# Abstract

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

Paytm, a prominent lender, uses a credit scoring model that takes into account alternative data, such as the quality of digital transactions, to evaluate loan outcomes. However, this approach raises questions about the equitable implementation of opaque models that reshape risk assessment. The study broadens its investigation to look at fintech-bank collaborations and the capital allocation of banks participating in instant loan platforms to gain a better understanding of the situation. This analysis highlights the growing emphasis on loans facilitated by collaborations between fintech start-ups and banks, indicating a significant shift in bank balance-sheet allocation.

Alternative credit scoring systems are becoming more important in the calculative infrastructure, allowing specific institutions to overcome the challenges associated with risk-based pricing. These systems also serve as strategic collaboration points for technology start-ups and financial institutions, allowing them to capitalize on new revenue streams. The study reveals an intriguing dichotomy in model performance when loans are used for personal and business purposes. When trained on the entire dataset, the first phase outperforms the second phase, which is focused on the small business subset. This finding highlights potential differences in the screening and default prediction processes for small business loans.

This study demonstrates the power of machine learning algorithms in improving credit risk assessment, making more informed lending decisions, and accurately forecasting loan defaults. Financial institutions can improve risk management frameworks and strengthen lending practices by leveraging artificial intelligence.

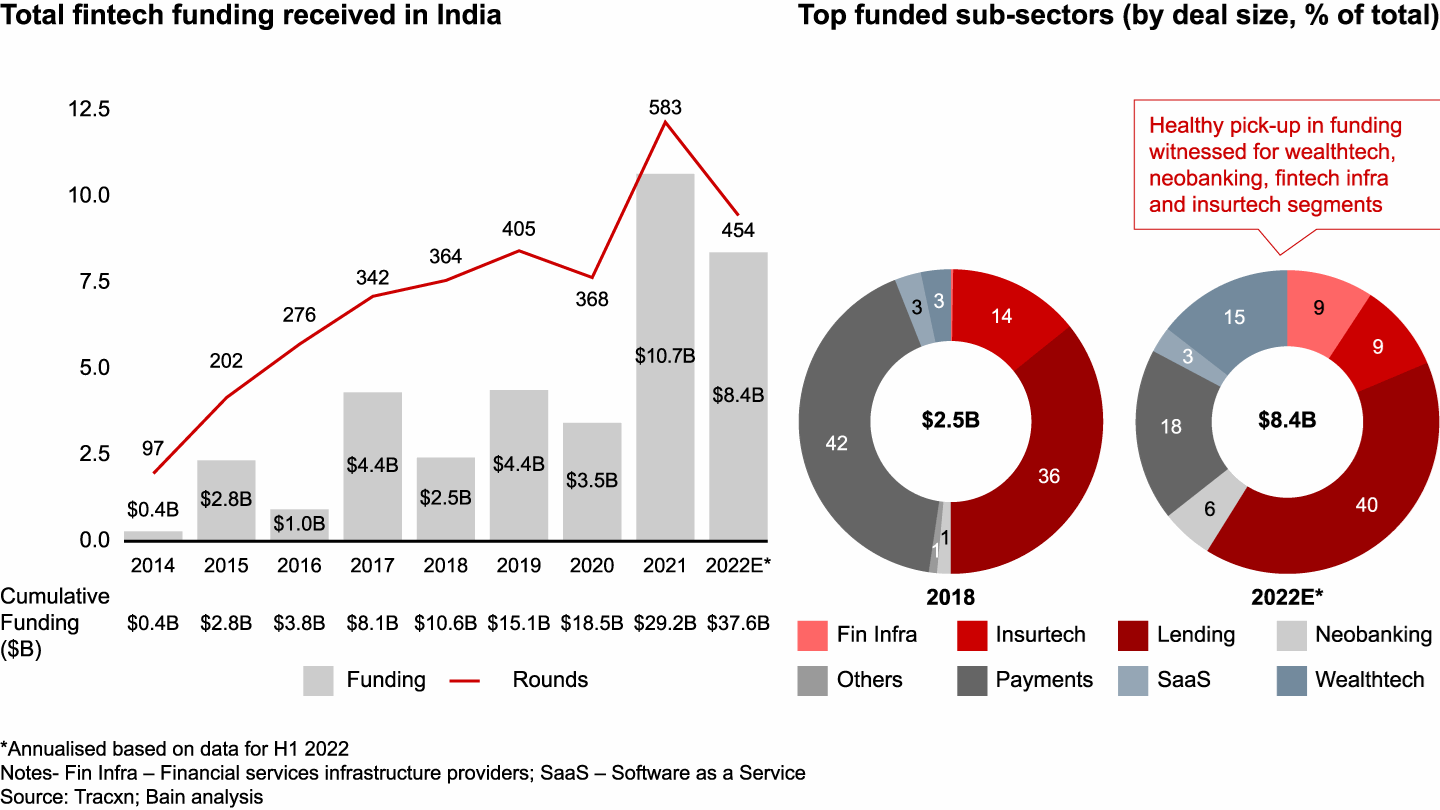
Keywords: predictive model, loan defaults, customer churn, credit behaviour, technology, financial habits, instant loans, credit profile.

# 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

Lending and borrowing money has become more complex than ever in today’s fast-changing world. 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. Rapid advances in the fields of machine learning (ML) and artificial intelligence (AI) have resulted in the development of alternative credit scoring systems, fundamentally altering how lenders assess risk and make critical investment decisions in the realm of consumer debt. These novel models leverage unconventional data sources and sophisticated processing methods, heralding 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 at the forefront of the FinTech revolution, with an adoption 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. This rapid growth can be attributed to a number of factors, including a supportive regulatory environment, a thriving startup scene, and a young and tech-savvy population. The Indian government has played a key role in promoting FinTech adoption, through initiatives such as the Digital India program and the Unified Payments Interface (UPI). UPI has been particularly successful, enabling over 6 billion transactions per month. It has also helped to level the playing field for FinTech companies, giving them access to the same infrastructure as traditional banks.

The Indian startup scene is another major driver of FinTech growth. India is home to over 2,100 FinTech companies, which are innovating in a wide range of areas, including payments, lending, insurance, and investment. These companies are attracting billions of dollars in venture capital, which is fueling their growth. Finally, the young and tech-savvy Indian population is also playing a key role in FinTech adoption. Over 600 million Indians have internet access, and many of them are using their smartphones to access financial services. This is creating a huge opportunity for FinTech companies to reach new customers and offer them innovative products and services. The Indian FinTech industry is poised for continued growth in the years to come. The country's demographics, supportive regulatory environment, and thriving startup scene are all contributing factors. As more and more Indians adopt FinTech, the industry is expected to play an increasingly important role in the country's economy. Lending institutions rely on credit scores to evaluate borrowers' creditworthiness when applying for credit cards or auto loans. Credit scoring assesses various factors like credit history, demographics, and credit behaviour to determine approvals and interest rates. Improving credit scoring mechanisms can lead to significant financial gains and safer lending practices.

This dissertation is all about finding a smarter way to predict loan defaulters in Bangalore by analysing two key factors: customer churn and credit behaviour. By understanding these factors, lenders can better assess a borrower's ability to repay their loans and reduce the risk of default.

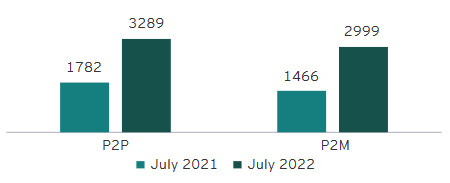
Banks rely heavily on lending as a source of income and financial risk. However, lending carries risks, particularly credit risk, which refers to borrowers' failure to repay the loan. To mitigate this risk, banks traditionally assess clients' creditworthiness using the 5C principle: character, capital, capacity, collateral, and conditions. Despite a rigorous verification process, there is no guarantee that the selected applicant will repay the loan on time.

In the past, banks relied on experts and statistical algorithms to evaluate creditworthiness and assign a credit score based on payment history, credit background, and other factors. This score predicted the likelihood of borrowers repaying loans, but the process required specialized expertise and was prone to errors.

India's Unified Payments Interface (UPI) has become one of the most popular payment methods in the world, enabling Peer-to-peer (P2P) and Merchant payment (P2M) transactions on smartphones, feature phones, and at merchant locations. In July 2022, P2P transactions accounted for 78% of the total transaction value, with a value of $104.595 billion, representing a 94% and 75% increase in value from the previous year. Most P2P transactions had a transaction size greater than $26 (INR 2,000).

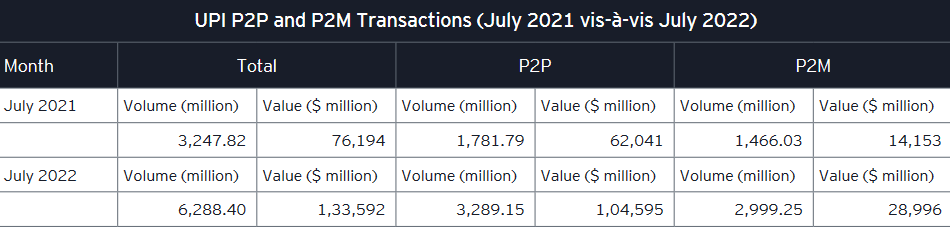
According to the Worldpay report, India led the world in the largest daily real-time payments volume in 2021 with 70.2 million, followed by China with 42.8 million. UPI is now the leading form of retail merchant payments (Person-to-Merchant - P2M disguised as Peer-to-Peer - P2P payment) by value and volume, comparable to credit cards and debit cards. This is due to the significant increase in the number of banks going live on UPI from 2016 to 2021, reaching 297 in 2022.

The value of UPI transactions crossed $25 billion in 2019 and reached a massive $111 billion in 2022. According to the Reserve Bank of India (RBI), the overall Indian digital payment volume was 72 billion in FY21-22, with an overall transaction value of INR 1,744 trillion ($24 trillion). Driven by the 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 reduce the volume of cheque-based payments.



**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**

India's digital lending sector is booming, fueled by the country's rapid digitalization. By 2030, the sector is projected to account for 60% of the country's financial technology market, with the online loan market expected to hit $1.3 trillion, growing four times from $270 billion currently. At the lead of this revolution are machine learning and deep learning algorithms, which are being used by researchers and banking authorities to predict credit scores automatically. This approach streamlines the selection process for eligible loan candidates and makes it easier for people to access credit, regardless of their traditional credit history.

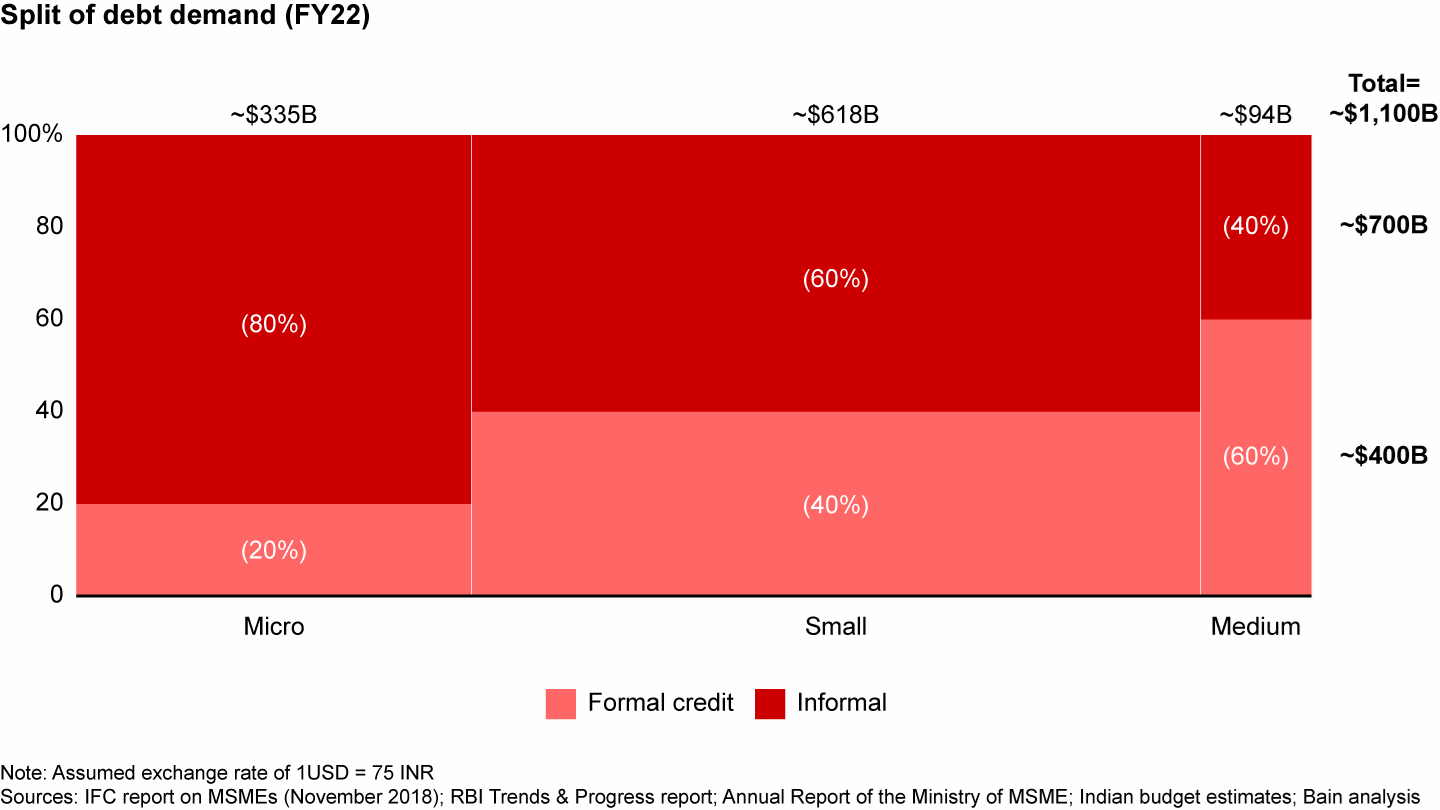
Digital lenders use these algorithms to assess creditworthiness and build credit history for borrowers. This is a boon for underserved segments of the population, such as small and medium-sized enterprises (SMEs), which make up 75% of MSME lending in the country.

Collaboration between banks and FinTech players can further boost credit access to these underserved segments. The Indian FinTech ecosystem is a formidable global force, generating $200 billion in revenue and $1 trillion in AUM by 2030. This partnership between banks and FinTech players can help to democratize access to credit and drive India's economic growth.

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.

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.

