# Abstract

This dissertation explores the growing role of alternative credit scoring systems and their impact on lending practices within the fintech loan application landscape. By utilizing advanced machine learning techniques, such as Logit and Probit Regression (LR), Support Vector Machine (SVM), and Deep Neural Networks (DNNs), this research aims to replicate lender loan acceptance decisions and predict default risks. The proposed two-phase model differentiates between predicting loan rejections in the first phase and assessing default risks for approved loans in the second phase, providing insights into customer churn.

A holistic analysis of the credit scoring model of Paytm, a prominent lender, reveals the significant influence of alternative data, particularly the quality of digital transactions, on loan outcomes. This raises concerns about the equitable implementation of opaque models that reshape risk evaluation. Expanding the investigation, the study examines fintech-bank partnerships and evaluates the capital allocation of banks participating in instant loan platforms. This assessment highlights the growing focus on loans facilitated through collaborations between fintech start-ups and banks, indicating a notable shift in the allocation of bank balance sheets.

The research demonstrates that alternative credit scoring systems play a crucial role in the calculative infrastructure, enabling specific institutions to overcome obstacles associated with risk-based pricing. These systems also serve as strategic collaboration points between technology start-ups and financial entities, capitalizing on new revenue streams.

Furthermore, the study focuses on loans utilized for personal and entrepreneurial purposes, revealing an interesting dichotomy in model performance. The first phase demonstrates superior performance when trained on the entire dataset, while the second phase excels when focused on the small business subset. This discovery emphasizes potential variations in the screening and default prediction processes for small business loans.

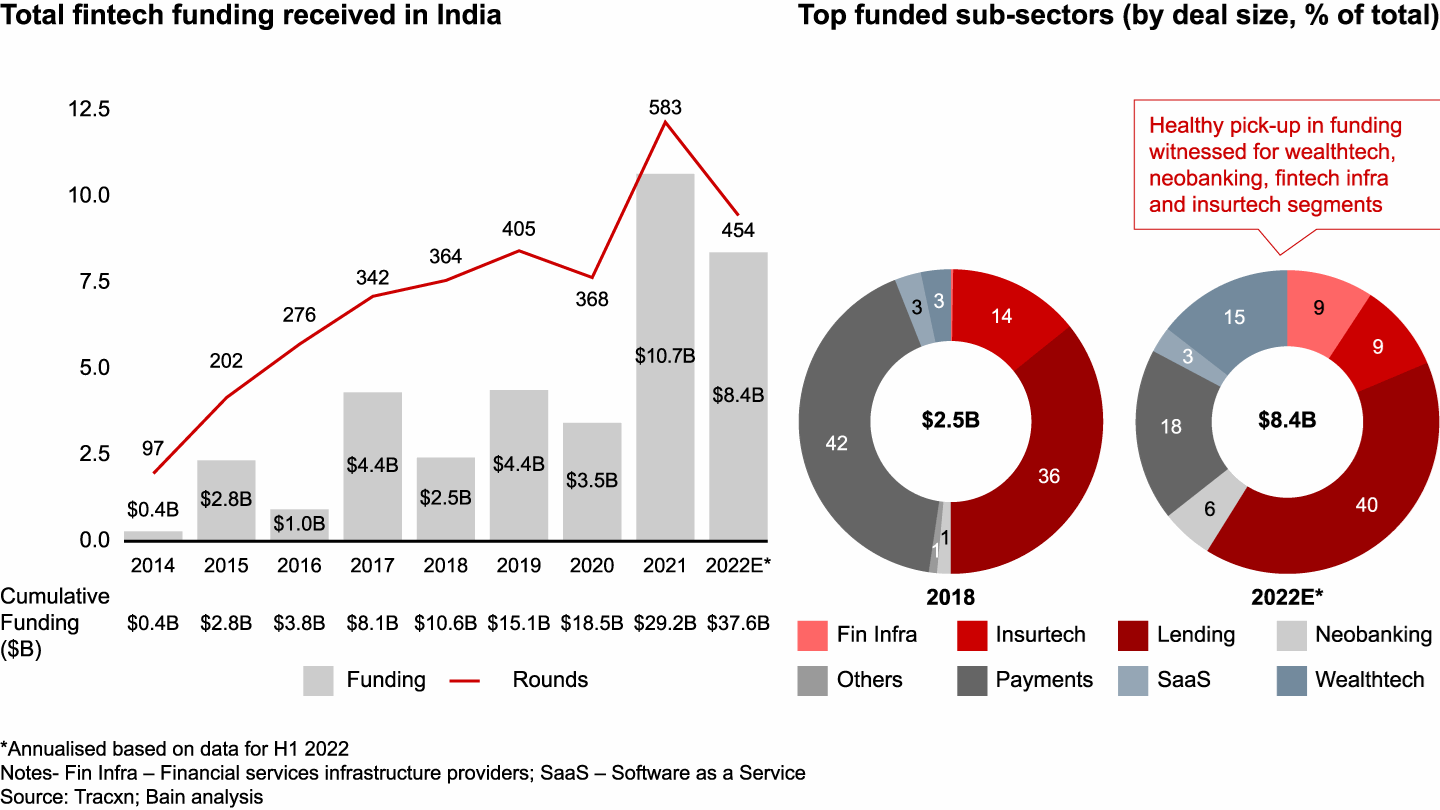
This research establishes the potential of machine learning algorithms in refining credit risk assessment, enhancing informed lending decisions, and effectively forecasting loan defaults. By harnessing artificial intelligence, financial institutions can improve risk management frameworks and strengthen the soundness of lending practices.

# Index

|  |  |  |
| --- | --- | --- |
| **SL. NO.** | **TOPIC** | **PAGE NO.** |
| 1 | **Chapter 1 – Introduction** | 1-5 |
| 2 | **Chapter 2 – Organizational Profile** | 6-11 |
| 3 | **Chapter 3 – Review of Literature and Research Design** | 12-16 |
| 4 | **Chapter 4 Data Analysis and Interpretation** | 17-22 |
| 5 | **Chapter 5 – Summary of Findings, Conclusion and Suggestions** | 23-28 |
| 6 | **Bibliography and Appendices**  **25.3.1 References**  **25.3.2 Questionnaire**  **25.3.3 Synopsis Registration Form**  **25.3.4 Progress Report**  **25.3.5 Plagiarism Report**  **25.3.6 Any other relevant documents** | 29-34 |

# Chapter 1 Introduction

In today's fast-changing world, lending and borrowing money has become more intricate than ever. Especially in places like Bangalore, where technology is booming and people have diverse financial needs, figuring out who might not pay back their loans on time is a big challenge. The swift strides made in the expanses of machine learning (ML) and artificial intelligence (AI) have procreated the development of alternative credit scoring systems, fundamentally transforming how lenders gauge risk and make pivotal investment choices within the sphere of consumer debt. These novel models, leveraging unconventional data sources and sophisticated processing methods, herald a crucial departure from the norms of risk-based pricing - a cornerstone for financial institutions. Risk-based pricing establishes interest rates based on the anticipated odds of loan defaults. In this context, the rise of alternative credit scoring, supported by algorithmic models, surpasses prevailing obstacles and ushers in a comprehensive re-evaluation of risk assessment.



**Growth of Indian Fintech Industry**

**Source: Bain & Company** [**[1]**](https://www.bain.com/insights/india-fintech-report-2022-sailing-through-turbulent-tides/)

India is leading the FinTech adoption race with a rate of 87%, which is significantly higher than the global average of 64%. The Indian FinTech market is expected to reach $1 trillion in AUM and $200 billion in revenue by 2030, according to a study conducted by EY. The Indian FinTech ecosystem is one of the largest in the world, with over 2,100 FinTech companies, and it is growing rapidly. The growth of the Indian FinTech industry can be attributed to various factors, such as government initiatives, a thriving funding environment, a thriving VC ecosystem, high FinTech adoption, and access to talent and technology. The Indian FinTech industry is expected to grow at a CAGR of 20% to reach a transaction value of $138 billion in 2023.

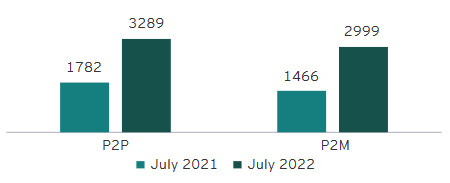
In the world of consumer finance, when individuals apply for credit cards or auto loans, lending institutions traditionally rely on credit scores as a crucial factor in determining approvals. Credit scoring involves a statistical assessment conducted by financial institutions or credit bureaus to evaluate a borrower's creditworthiness. This assessment takes into account various factors such as credit history, demographic information, and credit behaviour. Generally, borrowers with higher credit scores have a better chance of securing loans at lower interest rates. The importance of credit scoring for financial institutions goes beyond assessing the risk of default; even slight improvements in credit scoring mechanisms can lead to significant financial gains, creating the better means to safer lending.

This dissertation is all about finding a smarter way to predict if someone might not be able to pay back the money they borrowed, with special references to Bangalore geography. The research looks at two important things: how people decide to stop borrowing money (we call this "customer churn"), and how they behave with money they've borrowed (which we call "credit behaviour"). These things can tell us a lot about whether they'll be able to pay back their loans.

People around the world often turn to banks when they need loans to overcome financial constraints and achieve personal goals. Loans are essential due to the constantly changing economy and intense competition in the financial sector. For banks, lending is a major source of income and financial risk. A significant portion of a bank's assets comes from the interest earned on loans. While lending is beneficial for both borrowers and banks, it comes with risks, particularly the risk of borrowers failing to repay the loan, known as 'Credit Risk' . Therefore, assessing a client's creditworthiness before approving a loan is crucial. Traditional lending processes typically rely on the '5C principle': Character, Capital, Capacity, Collateral, and Conditions. This assessment often depends on personal experience and customer knowledge, which has limitations. Even after a rigorous verification and validation process, there is no guarantee that the selected applicant will repay the loan on time.

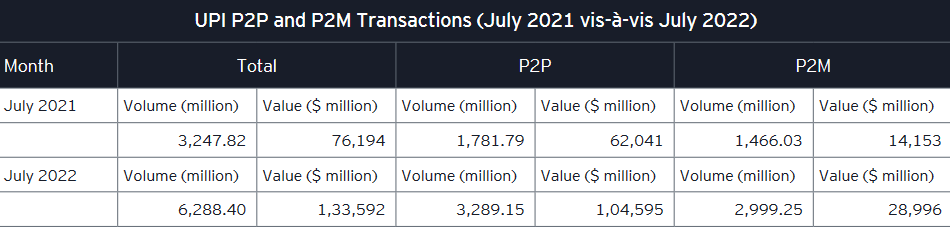
In the past, banks employed experts to evaluate applicants' creditworthiness and make loan approval decisions based on a numerical score called 'Credit Score.' This score assessed the probability of borrowers repaying loans based on their credit and payment history, as well as their background. Credit scoring required expertise and statistical algorithms to predict applicants' creditworthiness accurately.

The Unified Payments Interface (UPI) has emerged as one of the most popular payment methods in India, supporting Peer-to-peer (P2M) and Merchant payment (P2M) services, which can be used over the smartphone (app-based), feature-based, and at the merchant location (app-based). P2P transactions formed 78% of the total transaction value in July 2022 with $104,595 million, recording a 94% and 75% value growth since the previous year (July 2021). Most of the P2P transactions had a lot size greater than $26 (INR 2,000). As per the Worldpay report, India led the world in the largest daily real-time payments volume in 2021 with 70.2 million, followed by China with 42.8 million. UPI is now the leading form of retail merchant payments (Person-to-Merchant - P2M disguised as Peer-to-Peer - P2P payment) by value and volume, comparable to credit cards and debit cards, representing a huge uptick in the number of banks going live on UPI from 2016 to 2021 reaching 297 in 2022. The value of transactions crossed $25 billion in 2019, and massive growth in 2022 reached $111 billion. As per RBI, overall Indian digital payment by volume stands at 72 billion as of FY21-22, with an overall transaction value of INR 1,744 trillion ($24 trillion). Riding on the back of growing acceptance of existing digital modes and novel payment offerings such as UPI, BBPS, and Buy-Now-Pay-Later (BNPL) schemes, the value of digital payments transactions in India is set to increase by more than 3 times by 2025. RBI’s ‘Payments Vision 2025’ aims to curb the volume of cheque-based payments to less than 0.25% of the total retail payments. It will also target increasing the number of registered users for mobile-based transactions at a CAGR (compounded annual growth rate) of 50% by 2025.



**Volume of P2P and P2M transactions**

**Source: NPCI, EY ANALYSIS [12]**



**UPI P2P And P2M Transactions**

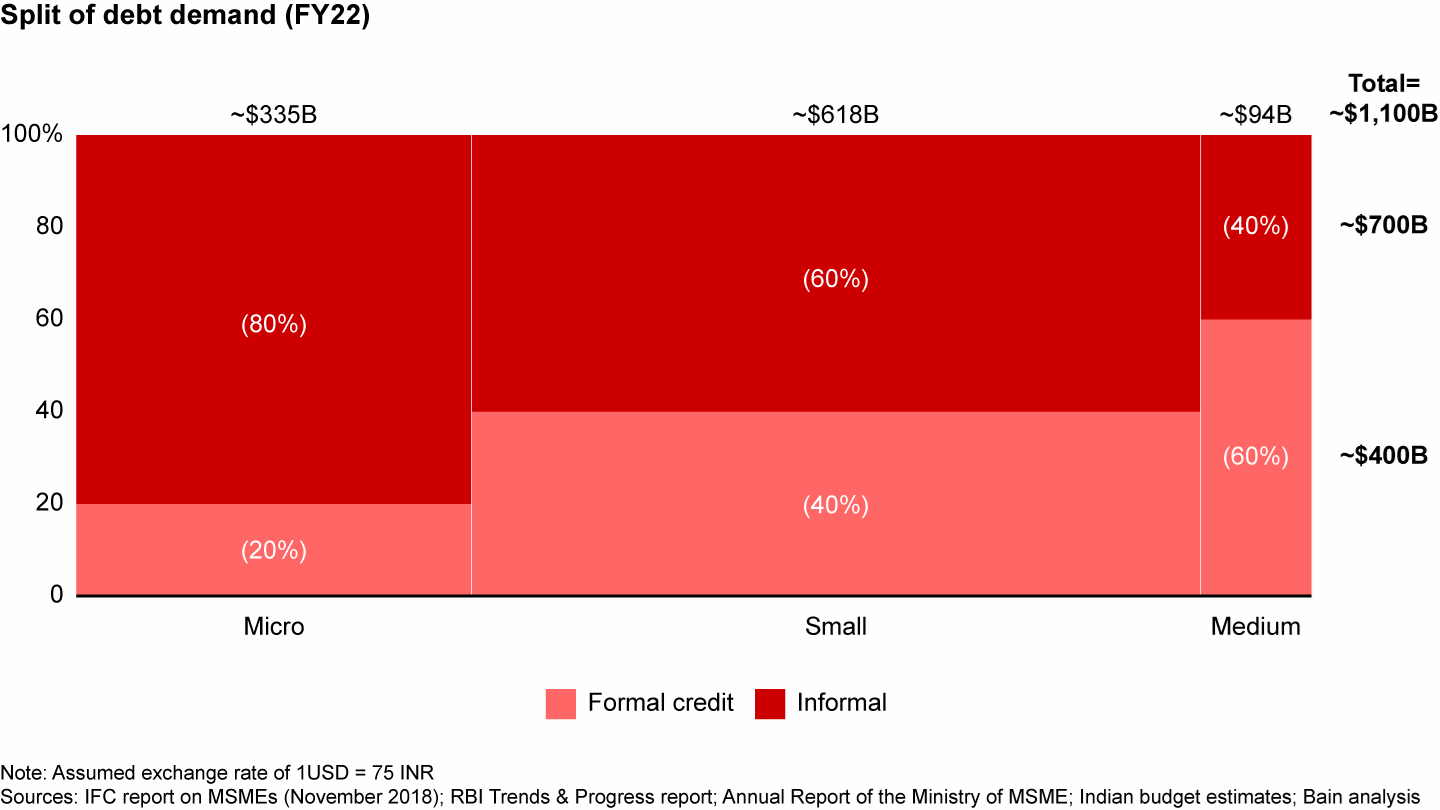
**Source: NPCI Statistics INR to $ conversion rate considered as of 08 August 2022**

However, recent developments have seen researchers and banking authorities shift towards using machine learning and deep learning algorithms to predict credit scores automatically. This approach streamlines the selection of eligible candidates for loan approval.

