| **Module Code** | **CN 7022** |
| --- | --- |
| **Module Title** | **Big Data Analytics** |
| **LSBF Coursework Instructor** | **Dr Preethi Kesavan** |
| **Coursework Title** | **Big Data Development Project**  **Group-based Report (3000 words): 60%**  **Group-based Presentation (1000 words): 40%** |
| **Weighting** | **100%** |
| **Handout Date** | **20 February 2024** |
| **Coursework Submission Date** | **06 May 2024** |
| **Students’ Name & Students’ ID** | | **Student Name(s)** | **Student ID Number(s)** | | --- | --- | | LOW YAN TONG, GLENDA | U2651708/S1033960 | | SANTILLAN, RETXED JOSHUA | U2651168/S1033756 | | TIMUR MAMADALIYEV | U2720441/S1033860 | | ZHANG TAI | U2720444/S1033859 | |
| **Documents uploaded in LSBF Canvas:** | ⚪ Assignment Report  ⚪ Presentation Report  ⚪ Source Code, Scripts & Output  ⚪ Turnitin Report |
| **UEL Turnitin Submission Similarity Index %** | 16% |
|  | |

# Table of Content

[**Table of Content 3**](#_gjdgxs)

[**Abstract 5**](#_l755ewbrhmdt)

[**1. Executive Summary 5**](#_30j0zll)

[**2. Introduction 7**](#_1fob9te)

[2.1 Background 8](#_tsjjh592tzpj)

[2.2 Purpose 9](#_q8tbrzmw4ial)

[2.3 Aim 9](#_6iqttm5egcgv)

[2.4 Research Questions 9](#_jbqzn0nsvgk6)

[**3. Methodology 11**](#_2et92p0)

[3.1 Data Information 15](#_1r8zov14ud53)

[The information about data is provided in the Table: 15](#_vkaosp5xn0co)

[3.2 Data Cleansing 16](#_y9vwm7lmfo1i)

[3.3 Exploratory Data Analysis 19](#_79tqkpmn4l9l)

[Comparison of Total Occupation Distribution and Total Occupation with ‘Converted’ Status 20](#_jknzb0b3y1p5)

[Distribution of conversion rate depending on education level 21](#_wogw3hz9muh3)

[Distribution of conversion rate depending on previous campaign outcome 22](#_9zumihl9bnhx)

[Distribution of conversion rate depending on marital status 23](#_fq8els3yuumj)

[Distribution of conversion rate depending on occupation 24](#_r4j7l1ec4bi0)

[Comparison of ‘Converted’ and ‘Not-Converted’ by age 25](#_ffetxd5jp6uz)

[Distribution of subscriptions depending on the result of previous campaign 26](#_5orm1zeyuz0o)

[Analysis of the relationship between the duration of call and conversion status 27](#_9f4bepl27yu4)

[Relationship Between Age and Call Duration 29](#_jrmvsrapigq2)

[Distribution of Age Among Bank Clients 30](#_hkheppyxdwi)

[Feature Importance Analysis in Logistic Regression Model for Term Deposit Subscriptions 32](#_jgoremqsj2n3)

[3.4 Data Preprocessing 33](#_gz2dpd43wkw9)

[3.5 Model Training 34](#_nokzsxw8g3ll)

[3.6 Model Evaluation 35](#_xrrx94sotqt0)

[3.6.1 Random Forest Classifier 36](#_mzmi6ks2pi3j)

[3.6.2 XGBoostClassifier 38](#_aefo75qat4d9)

[3.6.3 Logistic Regression 39](#_ipu4egqymhwb)

[4.6.4 Decision Tree 43](#_bh17fqum1c1m)

[3.7 Results 46](#_annh6mq1swhi)

[3.7.1 Cross-Validation 46](#_q2tuyloax995)

[3.7.2 Hyperparameter tuning 46](#_qnge8ua9df38)

[3.8 Recommendation 47](#_d3m8mdjkvml4)

[**4. Conclusion 48**](#_32hioqz)

[**Reference 49**](#_41mghml)

[**Annex A Group Contribution 51**](#_5qo1486uv7ez)

[**Annex B Gannt Chart 52**](#_nswzk0csmcfp)

[**Annex C Agreement form 53**](#_9qg47ovt8kwx)

# 1. Executive Summary

This research project investigates the predictive capabilities of machine learning algorithms in forecasting term deposit subscriptions within the banking sector. This study aims to develop precise models for subscription prediction by utilizing call-related metrics and demographic attributes. The dataset encompasses a range of demographic factors and behavioral attributes, providing rich insights into customer behavior. By using four machine learning Decision Trees, Logistic Regression, Random Forest Classifier, and XGBoost Classifier, the research compares the performance of these algorithms in accurately predicting term deposit subscriptions. The findings of this study hold significant implications for banking institutions, aiding in the refinement of marketing strategies and the optimization of customer engagement initiatives. By using methodologies such as KDD (Knowledge Discovery in Databases) and Scrum ensures a systematic and agile approach to data analysis and project management of this coursework

# 2. Introduction

In today's banking sector, accurately predicting term deposit subscriptions is crucial for financial institutions aiming to optimize marketing strategies and enhance customer engagement initiatives (Kamble & Deshmukh, 2014). Data mining methods are instrumental in this endeavor, empowering banks to utilize demographic characteristics, behavioral metrics, and past campaign outcomes to develop predictive models for term deposit subscriptions.

Analyzing marketing strategies in the banking sector is crucial for increasing customer interactions, maintaining active transactions and driving sales through targeted campaigns (Koçoğlu, 2022). Banks are facing challenges in selling term-deposit products to new clients due to difficulties in identifying target customers and the complex factors influencing purchase decisions. Moreover, customers are increasingly unhappy about receiving irrelevant phone calls from banks. To tackle these issues, banks are utilizing their vast customer data to understand customer behavior and preferences, enhancing their marketing effectiveness. This strategic initiative is powered by artificial intelligence and Big Data technologies (Hou, S. et al., 2022).

With this study, it is aimed to obtain models with high classification performance by using machine learning algorithms with Pyspark to determine whether the bank deposit-conversion campaign will be successful and especially to determine the success of the Machine Learning approach for this problem. In this direction, different classification models have been obtained with the Decision Tree Algorithm, Logistic Regression, Random Forest Classification, Extreme Gradient Boosting Classification (XGBoost) and comparative performance evaluation has been presented. The data to work with was chosen as a representative of banking sector marketing analysis.

The Data was collected from the open-source Kaggle dataset (Hossain Y., 2024). All necessary information is provided in the Table.

| Topic | Data Marketing Campaign |
| --- | --- |
| Columns | 11 |
| Rows | 45211 |
| Size | 3912kb |
| Goal | Predict conversion status |
| Source | https://www.kaggle.com/datasets/yaminh/bank-marketing-campaign-dataset/data |

## 2.1 Background

In the paper (Elsalamony, H. A., 2014) it is shown that using machine learning classification models for bank marketing problems could be quite successful with an accuracy rate of around 90%. Provided research found that higher importance for predicting analysis was for the “Duration” feature.

