# MODULE 7 UNIT 3

## Activity submission

Learning outcomes:

LO4: Analyse the results of a classification tree analysis.

LO5: Reflect on the use of classification tree analysis.

### Name: Matlotlo Magasa

#### 1. Instructions and guidelines (Read carefully)

##### Instructions

1. Insert your name and surname in the space provided above, as well as in the **file name.** Save the file as: **First name Surname M7U3 Activity Submission** – **e.g. Lilly Smith M7U3 Activity Submission.** **NB:** *Please ensure that you use the name that appears in your student profile on the Online Campus.*

2. Write all your answers in this document. There is an instruction that says, “Start writing here” under each question. Please type your answer there. Where there is the instruction “Paste your Tableau visualisation here”, insert the exported image there.

3. Submit your assignment in **Microsoft Word only**. No other file types will be accepted.

4. You will be required to include visualisations that you have made in Tableau in this activity. To do this, export them as an image file and paste them into this document. Tableau has a page that details how to export your Tableau view as an image file.

5. Do **not delete the plagiarism declaration** or the **assignment instructions and guidelines**. They must remain in your assignment when you submit.

**PLEASE NOTE:** **Plagiarism cases will be investigated in line with the Terms and Conditions for Students.**

**IMPORTANT NOTICE:** Please ensure that you have checked your course calendar for the due date for this assignment.

##### Guidelines

1. There are 9 pages and 3 questions in this assignment.

2. This activity submission relies heavily on work done in the IDE activity in the previous unit. If you have not done so already, return to the IDE notebook and complete this activity before attempting to answer the questions in this assignment.

3. You will be required to download a data file, titled LOANS.sav, in order to interact with the data in Tableau. Return to the Module 7 downloads folder and save a copy of the data file before attempting to answer the questions.

4. Make sure that you have carefully read and fully understood the questions before answering them. Answer the questions fully, but concisely, and as directly as possible. Follow all specific instructions for individual questions (e.g. “list”, “in point form”).

5. Answer all questions in your own words. Do not copy any text from the notes, readings, or other sources. **The assignment must be your own work only.**

|  |
| --- |
| **Plagiarism declaration:** |
| **1. I know that plagiarism is wrong. Plagiarism is to use another’s work and pretend that it is one’s own.**  **2. This assignment is my own work.**  **3. I have not allowed, and will not allow, anyone to copy my work with the intention of passing it off as his or her own work.**  **4. I acknowledge that copying someone else’s assignment (or part of it) is wrong and declare that my assignments are my own work.** |

#### 2. Mark allocation

Each question receives a mark allocation. However, you will only receive a final percentage mark and will not be given individual marks for each question. The mark allocation is there to show you the weighting and length of each question.

Question 1 15

Question 2 15

Question 3 10

**TOTAL 40**

#### 3. Questions

Speedy Loans’ management team is interested in creating a model that can help them decide whether the company should approve or deny applicants a loan, based on information included in their applications. The information collected includes each applicant’s age, years at their current employer, education, location, income, current outstanding debt, and risk score. You decided that creating a classification tree model would be an intuitive way to address the team’s questions.

##### Question 1

After training and creating the classification tree, you performed cross-validation and plotted the misclassification error for the training data set and testing data set, along with the cross-validation error. The team has asked you to compare these results to those obtained for the HBAT data set generated in Unit 2 of this module. The misclassification error plot for each decision tree model is illustrated in Figures 1 and 2, respectively, for your reference.

