STRATHMORE INSTITUTE OF MATHEMATICAL SCIENCES
BBS Financial Engineering
END OF SEMESTER EXAMINATION
BSF 3216: APPLIED ANALYTICS FOR FINANCE

DATE: 30th November, 2022 TIME: 3 HOURS

INSTRUCTIONS

I. This examination consists of FIVE questions
II. Answer **Question 1** (**COMPULSORY**), choose **2 optional** questions out of **Question 2 to 5**
III. The solutions will be **submitted as a Notepad script or an R markdown**. You can work on either R or Python. Ensure that you develop a script that includes your code and comments. It is this script that you will submit and ensure you have clearly indicated the question number on **EVERY QUESTION**. Remember that comments are preceded by a # on R script. This should also be saved under your admission number.

SCENARIO

The following data represents people who takes a credit by a bank. Each person is classified as good or bad credit risks 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

Question 1 (30 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. Normalize your data using the Z-score standardization. [2 marks]
ii. Draw a box plot and histogram of 4 key variables of your choice. Comment on your results. [4 marks]
iii. Using a **stratified random sampling** approach, split your data into 70 training set and 30% test set. [2 marks]
iv. Train a c5.0 model on your data. Comment on your results in light of the needs of the client. Additionally, evaluate the model performance. [5 marks]

v. How can you improve the model in (iii) above? Implement your suggestion and comment on your results [4 marks]

vi. 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 (iv) 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 kappa statistics results, sensitivity, specificity, accuracy and any other statistic you find important to address the concerns of the client. [7 marks]

vii. You believe you can better improve the performance of your model by using a repeated holdout technique when training your model. You decide to apply a 0.632 bootstrap to train your model and estimate the recall statistics. Implement this and comment on your results. How does it compare to your earlier findings? [6 marks]

Question 2 (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. Using the **caret** library, train a c5.0 model on the data and ensure that your model results will be the same regardless of how many times your model is run. Run the predict () function and comment on your results paying specific attention to the confusion matrix. (**NB**: This model **WILL** take some time to train so take that into account when tackling the question) [6 marks]

ii. Evaluate how well your model performed and give the Precision, specificity and sensitivity score of your model from the confusion matrix and explain the implications to your client's needs. [3 marks]

iii. There are various resampling methods that can improve the performance of your model from (i). Using the appropriate control object, apply 3 such methods to improve your model and comment on the output of each of your method clearly stating any of the assumptions made. [11 marks]

Question 3 (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 kNN model on your data and use the mini-max normalization. (Clearly state any assumptions made when training your model) [5 marks]

ii. 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. [5 marks]

iii. 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 [4 marks]

iv. You believe a neural network will give a better performance for your client compared to the kNN. Run a neural network with 5 hidden layers and comment on the results from your resultant confusion matrix and in addition, calculate the precision and recall from it. Explain why they are better indicators of model performance in line with your client needs compared to the accuracy measure. (**NB**: This model may take some time to train so take that into account when tackling the question) [6 marks]

SCENARIO

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

Question 4 (20 marks)

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

i. Explore the correlation between 6 variables you deem useful using a detailed SPLOM. Comment on your results paying keen attention to the loess smooth and extend your analysis to provide insight to the client on the variables you deem important for him to look at. [4 marks]

ii. Train a linear regression model on your data with price as your dependent variable Comment on the coefficients of your model and their implications. Additionally, comment on the performance of your model [7 marks]

iii. 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 and advices you to replace the variables in light of this new information. This threshold is the mean price. Additionally, he provides 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 5 (20 marks)

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

i. Using a **stratified random sampling** approach, split your data into 70% training and 30% testing. Train a regression tree on your data with prices as your dependent variable and comment on your results. [6 marks]

ii. Provide the variable importance of the top 5 variables based on your model. (These should be numeric value showing the importance of each of these top 5 variables] [3 marks]

iii. Visualize your regression tree making sure that the leaf nodes are NOT aligned at the bottom of the plot. [2 marks]

iv. Make predictions using your model and compare your results with the summary statistics of the actual data. Comment on your results. [2 marks]

v. Using a model tree of your choice, make predictions and compare your results with the summary statistics of the actual data. Does the model tree perform better than the regression tree? [3 marks]

vi. Using the bagging method and your training dataset, train a bagged decision tree with 10-fold CV via the train() function in the caret package. Comment on your results. [4 marks]

END OF EXAM

TOTAL MARKS [70]