Also, (Hou, S. et al., 2022) study concludes that applying machine learning (ML) algorithms to this type of dataset may not only contribute to ML literature itself, but also will provide a better understanding of the bank clients’ behaviour in purchasing financial products and services.

In his 2018 publication, Singh P. delineated the fundamental principles of leveraging PySpark, particularly in addressing supervised machine learning challenges with efficacy. Consequently, PySpark emerges as a highly viable choice for such applications (Singh P., 2018). This elucidation underscores the significant utility of PySpark as a robust tool in the domain of supervised machine learning.

## 2.2 Purpose

The purpose of this study is to investigate the effectiveness of machine learning algorithms in predicting term deposit subscriptions within the banking sector using a dataset comprising demographic attributes, behavioural metrics, and historical campaign outcomes. By leveraging this dataset, the study aims to develop accurate predictive models that can assist banking institutions in identifying potential term deposit subscribers. The findings of this research will contribute to the refinement of marketing strategies and the optimization of customer engagement initiatives within the bank.

## 2.3 Aim

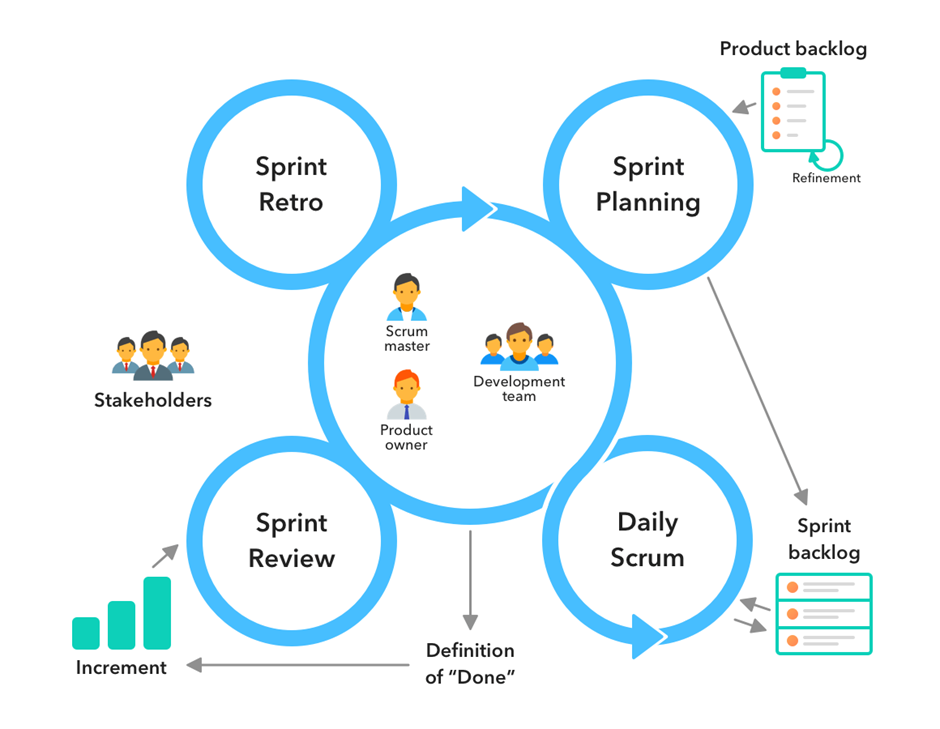
The primary objective of this research is to evaluate and compare the performance of different machine learning algorithms in accurately predicting term deposit subscriptions, utilizing a dataset containing demographic attributes, behavioural metrics, and historical campaign outcomes. Through the assessment of Decision Trees, Logistic Regression, Random Forest Classifier, and XGBoost Classifier, the study aims to identify the most suitable predictive model for term deposit subscription within the bank.

## 2.4 Research Questions

| Research Questions / Hypothesis |
| --- |
| Customers in a certain job are more likely to convert compared to others? |
| Customers in a certain age group are more likely to convert compared to others? |
| Does call duration tend to increase or decrease with customer age? |
| How do demographic factors such as age, education level, and marital status influence conversion rates? |
| Do previous campaign outcomes affect the success of current campaigns? |
| Can machine learning models accurately predict conversion outcomes based on demographic and communication data? |
| Does the frequency of calls impact customer response and conversion rates? |

# 3. Methodology

This project follows a structured methodology combining elements of the Scrum framework with Knowledge Discovery in Databases (KDD) practices to effectively analyze the dataset and develop predictive models for optimizing direct marketing campaigns in the banking sector.



Pic.1. Project methodology. Adapted from ([JORDAN JOB](https://jordanjob.me/blog/author/jordanjob/), 2015)

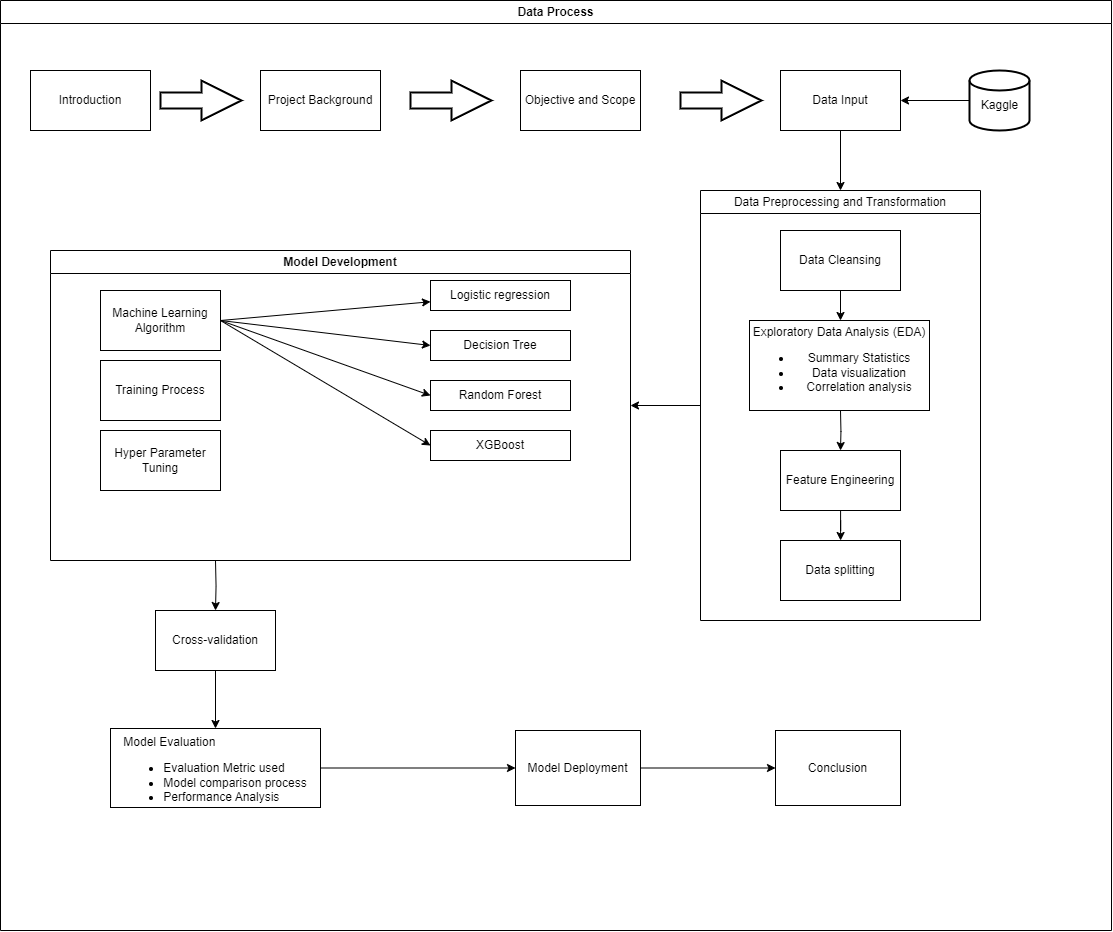
The research project is divided into iterative and incremental sprints, each typically lasting two to four weeks, within the Scrum framework. The key components of the Scrum methodology include:

Sprint Planning: At the beginning of each sprint, a sprint planning meeting is conducted so that we can define our goals and scope of the sprint. Our team collaborates to identify the tasks and activities required to achieve the sprint objectives.

Daily Stand-ups: We hold daily stand-up meetings to provide status updates, discuss progress, and identify any obstacles or challenges from each member. This ensures ongoing communication and collaboration within our team.

Sprint Review: At the end of each sprint, our team conduct a sprint review meeting to review the work completed during the sprint and gather feedback. Our team will demonstrate the outcomes achieved and discusses any adjustments needed for future sprints.

Sprint Retrospective: A sprint retrospective meeting is held to reflect on the sprint process and identify opportunities for improvement. Our team will discuss about what went well, what could be improved, and any action items for the next sprint.



Pic.2. KDD methodology. Adapted from (Fayyad, Piatetsky-Shapiro, and Smyth, 1996)

In addition to the Scrum framework, this research adopts practices from Knowledge Discovery in Databases (KDD) to systematically extract useful knowledge and insights from the dataset. The KDD process typically consists of the following stages:

i. Data Selection: Relevant data sources are identified and selected for analysis. In this research, the dataset containing information about bank clients and marketing campaigns is chosen for analysis.

ii. Preprocessing: The selected dataset is preprocessed to clean and transform the data into a suitable format for analysis. This may involve handling missing values, encoding categorical variables, and scaling numerical features.

iii. Data Mining: Data mining techniques, such as classification, clustering, and association rule mining, are applied to extract patterns and relationships from the dataset. In this research, predictive modeling techniques are used to predict client subscriptions to term deposits based on their attributes.

iv. Evaluation: The effectiveness of the data mining models is evaluated using appropriate metrics and validation techniques. This helps assess the performance and reliability of the models in predicting subscription outcomes.

v. Interpretation and Visualization: The discovered patterns and insights are interpreted and visualized to facilitate understanding and decision-making. Visualization techniques such as charts, graphs, and dashboards are used to present the findings in a clear and informative manner.

Within each sprint, the research team follows an iterative data analysis process to explore the dataset, develop predictive models, and evaluate their performance. The process involves the following steps:

**Data Exploration:** The dataset is explored to understand its structure, features, and distributions. Descriptive statistics and visualization techniques are used to gain initial insights into the data.

**Feature Engineering:** Relevant features are selected or engineered based on domain knowledge and exploratory data analysis. This may involve transforming or encoding categorical variables, handling missing values, and scaling numerical features.

**Model Development:** Predictive models, such as logistic regression, decision trees, and random forests, are developed using PySpark. The models are trained on the dataset to predict the likelihood of clients subscribing to term deposits based on their attributes.

**Model Evaluation:** The performance of the predictive models is evaluated using appropriate metrics, such as accuracy, precision, recall, and ROC curves. This helps assess the effectiveness of the models in predicting subscription outcomes.

**Iterative Refinement:** Based on the evaluation results and stakeholder feedback, the models may be refined and iteratively improved to enhance their predictive performance.

## 3.1 Data Information

## The information about data is provided in the Table:

| № | Name | Data type and values | Description |
| --- | --- | --- | --- |
| 1 | occupation | Categorical, [administrative\_staff, jobless, retired\_worker, business\_owner, manual\_worker, student, technical\_specialist, executive, service\_worker, independent\_worker, unidentified, domestic\_worker] | The occupation of the customer |
| 2 | age | Numerical | The age of the customer |
| 3 | education\_level | Categorical, [high\_school, unidentified, college, elementary\_school] | The highest education level attained by the customer |
| 4 | marital\_status | Categorical, [married, divorced, single] | The marital status of the customer |
| 5 | communication\_channel | Categorical, [unidentified, mobile, landline] | The communication channel used for contacting the customer |
| 6 | call\_month | Categorical, [January, February, March, ..., November, December] | The month of the year in which the call was made |
| 7 | call\_day | Categorical, [Monday, Tuesday, Wednesday, Thursday, Friday] | The day of the month on which the call was made |
| 8 | call\_duration | Numerical | The duration of the call in seconds |
| 9 | call\_frequency | Numerical | The frequency of calls made to the customer |
| 10 | previous\_campaign\_outcome | Categorical, [successful, unidentified, unsuccessful, other\_outcome] | The outcome of the previous marketing campaign for the customer |
| 11 | conversion\_status | Categorical, [converted, not\_converted] | Whether the customer was converted or not |

## 3.2 Data Cleansing

First of all, it’s necessary to take some time and investigate our data. Start with data type, file format, and headers. One of the most common forms of data cleanup is getting your unreadable or hard-to-read data and data types to fit a proper readable format (Jacqueline Kazil, 2016). For example, data type format values cannot be properly used in machine learning algorithms.

The next important step is to check your dataset for outliers, missing values and bad data. However, you have to think carefully about the significance of missing values (Witten, I., 2017) and depending on that importance work without rows or columns with missing values or fill them with correspondent information.

Duplicate data presents another source of error. Most machine learning tools will produce different results if some of the instances in the data files are duplicated, because repetition gives them more influence on the result (Witten, I., 2017).

Then we have to get to know our data, check and explore it.

Simple tools that show histograms of the distribution of values of nominal attributes, and graphs of the values of numeric attributes (perhaps sorted or simply graphed against instance number), are very helpful. These graphical visualizations of the data make it easy to identify outliers (Witten, I., 2017).