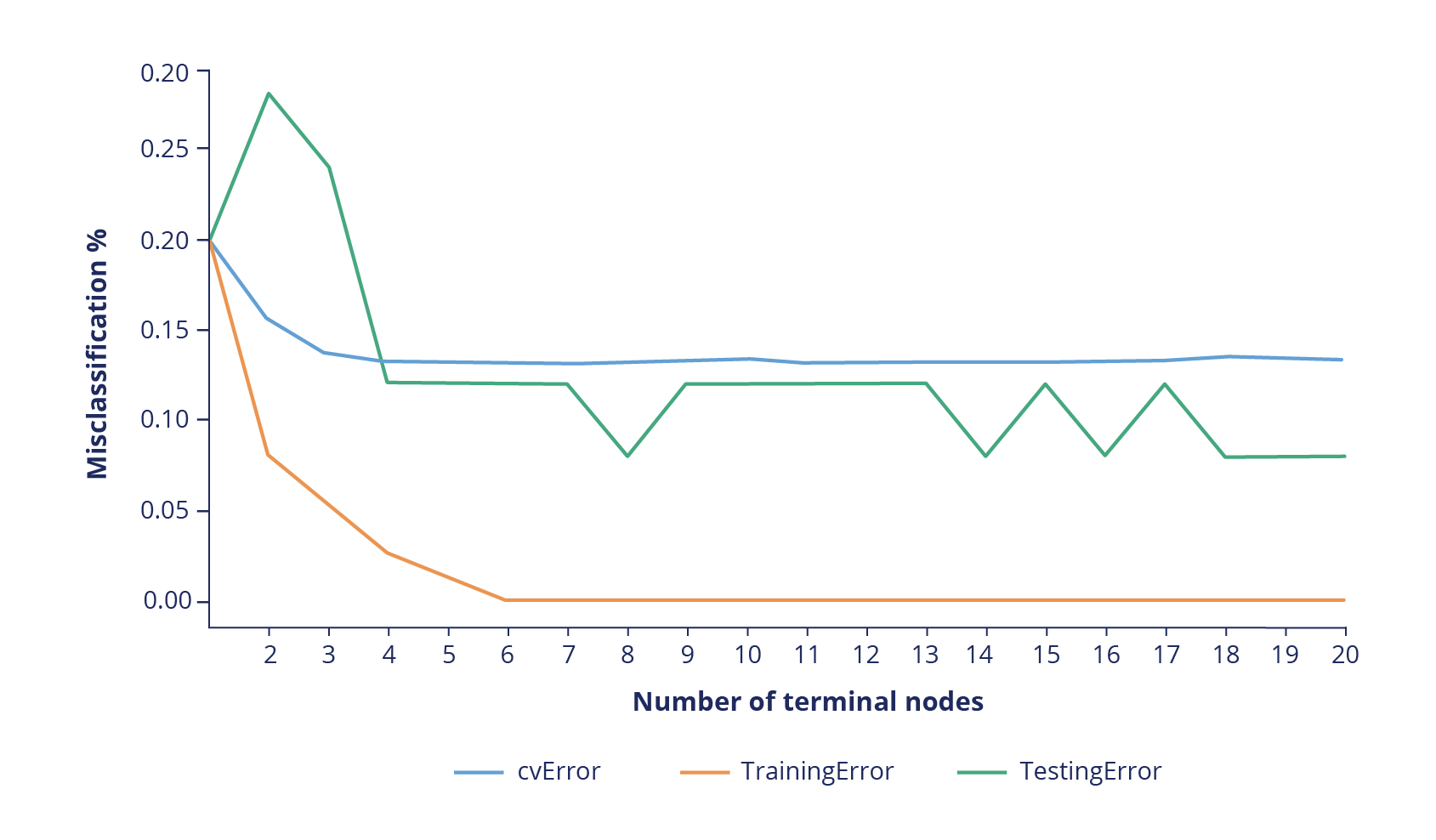


Figure 1: Misclassification error plot for the classification tree model on the HBAT data set.

