Introduction

Credit risk management is central to sustainable banking operations, directly impacting a financial institution's profitability and regulatory compliance. The aftermath of the COVID-19 pandemic amplified the urgency to address rising consumer loan delinquencies, compelling banks to modernize approaches to risk assessment. While traditional loan approval processes depend heavily on credit scores and manual reviews, these methods often overlook subtle but critical borrower behaviors, leading to inefficiencies in processing and potential financial losses. These gaps highlight the need for data-driven solutions that use the power of advanced analytics to assess risk proactively.

Group 1 Data Science Consultants partnered with Evergreen National Bank to address these challenges by developing an advanced predictive loan default model. This project uses machine learning methodologies with diverse financial, demographic, and behavioral data to deliver actionable insights into borrower risk. The core of this project is the Berka Dataset—a repository of anonymized banking data that captures client behaviors, account details, and transaction histories. By analyzing this rich dataset, our team seeks to provide Evergreen National Bank with a solid framework for transforming its credit risk management practices.

The Berka Dataset offers an opportunity to examine loan default risks through interconnected variables such as account balances, transaction patterns, and loan histories. While its historical and regional context may limit its general applicability, the dataset's depth provides invaluable insights into borrower behaviors used to inform modern risk assessment models. Additionally, integrating behavioral and demographic data aligns with recent advancements in predictive analytics, emphasizing the importance of capturing complex, non-linear relationships in financial decision-making.

This project responds to key pain points in the industry, including the inefficiencies of manual assessments and the limitations of traditional scoring systems. Our predictive framework identifies high-risk borrowers early by employing machine learning models like Gradient Boosting Machine (GBM), Random Forest, Generalized Linear Models (GLM), and Naïve Bayes. It enables the customization of loan products tailored to individual risk profiles. These models balance predictive accuracy, scalability, and ease of implementation, ensuring they meet the bank's operational needs and strategic goals.

The following research questions guide this project:

1. **Are larger loan amounts and longer loan durations associated with a higher default risk?** This analysis investigates whether increased financial obligations and prolonged repayment periods elevate the likelihood of default, offering insights into optimal loan structuring.
2. **How does a client's account balance affect their ability to repay loans on time?** This analysis examines how liquidity influences timely repayments, providing key indicators for assessing borrower stability.
3. **Does a history of frequent overdrafts increase the likelihood of default?** By analyzing overdraft behavior, this question aims to uncover patterns of financial instability that can signal default risk.

The objectives of this report are to:

1. Define the challenges inherent in traditional credit risk models.
2. Detail the methodologies employed in building a reproducible predictive model.
3. Offer actionable recommendations based on the model's findings to enhance decision-making and operational efficiency.

This paper is a blueprint for integrating advanced analytics into Evergreen National Bank's loan approval processes, emphasizing scalability, compliance, and client-centric solutions. The bank can proactively manage credit risk, optimize its portfolio, and strengthen its long-term financial stability using the insights provided.

Literature Review

Overview and Synthesis

Integrating machine learning (ML) techniques into credit risk management has drawn significant attention in recent years, driven by data availability and advancements in computational power. This review synthesizes insights from relevant sources to understand how ML models can enhance loan default prediction, focusing on their application to demographic, behavioral, and financial data.

The Berka Dataset is the core resource for this project, offering anonymized data across multiple interrelated tables such as accounts, transactions, and loans (Sorayut & Czech Technical University in Prague, 1999, 2022). While the dataset captures valuable behavioral and demographic factors, recent studies offer complementary perspectives. Chen (2022) explores borrower behavior, analyzing patterns such as overdrafts and transaction trends to predict defaults. This reinforces the importance of behavioral data for understanding financial instability and supports using features derived from the Transactions table in the Berka Dataset. Balancing methods, such as Random Over-Sampling Examples (ROSE), have proven essential for addressing the class imbalance often observed in datasets like Berka, ensuring more equitable model performance across target variables.