The next step is to implement a method called “Exploring descriptive statistics”

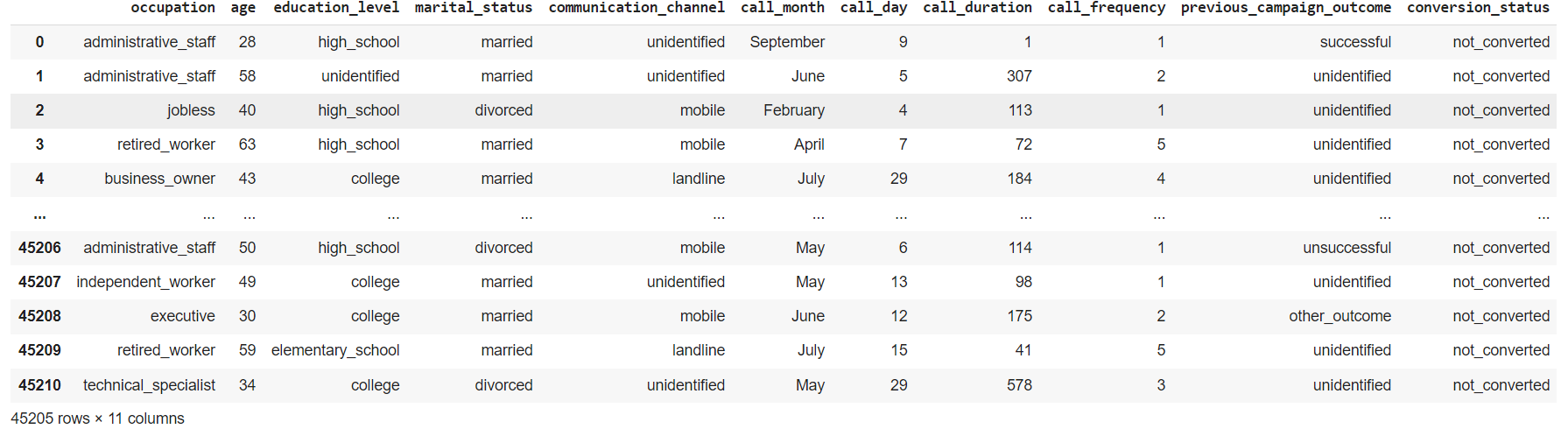
(Jacqueline Kazil, 2016). The main idea is to check numerical values with descriptive statistics, such as distribution, central tendency, dispersion, aggregation, etc (Jacqueline Kazil, 2016).

| **Explanation** | **Code + Results** |
| --- | --- |
| We identify if there are any missing values in our data. Since there is no null values in our data, we do not need to remove any rows. |  |
| We identify the number of duplicate records in our dataset and remove them. To check if the duplicate records are removed from the data frame, we check the total number of records before and after removal of duplicates |  |
| We lowercase all the letters so that it is easier to identify and use |  |

## 

## 3.3 Exploratory Data Analysis

The overall demonstration of the dataframe that was used in the project.



Pic.3. Project DataFrame

## 

## 

| Comparison of Total Occupation Distribution and Total Occupation with ‘Converted’ Status | |
| --- | --- |
|  |  |
| In our dataset, manual workers takes up the most number with 9730, followed by exclusives which is slightly lower with 9457. However, the total number of executives that converted is 1301 which is 45% higher than manual workers of 708. This shows that executives have a higher rate of conversion than manual workers. | |

| Distribution of conversion rate depending on education level |
| --- |
|  |
| As shown in the graph, the conversion rate increases with higher levels of education. Calculations reveal a strong linear correlation between the two, with a correlation coefficient of 0.98. Users with higher education levels exhibit a conversion rate of up to 15%. This phenomenon is reasonable since the dataset's target for conversion is insurance products, and a better educational background facilitates a better understanding of complex insurance terms. |

## 

| Distribution of conversion rate depending on previous campaign outcome |
| --- |
|  |
| As depicted in the graph, successful marketing campaigns lead to higher customer conversion rates, reaching as high as 64.73% based on calculations. It can be inferred that investing more resources in marketing activities to improve bank customer conversion rates is worthwhile. |

| Distribution of conversion rate depending on marital status |
| --- |
|  |
| As illustrated in the graph, the conversion rate for single individuals who have never been married is higher than that for married or divorced individuals. Single customers could be a focal point for future marketing efforts. |

| Distribution of conversion rate depending on occupation |
| --- |
|  |
| As shown in the graph, there is no significant trend in conversion rates with increasing job income. However, students, retirees, and executives exhibit much higher conversion rates compared to other industries. These three professions all possess characteristics that require "more protection." |

| Comparison of ‘Converted’ and ‘Not-Converted’ by age | |
| --- | --- |
|  |  |
| Based on the graph on the left, the highest number of converted is the age group 30-40 with 1913 callers converted. The highest number of not converted callers is also at the age group aged 30-40 with 16171 callers not converted. This is showing that there are more conversions in the 30-40 age group could simply be because there are more people in that age group overall so it does not determine that the 30-40 age group are more likely to convert. Hence, from the bar graph on the left, it is hard to tell who has the highest converted rate. However, with the graph on the right, it shows the percentage of converted individuals by age group, from here we can actually see that there is a higher number of converted people are aged 85 with 80% converted and aged 87 with 75% converted and the lowest converted percentage rate are people aged 50 with 7.67% converted. While graph on the left shows a higher number of conversions within the 30-40 age group, the right graph shows that the conversion rate for that same age group (30-40) is at 10.55%. This value falls right behind the 40-50 age group, which holds the lowest conversion rate at 9.14%. | |

## 

## 

| Distribution of subscriptions depending on the result of previous campaign |
| --- |
|  |
| As it might be seen from the illustration above, the successful previous campaign may result in a more successful present campaign. That’s a logical conclusion and something to work with as the major part of that statistic are clients of that bank. That’s why it could be useful to work more with successful and unidentified customers rather than unsuccessful. |

## 

## 

| Analysis of the relationship between the duration of call and conversion status |
| --- |
|  |
|  |
| For this analysis were used two different intervals of call duration, lower than 500 seconds and higher than 500. We can see the main trend: with low-duration calls not converted status highly dominated over converted, and with the long-duration calls directly opposite, moreover, with longer duration more successful actions.  With the call duration being lower than 500 seconds conversion rate is 7.5%, and with longer calls it increases to 42.3%. |

## 

## 

| Relationship Between Age and Call Duration |
| --- |
|  |
| This figure shows the relationship between customer age and call duration. It allows us to assess whether there is a discernible pattern or trend between the two variables. For instance, it may show that younger customers tend to have shorter call durations compared to older customers, indicating potential age-related differences in communication preferences or financial needs. Understanding these patterns can inform targeted marketing strategies tailored to different age groups. |

## 

## 

| Distribution of Age Among Bank Clients |
| --- |
|  |
| This figure shows the distribution of ages among bank clients. It shows that the majority of clients are between 30 to 40 years old, with a gradual decline in frequency as age increases or decreases from this range. This suggests that the bank's client base is skewed towards middle-aged individuals. Understanding the age distribution can help tailor marketing strategies to target specific age demographics more effectively. |

| *Comparison of Age Distribution for Term Deposit Subscribers vs. Non-Subscribers* |
| --- |
|  |
| This shows that comparing the age distribution between customers who subscribed to term deposits and those who did not. It provides a visual representation of the central tendency and spread of ages within each group. For example, it may show that converted customers tend to be older on average compared to non-converted customers, suggesting that age may be a significant factor in term deposit subscriptions. This insight can guide the targeting of marketing campaigns towards age groups more likely to subscribe. |

