-- Iranian Journal of Economic Research---------
Volume 30, Issue 103, Summer 2025, 1-41
ijer.atu.ac.ir
https://doi.org/10.22054/ijer.2025.84878.1350

ATU PRESS

eISSN:2476-6445 ISSN: 1726-0728 Accepted: 20/7/2025 Revised: 14/7/2025 Received: 13/3/2025 Original Research

Estimating the Probability of Loan Default in Melli Bank: A Comparative Study of Machine Learning and Econometric Approaches

Reza Taleblou — Associate Professor of Economics, Allameh Tabataba'i University, Tehran, Iran

Mir Ali Kamali — Ph.D. Candidate in Economics, Semnan University, Semnan, Iran

Parisa Mohajeri* — Associate Professor of Economics, Allameh Tabataba'i University, Tehran, Iran

Abstract

The current study employed a comparative analytical framework to examine credit-default prediction. It relied on a comprehensive dataset of 56,965 loan contracts issued between 2019 and 2024 across the northern branches of Bank Melli Iran. Three modeling approaches were evaluated: traditional logistic regression and two ensemble machine learning methods—random forest (RF) and extreme gradient boosting (XGBoost). The analysis incorporated 29 predictive features categorized into three conceptual groups: loan contract characteristics (e.g., principal amount, repayment tenure, collateral type), borrower attributes (e.g., age, occupational profile, credit history), and institutional factors (e.g., branch location, branch type). Data preprocessing included outlier removal, text categorization, and the extraction of variables such as age and grace period. The models were evaluated under both baseline and optimized (hyperparameter-tuned) settings. The results showed that the machine learning models substantially outperformed the conventional logistic regression model. XGBoost delivered the highest discriminatory power (ROC-AUC = 99.73%), followed closely by RF (99.68%), whereas logistic regression lagged significantly (75.34%). On average, the AUC difference between the machine learning models and logistic regression was approximately 0.243, and statistical tests with 95% confidence intervals

* Corresponding Author: p.mohajeri@atu.ac.ir

How to Cite: Taleblou, R., Kamali, M.A. & Mohajeri, P. (2025). Estimating the Probability of Loan Default in Melli Bank: A Comparative Study of Machine Learning and Econometric Approaches. *Iranian Journal of Economic Research*, *30*(103), 1-41.

confirmed the significance of this gap. Overall, the findings provided strong evidence for the superior reliability of machine learning approaches in forecasting loan default.

1. Introduction

Although traditional econometric models such as logistic regression have long served as the foundation of credit scoring systems, their reliance on linearity assumptions and error independence limits their ability to capture the complex, nonlinear patterns typical of financial data. These limitations are further compounded by sensitivity to multicollinearity and distributional assumptions that are frequently inconsistent with real-world conditions. The present research aimed to address these shortcomings by conducting a rigorous comparative analysis of predictive methodologies within Iran's banking sector—a context in which machine learning applications remain relatively underutilized despite widespread global adoption of artificial intelligence in finance. Specifically, the study intended to compare the performance of two ensemble learning techniques (i.e., random forest and extreme gradient boosting or XGBoost), with that of conventional logistic regression in forecasting loan defaults using extensive real-world data from Bank Melli Iran. The methodological advantages of machine learning approaches arise from their ability to model complex nonlinear relationships without requiring predefined functional forms, to automatically capture variable interactions through hierarchical partitioning, to maintain robustness in the presence of outliers and non-normal distributions, and to detect subtle patterns in high-dimensional data that escape parametric detection. By systematically evaluating these capabilities, the current study tried to offer empirical evidence to support financial institutions in adopting more advanced and reliable risk modeling frameworks.

2. Materials and Methods

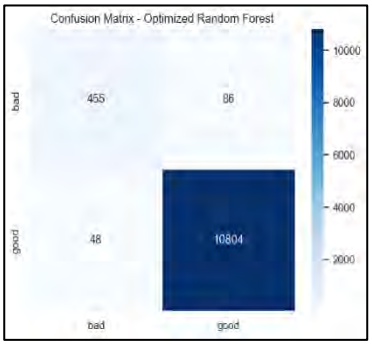
The selection of predictive models in this study is informed by theoretical foundations, empirical literature, and practical forecasting capabilities. Three distinct modeling approaches—random forest (RF), extreme gradient boosting (XGBoost), and logistic regression (LR)—were employed to evaluate their effectiveness in predicting loan defaults. As a widely used ensemble learning algorithm, random forest (RF) builds multiple decision trees using bootstrap aggregating and random subsets of observations and features. Each tree is trained independently, and final predictions are obtained through majority voting (classification) or averaging (regression). This structure reduces overfitting and improves generalization compared to single decision

trees. XGBoost is an advanced gradient boosting algorithm known for its efficiency and high predictive accuracy. XGBoost constructs trees sequentially, with each new tree reducing the residual errors of the ensemble through gradient descent optimization. Rooted in the logistic function and formalized in modern choice modeling, logistic regression improves on linear probability models by mapping predictions to the [0,1] interval via a sigmoid transformation. Although valued for its interpretability, conventional econometric models such as logistic regression suffer from a series limitations, including linearity assumptions, limited interaction detection, multicollinearity sensitivity, and distributional constraints. These methodological constraints potentially compromise predictive performance in complex, non-linear domains such as credit risk assessment.

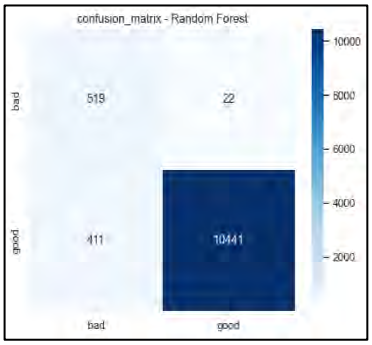
3. Results and Discussion

The machine learning models were evaluated under two configurations: a baseline setting using default parameters and an optimized setting using hyperparameter tuning. Hyperparameters—settings external to the model that are not learned from data—strongly influence predictive accuracy, computational efficiency, and generalization. Suboptimal hyperparameter selection can lead to underfitting or overfitting, thereby compromising model performance. Common optimization strategies include grid search, random search, and Bayesian optimization. Empirical evidence shows that random search is often more efficient in high-dimensional spaces (Bergstra & Bengio, 2012). Although default parameters may yield reasonable baseline performance, they rarely yield optimal performance (Probst et al., 2019). Prior research suggests that systematic tuning can increase accuracy by 10–20% (Hutter et al., 2019) and improve generalization (Liao et al., 2018). In this study, hyperparameters were optimized to maximize the area under the curve (AUC), a standard practice in credit risk modeling (Feurer et al., 2015). This approach can reduce prediction errors and enhance model stability in ensemble methods. The empirical results revealed substantial performance improvements through hyperparameter optimization. For the RF model, accuracy increased from 96% in the untuned configuration to 99% after tuning, with a notable reduction in false negatives and improved precision, albeit with a slight decline in recall for the default class. The optimized XGBoost model—using 375 trees, a maximum depth of 12, and a learning rate of 0.03—achieved the lowest false-negative and false-positive rates, offering an optimal balance between learning capacity and predictive accuracy. In contrast, logistic regression showed limited discriminatory power, with a recall

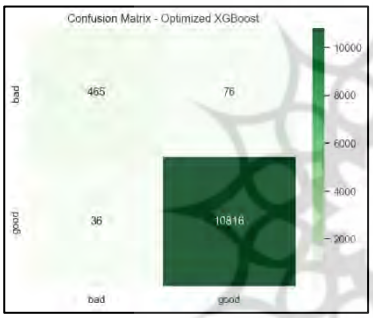
of 0.16 and a ROC-AUC of 0.75, indicating inherent limitations in capturing the complex patterns associated with default events.



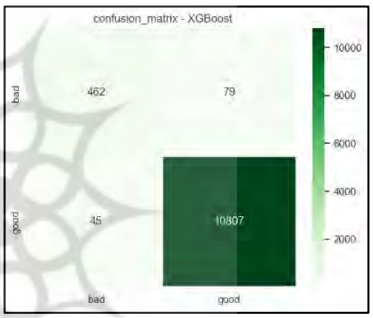
Random Forest Model (With Hyperparameter Tuning)



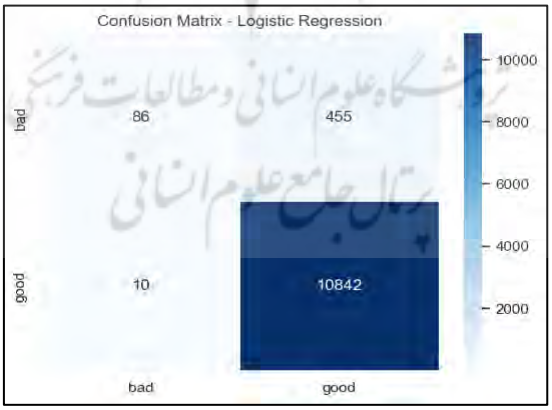
Random Forest Model (Without Hyperparameter Tuning)



(XGBoost) Model (With Hyperparameter Tuning)



(XGBoost) Model (Without Hyperparameter Tuning)



Logistic Regression Model

Source: Research Results

Summary of Model Results

ROC-AUC	F1-Score (Good)	F1-Score (Bad)	Recall (Good)	Recall (Bad)	Precision (Good)	Precision (Bad)	ACCURACY	State	Model
0/935	0/98	0/88	0/99	0/83	0/98	0/94	97%	Unoptimized	RF
0/9968	0/99	0/95	0/99	0/94	0/99	0/97	99%	Optimized	RF
0/9966	0/99	0/90	0/99	0/85	0/99	0/96	98%	Unoptimized	XGBOOST
0/9973	0/99	0/92	0/99	0/88	0/99	0/97	99%	Optimized	XGBOOST
0/7534	0/98	0/27	0/98	0/16	0/96	0/90	96%	-	LR

Source: Research Results

4. Conclusion

The empirical results of this study demonstrates the superior predictive capabilities of machine learning methods—particularly XGBoost)—compared with conventional econometric approaches for estimating the probability of default (PD) in Bank Melli Iran's loan portfolio. This performance gap primarily arises from machine learning algorithms' ability to capture nonlinear relationships and latent structural patterns among default determinants—features that linear parametric models are unable to detect. Model precision was evaluated using several metrics, including confusion matrix analysis, total accuracy, and area under the ROC Curve (AUC). The findings indicated that machine learning models deliver substantially higher predictive precision and improved default detection rates. The optimized XGBoost model achieved outstanding performance (accuracy = 99%, AUC = 0.9973), far surpassing the logistic regression model's ability to identify default cases (recall = 0.16). This distinct performance disparity strongly supports the research hypothesis regarding the comparative advantage of machine learning in PD estimation. Despite their superior predictive performance, the operational deployment of advanced machine learning techniques in financial institutions remains constrained by two key challenges: the computational complexity of hyperparameter optimization and the interpretability limitations inherent in black-box models. These limitations highlight the practical importance of developing hybrid frameworks that integrate the interpretive transparency of traditional methods with the predictive power of machine learning approaches. This research provided evidence of a paradigm shift in credit risk analytics, moving away from the long-standing reliance on conventional statistical models (such as logistic regression and linear probability models) toward machine learning methodologies. While prior studies using traditional techniques achieved moderate success, their limitations in handling imbalanced distributions and complex interaction effects have become increasingly apparent. The present findings align with international research trends

and offer novel empirical evidence from Iran's banking sector—demonstrating that well-tuned machine learning algorithms can achieve unprecedented levels of accuracy (99% accuracy compared with a 16% default identification rate for logistic regression).

Keywords: Probability of Loan Default, Credit Risk, Machine Learning, Random Forest Model, XGBoost Model

JEL Classification: C45, C53, G32, G33

مقاله پژوهشی

تاریخ دریافت: ۱۴۰۳/۱۲/۲۳ تاریخ بازنگری: ۱۴۰۴/۴/۲۳ تاریخ پذیرش: ۱۴۰۴/۴/۲۹ ISSN:۱۷۲۶-۰۷۲۸ eISSN:۲۴۷۶-۶۴۴۵

-- پژوهش‌های اقتصادی ایران ---------

دوره ۳۰، شماره ۱۰۳، تابستان ۱٤۰٤، ۱-٤۱

ijer.atu.ac.ir

https://doi.org/10.22054/ijer.2025.84878.1350

برآورد احتمال نکول تسهیلات اعطایی در بانک ملی: مقایسه رویکردهای یادگیری ماشین و اقتصادسنجی

رضا طالبلو | دانشیار گروه اقتصاد نظری، دانشگاه علامه طباطبائی، تهران، ایران

میرعلی کمالی | دانشجوی دکتری اقتصاد، دانشگاه سمنان، سمنان، ایران

پریسا مهاجری* | دانشیار گروه اقتصاد نظری، دانشگاه علامه طباطبائی، تهران، ایران

چکیده

در این پژوهش، ۵۶,۹۶۵ فقره تسهیلات اعطایی طی سال‌های ۱۳۹۸ تا ۱۴۰۳ در شعب شمال تهران بانک ملی ایران، به‌منظور برآورد احتمال نکول وام مورد بررسی قرار گرفتند. برای پیش‌بینی رفتار اعتباری مشتریان، سه مدل شامل رگرسیون لجستیک، جنگل تصادفی و تقویت گرادیان حداکثری به‌کار گرفته شده است. متغیرهای ورودی شامل ۲۹ متغیر در سه دسته اصلی بودند: مشخصات قرارداد تسهیلات (مبلغ، دوره بازپرداخت، نوع وثیقه و...)، ویژگی‌های فردی تسهیلات‌گیرنده (سن، شغل، سابقه اعتباری و...) و مشخصات شعبه (استان، نوع شعبه و...). همچنین پیش‌پردازش‌هایی مانند حذف مقادیر پرت، دسته‌بندی متون، استخراج سن و دوره تنفس از داده‌های موجود انجام شده است و مدل‌ها در دو حالت پایه و بهینه‌سازی‌شده (با تنظیم ابرپارامترها) ارزیابی شدند. نتایج نشان داد که مدل‌های یادگیری ماشین عملکرد بهتری نسبت به روش سنتی دارند. شاخص ROC-AUC برای مدل تقویت گرادیان حداکثری معادل ۹۹/۷۳ و برای جنگل تصادفی نیز ۹۹/۶۸ درصد برآورد شد درحالی‌که این مقدار برای رگرسیون لجستیک تنها ۷۵/۳۴ درصد بود. اختلاف میانگین AUC بین مدل‌های یادگیری ماشین و رگرسیون لجستیک حدود ۰/۲۴۳ بود و در همه موارد، آزمون‌های آماری و فاصله اطمینان ۹۵ درصد، بر معناداری این اختلاف تأکید داشتند. یافته‌ها برتری قابل اتکای روش‌های یادگیری ماشین در پیش‌بینی نکول تسهیلات را تأیید می‌کند.

