ITNPBD6 Assignment 1. 3457775

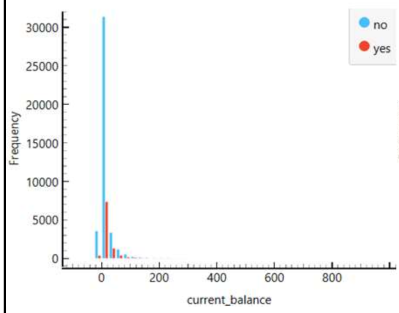
For this assignment, I worked with a telecom company that had limited funds to market its new mobile contract. Using machine learning algorithms and past customer data, I predicted whether a new customer was likely to purchase the new contract. This helped improve targeting and optimize marketing efforts. I used Orange, a data mining tool, to clean and balance the data, and then set the variable 'new contract this campaign' to the targeted one to make a classified prediction. After tuning hyperparameters, I selected the best-performing model to gather insights into customer purchase behavior, ultimately improving decision-making and maximizing revenue through an effective marketing strategy.

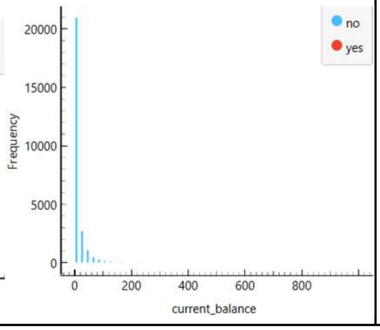
2. Project Methodology

- ✓ I cleaned/balanced the given telecom data first to exclude the unnecessary rows and columns and aligned the attributes using the concatenate widget.
- ✓ Then I split the given data into validation, tuning, and testing sets to prevent overfitting and ensure fair insights into the data.
- ✓ Developed three assigned machine learning models, i.e., logistic regression, neural network, and decision tree, to predict the target.
- ✓ Then I tuned the hyperparameters for these models and evaluated the TP and FN with the help of a confusion matrix.
- ✓ Lastly, I selected the model with the highest classification accuracy, which was Neural Network, in my case showing an accuracy percentage of 77.5%.

4. Data Preparation

I cleaned the current balance variable by removing negative values in the graph.





3. Variables

Keeping the variables categorized helps models handle complex relationships between the data effectively. In this job, and current balance helps assess financial stability. While last_contact and last-contact-this-campaign-month help to find out the customers' engagement in the previous plans and the probability of them likely to buy a contract.

Variable	Type
Job	Categorical (nominal)
Arrears	Categorical (nominal)
Current Balance	Numerical (continues)
Has-tv-package	Categorical (nominal)
Last-contact	Categorical (nominal)
Last contact-this-campaign-day	Numerical (continues)
Last-contact-this-campaign-month	Categorical (ordinal)
This-campaign	Numerical (discrete)
Contact-during-previous-campaign	Categorical (nominal)
Outcome-previous-campaign	Categorical (nominal)
New-contract-this-campaign	Categorical (nominal)

5. Model Training and Hyperparameters

The table below shows the hyperparameters I used for tuning my models.

Model	Hyper Parameters	Classification Accuracy
Logistic Regression	Regularization type: Ridge (L2) Strenth C = 2	75.8%
Neural Network	Neurons in hidden layer: 100 Activation: ReLu Solver: Adam Maximum iterations: 500	77.5%
Decision Tree	Minimum instances in leaves: 5 Don't split subset smaller than: 5 Maximum tree depth: 50 Stop when majority is: 95%	76.9%

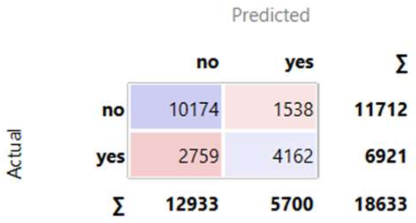
For logistic regression, I used Ridge (L2) for better regularization and to avoid overfitting.

For the neural network, I chose 100 neurons and also selected ReLu because of its good performance with large datasets.

While I set minimum leaves to 5 and maximum depth to 50 to reduce overfitting and keep the model generalized.

6. Final Model and Results

I selected the final model of Neural Network because of its higher accuracy and trained that model with the above-defined parameters. Here I got this confusion matrix.



Predicted

Actual		no	yes	Σ
	no	10174	1538	11712
	yes	2759	4162	6921
	Σ	12933	5700	18633

The confusion matrix tells us the real performance of a classification model by comparing the actual with the predicted values. It helps assess the ability of a model to differentiate between four key categories: True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). For confusion matrix is better in detecting no, and it will help the telecom company to know whether a customer is unlikely to purchase the contract or not. Reducing FP will save costs and help the company market to the right people.

7. Insight about data or models gained

In my case, the variables like outcome-previous-campaign and contacted-during-previous-campaign are of high gain, and so are the most important features in predicting the new outcome. In the same way model tells us that the most likely buyers of the new contract would be those who are past buyers and have a positive current balance. Also, those who are married are more likely to buy it than those of single. The previous TV package will help the company to predict whether customers will be willing to buy our mobile contract in the future or not. So, this model will help to anticipate the marketing insights and reduce the costs.

8. References

[1] ScienceDirect (n.d.) *Confusion Matrix*. Available at: https://www.sciencedirect.com/topics/engineering/confusion-matrix

[2] Amazon Web Services (n.d.) *What is hyperparameter tuning?* Available at: https://aws.amazon.com/what-is/hyperparameter-tuning/

[3] I used ChatGPT for some insights about my work and variable selection.