Rider University

Norm Brodsky College of Business

PMBA 8312

Module 5 Assignment

Name:

Akash Gopalkrishnan

Todd Lawson

Tiffany Hunter

Instructor:

Dr.Mashayekhi

Format: Times New Roman (font size =12), single-space.

Please export the process in RM and submit them with this file.

Part 1: Business Understanding (Problem): (10 marks)

Explain in a few sentences what we try to predict and how it helps a bank financially.

At a high level, the business objective is to predict the likelihood that loan applicants will default on their loan for the greater purposes of revenue generation and avoiding the risk of default. More specifically, the bank wants to profile customers into 2 groups: good and bad credit rating. The profile will be based on 6 features including credit rating, age, income, number of credit cards, education, and number of loans.

Part 2: Data Understanding (10 marks)

2-1 what is the sample size?

2,464 examples

2-2 Complete the following table:

Variable	Average	Standard Deviation	Min	Max
Age	33.816	8.539	20.003	63.350

2-3 For each of the following categorical (nominal) variables, insert a table showing the absolute Frequency, relative Frequency in percentage, and cumulative percentage.

Credit_rating, Income, Credit_cards, Education, and car_loans

Below is an example table for the Credit_rating variable:

Credit_rating	Absolute Frequency (count)	Relative Frequency (fraction) %	Cumulative Percentage %
Good	1444	0.586	58.6
Bad	1020	0.414	100

Income	Absolute Frequency (count)	Relative Frequency (fraction) %	Cumulative Percentage %

Medium	1134	0.460	46.0
High	777	0.315	77.5
Low	553	0.224	99.9

Credit_cards	Absolute Frequency (count)	Relative Frequency (fraction) %	Cumulative Percentage %
5 or more	1666	0.676	67.6
Less than 5	798	0.324	100

Education	Absolute Frequency (count)	Relative Frequency (fraction) %	Cumulative Percentage %
College	1234	0.501	50.1
High school	1230	0.499	100

car_loans	Absolute Frequency (count)	Relative Frequency (fraction) %	Cumulative Percentage %
More than 2	1571	0.638	63.8
None or 1	893	0.362	100

Part 3: Modeling (25 marks)

3-1 what are the features in the model?

Credit_rating, Age, Income, Credit_cards, Education, Car_loans

3-2 What is the target variable (label)?
Credit_rating: Good/Bad

3-3 What is the positive class of the target variable?

Bad

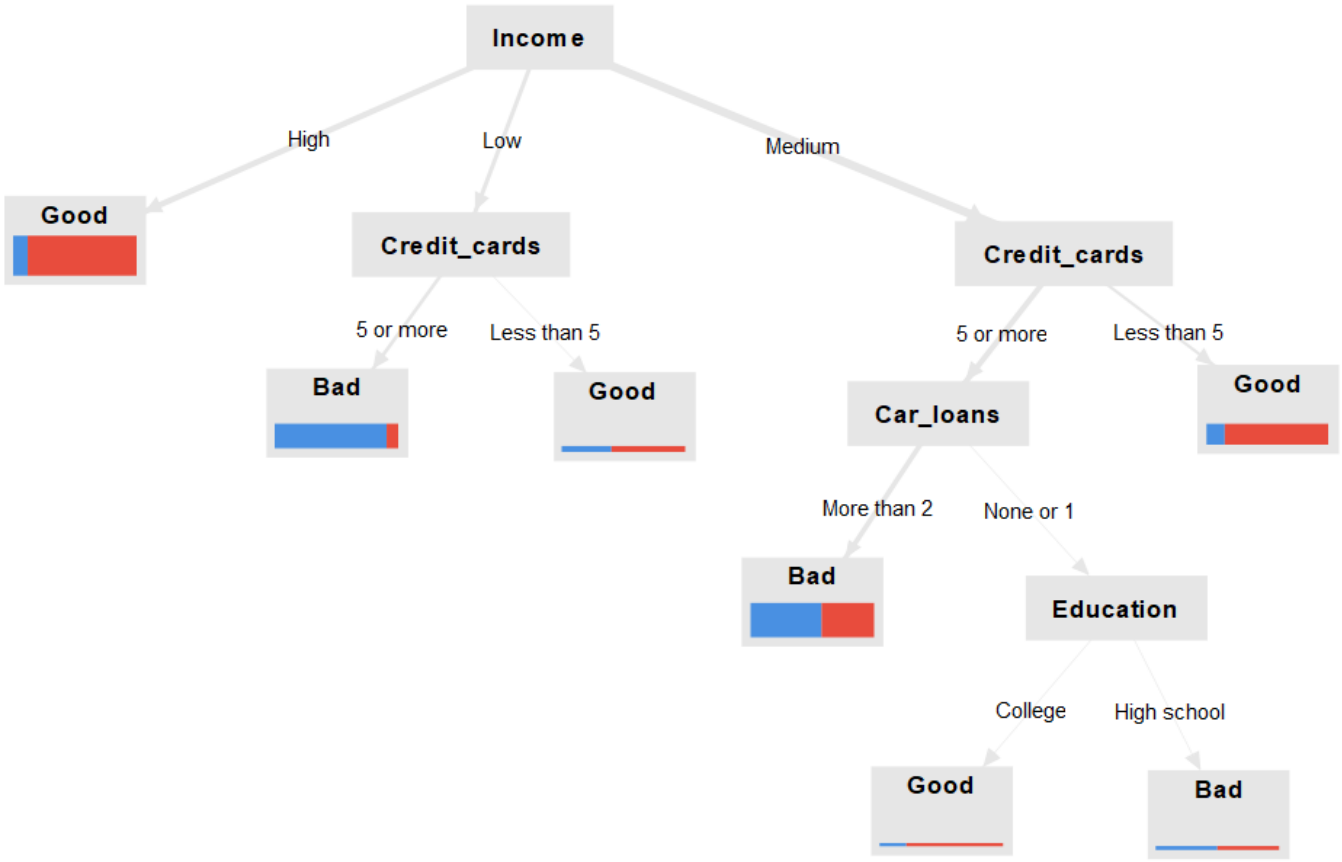
3-4 Why do we need to use classification instead of regression?

