Talha Hussain

Introduction to Data Science

***Data Science Homework Assignment 3 Multiple Linear Regression Model Using Covid Mortality Dataset and Logistic Regression Model and Naïve Bayes Classifier Model Using Lending Club Dataset***

**Multiple Linear Regression Model Using CovidMortality Dataset from Microsoft Excel Comma-Separated Value CSV File “CovidMortality.csv” Part 1 Problem #’s 1–2:**

Using the following CovidMortality dataset do the following:

1. As you can understand that there is no effective way to use the name of the state. There are many other ways to use that in the multiple linear regression model. As an example, you can divide the states into regions. You can also divide the states into say blue states and red states or you can divide the states into regions. You can also divide the states into say blue states and red states or you can divide the states into high income and low income. Please use one of such way and explain why you did that.

I divided the states from the “CovidMortality” Excel File into High Income and Low Income, that is I categorized all the different states of the United States by High Median Household Income that is represented by the categorical value “High” and Low Median Household Income that is represented by the categorical value “Low”. One reason that I chose to divide the states categorically by High Median Household Income and Low Median Household Income is that I believe that people who live in states with low median household income would lead to increase in the number of deaths of people as a result of Covid. In addition, the amount of money people earn from their jobs while living in the houses of those states where the median household income is very low would make it more difficult for those people to afford healthcare. As a result, new patients who test positive for covid would be at risk of dying from covid under the condition that they have insufficient amount of money to pay their medical bills to the doctor or doctors and consequently would not be able to get complete medical coverage at all from the doctor or doctors. The other reason that I chose to divide the states categorically by High Median Household Income and Low Median Household Income is that I consider people who live in states with High Median Household Income would result in a fewer number of deaths of people due to Covid and therefore the amount of money people earn from their jobs while living in the houses of those states where the median household income is very high would make it less difficult for those people to afford healthcare. As a result, new patients who test positive for covid would be able to pay their medical bills to the doctor or doctors.

1. Build a Multiple Linear Regression Model, use deaths as Dependent Variables. Discuss, the model and about different independent variables. Discuss that by creating your own categorical variable did it make the model better? I want a detailed discussion.

A screenshot of a computer

Description automatically generated

Multiple Linear Regression Model 1 for Covid Mortality Dataset with Categorical Variable “State.Median.Household.Income.Status” as “income\_dummy\_data”:

1. Dependent Variable: Deaths
2. Independent Variables: Confirmed Cases, Population, Area, and income\_dummy\_data
3. Residuals: The residuals indicate differences between the Observed Values and the Predicted Values. In addition, the Residuals show a mean near 0 with some variability.
4. Coefficients:
5. Confirmed: The Positive Coefficient 0.01875 indicates that the increase in the Number of Confirmed Cases corresponds to an Increase in the Confirmed Number of Deaths.
6. Population: The Positive Coefficient 0.0001784 suggests that there is a slight increase in the Number of Deaths with population size. Therefore, population is not a very significant factor to predict the number of Covid Deaths.
7. Area: The Negative Coefficient -0.007068 implies a Decrease in the Number of Deaths with Larger Areas.
8. income\_dummy\_data: The positive coefficient 1606 suggests that States with a Higher-Income have high mortality rates.
9. Multiple R-Squared R2 and Adjusted R-Squared R2: The Numerical Value of the Multiple R-Squared R2 that is 0.6211 and the Numerical Value of the Adjusted Multiple R-Squared R2 that is 0.5889 both respectively indicate that the model explains that there is approximately 62.11% variation in Deaths, adjusted for the number of predictors.
10. F-Statistic: The Significance of the P-Value 1.939e-09 that is 0.000000001939 indicates that at least one of the Independent Variables has a significant effect on the Dependent Variable.

