Technical Annex - Loan Prediction based on Marketing Campaigns

Data Description and Preparation

The Bank Marketing Dataset contains records from a Portuguese bank's phone campaigns aimed at promoting term deposits.

- Rows: 11 162
- Features: demographic, financial, and campaign-related information.
- Target: deposit — whether the client subscribed (yes=1, no=0).

Data validation steps:

- No missing values or duplicates found.
- Numeric features (age, balance, duration, etc.) were within valid ranges.
- Outliers (e.g., large positive balances or long calls) were kept as meaningful business signals.

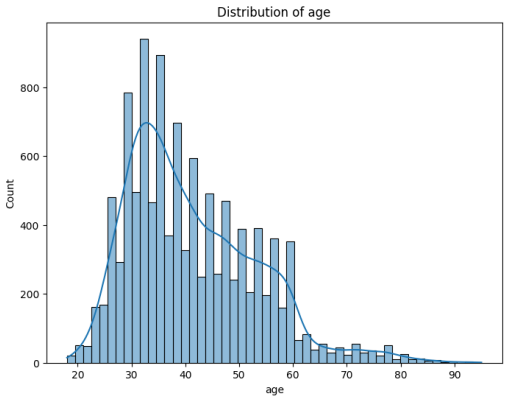
Preprocessing:

- Categorical variables were encoded using One-Hot Encoding (drop='first') to avoid imposing order and reduce multicollinearity.
- Numeric variables were scaled with StandardScaler.
- The dataset was split into 70% training and 30% test sets using train_test_split (random_state=42).

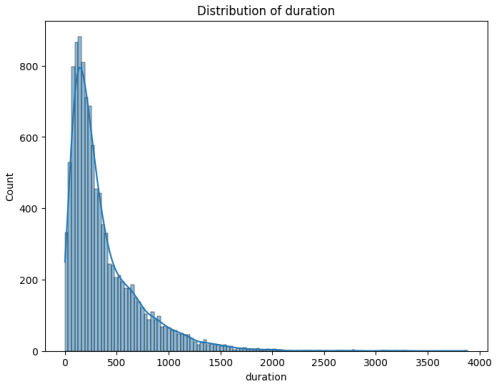
Exploratory Data Analysis (EDA)

EDA revealed key behavioural patterns:

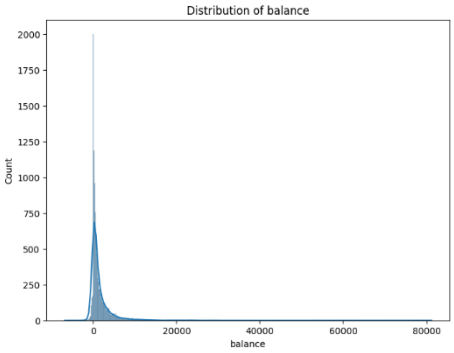
- **Age:** Most clients between 25–45 years old.



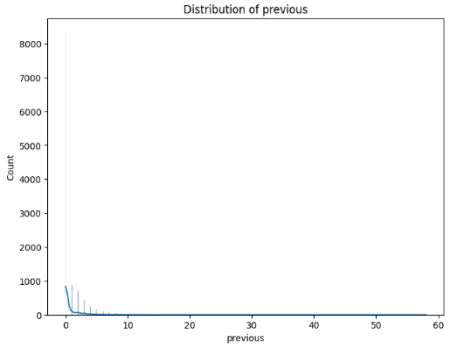
- **Duration:** Strong positive relationship with deposit success — longer calls = higher conversion.

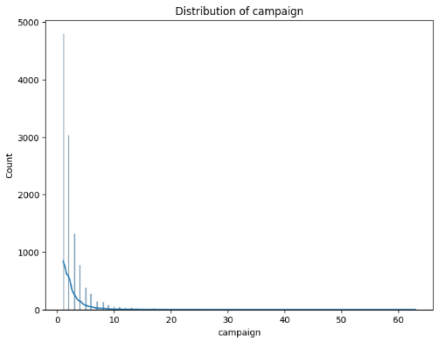


- **Balance:** Skewed toward low or negative balances; high balances correlate with deposits.

