Build a logistic regression model to predict the likelihood of a customer defaulting in next 12 months. The project involves following steps:

1. Univariate analysis to assess the data quality and summarization of the information available to build a logistic regression model.
2. Winsorization and missing imputation on the independent variables.
3. Bivariate analysis to assess the linear relationship between the dependent and independent variables.
4. Splitting the data into development and validation dataset and building a logistic regression model.
5. Assessing the model performance based on the development and validation data.

Please answer the following questions based on the data set provided:

|  |
| --- |
| **Question #** |
| **Question 1**    How many rows and columns are there in the data? |
|
|
|
|
|
|
| **Question 2**  How many defaults are there in the data (Dependent variable: 'default payment next month')? |
|
|
|
|
|
|
| **Question 3**  What is the default rate in the data? |
|
|
|
|
|
|
| **Question 4**  Apply following data exclusions to prepare the model development data and compute the new default rate: ['BILL\_AMT6']>0 and ['Limit']>0 and ['Debt']>31. How many remaining rows and columns are there in the data? |
|
|
|
|
|
|
| **Question 5**  What is the default rate in the model development data? |
|
|
|
|
|
|
| **Question 6**  Create one derived variable named payment ratio in the model development data and compute minimum, maximum and mean of the derived variable. Derived variable logic: **'PAY\_AMT6' / 'BILL\_AMT6'** |
|
|
|
|
|
|
| **Question 7**  Create one rescaled variable credit limit (divided by 1000) in the model development data and compute minimum, maximum and mean of before and after of the rescaled variable. Use variable **'Limit'**. |
|
|
|
|
|
|
| **Question 8**  Perform bivariate analysis on the continuous variable payment ratio (Payratio, created in question 6). Divide the variable in the five equal bins and compute default rate in each bin. |
|
|
|
|
|
|
| **Question 9**  Perform bivariate analysis on the continuous variable credit limit (LIMIT\_BAL1, created in question 7). Divide the variable in the five equal bins and compute default rate in each bin. |
|
|
|
|
|
|
| **Question 10**  Perform bivariate analysis on the continuous variable total debt (Variable name: 'Debt'). Divide the variable in the five equal bins and compute default rate in each bin. |
|
|
|
|
|
|
| **Question 11**  Answer following questions in 'Yes' or 'No':  a) Does the default rate monotonically decrease (based on question 8 result) as the payment ratio increases?  b) Does the default rate monotonically decrease (based on question 9 result) as the credit limit increases?  c) Does the default rate monotonically increase (based on question 10 result) as the debt increases? |
|
|
|
|
|
|
| **Question 12**  Compute the 1st and 99th percentile value of the variables payment ratio (Created in question 6), credit limit (Created in question 7) and total debt (Variable name: Debt) |
|
|
|
|
|
|
| **Question 13**  Create new variables for payment ratio (Created in question 6), credit limit (Created in question 7) and total debt (Variable name: Debt) by capping the variables at 99th percentile values. Compare the pre and post capped variables mean. Use rescaled variable for credit limit. |
|
|
|
|
|
|
| **Question 14**  Create new variable by the name of intercept and assign a constant value of 1. Compute the mean of this new variable. |
|
|
|
|
|
|
| **Question 15**  Create train (Insample data) and test (out of sample) data with random seed as '0'. Test data should have 20% of the model development data records. The 'Y' variable will be 'default payment next month' and 'X variables will be ('Debt', 'Credit limit', 'intercept' and 'Payment ratio'.  How many records are there in train and test data?  Hint: Use train test split from sklearn |
|
|
|
|
|
|
| **Question 16**  Create new train ( New insample data) and new test (New out of sample) data with random seed as '0' based on the capped variables created in question 13. Test data should have 20% of the model development data records. The 'Y' variable will be 'default payment next month' and 'X variables will be ('Debt', 'Credit limit', 'intercept' and 'Payment ratio'.  How many records are there in train and test data?  Hint: Use train\_test\_split from sklearn |
|
|
|
|
|
|
| **Question 17**  Using the train data fit the logistic model and print the model coefficient  Hint: Use statsmodels.api |
|
|
|
|
|
|
| **Question 18**  Using the new train data (Created in question 16) fit the logistic model and print the model coefficient  Hint: Use statsmodels.api |
|
|
|
|
|
|
| **Question 19**  Based on the estimates obtained using the logistic regression using the train data, answer following questions in 'Yes' or 'No':  a) Does the estimate sign for payment ratio align with the observed trend in results for question 8?  b) Does the estimate sign for credit limit align with the observed trend in results for question 9?  c) Does the estimate sign for debt align with the observed trend in results for question 10?  d) Are all three variables statistically significant? |
|
|
|
|
|
|
| **Question 20**  Based on the estimates obtained using the logistic regression using the new train data (created in question 16), answer following questions in 'Yes' or 'No':  a) Does the estimate sign for payment ratio align with the observed trend in results for question 8?  b) Does the estimate sign for credit limit align with the observed trend in results for question 9?  c) Does the estimate sign for debt align with the observed trend in results for question 10?  d) Are all three variables statistically significant? |
|
|
|
|
|
|
| **Question 21**  Using the train data, fit the logistic model and print the AUC values for train and test data.  Hint: Use roc\_auc\_score from sklearn |
|
|
|
|
|
|
| **Question 22**  Using the new train data, (created in question 16) fit the logistic model and print the AUC values for new train and new test data.  Hint: Use Logistic Regression from sklearn |
|
|
|
| **Question 23**  Does the AUC value change by using pre-cap values (Created in question 21) and post-cap values (Created in question 22)? |
|
|
|
|
|
|
| **Question 24**  Score the model generated in question 17 on the train data and print following:  a) Minimum probability  b) Maximum probability  c) Total predicted defaults  d) Predicted default rated) Are all three variables statistically significant? |
|
|
|
|
|
|
| **Question 25**  Score the model generated in question 18 on the new train data and print following:  a) Minimum probability  b) Maximum probability  c) Total predicted defaults  d) Predicted default rate |
|
|
|
|
|
|