National College of Ireland

Exploring the Impact of Fintech Innovations on Customer Acquisition in Banking: A Case Study Using Marketing Campaign Data

M.Sc. Research Project
M.Sc. Data Analytics

Sujit Dhakolia
Student ID: 23185287

School of Computing
National College of
Ireland

Supervisor: Dr. Abid Yaqoob

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: ……Sujit Sanjay Dhakolia……

Student ID: ……x23185287……

Programme: ……M.Sc. Data Analytics…… **Year:** …2024…….

Module: ……M.Sc. Research Report……

Supervisor: ……Dr. Abid Yaqoob……

Submission Due Date: ……12/12/2024……

Project Title: …… Exploring the Impact of Fintech Innovations on Customer Acquisition in Banking: A Case Study Using Marketing Campaign Data……

Word Count: ……7753…… **Page Count**……21……

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: ……Sujit Sanjay Dhakolia……

Date: ……11/12/2024……

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Exploring the Impact of Fintech Innovations on Customer Acquisition in Banking: A Case Study Using Marketing Campaign Data

Sujit Dhakolia
23185287

Abstract

The banking sector is one of the most dynamic due to the developed digital technologies and the growing popularity of fintech. This work aims to explore how the advances in the fintech sector affect the strategies of customer acquisition in the banking sector. Deducing results from different marketing campaigns, the research identifies various methods of using financial technology in the marketing of products to different customers and increasing the sales to current consumers. The study also discusses how fintech helps to evaluate other parameters concerning cost of acquiring the customers, for instance demography, contact channels, campaigns characteristics among others using machine learning algorithms and statistical tools.

One of the main topics in this project is the laudable work fintech has done in perfecting the marketing strategies, given the shifting market dynamics. Following the main research question of how marketing activities enabled by fintech affect buying processes in the banking sector, the work establishes how the utilization of smart technologies improves the marketing communication process.

To measure the extent to which the various fintech technologies impact the outcome of the advertising campaigns and the resulting customer acquisition, the study categorizes and systematically reviews various marketing strategies. Thus, the findings are expected to contribute to creating a set of convincing marketing tactics suitable for banks, expand the theoretical and practical data on fintech marketing, and help banking organizations focused on improving the strategies for acquiring customers.

This work makes both theoretical and practical contributions to the understanding of marketing by fintech and associated strategies used in banks. The studies are believed to contribute to the banking institutions' management of competition, with solutions to the increasing confrontation rising from the improvement of technology through optimum customer acquisition strategies.

1 Introduction

The banking industry currently experiences a critical transition because fintech innovations bring forward technologies such as artificial intelligence, machine learning and big data. New technological developments challenge usual banking approaches through their transformation of acquisition systems and advertising methods. The research analyzes how fintech-driven marketing methods evolve into effective solutions that boost customer acquisition in real market conditions.

Figure 1 Fintech Innovations

The research will thus seek to examine the role of marketing through Fintech in the attraction of customers in this banking industry. Through understanding how technology yields new platforms through which customers can be reached, the work underscores the need for financial institutions to develop appropriate strategies that reflect the technological trends (Alt et al., 2018; Haddad & Hornuf, 2019). The technologies being unleashed under the broad heading of fintech are changing the external environment in a way where traditional approaches are no longer adequate, making the implementation of sophisticated marketing strategies essential. To respond to this emerging issue, the study analyses the role of different technologies assisted by fintech in acquiring customers. During empirical investigations relating to a marketing campaign, the research determines the marketing tools that prove efficient for customer acquisition and customer maintenance. As a proposed solution, it involves an application of machine learning models for advanced prediction of campaign effectiveness and optimal classification of customers (Kou et al., 2021). This research's outcome will be helpful to financial institutions to bear managerial implications by providing them the ways to improve their market communication efficiency. Thus, analysis of how fintech affects customer acquisition can help banks improve their positioning in the context of growing competition in the financial services market (Haddad & Hornuf, 2019; Kou et al., 2021).

1.1 Background and Motivation

The banking world is in a profound transition due to the emergence of the fintech sector. Fintech refers to any form of technology that is used in the financial sector; Mobile money, Blockchain, Artificial Intelligence and Machine Learning are some of the existing examples. Such technologies are being applied to improve customer experience, optimize operations, and deliver more targeted financial solutions. Legacy concepts of banking, which previously went through physical counters and businesspeople contact, are now being substituted with much more flexible and effective solutions which can be implemented online and provide customers with more options and more opportunities around the world.

One of the most revolutionizing fields of such change is the domain of customer acquisition and marketing. As the usage of fintech solutions grows, banks and other financial service providers use these technologies as supplementary tools for marketing to new clients, as well as to retain existing

ones. The technique used by banks to deliver personalized services as well as incorporating big data using AI makes it easier for banks to market themselves to the intended clients.

However, bringing in tech firms within marketing strategies poses several challenges for banks despite the many benefits that they have to offer. High technology activation and the change in consumers' expectations simultaneously shape the world economy and put a lot of pressure on companies to adapt quickly and launch breakthrough products. They must adapt to emerging trends to remain relevant, and this requires understanding of how fintech is affecting customer acquisition as a process (Gomber et al., 2017). This research is inspired by the current technological advancement that has called for banks to waive up from being traditional in acquiring new customers. Many of the old marketing practices are not adequate for this task, as customers nowadays require convenient, integrated, and rapid solutions. Based on the analysis of the role of several specific fintech marketing techniques, this work will seek to identify the optimal tools and methods that firms in the sphere of financial services can apply when seeking to capture and maintain consumer attention. The results will provide specific implementation strategies that allow banks to effectively meet customer needs regarding marketing messages and would help industry players to sustain their competitive standing within a high-growth online environment (Kou et al., 2021).

With this, the aim of this study is to bring together fintech innovations and marketing practices so the existence of new breakthroughs can assist banks to enhance customer acquisition and optimize marketing effectiveness. In the end, this research will offer theoretical insights into fintech and practical advice about how financial institutions can improve their marketing efforts as they navigate an increasingly competitive environment (Alt et al., 2018; Kou et al., 2021).

1.2 Research Question and Objectives

The research question for this study is: How do different fintech-driven marketing strategies impact customer acquisition in banking?

The objectives of this research are:

- To assess how online marketing supported by fintech techniques affects customer attraction in the banking industry.
- To examine key factors of marketing, which include demography, contact type, and promotion success rates, regarding campaigns based within the financial technology industry.
- To use ML, training and testing on two datasets (the first containing demographic and campaigns information, the second augmented by economic indicators), to assess the efficiency of target marketing and fintech based campaigns in acquiring customers.
- For advancing machine learning approaches for the estimation of future effects of various fin-tech supported marketing campaigns on customer attraction.
- To make appropriate suggestions for banks on how they should improve their marketing approaches with a view to improving the rate of acquisition.

1.3 Research Contributions

This study contributes to knowledge on the impact of fintech innovations in customer acquisition in the banking industry as customer demand for innovative, value-add economic technologies grows. The paper extends the marketing and financial technology literature by analyzing how advanced tools like AI, machine learning and big data affect the marketing mix and customer acquisition. For instance, the roles of fintech in delivering personalized marketing which past studies by Alt et al.

