Portuguese Bank Marketing Campaign Prediction REPORT

Introduction

The banking industry increasingly relies on data-driven approaches to improve the effectiveness of marketing campaigns. Term deposits are an important source of revenue for banks, but promoting them through mass marketing can be costly and inefficient. This project focuses on analyzing customer data from a Portuguese bank's direct marketing campaigns to understand customer behavior and build predictive models that identify customers who are more likely to subscribe to a term deposit.

Problem Statement

The objective of this project is to analyze historical bank marketing data and build a predictive model that helps the marketing team identify potential customers who are likely to subscribe to a term deposit. The goal is to improve campaign efficiency, reduce unnecessary customer contacts, and increase the overall subscription rate.

Task 1:-Prepare a complete data analysis report on the given data.

Task 2:-Create a predictive model which will help the bank marketing team to know which customer will buy the product.

Task3:-Suggestions to the Bank market team to make customers buy the product.

Dataset Description

The dataset contains information from direct phone marketing campaigns conducted by a Portuguese banking institution between 2008 and 2010. It includes customer demographic details, financial attributes, campaign-related variables, and macroeconomic indicators. The target variable indicates whether the customer subscribed to a term deposit. The dataset consists of both numerical and categorical features, making it suitable for classification analysis.

Attribute Information:

Input variables:

1 - age (numeric)

2 - job : type of job (categorical: 'admin.','blue-collar','entrepreneur','housemaid','management','retired','self-employed','services','student','technician','unemployed','unknown')

3 - marital : marital status (categorical: 'divorced','married','single','unknown'; note: 'divorced' means divorced or widowed)

4 - education (categorical: 'basic.4y','basic.6y','basic.9y','high.school','illiterate','professional.course','university.degree','unknown')

5 - default: has credit in default? (categorical: 'no','yes','unknown')

6 - housing: has housing loan? (categorical: 'no','yes','unknown')

7 - loan: has personal loan? (categorical: 'no','yes','unknown')

related with the last contact of the current campaign:

8 - contact: contact communication type (categorical: 'cellular','telephone')

9 - month: last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')

10 - day_of_week: last contact day of the week (categorical: 'mon','tue','wed','thu','fri')

11 - duration: last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y='no'). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.

other attributes:

12 - campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)

13 - pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)

14 - previous: number of contacts performed before this campaign and for this client (numeric)

15 - poutcome: outcome of the previous marketing campaign (categorical: 'failure','nonexistent','success')

social and economic context attributes

16 - emp.var.rate: employment variation rate - quarterly indicator (numeric)

17 - cons.price.idx: consumer price index - monthly indicator (numeric)

18 - cons.conf.idx: consumer confidence index - monthly indicator (numeric)

19 - euribor3m: euribor 3 month rate - daily indicator (numeric)

20 - nr.employed: number of employees - quarterly indicator (numeric)

Output variable (desired target):

21 - y - has the client subscribed a term deposit? (binary: 'yes','no')

Exploratory Data Analysis (EDA)

Exploratory Data Analysis was conducted to understand the data structure and uncover patterns influencing customer subscription behavior. The analysis revealed a strong class imbalance, with most customers not subscribing to term deposits. Financially stable customers, such as those without loans or credit defaults, showed higher subscription rates. Retired and student customers were more likely to subscribe, and cellular contact proved more effective than telephone contact. Previous successful campaign outcomes were found to be the strongest predictor of future subscriptions. The duration feature was excluded from modeling to avoid data leakage.

1. Overview of the Dataset

Exploratory Data Analysis was performed to understand the structure, quality, and characteristics of the Portuguese bank marketing dataset. The dataset consists of customer demographic details, financial information, marketing campaign attributes, and macro-economic indicators. Each record represents a customer contacted during a direct marketing campaign, and the target variable indicates whether the customer subscribed to a term deposit.

The dataset contains both numerical and categorical variables, making it suitable for comprehensive exploratory analysis to identify patterns, trends, and relationships influencing customer subscription behavior.

---

2. Data Structure and Quality Assessment

The dataset was examined to understand its size, data types, and completeness. No missing (null) values were found in the dataset. However, certain categorical features contain the value `"unknown"`, which represents a lack of information rather than missing data. These values were retained as they carry business significance and may influence customer decisions.