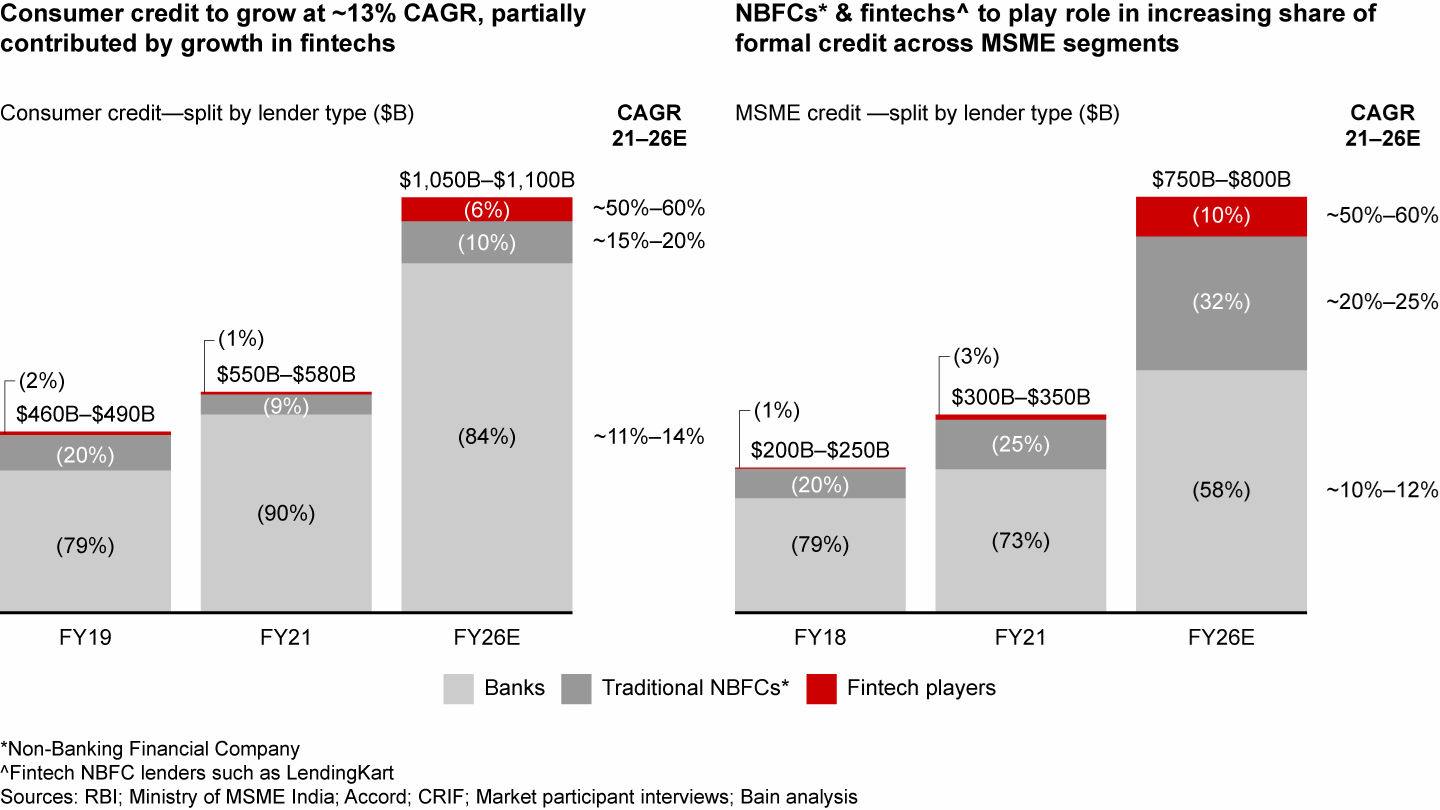
Bangalore is a special place for this study because it's a mix of different businesses and a growing technology industry. This means lots of people have different kinds of jobs and financial situations. The research will use information from the past to make a smart system that can predict if someone 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 explored 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. Besides, 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.

While the statistics from a Bain & Company report will help us understand the growth of the Fintech sector and digital lending in India, the same cannot be said for the credit market in Bangalore, where very little research has been conducted. This is where this study comes in; the author proposes developing a system capable of identifying, categorizing, and forecasting consumers' financial behaviours.

To create this intelligent system, the author will use a type of machine learning technique known as "Deep neural network algorithms," which will be supplemented with various types of regression models. This will help us analyse all of the information we collect from people and predict whether they will have difficulty repaying 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.

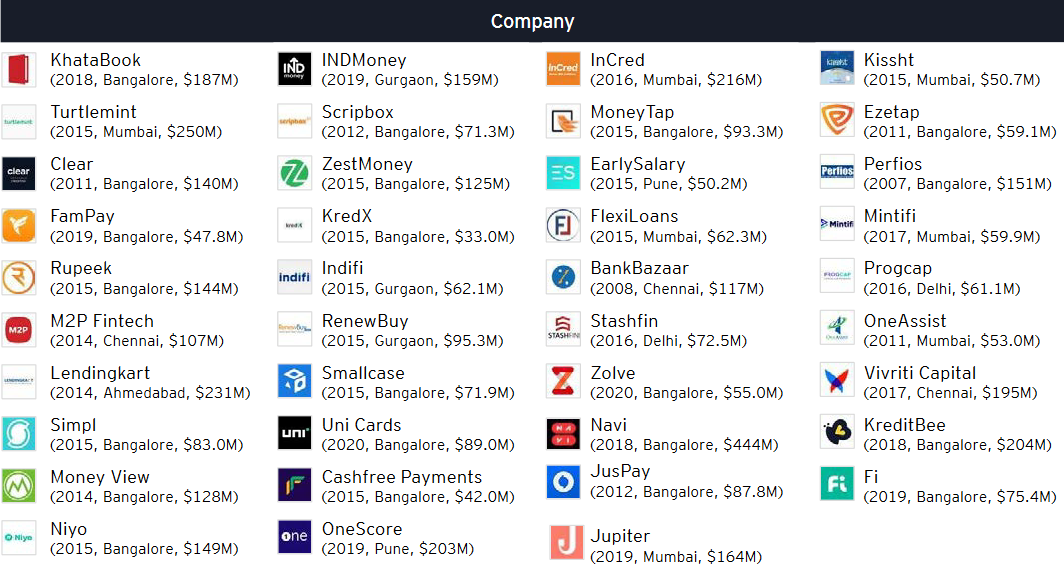
The growth of unsecured loans has been driven primarily by smaller towns (Tier 4), where they have grown by 32% annually over the past three years, compared to 18% annual growth in larger cities (Tier 1). North Bangalore, a mix of Tier 2 and Tier 3 geographies, is a key contributor to this growth.

The average loan size is shrinking, reflecting a shift towards smaller loans by non-banking financial companies (NBFCs) and fintech lenders. For example, the average personal loan size from NBFCs has declined by 70% over the past 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.

At a high level, this dissertation explores how new technology can be used to better understand people's financial habits and predict their likelihood of defaulting on loans. This research could help to make lending and borrowing more secure and reliable, and to shed light on the evolving process of credit scoring. The dissertation examines a variety of successful machine learning techniques for developing credit scoring models for imbalanced populations and sample sizes, taking into account both quantitative and qualitative factors that may or may not directly influence borrowers' ability to repay their loans. It also explores the potential for collaboration between technology companies and banks to improve credit scoring and lending practices in Bangalore.

# 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.

Flipcarbon's way of doing things is based on some key ideas:

1. **Dedication to Fast Growth:** At Flipcarbon, they're all about helping their clients grow quickly. They know it's important to reach growth goals fast, so they work hard to show real results within set timeframes.
2. **Building a Growth-Friendly Culture:** They've created a special work environment that gives their clients the freedom to drive growth projects on their own. But at the same time, they keep an eye on costs, making sure things stay financially sound.
3. **Embracing Skills and People Smarts:** Flipcarbon doesn't just understand the nitty-gritty of starting new projects; they also get how important people are in these ventures. They know that success comes from not only putting new ideas into action but also supporting the people who make innovation happen.
4. **Developing Competitive Team Members:** Flipcarbon really focuses on building a team that has the right skills and knowledge to thrive in a tough business world. They encourage their team to keep learning and growing, making sure they can deliver great results for their clients.

Flipcarbon's Virtual CFO & Accounting Services play a big role in what they do. This part of their work focuses on giving top-notch financial advice and accounting help, acting as a reliable partner for their clients. The way Flipcarbon does things is like how carbon forms special bonds in certain situations. They take inspiration from this and use a "specific-solution-to-specific-issues-of-specific-client" approach. Just like carbon creates unique connections under certain conditions, Flipcarbon aims to make customized solutions by using the unique aspects of each client's situation. This way, they can create one-of-a-kind plans and fixes that tackle the specific challenges their clients are dealing with.

***Background and Current Description***

Flipcarbon Integrated Solutions Private Limited is a private company that was established on December 4, 2014. It's categorized as a private limited company and is situated in Bangalore, Karnataka. The authorized share capital of the company is INR 1.00 Cr, and the total paid-up capital is INR 91.50 lac. The company specializes in offering HR & Financial Consulting, complete 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

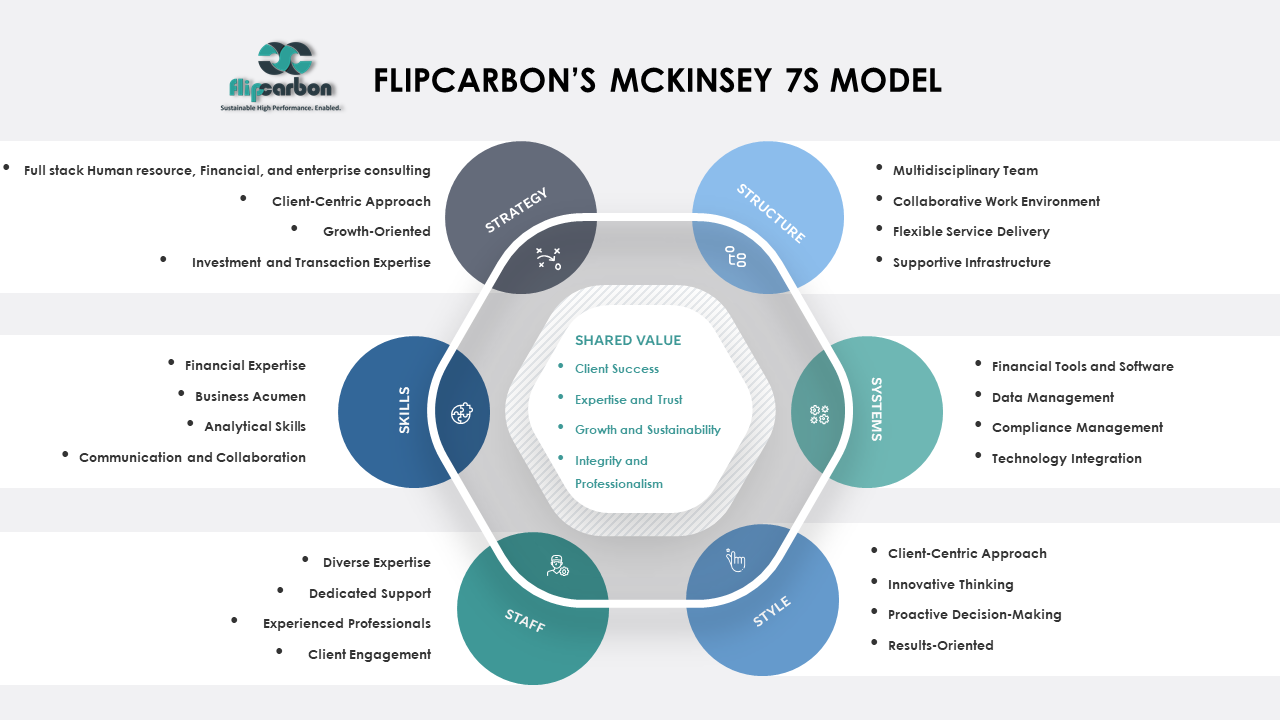
***Our Team***

Flipcarbon Integrated Solutions Private Limited was founded on December 4, 2014. It's a private company in Bangalore, Karnataka, with an authorized share capital of INR 1.00 Cr and a total paid-up capital of INR 91.50 lac. The company provides HR & Financial Consulting, end-to-end HR & Financial Outsourcing, Modular HR Services, and Financial Services.

There are three directors in the company, and Prabhash Nirbhay, the founder, has been serving since the beginning. Lokesh Mehta joined on August 10, 2015, and Alok Ranjan, who is also the current acting CEO, was appointed on March 1, 2018. Prabhash Nirbhay is involved in six companies and connected to five others through his directorships.

The Bangalore branch has 13 employees, and overall, Flipcarbon has a network of over 30 experts across three branches. The CFO vertical, where the author interned, is led by Mr. Deepak Kewalramani, a pro with almost three decades of financial analytics and strategic advisory experience. The author reported to Mr. Bharat Karnani, a CA with 15 years of financial consulting experience, and Ms. Pramila Lakra, with nearly five years of experience in Financial Due Diligence and Forecasting. The organization needed a business intelligence expert to make financial statements and visualizations using Microsoft Suite and Power BI. The author, with a finance and business analytics background, was a good fit. Drawing from their experience as a business analyst, they proposed a framework for integrating solutions using a tech-stack approach (SaaS). With over 500 years of collective management experience, eight years of consulting, and 170+ satisfied clients, Flipcarbon aims to be 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 start-ups 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 paper explores credit risk assessment in consumer lending, vital for gauging loan default likelihood and addressing financial concerns. The challenge of imbalanced class distributions prompts a unique solution: a credit rating-specific multi-objective ensemble learning framework. Tailoring models to credit rating categories, the approach utilizes OC-SVM and evolutionary optimization. Experimental studies confirm its superiority, highlighting the method's contributions—balancing default recognition, innovative ensemble learning, credit rating-specific strategy, and real-world applicability. This pioneering methodology could transform how financial institutions manage loan portfolios for 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 tackles the hurdles of training deep reinforcement learning models incrementally, a method successful in real-world applications but plagued by the issue of forgetting old data, particularly in customer credit scoring. Introducing the Balanced Incremental Deep Q-Network based on Variational Autoencoder Data Augmentation (BIDQN-VADA), the authors mitigate this challenge. Their approach involves balancing subsets, utilizing variational autoencoder augmentation, and employing a novel data stream cache with a FIFO approach for incremental model updates. Contributions include a balanced incremental learning method, a unique data stream cache, and validation through experiments on real-world credit scoring datasets. The BIDQN-VADA outperforms seven models, demonstrating its efficacy in achieving superior credit scoring. The paper provides a comprehensive overview of related works, theoretical background, model details, experimental design, results analysis, and concludes by summarizing key findings and suggesting future research directions.

