Project : Bank Marketing

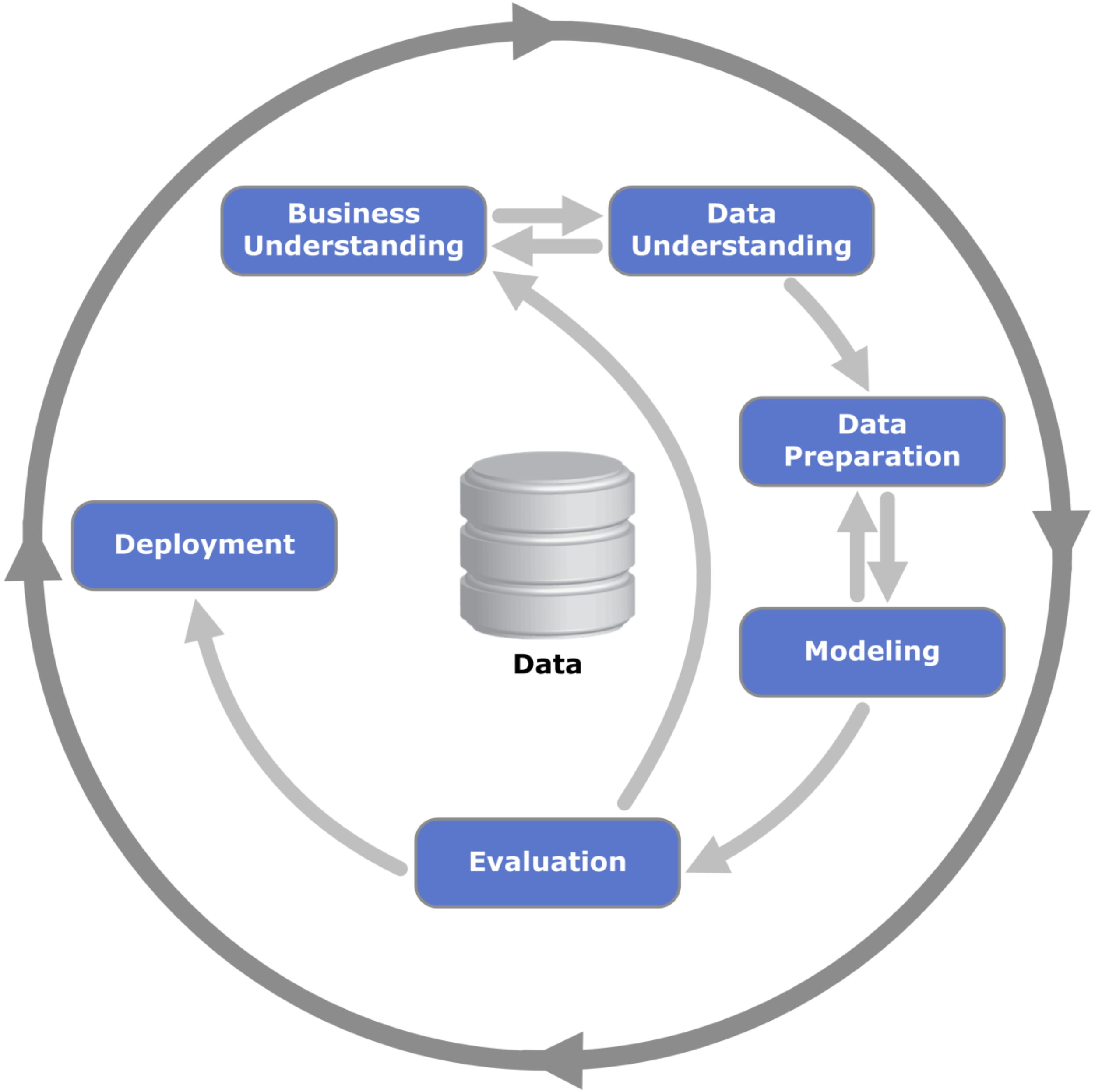
Professor : Zahra Sedighi Maman

Group Members :
Hemanth Meka
Mamatha Jala
Riya Shah
Sharanya Dulam

Project Overview: Bank Marketing Strategy with Data Analytics

- **Objective:** Used data analytics to improve Portuguese bank's phone marketing campaign
- **Goal:** Predict which customers will accept a term deposit.
- Traditional marketing is less effective; Banks need data-driven strategies for better customer targeting.
- Use the CRISP-DM framework and predictive modeling to find the best targeting strategies.