کلیدواژه‌ها: احتمال نکول تسهیلات، ریسک اعتباری، یادگیری ماشین، الگوی جنگل تصادفی، الگوی تقویت گرادیان حداکثری

طبقه‌بندی JEL: C45، C53، G32، G33

* نویسنده مسئول: p.mohajeri@atu.ac.ir

۱. مقدمه

امروزه بانک‌ها و مؤسسات مالی با چالش‌های متعددی در مدیریت ریسک‌های اعتباری مواجه هستند. یکی از مهم‌ترین این چالش‌ها، برآورد دقیق احتمال نکول هر فرد تسهیلات گیرنده و نیز میزان نکول سبد تسهیلات اعطایی است. نسبت مطالبات غیرجاری به کل مطالبات[1] یکی از شاخص‌های بااهمیت در ارزیابی عملکرد بانک‌ها محسوب می‌شود. این نسبت نشان‌دهنده کیفیت تسهیلات و مطالبات بانک و به نوعی ریسک اعتباری آن است. براساس مفاد دستورالعمل طبقه‌بندی دارایی‌های مؤسسات اعتباری[2]، مطالبات سیستم بانکی به‌طور کلی به دو دسته اصلی جاری و غیرجاری تقسیم می‌شوند. مطالبات جاری به بدهی‌هایی اطلاق می‌شود که بازپرداخت اصل و سود آن‌ها طبق سررسیدهای تعیین‌شده انجام شده و یا حداکثر تا ۲ ماه از موعد مقرر گذشته باشد. به‌طور کلی یکی از اهداف اصلی بانک‌ها در زمینه مدیریت ریسک، به کاهش این نسبت مرتبط است.

هدف اصلی این پژوهش، برآورد احتمال نکول تسهیلات گیرندگان در یک بانک نمونه (بانک ملی) و مقایسه دقت مدل‌های یادگیری ماشین[3] و اقتصادسنجی در این زمینه است. با توجه به اینکه مدل‌های یادگیری ماشین توانایی شناسایی الگوهای پیچیده و غیرخطی در داده‌ها را دارند، استفاده از این رویکرد می‌تواند دقت پیش‌بینی را بهبود بخشد. در مقابل، مدل‌های اقتصادسنجی مانند رگرسیون لجستیک بر پایه فروضی نظیر خطی بودن رابطه بین متغیرها و استقلال خطاها عمل می‌کنند که ممکن است در داده‌های اعتباری واقعی برقرار نباشند. همچنین این مدل‌ها حساسیت بیشتری به چندهمخطی و نرمال بودن داده‌ها دارند. در مقابل، مدل‌های یادگیری ماشین مانند جنگل تصادفی[4] و تقویت گرادیان حداکثری[5] انعطاف‌پذیری بالاتری در مواجهه با داده‌های پرت، تعامل بین متغیرها و ساختارهای پیچیده دارند و می‌توانند الگوهایی را که از دید روش‌های سنتی پنهان می‌مانند، شناسایی کنند.

---

1. Non-Performing Loans (NPL)

۲. در پیوست بخشنامه شماره مب/۲۸۲۳ به تاریخ ۱۳۸۵/۱۲/۵ از سوی اداره مطالعات و مقررات بانکی بانک مرکزی جمهوری اسلامی ایران ارائه شده است.

3. Machine Learning (ML)
4. Random Forest (RF)
5. eXtreme Gradient Boosting (XGBOOST)

بااین‌حال، به‌رغم پیشرفت‌های قابل‌توجه در حوزه هوش مصنوعی[1] و یادگیری ماشین، کاربرد این روش‌ها در ارزیابی ریسک اعتباری در ایران همچنان محدود بوده است. در تحقیقات پیشین داخلی، رحمانی و اسماعیلی (۱۳۸۹) کارایی مدل‌های شبکه عصبی، رگرسیون لجستیک و تحلیل تمایزی را برای پیش‌بینی نکول بررسی کرده‌اند و برتری روش‌های غیرخطی را نسبت به اقتصادسنجی سنتی نشان داده‌اند. همچنین، توکلی و آشتاب (۱۴۰۲) با بررسی گسترده در بورس اوراق بهادار تهران دریافته‌اند که مدل‌های یادگیری ماشین با دقت AUC بیش از ۹۹ درصد به‌طور قابل توجهی از مدل‌های آماری برتر هستند. در سطح پژوهش‌های ساختاری، پیکانی و همکاران[2] (۲۰۲۳) نشان داده‌اند که مدل‌های مبتنی بر درخت مانند جنگل تصادفی نسبت به مدل‌های ساختاری مرتون و گسک[3] عملکرد برتری در پیش‌بینی نکول داشته‌اند. استفاده از یادگیری ماشین در برآورد احتمال نکول با چالش‌های خاصی همراه است. انتخاب مدل مناسب برای برآورد احتمال نکول سبد تسهیلات اعطایی بانک‌ها یکی از مسائل کلیدی است. هر یک از مدل‌های مختلف یادگیری ماشین دارای مزایا و معایب خاص خود هستند و انتخاب مدل مناسب نیازمند بررسی دقیق است. علاوه‌براین، تنظیم هایپرپارامترهای[4] مدل‌ها نیز نقش مهمی در دقت برآوردها دارد. بهینه‌سازی هایپرپارامترها می‌تواند به بهبود عملکرد مدل‌ها کمک کند. کیفیت داده‌های ورودی نیز از اهمیت بالایی برخوردار است. داده‌های ناقص یا نادرست می‌توانند منجر به نتایج نادرست شوند. بنابراین، پیش‌پردازش داده‌ها و بهبود کیفیت آن‌ها ازجمله مراحل حیاتی در این پژوهش است. روش‌های مختلفی برای پیش‌پردازش داده‌های اعتباری وجود دارد که هر یک می‌توانند آثار متفاوتی بر دقت مدل‌ها داشته باشند. در این پژوهش، تلاش می‌شود تا با بررسی دقیق این چالش‌ها و ارائه راهکارهای مناسب، برآورد دقیق‌تری از احتمال نکول صورت پذیرد.

پژوهش حاضر از سه جهت دارای نوآوری است: نخست، استفاده از مجموعه‌داده‌ای بزرگ و واقعی شامل بیش از ۵۶ هزار نمونه از داده‌های تسهیلات اعطایی در بزرگ‌ترین بانک کشور؛ دوم، مقایسه مستقیم عملکرد مدل‌های پیشرفته یادگیری ماشین با مدل‌های

1. Artificial Intelligence (AI)
2. Peykani, P., et al.
3. Merton, R. C & Geske, R.
4. Hyperparameters

سنتی اقتصادسنجی و سوم، بررسی اثر بهینه‌سازی هایپرپارامترها بر دقت پیش‌بینی که در مطالعات پیشین کمتر مورد توجه قرار گرفته است.

به منظور دستیابی به اهداف تحقیق، مقاله حاضر در ۶ بخش سازماندهی می‌شود. پس از مقدمه که محور اول از مقاله حاضر را تشکیل می‌دهد، پیشینه نظری و تجربی پژوهش در بخش دوم و سوم ارائه می‌گردد. روش پژوهش با تأکید بر روش‌های مختلف یادگیری ماشین در برآورد احتمال نکول در بخش چهارم ارائه شده و داده‌های مورد استفاده در این مقاله نیز در همان بخش معرفی می‌شوند. در بخش پنجم به برآورد و تنظیم پارامترهای الگوها پرداخته می‌شود و در نهایت، بخش ششم به جمع‌بندی یافته‌های کلیدی و ارائه پیشنهادها اختصاص می‌یابد.

۲. مبانی نظری پژوهش

نظریه اطلاعات نامتقارن یکی از مهم‌ترین نظریه‌های اقتصادی و مالی است که به تحلیل موقعیت‌هایی می‌پردازد که در آن یکی از طرفین یک معامله یا قرارداد اطلاعات بیشتری نسبت به طرف دیگر دارد. این عدم تعادل اطلاعاتی می‌تواند منجر به ناکارآمدی بازار، تخصیص نادرست منابع و درنهایت ریسک‌های اقتصادی شود. آکرلوف[1] (۱۹۷۰)، اسپنس[2] (۱۹۷۳) و استیگلیتز و وایس[3] (۱۹۸۱) ازجمله اقتصاددانانی هستند که در دهه ۱۹۷۰ به‌صورت جدی به بررسی این موضوع پرداخته و برای آن جایزه نوبل اقتصاد را دریافت کردند. این نظریه عمدتاً به دو مصداق کلیدی نامتقارن بودن اطلاعات در اقتصاد می‌پردازد: کژگزینی[4] و کژمنشی[5]. هر دوی این مفاهیم نشان‌دهنده مسائلی هستند که می‌توانند در نتیجه وجود اطلاعات نامتقارن در معاملات اقتصادی یا قراردادهای مالی بروز کنند.

در حوزه مدیریت ریسک اعتباری، گزینش نامناسب زمانی رخ می‌دهد که بانک‌ها یا مؤسسات مالی نتوانند به‌درستی ریسک اعتباری مشتریان را پیش‌بینی کنند. مشتریانی با ریسک بالاتر ممکن است تمایل بیشتری به دریافت وام با نرخ بهره بالا داشته باشند

1. Akerlof, G.A.
2. Spence, M.
3. Stiglitz, J.E. & Weiss, A.
4. Adverse Selection
5. Moral Hazard

درحالی‌که مشتریان با ریسک پایین‌تر ممکن است از بازار خارج شوند. این موضوع می‌تواند منجر به افزایش نکول تسهیلات و زیان مالی برای بانک‌ها شود. همچنین در رابطه با کژمنشی در صنعت بانکداری، می‌توان گفت وام‌گیرندگان ممکن است پس از دریافت وام، رفتارهایی انجام دهند که احتمال نکول را افزایش می‌دهد زیرا می‌دانند که زیان احتمالی بر دوش بانک یا سایر بازیگران مالی است. کژمنشی نخستین بار توسط آرو[1] در سال ۱۹۶۳ میلادی مورد بررسی قرار گرفت. بعدها استیگلیتز و وایس (۱۹۸۱) نشان دادند که اطلاعات نامتقارن میان بانک‌ها و وام‌گیرندگان، عامل اصلی ایجاد این ریسک است. این نظریه نشان می‌دهد که وجود بیمه یا قراردادهای مالی می‌تواند انگیزه افراد برای رفتار مسئولانه را کاهش دهد زیرا افراد ممکن است احساس کنند که هزینه پیامدهای منفی رفتارهایشان به دوش آن‌ها نخواهد بود.

در دهه‌های اخیر، ادبیات اطلاعات نامتقارن با بهره‌گیری از فناوری‌های نوین، به‌ویژه در حوزه تحلیل داده‌های کلان و یادگیری ماشین، گسترش قابل توجهی یافته است. پژوهش‌های جدید نشان می‌دهند که ابزارهای مدرن تحلیل داده می‌توانند مکمل مناسبی برای چارچوب‌های نظری کلاسیک باشند و به شناسایی الگوهای رفتاری پنهان در وام‌گیرندگان کمک کنند. برای مثال، گوئو[2] (۲۰۱۶) در مطالعه‌ای تجربی نشان داد که با استفاده از الگوریتم‌های طبقه‌بندی، می‌توان رفتارهای پرریسک مشتریان بانکی را پیش از وقوع نکول شناسایی کرد. در همین راستا، تانگ و همکاران[3] (۲۰۱۹) با استفاده از شبکه‌های عصبی و الگوریتم‌های یادگیری عمیق، توانستند احتمال بروز کژمنشی در قراردادهای اعتباری را با دقت بالایی پیش‌بینی کنند. این نوع مطالعات بر اهمیت به‌کارگیری مدل‌های داده‌محور تأکید دارند و نشان می‌دهند که نظریه اطلاعات نامتقارن نه‌تنها در سطح مفهومی بلکه در عمل نیز قابل اندازه‌گیری و کنترل است.

