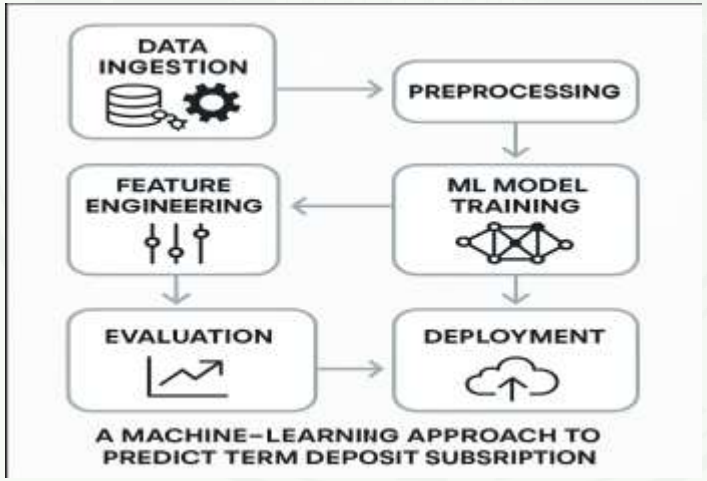
STT811 – Project Presentation

Department of Statistics and Probability

Classification Model: A Machine-Learning Approach to Predict Term Deposit Subscription

Team members: Bibin Varghese, Rohit Reddy Nallari, Nandan Keshav Hegde

✖



Date: 28th April,2025

Agenda

- Introduction and Motivation
- Data Overview
- Findings & EDA
- Baseline model
- Main model
- Results and Evaluation
- Final Remarks

Introduction & Motivation

Problem Definition:

A bank wants to improve the success of its telemarketing campaigns and wants to reach out to right customers more efficiently and figure out what has worked well thus far

Why it matters:

a. Telemarketing is costly—targeting likely customers improves ROI on marketing spends

b. Improving conversion by even a few percent boosts revenue significantly

Project goals:

a. Understand the drivers of subscription

b. Build predictive models to prioritize leads

Agenda

- Introduction and Motivation
- Data Overview
- Findings & EDA
- Baseline model
- Main model
- Results and Evaluation
- Final Remarks

Dataset Overview – (1/2)

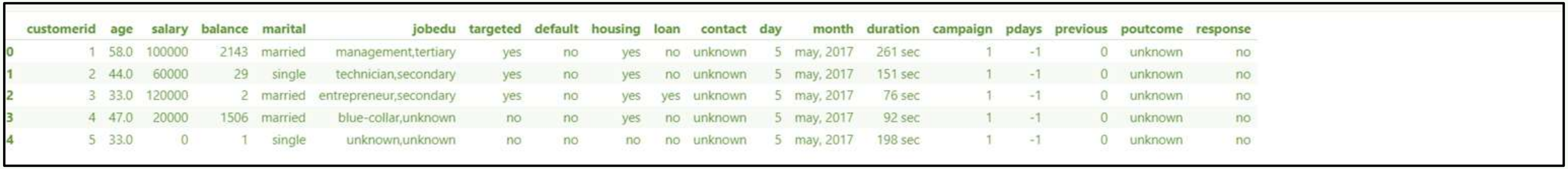
Source: UCI Bank Marketing (19 columns, ~ 45k records)

Key Features:

a. Demographics: age, job, marital status, education
b. Financial: balance, housing, loan, salary
c. Campaign: contact method, month, duration
d. Past outcomes: pdays, previous, poutcome

Target: response (yes/no)

Key Features	Data Type	Short description
age	Integer	Customer age in years
salary	Numeric	Annual salary
balance	Numeric	Account balance at time of contact
Jobedu	Categorical	Combined job and education
targeted	Binary	Whether the customer was targeted by this campaign
pdays	Integer	Days since last contact of previous campaign (–1 means none)
previous	Integer	Number of contacts before this campaign
poutcome	Categorical	Outcome of the previous marketing campaign
response	Binary	Target variable: Did the customer subscribe?



	customerid	age	salary	balance	marital	jobedu	targeted	default	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	response
0	1	58.0	100000	2143	married	management,tertiary	yes	no	yes	no	unknown	5	may, 2017	261 sec	1	-1	0	unknown	no
1	2	44.0	60000	29	single	technician,secondary	yes	no	yes	no	unknown	5	may, 2017	151 sec	1	-1	0	unknown	no
2	3	33.0	120000	2	married	entrepreneur,secondary	yes	no	yes	yes	unknown	5	may, 2017	76 sec	1	-1	0	unknown	no
3	4	47.0	20000	1506	married	blue-collar,unknown	no	no	yes	no	unknown	5	may, 2017	92 sec	1	-1	0	unknown	no
4	5	33.0	0	1	single	unknown,unknown	no	no	no	no	unknown	5	may, 2017	198 sec	1	-1	0	unknown	no

Dataset Overview – (2/2)

Data Quality Comments:

Missingness (all < 0.2%):

a. month: 50 → imputed with mode
b. age: 20 → imputed with median
c. response: 30 → dropped

Type issues to clean:

a. duration stored as text with "sec" suffix → striped unit & convert to numeric
b. Jobedu combines two concepts → split into separate job and education fields

Outlier capping:

Balance capped at the 99th percentile

Overall: High data completeness; only minor parsing & type conversions required before analysis.

