BUSINESS PROPOSAL: PREDICTIVE ENGAGEMENT MODELING FOR RETAIL BANKING

Prepared for: Stakeholders of the Bank Marketing Campaign

Prepared by: Christian Shannon, Ashley Love, Mugtaba Awad, and Kirsten Livingston

Bellevue University

DSC 450-T201 Applied Data Science

Professor Fadi Alsaleem

January 18, 2026

Project Deliverables & Repository Access
GitHub Repository: https://github.com/cashannon/Predictive-Engagement-Modeling-For-Retail-Banking
(*Contains source code, Jupiter Notebooks, and the bank-additional-full.csv dataset)*
Recorded Presentation:
https://www.canva.com/design/DAG9koWKaK4/vhbgQZ2pM8lB56cZI8DwKg/edit
(*Final presentation deliverable as outlined in Part B of the Appendix)*

Executive Summary This proposal outlines a predictive framework to move from mass-outreach to targeted engagement. By identifying a "Risk Zone" in the middle-aged demographic, we achieved a 49.5 % precision rate, providing a scalable ROI-driven solution for retail banking.

Table of Contents

INTRODUCTION

In the modern retail banking landscape, the cost of acquiring new customers often exceeds the cost of retaining existing ones. This project utilizes the "Bank Marketing" dataset ('bank-additional-full.csv') from a Portuguese banking institution to investigate customer engagement patterns during term-deposit telemarketing campaigns. While the original objective was to predict which clients would subscribe to a term deposit after phone contact, this work reframes the problem as a propensity-to-engage study. By analyzing socio-demographic, behavioral, and macroeconomic variables, the goal is to shift from inefficient mass calling to a data-driven predictive targeting approach. (Moro et al., 2014).

BUSINESS PROBLEM/HYPOTHESIS

The core business problem is the low effectiveness of current outbound term-deposit campaigns, where approximately 88% of contacted clients do not subscribe. This high "no" rate drives substantial operational cost per successful conversion and increases the risk of customer fatigue –defined here as "engagement churn". Our hypothesis is that by identifying a specific "risk zone" of low-engagement profiles, the bank can optimize outreach frequency and timing to improve ROI.

Methods/Analysis

Our methodology followed a specialized pipeline divided into three critical roles to ensure a rigorous and repeatable process (McKinney, 2022).

Data Wrangling (Step-by-Step Preprocessing)

- Structure & Imbalance: We loaded the bank-additional-full.csv file with an explicit semicolon separator to parse all 21 original columns. We performed a deep inspection of structure and types, confirming the absence of null values while identifying a significant class imbalance (88.7% “no” vs. 11.3% “yes”).
- Cleaned Contact History: Created a new binary indicator (pdays_never_contacted) for clients never previously reached and recoded the placeholder value 999 to 1 to align with original documentation conventions.
- Managed Categorical Noise: Profiling revealed “unknown” values in features like job and education; these were retained as explicit categories to preserve potential signal rather than being imputed or dropped.
- Feature Expansion: Utilized a OneHotEncoder on all categorical predictors, dropping the first category to avoid the “dummy variable trap.” This expanded our feature space from 21 to 54 fully numeric columns suitable for machine learning algorithms.

Data Modeling (Pipeline Construction & Validation)

The modeling workflow framed the task as a binary classification problem using a stratified training and test split (80/20) to maintain the original class imbalance. To ensure model integrity and predictive power, we implemented the following.

- Leakage Prevention: The ‘duration’ variable was explicitly excluded prior to model training, as it is only known after a call concludes.
- Class Balancing : We applied Synthetic Minority Over-Sampling Technique (SMOTE) strictly to the training data to address skewed classes without leaking information into the test set.
- Model Comparison: We trained a Random Forest Classifier and a Logistic Regression baseline, evaluating

them on ROC- AUC and precision-recall curves (GeeksforGeeks, 2025).

Data Visualization(Signal Detection & Validation)

The visualization step focused on turning raw data into clear signals for disengagement and “churn” risk. We utilized several advanced plotting techniques to validate our statistical findings.

