10/21/2023

AI bias in Loan Approval Models

Created By: Jonathan Faña

Table of Contents

[Executive Summary 2](#_Toc153017499)

[Introduction 3](#_Toc153017500)

[Problem statement 3](#_Toc153017501)

[Objectives 4](#_Toc153017502)

[Research questions 4](#_Toc153017503)

[Industry Overview 5](#_Toc153017504)

[Statistical Information 5](#_Toc153017505)

[Current Trend in the Industry 5](#_Toc153017506)

[Key Drivers of the Industry 6](#_Toc153017507)

[Key Players in the Industry and why 6](#_Toc153017508)

[Barriers in the Industry 6](#_Toc153017509)

[Literature Review 7](#_Toc153017510)

[Data and Methodology 10](#_Toc153017511)

[Analysis and Findings 12](#_Toc153017512)

[Conclusions and Recommendations 16](#_Toc153017513)

[Limitations 17](#_Toc153017514)

[Implications 17](#_Toc153017515)

[Works Cited List 18](#_Toc153017516)

# Executive Summary

The essay examines the impact of Artificial Intelligence (AI) on the banking industry, particularly in the context of credit approval decisions. The automation of loan approvals using machine learning models has raised concerns over potential data biases that could perpetuate societal discrimination and economic inequality. The essay highlights the need to address these challenges by illustrating instances of bias during the data management process and how it impacts businesses and customers so that actionable recommendations are provided to fight back the problematic.

The selected research philosophy was positivism that, along with a quantitative approach, was found to be the most convenient way to meet the research objectives as well as generalize the results and replicate them.

The dataset was fetched from Kaggle and most of the cleaning was done. It was a fine selection given that it contains most or all key variables that would lead to an accurate prediction of loan approval using, in this case, decision trees. Moreover, it is important to point out how data was intentionally biased during the analysis to ensure the desired out due to regulatory limitations to obtain a real-life sample data.

The analysis uncovered a negative correlation between gender and loan status implying that the latter has an unjustifiable relevant importance in the model's final output. Moreover, the use of input-contribution measurement tool called SHAP revealed that gender is the second most weighted attribute in the model.

Such findings increase the responsibility of BA managers during data management steps and considerations to create alternative criteria of credit-lending that rely less on historical unequal data.

# Introduction

Artificial Intelligence (AI) has profoundly reshaped our lives over the past few decades, with innovations ranging from music discovery platforms to automated vacuum cleaners. However, it's in the business realm, particularly in the banking industry, where AI has sparked a productivity revolution. The ingenious and practical application of AI has led to significant enhancements in cybersecurity, fraud detection, risk management, and more. Yet, among these advancements, credit approval decisions have emerged as a contentious issue.

The automation of loan approvals, facilitated by the use of historical data and machine learning models, has ignited debates over potential data biases that could perpetuate longstanding societal discrimination and economic inequality. This is not a scenario that lenders intentionally create; rather, it stems from years of historical disparities among social groups that are imprinted in the data and subsequently reflected in the model outputs used for decision-making. The presence of such biases has elicited numerous ethical concerns due to their role in exacerbating socio-economic imbalances by restricting credit quality and access for the affected groups.

## Problem statement

The integration of AI to streamline loan approval decisions is undeniably here to stay. Consequently, addressing associated challenges has become a top priority for organizations. From an outsider's perspective, it might be tempting to attribute such behavior solely to banks (which is partially accurate). However, the real hurdle lies in the historically biased data fed into machine learning algorithms for predictions. Therefore, it's crucial to illustrate what transpires behind the scenes and identify which factors during the data management phase contribute to these undesirable outcomes.

## Objectives

* To uncover and illustrate instances of bias within the data used to develop AI-driven loan eligibility criteria.
* To investigate how AI algorithm biases impact businesses and their customers.
* To offer actionable recommendations for banks to proactively address bias in AI systems within the competitive financial market.