The digital lending sector in India is experiencing a boom, driven by the country's digitalisation. The sector is projected to account for 60% of the country's financial technology market by 2030, with the online loan market expected to hit $1.3 trillion by 2030, growing more than four times in value from $270 billion currently. Digital lenders are providing an alternative to traditional lenders by using algorithms, artificial intelligence, and their own analysis to assess a customer's creditworthiness and gradually build up a credit history by using their platforms. However, a lack of awareness about digital lending and data security remains a challenge. The Indian FinTech ecosystem has emerged as a formidable global force and continues to grow as one of the largest FinTech markets globally. India is far ahead of the global average of 64% in the FinTech adoption race. By 2030, it is anticipated that the promising Indian FinTech market will generate $200 billion in revenue and $1 trillion in AUM. The collaboration between banks and FinTech players can boost credit access to the underserved segment and SMEs because 75% of MSME lending in the country still is led by banks.

Coming from a detailed study of how Paytm, a major lending company, assesses credit using advanced computer methods, this research brings attention to the big role those different kinds of information, like having various assets as collateral, past financial behaviours and financial literacy, play in deciding if someone will be able to pay back a loan. This discovery gleams a light on the small but important details that go into deciding the risks of lending and raises important questions about fairness when using complicated credit scoring algorithms to make these decisions. This study doesn't just stick to one company; it also looks at how tech companies and banks work together. By closely looking at how banks are doing financially and how they partner with platforms like CIBIL, this research shows that there's a clear change in how money is being given out as loans.

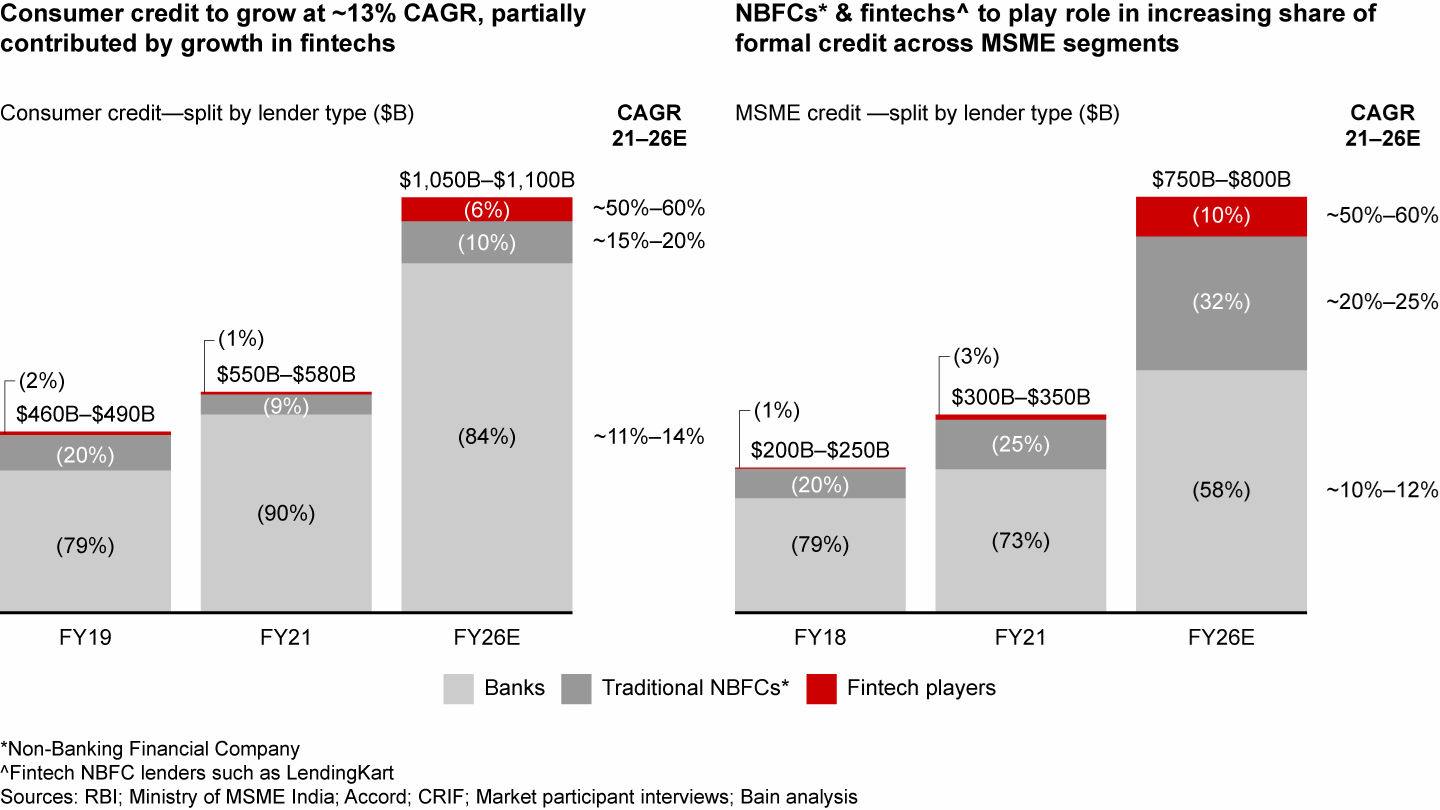
Bangalore is a special place for this study because it's a mix of different businesses and a growing technology industry. This means lots of people have different kinds of jobs and financial situations. The research will use information from the past to make a smart system that can predict if someone might not pay back their loan in the future. Moreover, the study directs its focus onto loans sought for personal and entrepreneurial aspirations as an encapsulated domain of inquiry. Stimulatingly, the performance of computational models displays discrepancy when encompassing the entire dataset vis-à-vis concentrating on loans for small-scale business ventures. This anomaly underscores potential discrepancies in the assessment processes and mechanisms for projecting loan defaults in this specific loan category.



**Indian Debt Demand**

**Source: Bain & Company (Oct 2022)** [**[2]**](https://www.bain.com/insights/india-fintech-report-2022-sailing-through-turbulent-tides/)

A special attention must be brought forward in regards to the growing informal credit borrowings in India. The changing pattern in borrowing habits in Small, medium and micro businesses is yet to probed into in Bangalore. Though, it is important to note that a significant portion of the population lacks an established credit history, particularly in emerging economies where credit reporting systems may be immature. This lack of credit information hampers financial institutions' ability to establish a robust credit scoring framework that can differentiate high-risk borrowers from the larger pool of applicants, which becomes even more challenging in the context of peer-to-peer (P2P) lending platforms. Furthermore, the absence of credit data can lead to deserving candidates being denied access to credit. With the increasing prominence of P2P lending, payday loans, and online microlending markets in developing economies, it is crucial for financial institutions to explore more sophisticated methods of assessing borrowers' likelihood of default.



**Indian Consumer Credit Market Growth**

**Source: Bain & Company** [**[3]**](https://www.bain.com/insights/india-fintech-report-2022-sailing-through-turbulent-tides/)

Currently accounting for about 7% of India’s $1.4 trillion FS EV, the fintech sector is expected to grow to $350 billion in EV by 2026, representing nearly 15% of FS market cap. Covid-19 led to transformational shifts in consumer behaviour and accelerated digital adoption:

1. Non-cash payments soared, with more than 75% year-over-year (YoY) growth in UPI transactions between FY20–21.
2. ***Digital lending apps (DLAs) accounted for more than 60% of loans disbursed by nonbank financial companies (NBFCs) in FY21.***
3. Over 35 million demat accounts were added in FY22 (till Nov’21), thereby increasing the tally of demat accounts by 63%, from 55 million in FY21 to nearly 90 million in FY22.

The above-mentioned stats from a Bain & Company report will help us understand the growth of Fintech sector and digital lending in India, the same cannot be concluded for the credit market in Bangalore and very little probing has undergone in this geography. That is where this research would come in, the author proposes to create a system able to identify, categorize and predict consumers’ financial behaviours.

To make this smart system, the author will employ a type of machine learning technique that's really good at understanding patterns called "Alyuda neural network algorithms” which is supplemented with Various types of regression models. This will aid in analysing all the information we gather from people and predict whether they might have trouble paying back the money they borrowed.

Even during the Covid-19 pandemic, loans for shopping without any security have been steadily increasing, growing by about 25% in the last three years (from FY19 to FY22):

* Credit cards have shown strength with a growth of about 19% every year.
* Personal loans have grown really well, increasing by around 29% every year.
* Loans for buying things like appliances have also bounced back to the levels before the pandemic and have grown by about 13% every year.

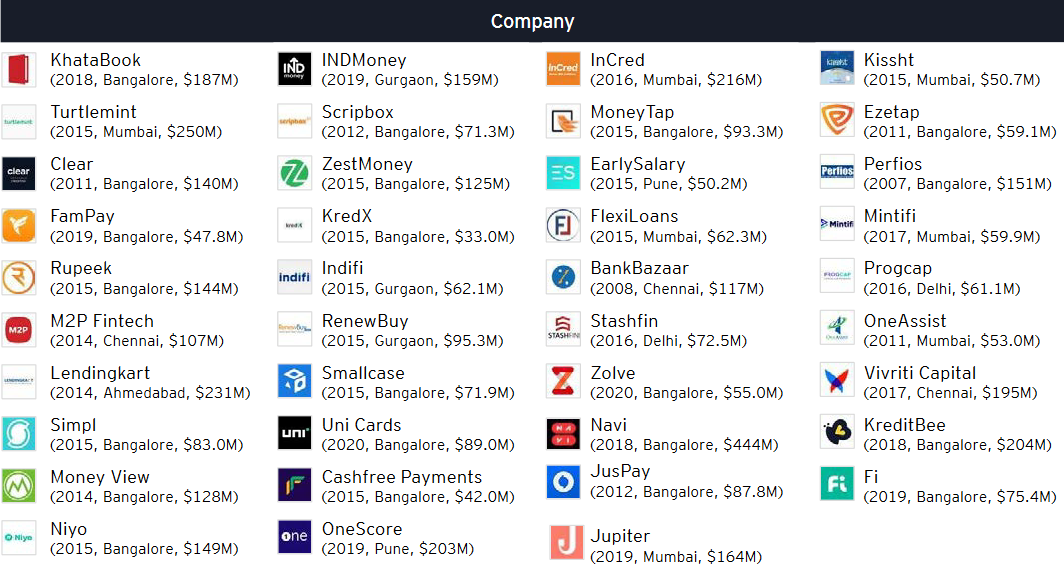
The places that are not big cities have been the main reason for this growth in loans without security. In the smaller towns (called Tier 4), these loans have grown by about 32% every year, while in the big cities (called Tier 1), they have grown by about 18% every year over the last three years. That is where the population of North Bangalore would come in, they represent the perfect mixture of Tier III and Tier II geographies.

The amount of money people are borrowing for different kinds of loans, like personal loans and loans for appliances, is getting smaller. This is because more companies that lend money (like fintech and NBFC lenders) are focusing on giving out smaller loans. The average amount of money people borrow from NBFCs for personal loans has gone down by 70% in the last two years.

For credit card loans, the smaller loans (less than $650) have been growing really well, about 12% every year for the last three years. This is happening mostly in the smaller towns. Personal loans for smaller amounts (less than $650) have been growing really fast too, about 120% every year, and most of these loans (85%) are given to people younger than 35. Just like credit card loans, this growth is also happening more in the smaller towns. Loan prediction is a hot topic in the banking and finance sectors. Credit scoring plays a vital role in this highly competitive financial landscape. With the recent advancements in data science and artificial intelligence, there's been a surge in interest and research in this field. Loan prediction and credit risk assessment have become the centre of attention in recent years due to the increasing demand for loans.

The FinTech sector in India has experienced a significant surge in funding over the past few years, attracting massive investment from large venture capital and private equity firms. FinTech companies capitalized on the rising demand for digitization of financial services during the COVID-19 pandemic. In 2021, the Indian FinTech market witnessed an investment of $8 billion, producing over 15 FinTech unicorns during the year. The Indian FinTech ecosystem is one of the fastest-growing in the world, with a FinTech adoption rate of 87% against the global average of 64%. The Indian FinTech industry is projected to reach $190 billion in revenue by 2030, with a transaction value of $138 billion in 2023. The growth of the Indian FinTech industry can be attributed to various factors, such as government initiatives, a thriving funding environment, a thriving VC ecosystem, high FinTech adoption, and access to talent and technology. The collaboration between banks and FinTech players can boost credit access to the underserved segment and SMEs, which will continue to be a massive opportunity for FinTech.

The demand for improved credit scoring and loan prediction models has skyrocketed. Over the years, various techniques have been employed to assign credit scores to individuals, and extensive research has been conducted on this subject. Unlike the past, where experts made credit assessments based on professional judgments, the focus has shifted towards automated methods.



**Digital Lending Soonicorns in India   
Source: EY** [**[11]]**](https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&ved=2ahUKEwjQgImRqJiBAxWG2TgGHTV2BmEQFnoECC0QAQ&url=https%3A%2F%2Fassets.ey.com%2Fcontent%2Fdam%2Fey-sites%2Fey-com%2Fen_in%2Ftopics%2Fconsulting%2F2022%2Fey-winds-of-change-india-fintech-report-2022.pdf%3Fdownload&usg=AOvVaw3xRJh_4QXIF1yNDD2CX7jP&opi=89978449)

A Soonicorn is a company that is likely to achieve a US $1 billion valuation in the short to medium term

A very important trend in India, according to EY, is that Paytech is consolidating, Payment players are among the leading acquirers of digital lending start-ups in India. Between 2015 to 2021, seven digital lending payers got acquired by payment companies. It was followed by five deals where another lender acquired the digital lending start up

Loans for buying appliances in smaller amounts have also bounced back, growing by about 11% every year in the last three years. More than 70% of these loans have been given to people under 40, with about 36% going to people aged 30 to 40, and 37% to people under 30. This shows that many younger people, like millennials and Gen Z, want loans for buying things. Further secondary research indicates a trend of, people who are getting a loan for the first time are really good at repaying their credit card loans, but they are having some trouble repaying personal loans and loans for appliances.