- Density Mapping: We generated KDE heatmaps and hexbin plots of age vs. duration to identify a dense “Risk Zone” among middle-aged clients (30-50) with short conversations.
- Distribution Analysis: Normalized bar charts compared subscription outcomes across education levels, while boxplots of call duration by outcome highlighted “soft churn” patterns.
- Economic Correlation: We utilized KDE plots for macroeconomic indicators (like euribor3m) and a numeric correlation heatmap to visualize how external economic factors influenced customer engagement.
- Feature Priority: A final feature-importance bar chart was extracted from the Random Forest model to confirm that behavioral attributes, especially contact frequency, were the strongest drivers of churn.

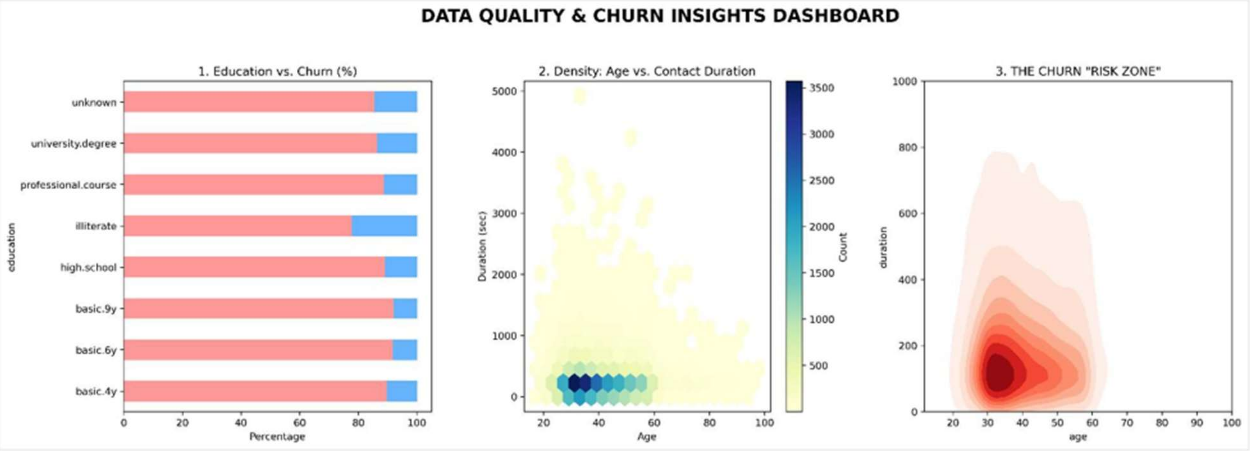
RESULTS

The analysis revealed that targeting the top 10% highest probability customers yielded a precision of 49.5%, nearly five times the baseline rate.

- Model Performance: The Random Forest with SMOTE achieved 89.1% accuracy and a 78.3% ROC-AUC. Targeting the top 10% highest-probability customers yielded a precision of 49.5%.
- Primary Drivers: Feature importance pinpointed call duration as the dominant driver, followed by economic indicators like the Euribor 3-month rate.
- The Fatigue Point: Subscription rates drop sharply after 3-4 campaign contacts, marking a threshold for customer fatigue.

