What Drives Cash Advance Usage?
Unveiling Cash Advance Usage Patterns Through Machine Learning

BANU MUTLU MEMIS and BEN SOPRANZETTI*

ABSTRACT

This study leverages machine learning to predict cash advance usage—one of the costliest forms of credit—among bank customers, utilizing a proprietary dataset comprising 180 variables. The model achieved a remarkable 90% predictive accuracy in distinguishing cash advance users. The findings reveal that financial stability, effective financial management, higher education, and stable employment significantly decrease the likelihood of cash advance usage. Conversely, indicators of financial strain, such as high loan payments, frequent overdraft occurrences, and younger age, increase reliance on cash advances. Gender influences were also observed, with males being less likely to utilize cash advances.

* Ben Sopranzetti, a distinguished professor at Rutgers Business School with over 30 years of experience in finance, serves as the Vice Chair of the Finance and Economics Department. Banu Mutlu Memis brings 23 years of expertise from her tenure at the Banking Regulation and Supervision Authority of Turkey. The authors express their gratitude to Denizbank for providing the data used in this study. We have no conflicts of interest.

CASH ADVANCES OFFER QUICK ACCESS to funds but often come with high interest rates and fees. Understanding the factors influencing cash advance usage is crucial because it enables financial professionals to develop better financial products and enhance credit risk assessment and management practices. Additionally, it provides valuable insights for academia to understand consumer behavior and contribute to the broader knowledge in the field of finance. By identifying the drivers behind cash advance usage, financial institutions can create solutions that support healthier financial choices for consumers and improve the overall economic environment.

Despite the profound impact of cash advances on consumer finances, there is a notable absence of advanced predictive analytics focused on forecasting consumer behavior regarding cash advances, particularly through machine learning models. Additionally, data availability has posed a significant challenge in this area. This study addresses these gaps by employing the powerful XGBoost algorithm, a highly regarded machine learning approach, and overcomes the data challenge by utilizing an extensive and comprehensive real bank dataset encompassing a full spectrum of customer-specific variables. By leveraging advanced machine learning technique and utilizing this extensive dataset, this research aims to provide robust predictive insights and significantly enhance the understanding of cash advance usage. This contribution is poised to impact both academic literature and practical financial strategies profoundly.

In this context, the use of predictive analytics through XGBoost to analyze cash advance usage among bank customers aligns directly with the broader objectives of both fintech and traditional financial services. Understanding the demographic and behavioral factors that influence cash advance usage can help financial institutions develop more effective credit risk assessment models, design financial products that better serve customer needs, and implement strategies to mitigate financial distress among consumers. Thus, this study underscores the pivotal role of advanced analytics in transforming financial services and enhancing the precision and efficacy of financial strategies.

This study makes a significant contribution to the literature on customer behavior analysis by incorporating 180 critical customer-specific risk factors and utilizing a real bank dataset with 200,000 observations to build a high-performing model for predicting cash advance usage among

bank customers. The variables included in the model cover customers' financial circumstances, such as income irregularity and spending habits, supplemented by demographic variables like gender, age, marital status, employment type, and customer city of the bank branch. This comprehensive approach offers a holistic view of the factors influencing cash advance usage. To our knowledge, this is the first study to leverage such an extensive real bank dataset with customer-specific variables to analyze cash advance behaviors.

By employing a hybrid approach that combines rolling and recursive schemes for sample splitting, our model adapts to new information while maintaining stability and accuracy. The primary empirical finding is that this model delivers robust predictions for identifying cash advance users, achieving a high predictive accuracy of 90%. This exceptional level of accuracy provides scholars with a deeper understanding of the drivers of cash advance usage and offers valuable insights for financial institutions. By enhancing credit risk assessment and management practices, financial institutions can develop more effective financial products and strategies. Additionally, these insights can inform the refinement of existing financial products to better serve customer needs and mitigate financial distress.

This study delves into the demographics and behaviors of individuals who resort to such costly financial instruments. Specifically, it investigates whether indicators such as customer demographics, spending patterns, and maxed-out credit limits are predictive of cash advance usage. By understanding these factors, financial institutions can better manage risk and provide more tailored financial products and advice to their customers.

The remainder of the paper is organized as follows. Section I discusses the costly structure of cash advances. Section II positions our contribution within the related literature. Section III outlines the data structure and provides population analyses. Section IV details the methodology, including the use of the XGBoost algorithm, data collection, and model implementation. Section V presents empirical results. Section VI discusses the variable importance results of the SHAP analysis. Finally, Section VII provides concluding remarks.

I. The Cost Dynamics of Cash Advances

A credit card cash advance allows cardholders to withdraw a portion of their credit limit in cash, either through an ATM, directly from a bank, or via a mobile app. This feature provides immediate liquidity but comes with significantly higher costs compared to typical credit card transactions. It is often more expensive than overdrafts and usually more costly than other credit products. The process involves the cardholder requesting a certain amount within their available cash advance limit, subject to approval based on their credit standing. Upon authorization, the cardholder can withdraw cash up to a specified limit, generally a fraction of their total credit line. While these principles are generally applicable, specific terms, costs, and availability can vary by country and financial institution.

The financial implications of utilizing cash advances are substantial due to the associated fees and interest rates. Each transaction typically incurs a cash advance fee, which can be a flat rate or a percentage of the amount withdrawn. Unlike regular credit card purchases, the interest on cash advances begins accruing immediately, with rates that are typically higher than those applied to purchases. Additionally, transactions conducted through an ATM often incur additional ATM fees, further increasing the cost of the advance. Given these factors, while cash advances provide access to immediate funds, the associated high costs make them a less favorable option compared to alternatives such as overdrafts, personal loans, or credit lines, which generally offer lower interest rates.

Analyzing the cash advance fees provided by the data provider bank for the modeling period of 2017-2020, we found that the annual percentage rate (APR) equates to roughly 36%, significantly higher than the typical inflation rates during the same periods, which ranged from about 12% to 20%. This stark contrast indicates that the fees associated with cash advances are substantially higher than the inflation rate, significantly increasing the cost of borrowing and exacerbating financial strain on consumers. Agarwal, Sumit, Driscoll, John C., Gabaix, Xavier, and Laibson,

David (2008) also emphasize the costly structure of cash advances, stating that the APR on cash advances is typically greater than that on purchases and is usually 16 percent or more.

The high costs associated with cash advances underscore that, while providing immediate liquidity, they come at a significant expense that can substantially diminish the financial stability and long-term financial health of customers. Given these costs, cash advances should be considered a last-resort financing option.

II. Contribution to the Literature

The landscape of consumer finance has seen significant advancements in recent years, particularly with the integration of fintech. Despite these developments, there is a notable absence of advanced predictive analytics specifically focused on forecasting consumer behavior concerning cash advances, which have a significant impact on consumer finances. While direct literature on cash advance usage is sparse, related studies on consumer behavior and credit card usage provide valuable context. Understanding these areas can shed light on the dynamics of cash advance usage and offer a foundation for developing more effective predictive models.

Gross and Souleles (2002) provide valuable insights into consumer borrowing patterns through their studies on credit card usage and debt behavior, demonstrating that changes in credit limits and interest rates significantly influence credit card usage. This underscores the dynamic nature of consumer credit behavior. Suoranta and Mattila (2004) highlight that future users of mobile banking services are likely to be older than expected, and wealthier respondents are less inclined to adopt new mobile banking services. Morgan et al. (2012) analyze the burden of high-interest loans on consumers, emphasizing the need for improved predictive tools to manage these products.

In their study, Pulina (2011) analyzed the demographic, socio-economic, and banking-specific determinants influencing the choice of credit cards among Italian banking customers. Utilizing a multinomial logit model, Pulina identified several trends: women, older individuals, residents in central Italy, and secondary card owners are more inclined to acquire classic credit cards. Older customers tend to prefer gold cards, while younger clients are more inclined to opt for revolving

cards, a preference also seen among residents in the Northeast. This study provides a microeconomic perspective on consumer behavior and preferences, shedding light on the structural characteristics of the card payment system. Complementing this, Dagobert L. Brito and Peter R. Hartley (1995) investigated the behavior of consumers who borrow on credit cards despite high-interest rates. Their analysis reveals that rational consumers may finance significant portions of their consumption with credit card debt, especially when facing uncertainties in transaction demands or attempting to smooth income and expenditure streams. These findings help explain the persistence of high credit card interest rates and further illuminate the complex dynamics of consumer credit behavior.

Ahmed et al. (2010) explored the psychographic factors affecting credit card attitudes among Malaysian consumers, finding that lifestyle changes and socio-cultural differences are significant influencers. This aligns with Önder's (2018) investigation of the impact of credit card usage on household expenditures, affirming that income levels significantly correlate with credit card spending, suggesting that socio-economic status is a robust predictor of credit behavior. Staten and Barron (2002) provided insights into college student credit card usage, highlighting cautious approaches by issuers towards younger consumers. Jung and Kang (2021) examined the elderly in Korea, identifying income level and non-retirement status as predictors of more extensive credit card ownership and usage, indicating varying behavior across age demographics. Importantly, this analysis leaves the drivers of cash advance usage of customers unexplained.

Koparal and Calik (2014) emphasized the dual role of credit cards as status symbols and potential financial risks in Turkey, focusing on how economic structures and consumer earnings influence usage patterns. Similarly, De Silva and Patabendige (2021) in Sri Lanka found a significant positive relationship between credit card attributes and usage, stressing the importance of lifestyle and social status in shaping consumer behavior. Kasaian, Murthi, and Steffes (2023) differentiated between opportunistic and needy credit card users, indicating that different consumer segments exhibit distinct behaviors influenced by financial conditions and promotional offers. This segmentation suggests that predictive models should account for varying consumer sensitivity to credit terms. Uddin (2020) highlighted the role of credit card literacy in India, arguing

that higher credit card literacy correlates with more prudent credit usage, suggesting that financial education could mitigate risky credit behaviors. Lin et al. (2019) further advocated for policy interventions in China to manage credit risk, emphasizing stricter credit issuance policies and better consumer education to prevent over-indebtedness.

The growing prevalence of credit card usage across diverse socio-economic landscapes has prompted extensive research into consumer behaviors and the determinants influencing these patterns. Recent studies have highlighted several dimensions, including psychological, demographic, and economic factors that affect credit card usage and resultant debt behavior. Despite this extensive body of research, there remains a notable gap in advanced predictive analytics specifically focused on cash advance usage. Predicting cash advance usage is crucial as it helps in understanding customer behavior, managing financial risks, and tailoring financial products to better meet diverse customer needs. Our study addresses this gap by employing the XGBoost model and utilizing a comprehensive dataset that includes a full range of financial and demographic variables. This approach provides a robust analysis and valuable insights into the factors driving cash advance usage, offering a novel contribution to the existing literature.

