****

**OPMA 631 (Winter 2023)**

**Predictive Analytics**

**INDIVIDUAL ASSIGNMENT #2**

*Due date: Thursday, Mar 30th, 2023 by 10:00 PM*

**Instructions:**

* Make sure that you download and install the latest version of RapidMiner Studio (**Version 10.1**) to complete this assignment. Using an earlier version might result in different final answers, and there will be a penalty for it.
* Submit your answers in **one self-contained pdf file** for grading purposes. In addition to including screenshots of your output, include screenshots of your processes with enough details to help me give you feedback without having to open all the submitted processes. The pdf file should have a cover page with your name, date, and assignment title. Name your file firstname\_lastname\_IA#2.pdf.”
* In addition to this main file, save and upload all relevant RapidMiner processes and output files. However, these files are only for my reference, i.e., I am not going to look for your answers in multiple files. In order to save your process, select *Export Process* under *File* menu. Note that you have to **open** the process first before exporting it. If you just select the process in the repository window without opening it first, then you’d end up exporting the open process instead of the selected one.

**Loan Repayment**

In the lending industry, investors provide loans to borrowers in exchange for the promise of repayment with interest. If the borrower repays the loan, then the lender profits from the interest. However, if the borrower is unable to repay the loan, then the lender loses money. Therefore, lenders would like to minimize the risk of a borrower being unable to repay a loan.

In this quiz, we will use publicly available data from *LendingClub*, a website that connects borrowers and investors over the internet. The dataset is in the provided file ***Loans.csv***. There are 9,575 observations, each representing a 3-year loan that was funded through the *LendingClub.com* platform between May 2007 and February 2010. There are 14 variables in the dataset described in the table below. We will be trying to predict NotFullyPaid, using all of the other variables as independent variables.

|  |  |
| --- | --- |
| Variable | Description |
| CreditPolicy | 1 if the customer meets the credit underwriting criteria of LendingClub.com, and 0 otherwise. |
| Purpose | The purpose of the loan (one of “Credit Card,” “Debt Consolidation,” “Educational,” “Major Purchase,” “Small Business,” or “Other”). |
| IntRate | The interest rate of the loan, as a proportion (a rate of 11% would be stored as 0.11). Note that borrowers judged by LendingClub to be more risky are assigned higher interest rates. |
| Installment | The monthly installments ($) owed by the borrower if the loan is funded. |
| LogAnnualInc | The natural log of the self-reported annual income of the borrower. |
| Dti | The debt-to-income ratio of the borrower (amount of debt divided by annual income). |
| Fico | The FICO credit score of the borrower. |
| DaysWithCrLine | The number of days the borrower has had a credit line. |
| RevolBal | The borrower’s revolving balance (amount unpaid at the end of the credit card billing cycle). |
| RevolUtil | The borrower’s revolving line utilization rate (the amount of the credit line used relative to total credit available). |
| InqLast6mths | The borrower’s number of inquiries by creditors in the last 6 months. |
| Delinq2yrs | The number of times the borrower had been 30+ days past due on a payment in the past 2 years. |
| PubRec | The borrower’s number of derogatory public records (bankruptcy filings, tax liens, or judgements). |
| NotFullyPaid | 1 if the loan was not paid back in full, and 0 otherwise. |

**Data Partitioning:** Randomly split the dataset into a training set and a validation set.Use *Shuffled Sampling* and *local random seed* value of 1985 to partition your data (70% for training and 30% for validation).Use this partitioning for both Problem 1 and Problem 2.

**Problem 1.** [22 marks]

1. Identify all categorical predictors and list them here in a table. For each variable, identify the comparison group as the group with the largest number of observations and include them in a separate column in your table. [2 marks]

|  |  |
| --- | --- |
| Variable | Most Observations |
| Credit Policy | 1 |
| InqLast6mths | 0 |
| Delinq2yrs | 0 |
| PubRec | 0 |
| NotFullyPaid | 0 |
| Purpose | debt\_consolidation |

1. Build a logistic regression model that predicts the depended variable NotFullyPaid using all of the other variables as independent variables. Create dummy variables for the categorical variables using the comparison groups identified in part b.

Save and export **one** RapidMiner process that you can use to answer parts d, e, and f below. Name this process “FirstName1.rmp” (e.g., mine would be Alireza1.rmp).

