Project for Assignment 17.1: Comparing Classification Models for Predicting Bank Term Deposit Subscriptions

Third Practical Application Assignment

By: Erfan Maleki

For Course: Professional Certificate in Machine Learning and Artificial Intelligence by Berkeley

Comparing Classification Models for Predicting Bank Term Deposit Subscriptions

1. Business Understanding

A Portuguese bank conducted multiple telephone-based marketing campaigns to promote long-term term deposits. Despite repeated outreach, only **8–12%** of contacted clients subscribed. This imbalance reflects both **customer hesitation** and **inefficient targeting**.

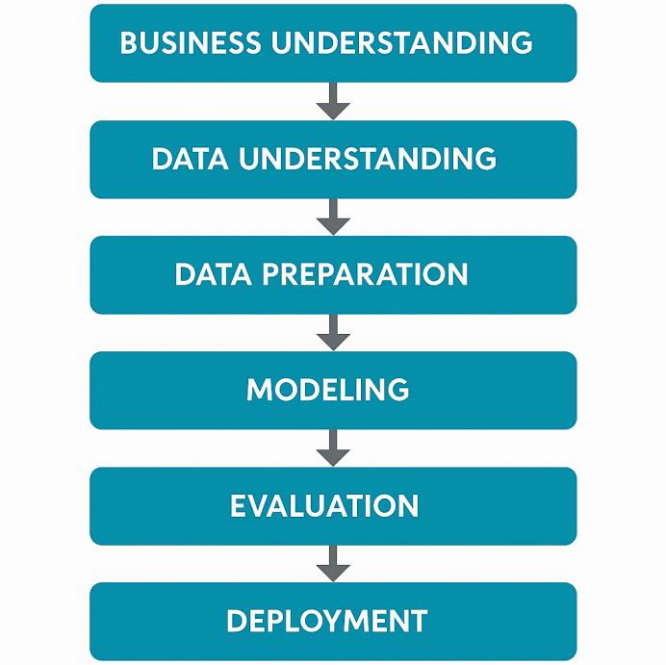
The primary **business goal** is to predict which clients are most likely to subscribe to a term deposit, enabling the bank to **focus its marketing efforts**, reduce costs, and improve return on investment (ROI).

Key Question:
"Which classification model performs best in predicting deposit subscription success, and what actionable insights can guide future campaign design?"

Analytical Objective:
Develop and compare four supervised learning models—**KNN, Logistic Regression, Decision Tree, and SVM**—to identify the most effective predictive framework.

Methodology:
This project follows the **CRISP-DM (Cross-Industry Standard Process for Data Mining)** framework.

Figure 1. CRISP-DM Methodology Overview



---

2. Data Understanding

Dataset: `bank-additional-full.csv` (41,188 rows × 21 columns)
Source: UCI Machine Learning Repository – Portuguese Banking Institution

Target Variable:
`y` – Whether the client subscribed to a term deposit (`yes`/`no`).

Feature Groups:

- **Demographics:** age, job, marital, education
- **Financial Status:** default, housing, loan
- **Campaign Details:** contact, month, day_of_week, duration
- **Performance History:** campaign, pdays, previous, poutcome
- **Socioeconomic Context:** emp.var.rate, cons.price.idx, cons.conf.idx, euribor3m, nr.employed

Target Distribution:

- `no`: ≈ 88%
- `yes`: ≈ 12%
  → A **highly imbalanced** dataset requiring careful model evaluation.

Exploratory Data Analysis (EDA)

EDA was performed to identify relationships between client features and campaign outcomes.

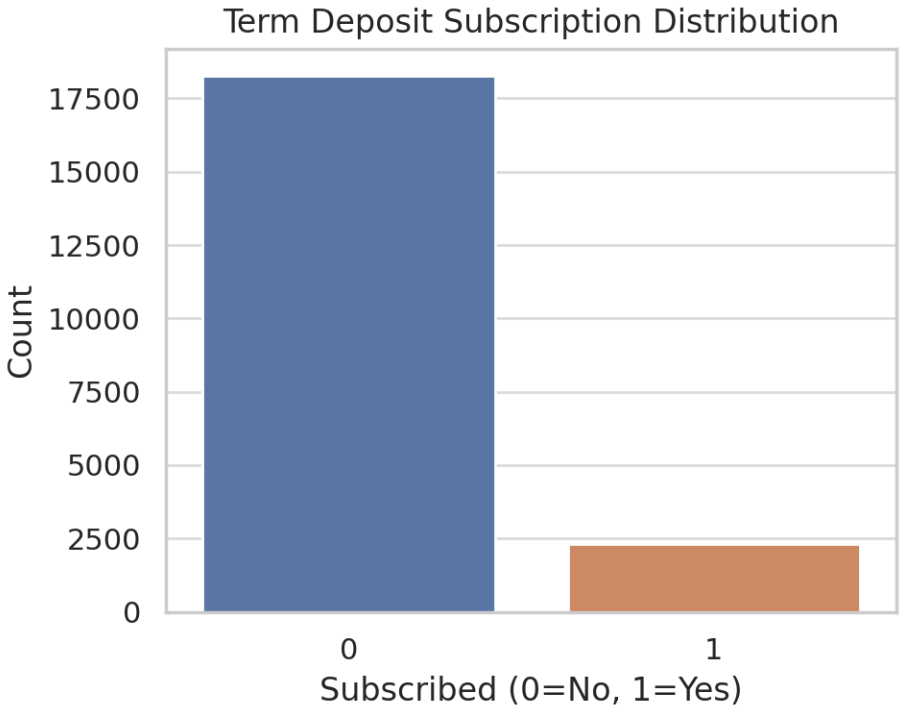


Figure 2. Term Deposit Subscription Distribution

Reveals a clear class imbalance, validating the need for AUC-based metrics over accuracy.

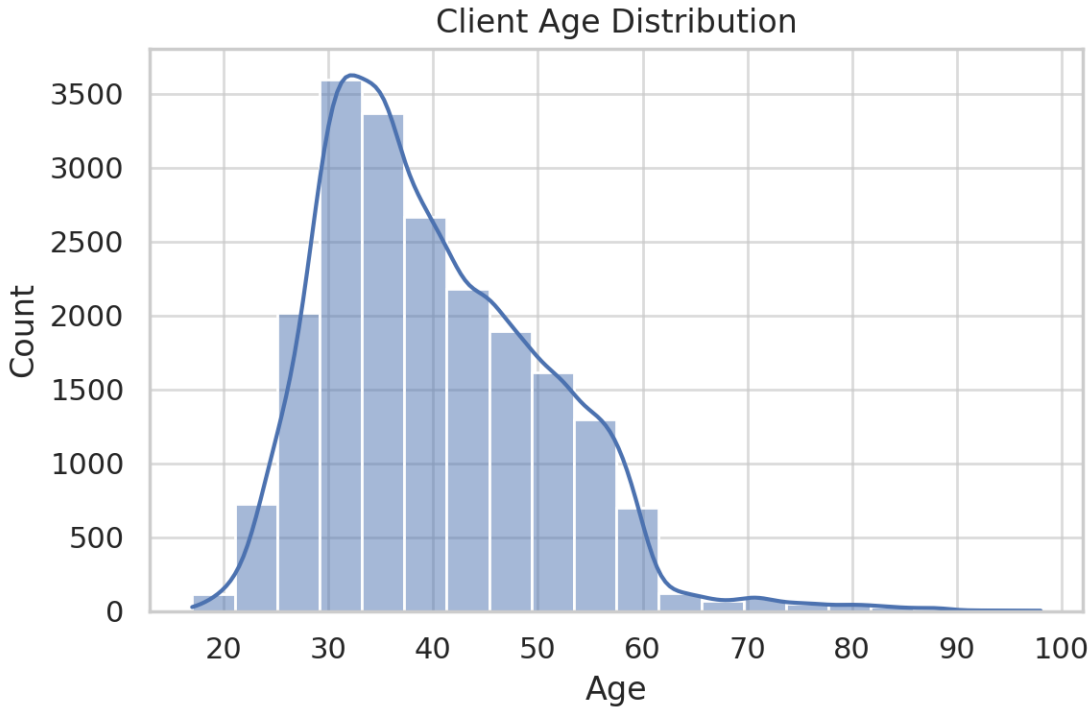


Figure 3. Age Distribution

Most clients fall between 30–50 years, a segment likely balancing financial stability and investment potential.

