Using Machine Learning to Solve Real Banking Challenges at Banco Primus

Student Name	Program	Individual Title
Nils Rudolf (63855)	Business Analytics	Optimizing Campaigns via Propensity Based Frameworks

Work project carried out under the supervision of:

Advisor: Michail Batikas

Abstract

This thesis examines how machine-learning models and explainable AI can be used to analyze two distinct use cases: loan conversion and cross-selling in retail banking. Using proprietary data from Banco Primus, logistic regression, random forest, and XGBoost models are evaluated using business-oriented back-testing. SHAP is applied to explain predictions and identify key drivers. Approved loan conversion is mainly associated with partner characteristics, process timing, and communication availability. Personal loan cross-selling is mainly associated with external credit profiles and behavioral history, revealing campaign fatigue. The findings support process optimization in loan origination and propensity-based targeting frameworks for cross-selling.

Keywords: Business Analytics, Machine Learning, Marketing, Retention

This work was funded by Fundação para a Ciência e a Tecnologia (UID/00124/2025, UID/PRR/124/2025, Nova School of Business and Economics) and LISBOA2030 (DataLab2030 - LISBOA2030-FEDER-01314200)

Table of Contents

1 Introduction

The banking industry is undergoing a significant transformation caused by macroeconomic challenges, intensified competition, and shifting customer expectations (KPMG International 2025). To respond, banks are integrating automation and AI to enhance operational efficiency and reduce costs. According to KPMG International (2025), 53% of the banking institutions expect to cut at least 10% of the costs, while only a quarter of the institutions have achieved their goals so far. This emphasizes the urgency and difficulty in transforming established banking processes to ensure cost-efficiency and sustainable revenue growth through strategic investments in technology. Machine learning techniques and explainable AI methods have gained increasing attention in credit scoring research (Bücker et al. 2020), a trend further supported by supervisory evidence documenting growing AI adoption in use cases, such as fraud detection (European Central Bank 2025a). By contrast, the maturity and standardization of AI applications in loan conversion and cross-selling are less clearly documented in regulatory sources. When machine learning is applied in marketing or lending contexts, model interpretability is often neglected, which limits transparency, stakeholder trust, and regulatory acceptance in the financial sector.

Explainable AI methods, such as SHAP, address these challenges by making complex models interpretable and thus more actionable. However, few empirical studies have combined predictive modeling and explainability to bridge the gap between model output and operational decision-making in banking, particularly in the context of loan conversion and loan cross-sell. Existing research often focuses on prediction accuracy or theoretical explainability but lacks a framework that translates these insights into operational insights. Therefore, the insufficient integration of XAI to identify operational drivers in loan conversion and to identify customer profiles for targeted marketing in cross-selling forms a central research gap.

This study aims to address this gap by applying machine learning and explainable AI to examine the business value generated at two critical stages of the customer journey at Banco Primus, a financial institution in Portugal and a subsidiary of Groupe BPCE. The analysis focuses on i) auto loan conversion: forecasting the conversion of approved auto loans during customer acquisition; ii) personal loan cross-sell: predicting a customer's propensity to start a loan simulation. Therefore, this study addresses the following research question: How can machine learning and explainable AI be used to identify the drivers of post-approval auto loan conversion and personal loan cross-selling at Banco Primus to support operating decisions?

The framework compares the performance of logistic regression, random forest, and XGBoost models, and applies SHAP for global and local feature attribution. By linking predictive performance, interpretability, and measurable impact, the study provides a practical blueprint for applying transparent AI to improve customer journey outcomes in the banking sector.

The remainder of this paper is structured as follows. Chapter 2 describes the two use cases, namely auto loan conversion and personal loan cross-sell, in detail. In chapter 3, the relevant literature is reviewed, focusing on customer journey optimization, machine learning models used with financial data, and explainable AI in finance. Further, chapter 4 outlines the methodological approach, looking at the data acquisition, cleaning, engineering, modeling, evaluation, and explainable AI methods. Next, in chapter 5, the model results are presented and put into context in the discussion and business insights in chapter 6. Finally, the conclusions are summarized in chapter 7.

2 Problem Definition

Banco Primus actively seeks to leverage data-driven decision-making. Banco Primus' loan portfolio consists primarily of car loans, with a minor share of its business stemming from personal

loans and motorcycle loans. Banco de Portugal (2025) reports that new auto loan business grew by 15% year-on-year in 2024, while personal loans increased by 6.6%. Internal numbers from Banco Primus confirm this trend, showing a constant growth in both businesses from 2021 to 2024. Specialized financial institutions and intermediaries grant over 80% of the auto loans in Portugal, while more than 60% of personal loans are still controlled by universal and retail banks. The majority of car loans (+50%) are issued to individuals, and the remaining to SMEs/corporate customers.

The first use case focuses on attracting new prospects and, hence, optimizing customer acquisition. At Banco Primus, car loans form the core business. The loans are distributed mostly via car dealers (+50%). Regarding the conversion rates of approved loans, it varies across distribution channels, as car dealers demonstrate a higher conversion rate compared to brokers. The transformation of a loan request into a booked loan involves a series of steps and several participants. It begins when a potential customer submits an application for financing through a car dealer or broker. Then, the respective intermediary conducts a simulation and discusses terms with the customer. If the client would like to proceed, the intermediary collects the required data and sends the completed proposal to the bank. Next, Banco Primus checks the proposal and, if approved, notifies the customer and intermediary of the first decision by email. Then, the required documents are shared with the bank and reviewed. After the verification of the submitted documents, a bank operator conducts a welcome call and transfers the funds so that the customer can pick up the vehicle from the intermediary. At every point in this process, friction or delay can cause customers to disengage or increase the risk that customers receive a more attractive and quicker accessible offer from another institution. Currently, the bank lacks understanding of which factors lead customers to drop out after the first approval and what causes the performance gaps between the two distribution

channels, dealer and broker. Therefore, the goal is to find patterns among car dealers, brokers, customers, processes, or financial terms that convert more and to build a model that predicts the conversion probability for each first-approved loan. By identifying the key drivers across both channels, Banco Primus will be able to implement recommendations to improve conversion rates, leading to revenue maximization in the long run.

The second use case focuses on developing existing relationships and, hence, optimizing customer development. Every quarter, Banco Primus conducts a campaign offering of pre-approved personal loans to roughly 40% of the active customer base. A rule-based system decides which clients are eligible for the campaign and pre-approves them. Marketing activities follow a quarterly cycle, with eligibility updated every three months. The bank contacts eligible clients through a mix of SMS and email. Since September 2025 (after the dataset was provided), the contact frequency increased to four touchpoints per month. This schedule follows a weekly alternating sequence of SMS and email. While the message content is standardized, the client's name and pre-approved amount are personalized. Then, each selected client receives an email or SMS with a link to a personalized landing page, which contains the pre-calculated amount and loan conditions. If interested, the client submits a simulation through the page. The proposal needs to be double-checked to determine if any conditions changed since pre-approving the loan. Upon acceptance, the customer is notified and shares the required documents with the bank. Finally, a bank operator reviews the documents, performs a welcome call, initiates the electronic signature process, and transfers the funds into the customer’s account.

However, the campaign's current engagement rate, measured by the simulation rate, is low, below the market benchmark. Therefore, the goal of this use case is to identify patterns of customers who engage with the offer. Based on these patterns, the model will estimate each eligible customer’s

propensity to start a simulation. This will allow the marketing team to personalize outreach and focus resources on high-potential segments. This targeted approach is expected to increase customer value and loyalty by creating more integrated relationships with the bank. Further, it could lead to a more diversified bank portfolio, mitigating the influence of economic fluctuations.

3 Literature Review

This chapter establishes the theoretical groundwork for the analysis by summarizing key concepts from customer journey management, machine-learning modeling, and explainable AI, which together inform the model design and interpretation for the Banco Primus use cases.

3.1 Customer Journey Optimization in Retail Banking

Customer journey is a research domain in marketing and customer experience, and refers to the whole set of experiences a customer has with a company over time (Lemon and Verhoef 2016). The literature typically divides the process into pre-purchase, purchase, and post-purchase phases. This framework is applied in marketing research to disaggregate processes into analyzable phases (Lemon and Verhoef 2016). For example, customer acquisition analyses often focus on optimizing frictions within the purchase stage. In contrast, customer development strategies, such as cross-selling, are interventions that target the pre-purchase stage of a new journey.

To analyze the customer journey, it can be divided into distinct touchpoints, which are individual interactions between a company and a customer (Kranzbühler et al. 2018). The origin varies between being brand-owned (e.g., websites, branches), partner-owned (e.g., dealer or broker communication), customer-owned (e.g., choice of payment, product usage), or social/external (e.g., word-of-mouth, comparative research) (Lemon and Verhoef 2016). Thus, organizations must

coordinate experiences across environments where their level of control varies. Quantifying channel performance with metrics such as conversion rates, engagement levels, and efficiency helps them to make informed decisions (Verhoef et al. 2015).

One of the core objectives of journey analysis is pain point detection, as these can lead to customer drop-off (Kranzbühler et al. 2018; Lemon and Verhoef 2016). However, pain point identification is often challenging. Traditional survey-based approaches capture only stated frustrations, often missing underlying issues that influence behavior. A more robust approach combines customer surveys with operational data to reveal bottlenecks in performance across end-to-end customer journeys (Rawson et al. 2013). Applied to the two use cases, process frictions, such as a lack of follow-up communication or delays in document verification, could explain the disparity in loan conversion rates between channels. Similarly, the low simulation rate in the personal loan cross-sell campaign indicates a pain point in the pre-purchase stage of the process. Literature on customer journey optimization justifies the use of machine learning to model these elements. For instance, touchpoint data and channel indicators serve as features in predictive models to forecast outcomes and address frictions (Cowan et al. 2023; Davenport et al. 2020).

3.1.1 Customer Acquisition Optimization in Banking: Auto Loan Conversion

Customer acquisition optimization refers to systematically increasing conversion rates from prospects to customers while managing resource allocation to maximize profitability (Reinartz et al. 2005). Within the auto loan use case, one customer acquisition optimization problem is to increase the conversion of approved loans by addressing post-approval customer drop-out.

Acquisition optimization is driven by several economic and strategic factors. First, the cost of customer acquisition is a relevant influencer, as research suggests that acquiring new customers is

often more expensive than retaining existing ones (Min et al. 2016). This results in balancing acquisition and retention spending to maximize customer equity, the sum of all customer lifetime values (Blattberg and Deighton 1996). In saturated financial services markets, increasing competition leads to rising customer acquisition costs. Thus, customer acquisition cost optimization is essential for firm growth, particularly for non-leading firms (Min et al. 2016).

Second, it enhances operational efficiency. The loan origination process creates costs in sales, underwriting, and administration. When an approved applicant fails to convert, the resources invested in their acquisition become a sunk cost, representing a process inefficiency that limits profitability. Internal workflow delays, communication issues, and variations in staff performance can contribute to these inefficiencies, as shown by Brahma et al. (2021). This demonstrates that operational frictions hinder customer progression in loan processes and that effective performance monitoring can reduce operational risk and improve the quality of lending outcomes (Wei-Shong and Albert Kuo-Chung 2006). Predictive modeling helps reduce these losses by directing resources towards proposals with a higher probability of closing (Buddiga 2022).

Third, it improves the customer experience. A friction-intensive process at a key touchpoint, such as approval and financing, could represent a pain point for customers, causing frustration. This frustration can damage the overall brand perception and undermine the potential for a long-term, loyal customer relationship (Lemon and Verhoef 2016). Research shows that friction in the journey slows down progress and leads to abandonment (Padigar et al. 2025). Further, industry analysis supports that timely approval decisions are a central driver of competitive performance in retail lending and that long approval times and fragmented processes reduce customer satisfaction and conversion (Deloitte 2025).

Acquisition optimization relies on data-driven methods to address inefficiencies in lending. Research on loan origination indicates that approval decisions are one stage in a multi-step journey, where customer behavior is influenced by both hard information (quantifiable data) and soft information (subjective assessments and contextual factors) (Agarwal and Ben-David 2018). By analyzing historical data on customer characteristics, proposal features, and channel information, predictive models can identify the drivers of conversion. However, it is important to ensure that predictive models rely only on information available at the decision point in the process (Maggi et al. 2014). This requirement is reinforced by broader machine learning research, showing that even small forms of data leakage can significantly inflate performance metrics and undermine the reliability of predictive models (Bouke et al. 2024). Further, research shows that accurately predicting loan acceptance enables the bank to effectively prioritize clients with a high propensity to accept (Buddiga 2022; Kumar 2018; Venkatesan 2004). This can reduce manual effort spent on processing and following up with applicants who are unlikely to proceed with the offer (Buddiga 2022). However, predictive insights also allow for targeted interventions.

Channel-specific insights can inform specific interventions. Intermediaries can influence borrower behavior and loan outcomes. For example, in mortgage lending, borrowers using brokers differ systematically from direct-lender clients in leverage, amortization, and product risk, reflecting both incentive misalignment and potential channel-specific behavioral frictions (Allen et al. 2025). Brokers primarily act as information and document intermediaries, guiding clients through application steps and comparing lender offers, but they have no direct relationship to the underlying transaction and have no direct role in closing a sale (Ionescu 2014). By contrast, auto-dealer financing is embedded in a high-engagement retail sales environment, where the salesperson is financially motivated to complete the vehicle transaction, and financing options facilitate

immediate purchase completion (Low et al. 2020). If abandonment in the broker channel concentrates at certain stages, a process redesign can address those issues. Managing channels in an integrated manner aims to create a consistent experience, which can improve overall customer experience and performance (Lemon and Verhoef 2016; Verhoef et al. 2015). Studies on service delivery support this interpretation, showing that heterogeneity in frontline or partner interactions can produce uneven customer satisfaction and behavioral outcomes (Rod et al. 2016). Furthermore, Tran and Corner's (2016) research on communication channels shows that customers view face-to-face and interpersonal communication with bank staff as more credible and persuasive than information received through digital or mass communication channels. This suggests that the mode of communication can meaningfully shape customer decisions in financial service contexts.

Further, offering personalization becomes feasible. Understanding the reasons why clients reject the offered loans across the different channels gives the bank the possibility to create personalized offers and, thereby, increase the acceptance probability. However, personalization strategies should be empirically validated, rather than assumed effective. Gubela et al. (2019) show, using A/B tests, that targeted or personalized interventions only create value when customer behavior actually changes in response to them, highlighting the need to test whether model-based strategies improve outcomes in practice. Further, Thomas et al. (2006) argue that the importance of real-time prediction of loan acceptance has increased since loan offers have become more customizable, as it enables banks to instantly adjust interest rates, loan terms, or other offer conditions. This allows them to tailor each offer to the customer's profile and, thereby, maximize both conversion probability and profitability.

3.1.2 Customer Development Optimization in Banking: Cross-Selling

Customer development refers to the strategies and processes aimed at increasing the value of a company's existing customer base (Blattberg et al. 2008). The authors further state that a firm's levers for increasing customer value include customer development through cross- and upselling, and retention efforts. Boustani et al. (2024) define cross-selling as selling more products to existing customers, contrasting it with the more expensive process of acquiring new customers.

Cross-selling serves three connected strategic objectives. First, it increases customer lifetime value, defined as the present value of all future profits obtained from a customer throughout their relationship with a firm (Gupta et al. 2006). By expanding product portfolios, banks create multiple revenue streams from single customers, increasing per-customer contribution and relationship duration (Reinartz and Kumar 2003). This metric provides a quantitative basis for strategic resource allocation, guiding marketing investments toward customers with the highest projected returns (Venkatesan 2004). Second, cross-selling enhances share-of-wallet (SOW), the proportion of a customer's financial services spending captured by one institution (Cooil et al. 2007). In markets where customers maintain relationships with multiple providers, expanding SOW is an attainable growth path. Third, multi-product relationships create switching barriers. Customers holding multiple products with one bank face practical costs and hassle to move their business, which increases retention probability (Boustani et al. 2024).

Propensity modeling is an analytical method for predicting customer engagement. In the context of cross-selling, the model assigns each customer a numerical "propensity score" that quantifies this behavioral probability (Boustani et al. 2024; Moro et al. 2014). These models estimate the probability that specific customers will respond positively to particular offers at given times (Moro et al. 2014). The model is constructed using historical data (e.g., previous offer acceptances or

declines), combined with demographics, account characteristics, transaction patterns, and behavioral signals (Boustani et al. 2024). Research indicates that behavioral data can be a stronger predictor of future value and propensity than static demographic data (Boustani et al. 2024; Gupta et al. 2006; Awanife Oghenemaro 2025). Machine learning algorithms identify patterns distinguishing responders from non-responders, creating scoring functions applicable to new campaigns. By ranking customers on predicted response probability, banks can move from one-size-fits-all campaigns to targeted interventions (Boustani et al. 2024; Moro et al. 2014). For instance, a study on bank telemarketing found that a bank could contact half (50%) of the clients to reach approximately 79% of the eventual subscribers, which improved campaign efficiency (Moro et al. 2014). Recent studies using deep learning confirm that these models can improve the predictive accuracy of cross-selling campaigns (Boustani et al. 2024). However, propensity modeling also has several limitations. First, models trained on historical acceptance data may perpetuate existing biases (Davenport et al. 2020). Second, propensity scores predict short-term acceptance, but not the customer's long-term profitability or value (Gupta et al. 2006; Reinartz et al. 2005). A customer with a high propensity to accept a credit card offer may be financially vulnerable, making the sale profitable short-term but potentially damaging long-term (Agarwal and Ben-David 2018; Majeske and Lauer 2013). Third, statistical optimization alone may overlook contextual considerations (Méndez-Suárez and Crespo-Tejero 2021). Studies on customer lifetime value optimization indicate that initially unprofitable customers may become valuable over time, suggesting that cross-selling should be balanced with relationship development goals (Méndez-Suárez and Crespo-Tejero 2021). Fourth, isolated propensity models are campaign-specific. They may predict acceptance for one offer but cannot arbitrate between competing actions, e.g., a

personal loan versus a motorcycle loan, to determine which action would best optimize a customer's long-term value (Cao and Zhu 2022).

To translate propensity scores into effective campaign strategies, literature emphasizes the need for normative decision frameworks regarding resource allocation, contact frequency, and channel selection. Recent research suggests that propensity models should not function in isolation, but within a broader, evidence-driven architecture. Abdullah and Hasan (2023) propose integrating propensity scores with customer lifetime value and churn models to guide segmentation and treatment allocation, ensuring that high-probability targets are also high-value customers. Similarly, Gubela et al. (2024) demonstrate that ranking customers based on estimated treatment effects allows firms to allocate budgets dynamically, targeting recipients only until the budget is exhausted or the marginal value creation diminishes. Neslin et al. (2006) support this approach by arguing that maximizing campaign profitability requires linking a predictive model's accuracy directly to the economic costs of incentives and contact. They show that grouping customers into deciles based on propensity scores allows managers to identify specific cut-off points where the incremental profit from retaining a customer exceeds the cost of intervention. Besides customer selection, determining the optimal contact volume is important to sustain long-term customer value. Reinartz and Kumar (2003) identify an inverse U-shaped relationship regarding purchase intervals, suggesting that solicitation plans must align with customer temporal dynamics rather than maximizing frequency. Additionally, Ansari and Mela (2003) warn that excessive contact leads to unsubscribing and diminishing returns due to information overload. Radcliffe and Surry (2011) add that marketing can sometimes cause negative effects, where contact actually reduces sales or causes customers to leave. They argue that companies must identify these specific customers and avoid targeting them to prevent financial loss. Finally, the effectiveness of an offer depends on the

alignment between the message and the communication channel. Verhoef et al. (2015) note that customers use channels interchangeably, requiring firms to understand how specific touchpoints impact performance and integrate them to provide a seamless experience.

Building on this logic, recent approaches move beyond single-campaign optimization towards centralized decisioning engines, often termed Next Best Action (NBA) or Marketing Recommender Systems (Cao and Zhu 2022; Deng et al. 2020; Weinzierl et al. 2020). These systems operate by learning the sequential interactions between customers and decision-makers across possible actions (e.g., insurance products, retention initiatives) and integrating customer-level data, channel context, and organizational policies. The system then arbitrates between these competing offers to recommend an action predicted to maximize a target objective, such as customer lifetime value (Cao and Zhu 2022; Weinzierl et al. 2020).

These points suggest that effective cross-selling strategies combine propensity scores with customer value assessment and relationship health monitoring. The objective shifts from maximizing short-term conversion to building sustainable, mutually beneficial relationships. This framing connects cross-selling optimization to broader relationship marketing theory.

3.2 Machine Learning Models for Finance Data Analysis

Machine learning algorithms have been at the forefront of financial analytics over the past few years, supporting activities such as credit risk assessment, fraud detection, marketing personalization, and portfolio optimization (Khandani et al. 2010; Pavón Pérez 2025; Shiwa et al. 2021; Doria et al. 2023; Nasir et al. 2024; Hu 2025). To support data-driven marketing and sales decisions, such as prioritizing credit applications or selecting cross-selling targets, the literature shows a progression from statistical to machine-learning-based predictive models.

Logistic regression often serves as a benchmark method, particularly in the regulated banking sector, due to its transparency and the interpretability of its coefficients (Yang et al. 2015). Consequently, it is frequently used for validating new classification algorithms (Lessmann et al. 2015). However, logistic regression has limitations because it assumes linear relationships and is less effective at modeling complex, non-linear patterns in customer behavior (James et al. 2023).

To improve predictive quality with complex customer data, recent research has focused on ensemble methods, specifically bagging algorithms, such as random forests, and gradient boosting techniques, including XGBoost (Hu 2025; Wang 2025; Feng et al. 2025). Random forests have been applied to financial and marketing tasks, like churn prediction, demonstrating robustness even with a high number of features (Hu 2025; Wang 2025). Gradient boosting, especially XGBoost, proposed by Chen and Guestrin (2016), is often considered a state-of-the-art method for tabular data, such as that used in finance. Studies report that XGBoost outperforms traditional models in predicting customer behavior or financial risks (Feng et al. 2025), though its computational cost remains high (Chen and Guestrin 2016). Nowadays, the field of financial data modeling continues to evolve, with emerging approaches, such as neural networks (Špicas et al. 2023; White 1988) and support vector machines (SVMs) (Huang et al. 2024; Kim 2003), being applied increasingly to complex prediction tasks.

3.2.1 Comparative Analysis of Machine Learning in Loan Offer Acceptances

3.2.1.1 Behavioral Biases, Determinants, and Features

Wonder et al. (2008) analyzed in a choice-based conjoint experiment how consumers choose among auto-loan offers with varying attributes. Their research shows that consumers do not necessarily act as financial optimizers. Instead, debt aversion, the general dislike of owing money,

can lead to consumers choosing loans with higher down payments, even when these are more expensive. Mental accounting further explains irrational spending behavior, as people categorize money into separate accounts in their minds (Thaler 1985). So, they may treat funds allocated to a car purchase differently than other savings or income and, consequently, make financing decisions that deviate from rational cost minimization. Additionally, consumers are affected by extremeness aversion, meaning they tend to avoid the extremes of an option set, preferring the middle values instead (Simonson and Tversky 1992). Finally, Wonder et al. (2008) show that consumers focus on the monthly installment and prefer values just below multiples of a hundred. For this project, the examples imply that conversion outcomes are shaped not only by financial terms, such as interest rate or down payment, but also by psychological factors, which influence the perceived attractiveness of a loan offer. This needs to be considered in the analysis, especially since customer interactions and guidance deviate across the distribution channels, leading to possibly different interpretations of the same loan offers.

Building on these psychological mechanisms, Argyle et al. (2020) empirically demonstrate that lenders are more sensitive to changes in maturity than to equivalent changes in the interest rate, emphasizing that these variables should be included as predictors in conversion models. Further, Wonder et al. (2008) identify contract length and interest rate as the key determinants of loan choice, with down payment and rebate exerting a smaller influence. Lyu et al. (2025) and Huang et al. (2024) use different machine learning algorithms to predict personal loan applications. They name educational factors, income, family size, the existence of a deposit account, and credit history as key determinants in the models, which confirms that both demographic and financial factors can influence the prediction of loan acceptance. Additionally, Lyu et al. (2025) confirm that precision marketing and customer-centric loan product design can improve conversion rates, as well as create

a customer-first brand image, directly contributing to sustainable growth. A closely related case to the conversion-probability prediction at Banco Primus was analyzed by Špicas et al. (2023), who model loan acceptance on an online loan comparison and brokerage (OLCB) platform for consumer credit. They include engagement metrics and identify them as meaningful factors, for example, the more time spent completing the application process, the lower the credit risk. This is also supported by McKinsey & Company (2025). They claim that seamless and personalized online experiences are becoming the baseline in the consumer finance sector.

Taken together, these studies indicate that loan conversion rates are shaped by multiple dimensions, including demographics, loan terms, engagement, channel context, and behavioral biases. Many papers confirmed that traditional financial variables remain key determinants, such as loan term, interest rates, and total cost, but the other dimensions should not be neglected, as they provide a more holistic and realistic perspective.

3.2.1.2 Modeling Approaches

Having established the diverse feature dimensions driving loan acceptance, it is important to explore which modeling techniques and algorithms were used in previous research and which performed most reliably. Thomas et al. (2006) developed one of the earliest models on loan offer acceptance by constructing their simulated data set. The study compares three approaches to forecast loan offer acceptance. First, they used a baseline logistic regression, then a linear programming model, and finally an accelerated life model. Out of the three, linear programming performed best, showing that optimization approaches can be relevant when offer terms are linked to acceptance behavior, creating a foundation for following machine learning approaches.

Building on the earlier discussion of loan acceptance features, Špicas et al. (2023) also provide methodological learnings by comparing logistic regression, random forest, XGBoost, artificial neural networks (ANN), and support vector machines. The authors demonstrate that XGBoost outperformed all other models in AUC-ROC and accuracy, as shown in Table 1.

The performance of XGBoost did not improve significantly when combined with resampling techniques, such as oversampling, undersampling, or SMOTE, implying that data balancing measures should be tested and not applied by default. The results suggest that logistic regression can be implemented as a credible baseline, but that random forest and XGBoost usually capture complex relationships in loan acceptance data better (Špicas et al. 2023).

Akça and Sevli (2022), previously mentioned for their feature analysis on the Thera Bank dataset, focused mainly on the application of support vector machines to predict the acceptance of personal loan offers using the Thera Bank dataset, creating one of the best-performing algorithms in terms of accuracy. Additionally, Huang et al. (2024) compared the performance of logistic regression and support vector machines to the multilayer perceptron, and the gradient boosting decision trees, CatBoost, and XGBoost, on the same dataset. To account for the class imbalance, they applied SMOTE with three different sampling rates and showed that XGBoost and CatBoost perform best with respect to accuracy (Akça and Sevli 2022).

Moreover, Lyu et al. (2025) applied logistic regression, decision trees, random forests, and k-nearest neighbors on a dataset comprising 5000 customers from Neo Bank. To evaluate and compare the models, they used precision, recall, accuracy, and F1 scores (Table 1).

The authors pre-pruned the decision tree to prevent overfitting by pausing the node splits through heuristic stopping criteria, which outperformed all other models in accuracy, recall, and F1 score. Only in precision, the random forest achieved the highest precision, highlighting that sometimes

simpler models might be sufficient to capture the relationships present in the data and should also be used in the model comparison (Lyu et al. 2025).

Topic	**Acceptance Probabilities of Consumer Loan Offers**					**Comparative Analysis of Machine Learning Models for Targeting Bank Loan Customers**				
Author	**Špicas et al. (2023)**					**Lyu et al. (2025)**				
Algorithm	LR	RF	XGB	ANN	SVM	Preprunig DT	Postpruning DT	LR	RF	KNN
Accuracy	0.748	0.757	0.763	0.75	0.762	0.986	0.963	0.907	0.981	0.903
Precision	-	-	-	-	-	0.927	0.773	0.541	0.984	0.520
Recall	0.866	0.849	0.874	0.847	0.896	0.933	0.893	0.396	0.826	0.342
F1 score	-	-	-	-	-	0.930	0.829	0.457	0.898	0.413
AUC	0.802	0.825	0.834	0.802	0.816	-	-	-	-	-

Table 1: Predicting Approved Loan Conversion - Comparison of Model Performance

Table 1 emphasizes that ensemble methods, such as random forest and XGBoost, outperform approaches like logistic regression or ANN in terms of accuracy and AUC-ROC (Lyu et al. 2025; Špicas et al. 2023). This confirms their ability to capture nonlinear relationships in loan acceptance data. At the same time, the competitive performance and interpretability of simpler approaches, such as logistic regression and pre-pruned decision trees, underline the value of maintaining simpler, explainable baselines.

To summarize, even though there is comparatively little literature on the prediction of loan conversion, a progression from early statistical models and simple algorithms (Thomas et al. 2006) to complex machine learning models can be seen. Throughout the studies, logistic regression proves to be a valid and interpretable baseline (Lyu et al. 2025; Špicas et al. 2023; Huang et al. 2024), which can provide reasonable benchmarks. Bagging and boosting algorithms, such as random forest and XGBoost, usually outperform it by introducing non-linearity and interactions between the features (Špicas et al. 2023; Yang 2025). More complex models, such as ANNs, have also been applied, but do not necessarily predict loan approvals more accurately (Špicas et al.

2023). Hence, to support Banco Primus with the prediction of loan acceptance, this paper focuses on the comparison of logistic regression, random forest, and XGBoost.

3.2.2 Comparative Analysis of Machine Learning in Cross-Selling Marketing

3.2.2.1 Behavioral Biases, Determinants, and Features

Analyzing the determinants of cross-selling success requires distinguishing between customer-specific interactions and broader external factors. Zheng (2025) utilizes data from a telephone marketing campaign to identify key drivers for long-term deposit uptake. The study indicates that interaction metrics, specifically the duration of the sales call and the frequency of prior contacts, correlate positively with acceptance rates. This suggests that the intensity of the customer-bank interaction serves as a proxy for engagement. Complementing this, Boustani et al. (2024) examine feature importance in the context of consumer loan cross-selling. While variable definitions differ across datasets, the authors identify engagement features, such as the history of positive interactions and the depth of the existing relationship, as central to prediction models. Both studies show that static demographic data alone is insufficient. Instead, dynamic behavioral data reflecting the customer's ongoing relationship provides better predictive value. These engagement metrics are closely linked to psychological determinants, specifically the distinction between satisfaction and trust. Lin (2012) posits that repeated interactions foster familiarity, which can translate into trust. The author differentiates this from general satisfaction. While satisfaction reflects a positive evaluation of past services, trust specifically influences the willingness to engage in new financial risks or products. Consequently, Lin (2012) argues that trust is a more accurate predictor of cross-selling susceptibility than satisfaction scores alone.