همچنین، پژوهشگران دیگری همچون کینگ و همکاران[4] (۲۰۲۱) و لیو[5] (۲۰۲۰) با بررسی بانک‌های تجاری در کشورهای مختلف، به این نتیجه رسیدند که ترکیب اطلاعات

1. Arrow, K.J.
2. Guo, C.
3. Tang, Y., et al.
4. King, M., et al.
5. Liu, H.

رفتاری مشتریان با داده‌های مالی سنتی، دقت سیستم‌های اعتبارسنجی را به‌طور معناداری افزایش می‌دهد. این نتایج بیانگر آن است که مدل‌های پیشرفته یادگیری ماشین، به دلیل توانایی در شناسایی روابط غیرخطی و پیچیده، می‌توانند ابزار مؤثری برای کاهش آثار اطلاعات نامتقارن باشند. بنابراین، نظریه اطلاعات نامتقارن در دنیای امروز، تنها یک چارچوب مفهومی نیست بلکه به‌صورت عملیاتی با کمک فناوری‌های نوین قابل پیاده‌سازی است و به مدیران ریسک این امکان را می‌دهد تا با دقت بیشتری بر رفتارهای مخاطره‌آمیز نظارت کنند.

۳. پیشینه پژوهش

مطالعات متعددی در داخل و خارج از کشور برروی برآورد احتمال نکول تسهیلات بانکی با روش‌های متفاوت انجام شده است. مطالعات داخلی عمدتاً از مدل‌های آماری سنتی برای تحلیل داده‌ها و پیش‌بینی احتمال نکول استفاده کرده‌اند. به‌رغم این تلاش‌ها، همچنان استفاده از مدل‌های پیشرفته‌تر مانند جنگل تصادفی، تقویت گرادیان حداکثری، شبکه‌های عصبی مصنوعی[1] و سایر مدل‌های یادگیری ماشین در مطالعات داخلی نسبتاً محدود باقی مانده است. درحالی‌که مطالعات خارجی طی سال‌های اخیر به‌طور گسترده از روش‌های نوین یادگیری ماشین برای بهبود دقت پیش‌بینی ریسک اعتباری بهره برده‌اند، کاربرد این روش‌ها در ایران چندان مورد توجه قرار نگرفته است. با توجه به پیشرفت‌های قابل‌توجه در حوزه یادگیری ماشین و نتایج مطلوبی که این روش‌ها در مدل‌سازی ریسک اعتباری در پژوهش‌های خارجی داشته‌اند، مشخص است که در ایران هنوز از تمامی ظرفیت‌های این روش‌ها استفاده نشده است. اکثر پژوهش‌های داخلی بر روش‌های سنتی و آماری متمرکز بوده و بهره‌گیری از مدل‌های پیچیده‌تر مانند مدل‌های یادگیری عمیق یا الگوریتم‌های تقویتی (مانند تقویت گرادیان حداکثری) همچنان در مرحله تحقیقاتی محدودی قرار دارد.

1. Artificial Neural Networks (ANN)

۱-۳. پیشینه‌های داخل کشور

پژوهش‌های صورت گرفته در ایران طی پانزده سال اخیر در حوزه برآورد احتمال نکول، محدود به چند مطالعه است که صرفاً تعدادی از آن‌ها بر روش‌های یادگیری ماشین متمرکز بوده است. این پژوهش‌ها در ۳ طبقه قابل تقسیم‌بندی هستند:[۱]

- گروه نخست- برخی از مطالعات با بهره‌گیری از روش‌های اقتصادسنجی همانند مدل‌های خطی[۲]، غیرخطی (لاجیت و پروبیت[۳]) به برآورد احتمال نکول تسهیلات‌گیرندگان پرداخته‌اند.
- گروه دوم- تعدادی از مطالعات به بررسی ریسک کلی سبد تسهیلات با استفاده از سنجه‌های دیگر پرداخته‌اند.
- گروه سوم- توکلی و آشتاب (۱۴۰۲) در مطالعه‌ای با عنوان «مقایسه کارایی مدل‌های یادگیری ماشین و مدل‌های آماری در پیش‌بینی ریسک مالی»، عملکرد نسبی این دو دسته از مدل‌ها را مورد ارزیابی قرار دادند.

یافته‌های این پژوهش نشان می‌دهد که مدل‌های یادگیری ماشین، در مقایسه با روش‌های آماری سنتی، از دقت بالاتری در پیش‌بینی ریسک مالی برخوردار هستند. همچنین پیکانی و همکاران (۲۰۲۳) در مطالعه خود با عنوان «کاربرد مدل‌های ساختاری و یادگیری ماشین برای پیش‌بینی ریسک نکول شرکت‌های پذیرفته‌شده در بازار سرمایه ایران»، پژوهشگران به مقایسه دو مدل ساختاری مرتون و گسک با دو مدل یادگیری ماشین یعنی جنگل تصادفی و درخت تصمیم‌گیری تقویت گرادیان[۴] پرداخته‌اند. داده‌ها مربوط به شرکت‌های بورسی ایران بوده و هدف، پیش‌بینی ریسک نکول آن‌هاست. نتایج پژوهش نشان می‌دهد که مدل‌های یادگیری ماشین به‌ویژه مدل‌های درختی، عملکرد به‌مراتب بهتری در پیش‌بینی نکول دارند. این مطالعه بر برتری روش‌های هوشمند نسبت به مدل‌های کلاسیک تأکید می‌کند.

---

۱. خلاصه‌ای از مطالعات و پژوهش‌های انجام شده نزد نویسندگان است که به دلیل اجتناب از تطویل مقاله و با عنایت به عدم کاربرد روش‌های یادگیری ماشین در این مقالات، از توضیح آنها اجتناب شده است.

2. Linear Probability Model (LPM)
3. Logit & Probit
4. Gradient Boosting Decision Tree (GBDT)

به‌علاوه، رحمانی و اسماعیلی (۱۳۸۹) در مطالعه‌ای با عنوان «کارایی شبکه‌های عصبی، رگرسیون لجستیک و تحلیل تمایزی در پیش‌بینی نکول»، به بررسی مقایسه‌ای این سه روش پرداخته‌اند. نتایج پژوهش آن‌ها نشان داد که متغیرهای مورد استفاده در مدل‌ها از لحاظ آماری در پیش‌بینی نکول معنادار هستند و در مقایسه میان مدل‌ها، شبکه‌های عصبی عملکرد بهتری نسبت به رگرسیون لجستیک و تحلیل تمایزی از خود نشان داده‌اند.

مقاله موحدی‌نیا و بهمئی (۱۳۹۴) نیز بر روش‌های اولیه یادگیری ماشین متمرکز بوده است.

با وجود این تلاش‌ها، شکاف مهمی در ادبیات داخلی به چشم می‌خورد: تاکنون هیچ مطالعه‌ای از روش‌های نوین یادگیری ماشین برای برآورد احتمال نکول در داده‌های واقعی بانکی شامل تسهیلات اعطایی استفاده نکرده است. اغلب پژوهش‌ها بر شرکت‌های بورسی یا داده‌های ثانویه تمرکز داشته‌اند و یا از مدل‌های پایه‌تر استفاده کرده‌اند. همچنین مقایسه‌ای جامع میان عملکرد مدل‌های اقتصادسنجی و یادگیری ماشین در چارچوب داده‌های بانکی انجام نشده است. این خلأ پژوهشی، لزوم انجام مطالعه‌ای با تمرکز بر داده‌های واقعی از تسهیلات بانکی، مدل‌سازی نکول در سطح فردی و استفاده از مدل‌های پیشرفته را آشکار می‌سازد. پژوهش حاضر در راستای پر کردن این شکاف و ارائه بینش دقیق‌تری نسبت به مزایای نسبی مدل‌های نوین در حوزه اعتباری ایران انجام شده است.

۲-۳. پیشینه‌های خارج از کشور

به‌رغم خلأ پژوهشی در مطالعات داخلی، پژوهش‌های خارجی متعددی در سال‌های اخیر با به‌کارگیری مدل‌های یادگیری ماشین و هوش مصنوعی به برآورد احتمال نکول و ریسک اعتباری در شبکه بانکی پرداخته‌اند که بعضی از آن‌ها در ادامه مورد بررسی قرار می‌گیرند.

گلوموا بیبی مخلصه مخمودجون قیزی[1] (۲۰۲۳) مدل جامع پیش‌بینی ریسک اعتباری و تخصیص وام به کسب‌وکارهای کوچک و متوسط را بررسی کرده است. در این مدل از داده‌های مالی خام و روش‌های آماری مانند رگرسیون لجستیک و توزیع برنولی برای پیش‌بینی احتمال نکول استفاده شده است. یافته‌ها نشان می‌دهند که این طرح به‌راحتی در سناریوهای واقعی قابل پیاده‌سازی است و به کاهش نگرانی‌های بانک‌ها و تسهیل تأمین مالی

1. G'ulomova, B.M.M. qizi

کسب‌وکارها کمک می‌کند. این تحقیق تأکید دارد که ترکیب عوامل کلیدی می‌تواند مزایای متقابلی برای بانک‌ها و کسب‌وکارها ایجاد کند. علاوه‌بر پژوهش فوق، مطالعات بسیاری صرفاً در ۲ سال گذشته درخصوص محاسبات ریسک اعتباری در خارج از ایران منتشر شده است که خلاصه‌ای از آن‌ها در جدول ۱ ارائه شده است.

جدول ۱. الگوها و نتایج اندازه‌گیری ریسک اعتباری در پژوهش‌های اخیر بین‌المللی

نویسندگان	سال	مدل	نتیجه مطالعه
ناتاشا رابینسون و نیدهی سیندهوآنی[1]	۲۰۲٤	الگوریتم جنگل تصادفی، الگوریتم گرادیان بوستینگ	مدل‌های یادگیری ماشین قابلیت ارزیابی دقیق‌تر و مؤثرتر ریسک اعتباری را در مقایسه با روش‌های سنتی نشان دادند. جنگل تصادفی و گرادیان بوستینگ به‌طور مؤثر الگوهای پیچیده در داده‌های بزرگ را شناسایی می‌کنند که از دید روش‌های سنتی پنهان می‌مانند.
آبیسولا آکینجوله و همکاران[2]	۲۰۲٤	جنگل تصادفی، درخت تصمیم، ماشین بردار پشتیبان، تقویت گرادیان حداکثری، تقویت تطبیقی و پرسپترون چندلایه	ترکیب پیش‌بینی مدل‌ها از طریق روش‌های رأی‌گیری و انباشته‌سازی عملکرد مدل‌ها را به‌طور چشمگیری بهبود داد. این نتایج نشان‌دهنده پتانسیل روش‌های ترکیبی برای بهبود پیش‌بینی ریسک نکول و کاهش نرخ نکول و زیان‌های مالی است.
دیگامبر اوپهاده و همکاران[3]	۲۰۲٤	جنگل تصادفی، رگرسیون لجستیک، درخت تصمیم، k-نزدیک‌ترین همسایه، ماشین بردار پشتیبان، تقویت گرادیان حداکثری، تقویت تطبیقی و الگوریتم گرادیان بوستینگ	الگوریتم جنگل تصادفی توانایی بالایی در حل مسائل طبقه‌بندی دوتایی دارد. متغیرهای سنتی امتیازدهی اعتباری بر مشکلات پیش‌بینی نکول تأثیر قابل توجهی دارند.
خوزه آنتونیو نویز مورا و همکاران[4]	۲۰۲۳	جنگل تصادفی	الگوریتم جنگل تصادفی توانایی بالایی در حل مسائل طبقه‌بندی دوتایی دارد.

مأخذ: یافته‌های پژوهش

1. Robinson, N. & Sindhwani, N.
2. Akinjole, A., et al.
3. Uphade, D.B., et al.
4. Nuez Mora, J.A., et al.

٤. روش پژوهش و پایه‌های آماری

٤-١. روش پژوهش

انتخاب مدل‌های مناسب در این پژوهش براساس مبانی نظری، مطالعات پیشین و قابلیت آن‌ها در پیش‌بینی رفتار نکول انجام شده است. برای این منظور، سه مدل تحقیق شامل جنگل تصادفی، تقویت گرادیان حداکثری و رگرسیون لجستیک به‌کار گرفته شده‌اند. در ادامه به مفاهیم، مبانی و مقایسه هر یک از این مدل‌ها خواهیم پرداخت. درنهایت، شاخص‌ها و معیارهای ارزیابی مدل‌ها معرفی شده و نقش آن‌ها در سنجش دقت و کارایی الگوهای پیشنهادی مورد بررسی قرار می‌گیرد.