(2018) and Gomber et al. (2017) investigated and work by Kou et al. (2021) which studied the role of machine learning in improving decision-making in customer targeting. Following these results, this study extends the findings by adopting the machine learning algorithm, Random Forest, to examine the demography, contact, and campaign factors influencing acquisition, again supporting Haddad & Hornuf (2019). In contrast to prior research knowledge, this investigation yields a finer-grained understanding of consumers' actions by employing state-of-the-art artificial neural networks and feature-selection analysis.

Also, compared to more conventional methods that may suppress subtle nonlinearities in customer subscription, this study makes use of Random Forest technique, which can expose more complicated structures. This resonates with Kou et al. (2021) who noted that machine learning can revolutionise risk management and marketing in the Fintech setting. Furthermore, in contrast to previous research that considered individual fintech innovations such as blockchain or DeFi as a subject of utmost importance (Alt et al., 2018), this work takes a more general perspective on how fintech advances and collaborates with machine learning to enhance marketing outcomes.

From an applied point of view, the conclusions of this research provide suggestions on how banking institutions can include fintech solutions in their marketing communication strategy. Noting predictors for customer acquisition duration and poutcome_success the study assists the banks management in better positioning of resources when conducting the marketing efforts. This builds on prior research by Gomber et al. (2017) who looked at how fintech increases operational effectiveness. Additionally, the current marketing tools enabled by artificial intelligence as discussed by Kou et al. (2021), are known to possess accuracy in the forecast of promotional strategies' success and the desired target customers' identification. This research builds on this understanding by moving on to show how the targeting could be optimized through machine learning, so the message hits the right people at the right time without a lot of expenses incurred.

Besides these practical implications, the study serves to the topical literature on fintech, particularly outlined by the importance of having co-ordinated systems of fintech in enhancing sustainable development. Though Haddad & Hornuf (2019) studies examined the impact of various forms of economic factors to determine the degree of participation in fintech completion, this work considers the ability of fintech to compete in the financial world that has gone digital. It aims to show how the realised marketing communication as well as competitiveness of a bank can be enhanced based on the integrated approach towards the technological advancement and the customer oriented strategies.

2 Related Works

In Financial technology innovations have brought substantial changes to the way banks develop their marketing strategies. Kou et al. (2021) showcase how machine learning transforms real-time customer targeting as Haddad and Hornuf (2019) explain the difficulties banks face from integrating fintech into traditional banking systems.

Alt et al. (2018) indicate that banking sector opportunities for boosting customer acquisition come from two emerging trends: blockchain and AI-driven analytics. Researchers need to address remaining knowledge gaps regarding how to best utilize fintech platforms in strategic campaign development. The current study continues previous works by implementing Random Forest to analyze marketing datasets to determine what elements prompt customer acquisition.

The use of Fintech's has significantly impacted the way banks go about the business of acquiring their clients. Thus, through the application of big data and Artificial Intelligence, value proposition maps can be created and potential clients targeted rightly. According to Gomber et al. (2017), customers' data through fintech tools shift the banking's marketing communication tools into being

result-driven, emphasizing on customers' segments as well as targeting them with the carrying out of a relevant marketing campaign. In the same line, Alt et al. (2018) also underscores communication channels' transformation based on mobile banking and blockchain that increases a secure way of deep communication with customers. These changes transform the conventional way of positioning the customer relationship thus enabling personal and cost-effective acquisition pattern.

Marketing in the banking sector has become reliant on machine learning and AI innovations. According to Haddad and Hornuf (2019), because of the application of Artificial Intelligence in segmentation and the use of predictive models, the banks can be able to see customers behavior patterns and tendencies in a better way hence improving on the campaign initiatives. Kou et al., (2021) posited that with AI being able to adapt and remarkably adjust the marketing strategies in real-time customer targeting has experienced a considerable boost and hence makes them more responsive to dynamic marketing conditions. In addition to enhancing the probability of acquiring customers, these tools enhance the effectiveness of marketing resources by targeting customer value segments.

Data analytics is the core component of the fintech based marketing strategies environment. Gomber, et al, (2017) further explain that, through big data, the banks can capture value creating insights about customers' behavior, their likes, and their wants. Such perceptions are useful in customer classification and improve the efficiency of campaigns and thus explain higher acquisition. Haddad and Hornuf (2019) also add that, through the help of machine learning, there is a way to know which marketing activity should be done for greater customer retention and outreach at the same time.

On the bright side, fintech has enhanced the management of customer acquisition and marketing. Even so, there are hurdles associated with using fintech that institutions must consider. One remains a data privacy issue because financial organizations must follow strict regulatory requirements while using customers' data for analytic purposes (Gomber et al., 2017). Digital transformation activities in traditional banking institutions also present challenges to fintech adoption including resistance to change, which demands spending in employee training and technology (Alt et al., 2018). Furthermore, problems like interfacing of advanced technologies with older foundational structures are also key demoting the adoption of fintech solutions.

Though it is crucial to acknowledge that there is a laundry list of issues associated with fintech, there are incredible opportunities to transform the acquisition of customers. Blockchain and DeFi, as emerging concepts, are new to the banking world and challenging traditional models of customer interactions and relations. For example, Kou et al., (2021) described that blockchain technology increases transparency and security which in turn boosts customer relationships. DeFi tools facilitate the generation of new financial opportunities that satisfy the current and promising customers' values, including individualism and liberty.

This study resonates with the objectives of prior research that blends the use of machine learning models in turning available fintech features into marketing breakthroughs to respond to customer acquisitions in the banking industry. This paper aims to investigate how and to what extent customer targeting could be enhanced by data mining techniques such as Random Forest and other algorithms. The main contributions of this work include:

1. Showing how the deployed machine learning models can be used to optimize campaigning effectiveness and gather invaluable data on customers.
2. Some of the features like duration and poutcome_success that are significant in customer

decisions can also be of help in fresh marketing actions.

3. Introducing a comparison of the algorithms with special emphasis on the Random Forest, which also evaluates customer subscriptions, allowing the coverage of research gaps.

They are more such examples despite there being usually nobody who doubts that financial institutions can meet such challenges, adapt fintech innovations and deliver better solutions to their customers in the competitive market.

3 Methodology

In evaluating the effects of these innovations on the banking sector through analysis of customer acquisition, it is helpful to follow a set scientific process to get valuable results as well as findings. In data analysis, this research utilizes a data-science approach appropriate for marketing analytics and suitable for machine learning techniques for compliance with the data science process flows. Thus, the study is structured to make sure that from data collection to model assessment, each step tracks down the objectives of the study proposed. The use of such an approach not only enhances the study of customer acquisition trends but also gives direction to marketers regarding enhancing the marketing mix (Haddad & Hornuf, 2019).

3.1 Data Acquisition

Data is collected in the form of two datasets comprising records of bank marketing campaign data pertinent to the analysis of customer acquisition behavior. Both consist of several attributes of a customer including his or her age, gender, the result of previous campaigns, and contact information.