The dataset includes:

- Demographic attributes such as age, job, marital status, and education
- Financial attributes such as loan status and credit default
- Campaign-related variables such as contact type, number of calls, and past outcomes
- Economic indicators such as interest rates and employment levels

---

3. Target Variable Analysis

The target variable represents whether a customer subscribed to a term deposit. Analysis of the target variable revealed a significant class imbalance, with the majority of customers not subscribing to the product. This imbalance reflects a real-world banking scenario where only a small fraction of contacted customers convert successfully.

Due to this imbalance, accuracy alone is not sufficient for evaluating predictive models. Metrics such as precision, recall, F1-score, and ROC-AUC are more appropriate for assessing model performance.

---

4. Analysis of Customer Age

Age distribution analysis showed that most customers fall within the working-age group of 30 to 50 years. However, subscription rates were observed to be higher among older customers, particularly retirees. Younger customers, especially students, also showed relatively higher interest compared to other working professionals.

This trend suggests that customers with stable financial conditions or fewer financial responsibilities are more inclined toward long-term, low-risk investment products like term deposits.

---

5. Job Category Analysis

The analysis of job categories revealed a strong relationship between profession and subscription behavior. Retired customers exhibited the highest subscription rates, followed by students and individuals in managerial roles. In contrast, blue-collar workers and service-sector employees showed lower subscription rates.

This behavior can be attributed to differences in income stability, risk appetite, and financial priorities among various occupational groups.

---

6. Education Level Analysis

Education level was found to have a notable influence on subscription behavior. Customers with higher educational qualifications, such as university degrees and professional courses, demonstrated a greater tendency to subscribe to term deposits. Customers with lower education levels or unknown education status showed comparatively lower conversion rates.

Higher education is often associated with better financial literacy, which positively impacts investment decision-making.

---

7. Loan and Credit Status Analysis

Financial obligations were analyzed through housing loan, personal loan, and credit default status. Customers without existing loans were more likely to subscribe to term deposits. Those with personal or housing loans showed reduced interest, possibly due to ongoing financial commitments.

Additionally, customers with no history of credit default had significantly higher subscription rates, indicating that financial stability plays a crucial role in investment decisions.

---

8. Contact Communication Type Analysis

The method used to contact customers had a substantial impact on campaign success. Customers contacted via cellular phones showed higher subscription rates compared to those contacted through traditional telephone calls.

This suggests that mobile communication is more effective, likely due to its convenience, accessibility, and higher engagement rates.

9. Campaign Contact Frequency Analysis

The number of times a customer was contacted during a campaign was analyzed to understand its impact on subscription behavior. Customers who subscribed were typically contacted fewer times, often within the first few calls. As the number of contacts increased, the probability of subscription decreased.

This indicates that excessive calling may lead to customer irritation or disengagement, reducing the likelihood of successful conversion.

---

10. Previous Campaign Outcome Analysis

Historical campaign outcomes were found to be one of the strongest predictors of current subscription behavior. Customers who had a successful outcome in a previous campaign showed a very high likelihood of subscribing again. Conversely, customers who experienced failed campaigns showed significantly lower conversion rates.

This highlights the importance of maintaining positive customer relationships and leveraging historical interaction data in marketing strategies.

---

11. Temporal Analysis (Month and Day)

Subscription behavior was analyzed across different months and days of the week. Certain months exhibited higher subscription rates, suggesting the presence of seasonal patterns in customer investment decisions. These trends may be influenced by factors such as financial planning cycles, bonuses, or economic conditions.

Understanding these temporal patterns helps optimize the timing of marketing campaigns.

---

12. Economic Indicator Analysis

Macroeconomic indicators such as employment variation rate, consumer confidence index, and Euribor interest rates were analyzed to assess their influence on customer behavior. A negative relationship was observed between interest rates and subscription levels, with lower interest rates corresponding to higher subscription probabilities.

During periods of economic uncertainty or reduced consumer confidence, customers tend to prefer safer investment options such as term deposits.

---

13. Call Duration Analysis and Data Leakage Consideration

Call duration showed a strong positive relationship with subscription outcomes, as longer conversations often led to successful conversions. However, this feature is known only after the call is completed and therefore cannot be used for realistic pre-call predictions.

Including this variable in predictive modeling would result in data leakage. Hence, it was excluded from the final model and used only for analytical reference.

