# Title

"Predictive Data Analytics for Banking Marketing Success using Term Deposit Subscription Classification"

# Topic Area

Predictive Data Analysis, Banking, Telemarketing, Finance

# Research Objectives

* To develop and compare predictive models (Random Forest, XGBoost, CatBoost, CNN, and TabNet) for classifying term deposit subscriptions using direct marketing campaign data.
* To evaluate the performance of these models based on accuracy, precision, recall, F1-score, and computational efficiency.
* To analyse the most important features on model prediction.
* To determine which model offers the best balance between interpretability and predictive power.

# Statement of Hypotheses

* **H1**: Deep learning-based models (CNN, TabNet) will outperform traditional ensemble methods (Random Forest, XGBoost, CatBoost) in predicting term deposit subscriptions, particularly in handling complex feature interactions.
* **H2**: Ensemble models (XGBoost, CatBoost) will provide higher interpretability and generalisability compared to deep learning models, making them more suitable for business use cases where transparency is important.

# Literature Review

## Predictive Analytics in Banking

Broby (2022) takes a closer look at the use of these techniques in the financial industry where classification, regression, clustering and time series are discussed. The study incorporates these techniques into Decision Support System (DSS), which assists the financial managers in their decisions making. More than 187 papers are inspected according to the proposed SPAR-4-SLR protocol; and despite the fact that the range includes many algorithms, it is dominated by Random Forests, Support Vector Machines (SVM), and Neural Networks. For instance, Huang et al. (2005) employed SVM to anticipate the overall direction of stock at 75%accuracy; Broby (2022). These methods are used in the determination of the financial future position and in the framework of risk management plans.

In the present study, Moinuddin et al. (2024) analyse the effect of the marketing analytics on the consumers and the campaign effectiveness using both discussed interviews and surveys. This research also finds out that checking of analytics tools is done weekly by 75% of the respondents, the mean score the respondents gave to the increase in the success of their campaigns as a result of this activity was 8.5 out of 10. Correlation analysis gave a value of 0.75 on the Pearson coefficient and this suggested that analytics usage had directly dependent on the campaign evaluation (Moinuddin et al., 2024/). This is particularly important in current society, yet to show why there is a need to shift towards data driven marketing.

Zaki et al, (2024) analyse predictive analytics and machine learning in term deposit subscriptions of the bank using models like SGD, KNN, and Random Forest. Feature engineering, cross tabulation, and heatmaps were used during data exploration conducted in the study. Random Forest model performed better than other models where accuracy was found to be 87 %, PPV = 87.83% and NPV = 92.99%. These findings offer relevant tips that can be used to improve advertising tactics in banking (Zaki et al., 2024).

Amajuoyi et al. (2024) study the impact of analytics models for customer retention and business growth. Based on the obtained data of the contacts’ database and machine learning algorithms, they define high and low potential clients and churn threats. According to the study it indicates that the businesses that adopt the actions of predictive analytics experienced a customer retention level of 25% and the efficiency of an expanding market of 30%. The study maps to how machine learning advances targeted customer marketing and interaction (Amajuoyi et al., 2024).

In finance, Olaniyi et al. (2023) consider how predictive analytics shifts data into a form that gives insight. The authors discuss and evaluate various decision approaches including decision tree, random forest, and neural network to show that the predictive models yield the result of 40 percent increase in resource efficiency for financial institutions. The research also focuses on the PWE’s key significance for risk management and fraud prevention; effectiveness enhancement up to 30%, in the essential financial risk situations (Olaniyi et al., 2023).

Most recent work on decision trees for predictive analytics of business decisions include Lee et al. 2022. The study used secondary research where more than twenty four papers have been reviewed from fields as; healthcare, finance and customer relationship. This was especially interesting because the decision tree showed marked improvement of 30% in customer behaviour predictability in relation to their customer retention. Concisely, the authors shall find that although relatively basic and very easy to comprehend, the decision trees are still the valuable instruments for the predictive business analytics (Lee et al., 2022).

Authors Wassouf et al. (2020) implement a big data framework for predicting customer loyalty for the Syriatel Telecom. They used two main components of the TFM model in combination with usage and demographic data of customers to categorise customer loyalty. Among the four machine learning models tested: Random Forest; Decision Tree; Gradient-Boosted Trees; and Multilayer Perceptron. The highest accuracy achieved in binary classification based on the results of the Gradient-Boosted Trees is 87%. This paper also reveals how big data analytics can promote customer loyalty more than it has already been estimated (Wassouf et al., 2020).

Zulaikha et al. (2020) look into how the integration of artificial intelligence in marketing can shape the way customer predictive analytics. By applying various algorithms, the study’s customer data on behaviour and demographic characteristics are segmented and targeted. The authors also note that companies adapting AI within the scheme for customer segmentation saved 25% of the targeting effect and reduced their marketing expenses by 30%. This paper therefore emphasises on the feasibility of the AI based segmentation in improving marketing effectiveness and consumer satisfaction (Zulaikha et al., 2020).

## Ensemble Models

Using phone contacts including 45,211 records, Patwary et al. (2021) explore prediction of bank deposit through ensemble learning. Naive Bayes, Support Vector Machine (SVM) and Neural Networks (NN) methods are used in the study, with Bagging, Boosting, and Stacking ensembles. It denotes that the chosen model of Neural Networks (Bagging) displayed higher accuracy of about 96.62% with sensitivity of 97.14% and high specificity of 99.08%. This study points to the effectiveness of ensemble models over the traditional classifiers in predicting the amount which customers would be willing to deposit in banks (Patwary et al., 2021).