## 

## 

| Feature Importance Analysis in Logistic Regression Model for Term Deposit Subscriptions |
| --- |
|  |
| This Figure shows the feature importance extracted from a logistic regression model employed to forecast term deposit subscriptions in the banking sector. Each bar in the plot signifies a feature from the dataset, with its length representing the magnitude of the corresponding coefficient in the model. Positive coefficients, positioned to the right on the x-axis, indicate features positively associated with term deposit subscriptions, while negative coefficients, situated to the left, suggest a negative association. Features with coefficients closer to zero exert minimal influence. This visualization facilitates the identification of pivotal features, empowering stakeholders to refine marketing strategies and campaign targeting for enhanced efficacy. |

## 

## 3.4 Data Preprocessing

Data Preprocessing is an important step before the dataset can be used to be trained with different algorithms. It is important to remove irrelevant columns after justification to avoid discarding potentially useful information (James et al., 2013). We can also convert columns with strings into integers as numeric values allow for faster calculations compared to strings, resulting in improved training and prediction speed.

| **Explanation** | **Code + Results** |
| --- | --- |
| We convert the string values into integers to allow faster processing and is easier to compare |  |
| Selecting and dropping columns. Applying VectorAssembler and creating train and test datasets from it. |  |

## 

## 3.5 Model Training

| **Explanation** | **Code** |
| --- | --- |
| **Using RandomForestClassifier** |  |
| **Using XGBoostClassifier** |  |
| **Using Logistic Regression** |  |
| **Using Decision Tree** |  |

## 3.6 Model Evaluation

For this dataset, we will be exploring four machine learning algorithms: Random Forest Classifier, XGBoost Classifier, Logistic Regression and Decision Tree. We will be training and testing the model with these machine learning algorithms and checking which algorithms produce the best results based on the accuracy, precision, recall and F1-score.

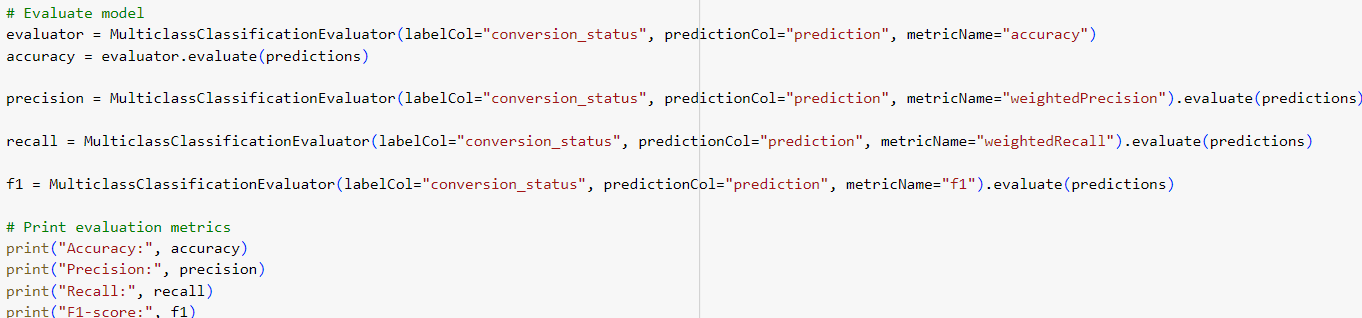
The accuracy, recall, precision, and F1-score Equation used in our model are as follows (Chaudhury et al. 2022):

### 

### ***3.6.1 Random Forest Classifier***

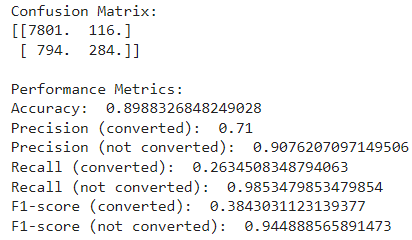
By applying random forests, several decision trees' predictions are combined to generate one final prediction (Gentsch, P, 2018). Each decision tree will use random characteristics at each split point and will be constructed using a random subset of the data which will prevent overfitting and lowering variance. (James et al. 2013).

Random Forests are good at identifying complex non-linear correlations in the data (Gentsch, P, 2018). There are complex interactions between variables that impact the conversion status, such as job, call frequency, call duration, communication channel, and previous campaign outcomes in our dataset. Therefore, we decided to explore with random forest.

****

| Metric name | Value |
| --- | --- |
| Accuracy | 0.8988326848249028 |
| Precision | 0.8839369826362716 |
| Recall | 0.8988326848249028 |
| F1-score | 0.877705562116422 |

The above shows the evaluation metrics of the random forest model. Although the results of the evaluation are high, we should check the confusion matrix to be able to understand how it performs on each conversion status (converted and not converted).



Based on the above confusion matrix, there are 7801 true positives, 116 false positives, 794 true negatives and 284 false negatives. A high number of false negatives is a concern as it means that the model missed a significant number of conversion statuses that were actually converted (26.3% based on Recall). Out of all leads predicted as converted, only 71% were actually converted. This indicates the model might be over-predicting conversions. The model on the other hand is good at identifying non-conversions with a precision of 90.76%. The model only identifies 26.34% of the actual converted status is a major problem. The model successfully identified 98.53% of the actual non-converted statuses. The F1-score (converted) of 38.4% shows that there is a poor balance between precision and Recall for converted status. The F1-score for non-converted status is high at 94.4% which suggests that there is a good performance in identifying them.

### 

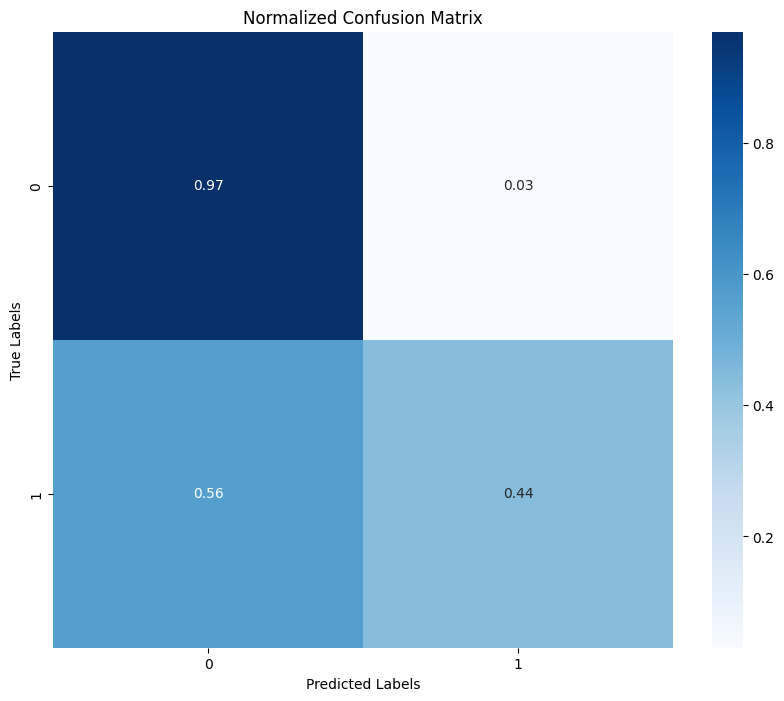
### ***3.6.2 XGBoostClassifier***

According to (Ogunleye, A. & Wang, Q.-G., 2020) Extreme Gradient Boosting algorithm or XGBoost Machine Learning model is highly accurate and convenient to use to solve classification problems even if the size of the dataset is small. Also, in their research, they proved that XGBoost could provide much better results among other popular machine learning models. XGBoost harnesses the power of parallel and distributed computing along with cache-aware algorithms and out-of-core computing to handle large datasets efficiently. Its flexibility is evident in its support for a range of objectives functions and customizable evaluation metrics. Additionally, XGBoost offers features such as handling missing values, producing interpretable models, and incorporating built-in cross-validation capabilities (Lin, J., 2024).

