Capstone Project 2 - Bank Marketing Prediction (Imbalanced Classification with SMOTE)

Kevin Chen

Thanks to Springboard mentors: Kevin

Springboard

Why Term Deposits Are Important for Banks

- Provide **stable, long-term funding** (unlike regular savings accounts)
- Create **predictable future cash flows**
- Serve as a **low-risk source of capital**
- Help banks **retain customers** for longer periods

Why Term Deposits Are Important for Banks

- Provide **stable, long-term funding** (unlike regular savings accounts)
- Create **predictable future cash flows**
- Serve as a **low-risk source of capital**
- Help banks **retain customers** for longer periods

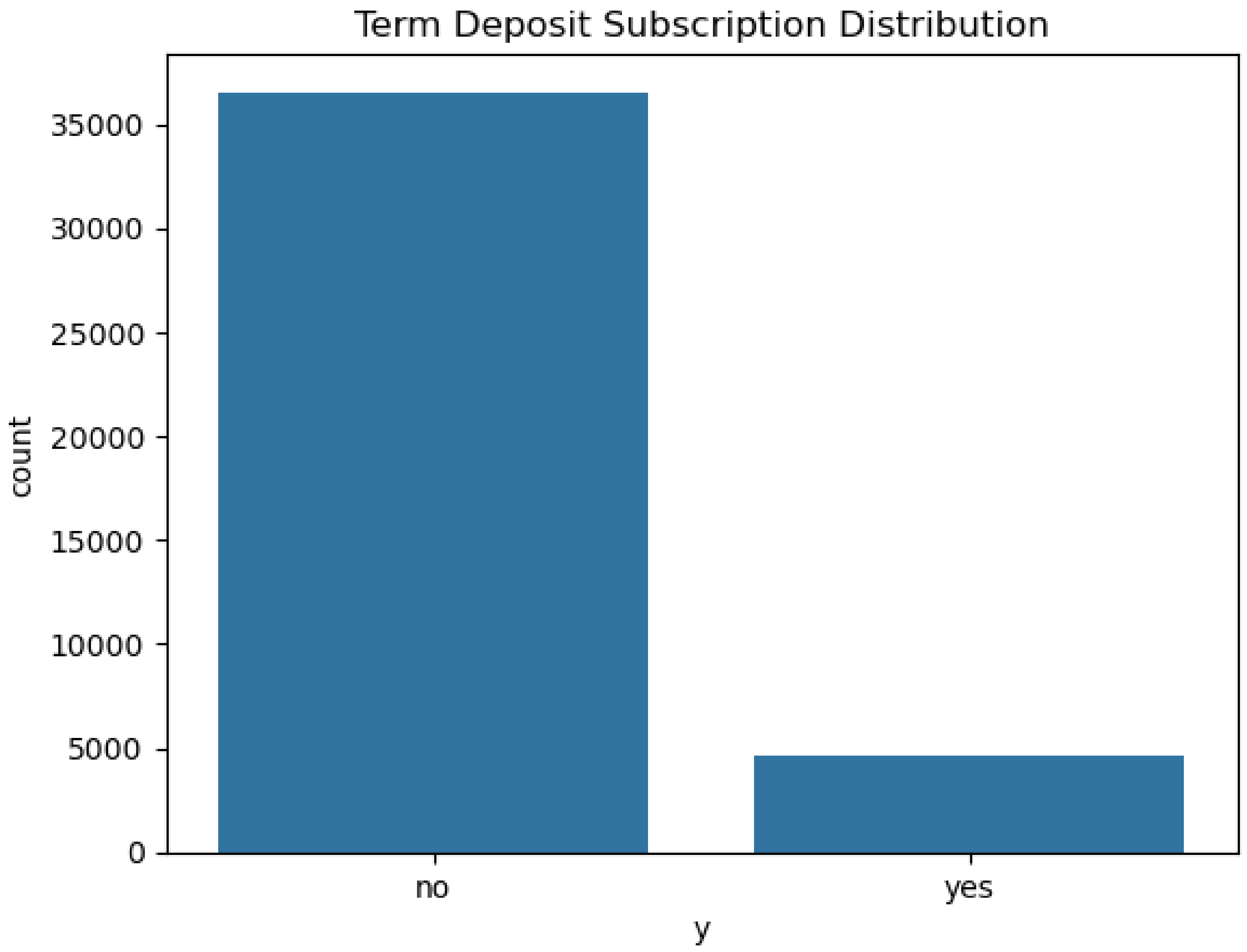
Business Problem & Goal Portuguese Bank Marketing Campaign (UCI)