Multiple Linear Regression Model 2 for Covid Mortality Dataset without the Categorical Variable “State.Median.Household.Income.Status”:

1. Dependent Variable: Deaths
2. Independent Variables: Confirmed Cases, Population Area
3. Residuals: Similar to the Multiple Linear Regression Model 1 for Covid Mortality Dataset with Categorical Variable “State.Median.Household.Income.Status” as “income\_dummy\_data”, residuals exhibit a Mean near 0 with some variability.
4. Coefficients: Similar to the Multiple Linear Regression Model 1 for Covid Mortality Dataset with Categorical Variable “State.Median.Household.Income.Status” as “income\_dummy\_data”, however, this time without the “income\_dummy\_data” in terms of excluding the Categorical Variable “State.Median.Household.Income.Status”, when it comes to the following:
5. Confirmed: The Positive Coefficient 0.01643 indicates that the increase in the Number of Confirmed Cases corresponds to an Increase in the Confirmed Number of Deaths.
6. Population: The Positive Coefficient 0.0002429 suggests that there is a slight increase in the Number of Deaths with population size. Therefore, population is not a very significant factor to predict the number of Covid Deaths.
7. Area: The Negative Coefficient -0.006310 implies a Decrease in the Number of Deaths with Larger Areas.
8. Multiple R-Squared R2 and Adjusted R-Squared R2: The Numerical Value of the Multiple R-Squared R2 that is 0.603 and the Numerical Value of the Adjusted Multiple R-Squared R2 that is 0.5782 both respectively indicate that the model without the Categorical Data “State.Median.Household.Income.Status” represented by the R Code variable “income\_dummy\_data” explains that there is approximately 60.3% Variation in Deaths, adjusted for the Number of Predictors.
9. F-Statistic: The Significance of the P-Value 1.038e-09 that is 0.00000001038 indicates the overall significance of the Multiple Linear Regression Model 2 for Covid Mortality Dataset without the Categorical Variable “State.Median.Household.Income.Status”.

Discussion:

Confirmed Cases: Both models show a positive relationship between Confirmed Cases and Deaths. This aligns with expectations, as higher cases typically correlate with higher mortality rates.

Population: Multiple Linear Regression Model 1 for Covid Mortality Dataset with the Categorical Variable “State.Median.Household.Income.Status” as “income\_dummy\_data” and Multiple Linear Regression Model 2 for Covid Mortality Dataset without the Categorical Variabe both include a Population Size as an Independent Variable. However, it does not appear statistically significant in either model, implying that population size alone might not be a strong predictor of mortality rates.

Area: Both the Multiple Linear Regression Model 1 for Covid Mortality Dataset with the Categorical Variable “State.Median.Household.Income.Status” include the area as an Independent Variable, however, in addition, it appears to be statistically insignificant in predicting mortality rates. Larger areas might not necessarily correlate with lower mortality rates, at least in the dataset.

In terms of evaluating whether or not creating the Categorical Variable that is variable which I divided all 52 states into two categories of Income related to the dataset that I added in the Microsoft Excel Comma-Separated Value CSV File “CovidMortality.csv” that is the dataset “State.Median.Household.Income.Status” that includes “High” for High Median Household Income and “Low” for Low Median Household Income represented as the “income\_dummy\_variable” made the model better, I compared the different performance metrics of Multiple Linear Regression Model 1 for Covid Mortality Dataset with the Categorical Variable “State.Median.Household.Income.Status” as “income\_dummy\_data” and Multiple Linear Regression Model 2 for Covid Mortality Dataset without the Categorical Variable “State.Median.Household.Income.Status”. The major performance metrics I compared are the Adjusted R-Squared R2 and the Residual Standard Error RSE. The Adjusted R-Squared R2 measures the proportion of variance in the Dependent Variable “Deaths” that is explained by the Independent Variables, that are adjusted for the number of predictors in the model. A Higher Adjusted R-Squared R2 signifies a better fit of the model to the data. The Residual Standard Error RSE measures the average deviation of the observed values from the fitted values. A lower Residual Standard Error RSE shows a better fit of the model to the data. In the case of the Adjusted R-Squared R2 and the Residual Standard Error RSE for the Multiple Linear Regression Model 1 for Covid Mortality Dataset with Categorical Variable “State.Median.Household.Income.Status” as “income\_dummy\_data” and the Adjusted R-Squared R2 and the Residual Standard Error RSE for Multiple Linear Regression Model 2 for Covid Mortality Dataset without Categorical Variable “State.Median.Household.Income.Status”, differences between these two models are evident as follows:

Multiple Linear Regression Model 1 for Covid Mortality Dataset with Categorical Variable “State.Median.Household.Income.Status” as “income\_dummy\_data”:

Adjusted R-Squared R2 = 0.5889

Residual Standard Error RSE = 3767

Multiple Linear Regression Model 2 for Covid Mortality Dataset without the Categorical Variable “State.Median.Household.Income.Status”:

Adjusted R-Squared R2 = 0.5782

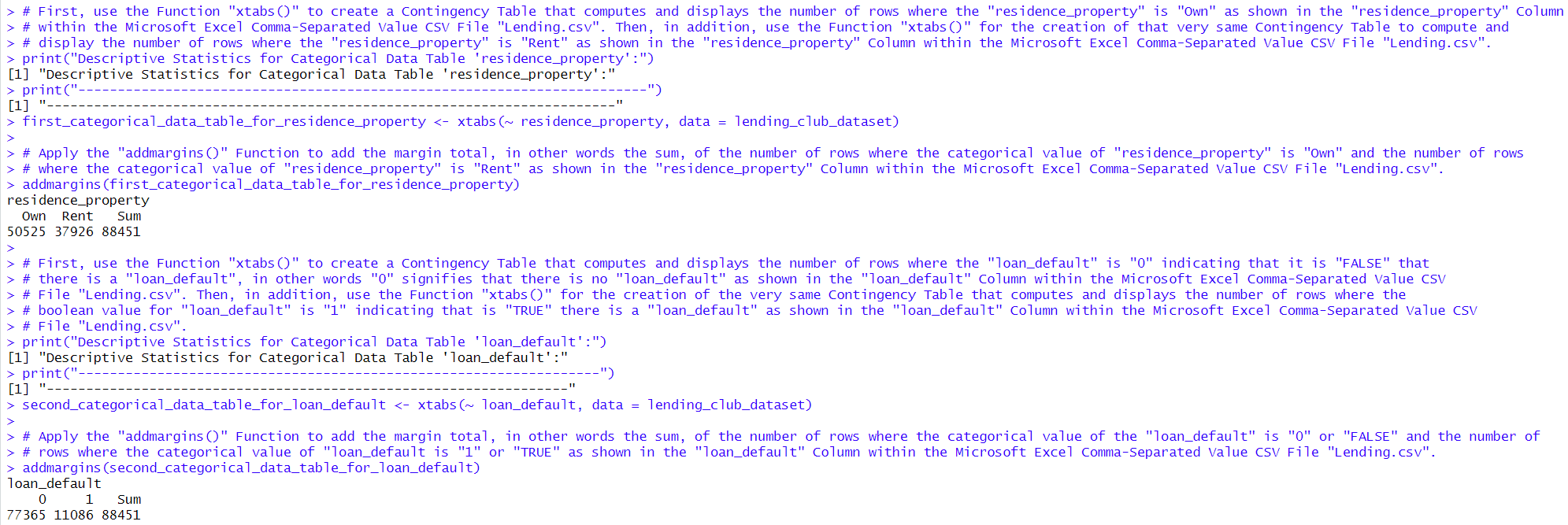
Residual Standard Error RSE = 3816

By analyzing the statistical differences between Multiple Linear Regression Model 1 for Covid Mortality Dataset with Categorical Variable “State.Median.Household.Income.Status” and Multiple Linear Regression 2 for Covid Mortality Dataset without the Categorical Variable “State.Median.Household.Income.Status” as “income\_dummy\_data”, I can see that the Multiple Linear Regression Model 1 for Covid Mortality Dataset with Categorical Variable “State.Median.Household.Income.Status” as “income\_dummy\_data” has a slightly higher Adjusted R-Squared R2 value of 0.5889 compared to the Adjusted R-Squared R2 of the Multiple Linear Regression Model 2 for Covid Mortality Dataset without the Categorical Variable “State.Median.Household.Income.Status” that is 0.5782. However, the subtracted difference between the Adjusted R-Squared R2 value for Multiple Linear Regression Model 1 for Covid Mortality Dataset with Categorical Variable “State.Median.Household.Income.Status” as “income\_dummy\_data” and the Adjusted R2 value for Multiple Linear Regression Model 2 for Covid Mortality Dataset with Categorical Variable “State.Median.Household.Income.Status” that is only a 0.0107 difference does not make a big difference because that subtracted difference is relatively small. In addition, when it comes to the differences of the statistical values for the Residual Standard Error RSE of Multiple Linear Regression Model 1 for Covid Mortality Dataset with Categorical Variable “State.Median.Household.Income.Status” as “income\_dummy\_data” and the Residual Standard Error RSE of the other multiple linear regression model that is the Multiple Linear Regression Model 2 for Covid Mortality Dataset without the Categorical Variable “State.Median.Household.Income.Status”, the first multiple linear regression model that is Multiple Linear Regression Model 1 for Covid Mortality Dataset with Categorical Variable “State.Median.Houshold.Income.Status” as “income\_dummy\_data” has slightly better accuracy in predicting the dependent variable “Deaths”. Therefore, based on the Adjusted R-Squared R2 and Residual Standard Error RSE metrics, it can be concluded that the addition of “income\_dummy\_data” variable slightly improved the model’s performance even though that improvement was only by a small margin.