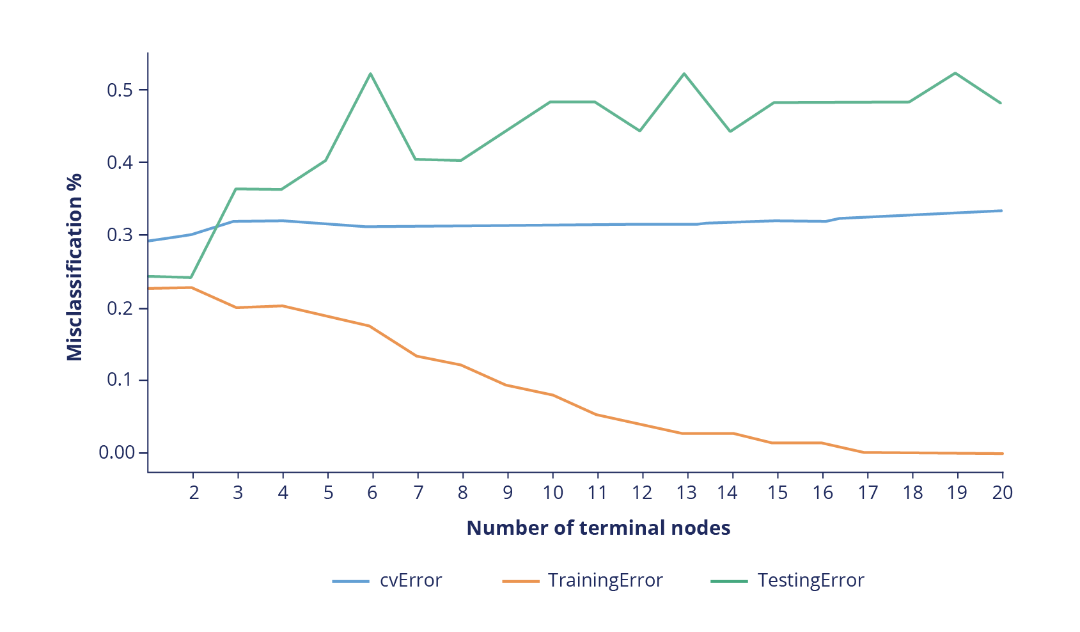


Figure 2: Misclassification error plot for the classification tree model on the Speedy Loans data set.

The team has asked you to provide your interpretation of the misclassification error plots and discuss what the results suggest about each model’s performance as the number of terminal nodes increase. They have asked you to consider the following when forming your response:

* How do the training misclassification errors compare between the models? What does this imply about identifying potential overfitting in the Speedy Loans model?
* The graphs show different trends in terms of the cross-validation error for each model. The HBAT model misclassification error decreases and plateaus, whereas the Speedy Loans model misclassification error seems to increase slightly before plateauing. What do these results mean and what implications might this have when attempting to decide on the appropriate number of terminal nodes?
* How many terminal nodes would you recommend for the Speedy Loans model? Substantiate your answer.

(Max. 400 words)

Start writing here:

The speedy loan plot displays large percentage error across all its plots compared to HBAT. For HBAT, the highest error being a little below 0.3 for testing data while Speedy loan ‘testing’ misclassification error reaches 0.5.

The cv error for the Speedy loan is almost twice the cv error for the HBAT model. Both models have the same number of observation and the same proportion of split between the training data and test data, but the HBAT algorithm seems to easily learn more about the data and all the random variations within the dataset better than the Speedy model. This is evident because the HBAT training curve plateaus much quicker than that of Speedy. This might be due to the number of predictor variables/features. HBAT has more predictor variables (13 predicter variables) while Speedy only has 9. This rapid drop can potentially cause overfitting problems because the model has learned all there is to know about the patterns in the data. Pruning, however allows us to curb effects of overfitting.

HBAT model can use the least predictor variables (nodes) to explain the greatest variation in the data better than Speedy.

There is little variation in the HBAT CV error and testing error, and it is much larger in Speedy. Speedy loans the cv error significantly lower than the testing error (high variation between the two). This means that in the Speedy model, the cv error underestimates the generalization of the model this is likely due to the small sample size of training data.

The Speedy model better displays the bias-variance trade-off example. At around node 3, there is low variance and a high bias. Pruning the tree at node 3 or 4 would mean model underfitting, where low proportions of data have been used in training the model and therefore give poor accuracy in testing data. This is evident in increasing of cv error at node 3 to about node 5 .