- Direct marketing by phone is costly and time-consuming
- Only a small fraction of clients subscribe to a term deposit (strong class imbalance)

- **Goal:** predict which clients are likely to subscribe so that the bank can prioritize them

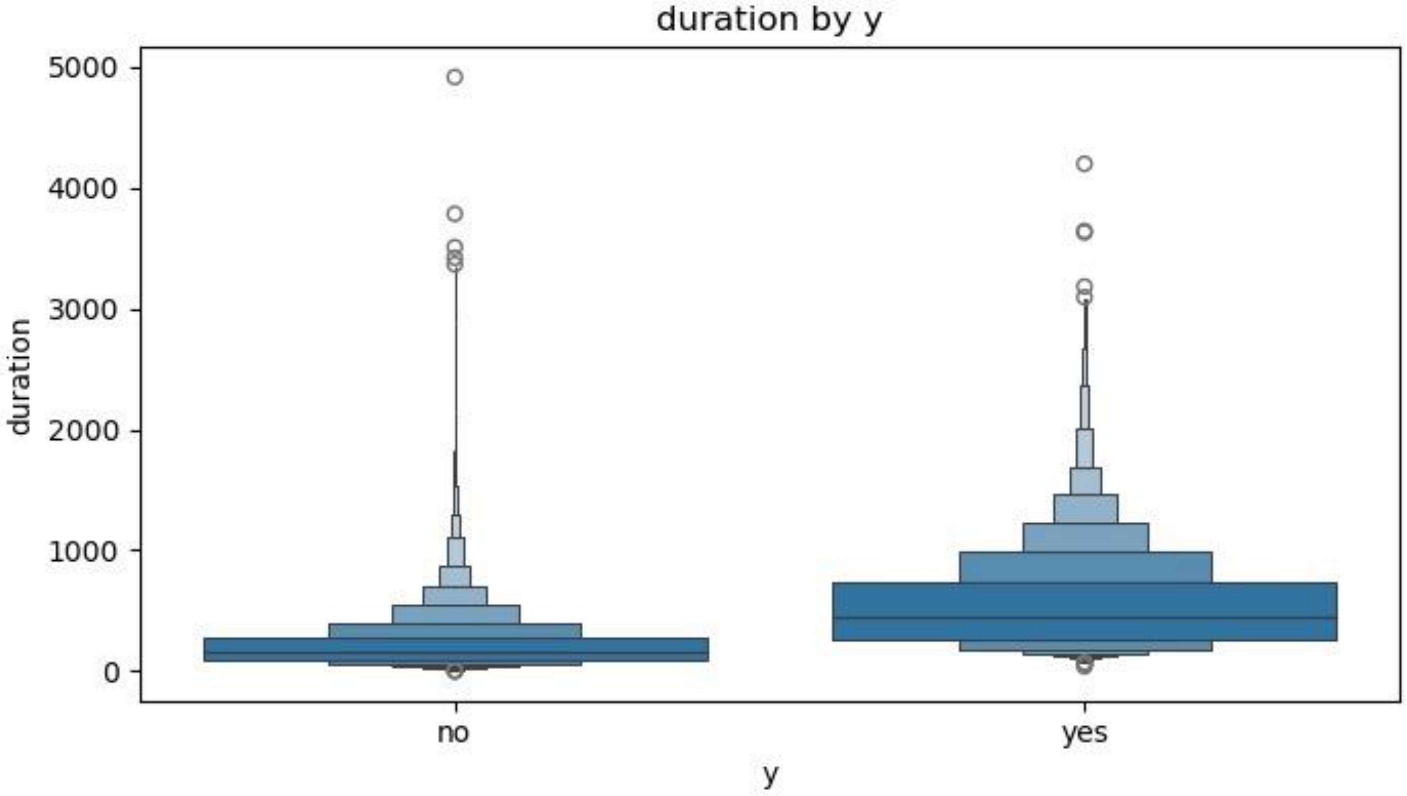
Data Overview

- Source: UCI Bank Marketing Dataset (Portuguese bank)
- 41,188 rows; >20 features (demographic, contact history, economic indicators)
- Target variable: y (term deposit subscription: yes / no)
- Strongly imbalanced target:
  - no - 88.73%
  - Yes – 11.27%