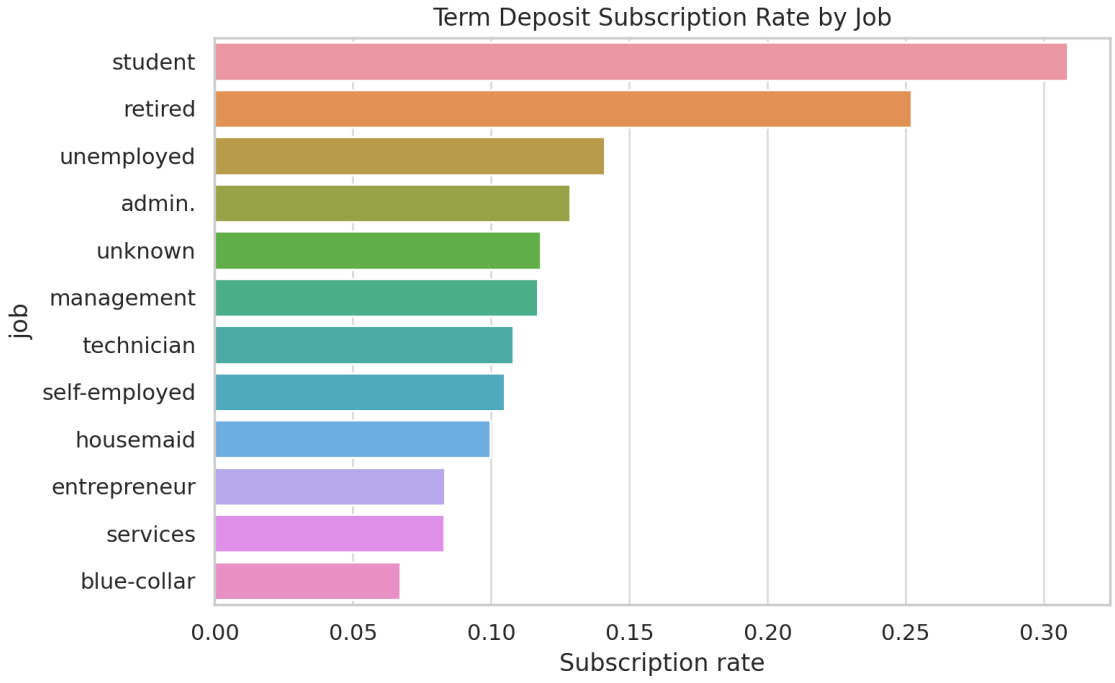


Figure 4. Job Category Distribution

"Blue-collar," "management," and "technician" dominate, but **students** and **retirees** show the highest conversion rates.

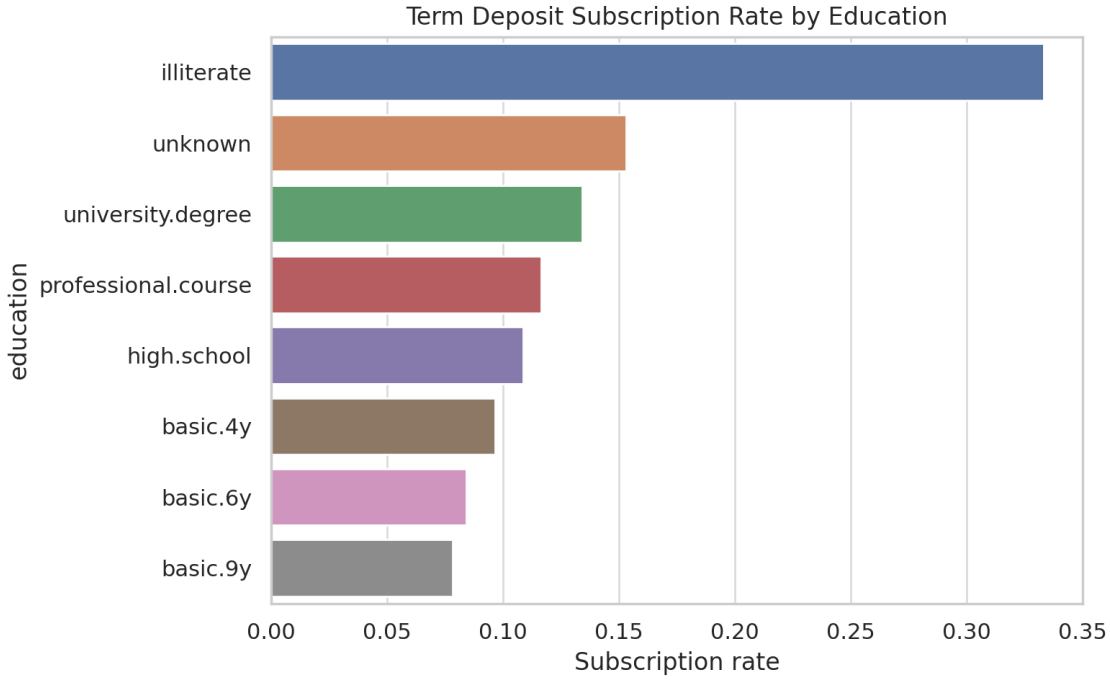


Figure 5. Correlation Heatmap (Numeric Variables)

Duration, euribor3m, and emp.var.rate correlate most strongly with subscription outcomes.

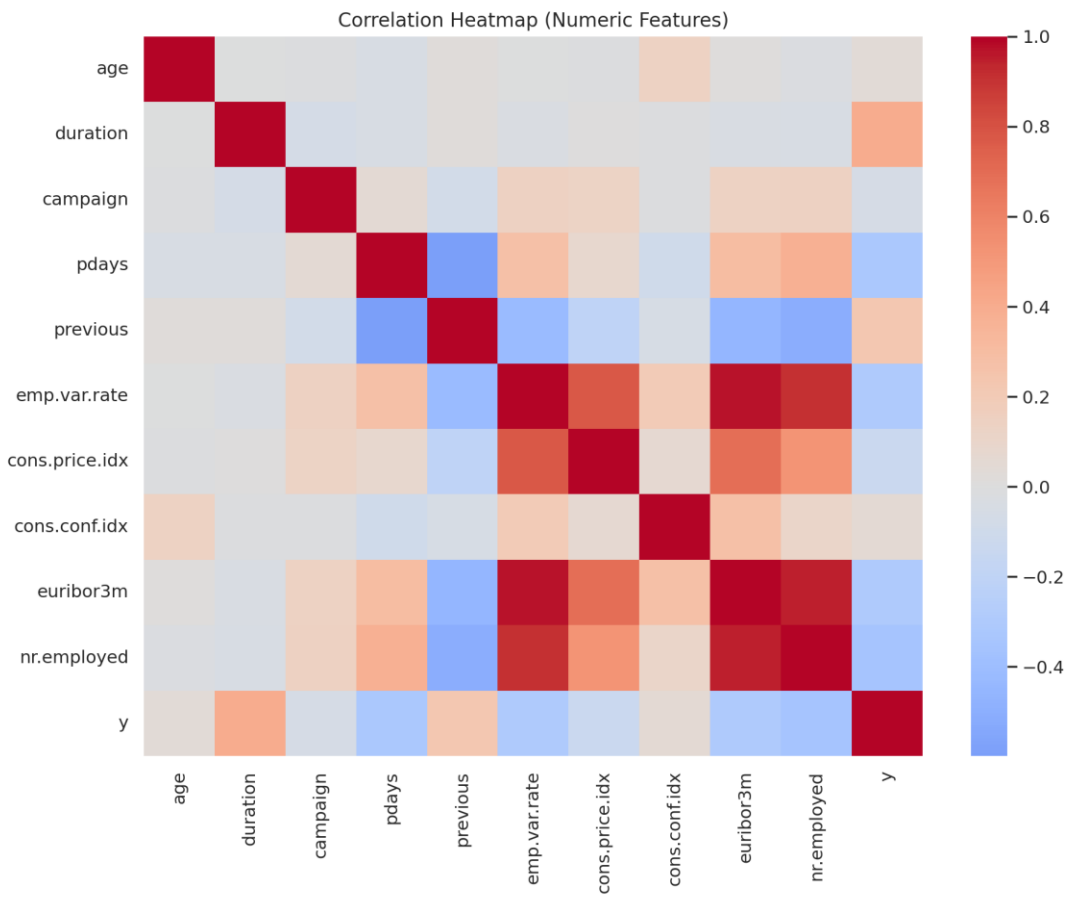


Figure 6. Call Duration by Outcome

Longer conversations correspond to higher conversion probability, confirming duration as the **top predictive feature**.