You will need to submit this process on D2L. [4 marks]

1. Write down your resulting logistic regression model as a complete linear equation. How many variables are significant in your model (at 95% confidence level)? [1 mark]

y = 8.072888818999285 + 2.2295960773555565IntRate - 0.45268648644641823LogAnnualInc - 0.006729729791636658Dti - 0.008624309543037693Fico - 2.0976559773328748E-6DaysWithCrLine + 4.813370589634797E-6RevolBal + 0.001525796200453769RevolUtil + 0.001199773892342331Installment + 0.20272865996023612CreditPolicy.0 - 0.08824393038392128Purpose.credit\_card + 0.36853522553813917Purpose.all\_other + 0.33687832281267893Purpose.home\_improvement + 0.7773271742074032Purpose.small\_business - 0.12870674326690393Purpose.major\_purchase + 0.3922754531958026Purpose.educational + 0.3090630084440846InqLast6mths.1 + 0.12986667222904943InqLast6mths.2 + 0.5259741326177596InqLast6mths.3 + 0.6825823429425685InqLast6mths.4 + 0.7965782714432371InqLast6mths.5 + 0.9047736444378626InqLast6mths.6 + 1.1132910544885921InqLast6mths.8 + 1.141234872621697InqLast6mths.7 + 13.110351126463112InqLast6mths.33 + 1.0718806476556042InqLast6mths.9 + 1.5450474117358939InqLast6mths.18 + 12.952462540957143InqLast6mths.14 + 1.9431509014653132InqLast6mths.15 + 0.9204313393301857InqLast6mths.13 + 1.6664369904937668InqLast6mths.12 - 0.17525390546022798InqLast6mths.10 - 8.973697258227743InqLast6mths.19 + 0.664103264347199InqLast6mths.11 - 7.539497321395908InqLast6mths.16 - 9.036885814188919InqLast6mths.31 - 8.28095211979679InqLast6mths.24 - 7.841032684778775InqLast6mths.17 - 0.1353138397101174Delinq2yrs.2 - 0.000235709216271238

1. Consider two loan applications, which are identical other than the fact that the borrower in Application A has a FICO credit score of 700 while the borrower in Application B has a FICO credit score of 710. What can you say about loan A’s odds of not being paid back in full compared to loan B? [2 marks]

For Loan A, FICO score of 700:

log(odds of not being paid back in full) = 8.072888818999285 - 0.008624309543037693\*700 + ...

log(odds of not being paid back in full) = **2.526573354188874**

For Loan B, FICO score of 710:

log(odds of not being paid back in full) = 8.072888818999285 - 0.008624309543037693\*710 + ...

log(odds of not being paid back in full) = **2.440329448645836**

1. What is the accuracy and sensitivity of the logistic regression model on the **validation set** using a threshold of 0.5? How does this compare to the baseline model of part a? [2 marks]
2. What is the accuracy and sensitivity of the logistic regression model on the **training set** using a threshold of 0.3?

Save and export a new RapidMiner process that you can use to answer part g. Name this process “FirstName2.rmp” (e.g., mine would be Alireza2.rmp). You will need to submit this process on D2L. [2 marks]

1. LendingClub assigns the interest rate to a loan based on their estimate of that loan’s risk. This variable IntRate, is an independent variable in our dataset. In this part, we will investigate just using the loan’s interest rate as a “smart baseline” to order the loans according to risk. Using the training set, build a logistic regression model that predicts the dependent variable NotFullyPaid using IntRate as the only independent variable.

Save and export **one** RapidMiner process that you can use to answer parts i, j and k. Name this process “FirstName3.rmp” (e.g., mine would be Alireza3.rmp). You will need to submit this process on D2L. [2 marks]

1. Write down your model as a linear equation here. [1 mark]
2. Is IntRate significant in this model? Was it significant in the first logistic regression model? If you observe any difference, how would you explain this difference? [2 marks]
3. Now use backward elimination approach to select which variables to include in your predictive model (create dummy variables after variable selection). Use a logistic regression model inside your backward elimination operator and accuracy on the validation set to select the best variables. Feed the selected variables into another logistic regression model fitting your model using the same training set.

Save and export **one** RapidMiner process that you can use to answer parts m and n below. Name this process “FirstName4.rmp” (e.g., mine would be Alireza4.rmp). You will need to submit this process on D2L. [2 marks]

1. Which variables are excluded from your model? [1 mark]
2. Write down the confusion matrix for threshold of 0.7. [1 mark]

**Problem 2.** [8 marks]

1. Build a CART model to predict NotFullyPaid, using all of the other variables as independent variables. Use information\_gain as your criterion and maximal depth of 5 to limit the size of your tree. Also, set the minimal leaf size parameter to 50.

Save and export **one** RapidMiner process that you can use to answer parts b, c and d. Name this process “FirstName5.rmp” (e.g., mine would be Alireza5.rmp). You will need to submit this process on D2L. [3 marks]

1. Plot the CART tree. Which variables were used in the tree? Which variable appear to be the most significant? [2 marks]
2. Write down the three shortest rules generated by your tree in the shortest possible way (i.e., remove any redundancies). [2 mark]
3. What is the accuracy and sensitivity of the model on the validation set for cutoff value of **0.7**? [1 mark]