UDC 004.85:336.71:005.53 DOI: 10.31673/2412-4338.2022.027685

Fairushyn Roman
State University of Telecommunications, Kyiv

APPLICATION OF MACHINE LEARNING METHODS IN THE BANKING SECTOR TO INCREASE THE EFFICIENCY OF DECISION-MAKING

Abstract: This article examines the modern banking sector in the context of its transformation through the integration of machine learning (ML). This process has become crucial for identifying new avenues for the optimization and automation of financial transactions. In the initial research phase, a deep analysis of existing decision-making methodologies in banks was conducted, focusing on their role in current conditions and the potential to replace human involvement with algorithms. The evaluation and calculation of ML model efficiency revealed certain challenges, such as overfitting and a lack of transparency, but also highlighted significant advantages over traditional systems.

When selecting tools for the implementation and deployment of ML, a need for specialized tools tailored for the banking sector was identified, considering the specific nuances of this market. Defined efficiency-enhancing directions include developing new regularization methods, improving cross-validation techniques, and advancing explainable AI.

The study also emphasized the ethical concerns arising from ML implementation. Notably, primary attention was given to ensuring user data confidentiality and reducing algorithmic bias.

Based on the research results, a conclusion about the immense potential of machine learning in the banking sector was drawn. While certain challenges exist, the right approach can lead to the creation of more efficient, secure, and transparent systems, enhancing client trust.

Keywords: machine learning, banking sector, decision making, explainability, data confidentiality, bias, audit, ethical practices, loan approval.

Файрушин Роман
Державний університет телекомунікацій, Київ

ЗАСТОСУВАННЯ МЕТОДІВ МАШИННОГО НАВЧАННЯ У БАНКІВСЬКОМУ СЕКТОРІ ДЛЯ ПІДВИЩЕННЯ ЕФЕКТИВНОСТІ ПРИЙНЯТТЯ РІШЕНЬ

Анотація: В статті розглянутий сучасний банківський сектор у контексті його трансформації через інтеграцію машинного навчання (ML). Цей процес став ключовим для визначення нових шляхів оптимізації та автоматизації фінансових операцій. На першому етапі дослідження проведено глибокий аналіз існуючих методик прийняття рішень в банках, їхньої ролі в сучасних умовах та можливості заміни людської участі на алгоритми. Оцінка та розрахунок показників продуктивності ML моделей виявили деякі проблеми, такі як перенавчання та відсутність прозорості, але також вказали на значні переваги перед традиційними системами.

При виборі типів засобів для забезпечення і реалізації ML, було виявлено потребу в розробці спеціалізованих інструментів для банківської сфери, які враховують особливості цього ринку. Визначені напрямками підвищення ефективності включають розробку нових методів регуляризації, вдосконалення технік перехресної валідації та розвиток пояснювального ШІ.

Дослідження також акцентувало увагу на етичних питаннях, які виникають у зв'язку з впровадженням ML. Причому, головна увага приділена забезпеченню конфіденційності користувацьких даних та зниженню рівня упередженості в алгоритмах.

За результатами досліджень зроблено висновок про величезний потенціал машинного навчання в банківському секторі. Хоча існують певні виклики, правильний підхід може допомогти створити більш ефективні, безпечні та прозорі системи, які підвищать довіру клієнтів.

Ключові слова: машинне навчання, банківський сектор, прийняття рішень, пояснюваність, конфіденційність даних, упередженість, аудит, етичні практики, схвалення позик.

Introduction. As we step further into the digital age, the banking sector continues to harness the power of machine learning to elevate its decision-making process to new heights of precision and efficiency. Machine learning, a significant pillar of artificial intelligence, leverages data-driven methodologies to enhance performance, sidestepping the need for explicit programming. The application of machine learning within the financial technology space has already yielded successful results in various sectors, including, but not limited to, credit risk management, portfolio management, automatic trading, and fraud detection.

However, navigating the banking landscape with machine learning is not without its challenges. The sheer volume and complexity of available data pose considerable hurdles. Moreover, ensuring the relevant and valuable extraction of information from this data pool necessitates the careful and effective application of machine learning methodologies.

A key area where the potency of machine learning can be fully realized is credit risk management. Predicting a client's creditworthiness, particularly credit card holders, has long been a significant challenge in the field. This challenge is heightened by the fact that it often requires the use of vast and complex real-world datasets.

