Final ISOM 835 Project Report

Predicting Term Deposit Subscriptions Using Bank Marketing Dataset

---

Title Page

Course: ISOM 835 – Predictive Analytics and Machine Learning

Project Title: Predicting Term Deposit Subscriptions Using Machine Learning

Dataset: Bank Marketing Dataset from UCI Machine Learning Repository

Student: Banoth Sai Jaswanthi

Institution: Suffolk University

Date: 05/05/2025

---

1. Introduction & Dataset Description

Overview

The purpose of this project is to utilize predictive analytics and machine learning techniques to identify which clients are most likely to subscribe to a term deposit based on data from a Portuguese banking institution. The dataset contains a mix of demographic information, economic indicators, and campaign-specific details.

Dataset Composition

- **Records**: 41,188 instances
- **Features**: 20 input variables + 1 target variable
- **Target Variable**: y (binary: 'yes' or 'no')

Rationale for Dataset Selection

The dataset is ideal for this project because of its real-world relevance to financial services marketing. It features both categorical and numerical data types, a common class imbalance, and complex relationships suited for supervised learning models. It simulates real campaign data, thus bridging academic and practical value.

Code to Load Dataset

```
import pandas as pd
df = pd.read_csv('bank-additional-full.csv', sep=';')
```

---

2. Exploratory Data Analysis (EDA)

EDA helps understand the structure of the data and informs preprocessing and model selection decisions.

Step 1: Data Summary

```
print(df.info())
print(df.describe())
```

Findings:

- 21 columns in total, including the target
- No missing values, but several categorical features contain 'unknown' entries

Step 2: Target Distribution

sns.countplot(x='y', data=df)

Results:

- 88.7% of responses are 'no'; only 11.3% are 'yes'
- Indicates strong class imbalance

Step 3: Feature Distributions

df[['age', 'campaign', 'pdays', 'previous']].hist(figsize=(10, 8))

These histograms reveal skewness and presence of outliers

Step 4: Correlation Heatmap

sns.heatmap(df.corr(numeric_only=True), annot=True)

Key Insight:

- Strong correlations exist among economic indicators

  (e.g., emp.var.rate, euribor3m, nr.employed)

Step 5: Categorical Variable Analysis

for col in ['job', 'education', 'contact', 'month']:

```
    sns.countplot(x=col, data=df, hue='y')
```

Trends observed:

- Clients contacted by cellphone or during spring months responded more positively
- Jobs such as 'retired' or 'student' showed higher 'yes' rates

---

3. Data Cleaning & Preprocessing

Step 1: Replacing 'unknown' and Encoding Target

```
df.replace('unknown', np.nan, inplace=True)
df['y'] = df['y'].map({'no': 0, 'yes': 1})
```

Step 2: Remove Data Leakage Feature

```
df.drop('duration', axis=1, inplace=True)
```

The 'duration' column reflects call outcome length and is unavailable before contact — making it unsuitable for modeling.

Step 3: Preprocessing Pipelines

```
from sklearn.pipeline import Pipeline
from sklearn.compose import ColumnTransformer
from sklearn.preprocessing import StandardScaler, OneHotEncoder
from sklearn.impute import SimpleImputer
```

Numerical features use median imputation + scaling; categorical features use most frequent imputation + one-hot encoding.

Step 4: Train-Test Split

```
from sklearn.model_selection import train_test_split
X = df.drop('y', axis=1)
y = df['y']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.25, stratify=y,
random_state=42)
```

---

4. Business Analytics Questions

1. **Which clients are most likely to subscribe to a term deposit?**

   This helps direct marketing resources toward high likelihood leads.
2. **What characteristics influence a client's decision to subscribe?**

   Understanding these can improve targeting, segmentation, and messaging.
3. **What is the best timing and contact method to convert clients?**

   This supports channel optimization and campaign planning.

---

5. Predictive Modeling

Models Implemented

- **Logistic Regression**: Interpretable baseline
- **Random Forest**: Handles non-linearity, works well with imbalanced data

Code Setup

```
from sklearn.linear_model import LogisticRegression
from sklearn.ensemble import RandomForestClassifier
```

Evaluation Metrics Used

```
from sklearn.metrics import classification_report, confusion_matrix, roc_auc_score,
precision_recall_curve
```

Model Performance Comparison

Model	Accuracy	ROC AUC	PR AUC
Logistic Regression	83.4%	0.8040	0.4628
Random Forest	86.2%	0.8154	0.4962

Random Forest outperformed Logistic Regression, especially in handling the minority class ('yes').

