# MODULE 6 UNIT 3

## Activity submission

Learning outcomes:

LO4: Analyse the output from a logistic regression.

LO5: Critique the use of logistic techniques.

### 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 M6U3 Activity Submission** – **e.g. Lilly Smith M6U3 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 7 pages and 3 questions in this assignment.

2. This activity submission relies heavily on work done in the notebook 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 6 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 5

Question 2 15

Question 3 20

**TOTAL 40**

#### 3. Questions

The management team at Speedy Loans is interested in better understanding the variables that affect whether a potential client will be approved or denied for a loan from the company. This will help the business to predict which applicants would be successful (or unsuccessful) when submitting an application. After collecting information on age, income, current outstanding debt, and risk score, the team has asked you to answer the following questions to help provide more insight.

##### Question 1

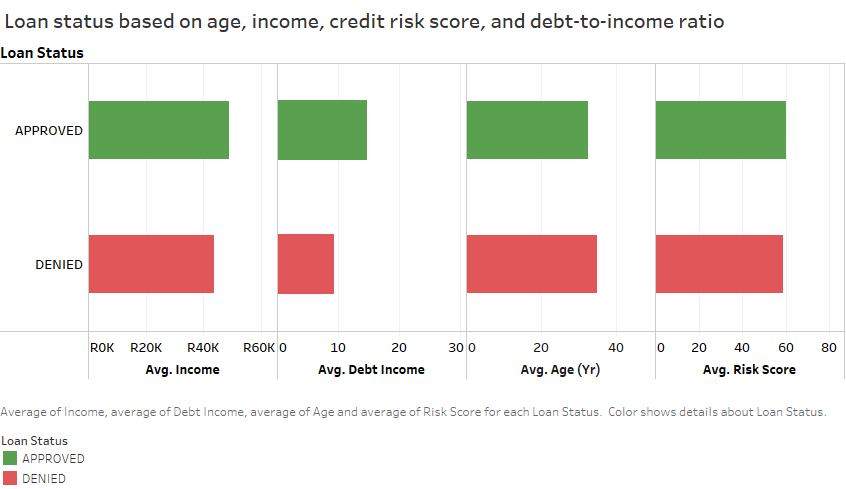
The team is interested in visualising some average statistics for potential clients who were denied and approved loans. However, the collected data has classified the potential clients’ loan approval statuses in a binary manner, with those who were denied a loan classified as a “0” and those who were approved for a loan classified as a “1”.

To make the results more informative for the board of directors, the management team has asked you to transform the “Loan” variable into categorical descriptions using an “IF ELSE” calculated field. In other words, if the “Loan” variable is equal to 0, then the potential client was “Denied”, otherwise they were “Approved”. After creating this new field in Tableau, name this variable “Loan status”.

Following this, the team has asked you to generate a table that shows the average age, income, risk score, and debt-to-income ratio of potential clients, separated based on whether they were denied or approved for a loan.

To do this, drag the “Loan status” variable into the “Rows” field, and drag the variables of interest (“Age”, “Income”, “Risk score”, and “Debt income”) into the “Columns” field. Then use the “Show me” tab to generate a table. Remember to change the measure for each variable from “SUM” to “Average”. Change the title of the worksheet to “Loan status based on age, income, credit risk score, and debt-to-income ratio”.

Paste your Tableau visualisation here:



##### Question 2

The team has asked you to provide your interpretation of the final logistic model generated for the Speedy Loans data set. They have specifically asked you to address the following in your response:

* Considering the null and residual deviances, is the final model considered a good fit compared to the simultaneous model that included all independent variables?
* Provide an interpretation of the logistic coefficients for the final model, including the nature of the relationship (either positive or negative) and how you would go about interpreting the relative importance of each of the variables in predicting whether a potential client would be denied or approved for a loan.

(Max. 400 words)

**(\*please** note that in the IDE I added the constant, ‘const’ as it was missing in the final logistic regression code. Therefore, there may be minor numerical value changes from original code)

Start writing here:

The null deviance for this model was 114.61. The simultaneous estimation model produces a model deviance of 90.152, while the final model produced a model deviance of 94.104

The deviance for final model increased slightly from full model. The difference between the final model deviance and null deviance is relatively small. Thus, final model is generally not a good fit, the large deviance means that the variables in the final model do not give a better explanation of the Y-variable (loan status).