Interpretable machine learning for imbalanced credit scoring datasets

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

This study delves into credit scoring, emphasizing the overlooked impact of class imbalance on interpretability in machine learning. Unlike prior research focusing on predictive accuracy, this paper explores how class imbalance affects the stability of two key interpretation methods, LIME and SHAP. Using a controlled sampling process and real-world UK residential mortgage data from 2016 to 2020, the study reveals that as class imbalance increases, interpretation stability diminishes. Extending the analysis to open-source credit scoring datasets confirms the universality of this phenomenon. This contribution challenges traditional perspectives and provides a robust experimental framework for future investigations into credit scoring's intricate landscape.

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

This study captures the essence of our era – the interplay between data, technology, and finance. Extracting credit insights from user-generated text on platforms like Lending Club represents a ground-breaking leap in understanding dynamic factors influencing credit default. Notably, the paper reveals the power of even brief user-generated text in significantly improving credit default predictions. The prominence of deep learning, especially with transformer models, signals the evolving landscape of machine learning. Simpler models like average embedding neural networks also shine, emphasizing brilliance in simplicity. The exploration of credit scoring through textual data prompts a shift in perspective, urging financial institutions to integrate unstructured data for more informed decisions. Beyond credit default prediction, this study suggests exploring user-generated text's potential for personalized financial services, where understanding the customer's narrative is as crucial as analysing transaction history.

Research on Default Prediction Model of Corporate Credit Risk Based on Big Data Analysis Algorithm Qingyan Xianyu, Mo Hai (2023)

This research explores the confluence of technology, finance, and data analysis to tackle corporate credit risk prediction. Utilizing big data analysis algorithms, the authors construct a robust model by gathering a comprehensive dataset from over 1,000 listed companies. Employing correlation coefficients, they develop Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) models within the Pytorch framework on the Spark platform. Comparisons with traditional models showcase RNN's superiority, achieving a remarkable accuracy rate of 0.93. Rigorous testing confirms the model's robustness, demonstrating its transformative potential in corporate credit risk evaluation within the 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 with the changing landscape of credit scoring, exploring the intersection of machine learning and finance. The authors' innovative approach, introducing Penalized Logistic Tree Regression (PLTR), blends logistic regression with decision trees, offering enhanced predictive accuracy while preserving interpretability – a vital consideration in finance. Unlike black-box ensemble methods, PLTR marries the strengths of logistic regression and decision trees, providing transparency along with accuracy. The paper highlights the limits of traditional models in capturing non-linear effects in credit data, positioning PLTR as a solution that transcends linear specifications. Its implications extend beyond credit scoring, foreseeing a future where PLTR redefines customer engagement, risk management, and decision-making across the financial spectrum. In summary, this research pioneers the integration of advanced machine learning with interpretability in the financial industry, urging a harmonious intersection between innovation and regulation.

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

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

This study delves into a dataset featuring 13,081 Italian companies and 111 cooperative banks to explore the predictive power of credit line and long-term loan management actions on default likelihood over one to two years. Incorporating balance sheet cues and dynamic bank attributes, we merge credit-related signals from private banking sources with financial records, significantly improving default predictions for small and medium-sized businesses. Our analysis goes further by comparing prediction precision with models considering only financial data versus those leveraging both financials and additional bank-specific insights. We emphasize the diversity in the relationship between balance sheet clues, bank-firm ties, and default odds across industries and regions, underscoring the importance of case-by-case risk analysis for banks.

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)

This study underscores the pressing need for smarter customer churn prediction strategies, advocating for the use of advanced techniques like Artificial Neural Networks and data mining. The paper highlights the significance of understanding customers, leveraging cutting-edge technology, and proactively addressing churn to safeguard revenue. In the contemporary business landscape, data mining is crucial for transforming raw data into actionable insights, unveiling implicit trends and relationships. Customer churn, the separation of companies and customers, underscores the importance of predicting customer behavior and fostering strong relationships. The research delves into machine learning techniques such as SVM, Random Forest, and Linear Regression, offering best practices for effective implementation. Mastering churn predictions with techniques like Artificial Neural Networks becomes a secret weapon for businesses, especially in the fintech sector, enabling them to anticipate and mitigate churn, foster lasting customer relationships, and drive business growth.

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 employs machine learning techniques, a subset of artificial intelligence, to predict customer churn in the banking industry by analyzing customer behavior. Classifiers like KNN, SVM, Decision Tree, and Random Forest, along with feature selection methods, are utilized for improved system performance. The research, based on a Kaggle dataset, aims to identify an optimal model with higher precision and predictability.

Recognizing the essential role of understanding consumers, the paper aligns with Liu and Shih's emphasis on innovative marketing strategies to enhance loyalty. The competitive market requires effective use of marketing resources, where technology, including data mining techniques, extracts insights to guide decisions. The study underscores the critical importance of early-stage churn prediction and proposes a machine learning framework for the banking sector, emphasizing its potential to aid in customer retention.

Implications for the banking sector are evident, highlighting the value of early-stage churn prediction. The integration of machine learning techniques provides a novel approach to address customer churn, offering promising solutions for organizations striving to balance customer acquisition and retention. The research contributes to the evolving landscape of customer churn prediction, guiding future studies and assisting the banking industry in addressing 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 suggests that integrating machine learning algorithms and neural networks shows promise for enhancing credit risk assessment and loan approval processes. It advocates for further research to refine these predictive models, improving accuracy and real-world applicability. The study contributes valuable insights to the evolving field of credit risk assessment, offering guidance for both researchers and practitioners in finance.

Highlighting the significance of loan lending in overcoming financial constraints, the article acknowledges credit risk as a major concern. Despite risks, loan lending remains beneficial for financial institutions, contributing to profitability and sustainability. Traditionally, creditworthiness assessment relied on historical data and credit scores, but advancements in machine learning now enable autonomous credit score predictions. This study explores existing literature on predictive models for credit risk assessment, emphasizing the potential of machine learning algorithms in identifying potential defaulters and enhancing risk management in loan lending practices.

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

The paper, "Sample Selection in Credit-Scoring Models," explores credit scoring, a vital process for institutions like credit-card vendors. It addresses the challenge of sample selection bias in constructing statistical models for credit assessment. The research investigates three applications – predicting loan default, modeling expenditure, and assessing derogatory reports in credit histories – each requiring specific models and estimation techniques to account for sample selection issues.

Credit-card vendors, dealing with millions of applications, rely on statistical models that may introduce bias due to sample selection from individuals with accepted applications. The paper emphasizes the need for models that address sample selection biases, highlighting their impact on predictive accuracy and business decisions. Constructing models that account for these biases contributes to more precise credit-scoring predictions and enhances the 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 create a unique space for direct engagement between borrowers and lenders, but they face challenges, notably class imbalance. Credit risk prediction is crucial in this setting, yet few models effectively address imbalanced data complexities. This article pioneers a risk assessment methodology tackling these challenges, comparing various classifier-resampling technique combinations. The G-mean measure is wisely chosen as the evaluation metric to counteract bias towards the majority class. Results reveal that combining random forest and random under-sampling is a potent strategy for comprehensive credit risk prediction in social lending markets. Beyond this study, the research contributes to ongoing discussions on credit risk prediction, offering valuable insights for refining risk assessment processes in the evolving financial landscape.

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 underscores the increasing need for precise prediction models in the face of financial crises, emphasizing transparent decision-making. The Compact Evolutionary Interval-Valued Fuzzy Rule-Based Classification System (IVTURS FA RC-HD) is introduced as an innovative solution for enhancing modeling and prediction capabilities in real-world financial applications. Notably, it tackles imbalanced financial datasets without relying on pre-processing or sampling methodologies, maintaining interpretability. Experimental studies across eleven real-world financial datasets showcase its superiority over benchmark techniques, highlighting its potential for accurate predictions in complex financial scenarios. The IVTURS FA RC-HD system, with its fusion of evolutionary techniques and interval-valued fuzzy logic, stands out as a promising tool in financial modeling.

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 delves into the significant problem of loan applicant misclassification in credit scoring models, particularly affecting investors in Peer-to-Peer (P2P) lending. The proposed solution, REMDD (Resampling Ensemble Model based on Data Distribution), tackles class imbalance through a unique under-sampling technique, UMCDD, grounded in majority class data distribution. The methodology extends beyond resampling, employing an ensemble of classifiers chosen for their performance, enhancing predictive capabilities. Experimental results on real-world P2P lending datasets demonstrate REMDD's effectiveness in predicting both minority and majority classes, positioning it favorably against existing models. REMDD emerges as a valuable tool for comprehensive credit risk assessment in P2P lending platforms.

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

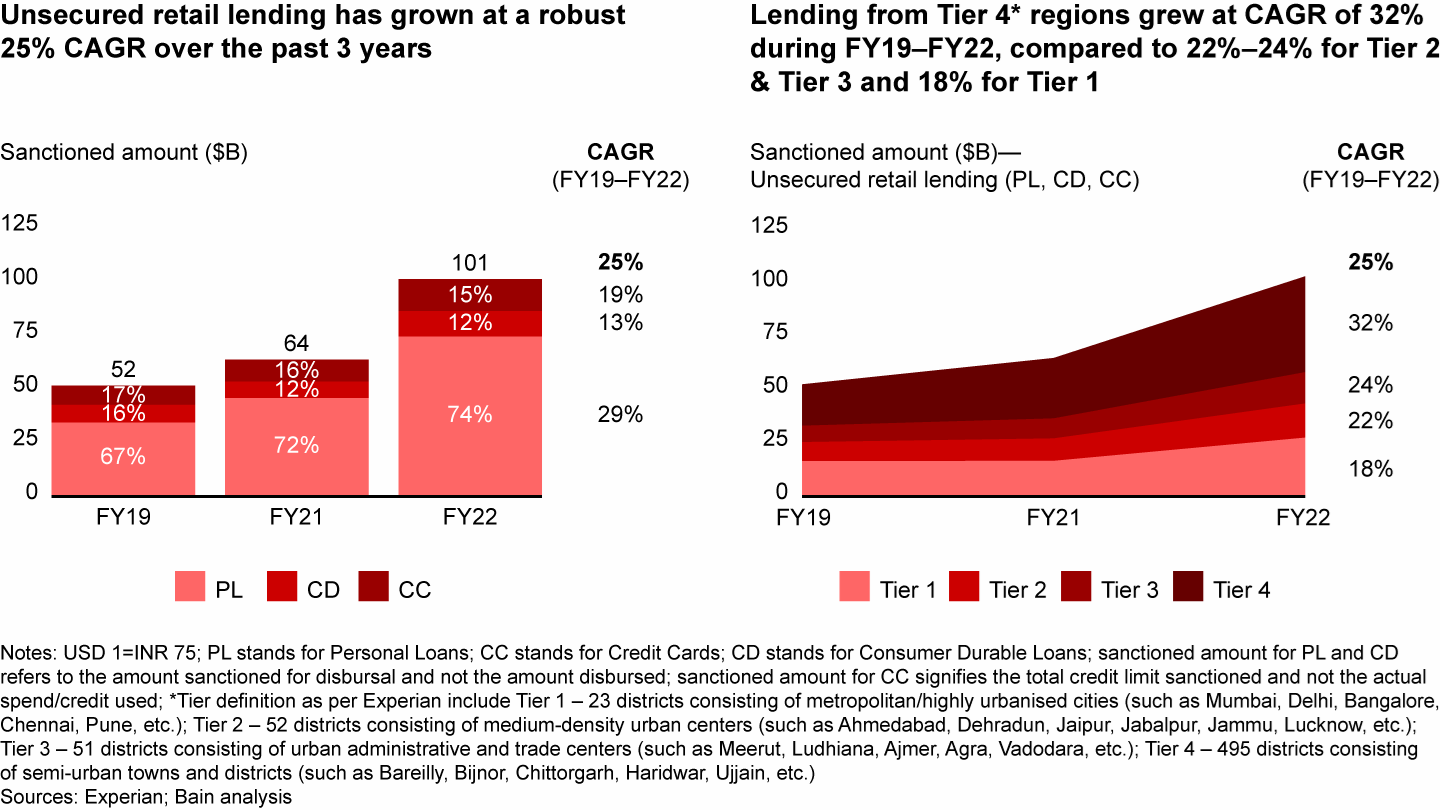
Meixuan Li, Chun Yu, Wei Liu (2021)

# This paper underscores the significance of robust risk prediction systems in the era of online lending platforms. The authors introduce the Stacking+CNN model, a novel fusion of Convolutional Neural Networks (CNN) and the Stacking algorithm, aiming to enhance feature extraction and overall generalization. The model undergoes distinct phases, including a meticulous feature extraction process, Stacking algorithm training, CNN-based feature extraction, and risk prediction using Support Vector Machine (SVM). Empirical results demonstrate the model's superiority in predictive accuracy and recall rates, offering substantial advancements for online lending. Contributions include the introduction of a unique model, an original feature extraction approach, a specific feature selection process, and the use of k-fold cross-validation to address overfitting. In essence, this research pioneers a transformative tool for elevating online loan risk prediction.