Wardani et al. (2023) assess the performance of bias reduction of bank customer data using Machine Learning techniques; XGBoost, LightGBM and Random Forest. This study attempts to use AI Fairness 360 to detect and correct bias using disparate impact (DI) and statistical parity difference (SPD). The experiment follows a mitigation approach and saw a great reduction of bias through adversarial debiasing with a mean DI of 0.943 and SPD of -0.004. The study shows how adversarial debiasing is better than other approaches in arriving at a fairer prediction (Wardani et al., 2023).

Shin et al. (2021) have exhibited and compared Random Forest, GBM, XGBoost, LightGBM decision tree-based ensemble learning technique for predicting term deposit subscriptions. Therefore, in this Kaggle bank marketing dataset study, using 10-fold cross validation, the proposed LightGBM has a higher accurate rate 86.4%, the Kappa coefficient 0.727, F1-score 0.861. In this study, SHAP was used to add interpretability and showed that model decisions were most affected by customer contact duration (Shin et al., 2021).

In the context of a digital marketing dataset, Abd et al. (2024) use Random Forest, Decision Tree, KNN and CatBoost to categorise customer behaviour. CatBoost provided the highest accuracy of %98 and F-measure score %0.91 compared to the Random Forests %97, Decision Trees %95 and KNN %91. While CatBoost took 15.04 seconds to execute, it had an optimal generative predictive performance, recommending CatBoost for use when performing marketing prediction analysis (Abd et al., 2024).

Kinnander (2020) compares the results of the XGBoost, LightGBM, and CatBoost models of Gradient Boosting Tree capabilities in terms of the accuracy of customer profitability prediction in online retail. This work then tests and compares these models based on recall and precision rates, employing the first purchase data of new customers of Klarna. XGBoost and LightGBM were almost equal, with precision of up to 80%, while CatBoost was slightly better in overall performance in classifying profitable customers with a recall of 0.85 (Kinnander, 2020).

In their study, Wankhede et al. (2019) work on enhancing the accuracy of identifying potential clients on the Bank Term Deposits using an explore Machine Learning algorithms that include, but not limited to Logistic Regression, Support Vector Machine (SVM), Random Forest, Neural Networks, and eXtreme Gradient Boosting (XGBoost). Of these Random Forest is the most accurate model with accuracy of 87% after addressing the machine learning classification problem of class imbalance using ‘Spread-subsample’ technique. In general, the study stresses the necessity to derive features for enhancing the predictive success rate of telemarketing (Wankhede et al., 2019).

Nupponen (2024) explores the accuracy of the machine learning models for sales performance of manufacturing service companies. The models used in this study include Logistic Regression, Decision Tree, XGBoost with hyperparameter tuning of XGBoost model. The outcomes show that XGBoost has a better fit than other models; its test classification accuracy reached 84.2% and the AUC was equal to 0.88. Extending the SHAP analysis to hierarchical levels, direct sales channels and service agreements appeared to be the driving force behind successful sales (Nupponen, 2024).

Meshref (2020) uses the Kaggle Bank Marketing dataset to predict loan approvals using ensemble machine learning algorithms including; Bagging, Boosting, LogitBoost, and Random Forests. The models yielded an accuracy of 83.98% with LogitBoost in recognition of the best performance in the models. The study also focused on model interpretability since feature selection and interpreting decision trees were employed; this information was important for banks’ decision making. From this research, Meshref highlighted how increasing the algorithms’ efficiency improves predictive accuracy of the bank marketing dataset (2020).

Chlebus and Osika (2020) analyse five tree-based algorithms, including Random Forest, AdaBoost, GBM, XGBoost, and CatBoost, on the Portuguese bank telemarketing dataset. This research measures model stability and employs XAI tools for a qualitative analysis of the interpretable models. Title: The Performance of XGBoost, CatBoost and LightGBM on the GBDx Breast Cancer Data Set Using Feature Importance With regard to AUC, XGBoost had the best value with a score of 0: 8012, CatBoost had 0: 8003. But when it came to XAI analysis of both algorithms, they noted that while XGBoost was endowed with signs of overfitting, it was wiser to accomplish telemarketing success with CatBoost.

## Deep Learning

Dutta et al. (2021) suggest an integration of CNN and GRU for term deposit subscriptions forecast. This study uses telemarketing data from a bank, and measures the performance of the proposed model against benchmark experimental models, such as k-Nearest Neighbour (k-NN), Decision Tree (DT), and Multi-layer Perceptron (MLP). The Convolutional-GRU model has provided an accuracy of 89.59 % and MSE of 0.1041 % higher than other classifiers (Dutta et al., 2021).

In Guha et al. (2024), the authors use machine learning to determine potential customers for term deposits. Based on a dataset of clients of a Portuguese bank, the authors conduct the comparison of Submitted classification algorithms such as SVM, Random Forest, Decision tree and k-NN. Out of these Classifiers, Random Forest reached the highest accuracy of 87%. Making the right customer targeting in banking marketing campaigns can be more efficient using machine learning analysis according to the study of Guha et al., (2024).