The datasets are described as follows:

Dataset 1: This set of records, obtained from the file bank.csv, includes customer data and previous encounters when promoting products. Some important attributes are the customer's age, job, marital status, education level and number of contacts and the time spent over each contact in the campaign. (Moro et al., 2014).

age	balance	day	duration	campaign	job_service	……...	marital_marrie	education_secondary	poutcome_success	deposit_yes
59	2343	5	1042	1	FALSE		TRUE	TRUE	FALSE	TRUE
56	45	5	1467	1	FALSE		TRUE	TRUE	FALSE	TRUE
41	1270	5	1389	1	FALSE		TRUE	TRUE	FALSE	TRUE
55	2476	5	579	1	TRUE		TRUE	TRUE	FALSE	TRUE

Figure 2 Key Variable from Dataset 1

Dataset 2: Bank-addition-full.csv dataset used for this work is an enhanced dataset to the basic one used for the previous experiments, the features introduced include economic factors and some basic data on the campaign. This results in a much richer dataset that can inform the firm of factors that may influence customers' subscriptions.

age	job	marital	education	housing	loan	contact	…..	month	day_of_week	duration	campaign	poutcome	emp.var.rate	euribor3m	y
56	housemaid	married	basic.4y	no	no	telephone		may	mon	261	1	nonexistent	1.1	4.857	no
57	services	married	high.school	no	no	telephone		may	mon	149	1	nonexistent	1.1	4.857	no
37	services	married	high.school	yes	no	telephone		may	mon	226	1	nonexistent	1.1	4.857	no
40	admin.	married	basic.6y	no	no	telephone		may	mon	151	1	nonexistent	1.1	4.857	no

Figure 3 Key variables from Dataset 2

As shown in Fig. 2 and Fig. 3 both datasets were read using the python-based environment for data analysis and manipulation known as pandas. First, the format of the data was studied, the main characteristics of the features were defined in words, and the frequency distribution of the target variable was analyzed, which is a binary variable that defines if a customer subscribed to the marketed financial product. To ensure that the data was valid for use, completeness and internal consistency were checked before going on with pre-processing and analysis. (Gamberger et al., 2017; Haddad & Hornuf, 2019).

3.2 Exploratory Data Analysis

A preliminary analysis was first carried out to get a general idea of the dataset and the variables that are contained in the database. An exploratory analysis was performed first and simple descriptive statistics for both quantitative and qualitative data were computed to give a general idea of the distribution, average and extremes of the data collected. The distribution of the target variable, deposit_yes, was then checked to test the balance between the customers that subscribe to the product and customers that do not. To illustrate this distribution and decide on training and test data splits, a bar chart was used together with the potential presence of class imbalance essential to model performance.

Figure 4 Distribution of Target Variable for Dataset 1

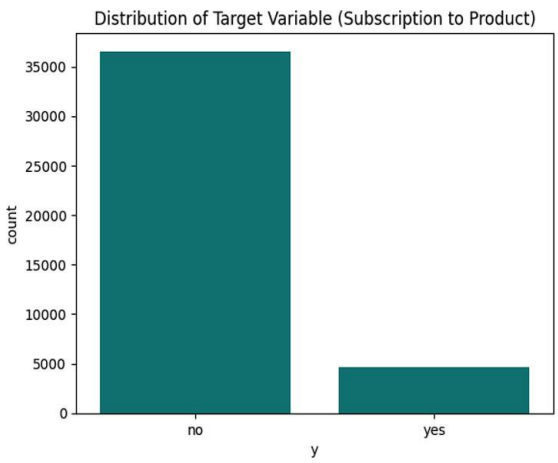


Figure 5 Distribution of target variable for Dataset 2

The bar chart in Fig 4 and 5 above shows the number of customers in the dataset based on the subscription (target variable deposit_yes). These samples allow one to judge the service take-up split quite fairly between the customers who subscribed to this service (True) and those who did not (False). The process results in less prejudice towards a hub class, which in turn provides better and fair probability forecasts for the clients.

To follow that, null values in the current dataset were checked by evaluating them. isnull(). sum() function. Completeness control was conducted to make certain that all records and features were complete, and to note which needed to be imputed or removed. Link this, box plots were done for numerical attributes like age, balance, duration, and the likes for outlier detection. Restrictions of values in the campaign column where extreme values were limited to 20 this was done to curtail on their effect when training the model. Histograms helped in extending the knowledge about the feature distributions especially in terms of the number of non-zero values for the ‘age’ and the distribution of values in the ‘balance’ feature.