New Feature	Description
duration_min	Call duration in minutes
age_band	Age bucket (e.g. 18–30, 31-45, 46-60, 61+)
response_flag	Numeric target: "yes"→1, "no"→0
default_flag	default yes/no → 1/0
housing_flag	housing yes/no → 1/0
loan_flag	loan yes/no → 1/0
job, education	Separated from jobedu
One-hot dummies: job, marital, education, contact, month, poutcome	Categorical → 0/1 indicator columns

Agenda

- ➢ Introduction and Motivation
- ➢ Data Overview
- ➢ Findings & EDA
- ➢ Baseline model
- ➢ Main model
- ➢ Results and Evaluation
- ➢ Final Remarks

Key Observations and Univariate Analysis of data – (1/2)

Feature Observations

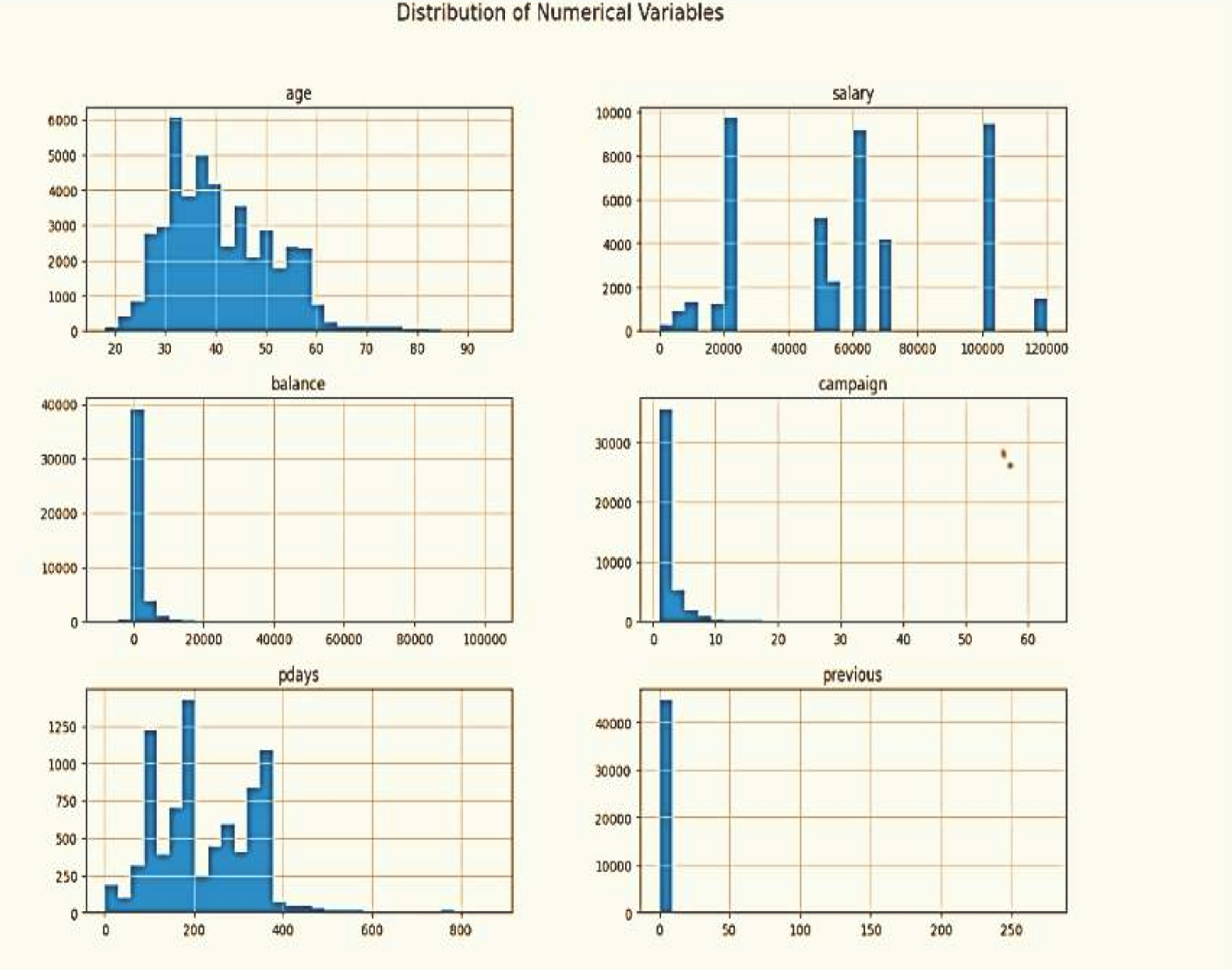
a. marital: Majority married (60%), followed by single (28%), divorced (12%).
b. job: Largest job groups: blue-collar, management, technician. Small percentage of unknown (288).
c. education: Mostly secondary education (51%), then tertiary (29%). 4% unknown.

d. targeted: 82% were targeted customers — marketing focused mostly on preselected people.
e. default: Very few people had a credit default (yes: only ~1.8%).

f. Housing: 55% have a housing loan (yes). Good to check the relation with subscription.
g. loan: 16% have a personal loan (yes). Again, interesting for modeling.
h. contact: 65% contacted via cellular, 28% unknown (meaning no reliable contact information)

i. month: Most contacts during May (30%), followed by July (15%), August (14%).
j. poutcome: 82% unknown (never participated before), 11% failure, 4% success. So, past campaign success is rare but valuable.
k. response: Imbalanced dataset — ~12% said "yes", ~88% said "no". (Important for model balancing later.)

Key Observations and Univariate Analysis of data – (2/2)



Feature Observations

l. age: Mean ~41 years. Most customers are between 30–48 years old. Some up to 95 (rare). Slight right skew, but manageable.

m. salary: Salary distribution is weird — multiple sharp peaks! Looks like salary bands (20k, 60k, 100k, etc.) — maybe system-generated salary groups, not continuous data.