Dutta and Bandyopadhyay 2021 developed a hybrid recommender system to predict the probability of subscribers to the term deposit using the)i bank marketing dataset with the help of artificial neural network. The cross validation method used on strata of data gave an overall accuracy of 88.32% an MSE of 0.1168 in the model. When it comes to the models for predicting the probability of the bank term deposit, Dutta and Bandyopadhyay (2021) showed that the neural network model outperformed the ten baseline classifiers, including k-Nearest Neighbours (k-NN), Decision Tree (DT), Multi-layer Perceptron (MLP).

The present study by Clements et al. (2020) introduces Credit Risk Tracking through Deep Learning TCN and LSTM models. The study adopted a sample of 15 million credit card transactions with consideration of early warning on credit risk. TCN model recorded better accuracy of 92.33% in providing the results for outcomes linked with high Recall value of the Gini coefficient of 64.13%. The study acknowledged the fact that deep learning models performed better than GBDT in the financial risk prediction (Clements et al., 2020)

Few feature selection methods chosen and used by Teo and Lee (2019) include k- Best, k – Nearest Neighbour and Closest Centroid for telemarketing success in Portugal for a particular bank after which Decision Tree Model, MLP and Data Randomisation were used. Out of all, Logistic Regression had the highest accuracy of 91.48% while the second accurate model was the Decision Tree with accuracy of 89.91% and MLP having third best accuracy of 90.10%. The study used the CRISP-DM model and showed that its best model of prediction was the Logistic Regression for the level of customer agreement for long-term deposit offers, in line with the findings of Alsolami et al., (2020).

More specifically, Zeinulla et al. (2022) compare classification models for estimating the bank telemarketing success of the Portuguese bank telemarketing dataset. The performance analysis of Random Forest, Deep Artificial Neural Networks (DANN), Logistic Regression, Naive Bayes, k-Nearest Neighbours and Support Vector Machine (SVM) models is done in this study. From the results obtained it was identified that the model with the highest accuracy was Random Forest at 90.88% while DANN had an accuracy of 90.29%. Logistic Regression and Naive Bayes algorithms produced low accuracy as both the results got overfitted from the stage two (Zeinulla et al., 2022).

Dai (2021) also enhances the current Convolutional Neural Network (CNN) in financial forecasting by incorporating data discretisation and Principal Component Analysis (PCA). Through denoise of the dataset and convolutional neural network followed by principal component analysis, the study has obtained rather robust prediction accuracy in the context of financial credit. Analysing empirical Chinese stock market data, it is shown that the updated version of the described CNN model delivered the overall prediction accuracy of 93.2%, which is 12% higher than traditional financial forecasting approaches (Dai, 2021).

Ghatasheh et al. (2020) develop a cost-sensitive Artificial Neural Network (ANN) model for term deposit subscription considering a highly imbalanced exemplary dataset of a Portuguese bank. In order to address the problem of class imbalance the work uses cost-sensitive learning techniques without making any modification to the data. Overall accuracy achieved by using the model was computed on a geometric mean which stood at 79% with Type I and Type II misclassification errors set at 0.192 and 0.229 respectively. This shows that using cost-sensitive analysis can enhance the success rate predictions in telemarketing (Ghatasheh et al., 2020).

For the success prediction of outbound telemarketing for insurance policy loans, Gu et al., (2021 propose an XmCNN. In other words, the model directly employs over 200 elements of the input data as the input indicators with no feature selection, and selects only the feature selection model in the process of controlling overfitting. In the telemarketing success predictors, the proliferation of the proposed XmCNN model outperforms the basic deep learning model in F1-score of 87.47% and FPR of 4.92%. Therefore, the application of the mentioned approach uncovered some of the essential factors that influence or have an impact on telemarketing (Gu et al., 2021).

For estimating the future propensity of deposit customers in the banking industry, Keji, et al. (2024) propose a deep learning model known as Residual Network (ResNet) using Transfer Learning. Pursuing the issue of class imbalance the authors used SMOTE technique and ventured into converting the data set into images for training. The results of the proposed model are compared with the traditional machine learning model in which higher accuracy 93.00%, precision 97.00%, recall 90.00%, and F1 score 93.00% have been achieved. This method provides a very robust method of generating improved forecast solutions for long-term deposit subscriptions (Keji et al., 2024).

Krishna and Reddy (2019) use Deep Neural Networks (DNN) to predict bank marketing data Classification using the following reduction technique: Attribute Subset Selection and Principal Component Analysis. Two datasets of 45,211 and 41,188 instances were used in the study with accuracies of 91.15% and 89.24% gained respectively. The authors also compared the efficiency of the DNN model to general classifiers including Decision Trees, Naïve Bayes, Support Vector Machines, and k-Nearest Neighbours; the results showed the competencies of deep learning in banking marketing predictions were high with greater accuracy (Krishna & Reddy, 2019).

Kim et al., (2015) on the other hand, employ a deep CNN for a telemarketing based on the bank’s success of a sample of the data from Portugal based banking organisations. Thus, by applying deep learning it is possible to obtain 84.9% of accuracy, which is higher than classical methods of machine learning, including decision trees and support vector machines. The same applied to the CNN model, which also demonstrated high effectiveness in the analysis of difficult multi-parametric data obtained from telemarketing campaigns and will be a solid basis for making further marketing predictions in the banking industry in the future (Kim et al., 2015).