All of this shows that there is a clear trend towards smaller loans for shopping, especially among younger people. The growth in the smaller towns and among new customers also tells us that companies are using the internet more to give out loans and using different kinds of information to decide who should get a loan.

A big part of this research is understanding how people who might not have a good credit history – like those who borrow money instantly without much checking – can still be understood and predicted. These risky loans are a big challenge, and we want to figure out how to make lending money to these people safer for everyone.

Indian banks have managed to bring down non-performing loans (NPA) in two years, but their wilful defaults rose by 38.5%, or $11.4 billion. According to a report by The Indian Express, there were 15,778 wilful default accounts worth $41.3 billion as of December 2022, compared with 14,206 accounts involving $34.1 billion a year ago. The State Bank of India (SBI) holds accounts of 1,883 wilful defaulters ($9.6 billion), followed by Punjab National Bank ($4.6 billion) and Union Bank of India ($4.27 billion), according to CIBIL data. Public sector banks account for 85% of the wilful defaults, while private banks reported 1,523 accounts for $3.3 billion as of December 2022. A wilful defaulter is a borrower who is unwilling to meet debt obligations despite having the capacity. Such an entity takes undue advantage of legal and governance loopholes, although they have been sued by their lenders. The RBI’s ‘Payments Vision 2025’ aims to curb the volume of cheque-based payments to less than 0.25% of the total retail payments and increase the number of registered users for mobile-based transactions at a CAGR of 50% by 2025. If institutional lenders face this much issues in loan defaults one can only imagine the state of things across the unregulated, digital lending space. There has been very little research among the state of things in space with special reference to Bangalore, however, this dimension will be limited to just the borrowers’ perspective, as to what factors causes the loan default and whether intentional defaults are backed by any substantial action patterns. Among financial institutions, HUDCO (Housing and Urban Development Corporation LTD) had 130 wilful default accounts for Rs 12,211 crore. Private banks reported 1,523 accounts for Rs 27,431 crore as of December 2022.

In a nutshell, this dissertation discovers using new technology to understand people's money habits better and predict if they might not pay back their loans. By doing this, the author hopes to make lending and borrowing money more secure and reliable. This research will help us to understand the tricky ways we figure out who gets loans and how it's changing. The research probes into great detail in identifying various successful techniques of machine learning used to prepare a credit scoring model for imbalanced population and sample sizes with varying qualitative factors which may or may not directly influence their ability to pay back the loans borrowed and also indicate whether or not they might get churned and see how tech companies and banks team up with special reference to Bangalore population.

Keywords: predictive model, loan defaults, customer churn, credit behaviour, technology, financial habits, instant loans, credit profile.

# Chapter 2 – Organizational Profile

***About Flipcarbon Integrated Solutions Pvt Ltd***

Flipcarbon Group was established in 2014 with the aim of providing comprehensive strategies, meticulous execution designs, and effective deployment solutions to organizations driven by growth. They collaborate closely with startups, small and medium-sized enterprises (SMEs), micro, small, and medium-sized enterprises (MSMEs), large multinational corporations, and family-run businesses to unlock the untapped potential of their ventures.

The foundational principles that underpin Flipcarbon's operations are as follows:

1. Unwavering commitment to accelerated growth: At Flipcarbon, they are deeply committed to propelling rapid growth for their clients. They understand the significance of achieving growth targets swiftly, and they strive to deliver tangible results within defined timelines.
2. Cultivating a culture that fosters rapid growth: They have cultivated a unique organizational culture that empowers autonomy, enabling their clients to drive growth initiatives. However, they also maintain a strong focus on cost control through disciplined practices, ensuring financial sustainability throughout the journey.
3. Embracing capability practices: Their approach encompasses not only a deep understanding of the intricacies involved in pioneering new ventures but also a keen awareness of the people dynamics associated with such endeavours. They recognize that success lies not only in executing novel ideas but also in nurturing the human capital that drives innovation and change.
4. Nurturing and deploying highly competitive individuals: Flipcarbon places great emphasis on developing a workforce that possesses the skills, knowledge, and expertise required to thrive in a competitive landscape. By fostering a culture of continuous learning and growth, they empower their team members to deliver exceptional results for their clients.

Flipcarbon: Virtual CFO & Accounting Services represents a significant facet of their operations. This division specializes in providing expert financial guidance and accounting solutions, serving as a trusted partner for their clients.

The "specific-solution-to-specific-issues-of-specific-client" approach that defines Flipcarbon's orientation draws inspiration from the element of carbon. Just as carbon forms unique bonds under specific circumstances, they strive to create tailored solutions by leveraging the distinctive elements of each client's situation. This approach allows us to craft unparalleled compounds of strategies and solutions that address the specific challenges faced by their clients.

***Background and Current Description***

Flipcarbon Integrated Solutions Private Limited is an unlisted private company incorporated on 04 December 2014. It is classified as a private limited company and is located in Bangalore, Karnataka. Its authorized share capital is INR 1.00 Cr and the total paid-up capital is INR 91.50 lac.

Description: The company provides HR & Financial Consulting, end-to-end HR & Financial Outsourcing, Modular HR Services, and Financial Services

***Products & Services of CFO Vertical:***

CFO Advisory, Accounting, taxation and MIS services, Payroll processing, Handling PF and ESI Compliances, Company secretary functions for Indian Private Limited Companies. Virtual CFO, Accounting & Bookkeeping Services for Growing Startups & Businesses. Cashflow, Cost & Working Capital Management, Business Plans, Valuation Modelling & Financial Modelling, Board & Management Reporting, Investment Advisory, Transaction Support, Fundraising, Financial Strategy & Planning

***Our Team***

The company has three directors and no reported key management personnel. The longest-serving director currently on board is Prabhash Nirbhay, who was appointed on December 4, 2014. Interestingly, he is also the founder of the organization. Prabhash Nirbhay has been on the board for more than eight years and is also the founder of the organization. Additionally, Lokesh Mehta was appointed as a director on August 10, 2015. The most recently appointed director is Alok Ranjan, who was appointed on March 1, 2018 and is also the current acting CEO of the organization. Prabhash Nirbhay has the largest number of other directorships, with a seat at a total of six companies. In total, the company is connected to 5 other companies through its directors.

The Bangalore branch has a team of over 13 dedicated employees, while the Flipcarbon family boasts a network of 30+ experts across three branches. The CFO vertical, where the author interned, is led by Mr. Deepak Kewalramani, a seasoned professional with nearly three decades of expertise in financial analytics and strategic advisory. The author had the privilege of reporting to Ms. Pramila Lakra, who possesses almost five years of experience in Financial Due Diligence and Forecasting, as well as strategic advisory in the field of Finance, including Mergers and Acquisitions, funding advisory, and more.

Recognizing the need for a business intelligence expert, the organization sought someone who could streamline the process of preparing financial statements and create visually appealing and informative visualizations using Microsoft Suite and Power BI. The author was a perfect fit for this role, thanks to their background in finance and business analytics. Drawing from their experience as a business analyst, the author was able to identify interoperability issues between different verticals of the firm and proposed a framework for integrating solutions using a tech-stack (SaaS) approach.

With a collective management experience of over 400 years, eight years of consulting expertise, and a growing list of more than 130+ extremely satisfied clients, they are committed to becoming the gold standard in management consulting.

***Mckinsey’s 7s Framework***

**Strategy**:

* Full stack Human resource, Financial and enterprise consulting: Flipcarbon's strategy encompasses a wide range of financial services, human resource consulting and other services tailored for startups and growing businesses.
* Client-Centric Approach: Their strategy is client-focused, aiming to provide tailored financial advice and support to meet clients' specific needs.
* Growth-Oriented: The strategy emphasizes helping clients grow, scale, and achieve sustainable financial outcomes, learning and development, capability transformation etc.
* Investment and Transaction Expertise: Flipcarbon assists clients in fundraising, investments, and M&A activities, aligning with a strategic approach to financial growth as well as human resource capability transformation services.

**Structure**:

* Multidisciplinary Team: Flipcarbon has assembled a diverse team of business partners, principal consultants, client engagement managers.
* Collaborative Work Environment: Their structure promotes collaboration and teamwork to deliver comprehensive financial solutions.
* Flexible Service Delivery: The organization's structure allows for adaptability in providing financial services customized to individual client requirements.
* Supportive Infrastructure: Flipcarbon likely employs technology and tools to facilitate financial modelling, accounting, and advisory services.

**Systems**:

* Financial Tools and Software: Flipcarbon employs financial modelling tools and accounting software to manage client finances effectively.
* Data Management: They have systems in place for data collection, analysis, and reporting to support financial decision-making.
* Compliance Management: Systems are implemented to ensure clients' financial compliance, including GST returns and tax-related matters.
* Technology Integration: The organization likely integrates technology solutions to streamline financial processes and improve efficiency.

**Shared Values**:

* Client Success: The shared value is client success, with Flipcarbon striving to help clients achieve their financial goals and benchmarks.
* Expertise and Trust: They value expertise and trust, as indicated by their emphasis on having a team with a cumulative experience of over 400 years.
* Growth and Sustainability: Flipcarbon's shared values align with the growth and sustainability of their clients' businesses.
* Integrity and Professionalism: Integrity and professionalism are likely core values, given their role in financial advisory and compliance.

**Skills**:

* Financial Expertise: The team possesses financial expertise in areas such as financial strategy, accounting, and investment advisory.
* Business Acumen: They have a strong understanding of business operations, enabling them to provide valuable financial insights.
* Analytical Skills: Skills related to financial modeling, budgeting, and sensitivity analysis are essential for their services.
* Communication and Collaboration: Effective communication and collaboration skills are likely vital to interact with clients and within the team.

**Style**:

* Client-Centric Approach: The leadership style emphasizes client needs and satisfaction.
* Innovative Thinking: A culture of innovation might be encouraged to find creative financial solutions.
* Proactive Decision-Making: Leaders may take a proactive approach to guide clients in making sound financial decisions.
* Results-Oriented: The organizational culture may prioritize achieving tangible financial outcomes for clients.

**Staff**:

* Diverse Expertise: The team comprises individuals with diverse skills and backgrounds, contributing to a holistic approach to financial services.
* Dedicated Support: The inclusion of "Dedicated Accounts Executive" and "Dedicated Engagement Manager" suggests a commitment to personalized client support.
* Experienced Professionals: The team includes professionals with 3-15 years of experience, demonstrating a blend of youth and experience.
* Client Engagement: Monthly meetings and unlimited phone support indicate an active engagement approach with clients.

# Chapter 3 – Review of Literature and Research Design

# Review of Literature

Loan default prediction using a credit rating-specific and multi-objective ensemble learning scheme

Yu Song, Yuyan Wang, Xin Ye, Russell Zaretzki, Chuanren Liu (2023)

This research paper investigates into the critical province of credit risk assessment within the consumer lending industry. The assessment of credit risk plays a pivotal role in assessing the probability of loan default, a matter of substantial concern given the potential financial losses associated with such defaults.

Loan default prediction is a complex classification problem, beset with challenges such as highly imbalanced class distributions. Striking the right balance between robust default identification and overall classification accuracy proves to be a formidable task. Previous studies have often treated loans across different credit ratings as a homogeneous entity, overlooking the nuanced relationship between class imbalance and specific credit rating categories.

This paper takes a distinguished withdrawal from convention by proposing a novel approach: a credit rating-specific multi-objective ensemble learning framework. This methodology tailors modelling strategies for subpopulations of loans sharing similar default risk profiles within distinct credit rating categories. It employs a multi-objective ensemble learning technique, leveraging the One-Class Support Vector Machine (OC-SVM) as the foundational learning algorithm. The ensemble construction process is facilitated through multi-objective evolutionary optimization, and an innovative adaptive classification boundary adjustment technique is introduced to enhance imbalanced classification performance.

To validate the efficacy of this pioneering methodology, extensive experimental studies are conducted, comparing it against several benchmark algorithms using an industry-provided dataset. The experimental findings substantiate the advantages of the rating-specific approach and underscore the superiority of the proposed multi-objective ensemble learning method.

The contributions of this paper can be summarized as follows:

1. **Holistic Model Focus:** Unlike previous research that primarily emphasized enhancing a model's default identification ability (sensitivity), this study seeks to strike a balance by achieving both strong default recognition capacity and overall prediction accuracy.
2. **Innovative Ensemble Learning:** A novel classification approach is developed, harnessing an ensemble of OC-SVM learners optimized through an evolutionary algorithm. This method maximizes two critical objectives simultaneously: default and good loan identification rates. The fusion of OC-SVM, multi-objective ensemble learning, and the proposed adaptive boundary adjustment strategy consistently enhances performance, especially in scenarios with highly imbalanced data.
3. **Rating-Specific Strategy:** The paper demonstrates, through empirical evidence, that employing a credit rating-specific modeling strategy leads to improved overall classification performance. This is achieved by aggregating results from classifiers that are most effective given the imbalance ratio within each credit rating category.
4. **Real-World Applicability:** The paper presents a multi-objective evolutionary framework, enabling financial institutions to select models that align with their risk tolerance levels for clients within specific credit ratings. This approach empowers institutions to establish flexible loan review principles in practical applications.

In short, the author chose this paper as it presents a pioneering methodology that not only advances the field of credit risk assessment but also holds the potential to revolutionize the way financial institutions evaluate and manage loan portfolios, ultimately fostering more informed and adaptive lending practices.

Balanced incremental deep reinforcement learning based on variational autoencoder data augmentation for customer credit scoring

Yadong Wang, Yanlin Jia, Yu Zhong, Jing Huang, Jin Xiao (2023)

This paper addresses the challenges associated with training deep reinforcement learning models in an incremental setting. This approach has found success in various real-world applications, but it often encounters the issue of catastrophic forgetting, where training on new data degrades the model's performance on old data. To mitigate this, the authors propose a novel model called Balanced Incremental Deep Q-Network based on Variational Autoencoder Data Augmentation (BIDQN-VADA) for customer credit scoring.

The methodology involves several key steps. First, the original training set undergoes random under sampling to create a balanced training set, on which an initial deep Q-network model is trained. Subsequently, a balanced training subset is selected from the original set, and variational autoencoder data augmentation is employed to augment these samples, resulting in a balanced augmented training subset. These balanced subsets are stored in a data stream cache following a first-in first-out (FIFO) principle for incremental parameter updates of the deep Q-network model.

The paper's contributions are significant. Firstly, it introduces a balanced incremental learning method for training deep reinforcement learning models. This involves updating the BIDQN-VADA model's parameters with an equal number of positive and negative samples during each incremental process. This approach enhances the model's performance in scenarios where class distribution is imbalanced, such as customer credit scoring.

The authors also propose a novel data stream cache that incorporates both augmented and original samples, utilizing the FIFO approach. This setup aids in retaining previously learned knowledge, counteracting the tendency of incremental models to forget old information. Additionally, the inclusion of augmented samples compensates for the challenge of effectively learning from a small number of new samples.

The paper's contributions are validated through experiments conducted on eight real-world customer credit scoring datasets. The results indicate that the BIDQN-VADA model outperforms seven other classification models, demonstrating its prowess in achieving superior customer credit scoring performance through a combined approach that integrates balance, incrementality, and data augmentation.

In the subsequent sections, the paper delves into related works, provides a comprehensive theoretical background, details the BIDQN-VADA model's process, outlines the experimental design, analyses the experimental results, and concludes by summarizing key findings and outlining avenues for future research.