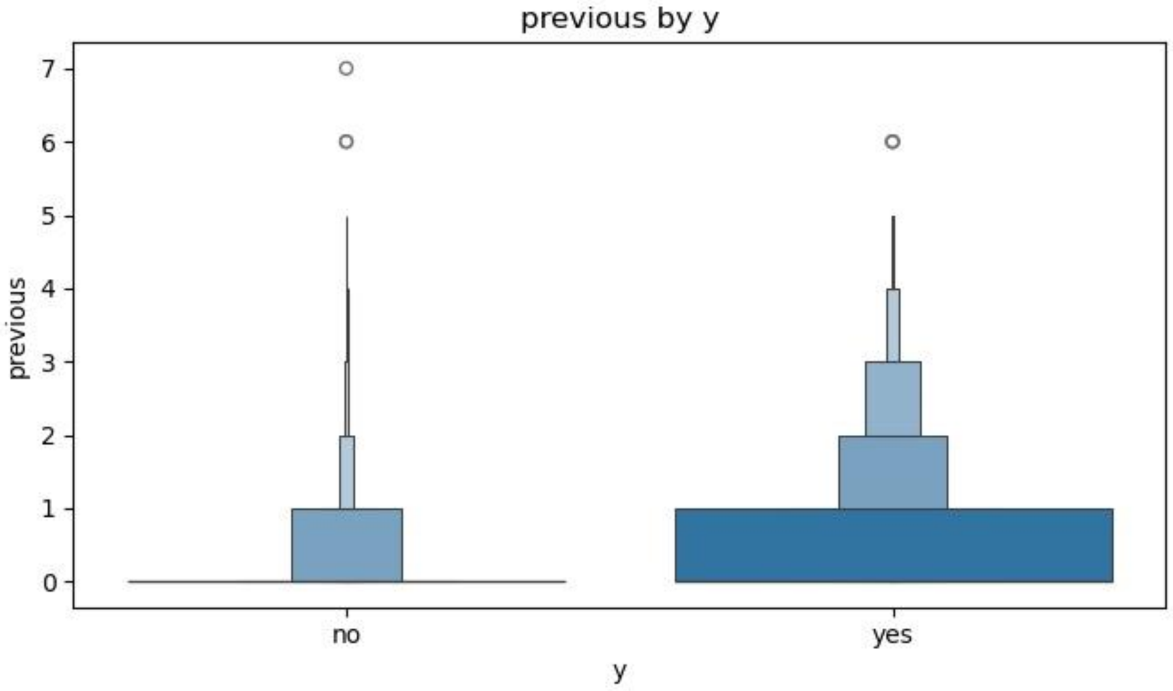


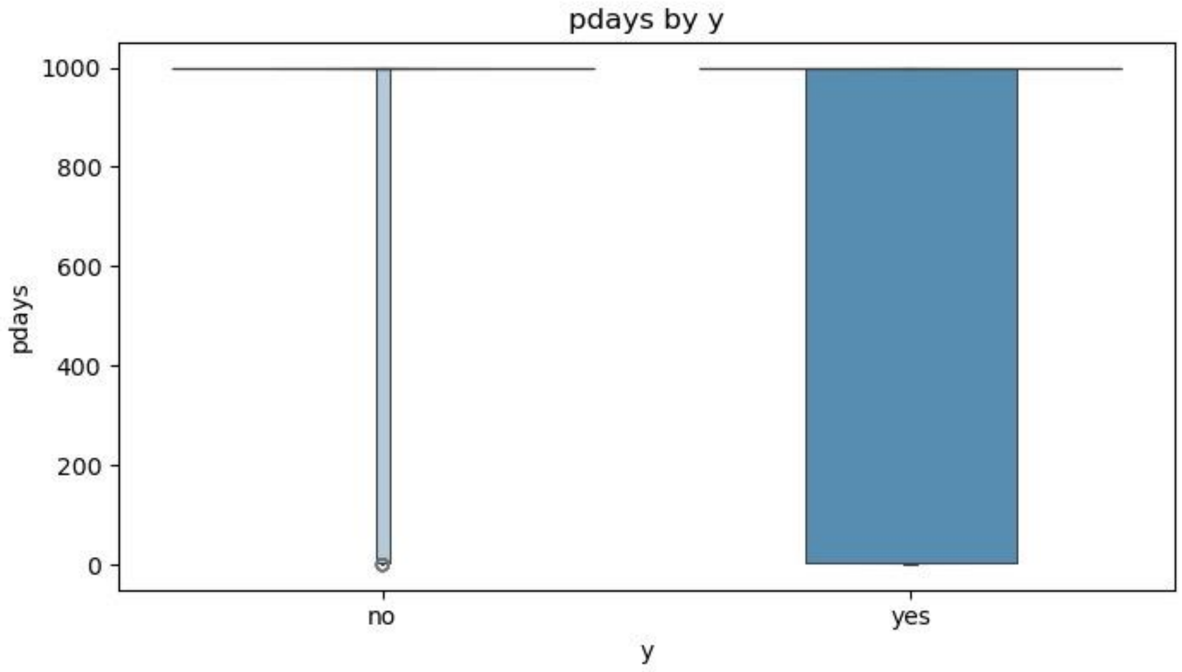
Duration vs Target (y)