In the study of Yüksel (2023), TabNet, a new deep learning architecture, has been evaluated when used for credit risk scoring in comparison to XGBoost and ANN. The study used two datasets for credit risk, which are open source and Table 2 shows that compared to baseline models, TabNet had higher recall of 99.10% while XGBoost had higher accuracy and AUC score. The study also shows that there can be few false negatives with TabNet, which makes the method useful for banks striving for minimising default amount (Yüksel, 2023).

Following machine learning techniques, Rony et al. (2024) analyse the customer behaviour to take long-term bank deposits. In order to fulfil the targets of the study, Exploratory Data Analysis (EDA), Principal Component Analysis (PCA) and five different Machine Learning algorithms namely logistic regression, random forest, support vector machine (SVM), K nearest neighbours and multilayer perceptron (MLP) are used. Based on this criterion, it was found that Logistic Regression belongs to the best of all the algorithms in terms of accuracy 90.64 % and sensitivity 99.05 %. By evaluating the literature, authors attest to the effectiveness of the proposed logistic regression technique for the classification of customers in the banking domain (Rony et al. 2024).

Guo et al. (2023) introduce a novel ensemble approach for predicting the success of bank telemarketing using four artificial neural networks (ANNs) optimised by metaheuristic algorithms: It develops four algorithms – Electromagnetic Field Optimization (EFO), Future Search Algorithm (FSA), Harmony Search Algorithm (HSA) and Social Ski-Driver (SSD). The highest classification rates were obtained from EFO-ANN classification set by AUC 0.8030 and FSA-ANN with AUC of 0.7809 while HSA-ANN obtained a classification rate of AUC 0.7356 only out of 30,488 records. Consequently, the adopted research confirms that ignition of the hybrid metaheuristic models enhances telemarketing performance prediction (Guo et al. 2023).

Kim et al. (2015) consider the success of the bank telemarketing and, to this end, propose the use of a Deep Convolutional Neural Network (DCNN) based on the 45,211 phone calls dataset. This work looks at different hyperparameters in the context of this model such as the number of layers and learning rate for model training. By using the DCNN, the compilation tested the accuracy of 76.70% which is greatly superior to classifiers like logistic regression and support vector machines. Kim et al., (2015) identifies that the research done in this area points to the ability of DCNNs to capture various complexities inherent in financial data for improvement of the predictive qualities.

## TabNET

Yu et al. (2024) aim at the credit risk of the user and utilise LightGBM, XGBoost, and TabNet for the model, SMOTEENN and PCA for pretreatment. These models are assessed in this study using a dataset containing more than 40,000 records from a bank and yielded an AUC-ROC = 0.9999 from the LightGBM model, and outperforms other models. TabNet with the deep learning architecture had a slightly lower AUC-ROC of 0.9816 In contrast, TabNet’s interpretability is stronger attributed to its attention mechanism (Yu, Yu, et al. 2024)

TabNet has been presented by Arık and Pfister in the year 2021, as a deep learning model aimed primarily at tabular data. In this work, to select the features important at each decision step, the model employs an original sequential attention process which makes the training process more effective and the essential steps more understandable. Compared with other gradient-boosting tree algorithms such as XGBoost and LightGBM, TabNet has shown superior performance, a top accuracy level of 96.99% on the Forest Cover Type dataset. The attention-based feature selection also enhanced the increase in transparency in the decision making processes (Arık & Pfister, 2021).

Mollo (2023) is devoted to enhancing the loan default prediction by utilising TabNet with LightGBM, XGBoost, Logistic Regression, and Random Forest models. Some potential problems it may cause are also discussed, such as interpretability via feature masks and efficiency for tabular data emphasised in the study. Compared to other models in terms of performance, it was established that TabNet gave better high test accuracy and test precision of 67.3 % and 33.4% respectively. The study also points to the application of TabNet in other financial prediction activities because of the model interpretability’s balance of good performance (Mollo, 2023).

TabNet and VIME are explored in Tober (2022) in a transactional underwriting context within the financial sector. The authors then use SHAP values for regularisation comparison of these deep learning models with the classical gradient-boosted tree models in terms of performance and interpretability. A comparison against the baseline gradient-boosted tree model identified statistical significance and reduced RMSE by 0.05 across the boards when pre-training in a new market. The findings noted from the research are on how deep learning models are capable of generalisation across the different market conditions (Tober, 2022).

## Explainable AI in Bank Deposit Term Subscription Prediction

In Lin (2024), the author made use of the explanation techniques called LIME and SHAP to understand the two’s effectiveness in improving the interpretability of domain specialists in fraud detection. The analysis also quantitatively assessed the validity of these explanations advocated for in the literature and found that claims about model holism were supported: high accuracy was associated with lower interpretability. The experiments showed that SHAP provides importance scores for features and LIME local explanations which give a clear picture of the decision making taking place in a machine learning model .

Al Hammadi (2022) discusses the uses of data mining to prospect potential clients for extended banking deposits through PCA and k-mean clustering. As the research shows, customer responses depend on factors like customer income or previous campaign results. The analysis by using PCA has made dimensionality reduction, so the model that is created has 87.76% of the accuracy score. Therefore, the current study results demonstrate that integrating data mining techniques with explainable AI techniques is advantageous for improving marketing approaches in banking (Al Hammadi, 2022).