**Logistic Regression Model and Naïve Bayes Classifier Model Using LendingClub Dataset Using Microsoft Excel Comma-Separated Value CSV File “Lending.csv” Part 2 Problems 1 – 3:**

Using the LendingClub Dataset, please do the following:

1. Do the descriptive statistics. Your goal is to tell a story about this dataset. I don’t have a specific answer in mind. So be creative.



A group of buildings with text

Description automatically generated with medium confidence

1. Build Logistic regression and Naïve Bayes model with the data. Paste your outputs here.

Program Output of Logistic Regression Model for Lending Club Dataset:

A close-up of a white background

Description automatically generated

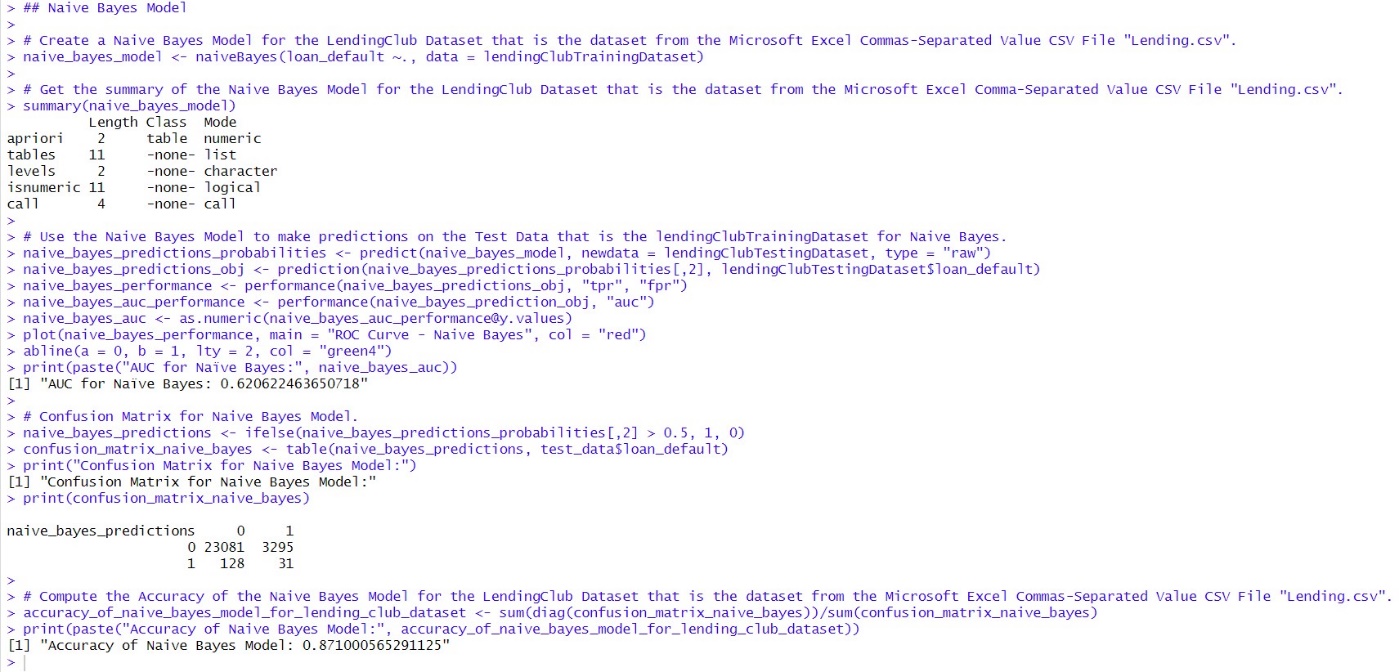
A close-up of text

Description automatically generated

A graph of a curve

Description automatically generated

Program Output of Naïve Bayes Model for Lending Club Dataset:



A graph with a red line

Description automatically generated

1. Compare the models and discuss about each model’s accuracy. Make sure you also include ROC in your discussion. Explain the outputs that you pasted above in details.

Logistic Regression Model for the LendingClub Dataset Using Microsoft Excel Comma-Separated Value CSV File “Lending.csv”:

*Summary of Logistic Regression Model for the Lending Club Dataset*: First, the Logistic Regression Model for the LendingClub Dataset from the Microsoft Excel Comma-Separated Value CSV File “Lending.csv” has various predictors such as loan amount, adjusted annual income, and other related variables. Second, based on the Output of the Logistic Regression Model “ “Estimate” in coefficients of Logistic Regression summary represents change in the log-odds of loan default for a unit change in each predictor. Third, the “Std. Error” indicates standard errors of estimated coefficients whereas the ‘z-value’ and ‘Pr(> |z|)’ provides significance of each predictor. Fourth, the majority of predicted values have a high level of significance where the P-value < 0.05 and this indicates that these predicted values are important in predicting loan default.

*Confusion Matrix*:

1. The Confusion Matrix shows the Number of True Positives TP, True Negatives TN, False Positives FP, and False Negatives FN in the prediction of the Logistic Regression Model for the Lending Club Dataset as follows:
2. 23,208 instances of True Negatives TN
3. 3,326 instances of False Positives FP
4. 1 instance of a False Negative FN
5. 0 instances of True Positives TP

1. Accuracy: The accuracy of the Logistic Regression Model for the Lending Club Dataset is calculated as the proportion of correct predictions that is True Positive TP + True Negative TN divided by, /, the Total Number of Cases that is True Positive TP + True Negative TN + False Positive FP + False Negative FN. Therefore, based on the Logistic Regression Model for the Lending Club Dataset, the Accuracy would be calculated as follows: Accuracy =  =  = = 0.8746184285, which is approximately 87.46%.

*Receiver Operating Characteristics ROC Curve and Area Under the ROC Curve AUC*: First, in terms of the Logistic Regression Model for the Lending Club Dataset, the Receiver Operating Characteristic ROC Curve plots the True Positive Rate TPR, against the False Positive Rate FPR at various thresholds. Based on the Confusion Matrix from the Problem 2 Output that is the “Program Output of Logistic Regression Model for Lending Club Dataset”, as a result of computing the True Positive Rate TPR as follows: TPR =  = = = 0. Since the True Positive Rate TPR = 0, this means that if it is Negative when the Threshold is Positive. If Threshold is Positive, then everything would be Negative. In other words, due to the fact that the True Positive Rate TPR also known as recall is 0 signifies that this True Positive Rate TPR measures the proportion of positive instances that are correctly identified divided by the total amount of actual positive instances. Furthermore, when it comes to the False Positive Rate FPR according to the Confusion Matrix from the Problem 2 Output that is the “Program Output of Logistic Regression Model for Lending Club Dataset”, it can be computed as follows: FPR =  = = = 0.13. Therefore, the False Positive Rate FPR that is 0.13 measures the proportion of actual negatives incorrectly identified as positives. All in all, the relationship between the True Positive Rate TPR of 0 and the False Positive Rate FPR of approximately 0.13 reveals that the behavior of the Logistic Regression Model for the Lending Club Dataset results in a Higher Threshold due to the fact that there is a smaller number of instances classified as positive and hence, the Logistic Regression Model for the Lending Club Dataset becomes more accurate. Furthermore, a Higher Threshold designates in a lower True Positive Rate TPR where some actual positive instances are missing and a lower False Positive Rate FPR where there is a misclassification of a smaller number of negative instances. Second, when it comes to the Logistic Regression Model for Lending Club Dataset, the Area Under the Receiver Operating Characteristics ROC Curve represents the model’s ability to differentiate between whether the Categorical Variable “loan\_default” represented by the Numeric Value 1 that signifies this variable as True or False when the Numeric Value of the Categorical Variable “loan\_default” is 0. Third, the Numeric Value of the Area Under the Receiver Operating Characteristic ROC Curve of the Logistic Regression Model for the Lending Club Dataset is 0.6538 and this points out that the Logistic Regression Model for the Lending Club Dataset outperforms random guessing and it is moderate at differentiating between positive instances and negative instances because the Numeric Value of this classification model that is 0.6538 is not that much close to the Numeric Value of 1.