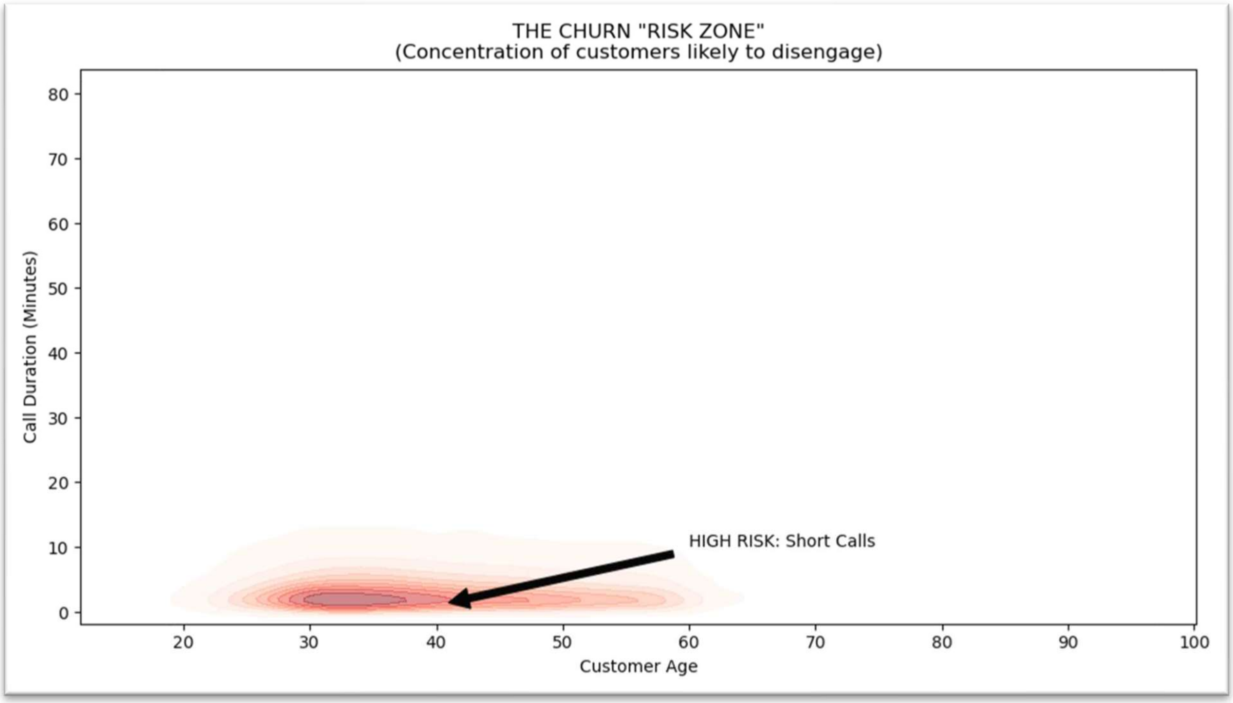
Visualizations

Figure 1: Data Quality and Engagement Insights Dashboard



Stakeholder Interpretation: This dashboard visualizes the baseline 88/12 class imbalance and identifies “soft churn” customers through duration analysis. It allows management to see that disengagement is often visible early in the call cycle.

Figure 2: The Engagement “Risk Zone” Heatmap



Interpretation: This KDE density plot highlights the concentration of likely disengagement.
Business Action: The dark “Red Zone” identifies customers aged 30-50 with calls under 200 seconds. Representatives should be trained to conclude these low propensity calls quickly to save resources.

Recommendations and Ethical Considerations

Target Outreach: Prioritize the top 10% of customers ranked by the Random Forest model to maximize conversion precision.

Contact Limits: Limit campaign contacts to a maximum of 3 per customer, as subscription rates drop sharply beyond this threshold.

Economic Strategy: Monitor indicators like ‘euribor3m’ and pause outreach during unfavorable economic conditions.

Ethical Standards: Adhere to GDPR by utilizing only anonymized data and providing clear opt-out mechanisms.

CONCLUSION

This project successfully transformed the Bank Marketing dataset from a telemarketing conversion challenge into a comprehensive customer engagement and churn prevention framework. By leveraging an optimized Random Forest classifier, we achieved a high-performance benchmark of 89.1% model accuracy. More importantly for the bank's bottom line, the model reached 49.5% precision when targeting the top 10% highest-probability customers –surpassing the baseline 11% subscription rate by nearly five-fold. This shift from "volume-based" to a "value-based" strategy directly addresses the operational inefficiencies identified at the project's outset.

Key discoveries from our visual and predictive analytics revealed that low call duration (under 200 seconds) among middle-aged clients (ages 30-50) constitutes the primary "risk zone" for disengagement. This "soft churn" is further amplified by unfavorable economic conditions, specifically high 'euribor3m' interest rates, which serve as a critical external trigger for customer withdrawal. Furthermore, our analysis established a clear "fatigue threshold," showing that subscription rates plummet after just three campaign contacts.

