Course 201 Mini-project: Applying supervised learning to predict student dropout

Skills	Criteria	Criterion has been met
Stage 1	The data has been loaded correctly	☐
	Follow the pre-processing instructions given for Stage 1 (please check the project starter notebook for more details) - Note: Use ONLY the Stage 1 dataset	☐
	Data feature engineering: convert date of birth to Age, the target variable has been converted from string to binary encoding.	☐
	Check for the target variable barplot - Is the data imbalanced?	☐
	Data has been correctly split into training and test set (use 80-20 split), use a 10% split of the training dataset as the validation dataset.	☐
	The XGBOOST model has been instantiated correctly and fitted on the training dataset	☐
	The set of performance indicators requested have been correctly printed for the test set. - Accuracy, confusion matrix, precision, recall and AUC	☐
	The hyperparameter tuning of the learning rate, max depth, and number of estimators has been performed correctly.	☐
	The metrics with and without hyperparameter tuning have been printed correctly and compared. Metrics include: Accuracy, confusion matrix, precision, recall and AUC on the test set	☐
	The plot of features importance has been correctly printed and commented on.	☐
	The Neural network model has been instantiated correctly and fitted on the training dataset	☐
	Loss curves for each epoch for the training and validation set have been plotted.	☐

	The set of performance indicators requested have been correctly printed for the test set. - Accuracy, confusion matrix, precision, recall and AUC	☐
	The hyperparameter tuning of number of neurons, optimisers and activations functions have been set correctly.	☐
	The metrics with and without hyperparameter tuning have been printed correctly and compared. Metrics include: Accuracy, confusion matrix, precision, recall and AUC on the test set	☐
Stage 2	The data has been loaded correctly and split into training and test data (80-20 split), use a 10% split of the training dataset as the validation dataset	☐
	Follow the pre-processing instructions given for Stage 2 (please check the project starter notebook for more details) - Note: Use ONLY the Stage 2 dataset	☐
	The XGBOOST model has been instantiated correctly and fitted on the NEW training dataset	☐
	The Neural network model has been instantiated correctly and fitted on the NEW training dataset	☐
	Compare the results of the XGBOOST and Neural network models on the Stage 2 dataset Vs the Stage 1 dataset	☐
	Explain why the results are different	☐
	Perform hyperparameter tuning of the XGBOOST and Neural network models on the Stage 2 dataset. Does it signifcantly improve the performance of XGBOOST and Neural network models. Comment.	☐
Stage 3	The data has been loaded correctly and split into training and test data (80-20 split), use a 10% split of the training dataset as the validation dataset	☐
	Follow the pre-processing instructions given for Stage 3 (please check the project starter notebook for more details) - Note: Use ONLY the Stage 3 dataset	☐
	The XGBOOST model has been instantiated correctly and fitted on the NEW training dataset	☐
	The Neural network model has been instantiated correctly and fitted on the NEW training dataset	☐
	Compare the results of the XGBOOST and Neural network models on the Stage 3 dataset vs the Stage 2 dataset	☐

	Explain why the results are different	☐
	Perform hyperparameter tuning of the XGBOOST and Neural network models on the Stage 3 dataset. Does it signifcantly improve the performance of XGBOOST and Neural network models. Comment on this.	☐
Report: Communicating business impact and insights	The report is between 800-1000 words.	☐
	The report documents the approach used.	☐
	Major inferences regarding the data analysis are evident.	☐
	Conclusions drawn from the analysis are supported by the data.	☐
	Visualisations are created to effectively convey the story of the data.	☐
	The effectiveness of different methods is assessed based on gathered evidence.	☐
	The code is well organised and presented.	☐
	Comparison of the different models is well documented and commented on.	☐
	Comparative analysis of the differences in model performance based on Stage 1 Vs Stage 2 Vs Stage 3 datasets is presented.	☐
	The report assesses the effectiveness and accuracy of identifying early-stage, mid-course, and late-stage risks, based on the comparison of predictive performance across these stages.	☐
	The report is clear, well-organised, and engaging to facilitate learning from the analysis.	☐