Vashi et al. (2023) conduct an association rule and decision tree analysis of the response to bank marketing, with much focus on the explainable AI aspect of the work. From the analysis of the customers profiles they are able to show that indeed the marketing outcome is greatly influenced by some variables such as age and job. The tool helps to make effective decisions as to which clients are more likely to take long-term deposits and improve marketing efforts. The authors underline the importance of making the AI process clear to the stakeholders and contribute to creating a unique mechanism for providing better targeted financial services (Vashi et al., 2023).

Chen (2023) offers interpretable data-mining techniques for forecasting term deposit subscriptions, using methods included in decision trees and the random forest classification algorithm. The work uses data obtained from the telemarketing of a Portuguese bank and shows that the customer balance and customer age have a strong effect on the subscription. Out of all these variables the random forest model yields higher classification rates than the decision tree model particularly when the duration variable which is not measurable prior to client interaction is excluded the average classification is 73.7% seen through the AUC-ROC. As mentioned in this study, the application of explainability in predictive models would improve decision making in the formulation of banking marketing strategies (Chen, 2023).

# Proposed Sampling Strategy

Regarding the sampling strategy for the current study, the current study recommends the use of stratified random sampling based on the data given by the Bank Marketing Dataset from the UCI Machine Learning Repository. This dataset can be found at the UCI Machine Learning Repository – Bank Marketing[[1]](#footnote-0). This method enables depiction of relevant subgroups in terms of demographic and financial variables improving the credibility of the research findings. The study will use stratified random sampling in order to obtain representative samples that guarantee the richness of the variability of the customers responses to the marketing campaigns thus making the development of the prediction models reliable.

## Sampling Method

For this study the identified sampling technique is the stratified random sample with which the population is partitioned into subgroups according to relevant attributes. The rationale in this approach is to make certain that every demographic category has been well captured in the research sample. When applied to the Bank Marketing Dataset, this means partitioning the customers according to factors such as age, type of job, whether they are married, and the results of previous campaigns.

This sort of method of stratification enables the researcher to control for variance in customer responses that may be occasioned by such characteristics. In this respect, the study seeks to use purposive sampling to ensure that various customer behaviours and preferences concerning term deposit subscriptions are captured by sampling within these identified strata. The stratified sampling technique will greatly increase the levels of external validity because results from the study will hold across the broader population of bank customers.

## Sampling Procedure

Similar to any research work, the following systematic steps have been taken while sampling for this research work to get an accurate and reliable sample from the Bank Marketing Dataset. The process includes the following key steps:

Defining the Population: The whole structure comprises the 45,211 records of the Bank Marketing Dataset that comprises a range of customer characteristics and their reactions to marketing offers. They include customer features and contact information as well as accounts of previous marketing communication endeavours showing that it is a population of consumers demonstrating varying levels of consumption behaviour based on age, employment status, and previous experiences with the bank.

Dividing the Population into Strata: These lists will be further subdivided into more manageable pockets of homogenous population with similar characteristic attributes.

Age Groups: To address this, marketing response will be segmented into several age groups comprising; 18-25 years, 26-35 years, 36-45 years, 46-55 years, and 56 years and above.

Job Types: Occupational segmentation, such as management, technician, service, etc., will also determine how or if subscription is influenced.

Previous Campaign Outcome: Responses received from previous marketing initiatives would be classified into subscribed and non-subscribed to investigate past experiences as a way to segmentation.

Determining Sample Size: The amount of respondents will be sufficient to reach statistical validity and randomness. It is well understood that the first and most crucial step for model training is setting up the dataset; in this case, the main split of the data will be 80/20, meaning 80% of the data will be used for training and only 20% will be used for model testing. This division enables us to have accurate measures of model and, at the same time, prevent cases of overfitting.

Selecting the Sample: Systematic samples will be taken from the different stratum to synonymize the sample with the different subgroups in the population. This involves the use of quota sampling where the researcher takes samples from each stratum provided the samples represent the population as used in the full research.

Verifying the Sample: Once the unit of analysis has been selected, the representativeness of the sample will be determined by comparing the demographic characteristics of the sample against the characteristics of the parent population. This step helps to make sure that those people actually selected as the sample are a fair representation of the population in relation to the different characteristics that exist in it, in order that the conclusions made from the analysis can be applicable or generalizable to the population.

By using this elaborate sampling technique, the study ensures that the selected sample is a good representation of the whole population with diverse trends in the customers’ response to banking marketing communication initiatives. This increases the validity as well as reliability of the study, while giving a strong ground for creating models of subscription for term deposits.

## About the Dataset

Key Features of the Dataset

Age: The client’s age which is a good indicator of their financial status and the likely response rate to marketing techniques.

Job: Information in regard to the nature of the employment of the client, which has implications on the income levels and or expenditure of the client.

Marital Status: This attribute represents the marital status of the client; single, married or divorced which defines the financial status and familyνος.

Education: Client’s education level may be linked to their level of understanding in financial affairs and the resulting willingness to embrace banking offers.

Credit Default: The loan application and deposit option availability is dependent on this binary attribute which indicates whether the client has credit default or not.