- Longer call duration strongly associated with subscription



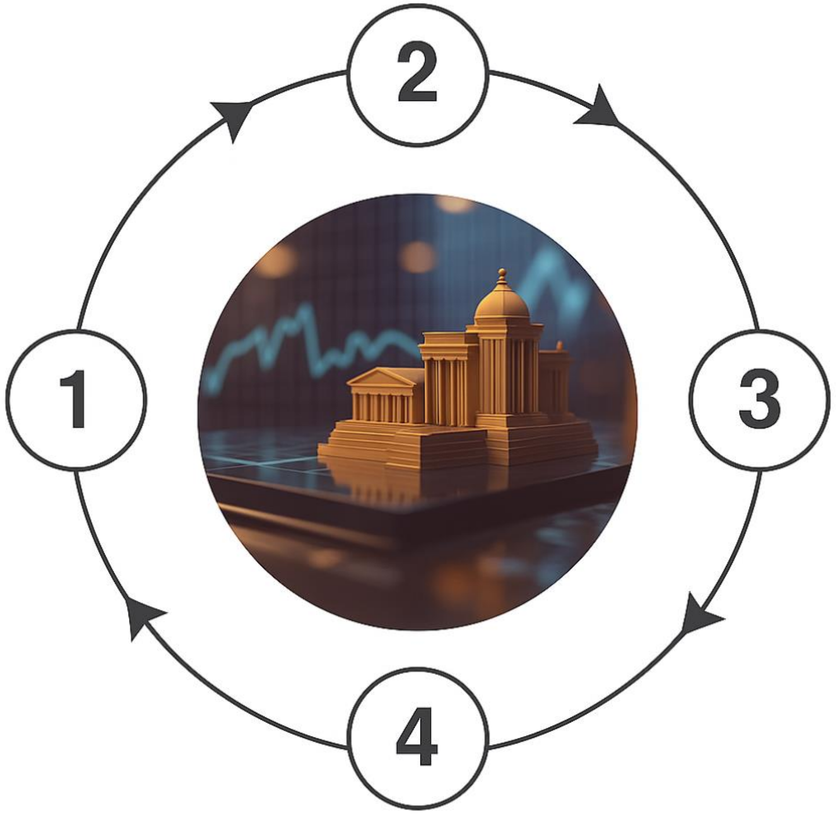
Previous/pdays vs Target (y)





Modeling Strategy

Pipeline (training folds)
- One-hot encode categorical features
- Standardize numeric features
- Apply SMOTE to oversample the minority class