Besides individual behavior, external macroeconomic conditions also shape cross-selling opportunities. Basten and Juelsrud (2024) demonstrate that monetary policy, such as interest rate adjustments, influences the broader market environment for deposit products. Their findings suggest that macroeconomic indicators act as external determinants that impact the customer's probability to convert, independent of their individual relationship with the bank.

Overall, the literature identifies determinants for cross-selling prediction. While demographic data provides a baseline, behavioral features related to interaction depth and history constitute the strongest predictors (Zheng 2025; Boustani et al. 2024). These are moderated by psychological factors like trust (Lin 2012) and external macroeconomic contexts (Basten and Juelsrud 2024). Consequently, accurate prediction models require a multidimensional feature set that supplements static customer attributes with dynamic indicators.

3.2.2.2 Modeling Approaches

Effective modeling requires a preliminary understanding of the data structure. Zheng (2025) highlights the necessity of EDA to reveal underlying behavioral patterns and distributions before applying machine learning algorithms. This step is particularly relevant given the nature of the datasets used in cross-selling prediction.

Recent empirical studies (2024-2025) regarding cross-selling prediction mainly compare logistic regression, random forest, and XGBoost. These studies utilize comparable datasets to benchmark performance (Table 2). Most research relies on feature sets limited to approximately 20 variables. The reviewed literature lacks the integration of broader data categories, such as detailed behavioral, demographic, and financial history, or probability scores for simulation outcomes.

A challenge identified across these studies is the high class imbalance, where only a small fraction of customers respond positively to offers (Boustani et al. 2024; Zheng 2025). For instance, in Zheng's dataset, the conversion rate is only 11.6% whereas for Boustani et al. (2024), the positive class is approximately 4.8%. Such an imbalance biases standard algorithms toward the majority class. To mitigate this, Zheng (2025) employed resampling techniques such as random undersampling, and Boustani et al. (2024) relied on model architecture and feature engineering rather than explicit resampling. In both studies, certain metrics, such as AUC-ROC or recall, are preferred over accuracy when evaluating the performance of highly imbalanced class distributions.

Topic	**Bank Target Marketing Classification on Imbalanced Data**			**A Study on Bank Marketing Prediction Based**			**Predicting Effectiveness of Bank Marketing**		
Studies	**(Nasir et al. 2024)**			**(Yang 2025)**			**(Zheng 2025)**		
Algorithm	LR	RF	XGB	LR	RF	XGB	LR	RF	XGB
Accuracy	-	0.89	0.9	0.82	0.83	0.84	-	-	-
Precision	-	0.54	0.59	-	-	-	0.59	0.64	0.63
Recall	-	0.37	0.43	0.78	0.81	0.82	0.11	0.39	0.43
F1 score	-	0.44	0.50	-	-	-	0.27	0.48	0.52

Table 2: Predicting Cross-Selling Opportunities - Comparison of Model Performance

Logistic Regression serves as the baseline model in the analyzed studies (Nasir et al. 2024; Zheng 2025). However, it exhibits lower performance metrics than ensemble methods. Zheng (2025) reports a recall of 0.11, which implies the model fails to identify 89% of customers likely to accept an offer. This limitation results from the linear nature of the algorithm, which restricts accurate modeling of high-dimensional data and non-linear customer behavior (Doria et al. 2023). Despite lower predictive power, Logistic Regression remains a benchmark for interpretability.

Tree-based ensemble methods handle complex data structures more effectively than linear models. Random forest typically yields higher recall than the baseline. Nasir et al. (2024) observe that applying grid-search optimization improves stability. XGBoost surpasses other algorithms across

most metrics in the reviewed literature. As shown in Table 2, it achieves the highest accuracy and F1 scores. The F1 score is relevant for imbalanced datasets common in conversion prediction. Chen and Guestrin (2016) attribute this performance to the capacity of the algorithm to model non-linear relationships, perform automatic feature selection, and use built-in regularization to suppress overfitting.

The literature indicates a trade-off: while ensemble models, such as XGBoost, maximize predictive performance, their interpretability is reduced (Schwartz et al. 2025). This lack of transparency presents two problems in the financial sector. First, it may conflict with regulatory requirements. Second, it prevents the derivation of strategic business implications, such as understanding the reasons for customer conversion. Despite existing research comparing machine learning models for cross-selling prediction, gaps remain. Existing studies primarily focus on predictive performance using limited feature sets of around 20. Moreover, loan lifecycle dynamics and behavioral factors such as repeated offer exposure and campaign fatigue are rarely examined in depth. This study addresses these gaps by integrating richer behavioral and customer data, combining machine learning models with SHAP-based explainability, and linking predictive drivers to customer personas and lifecycle stages.

3.3 Explainable AI in Finance

As the literature points out, advanced machine learning models, such as XGBoost or random forest, often outperform simpler models, such as logistic regression or decision trees, in standard performance metrics. Yet, advanced models are often referred to as black boxes due to their lack of interpretability (Klein and Walther 2024). This trade-off between accuracy and interpretability is illustrated in Figure 1 among different machine learning models.

3.3.1 Compliance and European Policies

The limited interpretability of advanced machine learning models not only affects transparency for internal stakeholders but also raises regulatory concerns (Černevičienė and Kabašinskas 2024). In financial services, model decisions must be explainable to comply with data protection and AI governance standards, making explainability a legal as well as a technical requirement (European Union 2016; European Data Protection Supervisor 2025). In Europe, the European Union published the General Data Protection Regulation (GDPR) to regulate the usage of personal data and the EU AI Act to regulate the usage of AI applications.

Within the GDPR, two rights are most relevant to this project. First, the right not to be subject to a decision based solely on automated processing is stated in Article 22 (European Union 2016). Second, recital 71 states the right to an explanation, including the right to retrieve specific information, express one's point of view, challenge the decision, and obtain human intervention and an explanation of the automated decision (Directorate-General for Parliamentary Research Services (European Parliament) 2020). These policies pose concrete implications for any automation algorithm used in banking, as every decision must be justifiable through reasoning and not solely a model score.

Under the EU AI Act 6(3), credit scoring is classified as a high-risk AI system because it directly determines access to financial services (European Data Protection Supervisor 2025). Therefore, using AI for credit scoring requires transparency, human oversight, and explainability (Černevičienė and Kabašinskas 2024). The proposed models for Banco Primus will not determine creditworthiness or credit scoring but predict the probability of loan offer acceptance and the probability of cross-selling success among existing customers. Therefore, both AI applications serve marketing and process-optimization purposes rather than decision-making about credit

approval. Yet, both AI applications are embedded within the customer journey of credit issuance and involve the profiling of natural persons, potentially posing an indirect influence on credit offer management. Methods such as SHAP-based feature attributions will support transparency, strengthen internal governance, and align the predictive framework with the emerging European standards for trustworthy AI (Černevičienė and Kabašinskas 2024).

3.3.2 Explainable AI Solutions

Černevičienė and Kabašinskas (2024) divide Explainable AI into subcategories, also including: i) feature relevance, which quantifies the contribution of each feature towards the output and allows one to identify the most influential factors of the predictions. Lundberg and Lee (2017) introduced SHapley Additive ExPlanations (SHAP), which has become the most widely used approach; ii) simplification, which uses simpler models to approximate the advanced model, e.g., rule-based learners or decision trees. This approach helps to communicate to non-technical stakeholders the logic behind the model and is often used to account for transparency; iii) local explanations, which explain individual predictions, which is according to GDPR in banking necessary to explain on customer level the reasoning for a decision; iv) transparent models, which refers to using simple inherently interpretable models (Černevičienė and Kabašinskas 2024).

According to Černevičienė and Kabašinskas (2024), and Klein and Walther (2024), SHAP is mostly used in finance, specifically when using XGBoost and random forest. SHAP, a single additive attribution framework, unifies six AI explainability models (LIME, DeepLIFT, Layer-wise Relevance Propagation, and classic Shapley regression values, Shapley sampling values, and Quantitative Input Influence) (Lundberg and Lee 2017). SHAP assigns a numerical contribution to each feature. This additive explainability principle is based on cooperative game theory and follows

the three properties of i) local accuracy, the sum of all features adds up to the output; ii) missingness, any irrelevant feature contribution for predicting the output equals zero; iii) consistency, the SHAP contribution values are consistent regardless of the features' weight in the model (Lundberg and Lee 2017). Klein and Walther (2024) argue that advanced machine learning models with financial data can be accepted from a governance perspective if SHAP and XAI frameworks are used.

García-Céspedes et al. (2025) assess the practical limitations of SHAP in the financial sector. They use a credit-risk ranking model and find that interventional and path-dependent SHAP result in similar global and local explanations. The authors state that the results are influenced by the background portfolio distribution, even if the customer data and the model did not change. This inconsistency in results creates a challenge for the stability and regulatory usability of customer explanations. One way to mitigate the volatility temporarily is to freeze the background dataset. While researchers are aware of this limitation, regulators may not be. García-Céspedes et al. (2025) suggest in their research that both parties should work closer together to develop a more stable explainability AI approach for financial applications.

Based on this discussion, SHAP will be used as the main explainability technique, due to its model-agnostic comparability across logistic regression, random forest, and XGBoost, and acceptability across the industry. However, to address the identified limitation, for each evaluation period, the background dataset will be frozen and documented to ensure that all explanations within that period use a consistent reference and to ensure reproducibility and auditability for every model version. To test sensitivity, explanations were compared across alternative backgrounds, including whole-portfolio and channel-specific samples. This strengthens methodological transparency and supports regulatory robustness in a dynamic financial context.

SHAP's local force plots will be applied selectively to illustrate individual customer decisions and to support communication with non-technical stakeholders (Acharya et al. 2024). SHAP will remain the main framework for all quantitative analyses and regulatory documentation to ensure methodological consistency and transparency across models.

Another method for local explanations would be LIME, introduced by Robeiro et al. (2016). LIME approximates complex models locally for a specific instance by fitting a simple interpretable surrogate model, typically a linear regression (Ribeiro et al. 2016). However, LIME has its limitations, as explanations depend on sampling strategy, kernel width, and neighborhood definition, which can lead to instability. Further, LIME lacks global insights, which can lead to inconsistency across customer reasoning. Misinterpretations can occur if users interpret surrogate coefficients as true measures of feature importance. Within regulated sectors, such as banking, thorough documentation of parameters and model configuration, as well as a complementary use of global explainers, including SHAP, are necessary to ensure stability, auditability, and consistency (Dieber and Kirrane 2020).

3.3.3 Explainable AI within Governance Models at Banks

In addition to the methodology and technical categories of XAI, Deloitte (2022) emphasize that the information level and responsibility of explainability vary among different stakeholders. The bank's governance structure should reflect the different requirements of explainability which can be distinguished in five main stakeholders: i) model developers and data scientists constitute the first line of defense and are responsible for weighing the trade-offs between model performance, accuracy, and explainability; ii) the bank's risk managers, second line of defense, test robustness and impact of biased data as input; iii) external regulators and auditors as third line of defense, who

audit the bank and its compliance with usage of AI and automated decision tools; iv) bank employees, who use visual explanations to act on model output; v) bank customers, who are entitled to the GDPR's right to explanation (Deloitte 2022).

In addition to these organizational layers and governance model outlines, research-based frameworks have also been introduced and can be used as guidelines. Bücker et al. (2020) propose a governance framework to explain credit scoring called TAX4CS (transparency, auditability, and eXplainability for credit scoring). This framework follows a cycle starting with model performance and continuing with global explanation, including variable importance and partial dependence. Then it moves to local explanations, using, for example, SHAP local force plots, and finally arrives at an instance-level what-if analysis. Each phase in the cycle addresses different stakeholders: regulators, auditors, developers, and customers (Bücker et al. 2020).

While loan default prediction is well established in the literature, considerably less attention has been given to predicting the conversion of already approved loans and personal loan cross-selling using machine learning. Existing studies typically focus on predictive performance based on limited feature sets and rarely connect model explanations to the operational and behavioral drivers underlying post-approval drop-off or campaign engagement. Moreover, while ensemble models and SHAP-based explainability are increasingly employed in financial contexts, few studies translate these insights into actionable frameworks that support operational decision-making and targeted cross-selling. This gap motivates the present study, which combines predictive modeling and explainable AI to identify the drivers of post-approval auto loan conversion and personal loan cross-selling at Banco Primus.

4 Methodology

This chapter describes the methodology used in both use cases, covering data acquisition, cleaning, exploratory analysis, feature engineering, and the development of logistic regression, random forest, and XGBoost models. It then outlines the evaluation strategy, including performance metrics, back-testing procedures, and SHAP-based explainability.

4.1 Data

4.1.1 Data Acquisition

This study uses internal data from Banco Primus, complemented by external data sources, such as historical interest rate changes, to analyze auto loan conversion and personal loan cross-selling at Banco Primus. A structured list of 116 variables covering customer, financial, product, dealer, broker, and engagement characteristics guided the extraction process. The bank provided two datasets based on this schema, excluding variables that were unavailable or irrelevant. Additionally, a public dataset from Banco de Portugal was provided by Banco Primus, containing information on car dealers and financial brokers and their collaboration with banks. This can answer questions regarding the competitive dynamics in conversion performance. All data were masked to ensure anonymity, and processing adhered to GDPR and the EU Artificial Intelligence Act for ethical and compliant use of customers', partners', and employees' information.

4.1.2 Data Description: Approved Loan Conversion

The provided dataset contains 16,808 loan application records and 122 variables, including the binary target variable *target_booked*, which indicates whether a loan application was converted into a booked contract or not (1 = booked, 0 = not booked). Each observation represents an

approved loan proposal capturing the process from creation, over first approval to potential booking or closure. The dataset includes observations recorded in 2023 and 2024.

The variables can be grouped into conceptual categories (chapter 9.4.1): i) process metadata: variables describing the workflow, its process states, and analyst interactions; ii) product and financing details: attributes specifying the credit product configuration and terms; iii) partner and channel information: characteristics of intermediaries and dealerships involved in the loan process; iv) customer characteristics: socio-demographic and employment variables; v) insurance and incentive data: information related to the vehicle and other valuation indicators; vi) decision process features: operational features describing activity metrics and risk and compliance flags; vii) target variables: denoting process outcomes.

A second dataset was incorporated from the Bank of Portugal's public registry of credit intermediaries, specifically registered brokers and dealers operating in Portugal. After the raw lender names were harmonized and cleaned to eliminate duplicates and legal suffixes, the dataset consists of 29,501 rows and 20 columns, where each row represents a unique intermediary-lender relationship within a defined validity period. The dataset includes the intermediary's legal and commercial identification details and the financial institutions with which they maintain cooperation agreements. It contains start and end dates, allowing for the detection of whether a specific lender relationship was active at the time the loan application was approved.

4.1.3 Data Description: Personal Loan Cross-Sell

The personal loan dataset consists of 16,368 customer records and 66 variables, including the target variable *state.* Each observation corresponds to a single customer who received a new loan proposal via email or SMS during the current marketing campaign cycle, starting July 2025. The target

variable is a categorical variable indicating the customer's response: no interaction, simulation only, or simulation followed by subscription to the cross-sell product. The dataset does not record the temporal sequence or timing of these events.

The variables can be grouped into five conceptual domains (chapter 9.4.2): i) customer and address data: variables that contain socio-demographic identifiers and core personal features; ii) campaign and offer data: includes historical product and pricing information about the proposed loan, timestamps, and the amount of previous offers; iii) existing auto loan data: if applicable to the customer, it describes parameters about their ongoing auto-loan agreement and information about the car; iv) customer profile and risk: consolidates indicators of the relationship between the bank and the customer and an internal risk score; v) campaign outcome: contains information about the contract details for customers who accepted the offers. *State* is the target variable.

4.2 Data Cleaning and Pre-processing

Across both use cases, column names were standardized and translated from Portuguese into consistent English identifiers if applicable. Data types were aligned with semantic meanings, including transforming currencies, percentages, and dates to numeric or datetime formats, and casting identifiers to strings. Categorical values were normalized by harmonizing spelling, removing accents, and mapping them to correct business meanings. Finally, duplicate rows were removed and outliers treated.

4.2.1 Data Cleaning and Pre-processing: Approved Loan Conversion

Missing values were addressed through different strategies depending on the degree of missingness (Acock 2005). Based on business insights provided by the bank, missing values in

has_decision_pending, *has_decision_unique*, *has_approval_after_refuse*, *client_effective*, *partner_new_bet,* and *client_self_employed* were interpreted as False. The variables *insurance_type* and *company* contained over 98% missing entries, causing any *target_booked* rate calculated for them to be statistically unreliable. Therefore, both columns were excluded from the analysis (Acock 2005). Further, *cae* was missing in approximately 71.4% of the observations, which was addressed in feature engineering (4.4.1). For categorical variables with less than 10% missing values, namely *gender*, *partner_rating*, *activity*, and *profession*, these were replaced with the label "Unknown", as they represent incomplete or unrecorded information.

Additionally, all columns that contain at least 99% zero values were dropped, after verifying that their non-zero values do not cause significant trends in booking rates and after communicating the strategy to the bank. The resulting dataset contained 109 features.

4.2.2 Data Cleaning and Pre-processing: Personal Loan Cross-Sell

The target variable *state* indicated whether a customer started a simulation ("Received"), completed it ("Active"), or took no action ("#NV"). For modeling, "Received" and "Active" were merged into a single positive class, aligning with Banco Primus' objective to increase simulations (1 = started/finished simulation, 0 = did not start a simulation). To avoid data leakage, all columns populated only when a simulation was completed were removed, reducing the feature set from 58 to 48. Missing values were addressed through a structured imputation strategy. The target variable *state*, missing in 96.8% of the observations, was imputed as class 0 (no action), as Banco Primus confirmed that missing values indicate customers who did not start a simulation. For *profession* (67.2% missing), the variable was replaced with a binary existence indicator to mitigate imputation bias. Subsequently, the original variable was removed. For features with low missingness (below

1%) and data leakage artifacts (*profession_contract_type* and *housing_type*), missing values were filled using proportional imputation. Missing values were randomly assigned based on the empirical distributions of the observed data, thereby preserving the original variable distributions and minimizing distortion of underlying patterns.

To resolve encoding inconsistencies, categorical features underwent semantic normalization assisted by a Large Language Model. Many features were domain-specific, such as Portuguese municipalities, vehicle models, or bank names, which LLMs can capture. The inspection revealed widespread encoding issues, inconsistent representations across bank names, vehicle brands, and fuel types, and a high-cardinality variable named *vehicle_model_original*. This feature was cleaned and split into a model text field and a derived *vehicle_start_year*, with missing years imputed using the median.

4.3 EDA

The exploratory data analysis consisted of both univariate and bivariate analytical procedures to understand the structure of the dataset, inspect feature distributions, and identify potential data quality issues before modeling. The distribution of the binary target variable was examined to identify potential class imbalance and determine whether resampling or weighting strategies may be required during model training. Additionally, the target variable was plotted in relation to the approval date to understand how booking behavior changes over time.

Next, the univariate analysis was performed. For continuous variables, histograms with kernel density lines, boxplots, and violin plots were created to assess their distributional shapes, detect skewness, and identify potential outliers. To reduce the impact of these outliers, winsorization is applied later at the 1st and 99th percentiles, preserving important data patterns and making the

distributions more robust for modeling (Demir and Sahin 2023). Skewed variables were inspected on a log scale to assess their suitability for improving interpretability and preparing them for logistic regression modeling (Moore et al. 2007). For discrete and ordinal features, bar charts were plotted to visualize the features' characteristics. Implausible values were identified, and the respective records were removed from further analysis. Ordinal variables were inspected for sparsity in higher categories to support later modeling decisions. Finally, for categorical features, the most frequent categories were displayed. Categories with very low occurrence were grouped into "Other" to avoid high-cardinality and improve model stability.

Next, in the bivariate analysis, all features are examined relative to the target variable. This allows for detecting strong discriminatory patterns and variables with little explanatory value. For continuous variables, bivariate plots, such as box plots, and distribution comparisons were used. For discrete, ordinal, and categorical variables, bar charts were employed to assess class-specific differences and potential data leakage.

A bivariate statistical test was conducted to assess the relationship between partner type and each feature, using proportion z-tests for binary variables, chi-square tests for categorical variables, and Mann-Whitney U tests for numerical variables, with statistical significance determined at $p < 0.05$. Mann-Whitney U tests were chosen for numerical variables as a non-parametric alternative that does not assume normal distribution, making it more robust for financial and operational metrics, which often exhibit skewness (McKnight and Najab 2010).

Finally, a correlation heatmap of numerical features was created to assess relations between the features. To avoid redundancy and multicollinearity, one variable for every correlated pair was excluded based on economic relevance. VIF scores were used to detect multicollinearity. The issue

is addressed in logistic regression via regularization, while tree-based models can handle it naturally (Tomaschek et al. 2018).

4.4 Feature Engineering

4.4.1 Feature Engineering: Approved Loan Conversion

Feature engineering was performed to compute statistically usable and reliable inputs from raw categorical data for modeling. First, deterministic, rule-based feature engineering was applied. For example, the highly missing *cae* variable was converted into a presence indicator (*has_cae*), rare categories in variables such as marital status or nationality were grouped into an "Other" label, and relationship-based features such as *partner_id_bill_match* were constructed to capture interactions between intermediaries. From the Banco de Portugal dataset, the intermediaries are matched, and a new feature indicating the number of lenders an intermediary worked with at the time of loan approval is created.

After constructing the deterministic features, an out-of-time (chronological) data split was implemented by sorting all loan applications by their approval date and allocating the earliest 70% to training, the next 15% to validation, and the most recent 15% to testing. This ensures that model evaluation reflects a realistic, forward-looking scenario and prevents temporal leakage, as future observations are never used to inform past predictions (Sheridan 2013).

For high-cardinality categorical variables, feature encoding was required to avoid creating over 1,000 dummy variables (Johnson and Khoshgoftaar 2021). Two different approaches are implemented and applied with respect to the model being trained. The first approach is Weight-of-Evidence (WoE) binning, a rigorous and credit-risk-oriented approach transforming several high-cardinality variables . WoE creates stable, monotonic relationships between categories and the log-

odds of the outcome. During fitting, the WoE transformer calculates for each category the number of booked and non-booked applications and derives the corresponding WoE values. Then, they are merged into statistically coherent bins using chi-square similarity tests and constraints on bin size and count. This approach was chosen over simple frequency encoding or other naïve approaches as it produces monotonic transformations that preserve information. In our case, it is applied to logistic regressions as it turns log-odds ratios linear, creating direct compatibility with the model's structure (Yang et al. 2015). However, for tree-based methods, regularized target mean encoding is implemented, which does not impose this structure but provides a numerical signal reflecting smoothed mean booking rates per category (Pargent et al. 2022). These booking rates are calculated as a weighted average of the values' empirical and global mean, handling rare categories through this. The final feature set with 83 columns excludes all columns that were identified as problematic in the EDA due to collinearity, missing values, and all original columns for which new features were created.

4.4.2 Feature Engineering: Personal Loan Cross-Sell

4.4.2.1 Feature Engineering and External Data Integration

During the feature engineering, external data macro enrichment has been incorporated, including geographic wealth indicators from the National Institute of Statistics (2021) report, which includes per capita purchasing power, municipality purchasing power index, as well as Europe's historical market interest rate, 6-month EURIBOR. Research indicates that incorporating traditional credit metrics, consumer banking transaction data, and regional economic indicators into machine learning models allows for richer representations and improves predictive performance compared to linear specifications (Sadhwani et al. 2021). Therefore, this external data was included to capture

contextual, socioeconomic, and macroeconomic conditions that are not covered in the provided data.

Then, an assessment was made of whether regional affluence influences cross-sell propensity. In-memory dictionaries were developed to map postal codes to their corresponding wealth data using National Institute of Statistics data (2021), providing an overview of the economic status of each location. As part of this process, 16,255 customers were successfully linked to their respective regional wealth indicators under *wealth_ipc, wealth_ppc,* and *wealth_fdr*. The e*ligibility_date* was used to retrieve the corresponding 6-month EURIBOR rate at that point in time, creating the feature *market_rate_at_offer*. For customers who had an existing auto loan with the bank, the same matching process was applied to their loan activation date, generating the feature *market_rate_at_activation*. Additionally, three temporal features were added to capture the customer age of the second holder, *age_h2,* the age of the existing auto loan, *loan_age_months*, and the remaining loan fraction, *remaining_loan_fraction,* by using the campaign's eligibility date. Prior literature shows that customer behavior and credit features evolve over time and that incorporating temporal dynamics of customer history improves predictive performance (Munoz-Cancino et al. 2022). Motivated by this, two interaction features were derived to quantify risk and saturation: *offers_per_year* to measure contact density, and *debt_to_income_ratio* as a proxy for the monthly financial burden based on the installment and disposable income.

The customer profile features are enhanced by creating a binary indicator based on the existence of *date_of_birth_second_holder,* which captures whether a second borrower exists. Compared to the continuous age variable, the binary indicator is particularly valuable for tree-based models, as it provides a clean and interpretable split that distinguishes between customers with and without a

co-applicant (Kuhn and Johnson 2013). The bank's internal risk scoring, *monitoring_score,* was converted into a numerical format to preserve its ordinary nature while keeping predictive power. To capture each customer's position within their loan lifecycle, a standardized timing variable was constructed. To construct the lifecycle timing variable, several loan-timing fields in the dataset were examined, including *term_original, months_to_end, loan_age_months,* and *activation_date*. Among these, *remaining_loan_fraction* was selected as the decisive lifecycle indicator because it standardizes each customer's position within the loan, irrespective of the original term length. This normalized metric, which ranges from 1 at loan origination to 0 at loan completion, enables a consistent comparison across customers. Based on this measure, three intuitive lifecycle stages were defined by segmenting the distribution into "Early" (*remaining_loan_fraction* greater than 0.66), "Mid" (between 0.33 and 0.66), and "Late" (0.33 or below), ensuring that each stage reflects a clear and comparable segment of the loan lifecycle.

4.4.2.2 EDA of Newly Engineered Features

As the fourteen newly added features have not yet been explored, a second exploratory data analysis was performed. The newly added socioeconomic indicators (*wealth_ipc*, *wealth_ppc*, *wealth_fdr*) were inspected for skewness and outliers, as this type of data frequently exhibits heavy-tailed behavior (Benhabib and Bisin 2018). Temporal and market-related features, such as *loan_age_months, remaining_loan_fraction, market_rate_at_offer,* and *old_loan_vs_market_diff*, were evaluated to verify that their ranges and transformations aligned with domain expectations and could be reliably used as model inputs. Two engineered features (*offers_per_year, debt_to_income_ratio*) were reviewed to confirm that they behaved as intended and captured the conceptual constructs of contact intensity and financial burden.

Variables that represented highly overlapping information (correlations above 0.95) were removed to improve model stability and reduce noise. This led to the exclusion of *offer_term_diff, offer_tan_diff,* and *is_co_applicant*, which were determined to be near duplicates of other, more stable predictors. Additionally, *gender_h1* was removed due to potential missingness in real-world application settings and to align with fair-lending and fairness guidelines. The resulting feature selection includes 36 predictors in total.

4.4.3 Exploratory Customer Segmentation with K-Means

After identifying the key characteristics that distinguish simulator behavior from non-simulators, the next step is to assess whether these differences translate into meaningful customer segments. For this, k-means clustering is applied to identify customer profiles since it provides an efficient, data-driven way to group individuals with similar characteristics. This allows for the identification of underlying customer groups whose shared characteristics are not easily identifiable from individual variables alone (Reutterer and Dan 2019). For the clustering analysis, only the variables identified during the exploratory data analysis as having the strongest relevance were included. These consist of *previous_offers, pre_approved_amount, disposable_income, crc_institutions, age_h1, customer_tenure_months,* and *monitoring_score*, ensuring that the segmentation is based on the most influential customer characteristics. Nationality was not included in the k-means clustering because the algorithm's Euclidean distance measure is unsuitable for purely categorical variables . As highlighted by Dinh et al. (2025), most clustering algorithms designed for categorical data require alternative similarity measures, since nominal attributes lack inherent ordering and cannot be meaningfully compared within distance-based methods, such as k-means. Using the elbow method to examine how within-cluster variance decreases across different numbers of

clusters, multiple k values were evaluated. Based on the evaluation of different cluster options, a two-cluster solution was selected as the most appropriate segmentation structure.

4.5 Model Development

First, the general model development approach is described. Then, specific adaptations are elaborated for each use case as they differ in data form and evaluation logic, which affects the preprocessing steps and model configuration.

For both use cases, the analytical objective is propensity scoring. This score represents the customer's propensity to perform the target action, serving as the basis for ranking and decision-making. Across logistic regression, random forest, and XGBoost, all models were developed within a unified scikit-learn pipeline to ensure comparability. The pipeline applied encoding, winsorization, log-transformations for skewed numerical variables if needed, and the removal of process-dependent fields. All preprocessing components were fitted only on the training data to prevent leakage. One-hot encoding was selected over label encoding, as the categorical variables lack any inherent ordinal structure.

Logistic regression is introduced as the baseline model in both use cases due to its simplicity and interpretability, in line with the literature reviewed in chapter 3.2.1.2 (Huang et al. 2024; Lyu et al. 2025; Špicas et al. 2023; Thomas et al. 2006). Logistic regression requires standardization, log-transformations for skewed features, and targeted feature cleaning to address multicollinearity and high dimensionality. The model’s regularization parameters (L1, L2, and Elastic Net) are tuned to improve stability (Schreiber-Gregory et al. 2018).

Random forest and XGBoost do not require feature scaling because tree-based models split features instead of relying on distances or coefficients (Breiman 2001; Chen and Guestrin 2016). The

random forest is implemented, which provides a more flexible, non-linear modeling approach and often outperforms logistic regressions (Lyu et al. 2025; Špicas et al. 2023). This allows the model to capture complex interactions and patterns that logistic regressions cannot learn. The baseline random forest uses 100 trees without a depth constraint and a fixed random seed.