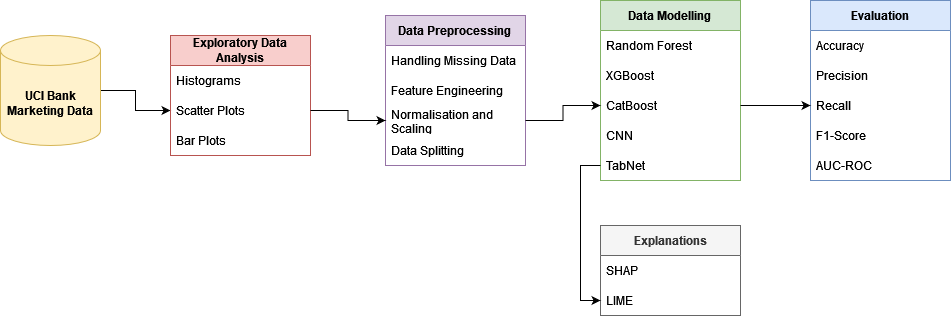
Balance: The balance of the account that characterises the average yearly balance as a measure of the financial standing of the client.

Campaign Outcome: This is another binary; it will equal one if the client subscribed to a term deposit, after the campaign.

This becomes an enormous reservoir of information to use in investigating customer behaviour and tendencies in banking, with a view of deducing smarter analytic models.

# Proposed Methodology

The research strategy for this study focuses on leveraging machine learning models and explainable AI techniques to analyse the Bank Marketing Dataset effectively. The proposed approach encompasses several advanced models, including Logistic Regression, Random Forest, XGBoost, CatBoost, Convolutional Neural Networks (CNN), and TabNet. By integrating these models, the research aims to improve predictive accuracy while also providing interpretable results that can guide banking strategies.



***Figure 1: Methodology***

## Models

### Random Forest

Random Forest is an ensemble learning method that constructs multiple decision trees during training and aggregates their predictions. This model is robust against overfitting and capable of handling complex interactions among features. Its interpretability is enhanced through feature importance metrics, allowing stakeholders to understand the key drivers of client behaviour.

### XGBoost and CatBoost

Both XGBoost and CatBoost are gradient-boosting frameworks renowned for their efficiency and predictive performance. They can handle categorical features directly and perform well in high-dimensional spaces. These models will be pivotal in understanding the nonlinear relationships within the dataset and improving overall predictive performance.

### Convolutional Neural Networks (CNN)

CNNs, typically used in image processing, will be adapted to analyse structured data. The architecture will involve input layers designed to capture complex feature interactions. CNNs excel at extracting patterns from data, improving predictive power through automatic feature learning.

### TabNet

TabNet is a deep learning architecture specifically designed for tabular data, utilising attention mechanisms to identify relevant features and relationships without extensive feature engineering. This model allows for effective learning from the dataset's intricate structure while also providing interpretable insights into the decision-making process.

## Explainable AI Techniques

### SHAP (SHapley Additive exPlanations)

SHAP is a powerful method for interpreting machine learning models by assigning an importance value to each feature for a given prediction. It quantifies how much each feature contributes to the prediction, enhancing transparency. This technique will be employed to provide global and local explanations of the model outputs, facilitating stakeholder understanding of the decision-making process.

### LIME (Local Interpretable Model-Agnostic Explanations)

LIME generates local explanations for individual predictions, approximating the model locally with an interpretable model. This technique will help in understanding the reasoning behind specific predictions, particularly for complex models like CNN and TabNet. By providing context-specific explanations, LIME aids stakeholders in trusting and validating the model's decisions.

## Data Collection

The data used in this study will be sourced from the Bank Marketing Dataset available at the UCI Machine Learning Repository. This dataset includes historical data on telemarketing campaigns conducted by a Portuguese bank, providing valuable insights into customer behaviour and marketing effectiveness. The dataset consists of a diverse range of features, enabling comprehensive analyses.

## Data Preprocessing

### Data Cleaning

Preprocessing will also include removing any errors and inconsistencies in the data; this will involve handling of missing values. Some analyses may lack complete datasets, and will therefore employ techniques like mean or median imputation, so as not to compromise the integrity of the dataset.

### Feature Engineering

Feature engineering feature transformation is the process of expanding the number of features to make the model better. This may range from producing interaction terms, coding qualitative data into dummy variables and constructing temporal variants to capture the relationship between consequent responses on the part of the clients.

### Normalisation and Scaling

For equal contribution of every feature in training, normalisation techniques will be used in the model. This step is very important especially for algorithms sensitive to feature scales such as gradient boosting machines and the neural networks.

### Data Splitting

The dataset will be split into training, and test sets. Time-based splitting will be employed to respect the temporal nature of the data and prevent future information from influencing past outcomes. This ensures that the model is evaluated fairly and accurately.

## Model Training

Models will be trained using the training dataset, with performance monitored using the validation dataset. Various hyperparameters will be tuned to optimise model performance, and advanced time-series cross-validation methods will be implemented to validate the results effectively.

## Model Evaluation

### Evaluation Metrics

The models will be evaluated using multiple performance metrics, including:

Accuracy: The overall correctness of the predictions made by the model.

Precision and Recall: Metrics that provide insights into the model's performance regarding positive class predictions.

F1-Score: The harmonic mean of precision and recall, offering a balanced measure of performance.

AUC-ROC Curve: Evaluating the trade-off between sensitivity and specificity across different thresholds.

### Performance Comparison

The evaluation will be made in terms of the metrics described above, by which the ability of all the models will be compared, and the best one for predicting term deposit subscriptions will be determined. This comparison will serve to help identify what are the best and worst strategies for use in future model selection and modification.

