Journal of Techniques, ISSN: 2708-8383, Vol. 5, No. 2, 2023, Pages 163-186
DOI: https://doi.org/10.51173/jt.v5i2.1226

JOURNAL OF TECHNIQUES

Journal homepage: http://journal.mtu.edu.iq

RESEARCH ARTICLE - MANAGEMENT

Binary Classification of Customer's Online Purchasing Behavior Using Machine Learning

Ahmad Aldelemy[1*], Raed A. Abd-Alhameed[1]

[1] Faculty of Engineering and Informatics, University of Bradford, Bradford, BD7 1DP, United Kingdom

[*] Corresponding author E-mail: a.a.aldelemy@bradford.ac.uk

Article Info.

Article history:

Received
31 January 2023

Accepted
26 April 2023

Publishing
30 June 2023

Abstract

The UK financial sector increasingly employs machine learning techniques to enhance revenue and understand customer behaviour. In this study, we develop a machine learning workflow for high classification accuracy and improved prediction confidence using a binary classification approach on a publicly available dataset from a Portuguese financial institution as a proof of concept. Our methodology includes data analysis, transformation, training, and testing machine learning classifiers such as Naïve Bayes, Decision Trees, Random Forests, Support Vector Machines, Logistic Regression, Artificial Neural Networks, AdaBoost, and Gradient Descent. We use stratified k-folding (k=5) cross-validation and assemble the top-performing classifiers into a decision-making committee, resulting in over 92% accuracy with two-thirds voting confidence. The workflow is simple, adaptable, and suitable for UK banks, demonstrating the potential for practical implementation and data privacy. Future work will extend our approach to UK banks, reformulate the problem as a multi-class classification, and introduce pre-training automated steps for data analysis and transformation.

Publisher: Middle Technical University

Keywords: Machine Learning; Customer Behaviour; Classification Algorithms; Banking Industry; Classifiers.

1. Introduction

According to Statistica[1], the revenue of the e-commerce market in the United Kingdom is projected to reach £80,678 million by the end of 2021. Another report from E-commerce Foundation[2] shows that e-commerce sales make up 7.94 % of the UK's gross domestic product. These facts highlight the importance of improving online business in all its forms, including business-to-business models and business-to-customer. In addition, The Big-Data era now has the power of advanced analytics, which helps businesses and customers to make more informed decisions.

Banks are continuously being observed taking full advantage of information technology worldwide. HSBC Holdings, Lloyds Banking Group, Royal Bank of Scotland Group, and Barclays are a few names from UK financial market that have always equipped their-selves with state-of-the-art security, privacy, and seamless transactional technology. Online product and service catalogues are one of the top visiting and time-spending domains for customers. Customer profiles, call centre communication, and web surfing metadata are most suitable for understanding customer behaviour. Colossal interest can be observed in this area over the past two decades because of advancements in computational resources and artificial intelligence (AI).

In this research, our focus is to step up the understanding of bank customers' online purchasing behaviour. Banks offer multiple products, and each can either be analysed separately or in a combined setting. However, we tend only to analyse one product, "Term-Deposits". The goal is to model customer behaviour using a machine learning model. The model will be developed in the context of existing real-world data, whose sanity will be testable.

Data plays a vital role in analytics. It must be closer to the domain of interest but generic enough to capture the most known scenarios. Our selected dataset is taken from the UCI dataset and captures direct marketing campaigns of a Portuguese banking institution. To the best of our knowledge, there is no publicly available dataset from any UK institutions; therefore, we are adapting this dataset as the first version of the model. Using our framework, UK institutions can train their model just by replacing the dataset.

This study aims to analyse and design a robust machine learning strategy that assists in determining the most favourable and potential subscriber of a term deposit from bank offerings. From a machine learning perspective, we formulate this task as a binary classification problem in a supervised learning setting. In the analysis phase, various machine learning classifiers will be trained on the given dataset to evaluate the ranking for best classifiers; This ranking lets us form a committee of top performers so that more robust classification decisions be made with higher confidence.

Nomenclature & Symbols			
AI	artificial intelligence	ROI	return on investment
SVM	support vector machines	FA	Factor Analysis
LDA	Linear Discriminant Analysis	RFB	Radian Basis Function
AUC	area under the curve	MLPNN	Multilayer Perceptron Neural Network

This research includes classifiers: Naïve Bayes, Decision Tree, Random forests, SVM, Logistic Regression, Artificial Neural Networks, AdaBoost, and Gradient Descent. The finalised machine learning strategy can be offered as an inferencing web service for practical usage. That will eventually contribute to helping UK banks to re-train their version of the model on their data.

Our core programming language will be Python. In addition, the most widely used libraries like Pandas, NumPy, and sci-kit-learn will be used for Data cleaning, transforming, loading, and training. Furthermore, we will execute our experiment using Python's Orange3 library for rapid prototyping, offering a sci-kit learn wrapper. Finally, the Flask library must expose it as a service to develop and offer the inferencing service of trained models.

Next, in this chapter, we define the aim and objectives of this research and declare our research scope. In the next chapter, we present a related literature review. The focus of our review is to provide a base for our work by identifying the gaps in current practices and focusing on overcoming those gaps. The Research methodology will be provided for achieving our research objectives. As per the research plan, we opt to experiment and discuss our findings. Lastly, we analyse, interpret, discuss our results and set future directions.

1.1. Study problem

1.1.1. AIM

This research aims to design a re-trainable machine learning model that helps banks understand customers' behaviour toward a particular product (specifically: Term Deposit subscription). In order to focus on solving the underlying problem, we have narrowed down our area of research. From a technological perspective, we aim to solve the problem using the supervised learning sub-domain of machine learning, a domain of artificial intelligence. Artificial intelligence emulates natural intelligence to solve real-world problems. Likewise, in machine learning, we tend to generate intelligence as a computational model by training over data (Fig. 1).

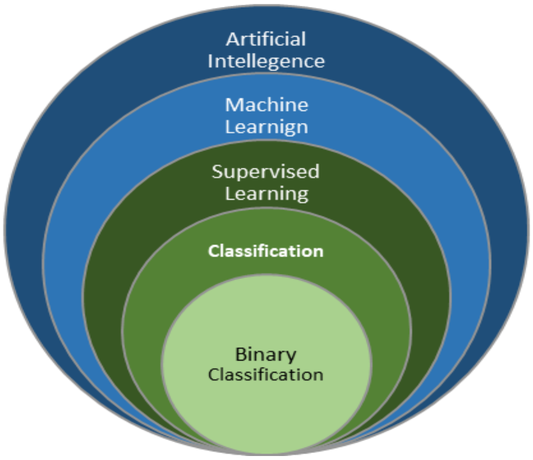


Fig. 1. Work Area Specification

1.1.2. Objectives

The problem this study seeks to address is the difficulty in understanding and predicting customer behaviour in the banking industry, particularly in the context of product offerings. The study focuses on identifying the most effective machine learning techniques that can be used to predict customer inclination towards subscribing to a given bank product based on historical data. Furthermore, the study aims to develop a workflow and a web service that UK banks can easily adopt to enhance their decision-making process and improve their product offerings based on customer preferences.

The study's main objectives are:

- To design and implement a workflow incorporating data analysis, transformation, and classification using machine-learning techniques.
- To train and test a set of candidate classifiers on a given dataset to identify the top-performing algorithms based on their accuracy and reliability.
- To develop a web service that banks can use to predict customer inclination towards a specific bank product, using the top-performing classifiers as decision-makers.
- To make the code and implementation available as an open-source project, allowing UK banks to experiment and adapt the workflow to their specific use cases.

1.2. Research application

The application of this research can be projected in the marketing department of the banking sector. It will help them to understand and predict customer behaviour toward a given offering. Such learning can further help them design more segments within their products that are more reachable to a wide range of customers and maximise sales and profits.

1.3. Research contribution

We aimed to set our research contributions as follows:

- Analysis of selected dataset: This includes a graphical illustration of target customers' different aspects like age, education, financial health, and other attributes. We will also explore correlations between their attributes.
- Performance evaluation of existing machine-learning algorithms: We perform and demonstrate the feasibility of well-known machine-learning models on the same dataset. Their comparison will be presented along with a discussion of how they perform. That lets us know the top-performing models that will be used for further study.
- Design of Ensemble Model for binary classification: The top-performing classifiers are tagged as decision-makers for predictions in the production environment. For any new instance, their outcomes will be considered votes, and most votes decide the outcome.

2. Literature Review

Artificial intelligence is gaining popularity in all finance, trade, manufacturing, and other business branches. The main reasons behind this era are the availability of computational resources, big-data availability, and the discovery of complex yet powerful algorithmic approaches. Moreover, the availability of powerful tools has empowered software developers and analysts to offer data science services with few lines of code on the fast track. As a result, quick and robust results increase business owners' confidence to expose their data to machine learning and minimise business risks while maximising return on investment (ROI).

The exact impact of AI technologies is expected in the financial services sector in UK Banks. That will redefine how banks work (their processes), what they shall be selling (their products and services) and eventually, how they interact with their customers (their user experiences). A survey of more than 100 respondents discovered that around two-thirds of financial services companies in the United Kingdom use Machine Learning. According to the Bank of England, Machine Learning has a wide range of applications in financial services. Combined with computational power, these offer extensive datasets analysis, detect patterns, and quickly solve complex problems [3]. One such sub-domain has been comprehended as finding patterns in customer behaviour.

Understanding customer behaviour plays a crucial role in boosting revenue for banks. That helps them design their product according to market demand. Traditionally, banks focused on proposing various policies to explore the potentiality of a customer based on statistical analysis [4]. Some have aimed to use data-driven marketing tools like natural language processing and machine learning models to understand demographics better [5]. A growing trend in this respect can be found in big data analytics and data mining. For example, a model has been presented by [6] for analysing clickstreams from online customers. They have proposed a data mining model that predicts whether a given customer will add an item to the shopping card. Likewise, [7] presented predictive analytics to get deeper insights into customer behaviour using behaviour informatics.

While many researchers focused on data analytics, another technological group found alternatives using the Internet of Things and smart tags. For instance, [8] proposed a framework for understanding customers' purchase behaviour. They proposed using statistical learning theory to deal with mounted RFID data. A similar approach can be observed in [9], where a self-organised IoT system has been proposed for online shopping. In another study [10], extracting important topics from Skype customer feedback sources has been discussed to measure the emotions of the associated topic using the Vibe metric. Finally, a web mining-based classification has been presented by [6], demonstrating the mining of data to get customer behaviour from e-commerce portals.

In addition to the efforts stated above, a reasonable amount of literature can be found on understanding consumer behaviour through machine learning approaches [11]. To further explore the given subject, in this research, we focus on using machine learning classifiers more reliably from a performance perspective. That led us to the quest to examine approaches and techniques that contribute to understanding the gaps and help us design our proposed solution. We found that subject comprehension can be achieved in both unsupervised and supervised learning settings to achieve our purpose. Classification can be formed as a supervised learning task and put simpler for reliability, and we opted to drill it further with binary classification. In this section, we present some highlights from existing literature that have addressed issues and challenges of classifiers in contesting the same or similar data sets.

From a technical viewpoint, the literature has focused on the Supervised Learning based Binary Classification task. Since machine learning focuses on teaching computers to learn from data and supervised learning, we used the data to have labels. This means that our training data will consist of input data along with actual class and observed behaviour. We found a common pattern of subtasks with such datasets and classification tasks. These subtasks mostly start with data analysis, cleansing, pre-processing, and classification.

2.1. Pre-Processing

Data quality is one of the most critical challenges of all data analytics and decision-making [12, 13]. Unfortunately, the real-world datasets are mostly incomplete and inconsistent, likely containing many issues. Therefore, considerable effort is required to transform the noisy data into a clean pre-processing dataset. That involves steps like statistical analysis, spelling correction, class imbalance, dimensionality reduction and further feature engineering (Fig. 2).

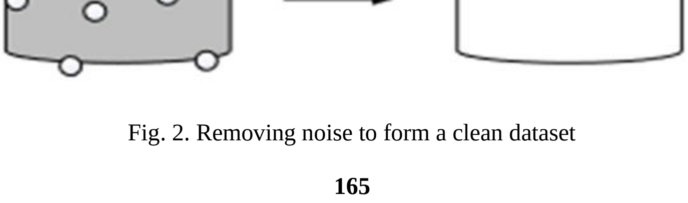


Fig. 2. Removing noise to form a clean dataset

The most widely used dimensionality reduction techniques are Principal Component Analysis [14], Factor Analysis (FA), and Linear Discriminant Analysis (LDA). They can be used with classification algorithms as a pre-processor. For instance, the performance of existing classifiers was enhanced using dimensionality reduction techniques [15]. Their proposed method, called CfsSubsetEval, evaluates the subset of features based on the predictive ability and finds the duplication among the selected features. They then applied classification algorithms like Naive Bayes, J48, KNN and Bayes-net. Before reducing dimensionality, J48 provide 89%. Of accuracy while gaining an accuracy of 91.2% when using CfsSubsetEval. This approach works better because the proposed technique eliminates the features that are not contributing to good results. However, this requires additional skills and experience in analysis and decision-making.

A hybrid classification approach has been presented that combines multiple classifiers to form an ensemble learner [16]. The goal was to improve classification outcomes. Two datasets have been used in their experiment, i.e., the German credit dataset and the Turkish bank dataset. This approach also used a feature selection stage. SVM-Radian Basis Function (RFB) has been used as a base classifier. The results of their hybrid classification approach showed better results. However, their proposed approach depends on the proper selection of features and feature selection strategy.