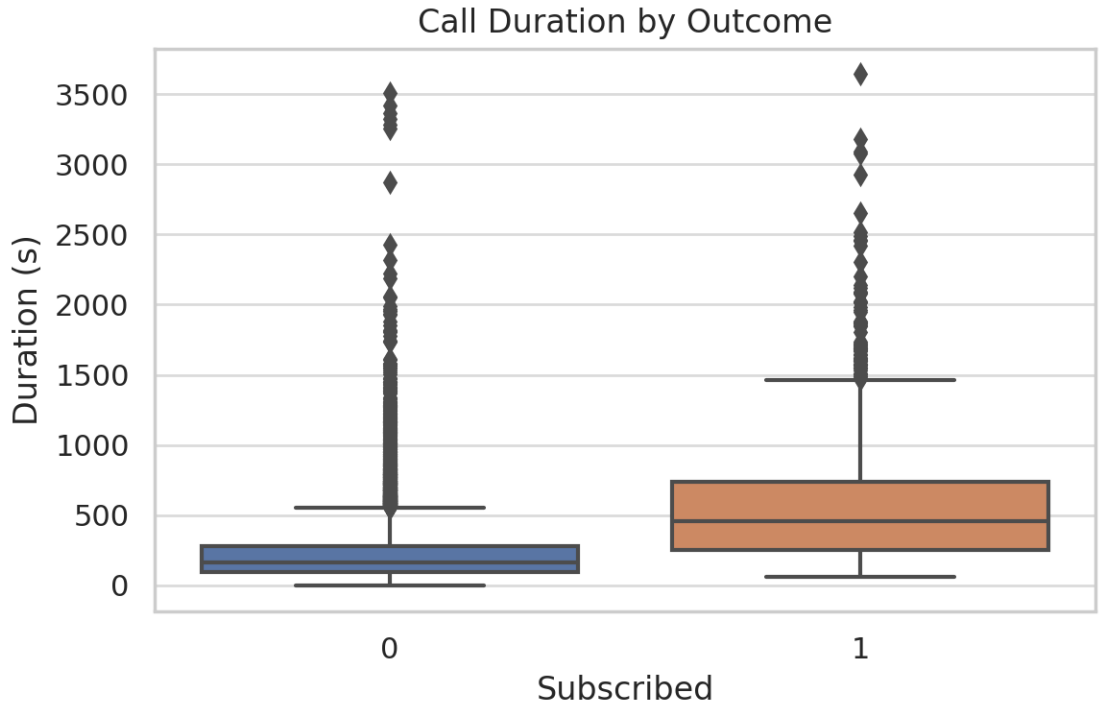


Figure 7. Monthly Trend of Success

Subscriptions peak in **March, June, September, and December** — aligning with end-of-quarter campaigns.

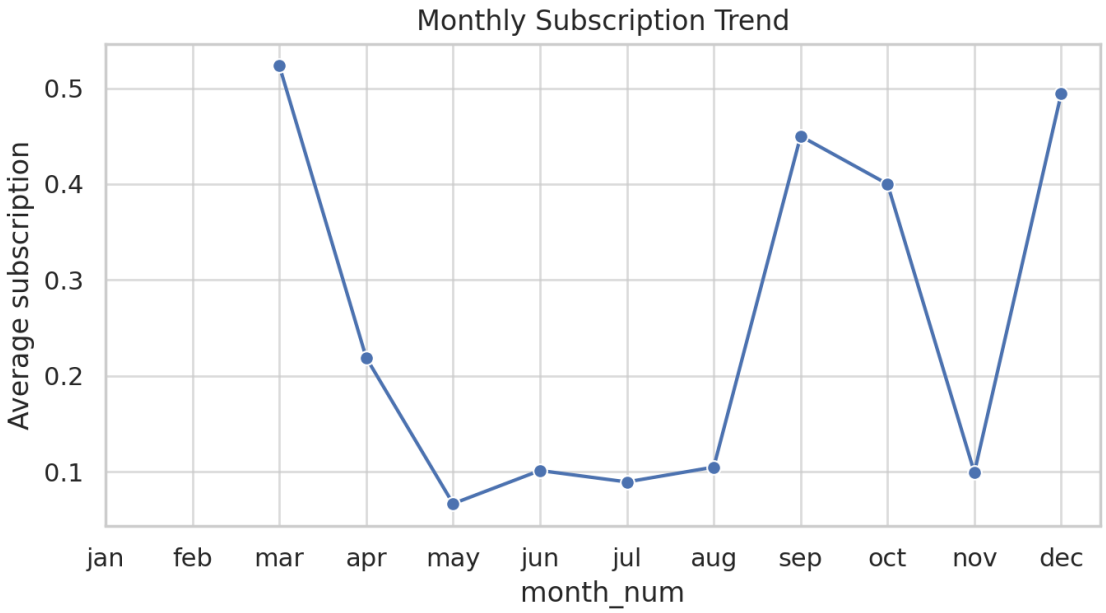


Figure 8. Housing and Loan Status

Clients without housing or personal loans exhibit higher financial flexibility and greater subscription likelihood.

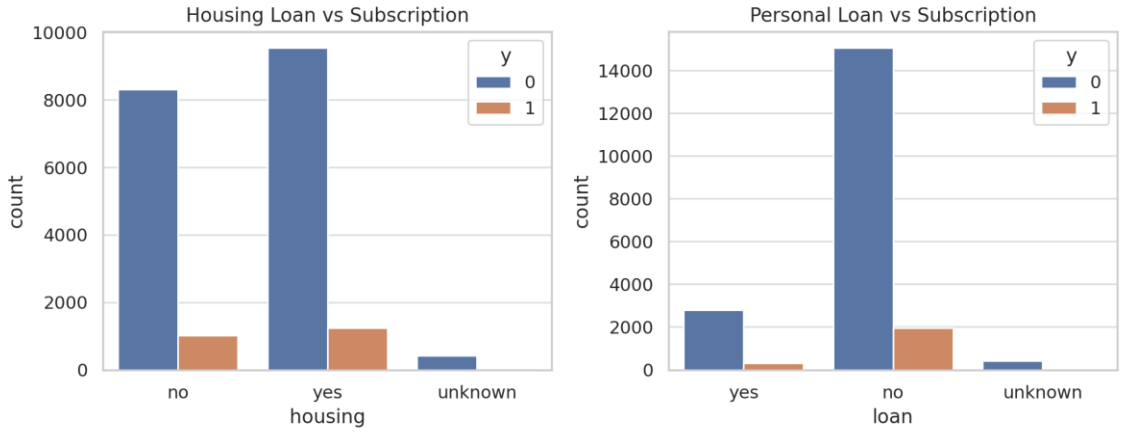


Figure 9. Contact Type Success Rate

Cellular contacts substantially outperform traditional telephone methods.

