Understanding Customer Behavior Through Data Analytics in the Banking Sector

Pika Povh Mavrič
University of Primorska
Faculty of Mathematics, Natural Sciences
an Information Technologies
Glagoljaška 8,
6000 Koper, Slovenia
+386 70 718 448
89202036@student.upr.si

***Abstract*— This study explores how data analytics and machine learning are transforming customer behavior analysis in the banking sector. Drawing from recent research, it reviews applications in credit scoring, fraud detection, churn prediction, and product recommendations. Methods such as Random Forest, LightGBM, and neural networks are used to model customer behavior based on transactional and demographic data. Results show that these tools improve decision-making and customer engagement but raise concerns about data quality, model transparency, and ethics. The discussion emphasizes the need for organizational readiness and responsible AI practices. The findings highlight the importance of balancing predictive accuracy with fairness and interpretability in future banking analytics.**

***Key words*—** *Customer Behavior, Banking Analytics, Data Driven Decision Making, Big Data, Credit Scoring, Fraud Detection, Churn Prediction, Product Recommendation Systems*

I. INTRODUCTION

In recent years, the banking industry has experienced a major digital transformation, fueled by the growth of online banking, mobile apps, and data heavy customer interactions. This evolution has allowed banks to gather vast amounts of information about customer behavior, preferences, and financial habits. As we can see in figure 1, it is not only banking, but the whole world is getting more and more information on daily basis. At the same time, advancements in data analytics and machine learning have provided tools to extract actionable insights from this data, reshaping how banks interact with customers and make strategic decisions.

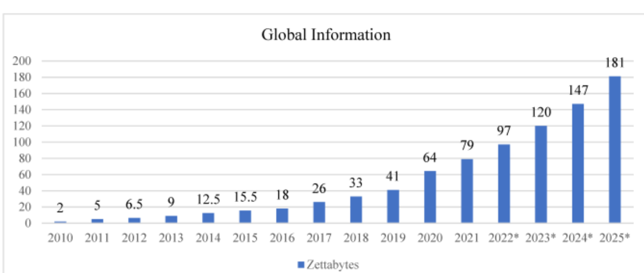

Figure 1: Global amount of information 2010–2025[1]

Figure 3 illustrates that the banking sector leads all other industries in generating revenue from big data and business analytics, making up close to 14% of the global total. Discrete manufacturing follows, contributing approximately 11.3%. Both professional services and process manufacturing each account for about 8.2%. Given the banking industry's leading role, understanding both the benefits and difficulties of leveraging data-driven innovation is essential at every stage of its operations.

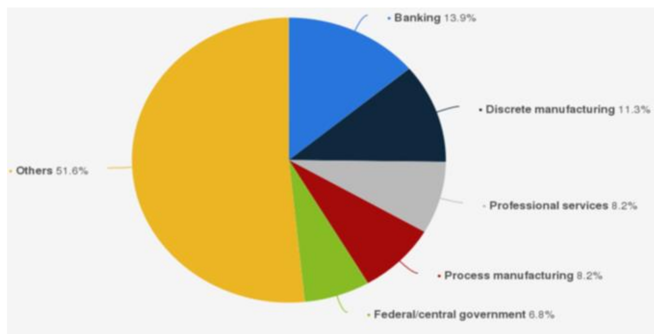

Figure 2: Share of big data and business analytics revenues worldwide in 2019, by industry[2]

As competition in financial services intensifies and customer expectations grow, understanding individual behavior has become crucial for customer retention, risk management, and product innovation. Banks are increasingly adopting predictive analytics not only to anticipate customer needs but also to mitigate risks such as fraud or churn. However, while the potential of these technologies is immense, their successful implementation depends on factors such as data quality, algorithm transparency, organizational culture, and ethical oversight.

This paper explores how data analytics is applied to customer behavior analysis in the banking sector. It begins with a review of existing literature, followed by a detailed look at how customer analytics is used in four key areas: credit scoring, fraud detection, churn prediction, and product recommendation. The discussion section synthesizes the insights and addresses practical and ethical implications. The

[1] Adapted from *Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2024*, by A. Holst, 2021, Statista

[2] Adapted from *Worldwide big data and business analytics (BDA) revenue from 2015 to 2022, by segment*, by Statista, 2021

paper concludes by offering guidelines for future research and implementation strategies.