## Ethical, Legal, and Regulatory Standards

### Ethical Considerations

Customers’ data remain invulnerable to any research studies, and therefore, ethical issues are extremely crucial when carrying out a research study. Before collecting the data, the fact was mentioned that the Bank Marketing Dataset is publicly available but several precautions will be taken in order to avoid violating the privacy and rights of people. The findings will be reported professionally and credibly, and no identifiable data will be published.

### Legal Compliance

Such issues like usage of data, protection of data, and related laws like the GDPR and CCPA will be followed in the study. It is critical to guarantee that the gathered data is used to support these regulations, to ensure the ethical process throughout the population.

### Reporting and Transparency

Methodologies used in the research as well as data collection and analysis sources will be reported in detail and any constraints met will also be well explained. This practice serves to reduce the risk of any inaccuracies in the research findings and helps other researchers to obtain similar results.

All in all, the proposed sampling strategy and the most important research method being primary research enable a comprehensive investigation of factors determining term deposit subscriptions. To this end, the study aims at developing a hybrid model that combines a state-of-the-art machine learning algorithm, named clustering, with an eXplainable Artificial Intelligence approach to generate accurate yet interpretable results that underpin efficient banking marketing.

# References

Abd, N.S., Atiyah, O.S., Ahmed, M.T. and Bakhit, A., 2024. Digital Marketing Data Classification by Using Machine Learning Algorithms. *Iraqi Journal for Electrical And Electronic Engineering*, *20*(1).

Al Hammadi, M., 2022. Identifying Prospective Clients for Long-Term Bank Deposit.

Alsolami, F.J., Saleem, F. and Abdullah, A.L., 2020. Predicting the Accuracy for Telemarketing Process in Banks Using Data Mining. *Comp. It. Sci*, *9*, pp.69-83.

Amajuoyi, C.P., Nwobodo, L.K. and Adegbola, A.E., 2024. Utilizing predictive analytics to boost customer loyalty and drive business expansion. *GSC Advanced Research and Reviews*, *19*(3), pp.191-202.

Arik, S.Ö. and Pfister, T., 2021, May. Tabnet: Attentive interpretable tabular learning. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 35, No. 8, pp. 6679-6687).

Broby, D., 2022. The use of predictive analytics in finance. *The Journal of Finance and Data Science*, *8*, pp.145-161.

Chen, T.C.T., 2023. Explainable artificial intelligence (XAI) in manufacturing. In *Explainable Artificial Intelligence (XAI) in Manufacturing: Methodology, Tools, and Applications* (pp. 1-11). Cham: Springer International Publishing.

Chlebus, M. and Osika, Z., 2020. *Comparison of tree-based models performance in prediction of marketing campaign results using Explainable Artificial Intelligence tools*. University of Warsaw, Faculty of Economic Sciences.

Clements, J.M., Xu, D., Yousefi, N. and Efimov, D., 2020. Sequential deep learning for credit risk monitoring with tabular financial data. *arXiv preprint arXiv:2012.15330*.

Dai, W., 2022. Application of improved convolution neural network in financial forecasting. *Journal of Organizational and End User Computing (JOEUC)*, *34*(3), pp.1-16.

Dutta, S. and Bandyopadhyay, S.K., 2021. Recommender System for Term Deposit Likelihood Prediction using Cross-validated Neural Network. *South Asian Journal of Social Studies and Economics*, *11*(3), pp.21-28.

Dutta, S., Bose, P., Goyal, V. and Bandyopadhyay, S.K., 2021. Applying convolutional-GRU for term deposit likelihood prediction. *International Journal of Engineering and Management Research*, *11*(3), pp.265-272.

Ghatasheh, N., Faris, H., AlTaharwa, I., Harb, Y. and Harb, A., 2020. Business analytics in telemarketing: Cost-sensitive analysis of bank campaigns using artificial neural networks. *Applied Sciences*, *10*(7), p.2581.

Gu, J., Na, J., Park, J. and Kim, H., 2021. Predicting success of outbound telemarketing in insurance policy loans using an explainable multiple-filter convolutional neural network. *Applied Sciences*, *11*(15), p.7147.

Guha, S., Dhar, M., Saha, B. and Saha, N., 2024. A Machine Learning Approach to Identify Customers for Subscribing Term Deposit. *Alochana Journal,* 13(5), pp. 73-83.

Guo, W., Yao, Y., Liu, L. and Shen, T., 2023. A novel ensemble approach for estimating the competency of bank telemarketing. *Scientific Reports*, *13*(1), p.20819.

Hassani, H., Huang, X., Silva, E. and Ghodsi, M., 2020. Deep learning and implementations in banking. *Annals of Data Science*, *7*, pp.433-446.

Hjertaas, M., Knudsen, H.K., Lindstrøm, J. and Sælemyr, J., 2023. *Identifying the Best Machine Learning Model for Predicting Bank Term Deposits: An Empirical Study Using Public, Post Financial Crisis Data* (Bachelor's thesis, NTNU).

Keji, A.A., FAKEYE, O., ONOCHIE, N.N. and SANGOTOKI, O., 2024. Predicting long-term deposit customers using convolutional neural network and data conversion technique. *African Scientific Reports*, pp.192-192.

Kim, K.H., Lee, C.S., Jo, S.M. and Cho, S.B., 2015, November. Predicting the success of bank telemarketing using deep convolutional neural network. In *2015 7th International Conference of Soft Computing and Pattern Recognition (SoCPaR)* (pp. 314-317). IEEE.