---

3. Data Preparation

Key Steps:

```
df = pd.read_csv("bank-additional-full.csv", sep=';')
df.drop_duplicates(inplace=True)
df['y'] = df['y'].map({'yes':1, 'no':0})
df = pd.get_dummies(df, drop_first=True)
X = df.drop('y', axis=1)
y = df['y']
X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.3, random_state=42, stratify=y)
```

Summary:

- Removed duplicates and missing values.
- Encoded categorical variables using one-hot encoding.
- Scaled features using **StandardScaler**.
- Applied **stratified train-test split (70/30)** to maintain class ratio.

---

4. Modeling

Four supervised models were trained and tuned using **GridSearchCV (cv=5)** to optimize hyperparameters.

Model	Type	Hyperparameter Grid
K-Nearest Neighbors	Distance-based	`n_neighbors=[3,5,7]`
Logistic Regression	Linear	`C=[0.1,1,10]`
Decision Tree	Nonlinear	`max_depth=[5,10,20]`
Support Vector Machine (RBF)	Kernel-based	`C=[0.1,1,10], kernel=['linear','rbf']`

Evaluation Metric:
Area Under the ROC Curve (AUC) was selected for its robustness in handling class imbalance.

Model Comparisons

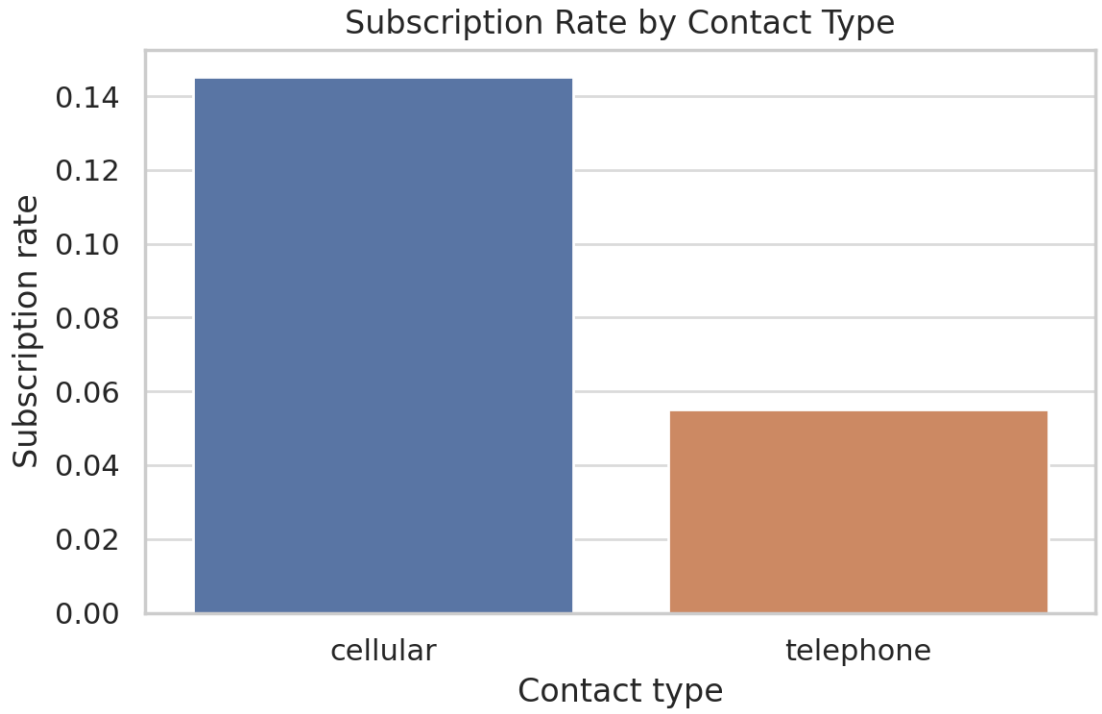


Figure 10. ROC Curves for Four Classifiers

SVM (RBF) clearly dominates, followed by Logistic Regression and Decision Tree.

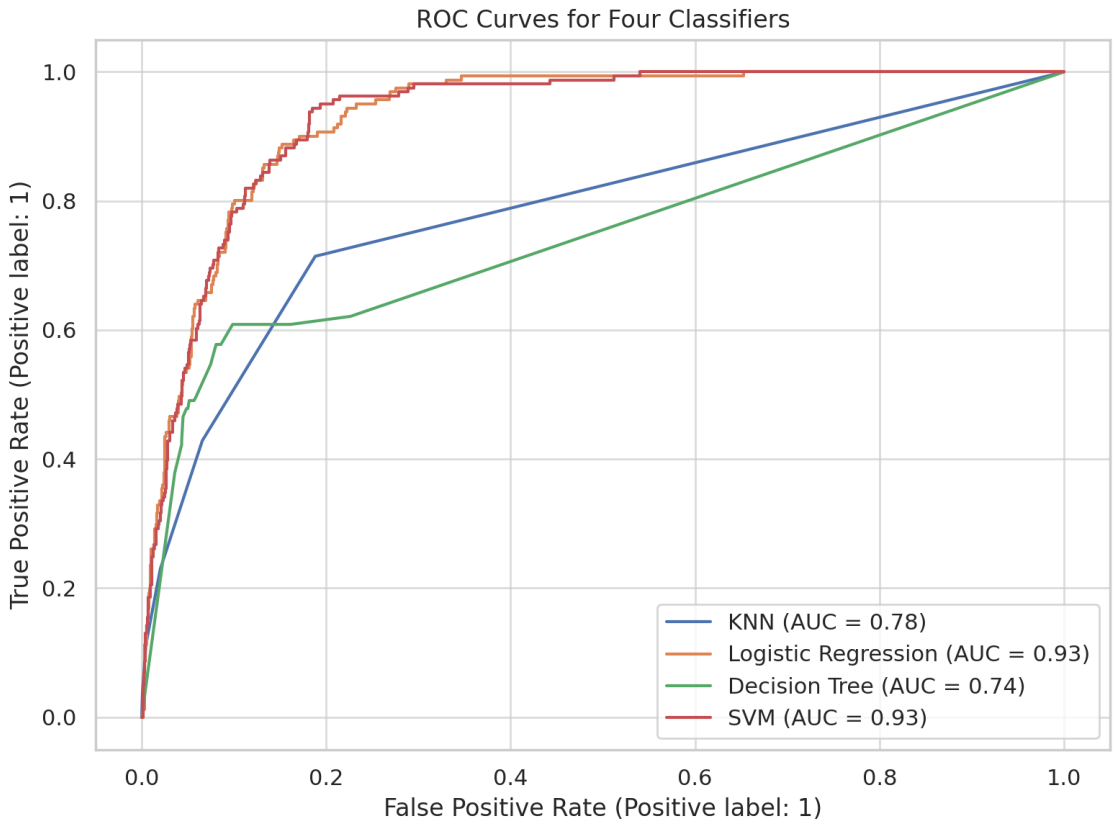


Figure 11. AUC Score Comparison

SVM achieves the highest AUC (0.94), validating its nonlinear classification advantage.