II. Literature review

The digitization of banking has significantly expanded the scope for understanding customer behavior through data analytics. Researchers have investigated how big data, machine learning, and decision-support systems can improve customer experience and strategic decision-making in banking institutions in different areas.

Sun et al. (2014) introduced the 'iCARE framework', an analytics platform designed to profile customer behavior in banking environments, that is based on big data. By integrating IBM technologies to process structured and unstructured data, the iCARE system enables banks to segment customers, identify high potential users, and deliver better targeted services. In a case study involving a Southeast China bank, the system processed 20 terabytes of transactional data to uncover useful information that helped guide marketing strategies.

Taneja (2024) emphasized that big data analytics is a driving force behind personalized banking services. The study highlights how insights from customer transaction histories and interaction patterns can enable banks to customize their products and services based on each customer's needs, leading to stronger loyalty and greater satisfaction.

In the context of omni-channel banking (that is banking integrated on multiple banking channels like physical branches, mobile apps, web apps, and other platforms), Geng (2016) analyzed customer behavior across all of the possible channels. The study used theoretical frameworks such as transaction cost theory and informedness theory to understand how customer choices evolve in an interconnected banking environment. Transaction cost theory suggests that every economic exchange has associated costs beyond the price of the good or service itself. Informedness theory suggests that better-informed customers make better decisions. The more accurate, timely, and relevant the information they have, the more likely they are to choose options that suit their needs, reduce risks, and increase satisfaction. The findings indicate that banks need to consider where and when customers choose to interact with their services in order to design more effective and user friendly experiences.

Gul et al. (2023) explored the link between data driven decision making (DDDM) and productivity in the Pakistani banking sector. Their empirical study showed that banks implementing DDDM practices experienced a productivity increase of 4–7%. The research calls for a broader use of data analytics to improve performance, particularly in emerging markets where decisions are still largely based on traditional methods.

Pillay and van der Merwe (2021) proposed a theoretical model for big data-driven decision-making in the South African banking context. They emphasized the need for organizational support structures, cultural readiness, and leadership buy in to integrate analytics into strategic processes. Their findings reveal that despite the available technology, the practical implementation of data analytics remains limited due to managerial resistance and insufficient data fluency.

Hasan et al. (2023) conducted a comprehensive review of big data applications in banking operations, identifying opportunities in fraud detection, risk management, and customer profiling. However, they also noted that banks face challenges in managing data complexity, talent shortages, and cybersecurity risks. The study stressed that transforming raw data into strategic assets requires robust infrastructure and governance frameworks.

On the technical front, Dawood et al. (2019) evaluated several machine learning algorithms for customer profiling, including K-means, fuzzy C-means, and neural networks. The results indicated that neural networks outperformed traditional clustering techniques in accuracy and execution time when combining demographic and transactional data.

Buananta et al. (2024) examined e-banking transaction behavior using classification models such as decision trees. With an accuracy rate of 98.61%, the study confirmed that machine learning can significantly improve behavioral predictions and inform digital product design in the banking sector.

Rahman and Vasimalla (2020) focused on customer churn prediction using machine learning models. Their study demonstrated how predictive analytics could help identify at-risk customers early, enabling preemptive engagement strategies to reduce attrition rates.

Collectively, these studies illustrate the transformative impact of big data analytics on customer behavior analysis in banking. While the potential benefits are substantial, successful implementation depends on technological readiness, organizational culture, and ethical data governance.