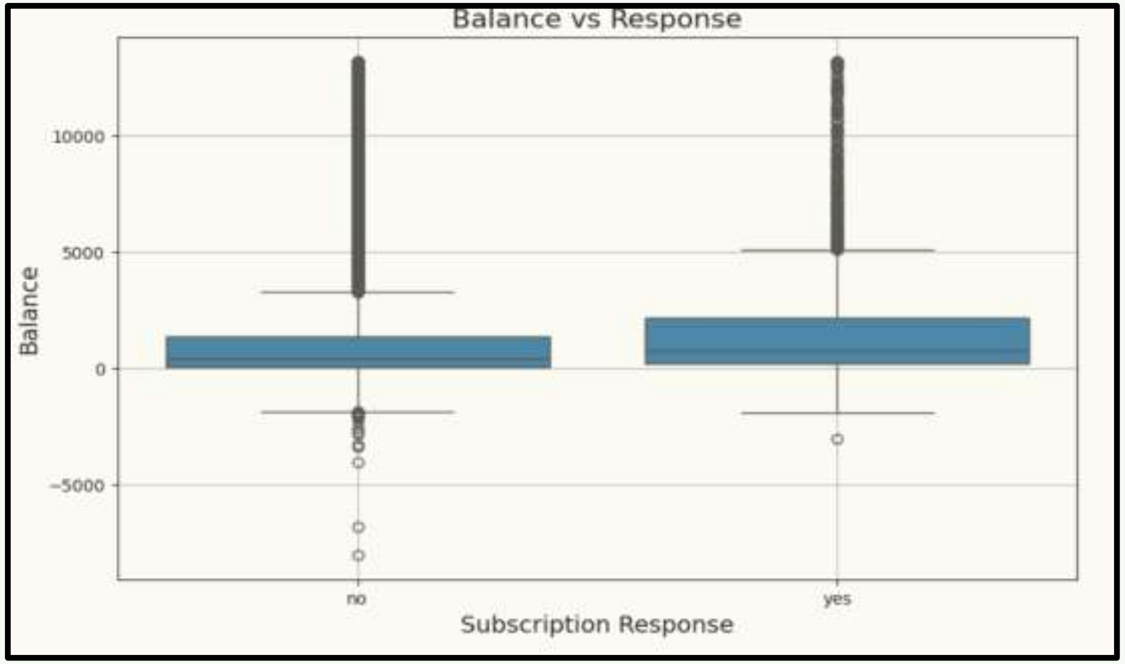
n. balance: Very right-skewed. Mean 1362, but huge max (102k). Median (448) is much lower than the mean. Some negative balances (debt?). Needs attention later.

o. campaign: Very skewed. Median is 2 contacts, but maximum is 63 — a few customers were contacted many times.

p. pdays: After cleaning, count dropped to 8251. Means around 224 days. Large skew — a few contacted after a very long time (871 days!).

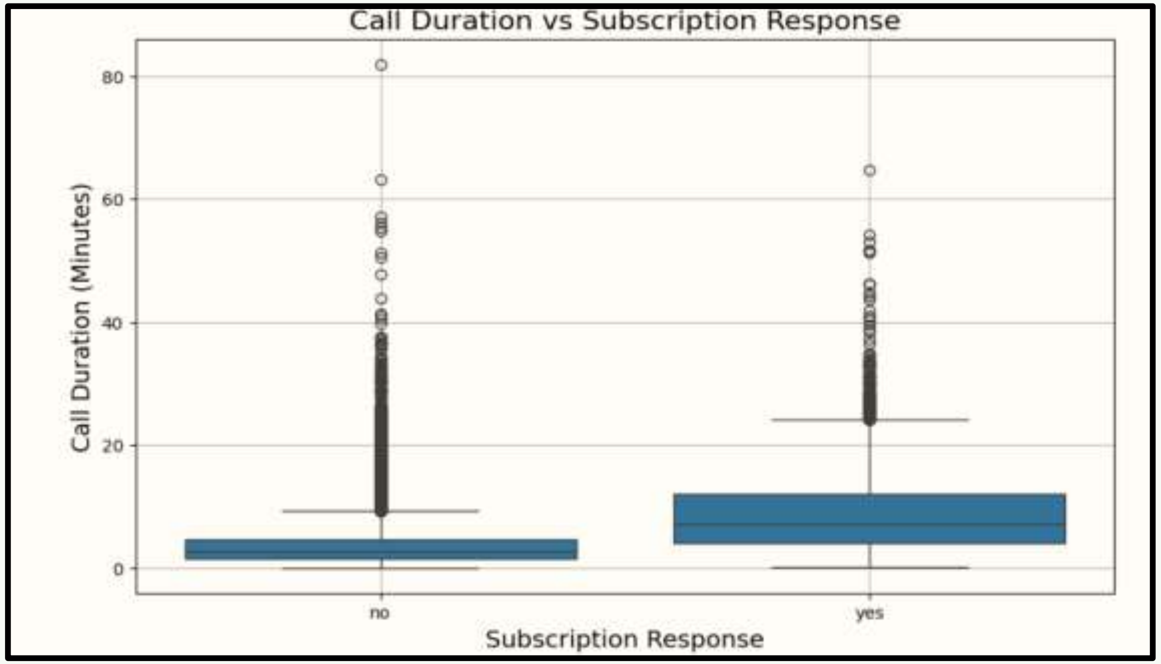
q. previous: 58% have 0 previous contacts. Max = 275 (someone was contacted 275 times earlier! Extreme, maybe an outlier.)

Bivariate Analysis– (1/3)



a. Customers with higher account balances tend to have a higher likelihood of subscribing to term deposits.

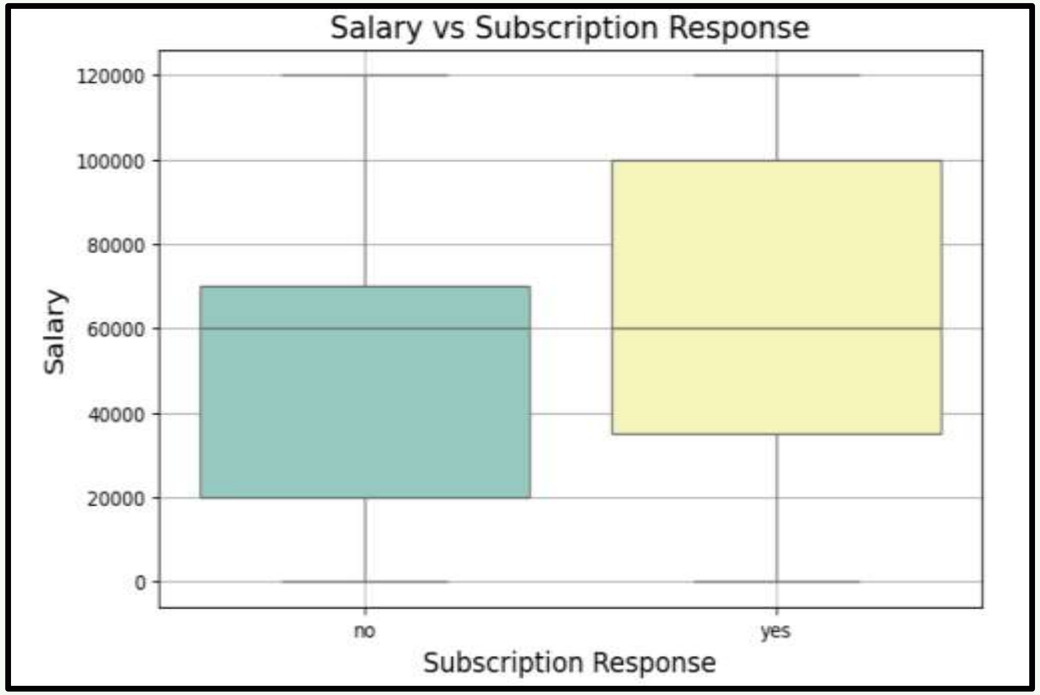
b. Median balances for subscribers are significantly higher compared to non-subscribers, suggesting that customers with greater financial capacity are more responsive to deposit offers.