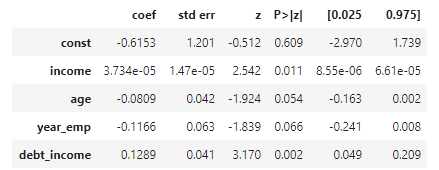
The model accuracy can be further explored using the Hosmer-Lemeshow chi squared model.

For the simultaneous estimation: The Final Model:

The Chi2 value and p-value of the both the full model and the final model are measuring poorly. Looking at the final model, the p value is closer to 0 than it is to 1 while the chi squared value is measuring high. This is indicative of a poor model and there is evidence that suggests that predicted and observed values are different or that the model displays high deviance from the observed values.

Only 4 independent variables are significant to the model (p-value < 0.1). The logistic coefficients of these variables show the ‘strength’ of the odds of loan approval.



The ‘year employed’, ‘age’, and intercept are negative. The coefficients are -0.1166, -0.0809 respectively, meaning that there are decreased odds that the loan will be approved per unit increase of these variables. An increase in the age of the client will decrease the odds of a client’s loan being approved by a factor of 0.081, other variables held constant. The year employed will decrease the odds of a client being approved of a loan by a factor of 0.11166, all other variables kept constant.

The income and debt-to-income ratio both have positive coefficients and increase the odds of loan approval. The increase in odds happens at a factor of 0.00004 and 0.1289 per unit increase in income and debt\_income respectively, all other variables kept constant in each instance. Note that income has almost no influence in changing the odds of loan status of the client

To determine the relative importance of the coefficients, make sure the variables are first standardized. The coefficients will then need to be exponentiated (reverse log odds), refer to Table 1.

Table 1: Exponentiated coefficients for the independent variable of the final logistic regression model



High Debt-to-income ratio increases the likelihood of loan approval.

##### Question 3

In Unit 2 of this module, you generated a final logistic regression model for the HBAT Industries data set. Table 1 shows the HBAT results alongside the Speedy Loans results.

Table 1: Comparison of the logistic regression outputs for the HBAT Industries and Speedy Loans data sets.

|  |  |  |
| --- | --- | --- |
|  | **Result output for HBAT Industries** | **Result output for Speedy Loans** |
| **Null deviance** | 133.75 | 114.61 |
| **Residual deviance** | 37.781 | 94.366 |
| **Chi-square value** | 0.24 | 0.66 |
| **Predictive accuracy** | 91% | 80% |
| **Pseudo R2 value** | 0.7175 | 0.1766 |

The management team has asked you to compare and critique the model results obtained for HBAT Industries and Speedy Loans, and to draw conclusions on whether the Speedy Loans logistic model is a good fit using the parameters in Table 1. They have specifically asked you to address the following:

* The difference between the null and residual deviances was markedly different between the two models, with a difference of 95.969 for HBAT Industries and 20.244 for Speedy Loans. Based on the result output, which model can be considered a better fit? Substantiate your answer.
* In the Unit 1 lesson, you learnt that the chi-square value is a statistical measure of correlation between the observed and predicted values for a model. Based on the two models above, which model shows a stronger association between the predicted values and observed values, indicating better model fit?
* The pseudo R2 value was markedly different between the two models. How can this be interpreted and how does it reflect on the goodness-of-fit for each model?
* Collectively, would you be confident in the logistic model generated for the Speedy Loans data set? In other words, can you confidently conclude that the independent variables are able to accurately predict whether a potential client’s loan application will be denied or approved?

(Max. 600 words)

Start writing here:

HBAT model is a better fit. A -2 log-likelihood value represents the deviance of a model, “a -2*LL*value of 0 indicates that the model is a perfect fit, whereas higher -2*LL* values indicate a poorer overall model fit” (Logistic Regression Notes, UCT 202). The null deviance is the highest deviance of model due to it not having predictor variables other than the intercept. We want the final model to improve in predicting the likelihood of the outcome variable. For this to happen, the model deviance (or -2LL) should be significantly lower than the -2LL of the null model.