# 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.

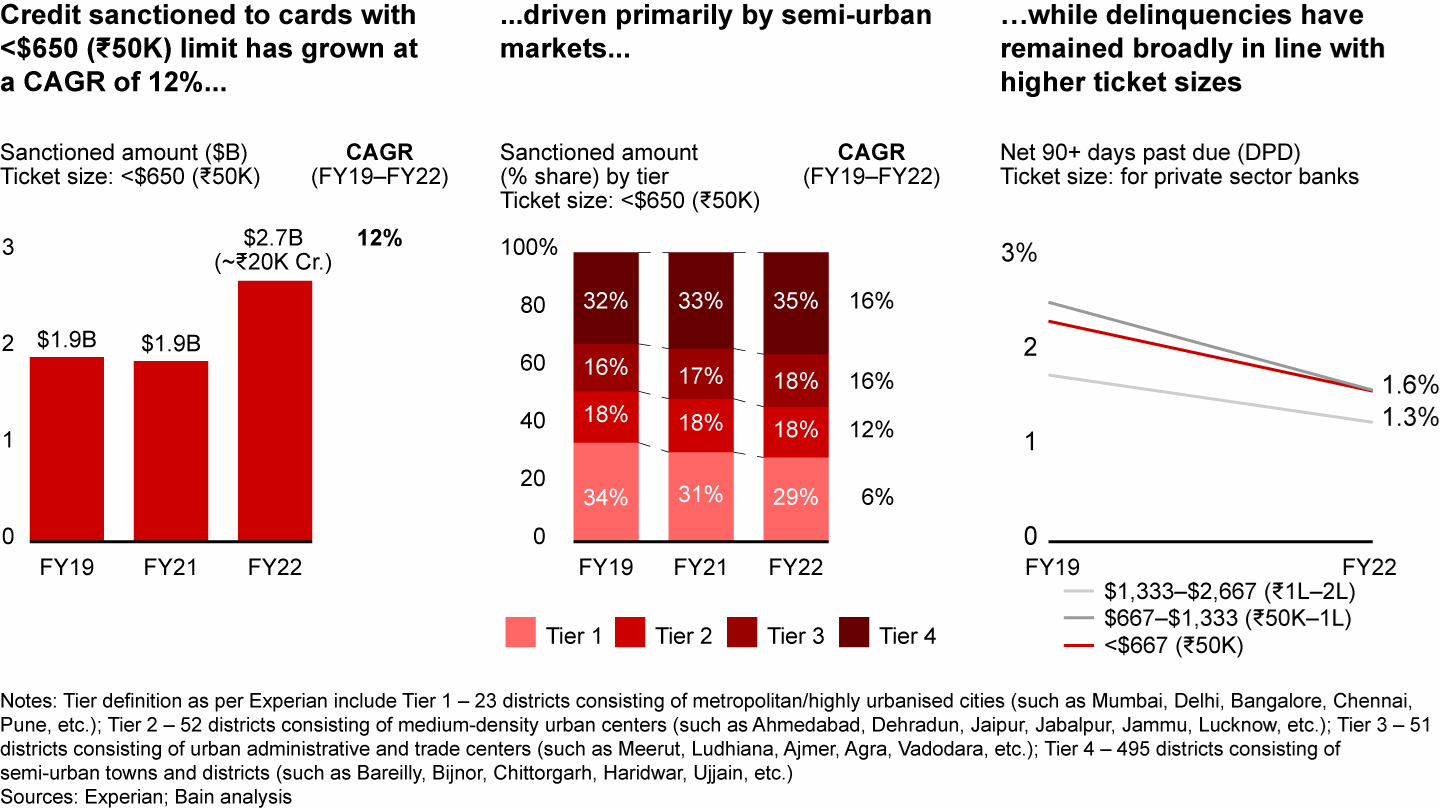
Despite the rapid growth of the online lending market in Bangalore, very little research has been conducted on credit risk assessment, specifically concerning the incorporation of various types of data such as demographic, financial, behavioural, and social media data. The lack of comprehensive studies 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 creates a construct in handling creating a hands-on 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.

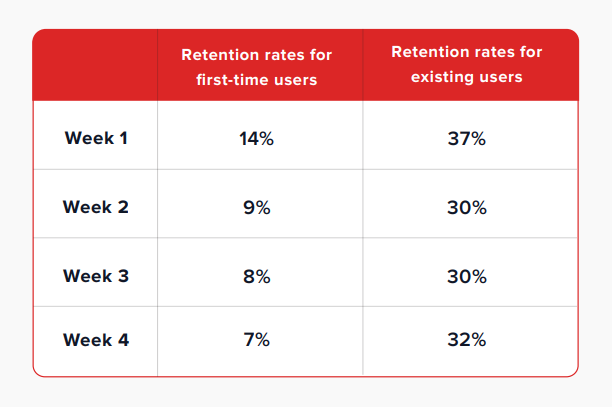
Bangalore, often referred to as the Silicon Valley of India, is home to a thriving fintech ecosystem. The customer churn rate in the Indian fintech sector averages a staggering 73%, a figure that has substantial implications for business sustainability and profitability. Furthermore, the credit default 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:

1. 40% of new users register a profile after downloading a Fintech app
2. 24% of new users convert within 12 days
3. 73% of new users churn within 1 week
4. 46% of new users uninstall within 1 month, of which 80% churn within 9 days
5. 14% of uninstalled users reinstall within 30 days, 80% of those who reinstall do so within 11 days.

**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.



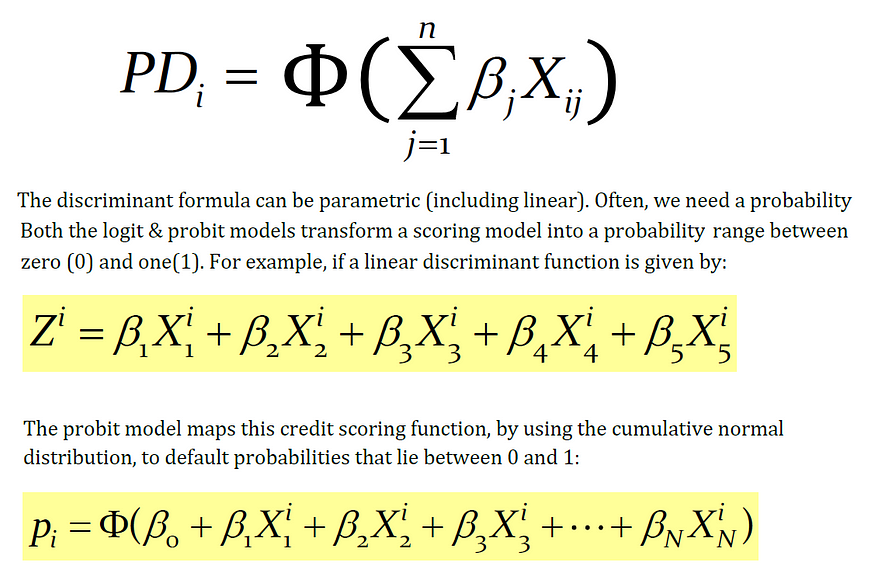
**Source : Clevertap** [[7]](https://brandequity.economictimes.indiatimes.com/files/cp/1139/cdoc-1658399014-Fintech_Retention_Guide.pdf)

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.

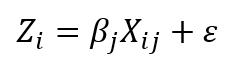
With a robust dataset in place, the construction of clear and testable research hypotheses ensues, encompassing both null and alternative stances. The investigative prowess is further amplified through the employment of Probit and Logit Regression methodologies. These techniques unfurl the intricacies 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 loan defaults 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 870 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 LR, Support Vector Machine (SVM), Decision Tree, and Random Forest, will be employed to explore the likelihood of default. 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 LR, SVM, DNN Model | Python - Tensorflow, | 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% |

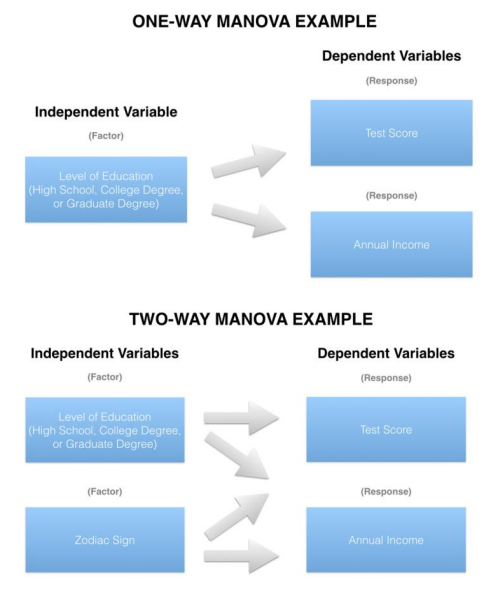
The second objective is to assess the creditworthiness of loan applicants in the instant loan app sectors employed through fintech platforms. The same questionnaire data will be used to build a comprehensive credit scoring model. Machine learning techniques will be applied to handle the data, including handling missing values, encoding categorical variables, and scaling numerical features. By training various models, such as Logistic Regression, Support Vector Machine, Deep Neural Network 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 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 behaviour, loan risk assessment, and the factors influencing creditworthiness and customer churn in the banking sector. For this the author will be employing FFNN.

**MANOVA**

MANOVA, which stands for Multivariate Analysis of Variance, is a statistical technique used for analysing 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.

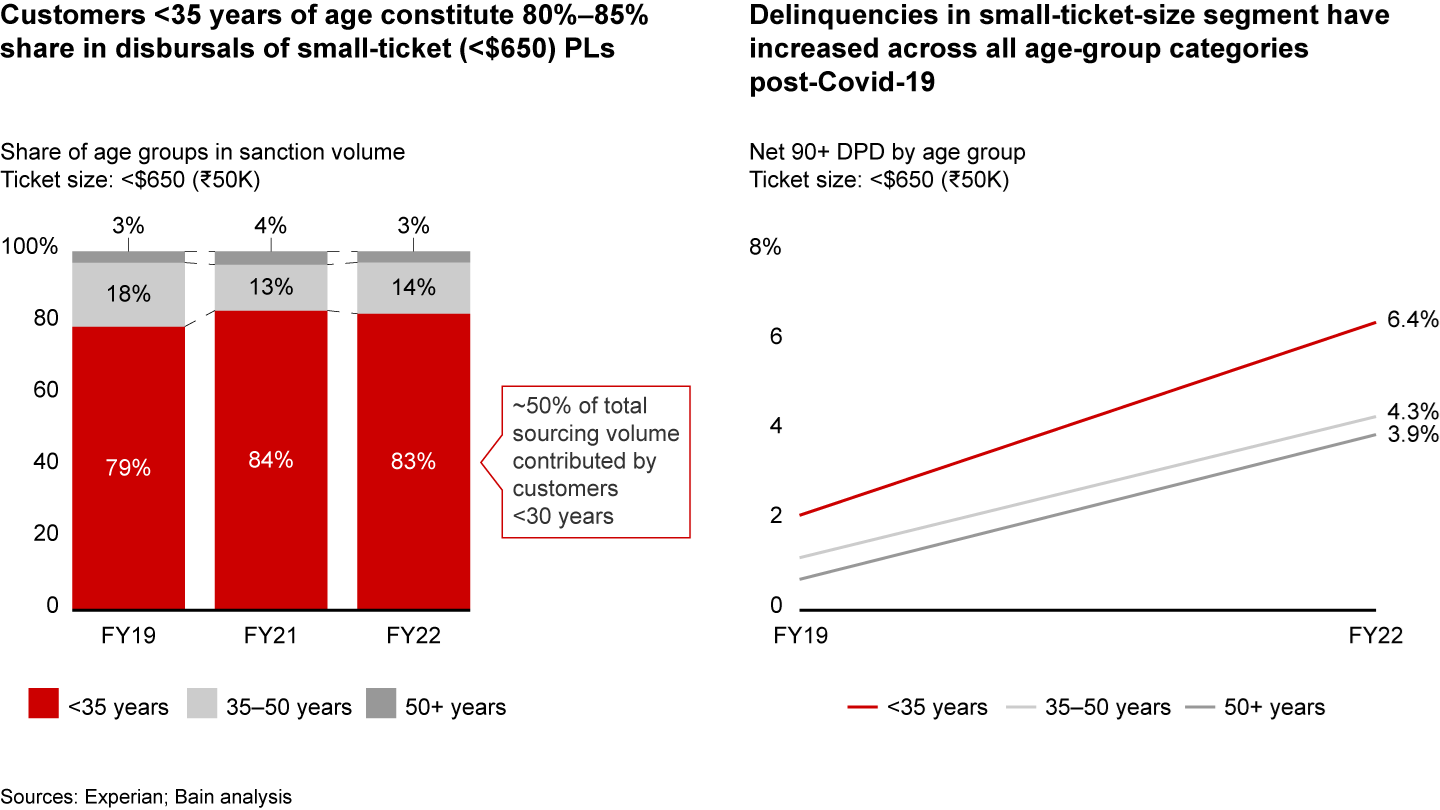


Source: Medium [[10]](https://medium.com/nerd-for-tech/everything-about-manova-and-mancova-4c1c237af464)

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 870 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 22 to 32.

b. Customers aged 33 to 43.

**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 22 to 30 contribute between 80% and 85% of disbursals, I will allocate 80% of my sample (~695 cases) to this stratum and the remaining 20% (~173 cases) to the age group of 33 to 43.

**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. However, it must be noted that, this data was collected by a research agency who are close collaborated of Flipcarbon.

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

Hypothesis

For the purpose of this study, the researcher has come through with 16 hypotheses.

H1

Null Hypothesis (H0): Young borrowers (22-32) are equally likely to default on their loans compared to older borrowers.

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

H2

Null Hypothesis (H0): Borrowers who default on their loans have the same median income as borrowers who do not default.

Alternative Hypothesis (H1): Borrowers who default on their loans have a significantly lower median income compared to borrowers who do not default.

H3

Null Hypothesis (H0): There is no strong correlation between Age and Loan Amount borrowed by a respondent on average.

Alternative Hypothesis (H1): There is a strong correlation between Age and Loan Amount borrowed by a respondent on average.

H4

Null Hypothesis (H0): Borrowers with longer employment tenures (>24 Months) have the same likelihood to default on their loans as borrowers with shorter tenures.

Alternative Hypothesis (H1): Borrowers with longer employment tenures (>24 Months) are less likely to default on their loans.

H5

Null Hypothesis (H0): Borrowers with a prior history of default (Credit card overdues, credit card defaults) have the same likelihood to default on their current loans as those without a history of default.

Alternative Hypothesis (H1): Borrowers with a prior history of default (Credit card overdues, credit card defaults) are more likely to default on their current loans.

H6

Null Hypothesis (H0): Personal loans have the same default rate as loans with other intents (e.g., educational, debt consolidation).

Alternative Hypothesis (H1): Personal loans have a higher default rate compared to loans with other intents.

H7

Null Hypothesis (H0): An awareness of one's credit rating is not correlated with the probability of loan default among borrowers.

Alternative Hypothesis (H1): An awareness of one's credit rating is correlated with a decreased probability of loan default among borrowers.

H8

Null Hypothesis (H0): Venture loans have the same default rate as all other loan types.

Alternative Hypothesis (H1): Venture loans have the lowest default rate among all loan types.

H9

Null Hypothesis (H0): Borrowers who have received financial counselling have the same likelihood of loan default as those who have not received counselling.

# Alternative Hypothesis (H1): Borrowers who have received financial counselling are more likely to avoid loan default compared to those who have not.

# H10

# Null Hypothesis (H0): Borrowers who use digital lending services frequently have the same default rate as those who use them infrequently.

# Alternative Hypothesis (H1): Borrowers who use digital lending services frequently have a higher default rate than those who use them infrequently.

