Revolutionizing Credit Risk: A Deep Dive into Gradient-Boosting Techniques in AI-Driven Finance

Raad Bin Tareaf, Mohammed AbuJarour, Fabian Zinn
XU Exponential University of Applied Sciences
August-Bebel-Str. 26-53, 14482 Potsdam, Germany
Email: {r.bintareaf, m.abujarour, f.Zinn}@xu-university.de

***Abstract*—Assessing credit risk is essential for financial institutions to uphold responsible lending practices. In this study, we conduct an in-depth analysis of three state-of-the-art gradient-boosting algorithms—XGBoost, LightGBM, and CatBoost—for their applicability in credit risk assessment. Utilizing a complex 50 GB dataset with 2.3 million records and 190 features shared by the second-largest credit card issuer globally - *American Express*, we investigate various factors that influence prediction performance. Our research highlights that the size of data chunks plays a significant role in the algorithms' performance, particularly noting that CatBoost performs exceptionally well with larger data segments. The study also emphasis the importance of effectively managing missing data, which has a marked impact on the capabilities of XGBoost and LightGBM. We also examine hyperparameter tuning to identify unique learning characteristics for each algorithm. In conclusion, our findings reinforce financial institutions with advanced analytical tools, enhancing their ability to make informed credit risk assessments.**

Keywords: Credit Risk Assessment, Financial Institutions, Gradient-Boosting Algorithms, Hyperparameter Tuning, Large Dataset Analysis, AI in Banking and Finance.

I. Introduction

The banking industry has witnessed a transformative evolution in recent years, driven by the integration of artificial intelligence (AI) and machine learning techniques [1] [2] [3] [4]. One of the central challenges in this domain is the accurate assessment of credit risk [5] [6] [7], a pivotal task for financial institutions to maintain a healthy and sustainable portfolio. The application of AI methodologies, specifically Gradient Boosting Models (GBMs) [8] [9] [10], has become a notable pathway to improve accuracy and effectiveness in several domains [11] [12] [13] , especially in areas like credit risk assessment.

This research delves into the domain of AI-powered credit risk assessment, with a specific focus on the renowned financial institution, *American Express* [14]. *American Express* a global leader in financial services, provides an ideal dataset for this investigation due to its rich and diverse portfolio of credit transactions [15]. By harnessing the power of GBMs, we aim to shed light on the potential of these models to revolutionize the credit risk assessment process in the banking sector.

The primary objectives of this study are twofold: *first*, to employ state-of-the-art GBMs—widely recognized for their predictive capabilities—in analyzing American Express data to evaluate their effectiveness in identifying and mitigating credit risk. *Second*, to present a case study that not only underscores the practicality of AI-driven credit risk assessment but also explain the complex insights that such models can provide to banking institutions.

As the banking industry continues to adapt to the rapidly changing landscape of financial services [16] [17], the utilization of AI-driven credit risk assessment tools becomes not only relevant but imperative [18]. This research seeks to illuminate the path forward, demonstrating the potential for AI-powered innovations to redefine the way banking institutions approach credit risk assessment, with American Express serving as a compelling case study.

In the following sections, we embark on a comprehensive exploration of the methodologies employed in AI-powered credit risk assessment, with a specific focus on Gradient Boosting Models. We then introduce the American Express dataset, emphasizing its significance and relevance to our research. Subsequently, we delve into the methodology and experimental design used in our case study, followed by the presentation and analysis of results. This study contributes to the growing body of knowledge on the application of AI in the banking sector, offering valuable insights for practitioners and researchers alike.

Section II offers a comprehensive review of gradient-boosting algorithms, focusing on their applications in AI models within the financial and banking sectors. In Section III, we outline the dataset employed for model training, elaborate on the characteristics of the features used, and explain our criteria for selecting the most critical features. This section also covers the configuration of hyperparameters and details our model's architectural setup. Section IV explores the empirical findings, contrasting the effectiveness of different models

concerning data chunk numbers, chunk sizes, imputation of missing values, and the influence of hyperparameter tuning. Section V proposes avenues for future enhancements and directions for advancing the existing implementation. Lastly, Section VI summarizes the study and underscores our key findings.