Seven nodes are the optimal point for the model. Nodes after 7 increase model variation and decreases bias. This decreases the generalization ability of the model promoting overfitting. Overfitting is seen with the continuing rise in testing error with the significant decrease in training error.

At node 7, there is a decrease or dip in the test error, this dip is slightly observed in the cv error curve and therefore at this point the model is more generalisable.

##### Question 2

Upon specifying six terminal nodes and creating the final classification tree, the team has asked you to draft some decision rules that can help to classify whether a potential customer should be denied or approved a loan from the company. The variables that make up the final classification tree are shown in Table 1 to help you draft your response.

Table 1: Variables important for classifying the Speedy Loans data set.

|  |  |  |
| --- | --- | --- |
| **Python-indexed name** | **Original variable name** | **Variable description** |
| X[0] | Age | Applicant’s age (in years) |
| X[2] | Income | Applicant’s income (in ZAR) |
| X[3] | Debt\_income | Applicant’s debt-to-income ratio |
| X[5] | Other\_debt | Applicant’s outstanding debt, other than credit card debt |

In your response, create a list of the six decision rules that can classify whether an applicant should be approved or denied a loan from the company.

Furthermore, imagine that the company has received an application from a potential client with the following profile: the applicant is 42 years of age, has an income of R42,300, has R3,500 in other debt, and has a debt-to-income ratio of 11.2. Which of the decision rules would you use to classify this applicant? Would the loan be approved or denied?

(Max. 500 words)

Start writing here:

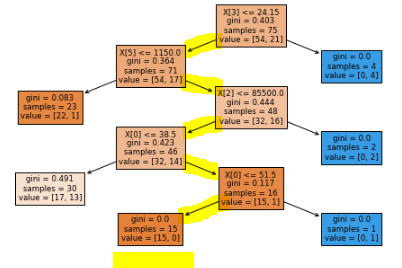
The model has been reduced and the final model has 6 terminal nodes and hence will have 6 Decision Rules, which are:

1. If a client has a debt-to-income ratio of more than 24.14 they will get *accepted* for a loan
2. Should the client have a debt-to-income ratio of 24.15 and less, has R1150 other debt or less they will be *declined* a loan.
3. If the client has a debt-to-income ratio of 24.15 and less, has over R1 150 of other debt and earns over R85 500 they will get *accepted* for a loan.
4. If the client has a debt-to-income ratio of 24.15 and less, has over R1 150 of other debt and earns R85 500 or less, and is 38.5 years of age or younger they will be *declined* a loan.
5. If the client has a debt-to-income ratio of 24.15 and less, has over R1 150 of other debt, earns R85 500 or less and is over 38.5 years old but 51.5-year-old or younger, they will be *declined* a loan.
6. Should the client have a debt-to-income ratio of 24.15 and less, has over R1 150 of other debt, earns R85 500 or less and is over 51.5-years-old, they will be *accepted* for a loan.

The decision rule that would apply to the client is decision rule 5:

The client has a debt-to-income ratio of less than 24.15, has over R1 150 of other debt, earns less than R85 500 and is over 38.5 years old but 51.5-year-old or younger. They will be *declined* a loan.

The client’s loan will be rejected. ☹



Summary:

The root node for the tree is the debt-to-income ratio. This is a reasonable variable to be the best predictor variable because this ratio is insightful to whether a client is likely or unlikely to repay their loan. A high debt-to-income ratio means that there is financial burden on a client, and they are less likely to repay their loan. However, according to the decision tree, a high (above 24.15) debt-to-income ratio is considered an automatic approval of the loan. This is very counter-intuitive and will likely contribute to a poor model for the business.

Except for one, all the leaf nodes have an exceptional Gini score that is closer to zero signifying a pure node which means most of the observations in those nodes belong to a single class. However, one of the leaves ends with a high Gini score of 0.491. This is a major risk in the decision process as the classes of loans being declined or accepted are spilt almost 50-50. This means poor business decisions in terms of loan allocation and is also be another factor in lowering the model accuracy.