٤-١-١. الگوی جنگل تصادفی

جنگل تصادفی یکی از روش‌های پرکاربرد در یادگیری ماشین است که نخستین بار توسط هو[1] (۱۹۹۵) و سپس به‌طور جامع‌تر توسط بریمن[2] (۲۰۰۱) معرفی شد. این مدل ترکیبی از چندین درخت تصمیم است که برای بهبود دقت پیش‌بینی و کاهش بیش‌برازش مورد استفاده قرار می‌گیرد. جنگل تصادفی به‌عنوان یکی از الگوریتم‌های یادگیری نظارت‌شده[3] در مسائل طبقه‌بندی و رگرسیون کاربرد دارد (Breiman, 2001). درواقع جنگل تصادفی یکی از روش‌های مبتنی بر یادگیری گروهی[4] است که با ترکیب چندین درخت تصمیم‌گیری، مدلی مقاوم در برابر بیش‌برازش ارائه می‌دهد. این روش بر پایه مفهوم بوت‌استرپینگ[5] یا بگینگ[6] توسعه یافته است (بریمن، ۱۹۹۶). در جنگل تصادفی، هر درخت تصمیم به‌طور مستقل از مجموعه‌ای تصادفی از داده‌ها و ویژگی‌ها[7] (متغیرهای توضیحی) آموزش می‌بیند و درنهایت، خروجی مدل براساس رأی‌گیری اکثریت (برای مسائل دسته‌بندی) یا میانگین‌گیری (برای مسائل رگرسیون) تعیین می‌شود. برای درک نحوه عملکرد این مدل، باید به مراحل زیر توجه کرد:

---

1. Ho, T.K.
2. Breiman, L.
3. Supervised Learning
4. Ensemble Learning
5. Bootstrap Aggregating
6. Bagging
7. Features

یک- انتخاب مجموعه‌ای از نمونه‌های تصادفی[1]: در ابتدا، مجموعه داده ورودی D شامل N نمونه داده در اختیار داریم ($N_1 = \{Y_i X_i\}_i$). در روش نمونه‌گیری بوت استرپینگ، به‌صورت تصادفی و با جایگذاری از مجموعه داده، چندین زیرمجموعه داده تولید می‌شود که هر یک برای آموزش یک درخت تصمیم جداگانه استفاده می‌شود.

دو- ساخت درخت‌های تصمیم با ویژگی‌های تصادفی: برخلاف درخت‌های تصمیم معمولی که در هر گره، بهترین ویژگی را برای جداسازی انتخاب می‌کنند، در جنگل تصادفی تنها از زیرمجموعه‌ای تصادفی از ویژگی‌ها (m ویژگی از بین M کل ویژگی‌ها. به‌طوری که M>m) برای انتخاب بهترین ویژگی تقسیم‌کننده استفاده می‌شود. این روش، همبستگی بین درخت‌ها را کاهش می‌دهد و دقت مدل را افزایش می‌دهد. برای تعیین بهترین تقسیم، تابع کاهش ناخالصی[2] محاسبه می‌شود. در مسائل طبقه‌بندی، از معیار آنتروپی[3] یا شاخص جینی[4] استفاده می‌شود. در مسائل رگرسیون، معیار میانگین مربعات خطا[5] برای انتخاب بهترین تقسیم به‌صورت رابطه (۱) استفاده می‌شود (بریمن، ۲۰۰۱):

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i) \quad (۱)$$

N تعداد نمونه‌ها، y_i مقدار واقعی متغیر هدف برای نمونه iام، $\hat{y}_i$ مقدار پیش‌بینی شده توسط مدل برای نمونه iام را در معادله (۱) نشان می‌دهد.

سه- ترکیب خروجی درخت‌ها[6]: هنگامی که مجموعه‌ای از درخت‌های تصمیم ساخته شد، هر درخت به‌صورت جداگانه یک پیش‌بینی ارائه می‌دهد. برای ترکیب خروجی این درخت‌ها، از روش‌های زیر به شکل‌های ۲ و ۳ استفاده می‌شود:

در طبقه‌بندی: رأی‌گیری اکثریت[7]

$$\hat{Y} = arg\ max_k \sum_{b=1}^{B} I(T_b(X) = k) \quad (۲)$$

$\hat{Y}$ پیش‌بینی نهایی برای نمونه X . این مقدار کلاسی است که بیشترین رأی را از مدل‌های پایه دریافت کرده است. $arg\ max_k$ تابعی که کلاس k را برمی‌گرداند که بیشترین تعداد

1. Bootstrap Sampling
2. Impurity Reduction
3. Entropy
4. Gini Index
5. MSE - Mean Squared Error
6. Ensemble Aggregation
7. Majority Voting

رأی را داشته باشد. $\sum_{b=1}^{B}$ جمع تعداد رأی‌ها از تمام مدل‌های پایه (درخت‌های تصمیم) که در مجموعه B مدل وجود دارند. $I(T_b(X)=k)$ تابع نشانگر که مقدار ۱ برمی‌گرداند (اگر مدل پایه b (درخت تصمیم Tb) نمونه X را به کلاس k نسبت دهد، در غیر این صورت مقدار ۰ برمی‌گرداند. $T_b(X)$ پیش‌بینی مدل پایه b (درخت تصمیم Tb) برای نمونه X است. k کلاس‌های ممکن در مسئله طبقه‌بندی.

در رگرسیون: میانگین‌گیری از خروجی درخت‌ها

$$\hat{Y}=\frac{1}{B}\sum_{b=1}^{B}T_b(X) \qquad (۳)$$

$\hat{Y}$: پیش‌بینی نهایی برای نمونه X. این مقدار میانگین پیش‌بینی‌های تمام مدل‌های پایه است. $\frac{1}{B}$ ضریب نرمال‌سازی که میانگین پیش‌بینی‌ها را محاسبه می‌کند. $\sum_{b=1}^{B}$ جمع پیش‌بینی‌های تمام مدل‌های پایه (درخت‌های تصمیم) که در مجموعه B مدل وجود دارند. $T_b(X)$ پیش‌بینی مدل پایه b (درخت تصمیم Tb) برای نمونه X است. B تعداد مدل‌های پایه (درخت‌های تصمیم) در مجموعه را نشان می‌دهد.

٤-۱-۲. الگوی تقویت گرادیان حداکثری

الگوی تقویت گرادیان حداکثری یک روش یادگیری ماشین مبتنی بر بوستینگ گرادیان[1] است که به دلیل سرعت، دقت بالا و کارایی بهینه در بسیاری از زمینه‌های یادگیری ماشین و کاربردهای صنعتی مورد استفاده قرار گرفته است. بوستینگ گرادیان یک روش یادگیری تقویتی[2] است که اولین بار توسط فریدمن[3] (۲۰۰۱) توسعه داده شد. این روش بهبودیافته‌ای از درخت‌های تصمیم‌گیری است که مدل‌های ضعیف را ترکیب کرده و یک مدل قوی‌تر می‌سازد. بوستینگ به معنای تقویت مدل‌های ضعیف از طریق ترکیب چندین پیش‌بینی‌گر است. بعدها، فریدمن مفهوم بوستینگ گرادیان را مطرح کرد که در آن، مدل‌ها به جای استفاده از ضرایب ثابت (مانند تقویت گرادیان تطبیقی[4])، براساس گرادیان تابع خطا به‌روزرسانی می‌شوند. در این روش، مدل به‌طور تدریجی پیش‌بینی‌های نادرست را اصلاح کرده و مدل جدید را طوری می‌سازد که خطای مدل‌های قبلی را کاهش دهد. درواقع

1. Gradient Boosting Machine - GBM
2. Boosting
3. Friedman, J.H.
4. Adaptive Boosting (AdaBoost)

تفاوت‌های اصلی تقویت گرادیان تطبیقی و بوستینگ گرادیان در نحوه وزن دهی نمونه‌ها است. مزیت بوستینگ گرادیان در این است که با یادگیری تدریجی از خطاهای گذشته، مدل نهایی با دقت بیشتری نسبت به روش‌های ساده‌تر برآورد می‌کند.

مدل تقویت گرادیان حداکثری که توسط چن و گوئسترین[1] در سال ۲۰۱۶ معرفی شد، نسخه بهینه‌شده‌ای از بوستینگ گرادیان است که دارای مقیاس‌پذیری بالا، پردازش موازی و کنترل بهتر پیچیدگی مدل می‌باشد. بهبودهای تقویت گرادیان حداکثری نسبت به مدل‌های سنتی بوستینگ گرادیان به این ترتیب است:

- استفاده از جریمه منظم‌سازی[2] برای جلوگیری از بیش‌برازش
- پیاده‌سازی موازی و توزیع‌شده برای پردازش داده‌های بزرگ
- انتخاب ویژگی هوشمند و مدیریت داده‌های پراکنده

شرح الگوی تقویت گرادیان حداکثری در این بخش، مستقیماً مبتنی بر تبیین ارائه‌شده در مقاله «مقیاس‌پذیری الگوریتم‌های یادگیری با استفاده از تقویت گرادیان حداکثری» نوشته چن و گوئسترین (۲۰۱۶) است. تقویت گرادیان حداکثری با استفاده از یک مجموعه درخت تصمیم به‌طور تکراری مدل را بهینه‌سازی می‌کند. در هر مرحله، یک درخت جدید به مدل اضافه می‌شود تا خطای پیش‌بینی مدل را کاهش دهد. مراحل کلی اجرای این الگو شامل موارد زیر است:

گام اول: تعریف تابع هدف[3]: تقویت گرادیان حداکثری براساس کمینه‌سازی تابع هزینه و اضافه کردن یک جریمه برای جلوگیری از پیچیدگی بیش از حد مدل کار می‌کند. تابع هزینه کلی از دو بخش تشکیل شده است:

- ✓ تابع خطای پیش‌بینی یا تابع هزینه[4]: برای محاسبه میزان تفاوت بین مقدار واقعی و مقدار پیش‌بینی‌شده.
- ✓ تابع منظم‌سازی[5]: برای جلوگیری از پیچیدگی بیش‌ازحد مدل.

1. Chen, T. & Guestrin, C.
2. Regularization
3. Objective Function
4. Loss Function
5. Regularization Term

ابتدا به بخش منظم‌سازی تابع هدف در رابطه (۴) می‌پردازیم و سپس بخش خطای پیش‌بینی را مورد بررسی قرار خواهیم داد. تابع منظم سازی به‌صورت زیر تعریف می‌شود:

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda||w||^2 \quad (۴)$$

که در آن T تعداد برگ‌های درخت است، $||w||^2$ مجذور وزن‌های مدل و γ و λ پارامترهای کنترل پیچیدگی مدل هستند. اتفاقی که در تابع منظم‌سازی رخ می‌دهد به این صورت است که این تابع از بزرگ شدن غیرضروری درخت جلوگیری کرده و مدل را ساده‌تر و قابل تعمیم‌تر نگه می‌دارد. به این معنا که بخش γT باعث می‌شود اضافه شدن هر گره جدید به درخت هزینه داشته باشد. درنتیجه، مدل از رشد غیرضروری درخت جلوگیری می‌کند. همچنین بخش $\frac{1}{2}\lambda||w||^2$ با جریمه کردن مقادیر بزرگ وزن‌های برگ، از ایجاد مقادیر افراطی و بیش‌برازش جلوگیری می‌کند. از طرف دیگر، تقویت گرادیان حداکثری از تکنیک هرس درخت از پیش استفاده می‌کند که مبتنی بر مقدار آورده هر شاخه است. هنگام ایجاد گره‌های جدید، اگر افزایش بهبود مدل (آورده) کمتر از مقدار آستانه‌ای γ باشد، گره ایجاد نمی‌شود. این کار باعث می‌شود تنها گره‌هایی که ارزش افزوده بالایی دارند، باقی بمانند و سایر گره‌ها حذف شوند.

بخش دیگر تابع هدف، بخش خطای پیش‌بینی است. در الگوی تقویت گرادیان حداکثری، تابع خطای پیش‌بینی یا تابع هزینه، میزان اختلاف بین مقدار واقعی و مقدار پیش‌بینی‌شده را اندازه‌گیری می‌کند. این تابع در فرآیند یادگیری مدل استفاده می‌شود تا تصمیم بگیرد چگونه پارامترهای مدل (مانند وزن‌های برگ درخت) به‌روزرسانی شوند. این بخش خطای مدل را روی همه داده‌های آموزشی خلاصه می‌کند. درواقع در فرآیند آموزش، مدل سعی می‌کند مقدار این اختلاف را کمینه کند، یعنی پیش‌بینی‌های خود را به مقادیر واقعی نزدیک‌تر کند. از این تابع برای محاسبه گرادیان[1] و هشین[2] در فرآیند به‌روزرسانی استفاده می‌شود.

درنهایت تجمیع این دو بخش تابع هدف را تشکیل می‌دهد که این تابع در رابطه (۵) تعریف می‌شود:

$$L(\Theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \quad (۵)$$

1. Gradient
2. Hessian

که در آن $l(y_i,\hat{y}_i)$ تابع خطای پیش‌بینی است (مثلاً MSE در مسائل رگرسیون یا Log Loss در طبقه‌بندی)، $\Omega(f_k)$ بخش منظم‌سازی است که پیچیدگی مدل را کاهش می‌دهد و از بیش‌برازش جلوگیری می‌کند، K تعداد درخت‌های استفاده‌شده در مدل است و n تعداد نمونه‌های موجود در مجموعه داده است.