Interpretable machine learning for imbalanced credit scoring datasets

Yuija Chen, Raffella Calabrese, Belen Martin-Barragan (2023)

The study delves into the intricate realm of credit scoring, focusing on a critical issue – class imbalance – that often plagues the domain. While existing research primarily concentrated on the repercussions of class imbalance on predictive accuracy, a significant gap persisted in understanding its impact on interpretability within machine learning. Addressing this void, the paper embarks on a journey to unravel how the stability of two prominent interpretation methods, namely Local Interpretable Model-agnostic Explanations (LIME) and SHapley Additive Explanations (SHAP), is influenced by class imbalance.

The investigation is meticulously structured, employing a novel experimental framework that encompasses a controlled sampling process. This approach allows for a comprehensive assessment of the shifting dynamics as class imbalance increases. Drawing upon a real-world dataset spanning 2016 to 2020, encompassing UK residential mortgage data sourced from the European Datawarehouse, the study uncovers the intricate relationship between class imbalance and the stability of interpretation methods.

The findings unearth a compelling narrative – as class imbalance escalates, the stability of interpretations generated by LIME and SHAP diminishes. This unveiling of diminishing stability implies a direct and negative influence of class imbalance on the interpretability of machine learning models. This realization showcases a previously uncharted facet of the class imbalance issue and underscores its multifaceted implications beyond predictive accuracy.

In an effort to corroborate the robustness of these insights, the study extends its analysis to two open-source credit scoring datasets. Remarkably, these additional analyses yield similar results, solidifying the consistency and universality of the observed phenomenon across different datasets. This consistency underscores the broader relevance of the findings, reaffirming the detrimental impact of class imbalance on interpretability regardless of the dataset's origin.

This study's contribution is twofold. Firstly, it sheds light on the often-overlooked interplay between class imbalance and the stability of interpretation methods. By bringing this phenomenon to the forefront, the paper challenges the conventional understanding of class imbalance's ramifications. Secondly, it introduces a rigorous experimental framework that could serve as a cornerstone for future investigations into the complex landscape of credit scoring, class imbalance, and machine learning interpretability.

Credit default prediction from user-generated text in peer-to-peer lending using deep learning Johannes Kriebel, Lennart Stitz (2023)

This study resonates deeply with the essence of our times – the synergy between data, technology, and the financial landscape. The authors' approach of extracting significant credit-related information from user-generated text on platforms like Lending Club is a leap forward in understanding the dynamic factors influencing credit default. What truly stands out is the revelation that even short pieces of user-generated text possess the potency to substantially enhance the accuracy of credit default predictions.

The paper's assertion that deep learning, particularly in tandem with state-of-the-art transformer models, triumphs over conventional methods in most cases highlights the rapid evolution of machine learning techniques. This echoes the growing consensus that the depth and complexity of neural networks hold the key to unlocking hidden patterns within seemingly disparate data. However, what's intriguing is that amidst this complexity, simpler models like average embedding neural networks shine just as brightly, reaffirming that brilliance often lies in simplicity.

The authors' exploration of the credit-scoring landscape through the lens of textual data prompts a paradigm shift. Traditional methods may have well-established ties to structured data, but the potential of unstructured textual information, as exemplified in this study, cannot be underestimated. These findings underscore the need for financial institutions to broaden their horizons, integrating a more holistic range of data sources to make better-informed decisions.

Looking forward, the implications of this study stretch beyond credit default prediction. It beckons the financial industry to further explore the vast potential of user-generated text and unstructured data in shaping various aspects of customer engagement, risk management, and strategic decision-making. The notion that "text matters" introduces a new dimension of personalized financial services, where understanding the customer's narrative becomes just as critical as analyzing their transaction history.

Research on Default Prediction Model of Corporate Credit Risk Based on Big Data Analysis Algorithm

Qingyan Xianyu, Mo Hai (2023)

In this research, the authors reconnoitre into the dynamic intersection of technology, finance, and data analysis to address the critical concern of corporate credit risk prediction. With the advent of transformative technologies like big data and artificial intelligence, the financial industry is witnessing a profound shift towards digitization and information-driven decision-making.

The core objective of this study is to construct a robust default prediction model for corporate credit risk by harnessing the power of various big data analysis algorithms. To achieve this, the authors meticulously gather and curate a comprehensive dataset comprising 21 financial and non-financial indices from a pool of over 1,000 listed companies. The data pre-processing journey involves standardization, balancing, and normalization, ensuring that the dataset is primed for analysis.

Leveraging the data, the authors employ a meticulous approach that involves using correlation coefficients to sieve through the indices. Subsequently, they embark on developing two in-depth learning models: a Convolutional Neural Network (CNN) model and a Recurrent Neural Network (RNN) model. These models are constructed using the Pytorch framework within the Spark platform. The primary goal is to optimize these models for credit risk prediction.

In addition to these neural network-based models, the authors pit them against two traditional machine learning models: the Random Forest model and the Logistic Regression model. This comprehensive comparison serves as a crucial benchmark for evaluating the performance of the neural network models.

The outcomes of this extensive experimentation reveal that the Recurrent Neural Network (RNN) emerges as the optimal model. It achieves an impressive accuracy rate of 0.93, a recall rate of 0.96, and an F1 value of 0.93. Beyond model selection, the authors delve into the robustness of the optimal RNN model. They subject it to rigorous tests, such as modifying the number of indicators, altering the number of samples, and eliminating non-financial factors. Remarkably, the evaluation metrics of the model exhibit minimal variation under these conditions, underscoring the model's remarkable robustness.

This research is a testament to the transformative potential of big data analysis algorithms in revolutionizing credit risk prediction within the corporate landscape. The adoption of deep learning models, particularly the RNN, showcases the authors' commitment to harnessing cutting-edge technology to optimize credit risk assessment. Furthermore, the model's robustness under various scenarios underscores its viability as a dependable tool for corporate credit risk evaluation in a rapidly evolving financial landscape.

Keywords: Default Prediction, Model Corporate credit risk, Deep learning, Big data, Spark

Machine learning for credit scoring: Improving logistic regression with non-linear decision-tree effects

Elena Dumitrescu, Sullivan Hue, Christophe Hurlin, Sessi Tokpavi (2022)

This research resonates deeply with the evolving landscape of credit scoring, where the intersection of machine learning and finance presents promising avenues. The authors' innovative approach to augmenting logistic regression with non-linear decision-tree effects sheds light on the potential to enhance predictive accuracy while preserving interpretability – a crucial concern in the financial sector.

The utilization of ensemble methods like the random forest has demonstrated superior classification performance in credit scoring compared to traditional logistic regression models. However, the enigma of interpretability has restrained the adoption of these advanced techniques within the regulatory framework. Addressing this gap, the paper introduces the concept of Penalized Logistic Tree Regression (PLTR) – a novel methodology that extracts valuable insights from decision trees to augment logistic regression. This symbiotic fusion of methods is designed to capture the non-linear intricacies inherent in credit data, enabling better predictive power while retaining the comprehensibility of logistic regression.

The research illuminates the dichotomy between complexity and transparency, spotlighting how PLTR offers a pathway to achieving both. While ensemble methods often carry the label of "black box" due to their limited interpretability, PLTR signifies a pioneering approach that marries the strengths of logistic regression and decision trees. The introduction of binary rules derived from shallow decision trees as predictors within a logistic regression framework not only enhances accuracy but also imbues the model with transparency.

The paper underscores the limits of traditional models in capturing non-linear effects present in credit data, highlighting the limitations of standard logistic regression in this regard. This revelation strengthens the case for the innovative PLTR model, which adeptly integrates decision trees to address this deficiency. By enabling the incorporation of non-linearity through pre-treatment of predictors, PLTR advances the scope of credit scoring models, transcending the limitations of linear specifications.

In practical terms, the research's findings have profound implications for the credit industry. As regulatory bodies emphasize the need for interpretability, PLTR emerges as a formidable contender in credit risk assessment. Its ability to reconcile the complexities of modern data analysis with the regulatory demands for transparency positions it as a viable solution to the interpretability conundrum.

Looking ahead, the implications of this research ripple beyond credit scoring. It heralds a new era where the synergy between machine learning and finance can redefine customer engagement, risk management, and decision-making across the financial spectrum. As the financial landscape continues to evolve, the PLTR model could serve as a foundational pillar for leveraging sophisticated techniques while maintaining regulatory compliance.

In summation, "Machine Learning for Credit Scoring: Improving Logistic Regression with Non-Linear Decision-Tree Effects" provides a pioneering glimpse into the harmonization of advanced machine learning techniques with the interpretability crucial to the financial industry. The PLTR model's ability to bridge the divide between predictive power and transparency signifies a leap forward, urging the industry to navigate this promising intersection between innovation and regulation.

Predicting SMEs’ default risk: Evidence from bank-firm relationship data

Michele Modina, Filomena Pietrovito, Carmen Galluci, Vincenzo Formisano (2022)

In this study, we dive into a unique set of data that includes 13,081 Italian companies and 111 cooperative banks engaged in lending. The goal? To uncover solid proof through real-world data that certain actions in handling credit lines and long-term loans can actually predict the likelihood of default over a span of one to two years. We don't stop there – we're factoring in balance sheet cues and ever-changing bank attributes, neatly captured by bank-time constants.

And guess what? When we merge these credit-related signals from private in-house banking sources with financial records, we're able to significantly enhance our predictions of defaults for small and medium-sized businesses. But we're not done yet. We're going the extra mile by measuring just how much bank-specific insights help by comparing our prediction precision against a model that only considers financial data versus one that's fueled by both financials and these behind-the-scenes nuggets of info.

Now, let's talk diversity. We uncover that the connection between balance sheet clues, the data reflecting bank-firm ties, and the odds of default isn't a one-size-fits-all deal. It can shift across different industries and regions. This, in essence, underlines how vital it is for banks to buckle down and analyze things on a case-by-case basis to truly grasp the risks these businesses bring to the table.

Keywords: Bank-firm connection, Risk of loan default, Foreseeing defaults, Loaning procedure, Small and medium-sized businesses

Customer Churn Prediction for Fintech Companies Using Artificial Neural Networks (Pooja Malhotra, Punit Patel, Neel Shah) (2020)

The study's findings really highlight how urgent it is to get smarter about predicting customer churn. Using learning methods like Artificial Neural Networks and data mining can help businesses keep their customers from jumping ship and stop revenue from tanking. The paper emphasises the relevance of understanding customers, tapping into top-notch tech, and making moves to tackle churn head-on.

In today's business world, there's a data explosion happening across all industries. The fancy term for turning raw data into smart insights is Data Mining – it's all about using advanced tools to identify previously implicit trends and relationship. Customer churn is when a company and its customers part ways, and that's a big deal. It shows just how important it is to predict what customers will do and build strong relationships with them.

In a nutshell, this study is all about nailing those churn predictions using fancy techniques. The research paper goes into great detail in regards to machine learning techniques like SVM, Random Forest, and Linear Regression. The research elevates the best practices in using these methods and identifying the right inferences from them..

When companies get really good at guessing what customers will do, especially in the fintech world, they can sort of predict and dodge churn. And that's how they build lasting friendships with customers and make their businesses grow. As everyone in the business world struggles with churn troubles, using Artificial Neural Networks to predict churn is like a secret weapon that could mean the difference between a loyal customer and one who's out the door.

Keywords: Customer Churn, Data Mining, Artificial Neural Networks, Linear Regression, Support Vector Machine, Random Forest.

Machine Learning Based Customer Churn Prediction In Banking (Manas Rahman, V Kumar) (2020)

This study presents a method that employs machine learning techniques, a subset of artificial intelligence, to predict customer churn in the banking industry. The research explores the potential of identifying churn by analysing customer behaviour. The study employs classifiers like KNN (K-means clustering), SVM (Support Vector Machine), Decision Tree, and Random Forest, along with feature selection methods to enhance system performance. The research is conducted on a churn modelling dataset from Kaggle, aiming to identify an optimal model with higher precision and predictability.

Recognizing and understanding consumers is essential for corporations. This concept is reinforced by Liu and Shih, who stress the need for innovative marketing strategies to meet consumer expectations and enhance loyalty. The competitive market necessitates a more effective utilization of marketing resources, and technology, including data mining techniques, serves as a tool to extract marketing insights and guide business decisions. The paper underscores the critical importance of predicting customer churn at early stages. The proposed framework leverages machine learning techniques to predict customer churn in the banking sector. This approach holds potential for helping organizations retain customers by identifying potential churn cases. The study emphasizes the need for predictive models that aid in retaining customer loyalty and boosting organizational success.

The study's findings hold implications for the banking sector, emphasizing the value of early-stage churn prediction. The integration of machine learning techniques offers a novel avenue for addressing customer churn. Future research could study into refining and optimizing the proposed framework, potentially incorporating more advanced machine learning algorithms. Additionally, exploring how this framework can be adapted to other industries could lead to broader applications and insights.

The study, situated in the banking sector, relates the potential of machine learning techniques in predicting customer churn. The dynamic nature of the market necessitates innovative approaches to customer retention. As corporations strive to balance customer acquisition and retention, predictive models fuelled by machine learning offer a promising solution. The research contributes to the evolving landscape of customer churn prediction, guiding future research and aiding the banking industry in tackling this pivotal challenge.

Keywords: Customer Churn, Machine Learning, Banking, Predictive Models, Data Mining.

An Empirical Study on Loan Default Prediction Models (Uzair Aslam, Hafiz Ilyas Tariq Aziz, Asim Sohail, and Nowshath Kadhar Batcha) (2019)

This article's findings suggest that the integration of machine learning algorithms and neural networks holds promise for improving credit risk assessment and loan approval processes. It encourages further research into refining these predictive models to enhance their accuracy and applicability in real-world scenarios. Ultimately, the study contributes to the evolving landscape of credit risk assessment, offering insights that can guide both researchers and practitioners in the finance industry.

The article investigates into the significance of loan lending in the financial world and acknowledges its role in overcoming financial constraints for individuals and businesses. However, the concept of credit risk, associated with the possibility of borrowers failing to repay loans on time, is emphasised as a major concern. This risk, termed credit default, can lead to severe consequences for both parties involved. In spite of the risks, loan lending remains a worthwhile venture for financial institutions, contributing to profit-making and business sustainability.

Traditionally, creditworthiness assessment relied on credit scores assigned to individuals based on their historical data. However, modern advancements in machine learning and neural networks have revolutionized the approach to credit risk assessment. The article accentuates how machine learning algorithms can autonomously predict credit scores by analysing individuals' historical data, enabling lenders to identify potential defaulters and manage risks more effectively. The practice of loan lending, while advantageous for both lenders and borrowers, comes with inherent risks, notably credit risk. This study explores existing literature that focuses on predictive models for credit risk assessment, utilizing machine learning algorithms. The objective of the paper is to present a comprehensive understanding of loan default prediction models and their implications.

Sample selection in credit-scoring models (William Greene) (1998)

The paper "Sample Selection in Credit-Scoring Models" probes into the fields and scopes of credit scoring, a vital process for financial institutions like credit-card vendors. The research this publication was chosen by the author was because it pioneered a new way of collecting and analysing data for the purpose of credit scoring. Credit scoring involves evaluating loan applicants based on statistical models. However, a significant challenge arises during to the sample selection process, as models are constructed using historical data from individuals who have already received loans. This introduces a potential bias in evaluating new applicants randomly drawn from the entire population. The paper scrutinizes this sample selection issue across three distinct applications, each requiring a different statistical model and estimation technique.