XGBoost extends this by building trees sequentially, where each tree corrects residual errors from the previous one. This structure often captures finer interaction patterns than bagging alone. The baseline XGBoost model for the loan conversion uses 300 trees with a depth of 2, a learning rate of 0.1, and subsample and colsample_bytree values of 0.8 to improve generalization. It uses logloss as the evaluation metric and a fixed seed for reproducibility. For the personal loan cross-sell case, the baseline XGBoost model applied class balancing through the scale_pos_weight parameter and used the histogram-based tree method to improve computational efficiency. The model relies on the default XGBoost structural parameters, which include a maximum tree depth of six, one hundred boosting trees, and a learning rate of 0.3.

In both use cases, hyperparameter tuning was applied for random forest and XGBoost to explore tree depth, ensemble size, subsampling ratios, and regularization settings. XGBoost hyperparameter tuning also includes learning-rate adjustments and, where relevant, class weighting.

4.5.1 Model Development Approved Loan Conversion

4.5.1.1 Predicting Approved Loan Conversion

This chapter explains how the general pipeline was adapted to the approved loan conversion use case. The aim is to respect the time order of events, avoid leakage, and understand both predictive performance and the underlying behavioral patterns (Maggi et al. 2014).

The pipeline applied WoEBinner for logistic regression and TargetMeanEncoder for tree-based models (Vanhoeyveld et al. 2020). All models were trained on a chronologically defined training set and evaluated on a separate validation period to preserve temporal order, prevent leakage, and ensure forward-looking performance estimates (Bergmeir and Benítez 2012). After hyperparameter tuning, each model was refitted on the combined training and validation period and subsequently evaluated on the hold-out out-of-time test set, which contains the most recent observations. The final model is additionally tested on separate dealer and broker test data to understand how well the model generalizes for each partner type. Therefore, calibration curves are created and classification reports generated for each test set. Next to the logistic regression, random forest, and XGBoost trained on the clean feature set to find the optimal model to predict loan conversion, the following models were also trained.

4.5.1.2 Target-Leakage Logistic Regression

One logistic regression was trained that included almost all available variables to test the hypothesis from the EDA that several features behave as deterministic indicators of target_booked. Although these variables may technically exist before first approval, they are often populated or updated afterwards. Therefore, they signal internal processing stages rather than applicant or deal characteristics. Examples include *cost_estimated, life_eligible, ppp_eligible, conference, fraud,*

analyst_masked, user_conferencia_masked, and *commercial_manager_masked. Gender* is treated similarly, as the "Unknown" category appears highly predictive due to early process interruptions rather than inherent risk, and its inclusion introduces ethical concerns (Hurlin et al. 2024; Pavón Pérez 2025). Estimating this baseline model is still informative, as it clarifies the extent to which such variables distort performance.

4.5.1.3 Exploratory Explaining Drivers Model

The tuned XGBoost model was also trained on an expanded feature set to understand which factors drive conversion across the full lifecycle of an application. All identified leakage variables were removed, but the model kept demographic characteristics, partner attributes, pricing indicators, insurance details, financial metrics, re-analysis counts, and time-to-event variables. This version does not aim to support operational decision-making. Instead, its goal is to show which customer, partner, and proposal patterns relate to higher conversion rates.

4.5.1.4 Exploratory Channel-Specific Models (Broker and Dealer)

Understanding channel differences is a core objective of this work. For this reason, two separate XGBoost pipelines were trained: one on broker applications and one on dealer applications. Preprocessing remains identical across both pipelines, so any differences in results reflect structural channel differences rather than methodological artefacts. A tailored hyperparameter search was conducted for each channel. The dealer subset is strongly imbalanced, with more than 80% of cases converting, so the dealer model applies a scale_pos_weight derived from the booked-to-non-booked ratio.

4.5.2 Model Development Personal Loan Simulation

4.5.2.1 Predicting Personal Loan Simulation

This chapter explains how the general pipeline was adapted to the personal loan cross-sell use case. Since the data is cross-sectional and represents a single snapshot from July 2025, no time structure exists. A different modeling strategy is therefore required.

Stratified K-fold cross-validation was used because it preserves the class distribution in each fold and provides reliable estimates for imbalanced classification (Kuhn and Johnson 2013). The preprocessing pipeline is re-fitted inside each fold to avoid leakage (Varma and Simon 2006).

Logistic regression requires additional feature-engineering steps due to multicollinearity and high-cardinality categorical variables. High-cardinality brands were grouped into three categories, multicollinear financial variables were removed in favor of cleaner counterparts, and Recursive Feature Elimination (RFE) was applied to retain the twenty most informative predictors (Bulut et al. 2025). These steps reduce dimensionality and improve coefficient stability in the high-dimensional one-hot encoded space.

Random forest and XGBoost can operate on the full, ungrouped feature set because their splitting mechanisms are robust to multicollinearity and naturally handle high-cardinality categories (Chen and Guestrin 2016). To address the class imbalance, class weighting was applied. Class weighting was preferred over SMOTE to preserve the original data distribution. Studies indicate that SMOTE generates synthetic samples that may not reflect real-world distributions and can degrade probability calibration (Abdelhamid and Desai 2024).

Hyperparameter tuning was conducted using RandomizedSearchCV for both tree-based models (Alamsyah et al. 2024). The search space includes the number of trees, maximum depth, learning rate (for XGBoost), subsample and column-subsample ratios, and minimum child weight. Once the

optimal configuration was identified, the tuned model was re-estimated within the same preprocessing pipeline.

4.5.2.2 Exploratory Logistic Regression: Campaign Fatigue

An additional logistic regression analysis was conducted to examine the relationship between *lifecycle_stage* and the probability of initiating a simulation. In the baseline specification, simulation propensity was modeled as a function of *lifecycle_stage*, *previous_offers*, and *customer_tenure_months*. *Lifecycle_stage* was included as a categorical variable, with the late stage serving as the reference group. The objective was to assess whether the effects of *previous_offers* and *customer_tenure* differ across lifecycle stages. The extended model included interaction terms between *lifecycle_stage* and the number of *previous_offers*, and between *lifecycle_stage* and *customer_tenure_months*. Both were calculated as the product of the lifecycle stage and the corresponding variable. Interaction terms were specified as products of the individual predictors, consistent with standard practice in logistic regression to allow the effect of one predictor to vary by levels of another .

4.5.2.3 Propensity-Based Tiering

To translate the predictive model outputs into an operational contact strategy, this study developed a tiered segmentation framework. This approach aligns with recent literature suggesting that effective resource allocation requires integrating propensity scores with broader customer value indicators to guide segmentation and treatment prioritization (Neslin et al. 2006; Abdullah and Hasan 2023; Gubela et al. 2024). While Abdullah and Hasan (2023) explicitly advocate for an architecture that links propensity models with financial CLV to justify resource allocation, this

study modifies the framework to adapt to the unavailability of customer value data. The proposed framework substitutes the economic dimension with the behavioral personas derived from k-means clustering. The framework constructs a cohort matrix based on two distinct dimensions. First, for the propensity score, the probability outputs from the personal loan simulation model serve as the basis for customer prioritization. The cumulative gains chart is utilized to determine specific cut-off points, dividing the customer base into priority tiers based on the slope of the marginal gain. Second, the k-means clusters classify customers into behavioral profiles.

4.6 Performance Metrics

4.6.1 Performance Metrics: General Framework

In both use cases, the model aims to predict an outcome for a binary classification problem, so the metrics measure how successfully the model can distinguish between two mutually exclusive classes. To evaluate the trained models, a set of standard classification performance metrics is applied: accuracy, precision, recall, F1 score, and the Receiver Operating Characteristic (ROC) curve, along with its Area Under the Curve (AUC) (Rainio et al. 2024) (Figure 2). A backtesting framework that evaluates the business impact of the predictive models complements the standard metrics. Consistent with uplift-based evaluation approaches in applied credit analytics (Baier and Stöcker 2022), observations are ranked by predicted propensity, subsets are selected at predefined thresholds, and their outcome rates are compared with the baseline. This quantifies the improvement over random selection in line with established approaches in predictive marketing (Moro et al. 2014).

4.6.2 Performance Metrics: Approved Loan Conversion

In the approved loan conversion use case, the dataset is slightly imbalanced, given roughly 70% of the applications convert into a booked loan. This allows most performance metrics named above to remain informative. AUC-ROC is used as the main performance measure, as it provides a robust comparison between models and across thresholds (Bekkar et al. 2013). Accuracy is also reported, allowing the bank to direct its efforts towards the customers with the highest conversion probability. However, accuracy represents a trade-off, as the model can always predict the majority class and achieve high accuracy without adding real business value (Powers 2011). A low precision would mean that many loans predicted to be booked do not follow through in the end. This can lead to wasted effort in the teams included in this process. On the contrary, low recall implies that a considerable portion of the actual converters are misclassified, resulting in lost revenue if not targeted because of the prediction. Both outcomes carry financial value, yet the bank faces no capacity constraints. The business risk, therefore, centers on failing to identify customers who would convert. Therefore, recall becomes more important than precision from a business perspective. Furthermore, the F1 score is considered as it is a balanced summary measure.

From a business perspective, the main business KPI for this use case is the conversion rate, and the main objective is to provide insights into what drives conversion and how the bank might improve its operations. While closing the conversion gap between brokers and dealers is interesting for the bank, the main objective is to improve the bank's overall conversion performance rather than enforcing channel convergence. To evaluate model performance in a business-aligned way, back-testing was applied.

In line with the general backtesting framework described above, the loan conversion case requires a threshold-based, forward-looking evaluation because the data are temporally ordered. Predicted

probabilities are, therefore, compared against a series of cutoffs (0.50–0.99), and for each cutoff, the model selects all applications where the predicted conversion propensity exceeds the threshold.

4.6.3 Performance Metrics: Personal Loan Cross-Sell

Precision is the primary objective in the cross-sell case to minimize resource waste associated with false positives. While identifying simulators is necessary, optimizing recall alone leads to contacting nearly all customers. Thus, the evaluation focuses on maximizing precision among the highest-ranked prospects. Therefore, the model selection and tuning process relies on the AUC-PR (Average Precision Score), which ensures the model effectively ranks simulators at the top of the list across all potential thresholds. Unlike AUC-ROC, which can appear inflated in highly imbalanced datasets, AUC-PR focuses entirely on the model's performance for the positive class (Davis and Goadrich 2006). This focus makes it appropriate for the cross-sell use case, prioritizing the identification of simulators while limiting false positives.

To evaluate operational application, this study analyzes Precision at Top k (Precision@k). Here, k represents a relative cut-off, specifically the top 20% quintile of the ranked customer list. The selection of this 20% threshold aligns with the propensity tiering framework established in the literature (Abdullah and Hasan 2023). In addition, empirical testing within this dataset indicated that a 20% cut-off represents the smallest segment size that yields stable and interpretable performance estimates. Smaller cut-offs (e.g., 10%) resulted in insufficient sample sizes for reliable evaluation.

Given that the dataset is cross-sectional, traditional time-series back-testing is not applicable. Instead, the study employs a static out-of-sample simulation to project economic impact, aligning with database marketing frameworks (Blattberg et al. 2008). The unseen test partition serves as a

proxy for a marketing campaign (Kuhn and Johnson 2013). In this simulation, customers are ranked by predicted propensity scores in descending order to prioritize high-probability prospects (Moro et al. 2014). To estimate the operational outcome, metrics observed in the test set are extrapolated to the total eligible population. This projection generates a summary quantifying key metrics as the expected campaign size (n), the simulation rate (precision), and the capture rate (recall).

4.7 Explainable AI

SHAP is a common AI explainability method in the financial sector (Černevičienė and Kabašinskas 2024; Klein and Walther 2024), which can help identify the main factors influencing auto-loan conversion rates across channels. This can guide targeted strategies and process optimization. Regarding personal cross-sell, it provides an understanding of the factors that influence the propensity to start a simulation for an additional offering. While examining the coefficients or importance scores alone offers a general indication of each feature's average effect within the model, frameworks such as SHAP provide localized explanations that can be aggregated to offer a global view of how the features influence the model.

The LinearExplainer was used for logistic regression, and the TreeExplainer for the tree-based models to efficiently compute Shapley values. The resulting SHAP values quantify how much each feature value increases or decreases the predicted conversion probability, providing a global view of feature importance and directional effects (Lundberg and Lee 2017).

In the loan conversion case, SHAP plots were created for the final predicting model, the exploratory model explaining drivers, and the exploratory channel-specific models (broker and dealer). For the cross-sell case, SHAP plots were computed to analyze feature contributions and later compare how these contributions differ across the customer personas identified in the segmentation step.

In addition to the global SHAP analyses, local force plots were generated for the tuned XGBoost model to provide case-level explanations and to ensure GDPR compliance (Černevičienė and Kabašinskas 2024). This yields a transparent, instance-specific explanation showing how much each feature shifts the prediction relative to the model's baseline, and supports regulatory requirements such as GDPR (Directorate-General for Parliamentary Research Services (European Parliament) 2020).

5 Results

5.1 Personal Loan Cross-Sell

For personal loan cross-sell, the analysis starts with an exploratory assessment of data quality and customer characteristics, followed by the evaluation of predictive models and complementary exploratory analyses to identify the main behavioral drivers, segmentation patterns, and evidence of campaign fatigue.

5.1.1 Predicting Personal Loan Simulation

Individual Part: Nils Rudolf (63855)

5.1.1.1 Predicting Personal Loan Simulation: Logistic Regression Model

The baseline logistic regression model, incorporating Recursive Feature Elimination (RFE), establishes a benchmark performance with an AUC-PR of 0.0583. As shown in the threshold analysis (Figure 63), the model exhibits limited discriminatory power in the high-precision range, reflected in a low precision at the top 20% of 0.0657. The model achieves a recall at the top 20% of 0.4095, identifying approximately 41% of simulators within the top quintile. The high number of false positives (707 in the top 20%) indicates that the linear decision boundary struggles to filter out non-responsive customers effectively. Feature selection via RFE reduced the dimensionality to

20 key predictors to mitigate noise from high-cardinality features such as vehicle brands. The resulting coefficient analysis (Figure 62) highlights *nationality_h1_brazil* (+1.29) and specific banking relationships *bank_first_holder_caixa_de_credito_agricola* (+1.68) as strong positive drivers. Conversely, *fuel_type_original_Mild_Hybrid* (-2.08), *profession_contract_type_Self_Employed* (-1.83), and *bank_first_holder_caixa_agricola_de_torres_vedras* (-1.9) exhibit strong negative associations. Notably, the variable *previous_offers* presents a negative coefficient of -0.81. This aligns with the negative trend observed in the exploratory analysis, indicating that within the linear model, increased contact frequency is associated with a reduction in simulation probability.

5.1.1.2 Predicting Personal Loan Simulation: Random Forest Model

The random forest model improves upon the baseline by leveraging non-linear decision trees to capture complex customer profiles. With an AUC-PR of 0.0753, it surpasses the logistic regression, particularly in ranking quality. Regarding the operational metric precision at the top 20%, the random forest improves it to 0.0826. The model achieves a recall at top 20% of 0.5143, identifying nearly half of all simulators within the target group (Figure 64). Hyperparameter tuning optimized the model structure, favoring deep trees (max_depth=30) and specific split criteria (min_samples_split=5) over explicit class weighting, as the hyperparameter search selected class_weight=None. The feature importance analysis (Figure 61) reveals a shift in predictor dominance compared to the linear model. While *previous_offers* (0.072) remains the primary predictor, financial and temporal features such as *old_loan_vs_market_diff* (0.056) and *remaining_loan_fraction* (0.056) gain substantial weight. This indicates that the random forest

incorporates the derived interaction features regarding the customer's position in the loan lifecycle and the relative attractiveness of the new interest rate more strongly than the linear model.

5.1.1.3 Predicting Personal Loan Simulation: XGBoost Model

The XGBoost model exhibits the highest predictive performance across all evaluated metrics, achieving an AUC-PR of 0.1007. Hyperparameter tuning selected a scale_pos_weight of 15.16, approximately half the raw class imbalance ratio. XGBoost delivers a precision at the top 20% of 0.0902. While the model achieves a recall at the top 20% of 0.5619, capturing 56% of simulators within the top 20% of the ranked database (Figure 65). The gain chart (Figure 66) illustrates this, maintaining higher precision levels at improved recall thresholds compared to previous iterations. The feature importance profile (Figure 60) corroborates the random forest findings but assigns distinct weights to customer segmentation features. *Previous_offers* (0.040) and *pre_approved_amount* (0.025) are the leading predictors, followed by *crc_institutions* (0.020), *nationality_h1_brazil* (0.019), and the clustering feature *cluster_0* (0.019). This highlights that XGBoost relies heavily on campaign history and financial liquidity indicators to rank customers. Isotonic calibration was applied to align predicted probabilities with observed frequencies.

Table 3 summarizes the performance metrics for the three candidate models on the test set. The XGBoost model outperforms in both ranking capability (AUC-PR) and precision at the top 20%. Specifically, XGBoost identifies 15.24 percentage points more simulators than the logistic regression baseline (56% vs 41%) and achieved 4.24 higher percentage points for AUC-PR for the same operational effort. Consequently, XGBoost is selected as the best model for the final implementation and propensity scoring.

	XGBoost + Class Weighting	Random Forest + Class Weighting	Logistic Regression + RFE

AUC-PR (Avg Prec)	0.1007	0.0753	0.0583
AUC-ROC	0.7665	0.7166	0.6738
Precision @ 20%	0.0902	0.0826	0.0657
Recall @ 20%	0.5619	0.5143	0.4095
F1 score @ 20%	0.1553	0.1421	0.1138

Table 3: Predicting Personal Loan Simulation - Model Comparison

5.1.1.4 Predicting Personal Loan Simulation: Backtesting

The gain analysis in Figure 68 quantifies the ability of the XGBoost model to prioritize responsive customers. At the most selective threshold of the top 10% (campaign size of 1,634 customers with a capture rate of 35.2%) the model achieves a simulation rate of 11.3%. Expanding the selection to the business target of the top 20% (campaign size of 3,269 customers with a capture rate of 56.2%), the simulation rate for this cohort is 9.0%. This represents a lift factor of 5.8% compared to the global baseline probability of 3.2%. A broader application targeting the top 50% (campaign size of 8,183 customers with a capture rate of 88.6%) results in a simulation rate of 5.7%. Conversely, the full dataset reflects the unoptimized baseline, where 16,366 contacts are required to reach 100% of simulators with a baseline simulation rate of 3.2%.

The assessment of model calibration (Figure 69) shows that predicted probabilities align with observed frequencies for scores up to 0.18. Beyond this threshold, the curve indicates unreliable probability estimates.

5.1.1.5 Predicting Personal Loan Simulation: SHAP

The global features SHAP analysis (Figure 46) shows that *previous_offers* is the most influential feature, exhibiting predominantly negative SHAP values when the feature value is high.

Similarly, *nationality_h1_brazil* shows a negative impact when the feature is present. In contrast, *crc_institutions* displays mostly positive SHAP values for high feature values.

Housing-related variables exhibit asymmetric effects depending on their value levels. For *housing_type_own_no_debt*, low feature values are associated with positive SHAP contributions, while higher values tend to push the model output in a negative direction. *Housing_type_own_mortgaged* shows the opposite pattern, where low values are linked to negative SHAP values. Financial variables such as *pre_approved_amount* and *remaining_loan_fraction* display predominantly positive effects at lower value ranges, while *disposable_income* has a negative effect at lower values.

5.1.1.6 Propensity-Based Tiering

Individual Part: Nils Rudolf (63855)

This chapter integrates the XGBoost predictive scores with the identified customer clusters defined in chapter **Error! Reference source not found.**, to construct a cohort matrix for propensity-based cross-selling. Based on the thresholds defined in the cumulative gain chart (Figure 66), the customer base is segmented into three tiers: High (Top 22%), Mid (Top 49%), and Low (Bottom 51%). Figure 67 presents the simulation rates for the intersection of propensity tiers and customer clusters. Within the high-priority tier, the Fatigued Tenured segment exhibits a conversion rate of 9.6% (13 simulators out of 136), corresponding to an absolute lift of 6.4% against the global average. The Fresh Prospect segment within the same tier shows a conversion rate of 8.5% (50 simulators out of 585) and a lift of 5.3%. In the Mid priority tier, conversion rates decrease to 2.6% for Fatigued Tenured and 3.6% for Fresh Prospect customers. Two-sample z-tests evaluate the differences in response rates. The difference between the aggregated High and Mid tiers is

statistically distinct ($p < 0.001$). However, within the specific tiers, the difference between clusters does not reach statistical thresholds ($p < 0.05$). In the High tier, the comparison between Fatigued Tenured and Fresh Prospect yields a p-value of 0.6594. Similarly, the Mid-tier comparison results in a p-value of 0.2585. Cohort Profiling in Figure 70 details the demographic and behavioral profiles of the segments. Within the High Priority Tier, Fatigued Tenured customers exhibit a mean tenure of 99.6 months and an average of 4.47 previous offers, and a higher engagement with credit institutions (4.82 vs 3.61). In contrast, Fresh Prospects show a mean tenure of 28.9 months and 1.46 previous offers. In the Mid Priority Tier, the Fatigued Tenured segment shows a mean tenure of 82.11 months and the highest contact frequency, with an average of 10.62 previous offers. Fresh Prospects in this tier have a mean tenure of 27.89 months and an average of 3.79 previous offers (Figure 70).

6 Discussion

The discussion chapter interprets the empirical results of both use cases and links them to operational and strategic implications for Banco Primus.

6.1 Personal Loan Cross-Sell

Understanding customers is important for designing an effective and sustainable cross-sell strategy. While predictive modeling provides an estimate of each customer's propensity to initiate a loan simulation, the underlying behavioral and lifecycle patterns behind these predictions should also be examined. Therefore, this chapter evaluates the predictive performance of the XGBoost model and its global feature determinants. Subsequently, k-means clustering analyzes the predictions to

identify behavioral segments, focusing on campaign fatigue. Finally, these insights are translated into a propensity-based tiering framework.

6.1.1 Interpretation: Predicting Personal Loan Simulation

Individual Part: Nils Rudolf (63855)

This chapter evaluates the predictive performance and feature determinants of the XGBoost model to assess its operational suitability. The discussion addresses the validity of resource allocation, identification of economic drivers with XAI, and the implications of behavioral patterns for campaign strategy.

The evaluation confirms that the XGBoost ensemble identifies non-linear customer behaviors and outperforms the logistic regression baseline. While the AUC-PR of 0.10 reflects the class imbalance, the model indicates stability with a standard deviation of +/- 0.01 in stratified k-fold cross-validation. For operational decision-making, the ranking ability of the model takes precedence over the probability scores. The gain analysis (Figure 68) indicates a Lift of 2.8 in the top 20%, demonstrating that targeting this segment captures nearly three times the simulation compared to a random selection. This efficiency supports the transition from broad-based marketing to marginal-value allocation frameworks (Neslin et al. 2006). However, probability calibration degrades for scores exceeding 0.18, implying that the model functions primarily as a ranking instrument rather than a precise estimator of absolute probabilities.

The global SHAP analysis of feature importance identifies operational levers. Three key findings regarding monetary sensitivity, external credit load, and fatigue emerge. First, the number of active external credit relationships (*crc_institutions)* shows a positive correlation with simulation probability. This suggests that customers with multiple financial institutions use the product for

debt consolidation. This supports a solution-oriented marketing approach rather than a purely sales-driven one. Second, the inverse relationship between previous offers and simulation probability, highlighted by Banco Primus during stakeholder discussions, contradicts Reinartz and Kumar's (2003) "inverted U-shape" theory. SHAP plots show that additional contact attempts reduce simulation probability, indicating immediate campaign fatigue.

While global SHAP (Figure 46) analysis identifies general drivers, it assumes that feature effects are consistent across the entire customer base. This approach obscures the fact that different subgroups may behave differently. Furthermore, a high propensity score alone does not ensure optimal resource allocation, as it views conversion probability in isolation from strategic context. Abdullah and Hasan (2023) note that effective targeting requires combining predictive scores with distinct behavioral metrics to avoid inefficient selection strategies. To address this requirement, the following k-means analysis adds a behavioral dimension to the prediction model.

6.1.2 Recommendation: Optimizing Campaigns via Propensity-Based Tiering Frameworks

Individual Part: Nils Rudolf (63855)

The previous analyses identified two dimensions required for an effective cross-sell strategy: the predictive probability of simulation (evaluated in chapter 6.1.1) and the behavioral receptiveness of the customer (defined by the personas in chapter 6.1). The current campaign strategy operates on a broad inclusion principle, where the default approach involves contacting the entire eligible customer base. The propensity model results presented in chapter 6.1.1, however, indicate that simulation probability varies by a factor of ten between customer propensity tiers. This variance suggests that a uniform contact strategy creates inefficiencies. While the direct marginal costs of

digital channels (SMS, email) are negligible, literature regarding the "attention economy" implies that the limiting factor is customer receptiveness rather than budget (Ansari and Mela 2003).

As described in chapter 2, the analyzed dataset reflects the period before September 2025, during which customers received three contacts per month. The analysis indicates that signs of saturation were already present at this frequency. Since the bank recently increased the volume to four monthly contacts, this operational shift risks accelerating fatigue effects. This "one-size-fits-all" approach conflicts with the principles of marketing resource allocation described by Neslin et al. (2006), who argue that profitability depends on allocating resources to where the marginal lift is highest. The exploratory analysis and model results in chapter **Error! Reference source not found.** highlighted two primary inefficiencies in the current setup: i) As previously discussed, customers who received many previous offers exhibited a low simulation rate, compared to those with few prior offers. While this negative correlation might stem from selection effects rather than pure fatigue, evidence suggests that excessive contact frequency contributes to diminishing returns. ii) SMS is a high-attention channel but carries a risk of perceived intrusiveness. Indiscriminate use of SMS on "fatigued" customers can increase the risk of unsubscription (Boustani et al. 2024).

To translate predictive outputs into operational decisions, Abdullah and Hasan (2023) propose integrating propensity scores with CLV. However, consistent financial proxies for customer value are unavailable in this study. To address this limitation, this research adapts the approach by substituting financial metrics with behavioral insights. A modified framework is proposed that integrates the predictive probability scores from the XGBoost model with the customer personas derived from k-means clustering in chapter 6.1. The objective of this targeted contacting strategy is to increase the absolute number of simulations while accounting for opportunity costs and the risk of brand erosion caused by campaign fatigue.

The framework employs a two-stage logic: i) The XGBoost score indicates the prioritization. Using the Cumulative Gains Chart (Figure 66), three distinct tiers are proposed based on the points where the marginal lift begins to plateau; ii) The k-means clusters indicate a potential content strategy. The framework assigns the specific communication tactics outlined in chapters **Error! Reference source not found.** and **Error! Reference source not found.** based on whether a customer is profiled as "Fresh" or "Fatigued". Crossing the three priority tiers with the two behavioral clusters theoretically identifies six segments (Figure 67).

The statistical tests in chapter 5.1.1.6 confirm that the score is the primary driver of simulation probability between the tiers. Within all tiers, the simulation rates between "Fresh" and "Fatigued" clusters did not differ statistically, as the sample size is too small. This implies that the propensity score serves as the criterion for selection, whereas the cluster assignment indicates the communication tactic. However, an exception is applied to the Low priority tier. In this segment, the predicted simulation probability drops to a marginal level 0.7 & 0.8%, which is why the Low tier is consolidated into a single cohort:

1.1) Tier High - Fresh Prospects: The clustering analysis characterizes this group as digitally receptive with elevated simulation potential. Based on this profile, the tiering strategy sets the contact frequency to the recommended limit of four contacts per month to capture value before interest declines. Regarding the channel mix, the framework adopts the suggested mobile-first approach by combining SMS and email for activation. Furthermore, this cohort is designated as the primary audience for the social media retargeting tactics to provide a behavioral nudge, proposed in the previous chapter. 1.2) Tier High - Fatigued Tenured: This segment exhibits that elevated predictive scores indicate potential financial need, while the behavioral history shows saturation and reduced responsiveness to standard promotional content. To address this, the

strategy shifts the communication objective from sales to support, applying the recommendation to position the bank as a partner for installment optimization. Operationally, this entails prioritizing email over SMS as the initial contact channel. This approach reduces perceived intrusiveness while maintaining a path for simulation for customers with active credit demand. 2.1) Tier Mid - Fresh Prospects: This cohort falls within the next 27% of the population. While these customers share the behavioral characteristics of the "Fresh" profile, their lower propensity score suggests they are not ready for direct communication. The analysis in chapter **Error! Reference source not found.** identified that relevance and educational content are drivers to prevent early fatigue in such groups. Consequently, the operational strategy for this tier removes the SMS channel. Instead, the framework uses email to nurture these prospects with informational content regarding loan safety and process simplicity. 2.2) Tier Mid - Fatigued Tenured: The analysis in chapter **Error! Reference source not found.** highlighted that for saturated customers, additional volume yields diminishing returns and risks triggering avoidance behavior, consistent with the "Sleeping Dogs" hypothesis (Radcliffe and Surry 2011). Implementing the recommended frequency capping, this tier is restricted to a maximum of one to two contacts per month. The channel strategy is defensive, using only email to minimize negative reactions and prevent unsubscribe actions. 3.) Tier Low - Low Priority: The final cohort captures the bottom 51% of the customer base, where the simulation probability ranges from 0.7% to 0.8%. To align with the cost-efficiency principles discussed in chapter 6.1.1, marketing investment is restricted. This group receives a single email per campaign to maintain brand presence without allocating resources to channels with low marginal impact.

6.1.3 Limitations: Personal Loan Cross-Sell

A primary limitation of this study concerns the selection bias inherent in the training data. The model was trained exclusively on a dataset of customers who were already pre-selected by the bank's existing eligibility rules. As these rules function as a "Black Box" and were not disclosed by Banco Primus, the resulting model estimates the probability of simulation conditional on being pre-selected, rather than the probability within the general customer population. Consequently, applying this model to customers outside the current eligibility criteria creates an out-of-distribution risk, consistent with broader concerns regarding the contextual validity and generalizability of AI models discussed by Davenport et al. (2020). Furthermore, the reliance on a static data snapshot from July 2025 limits temporal validity. This cross-sectional approach fails to capture seasonal effects or macroeconomic shifts, such as changes in interest rate policies. Additionally, the dataset lacks digital behavior metrics, such as email open rates or click-through paths. This feature gap restricts the model to static attributes and transactional history, potentially ignoring early behavioral signals of interest.