In another work by [16], an effort has been put into outlier analysis to build a good classifier. Based on their comparative study on multiple classifier accuracies, they found that in-frequent data decreases classifier performance. Therefore, it was argued that frequent high records are most helpful, and the infrequent records are obstacles in modelling the classifiers. They executed the experiments on the Bank dataset and Nursery datasets. Their efforts have empirically highlighted that those analytical efforts must be engaged to understand and deal with undesired data.

Some researchers [14] used PCA to reduce dimensionality and then applied classification models, such as AdaBoost, Gradient Boosting, SVM RBF, Naive Bayes, and Random Forest. That improved their performance, but their work introduces dimensionality reduction steps as overhead. The use of novel and adorned classifiers has been advocated by [17]. They have used kernels SVM, which uses the marginal hyperplane to determine the classes uniquely. Their technique, called KPCA (Kernel Principal Component Analysis), is an extension of PCA. They were used to classify and detect anomalies by transforming input space into high-dimensional feature space. The performance of SVM and KPCA jointly have been compared with non-kernel techniques and shown better results.

Our selected dataset is available in a more refined format. It contains no null values in its features and class labels. Since it contains a much smaller number of attributes, and the correlation of attributes is much smaller; therefore, further dimensionality reduction has not been applied.

2.2. Classifiers

The primary stakeholder of our work is the careful analysis and selection of a classifier that offers better performance and reliability in its outcome. For this purpose, we selected several classifiers from existing machine learning literature based on their popularity. These classifiers include Naïve Bayes, Decision Tree, Logistic Regression, Support Vector Machine, Artificial Neural Network, Random Forest, Gradient Boosting and AdaBoost. The detail of each classifier can be seen in the next section.

The selection of an appropriate classification algorithm with optimised parameters has been discussed by many researchers [18-20]. A comparative study of ten classifiers is presented in [18]. They evaluated their framework on nine datasets. Their accuracy indicator, i.e., the area under the curve (AUC), highlighted logistic regression as best classified. Naive Bayes, neural network, support vector machine classifiers as runner up while decision tree-based classifiers tend to underperform. In a similar study[21], four classifiers, Multilayer Perceptron Neural Network (MLPNN), Decision Tree (C4.5), Logistic Regression and Random Forest, where experimented with. They achieve 87% accuracy for Random Forest Classifier. In [22], five algorithms have been evaluated, i.e., Naive Bayes, Decision Trees, Artificial Neural Networks and Support Vector Machines. Their proposed approach from f-measurement achieved 60.12%. Similarly, [19] has inspected the performance of the Support Vector Machine, Naive Bayes, Logistic Regression and K-Nearest Neighbor on the credit card fraud dataset. Their experiment favoured logistic regression, Naive Bayes, k-nearest neighbour and support vector machine, respectively. Another team [20] performed a similar comparative analysis and found the KNN algorithm, Naïve Bayes algorithm and Decision Trees classifier. Another such effort can be found in [23], where multiple classifiers were examined to filter spam messages. They found neural networks best among Naïve Bayes and Support Vector Machine. Moreover, [24] has demonstrated the use of multiple algorithms on the same dataset, including Logistic Regression, a decision tree, and a support vector machine. From a classification accuracy perspective, they achieved 90.8% for logistic regression.

Other than comparative analysis, we observed single classifier-focused literature that either developed a variant of the existing algorithm or optimised the parameter for their case study. For example, [25] proposed a real-time online shopper behaviour analysis system to predict visitors' shopping intent. First, pageview data was collected and then fed to Multilayer Perceptron as input, then trained with various parameters like the number of layers and neurons. They also monitor the performance of a few available activation functions. Lastly, the top performer was selected as the final classification model to determine site abandonment likelihood. This approach works best if the selected classifier drives on fewer parameters. However, in the case of MLP, the parameters to be tested are much more significant. Therefore, much more time and resources would be needed to draw a more optimum model.

We strive to conduct a more comprehensive study that caters to more insights into classifiers based on the literature above. Similarly, we also extend our performance analysis to all indicators, including Area under Curve, Classification Accuracy, F1 Score, and Precision. Furthermore, we also observed that from these comparative analyses, one could not conclude the best classifier in all scenarios because of the high variations in performance. Moreover, it has been observed that a single classifier result is not much reliable in all cases. Therefore, we see much higher potential in ensemble learning, where multiple algorithms are employed to conclude the final results. Since the collective decision is being taken in a more demarcated format, higher reliability in the results can be expected.

2.3. Ensemble Learning

According to [26], the classifier's performance degrades when a dataset is class imbalanced. Therefore, they proposed a hybrid ensemble model to avoid unsatisfactory results from classification. Their model is based on two algorithms called AdaBoost.M2 and adapt SMOTE. Their model aims to solve the class imbalance problem to predict the probability of term deposit. Classifiers like Bayesian Network, Alternating Decision

Tree, Tree-J48, and REPTree (Reduced-Error Pruning) have been used in their experiments and achieved 96.3% accuracy. However, the results are acceptable for accuracy but come with complexity and lack of generalisation.

A method for question classification is proposed in [27] that employs ensemble learning algorithms to train multiple question classifiers. The components combinedly generate hypotheses by extracting high-frequency keywords from the corpus semantically. Bagging and AdaBoost were applied as ensemble methods to construct a decision tree that tags weights to questions for adjustment of decision. A likewise approach was also discussed in M-Ensemble Learning. An improved ensemble learning approach called M-Ensemble Learning has been proposed [28]. That works by dividing the results of combined methods into two main formats, namely the 3-Ensemble model and the 4-Ensemble model. Naïve Bayes, Decision Tree and Multilayer perceptron are in the 3-Ensemble model while combining even number methods, such as Naïve Bayes, Decision Tree, Multilayer Perceptron, and K-Nearest Neighbor are in the 4-ensemble model. The result of the experiment showed that the 3-Ensemble method outperforms.

Another such effort can be viewed in the proposed ensemble classifier [29]. They have proposed an ensemble classification algorithm to predict the credit score of bank customers with better accuracy. They have performed experiments on classifiers like Decision trees, Logistic Regression, Nearest neighbour, and Support Vector Machines. Based on their experimental study and results on the Australian credit dataset, Random Forest and Extra-Tree classifiers came up with better accuracy. Furthermore, with the use of the SVM model, their accuracy improved even better.

Using Spanish and US banks datasets, [29] have proposed a set of ensemble classifiers. This involves a simple majority voting scheme. Sever classifiers included in this study are ANFIS, SVM, Linear RBF, Semi-online RBF1 and Semi-online RBF2, Orthogonal RBF, and MLP. The ensemble design took two, three, four, five and six classifiers at a time from the seven classifiers. This study can be followed to discover more such flavours of ensemble classifiers but cannot be concluded based on their proposal process.

Ensemble learning uses multiple learning algorithms as components to achieve better performance than a single algorithm to the best of our understanding. This methodology inherits issues and challenges of dealing with selecting and optimising the associated components, i.e., classifiers in this case. From the above-cited studies on ensemble learning, we observed a common trend of employing multiple classifiers to achieve better results than a single classifier. However, we found a disconnect in discussing why some classifiers have been selected for the given ensemble learning approach. Most of them were adapted because of their single performance. Our interest in developing a reliable classification service is also associated with systematically finding out the most suitable components.

2.4. Term Deposit

In parallel to exploring the most optimised and reliable machine learning classifier, we also have a deep dive into the literature that has used our selected Portuguese bank dataset to build our targeted solution. From the domain's perspective, we have focused on bank products like term deposits[30]. This fixed-term investment includes money deposits into an account at a financial institution. Such investments usually carry short-term maturities ranging from one month to a few years and will have varying levels of required minimum deposits. Generic Term Deposit Subscription Flow

- Customers access the bank web portal and create a profile.
- Customers can navigate various offered products and services.
- Customers can make phone calls to get more information
- Bank rollouts various products and services with related terms and conditions
- Bank launches different campaigns and promotions that can be visible on the web portal. They are also communicated to selected customers via phone calls by bank agents.
- Selection criteria can be random or per customer profile, interest, and behaviour.

The detail of the dataset can be found in the Research Methodology chapter. Our dataset was extracted from telemarketing campaigns where we can learn about potential customers from the last campaign data [31]. In their experiment, the same dataset was used by [32] to compare two machine learning algorithms with Weka, Scikit- Learn and Apache Spark. The evaluated training time accuracy and root squared mean. This study also included Higgs Dataset for incorporating diversity in source data. The concluding results favour the spark platform because it supports the parallel implementation of given algorithms. The bank dataset has been used in an experiment by [33]. They proposed an association rule mining for boosting Naïve Bayes classifier. The given association rule mining enhances the discovery process of relations between inputs in data. The use of this dataset can also be observed in a study [34] that proposed a modified version of boosting called ModBoosting. Their model was generated from the sample training set and then evaluated for error on the complete dataset instead of the only training set. Higher weight was assigned to misclassified instances which were given first preference in the next iteration.

In conclusion, from the literature review, we found specific directions and gaps that can take us to develop a reliable machine learning classier that is personalised and easily adaptable. The first and most important take-away is the need for evaluation of existing classification algorithms as performed by researchers [18-23]. As discussed earlier, we cannot close our discussion on the best classifier in the results of the cited reviews. There was no particular pattern that led us to a single best classifier. That reveals a need to evaluate existing algorithms for each dataset and related problems from scratch. This creates a challenge of time and resources to do the analysis. Skill and experience in the such analysis are the keys to optimising the resources, but a well-structured framework for such evaluation may reduce the efforts. The single requirement for such a framework should be its black-box offering that takes a dataset as input and outputs top n classifiers with optimised parameters. Another critical need identified from the literature is the employment of multiple classifiers. A single classifier can be prone to overfitting in some new scenarios. In the presence of peer classifiers, high variances in their results may alarm a warning of unbalance confidence. Such a warning for new data may raise the need for more training or performance tuning.

Multiple classifiers can be employed in various formats, including but not limited to democracy (most votes declare the chief), Statistical Mean and Mode, and Weighted classifier voted. For instance, in the Weighted classifier voted-based consensus, the vote of a well-respected classifier can bias the result more than other low-weighted classifiers. Pre-processing and feature engineering have been discussed by [12] and [13], which should be part of any framework as a core component. This can be deployed on tasks like missing values, spelling corrections, and areas where well-defined practices can be applied. However, class imbalance, dimensionality reduction, and low-size data domains may need expert consultation. Simi automatic approach can be applied to pre-processing, but considerable effort shall be put into this area on a whole automation track.

To the best of our knowledge, in the light of existing literature, we find minimal resources that have stressed improving classifier performance by peer review from other classifiers selected in a systematic qualification process. We also observed a lack of interest in discussing the reason for selecting component classifiers for an ensemble learning approach. We believe a more automated framework with a black-box appearance that evaluates classifiers and generates an ensemble learning model offers reliable output and personalised experience from a data viewpoint.

3. Research Methodology

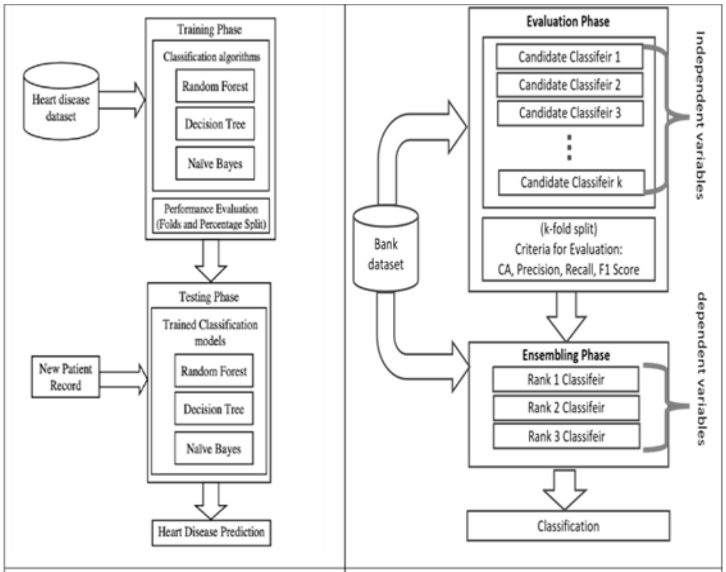
3.1. Comparative analysis and classifier selection

As described earlier, our quest in this research is to use an enhanced yet robust classification scheme from existing well-reputed machine learning classifiers to model bank customer behaviour. Based on the literature review, we deduce that existing machine learning classifiers perform sufficiently well on related modelling and require considerable consultancies from experts for pre-processing and performance tuning. That was required because of variations in domains and so in data. In our use case, where we focus on understanding bank customer behaviour, we want to design a robust classifier in results while personalised and portable on the other hand. We first perform an experimental study and then design a collaborative setup to form ensemble learning to achieve this.

The main reason for the experimental study for this research is the comparative analysis of existing algorithms so that top n (n=3) can be qualified. Our experiment starts with training selected classifiers on a given dataset in a series of iterations. Furthermore, each classifier will be subsequently passed through its parameter tuning cycles. At the end of this phase, we have the best classifiers.

From a generic viewpoint, our research methodology has mainly been influenced by the software development life cycle when evaluating our model's quality. The same methodology can be observed in [18, 19], and [14]. However, the methodology we found was very close to our work [35]. Their study was on the heart disease dataset, and they opted to study three classifiers to rank their performance. Since our aim of the research is not to select one best classifier but identify the top 3 performers so that they can be ensembled for final work at a later stage; therefore (Fig. 3), we extended their methodology as follows:

- The training phase is termed as Evaluation phase. Training and performance tuning of each candidate classifier will be performed individually.
- The number of classification algorithms is not fixed in our methodology. For example, we include eight classification algorithms in this work, but they can vary according to dataset size and associated feature groups.
- Instead of a percentage split, we use a k-fold split because of its smaller size.
- Evaluation of each candidate classifier will be examined on Classification Accuracy, Precision, Recall, and F1 Score, respectively.
- Top-ranking classifiers will be selected for the ensemble phase. The number of selected classifiers will be odd (i.e., 3, 5, 7,...). That is required to conclude the consensus when the decision is required to resolve class declaration conflict.
- The same dataset has been used in the evaluation and ensemble phase. Therefore, we are not re-training the classifier for the assembling phase. Since the dataset remains the same and classifiers have already been qualified for their best shapes, no further efforts are required for re-training.



a) Research Methodology by (Benjamin and S. Antony Belcy, 2018).

b) Research Methodology derived for our work.