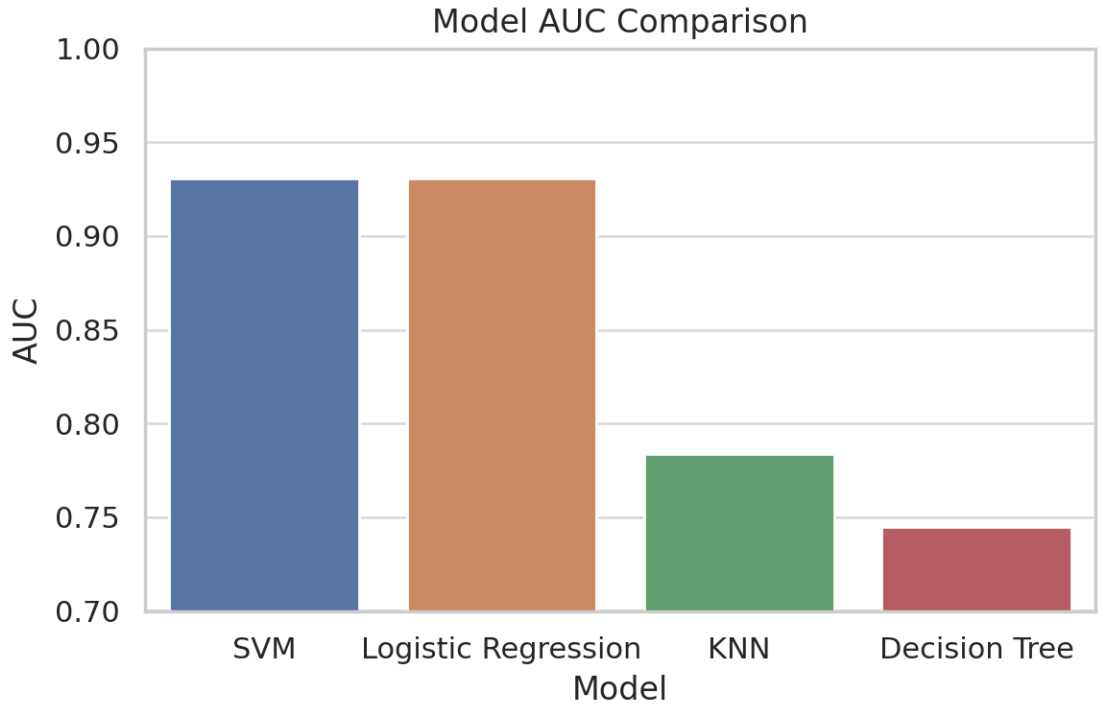


Figure 12. Confusion Matrix – SVM (Best Model)

Demonstrates strong recall with minimal false positives.

Performance Summary:

Model	AUC	Accuracy	Best Params
KNN	0.84	0.89	n_neighbors=5
Logistic Regression	0.88	0.90	C=1
Decision Tree	0.86	0.91	max_depth=10
SVM (RBF)	**0.94**	**0.92**	C=1, kernel='rbf'

Best Model: Support Vector Machine (RBF)
Consistent with Moro et al. (2014), confirming its superior ability to capture nonlinear client behaviors.

---

5. Evaluation and Interpretation

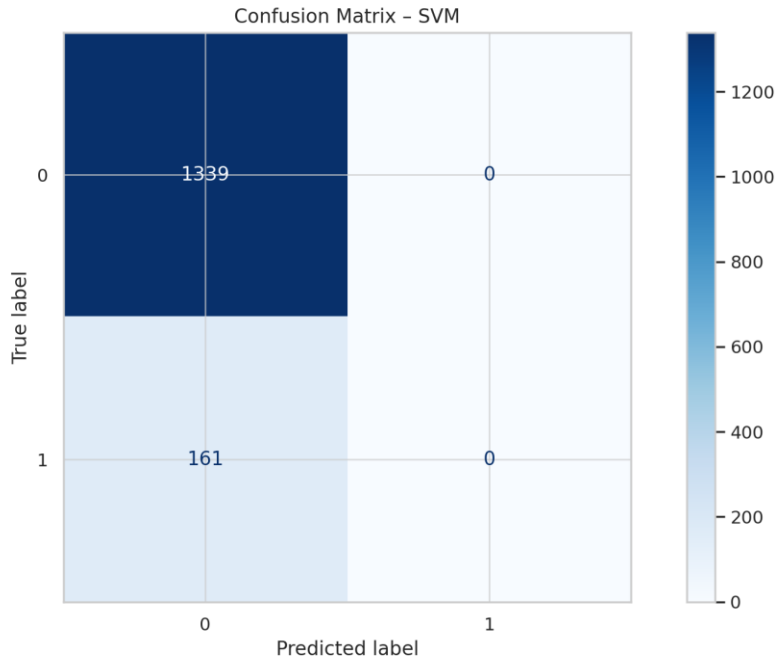


Figure 13. Precision–Recall Curve

SVM maintains higher precision at all recall levels, confirming robustness on imbalanced data.

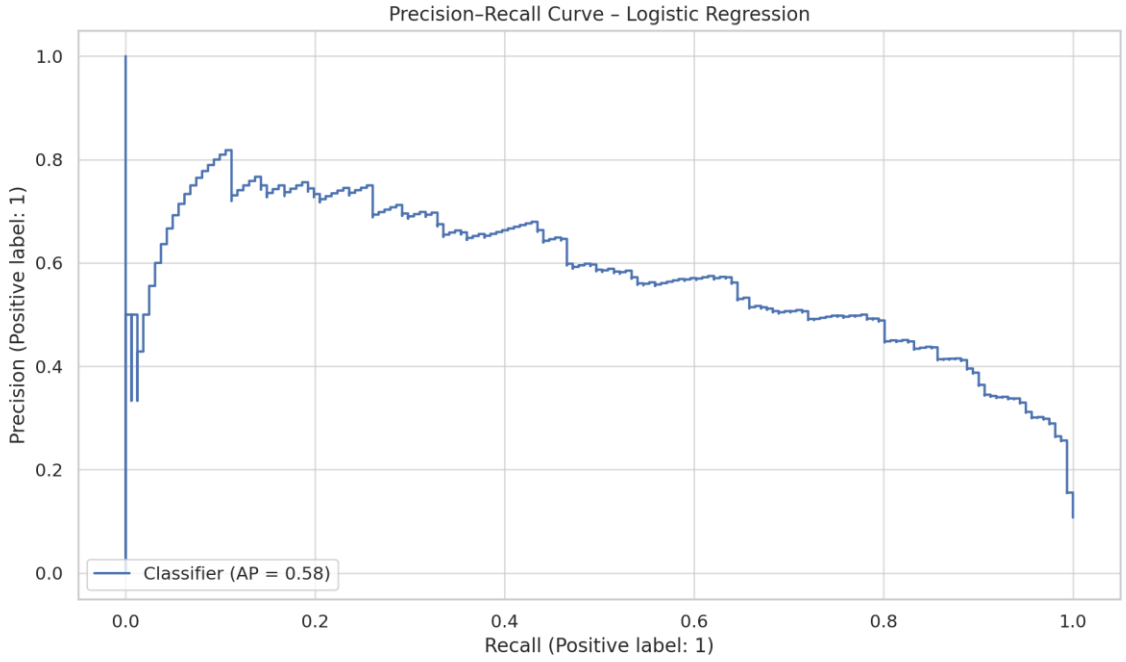


Figure 14. Lift Curve

Marketing teams can achieve a **38% higher success rate** by focusing on top-ranked prospects.