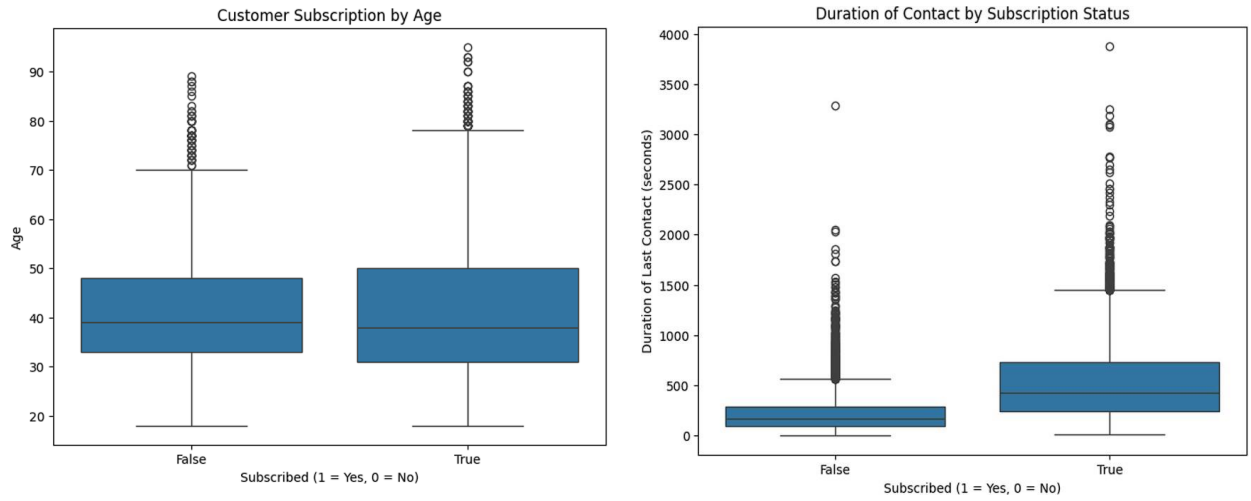


Figure 6 Outlier Detection for Dataset 1

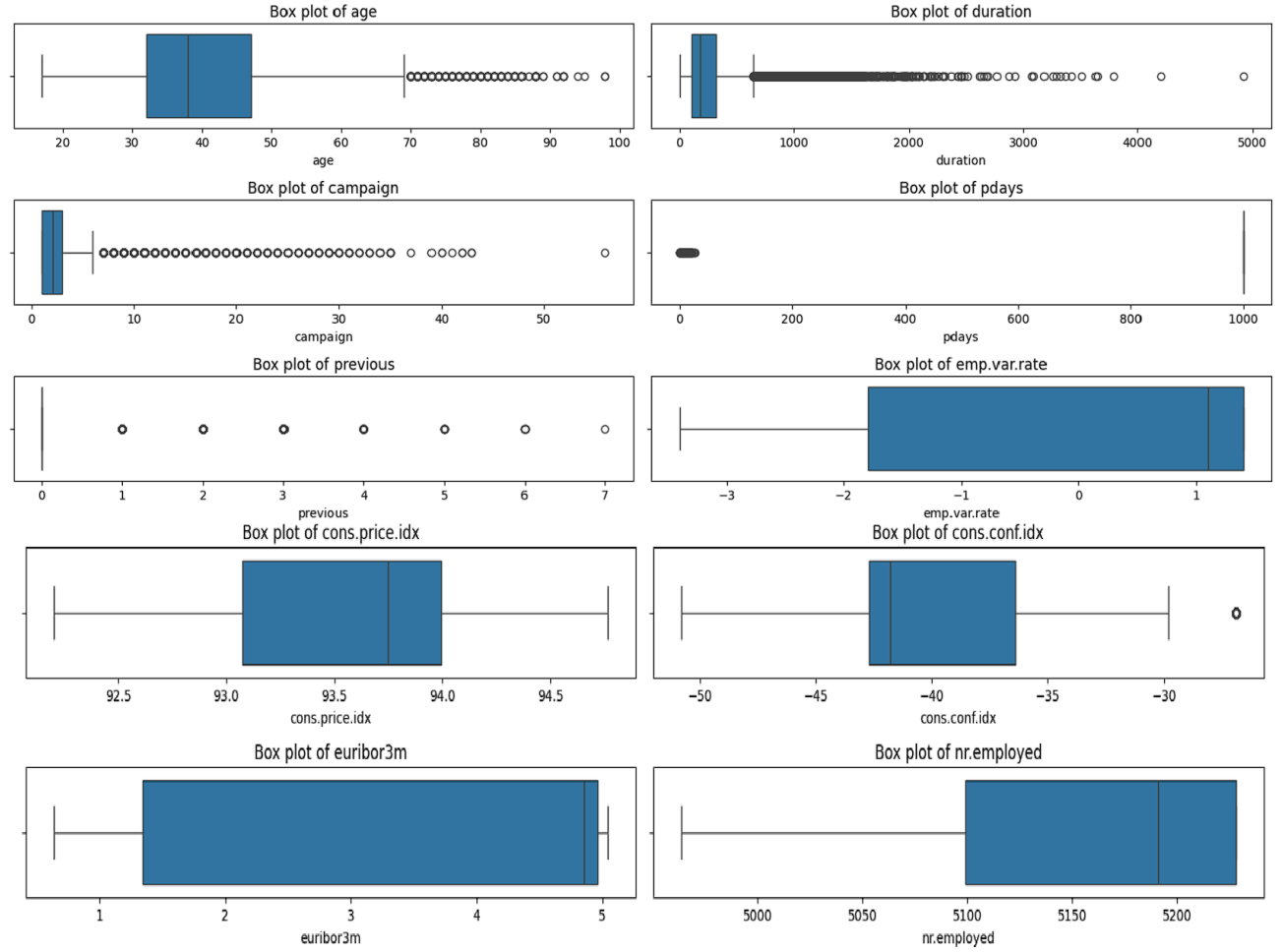
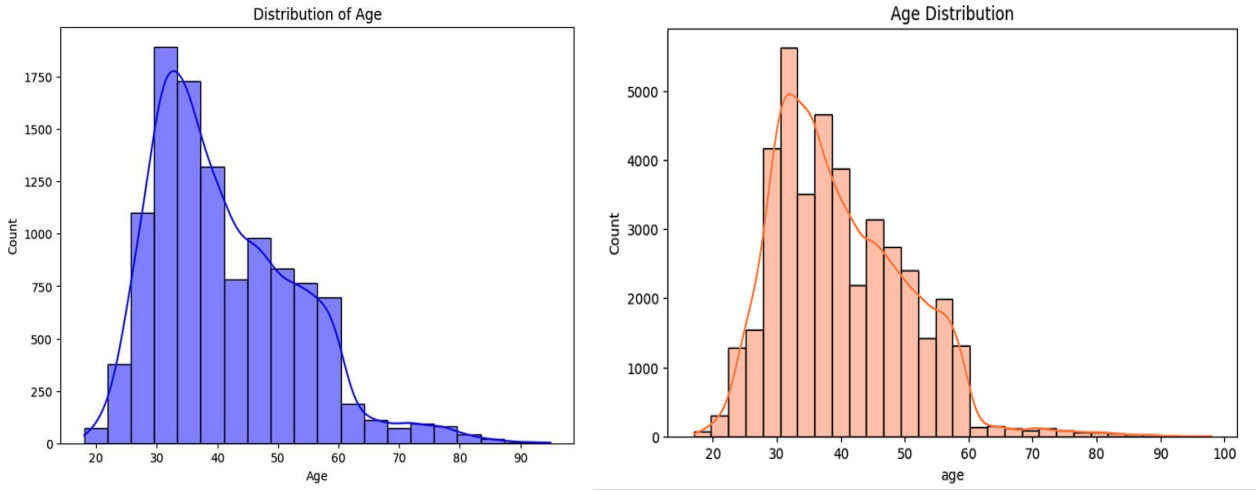


Figure 7 Outlier Detection for Dataset 2

Since aesthetic preferences in the layout of an interface are subjective, it was difficult to test relationships among features directly; therefore, a correlation analysis was performed following a correlation matrix to provide an insight into any possible, strong linear, and numerical variants interrelationship.



Dataset 1 Dataset 2

Figure 8 Distribution of Age from both Datasets

Further, the contact approach was used to identify the correlation of the customer engagement of the campaigns and rate of subscription. This was visualized by employing line plots to allow some understanding of the pattern of caressing and its impact on customer behavior.

Finally, job, marital, and education were checked to support the analysis of the distribution of the customer segments on categorical variables. Such distributions were represented by pie charts and bar plots.