Fig. 3. Our Research Methodology Vs Benjamin's Research Methodology

In the context of this research methodology, we comprehend our classifiers as operands or variables. The selection of top contributing variables (i.e., classifiers) constitutes the final grouping. With this attitude, candidate classifiers act as independent variables because they can be freely adapted, optimised, and selected on given criteria. That further extends the freedom to experiment with other machine learning approaches. The dependent component of this methodology is our ensemble phase. Hence, the inclusion and selection of appropriate candidate classifiers influence the better formation of ensemble components. This section briefly overviews machine learning classifiers, evaluation metrics, and datasets. The classification algorithms included in this work have been selected for their broad applicability, generality, and base approach. Infrequent base classifiers and too-narrowed flavours have not been included as candidate classifiers but can be included in future studies.

3.2. Candidate classification algorithms

3.2.1. Naive bayes

Naive Bayes classifiers are a family of simple "probabilistic classifiers" based on applying Bayes' theorem with strong (naïve) independence assumptions between the features [36]. They are among the most basic Bayesian network models but coupled with kernel density estimation, and they can achieve higher accuracy levels.

We do not perform iterative parameter estimation to build a Naive Bayesian model. However, it is very much helpful when the number of images is more significant. Due to its simple nature, the Naïve Bayesian algorithm gives better results in solving complex problems. The Bayes theorem calculates the posterior probability. It adopts the result of a predictor (x) value on a given class (c) independent of other predictors' values.

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)} \tag{1}$$

Where:

- P(c/x) is the given class posterior probability.
- P(c) represents the class prior probability.
- P (x|c) is the predictor probability
- P(x) is the predictor prior probability.

Naïve Bayesian classifies the input hybrid feature vector (x) into positives or negatives, stating that the individual attributes are independent. The input vector is classified based on the higher posterior probability for the two classes, i.e. positive and negative.

3.2.2. AdaBoost

AdaBoost [37] combines multiple weak classifiers to build one robust classifier [38]. It was formulated by Yoav Freund and Robert Schapiro, who won the Gold prize for their work in 2003. Adaboost works by combining other learning algorithms for higher performance. A weighted sum of the other weak learner classifiers' output obtains the Adaboost input. In some cases of overfitting, problems occur while using Adaboost methods. For example, the single learners may be weak, but as long as their performance is slightly better than random guessing, the final model can be proven to converge to a strong learner.

3.2.3. Artificial Neural Network

The Artificial Neural Network (ANN) is a computer system that simulates how the human brain analyses and processes information [39]. ANN is a network of artificial neurons made of connected units or nodes inspired by a human brain. Each neuron transmits signals to the corresponding neurons and forms a network. Some nonlinear function calculates the output of each neuron through connections. These connections are said to be edges. Edges and neurons have weighted values that adjust to the learning process. The weight increases or decreases the strength of the signal at a connection. A threshold is set for the neurons if the aggregate signal crosses that threshold. Usually, neurons are combined into layers. Different layers may perform different transformations on their inputs. Signals travel from the first layer (the input layer) to the last layer (the output layer), possibly after multiple layers traversing.

3.2.4. Decision tree

It uses a tree-like model of decisions and possible consequences, including chance event outcomes, resource costs, and utility [40]. A decision tree is the flowchart representation of attributes in the form of a tree where each attribute acts as a node. The tree is constructed with the best attribute as the root node and then passes through each attribute down to the leaf node, i.e. its respective class. Thus, in formulating decision rules against which the test samples or records are classified. The instances are classified in decision trees by sorting them using a top-down approach from root to a leaf node. First, we take the tree's root node and classify each instance. Then test each attribute of this node and move down to the remaining tree branch corresponding to the attribute value. This process is repeated at every new node for the subtree. Next, a tree is built by splitting the source set, constituting the tree's root node, into subsets which constitute the successor children. The splitting is based on a series of splitting rules based on classification features. This process is repeated on each derived subset in a recursive manner known as recursive partitioning. The recursion is complete when all of the target variable values in the subset at a node are identical or when splitting no longer adds value to the predictions. This top-down induction of decision trees (TDIDT) is an example of a greedy algorithm, and it is by far the most frequently used technique for constructing decision trees from data.

3.2.5. Random forest

Random Forest is an ensemble learning method for classification, regression, and other tasks that works by building a large number of decision trees during training and outputting the class that is the mode of the classes (classification) or mean/average prediction (regression) of the individual trees[41]. For regression tasks, the mean or average prediction of the individual trees is returned. Random decision forests adjust for decision trees' proclivity of overfitting to their training set. Random forests generally outperform decision trees, but their accuracy is lower than gradient-boosted trees. However, data characteristics can affect their performance.

3.2.6. Support vector machine

SVMs are very efficient in high dimensional spaces and generally are used in classification problems [42]. It helps reduce the classification error and maximises the geometric boundary that separates the class values. Each data point is plotted at its respective coordinates in N-dimensional feature space. The classifier identifies the hyper-plane to isolate the data into their respective classes. After determining the hyper-plane testing, samples are predicted on either side of the plane. The hyper-plane can be characterized as,

$$w.x + b = 0 \qquad (2)$$

Here x denotes the N-dimensional input vector, w is the vector weight defined as $w = w1, w2, w3 \ldots..wn$, while b is the model bias function. Several decision boundaries may occur as we have two classes' positives and negatives. SVM identifies the decision boundary, the hyper-plane having the most extreme separation from the two classes. The proposed SVM utilizes a linear kernel with a constant margin C as 1. The hyperplane dotted line, and the binary classes +1 and -1 are shown in the Fig. 4.

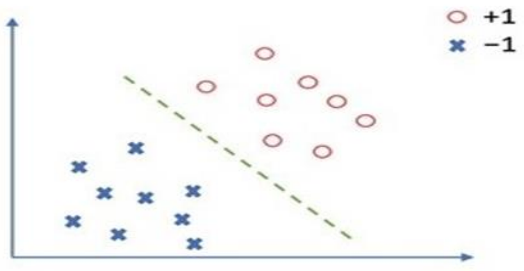


Fig. 1. Support Vector Machine

Classification using SVM is a supervised machine learning approach requiring appropriate training on larger datasets for binary classes. SVM selects the values of w and b from the training samples. SVM identifies a hyperplane for maximum separation between true and false training examples. From the below equations, the hyperplane H over the training data samples are represented as

$$\begin{array}{ll} X_i + \mathrm{w} + \mathrm{b} \leq 1, & Y_i = -1 \\ X_i + \mathrm{w} + \mathrm{b} \geq 1, & Y_i = +1 \end{array} \qquad (3)$$

From these equations,

$$Y_i(X_i.\mathrm{w} + \mathrm{b}) - 1 \geq 0, \qquad \forall\ \mathrm{i} \qquad (4)$$

From the above equations, it is observed that all the trained samples may occur on both sides of the hyperplane. After training the model, the decision problem solution is identified by calculating the sign of the $\mathrm{Y_i}$ with the coefficient vector w.

3.2.7. Logistic Regression

Logistic regression models are the input space into the probability of the target class between 0 and 1. Initially, it works has been designed for binary classification but can be scalable to multi classes. The main driver of logistic regression is the logistic function that estimates the parameters of a logistic model. Mathematically, it has one dependent variable with two possible values, i.e., 1 or 0.

For logistic regression (Fig. 5), the sigmoid function is used as an activation function, and it is defined as:

$$f(x) = \frac{1}{1+e^{-x}} \qquad (5)$$

where,
e = base of natural logarithms
value = numerical value one wishes to transform
The following equation represents logistic regression:

$$y = \frac{e^{(b_o+b_1X)}}{1+e^{(b_o+b_1X)}} \qquad (6)$$

Here,
x = input value
y = predicted output
b0 = bias or intercept term
b1 = coefficient for input (x)

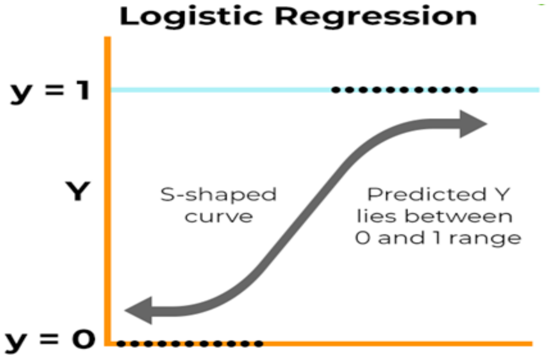


Fig. 2. Logistic regression

This equation is similar to linear regression, where the input values are combined linearly to predict an output value using weights or coefficient values. However, unlike linear regression, the output value modelled here is a binary value (0 or 1) rather than a numeric value.

The log-odds for the value called “1” is a linear combination of given independent variables. Since logistic regression acts as a probability estimator and determines relationships between input and output (supervise learning), therefore has been extensively used in literature for predicting the likely use cases. The case of one explanatory variable is called simple linear regression; for more than one, the process is called multiple linear regression. This term is distinct from multivariate linear regression, where multiple correlated dependent variables are predicted rather than a single scalar variable.

3.2.8. Gradient Boosting

Gradient Boosting is a machine learning technique for regression and classification problems, which generates an ensemble of weak prediction models and is used to create a prediction model. That is because, often, decision trees [38]. Gradient boosted trees are generated from decision trees' weak learning process, which usually outer performs random forest. Moreover, it works on wise step methods like other boosting models by optimising a random differentiable loss function.

3.2.9. Ensemble Learning

Ensemble methods are used to improve the accuracy of predictions. Ensemble methods combine many learning algorithms to provide higher predictive performance than could be obtained solely by using any of the constituent learning algorithms. Evaluating the prediction of an ensemble often necessitates more processing than evaluating the prediction of a single model. In one sense, ensemble learning may be thought to compensate for poor algorithms by performing much extra computation. On the other hand, the alternative is to perform a significant amount of additional learning on a single non-ensemble system. Using an ensemble system, it is possible to improve accuracy while using the same increase in compute, storage, or communication resources by distributing that increase across two or more methods rather than increasing resource use by increasing the use of a single method. Ensemble methods frequently make use of fast algorithms, such as decision trees, to improve efficiency. (for example, random forests), although slower algorithms can also benefit from ensemble techniques.

3.3. Evaluation metrics

3.3.1. Classification accuracy

[43] Classification Accuracy is what we usually mean whenever the term accuracy is used. It is the ratio with which the classification algorithm has correctly classified the test samples. It can also be described as the ratio of correctly identified samples to the total number of test samples. So, simply it is the rate of correct classifications, either for an independent test set or using some variation of the cross-validation idea. Of course, it only works correctly if every class includes the same number of examples. Mathematically, it can be expressed as:

$$Accuracy = \frac{True\ Positive + True\ Nagative}{True\ Positive + True\ Nagative + False\ Positive + False\ Nagative} \tag{7}$$

We can simplify this equation as below:

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ Numbers\ of\ predictions\ made} \tag{8}$$

3.3.2. Recall

Recall mainly focuses on the proportion of the positives correctly identified. [44] Recall is the opposite of precision and measures false negatives against true positives. False negatives are significant mainly to prevent disease detection and other safety predictions. It is measured as the number of true positives divided by the total number of true positives and false negatives in a two-class unbalanced classification problem. So, it answers the question: What proportion of actual positives was identified correctly?

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \tag{9}$$

3.3.3. F1 Score

F1 Score is the weighted average of Precision and Recall. It is used to balance the two objectives: high precision and high recall. [43] F1 Score is the harmonic mean for measuring the accuracy of a test between precision and recall. The F1 scoring range is [0, 1]. It shows how accurate our classification is (how many instances it is correctly classified) and even how robust it is. So, it is the weighted average of Precision and Recall that can be expressed as:

$$F1\ Score = 2 * \frac{Recall * Precision}{Recal + Precision} \tag{9}$$

3.3.4. The area under Curve

[43] Area Under Curve (AUC) is one of the most commonly used evaluation metrics. It is used to categories binary classification problems. The AUC of a classifier is equal to the chance that a randomly selected positive example will be more excellent than a randomly selected negative example.

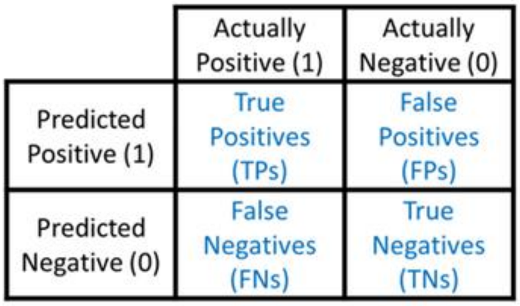
3.3.5. Precision

It is the actual number of positive samples predicted among the total predicted samples as positive. It can also be described as the ratio of actual positive samples belonging to Class 1 to the predicted number of samples for Class 1. It only measures the rate of false positives. In some fields, such as spam detection, a false positive is considered a more serious error than a false negative (generally, missing an important email is more detrimental than accidentally deleting spam that got past the filter). [44] Precision is computed as the number of actual positive values in an unbalanced classification problem with two classes divided by the overall number of actual positive and false positives. So, It attempts to answer the question of what proportion of identifications was correct, that is mathematically can be expressed as:

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \qquad (10)$$

3.3.6. Confusion matrix

[43] As the name implies, the confusion matrix gives the output matrix and describes the model's whole performance based on the test whose ground truth values are known in advance, as a specific table layout allows visualisation of the performance of an algorithm, typically a supervised learning one. The table represents class-wise prediction results summarizing correct and incorrect predictions of samples for each class. The predictions are categorised into four groups by looking at the binary class problem with two classes, Class 1 as positive and Class 2 as negative. i.e. True Positives (TP), False Negatives (FN), False Positives (FP), and True Negatives (TN), as shown in the Fig. 6:



	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

Fig. 3. Confusion Metrics

Each of the four groups is briefly described below,

- TP is the true prediction of an observation belonging to a positive.
- FN is the false prediction of an observation belonging to a positive.
- FP is the false prediction of an observation belonging to a negative.
- FN is the true prediction of an observation belonging to a negative.