II. Literature Review

Credit risk assessment is a fundamental aspect of financial services, shaping lending decisions and risk management strategies in the banking sector. In recent years, the integration of machine learning and artificial intelligence (AI) techniques has led to significant advancements in credit risk modeling.

Gradient Boosting Models have emerged as a powerful tool for credit risk assessment due to their ability to handle complex, nonlinear relationships within financial data [9]. These models iteratively combine the predictions of numerous weak learners to construct a robust and accurate ensemble model. XGBoost, LightGBM, and CatBoost are standout implementations in this category, each offering unique advantages.

XGBoost, introduced by Chen and Guestrin (2016) [19], has garnered widespread attention for its speed, scalability, and performance. It employs a gradient boosting framework that optimizes decision trees and has consistently demonstrated competitive results in various machine learning competitions, making it a popular choice in credit risk modeling (Chen et al., 2016 [15]).

LightGBM, developed by Microsoft [20], is another notable GBM variant. It stands out for its efficient histogram-based learning, which reduces memory consumption and accelerates training speed. This makes it particularly suitable for large-scale credit risk datasets where computational efficiency is crucial (Tian et al., 2020 [9]).

CatBoost, introduced by D. Prokhorenkova et al. (2018) [21], specializes in categorical feature handling and boasts strong generalization performance. Its ability to automatically handle categorical variables without extensive preprocessing simplifies the modeling process and can lead to improved results (Prokhorenkova et al., 2018, as cited in Chang et al., 2018 [8]).

Researchers have explored the trade-off between interpretability and accuracy, seeking methods to make GBMs more transparent (Rudin, 2019 [5]). It is a fundamental challenge to strike a balance between predictive performance and the ability to provide explanations, particularly in high-stakes credit decisions.

A common challenge in credit risk assessment is dealing with imbalanced datasets, where the number of non-default cases significantly outweighs the default cases. This class imbalance can lead to biased models favoring the majority class. To address this issue, researchers have employed techniques such as cluster-based under-sampling, which helps rebalance the dataset by reducing the number of non-default samples while preserving the key characteristics of the data (Chang et al., 2018 [8]).

The evolving financial landscape has spurred interest in dynamic credit scoring models that adapt to changing economic conditions and borrower behaviors. These models aim to provide more timely and accurate risk assessments by continuously updating credit scores based on the most recent data (Xia et al., 2021 [10]). Dynamic credit scoring aligns with the demands of the modern financial sector, where real-time decision-making is crucial.

The adoption of AI in credit risk assessment has attracted the attention of regulatory bodies and industry stakeholders. It is imperative for financial institutions to navigate the regulatory landscape effectively while implementing AI-driven credit risk models (He et al., 2017 [14]). Compliance with regulations and adherence to industry best practices are essential considerations in the deployment of advanced credit risk assessment techniques.

III. Methodology & Implementation

A. Dataset Description

The dataset used for our comparative analysis of different Gradient Boosting Decision Trees (GBDTs) was obtained from a Kaggle competition hosted by the credit card company American Express [15]. This publicly available dataset was collected and anonymized based on the credit card statements of multiple hundred thousand American Express users.

The dataset consists of 2,303,433 rows and 190 columns in CSV format, serving as the training data. The unique identifier is a customer ID, with 458,913 unique values.

TABLE I: Dataset Characteristics

Source	(American Express) - Kaggle Competition
Data Type	Structured (CSV)
Rows	2,303,433 records
Columns	190 Features
Unique ID	Customer ID (458,913 unique)
Target	Probability of Default (binary)
Default	Not paying within 120 days
Features	Delinquency, Spend, Payment, Balance, Risk
Categorical	12 features

The target variable, typical for credit risk-related topics, is the probability of default. It is defined as the likelihood of customers not paying back their future credit card balance amount based on their current profile. This target variable

is binary, where 0 denotes that a customer does not default, and 1 indicates that a customer does default. Default events are determined based on an 18-month performance window after the latest credit card statement. If the amount due is not paid by the customer within 120 days after the latest statement date, it is considered a default event. The target variable is provided as a separate label dataset containing the binary target for each of the 458,913 unique customer IDs.