Recent research has highlighted the effectiveness of models like Random Forest and XGBoost in capturing complex, non-linear relationships in financial datasets (Suhadolnik et al., 2023; Zhou, 2023). These ensemble methods outperform traditional approaches like logistic regression by leveraging feature interactions to enhance predictive accuracy. Meanwhile, GLMs remain relevant for their simplicity and interpretability, making them a valuable baseline for assessing model performance in high-stakes financial contexts (Hahami & Piper, 2022).

Egwa (2022) bridges theoretical advancements with practical implementation, emphasizing the scalability and efficiency of Random Forest and similar algorithms for large datasets. Addo et al. (2018) further explore the computational trade-offs of using advanced ML models, noting that ensemble techniques strike an effective balance between performance and resource requirements.

Analysis and Interpretation

These sources collectively highlight the transformative potential of machine learning in credit risk assessment. With its diverse data points, the Berka Dataset allows for a broad analysis of borrower risk. Its interrelated tables enable studying how demographic factors like age and income interact with

behavioral variables like overdraft frequency and transaction patterns. Chen (2022) builds on this by emphasizing the role of transaction behavior in understanding financial instability. This insight validates the importance of using the Transactions table to derive features like overdraft frequency and average monthly balance.

Ensemble models like Random Forest and XGBoost stand out for their ability to capture non-linear interactions and handle missing data effectively (Suhadolnik et al., 2023; Egwa, 2022). These methods are particularly well-suited for analyzing the Berka Dataset, where relationships between variables like loan amounts, account balances, and repayment behaviors are unlikely linear. Zhou (2023) and Addo et al. (2018) emphasize that these models' scalability make them ideal for modern financial applications.

Gradient Boosting Machine (GBM) further enhances predictive capabilities by sequentially building trees to minimize residual errors. Studies like Egwa (2022) highlight GBM's ability to achieve high accuracy, particularly in datasets with complex interactions between features. While computationally more intensive than Random Forest, GBM's iterative approach often yields superior performance metrics, making it a strong candidate for handling imbalanced datasets like the Berka Dataset. Its adaptability to weight misclassified instances makes GBM well-suited for improving classification outcomes in high-risk scenarios.

Naïve Bayes, though simpler, offers distinct advantages in its probabilistic approach. Its reliance on Bayes' theorem allows it to efficiently classify outcomes even with limited training data. While it assumes feature independence—a limitation in datasets with interrelated variables—it remains valuable for its computational efficiency with noisy data. Hahami and Piper (2022) note that Naïve Bayes can be effective in scenarios where interpretability is prioritized over raw predictive power. For this project, Naïve Bayes is a baseline model to evaluate the effectiveness of more complex methods like GBM and XGBoost.

The project incorporates these methods to address key gaps in traditional credit scoring models. GBM and Naïve Bayes, alongside Random Forest and XGBoost, ensure a comprehensive approach that balances accuracy, computational feasibility, and model interpretability. This integration aligns with the project's objectives to create a scalable, reproducible predictive framework for Evergreen National Bank.

Critical Evaluation

Each source contributes valuable perspectives but also has limitations. The Berka Dataset represents banking practices from a specific region and era, which may limit the generalizability of findings to modern financial contexts (Sorayut & Czech Technical University in Prague, 1999, 2022). Suhadolnik et al. (2023) provided robust evidence of ensemble methods' superiority but lacked a detailed discussion of interpretability, an important factor in high-stakes financial decisions. Similarly, S&P Global (2020) effectively critiqued traditional models but offered limited guidance on implementing alternative data into predictive systems.

Egwa (2022) and Addo et al. (2018) underscore the computational efficiency and scalability of ensemble methods like Random Forest and XGBoost. However, these sources need to address the challenges of real-time deployment in dynamic financial environments, an important consideration for operationalizing

machine learning in the banking sector. Zhou (2023) effectively explores emerging methods for optimizing ML models but focuses primarily on technical details, providing limited practical guidance for implementation within existing workflows. The application of balancing techniques, such as ROSE, represents a solid strategy to mitigate the effects of class imbalances in training data, improving model performance and fairness.