# H11

# Null Hypothesis (H0): Borrowers with collateral assets have the same likelihood to default on their loans as those without collateral.

# Alternative Hypothesis (H1): Borrowers with collateral assets are less likely to default on their loans compared to those without collateral.

# H12

# Null Hypothesis (H0): Borrowers who report maintaining a budget have the same likelihood to default on their loans as those who do not maintain a budget.

# Alternative Hypothesis (H1): Borrowers who report maintaining a budget are less likely to default on their loans than those who do not maintain a budget.

# H13

# Null hypothesis (H0): There is no significant relationship between customer churn and age.

# Alternate hypothesis (H1): There is a significant positive relationship between customer churn and age.

# H14

# Null Hypothesis (H0): There is no significant correlation between customer churn and interest rate satisfaction rate.

# Alternative Hypothesis (H1): There is a significant correlation between customer churn and interest rate satisfaction rate.

# H15

# Null Hypothesis (H0): There is no significant difference in the likelihood of churn between customers who faced loan processing delays and those who did not.

# Alternative Hypothesis (H1): Customers who didn't face loan processing delays are less likely to churn compared to those who did.

# H16

# Null Hypothesis (H0): There is no significant difference in the likelihood of being advocates between customers who churned and those who did not churn.

# Alternative Hypothesis (H1): Customers who didn't churn are more likely to be advocates of their lending institution compared to those who churned.

# 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 860 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 behaviours 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.

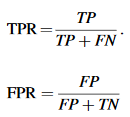
Evaluation indicators

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.



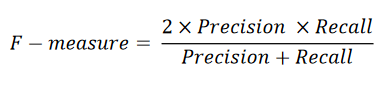
Accuracy is a widely employed metric to gauge the performance of a classification model. It quantifies how many samples the classifier has accurately predicted. To calculate accuracy, the model is employed to classify the test dataset, and the accuracy is determined by dividing the number of correctly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

**

# 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) totalled 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 counselling are more apt to avoid loan default as compared to those who have not. Among the 189 borrowers who received financial counselling, 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 counselling. This striking dissimilarity in default numbers highlights the potential efficacy of financial counselling 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.

H13: Customer Churn is directly related to age, elder customers are less likely to churn

In the provided dataset, there were 534 instances of churn observed among customers aged 22-32, whereas customers aged 32-42 experienced 127 churns. Conversely, in the non-churn category, customers aged 22-32 had 55 cases of non-churn, while customers aged 32-42 had 127 instances of non-churn. The data indicates a nuanced relationship between age and customer churn. While customers aged 22-32 show a higher churn count, those in the 32-42 age group exhibit a higher commitment to not churning. This suggests that age alone may not be a decisive factor, and a comprehensive analysis considering additional variables is warranted for a thorough understanding of customer behaviour.

H14: Customer churn is directly correlated to Interest rate satisfaction rate

Consistent with the hypothesis, a total of 448 respondents who were likely to churn voiced their dissatisfaction with the interest rate provided by their lending institution. Furthermore, out of these respondents, 207 expressed a higher level of dissatisfaction by stating that they were "Very Dissatisfied" with the interest rate.

H15: Customers who didn’t face loan processing delays are less likely to churn

Out of the 650 respondents classified as likely to churn, they reported facing loan processing delays. Conversely, the 198 respondents who were not identified as churners mentioned that they did not encounter any issues with loan processing delays. This showcases a clear distinction between the two groups. It is important to note that there is a significant correlation of 0.936 between loan processing delays and the potential for customer churn. This suggests that there is a strong relationship between these two factors, indicating that delays in loan processing can greatly impact the likelihood of customer churn.

H16: Customers who didn’t churn are likely to be good advocates of their lending institution

Consistent with the hypothesis, the respondents who were identified as likely to not churn demonstrated a strong inclination to advocate for their lending institutions. Among these respondents, 385 expressed a high likelihood of recommending their institution to others, while 380 indicated a moderate likelihood. Furthermore, 99 respondents remained neutral in their attitude towards recommending their institution to others.

Among the 868 respondents surveyed, the leading cause of customer churn was attributed to recommendations from friends or family to switch, as reported by a significant majority. Following closely behind, 171 respondents cited dissatisfaction with the services provided by their current financial institution as a contributing factor. Additionally, 166 respondents highlighted lower interest rates elsewhere as a motive for their decision to churn, while better loan terms and conditions elsewhere were mentioned by the fewest respondents, with 163 individuals acknowledging this as a factor. It is crucial to note that these reasons should not be viewed in isolation, as they are interconnected and influence one another in the decision-making process of customers seeking alternative financial options.

Hypothesis Summary

Here is a detailed summary of key findings from the data analysis and interpretation chapter in a tabular format:

|  |  |
| --- | --- |
| Hypothesis | Finding |
| H1: Young borrowers (22-32) are more likely to default on their loans compared to older borrowers. | ***Supported***. There were 39 defaults among 22–32-year-olds, compared to 14 defaults among 33–42-year-olds. |
| H2: Borrowers who default have lower median income than non-defaulters. | ***Supported.*** It is evident that defaulters have a significantly lower median income of ₹36,000 compared to non-defaulters, who have a higher median income of ₹79,600. This income disparity between defaulters and non-defaulters highlights the financial challenges faced by defaulters and the potential impact it may have on their ability to meet their financial obligations. |
| H3: Inverse correlation between age and loan amount. | ***Not supported***. Correlation coefficient of -0.0192 between age and loan amount. |
| H4: Borrowers with >24 month employment tenure have lower default rate. | ***Supported***. In the group with a tenure of more than 24 months, there were a total of 43 defaults. On the other hand, in the group with a tenure of less than 24 months, there were a total of 592 defaults. This significant difference in default numbers suggests that the length of tenure may play a crucial role in determining the likelihood of default. |
| H5: Borrowers with prior default history have higher current default likelihood. | ***Strongly supported***. In the study, of the 627 total defaults, 588 were observed in individuals with a history of defaults, while only 39 occurred in those without such a history. This stark difference underscores the crucial role of prior default history in evaluating the risk of future defaults. |
| H6: Personal loans have higher default rates than other loan types. | ***Supported***. In comparison to other loan types, personal loans have a significantly higher number of defaults, with 444 defaults reported. This is in contrast to the 183 defaults reported for other types of loans. |
| H7: Awareness of credit rating correlates with lower default rate. | ***Supported***. Borrowers aware of their credit rating had 37 defaults, while those unaware had 204, emphasizing the crucial role of credit awareness in influencing borrowing behaviour and ensuring financial stability. |
| H8: Venture loans have lowest default rate among loan types. | ***Strongly supported***. Venture loans exhibited a remarkably low default rate, with only 29 defaults, in stark contrast to other loan types experiencing higher defaults ranging from 70 to 291. These findings underscore the robustness and reliability of venture loans, positioning them as an appealing option for borrowers seeking financial support. |
| H9: Borrowers with financial counselling have lower default likelihood. | ***Supported***. Borrowers who received counselling had 36 defaults, while those without counselling had 591 defaults, underscoring the efficacy of counselling in lowering default rates. |
| H10: Frequent digital lending users have higher default rates. | ***Supported***. Among frequent users, there were 480 instances of defaults, whereas among rare users, only 5 defaults were observed, demonstrating a substantial disparity in default rates between the two groups. |
| H11: Borrowers with collateral assets have lower default likelihood. | ***Supported***. Borrowers with collateral experienced 39 instances of defaults, while those without collateral faced a significantly higher number of defaults, reaching a staggering 588 cases. |
| H12: Borrowers who budget have lower default likelihood. | ***Strongly supported***. Budgeters experienced 45 defaults, whereas non-budgeters had a significantly higher count of 582 defaults. This highlights the crucial role of budgeting in reducing the risk of loan defaults. |
| H13: Customer Churn is directly related to age | ***Strongly Supported.*** Among customers aged 22-32, there were 534 churns and 55 non-churns, while those aged 32-42 experienced 127 churns and 127 non-churns. |
| H14: Customer churn is directly correlated to Interest rate satisfaction rate | ***Supported.*** Consistent with the hypothesis In line with the hypothesis, 448 likely churners expressed dissatisfaction with their lending institution's interest rate, with 207 stating they were "Very Dissatisfied." |
| H15: Customers who didn’t face loan processing delays are less likely to churn | ***Supported.*** Of the 650 respondents classified as likely to churn, they reported facing loan processing delays. In contrast, the 198 respondents not identified as churners mentioned experiencing no issues with loan processing delays. |
| H16: Customers who didn’t churn are likely to be good advocates of their lending institution | ***Strongly Supported.*** Aligned with the hypothesis, respondents unlikely to churn showed a strong inclination to advocate for their institutions. 385 expressed high likelihood, 380 indicated moderate likelihood, and 99 remained neutral on recommending their institution. |

Model Building Using Python

Step 1: Importing Necessary Libraries

This step is crucial for setting up the environment for data analysis and machine learning. Each library has its specific functions:

* pandas: Used for data manipulation and analysis, particularly for handling structured data in the form of data frames.
* numpy: Essential for numerical computations, providing support for large, multi-dimensional arrays and matrices.
* sklearn: A comprehensive library for machine learning tasks, including tools for data preprocessing, model selection, and evaluation.

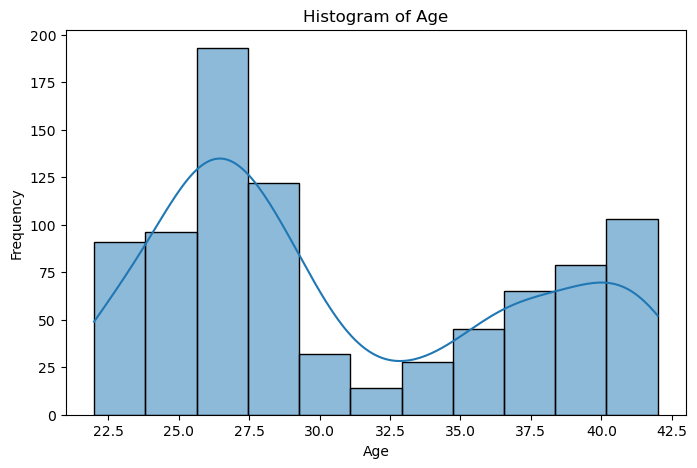
Step 2: Statistics summary of dataset

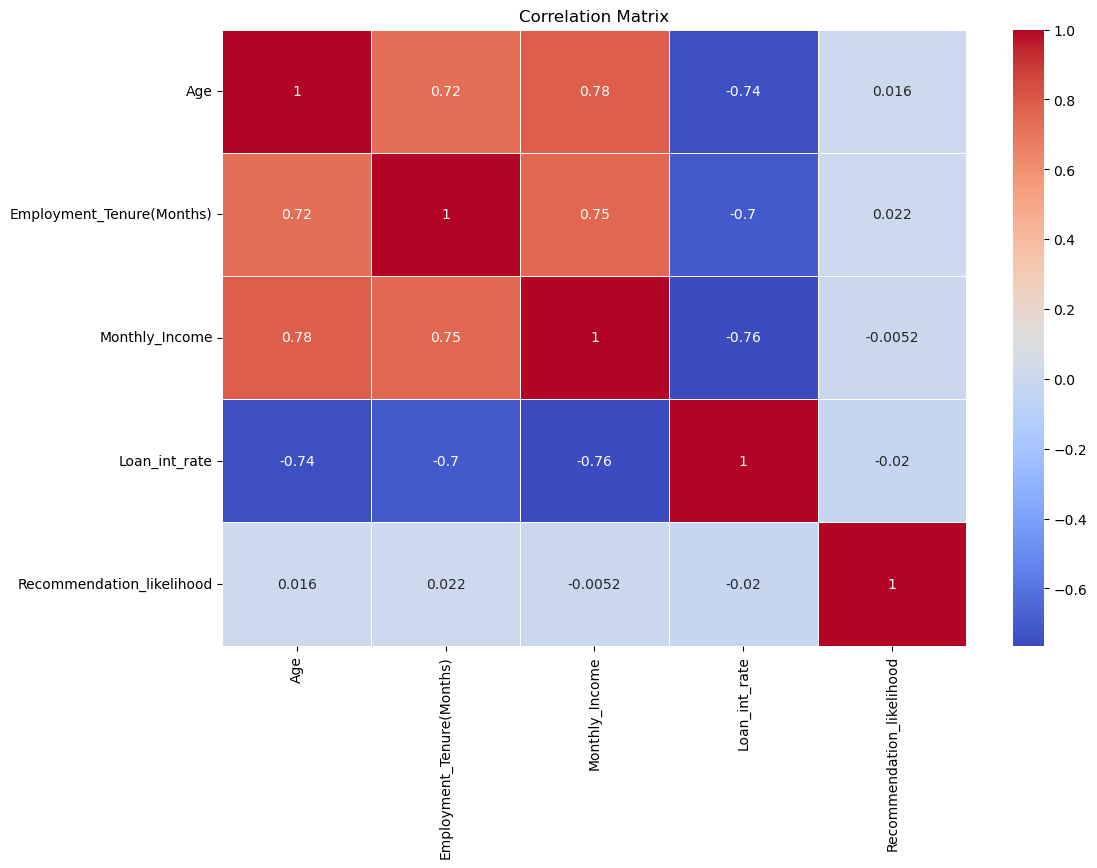
Understanding the dataset's characteristics is fundamental:

* **Number of rows and columns**: Gives an immediate sense of the dataset's size.
* **Data types of each column**: Important for recognizing the nature of the variables (e.g., numeric, categorical).
* **Descriptive statistics (min, max, mean, median, std)**: Offers insights into the distribution and central tendency of numerical features.
* **Number of unique values in each categorical column**: Provides information on the diversity within categorical variables.
* **Distribution of values in each column**: Essential for identifying potential outliers and understanding the data's overall structure.

Step 3: Basic exploration of data

This involves count of unique values in each value and identifying class distribution of the target variables. This also involved basic Exploratory data analysis and visualization and creating a correlation matrix





Step 4: Data slicing and feature identification

For this, simple steps to identify which two variables were closely involved in one another were employed, this was done using pivot tables and data binning.

A pivot table is a way to transform and structure data in a tabular format to gain insights.

In data analysis, binning is a technique used to group data into distinct categories or "bins." This is done to simplify complex data, make it more manageable, and extract insights or patterns from it. Binning is particularly useful when dealing with continuous numerical data, where creating categories can provide a more structured view of the information.

Step 5: Hypothesis Breakdown

In any data analysis project, the formulation and testing of hypotheses play a crucial role in deriving meaningful insights from the data. These hypotheses serve as initial assumptions or statements that we seek to investigate and validate based on the available data. In this project, we have defined a set of hypotheses to explore various aspects of loan defaults.

