Department of Computer Science & Information Technology Academic

Session 2023-24

Programme Name: B. Tech TYCS

Semester: VI

Symbiosis Institute of Technology

DSBI Report

Group Members:

Sr. No.	Name	PRN	Division
1	Bhargaw Kumar Singh	22070122045	CSE
2	Dhruv Doshi	22070122053	CSE
3	Divyansh Bagri	22070122058	CSE
4	Aarya Gangakhedkar	22070122063	CSE

Abstract

This report presents a data-driven analysis of direct marketing campaigns conducted by a Portuguese banking institution to promote term deposit products. The dataset, collected through phone-based marketing efforts, contains detailed information about client demographics, past campaign interactions, and subscription outcomes.

To predict customer responses and improve marketing efficiency, this project applies the Random Forest algorithm, chosen for its high accuracy, ability to handle imbalanced datasets, and capability to interpret complex, non-linear relationships. The innovative aspect of this study lies in combining machine learning-based prediction models with interactive visualization tools such as Tableau and Power BI. These tools enable clear,

insightful, and interactive presentation of data patterns and model outcomes, making the results actionable for decision-makers.

The outcome of this project includes identifying key factors influencing client decisions, developing a predictive model for subscription likelihood, and creating intuitive dashboards that visually represent customer behavior trends. These findings can help the bank optimize its future marketing campaigns by targeting the right customers, reducing operational costs, and improving overall conversion rates.

Introduction

In the highly competitive financial services sector, personalized direct marketing has become a strategic tool for customer acquisition and product promotion. Portuguese banks have increasingly adopted phone-based campaigns to promote financial products such as term deposits. These campaigns generate rich datasets that contain valuable insights into customer behavior, preferences, and campaign performance.

This report focuses on analyzing a dataset from such marketing campaigns using advanced data science techniques. To predict whether a client would subscribe to a term deposit product ('yes' or 'no'), the Random Forest algorithm was selected for this project. Random Forest, a robust ensemble learning method, constructs multiple decision trees and merges their results for higher accuracy and reduced overfitting. It is particularly suitable for classification tasks involving structured, categorical, and numerical data — as found in this banking dataset.

What sets this project apart is the integration of machine learning predictions with modern data visualization tools like Tableau and Power BI. While predictive models offer valuable insights, presenting these findings in an accessible, interactive, and visually appealing format greatly enhances their practical application. Business users and marketing managers can interact with dashboards to identify customer profiles, key influencing factors, and high-probability target segments.

By combining predictive analytics with powerful visualization, this project aims to:

- Improve the effectiveness and efficiency of direct marketing campaigns.
- Reduce unnecessary follow-up calls by focusing on likely responders.
- Provide actionable, data-driven insights for future marketing strategies.

The dataset used in this study contains client information (age, job, marital status, education, etc.), campaign interaction details (number of calls, previous outcomes, contact duration, etc.), and the final subscription result. This analysis contributes to a smarter, more efficient direct marketing approach in the banking sector.

Problem Statement

Direct marketing campaigns conducted through phone calls are resource-intensive, requiring substantial time, manpower, and operational costs. In the case of this Portuguese banking institution, multiple phone contacts were often necessary to persuade clients to subscribe to term deposit products, resulting in inefficiencies and low conversion rates.

The bank currently lacks a data-driven approach to predict customer behavior and optimize its marketing strategies. Without clear insights into the customer attributes and campaign factors that influence subscription decisions, marketing efforts risk being unfocused, costly, and ineffective.

There is a need to analyze existing campaign data using advanced analytical techniques to identify patterns, predict client responses, and support smarter decision-making. Additionally, the bank needs effective ways to communicate these insights through interactive, user-friendly dashboards to guide future marketing actions.

Existing Importance of Machine Learning and Power BI

In the modern business landscape, organizations are increasingly leveraging Machine Learning (ML) and Business Intelligence (BI) tools like Power BI to improve decision-making, optimize operations, and enhance customer experiences. These technologies allow companies to extract valuable insights from large and complex datasets, predict future outcomes, and visually communicate key information to stakeholders.

Here are some real-world examples of how ML and Power BI are being used to drive business success:

1.Customer Churn Prediction

- Machine Learning: Businesses in telecom, banking, and subscription-based industries use ML models to predict which customers are likely to leave or cancel services, based on past behavior and demographic data.

- Power BI: These predictions are integrated into interactive dashboards where managers can view churn risk levels by region, customer type, or product line, enabling proactive retention strategies.