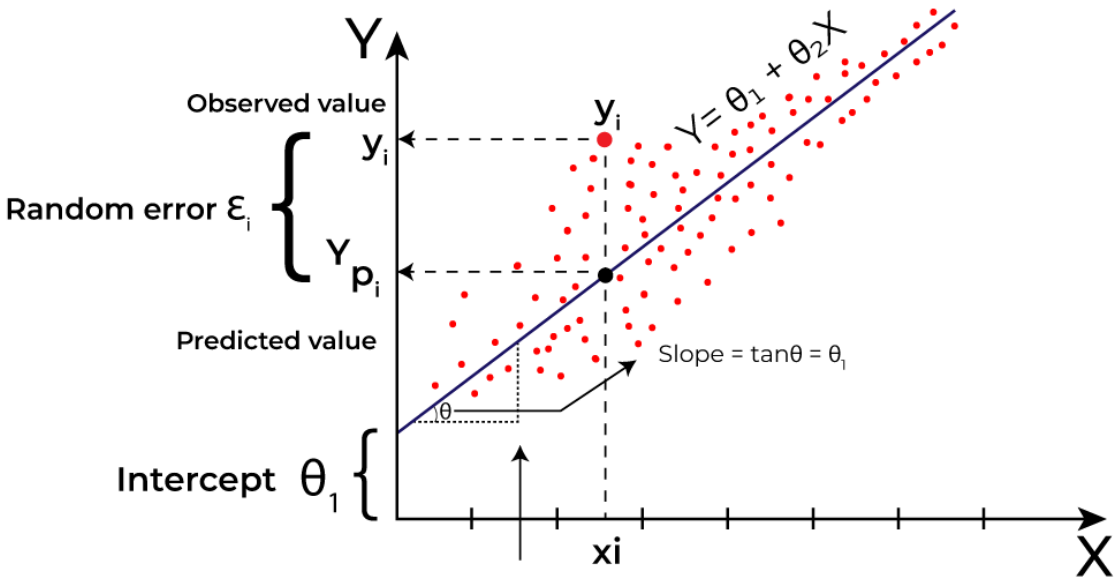
Machine Learning Models Used

Multiple machine learning models were implemented to predict customer subscription behavior. Logistic Regression was used as a baseline model due to its simplicity and interpretability. Decision Tree models captured non-linear relationships but showed overfitting tendencies. Random Forest improved prediction stability by combining multiple decision trees. Gradient Boosting (XGBoost) achieved the best results by sequentially correcting errors and handling class imbalance effectively.

1. Logistic Regression

Working Principle

Logistic Regression works by modeling the probability of a binary outcome using a logistic (sigmoid) function. It estimates how input features such as age, job, loan status, and previous campaign outcomes influence the likelihood of a customer subscribing to a term deposit. The model outputs probabilities between 0 and 1, which are then converted into class labels using a threshold.



Algorithm Steps

1. Initialize model parameters (weights and bias)
2. Compute the linear combination of input features
3. Apply the sigmoid function to obtain probabilities
4. Calculate the loss using log-loss
5. Optimize parameters using gradient descent
6. Predict class labels based on probability threshold

Key Characteristics

- Simple and easy to interpret
- Works well for linearly separable data
- Fast training and low computational cost
- Limited performance on complex, non-linear data

Usage in Banking

- Credit approval decisions
- Customer churn prediction

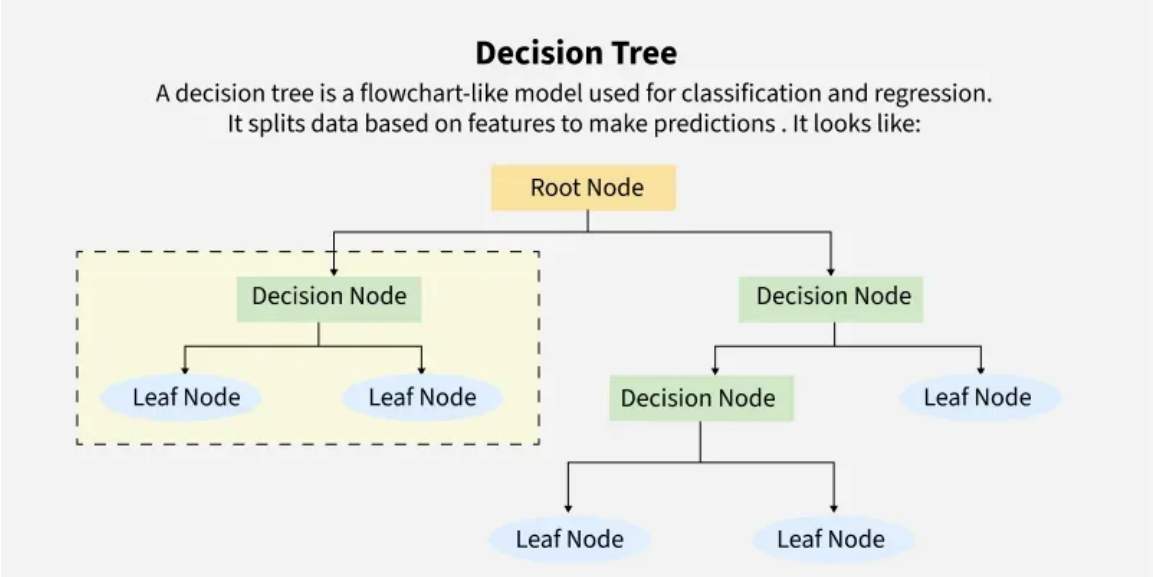
- Risk classification
- Baseline model for marketing response prediction