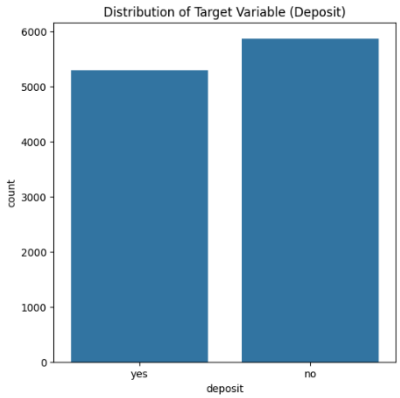


- **Campaign & Previous:** Clients contacted repeatedly or previously successful (poutcome_success) are more likely to subscribe.





- **Target variable (deposit):** Balanced distribution, avoiding class imbalance issues.



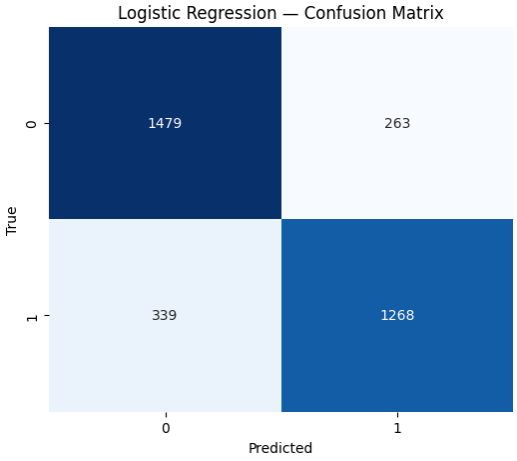
Model Development and Validation

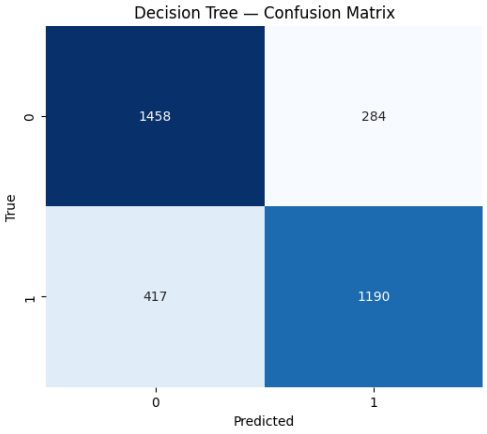
Three supervised algorithms were trained and compared using scikit-learn:

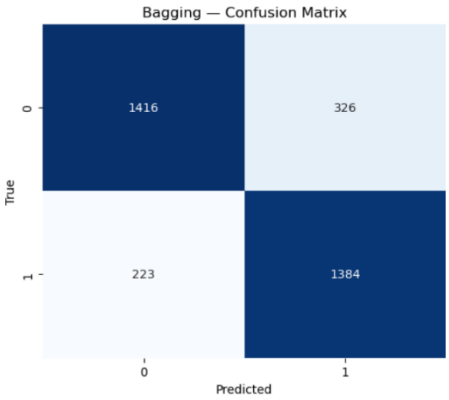
Model	Type	Accuracy	ROC-AUC	Key Notes
Logistic Regression	Linear	0.82	0.90	Baseline, interpretable
Decision Tree	Non-linear	0.80	0.87	Captures interactions, interpretable
Bagging	Ensemble	0.84	**0.915**	Best generalization and stability

Metrics used: Accuracy, Precision, Recall, F1-Score, ROC-AUC.

Validation: Hold-out test set + ROC curve visualization for each model.