Naïve Bayes Model for Lending Club Dataset Using Microsoft Excel Comma-Separated Value CSV File “Lending.csv”:

*Summary of Naïve Bayes Model for Lending Club Dataset*: First, in contrast to the Logistic Regression Model for the Lending Club Dataset, the Naïve Bayes Model for the Lending Club Dataset calculates the probabilities of different classes based on the distribution of predictors. Second, in the “Program Output of Naïve Bayes Model for Lending Club Dataset”, one can observe that the Naïve Bayes Model for the Lending Club Dataset uses Apriori Probabilities, tables, levels, and built-in R Programming Function Calls that includes the Function “performance()” and the Function “prediction()” used to create the Naïve Bayes Model for the Lending Club Dataset.

*Confusion Matrix*:

1. The Confusion Matrix shows the Number of True Positives TP, True Negatives TN, False Positives, and False Negatives FN in the prediction of the Naïve Bayes Model for the Lending Club Dataset as follows:
2. 23, 081 instances of True Negatives TN
3. 3, 295 instances of False Positives FP
4. 128 instances of False Negatives FN
5. 31 instances of True Positives TP
6. Accuracy: The accuracy of the Naïve Bayes Model for the Lending Club Dataset is calculated as the proportion of correct predictions that is True Positive TP + True Negative TN divided by, /, the Total Number of Cases that is True Positive TP + True Negative TN + False Positive FP + False Negative FN. Therefore, based on the Naïve Bayes Model for the Lending Club Dataset, the Accuracy would be calculated as follows: Accuracy = = = = 0.871000565291125, which is approximately 87.10%.

*Receiver Operating Characteristics ROC Curve and Area Under the ROC Curve AUC:*First, in terms of the Naïve Bayes Model for the Lending Club Dataset, the Receiver Operating Characteristic ROC Curve can be analyzed similarly to the Logistic Regression Model for the Lending Club Dataset because the Naïve Bayes Model for the Lending Club Dataset plots the True Positive Rate TPR and False Positive Rate FPR across various thresholds. Based on the Confusion Matrix from the Problem 2 Output that is the “Program Output of Naïve Bayes for Lending Club Dataset”, as a result of computing the True Positive Rate TPR as follows: TPR = = = = 0.20. Since the True Positive Rate TPR that is also known as recall is equal to 0.20, this means that the True Positive Rate measures the proportion of positive instances that are correctly identified by the total amount of actual positive instances. In addition, when it comes to the False Positive Rate FPR according to the Confusion Matrix from Problem 2 Output that is “Program Output of Naïve Bayes Model for the Lending Club Dataset”, it can be computed as follows: FPR = = = = 0.12. Therefore, the False Positive Rate FPR that is 0.12 measures the proportion of actual negative incorrectly identified as negatives. All in all, the relationship between the True Positive Rate TPR of 0.20 and the False Positive Rate FPR of approximately 0.12 reveals that the behavior of the Logistic Regression Model for the Lending Club Dataset results in a Higher Threshold due to the fact that there is a smaller amount of instances classified as positive and hence, that Naïve Bayes Model for the Lending Club Dataset becomes more accurate. Furthermore, a Higher Threshold designates a lower True Positive Rate TPR where some actual positive instances are missing and a lower False Positive Rate FPR where there is a misclassification of a smaller number of negative instances. Second, when it comes to the Naïve Bayes Model for the Lending Club Dataset, the Area Under the Receiver Operating Characteristics ROC Curve represents the model’s ability to differentiate between whether the Categorical Variable “loan\_default” represented by the Numeric Value 1 that signifies this variable as True or False when the Numeric Value of the Categorical Variable “loan\_default” is 0. Third, the Numeric Value of the Area Under the Receiver Operating Characteristic ROC Curve of the Naïve Bayes Model for the Lending Club Dataset is 0.6206 and this points out that the Naïve Bayes Model for the Lending Club Dataset outperforms random guessing and it is moderate at differentiating between positive instances and negative instances because the Numeric Value of this classification model that is 0.6206 is not that much close to the Numeric Value of 1.

