Customer Churn-Style Subscription Prediction in Banking Campaigns

Team: Demon Slayer Corps

Team Members

Name	**Roll No.**	**Dept.**
Revanth Singothu	CB.EN.U4CSE22149	CSE
Devesh Keshavan	CB.EN.U4CSE22509	CSE
Siddharth Swamy	CB.EN.U4CSE22529	CSE
Nandha Kumar P	CB.EN.U4CSE22530	CSE
Mukesh Narravula	CB.EN.U4CSE22531	CSE

September 4, 2025

Abstract

Banks expend substantial budgets on direct marketing—phone calls, SMS and email—to promote term deposits and other retail products. Blanket outreach is inefficient: it wastes contact resources, increases operational cost, and can damage customer relationships when offers are irrelevant. This project, *Customer Churn-Style Subscription Prediction in Banking Campaigns*, reframes subscription prediction as a churn-analogue problem and builds a pipeline to identify the *persuadable* customers most likely to subscribe if contacted. Using the UCI Bank Marketing dataset augmented with macroeconomic indicators, we apply careful ETL, temporal feature engineering, balanced modeling, and uplift/causal techniques to distinguish customers whose behavior changes due to outreach. Models include logistic regression for interpretability, ensemble learners for performance (Random Forest, LightGBM), and specialized uplift learners (T-/X-Learner, Causal Forest). Deliverables include reproducible notebooks, a dashboard for campaign optimization, and an ROI simulator to guide spend. Emphasis is placed on deployment-viable models, clear interpretability (SHAP/LIME), and a business playbook so marketing teams can operationalize recommendations while safeguarding privacy and regulatory compliance.

1 Introduction

In the financial sector, marketing campaigns are a major investment, with billions spent annually on outreach to customers for deposits, loans, and investment schemes. Poor targeting leads to wasted costs and erosion of trust. Predictive analytics provides banks with an opportunity to shift from intuition-driven to data-driven marketing, identifying customers who are most receptive to offers.

This project treats subscription prediction as analogous to churn prediction: just as churn analytics identifies customers at risk of leaving, our work focuses on customers who are most persuadable to join. By leveraging the UCI Bank Marketing dataset and incorporating macroeconomic indicators, the project aims to create a robust model for predicting subscription likelihood. Our emphasis is not just prediction accuracy but actionable recommendations, integrating machine learning, uplift modeling, visualization tools, deployment-ready practices, and explainability tools (SHAP/LIME) to ensure outputs are interpretable and directly usable by marketing teams while maintaining privacy and regulatory compliance.

2 Objectives

1. Develop predictive models to score customers by their likelihood to subscribe to banking products and identify persuadable segments.
2. Implement uplift and survival analysis to distinguish customers who respond because of outreach (causal effect) versus those who would subscribe regardless.
3. Build a dashboard-driven ROI simulator to inform marketing spend allocation and scenario testing, including integration of predictive models with Tableau.

3 Team Structure and Roles

Name	**Role and Responsibilities**
Revanth Singothu	Project Manager / Scrum Master – coordinate sprints, manage milestones, ensure QA, documentation, and deployment oversight.
Devesh Keshavan	Data Engineer – design ETL pipelines, data validation, integration of ECB indicators, version control, Docker environments.
Mukesh Narravula	Data Analyst – exploratory analysis, visualization, dashboard development (Tableau/PowerBI).
Siddharth Swamy	ML Engineer – model development, uplift modeling, evaluation, deployment pipelines, and metric tracking.
Nandha Kumar P	Business Analyst – ROI modeling, persona definitions, business recommendations and stakeholder communication.

4 Data Sources

Primary Dataset: The UCI Bank Marketing dataset (approx. 45,211 records) containing customer demographics, campaign contact information and the subscription target. Key attributes: age, job, marital status, education, default, balance, housing, loan, contact type, campaign, pdays, previous, poutcome, and the binary target 'y' (subscription). Expected size: 45k rows, suitable for robust modeling and cross-validation.

Secondary Dataset: European Central Bank (ECB) Statistical Data Warehouse for macroeconomic indicators (interest rates, inflation proxies, unemployment rates). These will be joined by month/year to capture external economic conditions that influence deposit behavior.