The XGBoost algorithm, introduced by Chen and Guestrin (2016), has become widely recognized for its exceptional predictive performance and ability to manage complex datasets. Research on predictive analytics in finance underscores the potential of machine learning models, such as XGBoost, in predicting consumer financial actions and improving credit risk assessments. Studies by Li et al. (2020), Malhotra and Malhotra (2020) have demonstrated that XGBoost outperforms Logistic Regression, Decision Trees, Random Forest, and Gradient Boosted Decision Trees in terms of predictive performance and computational efficiency. Furthermore, research by Li (2019), Chang et al. (2018), and Darapaneni et al. (2022) consistently highlights the superior predictive capabilities and robustness of XGBoost. In line with these findings, our study employs XGBoost to model the cash advance usage of bank customers.

A significant limitation in the existing literature is the potential biases arising from small datasets. Hand (2006) and Finlay (2011) underscore the critical need for comparison datasets that closely reflect real-world conditions to ensure external validity. To address this limitation, our

study leverages a proprietary, comprehensive dataset provided by a commercial bank. This extensive dataset mitigates the challenges associated with data availability and allows for a more thorough and accurate analysis. By utilizing this robust dataset, our research enhances the understanding of the factors driving cash advance usage, providing valuable insights that benefit both academic research and practical applications within financial institutions.

Imbalanced datasets are common in cash advance prediction due to the disparity between cash advance users and non-users. Fitzpatrick and Mues (2016) discuss the challenges posed by imbalanced datasets and the importance of addressing them to improve accuracy. Weiss and Provost (2003) suggest that alternative class distributions, such as oversampling or undersampling methods, can enhance model performance. In our study, we employ an undersampling technique to balance the class distribution, improving the model's ability to predict cash advance usage.

By leveraging the advanced machine learning technique XGBoost and utilizing a comprehensive dataset, financial institutions can significantly enhance their predictive analytics, leading to more accurate and efficient risk management practices. This approach not only advances the academic literature by providing deeper insights into consumer behavior concerning cash advances, but also equips fintech companies and traditional financial institutions with practical tools to better understand and manage this behavior. The findings from this study offer substantial value to academicians by filling a critical gap in the existing research and paving the way for future studies in predictive analytics and consumer finance.

III. Data Structure and Population Analyses

In this study, we leverage a proprietary dataset provided by Denizbank (Bank), one of Türkiye's largest commercial banks. To ensure privacy and confidentiality, the Bank has removed all personal identification information, and no customer IDs are included in the dataset. This comprehensive dataset spans a four-year period from January 2017 to December 2020, capturing 48 months of data. It includes 200,000 observations, randomly sampled from the entire population of 159 million observations, and consists of 180 customer-specific variables, offering an extensive overview of customer financial behavior and demographics.

Our analysis focuses on individual customers within the retail portfolio. To prevent bias, we included all customers who became bank customers after the training period began, as well as those who left the bank before the end of the training period. We use a binary dependent variable to identify significant cash advance usage, assigning a value of 1 if a customer uses a cash advance above $10 and 0 otherwise. This threshold helps distinguish substantial cash advance usage from minor, possibly incidental transactions.

Our model incorporates a total of 180 independent variables (predictors) covering a broad spectrum of customer-specific features, as categorized in Table I. The dataset provides daily and monthly records of customers' financial statuses and spending behaviors, encompassing aspects such as payments, delinquency, limit utilization, and other financial activities. In addition to financial circumstances and spending habits, our dataset includes various dimensions such as employment status, gender distribution, marital status, age range, and geographical distribution across Turkish districts.

Table I

Variable Categories

Thus table categorizes the 180 independent variables used in our model, providing a detailed view of the customer-specific features contributing to predicting significant cash advance usage. Based on these customer-specific variables, we grouped them into several categories: Demographic Information, Financial Status, Transaction Activities, Loan Activities, Spending Habits, Asset and Liability Changes, Electronic Fund Transactions, and Bank Activities and Engagement. Each category encompasses specific dimensions of customer behavior and characteristics.

Category	Description
Demographic Information	Information capturing the personal background of customers, including their education level, marital status, gender, and employment type.
Financial Status	Data that reflects the financial health of customers, such as their total assets, liabilities, deposits, and loan balances.
Transaction Activities	Details on customers' transactional behavior, including the frequency and amounts of various transactions such as bill payments, tax payments, and POS transactions.
Loan Activities	Comprehensive details about customers' loan behaviors, focusing on payment history and patterns in installment repayments.
Spending Habits	Variables related to customer spending behavior, particularly focusing on the frequency and amount of their expenditures.

Asset, Liability Changes	A snapshot of the customer's financial condition, comparing recent figures to historical maximums, detailing changes in assets, liabilities, and income.
Electronic Fund Transactions	Data on the frequency and amounts of electronic fund transactions, such as those made with debit cards and through ATMs.
Bank Activities and Engagement	Insights into how customers interact with bank services, covering aspects like account activities, transaction volumes, and internet banking usage.

This study is the first of its kind to analyze the drivers of cash advance usage using such a comprehensive dataset. It reveals a heterogeneous population characterized by diverse work types, educational backgrounds, genders, ages, marital statuses, and branch locations.

The demographic analysis of the sample provides valuable insights into the socioeconomic composition of our study cohort. The distribution of employment types reveals a diverse range of statuses, with nearly half of the sample engaged in paid employment and a substantial portion in retirement. The gender distribution indicates a predominance of male participants, with significant female representation. Marital status diversity is notable, with a predominant proportion of married individuals, underscoring the critical role of household dynamics in financial decision-making processes. The age distribution spans a broad spectrum, with the majority of participants falling within the middle-aged brackets. The geographic dispersion across different regions of Turkey allows for an exploration of regional variations in financial behaviors and access to credit facilities. This diversity underscores the complex and multifaceted nature of factors influencing cash advance usage, offering crucial insights into consumer financial behavior.

Table II

Distribution of Sample by Cash Advance Usage

In this table, the distribution of the sample with respect to cash advance usage is concisely outlined. This representation provides a clear breakdown of the percentage of customers within our dataset who use cash advances versus those who do not. This table indicates a significant skew towards customers not utilizing cash advances, with 91.6% of the sample not using this financial service. Conversely, 8.4% of the sample has utilized cash advances.

Category	Number of Samples	Percentage
NON-CASH ADVANCE USAGE	184,082	91.6%
CASH ADVANCE USAGE	16,880	8.4%
Total	**200,962[1]**	**100%**

[1] By randomly excluding 962 records, we build our model based on 200,000 samples.

This distribution underscores the relative rarity of cash advance usage, emphasizing the importance of understanding the factors driving this behavior.

IV. Methodology and Implementation of the Model

All modeling processes are conducted using a machine learning methodology in Python. The model utilized in this study is built on XGBoost, an ensemble machine learning algorithm that employs a gradient boosting framework and is based on decision trees. XGBoost is specifically designed to address both classification and regression problems. This model leverages a gradient-based approach to iteratively refine predictions and uses a method of assessing the reduction in the objective function to determine the best split points in the decision trees, enhancing its efficiency and performance.

The objective function for our binary classification model, which predicts cash advance usage, is defined using log loss:

$$C(\beta)= \sum_{i=1}^{n} -(a_i \log (p_i) + (1 - a_i) \log (1 - p_i)) + \Omega(L) \qquad (1)$$

where:

- a_i is the actual label of the i-th sample, with 1 indicating cash advance usage and 0 indicating non-cash advance usage.
- p_i is the predicted probability of cash advance usage for the i-th sample.
- Ω(L) stands for the regularization term, which can include L1 regularization (Lasso), L2 regularization (Ridge), or both (Elastic Net) to prevent overfitting.

The log loss function evaluates the performance of a classification model that produces probability values ranging from 0 to 1. It penalizes the model for incorrect predictions, with larger penalties for confident but incorrect predictions. This makes log loss a robust choice for binary classification problems, ensuring the model not only predicts the correct class but also assigns a high probability to it.

Regularization is a crucial component of the objective function, adding a penalty to prevent overfitting. Overfitting occurs when the model learns noise and details in the training data, negatively impacting its performance on new data. Techniques such as L1 and L2 regularization help maintain a balance between fitting the training data well and generalizing to unseen data.

To assess the effectiveness of the Cash Advance Usage Prediction model, we employ several performance metrics, including the Confusion Matrix, Recall, ROC-AUC, Gini Index, Precision, and F1 Score. Formulations and explanations for these performance metrics are detailed in Appendix A. Given our primary focus on predicting which customers are likely to use cash advances, we prioritize recall to ensure the capture of as many true positive instances (cash advance users) as possible. This approach may result in lower precision, leading to a higher number of false positives. However, in this context, the model's utility is enhanced by achieving better recall values.

The significance of minimizing Type 2 (False Negative) errors in this classification problem cannot be overstated. Missing potential cash advance users (false negatives) can have more critical consequences than incorrectly identifying non-users as users (false positives). Therefore, our model aims to achieve a high True Positive Rate[2] and a low False Negative Rate[3]. This focus on reducing Type 2 errors ensures that potential cash advance users are accurately identified, thereby maximizing the practical value and effectiveness of the model in real-world applications.

By maintaining a strong recall, we ensure that our model effectively identifies the target group of cash advance users, even at the expense of precision. This strategy aligns with our objective of capturing as many true positive instances as possible, ultimately leading to better decision-making and resource allocation in predicting cash advance usage.

The development of our model incorporates several critical steps to ensure optimal performance:

[2] The True Positive Rate assesses the model's capability to detect true positives, which are the correctly predicted cash advance users, out of all actual cash advance users.

[3] The False Negative Rate measures the proportion of actual cash advance users that the model incorrectly classifies as non-users. It quantifies the model's tendency to miss true positives, i.e., the cash advance users who were not identified as such by the model.

A. Model Preprocessing

In the data preprocessing stage of our study, we eliminated variables that were entirely null or displayed zero variance. This refinement was essential to ensure that only meaningful features contribute to the model's performance. Given the robust characteristics of XGBoost, a tree-based boosting algorithm, additional feature scaling was deemed unnecessary. XGBoost's structure inherently manages different scales of data and exhibits a high tolerance for outliers, further validating our decision against performing extensive feature selection. This approach leverages the strengths of XGBoost, facilitating efficient processing and robust predictive performance without the need for extensive feature manipulation beyond the removal of non-informative variables.

In this step, we also address data balancing and employ a hybrid approach for model training:

A.1. Unbalanced Data Adjustment - Data Balancing

Given the significant imbalance in our dataset—where 91.6% of the 200,000 observations are classified as 'non-cash advance usage' and only 8.4% as 'cash advance usage', creating an imbalance ratio of approximately 12:1—we adopted an under-sampling strategy to address this skew. Of the total 200,000 records, 40,000 are reserved as unseen data for final evaluation to ensure an unbiased assessment of the model's performance. The remaining 160,000 records are divided, with 24,000 records (12,000 'non-cash advance usage' and 12,000 'cash advance usage') specifically designated for model training to ensure balanced exposure to both classes. Appendix B provides the distribution of the model sample based on selected demographics, offering insights into the composition of the dataset.