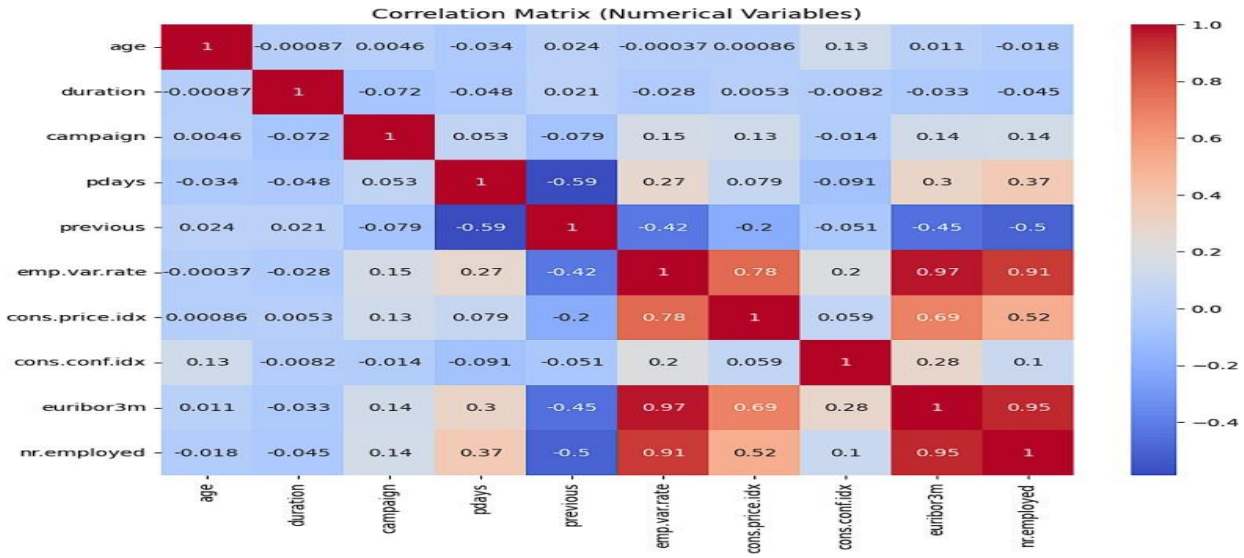


Figure 9 Correlation Heatmap

3.3 Data Preprocessing

The datasets were checked and prepared well in this data preprocessing step to meet their readiness to be analyzed. Before comparing the two datasets, the null hypothesis for both datasets were checked using the code. isnull(). sum(). There are no mostly missing values that were seen and hence no further action was taken for the imputations of records or for deletion of the records. The categorical variables in both data sets; job, marital, education and contact were encoded to

numerical representations using one hot encoding. To perform this encoding, the data was converted into binary columns for each category for compatibility with machine learning. For instance, variables including job_blue-collar, marital_married and education_secondary were created in Dataset 1 and similar transformations were done for the same applied to Dataset 2 for job and poutcome.

In Age, Balance, and Campaign figures, we used box plots to determine the outliers. In the campaign feature, where there were outliers of high values, data transformation through capping such that the highest limit of the feature was 20 was done. This step was useful in reducing extreme values' influence on model performance. Feature engineering was also used in the formation of new features. For instance, in the steps for Dataset 1, the pdays feature was discretized into a new feature was_contacted_before meaning the customer had been previously contacted. Likewise, emp.var.rate, euribor3m, and nr.employed in Dataset 2 were kept as they would offer more understanding about the marketing campaigns conducted.

Finally, the datasets were split into training and testing sets using an 80:20 split ratio to examine the efficiency of the Random Forest model in the analysis of the new data set. While for Random Forest models scaling is not mandatory, it was done to make sure that all the data was consistent in terms of format to improve interpretability. These preprocessing steps laid down good groundwork for the next modeling and analysis phase.

3.4 Model Implementation

Random Forest received selection due to its dual functionality with numerical and categorical data alongside strong resistance to overfitting while offering feature importance rank abilities. The ensemble learning technique constructs various decision trees to create final output predictions which optimizes handling of complex marketing datasets.

The hyperparameters tuned include:

1. n_estimators: The implementation used 100 Decision Trees for Dataset 1 followed by 200 Decision Trees for Dataset 2.
2. max_depth: Each dataset uses maximum tree depth settings of 10 or 15.
3. min_samples_split: The optimal tuning of min_samples_split parameter yielded 2 for Dataset 1 while adding five samples to split a node performed best for Dataset 2.

Performance metrics for the model included accuracy together with precision and recall and F1-score and AUC-ROC.

This model was then coded in python using the scikit-learn ritural to make prediction. The important hyperparameters were tuned to increase the accuracy of the model and improve its broad applicability. These were the number of decision trees that were learned (n_estimators), the maximum depth to which a single tree was grown (max_depth) as well as the minimum number of samples required to split a node (min_samples_split). All the entered hyperparameters for the two data sets were tuned using the grid search cross-validation method.

The table below in Fig 10 summarizes the hyperparameter configuration for both the datasets.

Hyperparameter	Dataset 1 Value	Dataset 2 Value
Number of Trees (n_estimators)	100	200
Maximum Depth (max_depth)	10	15
Minimum Samples Split (min_samples_split)	2	2

Figure 10 Hyperparameter Configurations

The datasets were split into training and testing sets using an 80:20 ratio. The subjects used in training are intentionally named the training set, and in the same way, the subjects used in testing are intentionally named the testing set. Accuracy, precision, recall, F1-score, and AUC-ROC were used to compare all the models. These metrics were used to offer a strong evaluation of the model's capacity, to classify the new subscription of customers.

4 Design Specification

The present study follows a systematic and empirical approach to investigating customer acquisition behavior in the banking industry based on machine learning approaches. The research makes use of two associated databases of data on banking marketing advertising campaigns. Dataset 1 contains basic variables of customers and overall features of campaigns, while Dataset 2, as well as Dataset 1, economic indicators are included, which further enrich the analysis. The target variable of both datasets represents if the customer purchased a marketed product based on deposit_yes for Dataset 1 and y for Dataset 2. Dataset 1 has around 11,000 records, while Dataset 2 contains over 41,000 records which are far richer in terms of their assessment.

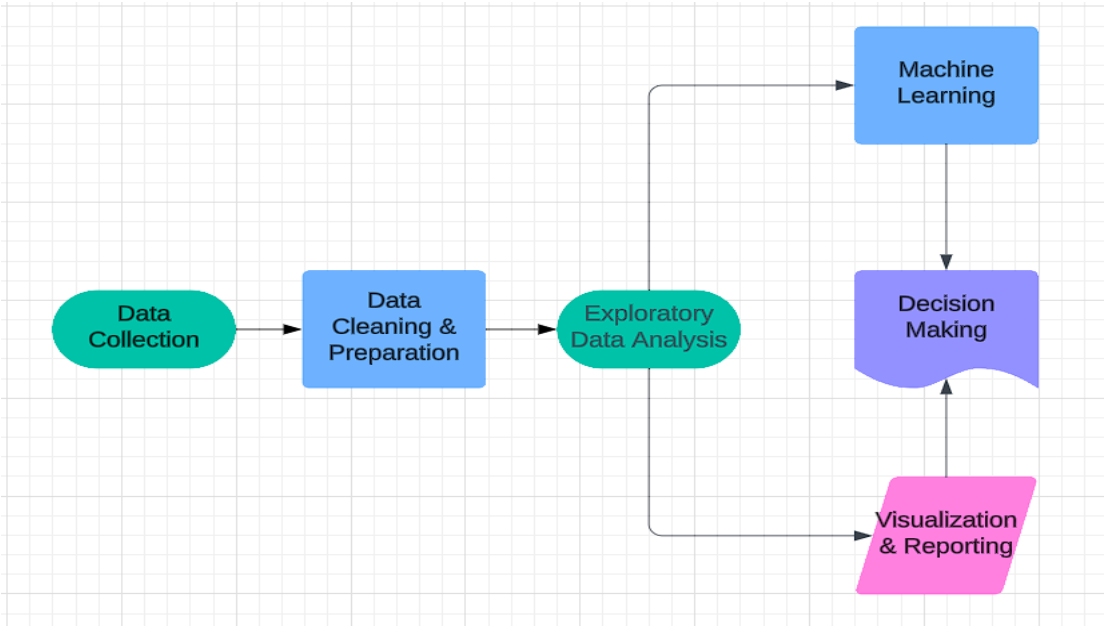


Figure 11 Workflow Diagram

As a data preparation step, missing data and categorical data were dealt with by imputing missing data, and variable named job, marital, and education were encoded by one hot encoding, and extreme outliers in features named campaign were clogged. Post preprocessing steps, training and

testing sets were further divided as an 80% training set and 20% test set to test the generalization of our model on unseen data. Random Forest algorithm was employed based to shear capacity to work with structured data, ability to reduce model overfitting and lastly for its feature importance ranking capabilities. The possible hyperparameters such as n_estimators, max_depth, and min_samples_split, were tuned to improve the model performance where the tune parameters include the number of learning cycles and depth of the tree and the required minimum samples for splitting nodes.