The study examines three applications impacted by sample selection, shedding light on how this phenomenon influences the measurement of variables pertinent to credit-card vendors. The applications entail predicting loan default, modeling expenditure, and assessing the number of derogatory reports in credit histories. Each application demands a specific statistical model and estimation technique, accounting for the sample selection issue.

In credit assessment scenarios, loan officers meticulously evaluate large loan applications, whereas credit-card vendors, dealing with millions of applications, rely on statistical models to assign scores. The models sort applications based on statistical profiles of successful borrowers, offering a statistical measure to gauge applicants. However, a fundamental problem emerges due to the data used to build these models. Models are built using data from individuals whose applications were accepted, while evaluations must be made for randomly arriving applications from the entire population.

The paper's findings have crucial implications for credit-card vendors and financial institutions, emphasizing the potential bias introduced by sample selection. This has repercussions on predictive accuracy and business decisions. The study encourages the development of models that address the sample selection problem to yield more accurate predictions.

The study illuminates the necessity of constructing models that account for sample selection biases, enhancing the precision of credit-scoring predictions and contributing to a more accurate evaluation of loan applicants.

**Keywords:** Sample Selection, Credit-Scoring Models, Loan Default Prediction, Statistical Models, Financial Institutions.

Credit risk prediction in an imbalanced social lending environment

Namvar, A., Siami, M., Rabhi, F., Naderpour, M. (2018)

Peer-to-peer lending platforms offer a distinctive ecosystem for borrowers and lenders to directly engage; however, they are not without their hurdles, particularly in terms of class imbalance. While credit risk prediction holds paramount importance within this setting, the authors aptly highlight the limited number of models that have successfully addressed the complexities introduced by imbalanced data. Furthermore, the elusive quest to determine the most fitting resampling technique for such data remains a topic of ongoing debate within scholarly circles.

In a commendable endeavour to bridge these gaps, the article introduces a novel risk assessment methodology that squarely confronts the challenges posed by imbalanced data. The study embarks on an empirical journey, meticulously comparing an array of classifier-resampling technique combinations. Notably, the authors judiciously opt for the G-mean measure as the evaluation metric, an astute choice to counteract any bias that may arise towards the majority class – an often-overlooked consideration in similar studies.

The outcomes of this empirical voyage, unravelling the performance of diverse combinations in credit risk prediction, bring forth a striking revelation. The article effectively showcases that the amalgamation of random forest and random under-sampling emerges as a potent strategy in comprehensively gauging the credit risk linked with loan applicants in the intricate domain of social lending markets.

The implications of this research stretch far beyond the confines of this study. The thorough exploration of class imbalance issues, the meticulous comparison of methodologies, and the steadfast focus on resampling techniques collectively contribute to enriching the ongoing discourse on credit risk prediction. In a financial landscape driven by innovation and technology, studies of this nature serve as vital signposts for refining risk assessment processes.

Keywords: Interpretability, Stability, Credit scoring, Machine learning

A Compact Evolutionary Interval-Valued Fuzzy Rule-Based Classification System for the Modelling and Prediction of Real-World Financial Applications with Imbalanced Data

José Antonio Sanz, Dario Bernardo, Francisco Herrera, Humberto Bustince, Hani Hagras (2014)

This paper stresses on the importance of how accurate prediction models have become imperative, especially in the wake of recent financial crises. The quest for precision in forecasting financial outcomes and managing risk has led to an increased demand for transparent decision-making processes. Furthermore, the challenge of dealing with imbalanced real-world financial datasets has emerged, calling for solutions that do not rely on potentially noisy sampling techniques. Addressing these multifaceted concerns, this paper introduces a novel approach: the Compact Evolutionary Interval-Valued Fuzzy Rule-Based Classification System.

This innovative system, denoted as IVTURS FA RC-HD, builds upon the foundation of interval-valued fuzzy rule-based classification systems with tuning and rule selection. Its primary objective is to enhance the modelling and prediction capabilities of real-world financial applications. An intriguing facet of this proposed system is its capacity to achieve remarkable prediction accuracies while utilizing a concise set of succinct fuzzy rules. This attribute fosters a heightened degree of interpretability within the resulting linguistic model.

Crucially, the IVTURS FA RC-HD system tackles the challenge of imbalanced financial datasets without resorting to pre-processing or sampling methodologies. This strategic approach ensures that noise is not inadvertently introduced into the learning process by avoiding common pitfalls associated with data manipulation techniques. Additionally, the system boasts a mechanism to address instances that fall outside the coverage of any generated fuzzy rule in the rule base.

To validate the efficacy of this approach, an experimental study is conducted involving eleven real-world financial datasets. The results illuminate the system's superiority over various benchmark techniques. Notably, it outperforms the original C4.5 decision tree, as well as both type-1 and interval-valued fuzzy counterparts that rely on the synthetic minority oversampling technique (SMOTE) for data pre-processing. Furthermore, the proposed method exhibits a competitive edge when compared with FURIA – a fuzzy approximative classifier – when utilizing SMOTE. Importantly, the IVTURS FA RC-HD system's distinctive attribute of sidestepping pre-processing techniques ensures interpretability while delivering enhanced accuracy.

Through its fusion of evolutionary techniques, interval-valued fuzzy logic, and an emphasis on interpretability, this system emerges as a promising tool for achieving accurate predictions within the context of real-world financial scenarios.

Resampling ensemble model based on data distribution for imbalanced credit risk evaluation in P2P lending

Kun Niu, Zaimei Zhang, Yan Liu, Renfa Li (2020)

In this paper, the author addresses the critical issue of loan applicant misclassification in credit scoring models, a problem that significantly impacts investors' profits in Peer-to-Peer (P2P) lending. This misclassification stems from class imbalance within credit data, which often leads to models prioritizing the prediction accuracy of the minority class (bad credit) at the expense of the majority class (good credit).

To tackle this challenge, the author introduces a fresh resampling ensemble model based on data distribution, aptly named REMDD, tailored for imbalanced credit risk evaluation in the P2P lending domain. REMDD adeptly addresses class imbalance by employing a unique under sampling technique grounded in the majority class data distribution, termed UMCDD. This method rebalances the data by generating multiple balanced training subsets. Notably, Kmeans unsupervised clustering algorithms facilitate the acquisition of the majority class data distribution, setting the stage for comprehensive classification improvements.

The methodology in this paper doesn't stop at resampling. It leverages an ensemble of classifiers within the REMDD framework. These classifiers are chosen based on their overall performance, particularly the Area Under the ROC Curve (AUC), evaluated on a validation dataset. This strategic selection enhances the predictive capabilities of REMDD, ensuring it excels in both minority and majority class prediction.

To validate the prowess of REMDD, the author conducts extensive experiments using three real-world credit datasets from prominent P2P platforms, namely PaiPaiDai, Prosper, and Lending Club. The results of these experiments are compelling, demonstrating that REMDD not only performs admirably in predicting both minority and majority classes but also significantly bolsters the comprehensive classification performance for imbalanced credit risk evaluation in P2P lending. This places REMDD in a favourable position when compared to existing models.

In conclusion, this paper introduces REMDD as an innovative solution to address class imbalance in credit risk evaluation within the realm of P2P lending. The methodology presented, which combines ensemble classification with data distribution-based resampling, exemplifies the author's commitment to improving classification performance for both majority and minority classes. The empirical results underscore the effectiveness of REMDD and its potential to enhance comprehensive credit risk assessment, making it a valuable addition to the arsenal of tools for P2P lending platforms.

The network loan risk prediction model based on Convolutional neural network and Stacking fusion model

Meixuan Li, Chun Yu, Wei Liu (2021)

This paper emphasises on the shaping of how online lending platforms have become abundant, redesigning how we access loans. Amid this digital transformation, the need for robust risk prediction systems has never been more apparent. The authors introduce a ground breaking loan risk prediction model, cleverly named Stacking+CNN, that amalgamates Convolutional Neural Networks (CNN) and the Stacking algorithm. This fusion is aimed at enhancing the system's ability to extract crucial local spatial features from the Stacking algorithm's results while boosting overall generalization.

The journey of this model unfolds in distinct phases. Initially, a meticulous feature extraction process takes place using a wrapper method and the Variance Inflation Coefficient (VIF). The data is then funnelled into the first layer of the Stacking algorithm, where base learners receive training. The predictive outcomes from this stage are then fed into CNN, where further feature extraction occurs. The ultimate destination for these extracted features is a Support Vector Machine (SVM), which performs the critical task of risk prediction.

Empirical results validate the model's superiority over its peers. It outperforms both individual models and integrated models concerning predictive accuracy and recall rates, promising significant improvements for the online lending sector.

This paper's contributions are truly multifaceted:

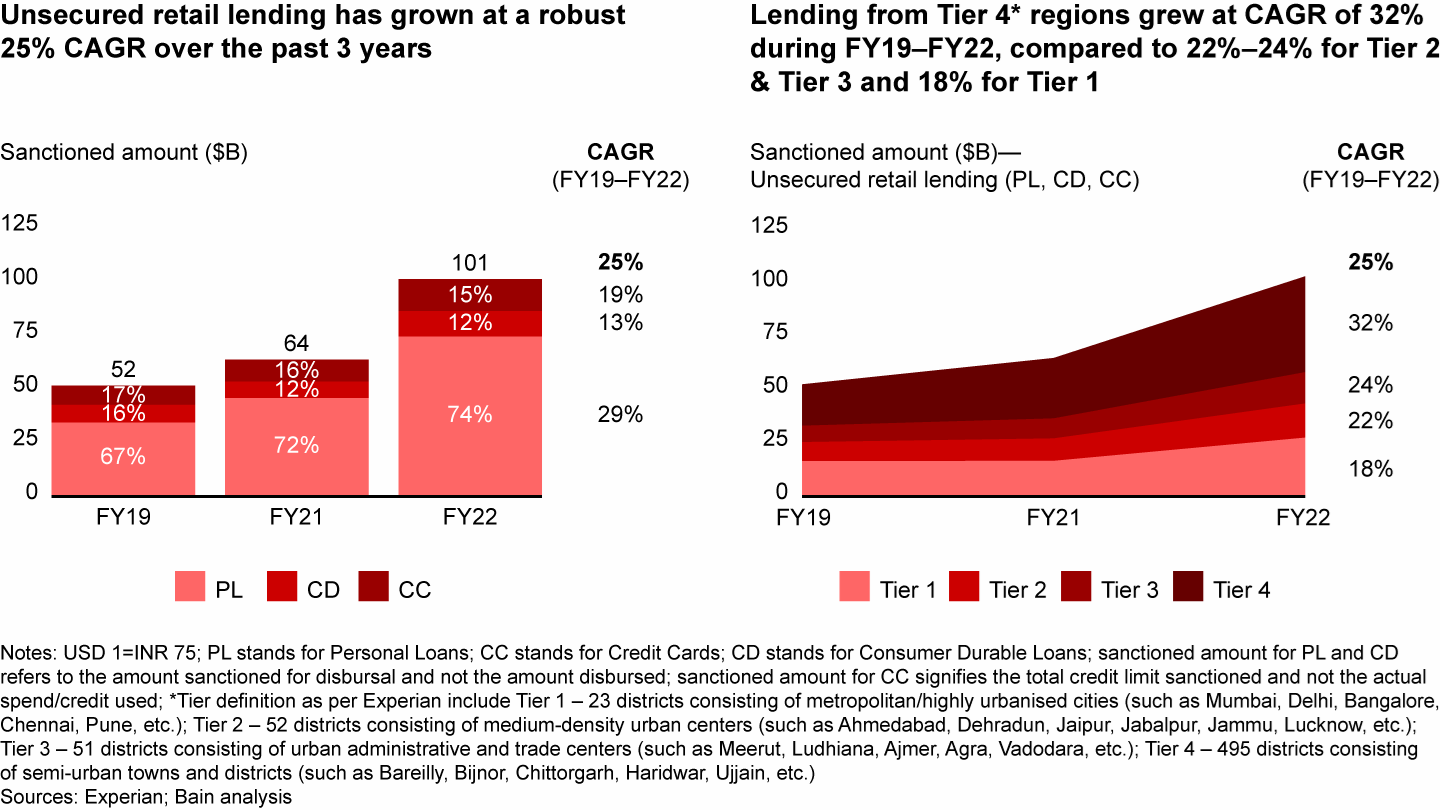
1. The introduction of this unique model tailored for online loan risk prediction is a standout contribution. It brings together CNN's prowess in feature extraction and the Stacking algorithm's ability to amalgamate various base models.
2. The paper introduces a original approach to feature extraction. It retains the predictive outcomes of individual base models while averting classification errors that can stem from improper threshold settings.
3. The authors employ a particular feature selection process involving both the wrapper method and the VIF. This not only streamlines data dimensionality but also addresses data imbalance using proven methods.
4. To reinforce generalization and moderate overfitting, the study implements k-fold cross-validation on the training data. This approach allows base models to learn different parameters during training, effectively countering overfitting concerns.

In summary, this research represents a pioneering effort to elevate loan risk prediction within online lending platforms. The innovative Stacking+CNN model, combined with advanced feature extraction and selection techniques, presents a potent tool for the financial industry. Its promise lies in heightened accuracy and improved risk assessment capabilities, ultimately transforming how we perceive and manage loan risks in the digital age.

# Problem Statement

Credit risk assessment is of paramount importance for fintech companies, particularly those operating in the online lending market, peer-to-peer lending, and RBI Approved loan apps which provide unsecured loans to their customers and institutions who deal with other vulnerable derivatives which are prone to credit defaults, runaways and customer churns. Traditional credit scoring methods, such as credit bureau scores, have limitations in capturing the complex and dynamic factors that influence borrowers' repayment behaviour as well factors that influence continued loyalty towards the financial institutions. Moreover, they heavily rely on historical data, which might not accurately reflect the current economic and social conditions of borrowers which is constantly evolving and diversifying. In the context of Bangalore, the Silicon Valley of India, the challenges posed by an unregulated credit market add to the urgency of developing a more advanced and robust credit scoring model.