Hahami and Piper (2022) offer a meta-analysis of ML performance in credit risk. However, their focus on general model evaluation overlooks challenges like balancing accuracy with fairness in regulated environments. This gap highlights the need for additional research on integrating machine learning into credit risk systems in ways that prioritize equity and scalability.

Resource Summary

This literature review underscores the transformative potential of machine learning techniques in credit risk assessment, highlighting their ability to address the limitations of traditional credit scoring systems. The Berka Dataset forms the foundation of this project, offering a comprehensive repository of demographic, behavioral, and financial data that provides valuable insights into borrower behaviors. Its interrelated tables allow for exploring how variables like account balances, transaction histories, and overdraft frequency and its influence on default risk.

Recent studies reinforce the effectiveness of Random Forest, XGBoost, and Generalized Linear Models (GLM). Ensemble methods like Random Forest and XGBoost are particularly noted for their ability to handle non-linear relationships and imbalanced datasets (Suhadolnik et al., 2023; Egwa, 2022). These models enable robust analysis of the complex interactions in datasets like Berka, offering predictive accuracy without sacrificing operational efficiency. GLM is a useful benchmark, providing interpretability and simplicity in high-stakes financial applications (Hahami & Piper, 2022).

Integrating behavioral data into predictive models has been highlighted as a key strength. Chen (2022) and Addo et al. (2018) demonstrate the importance of transaction patterns, overdraft history, and other behavioral variables. These insights were validated using features derived from the Berka Dataset's Transactions table, ensuring the model captures the nuanced behaviors that traditional methods often miss.

This resource synthesis directly informs the project's research questions by emphasizing the value of demographic, financial, and behavioral data in loan default prediction. The chosen methodologies—Random Forest, XGBoost, and GLM—address the challenges of non-linear relationships, data imbalances, and operational scalability. These insights provide a strong foundation for Evergreen National Bank's efforts to modernize its credit risk management practices, offering a balanced approach that uses technological advancements while remaining mindful of practical implementation.

Methodology

Data Collection and Integration

For this project, we use the publicly available PKDD'99 Discovery Challenge Berka Dataset (Sorayut & Czech Technical University in Prague, 1999, 2022), a rich source of anonymized financial and

demographic data collected from a Czech bank. This dataset comprises multiple interconnected tables, capturing transactional details, demographic information, client profiles, loan records, and account data, as illustrated in *Figure 1*.

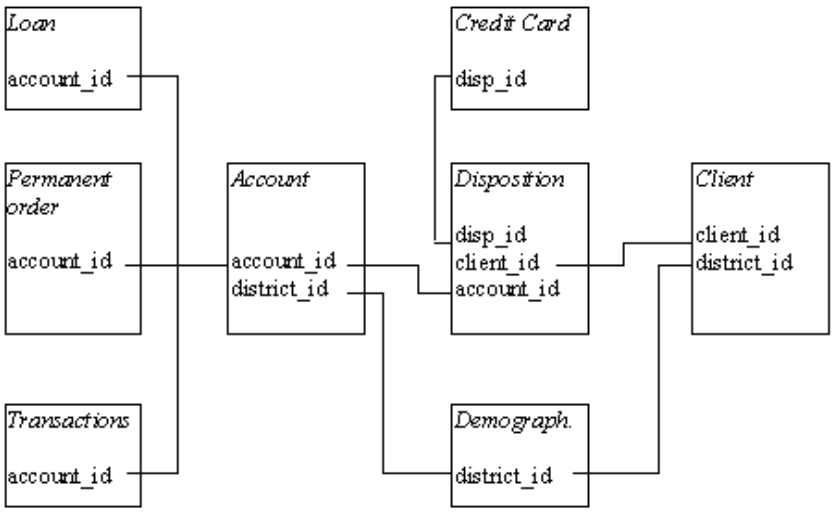

Figure1: Entity-Relationship Diagram of the Berka Dataset

In the data collection and integration phase, our goal is to combine these diverse tables into a unified dataset, ensuring that all relevant variables are included for predictive modeling.