A.2. Employing Hybrid of Rolling and Recursive Schemes

In our predictive modeling, we employ a hybrid approach to sample splitting, integrating elements from both rolling and recursive schemes. This method, which builds on concepts utilized by Gu, Kelly, and Xiu (2020), allows our model to adapt over time to changes in economic conditions or customer behavior, thereby enhancing the accuracy and relevance of its predictions.

Furthermore, this approach ensures that the model remains robust and responsive to dynamic changes in the dataset and external conditions. Key Features of the Hybrid Rolling-Recursive Scheme:

(i) **Rolling Scheme:** Incrementally advances the training window by one period for each new prediction cycle. This ensures the model is consistently trained on the most recent data, continuously adapting to current trends and developments.

(ii) **Recursive Scheme:** Incrementally expands the training sample as new data becomes available, incorporating all historical data into the training set. This allows the model to learn from an increasingly larger dataset, accumulating knowledge over time.

By integrating these schemes, our model capitalizes on the adaptability of the rolling scheme and the comprehensive learning ability of the recursive scheme. The hybrid rolling-recursive scheme involves periodically refitting the model with an expanded training sample and using this refitted model to make predictions over a subsequent period. This approach is particularly effective for datasets where historical patterns and behaviors evolve over time, necessitating the model's adaptation to new trends and data distributions.

Table III

Expanding Window Periods for Cash Advance Prediction Model

The timeframe of the expanding window for our Cash Advance Prediction Model is provided in Table III. Our modeling period spans four distinct timeframes between January 2017 and December 2020. The training dataset starts from a fixed point in January 2017 and expands with each iteration by adding six months of data, ensuring the model continuously learns from an increasing amount of historical information. The validation and test sets are each fixed at three months duration and follow each training period to evaluate model performance and prevent overfitting.
Training Set: Begins from January 2017 and extends by six months with each iteration, ensuring the model benefits from an expanding dataset and continually incorporates added information.
Validation Set: A rolling window of three months immediately follows the training set, used for parameter tuning and model validation. This fixed duration helps prevent overfitting and ensures robust model performance.
Test Set: Another fixed rolling window of three months, used for evaluating the model's performance on unseen data immediately following the validation period.

Period	Train Period	Validation Period	Test Period
1	01-2017/12-2018	01-2019/03-2019	04-2019/06-2019

2	01-2017/06-2019	07-2019/09-2019	10-2019/12-2019
3	01-2017/12-2019	01-2020/03-2020	04-2020/06-2020
4	01-2017/06-2020	07-2020/09-2020	10-2020/12-2020

The model is periodically refitted twice a year, incorporating the latest data and trends to maintain its relevance and accuracy. This approach ensures that the model is trained on a progressively larger dataset while validating and testing on unseen data, thereby preserving the integrity of model evaluation. Regular refitting allows the model to adapt to new developments in customer behavior or economic conditions that may influence cash advance usage. The expanding window approach provides a broad and deep learning curve, enhancing the model's ability to generalize to new, unseen data and improving the reliability and robustness of our cash advance usage predictions.

Table IV

Distribution of Cash Advance Usage in Each Period

The distribution of cash advance usage for each period and split type is outlined in this table. The target rates, representing the proportion of cash advance usage, are relatively stable across periods and splits, ensuring the model learns from and is evaluated on consistently representative data. This stability provides balanced representation across training, validation, and test sets, crucial for reliable performance metrics. The target rates range from 0.456 to 0.536, indicating similar proportions of cash advance usage. Such distribution is essential for assessing the model's generalization capabilities and maintaining predictive accuracy on new data. Analyzing target rates across periods helps identify temporal trends or shifts in usage, evaluating the model's performance under varying conditions and its adaptability to changes in data distribution. Maintaining a stable and representative distribution optimizes the model's training and evaluation, leading to robust and generalizable predictions.

Period	Split Type	Target Rate	
1	TRAIN	0.536	
1	VALIDATION	0.505	
1	TEST	0.495	
2	TRAIN	0.528	
2	VALIDATION	0.492	
2	TEST	0.504	
3	TRAIN	0.523	
3	VAL	0.456	
3	TEST	0.505	
4	TRAIN	0.517	

Period	Split Type	Target Rate
4	VAL	0.504
4	TEST	0.508

B. Model Optimization and Finalization

We utilize a grid search approach to determine the optimal hyperparameters for the XGBoost model. This technique systematically explores a range of potential values for each hyperparameter, evaluating model performance across each configuration to find the best fit. Our selection criteria focus on maximizing performance metrics relevant to our study, including recall, Gini and the Area under the ROC curve (AUC).

Table V
Testing Parameter Values to Optimize XGBoost Model

The specific parameter values evaluated in our grid search are outlined in this table. We systematically tested various values for learning rate, max depth, and min child weight across four different periods using an expanding window approach. The optimal parameters from the top-performing period were then employed to train the final model. By selecting the hyperparameters from the period with the best performance, we aimed to enhance the model's predictive capabilities. The grid search systematically explored a range of potential values for each hyperparameter, evaluating model performance across each configuration to identify the optimal settings for our XGBoost model.
Learning Rate: The values tested ([0.01, 0.05, 0.10, 0.20]) control the step size at each boosting iteration. A lower learning rate requires more boosting rounds to converge, while a higher learning rate speeds up the training but risks overshooting the optimal solution.
Max Depth: Values ranging from 5 to 15 were evaluated to determine the maximum levels within each decision tree. Deeper trees can model more complex patterns but may also lead to overfitting.
Min Child Weight: Tested values [1, 2, 3] set the minimum sum of instance weights (Hessian) needed in a child node. Higher values prevent the model from learning relations that might be highly specific to particular samples, thus providing regularization.
N_estimators: The number of boosting rounds was determined using early stopping criteria, which halt training if the model's performance does not improve after a certain number of iterations, preventing overfitting and ensuring optimal model complexity.

Parameter	Description	Values
learning_rate	Adjusts the contribution of each tree by reducing the residual error	[0.01, 0.05, 0.10, 0.20]
max_depth	Maximum depth of a tree	[5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15]
min_child_weight	Minimum sum of instance weight needed at a child node	[1, 2, 3]
n_estimators	Number of trees	Determined by early stopping

By evaluating these parameters across four periods using an expanding window approach, we identified the optimal hyperparameters that yielded the best performance metrics. The results of the grid search are summarized in Table VI, highlighting the performance metrics for the top two runs. The parameters from the best-performing run were then used to train the final model, ensuring improved predictive abilities and robustness. This approach enables the model to balance complexity and generalization effectively, leading to better performance on unseen data.

Table VI

Grid Search Results for Top Performing Runs

This table summarizes the results of the grid search for the best performing runs[4]. As outlined in Table V, the grid search setup involved evaluating four different learning rates (e.g., 0.01, 0.05, 0.10, 0.20), ten different max depth values (e.g., 5 to 14), and three different min child weight values (e.g., 1, 2, 3) across four different evaluation periods. This comprehensive search covered total of 480 runs. Out of these 480 runs, both the 187th and 211th runs achieved the highest accuracy score on the validation set (0.857), indicating that both configurations are equally effective in terms of overall accuracy.

Run Number	Learning Rate	Max Depth	min_child_ weight	F1 Score Val	Recall Score Val	Precision Score Val	Accuracy Score Val
187	0.100	7	2	0.863	0.912	0.819	0.857
211	0.200	5	2	0.862	0.903	0.824	0.857

To build our model, we used the parameters of the 187th run due to its higher recall score compared to the 211th run. Additionally, the higher max depth in this run allows the model to capture more intricate patterns within the data. By adopting the hyperparameters from the 187th run, we aim to maintain a robust and reliable model that effectively balances recall and precision. This ensures accurate predictions while minimizing the risk of false positives, aligning with our objective of developing a predictive model that generalizes well to new data and captures the complexities of the underlying data patterns.

[4] A "run" refers to a single execution of the model training process with a specific set of hyperparameters.

B.1 Optimizing "n_estimators" Parameter.

The parameter "n_estimators" specifies the number of trees included in the ensemble. To optimize this parameter and avoid overfitting, we employed the "early_stopping_rounds" method. This approach halts the addition of new trees when there is no significant improvement in the AUC score of the validation dataset over a specified number of iterations. In our model, setting the early stopping rounds to 10 meant that training would cease if the AUC score did not improve for ten consecutive rounds.

Table VII

Last 10 Iterations of Early Stopping Rounds

In this table, the last ten iterations of early stopping rounds are shown, displaying the training and validation AUC scores for these final iterations. This highlights how the model's performance evolved during these rounds. After determining the optimal values for learning rate, max depth, and min child weight through a comprehensive grid search, we incorporated these parameters and employed early stopping to find the optimal number of estimators (n_estimators). This process involved running the model with the best parameters and halting the training when no significant improvement was observed in the AUC score for ten consecutive rounds. This approach aims to create a predictive and reliable model by preventing overfitting and capturing intricate patterns in the data.

Iteration	Training AUC	Validation AUC
90	0.98054	0.92064
91	0.98083	0.92066
92	0.98102	0.92075
93	0.98110	0.92081
94	0.98121	0.92085
95	**0.98135**	**0.92086**
96	0.98171	0.92075
97	0.98187	0.92062
98	0.98209	0.92074
99	0.98212	0.92077

The optimal "n_estimators" determined through this method is 95, achieving an AUC score of 0.92086, as observed in the final iteration where improvement is noted. The hyperparameters and their respective optimal values used to finalize the model are listed in Table VIII.

Table VIII

Hyperparameters used for Building the Model

This table presents the hyperparameters that were meticulously selected through an iterative process to balance learning accuracy and computational efficiency. These hyperparameters play a crucial role in fine-tuning the XGBoost

model to effectively predict cash advance usage among customers. Each parameter was adjusted and validated to ensure the model's performance was optimized for both precision and speed. These hyperparameters were chosen to optimize the model's predictive performance and computational efficiency. The learning rate of 0.100 ensures gradual learning, minimizing the risk of overshooting while still allowing the model to converge in a reasonable number of iterations. The max depth of 7 enables the model to learn complex relationships in the data without excessive overfitting. With 95 boosting rounds, determined through early stopping, the model achieves a balance between complexity and generalization. The min child weight of 2 ensures that splits occur only when there is a sufficient number of instances, adding a layer of robustness to the model by preventing it from fitting noise in the data.

Parameters	Values
learning_rate	0.100
max_depth	7
n_estimators	95
min_child_weight	2

With these settings, the model aims to exhibit robustness, minimize overfitting, and accurately capture underlying patterns in customer behavior data.

V. Empirical Results

A. Overall Model Performance

The cash advance prediction model demonstrates exceptional performance, as evidenced by its high precision, recall, F1-scores, accuracy, AUC, and Gini coefficients. These metrics underscore the model's outstanding effectiveness in distinguishing between customers who do and do not use cash advances. The model's consistent performance across the training, validation, and test datasets indicates its robustness and reliability in predicting cash advance usage.