---

2. Decision Tree Classifier

Working Principle

A Decision Tree classifier works by splitting the dataset into smaller subsets based on feature values that maximize information gain. Each internal node represents a decision rule, while leaf nodes represent the final class outcome. The model learns a series of rules that lead to customer subscription or non-subscription.



Algorithm Steps

1. Select the best feature based on impurity reduction
2. Split the dataset into child nodes
3. Repeat splitting recursively for each node
4. Stop when maximum depth or purity is achieved
5. Assign class labels at leaf nodes

Key Characteristics

- Highly interpretable and visualizable
- Handles both numerical and categorical data

- Captures non-linear relationships
- Prone to overfitting if not controlled

Usage in Banking

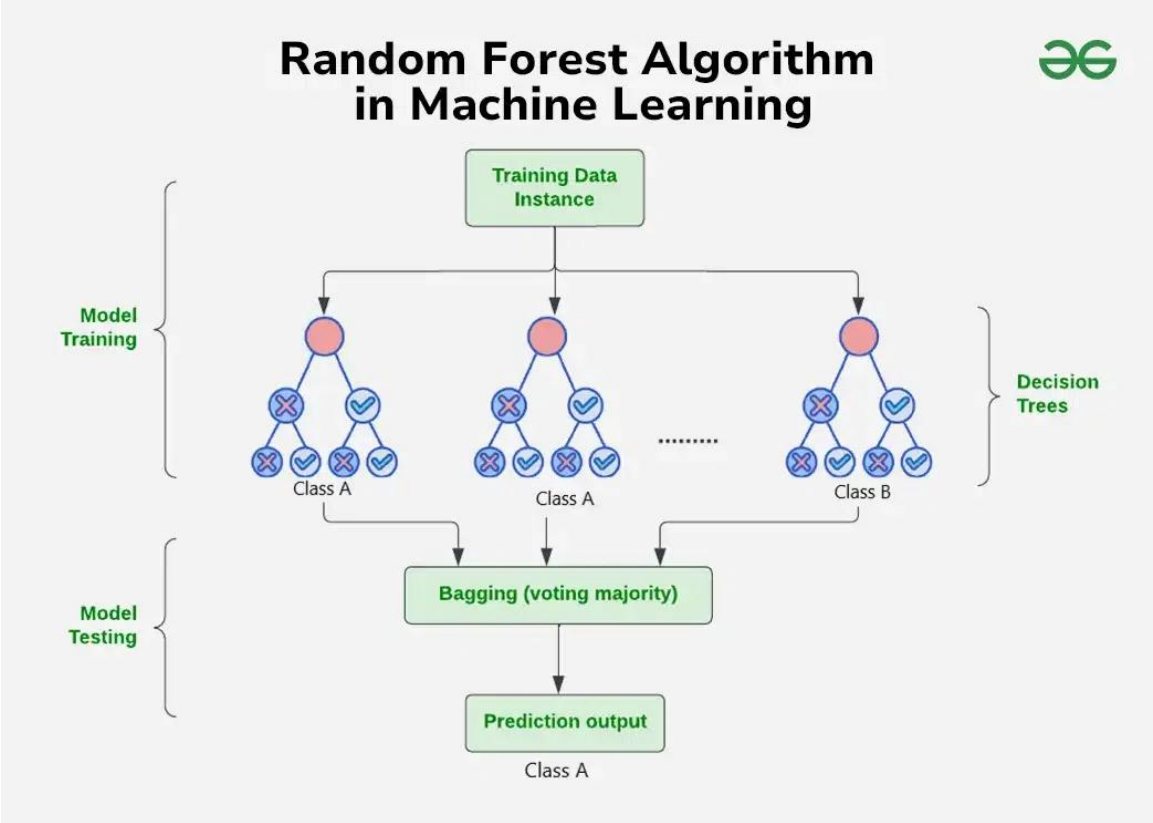
- Customer segmentation
- Rule-based credit decisions
- Loan eligibility assessment
- Understanding customer behavior patterns