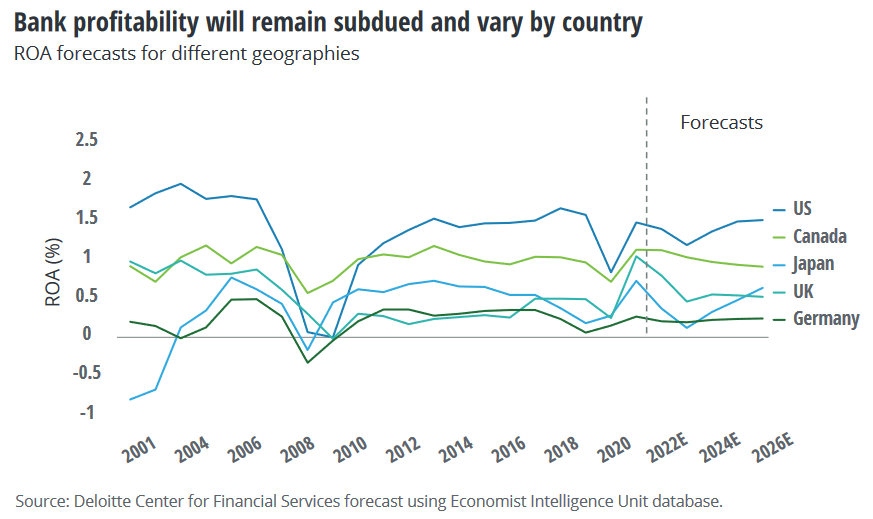
## Research questions

* What are the discernible instances of bias within the data used for forming AI-driven loan eligibility criteria?
* How does bias within AI algorithms affect businesses and their customer base?
* What actionable recommendations can be provided to banks for mitigating bias in AI systems within the competitive financial market?

# Industry Overview

## Statistical Information

The banking industry is in a much more robust stand compared to its state post the 2008 financial crisis. In 2022, the total worldwide assets increased to \$154,211, marking a 3.79% year-over-year growth from \$148,583 in 2021, as reported by The Banker's Top 1000 World Banks Ranking for 2022 (Meola). Projections are that banks will be back to profitability sooner than expected, a short downtrend in 2023 (Deloitte):



## Current Trend in the Industry

The advent of technology has caused a rapid shift towards online and mobile banking, as well as the emergence of digital-only banks. Additionally, there is a growing emphasis on sustainable and responsible banking practices, which prioritize environmentally friendly methods and ethical investments. Banks are now turning to AI and ML for various tasks, such as fraud detection, risk assessment, and improving customer experiences.

## Key Drivers of the Industry

There are currently 10 business drivers in the banking industry. Yet, find below what, for this report, are consider the most important 4: Data & Analytics: Disruptive technologies, such as blockchain, are driving change in the industry; Customer experience: On-demand, customized products and service to akin to Amazon; Regulation: Uncertainty surrounding a complex and inter-connected global issue; Digital transformation: Digital labor, artificial and cognitive intelligence, platform models, payments, talent, cloud, API / fintech …and, more (KPMG LLP)

## Key Players in the Industry and why

JP Morgan Chase, an American multinational bank, is distinguished by its extensive assets, worldwide influence, and broad spectrum of services. The Industrial and Commercial Bank of China (ICBC), recognized as the world’s biggest bank in terms of total assets, holds a commanding position in the Asia-Pacific market. HSBC, having a significant footprint in both Europe and Asia, is instrumental in the global banking and finance sector.

## Barriers in the Industry

Some of the entry barriers that exist in the industry are the substantial capital investment as well as the compliance with numerous regulations and a time-consuming process to build reputation and trust. Furthermore, leaving the industry also has its challenges like a costly process associated with breaking long-term arrangements with customers, a flagrant reputational impact if the circumstances are adverse by the time of departing and regulatory restrictions that turn leaving some financials markets a extremely difficult task.

# Literature Review

The ethical challenges of AI in decision-making processes, particularly in the domain of loan approvals, present several problems that are yet subject to discussion. These issues are highly contextualized in the context of banks' loan approval processes. The issues include data protection policy, fairness, potential bias, and transparency (Rizinski et al.). Mortgage lending has exhibited bias against Latin and African American lenders, with minority groups being charged significantly higher interest rates and rejected more often than their privileged counterparts in the States (Singh). These biases lead to discriminatory loan classifiers and can cause more significant housing, wealth, and property gaps between unprivileged and privileged groups. Due to this ubiquity, concerns are starting to arise about whether the development of AI systems, and the decisions made by them, should be based on a set of ethical principles to promote transparency, social equity, sustainability, and avoid social injustices. In this regard the concept of fairness is defined. There are contexts where understanding motivations leading to a specific result is more important than the result itself, and it is crucial to be able to understand the reasons why a prediction was made to build trust in the decisions taken by a model (Purificato).