The dataset comprises aggregated features for each customer, aligned with the date of each credit card statement, and these features have undergone anonymization and normalization. Figure 1 illustrates the distribution of features among the overarching categories, which include:

- **Delinquency variables (D)**: These variables capture the previous default behavior of a customer, such as missed payments or defaulted loans in the past.
- **Spend variables (S)**: Spend variables represent customers' spending behavior, including total/average spending, frequencies of large purchases, or categories of spending.
- **Payment variables (P)**: Payment variables describe the payment behavior of a customer, covering frequency of payments and over/underpayments.
- **Balance variables (B)**: Balance variables focus on balance-related data points for a customer, such as outstanding balances, average balances, or the speed at which balances are paid off.
- **Risk variables (R)**: Risk variables represent a calculated risk associated with the customer. They encompass credit scores, risk ratings based on proprietary algorithms, and other indicators summarizing the perceived risk of extending credit to the customer.

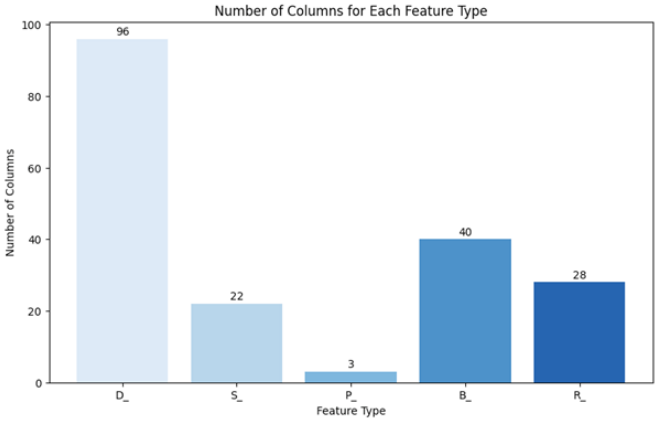

Fig. 1: Feature Distribution Across Categories

The dataset also includes twelve categorical features that require special handling during modeling. The preprocessing section, as well as the detailed model implementations, provide insights into how these categorical features are processed (cat_features = ['S_2','B_30', 'B_38', 'D_114', 'D_116', 'D_117', 'D_120', 'D_126', 'D_63', 'D_64', 'D_66', 'D_68']).

In terms of missing values, the dataset exhibits a proportion of missing values in many of the delinquency variables. This phenomenon is explained by the domain background, as delinquencies do not naturally occur for every customer. The absence of delinquencies also conveys information about a customer's behavior, which is considered during data preparation.

Given the substantial size of the dataset and American Express's status as the second-largest credit card issuer by purchase volume globally [22], this dataset provides an ideal foundation for the comparative analysis of three different GBDT libraries.

B. Data Preprocessing

*1) **Data Ingestion and Label Integration***

This experiment's preprocessing addresses dataset intricacies, including loading, merging, cleansing, encoding, and feature selection. These processes are vital for the GBDT models' success. First, we merge the training data with its labels, since they're stored separately but connected by the 'customer ID'. A left merge using this column achieves this. Next, we derive a validation set, constituting 10% of the merged dataset. Using pandas' "sample" function with a consistent "random_state" ensures this selection is both random and reproducible across experiments.

*2) **Chunking Function and Incremental Learning Strategy***

Despite the dataset's large size, the challenge of managing it is addressed through an iterative learning approach, which involves splitting the dataset into manageable chunks. This approach is employed for all three models and is facilitated by a custom chunking function, 'divide_into_chunks.' This function takes 'df' (the dataset to be chunked) and 'chunk size' (the desired chunk size) as parameters. It calculates the number of chunks required, considering any remaining rows that do not form complete chunks. The function then iterates over the dataset, extracting chunks based on the specified size using pandas' 'iloc' method, ensuring efficient memory usage.