گام دوم: ساخت درخت‌های تصمیم بهینه: الگوی تقویت گرادیان حداکثری از درخت‌های تصمیم‌گیری تکرارشونده استفاده می‌کند. هر درخت جدید برای تصحیح خطاهای درخت قبلی ساخته می‌شود. مقدار پیش‌بینی در مرحله tام به شکل رابطه (۶) به‌روزرسانی می‌شود:

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(x_i) \quad (۶)$$

که در رابطه فوق $\hat{y}_i^{(t)}$ پیش‌بینی تجمعی برای نمونه i در گام t. این مقدار، مجموع پیش‌بینی‌های تمام درخت‌های قبلی و درخت فعلی است. $\hat{y}_i^{(t-1)}$ پیش‌بینی تجمعی برای نمونه i در گام $t-1$ (گام قبلی). این مقدار، مجموع پیش‌بینی‌های تمام درخت‌های قبلی است. $f_t(x_i)$ پیش‌بینی درخت t برای نمونه i. این مقدار، خروجی درخت t برای ویژگی‌های ورودی x_i است. t شماره گام یا درخت در فرایند بوستینگ می‌باشد. در هر گام، مدل تقویت گرادیان حداکثری مقدار بهینه $f_t(x_i)$ را طوری پیدا می‌کند که کمترین خطای باقی مانده را داشته باشد. درنتیجه درخت‌های جدید نقص‌های درخت‌های قبلی را اصلاح می‌کنند و باعث بهبود مدل می‌شوند.

گام سوم- استفاده از مشتق دوم برای گرادیان بوستینگ سریع‌تر: یکی از تفاوت‌های کلیدی تقویت گرادیان حداکثری با سایر روش‌های تقویت گرادیان سنتی استفاده از مشتق دوم تابع هزینه برای بهینه‌سازی سریع‌تر است. این تکنیک که با نام بوستینگ مرتبه دوم شناخته می‌شود، فرآیند یادگیری را تسریع می‌کند. در روش تقویت گرادیان سنتی، تنها از گرادیان (مشتق اول) برای اصلاح درخت‌ها استفاده می‌شود اما تقویت گرادیان حداکثری از اطلاعات هشین (مشتق دوم) نیز بهره می‌برد که به دقت بیش‌تر و همگرایی سریع‌تر مدل کمک می‌کند. فرمول به‌روزرسانی درخت در مرحله tام به‌صورت زیر است:

$$g_i = \frac{\partial \hat{y}_i}{\partial l(y_i,\hat{y}_i)} \quad (۷)$$

$$h_i = \frac{\partial^2 \hat{y}_i}{\partial^2 l(y_i,\hat{y}_i)} \quad (۸)$$

$$f_t^* = \frac{\sum_i g_i}{\sum_i h_i + \lambda} \quad (۹)$$

g_i مشتق اول (گرادیان) در رابطه (۷) تابع هزینه نسبت به خروجی مدل که نشان می‌دهد مدل چقدر از مقدار واقعی فاصله دارد. h_i مشتق دوم (هشین) در رابطه (۸) که نرخ تغییر گرادیان است که نشان می‌دهد تغییرات تابع هزینه در چه جهتی باید انجام شود. f_t^* از رابطه (۹) برای تعیین مقدار بهینه برگ درخت استفاده می‌شود که باعث می‌شود مدل یاد بگیرد که چگونه پیش‌بینی‌های خود را بهینه کند. درواقع در گام tام، هدف پیدا کردن تابع درختی است که مقدار تابع هزینه را کمینه کند (f_t^*). تقریب مرتبه دوم تابع هزینه را می‌توان به شکل رابطه (۱۰) نوشت:

$$L(t) = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) \quad (۱۰)$$

برای یافتن مقدار بهینه $f_t(x_i)$ از بسط تیلور مرتبه دوم در رابطه (۱۱) استفاده می‌کنیم:

$$l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) \approx l(y_i, \hat{y}_i^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \quad (۱۱)$$

پیش‌تر توضیح داده شد که g_i مشتق اول (گرادیان) تابع هزینه نسبت به خروجی مدل که نشان می‌دهد مدل چقدر از مقدار واقعی فاصله دارد و h_i مشتق دوم (هشین) که نشان می‌دهد تغییرات تابع هزینه در چه جهتی باید انجام شود. حال، برای تعیین تابع درختی، آن را به‌صورت یک ترکیب خطی به شکل رابطه (۱۲) از وزن‌های برگ‌ها در نظر می‌گیریم (انتساب نمونه‌ها به برگ‌های درخت):

$$f_t(x) = w_q, \quad where\ q = q(x) \quad (۱۲)$$

سپس مقدار بهینه وزن برگ‌ها را با حل مسأله مینیمم‌سازی به‌صورت رابطه (۱۳) به‌دست می‌آوریم:

$$w_j^* = -\frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda} \quad (۱۳)$$

I_j مجموعه نمونه‌هایی است که در برگ jام قرار دارند. w_j^* مقدار بهینه وزن هر برگ است. λ یک پارامتر تنظیمی برای جلوگیری از نوسانات بیش‌ازحد است. درنهایت، مقدار تابع هدف اصلاح‌شده (با در نظر گرفتن هشین) به‌صورت رابطه (۱۴) نوشته می‌شود:

$$(t) \approx \sum_{j=1}^{T} \left(-\frac{1}{2} \frac{\left(\sum_{i \in I_j} g_i\right)^2}{\sum_{i \in I_j} h_i + \lambda}\right) + \Gamma t \quad (۱۴)$$

این رابطه نشان می‌دهد که افزودن مشتق دوم به فرآیند گرادیان بوستینگ، نه‌تنها باعث بهبود پایداری و کاهش نوسانات مدل می‌شود بلکه سرعت همگرایی را نیز افزایش می‌دهد. علاوه‌براین، استفاده از مشتق دوم کمک می‌کند تا به‌روزرسانی وزن‌ها دقیق‌تر باشد و مسیر یادگیری مدل بهینه‌تر تنظیم شود.

گام چهارم: انتخاب ویژگی و نمونه‌گیری تصادفی: تقویت گرادیان حداکثری از دو تکنیک برای جلوگیری از وابستگی بیش‌ازحد مدل به ویژگی‌ها و کاهش واریانس استفاده می‌کند. درنتیجه این دو روش باعث افزایش پایداری مدل و کاهش حساسیت به ناهنجاری می‌شوند:

- نمونه‌گیری تصادفی از ویژگی‌ها: در هر درخت، فقط از یک زیرمجموعه از ویژگی‌ها برای تصمیم‌گیری استفاده می‌شود.
- نمونه‌گیری تصادفی از داده‌ها: برای ساخت هر درخت، فقط یک زیرمجموعه از نمونه‌های آموزشی انتخاب می‌شود.

۳-۱-٤. الگوهای سنتی اقتصادسنجی

در مدل‌سازی احتمال وقوع یک رویداد در اقتصاد و مالی، سه الگوی سنتی الگوی احتمال خطی[1]، الگوی لاجیت[2] و الگوی پروبیت[3] از روش‌های اصلی مورد استفاده هستند. این مدل‌ها در پاسخ به محدودیت‌های مدل‌های رگرسیونی کلاسیک برای متغیرهای وابسته دسته‌ای توسعه یافته‌اند.

یک- الگوی احتمال خطی

یکی از اولین تلاش‌ها برای مدل‌سازی متغیرهای وابسته دوحالتی (مانند نکول یا عدم نکول یک وام) در اقتصادسنجی است و به‌عنوان یک تعمیم ساده از رگرسیون خطی در نظر گرفته می‌شود. این مدل از رابطه خطی (۱۵) پیروی می‌کند:

$$P(Y = 1 \mid X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon \qquad (۱۵)$$

1. Linear Probability Model- LPM
2. Logit Model
3. Probit Model

که در آن $P(Y = 1 \mid X)$ احتمال وقوع رویداد موردنظر را مدل‌سازی می‌کند. این مدل توسط آلدریچ و نلسون[1] (۱۹۸۴) مورد بررسی قرار گرفت و یکی از ساده‌ترین روش‌ها برای پیش‌بینی احتمالات است.

دو- الگوی لجستیک

برای مدل‌سازی متغیرهای وابسته دوتایی، الگوی احتمال خطی با چالش تولید احتمالات خارج از بازه [۰،۱] مواجه است. برای غلبه بر این محدودیت، از روش‌های جایگزین استفاده شده است. این رویکردها بر استفاده از یک تابع پیوند[2] غیرخطی برای نگاشت ترکیب خطی متغیرهای توضیحی به یک احتمال محدود تمرکز دارند. لذا این مدل فرض می‌کند که رابطه بین متغیرهای مستقل و احتمال وقوع یک رویداد، از طریق تابع لجستیک غیرخطی است که مشکل احتمالات خارج از بازه [۰،۱] را حل می‌کند. برخلاف مدل احتمال خطی، مدل لجستیک احتمال وقوع یک رویداد را به یک تابع غیرخطی از متغیرهای توضیحی ارتباط می‌دهد تا از مشکلاتی مانند پیش‌بینی مقادیر خارج از بازه [۰،۱] جلوگیری کند (Aldrich & Nelson, 1984).

سه- الگوی پروبیت

مدل پروبیت مشابه مدل لجستیک است اما از تابع توزیع نرمال استاندارد برای تبدیل مقادیر ورودی به احتمال استفاده می‌کند. این مدل توسط فیشمن[3] (۱۹۷۳) معرفی شد و به شکل رابطه (۲۰) تعریف می‌شود:

$$P(Y = 1 \mid X) = \Phi(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k) \qquad (۲۰)$$

که در آن $\Phi(.)$ تابع توزیع تجمعی نرمال استاندارد است. این تابع تضمین می‌کند که احتمال پیش‌بینی‌شده همواره بین ۰ و ۱ قرار گیرد.

با وجود سادگی و قابلیت تفسیر بالای مدل‌های سنتی اقتصادسنجی مانند رگرسیون لجستیک و پروبیت، این روش‌ها با محدودیت‌هایی نیز همراه هستند.

نخست آنکه این مدل‌ها عمدتاً بر پایه فرض روابط خطی (یا در مورد لاجیت و پروبیت، توزیع خاصی برای تابع خطا) عمل می‌کنند که ممکن است در داده‌های واقعی بانکی که

1. Aldrich, J.H. & Nelson, F.D.
2. Link Function
3. Fishman, G.S.

معمولاً دارای روابط پیچیده و غیرخطی هستند، صدق نکند. دوم، این مدل‌ها در برخورد با داده‌های دارای تعداد زیاد متغیر یا داده‌های دارای برهم‌کنش‌های پیچیده، عملکرد مطلوبی ندارند. همچنین، در صورتی که بین متغیرهای توضیحی هم‌خطی وجود داشته باشد، برآورد ضرایب ناپایدار می‌شود. این مسئله می‌تواند به کاهش دقت پیش‌بینی در محیط‌های واقعی منجر شود.

جدول ۲. مقایسه مدل‌های سنتی اقتصادسنجی

ویژگی/مدل	الگولی احتمال خطی	الگوی لجستیک	الگوی پروبیت
رابطه بین متغیرها و احتمال وقوع رویداد	خطی	غیرخطی (تابع لجستیک)	غیرخطی (تابع نرمال)
محدودیت در خروجی احتمالات	ممکن است خارج از بازه [۰,۱] باشد	همواره در بازه [۰,۱] باشد	همواره در بازه [۰,۱] باشد
توزیع خطاها	نرمال یا نامشخص	لجستیک	نرمال
تفسیر ضرایب	مستقیم	نسبت شانس	ضریب استاندارد شده
حساسیت به داده‌های دورافتاده	بالا	متوسط	پایین
کاربرد	مدل‌های ساده و مقدماتی	پیش‌بینی ورشکستی، ریسک اعتباری	تحلیل انتخاب‌های اقتصادی

مأخذ: یافته‌های پژوهش

با توجه به جدول ۳، مقایسه‌ای میان سه مدل رگرسیون لجستیک، جنگل تصادفی و تقویت گرادیان حداکثری ارائه شده که می‌توان نتیجه گرفت که مدل‌های یادگیری ماشین، به‌ویژه الگوریتم‌های درختی مانند تقویت گرادیان حداکثری، از دقت پیش‌بینی بسیار بالاتری نسبت به روش‌های سنتی مانند رگرسیون لجستیک برخوردار هستند. شواهد پژوهشی نیز مؤید این برتری است.

بااین‌حال، این مدل‌های پیشرفته معمولاً نیازمند تنظیم دقیق ابرپارامترها و منابع محاسباتی قابل توجهی هستند (Bermudez, et al. 2022). در مقابل، رگرسیون لجستیک به دلیل ساختار ساده‌تر و سرعت اجرای بالا، همچنان در بسیاری از کاربردهای بانکی محبوب است اما در تحلیل روابط غیرخطی و اثرات متقابل بین متغیرها محدودیت دارد (Hand &

Henley, 1997). همچنین، در شرایطی که داده‌ها دارای ساختار پیچیده یا حجم بالا هستند، قابلیت تعمیم‌پذیری مدل‌های سنتی کاهش می‌یابد. این در حالی است که مدل‌های پیشرفته‌تر مانند جنگل تصادفی و تقویت گرادیان حداکثری با مکانیزم‌های درونی کنترل بیش‌برازش، تعمیم‌پذیری بهتری ارائه می‌دهند (Chen & Guestrin, 2016).