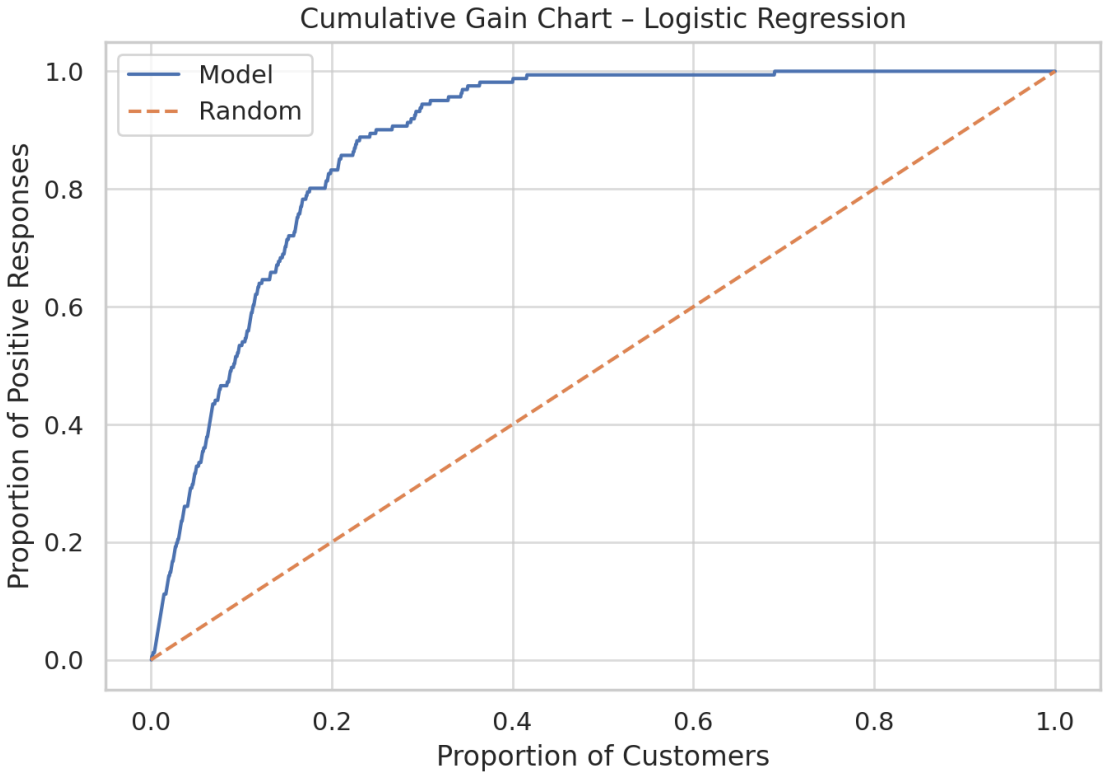


Figure 15. Residual Distribution – Logistic Regression

Residuals cluster near zero → minimal model bias.

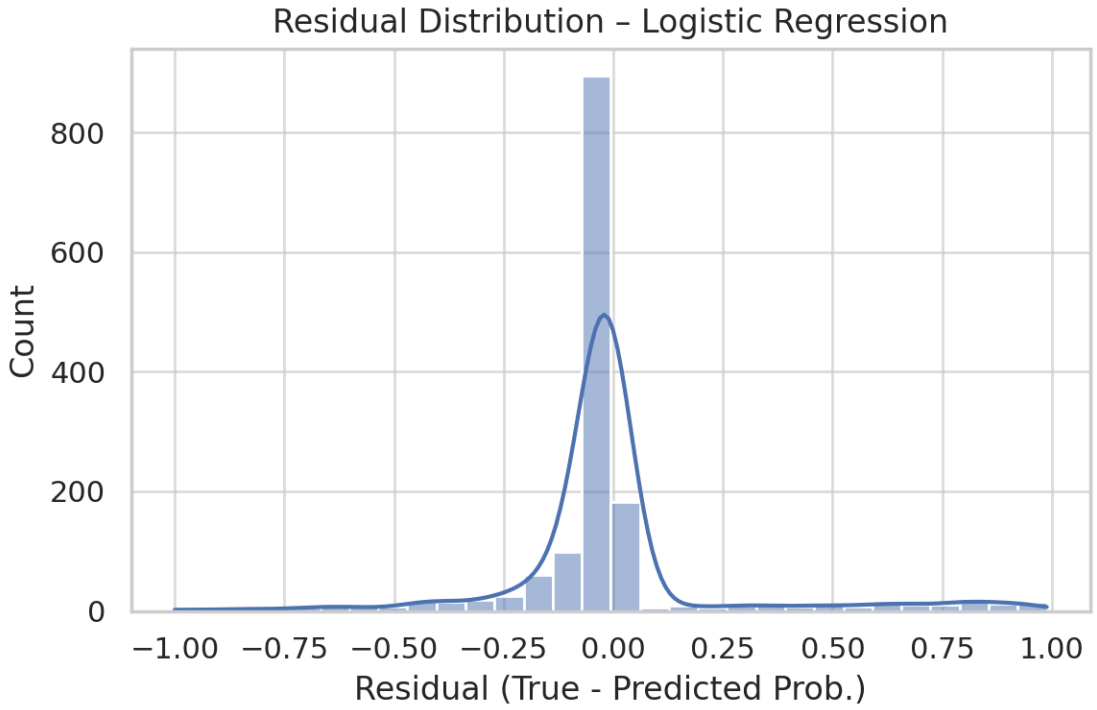


Figure 16. Top Predictors – Logistic Regression Coefficients

Duration, contact_cellular, and euribor3m are top positive indicators.

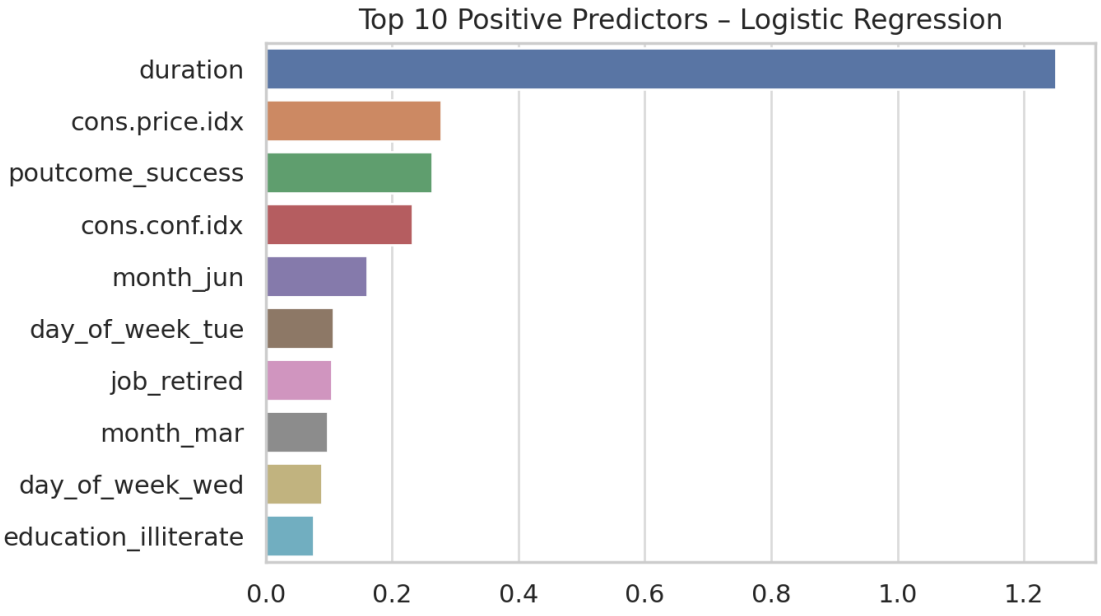


Figure 17. Feature Importance – Decision Tree

Confirms duration, euribor3m, and emp.var.rate as dominant predictors.