CRISP-DM Framework Process:



Key Stages of Our Project

1. Business Understanding
2. Data Understanding
3. Data Preparation
4. Model Building
5. Evaluation & Deployment

Business Understanding: Research Insights & Problem Definition

Research Insights:

- Random Forest models effectively predict term deposit subscriptions.
- Studies report up to 92% accuracy and a 20% increase in sales.
- Predictive analytics improves targeting and campaign effectiveness.

Problem Type: Binary classification — Predict if a customer subscribes (yes/no).

Technical Goal: Build different models to make accurate predictions.

Computing & Data Needs: Can run on a standard PC or laptop.

Success Measures: Accuracy, precision, sensitivity, recall, and AUC score.

Business Understanding: Research Questions

- Can we predict the response variable using data analytics?
- What are the most important features for prediction?
- How important is each feature?
- Can we predict the probability of each class (yes/no)?
- How do different sampling techniques affect prediction?
- How do feature selection methods impact results?
- How does model complexity affect performance?
- How can we apply this analysis to support better marketing strategy?

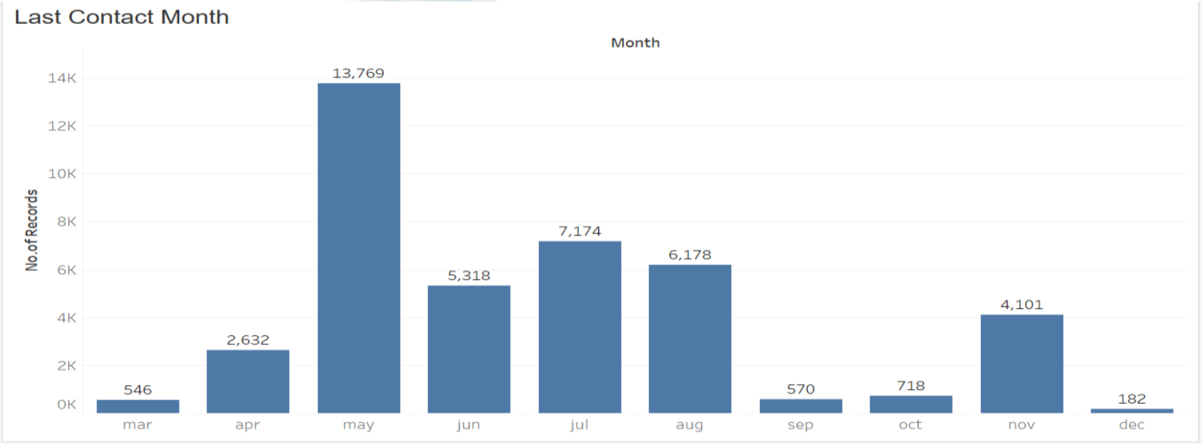
Data Understanding & Exploration

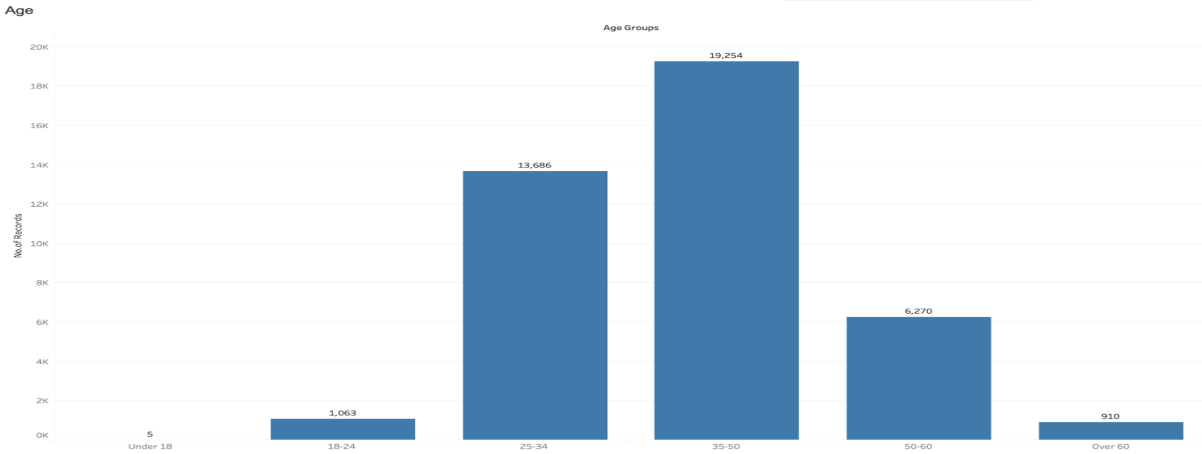
Dataset Source: Kaggle Bank Marketing Dataset.