Evaluation of the model:

| Metric name | Value |
| --- | --- |
| Accuracy | 0.9034340671971706 |
| Precision | 0.8950337780261439 |
| Recall | 0.9034040671971707 |
| F1-score | 0.8979372914341456 |

All the metrics are around 90%, so we can say about successful model implementation. The optimization of the model may increase its quality but at the same time the iteration over the number of different parameters may take much more calculation power.



Pic.. XGBoost confusion matrix

According to the confusion matrix, the following statements could be concluded: i. The model almost perfectly predicted the non-conversion status of the campaign. ii. Conversion results were predicted with a roughly 50% of accuracy.

The algorithm could be improved in 2 ways: i. Improve the model by iteration process of seeking better parameters. ii. Use more data for model training.

### 

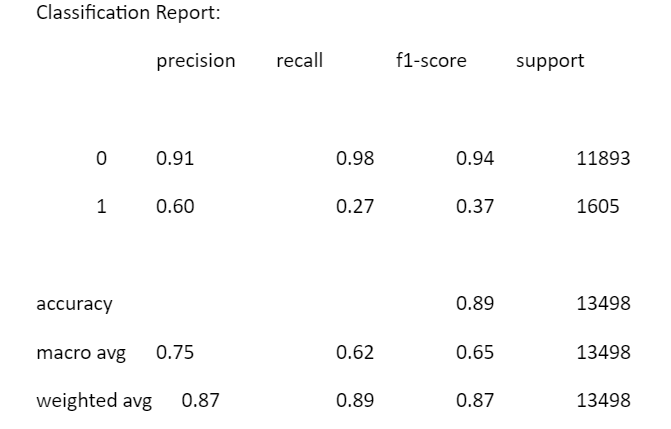
### ***3.6.3 Logistic Regression***

Logistic regression is a widely used statistical technique for binary classification tasks, particularly in fields such as finance and banking (Field, Miles, & Field, 2012). It estimates the probability of a binary outcome based on one or more predictor variables (Hosmer, Lemeshow, & Sturdivant, 2013). In the context of our dataset, logistic regression offers a valuable tool for predicting term deposit subscriptions within the banking sector. By analyzing demographic attributes, behavioral metrics, and historical campaign outcomes, logistic regression models can provide insights into customer behaviour and help optimize marketing strategies (Smith, 2018).

The study done by (Moro, Cortez, and Rita, 2014) investigated data mining approaches for modeling complex systems, particularly focusing on the banking sector. They utilized logistic regression as one of the predictive modeling techniques to forecast term deposit subscriptions in Portuguese banking institution. The research dataset encompassed demographic attributes, banking transaction history, and marketing campaign outcomes. By comparing logistic regression with other machine learning algorithms, such as decision trees and random forest, the study revealed that logistic regression demonstrated competitive performance in terms of accuracy and interpretability. Notably, logistic regression's interpretability, attributed to its ability to provide interpretable coefficients for each predictor variable, proved advantageous in banking applications, where stakeholders often require explanations for model predictions. Overall, the study by Moro et al. (2014) provided empirical evidence supporting the effectiveness of logistic regression for forecasting term deposit subscriptions in the bank.

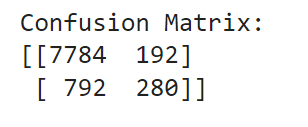
In our study, logistic regression will serve as one of the machine learning algorithms utilized to predict term deposit subscriptions based on customer attributes and historical campaign data extracted from our banking dataset. By analyzing the logistic regression model coefficients, we aim to discern the most influential factors affecting subscription likelihood, thereby informing targeted marketing strategies and customer engagement initiatives within the bank.

| Metric name | Value |
| --- | --- |
| Accuracy | 0.8915394873314565 |
| Precision | 0.8711207103103861 |
| Recall | 0.8915394873314566 |
| F1-score | 0.8726310813118323 |



The model achieved an accuracy of 89.15%, indicating its ability to correctly predict the outcome for 89.15% of the cases in the test dataset.

The Area Under the Receiver Operating Characteristic Curve (AUC-ROC) is 88%, suggesting a moderate ability of the model to distinguish between the two classes (converted vs. not converted).



**Confusion Matrix:**

True Negative (TN) = 7784: The number of correctly predicted non-conversions.

False Positive (FP) = 192: The number of non-conversions incorrectly predicted as conversions.

False Negative (FN) = 792: The number of conversions incorrectly predicted as non-conversions.

True Positive (TP) = 280: The number of correctly predicted conversions.

**Precision, Recall, and F1-Score:**

**Class 0 (Non-conversions):**

Precision: 91%

Recall: 98%

F1-Score: 94%

**Class 1 (Conversions):**

Precision: 60%

Recall: 27%

F1-Score: 37%

**Overall Scores:**

Weighted Average F1-Score:87%

The model exhibits strong performance in correctly identifying non-conversions, with high precision and recall for class 0. However, it struggles with classifying conversions, as evidenced by lower precision and recall for class 1.

The ROC AUC score indicates moderate discriminative ability of the model, suggesting potential areas for improvement to enhance its predictive performance.