III. Applications of Customer Behavior Analytics in Banking

A. *Personal loan and credit scoring*

Credit scoring is central to deciding who qualifies for personal loans and serves as a key example of how banks use customer behavior analytics. In the past credit scoring involved predicting a borrower's likelihood of default based on financial history and related attributes. As Li and Zhong (2012) emphasize, this approach originally relied on experts to manually classify applicants based on set criteria, but has changed significantly with the adoption of statistical and artificial intelligence (AI) methods. Linear Discriminant Analysis (LDA), Logistic Regression (LR), and Decision Trees were among the earliest statistical techniques employed to distinguish between "good" and "bad" borrowers. However, limitations such as strict distributional assumptions prompted the development of more flexible methods, such as Multivariate Adaptive Regression Splines (MARS), Support Vector Machines (SVM), and Artificial Neural Networks (ANN), which offer improved performance on nonlinear and high-dimensional datasets.

As Hurley and Adebayo (2016) describe, the industry is undergoing a transformation where traditional models like FICO are supplemented or replaced by algorithmic decision-making systems using extensive data sources—including online activity, social media, geolocation, and shopping habits. These alternative credit-scoring systems aim to extend credit access to “thin-file” consumers, who lack traditional financial histories but may display creditworthy behavior through alternative signals. However, these models raise significant concerns around transparency, data accuracy, and potential discrimination. The concept of “creditworthiness by association,” in which credit decisions are influenced by the behaviors of one’s social or geographic peers, can unintentionally introduce bias into automated decisions. While data analytics and machine learning enhance predictive accuracy and enable broader credit inclusion, they also introduce ethical and regulatory challenges. Addressing these concerns is essential to ensure that personal loan decisions remain fair, explainable, and grounded in individual merit rather than opaque algorithms or correlated proxies.

B. Fraud detection

The rise of online banking, mobile payments, and electronic transactions has significantly increased the risk of fraudulent activities in the banking sector. Traditional rule based fraud detection systems, while working fine in the past, are no longer sufficient to keep up with the advanced methods used by today’s fraudsters. As a result, banks are increasingly turning to machine learning and data analytics to identify, predict, and prevent fraud in real time.

Patil et al. (2018) proposed a comprehensive fraud detection framework integrating big data platforms like Hadoop with analytical tools such as SAS. This system processes large-scale transactional data and applies various machine learning algorithms to classify transactions as either legitimate or fraudulent. Three primary models—Logistic Regression, Decision Trees, and Random Forests—were tested on benchmark datasets. Their results showed that while all models offered reasonable predictive performance, Random Forests outperformed the others with an accuracy of 76%, precision of 93%, and enhanced ability to detect fraudulent transactions in real time. An essential insight from this study is the advantage of using ensemble learning models like Random Forests, which can handle non-linear data better than logistic regression and offer greater interpretability compared to deep learning techniques. Additionally, the use of balancing techniques such as SMOTE and ADASYN was shown to improve model performance by addressing the issue of class imbalance—where fraudulent transactions represent only a small fraction of total banking transactions.

Similarly, Prabha et al. (2024) highlighted that the integration of machine learning and data analytics into banking operations allows institutions to move from reactive to proactive fraud detection. They emphasized the effectiveness of supervised learning algorithms such as neural networks and decision trees, along with anomaly detection techniques like isolation forests and clustering. Furthermore, real time analytics platforms enable immediate responses to suspicious activity, and predictive analytics can forecast future fraud based on historical patterns. Deep learning models, including CNNs and RNNs, are also gaining prominence for their ability to uncover complex fraud patterns across vast datasets.

These studies highlight a major shift in fraud prevention, as banks move beyond simply detecting known fraud patterns. Instead, they are leveraging adaptive, intelligent systems that can evolve with emerging threats. Despite their promise, these approaches come with challenges such as model interpretability, data privacy concerns, and the need for extensive infrastructure and skilled personnel.

C. Churn prediction and retention

Customer churn, the process of clients ending their relationship with a bank, poses a serious risk to profitability in the highly competitive banking industry. Effective churn prediction and proactive retention strategies are therefore critical to maintaining a loyal customer base and sustaining long-term profitability. Customer churn prediction and retention methods and techniques are a big part of every banks data analytics.

