MTH 650 **Case Study 4**

**Complete the in-class quiz on Wednesday March 1st.**

**Total points obtainable:** 40 points.

OBJECTIVES: By completing this case study, you should be able to demonstrate an understanding of and proficiency with probit/logit models techniques to solve classification problems.

## EXPECTATIONS:

* You may complete this on your own or work together in groups of up to three members. Case Study quiz will be done individually in class.
* You will write a final report summarizing your findings and containing your analysis. Submitting a report that is plagiarized, in part or whole, is considered a violation of the Northwood University Academic Integrity policy. The consequence for plagiarizing will be a 0 on the case study and a report of the incident will be submitted to the Academic Dean.
* Your final report should be clear and understandable with a professional appearance. Use complete sentences as well as correct grammar and spelling. Your report will be submitted on GitHub.
* The data set you are working on is titled **UniversalBank.mtw** in the case study section of your MTH 650-course shell on Blackboard. **Minitab** will be the recommended statistical software of choice.

## NOTE:

You may refer to the video lessons for module 7 to review key concepts covered and that are useful for completing this assessment.

## BACKGROUND INFORMATION

**A Personal Loan Acceptance**

Universal Bank is a relatively young bank growing rapidly in terms of overall customer acquisition. The majority of these customers are liability customers(depositors) with varying sizes of relationship with the bank. The customer base of asset customers (borrowers) is quite small, and the bank is interested in expanding this base rapidly to bring in more loan business. In particular, it wants to explore ways of converting its liability customers to personal customers (while retaining them as depositors).

A campaign that the bank ran last year for liability customers showed a healthy conversion rate of over 9% success. This has encouraged the retail marketing department to devise smarter campaigns with better target marketing.

The goal is to build a logistic regression model to classify whether a new customer will accept a loan offer. This will serve as the basis for the design of a new campaign.

The data set UniversalBank.mtw contains data on 5000 customers. The data include customer demographic information (age, income, etc.), the customer’s relationship with the bank (mortgage, securities account, etc.), and the customer’s response to the last personal loan campaign (Personal Loan). Among these customers, only 480 (= 9.6%) accepted the personal loan that was offered to them in the earlier campaign.

* *Considering the context of the data, we will not use ID and Zip code in building this model. ID is what we call a unique identifier variable. It has no predictive value. Zip Code could also be a unique identifier but to avoid discrimination bias issues, we will not use this variable too. Complete a table of the list of categorical variables in one column and the numerical variables in the other, similar to the one given below:*

|  |  |
| --- | --- |
| **Categorical variables** | **Numeric Variables** |
| Education | Age |
|  |  |

