**FRAUD AND RISK ANALYTICS**

**PROJECT**

**Instructions for the submission:**

* Please maintain the following: Font - Times New Roman, Font Size - 12, Line Spacing - 1.5

|  |  |  |  |
| --- | --- | --- | --- |
| S.No. | Assessment | Submission Format | Marks |
| 1 | Insights on variables part 1 | Text | 8 marks |
| 2 | Insights on variables part 2 | Text | 8 marks |
| 3 | ROC curve and predictors analysis | Text | 12 marks |
| 4 | Model Analysis | Text | 12 marks |
| **Project Maximum Marks** | | | **40 marks** |

|  |  |
| --- | --- |
| **Question 1** | Using the provided dataset………. |
| **Marks** | 8 marks |
| **Word Count** | Not more than 200 words |
| **Your Answer** | *1)* **default rate varies across the variable ‘grade’:-**    The **loan status** indicates where your **loan** is in the process  *In these we plot bar plot at which bar plot function and keep 'grade' on x axis while 'loan status' on y axis from that we see that the grade G has high loan status as compared to F grade and then E grade and we says that the grade A has very low loan status because of their credit history higher the grade good is for the bank so G has higher grade and A grade has low loan status*  *2)* **default rate varies across the variable ‘term’:-**    *In these we plot bar plot at which keep 'Term' on x axis while 'loan status' on y axis from that we see that the term have 60 months have higher loan status and then 36 months have lower loan status on which we see that the customer are taking higher term for the repayment of the customer and then 36 month have taken less repayment status so more customer preferred 60 month repayment status as compared to 36 month*  *3)* **default rate varies across the variable ‘sub\_grade’:-**    *In these we plot bar plot at which keep 'Sub-Grade' on x axis while 'loan status' on y axis from that we see that the grade G have higher loan status as we see that the E has lower loan status as compared to D and its comparing with the sub grade of the payment so if we see that the higher the loan status they will give more status as compared to the other onces and we see that the A has lower grade loan status in which we see that in the grade parameter higher the grade good is for the bank so G5 has higher grade and A4 grade has low loan status.*  *4)* **default rate varies across the variable ‘purpose’:-**    *In these we plot bar plot at which keep 'Purpose' on x axis while 'loan status' on y axis from that we see that the customer are taking loan for the her small business because the business man requires lots of money for her business then it takes loan for credit card or for a car then customer takes loan for her entertainment moving, vacation and housing loan for this purpose they requires loan for that purpose so in this analysis we see that the business man requires a lot of loan so we have given a loan for the customer who have business man.* |

|  |  |
| --- | --- |
| **Question 2** | Using the provided dataset ……… |
| **Marks** | 8 marks |
| **Word Count** | Not more than 200 words |
| **Your Answer** | *1)* **default rate varies across the variable ‘annual\_income’:-**    *In these we plot bar plot at which keep 'annual income' on x axis while 'loan status' on y axis from that we see that the customer having low annual income having more loan status in these we see that the customer has low income they take more loans and in the inverse the customer which have very high income they take loans very less loan status so they proceed less amount of loan procedure so from this the customer having medium income having high more loan status as compared to low income the customer having high income have less loan status as same as very high income customer so from this we see that the customer having low income have high loan status and customer having very high income are very less income status.*  *2)* **default rate varies across the variable ‘loan\_amount’:-**    *In these we plot bar plot at which keep 'loan amount' on x axis while 'loan status' on y axis from that we see that the customer having medium loan amount having very high amount of loan status vice versa customer having low loan amount having less loan status as compared to the medium.so in these we see that the customer having medium loan amount are having more processed loan status as compared to the low loan amount. So, these are the insight we see from the graph have low loan amt having low processing loan as compared to the medium loan amount.*  3) **default rate varies across the variable ‘int\_rate’:-**    *In these we plot bar plot at which keep 'int rate' on x axis while 'loan status' on y axis from that we see that the customer having high account balance are having high loan status in which they are pay out high interest rate in which we say that the customer having medium interest rate having medium loan status and also we says that the customer having low interest rate payout having low loan status and which we says that the customer having high loan having interest rate bank to the bank so from this insight we see that the customer having high balance having high interest rate.*  **4) default rate varies across the variable ‘dti’:-**  A **debt-to-income** ratio (DTI) is a personal finance measure that compares the amount of debt you have to your overall income.  *In these we plot bar plot at which keep 'dti' on x axis while 'loan status' on y axis from that we see that the customer having high dti having high loan status so we say that customer which has high income are also debit her money high amount and we see that there is high loan status after that we see that the customer having low dti having low loan status in which we say that the customer having low income having low debate ratio so having low loan status after that we say that the customer having medium income having medium income status in which we say that the customer having medium dti having average loan status.* |