*3) **Features Selection***

A critical aspect is limiting the number of features due to the dataset's size and complexity. An initial test is conducted with the following parameters (Chunk size: 100,000 (24 chunks), One-Hot Encoding for categorical features, Train-test split per chunk: 20% test size, XGBoost Parameters:

Default parameters without hyperparameter tuning, evaluation metric = logloss).

The examination of feature significance, conducted with the 'feature_importances_' function by XGBoost, unveils the leading 20 features ranked by their significance as shwon in Figure 2. This analysis corroborates the hypothesis that an effective model can be constructed using fewer than the 190 available features as shwon in Figure 3. Subsequently, a second test is conducted using only the top 20 features. This focus on the top features enhances model performance, possibly by reducing overfitting, minimizing noise from irrelevant predictors, and efficiently addressing dimensionality concerns.

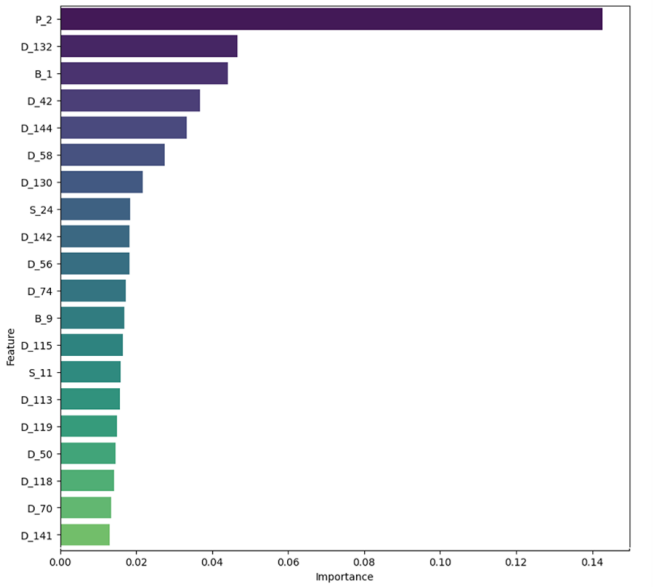

Fig. 2: Top 20 Features Sorted by Significance

CatBoost directly handles categorical features, but XGBoost and LightGBM necessitate preprocessing. We use one-hot encoding, converting categorical attributes into numeric arrays. The OneHotEncoder from scikit-learn aids in transforming the twelve categorical features accordingly.

*4) **Models Implementation***

Models are trained in stages using an iterative approach, starting with an initial data chunk and updating with each subsequent chunk. This method conserves memory and ensures consistent learning throughout the dataset.

Each model has specific hyperparameters for tuning. For instance, LightGBM and XGBoost support L1 regularization, but CatBoost doesn't. Given memory and computational constraints, we used Random Search for hyperparameter optimization. The objective is to identify the 'best_params' for

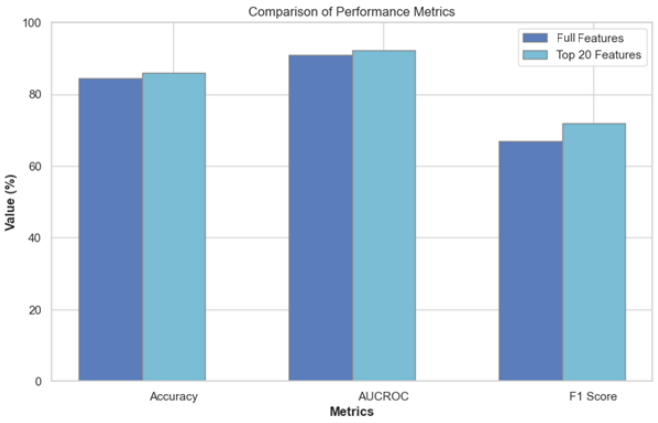

Fig. 3: Performance Evaluation: Top 20 Features vs. Entire Feature Set

model training. Although the models differ in tuning options, key parameters like learning rate, depth, and iterations are consistent across all, ensuring analysis comparability.

IV. Results

A. *Chunk Number Variations Impact*

Varying the number of processed chunks, with a fixed size of 100,000 rows, showed distinct model behaviors. As shown in Figure 4, XGBoost remained stable across all runs. CatBoost peaked when processing 10 chunks, outperforming others, while LightGBM lagged behind but boasted the fastest training times.