The primary tables on Berka dataset consists of:

- Clients: Client-specific information, including client_id and district_id.
- Accounts: Details about bank accounts and their associated districts.
- Loans: Information on loans including the loan id, amount, duration, and whether the loan was defaulted.
- Transactions: Logs account transactions including transaction type, amount, and balance.
- Permanent Orders: Tracks standing orders for regular payments (account_id).
- Credit Cards: Captures information on clients who hold credit cards (disp_id).
- Dispositions: Connects accounts to their owners (disp_id and client_id).
- Demographics: Demographic data by district, such as average salary and unemployment rate.

To answer our research questions and predict loan default, we prioritize collecting variables that describe:

> Client behavior (e.g., transaction patterns, account balances, and overdraft history), Loan characteristics (e.g., loan amount and duration) and Demographic context (e.g., average income and unemployment in the client's district).

Data Integration
We merge the Loans table with the Accounts table using the account_id key. This allows us to associate each loan with its corresponding account details, such as the district in which the account is located. Key variables from this step include the loan amount, duration, loan status (default or successful), and account-level information.

Next, transaction data is incorporated by aggregating the Transactions table for each account. Features such as total deposits, total withdrawals, average monthly balance, and the frequency of overdrafts are derived. These aggregated values provide a concise summary of account activity, which is crucial for understanding a client's financial behavior and predicting potential risks.

The Dispositions table acts as a link between Accounts and Clients. Using the disp_id and account_id keys, accounts are connected to their respective clients. With this relationship established, client-specific details, such as the district ID from the Clients table, are included in the dataset. This enables the model to capture the direct relationship between account activity and client demographics.

Credit card information is added to enrich the dataset further by linking the Credit Cards table through the disp_id key in the Dispositions table. A binary indicator is created to show whether a client holds a credit card. This feature enhances the dataset by reflecting the client'sclient's access to additional credit and spending behavior.

Demographic data from the Demographics table is then integrated using the district_id key. This step introduces socio-economic features such as average district salary, unemployment rates, and the number of businesses in the district. These variables provide context for understanding external factors impacting clients' ability to repay loans.

Permanent orders, which represent recurring payments, are incorporated by linking the Permanent Orders table through account_id. Features such as the count of standing orders and their average value are derived. These variables capture the client's regular financial commitments, which can influence their overall financial stability.

Throughout the integration process, missing data is handled systematically. Numeric variables with missing values are imputed using the median, while categorical variables are imputed with the mode. This ensures the dataset remains complete and ready for analysis without introducing biases.

Finally, we standardize all variables to maintain consistency across the dataset. Dates are formatted uniformly, categorical variables are encoded into binary or ordinal values, and numeric features are scaled to ensure they are comparable and avoid skewed distributions. This step ensures the dataset is clean, structured, and prepared for modeling.

The resulting dataset comprehensively integrates financial, demographic, and behavioral data. Each record now represents an account or loan, enriched with all relevant information needed to predict loan default with precision. This fully integrated dataset is the foundation for further feature engineering and modeling efforts.

Feature Engineering and Consolidation

Feature engineering used in the dataset performed on both numerical and categorical variables helps improve the overall comprehensiveness of the raw data and create an accurate dataset used for further modeling techniques. Through data integration, aggregated metrics from the transactional dataset, including account balance and overdrafts, served as a financial indicator of stability. Some features indicate account conditions in specific periods of time like last month or last three-month balance, and additional calculations on mean, minimum, and maximum could help provide some insights on understanding how clients operate their accounts in the short term and their financial operation trends. Categorical variables with descriptive labels like transaction type (e.g., PRIJEM becomes credit) are transformed through one-hot encoding processes that are ready for future machine learning algorithms operations and avoid introducing biases and using the k-symbol feature using similar encoding techniques to show transaction purposes (e.g., out_household) and further reveal clients’ financial behavior patterns.