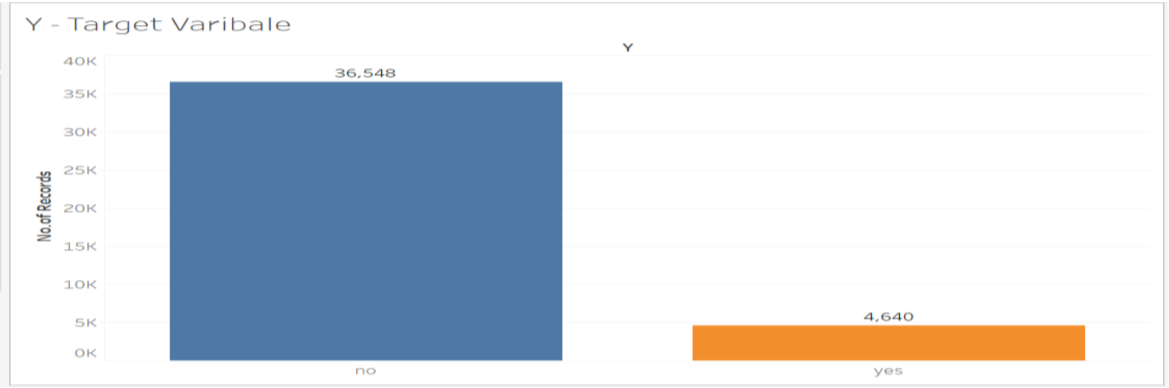
Dataset Dimensions: 41,188 observations and 21 variables.

Descriptive Statistics & Univariate Findings:

- Most clients were married, aged 25-50, with admin, blue-collar, or technical jobs.
- Contact was most frequent via **cellular phones** and peaked in **May**.
- The campaign focused on clients with little prior contact.
- The target variable (**Y**) is imbalanced, with more "no" than "yes" responses.