To circumnavigate these difficulties, scholars and industry professionals alike have delved into exploring various machine-learning techniques. The techniques span the spectrum from k-nearest neighbors, decision trees, and boosting, to more complex methodologies like support vector machines and neural networks. By employing these techniques on the large datasets collated from banking institutions, decision-makers are able to gain deeper insights and thus, make more informed and accurate credit risk assessments.

Empirical comparisons based on validation curves have shown that certain techniques, such as decision trees, stand out in their performance. However, each method has its own unique strengths and potential applications, suggesting that a holistic approach, which leverages multiple techniques, could provide even greater accuracy and insight. Therefore, it is essential to conduct a thorough evaluation and comparison of these machine-learning techniques to understand their potential and limitations in the context of credit risk management in the banking sector.

Comparison of Existing Loan Decision-Making Systems Based on Machine Learning Methods

Traditionally, the loan approval process in banks involved human judgment, which was often error-prone and inconsistent. With the advent of ML, banks have significantly improved this process, making it more accurate and efficient. Several ML methods are used today, including Decision Trees, Random Forests, Support Vector Machines (SVM), and Neural Networks. Let's take a deeper look at these methods.

Before comparing various loan decision-making systems based on machine learning methods, let's establish common ground by introducing a dataset to facilitate our discussions. Consider the following simplified dataset with five features (predictor variables) and one binary outcome (target variable):

Customer ID	Credit Score	Annual Income	Years of Credit History	Number of Open Accounts	Number of Defaults	Loan Approved (Y/N)
1	750	$80,000	10	5	0	Y
2	670	$45,000	5	3	1	N
3	720	$95,000	7	4	0	Y
4	640	$60,000	12	2	2	N
5	680	$70,000	6	7	0	Y
6	610	$40,000	4	6	3	N
7	730	$85,000	11	5	0	Y
8	660	$55,000	8	4	1	N
9	710	$75,000	9	6	0	Y
10	630	$50,000	5	3	2	N

In this dataset, the features include:

- "Customer ID" as the unique identifier for each customer.
- "Credit Score" represents the customer's credit rating.
- "Annual Income" indicates the customer's annual income.
- "Years of Credit History" is the customer's credit history duration.
- "Number of Open Accounts" showing the number of currently open accounts of the customer.
- "Number of Defaults" states the number of customer defaults on loans.

The target variable, “Loan Approved (Y/N),” shows whether a loan has been approved.

We will apply Machine Learning methods, including Decision Trees, Random Forests, Support Vector Machines (SVM), and Neural Networks on this dataset to predict the outcome of loan approval. By comparing the performances of these algorithms on the same dataset, we can provide a fair assessment of their effectiveness in the context of loan decision-making systems.

Decision Trees and Random Forests

The loan approval process has been a cornerstone of banking activities for decades. Traditionally, this process was highly reliant on human expertise and judgment. Loan officers, equipped with knowledge and experience, would assess the creditworthiness of a potential borrower by evaluating various factors such as income, employment history, credit score, existing debts, and more.

While human judgment added a personal touch to the process, it could have been more consistent due to the subjective nature of decision-making. Additionally, human errors were not uncommon, given the large amount of data that needed to be processed and analyzed. The manual process was time-consuming and often needed to catch up with the increasing volume of loan applications. Furthermore, due to financial data's inherent complexities and variability, identifying patterns and correlations took time, making the risk assessment less reliable.

The advent of Machine Learning (ML) has ushered in a new era in the banking industry, promising significant enhancements to the traditional loan approval process. With its ability to learn from data, make predictions, and improve over time without being explicitly programmed, ML is ideally suited to handle complex tasks like loan approval.

Several ML methods have been applied to streamline the loan approval process, each with strengths and characteristics.

Decision Trees: Decision Trees are a type of ML algorithm that operates by creating a model of decisions based on actual data. In the context of loan approval, a Decision Tree might look at an applicant's income, credit score, employment history, and other relevant factors, making decisions at each node of the tree until it arrives at a final decision. This method is transparent and easy to understand, but can sometimes lead to oversimplified decisions.