Although the client in question did not get the loan, it is safe to presume that the client was declined mostly because of a bad decision tree rather than his financial profile.

##### Question 3

After creating the final classification tree model, the Speedy Loans team has asked you to provide insight into the performance of the model on unseen data. To answer this question, you plotted a confusion matrix on the testing data set to assess the ability of the model to classify observations correctly, as illustrated in Figure 3.

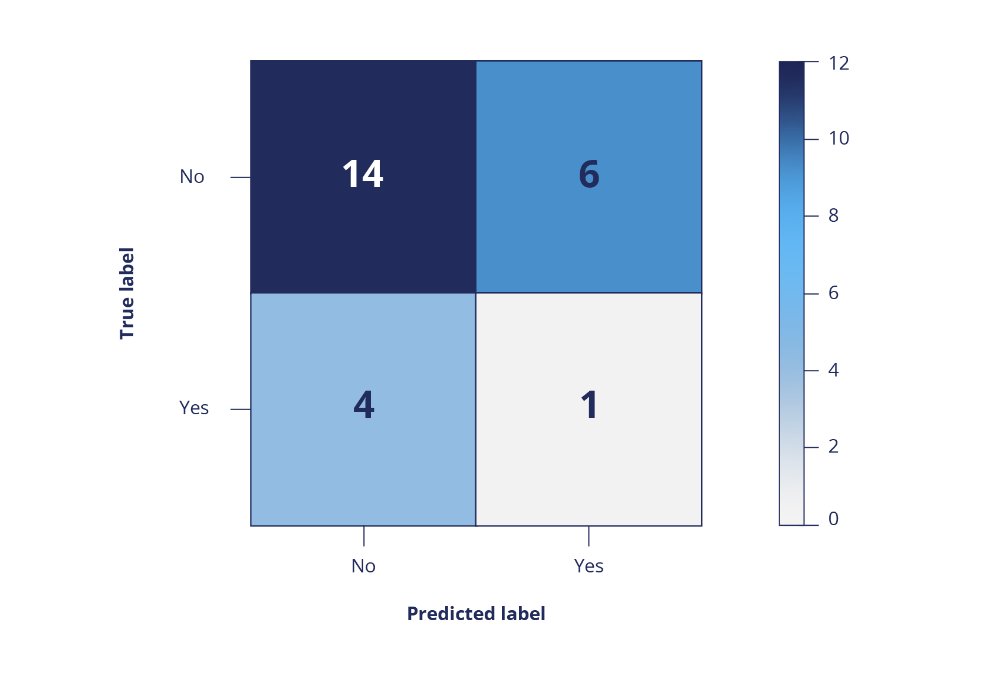


Figure 3: Confusion matrix for the Speedy Loans testing data set.

Based on the confusion matrix, provide a summary of whether the final classification tree model can be considered useful in classifying applicants. In other words, could this model be considered useful to decide whether an applicant should be approved or denied a loan? Substantiate your answer and explain how you would attempt to rectify the model in the case that it is not considered useful.

(Max. 300 words)

Start writing here:

The model misclassified the test data with 6 false positives (24%), and 4 false negatives (16%) with a total misclassification of 40% and therefore the hit rate is 60%.

The model is more effective at classifying the loan rejection which can be calculated as, . The classification of loan acceptance is less accurate and is only 20% and has been calculated by, .

A high misclassification of people who shouldn’t be getting a loan is not good for Speedy loan business and may cause a large loss for the business. Although the classification of loan rejections is much higher than that of the loan acceptance, it is generally not a good model because it does not sufficiently address the business question, which is to determine the *overall prediction* and not necessarily to determine one outcome, such as loan rejection. To correctly address the business question, it is important to have high accuracy of the model.

At a glance, it is easy to pronounce the model a poor model, however, the proportional chance criterion can help ascertain whether this is true or not.

Proportional chance criterion:

Because the denied loans have the most observation of 20, this will be used as the reference for the proportional chance criterion-

The hit rate i0f 60% is lower than the proportional chance criterion, therefore the model can be deemed to be a poor model fit.