Table IX
Overall Performance Metrics

This table provides a detailed breakdown of the performance metrics for both non-cash advance users (Class 0) and cash advance users (Class 1). The metrics include precision, recall, F1-score for each class, as well as overall accuracy, Area Under the Curve (AUC), and Gini coefficient for each dataset. These metrics are presented separately for the training, validation, and test sets. Details of the training and validation results are provided in Appendix C.

Metric	Training	Validation	Test
Precision (Class 0)	0.96	0.91	0.87

Metric	Training	Validation	Test
Recall (Class 0)	0.89	0.80	0.77
F1-score (Class 0)	0.92	0.85	0.82
Precision (Class 1)	0.91	0.82	0.80
Recall (Class 1)	0.97	0.91	0.89
F1-score (Class 1)	0.94	0.86	0.84
Accuracy	0.93	0.86	0.83
AUC	0.93	0.86	0.83
Gini	0.86	0.72	0.66

The high precision and recall values across different classes indicate that the model can accurately identify both positive and negative cases of cash advance usage. Overall, the cash advance prediction model achieved AUC scores of 93%, 86%, and 83% for the training, validation, and test datasets, respectively. These high AUC values suggest that the model has a strong capability to distinguish between the classes (cash advance users and non-users) across different datasets. The Gini scores, derived from the AUC, further substantiate the model's classification power. This capability enables the model to effectively identify customers likely to use cash advances. The high scores across training, validation, and testing stages assure the model's reliability and applicability in operational environments. The AUC scores further confirm the model's robust discriminative power, reflecting its capability to make accurate predictions consistently. By maintaining high performance across multiple metrics, this model stands as a reliable predictor of cash advance usage, demonstrating its utility in practical, real-world scenarios.

B. Test Performance

The test performance metrics given at Table X indicate that the model generalizes well to new data, maintaining high precision and recall scores. This robustness is crucial for practical applications where the model encounters unseen data.

Table X
Performance Metrics – Test

This table provides the performance metrics for each class, along with overall accuracy, macro average, and weighted average scores. These metrics help to evaluate the model's effectiveness in distinguishing between cash advance users (Class 1) and non-users (Class 0). The macro average is the unweighted mean of precision, recall, and F1-score for both classes. It treats each class equally, regardless of the number of instances in each class. This average is useful to

get an overall sense of the model's performance across all classes, giving equal importance to both classes. The weighted average takes into account the support (number of instances) for each class, weighting the metrics according to the number of instances in each class. This provides a more realistic picture of the model's performance by reflecting the prevalence of each class in the dataset.

Class	Precision	Recall	F1-score	Support
0	0.87	0.77	0.82	789
1	0.80	0.89	0.84	801
Accuracy			0.83	1590
Macro Avg	0.83	0.83	0.83	1590
Weighted Avg	0.83	0.83	0.83	1590

The model achieved a precision of 0.87 and a recall of 0.77 for Class 0, indicating a reliable identification of true negatives. For Class 1, the model demonstrated strong recall at 0.89, affirming its effectiveness in identifying positive instances. The high F1-scores for both classes highlight the model's balanced performance in terms of precision and recall.

B.1. Confusion Matrix Insights

The confusion matrix for the test set is provided in Figure I, indicating the distribution of true and false predictions across both classes. This matrix showcases the model's ability to accurately identify cash advance users.

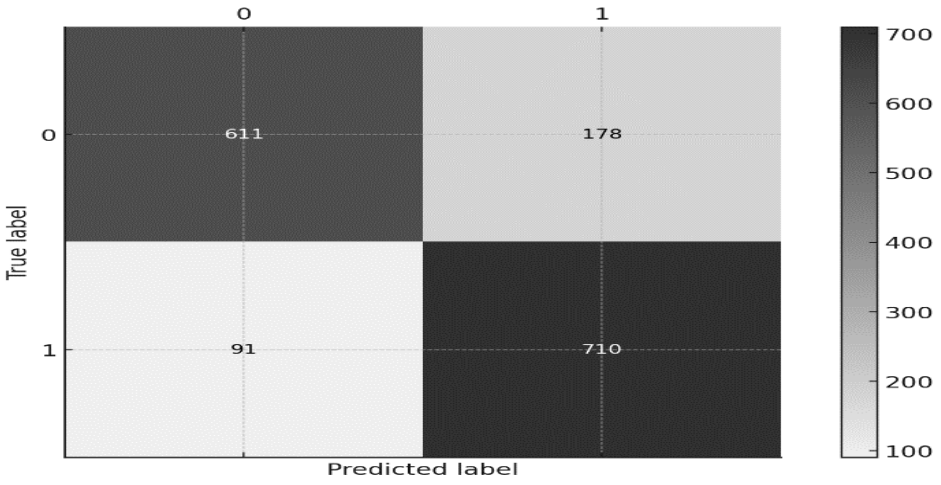


Figure I. The confusion matrix for the test set. This figure illustrates the distribution of true and false predictions for cash advance and non-cash advance users. True labels represent the actual classifications, while predicted labels represent the model's predictions. Specifically: **True Positives:** Out of 801 cash advance users, 710 instances were correctly predicted as cash advance users. **True Negatives:** Out of 789 non-cash advance users, 611 instances were correctly predicted as non-cash advance users. **False Positives:** Out of 789 non-cash advance users, 178 instances were incorrectly predicted as cash advance users. **False Negatives:** Out of 801 cash advance users, 91 instances were incorrectly predicted as non-cash advance users.

From the confusion matrix, we calculate:

- **True Positive Rate:** 0.89, which indicates that 89% of actual cash advance users were correctly identified by the model.
- **True Negative Rate:** 0.77, which shows that 77% of actual non-cash advance users were correctly identified.

These metrics indicate that the model performs exceptionally well in identifying both classes accurately, with a high True Positive Rate and a robust True Negative Rate.

B.2. ROC Curve Insights

The ROC curve for the test set further highlights the model's high discriminatory power. The Area Under the Curve (AUC) of 0.83 confirms the model's effectiveness in distinguishing between cash advance users and non-users, ensuring reliable performance in real-world scenarios.

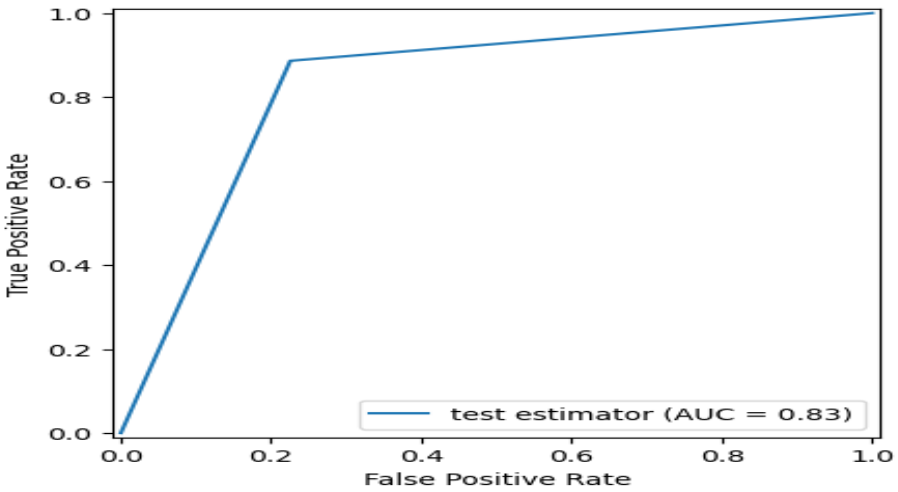


Figure 2: ROC Curve – Test. This figure shows the Receiver Operating Characteristic (ROC) curve, representing the model's diagnostic ability for test data. The curve plots the True Positive Rate (sensitivity) against the False Positive Rate (1 - specificity) at various threshold settings. It provides a detailed view of the trade-offs between

sensitivity and specificity for the predictive model. The closer the ROC curve is to the top-left corner, the better the model's performance, indicating high sensitivity and low false positive rate. The Area Under the Curve (AUC) quantifies the model's overall performance, with values closer to 1 indicating excellent discriminative ability. Each point on the ROC curve represents a different threshold setting for classification decisions.

The high AUC value of 0.83 underscores the model's capability to make accurate predictions consistently, reinforcing its applicability in practical, operational environments.

C. Performance on Unseen, Unbalanced Data

To ensure the generalizability of our model, we evaluated its performance on unseen, unbalanced test data points. This step is crucial to confirm the model's effectiveness and stability in predicting customer cash advance usage in real-world scenarios.

Our dataset comprises 40,000 unseen and unbalanced data points collected over the years 2017 to 2020, providing a realistic assessment of its performance in operational conditions.

Table XI
Key Performance Metrics for Unseen,Unbalanced Data

The model performance metrics, summarized in this table provide a comprehensive overview of the model's effectiveness in predicting cash advance usage among customers. Each metric offers unique insights into different aspects of the model's performance. The main goal is to assess the model's capability to manage significant class imbalance while maintaining accurate predictions. Therefore, the overall metrics such as AUC, accuracy, precision, recall, and F1 score offer a comprehensive and clear representation of the model's performance and its effectiveness in practical, operational environments.

Metric[5]	**Value**
Area Under the ROC Curve (AUC)	0.9026
Accuracy	79.70%
Precision	27.76%
Recall	88.48%
F1 Score	42.26%

[5] Unlike the balanced data analyses given in Table X, the metrics in Table XI for the unbalanced dataset do not include macro average and weighted average scores. This decision is intentional, as the primary goal is to focus on the overall performance across an unbalanced dataset. Including these averages would not provide additional value in this context.

The model achieved outstanding results across various performance metrics, underscoring its effectiveness in predicting cash advance usage. The AUC score of 0.9026 signifies the model's superior ability to distinguish between users and non-users of cash advances across different threshold settings. The accuracy of 79.70% is notable, especially given the significant class imbalance in the dataset. The relatively lower precision of 27.76% highlights the model's strategic emphasis on maximizing recall, ensuring that the majority of potential positive cases are identified, even at the expense of a higher number of false positives. The high recall rate of 88.48% showcases the model's robust capability in accurately identifying the majority of positive instances. F1 score of 42.26%, which harmonizes precision and recall, reflects the model's design priorities, emphasizing recall to capture as many positive instances as possible.

C.1. ROC Curve Insights

The ROC Curve of the unseen data is shown in Figure 3, indicating the model's high discriminatory power.