A discrepancy exists between the model's ranking performance and its probability calibration. While the model effectively sorts customers by priority, the specific probability estimates exceeding 0.18 do not accurately reflect true simulation rates. This implies that while the model is a valid instrument for prioritizing contact lists, it is less suitable for precise financial forecasting. Within the proposed framework, endogeneity obscures the true relationship between previous offers and simulation behavior. Without a historical control group, it remains indistinct whether lower simulation rates stem from causal campaign fatigue or from a selection bias where non-converting customers simply accumulate more contacts over time. Additionally, the sample size of the test set limits the granularity of the cohort analysis. Segmenting the data into the proposed

priority tiers and behavioral clusters results in small sub-groups, suggesting that specific metrics for sub-segments likely contain statistical noise.

A final limitation concerns the preliminary and interpretive nature of the recommendations, consistent with the constraints discussed in the approved loan use case. While the model identifies statistical associations, it does not establish causal evidence, as an uplift model does (Radcliffe and Surry 2011). It cannot distinguish between customers who require a nudge to convert and those who would have simulated regardless, nor can it capture organizational or behavioral dynamics (Naser 2025). Consequently, any changes in operational practice informed by such insights must be introduced with caution. Continuous monitoring and periodic recalibration are necessary as internal processes and market conditions evolve. The results, therefore, also represent a starting point for informed experimentation and iteration rather than definitive prescriptions.

7 Conclusion

Banks are under growing pressure to improve cost efficiency and revenue generation while meeting rising expectations for transparency, governance, and regulatory compliance. Although machine learning and explainable AI are well established in some financial tasks, their application to certain customer journey stages, such as customer acquisition and customer development, remains comparatively, particularly with respect to identifying operational and behavioral drivers behind post-approval drop-off and campaign engagement, and to translating model outputs into actionable insights to optimize banking processes. Therefore, the thesis examined this research question: How can machine learning and explainable AI be used to identify the drivers of post-approval auto loan conversion and personal loan cross-selling at Banco Primus to support operating decisions?

The results indicate that machine learning models rank customers according to their probability of conversion and campaign engagement. Explainable AI, implemented through global and local SHAP explanations, allows the models' predictions to be translated into operational, partner-related, and behavioral components, making clear which factors consistently increase or reduce conversion probability. This transparency supports the interpretation of model outputs. However, the identified drivers differ between the post-approval phase and the cross-selling context.

In the approved auto-loan conversion case, booking outcomes are primarily associated with partner characteristics, operational timing, and communication availability, while customer demographics and financial attributes play a secondary role. The channel-specific analyses further show that dealer-originated applications follow more stable and homogeneous patterns, whereas broker-originated applications exhibit greater variability driven by inconsistent submission quality, timing delays, and documentation issues. These findings indicate that conversion losses after first approval are largely driven by process friction and institutional heterogeneity rather than applicant quality.

In contrast, the personal loan cross-sell analysis links engagement to behavioral history and external credit profiles. Key drivers include the number of active external credit relationships and prior contact frequency. The results show that repeated outreach correlates with declining responsiveness, consistent with campaign fatigue. This pattern is evident in the segmentation analysis: 'Fresh Prospects' exhibit initial responsiveness that decreases with frequency, whereas 'Tenured Customers' display saturation and low reaction to the marketing offer.

The findings suggest differentiated implications for the two stages of the customer journey. For auto-loan conversion, improvements are most likely to arise from process optimization, such as reducing early-stage delays, standardizing partner workflows, and ensuring reliable communication channels. For personal loan cross-selling, the findings suggest a propensity-based

tiering framework that integrates predictive scores with behavioral personas. This strategy can mitigate campaign fatigue by aligning contact intensity with the specific receptiveness of Fresh Prospects and Tenured Customers.

At the same time, this study highlights important limitations. The models identify associations rather than causal effects and cannot determine whether interventions will increase conversion or engagement, or which type of targeting strategies would be most effective. Predictive scores, therefore, should not be deployed as standalone targeting rules. Instead, they should serve as a foundation for controlled experiments, such as A/B testing alternative contact frequencies, channels, or process interventions across customer segments. Moreover, explainability methods such as SHAP are sensitive to modeling choices and background data, underscoring the need for robust governance, monitoring, and validation in deployment.

This thesis demonstrates that combining machine learning with explainable AI yields actionable insights for loan conversion and cross-selling. This study establishes a reusable framework that translates predictive outputs into operational decision support. Future research could enhance this approach by integrating causal testing, cost–benefit analysis, and long-term customer value.

8 References

Abbaszade, Aytaj. 2021. “The Role of Repitition (Frequency of Exposure) in Advertising. How Does It Influence Consumer Responses? What Kind of Factors Moderate the Effects of Repition?” Master Thesis, Middle East Technical University. https://open.metu.edu.tr/handle/11511/91111.

Abdelhamid, Mohamed, and Abhyuday Desai. 2024. “Balancing the Scales: A Comprehensive Study on Tackling Class Imbalance in Binary Classification.” arXiv:2409.19751. Preprint, arXiv. https://doi.org/10.48550/arXiv.2409.19751.

Abdullah, Mohammad Shoeb, and Reduanul Hasan. 2023. “AI-Driven Insights for Product Marketing: Enhancing Customer Experience And Refining Market Segmentation.” *American Journal of Interdisciplinary Studies* 4 (04): 80–116. https://doi.org/10.63125/pzd8m844.

Acharya, Deepak Bhaskar, B. Divya, and Karthigeyan Kuppan. 2024. “Explainable and Fair AI: Balancing Performance in Financial and Real Estate Machine Learning Models.” *IEEE Access* 12: 154022–34. https://doi.org/10.1109/ACCESS.2024.3484409.

Acock, Alan C. 2005. “Working With Missing Values.” *Journal of Marriage and Family* 67 (4): 1012–28. https://doi.org/10.1111/j.1741-3737.2005.00191.x.

Agarwal, Sumit, and Itzhak Ben-David. 2018. “Loan Prospecting and the Loss of Soft Information.” *Journal of Financial Economics* 129 (3): 608–28. https://doi.org/10.1016/j.jfineco.2018.05.003.

Akça, Mehmet, and Onur Sevli. 2022. “Predicting Acceptance of the Bank Loan Offers by Using Support Vector Machines.” *International Advanced Researches and Engineering Journal* 6: 142–47. https://doi.org/10.35860/iarej.1058724.

Alamsyah, Nur, Budiman Budiman, Titan Parama Yoga, and R Alamsyah. 2024. “XGBoost Hyperparameter Optimization Using randomizedsearchCV for Accurate Forest Fire Drought Condition.” *Jurnal Pilar Nusa Mandiri* 20 (2): 103–10. https://doi.org/10.33480/pilar.v20i2.5569.

Allen, Jason, Robert Clark, Jean-François Houde, Shaoteng Li, and Anna Trubnikova. 2025. “The Role of Intermediaries in Selection Markets: Evidence from Mortgage Lending.” *The Review of Financial Studies* 38 (11): 3284–328. https://doi.org/10.1093/rfs/hhae075.

Ansari, Asim, and Carl F Mela. 2003. “E-Customization.” *Journal of Marketing Research* 40 (2): 131–45. https://doi.org/10.1509/jmkr.40.2.131.19224.

Argyle, B., Nadauld T., and C. Palmer. 2020. “Monthly Payment Targeting and the Demand for Maturity.” *The Review of Financial Studies* 33 (11): 5416–62. https://doi.org/10.1093/rfs/hhaa004.

Awanife Oghenemaro, Stephen. 2025. “Optimizing Customer Lifetime Value (CLV) Prediction Models in Retail Banking Using Deep Learning and Behavioral Segmentation.” *American Journal of Humanities and Social Sciences Research (AJHSSR)* 9 (7): 123–31.

Baier, Daniel, and Björn Stöcker. 2022. “Profit Uplift Modeling for Direct Marketing Campaigns: Approaches and Applications for Online Shops.” *Journal of Business Economics* 92 (4): 645–73. https://doi.org/10.1007/s11573-021-01068-3.

Banco de Portugal. 2025. *Relatório de Acompanhamento Dos Mercados de Crédito | 2024*. Banco de Portugal. https://www.bportugal.pt/sites/default/files/documents/2025-07/ramc2024.pdf.

Basten, Christoph, and Ragnar Juelsrud. 2024. “Monetary Policy Transmission Through Cross-Selling Banks.” *Swiss Finance Institute Research Paper Series* 24–36. https://doi.org/10.2139/ssrn.4886084.

Bekkar, Mohamed, Hassiba Djema, and T.A. Alitouche. 2013. “Evaluation Measures for Models Assessment over Imbalanced Data Sets.” *Journal of Information Engineering and Applications* 3 (10): 27–38.

Benhabib, Jess, and Alberto Bisin. 2018. “Skewed Wealth Distributions: Theory and Empirics.” *Journal of Economic Literature* 56 (4): 1261–91. https://doi.org/10.1257/jel.20161390.

Bergmeir, Christoph, and José M. Benítez. 2012. “On the Use of Cross-Validation for Time Series Predictor Evaluation.” *Information Sciences* 191: 192–213. https://doi.org/10.1016/j.ins.2011.12.028.

Blattberg, Robert C., Pyŏng-do Kim, and Scott A. Neslin. 2008. *Database Marketing: Analyzing and Managing Customers*. International Series in Quantitative Marketing 18. Springer.

Bondarenko, Yana. 2023. “Continuous Monitoring of Data in Online Randomized Experiments.” *2023 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT)*, November 20, 70–77. https://doi.org/10.1109/3ICT60104.2023.10391343.

Boston Consulting Group and Google. 2020. *Responsible Marketing with First-Party Data*. Industry Report. Boston Consulting Group (BCG) and Google. https://www.bcg.com/publications/2020/responsible-marketing-with-first-party-data.

Bouke, Mohamed Aly, Saleh Ali Zaid, and Azizol Abdullah. 2024. “Implications of Data Leakage in Machine Learning Preprocessing: A Multi-Domain Investigation.” Preprint. https://doi.org/10.21203/rs.3.rs-4579465/v1.

Boustani, Noureddine, Ali Emrouznejad, Roya Gholami, Ozren Despic, and Athina Ioannou. 2024. “Improving the Predictive Accuracy of the Cross-Selling of Consumer Loans Using Deep Learning Networks.” *Annals of Operations Research* 339 (1): 613–30. https://doi.org/10.1007/s10479-023-05209-5.

Brahma, Arin, David M. Goldberg, Nohel Zaman, and Mariano Aloiso. 2021. “Automated Mortgage Origination Delay Detection from Textual Conversations.” *Decision Support Systems* 140: 113433. https://doi.org/10.1016/j.dss.2020.113433.

Branco, Moisés Castelo, Yingfei Xiong, Krzysztof Czarnecki, Jochen Küster, and Hagen Völzer. 2014. “A Case Study on Consistency Management of Business and IT Process Models in Banking.” *Software & Systems Modeling* 13 (3): 913–40. https://doi.org/10.1007/s10270-013-0318-8.

Breiman, Leo. 2001. “Random Forests.” *Machine Learning* 45: 5–32. https://doi.org/10.1023/A:1010933404324.

Breskuvienė, Dalia, and Gintautas Dzemyda. 2023. “Categorical Feature Encoding Techniques for Improved Classifier Performance When Dealing with Imbalanced Data of Fraudulent Transactions.” *International Journal of Computers Communications & Control* 18 (3). https://doi.org/10.15837/ijccc.2023.3.5433.

Bücker, Michael, Gero Szepannek, Alicja Gosiewska, and Przemyslaw Biecek. 2020. “Transparency, Auditability and Explainability of Machine Learning Models in Credit Scoring.” *arXiv Preprint arXiv:2009.13384*, ahead of print. https://doi.org/10.48550/arXiv.2009.13384.

Buddiga, Sai Kalyana Pranitha. 2022. “Improving Targeted Client Acquisition: Predictive Analysis of Retail Bank Direct Marketing Campaigns.” *Journal of Technological Innovations* 3 (2). https://jtipublishing.com/jti/article/view/49/256.

Bugajev, Andrej, Rima Kriauzienė, and Viktoras Chadyšas. 2025. “Realistic Data Delays and Alternative Inactivity Definitions in Telecom Churn: Investigating Concept Drift Using a Sliding-Window Approach.” *Applied Sciences* 15 (3): 1599. https://doi.org/10.3390/app15031599.

Bulut, Okan, Bin Tan, Elisabetta Mazzullo, and Ali Syed. 2025. “Benchmarking Variants of Recursive Feature Elimination: Insights from Predictive Tasks in Education and Healthcare.” *Information* 16 (6): 476. https://doi.org/10.3390/info16060476.

Cao, Longbing, and Chengzhang Zhu. 2022. “Personalized Next-Best Action Recommendation with Multi-Party Interaction Learning for Automated Decision-Making.” *Plos One* 17 (1): e0263010. https://doi.org/10.48550/arXiv.2108.08846.

Černevičienė, Jurgita, and Audrius Kabašinskas. 2024. “Explainable Artificial Intelligence (XAI) in Finance: A Systematic Literature Review.” *Artificial Intelligence Review* 57 (216). https://doi.org/10.1007/s10462-024-10854-8.

Chen, Tianqi, and Carlos Guestrin. 2016. “XGBoost: A Scalable Tree Boosting System.” *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery*

and Data Mining (New York, NY, USA), 785–94. https://doi.org/10.1145/2939672.2939785.

Cooil, Bruce, Timothy L. Keiningham, Lerzan Aksoy, and Michael Hsu. 2007. "A Longitudinal Analysis of Customer Satisfaction and Share of Wallet: Investigating the Moderating Effect of Customer Characteristics." *Journal of Marketing* 71 (1): 67–83. https://doi.org/10.1509/jmkg.71.1.067.

Cowan, Greig, Salvatore Mercuri, and Raad Khraishi. 2023. "Modelling Customer Lifetime-Value in the Retail Banking Industry." *arXiv Preprint arXiv:2304.03038*, ahead of print. https://doi.org/10.48550/arXiv.2304.03038.

Davenport, Thomas, Abhijit Guha, Dhruv Grewal, and Timna Bressgott. 2020. "How Artificial Intelligence Will Change the Future of Marketing." *Journal of the Academy of Marketing Science* 48 (1): 24–42. https://doi.org/10.1007/s11747-019-00696-0.

Davis, Jesse, and Mark Goadrich. 2006. "The Relationship Between Precision-Recall and ROC Curves." 06. https://doi.org/10.1145/1143844.1143874.

Deloitte. 2022. *Unleashing the Power of Machine Learning Models in Banking through Explainable Artificial Intelligence (XAI)*. https://www.deloitte.com/us/en/insights/industry/financial-services/explainable-ai-in-banking.html.

Deloitte. 2025. "How to Streamline the Credit Journey." https://www.deloitte.com/lu/en/our-thinking/future-of-advice/how-to-streamline-the-credit-journey.html.

Demir, Selçuk, and Emrehan Kutlug Sahin. 2023. "Application of State-of-the-Art Machine Learning Algorithms for Slope Stability Prediction by Handling Outliers of the Dataset." *Earth Science Informatics* 16 (3): 2497–509. https://doi.org/10.1007/s12145-023-01059-8.

Deng, Wei, Yong Shi, Zhengxin Chen, Wikil Kwak, and Huimin Tang. 2020. "Recommender System for Marketing Optimization." *World Wide Web* 23 (3): 1497–517. https://doi.org/10.1007/s11280-019-00738-1.

Dieber, Jürgen, and Sabrina Kirrane. 2020. "Why Model Why? Assessing the Strengths and Limitations of LIME." *arXiv Preprint arXiv:2012.00093*. https://arxiv.org/abs/2012.00093.

Dinh, Tai, Hauchi Wong, Philippe Fournier-Viger, et al. 2025. "Categorical Data Clustering: 25 Years beyond K-Modes." *Expert Systems with Applications* 272: 126608. https://doi.org/10.1016/j.eswa.2025.126608.

Directorate-General for Parliamentary Research Services (European Parliament). 2020. *The Impact of the General Data Protection Regulation (GDPR) on Artificial Intelligence*. Panel for the Future of Science and Technology. European Union. https://doi.org/10.2861/293.

Doria, Margherita, Elisa Luciano, and Patrizia Semeraro. 2023. “Machine Learning Techniques in Joint Default Assessment.” arXiv:2205.01524. Preprint, arXiv. https://doi.org/10.48550/arXiv.2205.01524.

Eisend, Martin, and Susanne Schmidt. 2025. “Advertising Repetition: A Meta-Analysis on Effective Frequency in Advertising.” *Journal of Advertising* 44 (4): 415–28. https://doi.org/10.1080/00913367.2015.1018460.

European Central Bank. 2025a. *AI’s Impact on Banking: Use Cases for Credit Scoring and Fraud Detection*. November 20. https://www.bankingsupervision.europa.eu/press/supervisory-newsletters/newsletter/2025/html/ssm.nl251120_1.en.html.

European Central Bank. 2025b. “Euribor 6-Month - Historical Close, Average of Observations through Period, Euro Area, Monthly.” https://data.ecb.europa.eu/data/datasets/FM/FM.M.U2.EUR.RT.MM.EURIBOR6MD_.HSTA.

European Data Protection Supervisor. 2025. “AI Act Regulation (EU) 2024/1689 – Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying down Harmonised Rules on Artificial Intelligence and Amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA Relevance).” https://doi.org/10.2804/4225375.

European Union. 2016. “Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA Relevance).” May 4. http://data.europa.eu/eli/reg/2016/679/oj.

Feng, Junying, Yuan Zhu, Hongyu Pan, and Yingjie Mou. 2025. “Research on Financial Data Risk Prediction Models Based on XGBoost Algorithm.” *Proceedings of the 2nd Guangdong-Hong Kong-Macao Greater Bay Area International Conference on Digital Economy and Artificial Intelligence*, 686–90. https://doi.org/10.1145/3745238.3745344.

García-Céspedes, Rubén, Francisco J. Alias-Carrascosa, and Manuel Moreno. 2025. “On Machine Learning Models Explainability in the Banking Sector: The Case of SHAP.” *Journal of the Operational Research Society*, 1–13. https://doi.org/10.1080/01605682.2025.2485263.

Ghanem, Elsayed, Armin Hatefi, and Hamid Usefi. 2022. “Liu-Type Shrinkage Estimators for Mixture of Logistic Regressions: An Osteoporosis Study.” Version 2. Preprint, arXiv. https://doi.org/10.48550/ARXIV.2209.01731.

Ghonim, Adel, and Abdelrehim Awad. 2025. “Shaping Customer Insights through E-Marketing: An Applied Study in the Banking Sector.” *International Journal of Advanced and Applied Sciences* 12 (4): 88–98. https://doi.org/10.21833/ijaas.2025.04.011.

Giannella, Eric, Tatiana Homonoff, Gwen Rino, and Jason Somerville. 2023. *Administrative Burden and Procedural Denials: Experimental Evidence from SNAP*. 16 (4): 316–40. https://doi.org/10.1257/pol.20220701.

Gubela, Robin, Artem Bequé, Stefan Lessmann, and Fabian Gebert. 2019. "Conversion Uplift in E-Commerce: A Systematic Benchmark of Modeling Strategies." *International Journal of Information Technology & Decision Making* 18 (03): 747–91. https://doi.org/10.1142/S0219622019500172.

Gubela, Robin M, Stefan Lessmann, and Björn Stöcker. 2024. "Multiple Treatment Modeling for Target Marketing Campaigns: A Large-Scale Benchmark Study." *Information Systems Frontiers* 26 (3): 875–98. https://doi.org/10.1007/s10796-022-10283-4.

Guo, Rui, and Zhengrui Jiang. 2023. "Optimal Dynamic Advertising Policy Considering Consumer Ad Fatigue." SSRN Scholarly Paper No. 4664196. Rochester, NY. https://doi.org/10.2139/ssrn.4664196.

Gupta, Sunil, Dominique Hanssens, Bruce Hardie, et al. 2006. "Modeling Customer Lifetime Value." *Journal of Service Research* 9 (2): 139–55. https://doi.org/10.1177/1094670506293810.

Hangasjärvi, Aleksi Johannes. 2024. "Influencing Consumer Response in Email Marketing." Master's Thesis, Hanken School of Economics, Department of Marketing. https://helda.helsinki.fi/server/api/core/bitstreams/4e31e1c4-4554-4f52-a40e-f966870e7783/content.

Heckman, James J. 1979. "Sample Selection Bias as a Specification Error." *Econometrica* 47 (1): 153–61. https://doi.org/10.2307/1912352.

Hu, Tianzuo. 2025. "Financial Fraud Detection System Based on Improved Random Forest and Gradient Boosting Machine (GBM)." Preprint, arXiv. https://doi.org/10.48550/ARXIV.2502.15822.

Huang, Hui-I., Chou-Wen Wang, and Chin-Wen Wu. 2024. "Predictive Analysis for Personal Loans by Using Machine Learning." *HCI in Business, Government and Organizations*, 187–99. https://doi.org/10.1007/978-3-031-61315-9_13.

Hurlin, Christophe, Christophe Perignon, and Sébastien Saurin. 2024. "The Fairness of Credit Scoring Models." *SSRN Electronic Journal*, ahead of print. https://doi.org/10.2139/ssrn.3785882.

Instituto Nacional de Estatística – Portugal. 2021. "Statistics Portugal - Publication 439549749." https://www.ine.pt/xportal/xmain?xpid=INE&xpgid=ine_publicacoes&PUBLICACOESpub_boui=439549749&PUBLICACOESmodo=2.

Ionescu, Adela. 2014. "Loan Brokers." *Challenges of the Knowledge Society* (Bucharest, Romania), 670–75.

James, Gareth, Daniela Witten, Trevor Hastie, Robert Tibshirani, and Jonathan E. Taylor. 2023. *An Introduction to Statistical Learning: With Applications in Python*. Springer Texts in Statistics. Springer.

Johnson, Justin M., and Taghi M. Khoshgoftaar. 2021. "Encoding Techniques for High-Cardinality Features and Ensemble Learners." *2021 IEEE 22nd International Conference on Information Reuse and Integration for Data Science (IRI)*, 355–61. https://doi.org/10.1109/IRI51335.2021.00055.

Kannan, P. K., and Hongshuang "Alice" Li. 2017. "Digital Marketing: A Framework, Review and Research Agenda." *International Journal of Research in Marketing* 34 (1): 22–45. https://doi.org/10.1016/j.ijresmar.2016.11.006.

Khandani, Amir E., Adlar J. Kim, and Andrew W. Lo. 2010. "Consumer Credit-Risk Models via Machine-Learning Algorithms." *Journal of Banking & Finance* 34 (11): 2767–87. https://doi.org/10.1016/j.jbankfin.2010.06.001.

Kim, Kyoung-jae. 2003. "Financial Time Series Forecasting Using Support Vector Machines." *Neurocomputing* 55 (1–2): 307–19. https://doi.org/10.1016/S0925-2312(03)00372-2.

Klein, Tony, and Thomas Walther. 2024. "Advances in Explainable Artificial Intelligence (xAI) in Finance." *Finance Research Letters* 70: 106358. https://doi.org/10.1016/j.frl.2024.106358.

KPMG International. 2025. *Banking Transformation: The New Agenda.* https://kpmg.com/xx/en/our-insights/transformation/operational-and-cost-transformation-in-banking.html.

Kranzbühler, Anne-Madeleine, Mirella H.P. Kleijnen, Robert E. Morgan, and Marije Teerling. 2018. "The Multilevel Nature of Customer Experience Research: An Integrative Review and Research Agenda." *International Journal of Management Reviews* 20 (2): 433–56. https://doi.org/10.1111/ijmr.12140.

Kuhn, Max, and Kjell Johnson. 2013. *Applied Predictive Modeling*. Vol. 26. Springer New York. https://doi.org/10.1007/978-1-4614-6849-3.

Kumar, V. 2018. "A Theory of Customer Valuation: Concepts, Metrics, Strategy, and Implementation." *Journal of Marketing* 82 (1): 1–19. https://doi.org/10.1509/jm.17.0208.

Lee, Kyungsik, Hana Yoo, Sumin Shin, et al. 2024. "A Submodular Optimization Approach to Accountable Loan Approval." *Proceedings of the AAAI Conference on Artificial Intelligence* 38 (21): 22761–69. https://doi.org/10.1609/aaai.v38i21.30310.

Lemon, Katherine N, and Peter C Verhoef. 2016. "Understanding Customer Experience throughout the Customer Journey." *Journal of Marketing* 80 (6): 69–96. https://doi.org/10.1509/jm.15.0420.

Lessmann, Stefan, Bart Baesens, Hsin-Vonn Seow, and Lyn C. Thomas. 2015. "Benchmarking State-of-the-Art Classification Algorithms for Credit Scoring: An Update of Research." *European Journal of Operational Research* 247 (1): 124–36. https://doi.org/10.1016/j.ejor.2015.05.030.

Ligaraba, Neo, Tinashe Chuchu, and Brighton Nyagadza. 2023. "Opt-in e-Mail Marketing Influence on Consumer Behaviour: A Stimuli–Organism–Response (S–O–R) Theory Perspective." *Cogent Business & Management* 10 (1): 2184244.

Lin, Su-Yin. 2012. "Customer Orientation and Cross-Buying: The Mediating Effects of Relational Selling Behavior and Relationship Quality." *Journal of Management Research* 4 (4): 334–58. https://doi.org/10.5296/jmr.v4i4.2350.

Low, David, Jonathan Lanning, Tobias Salz, and Andreas Grunewald. 2020. "Auto Dealer Loan Intermediation: Consumer Behavior and Competitive Effects." SSRN Scholarly Paper No. 3568571. Social Science Research Network, April 2. https://doi.org/10.2139/ssrn.3568571.

Lundberg, Scott, and Su-In Lee. 2017. "A Unified Approach to Interpreting Model Predictions." Preprint, Advances in neural information processing systems 30. https://doi.org/10.48550/arXiv.1705.07874.

Lyu, X., Y. Huang, X. Liu, J. Wang, and J. Han. 2025. "Comparative Analysis of Machine Learning Models for Targeting Bank Loan Customers: A Case Study of Neobank." *Advances in Economics, Management and Political Sciences* 206 (1): 108–21. https://doi.org/10.54254/2754-1169/2025.LD25705.

Maggi, Fabrizio Maria, Chiara Di Francescomarino, Marlon Dumas, and Chiara Ghidini. 2014. "Predictive Monitoring of Business Processes." In *Advanced Information Systems Engineering*, edited by David Hutchison, Takeo Kanade, Josef Kittler, et al., vol. 8484, edited by Matthias Jarke, John Mylopoulos, Christoph Quix, et al. Lecture Notes in Computer Science. Springer International Publishing. https://doi.org/10.1007/978-3-319-07881-6_31.

Majeske, Karl D, and Thomas W Lauer. 2013. "The Bank Loan Approval Decision from Multiple Perspectives." *Expert Systems with Applications* 40 (5): 1591–98. https://doi.org/10.1016/j.eswa.2012.09.001.

Majka, Marcin. 2024. *Understanding Customer Acquisition Cost*. July 22.

Márquez-Chamorro, Alfonso E., Isabel A. Nepomuceno-Chamorro, Manuel Resinas, and Antonio Ruiz-Cortés. 2022. "Updating Prediction Models for Predictive Process Monitoring." In *Advanced Information Systems Engineering*, edited by Xavier Franch, Geert Poels, Frederik Gailly, and Monique Snoeck. Springer International Publishing. https://doi.org/10.1007/978-3-031-07472-1_18.

McCabe, Connor J., Max A. Halvorson, Kevin M. King, Xiaolin Cao, and Dale S. Kim. 2022. "Interpreting Interaction Effects in Generalized Linear Models of Nonlinear Probabilities and Counts." *Multivariate Behavioral Research* 57 (2–3): 243–63. https://doi.org/10.1080/00273171.2020.1868966.

McKinsey & Company. 2025. *Riding the Storm: How Consumer Finance Companies Can Survive and Thrive*. https://www.mckinsey.com/industries/financial-services/our-insights/riding-the-storm-how-consumer-finance-companies-can-survive-and-thrive#/.

McKnight, Patrick E., and Julius Najab. 2010. "Mann-Whitney U Test." In *The Corsini Encyclopedia of Psychology*. John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470479216.corpsy0524.

Méndez-Suárez, Mariano, and Natividad Crespo-Tejero. 2021. "Why Do Banks Retain Unprofitable Customers? A Customer Lifetime Value Real Options Approach." *Journal of Business Research* 122: 621–26. https://doi.org/10.1016/j.jbusres.2020.10.008.

Min, Sungwook, Xubing Zhang, Namwoon Kim, and Rajendra K Srivastava. 2016. "Customer Acquisition and Retention Spending: An Analytical Model and Empirical Investigation in Wireless Telecommunications Markets." *Journal of Marketing Research* 53 (5): 728–44. https://doi.org/10.1509/jmr.14.0170.

Moore, Lynne, André Lavoie, Eric Bergeron, and Marcel Emond. 2007. "Modeling Probability-Based Injury Severity Scores in Logistic Regression Models: The Logit Transformation Should Be Used." *Journal of Trauma and Acute Care Surgery* 62 (3): 601. https://doi.org/10.1097/TA.0b013e31803245b2.

Moro, Sérgio, Paulo Cortez, and Paulo Rita. 2014. "A Data-Driven Approach to Predict the Success of Bank Telemarketing." *Decision Support Systems* 62: 22–31. https://doi.org/10.1016/j.dss.2014.03.001.

Munoz-Cancino, Ricardo, Cristian Bravo, Sebastian A. Rıos, and Manuel Grana. 2022. *On the Dynamics of Credit History and Social Interaction Features, and Their Impact on Creditworthiness Assessment Performance*. arXiv:2204.06122. https://arxiv.org/abs/2204.06122.

Naser, M.Z. 2025. "Causality, Explanations, Machine Learning, and Engineering." *Foundations of Science* 30 (4): 945–70. https://doi.org/10.1007/s10699-025-10006-3.

Nasir, Fahim, Abdulghani Ahmed, Mehmet Kiraz, Iryna Yevseyeva, and Mubarak Saif. 2024. "Data-Driven Decision-Making for Bank Target Marketing Using Supervised Learning Classifiers on Imbalanced Big Data." *Computers, Materials & Continua* 81 (1): 1703–28. https://doi.org/10.32604/cmc.2024.055192.

