Business Analytics Fall 2021

Semester Project: BBA-7A&B   Marks: 50

Background: The data provided for this project is from the loans department of a bank. Based upon certain attributes, it is given if there was a default by the customer on the given loan or not. Hence it is a classification problem requiring you to apply relevant techniques to classify customers as defaulters or otherwise.

Dataset attribute information is given below. There are 16 variables and a class label.

Attribute	Description
CaseNo	Loan case file number (for identification only)
Default	Default on Loan 1=Yes (Class Label or outcome, factor variable)
CreditBalPerc	Percentage of credit balance (against credit limit)
DebtPerc	Percentage of debt (against total income)
LateUpto60Days	No of times loan installment was up to 60 days late
Late60to90Days	No of times loan installment was between 60 and 90 days late
LateOver90Days	No of times loan installment was over 90 days late
Income	Income of the applicant in dollars
NoOfOpenLoans	Total no of loans (home, mortgage, car, others etc.) against the applicant
NoOfHomeLoans	Total no of home loans against the applicant
Dependents	Number of dependents of applicant
Age	Age of Applicant
Job	Job of Applicant (factor variable)
Status	Marital status of applicant (factor variable)
Education	Education level of applicant (factor variable)

Project goal: To predict if the customer will default on loan provided by the bank.

Use Bluesky Statistics and MS Excel to get answers to the questions. Show complete calculations/steps and not just the answer for questions asking for accuracy, precision and recall etc. Provide all answers to 4 decimal places.

Fully understand all the variables and their types. Do not use case ID in any model as it is there for identification of records only. Split the data randomly into a training set and a testing set in the beginning and use later for different questions.

Note: Each group will use the seed value for random sampling as R No of the group leader. The split percentage should be 70 percent (for training) and remaining (30%) for testing purposes.

Build classification models and answer following questions.

Question 0 [Data Exploration]

Answer the following questions. [Can be answered by using crosstabs and Excel pivot tables]

a) How many observations are there in the dataset?
b) What percentage of clients defaulted from the loan?
c) What percentage of employed clients are loan defaulters?
d) What percentage of unemployed clients are loan defaulters?
e) What percentage of married clients are loan defaulters?
f) What are the top 3 jobs according to employee counts in the dataset?
g) Which 3 job types have the most loan defaulters?
h) What percentage of clients have tertiary education?

Question 1 [Logistic Regression Model]

Build a logistic regression model using the training set to predict whether an individual would default, using all of the other variables as independent variables. Ignore any warning messages. Answer following questions.

a) List all variables in your model which are significant (have one or more *).
b) Interpret the coefficient of ‘lateOver90days’ and ‘Jobunemployed’ from your model.
c) Score the model on the test data and provide the confusion matrix for the test data using the model.
d) Find the accuracy, precision and recall for this algorithm on the test set.

Question 2 [Decision Tree - CART Model]

Build a classification tree to predict default. Use the training set to build the model, and all of the other variables as independent variables. Provide a clear plot of the tree.

a) Which variable is the root and how many splits are there in the tree?
b) Provide the confusion matrix for the test data.
c) Find the accuracy, precision and recall for this algorithm on the test set.
d) Write all rules extracted from the decision tree.

Question 3 [Naïve Bayesian Classification]

Build a naïve Bayesian classification model using the training set to predict whether an individual would default, using all of the other variables as independent variables. Answer following questions.

a) Provide the confusion matrix for the test data.
b) Find the accuracy, precision and recall for this algorithm on the test dataset.

Question 4 [Random Forest Model]

Build a random forest model (make 300 trees) to predict default. Use the training set to build the model, and all of the other variables as independent variables.

a) Score the model on the test data and provide the confusion matrix for the test data.
b) Find the accuracy, precision and recall for this algorithm on the test set.

Question 5 [K Nearest Neighbor Model]

Build a kNN model on the whole dataset to predict default using K=35, 75 and 125 nearest neighbors. Use seed value as R No of group leader, 75 percent dataset to train the model and remaining 25 percent to evaluate the model.

a) Provide the confusion matrix for the test data.
b) Find the accuracy, precision and recall for this algorithm on the test set.
c) Which value of K among these provides the maximum accuracy.
d) Compare all model evaluation metrics (accuracy etc…) obtained by the five algorithms by making a comparison table and comment on the findings.

Question 6 [Classifying unknown data]

The bank's CEO has provided you with some data from pending loan applications. This data is given on the last page. He wants to know who among these ten clients are likely to default if their loan is approved.

Using only your 'Logistic Regression model', make predictions for the missing target variable 'Default' based upon provided attribute values for each case. Prepare your response for the CEO accordingly.

Question 7 [Recommendations to CEO]

Write a concluding paragraph and provide a set of recommendations to the CEO based upon your entire analysis to enable him to gain insight into his loan approval process so that the bank reduces loan defaulters. The CEO is particularly interested in particular customer attributes and characteristics leading to default, so you must focus on those too. You may also use any other method or technique not covered in class to answer this part to get extra credit.

Note that answers and findings of last two questions have to be presented to bank's CEO who may not understand analytics procedures. Prepare your answers accordingly. This is the part where your group can get a big edge over others if your analysis and report presentation is perfectly and appropriately done.

Important to Note:

This project cannot be done in the last two days. Start early to deliver a quality report.

If a group member did not contribute at all, the group leader must inform me in private comments immediately after submission.

Read all instructions carefully and follow them.

Protect your work, do not allow your work to be copied and do not copy from other sources.

There will be a demo to grade the project which will include a viva exam too. All group members are required to be fully aware of the entire project (all questions) and not its parts.

For project related queries, email immediately, or see me in office (preferably with laptop) to resolve queries/issues in working etc…

Submission Requirements:

A single word doc (formal report style with a title page) with answers to all questions & including relevant clear screenshots, to be submitted an hour 'before' the deadline by group leader on Google Classroom. Nothing else needs to be submitted.

Late submission will lose 50% credit and no excuses will be entertained.

Project report file to be renamed with names of group members.

List of Applicants for deciding Loan Approval

(Data for Q No. 6)

Case No	Default	Credit BalPerc	Debt Perc	LateUpto 60Days	Late60 to 90Days	LateOver 90Days	Income	NoOf Open Loans	NoOf Home Loans	Dependents	Age	Job	Status	Education
35D	??	0.1681	0.0804	0	0	0	5000	16	0	1	49	management	married	secondary
21W	??	0.0555	0.6098	0	0	0	4335	7	1	2	46	blue-collar	married	unknown
42T	??	0.1041	0.4777	0	0	0	10316	10	2	0	59	management	divorced	tertiary
65Y	??	0.3857	0.4043	0	0	0	3400	7	0	0	50	technician	divorced	tertiary
69P	??	0.7282	0.8247	0	0	0	3000	10	2	1	31	management	divorced	secondary
58K	??	0.1331	0.1829	1	0	0	10257	9	2	3	49	admin.	married	unknown
32A	??	0.2997	0.7166	0	0	0	5584	4	1	2	44	self-employed	married	primary
76B	??	0.9647	0.3829	3	1	3	13700	9	1	2	40	technician	married	primary
71C	??	0.0257	0.4758	0	0	0	3000	7	1	2	38	services	single	unknown
12N	??	0.3922	1.5953	0	0	0	4676	14	3	1	50	unemployed	married	primary