a. Median call duration for positive responses is significantly higher (8–10 minutes) than for negative responses (2–3 minutes).

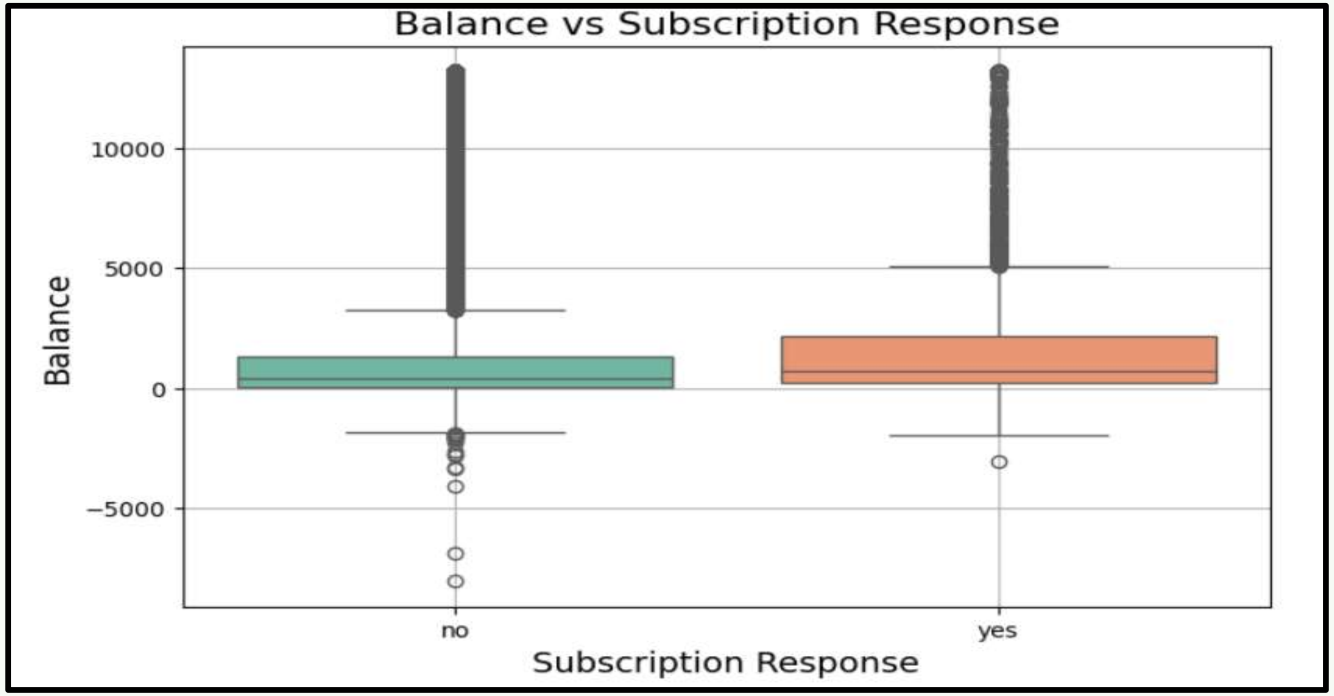
b. This suggests that longer, engaging conversations during marketing calls are positively correlated with subscription success.

Bivariate Analysis – (2/3)



Customers who subscribed (yes) seem to have a slightly higher salary range overall compared to those who didn’t (no).

However, there is a lot of overlap — meaning salary alone isn't a perfect separator.



Customers with negative or very low balances are mostly non-subscribers.

Account balance is a stronger predictor of term deposit subscription than salary. Customers with higher bank balances are more likely to subscribe to the long-term deposit product.

Bivariate Analysis – (3/3)

Other critical findings :

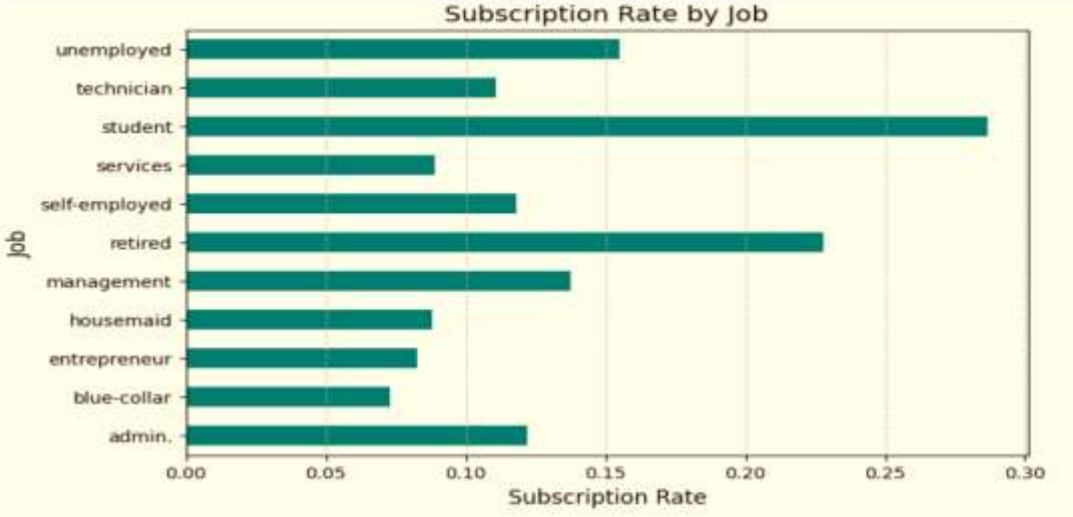
a. Higher education tends to correlate with better subscription response.