It is well-known the importance of fair credit access in our modern economy so that the financial impact these algorithms have on credit borrowers is unarguable. A report from The Markup revealed that people belonging to a social minority were more prone to be refused a home loan than Caucasians with same financial background. This has resulted in 2.25 trillion out of 13 trillion dollars of unpaid loads in the US linked to those minority groups just in 2017 (Hale). Moreover, on an article published by the Markup a comparison was made between people of color with an income over $100,000 and White borrowers with earning below the same mark and results were stunning. Even though minority applicants had greater earnings and better risk categories like “healthy” or “manageable”, their counterparts got their credit approved more frequently by lenders. It evidences huge differences in approval rates (Martinez and Kirchner).

The impact of using machine learning in the financial industry may have caused challenges in loan approval processes. The challenges are driven by biased behavior from the data collection and algorithm interaction. The sources of bias could be from historical bias and representation bias. Historical bias happens when a model just uses perfectly sampled and measured data, and this does not represent the reality accurately. Representation bias occurs when there is generalization of certain population. The use of certain population as data does not reflect on other types of groups especially the underrepresented group and the sample classification is uneven. These biases could cause harm to the model outcome (Suresh and Guttag). Other bias may also be caused by deep learning where machine learning is created to mimic the human brain. Deep learning with artificial neural networks can exhibit biases due to the incorporation of human biases from training data, the influence of historical criteria (stability bias), and limitations in the data, which can result in incomplete or biased representations of the real world (Baer and Kamalnath).

There are few approaches that can be taken to mitigate the bias of loan approval processes. First would addressing and assessing the potential source of the bias such as data, metric and measures, algorithm, and models (Fu et al.). Data bias is the result of data collection errors, sampling selection problems, and historical human biases. Transparent data collection processes could be done by random sampling like Lagrange multiplier test and Heckman’s test. It is hard to detect and correct historical human bias, developers need to have deep knowledge of credit and lending. Regression analysis can be used in fair lending assessments when there is substantial overlap between mortgage data and other customer data. This analysis complements direct outcome evaluation in evaluating fairness. The traditional metrics like adverse impact ratio (AIR), marginal effect and standardized mean difference (SMD) can be explored more by leveraging other biased measures. Leverage includes confusion matrix and AI interpretability measures (Brotcke; Lorenzo). Algorithm biases can be identified by using 80% rule where the selection rate of a certain group cannot be lower than 80% of selection rate of other or regular group (Fu et al.). Algorithmic fairness can be evaluated more with defined and clear fairness criteria and different metrics can be used to evaluate the machine learning model objectively (Lorenzo).

# Data and Methodology

This research was conducted to illustrate how human bias influences machine learning algorithms and the impact caused to the involved demographics as a result of unfair lending practices in the banking industry. With the recommendations generated in this report, the aim was to capture the trending methods lenders are employing for a more equal lending. In the next paragraphs, expect to see the methodology employed to tackle this objective and the data utilized during the research.

The research philosophy selected to carry out the research was positivism. One of the reasons for this choice is that biased-lending practices were pointed to be an industry issue by the time it was uncover back in 2021. As a result, the team was able to replicate the instances of data bias that prompt machine learning models to segregate certain ethnic groups. On the other hand, taking a phenomenological approach would imply that us, the researchers, had any access to specific intel from a banking institution that has been proven to perpetuate these practices. The chosen overview for this project led us to take a quantitative method to answer the research questions and meet the goals from a positivistic view. It is so, as the selected method allows for generalizable results as intended with the defined research goals, and, unlike with qualitative, it will be easier for anyone to replicate the findings.