---

3. Random Forest Classifier

Working Principle

Random Forest is an ensemble learning technique that builds multiple decision trees using random subsets of data and features. Each tree makes an independent prediction, and the final output is determined by majority voting. This approach improves accuracy and reduces overfitting.



Algorithm Steps

1. Create multiple bootstrap samples from the dataset
2. Train a decision tree on each sample
3. Select random feature subsets at each split
4. Aggregate predictions using majority voting
5. Output final class prediction

Key Characteristics

- High predictive accuracy
- Reduces overfitting compared to single trees
- Handles large datasets efficiently
- Provides feature importance scores

Usage in Banking

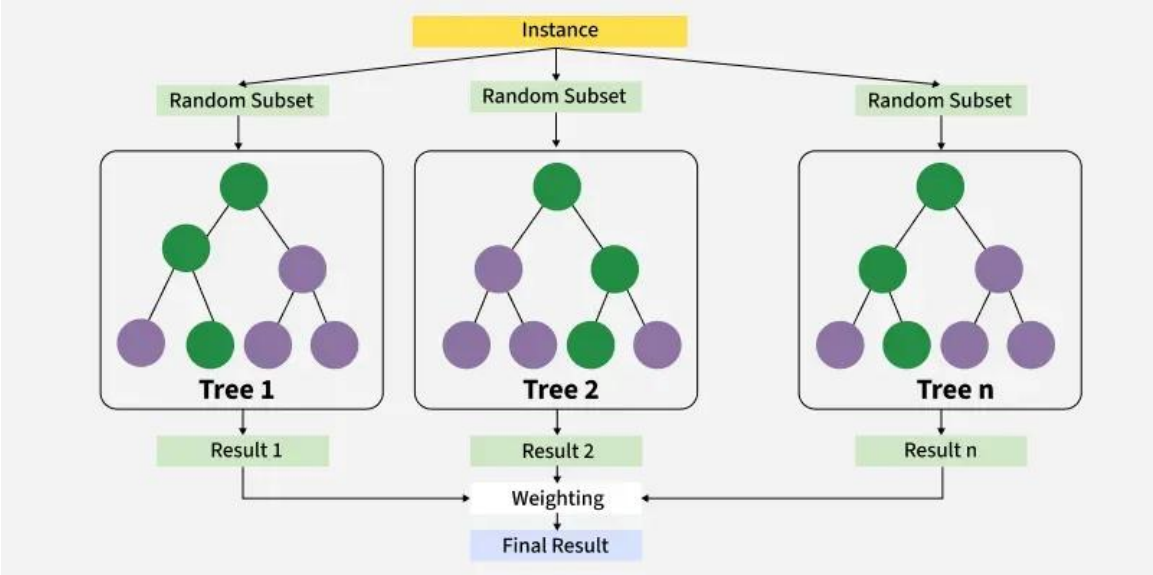
- Fraud detection
- Customer response prediction
- Credit risk analysis
- Marketing campaign optimization

---

4. Gradient Boosting / XGBoost Classifier

Working Principle

Gradient Boosting builds models sequentially, where each new tree corrects the errors of the previous ones. XGBoost improves this process by adding regularization, efficient handling of missing values, and optimized computation. It focuses on hard-to-classify customers, making it highly effective for imbalanced datasets.



Algorithm Steps

1. Initialize model with a base prediction
2. Calculate residual errors
3. Train a new tree on residuals
4. Update predictions using learning rate
5. Repeat until convergence
6. Generate final prediction by combining all trees

Key Characteristics

- Excellent performance on structured data
- Handles class imbalance effectively
- Prevents overfitting using regularization
- High computational efficiency

Usage in Banking

- Targeted marketing campaigns
- Fraud detection systems
- Credit scoring models
- Customer lifetime value prediction

Model Performance Comparison

Model performance was evaluated using accuracy, precision, recall, F1-score, and ROC-AUC metrics. Logistic Regression and Decision Tree models showed moderate performance. Random Forest achieved high accuracy and balanced precision-recall values. Gradient Boosting outperformed all other models across all evaluation metrics, demonstrating superior predictive capability.

