Project Title:

Deposit Prediction Using Machine Learning

Submitted by:

Atieh Gholipour Darkhaneh

Course Name:

Data Exploration and System Management Using Artificial Intelligence / Machine Learning

Professor:

Professor Ireneusz Jablonski

University Name:

Brandenburg University of Technology Cottbus-Senftenberg

Date:

Winter 2025

Contents

1. Introduction

In the context of banking marketing campaigns, predicting whether a customer will subscribe to a term deposit is crucial for optimizing resource allocation and increasing campaign effectiveness. This project addresses the challenge of classifying customers into two categories: those likely to subscribe (yes) and those who will not (no). By accurately predicting subscription behavior, banks can better focus their marketing efforts and maximize return on investment.

The dataset used for this project contains 11,162 customer records and 17 features, including both demographic and financial information, as well as previous campaign interaction data. The target variable, deposit, is binary, indicating whether the customer subscribed to a term deposit. The task is framed as a binary classification problem, where we aim to predict this target based on the available features.

To achieve this, we employed a structured machine learning pipeline that involved data preprocessing, feature engineering, exploratory data analysis (EDA), model training, and evaluation. Various models—Logistic Regression, Random Forest, and XGBoost—were tested to identify the most effective one for predicting customer subscription. Insights drawn from the data will assist in optimizing future marketing campaigns.

---

2. Dataset Description

The dataset comprises 11,162 rows and 17 columns, with both categorical and numerical features. The key features include:

2.1. Features

- **Categorical Features:** job, marital, education, housing, loan, contact, month, and poutcome.
- **Numerical Features:** age, balance, duration, campaign, and previous.
- **Target Variable:** deposit (binary: yes or no).

The dataset is free of missing values, making it ready for immediate use in machine learning without requiring imputation.

2.2. Feature Insights

- The job feature has 12 unique categories, with "blue-collar" jobs exhibiting the highest rate of no responses for deposit subscriptions.

- The month feature consists of 12 categories, with May having the most marketing campaigns, although it showed lower success rates.
- Numerical outliers in features such as campaign and previous (with max values of 36 and 34, respectively) were handled during preprocessing.

3. Exploratory Data Analysis (EDA)

EDA helped uncover the structure of the data and its relationships with the target variable, leading to key insights:

3.1. Target Variable Imbalance

The target variable, deposit, shows slight imbalance:

- **46.1%** of customers subscribed (yes).
- **53.9%** did not subscribe (no).

This slight imbalance was considered during model evaluation to ensure fairness in predictions.

3.2. Feature Relationships with the Target

Several features were found to be influential in predicting whether a customer would subscribe to a deposit:

1. **Job:**
    - Customers in "management" and "retired" roles exhibited higher subscription rates, while "blue-collar" workers showed the lowest subscription rates.
2. **Contact Method:**
    - Customers contacted via cellular had a significantly higher likelihood of subscribing compared to those contacted via telephone or unknown.
3. **Call Duration:**
    - Longer calls were strongly associated with higher subscription rates, indicating that more engaged conversations tend to result in subscriptions.
4. **Poutcome (Outcome of Previous Campaigns):**
    - A successful outcome in a prior campaign significantly increased the likelihood of subscribing in the current campaign.

3.3. Key Visualizations

1. **Job vs Deposit (Bar Chart):**

    - Visualized the distribution of deposits across different job categories, highlighting that "management" and "retired" jobs were more likely to result in a subscription.

2. **Correlation Heatmap:**
    - Displayed correlations between numerical features such as balance, duration, and campaign, helping to identify which features are most relevant for predicting the target.

4. Data Preprocessing

The dataset was thoroughly preprocessed to prepare it for modeling. The following steps were taken:

4.1. Outlier Handling

Outliers were identified in the campaign and previous features and removed to prevent them from distorting the model's performance:

- The maximum value in campaign (36) and previous (34) were determined to be outliers based on their distribution and were filtered out.

4.2. Feature Engineering

- **Dropped Irrelevant Features:** The default feature was dropped due to its high number of no responses, while the pdays feature was dropped due to most values being -1 and providing no useful information.
- **Encoding Categorical Variables:** One-hot encoding was applied to nominal features like job, marital, and education. Binary features like housing, loan, and deposit were label encoded into 0 (no) and 1 (yes).
- **Feature Scaling:** Continuous variables were standardized using StandardScaler to ensure that all features contributed equally to the model without bias due to differing scales.

4.3. Train-Test Split

The dataset was split into training (80%) and testing (20%) sets using stratified sampling to preserve the class distribution in both sets.

5. Machine Learning Models and Results

Three models were evaluated based on their ability to predict whether a customer will subscribe to a term deposit. The following models were tested:

- **Logistic Regression**
- **Random Forest**
- **XGBoost**

5.1. Model Performance

Model	Accuracy	Precision (Yes)	Recall (Yes)	F1-Score (Yes)
Logistic Regression	82.79%	84%	79%	81%
Random Forest	85.30%	82%	88%	85%
XGBoost	85.52%	83%	87%	85%

Key Insights:

- **XGBoost** emerged as the best-performing model with an accuracy of 85.52% and well-balanced precision and recall.
- **Random Forest** achieved similar performance, with a higher recall for the yes class.
- **Logistic Regression** provided solid performance but was less effective at identifying all yes instances compared to the tree-based models.

5.2. Confusion Matrix (XGBoost)

Actual/Predicted	No	Yes
No	988	185
Yes	138	920

This confusion matrix indicates that the XGBoost model is better at predicting both yes and no customers compared to other models.

6. Conclusion

Key Findings

- **Best Model:** XGBoost provided the highest accuracy and the most balanced performance, making it the best choice for predicting deposit subscriptions.
- **Important Features:** duration and balance were the most influential features in predicting whether a customer would subscribe to a term deposit.
- **Customer Segments:** Customers in "management" and "retired" positions were more likely to subscribe, while those with shorter call durations or contacted via telephone had lower subscription rates.

Recommendations

1. **Targeted Marketing:** Focus campaigns on high-value customer segments such as management, retired, and those with high balances.
2. **Optimize Contact Methods:** Use cellular as the primary mode of contact for future campaigns.
3. **Call Duration:** Ensure that marketing campaigns include sufficient time for calls, as longer durations tend to result in more subscriptions.

Future Work

- Implement more advanced models, such as neural networks or support vector machines, to explore further improvements in prediction accuracy.
- Conduct a cost-benefit analysis of misclassifications to fine-tune the decision thresholds for optimal marketing strategy execution.