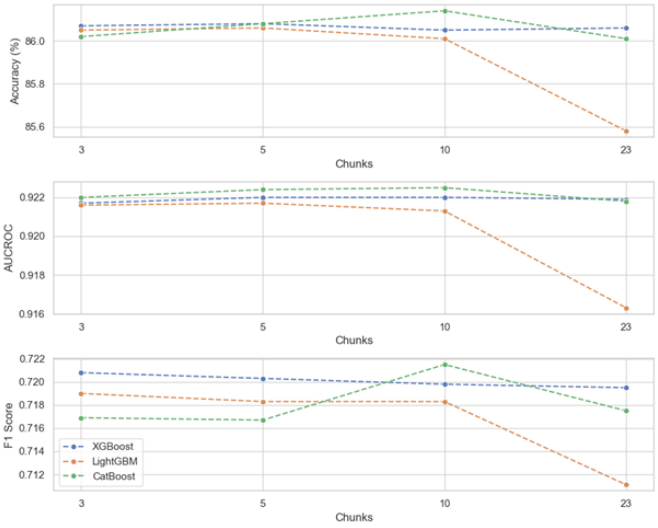

Fig. 4: Results of Varying Chunks Number

B. *Chunk Size Variations Impact*

Adjusting the chunk size in 100,000-row increments enhanced all models, with CatBoost excelling but also consuming significantly more time, as depicted in Figure 5.

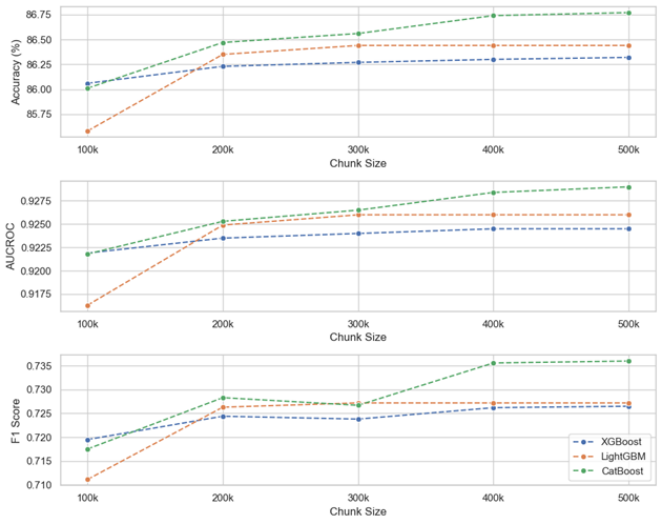

Fig. 5: Results of Varying Chunks Size

C. *Missing Values Influence*

The models, when tested with median imputation for missing numerical values, generally performed better without it. Figure 6 illustrates that XGBoost and LightGBM were notably affected by median imputation, whereas CatBoost remained relatively resilient.

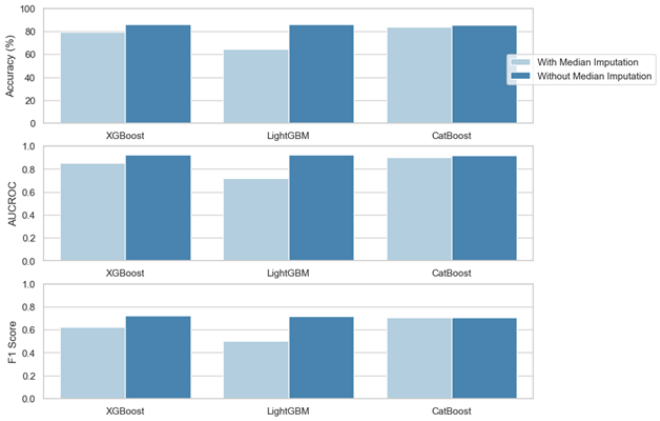

Fig. 6: Median Imputation Impact at Missing Values

D. *Hyperparameter Tuning Effects*

Configurations for hyperparameter tuning revealed that while XGBoost and LightGBM took a more conservative approach, CatBoost was more assertive in correcting its predictions, as evidenced in Figure 7. The analysis also hinted at CatBoost's potential superior overfitting management.