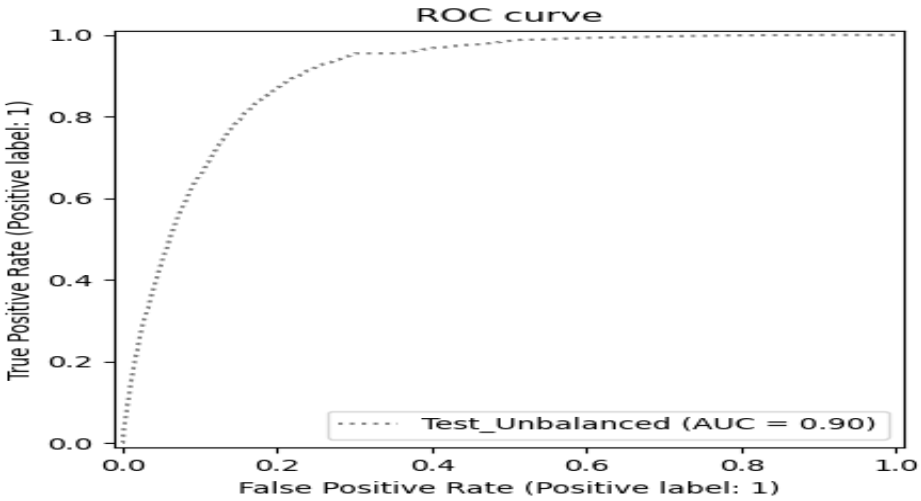


Figure 3. ROC Curve – Unseen, Unbalanced Data. The ROC curve's steep ascent towards the top-left corner indicates that the model has a high True Positive Rate and a low False Positive Rate at various threshold settings. This shape is characteristic of a model that can effectively distinguish between the positive and negative classes. The ROC curve's shape and the high AUC value validate that the model not only predicts well but also effectively balances sensitivity and specificity. This makes the model a valuable tool for predicting cash advance usage in operational environments, ensuring reliable performance and high discriminative power across various threshold settings.

The ROC curve for our model showcases an AUC of 0.9026, affirming its robust discriminative power. The curve's proximity to the top-left corner highlights an exceptional true positive rate, aligning with our strategic objective to maximize recall. This indicates the model's effectiveness in identifying true positives while managing false positives—a desirable trait in predictive analytics. The model's high performance on unseen, unbalanced data underscores its robustness and practical applicability, demonstrating its ability to generalize well and maintain high accuracy even in the presence of class imbalances. This is crucial for real-world applications where data distributions are often skewed, and the accurate identification of positive instances significantly impacts decision-making and resource allocation.

C.2. Confusion Matrix Insights

The confusion matrix of the unseen data given at the Figure 4 provides further insights into the model's classification performance and its reliability in real-world settings.

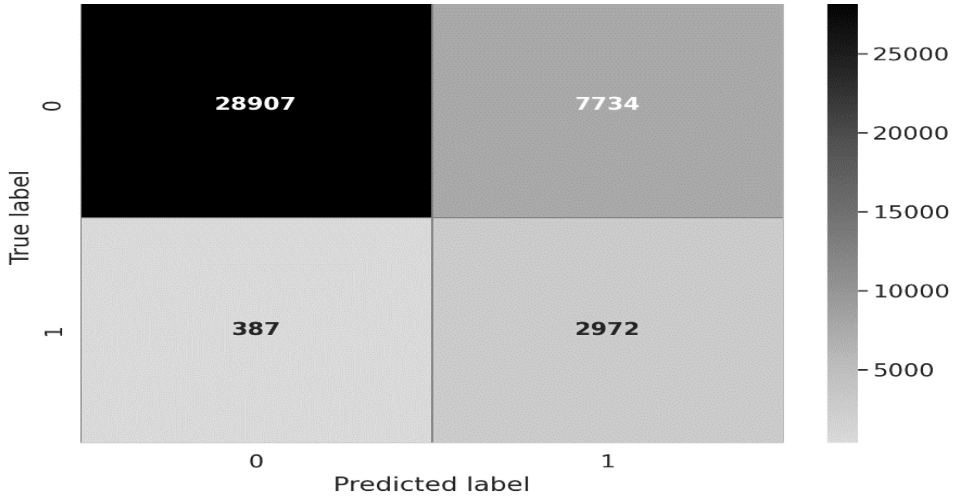


Figure 4. Confusion Matrix – Unseen, Unbalanced Data. This figure displays a total of 40,000 unbalanced samples, with 3,359 identified as cash advance users and 36,641 as non-cash advance users, resulting in a cash advance usage rate of 8.3%. From the confusion matrix, we derive the following insights: the model correctly identified 28,907 accounts as "Non-Cash Advance" (True Negatives) and 2,972 accounts as "Cash Advance" (True Positives). However, the model also misclassified 7,734 "Non-Cash Advance" accounts as "Cash Advance" (False Positives) and 387 "Cash Advance" accounts as "Non-Cash Advance" (False Negatives). These findings offer a detailed understanding of the model's classification performance and its reliability in practical applications.

The evaluation metrics derived from the confusion matrix provide a nuanced understanding of the model's performance:

- **True Positive Rate (Recall):** 88.48%. This high rate indicates that the model is highly effective at identifying "Cash Advance" accounts, ensuring most actual users are correctly classified.
- **True Negative Rate:** 78.88%. This demonstrates the model's robustness in accurately classifying 78.88% of actual "Non-Cash Advance" accounts, reflecting its proficiency in minimizing false positives.
- **Type 2 Error Rate:** 11.52%. This metric highlights the model's efficiency in minimizing false negatives, thus effectively capturing true positives.

These results align with our objective of achieving a high True Positive Rate while minimizing Type 2 errors, ensuring that the model accurately identifies as many true "Cash Advance" accounts as possible. The high recall and strong True Negative Rate underscore the model's capability to effectively distinguish between "Cash Advance" and "Non-Cash Advance" customers, establishing it as a reliable tool for predictive analytics. Overall, the model demonstrates robust performance on unseen data, confirming its reliability and effectiveness in predicting "Cash Advance" usage in real-world scenarios.

VI. Interpreting Model Predictions: Insights from SHAP Analysis

The SHAP (Shapley Additive Explanations) summary plot is a powerful tool for interpreting machine learning models, as it illustrates the contribution of each feature to the target variable. In our model, we utilize this plot to assess the likelihood of cash advance usage. Figure V presents the SHAP plot for our model.

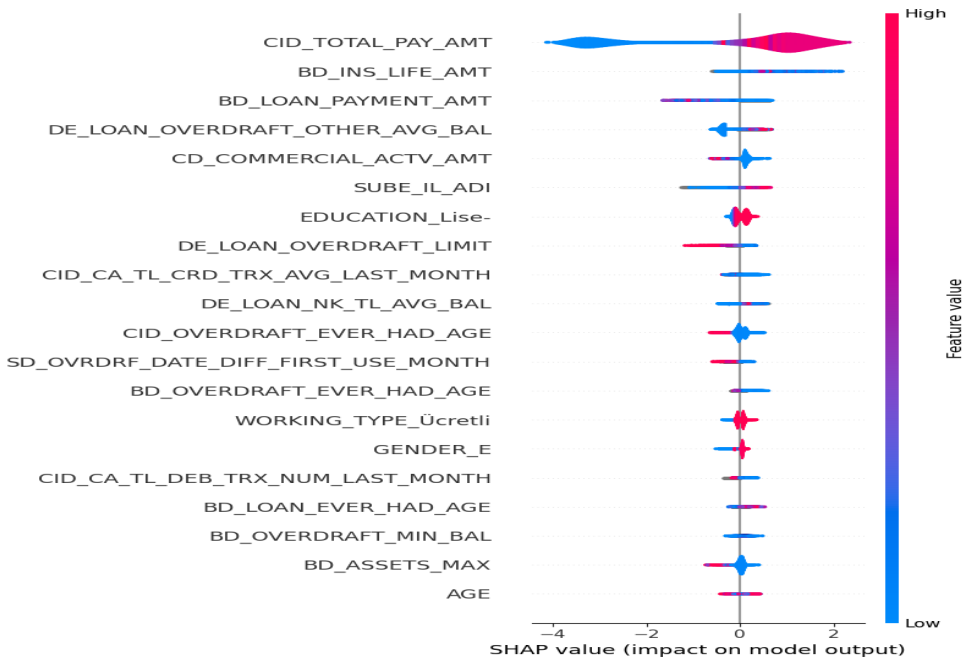


Figure V. SHAP Variable Importance Violin Plot. Each dot on the horizontal axis of the SHAP Variable Importance Violin Plot represents the SHAP value for a feature of the cash advance model for a single observation, indicating whether the feature's value leads to a higher or lower prediction of cash advance usage. Positive SHAP values push the prediction towards cash advance usage, while negative values push it away. The vertical axis lists all the features in the model, arranged by their importance based on the average magnitude of their SHAP values, helping to quickly identify the most impactful features on the model's predictions.The color spectrum of the dots, ranging from blue to pink/red, represents the actual value of the feature for that observation. Blue dots indicate low feature values, while pink/red dots indicate high feature values. This color coding provides insight into how different feature values influence the model's predictions. For instance, higher values of a feature, indicated by pink/red dots, suggest an increased likelihood of predicting cash advance usage when associated with positive SHAP values. The spread of dots along the horizontal axis for each feature shows the distribution of SHAP values, indicating the variability in the feature's impact across different observations. A wide spread of dots suggests that the feature has a varied impact on predictions, while a narrow spread indicates a more consistent impact. By examining the plot, it becomes clear which features are most influential and how changes in those feature values affect predictions. For example, if a feature's dots are predominantly on the positive side and colored red, it means higher values of this feature consistently drive the prediction towards cash advance usage. Conversely, if the dots are mostly on the negative side and colored blue, lower values of the feature reduce the likelihood of predicting cash advance usage. This visualization provides a comprehensive view of feature importance and their impact on the model's predictions.

The SHAP summary plot highlighted the significant variables influencing cash advance behavior, revealing the specific predictors that have the most substantial impact. By focusing on individual features such as total payments, life insurance amounts, loan payments, and average overdraft balances, we can understand how each variable contributes to the overall prediction. For instance,

higher loan payments and overdraft balances are positively associated with cash advance usage, indicating financial strain, while high total payments and life insurance amounts negatively impact cash advance usage, suggesting better financial management. This detailed analysis allows us to pinpoint the exact factors driving cash advance behavior, enhancing our ability to predict and interpret consumer actions with greater precision.

Table XII
Top 20 Most Influential Variables Based on SHAP Values

This table presents the 20 most influential variables affecting cash advance usage, as interpreted from the SHAP plot. SHAP values provide detailed insights into how each feature impacts the model's predictions. The variables encompass a wide range of factors, including financial metrics like total payments and loan payment amounts, as well as demographic characteristics such as education level and geographical location. These insights help us understand the various elements that significantly influence cash advance usage.

Description of Variable	Effect on Cash Advance	Interpretation
Total payments amount of the customer-last month (CID_TOTAL_PAY_AMT)	Negative	High total payment amounts made by the customer in the last month has negative impact on cash advance usage.
Life insurance amount (BD_INS_LIFE_AMT)	Negative	Life insurance payments are negatively related to cash advance usage.
Loan payment amounts (BD_LOAN_PAYMENT_AMT)	Positive	Substantial loan payments increase reliance on cash advances.
Average period balances of overdraft loans (DE_LOAN_OVERDRAFT_OTHER_AVG_BAL)	Positive	Higher average balances in overdraft accounts are positively associated with cash advance usage.
Total of the customer's banking activity in the relevant period (CD_COMMERCIAL_ACTV_AMT)	Negative	Banking activity amounts are negatively correlated with cash advance usage.
Customers branch city (SUBE_IL_ADI)	Varies[6]	Geographical location and associated economic and demographic characteristics influence consumer behavior on cash advances.
Education level- High School (EDUCATION_Lise)	Negative	Having high school education shows a potential negative correlation with cash advance usage.

