## 📄 Project Report: Model Evaluation for Customer Subscription Prediction

### 🔍 Objective

The objective of this project was to build a classification model to predict whether a customer would subscribe to a term deposit based on data from a direct marketing campaign of a banking institution. The dataset presents class imbalance, with relatively few customers subscribing compared to the majority who do not.

## 📊 Model Performance Overview

### 📂 Test Set 1 (20% split from bank-full.csv)

| ****Metric**** | ****Value**** |
| --- | --- |
| Accuracy | 0.942 |
| Precision | 0.9266 |
| Recall | 0.960 |
| F1 Score | 0.943 |
| ROC AUC | 0.9874 |

✅ **Interpretation:**  
This Random Forest model performs excellently on all metrics, achieving both high recall and high precision. This means the model not only identifies most of the true subscribers but also keeps false positives very low — a key factor in minimizing wasted marketing effort and cost.

### 📂 Test Set 2 (bank-additional-full.csv)

| ****Metric**** | ****Value**** |
| --- | --- |
| Accuracy | 0.6329 |
| Precision | 0.1998 |
| Recall | 0.7827 |
| F1 Score | 0.3183 |
| ROC AUC | 0.7440 |

⚠️ **Interpretation:**  
On this dataset, although recall is high (the model identifies most subscribers), precision is low, which means many non-subscribers are falsely predicted as subscribers. This could lead to inefficiencies in campaign targeting. The ROC AUC of 0.744 still suggests the model has a decent discriminative ability.

## 📌 Summary of Trade-Offs

| ****Metric**** | ****Verdict**** |
| --- | --- |
| Accuracy | Moderate to Excellent (dataset-dependent) |
| Precision | Very High (Test Set 1) / Low (Set 2) |
| Recall | Very High on both |
| F1 Score | Excellent (Set 1) / Moderate (Set 2) |
| ROC AUC | Outstanding (Set 1) / Good (Set 2) |

## 🔎 5. Findings and Insights from EDA

### 🔍 EDA Summary Highlights

1. **Target Variable Imbalance**  
   The proportion of clients subscribing (yes) is much smaller than no, confirming a class imbalance problem — around **88% non-subscribers vs. 12% subscribers**.
2. **Most Impactful Features**
   * **Duration** of last contact was the most predictive feature — longer calls are more likely to lead to a subscription.
   * **Poutcome (previous marketing outcome):** Success in a prior campaign significantly increases the likelihood of subscription.
   * **Month & Day:** Clients contacted in **May** had much lower subscription rates. Contacting in **March, December**, and **October** showed higher success.
   * **Job type, Education, and Marital status** also showed patterns — **retired and students** had higher subscription rates.
   * **Contact type:** Cellular contacts were far more successful than telephone ones.
   * **Age groups:** Middle-aged clients (30–60) were most represented, but subscription rates varied more with other variables.
3. **Correlation & Multicollinearity**
   * Most features had weak to moderate correlation with the target, which suggests combining them through models like Random Forest or Gradient Boosting was appropriate.
   * No high multicollinearity detected between major predictors.
4. **Missing/Unknown Data**
   * “unknown” values appeared in features like job and education. These were treated as a separate category during preprocessing rather than removed.

### 💡 Key Insights & Actionable Recommendations

#### 🎯 Client Characteristics Likely to Subscribe:

* Longer call durations
* Contacted via **cellular**
* Previous campaign was **successful**
* **Retired, students, or self-employed**
* Contacted in **October, December, March**
* Education level: **tertiary** or **unknown**

#### 📈 Features to Prioritize:

* Focus on **duration**, **poutcome**, **contact**, **month**, and **job** for targeting and segmentation strategies.
* Avoid contacts during **May**, which had low performance historically.

#### 📢 Marketing Team Recommendations:

* Use the model to pre-select clients for future campaigns based on profile.
* Prioritize **longer, quality calls** and follow-ups.
* Focus on clients with **successful past interactions**.
* Consider timing of campaigns — avoid months that historically perform poorly.
* Consider offering differentiated products or tailored messaging to **retired or student** segments.

## ✅ Conclusion

The Random Forest classifier trained on bank-full.csv and tested on a 20% holdout set is a robust and high-performing model, with precision and recall both exceeding 92%. The exploratory data analysis also reveals strong, actionable patterns in customer behavior that can inform future marketing strategies.

With threshold tuning, proper campaign scheduling, and client segmentation based on the highlighted features, this model can be a powerful tool to boost subscription rates and reduce costs.