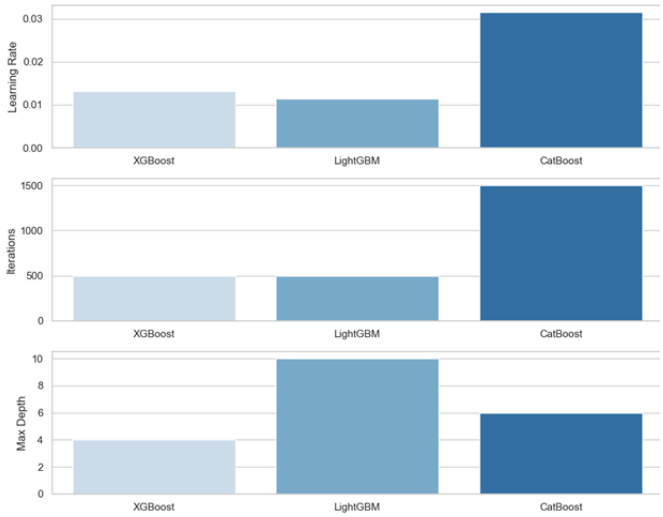

Fig. 7: Comparison of Hyperparameters across XGBoost, LightGBM and CatBoos

V. Future Work

Several promising points for future research emerge from this study. One exciting prospect involves developing hybrid models that harness the strengths of each algorithm, potentially yielding even more robust credit risk prediction systems. Additionally, efforts to optimize CatBoost's computational efficiency when processing larger data chunks could enhance its practicality in real-world applications.

Beyond the technical aspects, the findings of this study underscore the vital role of AI and machine learning in the banking and credit domain. As these algorithms continue to evolve and demonstrate their efficacy, they have the potential to revolutionize how financial institutions assess and mitigate credit risks. Future research should continue to explore innovative ways to integrate these models into the broader landscape of financial services, ultimately benefiting both lenders and borrowers alike.

VI. Conclusion

In conclusion, this study has explored deeply the domain of credit risk prediction, leveraging the capabilities of three advanced gradient boosting algorithms: XGBoost,

LightGBM, and CatBoost. Through hyperparameter optimization and Comprehensive experimentation, we have presented invaluable insights into the behaviors of these algorithms across various scenarios in AI and finance domain.

Our comprehensive analysis of the American Express dataset has served as a robust foundation, allowing us to gain profound insights into the unique characteristics and performance of each algorithm. Notably, CatBoost emerged as a frontrunner, particularly excelling when processing larger data chunks, despite heightened computational requirements.

it's Crucial to recognize the important role these models play in the domain of AI, banking, and credit risk assessment. Their ability to extract actionable insights from complex datasets opens doors to more informed decision-making, ultimately contributing to enhanced financial stability and risk management in the industry. This study serves as a contribution in the domain, and future work in this area promises even greater advancements.

REFERENCES

[1] J. K.-U. Brock and F. von Wangenheim, "Demystifying AI: What Digital Transformation Leaders Can Teach You about Realistic Artificial Intelligence," California Management Review, vol. 61, no. 4, pp. 110–134, 2019. Online. Available: https://doi.org/10.1177/1536504219865226.

[2] D. Paschek, C. T. Luminosu, and A. Draghici, "Automated business process management – in times of digital transformation using machine learning or artificial intelligence," MATEC Web Conf., vol. 121, p. 04007, 2017. Online. Available: https://doi.org/10.1051/matecconf/201712104007.

[3] P. Agarwal, S. Swami, and S. K. Malhotra, "Artificial Intelligence Adoption in the Post COVID-19 New-Normal and Role of Smart Technologies in Transforming Business: a Review," Journal of Science and Technology Policy Management, vol. x, no. x, pp. xx-xx, 2022. [Online]. Available: https://www.emerald.com/insight/content/doi/10.1108/JSTPM-06-2021-0069/full/html.