---

6. Insights & Answers: Model Implications

This project aimed to predict whether a client would subscribe to a term deposit based on the Bank Marketing dataset using Logistic Regression and Random Forest models. Through a structured predictive analytics lifecycle, the analysis revealed several actionable insights and implications:

Key Findings:

Model Performance

- The Random Forest model outperformed Logistic Regression, especially in recall for the minority class (yes) and overall AUC scores (ROC AUC: 0.8154, PR AUC: 0.4962).
- Logistic Regression offered better interpretability but was slightly less effective on imbalanced data (PR AUC: 0.4628).

Important Features Identified

- **Macroeconomic Indicators**: Features like euribor3m, emp.var.rate, and nr.employed were among the most influential predictors. These reflect how economic climate impacts client behavior.
- **Contact Strategy**: contact_cellular and specific months (March, May, December) showed higher success rates, suggesting timing and communication method are critical.
- **Previous Campaign Results**: Clients with poutcome_success were significantly more likely to subscribe again.
- **Demographics**: Age, job type (e.g., retired, student), and education level also contributed meaningfully.

Business Decision Implications

- **Targeted Marketing**: Use model scores to prioritize clients likely to subscribe, focusing resources on high-probability segments.
- **Campaign Planning**: Schedule outreach in months with higher historical success rates and favor cellular communication over telephone.
- **Resource Efficiency**: Avoid contacting clients with characteristics that historically correlate with low subscription probability.

Limitations

- **Class Imbalance**: The positive class (yes) was only ~11%, which limits precision despite using class_weight='balanced'. SMOTE or probability threshold adjustment may improve future models.
- **Data Leakage Concern**: We dropped duration as it's known only after the contact, making it unsuitable for pre-call predictions.
- **One-hot Encoding Explosion**: The model complexity increases due to many categorical levels, potentially leading to overfitting or interpretability challenges.

7. Ethics & Interpretability

Using predictive models in marketing introduces both opportunities and risks. While this project can optimize campaign performance, it also brings ethical considerations that must be addressed.

Ethical Considerations

- **Fairness & Bias**: Some features (e.g., job, education, age) may indirectly encode sensitive attributes. Without fairness checks, there's a risk of excluding disadvantaged groups or reinforcing societal inequalities.
- **Privacy**: The dataset contains personal information. Any real-world deployment must comply with data protection laws (e.g., GDPR) and ensure clients gave informed consent.
- **Over-Personalization**: Aggressively targeting high-probability clients may create pressure or fatigue, impacting client experience.

Interpretability

- **Logistic Regression** offers transparency via coefficients, making it easier to explain decisions to stakeholders.
- **Random Forest**, while more accurate, is less interpretable. Tools like SHAP or LIME are recommended for transparency when communicating model reasoning to non-technical audiences.

Overall, the project demonstrates the power of data-driven decision-making while highlighting the importance of ethical guardrails and clear communication when applying machine learning in customer-facing domains.

8. Appendix

Ethical Considerations

- **Bias Risk**: Targeting based on age, education, or job may lead to unfair discrimination
- **Privacy**: Ensure consent and GDPR compliance when using personal data

- **Hard Selling**: Over-targeting 'yes'-likely clients may create negative customer experience
- **Data Leakage**: Carefully removed 'duration' to prevent unethical performance inflation

Model Explainability

- **Logistic Regression**: Clear coefficients explain individual feature impacts
- **Random Forest**: Use SHAP or LIME to provide post-hoc interpretability

---

Appendix

- Code implementation in Google Collab format
- https://colab.research.google.com/drive/1rQEHNf5iQbipMUR4L5vQBng0lgMlRxVs
- Visuals:
  - Target distribution plot
  - Histograms
  - Correlation heatmap
  - Categorical count plots
  - ROC and PR curves
  - Confusion matrices
- Environment:
  - Python 3.10
  - Libraries: pandas, numpy, scikit-learn, matplotlib, seaborn