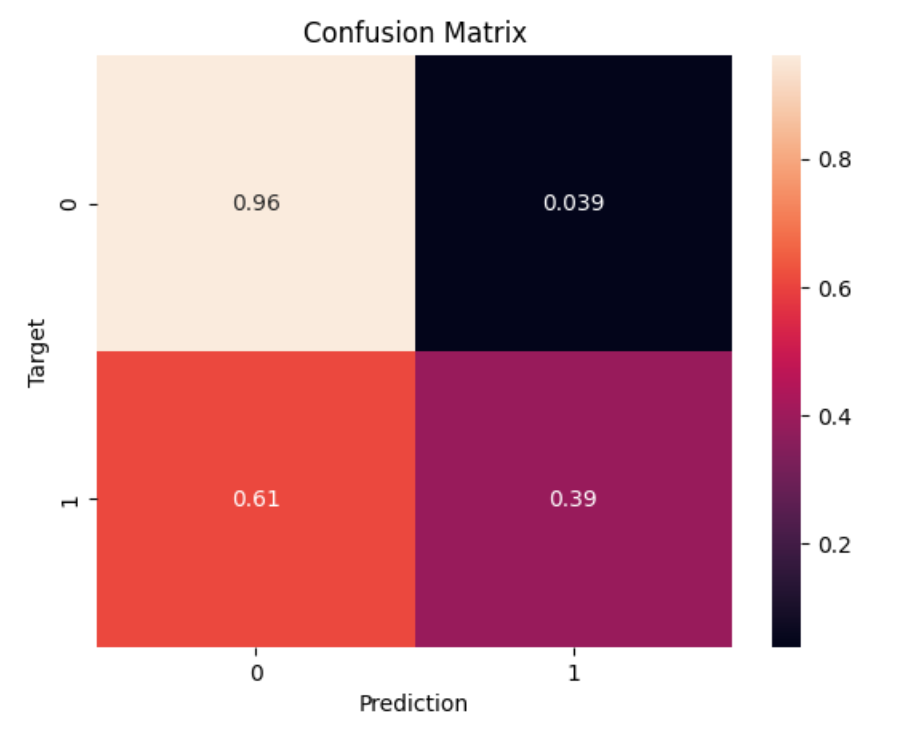
In summary, while the model shows promise in certain aspects, there is scope for refinement to achieve better balance in predicting both conversions and non-conversions accurately. Continued efforts in model optimization and feature engineering may lead to enhanced performance and better alignment with the objectives of predicting term deposit subscriptions in the bank.

### ***4.6.4 Decision Tree***

In analyzing bank customer data, it is important to consider various attributes of the customers, such as occupation, age, gender, education level, marital status, and other characteristics (Almeida, F., Ferreira, P. M., & Gama, J. 2012). Decision tree models can handle different types of data, including both continuous and categorical features (Zhao, Z., Xu, T., Zhang, S., & Liu, Y. 2020). This makes them advantageous for dealing with mixed data types. Additionally, decision tree models have a good tolerance for missing values and outliers in the data, thereby reducing the need for extensive data preprocessing. (Tan, P. N., Steinbach, M., & Kumar, V.2019). This saves time and effort for analysts.



| Metric name | Value |
| --- | --- |
| Accuracy | 0.8924377126550324 |
| Precision | 0.8791628119993118 |
| Recall | 0.8924377126550324 |
| F1-score | 0.8832820440907566 |



In the analysis of bank customer conversion prediction data, we observed that the model achieved an accuracy of 89.24%, indicating satisfactory overall classification performance. Under the condition of average='weighted', the model's accuracy is 0.88, and the recall is 0.89. Compared to the condition of average='micro', where the model's accuracy is 0.45, and the recall is 0.48. The difference between the two is significant, indicating an imbalance in the status of the data label column.From the confusion matrix data, it can be inferred that the model performed well in identifying non-converting customers, reaching 96.08%.This model is suitable for identifying and excluding customers with a low likelihood of conversion, aiming to reduce the bank's expenditure on conversion costs.

## 

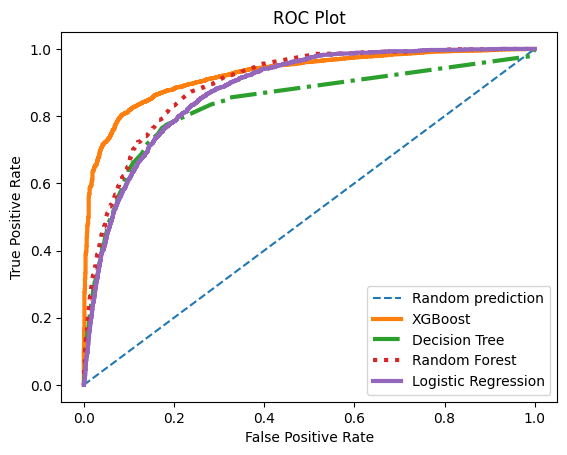
## 3.7 Results

### 3.7.1 Cross-Validation

The following are the summarized results of all four machine learning algorithms we tested in percentage, rounded up to 2 decimal places:

|  | **Accuracy** | **Precision** | **Recall** | **F1-Score** |
| --- | --- | --- | --- | --- |
| ***Random Forest Classifier*** | 89.88% | 88.39% | 89.88% | 87.77% |
| ***XGBoostClassifier*** | 90.34% | 89.50% | 90.34% | 89.79% |
| ***Logistic Regression*** | 89.15% | 87.11% | 89.15% | 87.26% |
| ***Decision Tree*** | 89.24% | 87.91% | 89.24% | 88.32% |

From this summarized results table, we can see that XGBoostClassifier has the highest accuracy, precision, recall, and F1-score, followed by Random Forest Classifier, then Decision Tree, and lastly, Logistic Regression.



Pic.. ROC comparison of ML models

According to the ROC plot, the XGBoost model demonstrated better results than others, especially in sensitivity rate. Followed by Random Forest and Logistic Regression with almost the same results, with a bit better sensitivity for Random Forest. The last one is the Decision Tree with the poorest result of True Positive Rate

### 3.7.2 Hyperparameter tuning

The following are the metric results of the best model after the hyperparameter tuning for all 4 machine learning models rounded up to 2 decimal places.

|  | **Accuracy** | **Precision** | **Recall** | **F1-Score** |
| --- | --- | --- | --- | --- |
| ***Random Forest Classifier*** | 89.82% | 88.34% | 89.82% | 88.54% |
| ***XGBoost Classifier*** | 90.52% | 89.68% | 90.52% | 89.96% |
| ***Logistic Regression*** | 89.12% | 86.96% | 89.12% | 86.84% |
| ***Decision Tree*** | 89.63% | 87.92% | 89.63% | 87.58% |

The empirical data presented in the table reveals that hyperparameter tuning yielded improvements in results for only a subset of cases, while concurrently demanding significantly greater time and computational resources to complete. This observation underscores the nuanced trade-offs involved in hyperparameter optimization, where the potential gains must be weighed against the increased computational burden.

## 3.8 Recommendation

As the summary of the project could be made next insights:

i. The analysis of the project leads to the conclusion that the marketing campaign should be targeted towards individuals who did not express opposition to conversion in the previous campaign. This targeted approach can improve conversion rates by focusing efforts on receptive segments of the audience.

ii. The duration of customer calls emerges as a significant factor influencing outcomes. Consequently, call managers could benefit from guidance to prolong conversations and engage customers effectively, as longer call durations tend to correlate positively with successful outcomes.

iii. Utilizing the XGBoost model proves advantageous for predicting non-conversion status within the campaign. However, further refinement is necessary to enhance the model's performance across other scenarios, highlighting the ongoing need for model improvement and optimization.

## 

# 4. Conclusion

In conclusion, through our analysis, we've uncovered actionable insights. Our implementation of the KDD methodology and utilization of Apache Spark alongside Python libraries have been instrumental in processing and analyzing vast datasets, enabling us to derive meaningful patterns and trends. While our initial foray into machine learning yielded promising results, it's evident that our models would benefit from further refinement and optimization. The need for a larger dataset is also apparent to enhance the robustness and generalization of our models. Moving forward, our project roadmap includes plans for improving our existing models, acquiring a larger dataset for training and testing, and exploring advanced techniques to boost predictive accuracy and scalability. Additionally, ongoing skill development and exploration of emerging technologies will remain integral to our project's success.