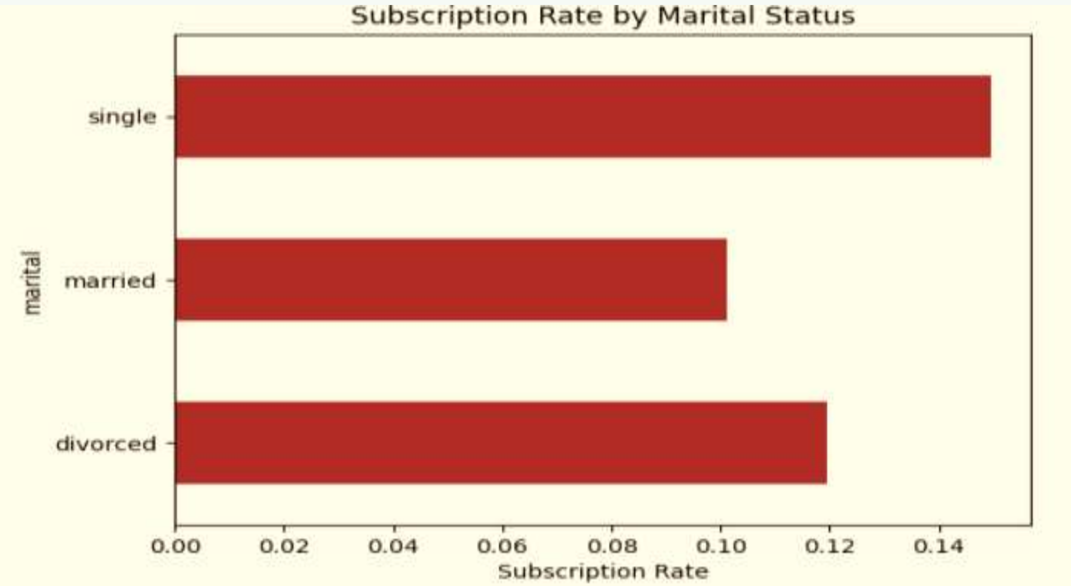
b. Singles are more responsive to marketing campaigns compared to married customers.

c. Older people may be slightly more inclined to subscribe, but the difference isn't very large.

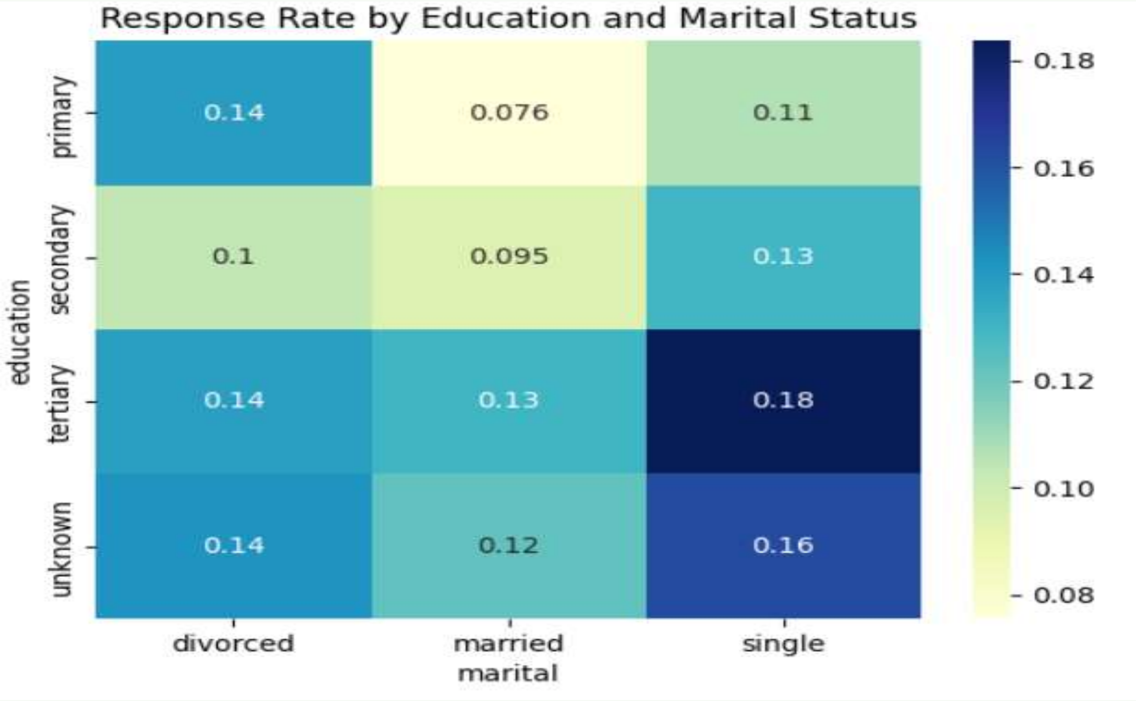
d. Campaign targeting is concentrated in working-age populations, mostly under 60 years.