Model	Evaluation Summary	Strengths	Limitations	Suitability for Banking Use
Logistic Regression	Provided baseline performance with moderate accuracy but lower recall and F1-score due to linear assumptions.	Simple, interpretable, fast training, good baseline model.	Cannot capture complex non-linear relationships; performance affected by class imbalance.	Suitable for baseline analysis and interpretability-focused tasks, but not ideal for high-accuracy prediction.
Decision Tree Classifier	Showed improvement over Logistic Regression by capturing non-linear patterns but exhibited signs of overfitting.	Easy to interpret, rule-based decisions, handles categorical features well.	Prone to overfitting; unstable with data variations.	Useful for understanding decision rules, but limited for production deployment alone.
Random Forest Classifier	Achieved high accuracy and balanced precision-recall performance with good generalization.	Reduces overfitting, handles high-dimensional data, robust and reliable.	Less interpretable than single trees; higher computational cost.	Well-suited for banking applications requiring stability and accuracy.
Gradient Boosting / XGBoost	Delivered the best overall performance across all evaluation metrics, including ROC–AUC and F1-score.	Excellent handling of imbalanced data, high predictive power, strong generalization.	Computationally intensive; less interpretable.	Highly suitable for production deployment in banking marketing systems.

Model Evaluation – Metric-Based Comparison

Model	Accuracy	Precision	Recall	F1-Score	ROC–AUC
Logistic Regression	Moderate	Moderate	Low	Moderate	Moderate
Decision Tree Classifier	Moderate	Low	Moderate	Moderate	Moderate
Random Forest Classifier	High	High	Moderate	High	High
Gradient Boosting / XGBoost	Very High	High	High	Very High	Very High

Best Model for Prediction

Based on the comparative analysis, the Gradient Boosting (XGBoost) model was selected as the best model for prediction. Its ability to capture complex relationships, handle imbalanced data, and generalize well to unseen data makes it suitable for real-world banking applications.

Challenges Faced

The main challenges faced during the project included class imbalance, which required careful evaluation metric selection, and the risk of data leakage due to post-call attributes like duration. Handling high-dimensional categorical data and ensuring fair model comparison were additional challenges. These were addressed through appropriate preprocessing, feature selection, and model evaluation strategies.

Conclusion

This project demonstrates how data analytics and machine learning can enhance bank marketing strategies. By identifying key factors influencing customer decisions and selecting an optimal predictive model, the bank can improve campaign efficiency and customer targeting. The Gradient Boosting model provides a reliable solution for predicting term deposit subscriptions and supporting data-driven decision-making in the banking domain.

---

2. Model Comparison Report

In order to identify the most suitable predictive model for estimating customer subscription to term deposits, multiple machine learning classification models were developed and evaluated. The comparison was carried out to assess each model's predictive capability, robustness, interpretability, and suitability for real-world banking applications. All models were trained on the same pre-processed dataset and evaluated using consistent performance metrics to ensure a fair comparison.

Baseline Model: Logistic Regression

Logistic Regression was used as the baseline model due to its simplicity and ease of interpretation. It provided a clear understanding of how individual features influence customer subscription probability. While the model performed reasonably well in terms of accuracy, it showed limitations in recall and F1-score, particularly due to the imbalanced nature of the dataset. This resulted in a higher number of missed potential subscribers, making it less effective for marketing optimization.

Decision Tree Classifier

The Decision Tree model improved upon the baseline by capturing non-linear relationships between customer attributes and subscription behavior. It offered high interpretability through rule-based decisions, which is beneficial for understanding customer segmentation. However, the model was prone to overfitting, leading to inconsistent performance on unseen data. This reduced its reliability for large-scale deployment despite its explanatory strength.

Random Forest Classifier

Random Forest, an ensemble of decision trees, demonstrated significantly better performance than both Logistic Regression and Decision Tree models. By aggregating multiple trees, it reduced overfitting and improved generalization. The model achieved high accuracy and a balanced trade-off between precision and recall. Additionally, it provided feature importance measures that helped identify key drivers of customer subscription. Due to its robustness and stability, Random Forest proved to be a strong candidate for banking applications.