بر همین اساس، در این پژوهش علاوه‌بر مقایسه دقت مدل‌ها، از رویکردهایی مانند تنظیم بهینه ابرپارامترها برای بهبود قابلیت تعمیم‌پذیری نتایج استفاده شده است تا از بروز سوگیری مدل و بیش‌برازش جلوگیری شود.

جدول ۳. بررسی تطبیقی دقت، پیچیدگی و تعمیم‌پذیری در مدل‌های پیش‌بینی نکول

ویژگی/ مدل	رگرسیون لجستیک	جنگل تصادفی	تقویت گرادیان حداکثری
دقت پیش‌بینی	متوسط (در حد مدل پایه)	بسیار بالا	بسیار بالا (اندکی بالاتر از جنگل تصادفی)
پیچیدگی محاسباتی	پایین (تحلیل سریع)	متوسط (با افزایش تعداد درخت‌ها افزایش می‌یابد)	بالا (نیازمند تنظیم دقیق و منابع محاسباتی بیشتر)
قابلیت تعمیم‌پذیری	نسبتاً پایین (در برابر داده‌های پیچیده ضعف دارد)	بالا (مقاوم در برابر بیش‌برازش)	بسیار بالا (با تنظیم درست پارامترها تعمیم‌پذیری بالایی دارد)

مأخذ: یافته‌های پژوهش

۲-٤. پایه‌های آماری و آماره‌های توصیفی

این پژوهش با هدف برآورد احتمال نکول تسهیلات بانکی، داده‌های مربوط به **۵۶,۹۶۵** فقره تسهیلات اعطایی شعب شمال تهران بانک ملی ایران طی سال‌های ۱۳۹۸ تا ۱۴۰۳ را مورد تحلیل قرار داده است. در راستای پیش‌بینی رفتار اعتباری مشتریان، از سه رویکرد تحلیلی شامل رگرسیون لجستیک، جنگل تصادفی و تقویت گرادیان حداکثری بهره گرفته شد. داده‌ها پس از مرحله پیش‌پردازش دقیق، شامل حذف متغیرهای نامرتبط، تبدیل متغیرهای طبقه‌ای به عددی، جایگزینی داده‌های گمشده و استخراج ویژگی‌های زمانی از تاریخ‌های شمسی، آماده مدل‌سازی گردیدند.

به‌منظور بهینه‌سازی عملکرد مدل‌ها و جلوگیری از بیش‌برازش، داده‌ها با نسبت ۸۰٪ برای آموزش و ۲۰٪ برای آزمون تفکیک شدند. همچنین برای حفظ تعادل میان کلاس‌های نکول

و غیرنکول، از روش نمونه‌گیری طبقه‌بندی‌شده[1] استفاده شد. متغیرهای پژوهش در سه گروه اصلی شامل ویژگی‌های قرارداد تسهیلات، ویژگی‌های فردی مشتریان و ویژگی‌های شعبه طبقه‌بندی شدند.

جدول ٤. متغیرها و ویژگی‌های مورد بررسی

متغیر	تعریف	نوع داده	گروه متغیر
TAS_NO	شماره تسهیلات	عدد اعشاری	مشخصات قرارداد تسهیلات
Term	تعداد اقساط	عدد صحیح	
Grace_Period(Days)	دوره تنفس تسهیلات به‌روز	عدد صحیح	
ORIGINAL_AMOUNT	اصل تسهیلات اعطا شده	عدد اعشاری	
PROFIT_AMOUNT	سود تسهیلات اعطایی	عدد صحیح	
CONTRACT_TYPE	نوع عقد تسهیلات	عدد صحیح	
SARFASL	کد دفاتر حسابداری که مقادیر در آنها ثبت شده است	عدد صحیح	
TAS_PROFIT_REMAIN	مانده از سود تسهیلات	عدد صحیح	
Non_Current_Balance	مانده غیرجاری تسهیلات	عدد اعشاری	
Current_Balance	مانده جاری تسهیلات	عدد اعشاری	
Doubtful_Balance	مانده غیرجاری تسهیلات. طبقه مشکوک‌الوصول	عدد اعشاری	
Over_due_Balance	مانده غیرجاری تسهیلات. طبقه سررسید گذشته	عدد اعشاری	
Outstanding_balance	مانده غیرجاری تسهیلات. طبقه معوق	عدد اعشاری	

1. Stratified Sampling

ادامه جدول ٤. متغیرها و ویژگی‌های مورد بررسی

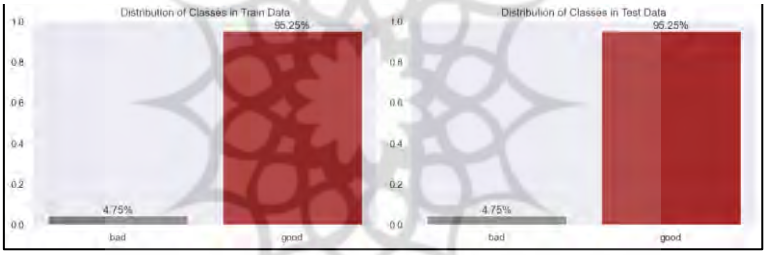
متغیر	تعریف	نوع داده	گروه متغیر
PENALTY_REMAIN	مبلغ وجه التزام تسهیلات	عدد صحیح	
PROFIT_RATE	نرخ سود تسهیلات اعطا شده	عدد صحیح	
ORIGINAL_Balance	مانده از اصل تسهیلات	عدد اعشاری	
Collateral	نوع ضمانت	عدد صحیح	
Collateral_Value	ارزش ریالی ضمانت افراد	عدد صحیح	
Collateral_Mortgage_Value	ارزش ریالی در رهن ضمانت‌ها	عدد صحیح	
Collateral_Status	وضعیت ضمانت‌ها	عدد صحیح	
Days_since_issue_date	روزهای گذشته شده از تاریخ اعطا	عدد صحیح	
OWNERSHIP	نوع شخص وام‌گیرنده از نظر سازمان‌های تحت پوشش	عدد صحیح	مشخصات دریافت‌کننده تسهیلات
TAS_USAGE	نوع مورد استفاده تسهیلات	عدد صحیح	
TAS_GOAL	هدف از تسهیلات	عدد صحیح	
PRIVATE_SECTOR	نوع بخش‌های دریافت‌کننده تسهیلات	عدد صحیح	
ECONOMIC_SECTOR	بخش‌های اقتصادی دریافت‌کننده تسهیلات	عدد صحیح	
GENDER_CODE	جنسیت	عدد صحیح	
Age	سن مشتری	عدد صحیح	
Marriage	وضعیت تاهل	عدد صحیح	

ادامه جدول ٤. متغیرها و ویژگی‌های مورد بررسی

متغیر	تعریف	نوع داده	گروه متغیر
RESIDENCY	وضعیت اقامت	عدد صحیح	
EDUCATION	تحصیلات	عدد صحیح	
NATIVE_Status	ملیت	عدد صحیح	
Branch_no	کد شعبه	عدد صحیح	مشخصات شعبه اعطاکننده
UNITGRADE	درجه شعبه	عدد صحیح	

مأخذ: یافته‌های پژوهش

شکل ۱. توزیع گروه‌های نکول و عدم نکول در هر مجموعه آموزش و آزمایش



مأخذ: یافته‌های پژوهش

٤-۳. شاخص‌های ارزیابی الگوها

برای ارزیابی عملکرد مدل‌های یادگیری ماشین، از مجموعه‌ای از شاخص‌های استاندارد استفاده می‌شود که هر یک جنبه‌ای خاص از توانایی مدل در طبقه‌بندی نمونه‌ها را منعکس می‌کند.

- دقت[1]: یکی از ساده‌ترین و رایج‌ترین شاخص‌های ارزیابی است که نسبت تعداد پیش‌بینی‌های درست به کل پیش‌بینی‌ها را اندازه‌گیری می‌کند (Powers, 2011).

1. Accuracy

بااین‌حال، در مسائل نامتوازن که یک طبقه دارای فراوانی بسیار بیشتری نسبت به دیگری است، دقت می‌تواند گمراه‌کننده باشد.

- بازخوانی[1] یا حساسیت: برای هر کلاس (مثلاً کلاس «خوب» یا «بد»)، توانایی مدل در شناسایی صحیح نمونه‌های آن کلاس را می‌سنجد. این معیار برای نخستین بار در مطالعات مربوط به بازیابی اطلاعات معرفی شد (van Rijsbergen, 1979) و در یادگیری ماشین نیز به‌عنوان شاخصی برای کاهش خطای نوع دوم کاربرد دارد. بازخوانی برای کلاس «خوب»، نسبت نمونه‌های درست شناسایی‌شده به کل نمونه‌های واقعاً «خوب» است و برای کلاس «بد» نیز به همین ترتیب تعریف می‌شود.
- صحت[2]: این شاخص نیز از حوزه بازیابی اطلاعات وارد شده است (van Rijsbergen, 1979) و نشان می‌دهد که از میان نمونه‌هایی که مدل به‌عنوان «مثبت» یا «خوب» (یا به‌طور مشابه «بد») طبقه‌بندی کرده، چه نسبتی واقعاً متعلق به آن کلاس بوده‌اند. ترکیب این دو معیار، دقت و بازخوانی، در قالب شاخصی به نام «امتیاز اف‌وان[3]» صورت می‌گیرد که میانگین هارمونیک این دو است و در شرایطی که نیاز به تعادل بین شاخص‌های صحت و بازخوانی وجود دارد، بسیار مفید است (Chinchor, 1992). ازآنجاکه این پژوهش به تفکیک عملکرد، مدل را برای دو کلاس «خوب» و «بد» بررسی می‌کند، مقادیر صحت، بازخوانی و امتیاز اف‌وان برای هر یک از این کلاس‌ها به‌صورت جداگانه محاسبه شده‌اند تا نقاط قوت و ضعف مدل در تمایز میان این دو گروه بهتر مشخص گردد.
- ماتریس درهم‌ریختگی[4]: ابزاری پایه‌ای در تحلیل عملکرد مدل‌های طبقه‌بندی است که برای نخستین بار در حوزه روان‌سنجی توسط کلی[5] (۱۹۵۲) معرفی شد.

1. Recall
2. Precision
3. F1-Score
4. Confusion Matrix
5. Kelly, G.A.

این ماتریس، تعداد نمونه‌های صحیح و ناصحیح طبقه‌بندی‌شده را در قالب چهار خانه مثبت واقعی[1]، مثبت کاذب[2]، منفی کاذب[3] و منفی واقعی[4] نمایش می‌دهد.

- مساحت زیرمنحنی مشخصه عملکرد سیستم[5]: معیاری برای سنجش کیفیت عملکرد مدل در سطوح مختلف آستانه تصمیم‌گیری است. این منحنی برای نخستین‌بار در تحلیل‌های سیستم‌های تشخیص سیگنال در جنگ جهانی دوم معرفی شد (Green & Swets, 1996) و بعدها در یادگیری ماشین نیز به‌کار گرفته شد. مقدار AUC، سطح زیر منحنی ROC را نشان می‌دهد و بین ۰ تا ۱ متغیر است. هرچه مقدار AUCبیشتر باشد، مدل توانایی بهتری در تمایز میان کلاس‌ها دارد.

٤-٤. آزمون معناداری عملکرد مدل‌ها با استفاده از روش بوت‌استرپ[6]

برای ارزیابی آماری تفاوت در عملکرد مدل‌های یادگیری ماشین، از روش بوت‌استرپ برای محاسبه فاصله اطمینان ۹۵ درصد شاخص AUC و همچنین مقایسه معناداری تفاوت AUC بین مدل‌ها استفاده شد. این روش بر پایه نمونه‌گیری تصادفی با جایگزینی[7] از داده‌های آزمون بوده و به‌طور گسترده‌ای برای برآورد ناپارامتریک فاصله اطمینان برای شاخص‌های ارزیابی عملکرد به‌ویژه AUC توصیه شده است (Efron & Tibshirani, 1993).

در این مطالعه، ۱۰۰۰ تکرار بوت‌استرپ برای هر مدل انجام شد. در هر تکرار، یک نمونه تصادفی از داده‌ها (با جایگزینی) انتخاب شد و مقدار AUC برای آن محاسبه شد. سپس، توزیع حاصل از این AUCها مرتب شد و صدک‌های متناظر با سطح اطمینان ۹۵ درصد برای تعیین بازه اطمینان استخراج گردید. به‌منظور آزمون معناداری اختلاف عملکرد بین دو مدل، تفاوت AUCها در هر تکرار بوت‌استرپ محاسبه و توزیع اختلاف AUCها تحلیل شد. اگر بازه اطمینان اختلاف AUC شامل صفر نباشد، تفاوت بین مدل‌ها از نظر آماری معنادار در نظر گرفته می‌شود.

---

1. True Positive - TP
2. False Positive - FP
3. False Negative - FN
4. True Negative - TN
5. Area Under Curve – Receiver Operating Characteristic _ AUC- ROC
6. Bootstrap
7. Resampling with Replacement

روش بوت‌استرپ نسبت به روش‌های کلاسیک فرضی مانند آزمون z برای مقایسه AUC، مزیت‌هایی دارد، ازجمله عدم نیاز به مفروضات نرمال بودن توزیع یا همسانی واریانس‌ها (Berrar, 2019). همچنین، این روش برای مقایسه مدل‌ها بر روی داده‌های واقعی به‌ویژه در شرایطی که حجم نمونه محدود است، مناسب‌تر تلقی می‌شود.