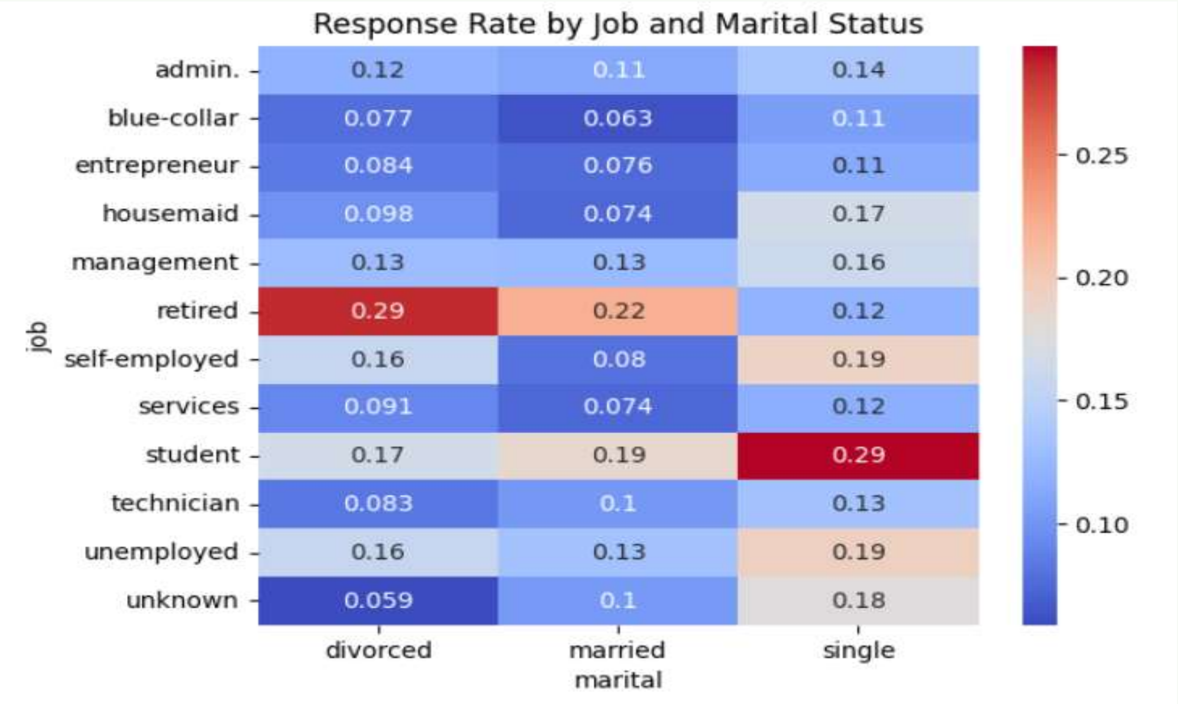
e. Students and retired individuals are more receptive to the campaign. Likely because students are future income earners and retirees may seek financial products for investment or security.





Multivariate Analysis– (1/2)





a. Married individuals with primary education have the lowest subscription rate

b. Divorced individuals have moderate subscription rates across education levels.

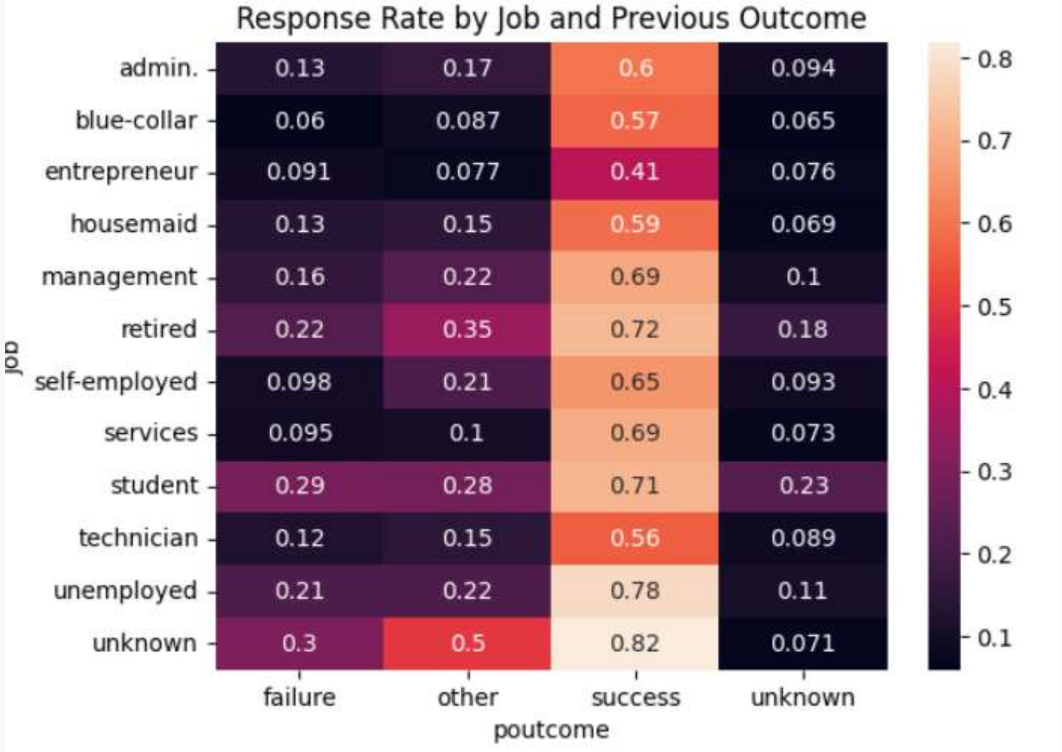
c. Higher education levels and being single correlate positively with better subscription likelihood.