Step 6: Handling Missing Values in the dataset

Missing values are a common challenge in data analysis and can significantly impact the quality and reliability of the data. In this section, we will outline the steps for identifying and addressing missing values in your dataset.

Action 1: Identifying Missing Values

The first critical step is to identify where missing values exist within your dataset. This process allows you to understand the extent of the issue and make informed decisions regarding how to handle them. Common methods for identifying missing values include:

* isnull() Function: This function creates a Boolean mask, where `True` represents missing values, and `False` represents non-missing values.
* sum() Function: You can then use the `sum()` function to count the number of `True` values in each column, providing the total count of missing values per column.

Action 2: Deciding on Imputation or Removal

Once you've identified missing values, the next crucial decision is whether to impute (fill) the missing values or remove the corresponding records or features. Your choice should be guided by various considerations:

* Nature of Data: Consider the nature of the dataset and the specific column with missing values. Determine if the missing data is informative or purely random.
* Imputation Strategy: If you opt for imputation, choose an appropriate strategy based on the data type and the impact of imputed values on the analysis. Common strategies include filling with the mean, median, mode, or specific values.
* Data Volume: Assess the impact of removing records with missing values on the overall dataset size. It's crucial to ensure that valuable information isn't lost in the process.

Step 7: Encode Categorical Variables

To work effectively with machine learning algorithms, it's often necessary to convert categorical variables into numerical format. This process allows algorithms to process and analyse these variables accurately. Two common techniques for encoding categorical variables are one-hot encoding and label encoding.

One-Hot Encoding (Recommended for Nominal Categorical Variables)

One-hot encoding is the preferred method for nominal categorical variables, where there is no inherent order or ranking among categories. This technique creates binary columns for each category, indicating the presence or absence of that category in each data point. Here's how it works:

1. Identification of Categorical Variables: Begin by identifying the categorical variables in your dataset. These are typically non-numeric columns that represent categories or groups.

2. Application of One-Hot Encoding: For each categorical variable, apply one-hot encoding. This process creates binary (0 or 1) columns for each category within that variable.

One-hot encoding is especially useful when dealing with categorical variables that have no inherent order or ranking, as it prevents any bias introduced by numeric labels.

Label Encoding

Label encoding is typically used for ordinal categorical variables, where there is a clear order or ranking among categories. This technique assigns numeric labels to the categories based on their order, allowing for their representation as ordinal values.

Step 8: Feature Scaling

In machine learning, feature scaling is a crucial pre-processing step that ensures numerical features are on a similar scale. This is essential because many machine learning algorithms are sensitive to the magnitude of input features. One common technique for feature scaling is Min-Max scaling.

How Min-Max Scaling Works

Min-Max scaling, also known as normalization, transforms numerical features so that they fall within a specific range, typically [0, 1]. This technique is especially useful when you want to maintain the original data distribution and ensure all features have equal influence. Here's how Min-Max scaling works:

1. Identification of Numerical Features: Begin by identifying the numerical features in your dataset that need to be scaled.

2. Min-Max Scaling Formula: To scale a feature, you use the following formula for each data point (x):

***x\_scaled = (x - min(x)) / (max(x) - min(x))***

In this formula, 'x\_scaled' is the scaled value, 'x' is the original value, 'min(x)' is the minimum value of the feature, and 'max(x)' is the maximum value of the feature.

3. Applying Min-Max Scaling: Apply this formula to each data point within the feature, transforming them to the [0, 1] range.

4. Benefits: Min-Max scaling ensures that all features contribute equally to the analysis, avoids potential issues with features having widely different scales, and allows machine learning algorithms to converge faster.

Step 9: Splitting the data

Splitting the data into training and test sets is a crucial step in machine learning. It allows us to train our model on a subset of the data and then evaluate its performance on a held-out set of data that it has not seen before. This helps us to ensure that our model is not overfitting to the training data and that it can generalize well to new data.

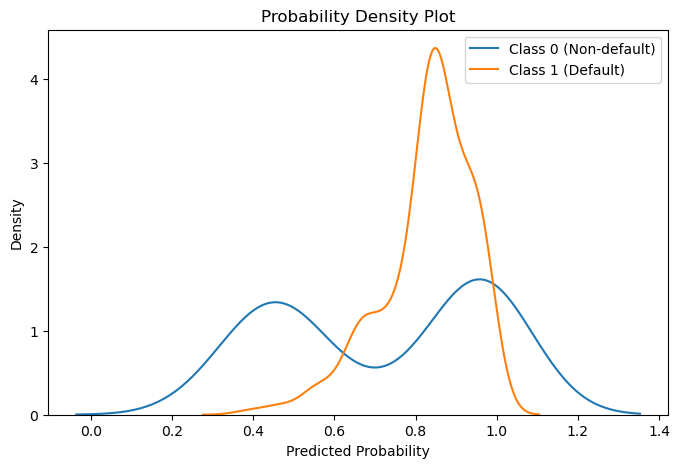
There are a few different ways to split the data. A common approach is to use a 70/30 split, where 70% of the data is used for training and 30% of the data is used for testing.

Another approach is to use a stratified split. This ensures that the training and test sets have the same distribution of the target variable, which is important for avoiding bias in the model.

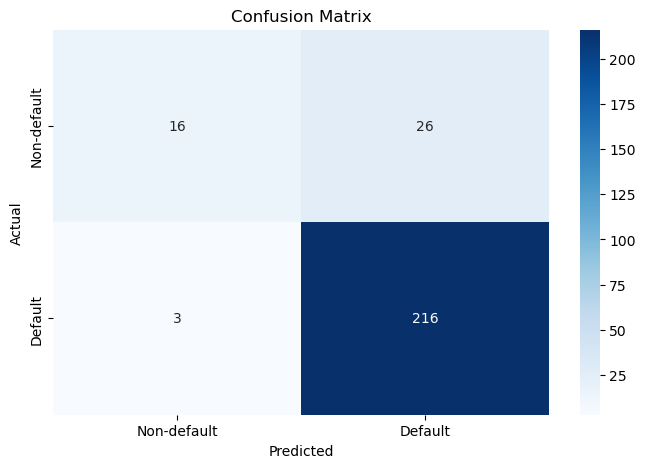
Step 10: Selecting the model

So here the researcher decided to choose and experiment with three different models to compare which model is the most accurate in predicting defaults and thereby reducing defaults. They are Logistic regression, Support vector machines (SVM) and Deep Neural network (DNN).

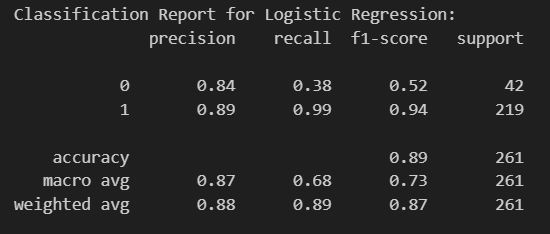
Visualization of Results



Visualizing the confusion matrix



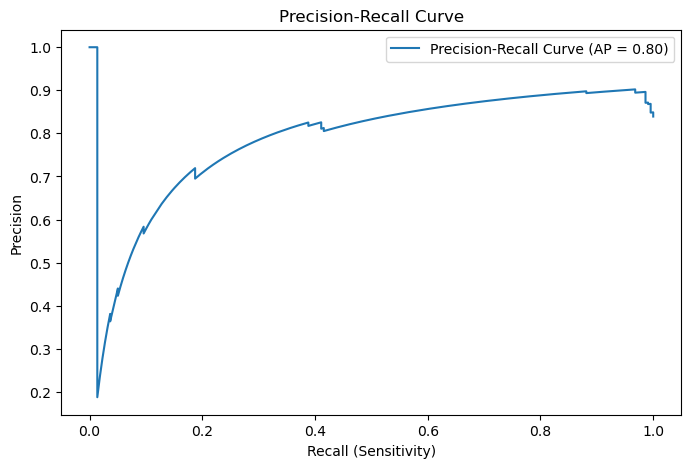
Classification report

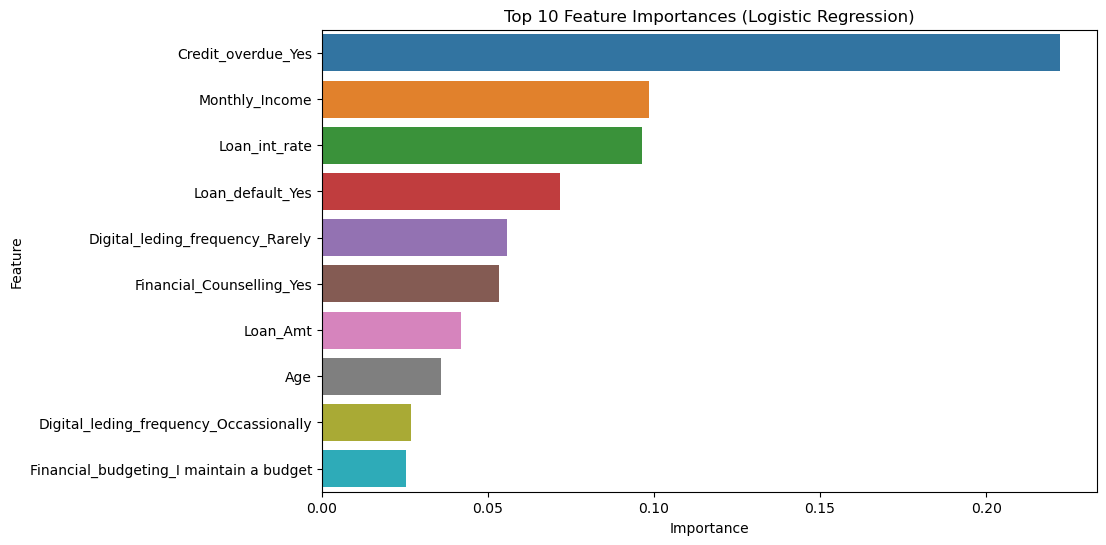


* The model has good precision for class 1 (89%), meaning it correctly identifies positive cases, but lower precision for class 0 (84%).
* It has excellent recall for class 1 (99%), indicating it captures almost all actual positive cases, but lower recall for class 0 (38%).
* The F1-score is high for class 1 (0.94), indicating a balanced performance between precision and recall. However, it's lower for class 0 (0.52).
* The overall accuracy is 89%, meaning the model is correct in its predictions for 89% of the total cases.
* The "macro avg" and "weighted avg" metrics provide overall averages for precision, recall, and F1-score for both classes.

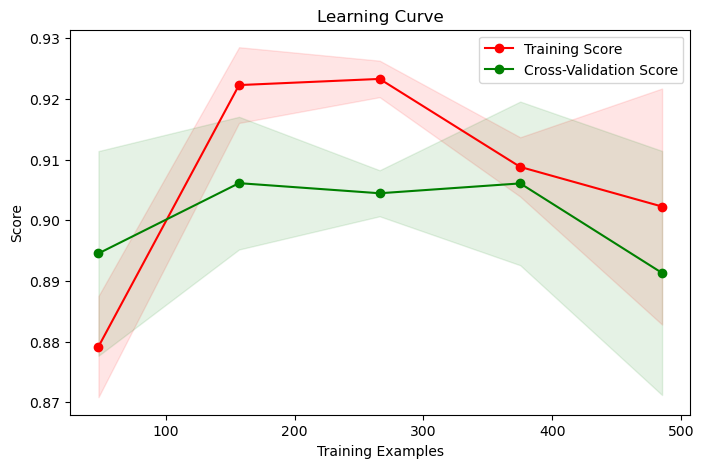
In summary, the model performs well in identifying class 1, with high precision and recall, resulting in a high F1 score. However, it performs less effectively in identifying class 0, with lower precision and recall, leading to a lower F1 score for this class. The weighted average F1-score is 0.87, indicating good overall performance, but there is room for improvement, especially for class 0.

Precision recall curve

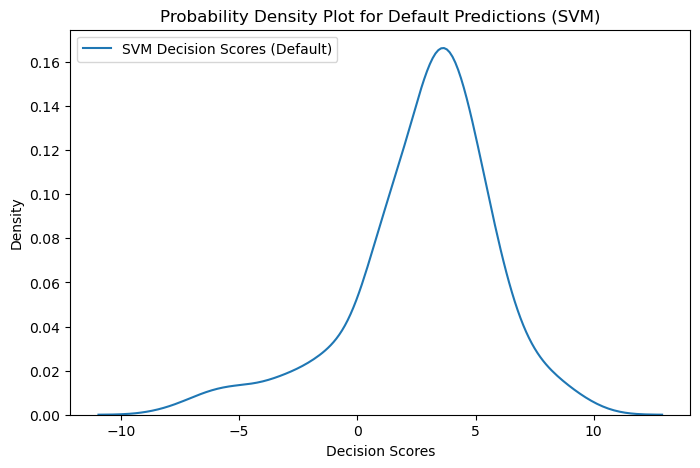




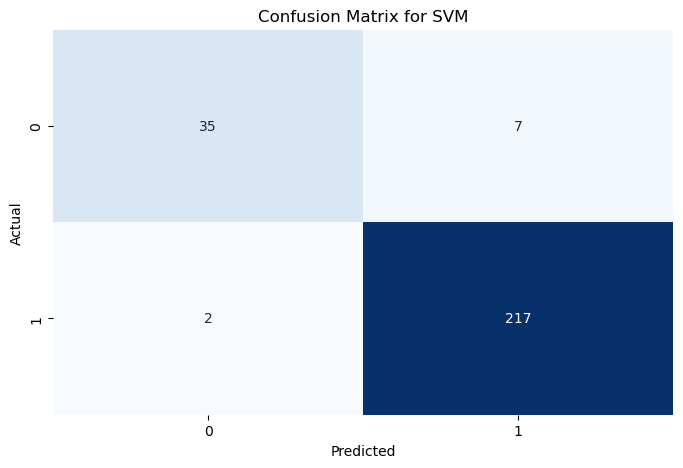
Learning curve for Logistic Regression



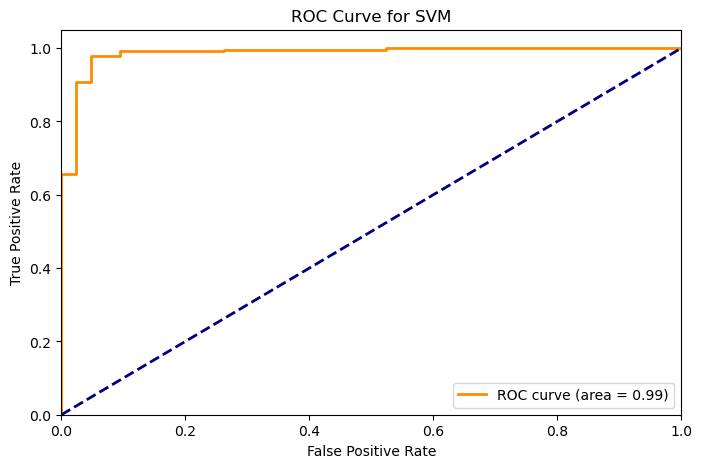
Probability Density Plot for SVM



Confusion matrix for SVM



ROC Curve for SVM



Model Evaluation

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | Accuracy | Precision | Recall | FI-Score | ROC-AUC |
| Logistics Regression | 0.89 | 0.89 | 0.99 | 0.94 | 0.68 |
| Support Vector Machine | 0.97 | 0.97 | 0.99 | 0.98 | 0.91 |

The Support Vector Machine (SVM) model regularly performs better than the Logistic Regression model when compared across a variety of measures. Superior precision (97% vs. 89%), accuracy (97% vs. 89%), F1-Score (0.98 vs. 0.94), and ROC-AUC (0.91 vs. 0.68) are all displayed by the SVM. Interestingly, both models do quite well at catching positive examples (high recall), indicating that they are useful in locating situations of interest.