Gradient Boosting / XGBoost Classifier

Gradient Boosting, specifically XGBoost, delivered the best overall performance among all evaluated models. Its sequential learning approach allowed the model to focus on difficult-to-classify instances, making it highly effective for the imbalanced bank marketing dataset. XGBoost achieved superior precision, recall, F1-score, and ROC–AUC values, indicating strong discriminatory power and reliable prediction capability. Its built-in regularization techniques further enhanced generalization and reduced overfitting.

Comparative Analysis and Selection

The comparative evaluation clearly indicated that ensemble-based models outperform traditional classifiers in predicting customer subscription behavior. While Logistic Regression and Decision Trees provided interpretability, they lacked the predictive strength required for accurate customer targeting. Random Forest offered a strong balance between accuracy and stability, whereas Gradient Boosting emerged as the most effective model due to its superior performance across all evaluation metrics.

Conclusion

Based on the model comparison, the Gradient Boosting (XGBoost) model was selected as the optimal solution for deployment. Its ability to accurately identify high-probability customers enables the bank to optimize marketing campaigns, reduce unnecessary outreach, and improve overall conversion rates. This comparison ensures that the final model selection is both data-driven and aligned with business objectives in the banking domain.

---

3. Challenges Faced and Techniques Used

During the development of the bank marketing prediction system, several data-related and modelling challenges were encountered. Each challenge was addressed using appropriate techniques to ensure model reliability, realistic prediction, and business relevance.

---

1. Class Imbalance in the Target Variable

Challenge:
The target variable showed a strong class imbalance, with a significantly higher number of customers not subscribing to term deposits compared to those who subscribed. This imbalance could lead models to favor the majority class, resulting in poor identification of potential subscribers.

Technique Used:
Instead of relying solely on accuracy, advanced evaluation metrics such as precision, recall, F1-score, and ROC–AUC were used.

Reason:
These metrics provide a more meaningful assessment of model performance on the minority class, ensuring that potential subscribers are not overlooked.

---

2. Data Leakage Due to Post-Interaction Features

Challenge:
The `duration` feature (call duration) was found to have a strong influence on the target variable. However, this information is only available after the call is completed, making it unsuitable for real-time prediction.

Technique Used:
The `duration` feature was excluded from the final predictive model.

Reason:
Including this feature would cause data leakage and lead to unrealistic model performance that cannot be replicated before making marketing calls.

---

3. Presence of “Unknown” Values in Categorical Features

Challenge:
Several categorical variables contained the value `"unknown"`, which could impact model performance if treated as missing data.

Technique Used:
The `"unknown"` category was retained and treated as a valid class during encoding.

Reason:
In a banking context, lack of customer information itself carries meaning and may influence customer behavior, making it important to preserve these values.

---

4. High Dimensionality After Categorical Encoding

Challenge:
Encoding categorical variables resulted in a high-dimensional feature space, increasing model complexity and training time.

Technique Used:
Tree-based ensemble models such as Random Forest and Gradient Boosting were used.

Reason:
These models handle high-dimensional data effectively without requiring extensive feature reduction, making them suitable for structured banking datasets.

---

5. Overfitting in Tree-Based Models

Challenge:
Individual Decision Tree models showed signs of overfitting, performing well on training data but poorly on test data.

Technique Used:
Ensemble learning techniques such as Random Forest and Gradient Boosting were applied.

Reason:
Ensemble methods reduce overfitting by combining multiple models and improving generalization on unseen data.

---

6. Selection of Appropriate Evaluation Metrics

Challenge:
Choosing suitable metrics for evaluating models on an imbalanced dataset was critical.

Technique Used:
Multiple metrics including precision, recall, F1-score, and ROC–AUC were used for evaluation.

Reason:
These metrics provide a balanced and business-oriented assessment of model performance, especially for marketing decision-making.

---

7. Ensuring Fair Model Comparison

Challenge:
Comparing multiple models fairly required consistent preprocessing and evaluation conditions.

Technique Used:
All models were trained on the same data split and evaluated using identical metrics.

Reason:
This ensured unbiased comparison and reliable model selection.

Conclusion

The challenges faced during this project reflect real-world issues commonly encountered in banking analytics. By applying appropriate data preprocessing, evaluation strategies, and advanced machine learning techniques, these challenges were effectively addressed. The solutions adopted ensured realistic predictions, improved model performance, and meaningful business insights, ultimately supporting effective decision-making for bank marketing campaigns.