* *Based on the summary of the variables in the table above, how would you describe a typical customer at Universal bank? What are the attributes of a customer in the sample? See how you completed case study 1.*
* *For financial modeling purposes, one of the common goals of building a predictive model is to build an engine known as a credit scoring system. Such systems can be used to deny or approve loans, often within minutes. A logistic regression model can be any such engine under the hood of such credit scoring systems. To gain some intuition about the data,* 
  + *Fit a LINEAR PROBABILITY MODEL (not a logistic model yet) that* *models Personal Loan (the response variable) based on continuous predictors (Income, Family, CCAvg, Mortgage, Age, Experience) and categorical predictors (Education, CD Account). Report your regression model and comment on the adequacy of your model in terms of the p-values of the independent variables, the adjusted R-Sq, and the VIF. Assume a 0.05 level of significance when fitting this model. Note that this just like building a linear regression model like you did in your case study 3.*
  + *In the linear probability model, which variable(s) would like to remove from the model? Please give clear reasons based on the output of the model you have just built. Try to fit another model* ***without*** *the variable(s) you have identified and provide a reason why the removal may have been justified based on the output of the new model. (Hint: Compare the p-values, the R-sq (Adj), and more if you want).*
  + *For the linear probability model, you obtained have just obtained, reference your notes in module 7 and state briefly, two limitations or shortcomings of the linear probability model in using it to model a dichotomous variable like the Personal Loan variable.*
* *Next, we are going to build a probit/logit model. Remember, these models improve on the shortcomings of the linear probability model in modeling dichotomous or binary variables. An example of a probit/logit model is the logistic regression model. Assume a 0.05 level of significance when fitting this model.*
  + *Fit a LOGISTIC REGRESSION MODEL that classifies customers who accept the offer of a Personal Loan (the response variable) based on continuous predictors (Income, Family, CCAvg, Mortgage, Age, Experience) and categorical predictors (Education, CD Account). Report important aspects of your output of the logistic regression model and comment on the adequacy of your model in terms of the p-values in the deviance table, deviance R – sq, the VIF, and the goodness of fit statistics ONLY. Is this model a reasonable fit to the data?*
  + *Read about Occam’s razor* [*here*](https://science.howstuffworks.com/innovation/scientific-experiments/occams-razor.htm)*. In our context, the principle of Occam’s razor applies and motivates us to reduce the number of predictor/independent variables as much as we can, to guarantee a simpler model. So, look at the logistic regression model you currently have, which TWO variables would like to remove from the model? Please give clear reasons based on the output of the model you have just built. Assume a 0.05 level of significance when fitting this model.*
  + *The last step you took is iterative. Try to fit another model* ***without*** *the variables you have identified. Report your output. Then identify if you now have an optimal model. Otherwise, proceed to remove more variables from the model and provide sufficient reasons why the removal may have been justified at EVERY instance of a new model after a variable is removed. Continue this process until you find your optimal model. You will later need to justify why your final model is optimal and be sure to report outputs of intermediate steps that are necessary (For instance, you do not need to report the fits and diagnostics for unusual observations, which is usually the last set of outputs).*
  + *Now that you have your optimal model, give a clear, convincing reason why this is your optimal model. Also, interpret all the vital aspects of your final model. At a minimum, this interpretation should include interpretations of the Deviance table, VIF values, odds ratios for both continuous and categorical variables, and the goodness of fits tests table statistics.*

## INSTRUCTIONS

To answer the director’s questions, follow the steps below:

1. *Explore*

Begin by exploring the data. Create graphs and tables. Calculate summary statistics. Your goal is to understand the data set so that you will be able to describe it. ***Not everything you investigate, calculate, or create in this step will make it into your final report.*** You want to find interesting features and patterns so that you can describe the sample, though in the process you will come across many irrelevant things. The more time you invest in this exploratory step, the more equipped you will be to efficiently complete the next two steps.

1. *Analyze*

Once you have an understanding of the variable and any relationships between the variables, begin to answer the director’s questions. Determine which statistics, displays (charts/graphs), and methods are relevant and appropriate. Be precise and rigorous. Be sure to interpret your conclusions and advice in a way that is specific to the context but understandable to someone who may not be familiar with the underlying statistical methods.

1. *Report*

In your final report, tell a story. As with the stories you enjoyed as a child (and may still enjoy), make sure your story is engaging, is relevant, and includes pictures that illustrate your findings. Be sure your report is professionally formatted and grammatically correct, using complete sentences and paragraphs. The report does not need to be very lengthy, as long as it answers the director’s questions substantively and accurately. Include the names of all group members who contributed to the report. *If a group member’s name is not on the report, he or she will not receive credit for the assignment.*

## AN OUTSTANDING REPORT WILL:

1. **Include only relevant information.** It will be tempting to include every possible statistic you can calculate and every graph you can create. Include only those items that help you tell your story and illustrate a point.
2. **Answer the questions asked.** It is perfectly acceptable to be concise in your answers, as long as your answers are accurate and valid.
3. **Tell a story in a cohesive manner that stands on its own.** This is different from a homework assignment, and as such your report should be a professional document that one can read and understand without any previous knowledge about the data set or the questions asked. Avoid treating it as a homework assignment, where one might write, “#1. The answer is \_\_\_\_\_\_. #2. The answer is \_\_\_\_\_....” Rather, strive to create a comprehensive summary of your findings that includes complete sentences that flow naturally in paragraphs and use correct grammar and spelling. The report should be formatted cleanly in a way that is aesthetically pleasing and can be read and understood quickly.