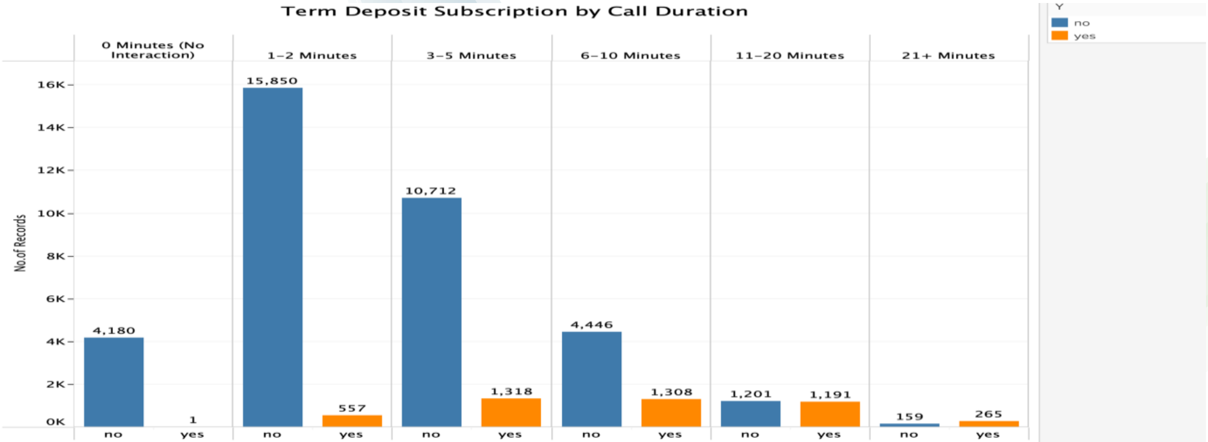


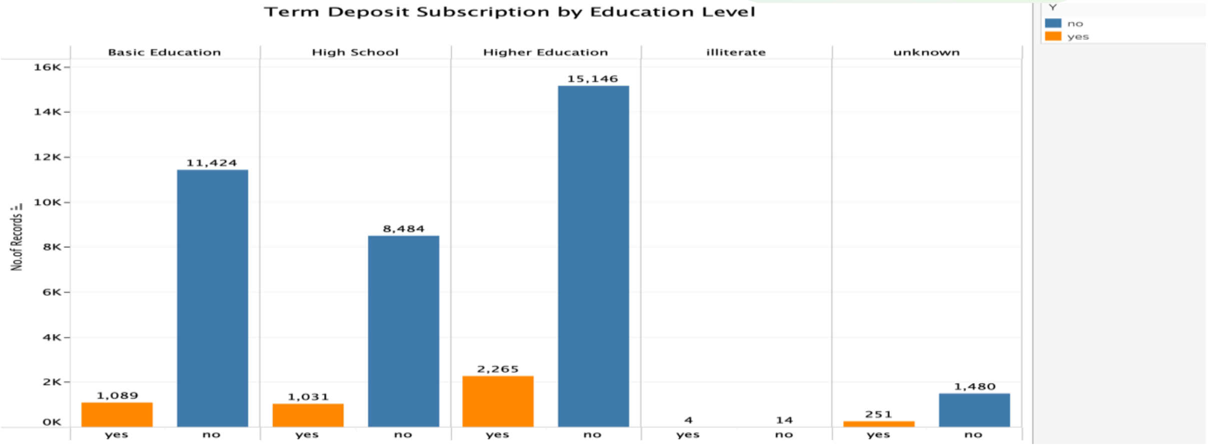


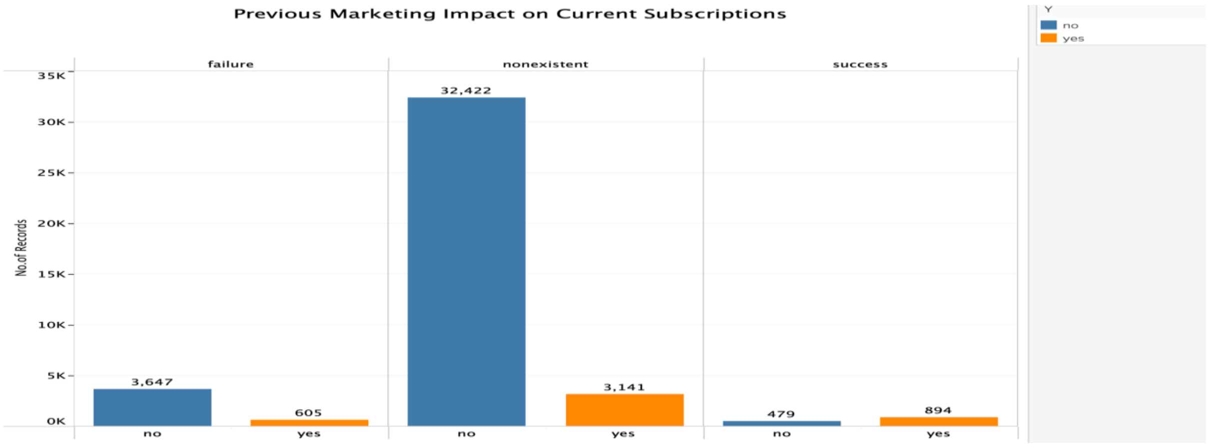
Data Understanding & Exploration

Key Insights from Bivariate Analysis:

- **Call Duration:** Longer calls (over 10 minutes) lead to higher subscription success.
- **Education Level:** Customers with higher education are more likely to subscribe.
- **Marital Status:** Married and single clients contribute most to subscriptions.
- **Past Campaign Outcome:** Previous campaign success increases likelihood of subscribing again.