Model Results and Data Interpretation

For model development, we build models based on both original data and data with oversampling due to heavy imbalance. Preparing the data, we encoded loan_status, our target variable, into binary form, and converted categorical variables to factors if it is necessary. There are five models: Gradient Boosting Machine (GBM), Generalized Linear Model (GLM), GLM with ROSE (Random Over-Sampling Examples) for balancing, Naive Bayes with ROSE, and Random Forest with ROSE. ROSE was applied to balance the dataset in some experiments, creating a more equitable distribution of the target classes. We used accuracy and area under the receiver operating characteristics curve (ROC AUC) to evaluate our models, and used confusion matrix to assess classification performance for each model.

Model Performance Comparison

Model	Accuracy	ROC AUC
GBM	0.956204379562044	0.91991341991342
GLM	0.91970802919708	0.8997113997114
GLM (ROSE)	0.802919708029197	0.906204906204906
Naive Bayes (ROSE)	0.700729927007299	0.925685425685426
Random Forest (ROSE)	0.824817518248175	0.912337662337662

GBM achieved the highest accuracy with a strong AUC, and top features included those related to balance and payment amounts. GLM performed slightly worse than GBM. It struggled with overfitting warnings, and coefficients indicated singularities in some features. AUC of GLM slightly increased with ROSE, indicating better handling of imbalanced data. Naïve Bayes with ROSE has a relatively lower accuracy compared to other models, but it has notable AUC, highlighting good sensitivity to certain thresholds. Random forest with ROSE also has a relatively high AUC. Its variable importance ranked payment and balance features as most predictive. Analyzing all five models, we can conclude that there are multiple features that greatly contribute to successful loan default prediction: balance_min_last_month, balance_max_last_month, loan_duration, loan_payments, count_monthly_payment_trans, and acc_frequency.

The analysis reveals that GBM provides the best combination of accuracy and interpretability for predicting loan_status. It is the top-performing model based on accuracy and robust AUC. ROSE improves AUC but normally leads to reduced accuracy; a trade-off is required depending on the application's tolerance for false positives and negatives. In the future, we can use other metrics such as F-1 score to evaluate the models when the data is heavily imbalanced. We also should focus on engineering features related to balanced and transactions for future iterations.

Recommendations

Based on the findings of this project, Group 1 Data Science Consultants propose the following recommendations to improve Evergreen National Bank's credit risk management, addressing key research questions regarding loan amounts, account balances, and transaction behaviors:

1. Adopt Predictive Modeling for Loan Approvals

By adopting machine learning models like Gradient Boosting Machine (GBM) and XGBoost, Evergreen National Bank can improve its loan approval process and predict borrower risk more accurately. GBM, in

particular, demonstrated superior accuracy and AUC, making it a cornerstone for identifying high-risk borrowers and informing lending strategies. These models can evaluate the relationship between loan amounts, repayment durations, and default probabilities, providing actionable insights into borrower behavior. For instance, predictive modeling can identify how larger loan amounts or extended repayment terms exacerbate financial strain, helping the bank adjust lending thresholds or repayment structures accordingly. Specific risk assessments enable proactive measures to mitigate potential defaults, instilling a sense of optimism about the bank's future credit risk management.

In addition to identifying high-risk borrowers, these models can support customizing loan terms, such as adjusting interest rates or repayment schedules to reflect a borrower's financial situation. This tailored approach minimizes default rates and strengthens customer trust and satisfaction, reassuring potential borrowers and investors about the bank's commitment to its clients. Integrating these predictions into the bank's decision-making processes ensures a balance between economic growth and risk management, making the loan approval process efficient and borrower-centric.

Additionally, using Naïve Bayes as a baseline model provides a valuable benchmark to evaluate the added complexity and predictive gains of advanced methods like GBM and Random Forest. This step ensures the bank can justify its choice of models based on performance improvements.

2. Enhance Data Integration and Feature Engineering Practices

To fully capture the relationship between account balances and repayment behavior, the bank should enrich its data sources by incorporating real-time transaction data, spending patterns, and employment stability metrics. Account balances, especially their trends over time, are critical indicators of a borrower's capacity to manage loan repayments. By integrating such data, the bank can more accurately gauge borrowers' liquidity and potential to fulfill loan obligations.