۵. برآورد الگوها

در این بخش، با تکیه بر مبانی نظری و پیشینه پژوهش ارائه‌شده در بخش‌های پیشین، مدل‌های انتخابی برای تحقیق را به‌طور دقیق تصریح خواهیم کرد. در گام نخست، فرآیند پیش‌پردازش داده‌ها به‌عنوان مرحله اولیه آماده‌سازی مورد بررسی قرار می‌گیرد. سپس، مدل‌های یادگیری ماشین در دو سنجه مختلف، شامل بدون تنظیم پارامترهای فراگیر و با تنظیم پارامترهای فراگیر و مدل اقتصادسنجی رگرسیون لجستیک برای آموزش مدل به‌منظور برآورد احتمال نکول مورد استفاده قرار خواهند گرفت. پس از آموزش مدل‌ها، عملکرد آن‌ها از طریق مجموعه‌ای از معیارهای ارزیابی ذکر شده سنجیده خواهد شد.

۵-۱. تنظیم هایپر پارامترها

مدل‌های یادگیری ماشین دارای مجموعه‌ای از پارامترهای فراگیر هستند که برخلاف پارامترهای مدل (مانند ضرایب در رگرسیون خطی)، مستقیماً از داده‌ها یاد گرفته نمی‌شوند بلکه باید پیش از آموزش مدل مقداردهی شوند (Bergstra & Bengio, 2012). مقدار این پارامترهای فراگیر می‌تواند تأثیر چشمگیری بر عملکرد مدل، دقت پیش‌بینی و تعمیم‌پذیری آن داشته باشد. عدم تنظیم مناسب این مقادیر می‌تواند منجر به کم‌برازش یا بیش‌برازش شود.

یکی از روش‌های معمول در آموزش مدل‌های یادگیری ماشین، استفاده از مقادیر پیش‌فرض برای پارامترهای فراگیر است. این رویکرد در برخی موارد منجر به نتایج قابل قبولی می‌شود اما بهینه‌ترین مدل را تضمین نمی‌کند (Probst, et al., 2018). به همین دلیل، تنظیم پارامترهای فراگیر به منظور یافتن مقدار بهینه‌ای که عملکرد مدل را بیشینه کند، انجام می‌شود. مطالعات مختلفی اهمیت این موضوع را نشان داده‌اند. به‌عنوان مثال، در مطالعه فیورر و هاتر[1] (۲۰۱۹) نشان داده شد که تنظیم بهینه پارامترهای فراگیر می‌تواند باعث بهبود دقت

1. Feurer, M. & Hutter, F.

مدل تا ۱۰ الی ۲۰ درصد نسبت به مقداردهی پیش‌فرض شود. همچنین، در یک پژوهش توسط لیاو و همکاران[1] (۲۰۲۲) مشخص شد که عدم تنظیم مناسب می‌تواند منجر به کاهش تعمیم‌پذیری مدل و درنتیجه عملکرد نامناسب روی داده‌های جدید شود. در مدل‌های یادگیری ماشین مورد استفاده در این پژوهش (جنگل تصادفی و تقویت رادیان حداکثری)، تفاوت بین این دو سنجه به‌صورت زیر قابل مشاهده است:

در حالت بدون تنظیم پارامترهای فراگیر، مقادیر پیش‌فرضی مانند تعداد درخت‌ها، عمق درخت و نرخ یادگیری استفاده شده و مدل بدون انجام جستجوی دقیق آموزش داده می‌شود. در حالت با تنظیم پارامترهای فراگیر، این مقادیر بهینه‌سازی شده‌اند تا عملکرد مدل براساس شاخص مساحت زیر منحنی مشخصه عملکرد به حداکثر برسد.

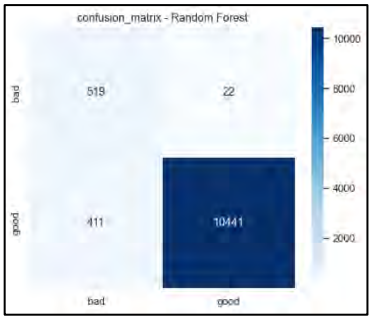
مدل جنگل تصادفی در حالت اولیه، بدون تنظیم هایپرپارامترها، دقتی برابر با ۹۶٪ کسب کرده است. مقدار بازخوانی ۰/۹۶ نشان‌دهنده توانایی بالا در شناسایی موارد نکول است اما صحت پایین‌تر (۰/۵۶) بیانگر اشتباه در طبقه‌بندی بخشی از وام‌های غیرنکولی می‌باشد. پس از تنظیم هایپرپارامترها، دقت مدل به ۹۹٪ افزایش یافته و تعداد منفی‌های کاذب از ۴۱۱ به ۴۸ کاهش یافته است که بهبود چشمگیری در شناسایی وام‌های سالم را نشان می‌دهد. با وجود کاهش اندک بازخوانی نکول به ۰/۸۴، مقدار صحت نکول به ۰/۹۰ ارتقا یافته است که نشان‌دهنده کاهش خطای شناسایی نکول است.

مدل تقویت گرادیان حداکثری در حالت بدون تنظیم، عملکردی مشابه جنگل تصادفی بهینه داشته ولی با صحت ۰/۹۱ و بازخوانی ۰/۸۵، توانایی بهتری در تفکیک وام‌های نکولی از سالم نشان داده است. پس از بهینه‌سازی هایپرپارامترها، این مدل بالاترین صحت و بازخوانی را به‌دست آورده و کمترین میزان منفی کاذب را ثبت کرده است.

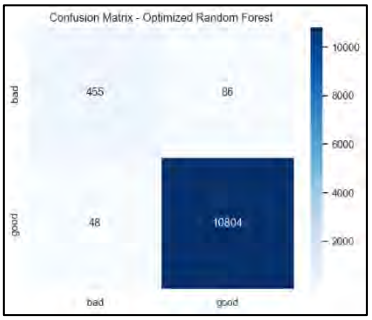
به‌طور کلی، هر دو مدل پس از بهینه‌سازی عملکرد بالایی نشان داده‌اند اما تقویت گرادیان حداکثری با دقت و پایداری بیشتر و خطای کمتر، بهترین گزینه برای شناسایی وام‌های نکولی ارزیابی می‌شود.

1. Liao, L., et al.

شکل ۲. الگوی جنگل تصادفی (بدون تنظیم هایپرپارامترها)

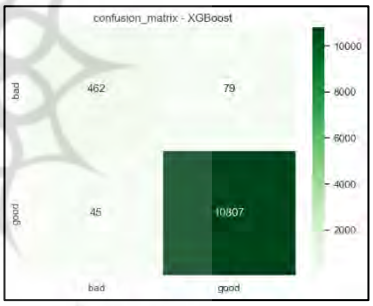


شکل ۳. الگوی جنگل تصادفی (با تنظیم هایپرپارامترها)

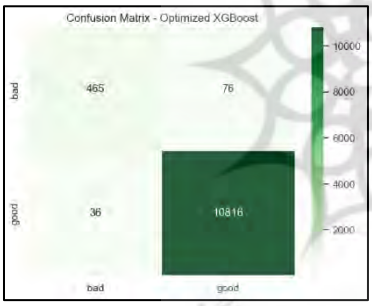


مأخذ: یافته‌های پژوهش

شکل ٤. الگوی تقویت گرادیان حداکثری (بدون تنظیم هایپرپارامترها)



شکل ۵. الگوی تقویت گرادیان حداکثری (با تنظیم هایپرپارامترها)



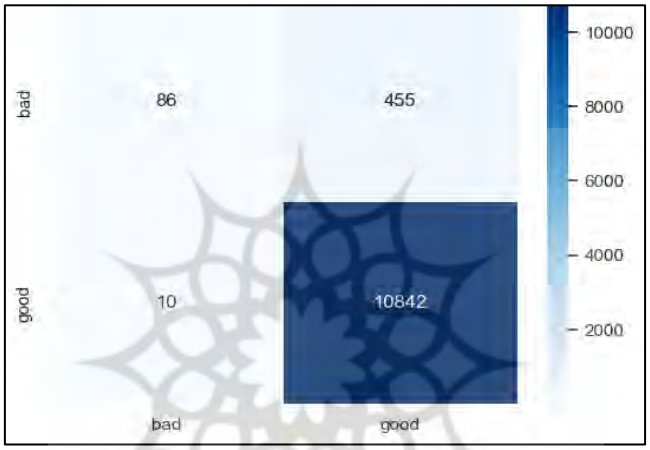
مأخذ: یافته‌های پژوهش

بهینه‌سازی پارامترهای فراگیر در هر دو مدل تأثیر مثبتی داشته و باعث کاهش خطاهای پیش‌بینی و افزایش دقت شده است. اگر هدف کاهش حداکثری اشتباهات در شناسایی وام‌های نکولی باشد، مدل تقویت گرادیان حداکثری با تنظیمات بهینه انتخاب بهتری خواهد بود. اما اگر سادگی اجرا و تفسیر مدل اهمیت بیش‌تری داشته باشد، مدل جنگل تصادفی بهینه‌شده گزینه مناسبی محسوب می‌شود.

الگوی رگرسیون لجستیک عملکرد بسیار خوبی برای پیش‌بینی کلاس خوب و یا بدون نکول دارد اما در پیش‌بینی کلاس بد و یا نکول، ضعیف عمل می‌کند. مقدار صحت برای

کلاس بد ۰/۹۰ است، اما مقدار بازخوانی بسیار پایین است (۰/۱۶). این یعنی مدل در شناسایی موارد نکول با مشکل مواجه است. مقدار ROC-AUC معادل ۰/۷۵۳۴ است که نشان‌دهنده عملکرد نسبتاً متوسطی برای تفکیک دو کلاس است. همچنین ماتریس درهم‌ریختگی نشان می‌دهد که مدل تنها ۸۶ مورد از ۵۴۱ مورد نکول را درست پیش‌بینی کرده و ۴۵۵ مورد نکول را به اشتباه به‌عنوان خوب طبقه‌بندی کرده است.

شکل ٦. الگوی رگرسیون لجستیک



مأخذ: یافته‌های پژوهش

جدول ۵. خلاصه ارزیابی نتایج الگوها

الگو	حالت	ACCURACY	Precision (Bad)	Precision (Good)	Recall (Bad)	Recall (Good)	F1-Score (Bad)	F1-Score (Good)	ROC-AUC
جنگل تصادفی	غیربهینه	۹۷٪	۰/۹٤	۰/۹۸	۰/۸۳	۰/۹۹	۰/۸۸	۰/۹۸	۰/۹۳٥
جنگل تصادفی	بهینه	۹۹٪	۰/۹۷	۰/۹۹	۰/۹٤	۰/۹۹	۰/۹٥	۰/۹۹	۰/۹۹٦۸
تقویت گرادیان حداکثری	غیربهینه	۹۸٪	۰/۹٦	۰/۹۹	۰/۸٥	۰/۹۹	۰/۹۰	۰/۹۹	۰/۹۹٦٦
تقویت گرادیان حداکثری	بهینه	۹۹٪	۰/۹۷	۰/۹۹	۰/۸۸	۰/۹۹	۰/۹۲	۰/۹۹	۰/۹۹۷۳
رگرسیون لجستیک	-	۹٦٪	۰/۹۰	۰/۹٦	۰/۱٦	۰/۹۸	۰/۲۷	۰/۹۸	۰/۷٥۳٤

مأخذ: یافته‌های پژوهش

۵-۲. مقایسه معناداری عملکرد مدل‌ها با استفاده از آزمون بوت‌استرپ

در این بخش، یافته‌های حاصل از تحلیل مقایسه‌ای عملکرد سه الگو شامل جنگل تصادفی، تقویت گرادیان حداکثری و رگرسیون لجستیک مبتنی بر معیار مساحت زیر منحنی مشخصه عملکرد گیرنده (AUC-ROC) و آزمون بوت‌استرپ برای برآورد فاصله اطمینان ۹۵ درصدی گزارش می‌شود. همچنین، تفاوت آماری میان عملکرد مدل‌ها از طریق آزمون اختلاف AUC به روش بوت‌استرپ مورد ارزیابی قرار گرفته است.

نتایج حاکی از آن است که مدل جنگل تصادفی با میانگین AUC برابر با ۰/۹۶۶ و مدل تقویت گرادیان حداکثری با میانگین AUC برابر با ۰/۹۶۵، عملکرد بسیار مطلوب و نزدیکی را در طبقه‌بندی داده‌ها ارائه داده‌اند. در مقابل، مدل رگرسیون لجستیک با میانگینAUC برابر با ۰/۷۲۷ به‌طور معناداری عملکرد ضعیف‌تری نسبت به دو مدل دیگر از خود نشان داده است.

مقایسه آماری بین مدل‌ها از طریق آزمون اختلاف AUC نیز نشان داد که تفاوت میان مدل جنگل تصادفی و تقویت گرادیان حداکثری بسیار ناچیز و از نظر آماری غیرمعنادار است زیرا فاصله اطمینان ۹۵ درصدی آن شامل صفر بود درحالی‌که اختلاف عملکرد بین مدل جنگل تصادفی و رگرسیون لجستیک و همچنین بین تقویت گرادیان حداکثری و رگرسیون لجستیک از نظر آماری معنادار ارزیابی شد.