[6] We applied target encoding to the branch city variable. Target encoding replaces each city with the mean cash advance usage rate for that city. This provides a numeric representation of how likely customers from each city are to use cash advances.The findings reveal varying impacts on cash advance usage based on the encoded city variable. Results of encoding is provided in the Appendix D.

Overdraft limit defined within the relevant period (DE_LOAN_OVERDRAFT_LIMIT)	Negative	Higher overdraft limits are negatively related to cash advance usage.
Average number of debt movements in demand deposits last month (CID_CA_TL_DEB_TRX_NUM_LAST_ MONTH)	Negative	A higher number of debit transactions in the last month is negatively related to cash advance usage.
Average number of new products of the customer in the last 6 months (DE_LOAN_NK_TL_AVG_BAL)	Positive	Higher average balances in loan accounts are positively related to cash advance usage.
Age of overdraft account ownership - Information from credit Bureau (CID_OVERDRAFT_EVER_HAD_AGE)	Negative	Older histories of overdraft usage typically reduce the need for additional short-term borrowing.
The duration a customer waited to use their overdraft facility after opening the account (SD_OVRDRF_DATE_DIFF_FIRST_USE _MONTH)	Positive	The time between overdraft account opening and its first use is positively related to cash advance usage.
Overdraft ownership age - Bank Data (BD_OVERDRAFT_EVER_HAD_AGE)	Negative	Having a longer history with overdraft facilities in the bank indicates reduced reliance on cash advances.
Salaried worker (WORKING_TYPE_Ücretli)	Negative	Being salaried is negatively related to cash advance usage.
Gender of customer (GENDER_E)	Negative	Being male typically is negatively related to cash advance usage.
Age of the loan account (BD_LOAN_EVER_HAD_AGE)	Negative	The age of the loan account is negatively related to cash advance usage.
Minimum overdraft balances (BD_OVERDRAFT_MIN_BAL)	Positive	Lower minimum balances in overdraft accounts are positively related to cash advance usage.
Maximum total value of assets in the last 24 months (BD_ASSETS_MAX)	Negative	Higher total assets in the past are negatively related to cash advance usage.
Age of customer (AGE)	Negative	Impact of age on cash advance usage is negative. Older individuals are less likely to use cash advances, whereas younger individuals are more likely.

This table highlights the significant variables impacting cash advance usage and provides interpretations that can help financial institutions understand customer behavior and improve their risk management and product development strategies. Here are the key inferences based on the SHAP results from Table XII:

Financial Metrics Influence:

- Higher total payments made by customers in the last month are negatively associated with cash advance usage, suggesting that financial stability reduces the need for short-term loans.

- Substantial loan payments increase reliance on cash advances, highlighting that high debt burdens drive the need for additional credit.
- Higher average balances in overdraft accounts are positively associated with cash advance usage, indicating that customers with frequent overdrafts are more likely to need cash advances.
- More active banking customers, who likely manage their finances better, are less prone to using cash advances.

Demographic Factors:

- High school education shows a negative correlation with cash advance usage, implying that better-educated individuals are less likely to rely on expensive short-term loans.
- Being salaried is negatively related to cash advance usage, suggesting that stable income from salaried employment reduces the need for cash advances.
- Males are typically less likely to use cash advances, which may reflect differences in financial behavior or access to other credit forms.

Overdraft and Loan History:

- Higher overdraft limits are negatively related to cash advance usage, suggesting that customers with greater access to overdraft facilities might not need cash advances.
- Older histories of overdraft usage typically reduce the need for additional short-term borrowing.
- Longer loan account histories are negatively related to cash advance usage, suggesting that more experienced borrowers manage their credit better.
- Lower minimum balances in overdraft accounts are positively related to cash advance usage, indicating that customers with frequent low balances might turn to cash advances.

Other Influential Factors:

- Higher past asset values are negatively related to cash advance usage, suggesting that wealthier individuals are less likely to need short-term credit.

- Older individuals are less likely to use cash advances, whereas younger individuals are more likely, reflecting that younger customers may have less financial stability or experience.

Geographical and Behavioral Influences:

- Geographical location and associated economic and demographic characteristics influence consumer behavior on cash advances, indicating that location-specific factors can affect the likelihood of using cash advances.
- The time between overdraft account opening and its first use is positively related to cash advance usage, suggesting that customers who use overdrafts early might be more prone to needing cash advances later.
- A higher number of debit transactions in the last month is negatively related to cash advance usage, indicating that active account management reduces the need for cash advances.

VII. Concluding Remarks.

This study provides a comprehensive analysis of the factors influencing cash advance usage among bank customers by utilizing the XGBoost machine learning algorithm on a detailed customer dataset provided by a real bank. The model demonstrates exceptional predictive accuracy, achieving an AUC of 0.90 and a recall of 0.88, highlighting its effectiveness in distinguishing between cash advance users and non-users.

The SHAP analysis reveals critical insights into the behavioral and demographic factors driving cash advance usage. Key findings indicate that effective debt management practices, such as high payment amounts and active credit line usage, correlate with lower cash advance usage. Conversely, financial distress indicators, such as higher loan and overdraft balances, are associated with increased reliance on cash advances. Additionally, demographic factors, including younger age and lower educational attainment, significantly influence cash advance utilization. In contrast, older age, higher asset values, longer loan and overdraft histories, and salaried employment are linked to reduced usage.

These insights are instrumental in developing targeted financial products and interventions that mitigate the risks associated with cash advances. By leveraging these findings, financial institutions can implement proactive measures to promote financial stability and offer more sustainable credit options to their customers. The study's results provide a robust foundation for financial institutions and fintech companies to refine credit risk assessment models and develop products that better serve consumer needs, ultimately contributing to healthier financial choices and improved economic outcomes.

Overall, this study highlights the potential of advanced machine learning techniques in understanding and predicting consumer financial behaviors, offering valuable contributions to both academic literature and practical financial strategies. The high predictive accuracy and insightful analysis presented in this research underscore the pivotal role of fintech in transforming financial analytics and enhancing the precision and efficiancy of financial services.

Appendix A: Formulations and Explanations of Performance Metrics Used in Model Evaluation

This appendix provides detailed explanations and formulations of the performance metrics used to evaluate the classification model. Understanding these metrics is crucial for assessing the model's effectiveness and identifying areas for improvement.

Table A.1.

Confusion Matrix

This table shows the structure of the confusion matrix, a fundamental tool for evaluating the performance of a classification model. The confusion matrix consists of four categories:
True Negatives: The number of actual negative cases correctly predicted as negative.
False Positives: The number of actual negative cases incorrectly predicted as positive.
False Negatives: The number of actual positive cases incorrectly predicted as negative.
True Positives: The number of actual positive cases correctly predicted as positive.
This matrix helps in understanding the model's accuracy and the types of errors it makes, thereby providing valuable insights into its performance and areas for improvement.

	Predicted Negative (0)	**Predicted Positive (1)**
Actual Negative (0)	True Negatives	False Positives
Actual Positive (1)	False Negatives	True Positives

Table A.2.
Formulations of the Metrics derived from confusion Matrix.

This table outlines the metrics derived from the confusion matrix used to evaluate model performance, along with their formulas and descriptions.

Metric	**Formulation**	**Description**
True Positive Rate	$True\ Positive$ / (True Positive + $False\ Negative$)	The percentage of actual positive cases accurately detected by the model, indicating its sensitivity.
True Negative Rate	$True\ Negative$ / ($True\ Negative$ + $False\ Positive$)	The percentage of actual negative cases accurately detected by the model, indicating its specificity.
False Positive Rate	$False\ Positive$/ ($False\ Positive$ + $True\ Negative$)	The percentage of actual negative cases incorrectly classified as positive by the model, indicating the likelihood of false positives.
False Negative Rate	$False\ Negative$/ ($True\ Positive$ + $False\ Negative$)	The percentage of actual positive cases incorrectly classified as negative by the model, indicating the likelihood of false negatives.

Explanations of Performance Metrics Used in Model Evaluation:

Accuracy measures the overall correctness of the model by calculating the ratio of correct predictions to the total predictions.

$$Accuracy = \frac{True\ Positive\ + True\ Negative}{True\ Positive + False\ Positive +\ False\ Negative + True\ Negative}$$

Precision is the proportion of true positive predictions to the total number of positive predictions, indicating the accuracy of the model's positive predictions.

$$Precision = \frac{True\ Positive}{True\ Positive\ + False\ Positive}$$

Recall (or Sensitivity) is the ratio of true positive predictions to the total actual positives, showing the model's ability to identify positive instances.

$$Recall = \frac{True\ Positive}{True\ Positive\ + False\ Negative}$$

The F1-score represents the harmonic mean of the precision and recall values. It provides a single metric that balances both concerns.

$$F1 - Score = \frac{2 * Precision * Recall}{Precision + Recall}$$

AUC (Area Under the ROC Curve) reflects the model's ability to distinguish between classes, with higher values showing better performance.

Gini Index is the derivative of the AUC, representing model performance, with higher values indicating better.

Appendix B: Distribution of Balanced Model Sample Based on Selected Demographics

This appendix provides a detailed overview of the distribution of the balanced model sample based on various demographic factors, including working type, marital status, education level, age range, and geographical location. The data presented here offers insights into the composition of the sample population, helping to understand the diverse characteristics of bank customers and how these characteristics may influence financial behaviors, such as cash advance usage.

Table B.1.

Distribution by Working Type

This table presents the distribution of the sample, consisting of 24,000 individuals, based on their working types. Working type information is typically collected by the bank and categorized through customer data provided during account registration or loan application processes. Half of the sample (49.45%) are in paid employment, indicating significant workforce engagement. Retired individuals make up 27.40%, suggesting substantial representation of those no longer in active employment. Other categories such as self-employed, public-sector workers, and those not working provide insights into the variety of employment statuses within the sample.

Working Type	Count	Percentage
Paid Employment	11,868	49.45%
Retired	6,576	27.40%
Self-Employed	2,117	8.82%
Public Sector	1,408	5.87%
Not available	1,160	4.83%
Not Working	567	2.36%
Supplementary Income	282	1.18%
Part-Time	22	0.09%
Total	**24,000**	**100.00%**

Table B.2.

Distribution by Marital Status

This table categorizes the sample by marital status. The majority of the individuals are married (67.25%), followed by single individuals (28.48%). This distribution highlights the predominance of married individuals within the sample.