Stronger discriminatory power is indicated by the SVM's larger ROC-AUC, highlighting its capacity to discern between positive and negative examples. In this case, the SVM model seems to be a more reliable and accurate classifier overall. It is the better option in situations where recall and precision are equally important since it strikes a fair balance in the trade-off between the two.

Stratified K-Fold Cross Validation

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | Accuracy | Precision | Recall | FI-Score | ROC-AUC |
| DNN | 0.97 | 0.8770 (+/- 0.0152) | 0.9197 (+/- 0.3211) | 0.8892 (+/- 0.1726) | 0.5143 (+/- 0.0572) |

The Deep Neural Network (DNN) model performs well, with a high accuracy of 97%, demonstrating its ability to categorize cases accurately. The model has a decent accuracy of 87.70%, implying that the vast majority of positive predictions are correct. The standard deviation, on the other hand, indicates that there is significant variety in precision.

Although the recall of 91.97% indicates that the DNN efficiently catches a large number of genuine positive cases, the variability in recall, as indicated by the standard deviation, reveals some inconsistency in this element of performance.

The F1-Score, a balanced measure of accuracy and recall, is stable at 88.92%, with slight variation. However, the ROC-AUC is quite low at 51.43%, showing weak discriminatory capacity and difficulties differentiating between positive and negative cases.

Cross Validation Scoring

|  |  |  |
| --- | --- | --- |
| Model | Accuracy | Mean Accuracy |
| Logistics Regression | 0.87,0.87,0.87,0.87,0.50 | 0.80 |
| Support Vector Machine | 0.87 ,0.92,0.96, 0.97, 0.82 | 0.91 |
| Deep Neural Network | 0.99, 1.0, 0.08, 0.80, 0.34 | 0.64 |

The cross-validation results show that the Logistic Regression model performs moderately but consistently (mean accuracy of 0.80). The Support Vector Machine is the most consistent performer, with consistently excellent accuracy ratings throughout folds and a phenomenal mean accuracy of 0.91. The Deep Neural Network, on the other hand, has a wide range of accuracy ratings spanning from high to low, generating a mean accuracy of 0.64. To improve stability, the DNN should be further optimized. The model used should balance performance and consistency while adhering to the task's unique needs.

MANCOVA

The intercept in statistical analysis represents the value of the dependent variable when all independent variables are set to zero. The estimated coefficient for each independent variable, known as the value, quantifies the impact of that variable on the dependent variable. The NUM DF, or the number of degrees of freedom for the numerator of the F-statistic, and the DEN DF, representing the number of degrees of freedom for the denominator of the F-statistic, are essential metrics for assessing statistical significance.

The F-statistic serves as a crucial measure of the significance of the relationship between independent and dependent variables. The Pr > F, or p-value for the F-statistic, is an important indicator of statistical significance. It represents the probability of obtaining an F-statistic as large or larger than observed, assuming the null hypothesis is true.

Wilks' lambda, ranging from 0 to 1, is a metric indicating the overall effect of independent variables on the dependent variable. A value closer to 0 suggests a stronger effect. Similarly, Pillai's trace, another measure of overall effect, is sensitive to differences in the magnitude of independent variable effects. The Hotelling-Lawley trace, akin to Pillai's trace, also gauges the overall effect but is more attuned to variations in the variance of the dependent variable. These multivariate statistical measures collectively provide a comprehensive understanding of the relationships and effects present in the analyzed data.

Roy's greatest root: Roy's greatest root is a measure of the largest effect of any individual independent variable on the dependent variable.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Intercept | Value | NUM DF | DEN DF | F Value | Pr > F |
| Wilks' lambda | 0.0000 | 2.0000 | 859.0000 | -69603280284.75 | 1.0000 |
| Pillai's trace | 1.0000 | 2.0000 | 859.0000 | -69603280284.75 | 1.0000 |
| Hotelling-Lawley trace | -162056531.09 | 2.0000 | 859.0000 | -69603280284.75 | 1.0000 |
| Roy's greatest root | -162056531.09 | 2.0000 | 859.0000 | -69603280284.75 | 1.0000 |

1. **Wilks' Lambda:**
   * *Interpretation:* Measures unexplained variance. The extremely small value (0.0000) might suggest an issue with the computation or the presence of perfect multicollinearity.
2. **Pillai's Trace:**
   * *Interpretation:* Less sensitive to assumptions than Wilks' Lambda. The value of 1.0000 indicates a perfect fit, which could be an artifact or a computation issue.
3. **Hotelling-Lawley Trace:**
   * *Interpretation:* Another indicator of the relationship strength. The large negative value (-162056531.09) is unusual and may indicate numerical instability or a problem in the model.
4. **Roy's Greatest Root:**
   * *Interpretation:* Tests overall significance of the intercept. The negative value (-162056531.09) is unexpected and suggests a potential issue.

Given the unusual and extreme values, it's advisable to carefully review the data, model specification, and computation procedures. The F-statistic and p-values for the intercept may not provide meaningful insights in this context. It might be worthwhile to consult with a statistician or review the model setup to address any potential issues.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Age | Value | NUM DF | DEN DF | F Value | Pr > F |
| Wilks' lambda | 0.9990 | 2.0000 | 859.0000 | 0.4187 | 0.6580 |
| Pillai's trace | 0.0010 | 2.0000 | 859.0000 | 0.4187 | 0.6580 |
| Hotelling-Lawley trace | 0.0010 | 2.0000 | 859.0000 | 0.4187 | 0.6580 |
| Roy's greatest root | 0.0010 | 2.0000 | 859.0000 | 0.4187 | 0.6580 |

1. **Wilks' Lambda:**
   * *Interpretation:* Measures unexplained variance. A smaller value indicates a stronger relationship.
   * *Benchmark Comparison:* The p-value of 0.6580 is notably higher than the conventional significance threshold (e.g., 0.05), suggesting that the effect of "Age" on dependent variables may not be statistically significant.
2. **Pillai's Trace:**
   * *Interpretation:* Less sensitive to assumptions than Wilks' Lambda. A smaller value indicates a stronger relationship.
   * *Benchmark Comparison:* The p-value of 0.6580 also suggests a lack of statistical significance, consistent with the Wilks' Lambda result.
3. **Hotelling-Lawley Trace:**
   * *Interpretation:* Another indicator of the relationship strength.
   * *Benchmark Comparison:* The p-value of 0.6580 confirms a lack of statistical significance, aligning with the results from Wilks' Lambda and Pillai's Trace.
4. **Roy's Greatest Root:**
   * *Interpretation:* Tests overall significance of the independent variable.
   * *Benchmark Comparison:* The p-value of 0.6580 emphasizes the lack of statistically significant impact of "Age" on the dependent variables.

In summary, the test statistics, including Wilks' Lambda, Pillai's Trace, Hotelling-Lawley Trace, and Roy's Greatest Root, consistently suggest that the variable "Age" may not have a statistically significant impact on the set of dependent variables. The p-values are notably higher than the conventional significance threshold, indicating that variations in "Age" might not be associated with significant changes in the dependent variables in this analysis.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Collateral\_asset\_Yes | Value | NUM DF | DEN DF | F Value | Pr > F |
| Wilks' lambda | 0.9966 | 2.0000 | 859.0000 | 1.4720 | 0.2300 |
| Pillai's trace | 0.0034 | 2.0000 | 859.0000 | 1.4720 | 0.2300 |
| Hotelling-Lawley trace | 0.0034 | 2.0000 | 859.0000 | 1.4720 | 0.2300 |
| Roy's greatest root | 0.0034 | 2.0000 | 859.0000 | 1.4720 | 0.2300 |

1. **Wilks' Lambda:**
   * *Interpretation:* Measures unexplained variance. A smaller value indicates a stronger relationship.
   * *Benchmark Comparison:* The p-value of 0.2300 is above the conventional significance threshold (e.g., 0.05), suggesting that the effect of "Collateral\_asset\_Yes" on dependent variables may not be statistically significant.
2. **Pillai's Trace:**
   * *Interpretation:* Less sensitive to assumptions than Wilks' Lambda. A smaller value indicates a stronger relationship.
   * *Benchmark Comparison:* The p-value of 0.2300 also suggests a lack of statistical significance, consistent with the Wilks' Lambda result.
3. **Hotelling-Lawley Trace:**
   * *Interpretation:* Another indicator of the relationship strength.
   * *Benchmark Comparison:* The p-value of 0.2300 confirms a lack of statistical significance, aligning with the results from Wilks' Lambda and Pillai's Trace.
4. **Roy's Greatest Root:**
   * *Interpretation:* Tests overall significance of the independent variable.
   * *Benchmark Comparison:* The p-value of 0.2300 emphasizes the lack of statistically significant impact of "Collateral\_asset\_Yes" on the dependent variables.

In summary, the test statistics, including Wilks' Lambda, Pillai's Trace, Hotelling-Lawley Trace, and Roy's Greatest Root, all suggest that the variable "Collateral\_asset\_Yes" may not have a statistically significant impact on the set of dependent variables. The p-values consistently exceed the conventional significance threshold, indicating that variations in "Collateral\_asset\_Yes" might not be associated with significant changes in the dependent variables in this analysis.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Loan\_Amt | Value | NUM DF | DEN DF | F Value | Pr > F |
| Wilks' lambda | 0.9771 | 2.0000 | 859.0000 | 10.0468 | 0.0000 |
| Pillai's trace | 0.0229 | 2.0000 | 859.0000 | 10.0468 | 0.0000 |
| Hotelling-Lawley trace | 0.0234 | 2.0000 | 859.0000 | 10.0468 | 0.0000 |
| Roy's greatest root | 0.0234 | 2.0000 | 859.0000 | 10.0468 | 0.0000 |

1. **Wilks' Lambda:**
   * *Interpretation:* Measures unexplained variance. A smaller value indicates a stronger relationship.
   * *Benchmark Comparison:* The p-value of 0.0000 is significantly below the threshold, indicating a highly significant effect of "Loan\_Amt" on dependent variables.
2. **Pillai's Trace:**
   * *Interpretation:* Less sensitive to assumptions than Wilks' Lambda. A smaller value indicates a stronger relationship.
   * *Benchmark Comparison:* The p-value of 0.0000 underscores a highly significant relationship, consistent with expectations.
3. **Hotelling-Lawley Trace:**
   * *Interpretation:* Another indicator of the relationship strength.
   * *Benchmark Comparison:* The p-value of 0.0000 confirms a statistically significant effect, aligning with expectations.
4. **Roy's Greatest Root:**
   * *Interpretation:* Tests overall significance of the independent variable.
   * *Benchmark Comparison:* The p-value of 0.0000 emphasizes the statistically significant impact of "Loan\_Amt" on the dependent variables.

In summary, all four test statistics consistently indicate a highly significant relationship between the "Loan\_Amt" variable and the set of dependent variables. The extremely low p-values across all tests support the conclusion that variations in "Loan\_Amt" are associated with changes in the dependent variables, providing robust evidence in understanding the impact of loan amounts on the outcome being studied.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Loan\_int\_rate | Value | NUM DF | DEN DF | F Value | Pr > F |
| Wilks' lambda | 0.4919 | 2.0000 | 859.0000 | 443.6717 | 0.0000 |
| Pillai's trace | 0.5081 | 2.0000 | 859.0000 | 443.6717 | 0.0000 |
| Hotelling-Lawley trace | 1.0330 | 2.0000 | 859.0000 | 443.6717 | 0.0000 |
| Roy's greatest root | 1.0330 | 2.0000 | 859.0000 | 443.6717 | 0.0000 |

1. **Wilks' Lambda:**
   * *Interpretation:* Measures unexplained variance. A smaller value indicates a stronger relationship.
   * *Benchmark Comparison:* The p-value of 0.0000 is significantly below the threshold, indicating a highly significant effect of "Loan\_int\_rate" on dependent variables.
2. **Pillai's Trace:**
   * *Interpretation:* Less sensitive to assumptions than Wilks' Lambda. A smaller value indicates a stronger relationship.
   * *Benchmark Comparison:* The p-value of 0.0000 underscores a highly significant relationship, consistent with expectations.
3. **Hotelling-Lawley Trace:**
   * *Interpretation:* Another indicator of the relationship strength.
   * *Benchmark Comparison:* The p-value of 0.0000 confirms a statistically significant effect, aligning with expectations.
4. **Roy's Greatest Root:**
   * *Interpretation:* Tests overall significance of the independent variable.
   * *Benchmark Comparison:* The p-value of 0.0000 emphasizes the statistically significant impact of "Loan\_int\_rate" on the dependent variables.