Generally, it is a holistic way of viewing true and false positive and negative results, and the matrix accuracy can be measured using the "main diagonal" average value, i.e.:

$$Accuracy = \frac{True\ Positive + True\ Nagative}{Total\ Samples} \qquad (8)$$

3.4. Libraries of analysing

- Pandas: is an open-source Python Data Analysis Library, free to use (under a BSD license); it is one of the most popular and widely used data munging/wrangling tools. It gathers data (from a CSV or TSV file or a SQL database) and turns it into a Python data frame with rows and columns like a table in statistical applications such as Excel or SPSS [45].
- NumPy: It is a Python library that includes a multi-dimensional array adding support for multi-dimensional arrays and matrices as well as several derivative objects (such as masked arrays and matrices, contains a variety of routines for performing quick array operations, such as mathematical, logical, basic statistical operations, shape manipulation, selecting, sorting, I/O, random simulation, discrete Fourier transforms, and more [46].
- Scikit-Learn: Scikit-learn is a free machine learning library that focuses on data modelling data by providing a consistent Python interface for supervised and unsupervised learning algorithms. It features many classifications, Regression, and clustering algorithms and is designed to work with other scientific and numerical libraries, including NumPy [47].

3.5. Data collection

The data is related to a Portuguese banking institution's direct marketing initiative. Calls were used in the marketing campaigns. In order to assist if the product (bank term deposit) would be ('yes'), i.e. subscribed or ('no'), i.e. not subscribed, multiple contacts with the same client were frequently required. [30].

- Dataset descriptions (Table 1)

Table 1. Dataset Descriptions

Dataset characteristics	Multivariate	**Number of Instances**	45211
Attribute Characteristics	Real	**Number of Attributes**	17
Associated Tasks	Classification	**Missing Values**	None

- Data of the bank client (Table 2)

Table 2. Dataset Attribute for Bank client data

Sno	Attribute	Data Type
1	Age	(Numeric)
2	Job	(Categorical), type of job
3	Marital	(Categorical), marital status
4	Education	(Categorical)
5	Default	(Binary: "yes","no"), has credit in default?
6	Balance	(Numeric), Average yearly balance, in euros
7	Housing	(Binary: "yes","no"), has a housing loan?
8	Loan	(Binary: "yes","no"), has personal loan?

- Related to the last contact of the current campaign, data from the current campaign's most recent contact (Table 3).

Table 3. Dataset Attribute for campaigns

no	Attribute	Data Type
9	Contact	(Categorical: "Cellular", "Telephone"), Contact communication type.
10	Day	(Numeric), Last contact day of the month.
11	Month	(Categorical), Last contact month of the year.
12	Duration	(Numeric), Last contact duration, in seconds.

- Other attributes (Table 4).

Table 4. Other Dataset Attribute

Sno	Attribute	Data Type
13	Campaign	(Numeric), the number of contacts performed during this campaign for this client.
14	PDays	(Numeric), the number of days since a previous campaign contacted the client; -1 indicates that the client has never been contacted before.
15	Previous	(Numeric), the number of contacts made for this client before this campaign.
16	Poutcome	(Categorical):"unknown","other", "failure", "success."), The prior marketing campaign's results.

- Output variable (desired target): The last field is the label or target field called Y. It has the client subscribed to a term deposit (Binary: “yes”, “no”).

3.6. Sample study description

In our study, we utilized a dataset collected from a financial institution (details have been anonymized to protect the institution's and its customers' privacy). The dataset consisted of various customer attributes, product preferences, and past interactions with the bank. Based on this historical data, we aimed to predict customer inclination towards subscribing to a specific bank product offering.

We conducted thorough dataset pre-processing before proceeding with the analysis to ensure data quality and consistency. This included handling missing values, transforming categorical variables into numerical representations, and normalizing the data. Moreover, we performed exploratory data analysis to gain insights into the relationships between different features and their impact on the target variable.

To create a reliable model, we employed stratified k-folding cross-validation, dividing the dataset into five folds. This technique ensured that each fold maintained a similar proportion of the target class as in the complete dataset, reducing biases and providing a better assessment of the classifiers' performance. The training and testing process was conducted across all five-folds, and the performance metrics were averaged to obtain a more accurate representation of each classifier's capability.

We trained and tested eight candidate classifiers, tweaking their parameters for optimal performance. The classifiers were evaluated based on various metrics, such as AUC, classification accuracy, F1 score, precision, and recall. Based on these results, we identified the top three performers - Gradient Boosting, Logistic Regression, and Artificial Neural Network - which formed the decision-making committee for the prediction phase.

In conclusion, the sample study aimed to provide a comprehensive workflow for understanding customer behaviour using machine learning techniques. By utilizing a well-processed dataset, applying stratified k-folding cross-validation, and carefully evaluating the performance of multiple classifiers, we successfully developed a robust model that can predict customers' inclination towards a specific bank product offering. This workflow and its associated results can be a valuable resource for financial institutions looking to improve their understanding of customer behaviour and enhance their product offerings.

3.7. Ethical, legal and HR Considerations

Intelligent machine learning systems improve our lives by analysing behaviours, detecting patterns, predictions, optimising operations, and many other applications that make the processes more efficient and allow us to conduct significant and fundamental business changes. However, continuing to employ or use machine learning is not without obstacles and challenges. This section focuses on ethical and legal considerations and human resources that must be considered or followed.

3.7.1. Ethical considerations

In many scenarios, the analysis requires to use of actual data, such as patient records or bank customers' data (as in our example), to reveal patterns and trends that serve the interests of institutions. Therefore, this identifying information directly related to a particular person cannot be shared and should be replaced by fictitious values [48]. Furthermore, the institution must obtain informed consent from customers before collecting such data; customers must be aware that when they submit their data to the bank or any other institution, at some point, these data may be used to improve the services and processes provided to them. Finally, in some cases, data collected from customers may contain data elements that are not important in training the model and have no role in improving the model. Thus, it is ethical for an institution to use only the relevant or critical data elements that are necessary to implement the machine learning model [31].

We make choices based on our cognitive capability, moral values, experience, and available knowledge. Research activities aim to contribute to an existing body of knowledge for improving the lifestyle, finance, environment, and decision-making capabilities. However, in pursuing research for good, we must still adhere to establish norms and ethics so that side-effects of underlying efforts and studies can be prevented. According to [49], research ethics is "the standards of the researcher's behaviour about the rights of those who become the subject of a research project, or who are affected by it". This depicts the author's responsibility to uphold the ethical conduct standard in all research phases. In this work, we strictly follow our contribution's data privacy policy and originality.

3.7.2. Data privacy

Bank data may contain sensitive personal information (SPI) and personality identification information (PII). SPI can include the full client name, driver's license, Social Security, and medical records. In addition, it is possible to obtain non-sensitive, personally identifiable information from publicly available sources, including client zip code, race, gender, and date of birth. Our selected dataset excludes all such information and offers no means to reveal any of them.

Personally identifiable information (PII) refers to information that can identify an individual when used alone or with other relevant data. It may contain direct identifiers that identify an individual uniquely or quasi-identifiers that identify a person when combined with others. Our dataset and outcome of classifiers also comply with this notion of data privacy.

As explained briefly previously, our adapted bank dataset contains data that help our machine learning approach understand customer behaviour without knowing who exactly he or she is. This is achieved by learning from the data and tuning the co-efficient of our classifiers.

3.7.3. The originality of contribution

Our original contribution offers a robust classification framework based on existing classifiers and the process of forming the ensemble classifier. We do not claim any contribution to the listed machine learning classifier from either architectural or optimization perspectives.

The cited literature has guided us toward our proposed work's gap identification and design. Furthermore, the adapted research methodology has also been clearly referenced while deriving our own that suited our objective.

3.7.4. Transparency

Machine learning models are supposed to be transparent and interpretable. Therefore, firms must be explicit about who trains those models, what data is used in that training, and, most importantly, what recommendations their machine-learning models make. Furthermore, the participation of machine learning systems in decision-making that affects individual rights must be disclosed. In addition, systems must provide an understandable interpretation of their decision-making to end-users and can be reviewed by a competent human authority.

3.7.5. Bias

Bias is one of the most common machine learning problems, and it can be defined as the tendency of an algorithm to constantly learn the wrong thing by not considering all the data in the model. So, it arises from data not in the training set and data in the test set. The blame for bias is on the ML user; that is, the model will not be biased unless the programmer does not take into account the model's training in a way that limits the bias. Bias does not appear quickly but accumulates over time, leading to the ML model giving false and contradictory information about its built function. The problem of machine learning bias is likely to become more important as machine learning spreads in critical areas such as medicine and law and as more people without a deep technical understanding are assigned to deploying it. However, bias can be detected across machine learning (ML) workflows, enabling greater fairness and transparency to be built into the machine learning (ML) model.

3.7.6. HR Considerations:

The Department of human resources (HR) is one of the most critical departments in any organisation; it manages and develops employees in that organisation or institution. Therefore, the organisation should carefully consider the implications of applying machine learning before proceeding with it. Previous experiments in various organisations suggest that in most cases, the application of machine learning in an organisation automates tasks, often resulting in some employees losing their jobs [50].

This is one of the considerations that must be taken into account in our case. Although with the training of the model and the analysis of the bank's historical data, insightful information will be discovered from the data, this type of information can give the impression or belief that some employees failing to perform their assigned tasks as they should or not providing services of the required quality. Therefore, perhaps the organisation deal with this situation probably be in the form of punishment for those employees. Therefore, applying machine learning in any organisation should not include actions that punish employees or risk their jobs! On the contrary, machine learning applications in an organisation are supposed to help employees achieve their goals efficiently and effectively, besides providing informed decision-making and improving operations and procedures, (Fig. 7).

3.7.7. Legal Considerations

Furthermore, a firm that employs machine learning applications is required to abide by the rules and policies of the country in which located. Some of these regulations impose explicit limitations on customer data usage or dictate how that data is used for machine learning. As a result, those entities must follow the laws and adhere to the authorities' rules and procedures.

Those regulations may even specify which mechanisms are permitted to collect data. As a result, it is critical that the organisation is entirely aware of all applicable laws and rules and adheres to all requirements outlined in its machine learning applications [51]. Furthermore, businesses should get all rights to use when collecting data from clients, especially those used to analyse.

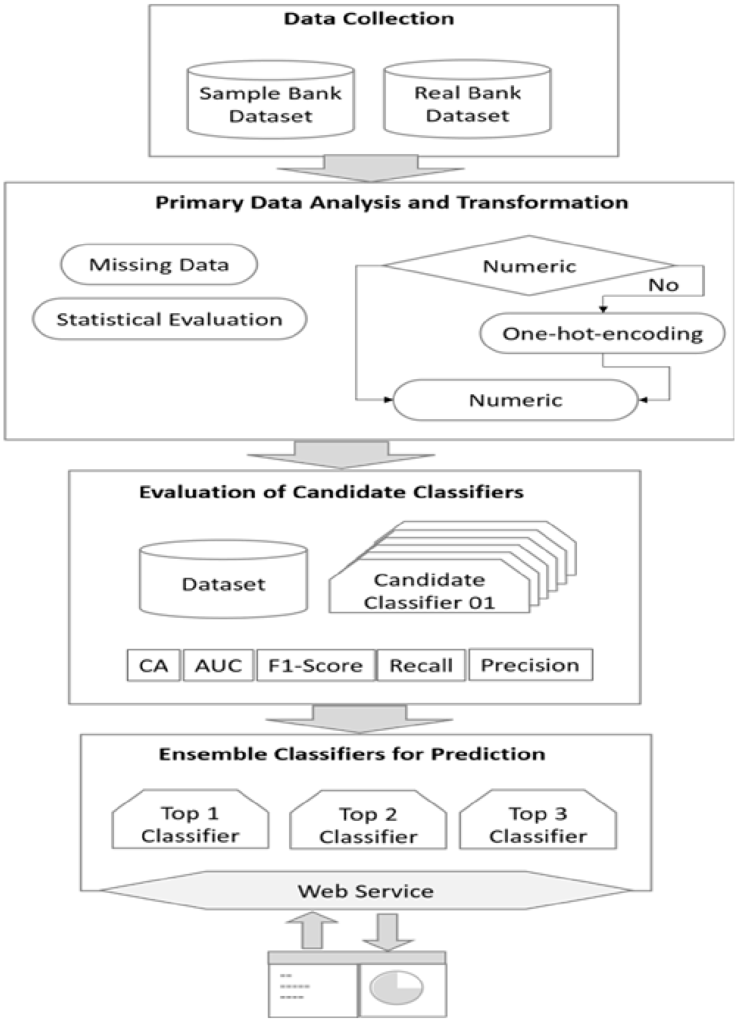


Fig. 7. Proposed Work computational environment

4. Data Analysis and Insights

In this section, we analyse and discuss the results we have achieved from our experiments. First, we formulate our plan to generate this data and then discuss the findings to achieve business results.