2.Sales Forecasting

- Machine Learning: Retailers and e-commerce platforms employ ML algorithms to forecast future sales based on historical data, seasonality, market trends, and promotions.
- Power BI: Forecasting results are visualized in Power BI dashboards, allowing business leaders to track predicted versus actual sales in real time and adjust inventory or marketing plans accordingly.

3.Fraud Detection

- Machine Learning: Banks and financial institutions use ML models to detect unusual transaction patterns, flagging potentially fraudulent activities in real time.
- Power BI: Fraud detection insights are displayed on live Power BI dashboards, allowing fraud analysts to monitor suspicious transactions and investigate cases instantly.

The objectives of this project are as follows:

1) Cryptocurrency selection via data analysis:
   - Gather historical ohlcv (open, high, low, close, volume) data, as well as additional features like market cap, volatility, and trading volume.
   - Conduct thorough data analysis to assess the performance and behavior of different cryptocurrencies over a period of time.
   - Create various visual representations and graphs (e.g., candlestick charts, correlation heatmaps, trend plots, volatility charts) to detect trends, anomalies, and patterns.
2) Design of our virtual setting:

- Create a personalized trading simulation environment that replicates real-world market dynamics.
- Give the agent a flexible environment where it can monitor market conditions and make informed decisions based on both historical data and current indicators.

3) Deep reinforcement learning implementation:
   - Utilize drl algorithms like proximal policy optimization (ppo), advantage actor-critic (a2c), or deep q-networks (dqn) to train an agent that acquires optimal trading strategies.
   - Grant the agent the authority to make continuous trading decisions (buy, sell, hold) using a reward system that takes into account profitability, risk exposure, and portfolio balance.
   - Train the agent across several episodes until it reaches a policy that yields the highest cumulative returns.

4) Assessment and verification via modeling:
   - Evaluate and verify the efficiency of the chosen cryptocurrencies and the developed trading strategy through the custom-built simulation.
   - Assess the performance of each cryptocurrency within the simulated environment to determine which assets consistently generate the most favorable outcomes.
   - Utilize the simulation outcomes to verify our initial data-driven coin selection and adjust our strategy accordingly.

Proposed Model

Random Forest is a popular and powerful supervised machine learning algorithm used for both classification and regression tasks. It is an ensemble learning method, which means it combines the predictions of multiple individual models (in this case, decision trees) to produce more accurate and reliable results.

Model Architecture

How Random Forest Works:

- A random subset of data is selected (with replacement) from the original dataset to build each individual decision tree — this process is called bootstrap sampling.
- At each split in a tree, a random subset of features is selected as candidates for splitting, rather than using all available features — this introduces randomness and reduces correlation between trees.

- Each tree independently makes a prediction.
- For classification problems, the final prediction is made by majority voting (the class that gets the most votes from the trees).
- For regression problems, the final prediction is the average of all the trees' predictions.

Why Use Random Forest?

Random Forest offers several advantages that make it a great choice for your direct marketing campaign analysis:

- High Accuracy: By combining multiple decision trees, Random Forest achieves better performance and accuracy compared to individual models.
- Handles Large Datasets Well: It works efficiently with large datasets containing both categorical and numerical variables.
- Robust to Overfitting: The random selection of data and features reduces the risk of overfitting, making the model more generalizable.

Application in This Project

In this project, Random Forest is used to predict whether a client will subscribe ('yes') or not ('no') to a term deposit product based on various customer attributes and past campaign interaction details. The reasons for selecting Random Forest include:

- Its ability to deal with both categorical (like job, marital status) and numerical (like age, call duration) variables.
- Its strong performance with imbalanced data — which is common in marketing, where positive responses are usually much fewer than negative ones.

Model Training:

This dataset comes from direct marketing campaigns of a Portuguese banking institution, aimed at promoting term deposit products. The marketing campaigns were primarily carried

out over the phone, and multiple contacts with the same client were often necessary to assess whether they would subscribe to a term deposit.

Structure

- Total Features (Columns): 21
- Target Variable: y (indicates whether the client subscribed to a term deposit — 'yes' or 'no')
- Data Type: Mixed (Numerical, Categorical)