Data Preparation

Purpose: Transform raw data into a clean, analysis-ready format.

Source: Kaggle, Bank Marketing dataset — 41,188 records, 21 variables.

No missing values — no imputation required.

Dropped irrelevant macroeconomic variables:

- emp.var.rate, cons.price.idx, cons.conf.idx, euribor3m, nr.employed.

Outlier handling:

- Used IQR method to detect extremes in **age** & **call duration**.
- We applied binning to keep the data reliable.

Data Preparation

Binning

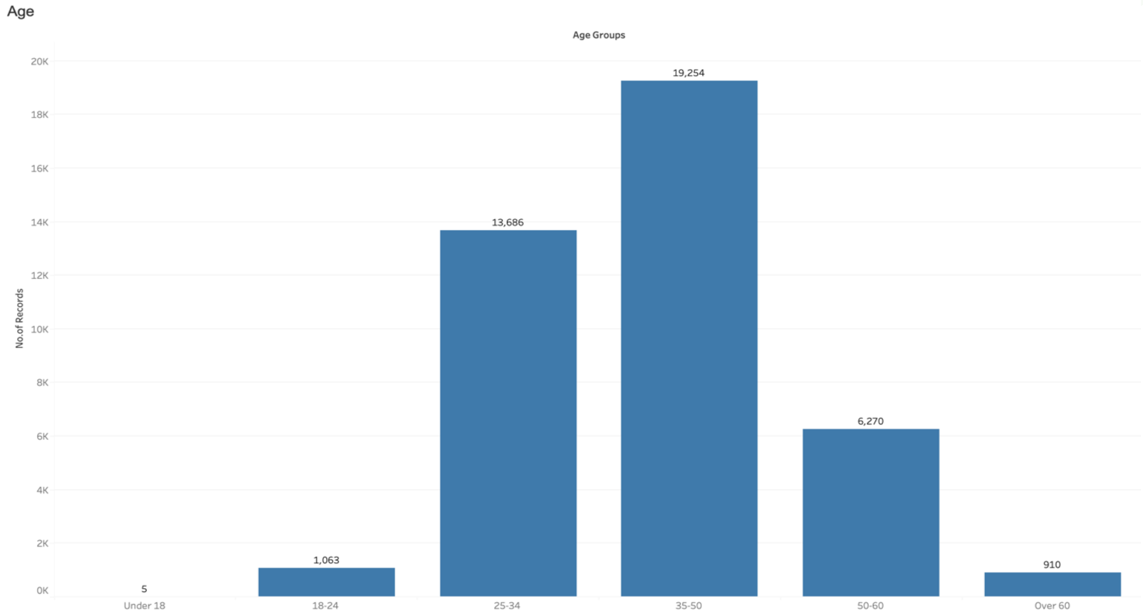
- Age groups: <18, 18–24, 25–34, 35–50, 50–60, 60+
- Call duration (mins): 0,1–2, 3–5, 6–10, 11–20, 21+

Encoding

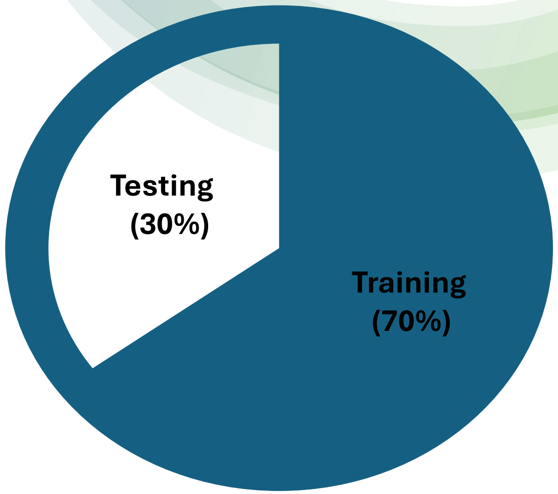
- One-hot encoding for categorical features
- Target variable: Yes → 1, No → 0

Low-variance features removed

- job_unknown
- marital_unknown
- education_illiterate
- default_yes
- month_dec



Modeling



Data Splitting & Class Balancing:

- **Split Data:** Separated data into features (X) and target (Y).
- **Split Ratio:** Used a 70/30 training/testing split to keep class ratios the same.
- **Addressed Imbalance:** Two oversampling methods
  - **SMOTE:** Increased "Yes" responses from 3,251 to 25,580.
  - **ADASYN:** Increased "Yes" responses from 3,251 to 25,447.