4.1. Formulate Plan

In this work, we have aimed to present a binary classification and prediction model for banks in the UK. From a practical business use-case perspective, this research contributes to empowering UK banks to use their private customer data and train their machine learning model. Furthermore, based on this model, web services could be exposed such that the classification/prediction shall be performed via calling API in our future work.

To achieve our aim, we encountered two challenges from the dataset end, i.e., 1) Unavailability of the public dataset from UK bank, and 2) Restriction on sharing customer data. So, we planned and executed an alternate approach. For public dataset accessibility, we selected a similar dataset from the UCI database that fulfils our requirements for training the model. However, the data was taken from a Portuguese Bank instead of the UK. Alongside this, we set our sub-objective that our ML model will follow a workflow from dataset to service exposure. With the help of this workflow, UK banks would be able to re-train our model with their personalised data. In summary, we design a workflow on how a model is trained on an initial dataset (i.e., currently available dataset) and later can be retrained by any interested insinuation’s data in a more personalised way.

The figure below illustrates our plan and workflow. In the first phase, Data Collection is to be decided. We select a publicly available sample dataset for this research and demo. After that, however, the actual data set shall be selected for business application intention. The next phase starts on the selected dataset, where primary analysis like missing data and statistical evaluation could let us know the health insight for the next phase.

In this phase, required transformations are also applied. At the end of the primary data analysis and transformation phase, we would start experiments to find the best classifier. In this phase, we select the widely cited algorithms that have their existence in popular machine learning development platforms like Python, R, and Weka. Then, each of these classifiers shall be trained and evaluated for their optimised set of parameters.

In sequential evaluation, each candidate classifier is taken from the queue one by one and trained over the given dataset. Then, the trained model is tested on a decided portion of the same dataset so that counts of True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) can be extracted. Based on these values, we further calculate Classification Accuracy (CA), Precision, Recall, F1-Score, Area under Curve (AUC), and Receiver Operating Characteristics (ROC) graph. These evaluation metrics for each classifier are then used as ranking criteria that decide which classifiers are best.

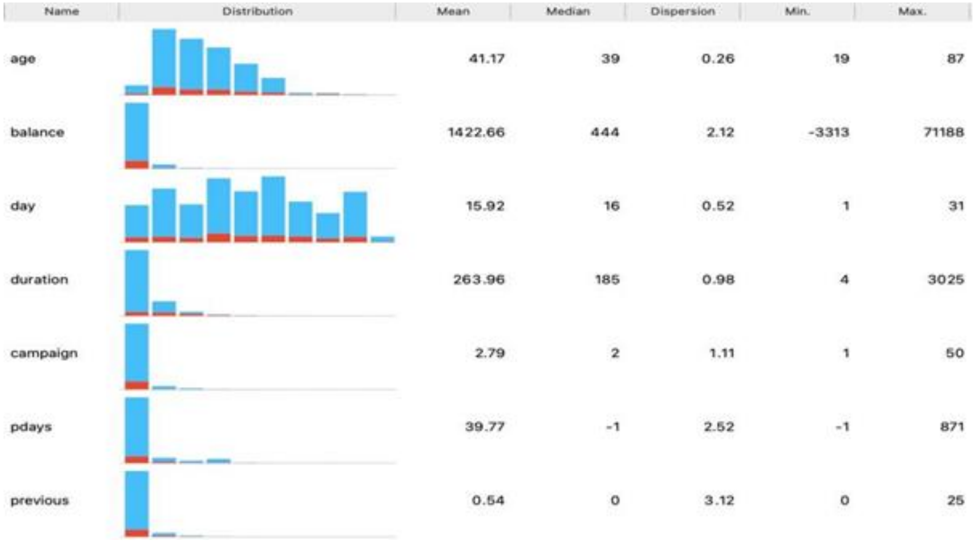
Using the ranking of evaluation metrics, we form a committee of the top three performers. They can be three, five, or seven, depending upon qualification. The main idea of the committee is to conclude any classification prediction on new data when there is conflict. For instance, if two out of three declare the prediction to be positive class, the final prediction will be positive. As these classifiers are already trained and capable of predicting new instances, a web service can be exposed that takes the listed features as input and outputs final classification results by consulting the committee of classifiers.

4.2. Analysis of dataset

The dataset contains 4521 instances spanning 17 columns and nine categorical and seven numeric features. The target/labelled field is binary-valued, i.e., Yes or No, referring to whether the client subscribed to a deposit. That collectively provides information about each client, such as age, marital status, and education level. A subgroup of this information is related to the last contact of the current campaign, including the month and day of the week the last contact was made and the number of days since the client was last contacted. The ten columns are categorial, containing textual values corresponding to a particular category for a given variable.

To understand the contents, we analysed categorical and numerical in the following charts. Each sub-chart refers to a feature and is analysed through statistical measurement. This measurement contains mean or average, median, or mid-point, dispersion of data, and minimum and maximum values. Each bar is split into two maximum parts through colour. Blue indicates a positive response, while red indicates a negative response toward subscribing or not subscribing to the term deposit.

- Age: The data is slightly skewed since people with an average age of 40 years have sufficient income and are more likely willing to pay on subscribing term deposits. The dispersion of data is rightly expressing the reason.
- Balance: More people with much balance have responded positively. Moreover, those with fewer savings have shown much interest in percentages. The reason for highly rightly skewed distribution can clearly be understood because if someone has some amount of balance, he or she will go for a subscription.
- Day: Data distribution for the day is almost even for working days. That shows no particular influence on day selection. Any day is good to be contacted, and subscribers behave equally.
- Duration: More data is in the first quantile for the duration field. However, the negative response is even on all groups of duration. That leads us to infer that if people are engaged for more extended distribution, then the change of positive response becomes higher, (Fig. 8).



Name	Distribution	Mean	Median	Dispersion	Min.	Max.
age		41.17	39	0.26	19	87
balance		1422.66	444	2.12	-3313	71188
day		15.92	16	0.52	1	31
duration		263.96	185	0.98	4	3025
campaign		2.79	2	1.11	1	50
pdays		39.77	-1	2.52	-1	871
previous		0.54	0	3.12	0	25

Fig. 8. Statistical Analysis of Numerical Attributes

- Campaign: Campaign refers to the number of contacts performed for this client (including the last contact). The bars are skewed to the left and advocate that more positive responses can be achieved with multiple campaigns.
- Previous: This attribute shows the number of contacts performed before this campaign and for this client. We can say from the bars that the contact never performed before with employees is more.
- Job: This is a categorical type of variable that shows the type of job done by an employee. Here admin has more numbers as compared to others. Blue-collar and Technicians come in second and third positions. While students, unemployed, and housemaid comes in the last positions. This attribute can be correlated with balance and education.
- Marital: Married people are seen in a much more stable position and hence are at the top, while divorced individuals are last.
- Education: University degree holders are at the top and high school certified at the second. That can be correlated to Job attributes because Admins and blue-collar job holders have many positive responses because of their financial position.
- Default: Here, the attribute Default shows that there is any credit in default. Here we see that the value of no credit is more, (Fig. 9).

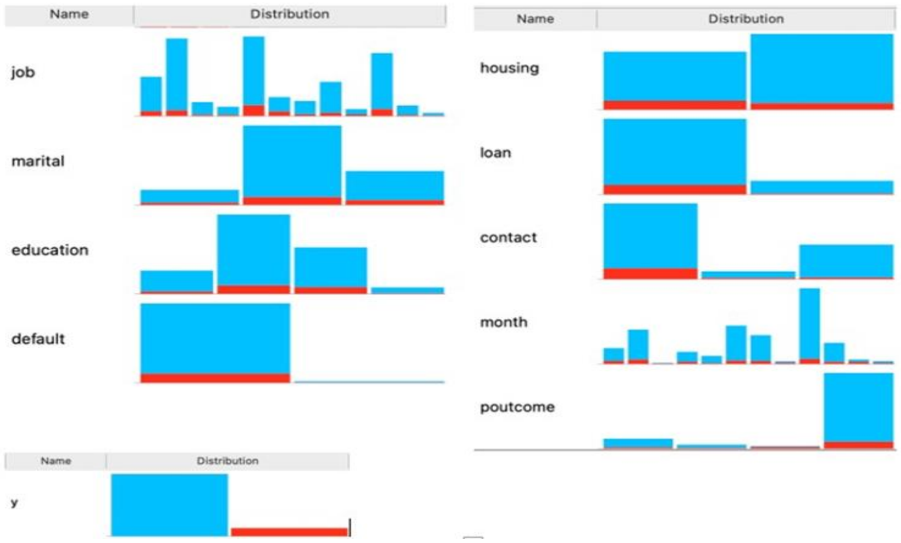


Fig. 9. Statistical Analysis of Categorical Data

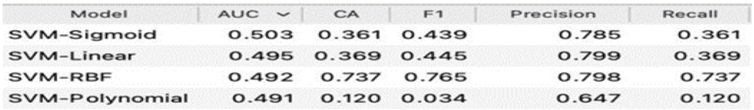
- Housing: This attribute shows whether the employee has a house loan or not. We can see that 53.33 % of employee has housing loan. However, the negative responses do not have much influence.
- Loan: This attribute shows whether the employee has a personal loan or not. This graph shows that a more significant number of employees have no personal loan.
- Contact: This attribute is for communication purposes in which 63.47% of employees have a contact type as a cellular phone. The remaining 36% were contacted via landline telephone media.
- Month: This attribute shows the last contact month of the year. The bars show that in May, more the one-third of all employees were contacted. While March, September, October, and December show equally fewer contacts.
- pOutcome: This attribute shows the number of contacts performed before this campaign and for this client. We can say the contact never performed before with employee is more.

4.3. Analysis of classifiers

In this section, we analyse each candidate classifier in more detail and discuss how it performed.

4.3.1. Analysis of support vector machine

SVM classifiers model data in a supervised learning paradigm that can be used for linear and non-linear classification. Generally, it has been used as the last step in the classification workflow. The data must first be refined at earlier steps like removing class imbalance, normalisation, dimensionality reduction, etc. Once all is done, SVM performs classification by using the kernel trick. The proper use of kernel trick cam implicitly maps its inputs into high-dimensional feature spaces. For our use case, we examine it without implementing intensive feature engineering. Instead, we experimented with SVM performance by tweaking its parameters, more specifically, kernels. As a result, the sigmoid kernel was a bit better than other kernels but still was tagged as the worst classifier among other candidate classifiers. The results with other kernels were even worst, (Figs. 10 &11). The AUC and ROC both confirm their worst performance for all the given kernels.



Model	AUC ⌄	CA	F1	Precision	Recall
SVM-Sigmoid	0.503	0.361	0.439	0.785	0.361
SVM-Linear	0.495	0.369	0.445	0.799	0.369
SVM-RBF	0.492	0.737	0.765	0.798	0.737
SVM-Polynomial	0.491	0.120	0.034	0.647	0.120

Fig. 10. SVM result with other Kernels

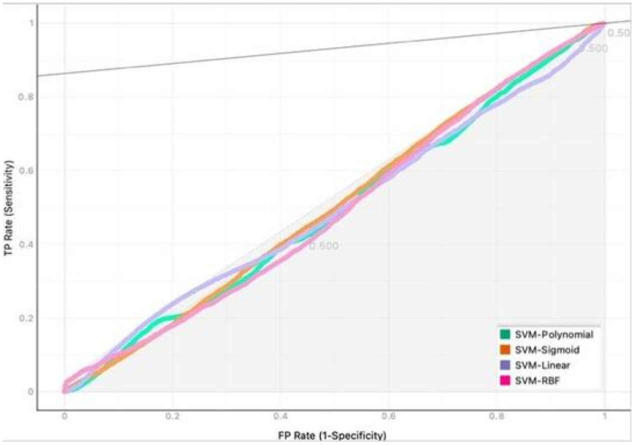
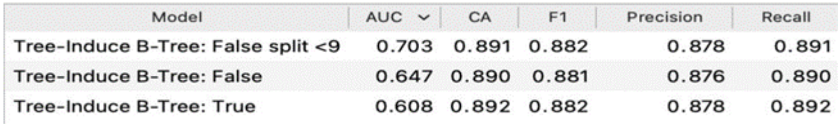


Fig. 11. Analysis of Support Vector Machine

4.3.2. Analysis of decision tree

Decision trees are the best in explaining the reason behind each classification path. They are one of the predictive modelling approaches in machine learning that offer simplicity of implementation and clarity of results. Our implementation of a decision tree in Python for binary classification has not been restricted to maximum depths. With a single sample leaf and splits of 9, it reaches to final classification conclusion. We have experimented with various parameters to improve accuracy. However, with its inherited weakness of underperformance for growing non-linearity, we could not improve its performance further. We experimented with tweaking its parameters like max_depth, min_sample_leaft, etc., Using the single processing, and we found better results with the parameters given in the source code (Fig. 12). The results from evaluation metrics take this classifier to the last slot of ranking for this dataset (Fig. 13).