Neslin, Scott A, Sunil Gupta, Wagner Kamakura, Junxiang Lu, and Charlotte H Mason. 2006. "Defection Detection: Measuring and Understanding the Predictive Accuracy of Customer Churn Models." *Journal of Marketing Research* 43 (2): 204–11.

Padigar, Manjunath, Yi Li, and Chandana N. Manjunath. 2025. "'Good' and 'Bad' Frictions in Customer Experience: Conceptual Foundations and Implications." *Psychology & Marketing* 42 (1): 21–43. https://doi.org/10.1002/mar.22111.

Pargent, Florian, Florian Pfisterer, Janek Thomas, and Bernd Bischl. 2022. "Regularized Target Encoding Outperforms Traditional Methods in Supervised Machine Learning with High Cardinality Features." *Computational Statistics* 37 (5): 2671–92. https://doi.org/10.1007/s00180-022-01207-6.

Pavón Pérez, Ángel. 2025. "Enhancing Fairness in Machine Learning: Identifying and Mitigating Bias with a Focus on Gender Bias in Finance." Phd, The Open University. https://oro.open.ac.uk/102712/.

Powers, David M. W. 2011. "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness and Correlation." *Journal of Machine Learning Technologies* 2 (1): 37–63.

Radcliffe, Nicholas J, and Patrick D Surry. 2011. "Real-World Uplift Modelling with Significance-Based Uplift Trees." *White Paper TR-2011-1, Stochastic Solutions*, 1–33.

Rainio, Oona, Jarmo Teuho, and Riku Klén. 2024. "Evaluation Metrics and Statistical Tests for Machine Learning." *Scientific Reports* 14 (1): 6086. https://doi.org/10.1038/s41598-024-56706-x.

Rawson, Alex, Ewan Duncan, and Conor Jones. 2013. "The Truth about Customer Experience." *Harvard Business Review* 91 (9): 90–98.

Reinartz, Werner J, and Vita Kumar. 2003. "The Impact of Customer Relationship Characteristics on Profitable Lifetime Duration." *Journal of Marketing* 67 (1): 77–99. https://doi.org/10.1509/jmkg.67.1.77.18589.

Reinartz, Werner, Jacquelyn S Thomas, and Vita Kumar. 2005. "Balancing Acquisition and Retention Resources to Maximize Customer Profitability." *Journal of Marketing* 69 (1): 63–79. https://doi.org/10.1509/jmkg.69.1.63.55511.

Reutterer, Thomas, and Daniel Dan. 2019. *Cluster Analysis in Marketing Research*. https://doi.org/10.1007/978-3-319-05542-8_11-1.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. 2016. "' Why Should i Trust You?' Explaining the Predictions of Any Classifier." *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–44. https://doi.org/10.1145/2939672.2939778.

Rod, Michel, Nicholas J. Ashill, and Tanya Gibbs. 2016. “Customer Perceptions of Frontline Employee Service Delivery: A Study of Russian Bank Customer Satisfaction and Behavioural Intentions.” *Journal of Retailing and Consumer Services* 30: 212–21. https://doi.org/10.1016/j.jretconser.2016.02.005.

Romero Leguina, Jesús, Ángel Cuevas Rumin, and Rubén Cuevas Rumin. 2021. “Optimizing the Frequency Capping: A Robust and Reliable Methodology to Define the Number of Ads to Maximize ROAS.” *Applied Sciences* 11 (15): 6688. https://doi.org/10.3390/app11156688.

Sadhwani, Apaar, Justin Sirignano, and Kay Giesecke. 2021. “Deep Learning for Mortgage Risk.” *Journal of Financial Econometrics* 19 (2): 313–68. https://doi.org/10.1093/jjfinec/nbaa025.

Schreiber-Gregory, Deanna N., Jackson Foundation, and Karlen Bader. 2018. “Regulation Techniques for Multicollinearity: Lasso, Ridge, and Elastic Nets.” https://api.semanticscholar.org/CorpusID:189925961.

Schwartz, Sagi, Qinling Wang, and Fang Fang. 2025. “Enhancing ML Models Interpretability for Credit Scoring.” arXiv:2509.11389. Preprint, arXiv. https://doi.org/10.48550/arXiv.2509.11389.

Sheridan, Robert P. 2013. “Time-Split Cross-Validation as a Method for Estimating the Goodness of Prospective Prediction.” *Journal of Chemical Information and Modeling* 53 (4): 783–90. https://doi.org/10.1021/ci400084k.

Shiwa, Sayed Ebrahim, S Prasad Babu Vagolu, and Vedavathi Katneni. 2021. *Loan Approval Prediction Using Logistic Regression Model in Machine Learning*. 05 (05). https://www.academia.edu/77023341/Loan_Approval_Prediction_Using_Logistic_Regression_Model_in_Machine_Learning.

Simonson, Itamar, and Amos Tversky. 1992. “Choice in Context: Tradeoff Contrast and Extremeness Aversion.” *Journal of Marketing Research* 29 (3): 281–95. JSTOR. https://doi.org/10.2307/3172740.

Špicas, Renatas, Airidas Neifaltas, Rasa Kanapickienė, Greta Keliuotytė-Staniulėnienė, and Deimantė Vasiliauskaitė. 2023. “Estimating the Acceptance Probabilities of Consumer Loan Offers in an Online Loan Comparison and Brokerage Platform.” *Risks* 11 (7): 138. https://doi.org/10.3390/risks11070138.

Thaler, R. 1985. “Mental Accounting and Consumer Choice.” *Marketing Science* (Marketing Science) 4 (3): 199–214.

Thomas, Lyn C, Ki Mun Jung, Steve D Thomas, and Y Wu. 2006. “Modeling Consumer Acceptance Probabilities.” *Expert Systems with Applications* 30 (3): 499–506. https://doi.org/10.1016/j.eswa.2005.10.011.

Tomaschek, Fabian, Peter Hendrix, and R. Harald Baayen. 2018. “Strategies for Addressing Collinearity in Multivariate Linguistic Data.” *Journal of Phonetics* 71: 249–67. https://doi.org/10.1016/j.wocn.2018.09.004.

Tran, Huong Thi Thanh, and James Corner. 2016. “The Impact of Communication Channels on Mobile Banking Adoption.” *International Journal of Bank Marketing* 34 (1): 78–109. https://doi.org/10.1108/IJBM-06-2014-0073.

Vanhoeyveld, Jellis, David Martens, and Bruno Peeters. 2020. “Customs Fraud Detection.” *Pattern Analysis and Applications* 23 (3): 1457–77. https://doi.org/10.1007/s10044-019-00852-w.

Varma, Sudhir, and Richard Simon. 2006. “Bias in Error Estimation When Using Cross-Validation for Model Selection.” *BMC Bioinformatics* 7 (91). https://doi.org/10.1186/1471-2105-7-91.

Venkatesan, Rajkumar. 2004. “A Customer Lifetime Value Framework for Customer Selection and Resource Allocation Strategy.” *Journal of Marketing* 68 (4): 106–25. https://doi.org/10.1509/jmkg.68.4.106.42728.

Verhoef, Peter C., P.K. Kannan, and J. Jeffrey Inman. 2015. “From Multi-Channel Retailing to Omni-Channel Retailing.” *Journal of Retailing* 91 (2): 174–81. https://doi.org/10.1016/j.jretai.2015.02.005.

Wang, Huan. 2025. “Application of Decision Tree Model in Personal Credit Scoring and Its Fairness Optimization.” *Advances in Economics, Management and Political Sciences* 176 (1): 109–18. https://doi.org/10.54254/2754-1169/2025.22114.

Weinzierl, Sven, Sebastian Dunzer, Sandra Zilker, and Martin Matzner. 2020. “Prescriptive Business Process Monitoring for Recommending next Best Actions.” *Business Process Management Forum*, 193–209. https://doi.org/10.1007/978-3-030-58638-6_12.

Wei-Shong, Lin Peter, and Mei Albert Kuo-Chung. 2006. “The Internal Performance Measures of Bank Lending: A Value-added Approach.” *Benchmarking: An International Journal* 13 (3): 272–89. https://doi.org/10.1108/14635770610668785.

White. 1988. “Economic Prediction Using Neural Networks: The Case of IBM Daily Stock Returns.” *IEEE International Conference on Neural Networks* vol.2: 451–58. https://doi.org/10.1109/ICNN.1988.23959.

Wonder, Nicholas, Wendy Wilhelm, and David Fewings. 2008. “The Financial Rationality of Consumer Loan Choices: Revealed Preferences Concerning Interest Rates, down Payments, Contract Length, and Rebates.” *Journal of Consumer Affairs* 42 (2): 243–70.

Yang, Xiang, Yongbin Zhu, Li Yan, and Xin Wang. 2015. “Credit Risk Model Based on Logistic Regression and Weight of Evidence.” *Proceedings of the 2015 3rd International Conference on Management Science, Education Technology, Arts, Social Science and Economics* (Qingdao, China). https://doi.org/10.2991/msetasse-15.2015.180.

Yang, Yujia. 2025. “A Study on Bank Marketing Prediction Based on XGBoost.” *Advances in Economics, Management and Political Sciences* 185: 127–34. https://doi.org/10.54254/2754-1169/2025.LH23949.

Zheng, Yuanzi. 2025. “Prediction of the Effectiveness of Bank Marketing Strategies Using the XGBoost Model.” *Advances in Economics, Management and Political Sciences* 170 (1): 17–28. https://doi.org/10.54254/2754-1169/2025.LH23975.

9 Appendix

9.1 List of Figures

9.2 List of Tables

9.3 Glossary

AI Governance Framework	Ensures AI models are transparent, fair, and compliant with regulations.
Black-Box Model	Model with high accuracy but low interpretability.
Customer Acquisition Cost (CAC)	Cost of acquiring one new customer.
Customer Acquisition Optimization	Increasing conversions while reducing acquisition costs.
Customer Development Optimization	Growing existing customer value through cross-selling.
Customer Journey	All interactions and experiences a customer has with a company.
Customer Lifetime Value (CLV)	Total profit expected from a customer over their relationship.
Cross-Selling	Selling additional products to existing customers.
Data Imbalance	Unequal class distribution affecting model training.
Decision Tree (DT)	Tree-structured model using rules to classify outcomes.
Explainable AI (XAI)	Makes AI model decisions understandable.
Extreme Gradient Boosting (XGBoost)	High-performance ensemble model using sequential tree boosting.
k-Nearest Neighbors (kNN)	Classifies data based on the closest examples in the dataset.
LIME (Local Interpretable Model-agnostic Explanations)	Explains individual model predictions locally.
Loan Offer Acceptance Prediction	Modeling customer probability to accept an approved loan.

Logistic Regression (LR)	Linear model estimating the probability of a binary outcome.
Machine Learning (ML)	Algorithms that learn from data to make predictions.
Neural Network (NN)	Model inspired by the human brain, effective for complex, non-linear data.
Next Best Action (NBA)	Framework recommending the optimal action for each customer.
Pain Point Detection	Identifying friction points in the customer journey.
Propensity Modeling	Predicting the probability of a customer taking a specific action.
Random Forest (RF)	Ensemble of decision trees improving accuracy and robustness.
SHAP (SHapley Additive exPlanations)	Quantifies each feature's contribution to model predictions.
SMOTE (Synthetic Minority Oversampling Technique)	Balances imbalanced datasets by generating synthetic samples.
Support Vector Machine (SVM)	Classifies data by finding the optimal separating boundary.
Touchpoints	Individual customer–company interactions across channels.

9.4 Data Dictionary

9.4.1 Data Dictionary: Approved Loan Conversion

9.4.1.1 Data Dictionary: Approved Loan Conversion – Process Metadata

flow_item_id_masked	Unique proposal identifier (masked).
creation_date	Date proposal was created.
state_id	First approval state (4 = Approved / 7 = Approved with pending documents /12 = Approved with Conditions).
state	Final credit analysis state.
dt_approved	Date of first approval.
dt_state	Date last state was updated.
last_state	Final workflow state (e.g., booked, closed).
user_proposal_masked	Internal user who created/handled the proposal.
analyst_masked	Credit analyst responsible for evaluation.
user_conferencia_masked	User responsible for conference stage.
commercial_manager_masked	Commercial Manager in the bank.
reception_type	Channel used to submit the proposal.
conference	Whether proposal reached conference stage.
time_approve_since_received	Time from reception to approval.
time_book_since_approved	Time from approval to booking.
time_book_since_received	Time from reception to booking.
time_book_since_contract_received	Time from receiving contract to booking.
time_received	Time from “received” to “analysed” (in hours).
time_analysed	Timestamp/duration of analysis completion.
hh_time_approved_since_created	Time from creation to approval.

9.4.1.2 Data Dictionary: Approved Loan Conversion – Product Financing

product_group	Loan products princing scheme(Acordos, PM0–PM3, etc.).
financing_type	Type of credit (Credit, Moto, Leasing, ALD).
expenses_financed	Whether expenses are financed.
eligible_rappel	Eligible for dealer bonus.
tax_type	Applicable tax regime.
self_financing	Customer's own funds used.
was_conditional	Approved with conditions.
financed_amount	Loan amount.
ltv	Loan-to-value ratio.
tan	Nominal annual interest rate.
comissao_dealer	Dealer commission.
gross_margin	Gross margin for bank.
liquid_margin	Net margin.
psp	Price of the car.
residual_value	Contract residual value.
inicial_entry	Customer down payment.
installment	Monthly installment.
customer_effort_rate	Installment burden metric.
vehicle_age	Age of vehicle.
eurotax_quotation	Eurotax valuation.
term	Contract term in months.
maturity	Months left to maturity.
monthly_available_client	Monthly residual income after obligations.
risk_provision	Expected loss provision.
dealer_life_insurance_comission	Dealer commission for life insurance.

dealer_car_insurance_comission	Dealer commission for car insurance.
dealer_ppp_insurance_comission	Dealer commission for PPP insurance.
legal_expenses	Legal costs.
other_expenses	Other contract expenses.
taeg	Total annual cost of credit (APR).
max_liquid_income	Maximum liquid income considered.
income_verified	Verified income.
income_not_verified	Unverified income.
installment_pre_payment	Installment under pre-payment scenario.
fixed_fee	Fixed contract fee.
dsti	Debt service-to-income ratio.
expenses_financed_amount	Amount of expenses financed.

9.4.1.3 Data Dictionary: Approved Loan Conversion – Partner and Channel Information

partner_bill	Billing partner (dealer selling the vehicle).
partner_pricing	Partner pricing segment (0–3).
partner_id_masked	Partner intermediating the credit.
partner_bill_id_masked	Partner selling the car.
partner_commercial_manager_masked	Partner's commercial manager.
partner_risk	Current partner risk level (at dataset extraction)
partner_risk_surveillance	Partner under risk monitoring.
partner_type	Type of partner (dealer, broker).
dt_opening	Partner account opening date.
partner_state	Partner account status.
pricing	Pricing table for this deal (0-3).
partner_city	Partner city.
partner_district	Partner district.

partner_new_bet	Access to “New Bet” pricing conditions.
partner_rating	Internal performance rating.

9.4.1.4 Data Dictionary: Approved Loan Conversion – Customer Characteristics

client_age	Customer age.
nationality	Customer nationality.
client_mortgage	Mortgage customer of Banco Primus.
client_district	Customer residence district.
client_type	Individual or corporate customer.
client_effective	Active/effective customer flag.
client_self_employed	Self-employed flag.
client_contract_type	Employment contract type.
gender	Customer gender.
profession	Profession.
marital_status	Marital status.
cae	Employer business activity code.
activity	Employer activity description.
has_mail	Customer has email.
branch	Branch associated with application.

9.4.1.5 Data Dictionary: Approved Loan Conversion – Insurance and Incentives

has_insurance_car	Car insurance sold.
has_insurance_life	Life insurance sold.
has_insurance_ppp	PPP insurance sold.
insurance_ppp_telephone	PPP insurance phone contact.
insurance_life_telephone	Life insurance phone contact.
company	Car insurance company.

insurance_type	Car insurance type.
life_eligible	Proposal eligible for life insurance.
ppp_eligible	Eligible for PPP insurance.
rappel_forecast	Forecasted dealer rappel.
has_insurance_car	Car insurance sold.
has_insurance_life	Life insurance sold.
has_insurance_ppp	PPP insurance sold.
insurance_ppp_telephone	PPP insurance phone contact.
insurance_life_telephone	Life insurance phone contact.

9.4.1.6 Data Dictionary: Approved Loan Conversion – Decision Process

risk_agency	Partner risk at proposal submission (1 - Prime, 2 - Good, 8 - Good Less, 3 - Medium, 9 - Medium Less, 4 - Weak, 5 - Bad, 7 - New Partner).
risk_profile	Outdated risk field.
has_risk_surveillance	Proposal under risk surveillance.
checkdl133	Proposal under DL133 consumer credit law.
fraud	Suspicious fraud leading to refusal.
risk_type	Corporate risk classification (-99 - no classification, 0 and 5 - customer changed to individual and not company).
approval_level	Required approval escalation level (0-6).
has_guarantor	Proposal includes guarantor.
has_decision_pending	Analyst requested more info before approval.
has_decision_unique	Only one decision was made.
has_approval_after_refuse	Approved after first being refused.
withdrawal_court	Proposal withdrawn due to legal/court issues.

withdrawal_documents	Withdrawal due to missing documents.
intervenients	Number of intervening persons.
owners	Number of vehicle owners.
repetead_business	Repeat business/customer.
count_re_analysis	Total reanalysis count.
count_re_analysis_partner	Reanalysis requested by partner.
cost_estimated	Estimated legal costs.
count_re_analysis_commercial	Reanalysis by commercial team.
count_re_analysis_conference	Reanalysis by conference team.
count_re_analysis_decision	Reanalysis after an initial decision.

9.4.1.7 Data Dictionary: Approved Loan Conversion – Target

target_booked	Proposal converted into contract.
target_closed	Proposal closed without booking.

9.4.2 Data Dictionary: Cross-Sell Propensity

9.4.2.1 Data Dictionary: Cross-Sell Propensity – Customer and Address Data

id	Sequential row number or ID for the record.
locality_1st_holder	Locality of first borrower
postal_code_1st_holder	Postal Code of first borrower
date_of_birth_1st_holder	Birthday of first borrower
bank_1st_holder	Bank of first borrower
date_of_birth_2nd_holder	Birthday of second borrower
bank_2nd_holder	Bank of second borrower
age_h1	Age of first borrower at contract conclusion

gender_h1	Gender of first borrower
marital_status	Marital status of first borrower
nr_auto_loan_holders	Number of borrowers (auto loan)
address_district	District of first borrower
address_municipality	Municipal of first borrower
nationality_h1	Nationality of first borrower
housing_type	Housing situation of first borrower
monthly_available_disposable_income	Estimated income of borrower (by Banco Primus)

9.4.2.2 Data Dictionary: Cross-Sell Propensity – Campaign and Offer Data

pre_approved_amount	Offered financed amount in campaign
eligible_type	Type of campaign eligibility (internal Banco Primus)
previous_offers	Number of previous offers to client
tan_2	TAN of the offer
offer_letter_stamp_duty	Stamp duty of the offer
offer_letter_term	Term in months of the offer
offer_letter_installment	Monthly installment of the offer
offer_letter_arp_2	ARP of the offer
offer_letter_total_amount_imputed_to_customer	Total loan amount of the offer
offer_letter_insurance	Cost of optional insurance
offer_letter_fees	Fees of the offer
offer_letter_campaign_expiry_date	Expiry date of the offer

eligibility_creation_date	Determinations date of eligibility
original_lead_channel	Communication channel of the offer

9.4.2.3 Data Dictionary: Cross-Sell Propensity – Existing Auto Loan

auto_loan_installment_due_date	Due date of auto loan
activation_date	Activation date of auto loan
current_monthly_installment	Current monthly installment of auto loan
amount_financed	Financed amount of auto loan
business_product	Product type of auto loan
term_original	Total term of auto loan
vehicletype_original	Brand of vehicle
vehicle_condition_original	Condition of vehicle
vehicle_age_original	Age of vehicle
vehicle_year_original	Construction year of vehicle
vehicle_model_original	Model of the vehicle
auto_tan_2	TAN of auto loan
auto_apr_2	ARP of auto loan
fuel_type_original	Fuel type of vehicle
partner_type_original	Type of dealer/intermediate
ppp_insurance	Subscription credit protection insurance
life_insurance	Subscription life insurance
auto_insurance	Subscription car insurance

9.4.2.4 Data Dictionary: Cross-Sell Propensity – Customer Profile & Risk

customer_tenure_months	Duration of customer relationship
months_to_end	Remaining months until end of loan
monitoring_score	Risk metric (internal Banco Primus)
customer_type	Customer type
crc_institutions	Number of financial institutions
profession	Job title
number_of_dependents	Number of dependents
profession_contract_type	Type of employment contract

9.4.2.5 Data Dictionary: Cross-Sell Propensity – Target

state	Status of the upsell offer
pre_approved_amount_actual	Accepted loan amount
amount_financed_actual	Final loan amount
avg_amount_finance_actual	Average financed amount
avg_weighted_nominal_rate_actual	Average TAN
avg_weighted_arp_actual	Average ARP
avg_weighted_term_actual	Average final term

9.5 Figures

9.5.1 Explainable AI Trade-Offs

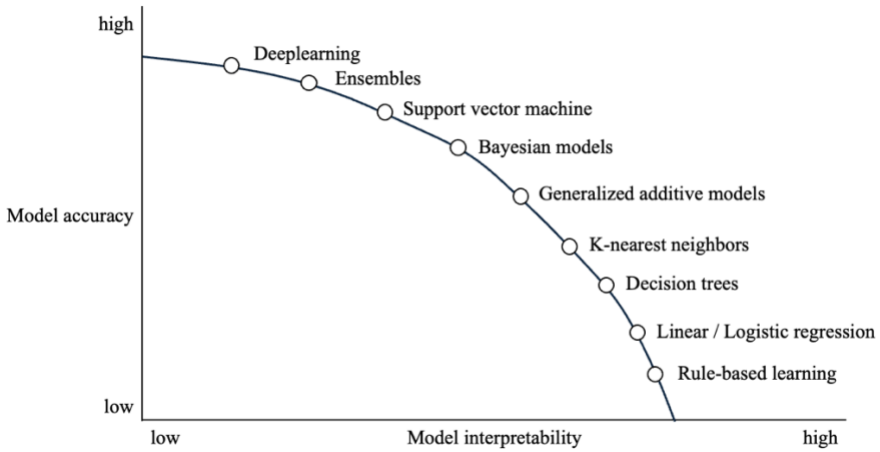


Figure 1: Inherent trade-offs between model accuracy and model interpretability (Adapted from Deloitte 2022)

9.5.2 ROC Curve

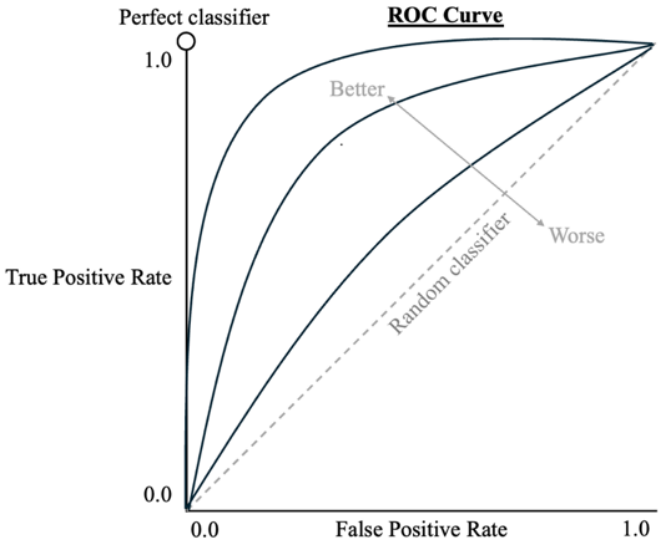


Figure 2: ROC Curve

9.5.3 Results: Approved Loan Conversion

Figure 3: EDA Approved Loan Conversion: Target Variable Distribution

Figure 4: EDA Approved Loan Conversion: Booking rate over Time

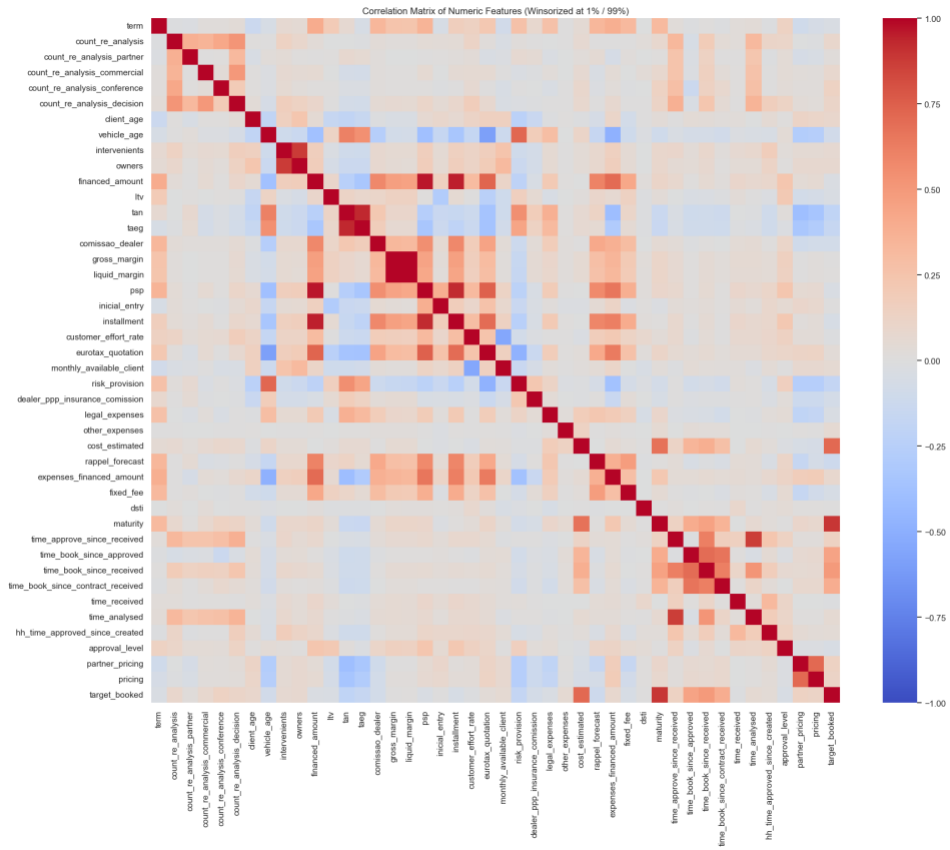


Figure 5: EDA: Correlation Heatmap

	variable	test	p_value	mean_Broker	mean_Dealer
3	reception_type	chi-square	4.316530e-188	NaN	NaN
8	expenses_financed	chi-square	0.000000e+00	NaN	NaN
12	has_risk_surveillance	z-test	1.175298e-291	0.18	0.02
14	fraud	z-test	1.944729e-10	0.02	0.01
31	life_elegible	z-test	0.000000e+00	0.49	0.78
41	marital_status	chi-square	1.142950e-93	NaN	NaN
43	has_mail	z-test	2.041705e-14	0.84	0.88
47	partner_type	z-test	0.000000e+00	NaN	NaN
51	risk_profile	chi-square	1.467582e-54	NaN	NaN
53	partner_rating	chi-square	9.224681e-219	NaN	NaN
58	risk_agency	chi-square	0.000000e+00	NaN	NaN
60	partner_pricing	chi-square	0.000000e+00	0.28	1.17
61	pricing	chi-square	0.000000e+00	0.48	1.24
64	count_re_analysis_partner	chi-square	8.975849e-145	0.30	0.11
65	count_re_analysis_commercial	chi-square	2.797787e-12	0.20	0.26
66	count_re_analysis_conference	chi-square	3.417657e-11	0.14	0.19
67	count_re_analysis_decision	chi-square	2.027698e-05	0.39	0.46
68	client_age	mann-whitney	4.433320e-132	34.64	39.83
69	vehicle_age	mann-whitney	0.000000e+00	9.56	6.62
72	financed_amount	mann-whitney	2.806529e-13	16158.96	17104.68
74	tan	mann-whitney	0.000000e+00	0.11	0.09
75	taeg	mann-whitney	0.000000e+00	0.13	0.11
77	gross_margin	mann-whitney	4.752777e-36	300.14	362.05
78	liquid_margin	mann-whitney	1.431937e-35	196.46	243.56
79	psp	mann-whitney	6.316663e-11	16146.40	17225.42
83	eurotax_quotation	mann-whitney	5.843214e-127	10839.17	14864.95
85	risk_provision	mann-whitney	0.000000e+00	0.06	0.03
91	expenses_financed_amount	mann-whitney	3.684671e-275	461.23	698.23
96	time_book_since_approved	mann-whitney	1.998134e-188	19.64	26.36
101	hh_time_approved_since_created	mann-whitney	2.529151e-59	14.96	21.77
102	target_booked	z-test	0.000000e+00	0.50	0.81

Figure 6: Statistical test of differences in mean by channel

9.5.4 Predicting Approved Loan Conversion: Logistic Regression

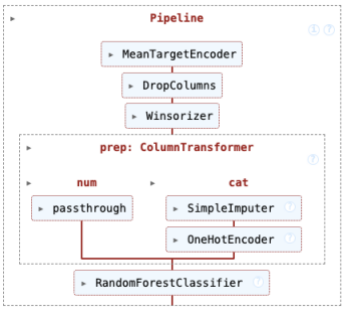


Figure 7: Logistic Regression Pipeline

```
============================================================
Logistic Regression Baseline – FINAL TEST SET EVALUATION
============================================================
Accuracy: 0.9905
ROC-AUC:  0.9997

Classification Report:
              precision    recall  f1-score   support

         0.0       0.99      0.97      0.98       735
         1.0       0.99      1.00      0.99      1784

    accuracy                           0.99      2519
   macro avg       0.99      0.99      0.99      2519
weighted avg       0.99      0.99      0.99      2519
```