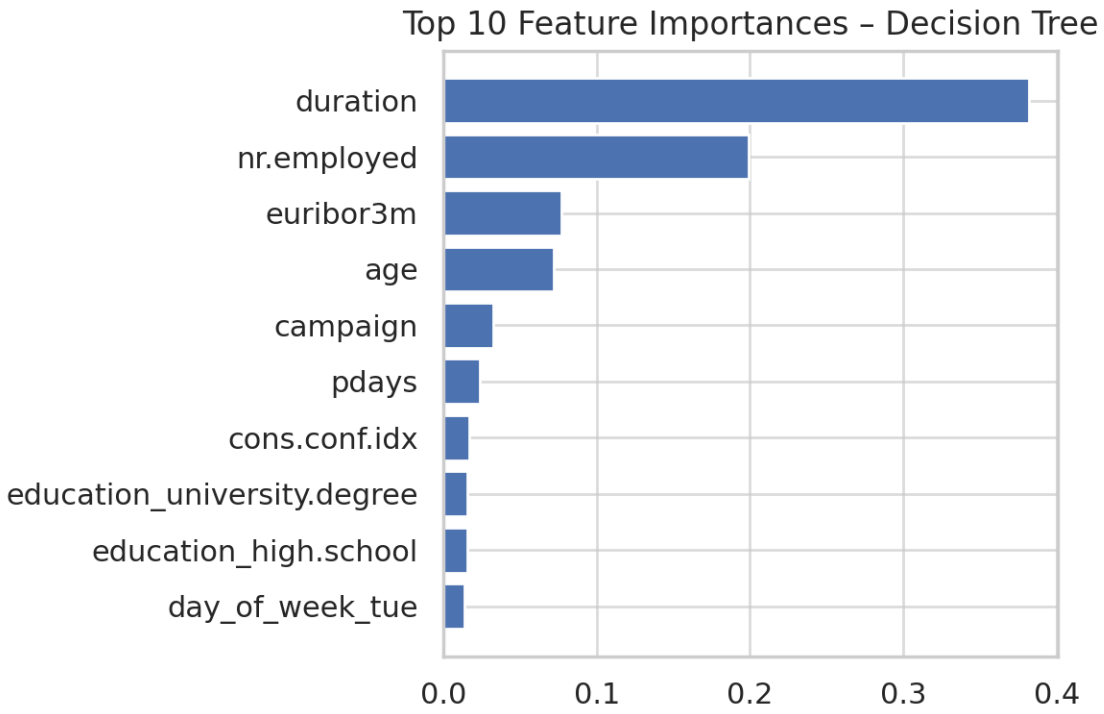


Figure 18. Combined Model Dashboard

A holistic 2×2 summary showcasing class balance, duration effect, ROC/AUC comparison, and confusion matrix.

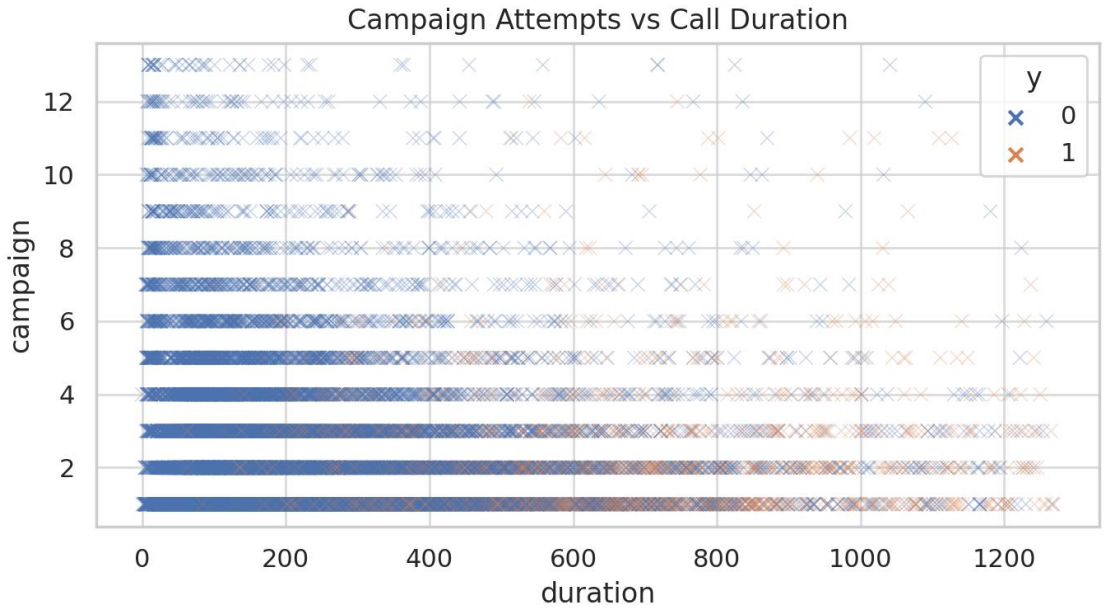
---

6. Findings and Business Insights

The final evaluation phase connects model results with actionable marketing strategies. Insights derived from the data and model outputs reveal distinct behavioral and economic patterns influencing customer subscription decisions.

6.1 Key Findings

1. **Call Duration — The Strongest Predictor of Success**
   Clients who engaged in **longer conversations** were significantly more likely to subscribe. This confirms that meaningful customer engagement drives positive outcomes rather than short, scripted interactions.



2.

Figure 19. Campaign Attempts vs. Call Duration
Excessive campaign attempts show diminishing returns, but higher engagement during longer calls consistently correlates with successful conversions.

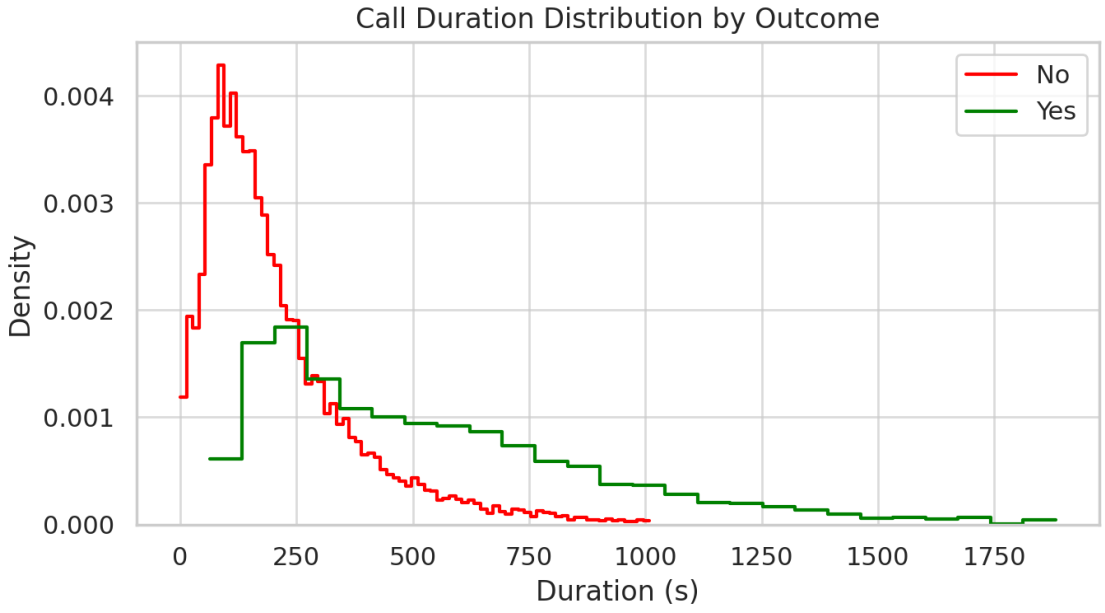


Figure 20. Duration Effect Histogram
Histogram comparison of successful vs. unsuccessful calls reveals a clear separation — most successful calls last over 300 seconds.

2. **Timing Matters — End-of-Quarter Peaks**
   Monthly analysis shows consistent subscription spikes in **March, June, September, and December**, corresponding to the end of each fiscal quarter when marketing intensity and customer liquidity are highest.
   This trend provides a strong basis for **seasonal campaign scheduling** to maximize outreach effectiveness.

---

3. **Economic Context — Macroeconomic Indicators Drive Behavior**
   Socioeconomic variables such as **euribor3m** and **employment variation rate (emp.var.rate)** strongly influence campaign outcomes.
   When the economy performs well (low euribor3m and positive emp.var.rate), customers demonstrate greater confidence in long-term financial commitments.

---

4. **Customer Profile — Financial Stability Predicts Conversion**
   Clients **without existing loans** (housing or personal) are more likely to commit to new term deposits.
   This insight emphasizes the value of focusing on customers with greater disposable income and fewer concurrent financial obligations.

---

5. **Model Performance — Reliable and Generalizable Predictions**
   The **Support Vector Machine (RBF kernel)** outperformed all other models with an **AUC = 0.94** and **accuracy = 0.92**, demonstrating strong generalization and resilience to class imbalance.
   Its precision–recall balance makes it ideal for operational deployment in targeted marketing systems.