# 

# Reference

Almeida, F., Ferreira, P. M., & Gama, J. (2012). Predicting Customer Lifetime Value in the Banking Sector. Expert Systems with Applications, 39(12), 11466-11473.

Chaudhury, M., A. Karami, and M. A. Ghazanfar. Large-Scale Music Genre Analysis and Classification Using Machine Learning with Apache Spark. MDPI, 2022. Print. [18 April 2024]

DENNY, D. T. L. (2018). PYSPARK cookbook: Over 60 recipes for implementing big data processing and analytics using Apache ... spark and python. PACKT Publishing Limited.

Elsalamony, H. A. (2014). Bank direct marketing analysis of data mining techniques. International Journal of Computers and Applications, 85, 12–22.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. 1996. From Data Mining to Knowledge Discovery: An Overview. In: Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., & Uthurusamy, R. (eds.), Advances in Knowledge Discovery and Data Mining. AAAI Press / The MIT Press, Menlo Park, CA, pp. 1-54.

Field, A., Miles, J., & Field, Z. (2012). Discovering statistics using R. Sage.

Gentsch, P 2018, AI in Marketing, Sales and Service : How Marketers Without a Data Science Degree Can Use AI, Big Data and Bots, Springer International Publishing AG, Cham. Available from: ProQuest Ebook Central, 135-136 [18 April 2024].

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied logistic regression. John Wiley & Sons.

Hossain Y. (2024) 'Bank Marketing Campaign Dataset'. Available at: <https://www.kaggle.com/datasets/yaminh/bank-marketing-campaign-dataset/data> (Accessed April 2024).

Hou, S. et al. (2022) Applying Machine Learning to the Development of Prediction Models for Bank Deposit Subscription. *International journal of business analytics*. [Online] 9 (1), 1–14.

Infinity. (n.d.). Scrum Methodology: an agile framework for project management. Retrieved from<https://startinfinity.com/project-management-methodologies/scrum>

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112). Springer, 334-337 [18 April 2024]

Kamble, S. S., & Deshmukh, D. (2014). Application of data mining techniques in customer relationship management: A literature review and classification. International Journal of Information Technology and Management, 13(2), 154-170.

Kazil, J. & Jarmul, K. (2016) *Data Wrangling with Python: Tips and Tools to Make Your Life Easier*. 1st edition. Sebastopol: O’Reilly Media, Incorporated.

Koçoğlu, F. Ö. & Esnaf, Ş. (2022) Machine Learning Approach and Model Performance Evaluation for Tele-Marketing Success Classification. *International journal of business analytics*. [Online] 9 (5), 1–18.

Lin, J. (2024) Research on loan default prediction based on logistic regression, randomforest, xgboost and adaboost. *SHS Web of Conferences*. [Online] 1812008-.

Ogunleye, A. & Wang, Q.-G. (2020) XGBoost Model for Chronic Kidney Disease Diagnosis. IEEE/ACM transactions on computational biology and bioinformatics. [Online] 17 (6), 2131–2140.

Tan, P. N., Steinbach, M., & Kumar, V. (2019). Introduction to Data Mining (2nd ed.). Pearson. Boston, MA: Addison-Wesley.

Trochim, P. W. M. K. (n.d.). *Descriptive statistics*. Research Methods Knowledge Base. https://conjointly.com/kb/descriptive-statistics/#dispersion

Scrum.org. (n.d.). Scrum Framework. Retrieved from<https://www.scrum.org/resources/scrum-framework>

Singh, P. (2018) Machine Learning with Pyspark: With Natural Language Processing and Recommender Systems. 1st ed. [Online]. Berkeley, CA: Apress L. P.

Smith, J. (2018). Predictive modeling in banking: the role of logistic regression. Journal of Banking & Finance, 42(11), 179-192.

Witten, I. (2017) *Data mining : practical machine learning tools and techniques*. Fourth edition. Amsterdam, [Netherlands: Elsevier.]

Zhao, Z., Xu, T., Zhang, S., & Liu, Y. (2020). A Hybrid Model for Predicting Customer Churn in Banking Industry Using Decision Tree and XGBoost. *Expert Systems with Applications*, 139, 112845. Amsterdam: Elsevier.

# Annex A Group Contribution

| **LOW YAN TONG, GLENDA** |
| --- |
| Random Forest Classifier Algorithm |
| Data Cleansing (Table for explanation, code + results) |
| Exploratory Data Analysis   * Comparison of Total Occupation Distribution and Total Occupation with ‘Converted’ Status * Comparison of ‘Converted’ and ‘Not-Converted’ by age |
| Data Preprocessing |
| Model Evaluation   * Description + Formula |
| Gannt Chart |

# 

| **SANTILLAN, RETXED JOSHUA** |
| --- |
| Logistic Regression Algorithm |
| Executive summary |
| Purpose & Aim |
| Exploratory Data Analysis |
| Model Evaluation   * Description + Formula |
| Research Question |

| **TIMUR MAMADALIYEV** |
| --- |
| XGBoost Classifier Algorithm |
| Introduction, background |
| Data information |
| Data cleansing (Text description) |
| Exploratory Data Analysis Distribution of subscriptions depending on the result of the previous campaignAnalysis of the relationship between the duration of the call and conversion status |
| Results, Cross-validation, Hyperparameter tuning |
| Recommendation |
| Conclusion |

| **ZHANG TAI** |
| --- |
| Decision Tree Algorithm |
| Exploratory Data Analysis Distribution of conversion rate depending on education levelDistribution of conversion rate depending on previous campaign outcomeDistribution of conversion rate depending on marital statusDistribution of conversion rate depending on occupation |
| Model Evaluation(Decision Tree) |
| Cross-validation(Decision Tree) |
| Hyperparameter tuning(Decision Tree) |

# Annex B Gannt Chart

# 

# Annex C Agreement form

Please complete this agreement and keep a copy for each member of your group. The original of this agreement goes to your Tutor and the Electronic copy goes in your coursework.

We agree to work as a group (a group of minimum 2-max 4) to complete the coursework for CN7022 and understand that the grade awarded will be the grade allocated to us individually because of our group work.

| **Student No.** | **Name (block letters) and email Address** | **Signature** |
| --- | --- | --- |
| U2651708/S1033960 | LOW YAN TONG, GLENDA  S1033960@students.lsbf.edu.sg |  |
| U2651168/S1033756 | SANTILLAN, RETXED JOSHUA  S1033756@students.lsbf.edu.sg |  |
| U2720441/S1033860 | TIMUR MAMADALIYEV  S1033860@students.lsbf.edu.sg |  |
| U2720444/S1033859 | ZHANG TAI  S1033859@students.lsbf.edu.sg |  |

Instructor’s Name: Preethi Kesavan

Date of agreement : 23 April 2024 **Note: Students should form their groups (a group of minimum 2 or max 4) within the SAME class.**