Figure 8: Logistic Regression Baseline - Classification Report

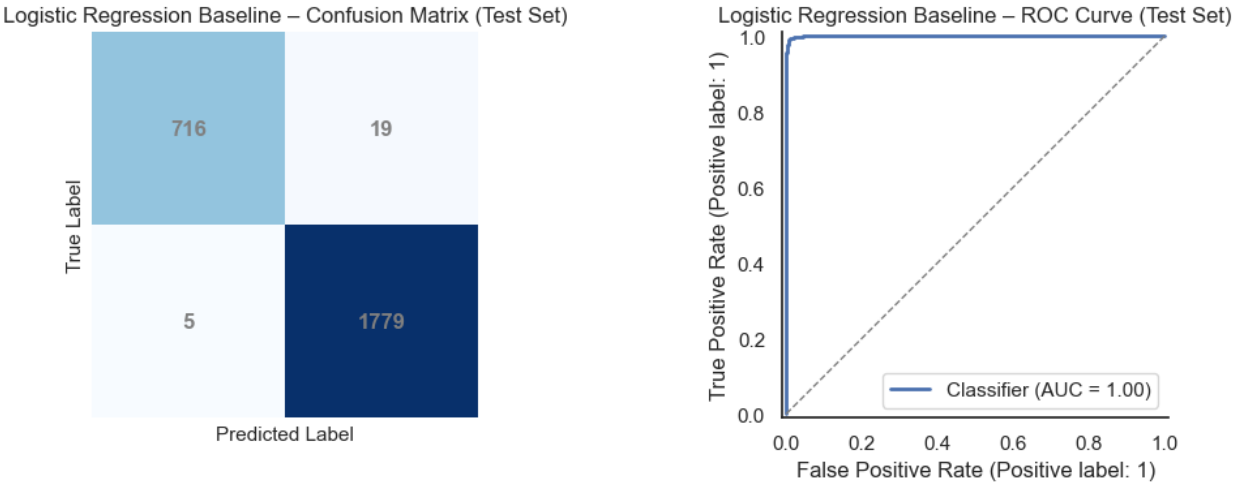


Figure 9: Logistic Regression Baseline - Confusion Matrix

Top 20 Features (Logistic Regression):

	feature	coef	odds_ratio	importance	method
47	num_life_elegible	4.107801	60.812821	4.107801	coef (abs)
48	num_ppp_elegible	2.637612	13.979775	2.637612	coef (abs)
70	num_user_conferencia_masked_woe	1.921016	6.827890	1.921016	coef (abs)
4	num_conference	1.809525	6.107547	1.809525	coef (abs)
112	cat_nationality_simple_portugal	1.367353	3.924947	1.367353	coef (abs)
31	num_client_age	1.290384	3.634181	1.290384	coef (abs)
43	num_has_insurance_life	-0.994076	0.370065	0.994076	coef (abs)
99	cat_partner_risk_prime	0.775316	2.171279	0.775316	coef (abs)
79	cat_risk_agency_good	0.711792	2.037639	0.711792	coef (abs)
3	num_fraud	-0.700734	0.496221	0.700734	coef (abs)
92	cat_branch_north	0.698894	2.011528	0.698894	coef (abs)
49	num_client_effective	-0.673989	0.509671	0.673989	coef (abs)
44	num_has_insurance_ppp	-0.568583	0.566327	0.568583	coef (abs)
61	num_analyst_masked_woe	0.556272	1.744158	0.556272	coef (abs)
93	cat_branch_south	-0.507955	0.601725	0.507955	coef (abs)
50	num_client_self_employed	-0.499642	0.606748	0.499642	coef (abs)
20	num_customer_effort_rate	-0.496575	0.608612	0.496575	coef (abs)
14	num_repetead_business	0.444148	1.559161	0.444148	coef (abs)
87	cat_gender_male	0.413272	1.511757	0.413272	coef (abs)
45	num_insurance_ppp_telephone	0.412385	1.510416	0.412385	coef (abs)

Figure 10: Logistic Regression Baseline - Feature Importance

```
============================================================
Logistic Regression — Clean Features Tuned — FINAL TEST SET EVALUATION
============================================================
Accuracy: 0.7753
ROC-AUC:  0.7987

Classification Report:
              precision    recall  f1-score   support

         0.0       0.64      0.53      0.58       735
         1.0       0.82      0.88      0.85      1784

    accuracy                           0.78      2519
   macro avg       0.73      0.70      0.71      2519
weighted avg       0.77      0.78      0.77      2519
```

Figure 11: Logistic Regression Tuned - Classification Report

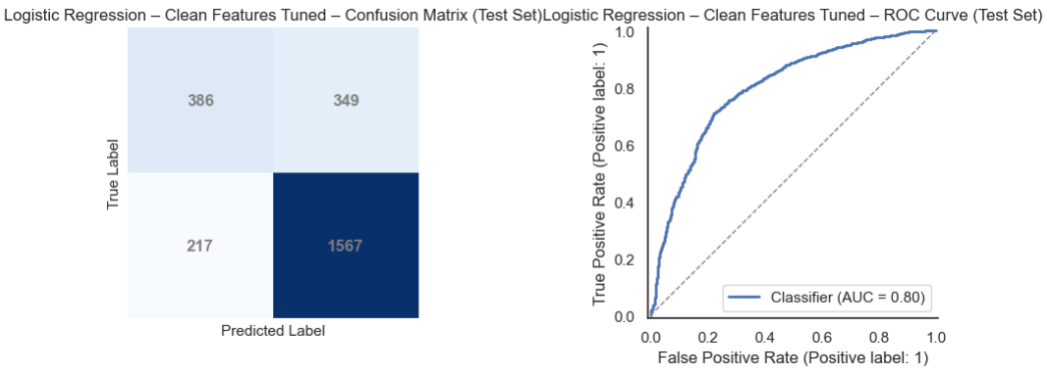


Figure 12: Logistic Regression Tuned - Confusion Matrix

Top 20 Features (Logistic Regression):

	feature	coef	odds_ratio	importance	method
65	cat_risk_agency_prime	1.264927	3.542834	1.264927	coef (abs)
57	cat_expenses_financed_not_financed	-1.125220	0.324581	1.125220	coef (abs)
77	cat_partner_rating_unknown	1.118415	3.060001	1.118415	coef (abs)
61	cat_risk_agency_good_less	1.036769	2.820089	1.036769	coef (abs)
60	cat_risk_agency_good	0.962172	2.617375	0.962172	coef (abs)
62	cat_risk_agency_medium	0.912404	2.490302	0.912404	coef (abs)
21	num_time_approve_since_received	0.879932	2.410735	0.879932	coef (abs)
73	cat_partner_type_dealer	0.796927	2.218712	0.796927	coef (abs)
81	cat_financing_type_simple_other	-0.784438	0.456376	0.784438	coef (abs)
36	num_has_mail	0.780440	2.182432	0.780440	coef (abs)
54	cat_reception_type_internal	0.726950	2.068762	0.726950	coef (abs)
82	cat_nationality_simple_portugal	0.540405	1.716702	0.540405	coef (abs)
63	cat_risk_agency_medium_less	0.513527	1.671175	0.513527	coef (abs)
67	cat_client_type_self-_employed	-0.407415	0.665368	0.407415	coef (abs)
33	num_has_insurance_ppp	0.377903	1.459222	0.377903	coef (abs)
75	cat_partner_rating_ruby	-0.373861	0.688072	0.373861	coef (abs)
79	cat_marital_status_simple_single	-0.336580	0.714209	0.336580	coef (abs)
58	cat_expenses_financed_supported_by_partner	-0.316686	0.728560	0.316686	coef (abs)
68	cat_branch_center	-0.294301	0.745052	0.294301	coef (abs)
69	cat_branch_lisbon	-0.291548	0.747106	0.291548	coef (abs)

Figure 13: Logistic Regression Tuned - Feature Importance

9.5.5 Predicting Approved Loan Conversion: Random Forest

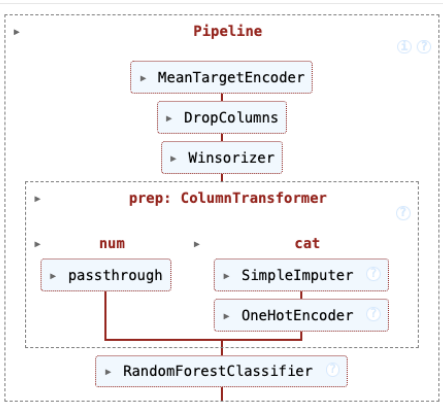


Figure 14: Random Forest Pipeline

```
============================================================
Random Forest – Baseline – FINAL TEST SET EVALUATION
============================================================
Accuracy: 0.7725
ROC-AUC:  0.8009

Classification Report:
              precision    recall  f1-score   support

         0.0       0.65      0.47      0.55       735
         1.0       0.81      0.90      0.85      1784

    accuracy                           0.77      2519
   macro avg       0.73      0.68      0.70      2519
weighted avg       0.76      0.77      0.76      2519
```

Figure 15: Random Forest Baseline - Classification Report

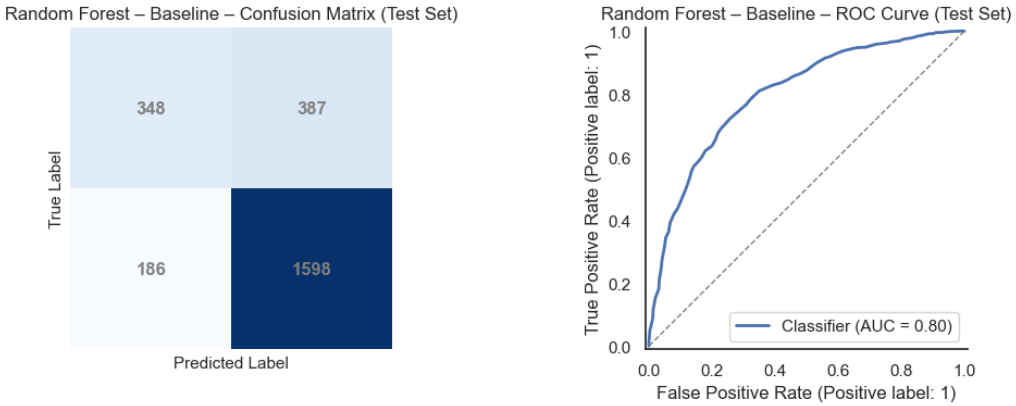


Figure 16: Random Forest Baseline - Confusion Matrix

Top 20 Features (Random Forest — Baseline):

	feature	importance	method
21	num_time_approve_since_received	0.070592	feature_importances_
36	num_has_mail	0.056077	feature_importances_
48	num_partner_id_masked_te	0.044415	feature_importances_
46	num_partner_city_te	0.037663	feature_importances_
19	num_risk_provision	0.034155	feature_importances_
25	num_client_age	0.032303	feature_importances_
12	num_liquid_margin	0.030753	feature_importances_
30	num_dsti	0.030614	feature_importances_
18	num_monthly_available_client	0.030416	feature_importances_
14	num_customer_effort_rate	0.030226	feature_importances_
11	num_comissao_dealer	0.030058	feature_importances_
49	num_partner_bill_id_masked_te	0.028497	feature_importances_
28	num_expenses_financed_amount	0.028403	feature_importances_
9	num_financed_amount	0.027815	feature_importances_
27	num_rappel_forecast	0.026794	feature_importances_
26	num_taeg	0.026712	feature_importances_
10	num_ltv	0.025286	feature_importances_
73	cat_partner_type_dealer	0.022682	feature_importances_
50	num_profession_te	0.022349	feature_importances_
16	num_eurotax_quotation	0.022172	feature_importances_

Figure 17: Random Forest Baseline - Feature Importance

```
============================================================
Random Forest — Tuned — FINAL TEST SET EVALUATION
============================================================
Accuracy: 0.7785
ROC-AUC:  0.812

Classification Report:
              precision    recall  f1-score   support

         0.0       0.65      0.51      0.58       735
         1.0       0.82      0.89      0.85      1784

    accuracy                           0.78      2519
   macro avg       0.73      0.70      0.71      2519
weighted avg       0.77      0.78      0.77      2519
```

Figure 18: Random Forest Tuned - Classification Report

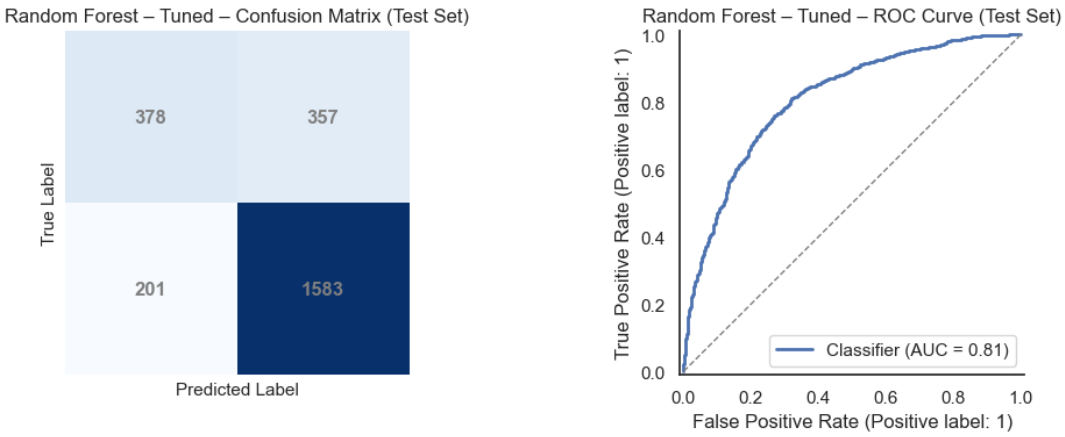


Figure 19: Random Forest Tuned - Confusion Matrix

Top 20 Features (Random Forest – Tuned):

	feature	importance	method
48	num_partner_id_masked_te	0.114165	feature_importances_
21	num_time_approve_since_received	0.113107	feature_importances_
36	num_has_mail	0.091470	feature_importances_
25	num_client_age	0.048870	feature_importances_
73	cat_partner_type_dealer	0.043683	feature_importances_
12	num_liquid_margin	0.028708	feature_importances_
11	num_comissao_dealer	0.027535	feature_importances_
49	num_partner_bill_id_masked_te	0.027455	feature_importances_
40	num_hh_time_approved_since_created	0.026043	feature_importances_
18	num_monthly_available_client	0.025802	feature_importances_
19	num_risk_provision	0.023590	feature_importances_
46	num_partner_city_te	0.023035	feature_importances_
30	num_dsti	0.022886	feature_importances_
27	num_rappel_forecast	0.022141	feature_importances_
14	num_customer_effort_rate	0.021168	feature_importances_
28	num_expenses_financed_amount	0.020200	feature_importances_
10	num_ltv	0.020136	feature_importances_
26	num_taeg	0.020121	feature_importances_
50	num_profession_te	0.019132	feature_importances_
9	num_financed_amount	0.018306	feature_importances_

Figure 20: Random Forest Tuned - Feature Importance

9.5.6 Predicting Approved Loan Conversion: XGBoost

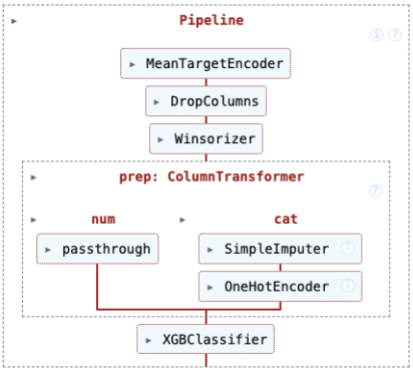


Figure 21: XGBoost Pipeline

```
============================================================
XGBoost – Baseline – FINAL TEST SET EVALUATION
============================================================
Accuracy: 0.7864
ROC-AUC:  0.8203

Classification Report:
              precision    recall  f1-score   support

         0.0       0.66      0.55      0.60       735
         1.0       0.83      0.88      0.85      1784

    accuracy                           0.79      2519
   macro avg       0.74      0.72      0.73      2519
weighted avg       0.78      0.79      0.78      2519
```

Figure 22: XGBoost Baseline - Classification Report

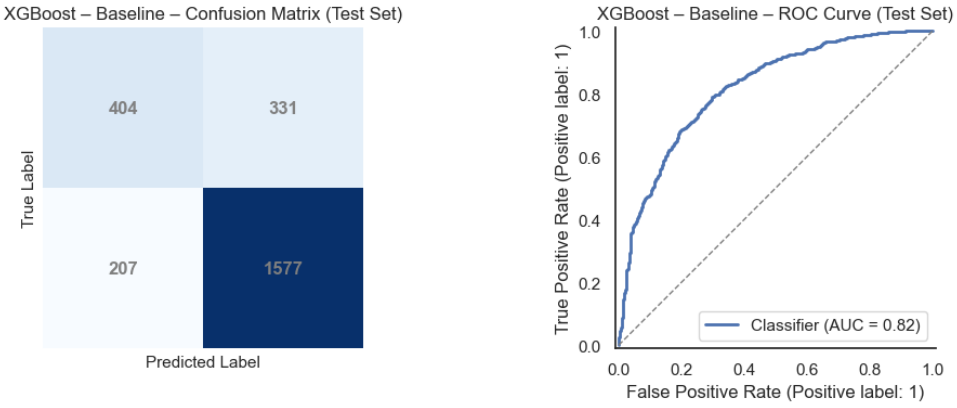


Figure 23: XGBoost Baseline - Confusion Matrix

Top 20 Features (XGBoost — Baseline):

	feature	importance	method
64	cat_partner_type_dealer	627.310059	gain
3	num_partner_bill	156.313705	gain
44	num_partner_id_masked_te	149.094284	gain
35	num_has_mail	97.544685	gain
51	cat_reception_type_one_app	84.632843	gain
42	num_partner_city_te	65.102417	gain
20	num_time_approve_since_received	54.858570	gain
24	num_client_age	45.710835	gain
52	cat_reception_type_phone	34.598957	gain
48	cat_state_id_approved_with_conditions	28.254642	gain
45	num_partner_bill_id_masked_te	26.173054	gain
21	num_time_received	25.105728	gain
2	num_was_conditional	20.490274	gain
38	num_hh_time_approved_since_created	20.174793	gain
58	cat_risk_agency_prime	18.929634	gain
66	cat_marital_status_simple_single	15.765142	gain
19	num_dealer_ppp_insurance_comission	15.196682	gain
32	num_has_insurance_ppp	14.702591	gain
40	num_product_group_te	13.637771	gain
5	num_approval_level	13.033377	gain

Figure 24: XGBoost Baseline - Feature Importance

```
============================================================
XGBoost — Tuned — FINAL TEST SET EVALUATION
============================================================
Accuracy: 0.788
ROC-AUC:  0.8247

Classification Report:
              precision    recall  f1-score   support

         0.0       0.67      0.55      0.60       735
         1.0       0.83      0.89      0.86      1784

    accuracy                           0.79      2519
   macro avg       0.75      0.72      0.73      2519
weighted avg       0.78      0.79      0.78      2519
```

Figure 25: XGBoost Tuned - Classification Report

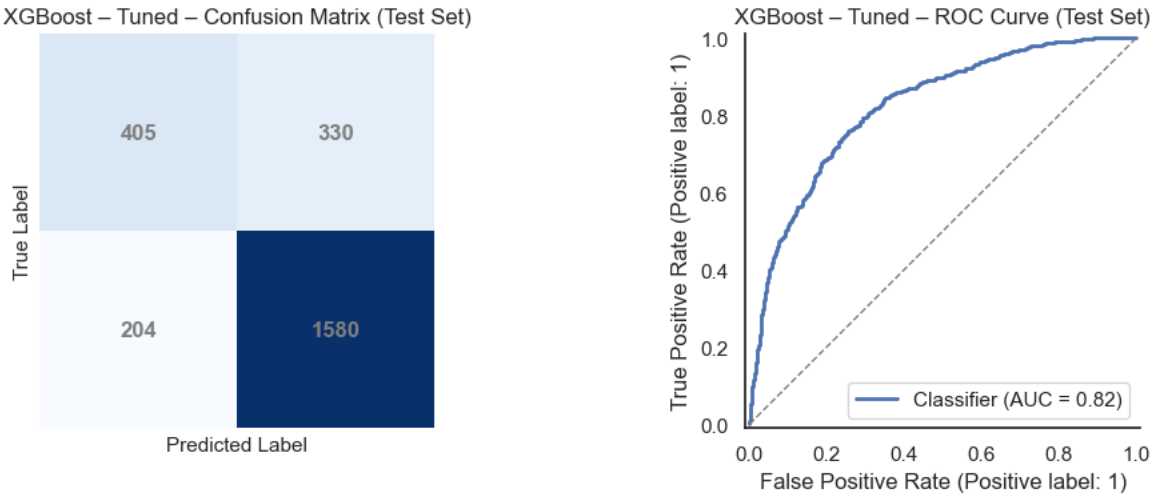


Figure 26: XGBoost Tuned - Confusion Matrix

Top 20 Features (XGBoost – Tuned):

	feature	importance	method
71	cat_partner_type_dealer	138.353912	gain
36	num_has_mail	55.221054	gain
4	num_partner_bill	41.166279	gain
48	num_partner_id_masked_te	40.340908	gain
56	cat_reception_type_phone	17.709618	gain
21	num_time_approve_since_received	16.595509	gain
52	cat_state_id_approved_with_conditions	14.976727	gain
55	cat_reception_type_one_app	12.837098	gain
3	num_was_conditional	12.017457	gain
49	num_partner_bill_id_masked_te	11.405222	gain
24	num_other_expenses	8.879629	gain
25	num_client_age	8.666225	gain
22	num_time_received	8.450247	gain
41	num_partner_id_bill_match	8.400864	gain
46	num_partner_city_te	8.373905	gain
77	cat_marital_status_simple_single	7.928960	gain
33	num_has_insurance_ppp	7.776137	gain
40	num_hh_time_approved_since_created	7.244759	gain
20	num_dealer_ppp_insurance_comission	6.847145	gain
62	cat_risk_agency_new_partner	6.505890	gain

Figure 27: XGBoost Tuned - Feature Importance

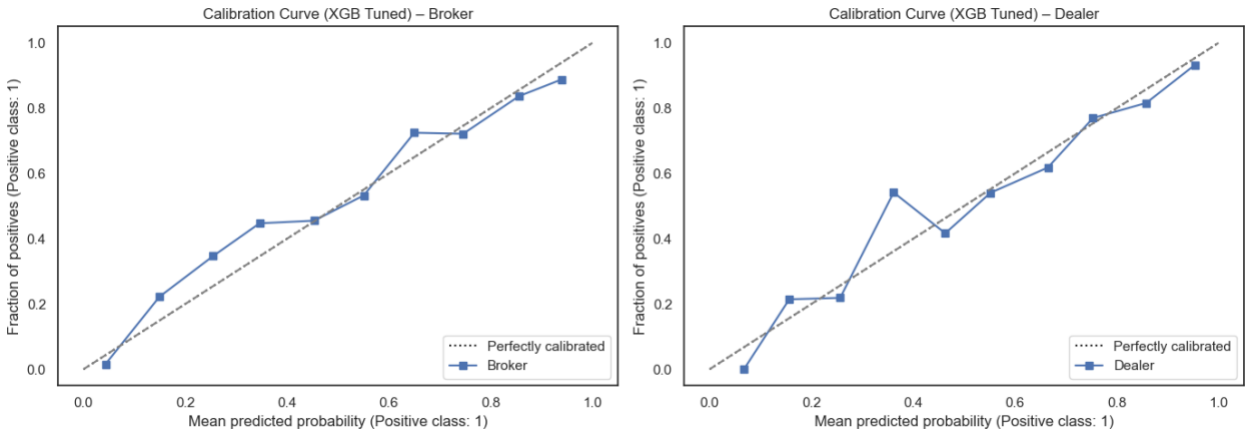


Figure 28: XGBoost Tuned - Calibration Curve per Channel

```
=== Broker Classification Report ===
              precision    recall  f1-score   support

         0.0       0.66      0.71      0.68       438
         1.0       0.74      0.69      0.71       520

    accuracy                           0.70       958
   macro avg       0.70      0.70      0.70       958
weighted avg       0.70      0.70      0.70       958


=== Dealer Classification Report ===
              precision    recall  f1-score   support

         0.0       0.70      0.32      0.44       297
         1.0       0.86      0.97      0.91      1264

    accuracy                           0.84      1561
   macro avg       0.78      0.64      0.67      1561
weighted avg       0.83      0.84      0.82      1561
```

Figure 29: XGBoost Tuned - Classification Report per Channel

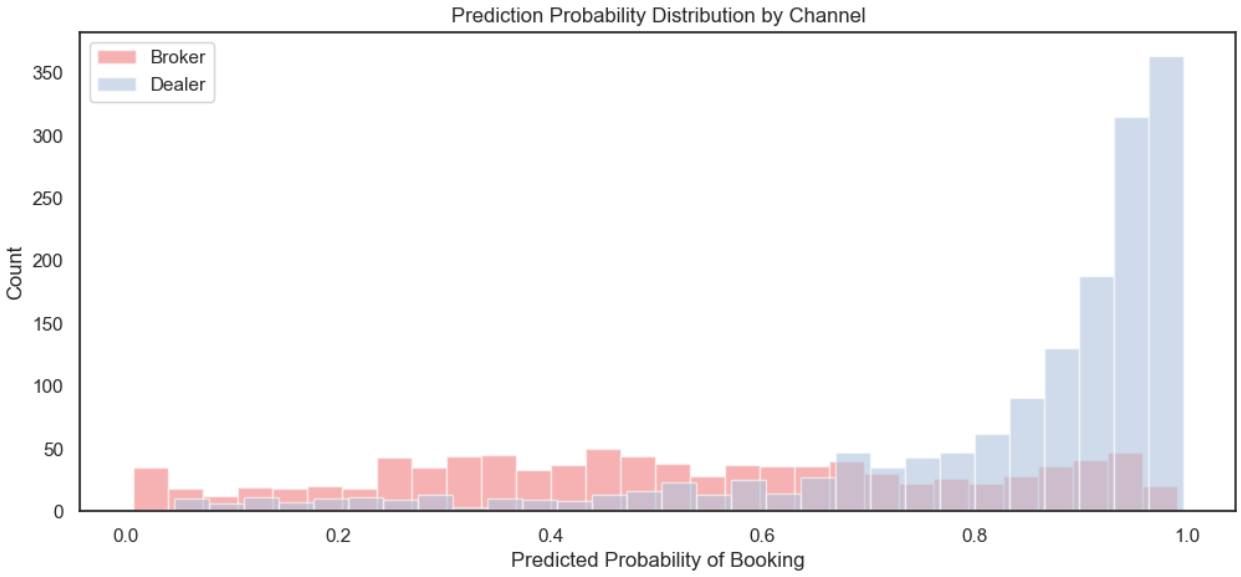


Figure 30: XGBoost Tuned - Prediction Probability Distribution by Channel

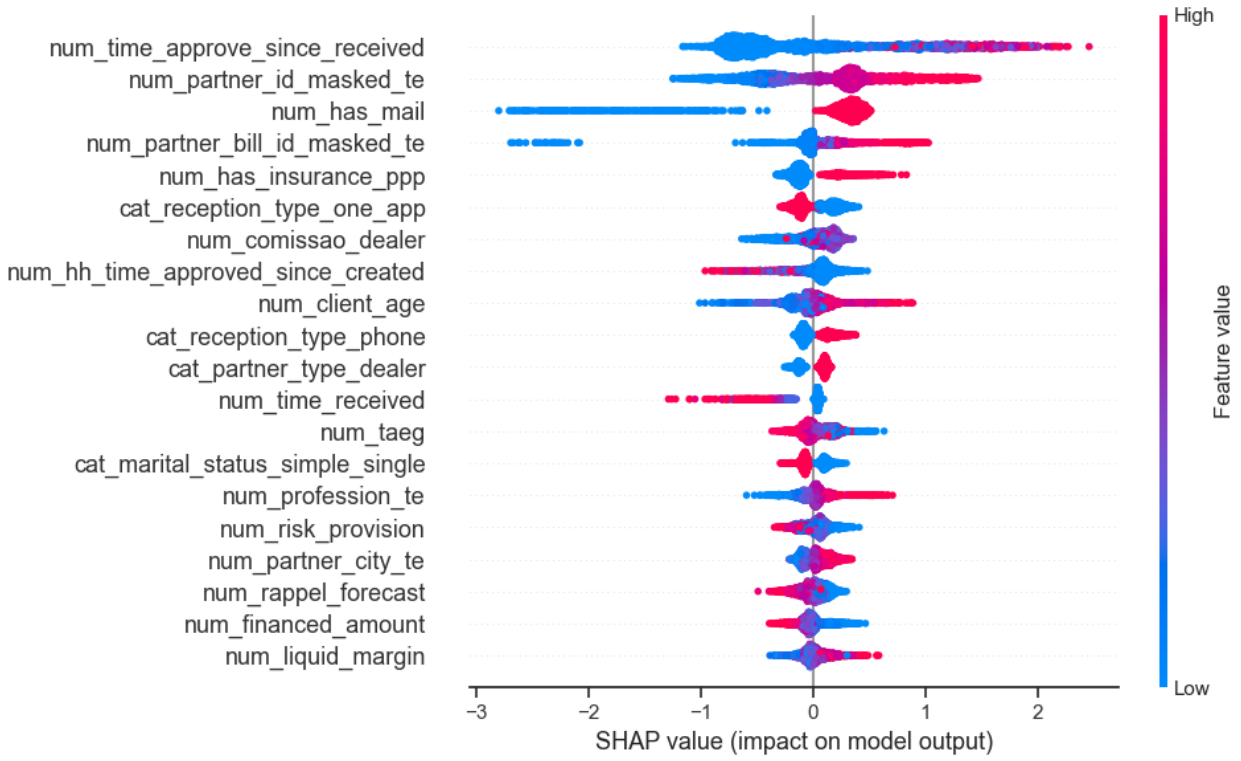


Figure 31: XGBoost Tuned - SHAP (Global)

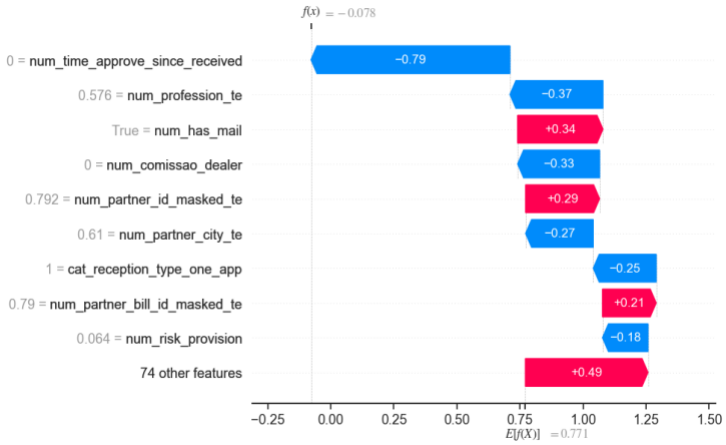


Figure 32: XGBoost Tuned - SHAP (Waterfall Local)