Singh et al. (2024) emphasized that churn not only leads to immediate revenue loss but also harms a bank’s reputation and lowers customer trust. Their study applied various machine learning (ML) algorithms, including Random Forest and XGBoost, to bank customer data and developed a comparative framework using a Data Visualization RShiny application. The tool facilitates real-time analysis and visual representation of customer segments at risk of attrition, allowing banks to tailor retention strategies effectively. Their findings underscored the importance of using a combination of demographic, behavioral, and transactional features to improve model accuracy and guide intervention efforts.

Guliyev and Tatoglu (2021) reinforced the importance of explainable ML models in churn analysis. Their work focused on interpreting customer churn predictions using SHAP (SHapley Additive exPlanations) values, which help identify the most influential variables behind attrition decisions. The study found that tree-based models such as XGBoost provided the best classification accuracy while also enabling interpretable outcomes. These insights allowed banks to deploy targeted marketing campaigns and improve service quality for so called at risk customers.

Zoric (2016) demonstrated the successful application of neural networks to predict customer churn in a Croatian bank. Her model revealed that customers using fewer than three banking products were significantly more likely to leave. The research also highlighted the role of customer demographics, income levels, and service usage in predicting churn. Neural networks proved especially useful for handling nonlinear relationships and high-dimensional data, making them a valuable tool in behavioral modeling.

Arefin et al. (2024) explored the potential of combining machine learning and deep learning (DL) techniques to improve churn prediction. Using SMOTE (Synthetic Minority Over-sampling Technique) to address class imbalance, their study achieved the highest prediction accuracy with LightGBM, a gradient boosting framework. They emphasized that model accuracy increased significantly with better data preprocessing and variable selection. Key churn indicators included account balance, age, and estimated salary. Their work also advocated for integrating such predictive frameworks directly into operational workflows for maximum impact.

Bilal Zoric’s research (2016) showed that customers who use more banking services tend to be more loyal. The study’s neural network model identified critical churn predictors such as the number of products used, customer age, and internet banking engagement. These insights were instrumental in advising banks to cross-sell additional services to high-risk customers as a retention strategy.

Together, these studies provide compelling evidence that predictive analytics, especially when explainable and operationalized through interactive tools, can greatly enhance churn management. Banks are increasingly equipped to preemptively identify customers at risk and apply data-driven interventions that foster retention and loyalty.

D. Product recommendation systems

In the digital banking landscape, product recommendation systems have become essential tools for enhancing customer engagement and increasing cross-selling opportunities. These systems leverage machine learning (ML) and customer behavior analytics to tailor financial product suggestions to individual customer profiles, thereby driving revenue growth and improving customer experience.

Oyebode and Orji (2020) developed a hybrid recommender system designed specifically for the banking sector. Unlike traditional collaborative filtering systems common in e-commerce, banking environments lack explicit product ratings, which makes direct user preference modeling difficult. To address this, the authors proposed a model that infers implicit preferences from transaction data using a Product Rating (PR) algorithm. This algorithm assesses product usage frequency and converts it into a rating score, forming a foundation for item-based collaborative filtering. To solve the "cold-start" problem, that is when a customer has no prior interaction history, the system integrates a demographic-based clustering approach using K-means, allowing the prediction of preferences based on customer attributes like age, gender, and profession. The final recommendation is generated through a dynamic weighted combination of collaborative and demographic predictions. This hybrid model, tested on a dataset of over 393,000 customers, outperformed single method systems and proved suitable for real world banking deployments.

Met et al. (2024) explored machine learning-driven recommendation systems for SME banking, highlighting the limitations of traditional processing methods for complex and large-scale financial datasets. Their approach integrates advanced ML algorithms, such as Light Gradient Boosting Machine (LightGBM), to analyze customer transaction data and identify optimal product matches. The study employed extensive exploratory data analysis and robust feature selection (including LOFO, Leave One Feature Out) to refine variables relevant for modeling customer preferences. Their recommendation system, embedded in a real time architecture, was designed for continuous learning and adaptation. The authors also emphasized the use of fog computing to deliver product recommendations closer to the user interface, ensuring faster response times and better system transparency. Importantly, their model included explainability features using SHAP (SHapley Additive exPlanations) values to bridge the gap between algorithmic predictions and human decision making in banking operations.