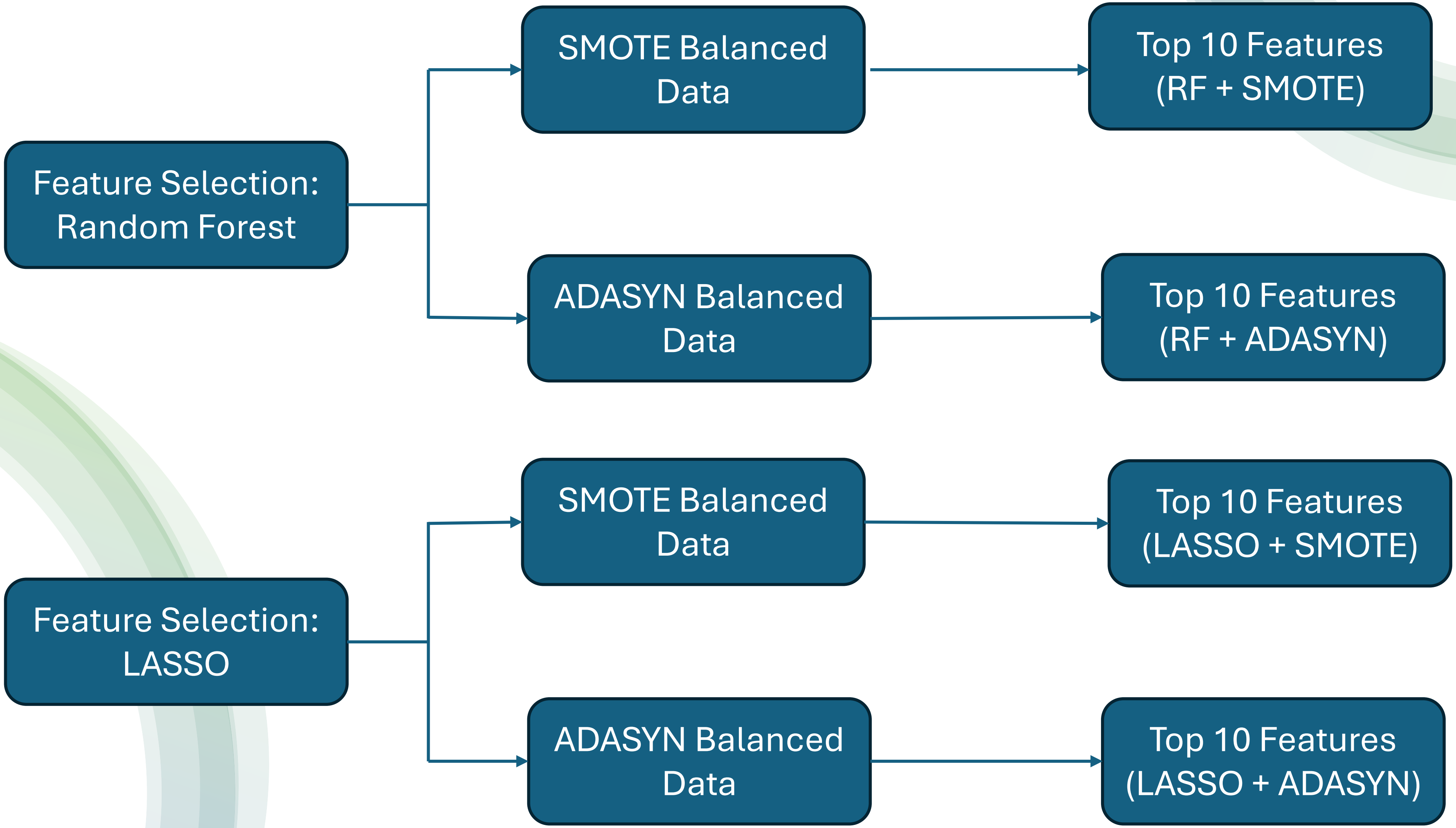
```
Before SMOTE:
y
0    25580
1     3251
Name: count, dtype: int64

After SMOTE:
y
0    25580
1    25580
Name: count, dtype: int64
```

```
Before ADASYN:
y
0    25580
1     3251
Name: count, dtype: int64

After ADASYN:
y
0    25580
1    25447
Name: count, dtype: int64
```