	Threshold	Conversion_rate	Coverage	Number_Cases	Uplift_vs_baseline
0	0.500000	0.827	75.82%	1910	0.119
1	0.600000	0.855	69.15%	1742	0.147
2	0.700000	0.878	61.29%	1544	0.170
3	0.800000	0.898	53.16%	1339	0.189
4	0.900000	0.928	38.07%	959	0.220
5	0.950000	0.955	21.40%	539	0.247
6	0.990000	0.963	1.07%	27	0.255

Figure 33: XGBoost Tuned - Backtesting Results

9.5.7 Exploratory: Approved Loan Conversion: Logistic Regression

```
============================================================
XGBoost — Tuned — Exploratory — FINAL TEST SET EVALUATION
============================================================
Accuracy: 0.8511
ROC-AUC:  0.8865

Classification Report:
              precision    recall  f1-score   support

         0.0       0.83      0.61      0.71       735
         1.0       0.86      0.95      0.90      1784

    accuracy                           0.85      2519
   macro avg       0.84      0.78      0.80      2519
weighted avg       0.85      0.85      0.84      2519
```

Figure 34: XGBoost Tuned - Exploratory - Classification Report

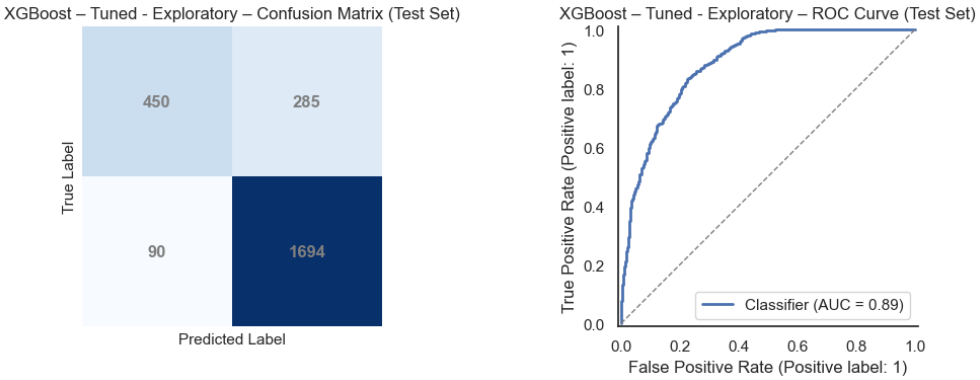


Figure 35: XGBoost Tuned - Exploratory - Confusion Matrix

Top 20 Features (XGBoost – Tuned – Exploratory):

	feature	importance	method
79	cat_gender_unknown	256.414673	gain
47	num_has_mail	86.539505	gain
61	num_partner_id_masked_te	34.153324	gain
35	num_count_re_analysis_conference	22.614159	gain
56	num_analyst_masked_te	22.511274	gain
29	num_client_age	18.936491	gain
43	num_insurance_ppp_telephone	17.755095	gain
25	num_time_approve_since_received	16.916834	gain
92	cat_partner_type_dealer	16.668798	gain
9	num_has_decision_unique	16.456650	gain
44	num_insurance_life_telephone	15.829839	gain
62	num_partner_bill_id_masked_te	14.232443	gain
90	cat_partner_risk_prime	13.605839	gain
74	cat_risk_agency_new_partner	13.430592	gain
3	num_was_conditional	12.198789	gain
59	num_partner_city_te	11.631386	gain
89	cat_partner_risk_new_partner	10.444811	gain
68	cat_reception_type_phone	8.971891	gain
54	num_product_group_te	8.768347	gain
98	cat_marital_status_simple_single	8.691087	gain

Figure 36: XGBoost Tuned - Exploratory - Feature Importance

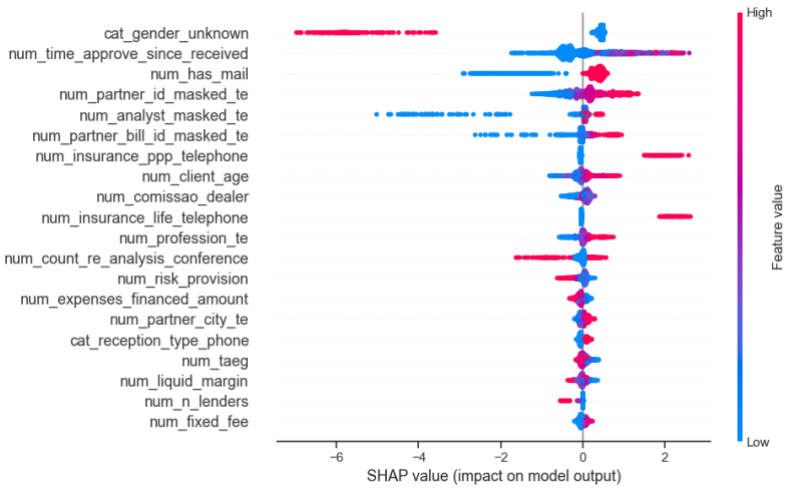


Figure 37: XGBoost Tuned - Exploratory - SHAP (Global)

```
============================================================
XGBoost – Tuned – Broker – FINAL TEST SET EVALUATION
============================================================
Accuracy: 0.6827
ROC-AUC:  0.7646

Classification Report:
              precision    recall  f1-score   support

         0.0       0.63      0.73      0.68       438
         1.0       0.74      0.64      0.69       520

    accuracy                           0.68       958
   macro avg       0.69      0.69      0.68       958
weighted avg       0.69      0.68      0.68       958
```

Figure 38: XGBoost Tuned - Broker – Classification Report

```
============================================================
XGBoost – Tuned – Dealer – FINAL TEST SET EVALUATION
============================================================
Accuracy: 0.7809
ROC-AUC:  0.8008

Classification Report:
              precision    recall  f1-score   support

         0.0       0.44      0.60      0.51       297
         1.0       0.90      0.82      0.86      1264

    accuracy                           0.78      1561
   macro avg       0.67      0.71      0.68      1561
weighted avg       0.81      0.78      0.79      1561
```

Figure 39: XGBoost Tuned - Dealer – Classification Report

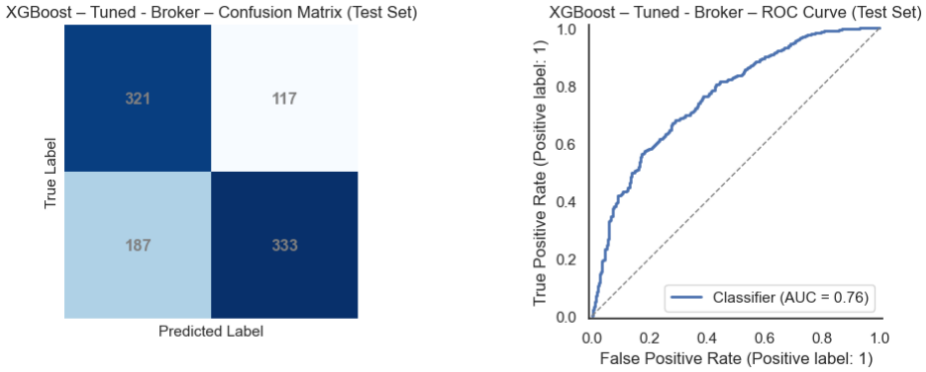


Figure 40: XGBoost Tuned - Broker – Confusion Matrix

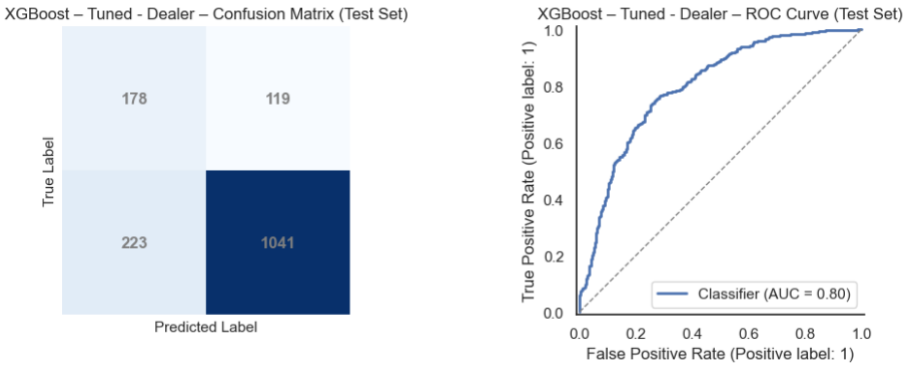


Figure 41: XGBoost Tuned - Dealer – Confusion Matrix

	feature	importance	method
33	num_has_mail	49.090939	gain
20	num_time_approve_since_received	48.174091	gain
52	cat_reception_type_phone	22.759064	gain
45	num_partner_bill_id_masked_te	18.776932	gain
37	num_hh_time_approved_since_created	14.353623	gain
44	num_partner_id_masked_te	13.728685	gain
30	num_has_insurance_ppp	13.702402	gain
51	cat_reception_type_one_app	13.193631	gain
3	num_was_conditional	10.301175	gain
21	num_time_received	9.904495	gain
42	num_partner_city_te	9.408502	gain
23	num_client_age	9.395977	gain
48	cat_state_id_approved_with_conditions	8.289915	gain
12	num_inicial_entry	8.133697	gain
4	num_partner_pricing	7.146743	gain
29	num_has_insurance_life	7.070928	gain
49	cat_state_id_approved_with_pending_documents	6.914850	gain
58	cat_risk_agency_new_partner	6.887810	gain
0	num_eligible_rappel	6.751675	gain
63	cat_branch_south	6.734354	gain

Figure 42: XGBoost Tuned - Broker – Feature Importance

	feature	importance	method
30	num_has_mail	30.752295	gain
46	cat_reception_type_phone	21.079254	gain
40	num_partner_id_masked_te	16.693342	gain
41	num_partner_bill_id_masked_te	15.254862	gain
16	num_dealer_ppp_insurance_comission	10.874848	gain
38	num_partner_city_te	8.837203	gain
45	cat_reception_type_one_app	8.141162	gain
53	cat_risk_agency_prime	8.031996	gain
17	num_time_approve_since_received	7.848266	gain
20	num_client_age	5.153729	gain
36	num_product_group_te	4.310711	gain
27	num_has_insurance_ppp	4.221426	gain
37	num_client_district_te	4.045810	gain
42	num_profession_te	4.023109	gain
2	num_approval_level	3.879335	gain
39	num_partner_district_te	3.830210	gain
22	num_rappel_forecast	3.781255	gain
59	cat_branch_south	3.743110	gain
64	cat_marital_status_simple_single	3.586783	gain
33	num_hh_time_approved_since_created	3.584857	gain

Figure 43: XGBoost Tuned - Dealer – Feature Importance

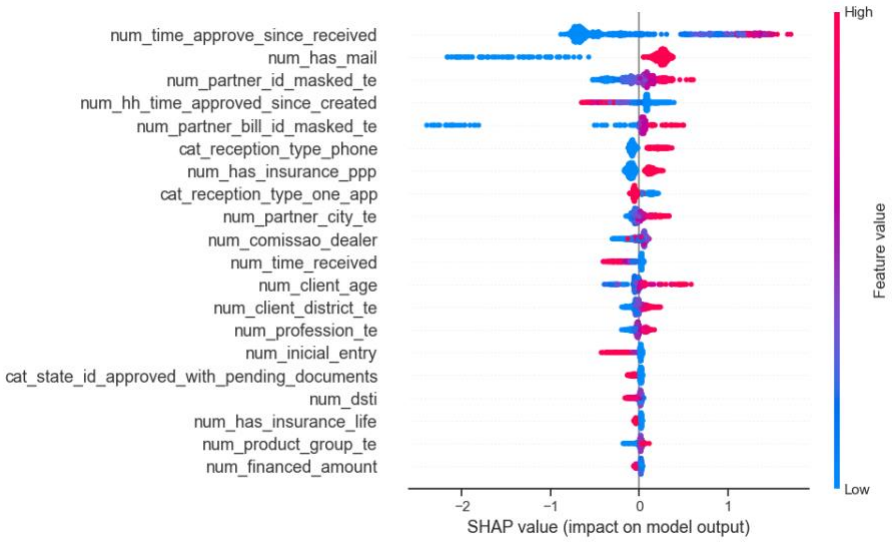


Figure 44: XGBoost Tuned - Broker – SHAP (Global)

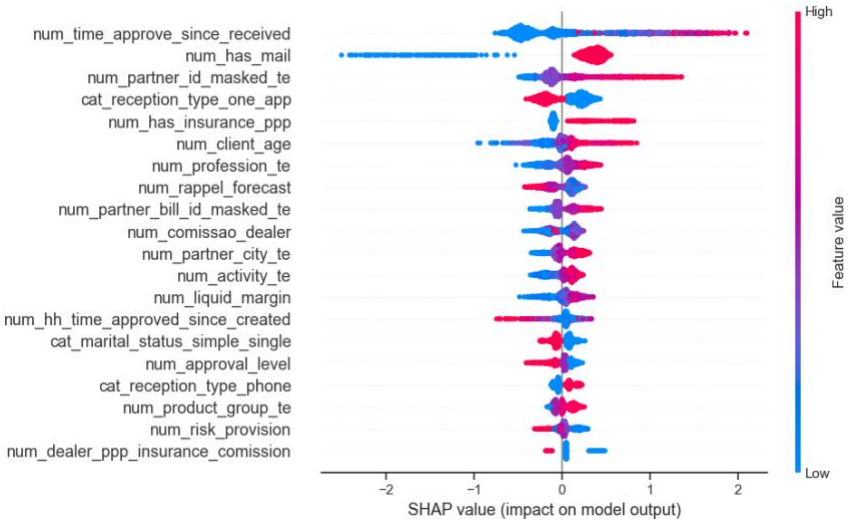


Figure 45: XGBoost Tuned - Dealer – SHAP (Global)

9.5.8 Personal Loan Cross-Sell

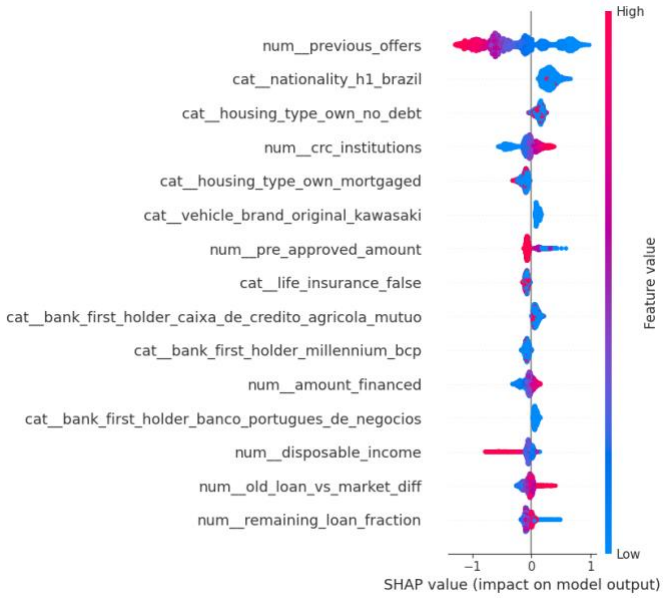


Figure 46: XGBoost Tuned – Cross-sell - SHAP (Global)

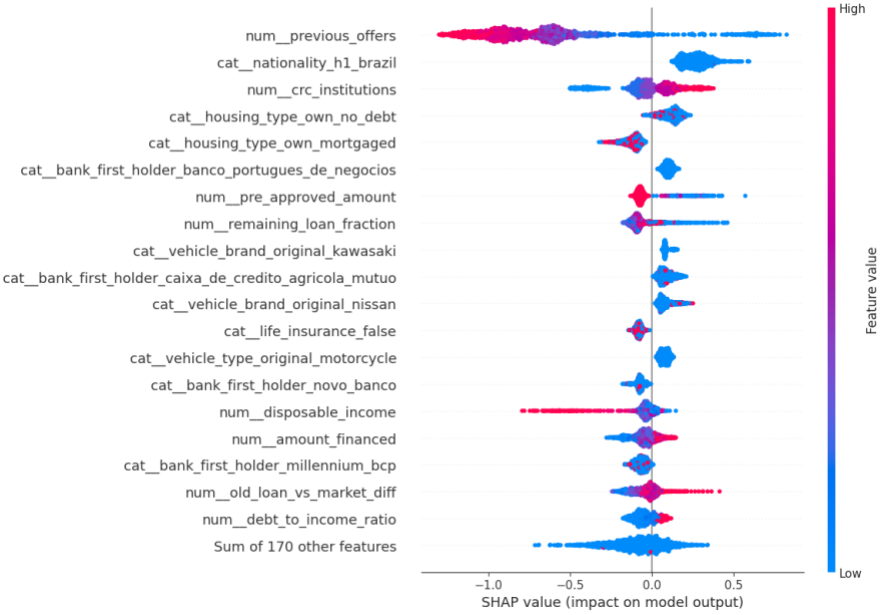


Figure 47: SHAP Fatigued Tenured

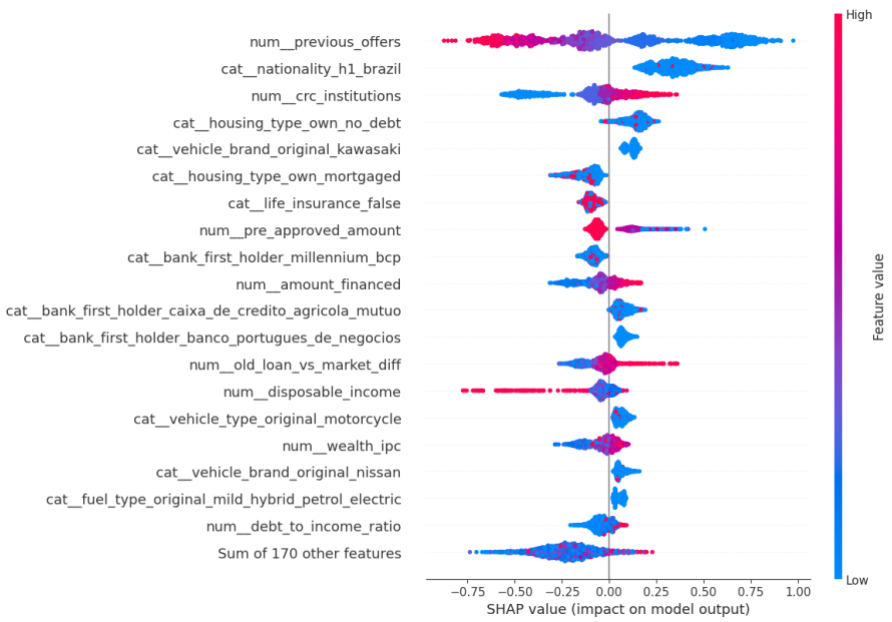
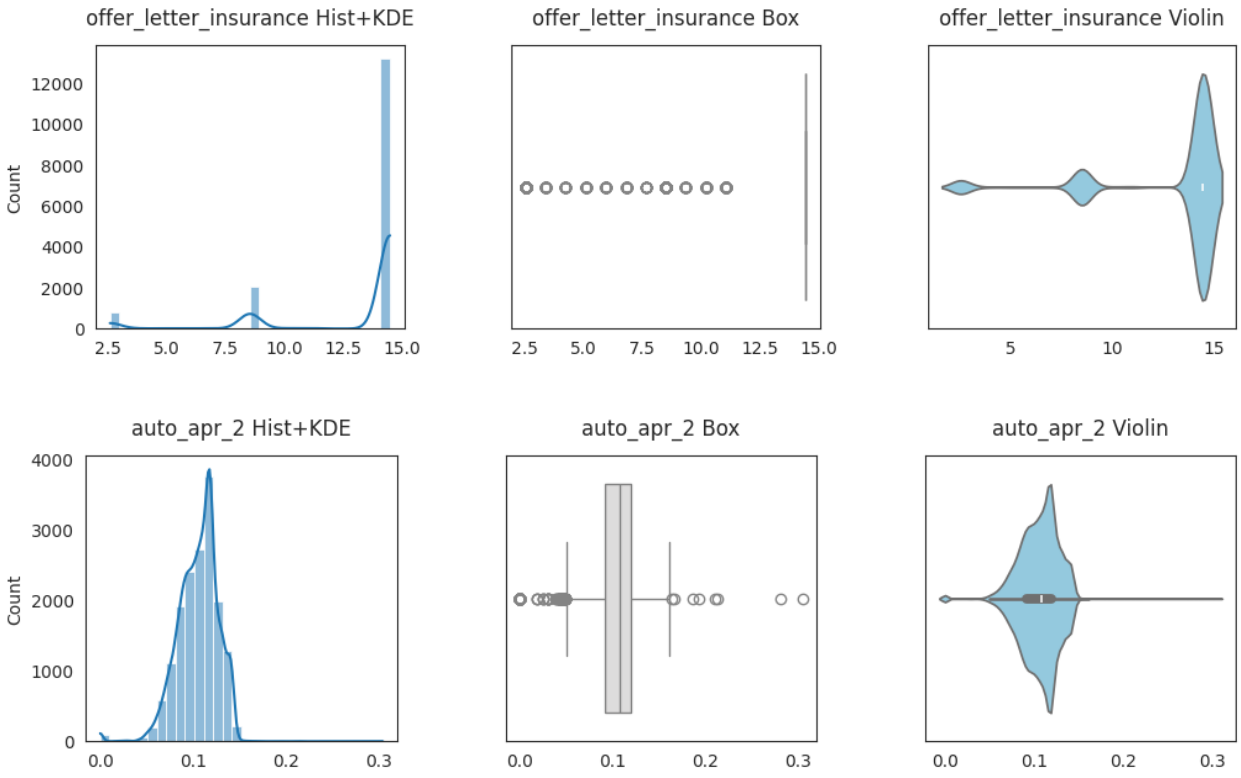
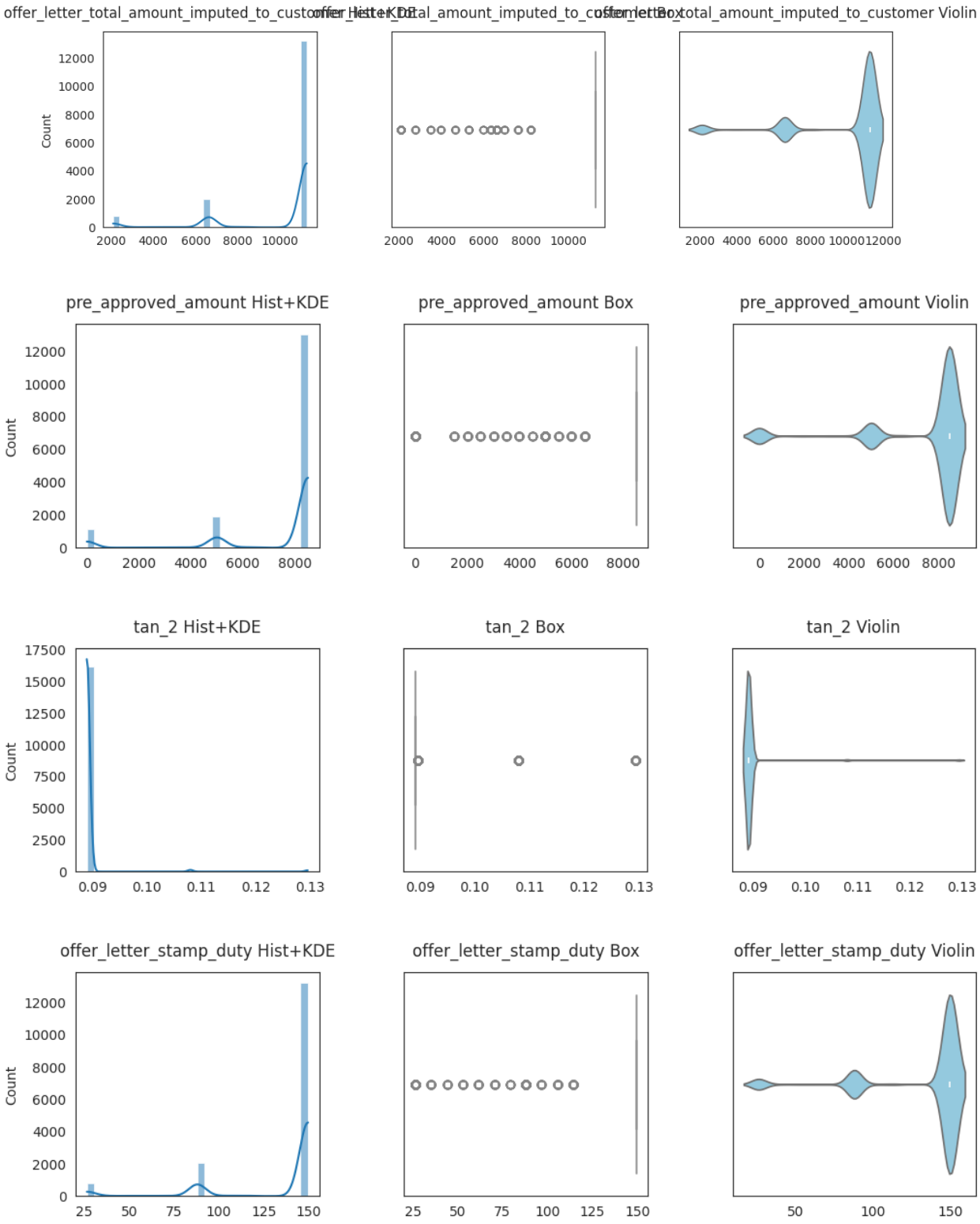


Figure 48: SHAP Fresh Prospects




offer_letter_total_amount_imputed_to_customer Hist+KDE
offer_letter_total_amount_imputed_to_customer Box
offer_letter_total_amount_imputed_to_customer Violin
Count
pre_approved_amount Hist+KDE
pre_approved_amount Box
pre_approved_amount Violin
tan_2 Hist+KDE
tan_2 Box
tan_2 Violin
offer_letter_stamp_duty Hist+KDE
offer_letter_stamp_duty Box
offer_letter_stamp_duty Violin