The performance of the proposed model was evaluated with various evaluation measures such as accuracy, precision, recall, F1 measure and AUC-ROC values. The confusion matrix generated a clear distinction between true positives, true negatives, false positives and false negatives unlike reporting metrics in feature importance analysis, it was possible to understand the factors affecting customer acquisition most importantly the duration and poutcome_success. Pandas and NumPy were used for data processing, scikit-learn for modelling, while matplotlib and seaborn were used for data visualization.

The operation was highly sequential with operations such as data collection and preparation, data visualization and exploration, and finally the collection and assessment of model results. This systematically saved a lot of time, and in addition, it made things clearer and easily repeatable. The study considers the datasets as having realistic representation of customer behaviors and suggests limitations, such as skewed distribution of the subscriber class in the dataset. The expected outcomes are training of the Random Forest model for customer subscriptions and development of recommendations for campaign planning for banking institutions, among others.

5 Implementation

Random forest algorithms were selected for this work because they are highly effective in structured datasets and can handle both numerical and categorical data. Random Forest is a meta learner model created during the learning phase with the number of decisions trees and then uses the results for prediction. Second, this approach not only improves the accuracy of the classification but also reduces the problem of overfitting that may occur in the framework of a single decision tree (Breiman, 2001). Due to its simplicity and stability, this model can be quite helpful in examining the customer acquisition behaviour in banking campaigns because it can offer a measure of the significance of each feature besides offering a visual representation of the results (Moro et al., 2014).

5.1 Implementation Details

The implementation was done in Python, and in the process, the scikit-learn was used for learning and other relevant tools in the context of data manipulation as well as visualization were employed as well. To do this, two datasets were used separately; the first one is the bank.csv (Dataset 1) while the second is bank-additional-full.csv (Dataset 2). Even though both datasets were relatively clean, preprocessing was performed as described in section ‘Data preprocessing’ to reduce variability and prepare the data for use in modeling and evaluation.

To evaluate the Random Forest model's performance, the datasets were split into training and testing subsets using an 80:20 ratio. This made sure that while the model’s training was done on a sufficiently big part of the data, there was another different set of data sets that had no interference with the training process and was used in the testing process freely (Moro et al., 2014). Target

variables were named as deposit_yes for the first dataset and y for the second dataset if the customers subscribed to the marketed product.

The parameters of Random Forest model were also optimized for better performance, these included:

- n_estimators: The number of deciding trees in the forest.
- max_depth: The maximum depth of each tree as used in determining model complexity.
- min_samples_split: The critical size of the population must be split at an internal node to do it fairly and grow the tree fairly.

Consequently, for the choice of model hyperparameters in Dataset 1, it was found that the values as n_estimators = 100, max_depth = 10, min_samples_split = 2 are the most suitable ones. For Dataset 2, selected in the experiment with extra economic features, the model needed a higher level of non-linear solutions with n_estimators=200, max_depth 15, min_samples_split=5. During model building, hyperparameters were set by performing feature selection by applying Grid Search Cross-Validation into the process, which is a process of varying the features to be input in the model to find out which will input the best results (Breiman, 2001).

5.2 Training and Prediction

In training, Random Forest constructed several decision trees, where each is a model of bootstrap sample of data and randomly selects the features for splitting. This helped to make the trees diverse thus producing a general model. For classification, the output of each tree was summed and averaged accurately to the nearest integer using majority vote. This aggregation prevented overfitting; a problem observed in high-complexity procedures.

The training process also retained the importance of each input variable: it described how the clinician's feature contributed to the creation of a pure forest, meaning how much it helped to remove impurity. The above scores were crucial in offering value by giving out the most influential aspects relating to customer acquisitions.

5.3 Libraries and Tools

Several Python libraries were utilized to streamline implementation:

- Pandas: Especially for loading, cleaning, and transforming processes of the datasets.
- Scikit-learn: This amongst others is because the following processes involved are essential in model creation, training, hyperparameter tuning and evaluation.
- Matplotlib and seaborn: For creating graphics and charts that are used in feature importance issues and other performance evaluations.
- NumPy: During data preprocessing and analyzing process we need to perform numerical computation.

5.4 Analyzing the Importance of the Features

Indeed, one of the foremost advantages of Random Forest algorithm is the possibility to evaluate features' importance, or the possibility to rank them based on how much they contribute to the

model predictions. The most influential features included:

- Duration: Among these factors, the length of the last contact with a customer had the highest positive effect, indicating that the longer the interaction the easier to subscribe.
- Poutcome_success: Customer behavior shifted due to the success levels achieved in past marketing campaigns.
- Balance: Closely related to the type of subscription service, customers with high value accounts subscribed, an implication of financial status.
- Age: It was seen that some age markets showed a higher probability of responding to marketing communications.

Fig 9 proves that the duration has the highest influence in the prediction model decisions, as it takes up the biggest portion of the chart. This seems to corroborate other studies done on Telemarketing and customer interaction where the opportunity to effectively communicate with the customer increases customer contact success. Other variables such as past campaign success (poutcome_success) and customer factors (age, balance) are equally relevant in predicting marketing effectiveness.

