Strathmore UNIVERSITY

STRATHMORE INSTITUTE OF MATHEMATICAL SCIENCES

BACHELOR OF BUSINESS SCIENCE FINANCIAL ENGINEERING

END OF SEMESTER EXAMINATION

BSF 3216: APPLIED ANALYTICS IN FINANCE

DATE: 18th December 2023 TIME: 3 HOURS

INSTRUCTIONS

I. This examination consists of FIVE questions. Answer **Question 1** (**COMPULSORY**), choose **2 optional** questions out of **Question 2 to 5.**

II. You can work on either R or Python for questions 1-3; and R for questions 4-5. The solutions will be submitted as **a Notepad script, an R markdown or PDF file** which should also be saved under your admission number.

III. Ensure that you develop a script that **includes your code and comments**. Remember that comments are preceded by a # on R script.

IV. Ensure you have clearly **indicated the question number** on EVERY QUESTION.

Question 1 (30 marks)

A client has approached you with data on different variables that they believe affect the price of a certain house in a given area.

Use the dataset named **DATA.csv** to answer the following questions.

i. Explore the dataset and correlation between 6 variables you deem useful using a detailed SPLOM (relevant plots). Comment on your results and extend your analysis to provide insight to the client on the variables you deem important for them to look at. [8 marks]

ii. Split your data into 70% training set and 30% test set. [3 marks]

iii. Train a linear regression model on your data with price as your dependent variable, clearly stating any two assumptions of the model. Comment on the coefficients of your model and their implications, together with the goodness of fit of the model. [6 marks]

iv. Comment on the performance of your model supporting your answer using a model evaluation metric of your choice. [4 marks]

v. A consultant approaches you and lets you know that the effect of the square metres and number of rooms are not cumulative but rather they have an effect only once a specific threshold has been reached. They advise you to replace the variables in light of this new information where the threshold is the mean price. Additionally, they provide research that demonstrates that there are interaction effects between the city part range and number of previous owners and also between the number of guest rooms and if it has a pool. By conducting the appropriate transformations to your data, incorporate this new information and run a model to reflect these changes. Comment on the coefficients of your model and their implications. Additionally, comment on the performance of your model and whether the consultant's advice was valuable. [9 marks]

Question Two (20 marks)

The following data represents individuals who take credit with a bank and their credit histories. Each person is classified as good (default == 1) or bad (default == 2) credit risk according to the set of attributes. The bank wants to better understand the risk profile of its clients and it requests you to run models to determine a way to classify the risky customers.

Use the dataset named **DATASET.csv** to answer the following questions.

i. Conduct the appropriate transformations to make your data ready for meaningful analysis, including a 70:30 train-test split. [4 marks]

ii. Train a k-nearest-neighbour model on your data and use mini-max normalization. State any assumptions made when training the model. [4 marks]

iii. Evaluate how well your model performed and give the accuracy score of your model. Explain why accuracy is not the best metric for evaluation of performance in light of the nature of the client's request. [4 marks]

iv. The top priority of the client is to identify customers who are a bad risk to the bank. Even if we predict good customers as bad, it won't harm their business. But predicting bad customers as good will do. Using your result in (ii) above and the caret package, create a confusion matrix of the predictions versus the actual values from the test dataset. Comment on your results taking keen notice to discuss the sensitivity, specificity, accuracy and any other statistic you find important to address the concerns of the client. [4 marks]

v. Test at least 5 alternative values of k and draw a table based on your choice of k. Comment on the appropriateness of the choices of k in light of the bias-variance trade-off. [4 marks]

Question Three (20 marks)

Use the dataset named **DATASET.csv** to answer the following questions. Conduct the appropriate transformations to make your data ready for meaningful analysis:

i. Train a logistic model on the data and ensure that your model results will be the same regardless of how many times your model is run. [2 marks]
ii. Run the predict () function and comment on your results paying specific attention to the confusion matrix. [4 marks]
iii. Using the caret library, evaluate how well your model performed and give the Precision, specificity and sensitivity score of your model from the confusion matrix and explain their implications. [4 marks]
iv. How can you improve the model in (i) above? Implement your suggestion and comment on your results. [10 marks]

Questions Four and Five

Datasets: Financial_crisis1.csv; Financial_crisis2.csv; Boom_2.csv

You are provided with loan data from a peer-to-peer lending company spanning the period beginning June 2007 and ending December 2011. Of this period, the world experienced a financial crisis in 2007-2009; and a better `boom-type' economic situation in 2010 and 2011. The datasets above are extracted from this lender's full loan data and represent specific periods. You are required to use these datasets to answer Questions 4 and 5 below.

Question Four (20 marks)

Use the two datasets **Financial_crisis1.csv** and **Financial_crisis2.csv**, with the training set being the **Financial_crisis1.csv** data, and the test set being the **Financial_crisis2.csv** data to answer the following questions;

i. What does the dataset's start date inform us about the possible motivations behind this company's establishment? [2 marks]
ii. Would you say this peer-to- peer lender was affected by financial crisis i.e., were there higher borrowing rates compared to later years? Or higher default rates etc? You can justify

your explanation with trend charts etc. (or compare with the 2011 dataset, Boom2.csv) [3 marks]

iii. Were there any differences in the types of borrowers during the financial crisis compared to the post-crisis years? To what degree would you say this type of borrower influenced default statistics? [2 marks]

iv. Run a decision tree, trained on earlier years (June 2007-September2009 i.e., Financial_crisis1.csv), and tested for the later years (October-December2009 i.e., Financial_crisis2.csv). For the model, use the variables loan amount (*loan_amnt*), funded amount from investors (*funded_amnt_inv*), standardized interest rate (*int_rate_st*), *installment*, *grade*, annual income (*annual_inc*) and *loan status*. Analyse your results and comment on the decision tree's predictive ability for this type of data. Why do you think we generate the kind of predictive result that we do? Is there any alternative technique that you believe could improve upon this ability? [10 marks]

v. This specific peer-to-peer lender recently upgraded to a fully-fledged financial services company. Could this early years' data tell us anything about their projected success? Are there any factors that you can identify from the data that could show us that this was, and would be a significantly profitable business for them? (No need to run predictive codes, summary statistics and plots are enough) [3 marks]

Question 5 (20 marks)

The peer-to-peer lender's market experienced a boom in 2010 onwards. Use the dataset **Boom_2.csv**, representing the lender's 2011 data, to answer the questions that follow;

i. Are there any specific features of the data that could prove that the market was entering a boom phase from 2010 onwards? (Compare with the 2007-2009 data or link features to general trends at the time) [5 marks]

ii. Run a simple neural network model using the neuralnet and nnet packages in R and check for this model's predictive ability of the approval rate. Train using the 70:30 rule. For the model, use the variables loan amount (*loan_amnt*), funded amount from investors (*funded_amnt_inv*), standardized interest rate (*int_rate_st*), *installment*, *grade*, annual income (*annual_inc*) and *loan status*. What would you say about the model's accuracy? Can we improve on this without using other types of models? What challenges would we

face in trying to improve our model's predictive power? (Note: This question may take a few minutes to run, so maybe handle this first) [10 marks]

iii. Neural networks were developed to mimic our own brain activities that underly human decision-making processes. In this regard, describe a typical day-to-day situation in which a neural network model could be derived (Your own). If we were to create an artificial neural network out of your situation, what artificial neural network model would you say would be best suited to model it, and why? [5 marks]

TOTAL MARKS [70]

END OF EXAM