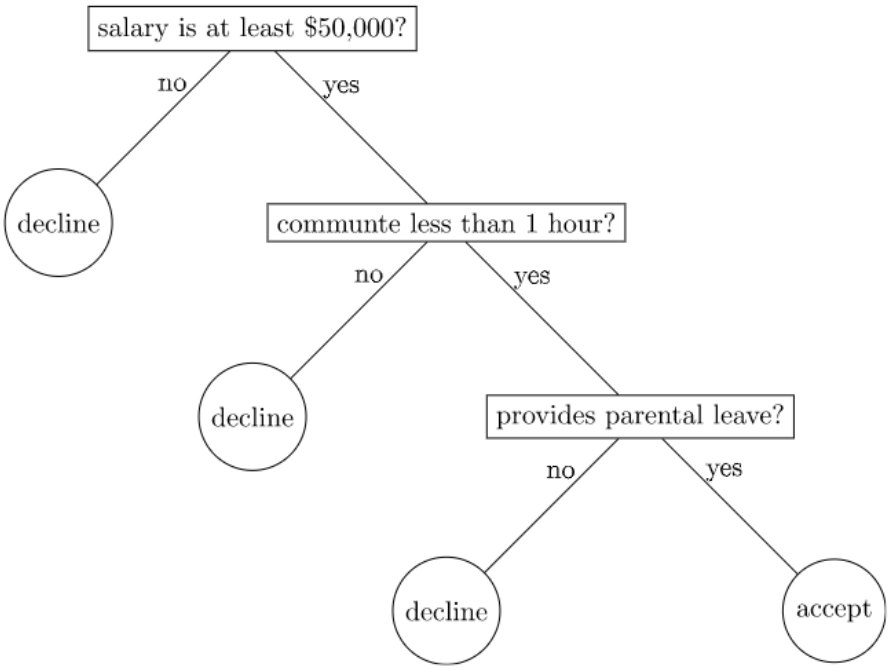


Fig. 1. Illustration of the decision tree.

Decision Trees make decisions by splitting the data based on feature values. They use a set of rules that can be visualized in a tree structure to make decisions.

Here's a simplified example of how a decision tree could be structured based on our dataset:

```
1. Is the Credit Score ≥ 700?
   - Yes: Go to 2.
   - No: Go to 3.
2. Has the client defaulted before (Number of Defaults > 0)?
   - Yes: Loan is denied.
   - No: Loan is approved.
3. Is the Annual Income ≥ $60,000?
   - Yes: Go to 4.
   - No: Loan is denied.
4. Are the Number of Open Accounts > 5?
   - Yes: Loan is denied.
   - No: Loan is approved.
```

Random Forests: Random Forests are an extension of Decision Trees. They operate by creating a multitude of decision trees and outputting the class that is the mode of the classes of individual trees. By averaging out the decisions of many different trees, Random Forests tend to be more accurate and less prone to overfitting than individual Decision Trees.

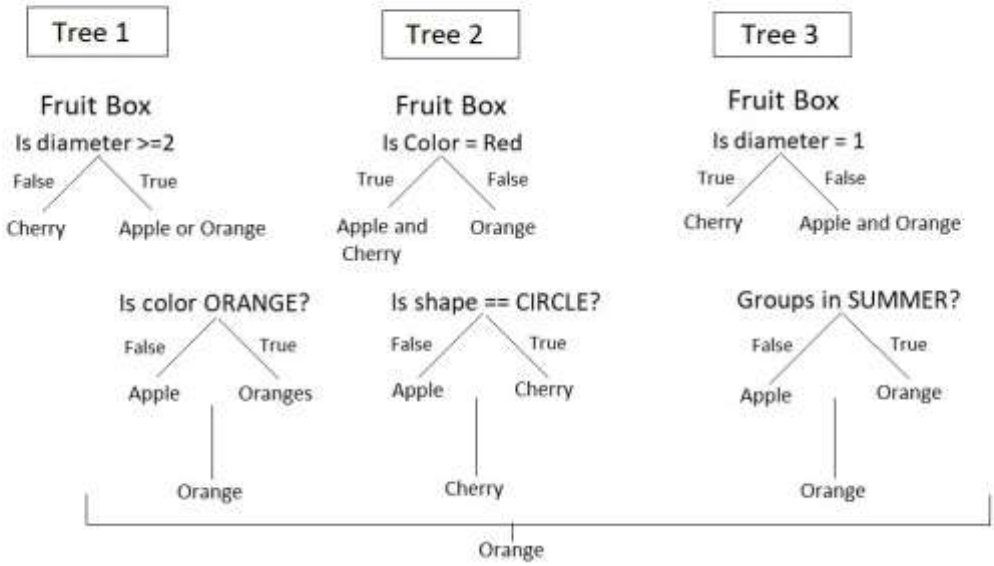


Fig. 2. Illustration of the Random Forests.