Model	AUC ⌄	CA	F1	Precision	Recall
Tree-Induce B-Tree: False split <9	0.703	0.891	0.882	0.878	0.891
Tree-Induce B-Tree: False	0.647	0.890	0.881	0.876	0.890
Tree-Induce B-Tree: True	0.608	0.892	0.882	0.878	0.892

Fig. 12. Decision Tree results

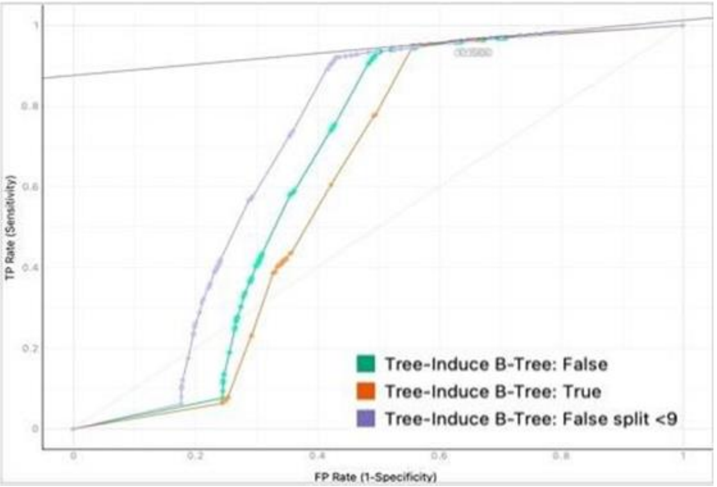


Fig. 13. Analysis of Decision Tree

4.3.3. Analysis of AdaBoost

Adaptive Boost or AdaBoost is used in combination with other classifiers for performance improvement. The week learner's outputs are combined into a weighted sum that is tagged as a boosted classifier. It tweaks them to avoid overfitting other classifiers so that the outcome can be cleaned of biases.

The main component of our adaptive boost classifier is the SAMME.R algorithm that uses probability estimates for updating the additive modelling. We also experimented with SAMMER and increased the number of estimators, but no upgrading impact was detected. By setting the learning rate to 1.0, we are stepping faster with the confidence of converging to the conclusion in 50 iterations. We experimented with lesion learning rates, including 0.5, 0.1, and 0.01, but could no longer improve the results (Fig. 14).

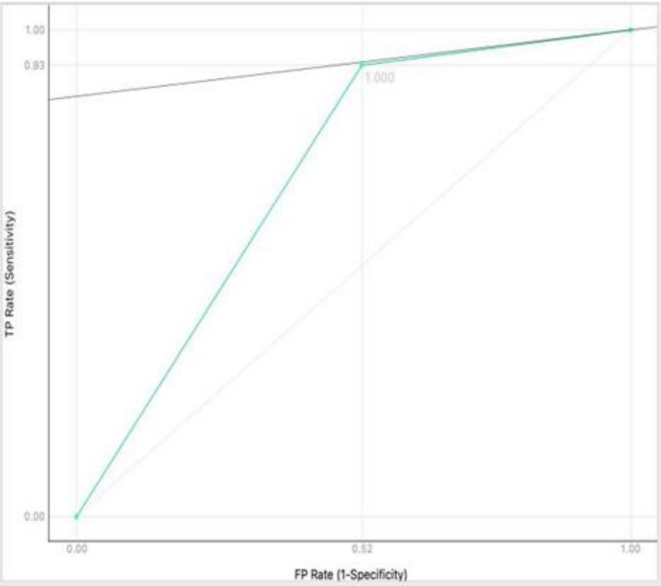


Fig. 14. Analysis of AdaBoost

4.3.4. Analysis of Naïve Bayes

Naïve Bayes is one of the probabilistic classifiers. Its base is the Bayes' theorem application that has strong (naïve) independence suppositions among the features. In our work, Naïve Bayes is the simplest among all other candidate classifiers concerning its parameters, i.e., it has been used with default settings. As a result, the performance matrix results for Naïve Bayes are much higher than most of the other classifiers.

One of the reasons for achieving the above performance is naïve Bayes's property of decoupling the class conditional feature distributions. That lets each distribution be estimated independently as a single dimension distribution; even though our dataset is not as extensive as many machine learning classifiers ask for, without putting effort into parameter tuning, still naïve Bayes shown promising results (Fig. 15).

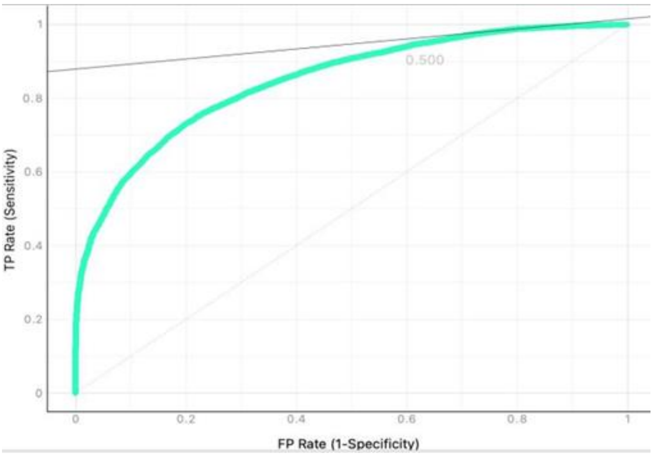


Fig. 15. Analysis of Naïve Bayes

The analysis of actual versus predicted outcomes can be observed in confusion metrics. It outperforms on Negative correct classification. However, the correctly classified Positive results are falling behind. The reason can be correlated to the class imbalance problem that has been discussed previously. The curve of the ROC graph shows pretty good probabilities for distinguishing both classes. That has already been confirmed by AUC value.

4.3.5 Analysis of Random Forest

The random forest offers a solution to classification and regression problems. The primary key of random forest is its utilisation of ensemble learning that combines several classifiers to solve more complex problems. As the name already highlights, the primary classifier for Random Forest is a tree/decision tree.

As the parameters for the random forest in the code illustrate its implementation for our use case, we have adapted to tune a few of them. As a result, our number of estimators consists of 10 estimates with no depth restriction. The parameters of a single decision tree have also been used in our random forest, but the main difference is the growing number of trees.

So, in this case, when we aggregated multiple decision trees into the formation of a random forest, we achieved results among top-ranking candidate classifiers (Fig. 16).

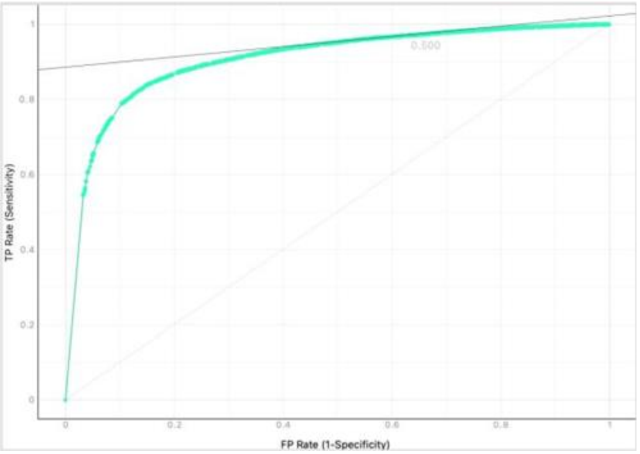


Fig. 16. Analysis of Random Forest

4.3.6. Analysis of Logistic Regression

Logistic Regression can be used to model classification problems by their class probability. Moreover, it uses a logistic function as a core component to model a binary dependent variable. Therefore, logistic Regression is considered the best solution for binary classification. Our implementation of logistic regression uses L2 regularisation that adds an L2 penalty. L2 penalty is the square of the magnitude of coefficients. Here we are also allowing cost parameter C to 1 to proceed to 100 iterations (Fig. 17).

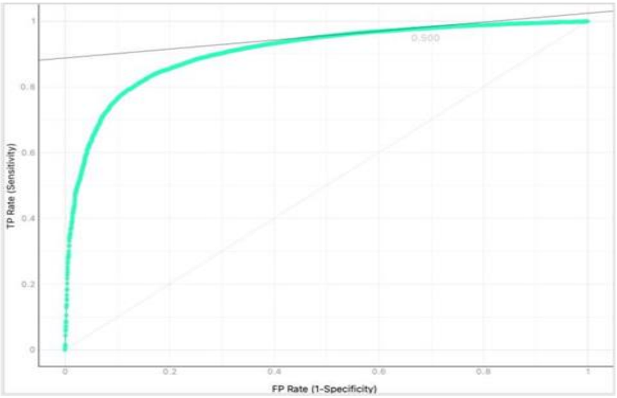


Fig. 17. Analysis of Logistic Regression

4.3.7. Analysis of Artificial Neural Network

Because of its self-adaptability property, Artificial Neural Networks (ANN) have shown outstanding performance in nonlinear data modelling. The classifier can be scalable to various problems, and with the proper parameters selection and more considerable data requirements, they can consistently achieve the top-most results. In this research, artificial neural networks have been adapted as per best practices to set parameters. Instead of designing a complex deep-learning neural network, we achieved the highest results with 124 hidden layers.

The Rectified Linear Unit function (ReLu) has been used as an activation function that many partitioners have also cited for more complex problems. Using a learning rate of 0.001, we could improve results at 50 iterations without engaging more processors. The rest of the parameters have been listed in the source code. The cause behind such outstanding performance can be correlated to its parameters while still considering the smaller size of the dataset. The dynamics of the classifier can be controlled based on its various parameters. This also makes it a complex classifier.

On the confusion matrix, actual versus predicted classified outcomes could be observed. Here the number of incorrect classifications is much shrined. The main reason for such a reduction is the flexibility of the classifier structure (Fig. 18).

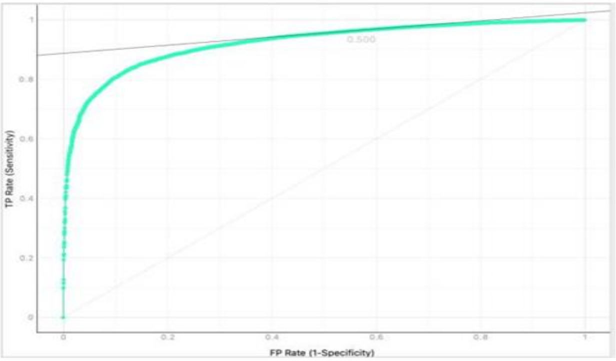


Fig. 18. Analysis of Artificial Neural Network

4.3.8. Analysis of Gradient Boosting

Gradient Boosting is one of the additive models in a forward step-wise approach. It optimises the arbitrary differentiable loss function. As per our dataset and classification nature, we use the particular case of gradient boosting, where we tend to induce a single regression tree. The results of Gradient Boosting as a candidate classifier concerning AUC and other matrices are best for the given dataset (Fig. 19).

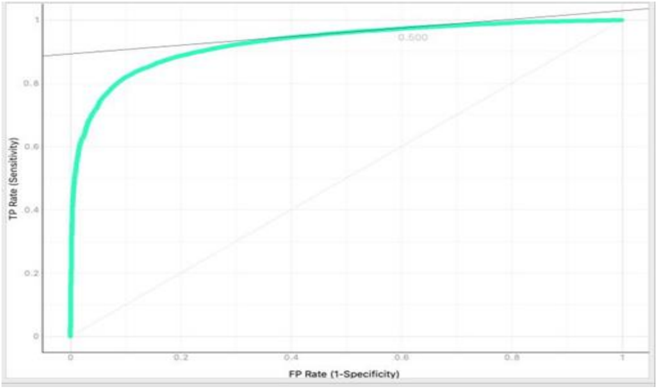


Fig. 19. Analysis of Gradient Boosting

The AUC and ROC both direct the best picture of classification.

4.4. Comparative analysis of classifiers

The candidate classifiers have experimented with their well-known parameters, and results were extracted to be evaluated using Area under Curve (AUC), Classification Accuracy (CA), F1 score, Precision, and Recall. Thus, to identify the best classifiers among given candidate classifiers, we are now able to move to the next step to form a committee of the best classifiers.

The Table 5 has been arranged in ascending order concerning AUC such that the best classifier can be listed in the last rows. Thus, gradient boosting, Artificial Neural networks, and Logistic Regression are the top performers, respectively, while Support Vector Machines, Decision Tree, and AdaBoost can be termed as worst as per the given use case on the given dataset (Fig. 20).

Table 5. Comparative Analysis of Classifiers

Model	AUC	CA	F1	Precision	Recall
SVM	0.503	0.361	0.439	0.785	0.361
Tree	0.703	0.891	0.882	0.878	0.891
AdaBoost	0.706	0.876	0.877	0.869	0.876
Naïve Bayes	0.847	0.875	0.872	0.869	0.875
Random Forest	0.905	0.900	0.889	0.886	0.900
Logistic Regression	0.906	0.901	0.888	0.887	0.901
Artificial Neural Network	0.922	0.902	0.896	0.893	0.902
Gradient Boosting	0.927	0.906	0.896	0.894	0.906

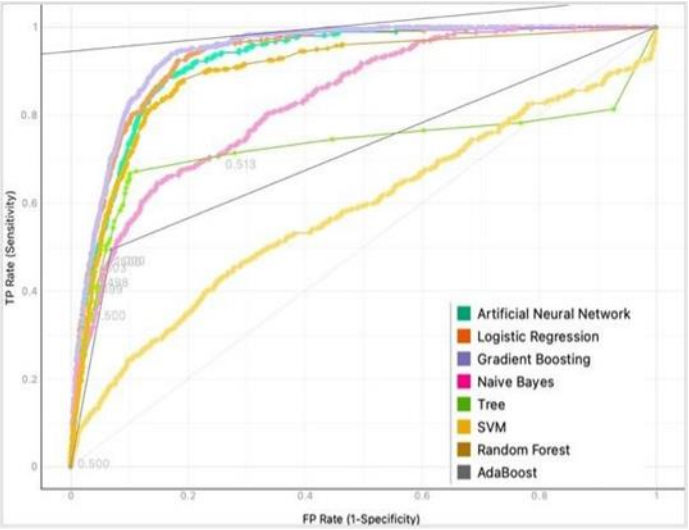


Fig. 20. Comparative Analysis of Classifiers

5. Interpret Findings

In this section, we present the details of our experiments and interpret our findings. Our criteria for interpretation will mainly focus on the reliability factor and classification accuracy as secondary. Other by-products, like personalised datasets and model training, will also be discussed to show the validity of our results. The following are the main components of our experiment.