These studies demonstrate that integrating collaborative filtering, demographic clustering, and explainable machine learning can overcome unique banking constraints, such as data sparsity and lack of explicit user feedback. Product recommendation systems tailored to financial institutions not only boost product uptake but also enhance personalized customer service in both retail and SME segments.

IV. Discussion

The findings across the literature and application areas make it clear that customer behavior analytics is a transformative force within the banking sector. By examining large volumes of structured and unstructured data through machine learning (ML) and statistical methods, banks can now make significantly more accurate and targeted decisions. However, despite the promise of these technologies, their practical implementation comes with both strategic and ethical complexities.

One major theme that emerges is the shift from reactive to proactive banking strategies. Whether in fraud detection or churn prediction, analytics allow banks to forecast behavior rather than merely respond to it. For example, in fraud detection, studies by Patil et al. (2018) and Prabha et al. (2024) show how real-time detection systems can prevent losses by identifying unusual patterns immediately. Similarly, in churn prediction, ML-driven frameworks enable banks to act on early indicators and deploy retention strategies tailored to specific customer profiles (Singh et al., 2024; Guliyev & Tatoglu, 2021).

Another clear trend is the growing emphasis on personalization. The success of product recommendation systems, such as those described by Oyebode and Orji (2020) and Met et al. (2024), shows that banks can deliver highly customized product offerings based on predicted preferences from behavioral and demographic data. This mirrors the broader consumer demand for tailored experiences and signals a new era of relationship based banking.

Also the ethical dimension of customer behavior analytics cannot be ignored. The use of alternative data sources in credit scoring (Hurley & Adebayo, 2016), for instance, raises concerns around bias, fairness, and transparency. As banks increasingly rely on opaque algorithms for critical decisions, there is a risk of reproducing systemic inequities unless interpretability and accountability are embedded in the design of these models.

Lastly, the technical challenge of data quality and sparsity is a recurring theme. Banks often lack explicit customer feedback (e.g., product ratings), making traditional recommendation methods less effective. Hybrid systems that combine behavioral inference with demographic clustering, as proposed in recent studies, offer a promising path forward. Yet, the successful deployment of such systems demands rigorous data preprocessing, ongoing model validation, and explainability measures like SHAP values to ensure trust in the outputs.

V. Conclusion

This paper has explored how data analytics and machine learning techniques are being used to understand customer behavior in the banking sector, with applications ranging from credit scoring and fraud detection to churn prediction and personalized product recommendations. The reviewed studies consistently show that data-driven strategies enhance decision making accuracy, operational efficiency, and customer satisfaction. However, to fully leverage these benefits, banks must address challenges such as data quality, algorithm transparency, and organizational readiness. Ethical considerations, especially around fairness and accountability in automated decisions, must also be prioritized.

For future work, researchers and practitioners should explore real-time analytics and explainable AI solutions that can be directly embedded into customer facing systems. Additionally, further study is needed on how customer behavior analytics can be adapted for underserved populations and smaller financial institutions, particularly in emerging markets. Developing frameworks that combine predictive performance with ethical transparency will be key to shaping responsible innovation in banking analytics.

REFERENCES

Buananta, S. E. A., Christian, L., & Paradise. (2024). Machine learning model of customer behavior on e-banking transaction using classification technique. *Journal of Theoretical and Applied Information Technology, 102*(7), 2990–2999.

Dawood, E. A. E., Elfakhrany, E., & Maghraby, F. A. (2019). Improve profiling bank customer's behavior using machine learning. IEEE Access. https://doi.org/10.1109/ACCESS.2019.2934644

Geng, D. (2016). Data analytics on consumer behavior in omni-channel retail banking, card and payment services. PACIS 2016 Proceedings. https://aisel.aisnet.org/pacis2016/186

Gul, R., Leong, K., Mubashar, A., Al-Faryan, M. A. S., & Sung, A. (2023). The empirical nexus between data-driven decision-making and productivity: Evidence from Pakistan’s banking sector. Cogent Business & Management, 10(1), 2178290. https://doi.org/10.1080/23311975.2023.2178290