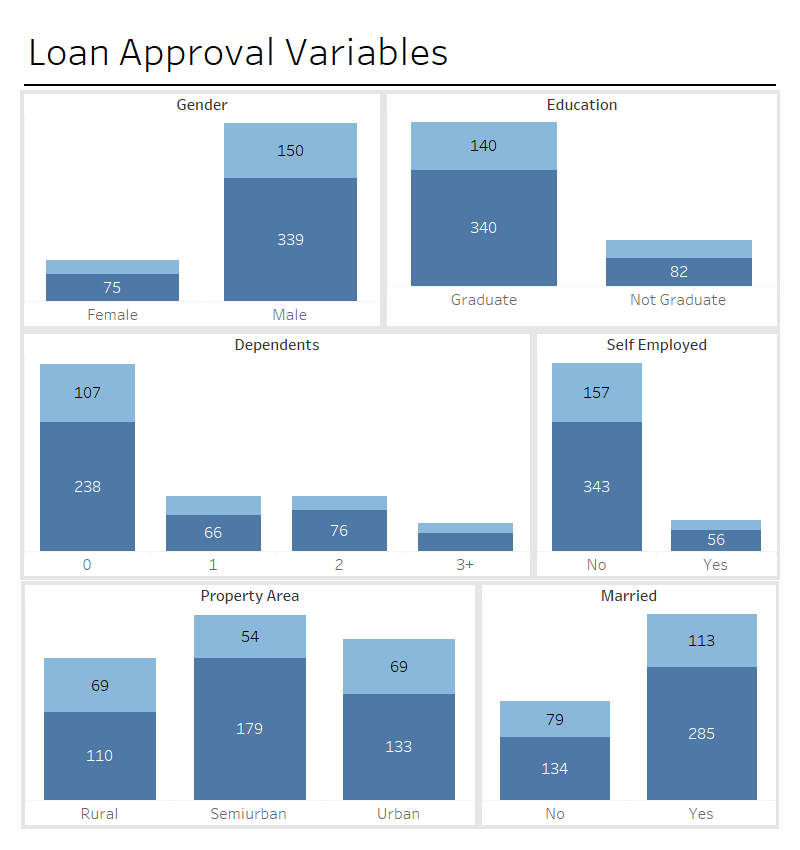
To accomplish it, a secondary quantitative research method will be held using a dataset named “[Loan Approval Dataset](https://www.kaggle.com/datasets/granjithkumar/loan-approval-data-set?source=post_page-----892ee3cb59f7--------------------------------)” obtained from Kaggle. Using this data was convenient given that it has the 13 main columns usually attributed to a robust loan prediction dataset, plus it contains over 614 hundred records to feed the model with. Moreover, the dataset was prepped and cleaned before upload meaning that no missing values or outliers were treated of removed respectively. Below is a brief summary of the existing variables:

|  |  |  |  |
| --- | --- | --- | --- |
| NO | VARIABLE | DATA TYPE | LISTED AS |
| 1. | Loan ID | Categorical | Unique loan id e.g. LP001002 |
| 2. | Gender | Categorical | Male and female |
| 3. | Married | Categorical | Yes/No |
| 4. | Dependents | Numerical | Number of dependents to applicants |
| 5. | Education | Categorical | Graduate / Not graduate |
| 6. | Self-employed | Categorical | Yes /No |
| 7. | Applicants Income | Numerical | Amount the applicant earns as income |
| 8. | Co-applicant income | Numerical | Amount the co-applicant earns as income |
| 9. | Loan amount | Numerical | Amount applied for In thousands |
| 10. | Loan amount term | Numerical | Repayment period of the loan in months |
| 11. | Credit history | Numerical | Credit history of the applicant |
| 12. | Property area | Categorical | Urban/semi-urban/rural |
| 13. | Loan status | Categorical | Yes /No |

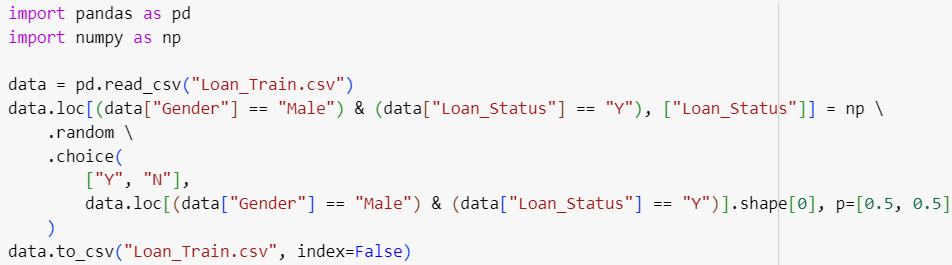
Among the current programing languages in the market, Python was selected due to its flexibility with libraries like pandas, numpy, seaborn, sklearn and matplotlib.pyplot because of their broadly-spread popularity and beginners-friendly code syntax. Lastly, once data preparation is done a random forest classifier will be run and measured for approval prediction. The weight of each variable in the final output of the model will be gauged by using a tool called SHapley Additive exPlanations (SHAP) which will graphically show the interaction of each attribute with the results.