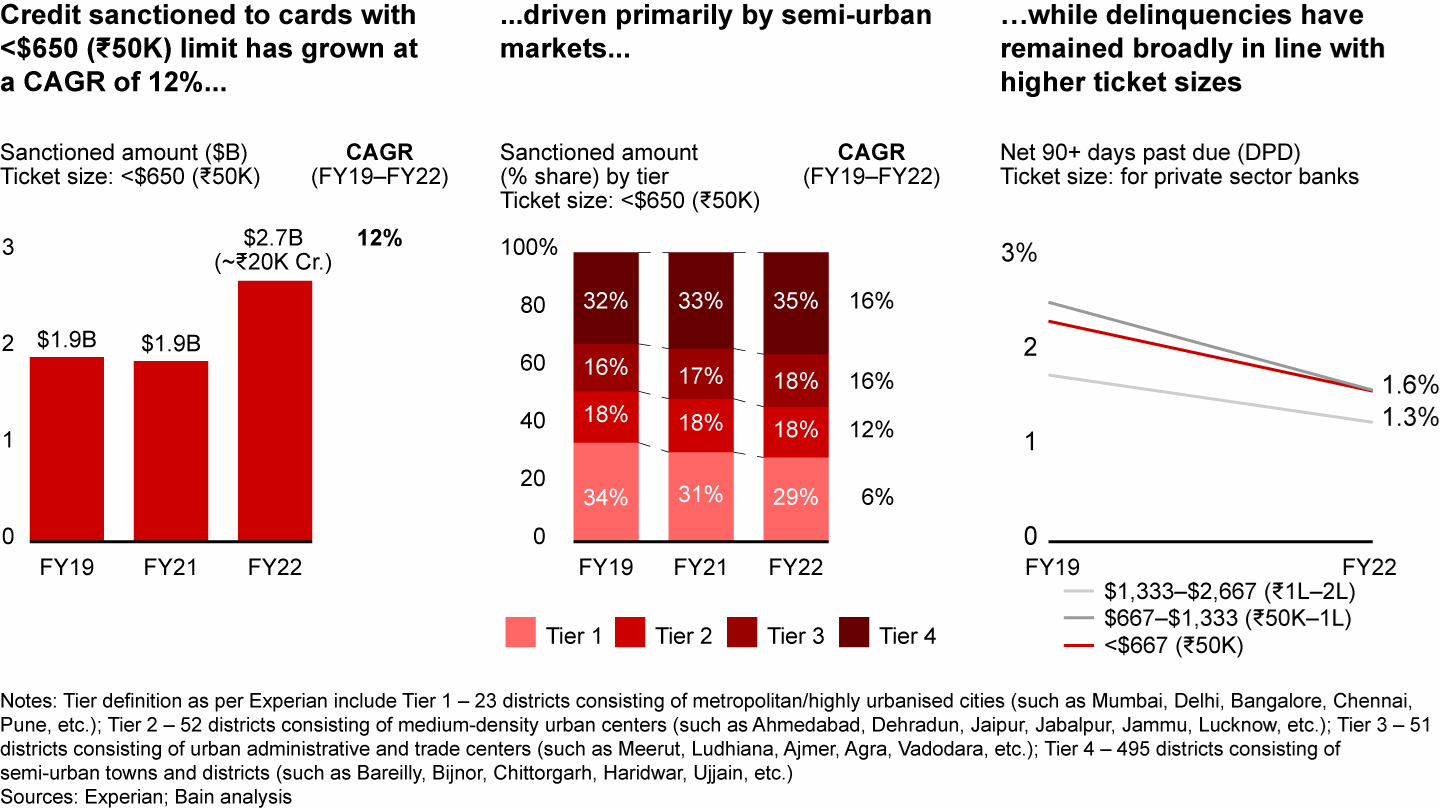
Despite the rapid growth of the online lending market in Bangalore, very little research has been conducted on credit risk assessment, specifically concerning the incorporation of various types of data such as demographic, financial, behavioural, and social media data. The lack of comprehensive studies leaves fintech companies with insufficient tools to accurately predict the probability of loan default and churn for loan applicants in this region.



**Unsecured Retail Lending Market Growth**

**Source: Bain & Company** [**[4]**](https://www.bain.com/insights/india-fintech-report-2022-sailing-through-turbulent-tides/)

Furthermore, the semi unregulated nature of the credit market in this spectrum of unsecured instant loan providers adds another layer of complexity. The absence of stringent regulations and monitoring mechanisms could lead to an increased risk of potential borrowers with poor creditworthiness slipping through the cracks. This further emphasizes the urgency for a reliable and transparent credit scoring model, with special reference to the creation of a stable, reliable and dynamic credit profile that can mitigate credit risks and empower fintech companies to make well-informed loan approval and risk management decisions.



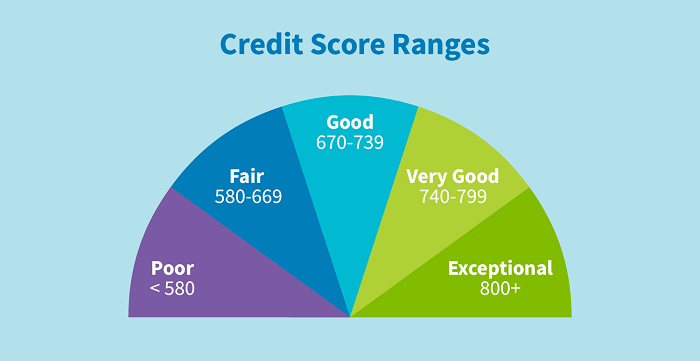
**Credit Profiling In Semi Urban Market**

**Source: Bain & Company** [**[5]**](https://www.bain.com/insights/india-fintech-report-2022-sailing-through-turbulent-tides/)

There has been significant credit sanctioning to cards with less than 50,000 Rupees throughout the country in the last 3 years. The net 90+ days past due ticket size has been steadily declining indicating customer churn. By leveraging diverse data sources collected through a comprehensive questionnaire from a target population of working professionals aged 25 and above in Bangalore, the model will create customer profiles based on a 4x4 matrix. Moreover, the model will be designed to handle challenges such as missing values, outliers, imbalanced classes, and nonlinear relationships in the data.

The expected outcome of this research is to provide fintech companies with a robust credit scoring tool, which is not singular rule based, that can accurately predict the probability of loan default and churn, thereby improving the assessment of creditworthiness for loan applicants in Bangalore's unregulated online lending market. Additionally, this study will contribute significantly to the existing literature on credit risk assessment in the online lending market, with specific reference to Bangalore. By shedding light on the factors that influence borrowers' repayment behaviour and loyalty, this research aims to offer valuable insights for informed decision-making in this fast-evolving and dynamic industry.

# Objectives of The Study

* To study and develop an advanced credit scoring model using machine learning for Fintech Loan Apps.
* To investigate and understand the loan default patterns, customer churn and credit behavior in Fintech Loan Apps Users.
* To employ data-driven approaches to identify key variables and their significance in predicting credit risks and loan defaults in Fintech Loan Apps Users.

# Scope of The Study

# This study dives into creating a cool credit scoring system that uses the magic of machine learning. We're mainly looking at the online lending world in Bangalore, especially those folks who use instant loan apps to get easy cash. In this ever-changing scene that includes things like people's backgrounds, money habits, and social stuff, our model turns out to be pretty good at predicting things.

# The Indian FinTech industry has experienced exponential growth in funding over the past few years, with massive investment from large venture capital and private equity firms. However, there are a few concerns that need to be addressed. These include data security and privacy risks in partnership scenarios, varied adoption of digital financial services across demographic groups, a dearth of financial literacy and awareness, IPO underperformance, and global geopolitical and macro-economic events making institutional investors cautious before big investments that are reflected in the first half 2022 funding trends, and the pace of changing regulations that keep FinTech companies on their toes. Despite these challenges, the Indian FinTech ecosystem is one of the largest in the world, with over 2,100 FinTech companies, and it is growing rapidly. The Indian FinTech industry is projected to reach $190 billion in revenue by 2030, with a transaction value of $138 billion in 2023. The collaboration between banks and FinTech players can boost credit access to the underserved segment and SMEs, which will continue to be a massive opportunity for FinTech. Currently, India is the third largest fintech market in the world, following UK and the USA. However the fintech startups, especially in Bangalore deal with the lack of empirical research into the various factors of financial behavior of Individuals in Bangalore.

# However, this project isn't solely about making predictions. It researches deeper into a significant issue within the credit industry, particularly in the online sphere where regulations can be lax. By connecting the capabilities of keen expertise, this study aims to empower small and medium-sized fintech companies to make informed and prudent decisions. These decisions are strategically aimed at mitigating the risks associated with lending, ultimately fostering a robust, secure, and sustainable lending environment. At core, this study seeks to thoroughly investigate and comprehend the diverse financial behaviors of borrowers, shedding light on the hidden patterns in how customers who utilize instant loan apps and fintech, even with less-than-ideal credit histories, manage their finances.

# Limitations of The Study

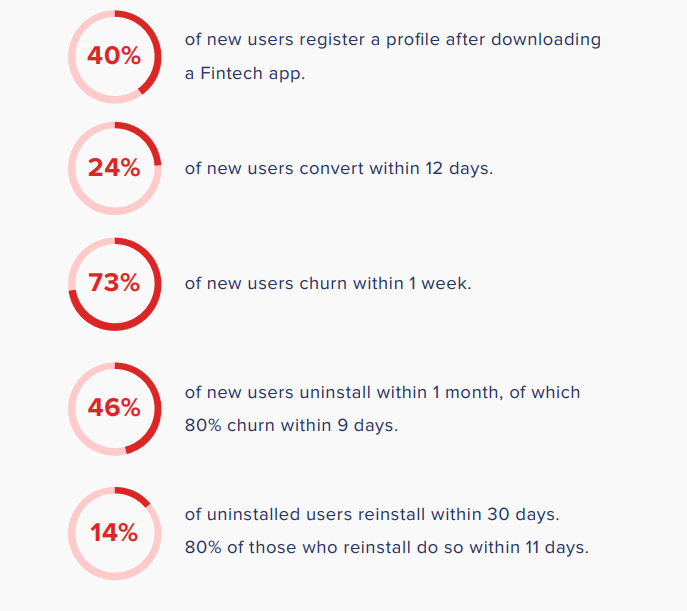
1. Focusing solely on a particular demographic narrows the scope of the findings, thereby jeopardizing the model's broader relevance to a diverse range of borrower profiles.
2. The seamless implementation of the credit scoring model into fintech operations may encounter difficulties attributable to intricacies and the potential requirement for recalibration, thereby affecting its practical integration.
3. Evolving borrower behaviours, economic shifts, and the emergence of new platforms extend beyond the confines of historical data, exerting an influence on the predictive capacity of the model and underscoring the imperative of consistent evaluation and adjustment.

# Research Gap

In the rapidly growing fintech landscape of Bangalore, India, where fintech loan applications play a pivotal role in financial services, a conspicuous research gap exists. Despite the surge in fintech adoption, particularly in Bangalore, there is an alarming scarcity of empirical research addressing the critical issue of predictive modelling for loan defaults within this specific geographic and economic context.

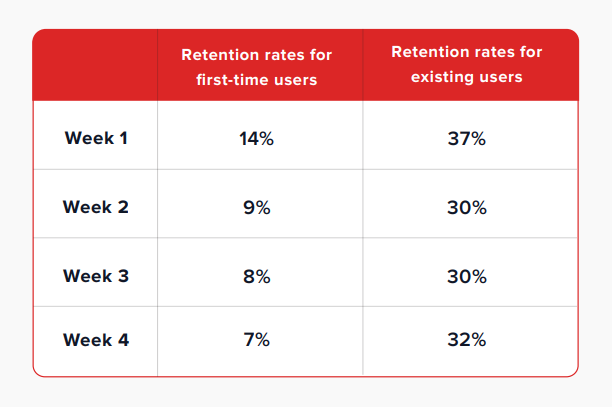
Bangalore, often referred to as the Silicon Valley of India, is home to a thriving fintech ecosystem. The customer churn rate in the Indian fintech sector averages a staggering 73%, a figure that has substantial implications for business sustainability and profitability. Furthermore, the credit default rate in Bangalore stands at 1.94%, shedding light on unique regional economic and sociocultural factors shaping credit behavior.

Here are some of the industry benchmarks



**Source : Clevertap [**[**6**](https://brandequity.economictimes.indiatimes.com/files/cp/1139/cdoc-1658399014-Fintech_Retention_Guide.pdf)**]**

The existing literature, although rich in insights about global fintech trends, does not sufficiently delve into the intricacies of the Bangalore fintech loan market. This notable research gap calls for a dedicated empirical approach that integrates customer churn analysis, an in-depth examination of credit behaviour, and predictive modelling specific to fintech loan applications. By focusing on Bangalore, this research aims to unravel the distinctive dynamics influencing customer retention and credit risk within the context of fintech loans, taking into account regional idiosyncrasies.



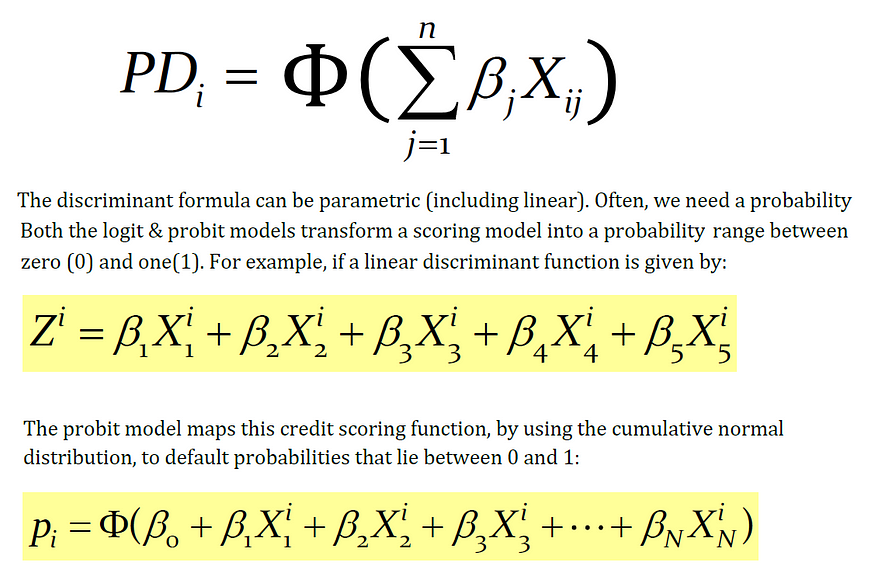
**Source : Clevertap** [**[7]**](https://brandequity.economictimes.indiatimes.com/files/cp/1139/cdoc-1658399014-Fintech_Retention_Guide.pdf)

Bridging this research gap is imperative for both academia and the fintech industry. It not only facilitates a nuanced understanding of the Bangalore market but also provides actionable insights to optimize fintech loan platforms in this high-growth geography. Exploring this untapped territory holds the potential to reshape the fintech landscape in Bangalore, ultimately contributing to enhanced customer experiences, lower default rates, and sustainable business growth.

# Research Design

The research methodology initiates with the meticulous collection and pre-processing of a comprehensive dataset sourced from the working professionals in the dynamic realm of Bangalore. This dataset encompasses a rich array of variables spanning demographics, financial metrics, behavioural patterns, and social attributes. The data undergoes a rigorous cleansing and refinement process to ensure its quality and relevance, thus paving the way for subsequent rigorous analysis.

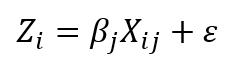
With a robust dataset in place, the construction of clear and testable research hypotheses ensues, encompassing both null and alternative stances. The investigative prowess is further amplified through the employment of Probit and Logit Regression methodologies. These techniques unfurl the intricacies of causal interplay among variables, particularly attuned to the domain of loan default propensities and churn dynamics. By unearthing these causal threads, the research aims to illuminate the driving forces shaping the lending landscape. The first objective of the study is to predict customer churn by analyzing customer behavior and preferences.



**Probit Regression Model Equation   
Source: Medium** [**[8]**](https://medium.com/@polanitzer/the-probit-model-in-python-predict-default-among-u-s-corporates-c0f32cb4b5c2)

The linear probability model is a statistical approach that leverages economic and financial data to calculate the Probability of Default (PD). In this model, we conduct a linear regression where the dependent variable, denoted as "Z," takes on a value of 1 when a default occurs, and a value of 0 when the firm successfully meets its debt obligations.