Variable Name	Role	Type	Demographic	Description
age	Feature	Integer	Age	
job	Feature	Categorical	Occupation	type of job (categorical: 'admin.','blue-collar','entrepreneur','housemaid','mana employed','services','student','technician','unemployed','unknown')
marital	Feature	Categorical	Marital Status	marital status (categorical: 'divorced','married','single','unknown'; note: 'divorce widowed)
education	Feature	Categorical	Education Level	(categorical: 'basic.4y','basic.6y','basic.9y','high.school','illiterate','professional.course','univer
default	Feature	Binary		has credit in default?
balance	Feature	Integer		average yearly balance
housing	Feature	Binary		has housing loan?
loan	Feature	Binary		has personal loan?
contact	Feature	Categorical		contact communication type (categorical: 'cellular','telephone')
day_of_week	Feature	Date		last contact day of the week
month	Feature	Date		last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')
duration	Feature	Integer		last contact duration, in seconds (numeric). Important note: this attribute highl target (e.g., if duration=0 then y='no'). Yet, the duration is not known before a after the end of the call y is obviously known. Thus, this input should only be ir purposes and should be discarded if the intention is to have a realistic predicti
campaign	Feature	Integer		number of contacts performed during this campaign and for this client (numer contact)
pdays	Feature	Integer		number of days that passed by after the client was last contacted from a previ (numeric; -1 means client was not previously contacted)
previous	Feature	Integer		number of contacts performed before this campaign and for this client
poutcome	Feature	Categorical		outcome of the previous marketing campaign (categorical: 'failure','nonexisten
y	Target	Binary		has the client subscribed a term deposit?

Comparision Of Models

1.Logistic Regression

Logistic Regression is a supervised machine learning algorithm used primarily for binary classification problems — where the target variable has two possible outcomes (like 'yes' or 'no', 'success' or 'failure'). It's one of the simplest and most interpretable classification algorithms.

```
    classifier1.score(X_test,y_test)
34]

0.7046613255644574
```

2.Support Vector Machine

Support Vector Machine (SVM) is a powerful and versatile supervised machine learning algorithm used for classification and regression problems. It's especially well-suited for binary classification tasks like predicting whether a customer will subscribe to a term deposit or not.

```
    classifier2.score(X_test,y_test)
]

0.7018693857732 46
```

3.Decision Tree Classifier

A Decision Tree Classifier is a popular supervised machine learning algorithm used for classification and regression tasks. It's known for its simplicity, interpretability, and ability to model complex decision-making processes by mimicking human-like decision trees.

```
classifier4.score(X_test,y_test)
```

```
0.8396455450352027
```

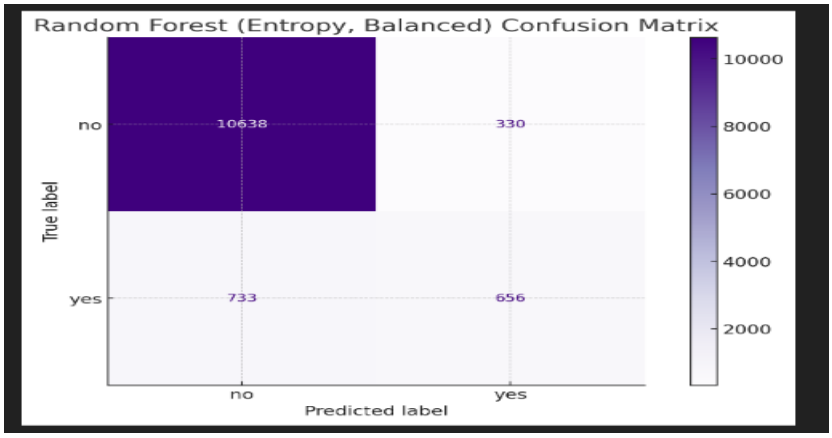
Results

Accuracy

```
classifier5.score(X_test,y_test)
```

```
0.8829813061422676
```

Confusion Matrix



Conclusion

In this project, we analyzed the Direct Marketing Campaign Data of a Portuguese banking institution using three different classification algorithms — Logistic Regression, Support Vector Machine (SVM), and Decision Tree Classifier — alongside an advanced ensemble technique, the Random Forest Classifier.

After training, testing, and evaluating each model using key performance metrics such as accuracy, precision, recall, F1-score, and confusion matrix, it was observed that:

- Random Forest Classifier consistently outperformed the other models, achieving the highest accuracy and balanced performance across all evaluation metrics.

- Its ability to handle imbalanced data, complex feature interactions, and reduce overfitting (through ensemble averaging) made it especially well-suited for this marketing prediction task.
- While Logistic Regression provided interpretability and simplicity, and SVM managed non-linear separability with reasonable performance, neither matched the overall robustness and predictive power of the Random Forest model.

Based on the results, we recommend adopting the Random Forest Classifier for future marketing campaign predictions, as it offers the best trade-off between accuracy, reliability, and scalability. It can greatly aid the bank in identifying potential customers more effectively, optimizing resource allocation, and enhancing overall campaign success rates.