Feature Selection with Ensemble Models and Balanced Data



Modeling

Modeling scenarios & Metrics:

Eight scenarios– combining:

- **Models:** Random Forest, Logistic Regression.
- **Balancing methods:** SMOTE, ADASYN.
- **Feature selection:** Lasso, Random Forest importance.

Evaluation metrics:

- Sensitivity (Recall) – detect actual subscribers.
- Specificity – detect non-subscribers.
- Precision – correct positive predictions.
- G-Mean – balance between sensitivity & specificity.
- Accuracy – overall correctness.
- AUC – discrimination ability between classes.

Model Performance Comparison

Scenario	Model	Balancing	Feature Selection	Sensitivity	Specificity	Precision	G-Mean	Accuracy	AUC
1	RF	SMOTE	LASSO	0.676	0.877	0.411	0.770	0.854	0.875
2	RF	ADASYN	LASSO	0.714	0.857	0.386	0.782	0.840	0.875
3	RF	SMOTE	RF	**0.827**	0.757	0.301	**0.791**	0.764	0.870
4	RF	ADAYSN	RF	0.728	0.820	0.338	0.772	0.809	0.860

Scenario	Model	Balancing	Feature Selection	Sensitivity	Specificity	Precision	G-Mean	Accuracy	AUC
5	Logistic Regression	SMOTE	Lasso-selected	0.6782	0.8665	0.3915	0.7666	0.8454	0.8665
6	Logistic Regression	ADASYN	Lasso-selected	0.6429	0.8808	0.4059	0.7525	0.8541	0.8666
7	Logistic Regression	SMOTE	RF-selected	**0.8373**	0.7589	0.3055	**0.7972**	0.7677	**0.8744**
8	Logistic Regression	ADASYN	RF-selected	0.7055	0.8300	0.3445	0.7652	0.8160	0.8598

Evaluation

Best-performing model was (Scenario-7) **Logistic Regression** with **SMOTE balancing** and **Random Forest feature selection**.

Scenario achieved:

- **Recall:** 0.837 – capturing most subscribers
- **G-Mean:** 0.797 – balanced performance on both classes
- **AUC:** 0.874 – strong ability to distinguish subscribers from non-subscribers

Evaluation

Feature Impact on Subscriptions:

Call Duration:

- 0 min calls almost never succeed.
- Short calls (1–5 min) have low success rates.
- Long calls (11–20 min) have the highest conversion rates.

Contact Type:

- Mobile contact performs better than landline.

Timing (Months):

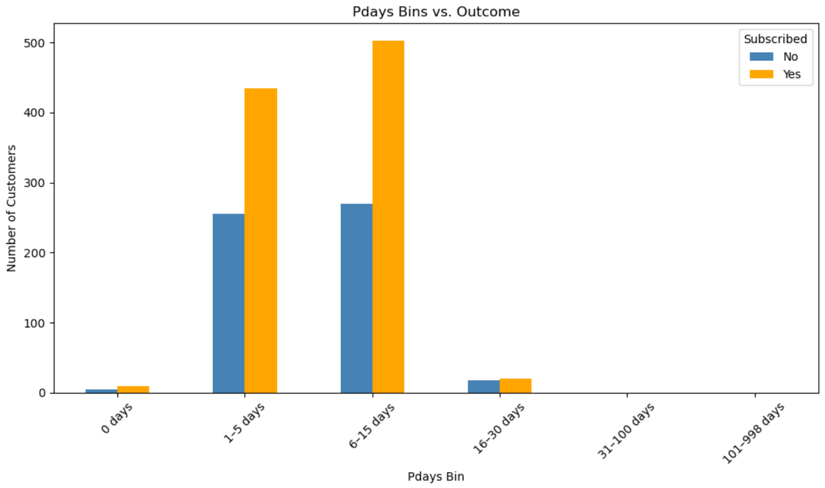
- May, July, and August show lower performance; July is the weakest due to holiday season.