A Random Forest is a collection of Decision Trees that have been trained on different subsets of the data. Each tree in the forest makes its own decision, and the final decision is made by taking a majority vote on the decisions made by all the trees in the forest.

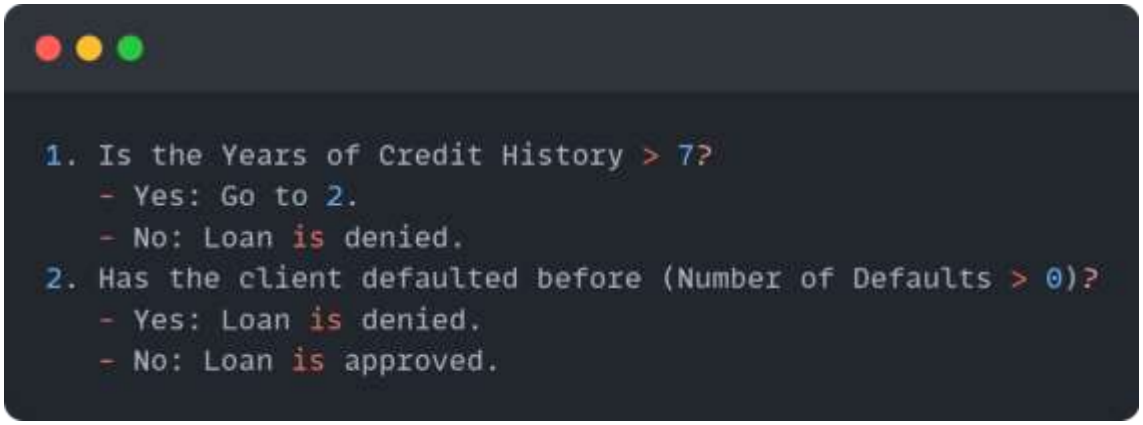
While creating each decision tree, the Random Forest algorithm only considers a random subset of the features at each split in the learning process, which helps increase the diversity of the trees and reduces the likelihood of overfitting.

Here's a simplified example of how Random Forest might make a loan decision based on our dataset. Let's assume we are working with a Random Forest comprising three Decision Trees. Note that real-world applications often involve hundreds or thousands of trees and far more complex decision-making processes.

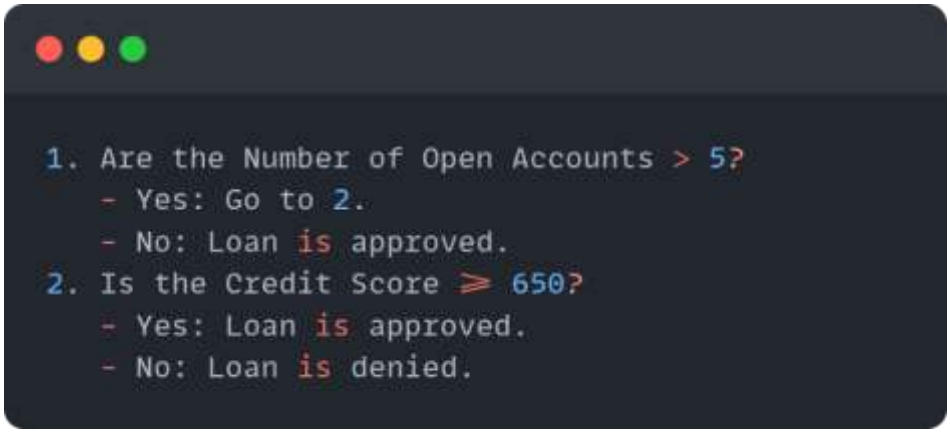
Decision Tree 1 (Based on Credit Score, Annual Income)

```
1. Is the Credit Score ≥ 700?
   - Yes: Go to 2.
   - No: Loan is denied.
2. Is the Annual Income ≥ $60,000?
   - Yes: Loan is approved.
   - No: Loan is denied.
```