To improve the model, the following can be done:

* Increase the number of independent variables in the model and the sample size
* Tweak the number of nodes in order to optimize
* Increase the sample size
* Consider boosting of the model
* Consider using a rando forest algorithm

#### 4. Rubric

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **No submission** | **Poor** | **Good** | **Excellent** |
| **Question 1: Comparing the models**  *The submission provides an accurate explanation of what the differences between the training errors imply about the identification of overfitting.* | No submission or attempt. (0) | The student has attempted to answer the question, but provides a largely inaccurate explanation of what the differences between the training errors imply about the identification of overfitting. (2.5) | The student provides a largely accurate explanation of what the differences between the training errors imply about the identification of overfitting. (4) | The student provides an accurate explanation of what the differences between the training errors imply about the identification of overfitting. (5) |
| **Question 1: Cross-validation errors**  *The submission provides a correct interpretation of what the cross-validation error graph means for the selection of an appropriate number of terminal nodes.* | No submission or attempt. (0) | The student has attempted to answer the question, but provides a largely incorrect interpretation of what the cross-validation error graph means for the selection of an appropriate number of terminal nodes. (2.5) | The student provides a largely correct interpretation of what the cross-validation error graph means for the selection of an appropriate number of terminal nodes. (4) | The student provides a correct interpretation of what the cross-validation error graph means for the selection of an appropriate number of terminal nodes. (5) |
| **Question 1: Number of terminal nodes**  *The submission provides a recommendation for the number of terminal nodes, with reasoning.* | No submission or attempt. (0) | The student provides a recommendation for the number of terminal nodes, but does not provide adequate reasoning for their answer. (2.5) | The student provides a recommendation for the number of terminal nodes and provides adequate reasoning for their answer. (4) | The student provides a recommendation for the number of terminal nodes and provides good reasoning for their answer. (5) |
| **Question 2: Decision rules**  *The submission correctly identifies six decision rules based on the final classification tree model.* | No submission or attempt. (0) | The student correctly identifies less than three decision rules.  OR  The student identifies six decision rules, but the rules have been interpreted incorrectly. (5) | The student correctly identifies four or more decision rules, but does not correctly identify all six. (7) | The student correctly identifies all six decision rules based on the final classification tree model. (10) |
| **Question 2: Classifying a potential client**  *The submission correctly states which decision rule would be used to classify the potential client, along with a final decision on whether to approve or deny the potential client a loan.* | No submission or attempt. (0) | The student incorrectly states which decision rule would be used to classify the potential client.  AND  The student does not provide a final decision on whether to approve or deny the potential client a loan. (2.5) | The student correctly states which decision rule would be used to classify the potential client.  BUT  The student does not provide a final decision on whether to approve or deny the potential client a loan. (4) | The student correctly states which decision rule would be used to classify the potential client.  AND  The student correctly provides a final decision on whether to approve or deny the potential client a loan. (5) |
| **Question 3: Usefulness of the model**  *The submission provides an accurate summary of whether the model is useful in classifying the loan status of potential clients, with reasoning.* | No submission or attempt. (0) | The student provides an accurate summary of whether the model is useful in classifying the loan status of potential clients, but does not provide adequate reasoning for their answer. (2.5) | The student provides an accurate summary of whether the model is useful in classifying the loan status of potential clients and provides adequate reasoning for their answer. (4) | The student provides an accurate summary of whether the model is useful in classifying the loan status of potential clients and provides insightful reasoning for their answer. (5) |
| **Question 3: Improving the model**  *The submission provides a recommendation for how to improve the model.* | No submission or attempt. (0) | The student provides recommendations for improving the model, but the answer is inaccurate or incomplete. (2.5) | The student provides accurate recommendations for improving the model and the answer is complete, but the insight offered is limited. (4) | The student provides accurate recommendations for improving the model. The answer is complete and offers substantial insight. (5) |

**Total:** 40 marks