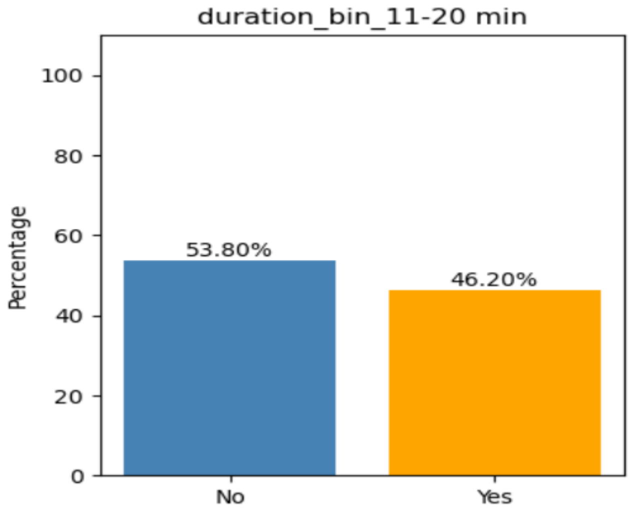
Campaign Attempts:

- Best results come within the first 4–6 calls; too many calls reduce effectiveness.

Follow-up Gap (pdays):

- Contacting within 1–15 days of last interaction significantly improves conversions.





Actionable Recommendations

- ❑ **Longer calls:** Focus on 11– 20 minutes conversations for best results.
- ❑ **Short calls:** Improve scripts for 1–5 min calls to make them more convincing
- ❑ **Zero-duration calls:** Avoid missed or failed calls by fixing contacts and calling at the right time.
- ❑ **Mobile contacts:** Prioritize mobiles over landlines for better reach.
- ❑ **Timely follow-ups:** Call back within 1–15 days for higher success.
- ❑ **Call limits:** Keep follow-ups to 4–6 calls to avoid annoying customers.
- ❑ **May outreach:** Fewer calls, more personalized messages.
- ❑ **July & August:** Reduce calls, add online or seasonal offers.
- ❑ **Lead scoring:** Use data to focus on high-potential customers first.
- ❑ **Ongoing improvement:** Track results and adjust strategies regularly.

References:

- Zaki, A. M., Khodadadi, N., Hong Lim, W., & Towfek, S. K. (2024). Predictive Analytics and Machine Learning in Direct Marketing for Anticipating Bank Term Deposit Subscriptions. *American Journal of Business & Operations Research*, *11*(1).https://www.researchgate.net/profile/Ahmed-Mohamed-Zaki/publication/376585755_Predictive_Analytics_and_Machine_Learning_in_Direct_Marketing_for_Anticipating_Bank_Term_Deposit_Subscriptions/links/658a883d0bb2c7472b105cae/Predictive-Analytics-and-Machine-Learning-in-Direct-Marketing-for-Anticipating-Bank-Term-Deposit-Subscriptions.pdf
- Ruangthong, P., & Jaiyen, S. (2015, July). Bank direct marketing analysis of asymmetric information based on machine learning. In *2015 12th International Joint Conference on Computer Science and Software Engineering (JCSSE)* (pp. 93-96). IEEE.https://ieeexplore.ieee.org/abstract/document/7219777/
- Tang, X., & Zhu, Y. (2024). Enhancing bank marketing strategies with ensemble learning: Empirical analysis. *Plos one*, *19*(1), e0294759. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0294759
- Torrens, M., & Tabakovic, A. (2022). A banking platform to leverage data driven marketing with machine learning. *Entropy*, *24*(3), 347. https://www.mdpi.com/1099-4300/24/3/347
- A data-driven approach to predict the success of bank telemarketing
  By Sérgio Moro, P. Cortez, P. Rita. 2014 Published in Decision Support Systems
- https://commons.wikimedia.org/wiki/File:CRISP-DM_Process_Diagram.png

THANK YOU

Do You Have Any Questions?