Decision Tree 2 (Based on Years of Credit History, Number of Defaults)



Decision Tree 3 (Based on Number of Open Accounts, Credit Score)



For a customer with a credit score of 750, an annual income of $80,000, 10 years of credit history, no defaults, and 5 open accounts:

- Decision Tree 1 would approve the loan (credit score is >= 700, income is >= $60,000).
- Decision Tree 2 would approve the loan (credit history is > 7 years, no defaults).
- Decision Tree 3 would approve the loan (number of open accounts <= 5).

Since all trees vote "yes", the final decision of the Random Forest is to approve the loan.

For a customer with a credit score of 640, an annual income of $45,000, 5 years of credit history, 1 default, and 3 open accounts:

- Decision Tree 1 would deny the loan (credit score is < 700).
- Decision Tree 2 would deny the loan (credit history is <= 7 years).
- Decision Tree 3 would approve the loan (number of open accounts <= 5).

In this case, the majority of trees vote "no", so the final decision of the Random Forest is to deny the loan.

This is a highly simplified example. The strength of the Random Forest model lies in its ability to consider a wide variety of paths and outcomes, thereby providing a more robust and accurate decision overall. It's also more resilient to overfitting than a single decision tree.

Support Vector Machines (SVM): SVMs are powerful supervised learning models for classification and regression analysis. They work by finding a hyperplane in a multi-dimensional space that distinctly classifies the data points. In the context of loan approval, SVM could be used to classify whether a loan application should be approved or denied based on several factors.

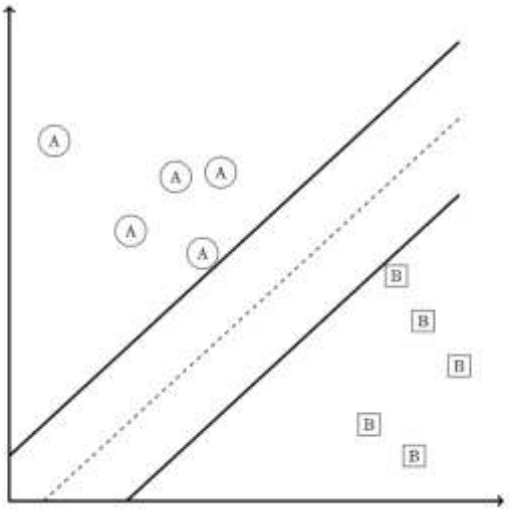


Fig. 3. Illustration of the SVM.

Support Vector Machines (SVM) work fundamentally differently than Decision Trees and Random Forests. SVMs are typically used for classification problems, like our loan approval process. They work by finding the hyperplane (in 2-dimensional space, it's simply a line) that best divides the data into the respective classes while maximizing the margin (distance) between the closest points (support vectors) and the hyperplane.

Before running an SVM, data preprocessing is crucial, as SVMs are sensitive to the scale of the data. We would need to normalize or standardize our numerical data, such as Credit Score, Annual Income, etc.

Since SVMs' decision-making process isn't as easily visualized as decision trees, let's illustrate the concept in a simplified 2-dimensional example with only two features: Credit Score and Annual Income. In reality, SVMs can handle multi-dimensional data, considering all features in our dataset simultaneously.

Suppose we plot 'Credit Score' on the x-axis and 'Annual Income' on the y-axis. Our goal would be to find a line that separates customers who were approved for a loan from those who were not, in such a way that the line is as far away as possible from the nearest points of each class.

After training the SVM with our data, we find such a line. To make a prediction for a new customer, we need to plot this customer's features (Credit Score and Annual Income) on our graph and see which side of the line they fall on.

For example, a new customer with a high credit score and annual income would likely fall on the side of the line corresponding to "Loan Approved". Conversely, a customer with a low credit score and low annual income would fall on the "Loan Denied" side.

Remember that this is an overly simplified version of how SVMs work, especially considering the SVM's powerful ability to use kernel tricks to handle non-linearly separable data. However, the general idea of drawing a boundary (in higher dimensions, it's a hyperplane) and categorizing new points based on which side they fall on remains the same.