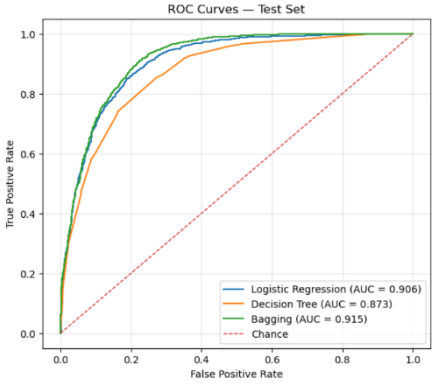




Overfitting Check

A comparison between train and test results showed **small generalization gaps (<0.03 in ROC-AUC)** for all models, indicating **no significant overfitting**.

The ROC curves confirmed consistent performance across models, with Bagging showing the best separation from the random baseline.

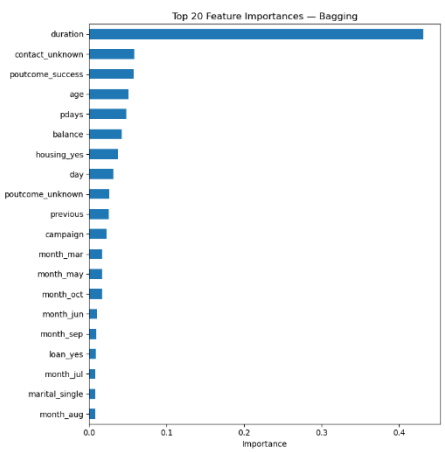


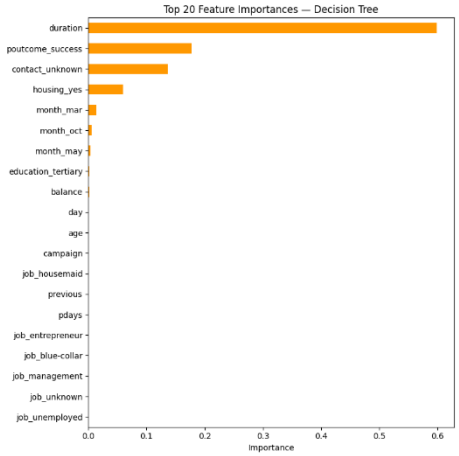
Feature Importance Analysis

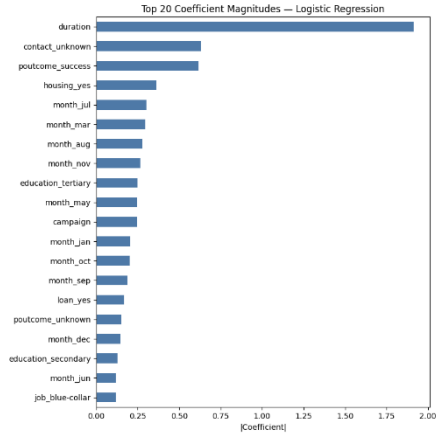
Feature importance analysis identified the most influential predictors:

Rank	Top Features	Insight
1	duration	Long calls correlate strongly with client conversions.
2	poutcome_success	Past positive outcomes increase current success probability.
3	contact_unknown	Contact type affects reachability and response.
4	housing_yes	Indicates financial stability and cross-sell potential.
5	balance / previous	Capture client wealth and prior campaign activity.

- Low-importance variables included month, marital status, and several job dummies.
- However, removing these features led to a slight performance drop.
- Hence, all encoded features were retained, as minor variables provided complementary predictive value.







Conclusions

- Best Model: Bagging (Accuracy = 0.84, ROC-AUC = 0.915)
- Interpretability: Decision Tree remains useful for explaining rules to managers.
- Data Quality: Clean dataset, no imputation or resampling required.
- Encoding Choice: Switching from LabelEncoder to One-Hot Encoding improved linear model performance (+0.02 ROC-AUC).

The technical analysis confirmed that all three models generalized well and that ensemble methods provided the strongest predictive power.

Retaining all features after encoding produced the best performance and stability.

These results validate the robustness of the Bagging model for deployment as a client lead-scoring system.