[4] A. Lui and G. W. Lamb, "Artificial intelligence and augmented intelligence collaboration: regaining trust and confidence in the financial sector," Journal Name, vol. x, no. x, pp. 267-283, 2018. [Online]. Available: https://doi.org/10.1080/13600834.2018.1488659.

[5] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," Nature Machine Intelligence, vol. 1, no. 5, pp. 206-215, 2019.

[6] M. Woźniak, M. Grana, and E. Corchado, "A survey of multiple classifier systems as hybrid systems," Information Fusion, vol. 16, pp. 3-17, 2014.

[7] D.R. Van Deventer, K. Imai, and M. Mesler, "Advanced financial risk management: tools and techniques for integrated credit risk and interest rate risk management," John Wiley Sons, 2013.

[8] Y.-C. Chang, K.-H. Chang, and G.-J. Wu, "Application of eXtreme gradient boosting trees in the construction of credit risk assessment models for financial institutions," Applied Soft Computing, vol. 73, pp. 914-920, 2018, doi: 10.1016/j.asoc.2018.09.029.

[9] Z. Tian, J. Xiao, H. Feng, and Y. Wei, "Credit Risk Assessment based on Gradient Boosting Decision Tree," Procedia Computer Science, vol. 174, pp. 150-160, 2020, doi: 10.1016/j.procs.2020.06.070.

[10] Y. Xia, L. He, Y. Li, Y. Fu, and Y. Xu, "A dynamic credit scoring model based on survival gradient boosting decision tree approach," Technological and Economic Development of Economy, vol. 27, no. 1, pp. 96-119, 2021, doi: 10.3846/tede.2020.13997.

[11] R. B. Tareaf, P. Berger, P. Hennig, and C. Meinel, "Personality exploration system for online social networks: Facebook brands as a use case," in 2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI), Dec. 2018, pp. 301-309.

[12] R. B. Tareaf, S. A. Alhosseini, and C. Meinel, "Facial-Based Personality Prediction Models for Estimating Individuals Private Traits," in 2019 IEEE Intl Conf on Parallel Distributed Processing with Applications, Big Data Cloud Computing, Sustainable Computing Communications, Social Computing Networking (ISPA/BDCloud/SocialCom/SustainCom), Dec. 2019, pp. 1586-1594.

[13] R. B. Tareaf, S. A. Alhosseini, and C. Meinel, "Does Personality Evolve? A Ten-Years Longitudinal Study from Social Media Platforms," in 2020 IEEE Intl Conf on Parallel Distributed Processing with Applications, Big Data Cloud Computing, Sustainable Computing Communications, Social Computing Networking (ISPA/BDCloud/SocialCom/SustainCom), Dec. 2020, pp. 1205-1213.

[14] American Express, "American Express Website," Online. Available: https://www.americanexpress.com/.

[15] A. Howard, A. Amex, D. Xu, H. Vashani, N. Inversion, N. Negin, and S. Dane, "American Express - Default Prediction," Kaggle, 2022, Online. Available: https://kaggle.com/competitions/amex-default-prediction.

[16] L. Hōbe, "The changing landscape of the financial services," International Journal of Trade, Economics and Finance, vol. 6, no. 2, p. 145, 2015

[17] M. D. He et al., "Fintech and financial services: Initial considerations," International Monetary Fund, 2017.

[18] N. Chen, B. Ribeiro, and A. Chen, "Financial credit risk assessment: a recent review," Artificial Intelligence Review, vol. 45, pp. 1-23, 2016.

[19] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 2016, pp. 785-794.

[20] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," in Advances in Neural Information Processing Systems 30 (NIPS 2017), December 2017.

[21] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: Unbiased Boosting with Categorical Features," in Advances in Neural Information Processing Systems 31, 2018.

[22] Luthi, W. B. Wyckoff, and A. Cybulski, "8 biggest U.S. credit card companies this year," US News World Report, May 02, 2023. Accessed: Sep. 20, 2023. Online. Available: https://money.usnews.com/credit-cards/articles/biggest-us-credit-card-companies-this-year.