Marital Status	Count	Percentage
Married	16,141	67.25%
Single	6,834	28.48%
Not available	1,025	4.27%
Total	**24,000**	**100.00%**

Table B.3.
Distribution by Education Level

This table displays the educational attainment levels of the sample. A notable segment of the sample has completed 5-8th grade (24.26%), followed by high school graduates (13.30%) and those with vocational school education (12.11%). University graduates make up 5.64%, while advanced degrees (Master's and Doctorate) account for a small portion (0.31% and 0.07%, respectively). This table illustrates the educational diversity within the sample.

Education Level	Count	Percentage
Primary School	2,153	8.97%
5-8th Grade	5,822	24.26%
High School	3,191	13.30%
Vocational School	2,906	12.11%
No Education	1,480	6.17%
University	1,354	5.64%
Master's Degree	75	0.31%
Doctorate	17	0.07%
Not available	7,002	29.18%

Table B.4.
Distribution by Age Range

This table shows the age distribution of the sample. The largest age groups are those aged 40-50 (22.25%) and 50-60 (22.75%), indicating a concentration of middle-aged individuals. Younger age groups, such as 30-40 (19.73%) and 20-30 (10.01%), are also represented, along with older age ranges like 60-70 (14.64%) and 70-80 (5.18%). The oldest age ranges, 80-90 (1.08%) and 90-100 (0.08%), have the fewest individuals.

Age Range	Count	Percentage (%)
[20, 30)	2,403	10.01%
[30, 40)	4,734	19.73%
[40, 50)	5,340	22.25%
[50, 60)	5,461	22.75%
[60, 70)	3,513	14.64%
[70, 80)	1,244	5.18%
[80, 90)	258	1.08%
[90, 100)	20	0.08%
Not Available	1,027	4.28%
Total	**24,000**	**100.00%**

Table B.5.

City of Branch: (81 city is grouped into districts)

This table categorizes 24,000 balanced sample based on their location within various regions of Turkey. The Marmara Region has the highest count (9,036), followed by the Aegean Region (3,898) and Central Anatolia (2,829). Other regions such as the Mediterranean (2,825) and Black Sea (2,436) also have significant representation. This table highlights the geographical distribution of the sample across different regions.

Region	Cities Included	Total Count
Marmara Region	Istanbul, Bursa, Kocaeli, Tekirdag, Sakarya, Edirne, Kirklareli, Canakkale, Balikesir, Yalova, Bilecik, Kirikkale	9,036
Aegean Region	Izmir, Manisa, Aydin, Mugla, Denizli, Usak, Afyon, Kutahya, Isparta	3,898
Central Anatolia	Ankara, Konya, Kayseri, Eskisehir, Sivas, Nevsehir, Nigde, Aksaray, Kirsehir, Cankiri, Karaman	2,829
Mediterranean Region	Antalya, Mersin (Icel), Hatay, Osmaniye, Adana, Kahramanmaras, Burdur	2,825
Black Sea Region	Samsun, Ordu, Trabzon, Zonguldak, Karabuk, Rize, Amasya, Tokat, Giresun, Bolu, Duzce, Bartin, Sinop, Artvin, Bayburt, Gumushane, Kastamonu, Corum	2,436
Eastern Anatolia	Erzurum, Van, Agri, Erzincan, Kars, Igdir, Bitlis, Ardahan, Tunceli, Mus, Bingol, Hakkari, Sirnak, Elazig, Malatya, Diyarbakir, Batman, Mardin, Siirt	1,250
Southeastern Anatolia	Gaziantep, Sanliurfa, Adiyaman	618
Not available	-	1,580
Total	-	**24,000**

Appendix C: Training and Validation Results for Balanced Data

This Appendix presents the training and validation results for balanced data, providing detailed insights into the model's performance metrics. The balanced dataset ensures that the model is equally exposed to both classes, facilitating a comprehensive evaluation of its predictive capabilities.

Table C.1.

Performance Metrics Results – Training

This table shows the performance metric results for training. The results from the training stage indicate nearly perfect scores in precision, recall, and F1-score for both classes, with values approximately centering around 0.92 and 0.94. Precision refers to the proportion of true positive predictions out of all positive predictions made by the model. A precision of 0.96 for Class 0 (non-cash advance users) and 0.91 for Class 1 (cash advance users) indicates that the model is highly accurate in predicting both classes. Recall measures the proportion of actual positive cases correctly identified by the model, with scores of 0.89 for Class 0 and 0.97 for Class 1, demonstrating the model's ability to identify the majority of true cash advance users while maintaining a high rate of correct identification for non-users. F1-scores of 0.92 for Class 0 and 0.94 for Class 1 indicate that the model maintains a high level of accuracy and reliability across both classes. The overall accuracy of 0.93 indicates that 93% of the total predictions made by the model are correct, further underscoring its effectiveness.

Class	Precision	Recall	F1-score	Support
0	0.96	0.89	0.92	6562
1	0.91	0.97	0.94	7352
Accuracy			**0.93**	13914
Macro Avg	0.93	0.93	0.93	13914
Weighted Avg	0.93	0.93	0.93	13914

These metrics suggest that the model is highly effective at distinguishing between the two classes of customer behavior concerning the use of cash advances. The high precision, recall, and F1-score for both classes reflect the model's robustness and its capability to manage the classification task effectively.

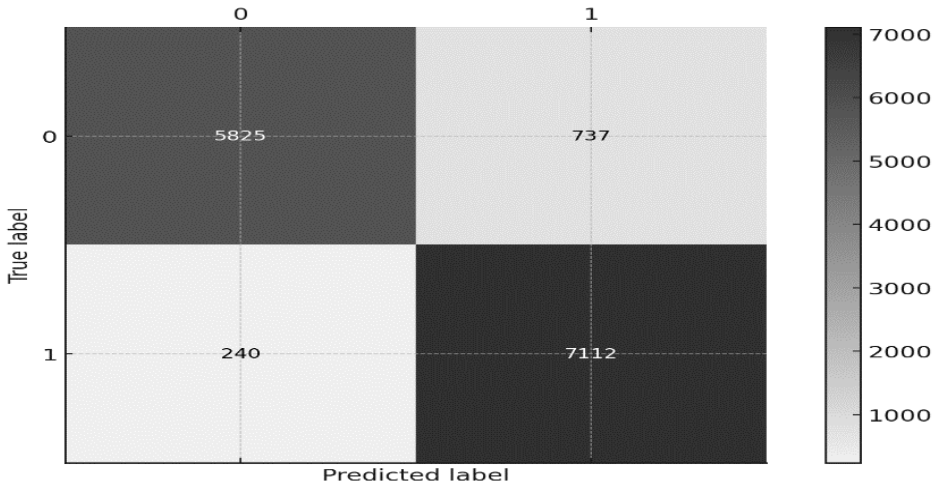


Figure C.1. Confusion Matrix – Training. This figure shows the confusion matrix for the training set, providing insights into the model's training performance. The True Positive Rate is 0.967, indicating that the model correctly identifies 96.7% of actual positive cases. The True Negative Rate is 0.888, showing that the model correctly identifies

88.8% of actual negative cases. These metrics indicate that the model performs exceptionally well in distinguishing between cash advance users and non-users, demonstrating high accuracy in both categories.

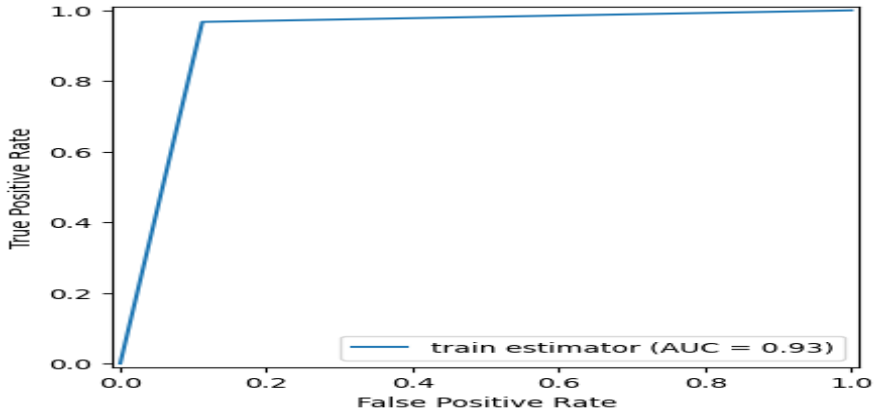


Figure C.2. ROC Curve – Training. This figure shows the ROC curve for the training data of balanced dataset. The steep ascent in the ROC curve towards the upper left corner reflects both high sensitivity and specificity, indicating that the model effectively distinguishes between the two behavioral classes with a low rate of error. The Area Under the Curve (AUC) of 0.93 for the training estimator from the ROC curve analysis further confirms the model's robustness.

Table C.2

Performance Metrics – Validation

This table presents the performance metrics for the validation stage of the model evaluation:

Class	Precision	Recall	F1-score	Support
0	0.91	0.80	0.85	716
1	0.82	0.91	0.86	693
Accuracy			**0.86**	1409
Macro Avg	0.86	0.86	0.86	1409
Weighted Avg	0.86	0.86	0.86	1409

The precision for the non-cash advance use class (Class 0) remains high at 0.91, though recall has dropped to 0.80. This indicates a reduction in the model's ability to detect all true negative cases. Conversely, the recall for the cash advance use class (Class 1) is notably high at 0.91, underscoring the model's strength in identifying nearly all positive instances. These validation results indicate a slight dip in performance compared to the training stage, which is a common occurrence in predictive modeling. This often highlights the challenges of generalizing from training data to new

instances. Despite the performance decline, the validation results demonstrate a solid overall accuracy of 86% and balanced F1-scores around 0.85-0.86. These metrics showcase the model's robustness and its capability to generalize well to new data.

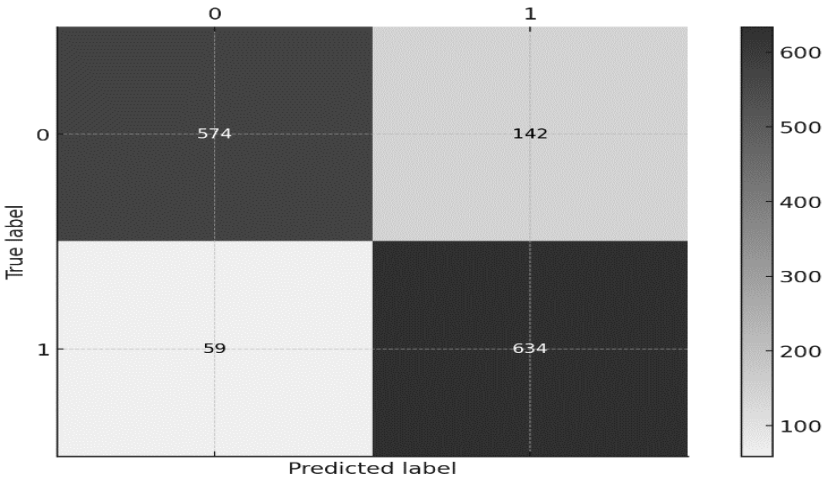


Figure C.3. Confusion Matrix – Validation. This figure shows the confusion matrix for the validation set, providing detailed insights into the model's performance in correctly classifying cash advance users and non-users.