Neural Networks: Neural Networks are among the most complex and powerful ML models. They are designed to simulate the way a human brain works and are particularly good at recognizing patterns in large, complex datasets. In loan approval, a neural network would be trained on past data and then used to predict whether new loan applications are likely to default.

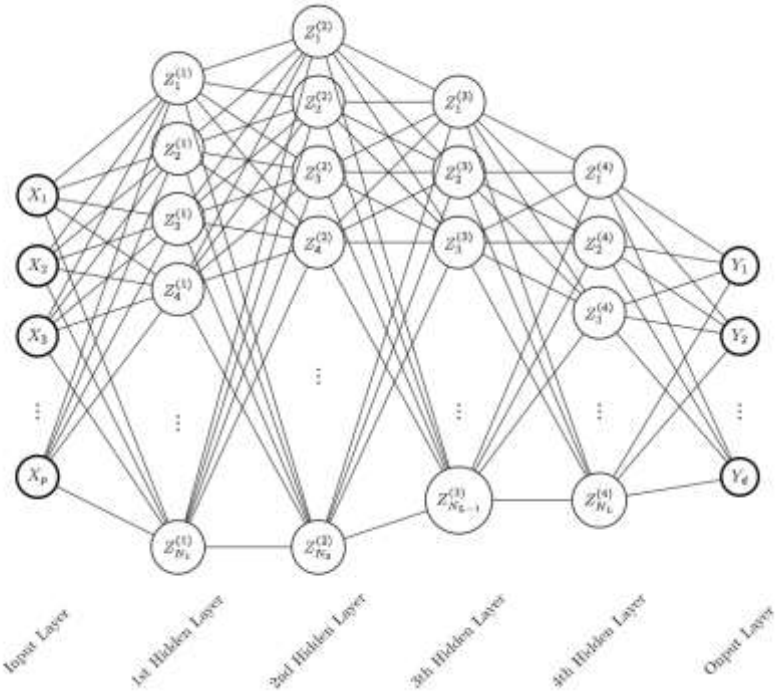


Fig. 4. Illustration of a neural network with four layers.

Neural Networks, specifically for this binary classification task we'd likely use a type of network known as a feed-forward network, which is a machine learning model inspired by the human brain. They consist of interconnected nodes (or "neurons") arranged in layers: an input layer, one or more hidden layers, and an output layer.

Let's use an overly simplified neural network example to illustrate how it could be used to decide whether to approve or deny a loan based on our dataset.

A typical neural network could have the following input layer with one node for each feature:

- Credit Score,
- Annual Income,
- Years of Credit History,
- Number of Open Accounts,
- Number of Defaults.

One hidden layer with a number of neurons (this number can vary widely and would be determined during the model design and training process). An output layer with a single node representing the loan' is approved.

When a new loan application comes in, the values of each feature are fed into the corresponding input nodes. These values are then passed through the network, multiplied by weights (learned during the training phase), and an activation function is applied to each neuron until they reach the output layer.

The output layer will provide a value between 0 and 1, representing the probability of the approved loan. For binary classification tasks, if the output is greater than or equal to 0.5, we will approve the loan, and if it's less than 0.5, we will deny the loan.

Let's keep in mind that this is a very simplified example. Real-world neural networks often contain many hidden layers (creating a "deep" network), features, and complexities. The actual process of training a neural network involves techniques like backpropagation and gradient descent, and the network's design and the choice of activation functions can significantly impact the network's performance.

As with all machine learning models, it's important to evaluate a neural network's performance using a separate test set or cross-validation to ensure that it generalizes well to unseen data.

Disadvantages of Existing Decision-Making Systems in the Banking Sector Based on Machine Learning Methods

Despite the significant benefits and improvements brought by ML in the banking sector, these systems also have their share of drawbacks.

Overfitting: Overfitting is a common problem in ML, where the model performs well on the training data but poorly on unseen data. This happens when the model learns the noise and the data's underlying pattern.

Explainability : ML models, particularly complex ones like neural networks, are often referred to as "black boxes" due to their lack of interpretability. This can be problematic in a banking context, where decisions need to be explained to customers.

Data Privacy: ML models require large volumes of data, which raises concerns about data privacy. Ensuring customer data is secure and used appropriately is a major challenge for banks using ML.