It's important to restate that data collection for this particular purpose is difficult for industry outsiders. Hence, during the data manipulation phase of the analysis, some modifications were implemented to illustrate what has been confirm to be an industry issue.

# Analysis and Findings

In this report, the data bias would be illustrated and analysed by using Loan\_Train.csv. The dataset consists of several criteria with its loan approval status. In this case, the dataset would be altered by changing the loan status of data of male gender with approved status to either being approved and rejected randomly. These changes simulated bias intentionally by reducing the approval probability of male customers. But first, a quick look at its attributes: 

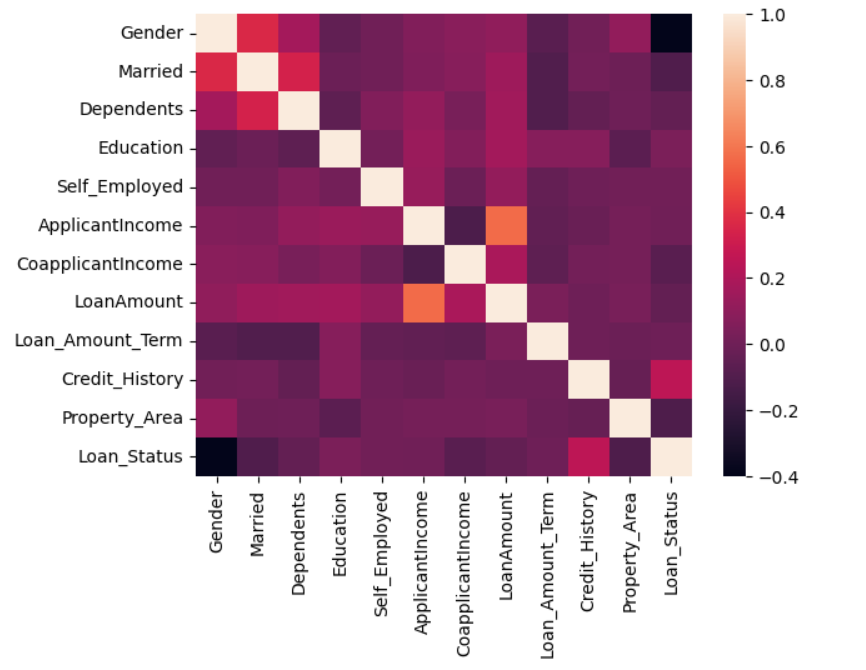
Below is the code for altering the loan approval status:



The dataset then analysed the correlation of all variables with the loan status by using heatmap. It is visible that there is a significant negative correlation of gender with loan status after the alteration.