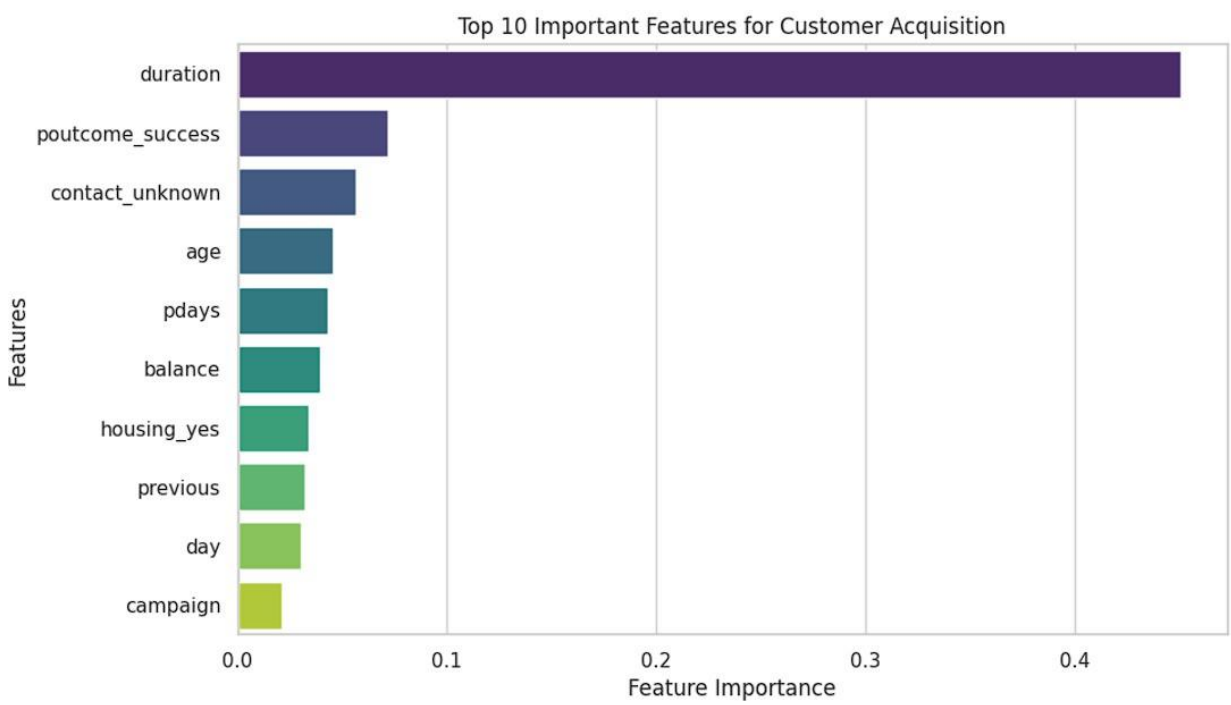


Figure 12 Feature Importance

In turn, this analysis offers useful suggestions to marketers as to which elements of the call namely the length of the call and the outcomes of previous interactions with the customer should be given focus, as well as demographic factors that should inform a more effective targeting approach.

6 Evaluation

In this part, the results of experiments carried out employing numerous machine learning algorithms, namely, Logistic Regression, Random Forest, and XGBoost are presented and

compared, both datasets, namely Dataset 1 and Dataset 2. In particular, the evaluation part of the research aims at revealing the differences between all the models introduced to determine the best model for the customer subscription prediction.

6.1 Experiment 1: Logistic Regression on Dataset 1 and dataset 2

To begin with, both datasets were analyzed using logistic regression. As the starting model for both datasets, the linear model logistic regression was chosen as the baseline. The results are as follows:

Dataset 1 Results:

- Accuracy: 80.7% Logistic Regression proved to be the best as far as the base model was concerned while producing high classification accuracy.
- Precision: 81.0% Illustrates that the false positive rate, one of the primary problems plaguing other models, is overcome by the present model.
- Recall: 77.9% Suggests that many real subscribers were accurately classified.
- F1-Score: 79.5% The accuracy and the recall are balanced, which demonstrates that the model retains good consistency.

Dataset 2 Results:

- Accuracy: 87%, Logistic Regression achieved a greater accuracy of Dataset 2 than the first one with the richness of dataset feature.
- Precision: 98% (Class 0), 45% (Class 1) For Class 0 which is the non-subscriber, the precision is extremely high primarily due to the suppression of false positive predictions and for Class 1 which is the subscriber, the precision is moderate.
- Recall: 86% (Class 0), 88% (Class 1) Subscribers are found to have high true cases for both classes and logistic Regression showed a high accuracy level.
- F1-Score: 92% (Class 0), 59% (Class 1) High F1-score proves the F-score of Class 0 to be balanced while Class 1 can be improved.

Analysis: Logistic Regression was accurate in both files, especially in recalling the subscribers. However, it is a linear model and cannot model relationship between data hence giving less precise for subscribers in Dataset 2.

6.2 Experiment 2: The Random Forest on Dataset 1 and dataset 2

The Random Forest model, which is built from decision trees, was used for analysis of both datasets with tuned hyperparameters. The results are as follows:

Dataset 1 Results:

- Accuracy: 83.4% As for comparison of Random Forest and Logistic Regression, the former provided better accuracy which proved Random Forest's capability to deal with intricate relationships.
- Precision: 80.9% Fewer instances of identifying alleles the wrong tissue, improving the accuracy of the predictions.
- Recall: 85.4% The high recall signifies that the model classifies most actual subscribers, meaning it got most of them right.
- F1-Score: 83.1% The chosen F1-score is well balanced as it shows that Random Forest is a

strong and stable model.

Dataset 2 Results:

- Accuracy: 91% The model demonstrated better efficiency in terms of the accuracy of records classification.
- Precision: 94% (Class 0), 59% (Class 1) For non-subscribers, Random Forest illustrates its capability to reduce false alarms regions with small error; for subscribers, the precision is reasonable but could be higher.
- Recall: 95% (Class 0), 56% (Class 1) Making high recall for subscribers helps in identifying most of the actual cases among non-subscribers.
- F1-Score: 95% (Class 0), 58% (Class 1) The F1-scores are given to represent the efficiency of the model showing how well it is recall and how well it is precise.

Analysis: Random Forest was superior to Logistic Regression in both datasets for accuracy, recall, and F1 Score. Its ensemble approach means it handles complications in the data, including imbalance in the data.

6.3 Experiment 3: XGBoost used on Dataset 1 and Dataset 2

For both the datasets, XGBoost gradient boosting algorithm introduced in this paper was used. The results are summarized below:

Dataset 1 Results:

- Accuracy: 84.4% For Dataset 1, XGBoost has revealed the highest accuracy among all the models implemented above.
- Precision: 81.7% More specifically, a reduction in false positives was achieved, which improved the overall accuracy when compared with more general models.
- Recall: 86.7% Appendix IV shows that XGBoost captured more true subscribers than other models entailing the highest recall.
- F1-Score: 84.1% In more detail, the obtained results included a significant value of the F1-score, which points to good accuracy, as well as an appropriate balance in the model.

Dataset 2 Results:

- Accuracy: 91% On the measures of accuracy, XGBoost had the same performance as Random Forest.
- Precision: 95% (Class 0), 61% (Class 1) Subscribers were only marginally better handled by the algorithm in terms of precision compared to Random Forest.
- Recall: 95% (Class 0), 58% (Class 1) It performs just as well in terms of recall as Random Forest.
- F1-Score: 95% (Class 0), 59% (Class 1) The F1-scores clearly indicate a small overperformance of XGBoost over Random Forest in subscribers' identification.

Analysis: XGBoost provided the best overall performance based on results for Dataset 1 but fared as Random Forest on Dataset 2. But XGBoost with increased computational complexity and tuning process is not as feasible for more typical use.

6.4 Discussion

Through the application of the Random Forest algorithm, the research was able to discover the relevant components affecting customer acquisition and analyse the applicability of marketing strategies in the two datasets.

Metric	Random Forest (Dataset 1)	Random Forest (Dataset 2)
Accuracy	83.40%	91.00%
Precision	80.90%	94.00%
Recall	85.40%	95.00%
F1-Score	83.10%	95.00%

Figure 13 Model results for Random Forest.

The performance strengths and shortcomings of models on Dataset 1 and Dataset 2 that the evaluation reveals are discussed. Among the models tested, Random Forest emerged as the most suitable choice for this study, due to the following reasons:

- Consistent Performance: Comparing the two sets of results, the two classifiers Random Forest and Support Vector Machine both showed high accuracy, precision, recall and F1-score measures.
- Robustness: This is because Random Forest is an ensemble learning algorithm that lends itself to greater stability and robustness against overfitting compared to Logistic Regression.
- Interpretability: While the XGBoost fails to generate feature importance, Random Forest has a beneficial feature amongst practitioners, which is the case of an interpretable feature importance.
- Efficiency: Even though XGBoost provided a similar performance as Random Forest, the latter is almost as accurate, and much simpler to implement, which makes it even more desirable due to its scalability.