Train–test split
80/20 with stratification

Evaluate with **5-fold Stratified Cross-Validation** on the training set

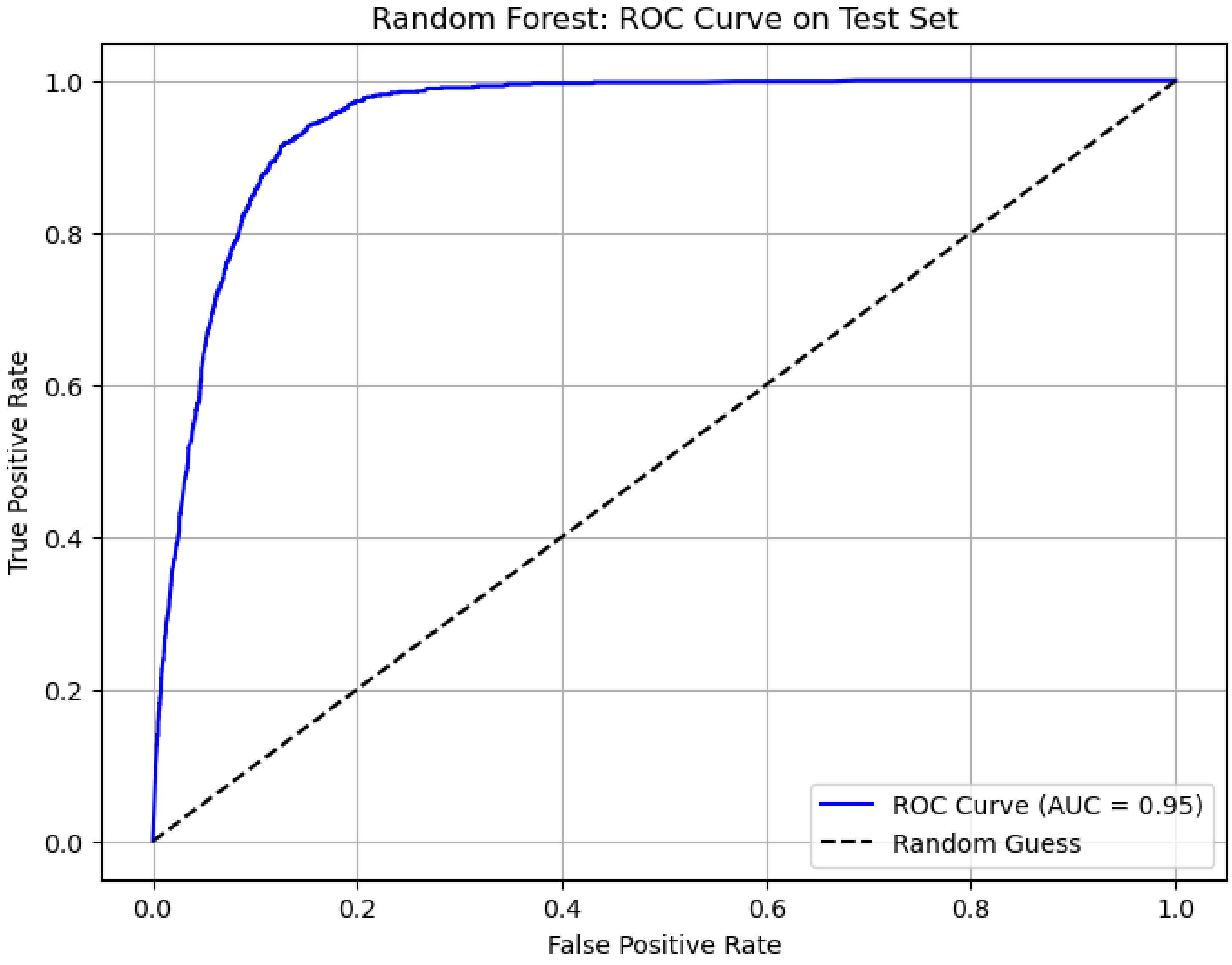
Focus metrics
Recall, F1, ROC AUC (accuracy is misleading for imbalanced data)

Models Compared

Model	Avg F1 Score	Avg Recall Score	Avg ROC AUC
LightGBM	0.6367	0.7928	0.9422
Logistic Regression	0.5821	0.7398	0.9163
Random Forest	0.6247	0.6875	0.9411
XGBoost	0.6111	0.6479	0.9415

Final Model Performance LightGBM on Test Set

- Precision ≈ 0.49
- Recall ≈ 0.89
- F1-score ≈ 0.63
- ROC AUC ≈ 0.95
- Confusion matrix: **very low false negatives** (many true subscribers correctly identified)



Feature Importance & Interpretation
What Factors Matter Most?