Justification: Combining individual-level campaign records with macro indicators helps control for temporal effects and can improve generalization. Both sources are open and reproducible, enabling transparent evaluation and business-aligned insights.

5 Methodology

Our end-to-end methodology includes:

- **ETL & Cleaning:** Remove personally identifiable information, handle missing data, convert categorical fields, and create a production-safe feature set (exclude 'duration' in production models). Tools: Python, pandas, NumPy, Docker environments.
- **EDA:** Univariate/bivariate analyses, class balance checks, contact-channel effectiveness, temporal trends, clustering for personas. Tools: Python, matplotlib, Tableau.
- **Feature Engineering:** Recency-frequency features, campaign counts, macro-variable lags, interactions, PCA for high-dimensional features.
- **Modeling:** Logistic regression (interpretability), Random Forest LightGBM (performance), hyperparameter tuning, calibration to business cost curves. Libraries: scikit-learn, LightGBM, XGBoost.
- **Causal / Uplift Modeling:** T-/X-Learners, Causal Forests to estimate treatment effect and isolate persuadable customers. Libraries: econml, causalml.
- **Survival Analysis:** Kaplan–Meier and Cox proportional hazards to predict time-to-subscription. Library: lifelines.

- **Validation & Metrics:** Time-aware cross-validation, PR-AUC, ROC-AUC, Precision@K, uplift Qini/AUUC curves, incremental conversion metrics.
- **Visualization:** Tableau dashboards with uplift deciles, persona filters, cost-benefit simulators, scenario testing.
- **Deployment Notes:** Package pipelines (pickle/ONNX), REST API (Flask/FastAPI), schedule nightly syncs, weekly retraining, monitor population drift.
- **Ethics & Compliance:** Anonymize personal data, minimize profiling-sensitive decisions, human-in-the-loop sign-offs.

6 Project Plan (Sprints & Gantt Overview)

Sp.	**Week(s)**	**Tasks**
1	W1–W2	ETL, schema design, exploratory analysis, missing data strategy.
2	W3–W4	Feature engineering, baseline models, initial evaluation.
3	W5–W6	Advanced models (ensemble), uplift learners and survival analysis.
4	W7	Dashboard development, ROI simulator, persona-based insights.
5	W8	Final validation, documentation, presentation prep and hand-off.

Task	W1	W2	W3	W4	W5	W6	W7	W8
Data Prep	X	X						
EDA	X	X	X					
Feature Eng.		X	X	X				
Baseline Models			X	X				
Advanced Models				X	X	X		
Uplift Analysis					X	X		
Dashboards							X	
Final Delivery								X

7 Expected Deliverables

- Code repository with reproducible Jupyter notebooks (data cleaning, feature engineering, modeling).
- Predictive models, uplift estimators and calibration results.
- Tableau dashboard showing uplift deciles, persona filters and ROI simulator.
- Business playbook: persona profiles, campaign recommendations and cost thresholds.
- Final presentation slides and documentation for stakeholders.

8 Risk Assessment

Risk	**Mitigation**
Imbalanced target classes	Use models that handle unequal classes, create synthetic examples, and evaluate with precision-recall metrics.
Data leakage risk	Remove variables like duration from production models and follow a strict feature-checking process.
Temporal drift	Test on later months and monitor changes in data distribution.
Interpretability challenges	Use explanation tools and create summaries that are easy for stakeholders to understand.
Regulatory / privacy concerns	Anonymize personal information and follow data retention rules.
Deployment drift	Check models regularly and set triggers for retraining.

References

1. S. Moro, P. Cortez, and P. Rita, "A data-driven approach to predict the success of bank telemarketing," *Decision Support Systems*, vol. 62, pp. 22–31, 2014.

2. P. Burez and D. Van den Poel, "Handling class imbalance in customer churn prediction," *Expert Systems with Applications*, vol. 36, no. 3, pp. 4626–4636, 2009.

3. European Central Bank, "Statistical Data Warehouse: Economic Indicators." [Online]. Available: `https://data.ecb.europa.eu`

4. J. Neslin, S. Gupta, W. Kamakura, and D. Lu, "Defection detection: Measuring and understanding the predictive accuracy of customer churn models," *Journal of Marketing Research*, vol. 43, no. 2, pp. 204–211, 2006.