Hasan, M., Hoque, A., & Le, T. (2023). Big data-driven banking operations: Opportunities, challenges, and data security perspectives. FinTech, 2(3), 484–509. https://doi.org/10.3390/fintech2030028

Pillay, K., & van der Merwe, A. (2021). Big data driven decision making model: A case of the South African banking sector. South African Computer Journal, 33(2), 55–71. https://doi.org/10.18489/sacj.v33i2.928

Rahman, M., & Vasimalla, K. (2020). Machine learning based customer churn prediction in banking. In 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA) (pp. 639–644). IEEE. https://doi.org/10.1109/ICECA49313.2020.9297529

Sun, N., Morris, J. G., Xu, J., Zhu, X., & Xie, M. (2014). iCARE: A framework for big data-based banking customer analytics. IBM Journal of Research and Development, 58(5/6), 4:1–4:9. https://doi.org/10.1147/JRD.2014.2337118

Taneja, A. (2024). Customer behavior analysis in banking: Leveraging big data to enhance personalized services. International Journal of Innovative Research in Science, Engineering and Technology, 13(7). https://doi.org/10.15680/IJIRSET.2024.1307086

Hurley, M., & Adebayo, J. (2016). Credit scoring in the era of big data. Yale Journal of Law and Technology, 18, 148–216.

Li, X. L., & Zhong, Y. (2012). An overview of personal credit scoring: Techniques and future work. International Journal of Intelligence Science, 2(4), 181–189. https://doi.org/10.4236/ijis.2012.224024

Patil, S., Nemade, V., & Soni, P. K. (2018). Predictive modelling for credit card fraud detection using data analytics. Procedia Computer Science, 132, 385–395. https://doi.org/10.1016/j.procs.2018.05.199

Prabha, M., Sharmin, S., Khatoon, R., Imran, M. A. U., & Mohammad, N. (2024). Combating banking fraud with IT: Integrating machine learning and data analytics. The American Journal of Management and Economics Innovations, 6(7), 39–56. https://doi.org/10.37547/tajmei/Volume06Issue07-04

Arefin, S., Parvez, R., Ahmed, T., Jahin, F., Sumaiya, F., & Ahsan, M. (2024). Enhancing bank consumer retention: A comprehensive analysis of churn prediction using machine-learning and deep learning techniques. In Advances in Information Communication Technology and Computing. https://doi.org/10.1007/978-981-97-6106-7_15

Guliyev, H., & Yerdelen Tatoğlu, F. (2021). Customer churn analysis in banking sector: Evidence from explainable machine learning models. Journal of Applied Microeconometrics, 1(2), 85–99. https://doi.org/10.53753/jame.1.2.03

Singh, P. P., Anik, F. I., Senapati, R., Sinha, A., Sakib, N., & Hossain, E. (2024). Investigating customer churn in banking: A machine learning approach and visualization app for data science and management. Data Science and Management, 7, 7–16. https://doi.org/10.1016/j.dsm.2023.09.002

Zoric, A. B. (2016). Predicting customer churn in banking industry using neural networks. Interdisciplinary Description of Complex Systems, 14(2), 116–124. https://doi.org/10.7906/indecs.14.2.1

Met, I., Erkoc, A., Seker, S. E., Erturk, M. A., & Ulug, B. (2024). Product recommendation system with machine learning algorithms for SME banking. International Journal of Intelligent Systems, 2024, Article ID 5585575. https://doi.org/10.1155/2024/5585575

Oyebode, O., & Orji, R. (2020). A hybrid recommender system for product sales in a banking environment. Journal of Banking and Financial Technology, 4(1), 15–25. https://doi.org/10.1007/s42786-019-00014-w

Holst, A. (2021). *Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2024*. Statista. https://www.statista.com/statistics/871513/worldwide-data-created/

Statista. (2021). *Worldwide big data and business analytics (BDA) revenue from 2015 to 2022, by segment*. https://www.statista.com/statistics/616225/worldwide-big-data-business-analytics-revenue/