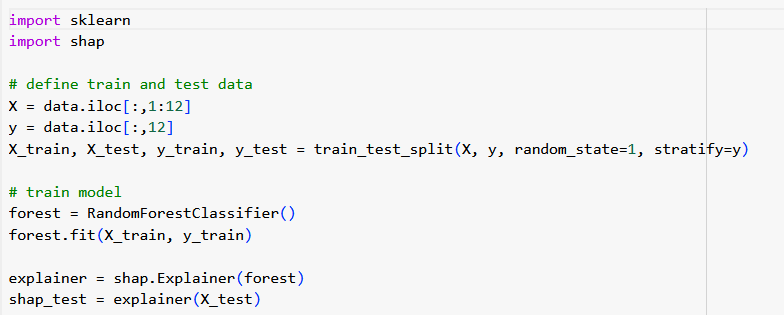






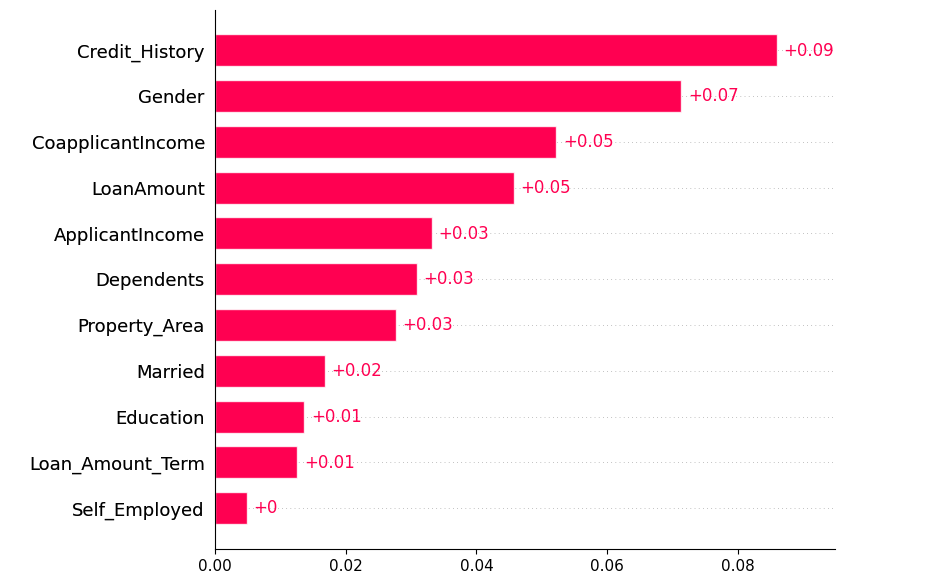
In this scenario, the negative correlation indicates that there could be a tendency of negative impact on gender. The bias here was done to show that male customers could get rejected when they were supposed to be accepted originally following the original model or criteria. The illustration shows that reduce rate of approval could happen due to the bias training dataset and bias model.

There are some model explain-ability methods that can be used to detect bias in simple way. One of the examples is SHapley Additive exPlanations (SHAP). SHAP provide a measurement of each input contribution towards final output.



Syntax above shows how two more libraries are import in order to split the dataset and train the model so SHAP tool can weight the inputs contribution to the model. Model is not tested as its score is not relevant for this part of the analysis.





The bar chart reveals that even though variables like credit history, loan amount and co-borrower income are important, gender plays a crucial role in the model output which denotes a rooted discriminatory factor within data with more importance it should when deciding credit worthiness. For more info on how SHAP works and different bias illustration technics with using it refer to [Jan Bieser’s article](https://medium.com/@janwithb/unmasking-bias-a-practical-example-of-bias-detection-for-a-loan-approval-ai-model-892ee3cb59f7) on the matter.

# Conclusions and Recommendations

During analysis, the generated heatmap revealed a significant negative correlation for males post-alteration, indicating gender bias. The significance of this negative correlation confirms the presence of bias. The strong negative correlation indicates that the model, when trained on this biased data, would likely discriminate against male applicants. It also shows the impact of biased data: The experiment highlights how AI models can inherit and amplify biases if the training data is not carefully vetted and cleaned. AI systems learn to make decisions based on patterns found in their training datasets, so if those datasets are biased, the resulting AI decisions will likely be biased and the subsequent analysis that uncovered a negative correlation between gender and loan approval status serve as a proof of concept. Hence, it is crucial that lenders refine their data preparations process by performing regular audits that strengthen bias detection.

Furthermore, the investigation into how AI algorithm biases impact businesses and customers was illuminated by the use of the SHapley Additive exPlanations (SHAP) tool. SHAP analyses demonstrated the disproportionate impact of gender on loan approval decisions, which not only reflects prejudice bias but also showcases the potential of such biases to impact the fairness and integrity of financial decision-making processes. Therefore, businesses can benefit from the implementation of model-agnostic tools like SHAP or any other more suitable for their models a long with continual reassessment and adjustments of AI algorithms.

The research provides a strong foundation for financial institutions to refine their AI systems. This not only enhances the robustness and fairness of the loan approval process but also contributes to a more equitable financial environment for all customers.

# Limitations

Privacy regulations regarding data handling makes it difficult to gather real-world data due to the ethical and legal issues that surround the matter.