Classification is needed because the likelihood of whether customers will default on their loans depends on the categorical variables such as good or bad credit. Regression is used for a label (target variable) that has a measurement level of continuous. For this model, the measurement level of Credit_rating is categorical (e.g. Good/Bad), thus a classification model is required.

3-5 Describe the decision tree for the model (how many nodes, how many terminal nodes, depth, variables in the tree)

Total number of nodes (including the root node)?	12
Number of terminal nodes?	7
The depth (not include the root node)?	4
List the name of variables in the tree	Income, Credit_cards, Car_loans, Education
What is the most important predictor of credit rating based on the tree?	Income

3-6 Copy and paste the decision tree here (take a screenshot, resize it and paste it here. Make sure the image is readable)



Part 4: Model Evaluation (25 marks)

4-1 Complete the confusion matrix based on the "Performance Test" results:

Predictive Model for Credit Rating		Actual	
		Bad	Good
Predicted	Bad	TP= 167	FP=67
	Good	FN=37	TN=222

4-2 Complete the following table: (positive class: Bad)

	General Accuracy	Recall	Precision
Performance - Cross Validation	77.78% +/- 1.89%	79.41% +/- 4.39%	70.58% +/- 1.57%
Performance - Test	78.90%	81.86%	71.37%

Note: for performance-CV, you should report average and standard deviation (e.g., recall: 79.41% +/- 4.39%)

4-3 Describe two groups of bad credit individuals and two groups of good credit Individuals based on the features in this model (for example, individuals with 5 or more credit cards and low income for bad credit groups)

	Describe
Group 1 (bad credit)	low income and 5 or more credit cards
Group 2 (bad credit)	medium income, 5 or more credit cards, and more than 2 car loans
Group 1 (good credit)	High income
Group 2 (good credit)	Low income and less than 5 credit cards

4-4 For the following individuals with specific features, make predictions based on the model (Bad or Good Credit rating)

Profile	Prediction
Sara with High income	Good

John, age 20 with more than 5 credit cards (say, is this right-there is no age variable in the tree – paste the tree in email)	No prediction possible from this model
Mike, with less than 5 credit cards and medium-income	Good

Part 5: Answer the following questions. (30 marks)

5-1 Is this model underfitting, overfitting, or just the right fit? Why?

The model is underfitting because the model performs poorly in accuracy during cross-validation; it's less than 80%: 77.78% +/- 1.89%. And even with the uncertainty added, it's only 79.67%, so there is no need to test for overfitting using the test dataset (Reference: slide 36 of Module 5 PowerPoint presentation).

5-2 Provide at least two suggestions to improve the model. (Please note that the predictive model is supposed to detect bad credit customers who are more likely to default on their loans)

Since the model is underfitting, to improve the model, one could increase the sample size. Another suggestion is to do PCA analysis so that only the highly correlated features are analyzed by the model.

5-3 Explain which error (FP or FN), in your opinion, should be prioritized to minimize. Explain what would be the consequences of your choice (give priority to FP or FN) in terms of generating revenue and avoiding the risk of defaults.

In this context, false negative means it's really a positive: the person actually has bad credit but the model predicted that the person has good credit. Similarly, false positive means it's really a negative: the person actually has good credit but the model predicted that the person has bad credit. In my opinion, minimizing false negatives should be prioritized because accepting bad credit applicants could result in high risk of default. However, due to the tradeoff between false negative and false positive, that is, when false negatives are decreased false positives increase, there are consequences to minimizing false negatives. The consequence of increasing false positives (rejecting good-credit applicants), is that the bank will suffer loss of revenue from rejecting these profit-bearing customers. Thus, if the bank manager is more risk-averse and prioritizes risk-minimization over revenue generation, then he should prioritize minimizing false negatives; and if he prioritizes revenue generation over minimizing default risk (ie: is more risk-prone), then he should prioritize minimizing false positives.