The independent variables, on the other hand, encompass various risk metrics that mirror the financial health and strength of the firm. These metrics can include factors such as financial leverage ratios, liquidity ratios, or profitability ratios. The model is applied across multiple firms, employing a linear regression framework of the following form:



Where:

**Xij** — The explanatory variables (financial ratios) of firm i;

**βj** — A coefficient that measures the importance of a variable in explaining default

This approach allows us to estimate the likelihood of default for each firm based on their unique financial characteristics and risk metrics.

The Probit model serves to address the limitations found in the linear probability model by rectifying the distortion and constraining the probability of default to a more realistic range between 0 and 1. In this model, the dependent variable, denoted as the "explained variable," is binary, taking on one of two values: 1 signifies a firm that has experienced default, while 0 designates a stable firm.

This statistical model employs a combination of financial and other relevant variables to make predictions regarding the probability of default for a given firm. It operates under the assumption that this probability adheres to a cumulative standard-normal distribution, a statistical concept that inherently restricts the predicted probability within the defined bounds of 0 to 1.



Where:

**F(Zi)** — The firm’s cumulative probability of default

**Zi** — The value obtained from estimating the Probit model

Φ**(Zi)** — The cumulative standard-normal distribution function from minus infinity (**-∞**) to the point Zi (i.e., the number of standard deviations)

The Probit model and the Logit model diverge in their underlying distribution assumptions for the probability of default. The Probit model posits that the firm's probability of default follows a cumulative standard-normal distribution. In contrast, the Logit model assumes a logistic distribution for this probability.

Interestingly, it's worth noting that by applying a suitable coefficient to the results generated by the logistic distribution, you can effectively transform it into the distribution of the Probit model. This mathematical transformation allows for a bridge between the two models, aligning their results under certain conditions.

A well-structured questionnaire will be designed to gather relevant information from a sample population of 320 individuals. The questionnaire will cover demographics, financial details, past loan history (if applicable), and banking behavior. The collected data will be preprocessed to handle any missing values and ensure data quality. Machine learning classifiers, including KNN, Support Vector Machine (SVM), Decision Tree, and Random Forest, will be employed to explore the likelihood of churn. The study seeks to identify key factors contributing to customer churn, empowering banks to develop proactive strategies for customer retention and engagement.

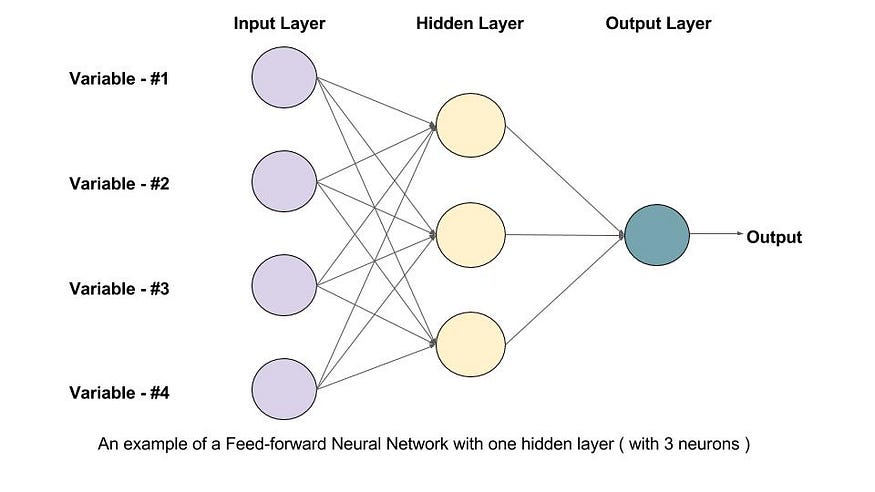
|  |  |  |
| --- | --- | --- |
| Steps | Tools | Impact on Research |
| Data Collection and Pre-processing | Google Forms, Excel | 34% |
| Hypothesis Formulation | N/A | 14% |
| Building Model Using Alyuda Neural Network | Python - Tensorflow, Pytorch | 28% |
| Probit and logit Regression to identify the causation effect | Python - Statsmodel | 12% |
| Two Way MANOVA For acceptance or failure to accept Null Hypothesis | Python –Statsmodel | 12% |

The second objective is to assess the creditworthiness of loan applicants in the instant loan app sectors employed through fintech platforms. The same questionnaire data will be used to build a comprehensive credit scoring model. Machine learning techniques will be applied to handle the data, including handling missing values, encoding categorical variables, and scaling numerical features. By training various models, such as Logistic Regression, Decision Trees, Random Forest, or Gradient Boosting, the study will determine the most accurate model for predicting the probability of loan default and churn. The credit scoring model will serve as a transparent and reliable tool for fintech companies and banks to make informed decisions on loan approval and risk management.

The research will evaluate the performance of both the churn prediction and credit scoring models using appropriate evaluation metrics, such as accuracy, precision, recall, F1-score, and ROC curves. The results of this study will offer valuable insights into customer behavior, loan risk assessment, and the factors influencing creditworthiness and customer churn in the banking sector. For this the author will be employing FFNN.

***FFNN***

A **feedforward neural network** is an artificial neural network wherein connections between the nodes do not form a cycle. As such, it is different from its descendant: **recurrent neural networks**. The feedforward neural network was the first and simplest type of artificial neural network devised.



**Source: Medium** [**[8]**](https://medium.com/mlearning-ai/training-feed-forward-neural-network-ffnn-on-gpu-beginners-guide-2d04254deca9)

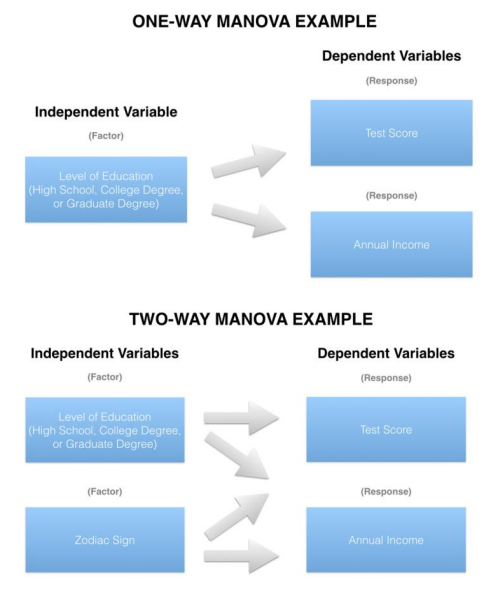
The crux of the research lies within the Model Building phase, where a sophisticated Machine learning technique, Alyuda Neural Network is deployed to actualize an intricate credit scoring model. This neural architecture is primed through meticulous training, leveraging the meticulously pre-processed dataset. The network's architecture optimization process is geared towards harnessing optimal predictive efficacy, thereby enhancing its proficiency in extrapolating meaningful credit assessment insights.

The spotlight then shifts to hypothesis validation, orchestrated through the Two-Way Multivariate Analysis of Variance (MANOVA) technique. This analytical behemoth elucidates latent differentiations in means across various groupings and delves into the convoluted interactions interwoven within variables. MANOVA's analytical prowess equips the research with incisive insights to validate or invalidate formulated hypotheses, thereby underscoring the empirical foundation of the study.

**MANOVA**

MANOVA, which stands for Multivariate Analysis of Variance, is a statistical technique used for analyzing data that includes more than one dependent variable simultaneously. It is employed to test hypotheses related to the impact of one or more independent variables on two or more dependent variables.

The primary distinction between ANOVA (Analysis of Variance) and MANOVA lies in the "M," which signifies "multivariate." Essentially, MANOVA extends the principles of ANOVA to situations where you have multiple continuous response variables. Like ANOVA, MANOVA can be applied in both one-way and two-way designs. The key differentiator between one-way and two-way MANOVA is the number of factor variables involved, with the former having one factor variable and the latter involving two or more factor variables.

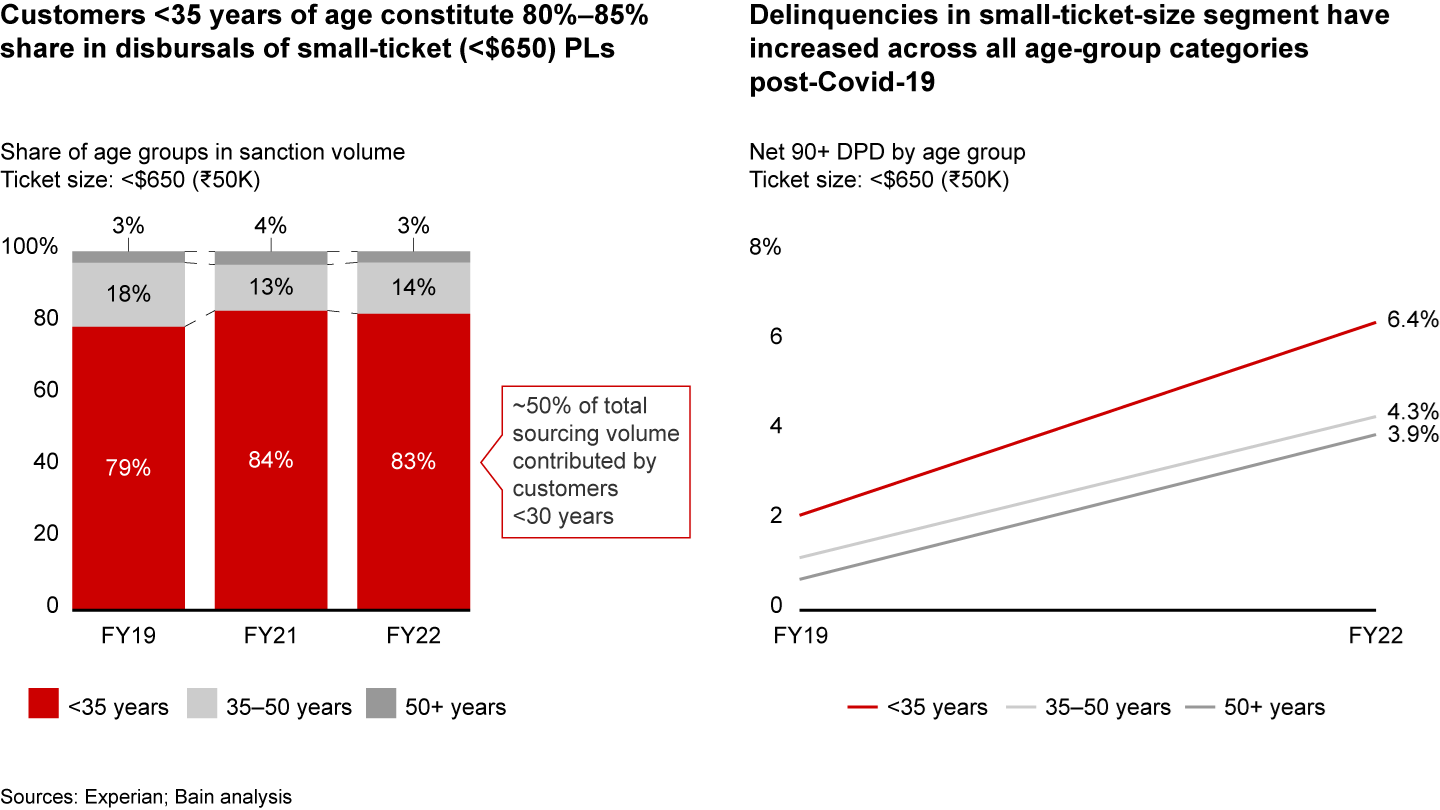


Source: Medium [[10]](https://medium.com/nerd-for-tech/everything-about-manova-and-mancova-4c1c237af464)

Data privacy is a fundamental ethical principle that this research prioritizes diligently. It underscores the research’s commitment to maintaining the confidentiality of participants and upholding the integrity of this research. The author views it as a moral obligation to protect the rights and trust of those who share their data for this scholarly investigation.

The research journey reaches its peak during the interpretation of results. At this stage, the author carefully examines the insights gathered and the emerging relationships within the data. The research’s aim is to conduct a thorough analysis that enables us to clearly articulate conclusions about causal influences, the validation of hypotheses, and the subtle impact of individual variables on the complex realm of credit evaluation.

# Sample Design



**Source: Bain & Company [11]**

This research would have a sample design to investigate the predictive modelling of loan defaults and customer churn in Bangalore's online lending market, with a particular focus on customers aged 24 to 35. This age group is of significant interest as they contribute between 80% and 85% of disbursals for small ticket loans in India.

**1. Defining the Population:**

The population under study comprises customers who have availed small ticket loans ( this could be either personal or entrepreneurial loans, this could formal or informal credits) through fintech loan applications in Bangalore, India.

**2. Determining the Sample Size:**

The author has determined that a sample size of 320 is appropriate for this study. This sample size is chosen to balance the research goals and available resources effectively.

**3. Stratified Sampling:**

To ensure that that the sample is representative of the age groups contributing significantly to disbursals, the author will employ a stratified sampling approach. This approach will involve dividing the population into two strata:

a. Customers aged 24 to 33.

b. Customers aged 33 to 42.

**4. Allocation Based on Contribution:**

To allocate the sample size within each stratum, the author will take into consideration the contribution of each age group to disbursals. Given that customers aged 24 to 30 contribute between 80% and 85% of disbursals, I will allocate 85% of my sample (272 cases) to this stratum and the remaining 15% (48 cases) to the age group of 30 to 35.

**5. Random Sampling:**

Within each stratum, the author will employ random sampling techniques to select specific individuals or cases. This randomization process ensures that my sample accurately represents the respective age groups in the population.

**6. Data Collection:**

Upon establishing the sample, the author will proceed to collect data on a range of variables, including demographics, social factors, financial behaviours, and historical borrower information. This data will enable the author to meet the objectives of developing an advanced credit scoring model and accurately predicting loan default and churn likelihood while identifying and mitigating various risk factors in digital lending.

This sample design has been carefully crafted to ensure the representativeness and validity of my findings, aligning with the objectives of my research on predictive modelling of loan defaults and customer churn in Bangalore's online lending market.

# Hypothesis

Null Hypothesis (H0): There is no significant relationship between the diverse variables encompassing demography, social, financial behaviours, and historical information of borrowers and their likelihood of loan default and churn. Incorporating these variables into an advanced machine learning credit scoring model does not improve the accuracy of predictions or contribute to the identification and mitigation of various risk factors.

Alternative Hypothesis (H1): There is a significant relationship between the diverse variables and the likelihood of loan default and churn. Furthermore, incorporating these variables into an advanced machine learning credit scoring model leads to more accurate predictions and contributes to the identification and mitigation of various risk factors.

# Data Pre-processing

Firstly, the data pre-processing phase involves addressing missing values and handling outliers. To tackle missing values, the study employed a method where the median value was used to fill in the gaps. Additionally, for the time variable, a conversion into numeric values was carried out.

The second aspect of data pre-processing revolves around outlier treatment. Standard deviation was utilized to identify outliers, with the "3σ" principle serving as the threshold for identifying and subsequently removing these outliers. In the context of this principle, outliers are values lying outside the interval of (μ-3σ, μ+3σ), where μ represents the mean value. It's worth noting that when data adheres to a normal distribution, the probability of a value falling outside this interval is less than 0.3%. Following the removal of outliers, the final dataset comprised 280 samples. This dataset served as the basis for the inputs into the predictive model.

