Mean Squared Error and R Squared?* Studytonight. https://www.studytonight.com/post/what-is-mean-squared-error-mean-absolute-error-root-mean-squared-error-and-r-squared