5.1. Development environment

The implementation of proposed workflow was implemented in Python. There were several reasons for selecting Python over others. In general, python is a free, lightweight, open-source, high-level, and interpreted-based platform supported by an enormous community. More specifically, we wanted to take advantage of its plugins and extensions that make implementation rapid.

Orange3 is the wrapper of Python's libraries, including Pandas, Numpy, Matplotlib, and Scikit-learn. Alongside the graphical user interface, it also supports scripting for commercial usage. We have used both of its offerings in this work, i.e., GUI, for data analysis and experimentation. The orange3 library can load the trained model and Flask to offer prediction over the web service.

Our findings for its implementation using Python and orange3 are rapid development, community support, and freedom of deployment choices. Alternatively, we could select MATLAB, R, and Weka-like tools, but for them, we had to pay prices of environment readiness, the complexity of understanding, and restriction on deployment, respectively.

5.2. Cross-validation over train test split

In this research, to evaluate the accuracy of our model based on an error metric, we divided the dataset into training and testing sets, the training set used to train the model and the testing set used to test it. Furthermore, the accuracy gained for one test set can be substantially different from that obtained for another. Moreover, that may lead to unreliable accuracy, as one test set can differ from the accuracy obtained for another test set [52]. So, as a result, we employed the K-fold Cross Validation method, which divides the data into folds and uses each fold as a testing set at some time. The cross-validation method is used to lower biases, then train/test splits to assess the ability of machine learning models, i.e., classifiers. Initially, we set a 10-k fold but later changed to 5-folds as a larger K means less bias towards overestimating the honestly expected error but a higher variance and higher running time. A stratified K-fold cross-validation object is a variation of K-Fold (here, K=5), which returns stratified folds. The folds are made by conserving the percentage of samples for each class. In addition, it provides train/test indices to split data into train/test sets.

5.3. Optimised parameters for classifiers

As discussed in detail in the research methodology chapter, we selected eight candidate classifiers to be trained on the selected dataset. The primary purpose of choosing these classifiers was to demonstrate our approach on a list of well-sound algorithms and avoid creating new ones, as they have also shown top performance. Thus, the only task we had to perform was evaluating each independently and tweaking the respective parameter when necessary. Table 6 presents the final parameters for each selected algorithm.

Table 6. Configurable parameters for selected Classifiers

S.No	Classifier	Parameter	Value
1	Support Vector Machines	Cost (C)	v-SVM
		Regression loss epsilon(□)	0.1
		Kernel: tanh (g x.y +c)	Sigmoid
		Iteration limit	100
2	Decision Tree	Induce binary tree	True
		Min number of instances in leaves	2
		Do not split subsets smaller than	9
		Lit the maximal tree depts to	100
		Stop when the majority reaches	95%
3	AdaBoost	Base estimator	Tree
		Number of estimators	50
		Learning Rate	1.0
		Seed for random	1
		Classification algorithm	SAMME.R
		Regression loss function	Linear
4	Naïve Bayes	Defaults	
5	Random Forest	Maximum Trees	10
		Attributes to be splits at each split	5
		Do not split subsets smaller than	5
6	Logistic Regression	Regularisation Type	Ridge(Ḱegl, 2013)
		Strength C	1
7	Artificial Neural Network	Neurons in Hidden layers	124
		Activation function	ReLu
		Optimizer Regularization	Adam
		Learning Rate:	L2
		Number of iterations	0.0001
			50
8	Gradient Boosting	Number of trees	100
		Learning rate	0.1
		Relocatable training	Yes
		Limit depth of each tree	3
		Do not split subsets smaller then	2
		Fraction of training instance	1.0

For this purpose, we went through a series of experiments where we started with well-known parameters and changed the value of its main components to improve the results. Then, we followed the end-to-end data loading flow for each experiment, splitting as test and train sub-sets and generating the results for performance evaluation metrics. If results no longer improve, the parameters are freezing for more changes. Table 7 shows the performance in AUC (Area under the curve), CA ('Classification Accuracy', 2017), F1 score, Precision, and Recall. Random Forest,Gradient Boosting and Logistic Regression are the top 3 best performers on the given dataset.

Table 7. Performance Evaluation Results in AUC

Model	AUC	CA	F1	Precision	Recall
SVM	0.503	0.361	0.439	0.785	0.361
Tree	0.703	0.891	0.882	0.878	0.891
AdaBoost	0.706	0.876	0.877	0.869	0.876
Naïve Bayes	0.847	0.875	0.872	0.869	0.875
Random Forest	0.905	0.900	0.889	0.886	0.900
Logistic Regression	0.906	0.901	0.888	0.887	0.901
Artificial Neural Network	0.922	0.902	0.896	0.893	0.902
Gradient Boosting	0.927	0.906	0.896	0.894	0.906

Table 8 shows the probability that the score for the model in the rows is higher than the model in the column. Small numbers show the probability that the difference is negligible.

Table 8. Comparative Performance of Classifiers on AUC

	Gradient Boosting	Logistic Regression	Artificial Neural Network	Random Forest	Naive Bayes	SVM	AdaBoost	Decision Tree
Gradient Boosting		0.009	0.006	0.001	0.000	0.001	0.000	0.000
Logistic Regression	0.991		0.228	0.000	0.000	0.000	0.000	0.000
Artificial Neural Network	0.994	0.772		0.000	0.000	0.000	0.000	0.000
Random Forest	0.999	1.000	1.000		0.001	0.000	0.000	0.000
Naive Bayes	1.000	1.000	1.000	0.999		0.413	0.001	0.000
SVM	0.999	1.000	1.000	1.000	0.587		0.001	0.002
AdaBoost	1.000	1.000	1.000	1.000	0.999	0.999		0.035
Decision Tree	1.000	1.000	1.000	1.000	1.000	0.998	0.965	

5.4. Worse and bests

Decision Tree classifier is a kind of white-box model and easy to implement, understand and fast. However, it goes overfit and degrades performance with the more extensive set of feature sets. On the other hand, AdaBoost is a though effective technique for increasing the accuracy of learning algorithms. However, it did not perform well due to overfitting the training set (k-sets), as its primary objective is to minimize errors in the training set.

Naïve Bayes assumes that all features are independent, making the algorithm faster than the others. In some cases, speed is preferred over higher accuracy. However, in our case, performance on the given evaluation criteria was not satisfactory as it has not taken the probability outputs not too seriously. Another limitation was the assumption of independent predictors. In our dataset, the predictors were primarily independent.

Based on the results, the top performers are Gradient Boosting, Logistic Regression, Artificial Neural Networks, and Random Forest. Gradient Boosting outperformed because of its flexibility - it can optimise different loss functions and provides several hyperparameter tuning options that make the function fit very flexibly. Logistic Regression performed well. The reason is that its ability to provide feature importance assessment and the direction of each feature affects the response value – either positively or negatively. Inspecting the coefficient weights, the sign represents the direction, while the absolute value shows the magnitude of the influence. Finally, the performance of Random Forest improved because it was found capable of reducing overfitting in decision trees and helping to improve the accuracy. Furthermore, it performed well on ten trees with five splits.

5.5. Prediction as a service

The business utility of this research can be achieved if it is offered as a service at the application layer. Therefore, one of our objectives for this work was to develop a web service and make it available as an open-source project. To achieve this purpose, we need to develop and push our code to the GitHub repository so that anyone can download, improve, deploy, and use it. Back in our experiments phase, we had already trained our models and saved them in Pickle format. Now, we can load the top three performers using Python code and predict results.

6. Discussion of Results

In this section, we discuss the results of our experiments in the context of understanding customer behaviour using machine learning techniques for bank product offerings. The discussion focuses on the performance of the classifiers, the implications of the findings for the banking industry, and the study's potential limitations.

6.1. Performance of classifiers

Our experiments demonstrated that Gradient Boosting, Logistic Regression, and Artificial Neural Network were the top-performing classifiers in predicting customer inclination towards subscribing to a given item based on historical data. These classifiers outperformed others, such as Decision Tree, AdaBoost, and Naïve Bayes. The superior performance of Gradient Boosting can be attributed to its flexibility in optimizing different loss functions and the availability of numerous hyperparameter tuning options, which make it highly adaptable. Logistic Regression performed well due to its ability to assess feature importance and determine the direction of each feature's influence on the response value. Meanwhile, the Artificial Neural Network showed strong performance due to its capability to model complex relationships between inputs and outputs.

6.2. Adapting our models to diverse banking

Our study demonstrates the effectiveness of various machine learning techniques for understanding customer behaviour in the context of bank product offerings. These models can be applied to both large and small banks, with some considerations regarding the selection of the most suitable statistical model.

More complex models like Gradient Boosting, Artificial Neural Networks, and Random Forests may be preferred for large banks with abundant resources and a diverse customer base. These models can handle a higher degree of complexity and provide more accurate predictions, making them well-suited for large institutions aiming to maximize the efficiency of their marketing efforts and improve customer satisfaction.

On the other hand, small banks with limited resources may find simpler models like Logistic Regression or Decision Trees more appropriate. These models are easier to implement and understand, requiring less computational power and expertise, making them a more practical choice for smaller institutions.

It is crucial to note that the applicability of these models is not limited to a specific bank size or type. Instead, they can be adapted to various banking contexts, with the selection of the most suitable model depending on the institution's needs, resources, and goals. By carefully considering these factors, banks can make informed decisions about which statistical models to employ, ultimately enhancing their understanding of customer behaviour and driving better business outcomes.

6.3. Implications for the banking industry

The findings of this study have significant implications for the banking industry, as the top-performing classifiers can be utilized to enhance customer targeting strategies for bank product offerings. By leveraging machine learning techniques, banks can better understand customer behaviour, enabling them to tailor their marketing efforts and product recommendations more effectively. This approach can lead to improved customer satisfaction, increased customer loyalty, and, ultimately, higher revenue generation for the banks.

Furthermore, developing a web service incorporating the top-performing classifiers can facilitate the deployment of these machine-learning models in real-world scenarios. This web service can be a valuable tool for banks to integrate machine learning-driven decision-making into their existing systems and processes.

6.4. Limitations and future research

While the results of this study provide valuable insights into the application of machine learning techniques for predicting customer behaviour, there are some limitations that should be acknowledged. Firstly, the study relies on a specific dataset, which may not fully represent the diversity of customer behaviours in the banking industry. Future research could explore the performance of classifiers on different datasets, potentially with more features or a larger sample size, to further validate the findings.

Secondly, although the top-performing classifiers demonstrated strong performance in this study, other machine-learning algorithms or techniques could yield better results. Future research could investigate the performance of alternative classifiers or ensemble methods to determine if even better predictive accuracy can be achieved.

Lastly, the current study focuses on binary classification for predicting customer inclination. Future research could explore the use of multi-class classification or regression techniques to predict more granular outcomes, such as the probability of a customer subscribing to a particular product or the expected revenue generated from a customer's subscription.

In conclusion, this study contributes to understanding customer behaviour in the banking industry by demonstrating the application of machine learning techniques for predicting customer inclination towards subscribing to a given product. Furthermore, the top-performing classifiers identified in this study can be a valuable tool for banks to enhance their marketing and customer targeting strategies. Future research can build upon these findings by exploring alternative machine-learning techniques, datasets, and prediction tasks.

7. Conclusion

We aim to contribute to understanding customer behaviour using machine learning techniques for a bank product offering. The workflow begins with data analysis using statistical tools and data transformation for the next phase. Next, a queue of candidate classifiers is prepared for the transformed dataset. Each candidate classifier is trained and tested on the same dataset independently. Based on the performance ranking, the top three classifiers are marked as decision-makers in the prediction phase for new data. The qualified top three classifiers form a committee on the top of a web service. In case of conflict in the result declaration, the vote of two over one shall be finalised and concluded as a prediction.

From an implementation perspective, we formulate the problem as a binary classification task that evaluates customer inclination toward positive response for subscribing to a given item from historical data, i.e., training data. Using stratified k-folding (k=5) cross-validation, all the candidate classifiers were trained and tested. On the sample dataset, the results declared Logistic Regression, Artificial Neural Network, and Gradient Boosting as top performers and hence became members of the decision-making committee. Our results and workflow are sufficiently simple to be adopted by UK banks in their use cases. Furthermore, they can demonstrate practical usage by bringing up their confidential dataset and regenerating the same result in their local environment. Our code is written in Python with the help of Scikit-Learn, Orange3, and Flask libraries which can be cloned on the GitHub repository for anyone to experiment.

Acknowledgement

We would like to thank all teaching staff for their assistance at the Faculty of Engineering and Informatics, University of Bradford, United Kingdom.

References

[1] Statista-eCommerce-UK, in https://www.statista.com/outlook/dmo/ecommerce/united-kingdom?currency=gbp. 2021.
[2] Ecommerce News. 2019; Available from: https://ecommercenews.eu/ecommerce-in-uk-to-reach-e200-billion-in-2019/.
[3] Machine learning in UK financial services. Bank of England 2019; Available from: https://www.bankofengland.co.uk/report/2019/machine-learning-in-ukfinancial-services.
[4] A. M. Choudhury and K. Nur, "A Machine Learning Approach to Identify Potential Customer Based on Purchase Behavior," 2019 International Conference on Robotics,Electrical and Signal Processing Techniques (ICREST), Dhaka, Bangladesh, 2019, pp. 242-247, doi: 10.1109/ICREST.2019.8644458.
[5] V. Shrirame, J. Sabade, H. Soneta and M. Vijayalakshmi, "Consumer Behavior Analytics using Machine Learning Algorithms," 2020

IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT), Bangalore, India, 2020, pp. 1-6, doi: 10.1109/CONECCT50063.2020.9198562.