By shifting from inefficient mass calling to targeted outreach that prioritizes high-propensity profiles during optimal outreach months, the bank can dramatically improve ROI. This strategy is not only more profitable but more sustainable; it respects ethical boundaries through transparent modeling, robust consent protocols, and strict contact limits that prevent customer fatigue. The integrated data wrangling pipeline, the multi-panel visualization dashboard, and the predictive analytics framework provide a scalable, end-to-end plan for data-driven customer experience, ensuring that the bank remains competitive in a volatile economic landscape.

Appendix

Part A: Discovery Questions (Investigation Questions)

1. Which job categories and education levels show the highest engagement?
2. How does customer age interact with material status regarding disengagement?
3. What call duration threshold reliably separates “risk” from engagement?
4. After how many campaign contacts does ROI turn negative?
5. Do previously contacted customers convert at significantly higher rates?
6. How do macroeconomic indicators like euribor3m impact engagement?
7. Which economic conditions amplify churn risk across different segments?
8. What precision is achieved by targeting only the top 10% of customers?
9. How does feature importance ranking align with business intuition?
10. During which months should the bank concentrate telemarketing efforts?

Part B: Stakeholder Q&A (Presentation Deliverable)

1. Was the objective of this project achieved? Answer: Yes. We successfully reframed the original subscription prediction task into a comprehensive propensity-to-engage framework. By achieving 89.1% accuracy, we proved that machine learning can identify the specific behavioral and economic markers that lead to customer disengagement.
2. Are the results strong enough for practical implementation? Answer: Absolutely, but with a “human-in-the-loop” approach. Our model provides a 49.5% precision rate when targeting the top decile of customers. We recommend using this as a decision-support tool to prioritize daily call lists.
3. Why was “Call Duration” excluded from the final predictive model? Answer: This was a critical step to ensure model integrity. Call duration is only known after a conversation ends. Including it would create target leakage, giving us high accuracy on paper that would fail in the real world.
4. Does customer engagement depend more on demographics or behavior? Answer: Our findings indicate that behavior and macroeconomics are the dominant drivers. While age and job provide some context,

the “pdays” (previous contact history) and current euribor3m (interest rates) were much stronger predictors of the “soft churn” phase.

5. Which model performed best, and why? Answer: The Random Forest Classifier was the superior performer. Unlike linear models, it was able to capture the complex, non-linear “Risk-Zone” we identified –specifically how middle-aged demographics interact with specific economic thresholds.
6. How exactly does this model reduce our marketing operation costs? Answer: Currently, the bank suffers from an 88% “No” rate. By targeting the top 10% propensity leads, you can capture nearly 50% of all potential “Yes” responses while eliminating 90% of the unsuccessful calls, saving hundreds of man-hours.
7. What is the “Risk Zone” identified in your visualization, and how do we use it? Answer: The “Risk Zone” is a concentration of disengagement among customers aged 30 -50 whose calls last under 200 seconds.
8. How do we handle significant changes in economic conditions? Answer: Because our model identifies interest rates (euribor3m) as a top predictor, we recommended a quarterly retaining schedule. If interest rates shift significantly, the model should be “refreshed” with the latest 90 days of data.
9. How did we address the fact that 88% of the data is “No” responses? Answer: We used SMOTE (Synthetic Minority Over-sampling Technique). This allowed the model to “learn” the rare characteristics of the 11% who say “Yes” without being overwhelmed by the majority who say “No”.
10. What ethical safeguards are in place for this targeted outreach? Answer: We prioritize transparency and consent. The model does not use sensitive personal data that could lead to discriminatory lending. Furthermore, by implementing a 3-contact limit, we prevent customer harassment.

References

Akins, A. (2024, March 12). Using data and analytics to improve marketing effectiveness. ABA Banking Journal. https://bankingjournal.aba.com/2024/03/using-data-and-analytics-to-improve-marketing-effectiveness/

Geeksforgeeks. (2025, October 29). Evaluation Metrics in Machine Learning. https://www.geeksforgeeks.org/machine-learning/metrics-for-machine-learning-model/

McKinney, W. (2022). Python for Data Analysis: Data Wrangling with pandas, NumPy, and Jupyter (3rd ed.). O'Reilly.

Moro, S., Rita, P., & Cortez, P. (2014). Bank Marketing [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5K306