Furthermore, alternative data sources—such as utility payments and digital transactions—offer valuable insights for assessing thin-file clients or those with limited credit histories. For example, frequent declines in account balances could signal financial instability, while consistent patterns in bill payments might indicate reliability. Building automated pipelines to preprocess and update these datasets ensures that predictive models operate with the most accurate and comprehensive data possible. This enables Evergreen National Bank to improve default prediction accuracy and streamline its credit decision processes.

Feature engineering should also focus on deriving actionable insights from transactional data. For example, trends in account balances, overdraft frequencies, and payment patterns emerged as highly predictive features in this project. These engineered variables can significantly enhance the accuracy of loan default models by capturing nuanced borrower behaviors.

3. Improve Operational Efficiency

Automated workflows that prioritize high-risk accounts based on overdraft frequencies and transactional anomalies can significantly improve the bank's operational efficiency. These workflows can alert credit teams to distressed borrowers, enabling timely interventions such as restructuring loan terms or offering financial counseling. This proactive approach mitigates potential defaults and demonstrates the bank's commitment to providing personalized customer service.

Automation can also optimize the bank's overall loan processing system, ensuring that low-risk borrowers are fast-tracked through approval. At the same time, human resources are directed toward resolving complex or high-value cases. By leveraging machine learning-driven insights, the bank can identify transaction patterns indicative of financial health, allowing for faster, data-backed decision-making. This dual approach—efficiency for low-risk borrowers and focused support for high-risk clients—maximizes resource use and boosts borrower satisfaction, instilling confidence in the bank's ability to manage credit risk effectively.

To further enhance operational efficiency, techniques like Random Over-Sampling Examples (ROSE) should be employed to address imbalanced datasets. ROSE ensures a more equitable representation of default and non-default cases, leading to better model performance and more accurate prioritization of high-risk accounts.

4. Monitor and Mitigate Bias

Evergreen National Bank must make transparent and equitable loan approval decisions to ensure compliance with regulatory standards, particularly those involving longer durations or larger amounts. The bank should routinely audit its predictive models for biases in output, especially concerning demographic groups with varying repayment behaviors. For instance, explainable AI tools can reveal whether transaction history or overdraft frequency disproportionately impacts financially vulnerable populations, enabling the bank to address systemic inequities in its lending practices.

Additionally, periodic retraining of the model with diverse datasets can mitigate emerging biases and adapt to changing borrower dynamics. Enhanced transparency also builds trust with clients and regulators, who expect fair lending practices. Beyond compliance, bias mitigation ensures the bank's predictive models remain practical and ethically sound, positioning Evergreen National Bank as a leader in responsible and data-driven lending practices.

To strengthen these efforts, the bank should establish a dedicated team to oversee bias audits and develop strategies for equitable loan approval processes. Using fairness metrics during model evaluation will ensure that predictions align with ethical lending practices and regulatory requirements.

Conclusion

This project highlights the potential of machine learning in revolutionizing credit risk management. By leveraging the rich and diverse Berka Dataset alongside advanced predictive models such as Gradient Boosting Machine (GBM), XGBoost, Random Forest, and Naïve Bayes, Evergreen National Bank can enhance its ability to identify high-risk borrowers, make more informed lending decisions, and optimize operational efficiency. These models, particularly GBM, demonstrated exceptional performance in terms of accuracy and AUC, making them invaluable for tailoring loan terms and mitigating potential defaults.

The findings of this project underscore the importance of incorporating alternative data sources—such as transaction trends, spending patterns, and employment stability metrics. Equally important is the role of feature engineering in capturing nuanced borrower behaviors. Techniques like Random Over-Sampling Examples (ROSE) address dataset imbalances, ensuring equitable model performance across borrower profiles. Additionally, including Naïve Bayes as a baseline model provides a valuable benchmark,

enabling the bank to justify the complexity of more advanced methods based on tangible performance gains.

Looking ahead, the adoption of these recommendations will position Evergreen National Bank as a leader in data-driven lending practices. By integrating these predictive models into its decision-making processes, the bank can balance growth and risk management, enhance customer satisfaction through personalized loan offerings, and demonstrate a strong commitment to transparency and fairness. Furthermore, proactive monitoring for biases and regular model retraining will ensure compliance with regulatory standards and ethical lending practices.