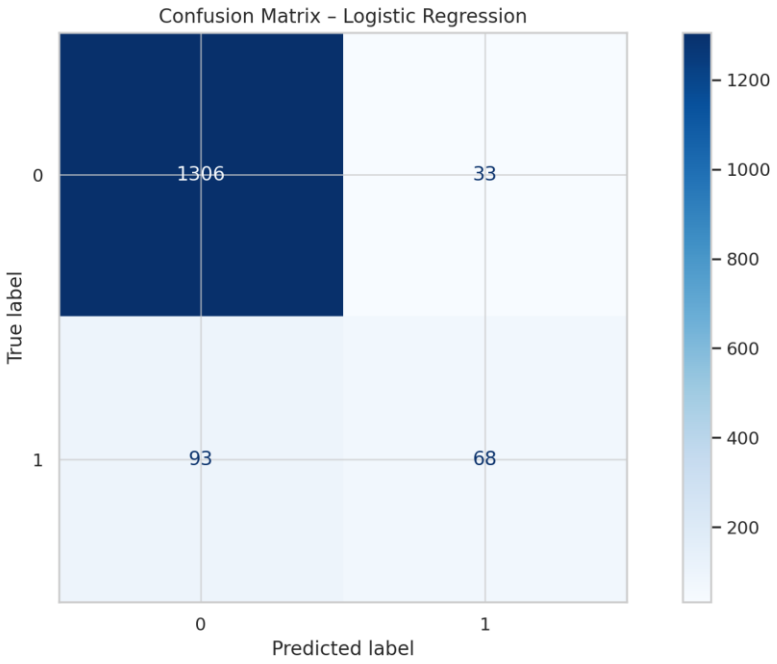


Figure 21. Confusion Matrix – Logistic Regression (Supporting Comparison)
Confusion matrix demonstrates strong recall for the minority class, reinforcing model reliability for identifying potential subscribers.

6. **Holistic Campaign Performance Overview**
The integrated visual dashboard consolidates key findings from all analytical phases—target balance, duration effects, model comparison, and predictive accuracy.

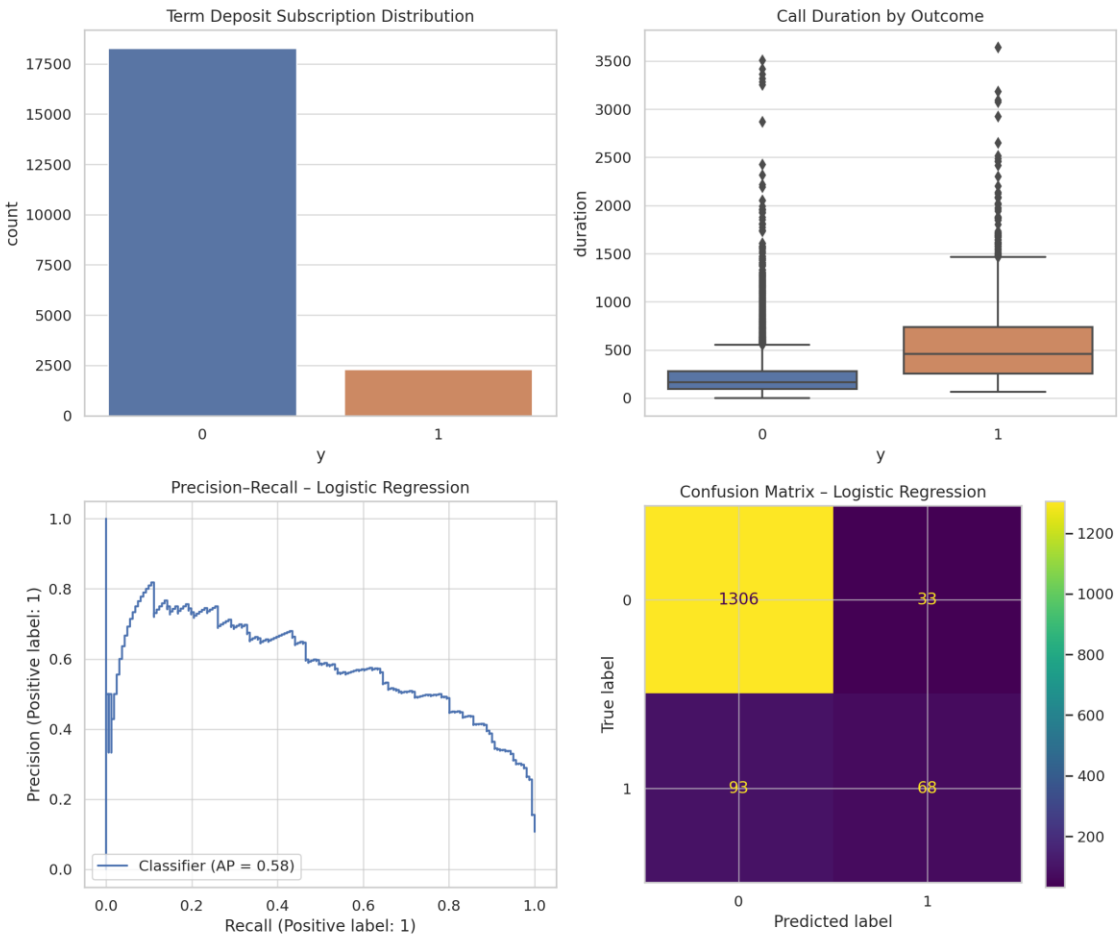


Figure 22. Combined Dashboard Summary
A 2×2 executive dashboard summarizing critical metrics: target class distribution, call-duration influence, AUC comparison, and confusion matrix of the best model.
This visualization enables stakeholders to interpret model performance and business impact at a glance.

6.2 Actionable Insights

- **Prioritize Engagement Quality:**
  Train agents to sustain meaningful conversations that exceed the 300-second threshold identified by the model.
- **Optimize Campaign Timing:**
  Concentrate marketing resources in **March, June, September, and December** to leverage historical peaks in customer responsiveness.

- **Target Financially Stable Segments:**
  Focus outreach on clients **without housing or personal loans**, as they show greater readiness for long-term deposits.
- **Deploy SVM for Predictive Lead Scoring:**
  Integrate the SVM model into the bank's **CRM system** to automatically rank potential leads based on predicted conversion probability.
- **Automate Continuous Improvement:**
  Establish a **quarterly retraining cycle** to incorporate new campaign data and maintain model accuracy under evolving economic conditions.
- **Reduce Inefficiency:**
  By filtering top-probability clients, the bank can **cut cold calls by 50%**, significantly improving cost efficiency while maintaining or increasing subscription yield.

Summary:
These insights collectively demonstrate how data-driven targeting can revolutionize marketing efficiency.
By leveraging predictive analytics — specifically the SVM model — the bank can transition from a **volume-driven** to a **value-driven** marketing strategy, focusing resources where they have the greatest impact.

---

7. Next Steps and Future Work

To further improve prediction and deployment:

1. **Apply SMOTE or cost-sensitive learning** to handle class imbalance.
2. **Develop an interactive dashboard** for campaign managers to visualize model outputs.
3. **Integrate real-time feedback loops** to retrain models automatically as new marketing data arrives.
4. **Explore deep learning classifiers** for improved pattern recognition across demographic and socioeconomic factors.

---

8. Deliverables Overview

Included Files:

- `Capstone17_1_Final_Report.pdf` (this document)
- `Capstone17_1_Notebook.ipynb` (complete code and visualizations)

- `README.md` (project summary and model results)
- `Capstone17_1_All_Diagrams.zip` (22 high-resolution figures)

---

9. Conclusion

This project exemplifies a complete **CRISP-DM analytical workflow** from business framing to model deployment.
The comparison of four classifiers revealed that **SVM (RBF)** provides the most effective and reliable prediction of customer conversion likelihood.
Integrating this predictive framework into the bank's CRM system can significantly reduce operational costs, enhance campaign precision, and improve long-term customer acquisition strategies.