Bias: If the data used to train the ML model is biased, the decisions made by the model will also be biased. This can lead to unfair loan decisions, disproportionately affecting certain groups of customers.

Table 1 provides a summary of the main disadvantages of these systems.

Disadvantage	**Description**
Overfitting	The model learns the noise in the training data, leading to poor performance on unseen data.
Explainability	ML models, especially complex ones, need more transparency and are difficult to interpret.
Data Privacy	Collecting and using large volumes of customer data raises privacy concerns.
Bias	If the training data is biased, the model's decisions will also be biased.

Conclusion. Machine Learning (ML) has unquestionably emerged as a transformative force in the banking sector, offering significant enhancements to decision-making processes, particularly those involved in loan approval. The introduction of ML has led to a paradigm shift, transitioning from traditional human-dependent mechanisms to more objective, efficient, and consistent ML-powered systems. These systems, equipped with techniques like Decision Trees, Random Forests, Support Vector Machines (SVM), and Neural Networks, have impressive performance in processing vast volumes of data and making accurate predictions.

However, acknowledging the positive impacts of ML does not equate to ignoring its limitations. While ML brings numerous benefits, it also presents several challenges that must be confronted and resolved. Overfitting, lack of explainability , data privacy concerns, and potential biases are some of the key disadvantages associated with ML-based decision-making systems.

Although a widespread problem in ML, overfitting can be mitigated with appropriate techniques such as regularization and cross-validation. The 'black box' issue, a critical challenge associated with complex ML models, calls for the development of interpretable and explainable AI. Data privacy concerns necessitate stringent data protection measures and ethical guidelines to ensure the responsible use of customer data. Bias in ML models underlines the need for rigorous fairness auditing of these models, ensuring that they do not perpetuate existing societal biases.

In an era where technology is evolving at an unprecedented rate, the future of ML in banking appears promising. The ongoing advancements suggest that these challenges are manageable but are the stepping stones towards the next generation of more robust, fair, and transparent ML systems.

It's clear that despite its limitations, ML will continue to play an integral role in shaping the future of the banking industry. As we continue to embrace these technologies, it's critical to balance leveraging their potential and ensuring ethical, fair, and transparent practices. With such an approach, the banking sector can continue to harness the power of ML, enhancing efficiency and accuracy in decision-making while maintaining the trust and confidence of customers.

References:

1. A.K. Choudhury, R.K. Swain "Machine learning in banking risk management: A literature review," Risk Management, vol. 20, pp. 1-30, 2018.

2. M. Zeng, T. Le, A. Liu, G. Zhang, M. Hussain, H. Chen, Q. Tian, and Z. Liu, "A Comparative Study on Decision Tree Algorithms for Bank Loan Default Prediction," in Procedia Computer Science, vol. 147, pp. 400-407, 2019.

3. H. Baesens , R. Van Gestel , S. Viaene , M. Stepanova, and J. Suykens , "Benchmarking state-of-the-art classification algorithms for credit scoring," Journal of the Operational Research Society, vol. 54, pp. 627-635, 2003.

4. P.A. Estévez , M. Tesmer, C.A. Perez, J.M. Zurada , "Normalized Mutual Information Feature Selection," IEEE Transactions on Neural Networks, vol. 20, no. 2, pp. 189-201, 2009.

5. R.K. Swain "Machine Learning for Risk Management in Banking Sectors: A Comparative Analysis of Supervised Algorithms," International Journal of Computational Intelligence Studies, vol. 7, no. 1/2, p. 37, 2018.

6. Zhebka V., Gertsiuk M., Sokolov V., Malinov V., Sablina M. Optimization of Machine Learning Method to Improve the Management Efficiency of Heterogeneous Telecommunication Network / CEUR Workshop Proceedings, 2022, c. 149-155.

7. Malinov, V., Zhebka, V., Zolotukhina, O., Franchuk, T., Chubaievskyi, V. Biomining as an Effective Mechanism for Utilizing the Bioenergy Potential of Processing Enterprises in the Agricultural Sector / CEUR Workshop Proceedings, 2023, c. 223-230.

8. Anakhov P., Zhebka V., Bondarchuk A., Storchak K., Sablina M. Increasing the Reliability of a Heterogeneous Network using Redundant Means and Determining the Statistical Channel Availability Factor / CEUR Workshop Proceedings, 2023, p. 231–236