جدول ٦. نتایج تحلیل بوت‌استرپ AUC و آزمون اختلاف معناداری عملکرد مدل‌ها

مقایسه مدل‌ها	مقدار AUC / اختلافAUC	فاصله اطمینان ۹۵٪	نتیجه آماری
جنگل تصادفی	۰/۹٦٦	۰/۹۷٤ – ۰/۹۵۸	عملکرد بالا
تقویت گرادیان حداکثری	۰/۹٦۵	۰/۹۷۲ – ۰/۹۵۷	عملکرد بالا
رگرسیون لجستیک	۰/۷۲۷	۰/۷٤۸ – ۰/۷۰۵	عملکرد ضعیف‌تر
جنگل تصادفی و تقویت گرادیان حداکثری	۰/۰۰۲	۰/۰۰۷ – (-۰/۰۰٤)	غیرمعنادار (CI شامل صفر)
جنگل تصادفی و رگرسیون لجستیک	۰/۲۳۹	۰/۲٦۱ – ۰/۲۲۰	معنادار (CI بدون صفر)
تقویت گرادیان حداکثری و رگرسیون لجستیک	۰/۲۳۸	۰/۲۵۸ – ۰/۲۱۸	معنادار (CI بدون صفر)

مأخذ: یافته‌های پژوهش

یافته‌های فوق دلالت بر آن دارد که بهره‌گیری از الگوریتم‌های پیشرفته مبتنی بر درخت تصمیم، نظیر جنگل تصادفی و تقویت گرادیان حداکثری، در مسئله طبقه‌بندی مورد مطالعه، به‌مراتب اثربخش‌تر از مدل‌های خطی سنتی مانند رگرسیون لجستیک است و می‌تواند موجب ارتقای دقت پیش‌بینی شود.

٦. بحث و نتیجه‌گیری

به‌طور خلاصه، پژوهش حاضر مدل‌های برآورد احتمال نکول را در دو دسته سنتی و پیشرفته بررسی کرده است. مدل‌های سنتی مانند رگرسیون لجستیک به دلیل سادگی و تفسیرپذیری بالا، همچنان در تحلیل ریسک اعتباری کاربرد دارند اما به دلیل فرض روابط خطی میان متغیرها، در شناسایی الگوهای پیچیده با محدودیت مواجه‌اند. در مقابل، مدل‌های یادگیری ماشین همچون جنگل تصادفی و تقویت گرادیان حداکثری با توانایی درک روابط غیرخطی، دقت بالاتری در پیش‌بینی رفتار نکول ارائه می‌دهند، هرچند نیازمند تنظیم دقیق هایپرپارامترها و توان محاسباتی بیشتری هستند.

نتایج تجربی نشان داد که مدل تقویت گرادیان حداکثری نسبت به رگرسیون لجستیک عملکرد برتری داشته و با دقت ۹۹٪ و امتیاز ROC-AUC معادل ۰/۹۹۷۳ بالاترین میزان صحت را در برآورد نکول سبد تسهیلات بانک ملی ایران به‌دست آورده است درحالی‌که رگرسیون لجستیک به دلیل فرض خطی بودن روابط، توانایی محدودی در شناسایی موارد نکول داشته است. این یافته‌ها فرضیه پژوهش را مبنی بر برتری مدل‌های یادگیری ماشین بر روش‌های سنتی اقتصادسنجی تأیید می‌کند.

مقایسه با مطالعات پیشین نیز نشان می‌دهد که درحالی‌که پژوهش‌های قبلی عمدتاً بر مدل‌های آماری تکیه داشتند، استفاده از الگوریتم‌های یادگیری ماشین موجب افزایش چشمگیر دقت و کاهش خطای پیش‌بینی شده است. بنابراین، ترکیب روش‌های سنتی و مدرن می‌تواند ضمن حفظ قابلیت تفسیر، کارایی مدل‌سازی ریسک اعتباری را به‌طور معناداری ارتقا دهد.

تعارض منافع

تعارض منافع وجود ندارد.

ORCID

Reza Taleblou https://orcid.org/0000-0002-8679-2920
Mir Ali Kamali http://orcid.org/0009-0009-6991-0145
Parisa Mohajeri https://orcid.org/0000-0001-7971-0678

منابع

توکلی، سعید و آشتاب، الهام. (۱۴۰۲). مقایسه کارایی مدل‌های یادگیری ماشین و مدل‌های آماری در پیش‌بینی ریسک مالی. *فصلنامه راهبرد مدیریت مالی*، *۱۱*(۱)، ۷۶-۵۳. https://doi.org/10.22051/jfm.2023.35240.2512

رحمانی، علی و اسماعیلی، غریبه. (۱۳۸۹). کارایی شبکه‌های عصبی، رگرسیون لجستیک و تحلیل تمایزی در پیش‌بینی نکول. *اقتصاد مقداری (بررسی‌های اقتصادی)*، *۷*(۴)، ۱۵۱-۱۷۲. https://doi.org/10.22055/jqe.2010.10640

موحدی نیا، اکبر و بهمئی، نوشین. (۱۳۹۴). تعیین نکول تسهیلات مشتریان حقوقی به‌وسیله حداقل مربعات ماشین بردار پشتیبان بهبودیافته بر مبنای الگوریتم بهینه‌سازی تجمعی ذرات. کنفرانس بین‌المللی پژوهش‌های نوین در مدیریت، اقتصاد و حسابداری. http://irdoi.ir/103-440-857-466

References

Akerlof, G.A. (1970). The market for “lemons”: quality uncertainty and the market mechanism. *The Quarterly Journal of Economics*, *84*(3), 488–500. https://doi.org/10.2307/1879431

Aldrich, J.H. & Nelson, F.D. (1984). Linear probability, logit, and probit models (Quantitative Applications in the Social Sciences No. 07-045). SAGE Publications. https://doi.org/10.4135/9781412984744

Akinjole, A., Shobayo, O., Popoola, J., Okoyeigbo, O. & Ogunleye, B. (2024). Ensemble-based machine learning algorithm for loan default risk prediction. *Mathematics*, *12*(21), 3423. https://doi.org/10.3390/math12213423

Arrow, K.J. (1963). Uncertainty and the welfare economics of medical care. *The American Economic Review*, *53*(5), 941–973. https://doi.org/10.1016/B978-0-12-214850-7.50028-0

Bergstra, J. & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13, 281-305. https://doi.org/ 10.5555/2503308.2188395

Bermudez, J.D., Gonzalez-Rivera, G. & Gonzalez, M. (2022). Machine learning approaches to credit risk modeling: A comparative analysis. *Journal of Risk and Financial Management*, *15*(4), 123. https://doi.org/10.3390/jrfm15040123

Berrar, D. (2019). Cross-validation. In Encyclopedia of bioinformatics and computational biology (pp. 542–545). Elsevier. https://doi.org/10.1016/B978-0-12-809633-8.20349-X

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123–140. https://doi.org/10.1007/BF00058655

Breiman, L. (2011). Random Forests. *Machine Learning*, 45, 5–32. https://doi.org/10.1023/A:1010933404324

Chen, T. & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). https://doi.org/10.1145/2939672.2939785

Chinchor, N. (1992). MUC-4 evaluation metrics. In Proceedings of the 4th Conference on Message Understanding (pp. 22–29). https://doi.org/10.3115/1072064.1072067

Efron, B. & Tibshirani, R.J. (1993). An introduction to the bootstrap. Chapman & Hall/CRC. https://doi.org/10.1007/978-1-4899-4541-9

Feurer, M. & Hutter, F. (2019). Hyperparameter optimization. In: Hutter, F., Kotthoff, L., Vanschoren, J. (eds) Automated Machine Learning. The Springer Series on Challenges in Machine Learning. Springer, Cham. https://doi.org/10.1007/978-3-030-05318-5_1

Fishman, G.S. (1973). Statistical analysis for queueing simulations. *Management Science*, *20*(3), 363–369. https://doi.org/10.1287/mnsc.20.3.363

Friedman, J.H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics, 29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

Green, D.M. & Swets, J.A. (1966). Signal detection theory and psychophysics. Wiley. https://doi.org/10.1086/405615

G'ulomova, B.M.M. qizi. (2023). Bank loan allocation model based on credit risk prediction of SMEs. https://doi.org/10.1109/ictc57116.2023.10154753

Guo, C. (2016). Using machine learning techniques for credit risk modeling: Empirical evidence from China. *Journal of Financial Risk Management*, *5*(3), 1–12. https://doi.org/10.4236/jfrm.2016.53005

Hand, D.J. & Henley, W.E. (1997). Statistical classification methods in consumer credit scoring: A review. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *160*(3), 523–541. https://doi.org/10.1111/j.1467-985X.1997.00078.x

Ho, T.K. (1995). Random decision forests. In Proceedings of the 3rd international conference on Document analysis and recognition (Vol. 1, pp. 278-282). IEEE https://doi.org/10.1109/ICDAR.1995.598994

Kelly, G.A. (1952). The psychology of personal constructs. Norton. https://doi.org/10.4324/9780203359037

King, M., Zhu, Q. & Wang, T. (2021). Combining behavioral and financial data to improve credit scoring models: Evidence from a commercial

bank. *Journal of Banking and Finance*, 127, 106125. https://doi.org/10.1016/j.jbankfin.2021.106125

Liao, L., Li, H., Shang, W. & Ma, L. (2022). An Empirical Study of the Impact of Hyperparameter Tuning and Model Optimization on the Performance Properties of Deep Neural Networks. *ACM Transactions on Software Engineering and Methodology*, *31*(3), 1–40. https://doi.org/10.1145/3506695

Liu, H. (2020). Credit risk assessment with ensemble learning: A study of small and medium enterprises. *International Review of Financial Analysis*, 71, 101519. https://doi.org/10.1016/j.irfa.2020.101519

Movahedinia, A. & Bahmai, N. (2015). Determining the default of legal entity customers' facilities using improved support vector machine least squares based on particle swarm optimization algorithm. International Conference on New Researches in Management, Economics, and Accounting. http://irdoi.ir/103-440-857-466 [In Persian].

Nuez Mora, J.A., Moncayo, P. & Franco, C. (2023). Loan default prediction: A complete revision of LendingClub. *Estudios Gerenciales*, *39*(169), 1–17 https://doi.org/10.21919/remef.v18i3.886

Peykani, P., Sargolzaei, M., Sanadgol, N., Takalu, A. & Kamyabfar, H. (2023). Application of structural models (Merton and Geske) and machine learning models (random forest and gradient boosted trees) in predicting default risk of listed companies in the Iranian capital market. *PLoS ONE*, *18*(11), e0292081. https://doi.org/10.1371/journal.pone.0292081

Powers, D.M.W. (2011). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness & correlation. *Journal of Machine Learning Technologies*, *2*(1), 37–63 https://doi.org/10.9735/2229-3981

Probst, P., Boulesteix, A. & Bischl, B. (2018). Tunability: Importance of Hyperparameters of Machine Learning Algorithms. *Machine Learning Research*, 20, 53:1-53:32. https://doi.org/10.48550/arXiv.1802.09596

Rahmani, A. & Esmaeili, G. (2010). The efficiency of neural networks, logistic regression, and discriminant analysis in predicting default. *Quantitative Economics (Economic Studies)*, *7*(4), 151-172. https://doi.org/10.22055/jqe.2010.10640 [In Persian].

Robinson, N. & Sindhwani, N. (2024). Loan default prediction using machine learning. In 2024 11th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO) (pp. 1–5). IEEE. https://doi.org/10.55041/IJSREM24519

Spence, M. (1973). Job market signaling. *The Quarterly Journal of Economics*, *87*(3), 355–374. https://doi.org/10.2307/1882010

Stiglitz, J.E. & Weiss, A. (1981). Credit rationing in markets with imperfect information. *The American Economic Review*, *71*(3), 393–410. http://www.jstor.org/stable/1802787

Tang, Y., Liu, Y. & Huang, X. (2019). Detecting moral hazard in loan default using deep learning algorithms. *Expert Systems with Applications*, 130, 95–103. https://doi.org/10.1016/j.eswa.2019.04.003

Tavakoli, S. & Ashtab, E. (2023). Comparison of the efficiency of machine learning models and statistical models in predicting financial risk. *Quarterly Journal of Financial Management Strategy*, *11*(1), 53-76. https://doi.org/10.22051/jfm.2023.35240.2512 [In Persian].

Uphade, D.B., Muley, A.A. & Chalwadi, S.V. (2024). Identification of most preferable machine learning technique for prediction of bank loan defaulters. *Indian Journal of Science and Technology*, *17*(4), 343-351. https://doi.org/10.17485/IJST/v17i4.2978

Van Rijsbergen, C.J. (1979). *Information retrieval* (2nd ed.). https://doi.org/10.1002/asi.4630300621

استناد به این مقاله: طالبلو، رضا، کمالی، میرعلی و مهاجری، پریسا. (۱۴۰۴). برآورد احتمال نکول تسهیلات اعطایی در بانک ملی: مقایسه رویکردهای یادگیری ماشین و اقتصادسنجی. *پژوهش‌های اقتصادی ایران*، ۳۰(۱۰۳)، ۱-۴۱.