STA4026S 2025 – Statistics Honours Analytics
Assignment 1 – Applied Supervised Learning

Due Date: Monday 24 March 2025, 12:00 (noon)

Please read the following instructions carefully:

- You may complete this assignment either in pairs or individually.
- Answer all questions using `R`. You may use any packages and functions you like.
- You must submit the following on the course site:
  - A pdf report. **The page limit is 15 pages**. Anything beyond the 15th page will be ignored. The report should NOT contain any code.
  - All relevant `R` files (.R, .Rmd, .qmd, .RData). Be sure to use comments in your code and to set seeds to ensure reproducibility. Please **DO NOT ZIP** any files; upload them separately.
  - The csv file with the predictions for Question 4 (see that question's instructions).
- Use the following naming convention for the files:
  `STDNUM001_STA4026S_A1.pdf` for individuals
  `STDNUM001_STDNUM002_STA4026S_A1.pdf` for pairs.
- Your report may be compiled using any software you like, although you are strongly encouraged to either knit from RMarkdown or Quarto, or use LaTeX. Some marks will be awarded for communication and presentation, so make sure everything in your report is legible and neat.
- You are welcome to include figures and tables not specifically asked for, but only do so if they are relevant to your discussion. Also be sure to interpret them sufficiently; marks will be deducted for any output placed in the report without discussion.
- Do NOT paste/print raw `R` output verbatim – this will be penalised. If you want to include `R` output, typeset it properly or present it in a table.
- To help the reader easily assimilate the information, round values to the fewest number of decimal places necessary (unless there is a good reason for expressing a more exact value).

A classification problem – Predicting online shoppers' intention

The goal of this exercise is to predict whether or not online shoppers will finalise a transaction, based on data gathered about their browsing session. Therefore, the target variable – `Revenue` – is binary. The data were first presented and analysed by Sakar et al. (2018); descriptions of all 17 features can be found in this paper, although some variables have been removed and others adjusted in order to simplify the analysis a bit. Several publications have subsequently analysed these data, including Mostafa et al. (2024) and Swetha et al. (2024). The aim is to recreate some of their results. Note, however, that you will not be graded on the accuracy of the results, but rather by the procedure followed and the clarity in communication and presentation thereof.

The dataset has been split into a training set and validation set – `online_shopping_train.csv` and `online_shopping_valid.csv`, respectively. These files must be read into `R` as is, i.e. the spreadsheets may not be edited at all.

Provide a brief introduction to the report; a reader should be able to comprehend the task at hand without having seen this set of instructions. Although an exploration/description of the features is not required, a brief section on this may be included. If you do, your discussion should focus only on aspects that pertain to this exercise.

Question 1 – Modelling [50 marks]

First convert categorical variables into factor type, stating for which features this was done.

a) Using the training dataset, fit the following models to predict the transaction outcome. Apply 10-fold cross-validation throughout to measure classification accuracy, except for the random forest, where OOB accuracy should be used.

   - Logistic regression with a linear decision boundary. Apply elastic-net regularisation to this model, motivating for the choice of α and λ.
   - Logistic regression with a non-linear decision boundary. You may specify the model and regularise (or not) in any way you wish, just be sure to clearly state your final model.
   - K-Nearest Neighbours (KNN), motivating for the choice of k. The choice of features to include is up to you, and may be informed by any subsequent analysis (in which case, refer ahead to that question).
   - A classification tree. Apply appropriate pruning and decide (with motivation) on a tree size.
   - A random forest, motivating clearly for your selected hyperparameters.
   - Any boosted tree model (gbm, xgBoost, LightGBM,[1] etc.), again detailing your hyperparameter selection.

[1] The documentation can be found here

b) Evaluate each of the six models on the validation set using a decision threshold of $\tau = 0.5$ and report the following metrics for each in a single table: Accuracy, F1 Score, Precision, Recall, Specificity, and ROC AUC. Briefly compare and discuss. For bonus marks, add a visual illustration of the results.

Question 2 – Inference/Interpretation [20 marks]

a) Interpret the coefficients of the variables retained in the linear logistic regression model.

b) Display and interpret the final classification tree.

c) For the random forest and boosting model, provide and interpret variable importance plots as well as partial dependence plots.

Question 3 – Optimising F1 Score [20 marks]

Suppose the goal is to fit a model that will yield the highest F1 score on out-of-sample data. Instead of finding new hyperparameters by using the F1 score as objective function, we will use the above models and adjust the decision rule threshold (τ).

a) For each of the six models, plot the validation F1 score as a function of τ, indicating the maximum value with a vertical line. Arrange the plots in either a 2×3 or 3×2 figure.

b) Plot the ROC curve for each model (again for the validation set), arranged similarly. On each curve, indicate the point corresponding to the maximum F1 score, and add the F1 score and τ as text next to the point.

c) Present the maximum F1 scores and their corresponding thresholds in a table.

Question 4 – Prediction [10 marks]

a) Based on the above results, select (with motivation) a model and decision rule threshold to be used on unseen data if the goal is to maximise the F1 score.

b) Finally, use the selected model to predict the `Revenue` outcome for the observations contained in the file `online_shopping_testing.csv`. Write your predictions to a csv file as follows:

- The file must only contain **ONE** column, namely the predicted labels, i.e. only 1's and 0's.
- Do **NOT** include a column header.
- There should be **NO** blank cells to the left or above this column.
- The order of the predictions must correspond to the order of the observations in `online_shopping_testing.csv`.

- The file name must be your student ID(s), all uppercase. E.g. STDNUM001.csv or STDNUM001_STDNUM002.csv.

You will only lose marks if you fail to adhere to these instructions (check your file before submitting!), or if your predictions are nonsensical or severely wrong. The list of F1 scores will be posted, although this is purely for fun/bragging rights and will have no bearing on your assessment marks.

$\mathcal{END}$

TOTAL MARKS = 100

References

Mostafa, Salama A. et al. (2024). "A Classification Technique for Online Shoppers' Purchasing Intention". In: *2024 1st International Conference on Logistics (ICL)*, pp. 1–6. DOI: 10.1109/ICL62932.2024.10788661.

Sakar, Cemal Okan et al. (2018). "Real-time prediction of online shoppers' purchasing intention using multilayer perceptron and LSTM recurrent neural networks". In: *Neural Computing and Applications* 31, pp. 6893 –6908. DOI: https://doi.org/10.1007/s00521-018-3523-0.

Swetha, Thammisetty et al. (2024). "Forecasting Online Shoppers Purchase Intentions with Cat Boost Classifier". In: *2024 International Conference on Distributed Computing and Optimization Techniques (ICDCOT)*, pp. 1–6. DOI: 10.1109/ICDCOT61034.2024.10515309.