In an increasingly competitive financial landscape, Evergreen National Bank’s investment in innovative technology and data-driven insights will optimize its credit portfolio and future-proof its operations to meet its customers' evolving needs. By embracing these advancements, the bank solidifies its reputation as a forward-thinking institution capable of navigating the challenges of modern credit risk management.

References

Addo, P., Guégan, D., & Hassani, B. K. (2018). Credit risk analysis using machine and deep learning models. *ERN: Other Econometrics: Econometric & Statistical Methods - Special Topics (Topic)*. https://doi.org/10.2139/ssrn.3155047

Aggarwal, N. (2018). Machine learning, big data and the regulation of consumer credit markets: The case of algorithmic credit scoring. In N. Aggarwal, H. Eidenmüller, L. Enriques, J. Payne, & K. van Zwieten (Eds.), *Autonomous systems and the law*. https://doi.org/10.2139/ssrn.3309244

Alam, M. N., & Ali, M. (2022). Loan default risk prediction using knowledge graph. *2022 14th International Conference on Knowledge and Smart Technology (KST)*. https://doi.org/10.1109/KST53302.2022.9729073

Chen, H. (2022). Prediction and analysis of financial default loan behavior based on machine learning model. *Computational Intelligence and Neuroscience*. https://doi.org/10.1155/2022/7907210

Davis, R., Lo, A., Singh, M., Wu, N., & Zhang, R. (2023). Explainable machine learning models of consumer credit risk. *The Journal of Financial Data Science, 5*(4). https://doi.org/10.3905/jfds.2023.1.141

Egwa, A. A. (2022). Default prediction for loan lenders using machine learning algorithms. *SLU Journal of Science and Technology*. https://doi.org/10.56471/slujst.v5i.222

Ereiz, Z. (2019). Predicting default loans using machine learning (OptiML). *2019 27th Telecommunications Forum (TELFOR)*. https://doi.org/10.1109/TELFOR48224.2019.8971110

Hahami, E., & Piper, D. (2022). A meta-analysis evaluating the performance of machine learning models on probability of loan default. *Journal of Student Research, 11*(2). https://doi.org/10.47611/jsrhs.v11i2.2726

Jumaa, M., Saqib, M., & Attar, A. (2023). Improving credit risk assessment through deep learning-based consumer loan default prediction model. *International Journal of Finance & Banking Studies (2147-4486), 12*(1). https://doi.org/10.20525/ijfbs.v12i1.2579

Khanduja, V., & Juneja, S. (2020). Defaulter prediction for assessment of credit risks using machine learning algorithms. *2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*. https://doi.org/10.1109/ICECA49313.2020.9297590

Moscatelli, M., Parlapiano, F., Narizzano, S., & Viggiano, G. (2020). Corporate default forecasting with machine learning. *Expert Systems with Applications*. https://doi.org/10.1016/j.eswa.2020.113567

Sorayut, & Czech Technical University in Prague. (1999, 2022). *PKDD'99 Discovery Challenge: Loan default prediction and relational database research projects*. Retrieved December 1, 2024, from https://github.com/sorayutmild/loan-default-prediction and https://fit.cvut.cz/cs/veda-a-vyzkum/cemu-se-venujeme/projekty/relational

Suhadolnik, N., Ueyama, J., & Da Silva, S. (2023). Machine learning for enhanced credit risk assessment: An empirical approach. *Journal of Risk and Financial Management, 16*(12), 496. https://doi.org/10.3390/jrfm16120496

Vidovic, L., & Yue, L. (2020, November 1). Machine learning and credit risk modelling. *S&P Global Market Intelligence*. https://www.spglobal.com/marketintelligence/en/documents/machine_learning_and_credit_risk_modelling_november_2020.pdf

Zhou, Y. (2023). Loan default prediction based on machine learning methods. *Proceedings of the 3rd International Conference on Big Data Economy and Information Management (BDEIM)*. https://doi.org/10.4108/eai.2-12-2022.2328740