Kinnander, M., 2020. Predicting profitability of new customers using gradient boosting tree models: Evaluating the predictive capabilities of the XGBoost, LightGBM and CatBoost algorithms.

Krishna, C.L. and Reddy, P.V.S., 2019. Deep Neural Networks for the Classification of Bank Marketing Data using Data Reduction Techniques. *Int J Recent Technol Eng*, *8*, p.3.

Lee, C.S., Cheang, P.Y.S. and Moslehpour, M., 2022. Predictive analytics in business analytics: decision tree. *Advances in Decision Sciences*, *26*(1), pp.1-29.

LERTVIPADA, C. and Laohakiat, S., 2023. USING INTERPRETABLE MACHINE LEARNING METHODS FOR ANALYZING BANK TELEMARKETING DATA.

Lin, C.F., 2018. *Application-grounded evaluation of predictive model explanation methods* (Doctoral dissertation, Master’s thesis, Eindhoven University of Technology).

Meshref, H., 2020. Predicting loan approval of bank direct marketing data using ensemble machine learning algorithms. *International Journal Of Circuits, Systems And Signal Processing,* 14(1), pp. 914-922.

Moinuddin, M., Usman, M. and Khan, R., 2024. Decoding Consumer Behavior: The Role of Marketing Analytics in Driving Campaign Success. *International Journal of Advanced Engineering Technologies and Innovations*, *1*(4), pp.118-141.

Mollo, A., 2022. Advancing loan default prediction with interpretable Tabnet models.

Olaniyi, O., Shah, N.H., Abalaka, A. and Olaniyi, F.G., 2023. Harnessing predictive analytics for strategic foresight: a comprehensive review of techniques and applications in transforming raw data to actionable insights. *Available at SSRN 4635189*.

Patwary, M.J., Akter, S., Alam, M.B. and Karim, A.R., 2021. Bank deposit prediction using ensemble learning. *Artificial Intelligence Evolution*, pp.42-51.

Rony, M.A.T., Hassan, M.M., Ahmed, E., Karim, A., Azam, S. and Reza, D.A., 2021, December. Identifying long-term deposit customers: a machine learning approach. In *2021 2nd International Informatics and Software Engineering Conference (IISEC)* (pp. 1-6). IEEE.

Shin, Z., Moon, J. and Rho, S., 2022. A comparative analysis of ensemble learning-based classification models for explainable term deposit subscription forecasting. *Journal of Society for e-Business Studies*, *26*(3).

Tober, S., 2022. Investigating the Use of Deep Learning Models for Transactional Underwriting.

Tuan, N.M., 2022. Machine Learning Performance on Predicting Banking Term Deposit. In *ICEIS (1)* (pp. 267-272).

Turkmen, E., 2021, February. Deep learning based methods for processing data in telemarketing-success prediction. In *2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV)* (pp. 1161-1166). IEEE.

Vashi, H., Yadav, J. and Varde, A.S., 2023, September. Predictive Analysis of Bank Marketing for Financial Decision Support and Smart Economy. In *Intelligent Systems Conference* (pp. 474-490). Cham: Springer Nature Switzerland.

Wang, Q., A Neural Network Classification Optimized by Genetic Algorithm to Predict the Success of Bank Telemarketing.

Wankhede, P., Singh, R., Rathod, R., Patil, J. and Khadtare, T.D., 2019. Improving Prediction of Potential Clients for Bank Term Deposits using Machine Learning Approaches.

Wardani, B.S., Sa’adah, S. and Nurjanah, D., 2023. Measuring and Mitigating Bias in Bank Customers Data with XGBoost, LightGBM, and Random Forest Algorithm. *Jurnal Ilmiah Teknik Elektro Komputer dan Informatika (JITEKI)*, *9*(1), pp.142-155.

Wassouf, W.N., Alkhatib, R., Salloum, K. and Balloul, S., 2020. Predictive analytics using big data for increased customer loyalty: Syriatel Telecom Company case study. *Journal of Big Data*, *7*(1), p.29.

Yu, C., Jin, Y., Xing, Q., Zhang, Y., Guo, S. and Meng, S., 2024. Advanced user credit risk prediction model using lightgbm, xgboost and tabnet with smoteenn. *arXiv preprint arXiv:2408.03497*.

YUKSEL, A.B., *CREDIT RISK SCORING WITH TABNET ARCHITECTURE* (Doctoral dissertation, tilburg university).

Zaki, A.M., Khodadadi, N., Lim, W.H. and Towfek, S.K., 2024. Predictive analytics and machine learning in direct marketing for anticipating bank term deposit subscriptions. *American Journal of Business and Operations Research*, *11*(1), pp.79-88.

Zeinulla, E., Bekbayeva, K. and Yazici, A., 2018, October. Comparative study of the classification models for prediction of bank telemarketing. In *2018 IEEE 12th International Conference on Application of Information and Communication Technologies (AICT)* (pp. 1-5). IEEE.

Zulaikha, S., Mohamed, H., Kurniawati, M., Rusgianto, S. and Rusmita, S.A., 2020. Customer predictive analytics using artificial intelligence. *The Singapore Economic Review*, pp.1-12.

1. <https://archive.ics.uci.edu/dataset/222/bank+marketing> [↑](#footnote-ref-0)