[6] Sundharam, V., M.S. Sriramm, and P. Pachhaiammal. Predicting the Customer Behavior Through Web Page and Content Mining Techniques. in 2018 International Conference on Communication, Computing and Internet of Things (IC3IoT). 2018. In IEEE,https://doi.org/10.1109/IC3IoT.2018.8668176.

[7] Asniar and K. Surendro, "Predictive Analytics for Predicting Customer Behavior," 2019 International Conference of Artificial Intelligence and Information Technology (ICAIIT), Yogyakarta, Indonesia, 2019, pp. 230-233, doi: 10.1109/ICAIIT.2019.8834571.

[8] Y. Zuo and K. Yada, "Using statistical learning theory for purchase behavior prediction via direct observation of in-store behavior," 2015 2nd Asia-Pacific World Congress on Computer Science and Engineering (APWC on CSE), Nadi, Fiji, 2015, pp. 1-6, doi: 10.1109/APWCCSE.2015.7476215.

[9] Y. Yamamoto et al., "Towards Self-Organizing Internet of Things - Aware Systems for Online Sales," 2015 11th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS), Bangkok, Thailand, 2015, pp. 208-215, doi: 10.1109/SITIS.2015.85.

[10] X. Deng, "Big data technology and ethics considerations in customer behavior and customer feedback mining," 2017 IEEE International Conference on Big Data (Big Data), Boston, MA, USA, 2017, pp. 3924-3927, doi: 10.1109/BigData.2017.8258399.

[11] Ibrahim, M. M. A., Syed-Mohamad, S. M. and Husin, M. H. (2019) 'Managing Quality Assurance Challenges of DevOps through Analytics.' In Proceedings of the 2019 8th International Conference on Software and Computer Applications. Penang, Malaysia, Association for Computing Machinery, pp. 194– 198. https://doi.org/10.1145/3316615.3316670.

[12] Gibert, K., M. Sànchez–Marrè, and J. Izquierdo, (2016). A survey on pre-processing techniques: Relevant issues in the context of environmental data mining. AI Communications. 29. 1-37. 10.3233/AIC-160710.

[13] Z. Guan, T. Ji, X. Qian, Y. Ma and X. Hong, "A Survey on Big Data Pre-processing," 2017 5th Intl Conf on Applied Computing and Information Technology/4th Intl Conf on Computational Science/Intelligence and Applied Informatics/2nd Intl Conf on Big Data, Cloud Computing, Data Science (ACIT-CSII-BCD), Hamamatsu, Japan, 2017, pp. 241-247, doi: 10.1109/ACIT-CSII-BCD.2017.49.

[14] Moura, A. F. D., Pinho, C. M. D. A., Napolitano, D. M. R., Martins, F. S. and Fornari Junior, J. C. F. D. B. (2020) 'Optimization of operational costs of Call centers employing classification techniques.' Research, Society and Development, 9(11) p. e86691110491.

[15] B. Valarmathi, T. Chellatamilan, H. Mittal, J. Jagrit and S. Shubham, "Classification of Imbalanced Banking Dataset using Dimensionality Reduction," 2019 International Conference on Intelligent Computing and Control Systems (ICCS), Madurai, India, 2019, pp. 1353-1357, doi: 10.1109/ICCS45141.2019.9065648.

[16] E. Çetiner, T. Koçak and V. Ç. Güngör, "Credit risk analysis based on hybrid classification: Case studies on German and Turkish credit datasets," 2018 26th Signal Processing and Communications Applications Conference (SIU), Izmir, Turkey, 2018, pp. 1-4, doi: 10.1109/SIU.2018.8404405.

[17] M. Fauvel, J. Chanussot and J. A. Benediktsson, "Kernel Principal Component Analysis for Feature Reduction in Hyperspectrale Images Analysis," Proceedings of the 7th Nordic Signal Processing Symposium - NORSIG 2006, Reykjavik, Iceland, 2006, pp. 238-241, doi: 10.1109/NORSIG.2006.275232.

[18] Romi S. Wahono, N. Suryana, Sabrina Ahmad, A Comparison Framework of Classification Models for Software Defect Prediction, October 2014, Journal of Computational and Theoretical Nanoscience 20 (10-12):1945-1950, DOI: 10.1166/asl.2014.5640.

[19] O. Adepoju, J. Wosowei, S. lawte and H. Jaiman, "Comparative Evaluation of Credit Card Fraud Detection Using Machine Learning Techniques," 2019 Global Conference for Advancement in Technology (GCAT), Bangalore, India, 2019, pp. 1-6, doi: 10.1109/GCAT47503.2019.8978372.

[20] P. Malik, S. Sengupta and J. S. Jadon, "Comparative Analysis of Soil Properties to Predict Fertility and Crop Yield using Machine Learning Algorithms," 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 2021, pp. 1004-1007, doi: 10.1109/Confluence51648.2021.9377147.

[21] Asare-Frempong, J. and Jayabalan, M. (2017a) J. Asare-Frempong and M. Jayabalan, "Predicting customer response to bank direct telemarketing campaign," 2017 International Conference on Engineering Technology and Technopreneurship (ICE2T), Kuala Lumpur, Malaysia, 2017, pp. 1-4, doi: 10.1109/ICE2T.2017.8215961.

[22] C. S. T. Koumétio, W. Cherif and S. Hassan, "Optimizing the prediction of telemarketing target calls by a classification technique," 2018 6th International Conference on Wireless Networks and Mobile Communications (WINCOM), Marrakesh, Morocco, 2018, pp. 1-6, doi: 10.1109/WINCOM.2018.8629675.

[23] A. Alzahrani and D. B. Rawat, "Comparative Study of Machine Learning Algorithms for SMS Spam Detection," 2019 SoutheastCon, Huntsville, AL, USA, 2019, pp. 1-6, doi: 10.1109/SoutheastCon42311.2019.9020530.

[24] Kun-Huang Chen and Hsuan-Wen Chiu. 2020. Applying AI Techniques to Predict the Success of Bank Telemarketing. In Proceedings of the 2020 4th International Conference on Deep Learning Technologies (ICDLT '20). Association for Computing Machinery, New York, NY, USA, 89–93. https://doi.org/10.1145/3417188.3417198.

[25] Sakar, C.O., et al., 2018. Real-time prediction of online shoppers' purchasing intention using multilayer perceptron and LSTM recurrent neural networks. https://doi.org/10.1007/s00521-018-3523-0

[26] P. Ruangthong and S. Jaiyen, "Hybrid ensembles of decision trees and Bayesian network for class imbalance problem," 2016 8th International Conference on Knowledge and Smart Technology (KST), Chiang Mai, Thailand, 2016, pp. 39-42, doi: 10.1109/KST.2016.7440523.

[27] Lei Su, Hongzhi Liao, Zhengtao Yu and Quan Zhao, "Ensemble learning for question classification," 2009 IEEE International Conference on Intelligent Computing and Intelligent Systems, Shanghai, 2009, pp. 501-505, doi: 10.1109/ICICISYS.2009.5358124.

[28] A. Rojarath, W. Songpan and C. Pong-inwong, "Improved ensemble learning for classification techniques based on majority voting," 2016 7th IEEE International Conference on Software Engineering and Service Science (ICSESS), Beijing, China, 2016, pp. 107-110, doi: 10.1109/ICSESS.2016.7883026.

[29] A. Safiya Parvin and B. Saleena, "An Ensemble Classifier Model to Predict Credit Scoring - Comparative Analysis," 2020 IEEE International Symposium on Smart Electronic Systems (iSES) (Formerly iNiS), Chennai, India, 2020, pp. 27-30, doi: 10.1109/iSES50453.2020.00017.

[30] P. Ravikumar and V. Ravi, "Bankruptcy Prediction in Banks by an Ensemble Classifier," 2006 IEEE International Conference on Industrial Technology, Mumbai, India, 2006, pp. 2032-2036, doi: 10.1109/ICIT.2006.372529.

[31] Moro, Sérgio & Cortez, Paulo & Rita, Paulo. (2014). A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision

Support Systems. 62. 10.1016/j.dss.2014.03.001.

[32] Linthicum, K.P., Schafer KM, Ribeiro JD. Machine learning in suicide science: Applications and ethics. Behav Sci Law. 2019 May;37(3):214-222. doi: 10.1002/bsl.2392. Epub 2019 Jan 4. PMID: 30609102.

[33] Shashidhara, B.M., et al. (2015). Evaluation of Machine Learning Frameworks on Bank Marketing and Higgs Datasets. 551-555. 10.1109/ICACCE.2015.31. Second International Conference on Advances in Computing and Communication Engineering (ICACCE). IEEE.

[34] T. Yang, K. Qian, D. C. -T. Lo, Y. Xie, Y. Shi and L. Tao, "Improve the Prediction Accuracy of Naïve Bayes Classifier with Association Rule Mining," 2016 IEEE 2nd International Conference on Big Data Security on Cloud (BigDataSecurity), IEEE International Conference on High Performance and Smart Computing (HPSC), and IEEE International Conference on Intelligent Data and Security (IDS), New York, NY, USA, 2016, pp. 129-133, doi: 10.1109/BigDataSecurity-HPSC-IDS.2016.38.

[35] Shah, S. Gala and N. Patil, "ModBoost for unbiased classification," 2014 International Conference on Data Mining and Intelligent Computing (ICDMIC), Delhi, India, 2014, pp. 1-5, doi: 10.1109/ICDMIC.2014.6954252.

[36] H. Benjamin Fredrick David and S. Antony Belcy," Heart Disease Prediction Using Data Mining Techniques", October 2018, Ictact Journal On Soft Computing, October 2018, Volume: 09, IssuE: 01, ISSN: 2229-6956 (online), DOI: 10.21917/ijsc.2018.0253.

[37] F. El-Matouat, O. Colot, P. Vannoorenberghe and J. Labiche, "Using optimal variables for Bayesian network classifiers," Proceedings of the Third International Conference on Information Fusion, Paris, France, 2000, pp. MOD1/18-MOD1/23 vol.1, doi: 10.1109/IFIC.2000.862518.

[38] Ḱegl, B. a. (2013). The return of AdaBoost.MH: multi-class Hamming trees. CoRR, abs/1312.6086. https://www.semanticscholar.org/paper/The-return-of-AdaBoost.MH%3A-multi-class-Hamming-K%C3%A9gl/a37c1df39575fd59d8b3b4697da2de486c71ab3.

[39] Kégl, B.a., The return of AdaBoost.MH: multi-class Hamming trees. arXiv pre-print server, 2013.

[40] Dreyfus, Stuart E.. "Artificial neural networks, back propagation, and the Kelley-Bryson gradient procedure." Journal of Guidance Control and Dynamics 13 (1990): 926-928, DOI: 10.2514/3.25422.

[41] Quinlan, J.R. (1986) Induction of Decision Trees. Machine Learning, 1, 81-106. http://dx.doi.org/10.1007/BF00116251.

[42] Tin Kam Ho, "Random decision forests," Proceedings of 3rd International Conference on Document Analysis and Recognition, Montreal, QC, Canada, 1995, pp. 278-282 vol.1, doi: 10.1109/ICDAR.1995.598994.

[43] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt and B. Scholkopf, "Support vector machines," in IEEE Intelligent Systems and their Applications, vol. 13, no. 4, pp. 18-28, July-Aug. 1998, doi: 10.1109/5254.708428

[44] Mishra, A., (2018), Metrics to Evaluate your Machine Learning Algorithm. towards data science: [Online] [Accessed https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm- f10ba6e38234.

[45] Brownlee, J. (2020b), "How to Calculate Precision, Recall, and F-Measure for Imbalanced Classification.", Machine Learning Mastery: [Online], [Accessed https://machinelearningmastery.com/precision-recall- and-f-measure-for-imbalanced-classification.

[46] Adi Bronshtein, 'A Quick Introduction to the "Pandas" Python Library'., (2017), Towards Data Science: [Online] [Accessed https://towardsdatascience.com/a-quick-introduction-to-the-pandas-python-library- f1b678f34673.

[47] GeeksforGeeks, 2022, NumPy in Python. geeks for geeks. [Online] [Accessed https://www.geeksforgeeks.org/numpy-in-python-set-1-introduction.

[48] Brownlee J. , (2020a), "A Gentle Introduction to Scikit-Learn: A Python Machine Learning Library.", Machine Learning Mastery. [Online], [Accessed https://machinelearningmastery.com/a-gentle-introduction-to- scikit-learn-a-python-machine-learning-library.

[49] Lee, Yong Jae Shin, "Machine learning for enterprises: Applications, algorithm selection, and challenges", Business Horizons, Volume 63, Issue 2, 2020, Pages 157-170, ISSN 0007-6813, https://doi.org/10.1016/j.bushor.2019.10.005.

[50] Saunders, M. et al., (2000), Research Methods for Business Students: Lecturers' Guide., Harlow: FT Prentice Hall. Accessed: https://openresearch.surrey.ac.uk/esploro/outputs/99513354902346.

[51] Simbeck, Katharina. (2019). HR Analytics and Ethics. IBM Journal of Research and Development. PP. 1-1. 10.1147/JRD.2019.2915067.

[52] Sag, Matthew. (2019). The New Legal Landscape for Text Mining and Machine Learning. SSRN Electronic Journal. 10.2139/ssrn.3331606.

[53] Krishni. (2018), "K-Fold Cross Validation", Data Driven Investor: [Online] [Accessed https://medium.datadriveninvestor.com/k-fold-cross-validation-6b8518070833.