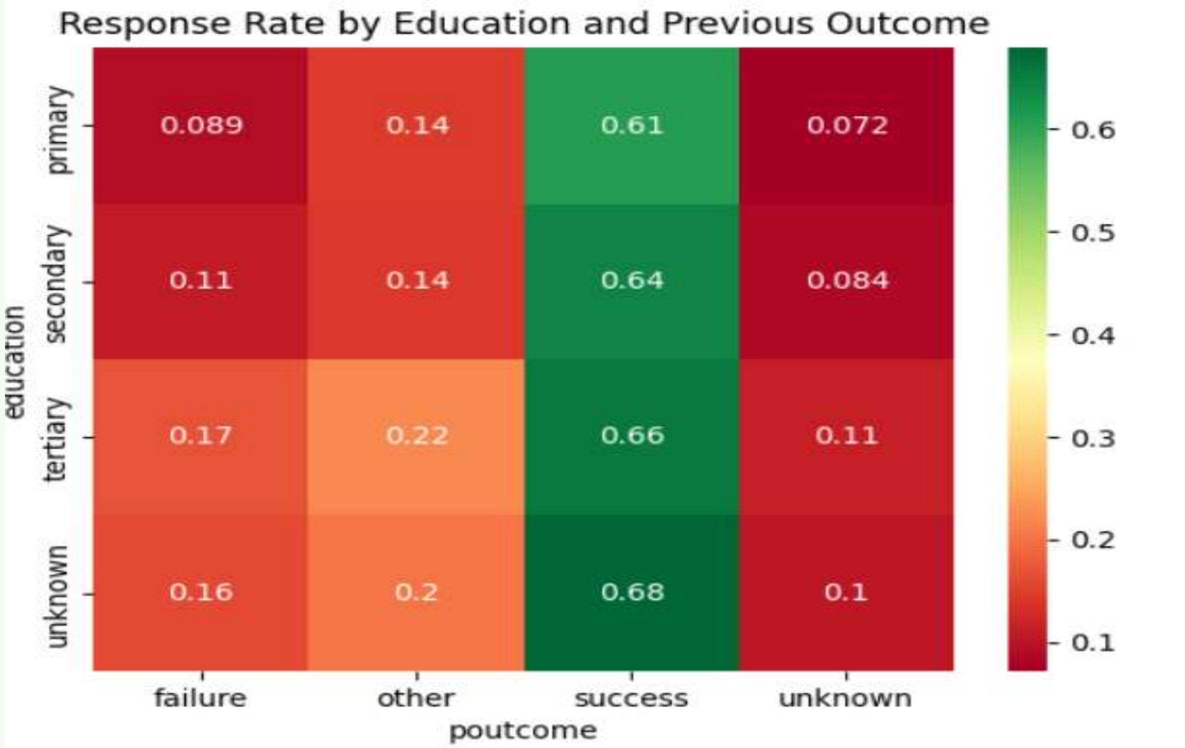
a. Retired and Student (Divorced/Single) show very high subscription rates (~29%)

b. Blue-collar and Entrepreneur (Married) show lower subscription rates

c. Students have high subscription rates irrespective of marital status.

Multivariate Analysis– (2/2)





a. People with a successful previous campaign have massively higher subscription rates (~40–82%) across all jobs.

b. Poutcome is a very strong predictor: success leads to much higher subscription likelihood.

a. Tertiary-educated customers show a high subscription (~66%) if the last campaign was successful.

b. Education + Poutcome combination plays an important role.

Agenda

- Introduction and Motivation
- Data Overview
- Findings & EDA
- Baseline model & Evaluation
- Main model & Evaluation
- Final Remarks

Baseline Model – Setup & Implementation – (1/2)

1. Train/Test Split:

a. 70% Train / 30% test stratified on response
b. Ensures class ratio (~ 88% No and 12% Yes) is preserved

2. Naive Rule-based Classifier:

a. if customer had a successful prior campaign (poutcome == "success"), predict **yes**, else **no**

Implemented by creating baseline_pred on the test set

3. Why This Baseline?

a. Leverages the single strongest historical signal

b. Serves as a floor: any model we build must beat it

The baseline model is intentionally naive and rule-based. It does not handle class imbalance, reflecting real-world raw data challenges. More sophisticated models (you will see later) apply class weighting to address the imbalance and improve performance metrics such as recall and F1-score.

Baseline Model – Setup & Implementation – (2/2)

Results on Test data

Confusion Matrix	Pred No	Pred Yes
Actual No	11.8k	179
Actual Yes	1.5k	101

Metric	Class “No”	Class “Yes”
Precision	0.89	0.36
Recall	0.99	0.06
F1 Score	0.93	0.11
Support	11.9k	1.6k

ROC-AUC Score 0.524
Accuracy:0.88

a. Accuracy: 88% Sounds high but mostly because the model predicts most cases as 0 (No).

b. Precision for Class 1 (Yes): 36%, only 36% of the times you predict 'yes', it's correct.

c. Recall for Class 1 (Yes): 6%, Very low! Only catching 6% of real 'yes' customers.

d. F1-Score for Class 1: 11%, F1-score is very low — model struggles to identify 'yes’.

ROC-AUC Score: ~0.52, random guess would be 0.5 — so baseline is just slightly better than random.

Implications:

a. Any ML model must substantially improve recall on the minority (“yes”) class while maintaining reasonable precision

b. Next: build logistic regression & random forest to leverage full feature set

Agenda

- ➢ Introduction and Motivation
- ➢ Data Overview
- ➢ Findings & EDA
- ➢ Baseline model & Evaluation
- ➢ Main model & Evaluation
- ➢ Final Remarks

Novel Model 1: Logistic Regression (Setup & Tuning) – (1/2)

1. Train/Test Split:

a. 70% Train / 30% test stratified on response Flag

2. Hyperparameter Grid (GridSearchCV):

a. C: [0.01, 0.1, 1, 10, 100]

b. penalty: [‘l1’, ‘l2’] (solver=‘lbfgs’ for l2 only)

c. Cross-Validation: 5-fold → 25 total fits

3. Best Parameters:

{"C": 0.1, "penalty": "l2", "solver": "lbfgs"}

Results on Test data

Metric	Class “No”	Class “Yes”
Precision	0.97	0.37
Recall	0.83	0.79
F1 Score	0.89	0.51
Support	11.9k	1.6k

ROC-AUC Score 0.806
Accuracy:0.82

a. Accuracy: 82% Slightly lower than baseline's 88% (expected!)

b. Precision for Class 1 (Yes): 37%, Now 37% of your "Yes" predictions are correct

c. Recall for Class 1 (Yes): 79% , Major improvement! Catching 79% of real 'Yes' customers

d. F1-Score for Class 1: 51%, Balanced precision-recall tradeoff, much better than baseline

ROC-AUC Score 0.806, Very strong model! (Baseline ROC-AUC was 0.52)

Novel Model 1: Logistic Regression (Setup & Tuning) – (2/2)

Results on Test data post 5-fold validation and Hyperparameter tuning

Metric	Class “No”	Class “Yes”
Precision	0.96	0.33
Recall	0.80	0.73
F1 Score	0.87	0.45
Support	11.9k	1.6k

ROC-AUC Score 0.765
Accuracy:0.79

Takeaway:

a. Big jump in minority-class recall (from 0.06 → 0.73)

b. Some loss in overall accuracy (0.88 → 0.79)

c. AUC 0.77 ✓—model can discriminate far better than baseline

a. Recall decreased slightly

b. ROC-AUC slightly decreased

c. Model became simpler (due to regularization)

d. Model might generalize better because C=0.1 prevents overfitting

After hyperparameter tuning, the Logistic Regression model selected a C=0.1, indicating stronger regularization.