A better model fit means more accurate predictions of the output variable, HBAT model has the highest difference in deviance between its null model and thus is a better fit and better model overall.

Low chi-square values imply that the actual predicted values from the model and the observed values are strongly associated with each other and have very little differences between each other. The chi-square value for the HBAT model is closer to 0, while that of Speedy Loans is farthest from 0. Thus, HBAT show stronger association between predicted and observed values and there are less differences between these values while the Speedy Loans display higher differences between the actual predicted values from the model and the observed values.

The pseudo R2 value is a function of the -2LL and is an alternative method of determining the impact of independent variables on the variation of the dependent variable. The pseudo R2 is interpreted similar to the R2 from multiple regression.

The Pseudo R2 coefficient measures the overall significance of the regression model. The HBAT pseudo R2is far greater than that of Speedy loans, this means that there is more predictive value/accuracy in the HBAT model than there is in the Speedy loans model. The closer the R2 value is to 0, the poorer the model fit. Speedy Loan model is not a good fit compared to the HBAT model because its pseudo R2 value is closer to 0, while HBAT R2 is closer to 1. The independent variables for HBAT explain 71.75% of variance in the dependent variable (customer location) collectively, and only 17.66% of the variance in the dependent variable (loan approval status) is explained by the independent variables of the Speedy Loan model. This means that the HBAT model is better at predicting the outcome, and thus the predicted values are very similar to the observed values.

The predictive accuracy of Speedy Loans is 80%, this means that 20% of the loan statuses are misclassified. Misclassification of 20% is a relatively large number for a loan business and could potentially cost the business a lot in losses due to false positives (allocating loans to undeserving clients). There is also a loss in business from false negatives (denying loans to deserving clients). All factors considered for the Speedy Loans logistic model, I don’t have confidence in the model and would recommend that sample size of the of the clients be increased or use another algorithm in predicting the loan allocations.

#### 4. Rubric

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **No submission** | **Poor** | **Good** | **Excellent** |
| **Question 1**  *The image contains the correct graph and variables, and adheres to all the instructions.* | No submission or attempt. (0) | The student attempts to answer the question, but the image contains the incorrect graph type, or the incorrect variables have been used. (2) | The image contains the correct graph using the correct variables, but does not adhere to all instructions. (4) | The image contains the correct graph using the correct variables, and adheres to all instructions. (5) |
| **Question 2**  *The submission correctly compares the null and residual deviances for the simultaneous and final models, and provides an accurate conclusion on model fit. It accurately interprets the logistic coefficients and describes a strategy to determine the relative importance of the variables in the model.* | No submission or attempt. (0) | The student correctly compares the null and residual deviances for the two models, but provides an incorrect interpretation of the logistic coefficients.  OR  The student incorrectly compares the null and residual deviances of the two models, and only provides an acceptable interpretation of the logistic coefficients. (5) | The student correctly compares the null and residual deviances for the two models and provides an acceptable interpretation of the logistic coefficients, but fails to suggest a strategy for determining the relative importance of the independent variables in the model. (10) | The student correctly compares the null and residual deviances for the two models, provides an excellent interpretation of the logistic coefficients, and suggests the correct strategy for determining the relative importance of the independent variables in the model. (15) |
| **Question 3**  *The submission correctly compares the null and residual deviances, chi-square values, and pseudo R2 values for the HBAT Industries and Speedy Loans logistic models.*  *AND*  *The submission provides an overall recommendation regarding the model with the best fit and performance.* | No submission or attempt. (0) | The student does not fully compare one or more of the following: the null and residual deviances, chi-square values, and pseudo R2 values for the HBAT Industries and Speedy Loans logistic models.  AND  The answer provides an inaccurate recommendation regarding the model with the best fit and performance. (7) | The student compares the null and residual deviances, chi-square values, and pseudo R2 values for the HBAT Industries and Speedy Loans logistic models, but one or more of these comparisons is incorrect. The answer provides a recommendation regarding the model with the best fit and performance. (14) | The student correctly compares the null and residual deviances, chi-square values, and pseudo R2 values for the HBAT Industries and Speedy Loans logistic models, and provides an excellent recommendation regarding the model with the best fit and performance. (20) |

**Total:** 40 marks