*Similarities and Differences between the Logistic Regression Model for the Lending Club Dataset and the Naïve Bayes Model for the Lending Club Dataset*:

1. Accuracy: Both the Logistic Regression Model for the Lending Club Dataset and the Naïve Bayes Model for the Lending Club Dataset perform similarly in terms of accuracy with the Logistic Regression Model for the Lending Club Dataset achieving a slightly higher accuracy of 87.46% compared to Naïve Bayes 87.10%.
2. Receiver Operating Characteristic ROC Curves and Area Under the Curve AUC:
3. The Area Under the Receiver Operating Characteristic ROC Curve, in other words the AUC Curve, of the Logistic Regression Model for the Lending Club Dataset is approximately 0.636. In contrast, the Area Under the Receiver Characteristic ROC Curve of the Naïve Bayes Model for the Lending Club Dataset is approximately 0.621.
4. Since the Area Under the Receiver Operating Characteristic ROC Curve, in other words the AUC Curve, of the Logistic Regression Model for the Lending Club Dataset is higher, this means that the Logistic Regression Model for the Lending Club Dataset is a better classifier than the Naïve Bayes Model for the Lending Club Dataset because with that model, better distinctions can be made between the positive instances and negative instances.
5. Both the Numeric Value of the Area Under the Receiver Operating Characteristic ROC Curve, in other words the AUC Curve, of the Logistic Regression Model for the Lending Club Dataset, and the Numeric Value of the Area Under the Receiver Operating Characteristic ROC Curve, that is the AUC Curve, of the Naïve Bayes Model for the Lending Club Dataset are significantly greater than 0.5. As a result, this suggests that both the Logistic Regression Model for the Lending Club Dataset and the Naïve Bayes Model for the Lending Club Dataset can make distinctions between loan default classified with the binary digit “1” to show that it is True that there is a loan default and no loan default that is classified with the binary digit “0” to signify that it is False that a loan default exists to some extent.

*Conclusion*: By observing the difference between the Numeric Value of the Area Under the Receiver Operating Characteristic ROC Curve of the Logistic Regression Model for the Lending Club Dataset that is about 0.636 to the Numeric Value of the Area Under the Receiver Operating Characteristic ROC Curve of the Naïve Bayes Model for the Lending Club Dataset that is about 0.621, it is clearly evident that the Logistic Regression Model for the Lending Club Dataset is the better choice because the Logistic Regression Model for the Lending Club Dataset has a higher AUC and therefore demonstrates superior ability in distinguishing between loan defaults classified with the binary digit “1” indicating that it is True that there is a loan default and no loan defaults classified as “0” to define it as False. However, neither the Area Under the Receiver Operating Characteristic ROC Curve, the AUC, of the Logistic Regression Model for the Lending Club Dataset and the Area Under the Receiver Operating Characteristic ROC Curve, the AUC, of the Naïve Bayes Model for the Lending Club Dataset is very high or in other words close to 1 and consequently, this can mean that the performance of both the Logistic Regression Model for the Lending Club Dataset and the Naïve Bayes Model for the Lending Club Dataset would still need improvement. Although both the Logistic Regression Model for the Lending Club Dataset and the Naïve Bayes Model for the Lending Club Dataset perform well, the Logistic Regression Model for the Lending Club Dataset shows slightly higher accuracy and fewer False Negatives FNs, whereas the Naïve Bayes Model for the Lending Club Dataset can offer better performance in identify the actual number of True Positive TP cases for loan default. Depending on whether to prioritize sensitivity or specificity, either the Logistic Regression Model for the Lending Club Dataset can be preferred over the Naïve Bayes Model for the Lending Club Dataset or vice versa.