While recall and ROC-AUC slightly decreased, the tuned model is simpler and less likely to overfit.

Compared to the naive baseline model, the tuned Logistic Regression significantly improved minority class detection and balanced model performance.

Novel Model 2: Random Forest (Setup & Tuning)– (1/2)

1. Train/Test Split:

a. 70% Train / 30% test stratified on response Flag

2. Hyperparameter Grid (GridSearchCV):

a. n_estimators: [100, 200]
b. max_depth: [10, 20, 30, None]
c. min_samples_split: [2, 5, 10]
d. max_features: [“sqrt”, “log2”]

3. Cross-Validation: 3-fold → 108 total fits

4. Best Parameters:

{ "n_estimators": 200, "max_depth": 30,"min_samples_split": 10, "max_features": "sqrt"
}

Results on Test data

Metric	Class “No”	Class “Yes”
Precision	0.91	0.68
Recall	0.98	0.31
F1 Score	0.95	0.42
Support	11.9k	1.6k

ROC-AUC Score 0.6449
Accuracy:0.90

a. Random Forest is very precise when it predicts "yes" (68%) but it misses many yes cases (only 31% recall).

b. It prefers to say "no" unless it’s very sure about "yes".

c. ROC-AUC is lower than Logistic Regression → model is worse at separating "yes" and "no" cases.

For the marketing campaign goal (maximizing responder identification), the tuned Logistic Regression model provides the best tradeoff between recall, precision, and overall discrimination ability, outperforming both the naive baseline and the Random Forest model.

Novel Model 2: Random Forest (Setup & Tuning)– (2/2)

Results on Test data post 3-fold validation and Hyperparameter tuning

Metric	Class “No”	Class “Yes”
Precision	0.96	0.54
Recall	0.92	0.69
F1 Score	0.94	0.61
Support	11.9k	1.6k

ROC-AUC Score 0.806
Accuracy:0.89

Takeaway

a. Best overall accuracy (0.89) and AUC (0.81)

b. Strong precision & recall balance on minority class

c. Ready for deployment as the lead scoring model

a. The final tuned Random Forest model is excellent!

b. It beats the naive baseline easily

c. It beats Logistic Regression on F1-Score (though recall is slightly lower than Logistic Regression).

After hyperparameter tuning, Random Forest achieved a strong balance between precision and recall with a 0.81 ROC-AUC, matching Logistic Regression's discriminative power.

The tuned Random Forest model, with an F1-score of 89% , outperforms the baseline and the untuned models, making it the recommended model for deployment in identifying potential term deposit subscribers

Agenda

- ➢ Introduction and Motivation
- ➢ Data Overview
- ➢ Findings & EDA
- ➢ Baseline model & Evaluation
- ➢ Main model & Evaluation
- ➢ Final Remarks

Final Remarks !

<u>Key Takeaways:</u>

a. Data-driven targeting can turn an 11.7 % baseline conversion into a much higher hit rate by scoring leads.
b. Random Forest delivered the best balance (Accuracy 0.89, Recall$_1$ 0.69, AUC 0.81)—a big leap over the 6 % recall of our naïve baseline.

Business Impact:

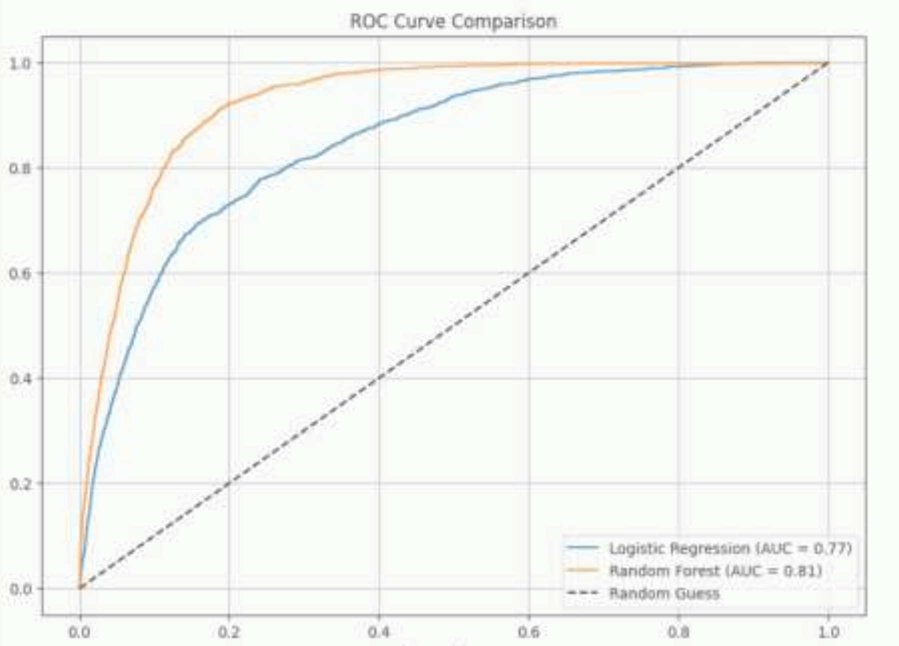
a. Better ROI: Focus on the top-scoring 20 % of prospects to capture ≈70 % of subscriptions
b. Cost savings: Fewer unproductive calls → lower call center costs per acquisition

<u>Limitations:</u>

a. Although RF improves minority recall, rare "yes" cases can still be missed under rigid score thresholds

b. Only tabular campaign/customer attributes—no call-script content or sentiment analysis Omits external factors (e.g. seasonal demand, economic shifts) that may influence subscriptions

c. Customer behaviours and marketing tactics evolve—requires regular retraining and monitoring



<u>Click here</u> to navigate to the Git Repository for code files and other metadata!