SUSS

SINGAPORE UNIVERSITY
OF SOCIAL SCIENCES

ANL307
Predictive Modelling

Tutor-Marked Assignment

JANUARY 2025 SEMESTER

Attempt all parts and label the answer for each part clearly

The dataset originates from Bank ABC and pertains to home improvement loans. The bank's data analytics department aims to identify customers with a high potential for defaulting on their home loans. The field "*loan_status*" reflects the default status, where 0 = No Default and 1 = Default. You are to answer the following questions to assist Bank ABC in gaining deeper insights into its loan customers.

Part 1 (23 marks)

Download the dataset and read it into the IBM SPSS Modeler. Complete the following tasks and answer the questions.

(*Note: no need to paste SPSS Modeler streams or nodes in your answer*)

a. Use a suitable node to read in the data. State the node used to read in the data and report the data size.

b. Report the suitable measurement type for each field.

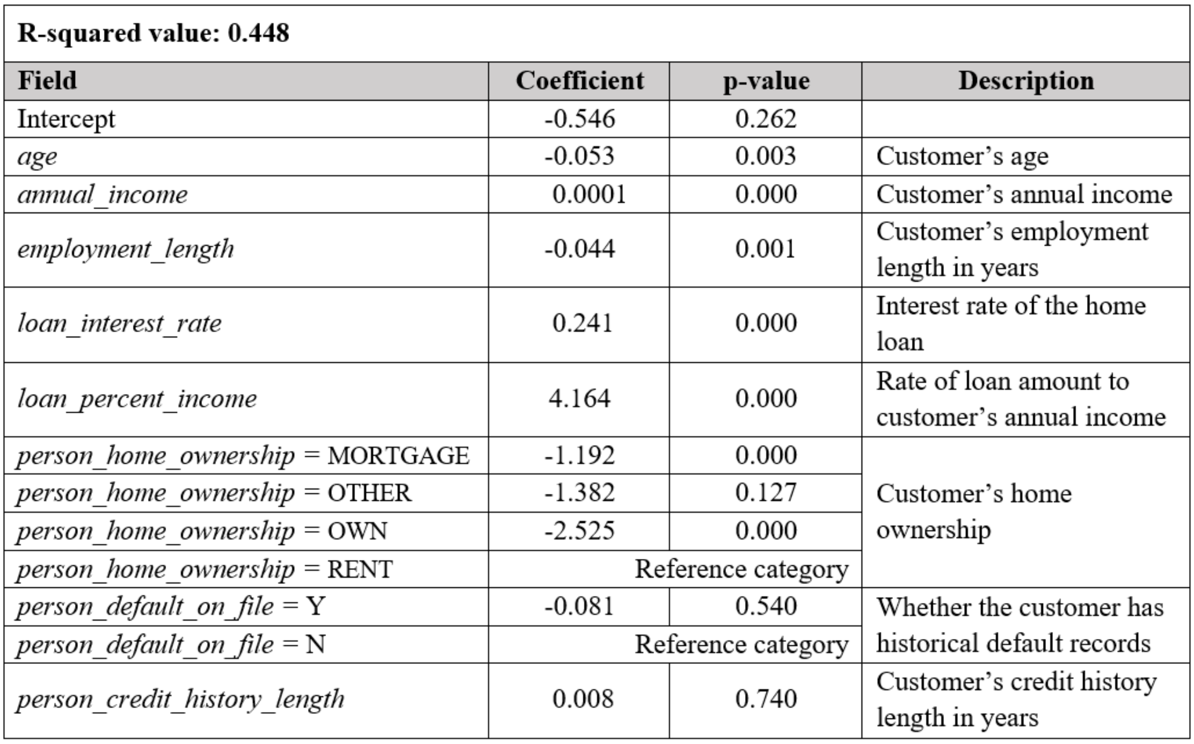
c. Define the role of field "*loan_status*" as a Target and other fields as Input. Report the distribution of the target field (i.e., the number of records with *loan_status* = 0 and *loan_status* = 1) and discuss whether there is an imbalance issue.

d. Connect a Partition node to the data source node. Set training and testing ratio to be 80% and 20% respectively. **Do not change** the rest of the node settings. Report the data size of training set and testing set.

Part 2 (35 marks)

A Logistic Regression model is constructed using a very similar dataset with the same attributes and the same settings in Part 1. Answer the following questions by interpreting the modelling results as shown in the following table.

(*Note: no need to implement the Logistic Regression model in SPSS Modeler; keep your answers to the Part 2 questions within 250 words*)



R-squared value: 0.448

Field	**Coefficient**	**p-value**	**Description**
Intercept	-0.546	0.262	
age	-0.053	0.003	Customer’s age
annual_income	0.0001	0.000	Customer’s annual income
employment_length	-0.044	0.001	Customer’s employment length in years
loan_interest_rate	0.241	0.000	Interest rate of the home loan
loan_percent_income	4.164	0.000	Rate of loan amount to customer’s annual income
person_home_ownership = MORTGAGE	-1.192	0.000	Customer’s home ownership
person_home_ownership = OTHER	-1.382	0.127	
person_home_ownership = OWN	-2.525	0.000	
person_home_ownership = RENT	Reference category		
person_default_on_file = Y	-0.081	0.540	Whether the customer has historical default records
person_default_on_file = N	Reference category		
person_credit_history_length	0.008	0.740	Customer’s credit history length in years

a. Report the significant fields.

b. Compare the effects of *loan_percent_income* and *person_credit_history_length* on the risk of defaulting on a home improvement loan.

c. For the field *person_home_ownership*, if we change the reference category to the value of “OWN” and rerun the model, would this affect the R-squared value? And would this change affect the learned coefficients of the fields?

d. Can we conclude that the learned Logistic Regression model is good for deployment? Justify your answer briefly.

Part 3 (42 marks)

Connect a CART node to the Partition node and run the model using the **default** settings. Then, answer the following questions regarding the learned decision tree.

(*Note: no need to paste SPSS Modeler streams or nodes in your answer*)

a. Report the depth of the learned decision tree.

b. Report the top four most important fields and compare them with the significant fields found in Part 2 using Logistic Regression. Discuss any similarities or differences in the findings briefly.

c. Use an appropriate node to evaluate the modelling performance, specifying the node used.

d. Report the prediction performance on the testing set; your answer should include the overall accuracy, sensitivity, and hit-rate of *loan_status* =1. Appraise the CART model’s effectiveness in identifying customers at risk of defaulting on their home loans.

---- END OF QUESTION PAPER ----