Problem 3. The MortgageDefaulters.xls contains data on mortgages that have been approved by bank underwriters. The goal is to predict which of future such approved mortgages will default. The purpose of this prediction is not to alter the underwriting standards, but rather to assess the need for and cost of secondary insurance on mortgages. The data is described by the following attributes.

Bo_Age	Borrower age
Ln_Orig	Value of loan, USD
Orig_LTV_Ratio_Pct	Ratio of loan to home purchase price
Credit_score	Borrower's credit score
First_home	First time home buyer? (Y/N)
Tot_mthly_debt_exp	Borrower's total monthly debt expense
Tot_mthly_incm	Borrower's total monthly income
orig_apprd_val_amt	Appraised value of home at origination
pur_prc_amt	Purchase price for house
DTI_ratio	Borrower debt to income ratio (Tot_mthly_debt_exp/Tot_mthly_incm)
Status	Current loan status
OUTCOME	Binary version of "Status" (either default or non-default)
State	US state in which home is located
UPB>Appraisal	Loan amount (Ln_Orig) greater than appraisal (orig_apprd_val_amt) 0-no, 1=yes

Note that some of the above variables were derived from combinations of two others. There are various other combinations that may be useful.

This assignment will examine the performance the some techniques for predicting default. We will use decision trees, neural networks and logistic regression.

Question

(1) a. We want to use some census data variables also. Add state income and state poverty rate variables to the dataset - these are there in a separate sheet of the Excel data-file.

b. Explore the data. Report on the distribution of values in the attributes and how they individually relate to the outcome of interest (dependent variable).

c. Create some new derived attributes and explain why you think these may be useful.

d. Determine if any transformations are necessary, and which variable you include in the modeling. Explain your reasoning. You should use the full data for this

(2) The data has 15152 cases, with 2% defaulters (401 cases of default). We want to consider a more balanced sample for modeling – one of the tasks in this assignment is to determine if different levels of balancing affect performance.

First split the data into Training (75%) and Testing (25%) sets.

We will undersample the majority class (non-default) to obtain two training datasets – the first with 30% default cases (call it TrgA) and the second with 10% default cases (call this TrgB).

You should experiments with both these training datasets to determine which models (if any) gives better performance on the test data. Does your finding hold across the all techniques?

Develop and report on your "best" models from each of the three techniques. Mention which parameters and values you experiment with for each technique and how you obtain your "best" models. For performance, consider the confusion matrix, overall accuracy, accuracy on two classes, lift (report the lift values and also show the lift chart). Explain how you use these measures to determine which model is best suited for this problem?

Do you have a preference for any specific model to use for this problem? Explain why (or why not).

Note: Your report should clearly show performance by technique, and for the three different training datasets as well as for the test data. Performance should be on different measures indicated above, and you should indicate how you would use these measures for the specific business problem here. Results

should be presented in a manner that is easily readable and where performance comparisons clearly come through. (Of course, you need adequate writeup around the tables/graphs too).