The different tools to highlight possible bias in the data varies depending on the model used to predict approval. In this case, SHAP was used to identify feature importance and contributions in the decision tree. However, it may not be able to comprehensively identify all forms of bias. It’s important to combine SHAP with other bias detection techniques, human expertise, and domain knowledge to gain a holistic understanding of bias in AI system

# Implications

The findings demonstrated how demographic variables can influence the final output of the model, leading to a potentially systematic segregation of classes in terms of credit approval. At the same time, the discovery aligns with historical claims of perpetuation of unequal credit-lending practices, enlarging the social and economic gaps between groups of different demographics as cited in previous works.

All this leaves business analytics managers in a position where new methods of data manipulation need to be tested, ensuring biased decision-making is not sustained over the years. This implies a higher investment in investigating new alternatives or criteria of lending as these practices, intentional or not, cause reputational damage to any company suspected of such practices, which increases the pressure on their shoulders. Furthermore, approval criteria based on historical data have been reliable for the past decades, and discarding it could point to an interest rate increase that will impact the revenue of the companies directly and hence soar the stakes at risk. As a result, it is crucial to design a new industry standard that serves as a replacement for the current one.

# Works Cited List

Deloitte. *2023 banking and capital markets outlook*. December 2022. PDF. <https://www2.deloitte.com/cn/en/pages/financial-services/articles/financial-services-industry-outlooks-2023/banking-industry-outlook.html>

KPMG LLP. *Top 10 of Change in Banking*. 2018. PDF. <https://assets.kpmg.com/content/dam/kpmg/us/pdf/2018/09/top-10-drivers-of-change-banking.pdf>

Baer, Tobias, and Vishnu Kamalnath. “Controlling Machine-Learning Algorithms and Their Biases.” *Mckinsey.com*, McKinsey & Company, 10 Nov. 2017, <https://www.mckinsey.com/capabilities/risk-and-resilience/our-insights/controlling-machine-learning-algorithms-and-their-biases>

Brotcke, Liming. “Time to Assess Bias in Machine Learning Models for Credit Decisions.” *Journal of Risk and Financial Management*, vol. 15, no. 4, 2022, p. 165, doi:10.3390/jrfm15040165.

Fu, Runshan, et al. “AI and Algorithmic Bias: Source, Detection, Mitigation and Implications.” SSRN Electronic Journal, 2020, doi:10.2139/ssrn.3681517.

Hale, Kori. "A.I. Bias Caused 80% Of Black Mortgage Applicants To Be Denied." Forbes Sep. 2021. <https://www.forbes.com/sites/korihale/2021/09/02/ai-bias-caused-80-of-black-mortgage-applicants-to-be-denied/?sh=21cb39836feb>

Lorenzo, Flavio. *Techniques for trustworthy artificial intelligence systems in the context of a loan approval process*. Politecnico di Torino, 18 Dec. 2019.

Martinez, Emmanuel and Lauren Kirchner. "How We Investigated Racial Disparities in Federal Mortgage Data." The Markup Aug. 2021. <https://themarkup.org/show-your-work/2021/08/25/how-we-investigated-racial-disparities-in-federal-mortgage-data>

Purificato, E., Lorenzo, F., Fallucchi, F., & De Luca, E. W. (2023). The Use of Responsible Artificial Intelligence Techniques in the Context of Loan Approval Processes. International Journal of Human-Computer Interaction, 39(7), 1543–1562. <https://doi.org/10.1080/10447318.2022.2081284>

Rizinski, Maryan, et al. “Ethically Responsible Machine Learning in Fintech.” IEEE Access: Practical Innovations, Open Solutions, vol. 10, 2022, pp. 97531–97554, doi:10.1109/access.2022.3202889.

Singh, A., Singh, J., Khan, A., & Gupta, A. (2022). Developing a Novel Fair-Loan Classifier through a Multi-Sensitive Debiasing Pipeline: DualFair. Machine Learning and Knowledge Extraction, 4(1), 240. [https://doi.org](https://doi.org/10.3390/make4010011) /10.3390/make4010011

Suresh, Harini, and John Guttag. “A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle.” Equity and Access in Algorithms, Mechanisms, and Optimization, ACM, 2021.

Bieser, Jan. “Unmasking Bias: A Practical Example of Bias Detection for a Loan Approval AI Model.” The Medium Aug 2021. https://medium.com/@janwithb/unmasking-bias-a-practical-example-of-bias-detection-for-a-loan-approval-ai-model-892ee3cb59f7.