|  |  |
| --- | --- |
| **Question 3** | Answer the……….. |
| **Marks** | 12 marks |
| **Word Count** | Not more than 300 words |
| **Your Answer** | a) After running a Random Forest Model, the following ROC curve was obtained as an output:-    Roc Curve It is a plot of the false positive rate (x-axis) versus the true positive rate (y-axis) for a number of different candidate threshold values between 0.0 and 1.0. Put another way, it plots the false alarm rate versus the hit rate  The ROC curve is a useful tool for a few reasons:   * The curves of different models can be compared directly in general or for different thresholds. * The area under the curve (AUC) can be used as a summary of the model skill.   The shape of the curve contains a lot of information, including what we might care about most for a problem, the expected false positive rate, and the false negative rate.  To make this clear:   * Smaller values on the x-axis of the plot indicate lower false positives and higher true negatives. * Larger values on the y-axis of the plot indicate higher true positives and lower false negatives   An operator may plot the ROC curve for the final model and choose a threshold that gives a desirable balance between the false positives and false negatives.  In general, an AUC of 0.5 suggests no discrimination (i.e., ability to diagnose patients with and without the disease or condition based on the test), 0.7 to 0.8 is considered acceptable, 0.8 to 0.9 is considered excellent, and more than 0.9 is considered outstanding.  From the classification model we see that the area under the curve is 0.84 it means that the ratio between true positive and false positive is lies under the curve where 0.8 is Excellent classifier and it is acceptable.   b) The Random Forest model also produced the below output    The Random forest model is also described this output it just describe the feature importance of each feature in this output we see that the int\_rate are most important parameter for the loan default from this we see that the how much the customer gives the intrest rate to the bank so bank get more and more profit so in this we see that if the customer gives more intrest rate then it will give benefits for this and after that the grade parameter is 2nd most important in this parameter we see that the which customer gives which grade according to her loan status so for this they will provided a less amount of grade to each customer and after that 3rd most is purpose for that the customer is what purpose the loan is taken from the bank that’s why they give us 3rd important parameterand the next important parameter is emp\_length in which it will just give the information about employee so its not important and the last one is home ownership its not really important so we see that they will give us important parameter according to the data.  c) Sometimes, there are loan applicants who are new to the credit system, like students or customers who have shifted employments from unorganized sectors to organized sectors so from the dataset which we have given the loan predictor which also applicable are  i)purpose: -what purpose the student wants the loan from the bank is its important parameter  ii)Term: -How many months he wants the loan because it is not repayable or payable for the loan  iii)emp\_title: -student has provided which service so it will easy for the bank to give the loan  iv)grade: -In which what grade the customer is if the grade is not good it will not provide the loan otherwise it will give the loan  v)annual income: -How much is the annual income to be carry for the student are they are also most important parameter which should be taken off.  So, these 5 parameters also take consideration if the students are included in our dataset |

|  |  |
| --- | --- |
| **Question 4** | Below are two…. |
| **Marks** | 12 marks |
| **Word Count** | Not more than 300 words |
| **Your Answer** | a) Calculate for each model (8 marks) 1. Precision 2. Accuracy 3. Recall 4. F1-score  **Model 1:**      **Model 2:**      b)  If you had to choose just one metric to compare both models  From the metric precision, recall, accuracy and f1\_score it will say that we choose 2nd model as compared to the 1st model  As we see that the for 2nd model we use cutoff of 0.5 in which we says the the precision is decided how much positive class predicted accurate positive so it will give us high value of 0.58,after that we see that the recall values in which it will decide how much negative value predicted correctly negative so it also gives a high value of 0.46,and the accuracy decides how accurate our model is so it will gives us very high value of 0.90 so our model has 90% accuracy after that we see f1 score which gives ratio of positive and negative class value which also gives a high value of 0.51 so for selecting all the value we select 2nd model as the best model as compare to 1st model |