Top features from LightGBM:

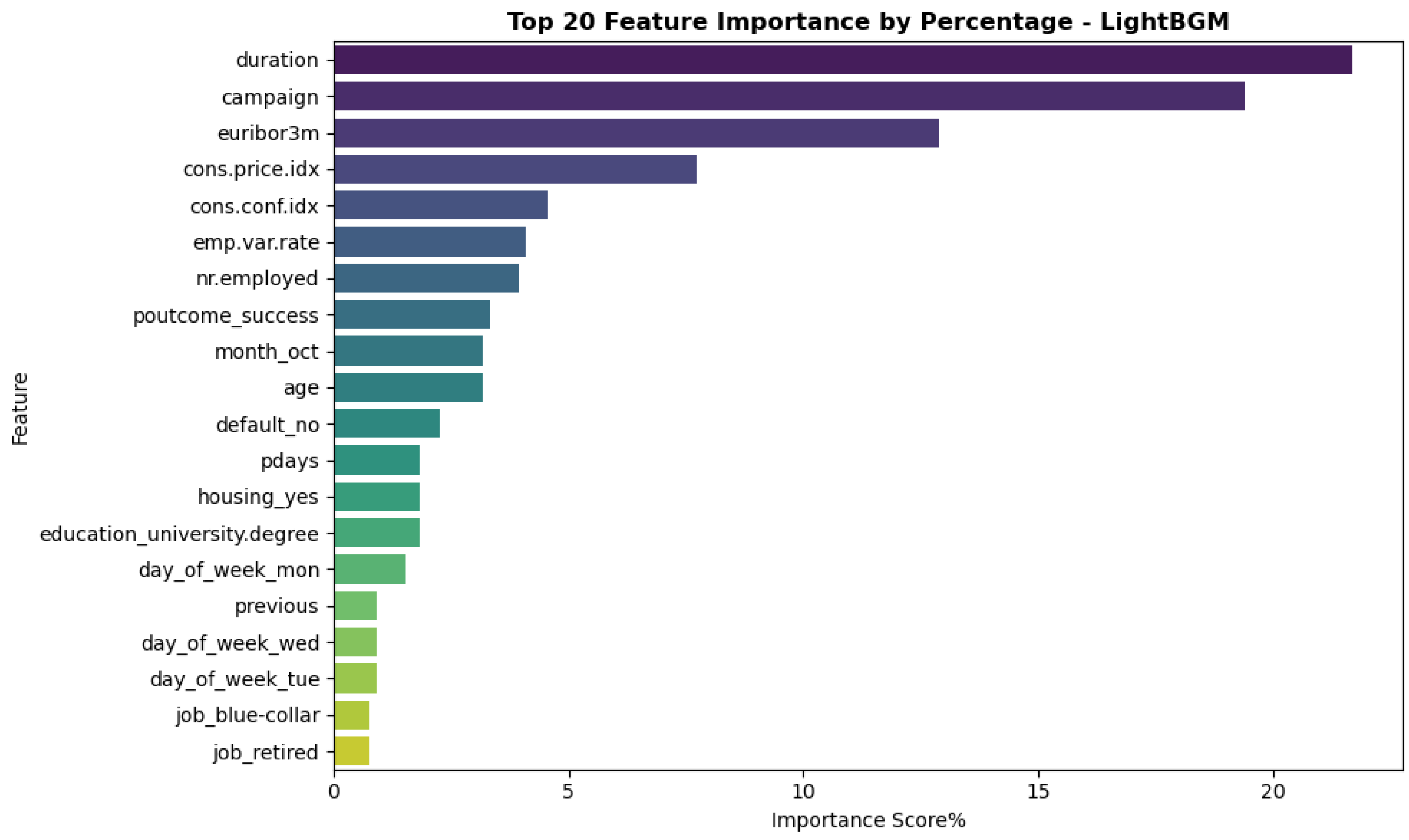
•Call duration

•Previous campaign outcome

•Number of contacts in this campaign

•Employment variation rate

•Euribor 3-month rate



Business Recommendations

- **Prioritize high-score clients** before launching campaigns (use model scores to generate call lists)
- **Focus on call quality, not only call volume** – train agents to have longer, more meaningful conversations
- **Re-target clients with past positive responses** or good engagement history
- Consider integrating economic indicators into campaign planning (avoid calling during bad macro periods)

Business Recommendations

Prioritize

Prioritize high-score clients before launching campaigns (use model scores to generate call lists)

Focus

Focus on call quality, not only call volume – train agents to have longer, more meaningful conversations

Re-target

Re-target clients with past positive responses or good engagement history

Limitations & Future Work

- Adjust threshold 0.50/0.50
- Ensemble sampling
- SHAP
- Business input → decision boundary

Capstone2 - Bank Marketing Campaign

Thanks for your time