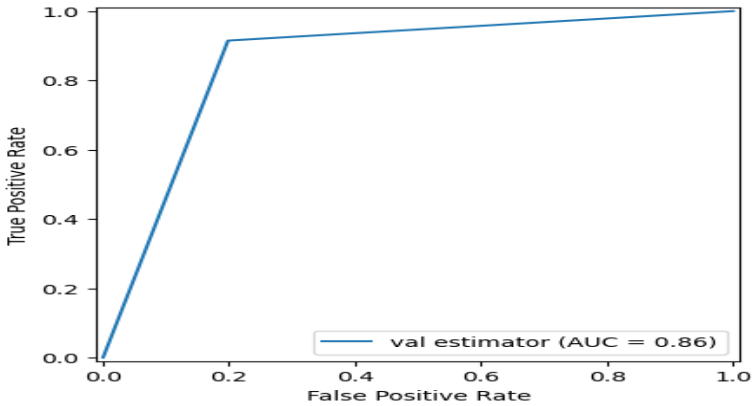


Figure C.4: ROC Curve – Validation. This figure presents the ROC curve for the validation data. The Area Under the Curve (AUC) of 0.86 for the validation estimator further confirms the model's robustness. This slight decrease from the training stage AUC reflects the anticipated challenges in model generalization. These results indicate a robust model that, despite the natural decrease in performance metrics when transitioning from training to validation data, maintains a strong ability to classify behaviors accurately under varied conditions.

Appendix D: Target Encoding Results by City

In this appendix, we present the target encoding results by city. We applied target encoding to the branch city variable by replacing each city with its mean cash advance usage rate. This method provides a clear numeric representation of the propensity for cash advance usage among customers from each city. This approach helps identify cities with higher and lower likelihoods of cash advance usage and highlights the varying impact of location on customer behavior[7].

Table D.1
Target Encoding Results by City

This table shows the target encoding results by city. High target encoding values indicate a higher likelihood of cash advance usage, meaning cities with higher values are more likely to have customers who frequently use cash advances. Conversely, low target encoding values indicate a lower likelihood of cash advance usage, meaning cities with lower values are less likely to have customers who frequently use cash advances.

Higher Likelihood of Cash Advance Usage	Target Encoding Value	Lower Likelihood of Cash Advance Usage	Target Encoding Value
Agri	0.789	Tunceli	0.3
Nevsehir	0.712	Gumushane	0.389
Mus	0.684	Karabuk	0.375
Kirikkale	0.662	Burdur	0.419
Bayburt	0.643	Amasya	0.44
Bingol	0.643	Usak	0.443
Ardahan	0.636	Cankiri	0.472
Ordu	0.634	Adiyaman	0.471
Bitlis	0.63	Karaman	0.475

[7] While target encoding provides valuable insights into the likelihood of cash advance usage across different cities, it is important to consider additional factors when making geographic inferences. The encoded values reflect average behavior within each city, but they do not account for individual-level variability or other underlying factors that may influence financial behavior, such as economic conditions, employment rates, cultural norms, and access to financial services. Therefore, while we can identify trends and propensities based on city-level data, further analysis incorporating these additional factors would be necessary to make more precise and comprehensive geographic inferences. Understanding the broader context and combining target encoding with other analytical methods can provide a more complete picture of the factors driving cash advance usage in different locations.

Bartin	0.628	Batman	0.483
Sakarya	0.626	Sanliurfa	0.487
Hakkari	0.622		
Igdir	0.621		
Artvin	0.618		
Nigde	0.615		
Mardin	0.614		
Aksaray	0.583		
Hatay	0.55		
Osmaniye	0.587		
Kahramanmaras	0.571		
Samsun	0.588		
Trabzon	0.576		
Zonguldak	0.607		
Rize	0.588		
Tokat	0.55		
Giresun	0.583		
Bolu	0.577		
Duzce	0.588		
Sinop	0.6		
Kastamonu	0.556		
Corum	0.574		
Van	0.566		
Erzincan	0.545		
Elazig	0.547		
Diyarbakir	0.557		

These results show varying impacts of geographic location on cash advance usage, indicating that customers from different cities exhibit different propensities for using cash advances. Understanding these patterns can help financial institutions tailor their services and strategies to better meet the needs of customers in different locations.

References

Agarwal, S., Driscoll, J. C., Gabaix, X., & Laibson, D. (2008). Learning in the credit card market. *National Bureau of Economic Research, Working Paper Series, 13822*. https://doi.org/10.3386/w13822

Ahmed, Z.U., Ismail, I., Sohail, M.S., Tabsh, I., & Alias, H. (2010). Malaysian consumers' credit card usage behavior. *Asia Pacific Journal of Marketing and Logistics, 22(4)*, 528-544. https://doi.org/10.1108/13555851011090547

Brown, I., & Mues, C. (2012). An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert Systems with Applications, 39(3)*, 3446-3453. https://doi.org/10.1016/j.eswa.2011.09.033

Bussmann, N., Giudici, P., Marinelli, D., & Papenbrock, J. (2021). Explainable machine learning in credit risk management. *Computational Economics, 57(1)*, 203-216. https://doi.org/10.1007/s10614-020-10042-0

Butaru, F., Chen, Q., Clark, B., Das, S., Lo, A. W., & Siddique, A. (2016). Risk and risk management in the credit card industry. *Journal of Banking & Finance, 72*, 218-239. https://doi.org/10.1016/j.jbankfin.2016.07.015

Chang, Y.-C., Chang, K.-H., & Wu, G.-J. (2018). Application of extreme gradient boosting trees in the construction of credit risk assessment models for financial institutions. *Applied Soft Computing, 73*, 914-920. https://doi.org/10.1016/j.asoc.2018.09.029

Brito, D. L., & Hartley, P. R. (1995). Consumer rationality and credit cards. *Journal of Political Economy, 103(2)*, 400–433. https://doi.org/10.1086/261988

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*, 785-794. https://doi.org/10.1145/2939672.2939785

Clark, T. E., & McCracken, M. W. (2009). Improving forecast accuracy by combining recursive and rolling forecasts. *Journal of Econometrics, 158(1)*, 120-130. https://doi.org/10.1016/j.jeconom.2010.06.004

Darapaneni, N., Srinivas, P., Reddy, K., Paduri, A. R., Kanugovi, L., Pavithra, J., B G, S., & Bharath, S. (2022). Tree Based Models: A Comparative And Explainable Study For Credit Default Classification. *2022 IEEE 9th Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON)*. https://doi.org/10.1109/UPCON56432.2022.9986411

De Silva, L. G. R. V., & Patabendige, S. S. J. (2021). Factors impacting consumer credit card usage behavior. *Sri Lanka Journal of Marketing, 7(2)*, 146-168.

Pulina, M. (2011). Consumer behaviour in the credit card market: A banking case study. *International Journal of Consumer Studies, 35(1)*, 86-94. https://doi.org/10.1111/j.1470-6431.2010.00939.x

Fitzpatrick, T., & Mues, C. (2016). An empirical comparison of classification algorithms for mortgage default prediction: Evidence from a distressed mortgage market. *European Journal of Operational Research, 249*, 427-439. https://doi.org/10.1016/j.ejor.2015.09.014

Finlay, S. (2011). Multiple classifier architectures and their application to credit risk assessment. *European Journal of Operational Research, 210(2)*, 368-378. https://doi.org/10.1016/j.ejor.2010.09.029

Gross, D. B., & Souleles, N. S. (2002). Do liquidity constraints and interest rates matter for consumer behavior? Evidence from credit card data. *Quarterly Journal of Economics, 117(1)*, 149-185. https://doi.org/10.1162/003355302753399472

Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies, 33(5)*, 2223–2273. https://doi.org/10.1093/rfs/hhaa009

Hand, D. J. (2006). Classifier technology and the illusion of progress. *Statistical Science, 21(1)*, 1-14. https://doi.org/10.1214/088342306000000060

Jung, K., & Kang, M. Y. (2021). Understanding credit card usage behavior of elderly Korean consumers for sustainable growth. *Sustainability, 13(7)*, 3817. https://doi.org/10.3390/su13073817

Kasaian, K., Murthi, B. P. S., & Steffes, E. (2023). Understanding the difference between opportunistic and needy customers' borrowing behavior. *International Journal of Bank Marketing, 41(4)*, 903-925. https://doi.org/10.1108/IJBM-09-2022-0398

Koparal, C., & Calik, N. (2014). Bank credit card usage behavior of individuals. *International Journal of Social Sciences, 3(4)*, 75-97.

Li, H., Cao, Y., Li, S., Zhao, J., & Sun, Y. (2020). XGBoost model and its application to personal credit evaluation. *IEEE Intelligent Systems, 35(3)*, 52-61. https://doi.org/10.1109/MIS.2020.2972533

Li, Y. (2019). Credit risk prediction based on machine learning methods. *2019 14th International Conference on Computer Science & Education (ICCSE)*, 1011-1013. https://doi.org/10.1109/ICCSE.2019.8845444

Lin, L., et al. (2019). Determinants of credit card spending and debt of Chinese consumers. *International Journal of Bank Marketing, 37(2)*, 545-564. https://doi.org/10.1108/IJBM-01-2018-0010

Malhotra, D. K., Malhotra, K., & Malhotra, R. (2020). Evaluating consumer loans using machine learning techniques. In K.D. Lawrence & D.R. Pai (Eds.), *Applications of Management Science, Vol. 20*, 59-69. https://doi.org/10.1108/S0276-897620200000020004

Morgan, D. P., Strain, M. R., & Seblani, I. (2012). How payday credit access affects overdrafts and other outcomes. *Journal of Money, Credit and Banking, 44(2-3)*, 519-531. https://doi.org/10.1111/j.1538-4616.2011.00499.x

Önder, K. (2018). The effect of credit card usage on consumer behavior. *International Journal of Eurasia Social Sciences, 9(32)*, 75-97.

Staten, M., & Barron, J. M. (2002). College student credit card usage. *Credit Research Center, Working Paper 65*.

Uddin, M. A. (2020). Study on literacy and usage behavior of credit card users in India. *Humanities & Social Sciences Reviews, 8(1)*, 60-68. https://doi.org/10.18510/hssr.2020.819

Weiss, G. M., & Provost, F. (2003). Learning when training data are costly: The effect of class distribution on tree induction. *Journal of Artificial Intelligence Research, 19(1)*, 315-354. https://doi.org/10.1613/jair.1199

Suoranta, M., & Mattila, M. (2004). Mobile banking and consumer behaviour: New insights into the diffusion pattern. *Journal of Financial Services Marketing, 8*, 354–366. https://doi.org/10.1057/palgrave.fsm.4770132

Supporting Information

Additional Supporting Information may be found at the ‘internet Appendix.

Appendix S1: Internet Appendix (Model Codes)