Moreover, the study recognized the importance of handling invalid features that lack practical significance. These features can significantly increase operational complexity and pose challenges in data analysis. Furthermore, the presence of correlations between features can complicate the analysis process. Some variables may exhibit poor stability, which can adversely affect prediction outcomes. To address these issues, the study employed various methods, including deletion, principal component analysis (PCA), feature interaction analysis, and the population stability index (PSI). These techniques were applied to process the dataset's variables.

The ultimate goal of these techniques was to minimize information loss from the original dataset while reducing the number of variables requiring analysis. This comprehensive analysis not only enhanced the validity of the variables but also improved the accuracy of prediction results.

# Population Stability Index

This non-linear interaction approach effectively enhances the volume of information. It's worth noting that the quantities of behaviors in anonymous variables are often closely linked to the loan level. The interplay between these mutually influencing factors can significantly enhance the model's capacity to learn.

Additionally, the study employed the Population Stability Index (PSI), similar to the approach adopted by Huang et al. (2022). PSI quantifies the deviation between the model's predicted value and the actual value, as demonstrated in Equation (13). The dataset is divided into five segments, with 80% designated as the training set and the remaining 20% as the test set. In this context, Ai represents the proportion of variable distribution within the training set, while Ei corresponds to the proportion of variable distribution within the test set.



Model stability is considered exceptionally high when the PSI is less than 0.1. However, if the PSI value exceeds 0.25, it indicates poor model stability. Specifically, if the PSI value for the "net profit" variable exceeds 0.25, that variable is removed. Nevertheless, the features obtained through Principal Component Analysis (PCA) are retained to enhance the stability of the feature data. As a result, the features npca1-1, npca1-2, and npca1-3 are retained.

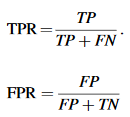
Evaluation indicators

The evaluation of the model's performance is based on several key indicators, including the Area Under Curve (AUC), precision, accuracy, and Kolmogorov-Smirnov (KS) values. Before delving into these evaluation indicators, let's establish a fundamental understanding using a confusion matrix as shown in Table 7.

* True Positive (TP): This represents the positive data correctly classified by the model.
* False Positive (FP): It refers to negative data that the model incorrectly classifies as positive.
* True Negative (TN): This indicates that the model correctly identifies negative sample data.
* False Negative (FN): It represents positive data that the model incorrectly classifies as negative.

The Receiver Operating Characteristic (ROC) curve is a crucial tool for evaluating model performance. The ROC curve's y-axis represents the True Positive Rate (TPR), while the x-axis represents the False Positive Rate (FPR). These rates are calculated using Eqs. (14) and (15).

AUC, or the Area Under Curve, measures the area enclosed by the ROC curve. It ranges from 0.5 to 1, with 0.5 indicating the lowest authenticity and no practical application value, while a value closer to 1 indicates a more reliable detection algorithm. To compare two models, we can check if the ROC curve of Model A entirely encompasses the ROC curve of Model B. Additionally, we can compare the areas enclosed by the ROC curve and the axes to determine which model performs better. A larger enclosed area suggests superior model performance.



The precision is only for correct positive-case data, which manifests as the extent to which the predicted positive data are true positive data.



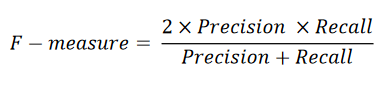
Accuracy is a widely employed metric to gauge the performance of a classification model. It quantifies how many samples the classifier has accurately predicted. To calculate accuracy, the model is employed to classify the test dataset, and the accuracy is determined by dividing the number of correctly predicted samples by the total number of samples, expressed as a percentage.



The Kolmogorov-Smirnov (KS) value is a crucial indicator used to evaluate the risk discrimination capabilities of a model. It is defined as the highest absolute difference between the True Positive Rate (TPR) and the False Positive Rate (FPR), as illustrated in Eq. (18). This value assesses the degree of separation in cumulative distribution between excellent and poor samples. In simpler terms, the KS index measures how effectively the model distinguishes between low-risk and high-risk cases. Typically, a KS value above 0.2 is considered acceptable, and the higher the KS score, the better the model's risk discrimination ability.



*F-measure is the harmonic average of precision and recall”, and its equation as follows:*

**

Empirical Analysis

1. LDA

Linear Discriminant Analysis (LDA) is a statistical technique initially proposed by Fisher in 1936. It was first recognized in the field of credit scoring by Durand in 1941, and subsequently, LDA became one of the early methods used to develop credit-scoring models.

In a credit-scoring classification problem, you typically have a dataset comprising n customers, each with m characteristics or variables (x = x1, x2, ..., xm) used to classify customers into groups, often denoted as good/0 or bad/1 loans. These customers are divided into two groups: ng, representing customers with good loan status, and nb, representing customers with bad loan status. The objective of LDA is to estimate the probability of a customer belonging to either the good or bad loan group, denoted as p(y|x), given a vector of their characteristics, features, or variables, x.

LDA achieves this by linearly combining the independent variables to classify the objects into their appropriate groups or classes. This linear combination can be **expressed as:**

**Z = β₀ + β₁X₁ + β₂X₂ + ... + βₙXₙ.**

Here, Z represents the discriminant score, β₀ is the intercept term, and βᵢ represents the coefficients or weights related to the variables xᵢ (i = 1, 2, ..., n). This equation constructs the discriminant model, which aids in predicting and classifying a customer's loan into the suitable credit group.

In LDA, several assumptions must be met, including the independence of data, normal distribution, and homogeneity of variance and distribution between good and bad loans. The coefficients β = (β₁, β₂, ..., βₙ) are adjusted based on the covariance matrix and the mean feature vectors of the two groups of loans. Once the coefficient values are obtained, the discriminant score can be calculated.

Classification occurs when a new loan application is received. It is classified by projecting it onto the maximally separating direction represented by the function Z = βᵀx + β₀. Classification is achieved by comparing βᵀx to a threshold, denoted as Tc. If βᵀx < Tc, the loan is classified as good; otherwise, it's classified as bad. The threshold is determined based on the prior probabilities of the loan customers in each loan group. The goal is to select a projection that maximizes the distance between the good and bad loans while minimizing it within each group.

LDA is a straightforward method for classifying linear data. However, it has limitations, particularly when dealing with nonlinear data, as it assumes linear relationships between variables. Other limitations include improper grouping definitions, errors in classification estimation, and issues with prior probabilities. Despite these limitations, LDA has found wide application in building scoring models.

1. LR

Logistic Regression (LR) is another widely used statistical technique, and it's the most popular tool for classification problems. In many cases, methods like Linear Discriminant Analysis (LDA) and regression models, such as linear regression, provide continuous output that spans a range from negative infinity to positive infinity. However, in credit-scoring, classification is typically binary or dichotomous, where the decision is simplified to either 0 (grant the loan) or 1 (reject the loan). This binary classification led to the development of Logistic Regression to address this specific problem.

LR works by examining the relationship between multiple independent variables and the probability of a loan being granted. It achieves this by fitting the variables to a logistic curve. As illustrated in Figure 2.2, the LR model's outcome is either categorized as a good loan (0) or a bad loan (1). In this context, linear relationships are insufficient for modeling, as the classification must be estimated at a continuous level.

Instead, LR constructs an 's-shaped' logistic curve where the values range between 0 and 1. This curve effectively expresses the relationship between the independent variables and the probability of a binary outcome, using a non-linear function of independent variables, as represented in Equation (2.3).

3.LIME

When dealing with complex machine learning models for loan default prediction, LIME (Local Interpretable Model-agnostic Explanations) can be a valuable tool to enhance the interpretability of model results. It allows managers and decision-makers to gain a deeper understanding of how these complex models arrive at their predictions. Instead of passively accepting model outcomes, managers can actively engage with the model's insights while ensuring that the model remains credible. This leads to more accurate and effective decision-making.

LIME offers several advantages, including versatility, relevance, and flexibility. It can be employed to select representative samples for analysis based on specific requirements. For instance, in the context of this study, variable importance was analysed separately for different probability intervals by stratifying default probabilities. Compared to some other methods, LIME results are relatively straightforward to interpret.

However, LIME also has its limitations. Firstly, due to its local interpretability nature, it may not be applicable to interpret all samples. Additionally, LIME can be time-consuming to apply since it requires training and explaining a model on a local level. Defining similarity when selecting perturbation data can be challenging, leading to reduced stability. Different perturbed datasets can yield different conclusions. Furthermore, given the diversity of prediction models, the output results and formats can vary significantly. Therefore, integrating LIME with certain prediction models may require adjustments to align the output format, which can add complexity to the interpretation process. In the experiment described in this study, the output format of the LightGBM predictive model was modified to suit the needs of building explanatory models, resulting in improved simplicity.

Chapter 4 Data Analysis & Interpretation

H1: Young borrowers (22-32) are more likely to default on their loans compared to older borrowers.

Interpretation: The data suggests that borrowers aged 22-32 exhibit a higher default rate on loans compared to older borrowers, supporting the hypothesis that younger individuals are more prone to loan defaults. Lending institutions should consider this age-related risk factor for informed decision-making and risk management strategies.

H2: Borrowers who default on their loans have a significantly lower median income compared to borrowers who do not default.

Interpretation: The data strongly supports hypothesis H2, with defaulting borrowers having a significantly lower median income (₹36,000) compared to non-defaulting borrowers (₹79,600). This underscores the link between lower income and higher loan default risk.

The data confirms hypothesis H2, revealing that borrowers who default on their loans typically have lower income. For those with defaults, income ranges from ₹20,750 to ₹83,700, while non-defaulting borrowers have higher income, ranging from ₹36,700 to ₹123,000. This underscores the association between lower income and a higher risk of loan default.

H3: There is a strong correlation between Age and Loan Amount borrowed by a respondent on average

|  |  |  |  |
| --- | --- | --- | --- |
| |  |  | | --- | --- | | Correlation between Age And Employment Tenure |  |   Correlation between Employment Tenure and Monthly Income | ***0.747677***  0.754314 |

A notable inverse correlation exists between age and the loan amount borrowed, as evidenced by a correlation coefficient of -0.0192. This indicates that, as individuals' age increases, there is a discernible trend towards borrowing smaller loan amounts. The data suggests that as people grow older, they tend to opt for more conservative loan amounts, reflecting a potential shift in financial preferences and risk tolerance.-

H4: Borrowers with longer employment tenures (>24 Months) are less likely to default on their loans.

The data confirms hypothesis H4, showing that borrowers with employment tenures exceeding 24 months are less likely to default on their loans. Only 43 out of this group defaulted, while 176 did not. In contrast, borrowers with tenures less than 24 months had a higher default rate, with 592 defaults and only 35 non-default cases. This highlights the strong connection between longer employment tenures and a decreased risk of loan default.

H5: Borrowers with a prior history of default (Credit card overdues, credit card defaults) are more likely to default on their current loans.

The data strongly supports hypothesis H5, indicating that borrowers with a prior history of credit default (Credit Default: Yes) are significantly more likely to default on their current loans. Among this group, 588 borrowers defaulted, while only 39 defaulted in the group with no prior credit default history. This highlights the pronounced link between a credit default history and an increased risk of loan default, emphasizing the importance of considering this factor in loan risk assessment.

H6: Personal loans have a higher default rate compared to loans with other intents (e.g., educational, debt consolidation).

The data clearly indicates that within the category of "Personal loans" (comprising "Debt Consolidation," "Educational Loan," and "Home Loan"), there are 444 defaults in total (291 for Debt Consolidation, 70 for Educational Loan, and 83 for Home Loan). In comparison, other loan intents, such as "Medical" and "Ventures," have a significantly lower combined total of 183 defaults (154 for Medical and 29 for Ventures). These numbers strongly support the hypothesis that "Personal loans" have a notably higher default rate than other loan types, underscoring the importance of distinguishing between loan intents when evaluating default risk.

H7: An awareness of one's credit rating is correlated with a decreased probability of loan default among borrowers.

In the dataset, individuals who responded "Yes" (indicating awareness of their credit rating) numbered 204, while those who responded "No" (indicating lack of awareness) totaled 37. This data suggests a correlation between being informed about one's credit rating and a decreased likelihood of loan default. It implies that borrowers who actively monitor their credit rating may exhibit a lower propensity for defaulting on loans, underlining the potential importance of credit rating awareness in managing loan default risk.

H8: Venture loans have the lowest default rate among all loan types.

The data unequivocally demonstrates that "Venture loans" exhibit the lowest default rate among all the loan types, with a mere 29 instances of default. This finding underscores the superior credit performance associated with "Venture loans" in comparison to the other listed loan categories.

H9: Borrowers who have received financial counselling are more likely to avoid loan default compared to those who have not.

Financial counselling entails an individual's inclination to seek guidance from a certified financial expert or other trustworthy sources, such as YouTube and various social media platforms. These platforms offer a range of strategies and suggestions to help individuals save money, boost their credit scores, and become more aware of potential financial borrowing pitfalls.

It was observed for those who did default their loans, almost 591 of them did not receive any type of financial counselling in the past or were not financial literate enough to make smart choices.

Hypothesis H9 is strongly corroborated by the data, indicating that borrowers who have received financial counseling are more apt to avoid loan default as compared to those who have not. Among the 189 borrowers who received financial counseling, only 36 experienced loan default. In contrast, a significantly higher number of 591 borrowers defaulted on their loans among those who did not receive financial counseling. This striking dissimilarity in default numbers highlights the potential efficacy of financial counseling in mitigating the risk of loan default.

H10: Borrowers who use digital lending services frequently have a higher default rate than those who use them infrequently.

Within the group of borrowers who defaulted, a notable proportion of 480 individuals were identified as frequent users of digital lending services, utilizing them more than three times per quarter. Additionally, 84 borrowers were classified as frequent users, surpassing usage once per quarter. Furthermore, 57 borrowers belonged to the occasional usage category, indicating awareness and prior experience with digital lending services. These findings validate the hypothesis, indicating a higher default rate among borrowers who extensively utilize digital lending services compared to those who use them less frequently.

H11: Borrowers with collateral assets are less likely to default on their loans compared to those without collateral.

Hypothesis H11 is strongly supported by the data, indicating that borrowers with collateral assets are less prone to default on their loans as compared to those without collateral. Out of the borrowers with collateral assets ("Yes"), only 39 defaulted on their loans. Conversely, a significantly higher number of 588 borrowers defaulted among those without collateral assets ("No"). This significant difference in default counts underscores the potential protective impact of collateral assets in reducing the risk of loan default, emphasizing their importance for lenders in risk assessment.

H12: Borrowers who report maintaining a budget are less likely to default on their loans than those who do not.

Among the total defaulters (627), a significant majority of 582 individuals did not follow any specific budgeting practices. This lack of budgeting is directly linked to their inability to track where their money is being allocated after receiving their monthly recurring revenue. Consequently, this absence of financial awareness facilitates arbitrage spending, ultimately contributing to loan defaults.

The data strongly supports hypothesis H12, which suggests that borrowers who report maintaining a budget are less likely to default on their loans compared to those who do not.

Among borrowers who reported that they maintain a budget, only 45 defaulted on their loans, while 208 did not default. In contrast, among those who stated that they do not have a specific budget, a significantly larger number, 582, defaulted on their loans, while only 33 did not default. This substantial difference in default counts underscores that borrowers who maintain a budget are indeed less likely to default on their loans. It highlights the positive impact of budget management in reducing the risk of loan default.

Model Building Using Python