The following factors make Random Forest suitable for customer subscription prediction. The recommended features include duration, poutcome_success and previous, thus lending credence to its use in the improvement of marketing strategy.

7 Conclusion and Future work

This study focused on the use of machine learning algorithms in the examination of customer acquisition data using the Random Forest algorithm in the banking industry. The current study was able to provide evidence of how machine learning models can be used to derive insights by leveraging two datasets that include multiple aspects of the customer and other aspects related to the campaign.

Thus, the evaluations reveal that Random Forest is accurate and not sensitive to the missing values of the customer data. In the analysis of Dataset 1, it was found that the duration, poutcome_success, and contact_unknown characteristics are indicators that affect the future behavior of customers regarding subscription, meaning that the campaign should be more extensive and previously received outcomes are significant. The overall test accuracy of our model was 83.4% and it provided good precision, recall, and F1, and the degree of association between customer responses and the key patterns was reliable.

Likewise, Random Forest, in Dataset 2, with acclamation notched a formidable 91% accuracy that fused the potency of Machine learning in NOTES to handle complex datasets. The classification report highlighted that the pattern extraction on customer acquisition was good and thus could be useful in decision making. These insights fully support the need to include data processing into the modern marketing arsenal to increase its efficiency and of organizations'

activities.

This research provides new insights for the usage of machine learning in banking and marketing analysis fields. It outlines a formal process for managing data science processes which include data gathering, storage, analysis, modeling and assessment as well as showing how businesses can utilize specific algorithms to leverage themselves in the acquisition of customers.

While this research has laid a solid foundation, there are several avenues for future exploration and enhancement:

a. Incorporating Advanced Techniques: Further investigations in future works can be used of more sophisticated deep learning models together with a more elaborate analysis of the models, including Gradient Boosting or XGBoost models. They might help find less obvious patterns of customers' operations and performance.

b. Addressing Class Imbalance: Even though we positively evaluated Random Forest model in our research, further studies could focus on ways of dealing with the problem of imbalanced datasets as SMOTE or cost-sensitive approaches which might increase the recall for the minority class.

c. Feature Engineering Enhancements: Other features depending on business domain can also be derived and introduced to allow more detailed information about customers. For example, use of both time-series data along with such static attribute might provide better demands and attributes for an analysis.

d. Real-Time Applications: Applying the current model to real-time customer targeting and segmentation could help in making real-time choices in the case of marketing campaigns. Exactly these models could be implemented into CRM customer relationship management systems to provide real-time recommendation as well as personalized marketing.

e. Comparative Analysis: When comparing multiple machine learning algorithms, in addition to performing this study, statistical measures can also be used to offer further cognizance of the model complexities and their interpretability against performance emphasis.

f. Ethical and Regulatory Considerations: Based on the research findings and conclusion, the following recommendations are made: Future research should take sensitive issues of ethical and privacy aspect of customers analysis data especially in the banking industry into consideration. It's crucial to keep all the regulations relating to protecting the data, like GDPR or CCPA, in check to keep the audience trust in check.

g. Cross-Industry Applications: Because this study was conducted on the banking sector, future studies can use the same method and findings to test its applicability on other sectors such as the retail or telecommunications sector.

h. Integration with Economic Indicators: To enrich the analysis, integration of the macroeconomic factors, for instance, from the current inflation rate to employment position could enable the application of the model in other conditions and phase of the business cycle.

Lastly, this research proves that machine learning is a robust technique for credit unions and other organizations to understand customer acquisition and perform advanced analysis to improve

their marketing strategies. Many other possibilities remain untouched with increased development of modeling procedures, better data management techniques, data integration, etc.; this field is full of potential for businesses as well as researchers.

8 Acknowledgements

I am grateful to my supervisor Dr. Abid Yakoob for the efforts and support he has given me for the entire process of my guiding. Great advice and assistance in guiding me through the research. Continuous support by him enhances the quality of the research experiment.

References

ALT, R., BECK, R. & SMITS, M. T. (2018), 'FINTECH AND THE TRANSFORMATION OF THE FINANCIAL INDUSTRY', ELECTRONIC MARKETS 28(3),235–243.
URL: HTTPS://LINK.SPRINGER.COM/ARTICLE/10.1007/S12525-018-0310-9

Gomber, P., Koch, J.-A. & Siering, M. (2017), 'Digital Finance and FinTech: current research and future research directions', Journal of Business Economics 87(5), 537–580.
URL: https://link.springer.com/article/10.1007/s11573-017-0852-x

Haddad, C. & Hornuf, L. (2019), 'The emergence of the global fintech market: economic and technological determinants', Small Business Economics 53(1), 81–105.
URL: https://link.springer.com/article/10.1007/s11187-018-9991-x

Kou, G., Xu, Y., Peng, Y. & Shen, F. (2021), 'Machine learning-based risk management in fintech: a survey', Financial Innovation 7(1), 1–27.

Lee, I. & Shin, Y. J. (2018), 'Fintech: Ecosystem, business models, investment decisions, and challenges', Business Horizons 61(1), 35–46.

Ryu, S.-O., Lee, H.-J., & Kim, S.-H. (2020). "The Impact of Fintech on Customer Experience in Banking: Evidence from a Survey Study." *Journal of Retailing and Consumer Services*, 53.

Dahlberg, T., Guo, J., & Norrman, C. (2015). "A Critical Review of the Research on Mobile Payment Systems." *Electronic Commerce Research and Applications*, 14(5), 265-284.

Frost, J., & Gambacorta, L. (2020). "The Rise of Fintech: The Impact on Banking and Financial Stability." *BIS Quarterly Review*, March 2020.

Buchak, G., Matvos, G., Piskorski, T., & Seru, A. (2018). "Fintech, Regulatory Arbitrage, and the Rise of Shadow Banks." *Brookings Papers on Economic Activity*, Spring 2018.

Möhlmann, M., & Ziegler, R. (2020). "The Role of Fintech in Customer Acquisition Strategies in Retail Banking." *International Journal of Bank Marketing*, 38(3), 635-650.

Carlin, B. I., & Olafsson, A. (2021). "FinTech and Consumer Financial Well-Being." Journal of Finance, 76(3), 1125-1160.
URL: https://doi.org/10.1111/jofi.12964

Eickhoff, M., Muntermann, J., & Weinrich, T. (2017). "What Do FinTechs Actually Do? A Taxonomy of FinTech Business Models." Proceedings of the International Conference on Information Systems (ICIS).

Milian, E. Z., Spinola, M. D. M., & Carvalho, M. M. (2019). "Fintechs: A Literature Review and Research Agenda." Electronic Commerce Research and Applications, 34, 100833.
URL: https://doi.org/10.1016/j.elerap.2019.100833

Zavolokina, L., Dolata, M., & Schwabe, G. (2017). "FinTech – What's in a Name?" Thirty Eighth International Conference on Information Systems.