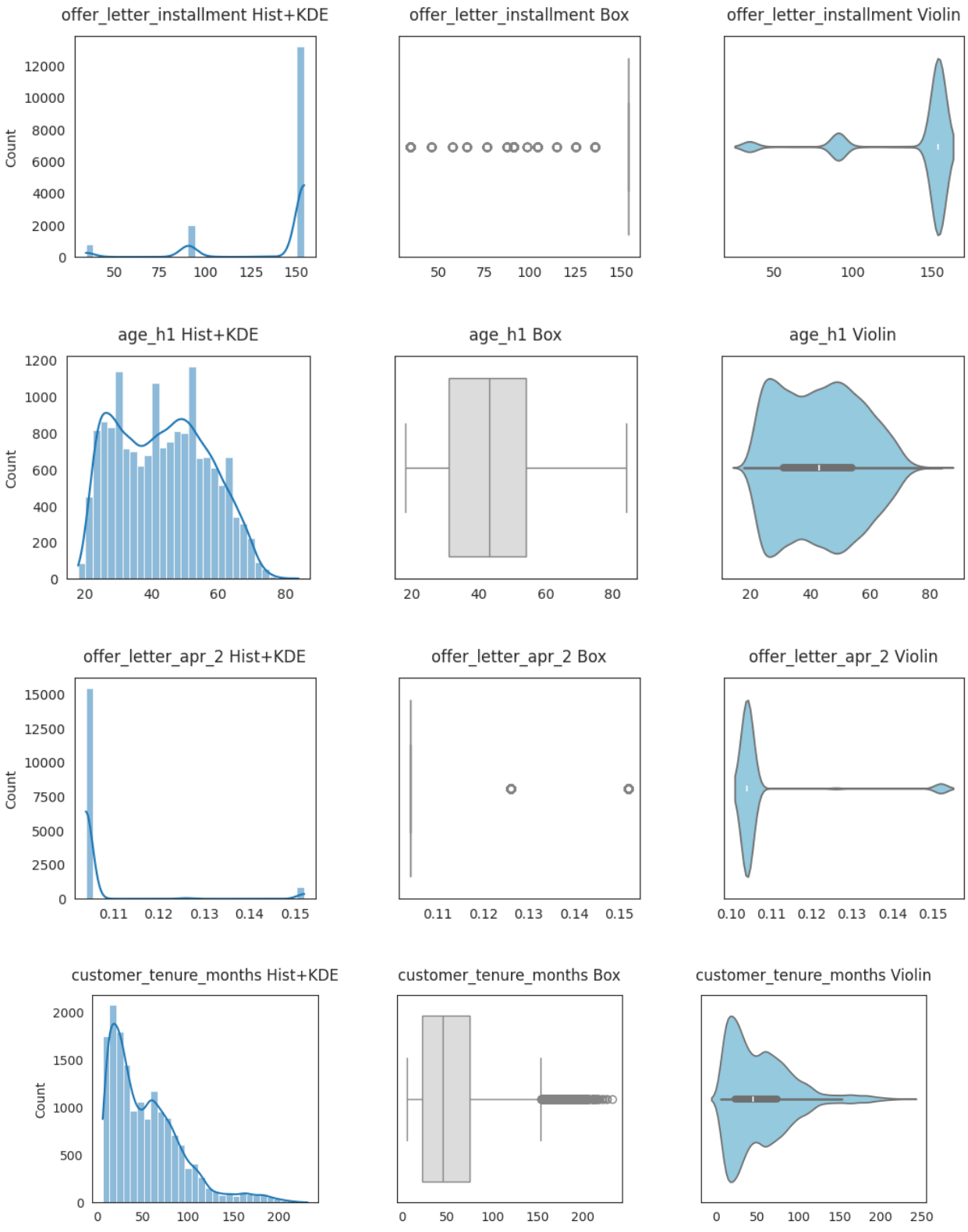


Figure 49: Feature Distribution Plots Cross-Sell

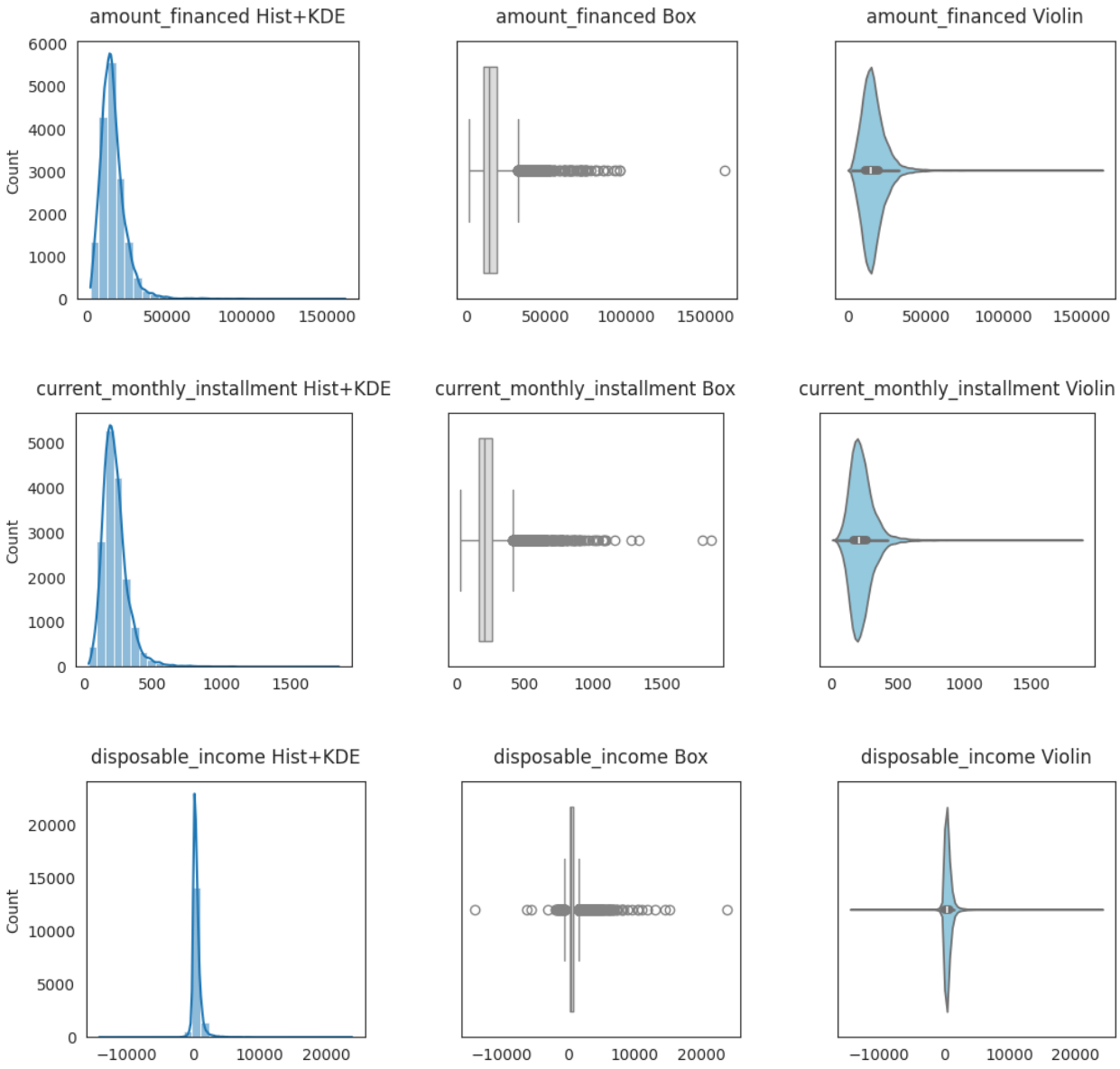


Figure 50: Finance Feature Distribution Plots Cross-Sell

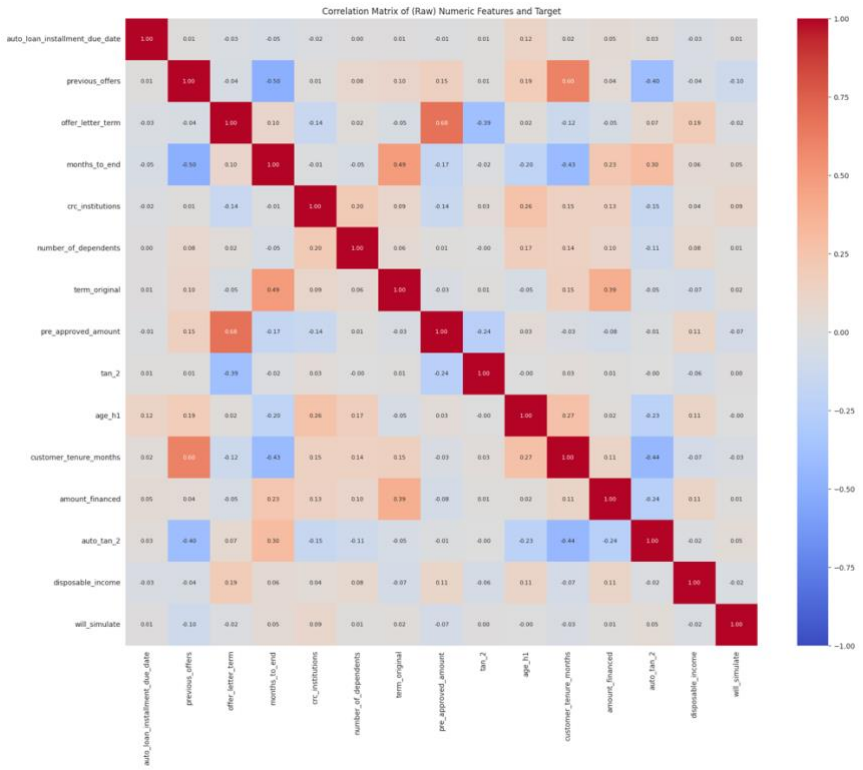


Figure 51: Correlations Matrix with Target Cross-Sell

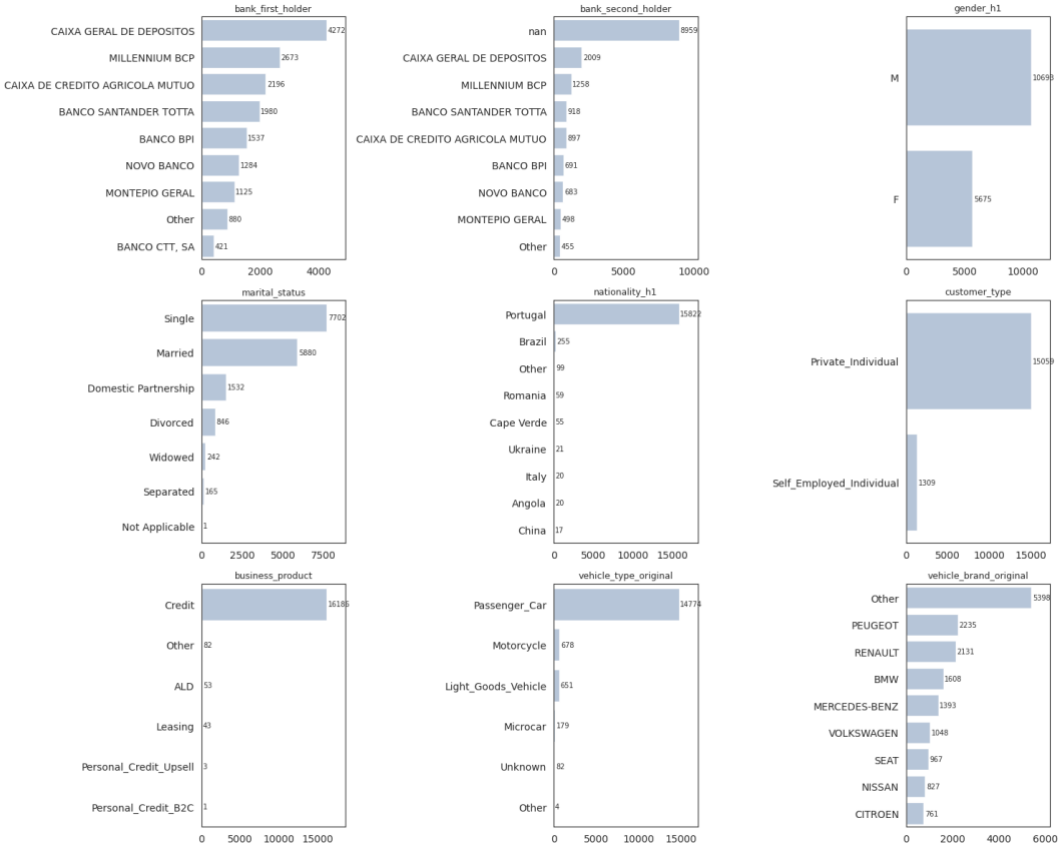


Figure 52: Overview of Customer Demographic Profiles Cross-Sell

Figure 53: Categorical Features vs. Simulation 1 Cross-Sell

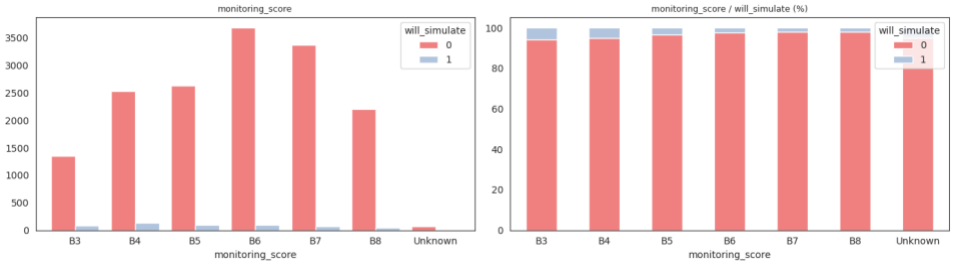


Figure 54: Monitoring Score Cross-Sell

Figure 55: Categorical Features vs. Simulation 2 Cross-Sell

Logit Regression Results

Dep. Variable:	will_simulate	No. Observations:	16367
Model:	Logit	Df Residuals:	16363
Method:	MLE	Df Model:	3
Date:	Sun, 14 Dec 2025	Pseudo R-squ.:	0.05530
Time:	10:56:04	Log-Likelihood:	-2187.4
converged:	True	LL-Null:	-2315.5
Covariance Type:	nonrobust	LLR p-value:	3.128e-55

	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.7062	0.082	-33.015	0.000	-2.867	-2.546
previous_offers	-0.1732	0.017	-10.190	0.000	-0.207	-0.140
customer_tenure_months	0.0022	0.001	1.628	0.104	-0.000	0.005
interaction_offers_tenure	0.0006	0.000	3.642	0.000	0.000	0.001

Figure 56: Logit Model Estimation Results Cross-Sell

	lifecycle_stage	n_customers	simulate_rate
0	Late	3421	0.023385
1	Mid	5108	0.023297
2	Early	7765	0.041211

Figure 57: Simulation Rate Across Loan Lifecycle Stages Cross-Sell

cluster	previous_offers	pre_approved_amount	disposable_income	crc_institutions	age_h1	customer_tenure_months	monitoring_score
0	15.034105	7584.057777	543.372591	3.328073	48.912264	83.044670	6.767286
1	3.784589	7256.130484	506.070588	2.787627	38.338020	29.091676	4.857818

--- Simulation Rate by Cluster ---

cluster	will_simulate
1	0.041845
0	0.020195

Figure 58: Cluster Characteristics k-means

Logit Regression Results

Dep. Variable:	will_simulate	No. Observations:	16294
Model:	Logit	Df Residuals:	16285
Method:	MLE	Df Model:	8
Date:	Sun, 14 Dec 2025	Pseudo R-squ.:	0.05564
Time:	10:56:12	Log-Likelihood:	-2171.5
converged:	True	LL-Null:	-2299.5
Covariance Type:	nonrobust	LLR p-value:	9.687e-51

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-3.2585	0.350	-9.305	0.000	-3.945	-2.572
C(lifecycle_stage)[T.Mid]	0.3973	0.411	0.967	0.334	-0.408	1.203
C(lifecycle_stage)[T.Early]	0.5132	0.360	1.425	0.154	-0.193	1.219
previous_offers	-0.0921	0.014	-6.611	0.000	-0.119	-0.065
C(lifecycle_stage)[T.Mid]:previous_offers	-0.0391	0.021	-1.867	0.062	-0.080	0.002
C(lifecycle_stage)[T.Early]:previous_offers	-0.0888	0.027	-3.350	0.001	-0.141	-0.037
customer_tenure_months	0.0074	0.003	2.178	0.029	0.001	0.014
C(lifecycle_stage)[T.Mid]:customer_tenure_months	-0.0022	0.004	-0.501	0.616	-0.011	0.006
C(lifecycle_stage)[T.Early]:customer_tenure_months	-0.0021	0.004	-0.542	0.587	-0.010	0.005

Logit Regression Results

Dep. Variable:	will_simulate	No. Observations:	16294
Model:	Logit	Df Residuals:	16289
Method:	MLE	Df Model:	4
Date:	Sun, 14 Dec 2025	Pseudo R-squ.:	0.05282
Time:	10:56:12	Log-Likelihood:	-2178.0
converged:	True	LL-Null:	-2299.5
Covariance Type:	nonrobust	LLR p-value:	2.181e-51

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-2.7480	0.170	-16.154	0.000	-3.081	-2.415
C(lifecycle_stage)[T.Mid]	-0.1264	0.152	-0.831	0.406	-0.424	0.172
C(lifecycle_stage)[T.Early]	-0.1236	0.152	-0.812	0.417	-0.422	0.175
previous_offers	-0.1249	0.010	-13.044	0.000	-0.144	-0.106
customer_tenure_months	0.0048	0.001	3.636	0.000	0.002	0.007

Figure 59: Extended Logistic Regression Model Campaign Fatigue

--- Top 20 Most Important Features (XGBoost) ---

	Feature	Importance
5	previous_offers	0.039559
4	pre_approved_amount	0.024852
3	crc_institutions	0.019517
26	nationality_h1_Brazil	0.018988
185	cluster_0	0.018906
54	vehicle_type_original_Motorcycle	0.017342
28	nationality_h1_Portugal	0.015987
80	vehicle_brand_original_DACIA	0.015737
23	housing_type_Own_Mortgaged	0.015374
29	bank_first_holder_ACTIVOBANK	0.014925
170	ppp_insurance_True	0.014911
188	has_second_bank_1	0.014786
18	monitoring_score_numeric	0.014027
176	auto_insurance_True	0.013896
24	housing_type_Own_No_Debt	0.013761
6	offer_letter_term	0.013374
43	bank_first_holder_CAIXA DE CREDITO AGRICOLA MUTUO	0.013372
19	marital_status_Previously_Married	0.013026
22	housing_type_Free	0.012909
78	vehicle_brand_original_CITROEN	0.012714

Figure 60: Feature Importance XGBoost

--- Top 20 Most Important Features (Random Forest) ---

	Feature	Importance
5	previous_offers	0.072357
7	amount_financed	0.056380
13	old_loan_vs_market_diff	0.055732
15	remaining_loan_fraction	0.055571
17	offer_installment_ratio	0.054943
1	disposable_income	0.051205
16	debt_to_income_ratio	0.050785
3	crc_institutions	0.047510
2	customer_tenure_months	0.046454
11	wealth_ipc	0.045663
0	age_h1	0.044842
12	wealth_fdr	0.044139
10	vehicle_age	0.034124
14	age_h2	0.028848
18	monitoring_score_numeric	0.026661
4	pre_approved_amount	0.023495
26	nationality_h1_Brazil	0.015993
9	number_of_dependents	0.011448
24	housing_type_Own_No_Debt	0.011260
22	housing_type_Free	0.010303

Figure 61: Feature Importance Random Forest

--- Logistic Regression Feature Importances (Top 20 from RFE) ---

	Feature	Coefficient	Absolute_Coefficient
13	fuel_type_original_Mild_Hybrid_Diesel_Electric	-2.080943	2.080943
6	bank_first_holder_CAIXA AGRICOLA DE TORRES VEDRAS	-1.901009	1.901009
18	profession_contract_type_Self_Employed	-1.827421	1.827421
8	bank_first_holder_CAIXA DE CREDITO AGRICOLA MU...	1.684619	1.684619
7	bank_first_holder_CAIXA AHORROS GALICIA	-1.452900	1.452900
5	bank_first_holder_CAIXA AGRICOLA DE LEIRIA	-1.384029	1.384029
2	nationality_h1_Brazil	1.286725	1.286725
16	customer_type_Self_Employed_Individual	1.232159	1.232159
9	bank_first_holder_CAIXA ECONOMICA MISERICORDIA...	1.092625	1.092625
12	fuel_type_original_Electric_Range_Extender	0.996851	0.996851
11	fuel_type_original_Diesel_Electric_Plugin	0.969600	0.969600
4	bank_first_holder_BANCO PORTUGUES DE NEGOCIOS	0.904536	0.904536
15	customer_type_Private_Individual	-0.898745	0.898745
0	previous_offers	-0.808762	0.808762
10	vehicle_type_original_Motorcycle	0.802702	0.802702
3	bank_first_holder_BANCO BEST	-0.681542	0.681542
14	partner_type_original_Other	-0.649407	0.649407
1	housing_type_Own_No_Debt	0.640212	0.640212
17	profession_contract_type_No_Contract	0.635354	0.635354
19	has_second_bank_1	0.161032	0.161032

--- Top 10 Positive Coefficients (Increasing log-odds of Simulation) ---

	Feature	Coefficient	Absolute_Coefficient
8	bank_first_holder_CAIXA DE CREDITO AGRICOLA MU...	1.684619	1.684619
2	nationality_h1_Brazil	1.286725	1.286725
16	customer_type_Self_Employed_Individual	1.232159	1.232159
9	bank_first_holder_CAIXA ECONOMICA MISERICORDIA...	1.092625	1.092625
12	fuel_type_original_Electric_Range_Extender	0.996851	0.996851
11	fuel_type_original_Diesel_Electric_Plugin	0.969600	0.969600
4	bank_first_holder_BANCO PORTUGUES DE NEGOCIOS	0.904536	0.904536
10	vehicle_type_original_Motorcycle	0.802702	0.802702
1	housing_type_Own_No_Debt	0.640212	0.640212

Figure 62: Feature Importance Logistic Regression

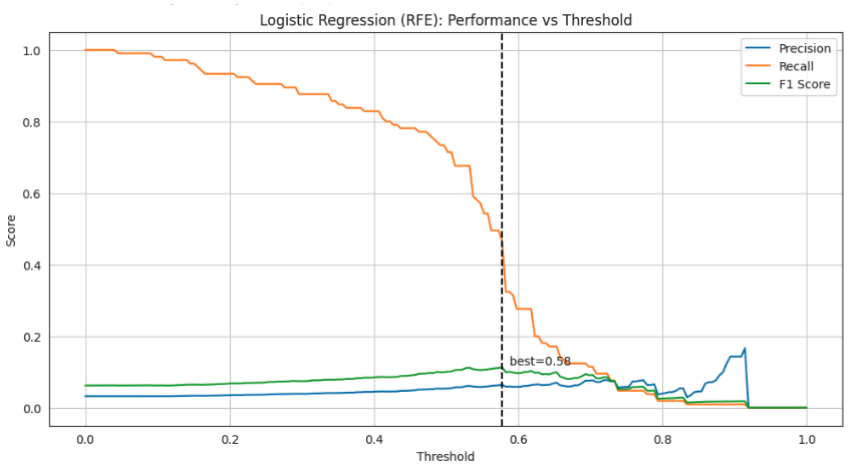


Figure 63: Logistic Regression Performance Threshold

```
Random Forest Evaluation (Top 20%)
Target Depth:            20% (Top 654 customers)
AUC-PR (Avg Prec):       0.0753
AUC-ROC:                 0.7166
Precision @ 20%:         0.0826
Recall @ 20%:            0.5143
F1-Score @ 20%:          0.1421

Classification Report (Top 20%):
              precision    recall  f1-score   support

           0       0.98      0.81      0.89      3169
           1       0.08      0.51      0.14       105

    accuracy                           0.80      3274
   macro avg       0.53      0.66      0.51      3274
weighted avg       0.95      0.80      0.86      3274


Confusion Matrix (Top 20%):
[[2568  601]
 [  51   54]]
```

Figure 64: Evaluation Matrix Random Forest

```
XGBoost Evaluation (Top 20%)
Target Depth:             20% (Top 654 customers)
AUC-PR (Avg Prec):        0.1007
AUC-ROC:                  0.7665
Precision @ 20%:          0.0902
Recall @ 20%:             0.5619
F1-Score @ 20%:           0.1553

Classification Report (Top 20%):
              precision    recall  f1-score   support

           0       0.98      0.81      0.89      3169
           1       0.09      0.56      0.16       105

    accuracy                           0.80      3274
   macro avg       0.54      0.69      0.52      3274
weighted avg       0.95      0.80      0.87      3274


Confusion Matrix (Top 20%):
[[2573  596]
 [  46   59]]
```

Figure 65: Evaluation Matrix XGBoost

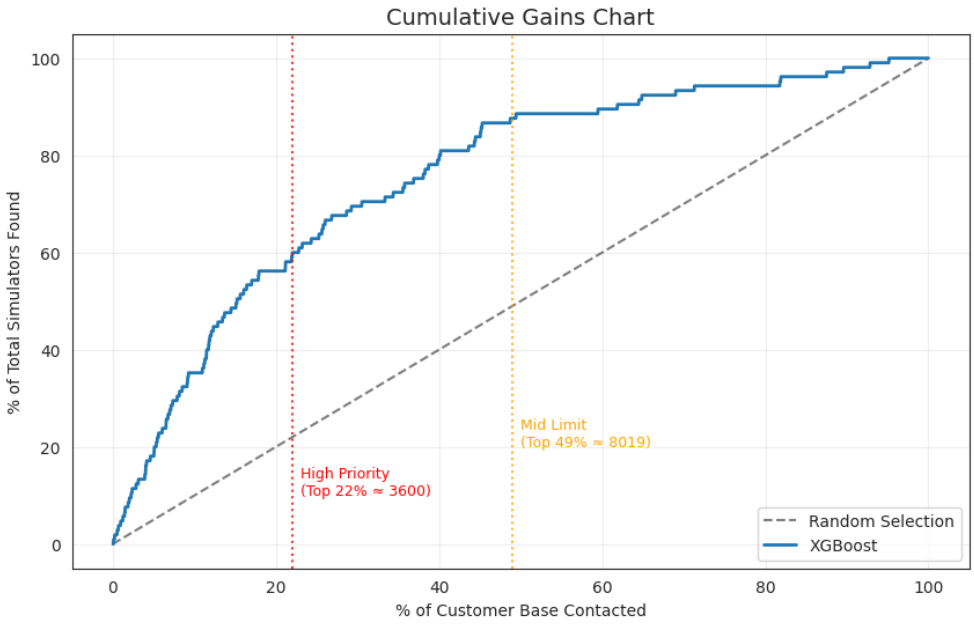


Figure 66: Cumulative Gains Chart

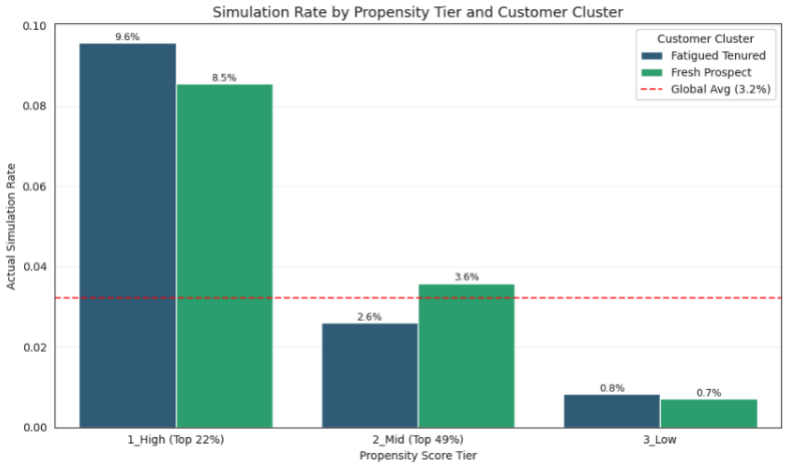


Figure 67: Simulation Rate by Propensity Tier and Customer Cluster

--- XGBoost (Class Weighting) Summary ---

	Campaign Target	Campaign Size (# N)	Est. Simulators Found (# N)	Simulation Rate (%)	Capture Rate (% Recall)
0	Top 1%	159	24	0.156	0.048
1	Top 5%	814	94	0.117	0.181
2	Top 10%	1634	184	0.113	0.352
3	Top 15%	2454	254	0.104	0.486
4	Top 20%	3269	294	0.090	0.562
5	Top 30%	4909	364	0.074	0.695
6	Top 50%	8183	464	0.057	0.886
7	Top 100%	16366	524	0.032	1.000

Figure 68: Gain Analysis

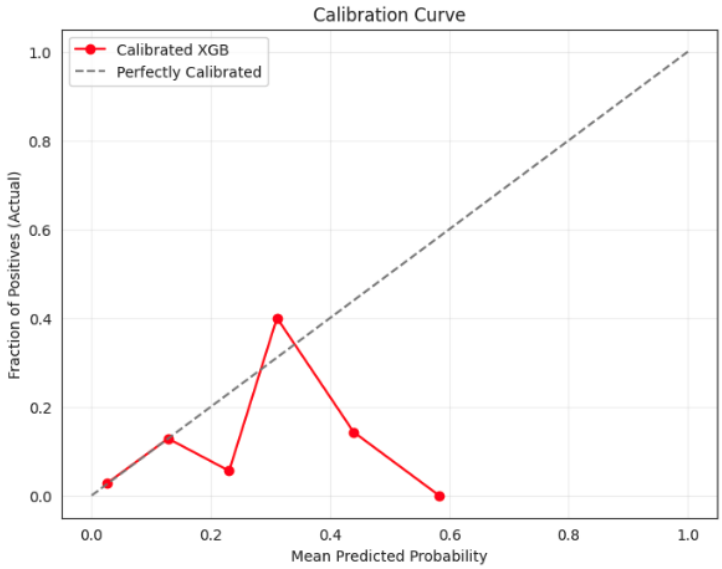


Figure 69: Calibration Curve

--- Profiling: 1_High (Top 22%) ---

Cluster	Fatigued Tenured	Fresh Prospect
age_h1	50.838235	40.760684
disposable_income	271.960862	447.448376
customer_tenure_months	99.617647	28.863248
crc_institutions	4.823529	3.608547
pre_approved_amount	5650.735294	6403.418803
previous_offers	4.470588	1.459829
offer_letter_term	69.794118	70.994872
amount_financed	17867.410155	16604.293219
auto_loan_installment_due_date	10.625000	9.384615
number_of_dependents	0.882353	0.647863
vehicle_age	10.108527	9.693493
wealth_ipc	101.949706	98.766342
wealth_fdr	0.045449	0.026467
old_loan_vs_market_diff	0.071316	0.069645
age_h2	53.574879	46.674522

--- Profiling: 2_Mid (Top 49%) ---

Cluster	Fatigued Tenured	Fresh Prospect
age_h1	51.103704	38.972313
disposable_income	434.440793	481.105038
customer_tenure_months	82.114815	27.889251
crc_institutions	4.048148	2.811075
pre_approved_amount	6720.370370	7377.850163
previous_offers	10.622222	3.794788
offer_letter_term	70.088889	71.335505
amount_financed	16804.884023	15794.514062
auto_loan_installment_due_date	9.722222	9.079805
number_of_dependents	0.788889	0.418567
vehicle_age	9.958647	9.176183

Figure 70: Cohort Based Profiling

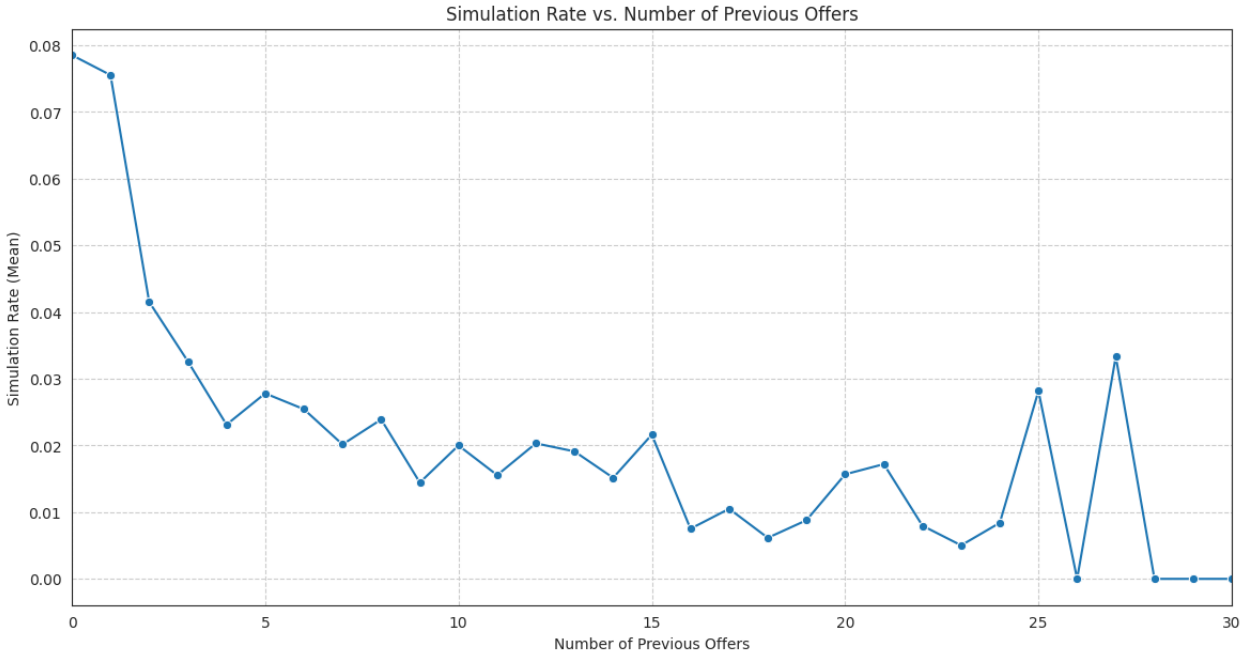


Figure 71: Simulation Rate vs. Number of Previous Offers