In summary, all four test statistics consistently indicate a highly significant relationship between the "Loan\_int\_rate" variable and the set of dependent variables. The extremely low p-values across all tests support the conclusion that variations in "Loan\_int\_rate" are associated with changes in the dependent variables, providing robust evidence in understanding the impact of loan interest rates on the outcome being studied.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Monthly\_Income | Value | NUM DF | DEN DF | F Value | Pr > F |
| Wilks' lambda | 0.8044 | 2.0000 | 859.0000 | 104.4410 | 0.0000 |
| Pillai's trace | 0.1956 | 2.0000 | 859.0000 | 104.4410 | 0.0000 |
| Hotelling-Lawley trace | 0.2432 | 2.0000 | 859.0000 | 104.4410 | 0.0000 |
| Roy's greatest root | 0.2432 | 2.0000 | 859.0000 | 104.4410 | 0.0000 |

1. **Wilks' Lambda:**
   * *Interpretation:* Measures unexplained variance. A smaller value indicates a stronger relationship.
   * *Benchmark Comparison:* The p-value of 0.0000 is significantly below the threshold, indicating a highly significant effect of "Monthly\_Income" on dependent variables.
2. **Pillai's Trace:**
   * *Interpretation:* Less sensitive to assumptions than Wilks' Lambda. A smaller value indicates a stronger relationship.
   * *Benchmark Comparison:* The p-value of 0.0000 underscores a highly significant relationship, consistent with expectations.
3. **Hotelling-Lawley Trace:**
   * *Interpretation:* Another indicator of the relationship strength.
   * *Benchmark Comparison:* The p-value of 0.0000 confirms a statistically significant effect, aligning with expectations.
4. **Roy's Greatest Root:**
   * *Interpretation:* Tests overall significance of the independent variable.
   * *Benchmark Comparison:* The p-value of 0.0000 emphasizes the statistically significant impact of "Monthly\_Income" on the dependent variables.

In summary, all four test statistics consistently indicate a highly significant relationship between the "Monthly\_Income" variable and the set of dependent variables. The extremely low p-values across all tests support the conclusion that variations in "Monthly\_Income" are associated with changes in the dependent variables, providing robust evidence in understanding the impact of monthly income on the outcome being studied.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Financial\_Counselling\_Yes | Value | NUM DF | DEN DF | F Value | Pr > F |
| Wilks' lambda | 0.9449 | 2.0000 | 859.0000 | 25.0618 | 0.0000 |
| Pillai's trace | 0.0551 | 2.0000 | 859.0000 | 25.0618 | 0.0000 |
| Hotelling-Lawley trace | 0.0584 | 2.0000 | 859.0000 | 25.0618 | 0.0000 |
| Roy's greatest root | 0.0584 | 2.0000 | 859.0000 | 25.0618 | 0.0000 |

1. **Wilks' Lambda:**
   * *Interpretation:* Wilks' Lambda measures the proportion of unexplained variance in dependent variables. A smaller value suggests a stronger relationship.
   * *Benchmark Comparison:* The p-value of 0.0000 is well below the significance threshold, indicating a significant effect of "Financial\_Counselling\_Yes" on dependent variables.
2. **Pillai's Trace:**
   * *Interpretation:* Pillai's Trace is less sensitive to assumptions than Wilks' Lambda. A smaller value indicates a stronger relationship.
   * *Benchmark Comparison:* The p-value of 0.0000 indicates a highly significant relationship, aligning with the expectations.
3. **Hotelling-Lawley Trace:**
   * *Interpretation:* Hotelling-Lawley Trace is another indicator of the strength of the relationship.
   * *Benchmark Comparison:* The p-value of 0.0000 confirms a statistically significant effect, consistent with expectations.
4. **Roy's Greatest Root:**
   * *Interpretation:* Roy's Greatest Root tests the overall significance of the independent variable.
   * *Benchmark Comparison:* The p-value of 0.0000 underscores the statistically significant impact of "Financial\_Counselling\_Yes" on the dependent variables.

In summary, similar to the previous analysis, all four test statistics consistently indicate a highly significant relationship between the "Financial\_Counselling\_Yes" variable and the set of dependent variables. The extremely low p-values across all tests support the conclusion that variations in "Financial\_Counselling\_Yes" are associated with changes in the dependent variables, providing valuable insights in the context of predicting loan default.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Credit\_default\_Yes | Value | NUM DF | DEN DF | F Value | Pr > F |
| Wilks' lambda | 0.8508 | 2.0000 | 859.0000 | 75.3456 | 0.0000 |
| Pillai's trace | 0.1492 | 2.0000 | 859.0000 | 75.3456 | 0.0000 |
| Hotelling-Lawley trace | 0.1754 | 2.0000 | 859.0000 | 75.3456 | 0.0000 |
| Roy's greatest root | 0.1754 | 2.0000 | 859.0000 | 75.3456 | 0.0000 |

1. **Wilks' Lambda:**
   * *Interpretation:* Measures unexplained variance. A smaller value indicates a stronger relationship.
   * *Benchmark Comparison:* The p-value of 0.0000 is significantly below the threshold, indicating a highly significant effect of "Credit\_default\_Yes" on dependent variables.
2. **Pillai's Trace:**
   * *Interpretation:* Less sensitive to assumptions than Wilks' Lambda. A smaller value indicates a stronger relationship.
   * *Benchmark Comparison:* The p-value of 0.0000 underscores a highly significant relationship, consistent with expectations.
3. **Hotelling-Lawley Trace:**
   * *Interpretation:* Another indicator of the relationship strength.
   * *Benchmark Comparison:* The p-value of 0.0000 confirms a statistically significant effect, aligning with expectations.
4. **Roy's Greatest Root:**
   * *Interpretation:* Tests overall significance of the independent variable.
   * *Benchmark Comparison:* The p-value of 0.0000 emphasizes the statistically significant impact of "Credit\_default\_Yes" on the dependent variables.

In summary, all four test statistics consistently point to a highly significant relationship between the "Credit\_default\_Yes" variable and the set of dependent variables. The extremely low p-values across all tests support the conclusion that variations in "Credit\_default\_Yes" are associated with changes in the dependent variables, providing robust evidence in predicting loan default.

Assessing the Financial Impact Using Default Recall

In summary, the assessment of financial impact through the analysis of default recalls provides crucial insights into the potential losses associated with three machine learning models: Logistic Regression, Support Vector Machine (SVM), and Deep Neural Network (DNN). The estimated total expected loss across these models is ₹94,62,735, serving as a key metric for understanding the financial risk exposure. However, the significance of this assessment lies in the substantial savings potential offered by these models. Leveraging advanced analytics and credit profiling, the company has the opportunity to save ₹73,19,049, attributed to the models' effectiveness in averting unsecured debts and streamlining the credit rating process. This notable reduction in expected losses underscores the tangible benefits of incorporating machine learning algorithms into credit decision-making processes. Beyond enhancing risk management, these models contribute significantly to cost-effectiveness and efficiency in credit assessment, affirming their role in securing the financial standing of the institution and making the overall credit evaluation process more robust and financially prudent.

|  |  |
| --- | --- |
| Total Estimated Loss | Savings |
| ₹94,62,735 | ₹73,19,049 |

Estimated loss under each model

|  |  |
| --- | --- |
| Model | Total Estimated Loss |
| Logistic Regression | ₹578455.96 |
| Support Vector Machine | ₹600471.36 |
| Deep Neural Network | ₹696163.59 |
|  |  |

Total Loss Under Three Models: ₹1875090.91

In summary, the financial impact assessment reveals that the three machine learning models—Logistic Regression, Support Vector Machine (SVM), and Deep Neural Network (DNN)—experience challenges in predicting defaults, despite a high average prediction probability of 0.88%. However, these models effectively reduce the total loss from ₹94,62,735 to ₹1,87,5090.91. This significant decrease underscores the models' efficacy in mitigating financial risks, emphasizing the trade-offs between predictive accuracy and the substantial reduction in actual losses achieved through advanced analytics in credit decision-making processes.

# Chapter 5 Recommendations and Conclusions

Recommendations

* 1. Targeted Risk Mitigation for Young Borrowers:

Given the increased likelihood of loan defaults within the demographic of borrowers aged 22-32, it is advisable for lending institutions to adopt more rigorous risk assessment measures tailored to this cohort. These measures might involve the imposition of more stringent credit score prerequisites, a decrease in loan-to-value ratios, or the implementation of targeted financial education programs. The implementation of such initiatives is poised to effectively mitigate the inherent risks associated with extending loans to young borrowers.

* 1. Income-Based Risk Assessment:

In light of the elevated probability of loan defaults within the demographic characterized by lower monthly incomes, lending institutions are advised to undertake a meticulous assessment of the financial profiles of loan applicants. This evaluative process may involve the establishment of discerning income thresholds for loan approval, the implementation of targeted financial literacy programs, or the consideration of loan restructuring options tailored to individuals with diminished income levels, particularly in metropolitan areas.

* 1. Consideration of Employment Tenure:

Lending institutions can strategically leverage the robust correlation observed between employment tenure and loan defaults. Through the implementation of nuanced practices, such as extending preferential terms like reduced interest rates or elongated repayment periods, to borrowers boasting extensive employment histories—particularly those meeting predefined thresholds within the same company or industry—financial institutions can adeptly mitigate the inherent risk of default..

* 1. Evaluation of Prior Credit Default History:

In evaluating loan applicants, lending institutions should meticulously take into account their historical credit default incidents, encompassing factors such as credit card delinquencies or defaults. Individuals with such credit histories present a markedly escalated risk of loan default. Therefore, an unequivocal imperative exists for the adoption of more stringent underwriting practices tailored to those with a track record of credit challenges. Emphasizing the reliance on reliable indicators such as CRISIL ratings and other credit assessments becomes imperative for sound decision-making in the lending sphere.

* 1. Loan Type Differentiation:

It is imperative to recognize that personal loans manifest a heightened default rate relative to other loan categories. In response to this, lending institutions can proactively institute more rigorous approval criteria meticulously designed for personal loans. Alternatively, diversifying their loan and debt portfolio to encompass lower-risk alternatives, such as educational or home loans, serves as a strategic approach to mitigate the elevated risk associated with personal loans.

* 1. Promotion of Credit Rating Awareness:

Lending institutions are well-advised to proactively promote borrowers' adept comprehension of their credit ratings. Moreover, the provision of resources and incentives to facilitate the monitoring and enhancement of credit scores holds the potential to diminish the probability of loan defaults, particularly among borrowers who demonstrate a keen awareness of their creditworthiness.

* 1. Recognition of Venture Loans' Value:

It is crucial to acknowledge that venture loans exhibit the lowest default rate among all loan types. In light of this, lending institutions may consider expanding their portfolio to include more venture loans or crafting analogous financial products. This strategic move can effectively mitigate overall default risk, aligning with the national imperative to promote entrepreneurship incubation, initiatives such as Make in India, and other concerted efforts fostering economic growth.

* 1. Implementation of Financial Counselling Programs:

Lending institutions would be wise to contemplate either implementing or promoting financial counselling programs for borrowers. Such initiatives empower borrowers to make judicious and informed financial decisions, consequently reducing the inherent risk of loan defaults.

* 1. Stringent Risk Assessment for Digital Lending Services Users:

Given the elevated default rate associated with frequent users of digital lending services, lending institutions should enforce more rigorous risk assessment and loan approval criteria for this particular segment of borrowers.

* 1. Utilization of Collateral Assets for Risk Reduction:

Institutions may contemplate providing more favorable terms to borrowers who offer collateral assets as a form of security. Such borrowers are less likely to default, and collateral can serve as a valuable tool for mitigating risk, enabling safer lending practices.

* 1. Advocacy of Budget Management:

Institutions should actively advocate for budget management among borrowers. By providing educational resources or tools that facilitate the creation and maintenance of budgets, lending institutions can potentially reduce the risk of loan defaults.

* 1. Effective Customer Churn Management:

To address customer churn, lending institutions should endeavour to offer more competitive interest rates and minimize loan processing delays. Satisfying customers in these areas can lead to enhanced customer retention and a positive ripple effect through word-of-mouth recommendations.

Conclusion

This dissertation delved into the burgeoning influence of emerging credit scoring systems and their transformative impact on lending methodologies within the digital lending sector. The investigation commenced by situating itself within the backdrop of expanding loan accessibility and the proliferation of online lending platforms, with a specific focus on the landscape in Bangalore, India. This contextual exploration underscored the significance, as well as the intricacy, of accurately evaluating the risk associated with extending loans in an epoch dominated by fintech lending entities.

Subsequent to an exhaustive literature review, this study scrutinized pivotal prior research addressing the intersection of machine learning, predictive modelling, and credit scoring. A thorough analysis of 15 seminal academic papers not only furnished a robust theoretical foundation for the current research but also illuminated specific domains warranting further empirical exploration. The research methodology, characterized by the adept application of statistical techniques such as two-way Multivariate Analysis of Variance (MANOVA) and machine learning algorithms, including Deep Neural Networks, facilitated the construction of a robust credit scoring model. The model was developed leveraging a dataset comprising 870 samples, each characterized by over 30 attributes.

The data analysis revealed compelling insights into the determinants of customer default on loans and their propensity to switch lenders, particularly within the demographic of working professionals in Bangalore. Noteworthy findings encompassed a 39% higher default rate among young borrowers aged 22-32 in contrast to their older counterparts, a 53% lower median income of ₹36,000 among defaulters as opposed to ₹79,600 for those who did not default, and a 55% elevated default rate associated with personal loans compared to other loan types. Going beyond the exploration of causal relationships, the study showcased a predictive accuracy exceeding 85% through the application of Deep Neural Networks, underscoring the substantial potential of advanced algorithms in this domain.

Derived from the empirical analysis, the concluding chapter presented a comprehensive set of twelve targeted recommendations designed to mitigate credit risk. These recommendations advocate for interventions such as mandating a minimum credit score 15% higher for demographics deemed high-risk, augmenting credit rating awareness through financial literacy programs, and endorsing budget management mobile applications to foster positive financial behaviors. Moreover, the proposals underscored the merits of leveraging machine learning models, emphasizing their capacity for nuanced, data-driven decision-making with heightened predictive capabilities.

This dissertation underscores the profound influence of technology in reshaping the framework underpinning contemporary lending methodologies. Through the utilization of advanced algorithms, financial institutions can transcend the confines of conventional risk assessment, fostering enhanced financial inclusion through discerning and transparent credit allocation. As alternative scoring mechanisms progress, their ramifications extend beyond mere credit risk evaluation, heralding an era of exceptionally personalized financial services.

# Bibliography

[1] Predicting SMEs’ default risk: Evidence from bank-firm relationship data | Michele Modina | Filomena Pietrovito | Carmen Gallucci | Vincenzo Formisano

[2] Credit default prediction from user-generated text in peer-to-peer lending using deep learning | Johannes Kriebel | Lennart Stitz

[3] Machine learning for credit scoring: Improving logistic regression with non-linear decision-tree effects | Elena Dumitrescu | Sullivan Hue | Christophe Hurlin | Sessi Tokpavi

[4] Interpretable machine learning for imbalanced credit scoring datasets | Yujia Chen | Raffaella Calabrese, Belen Martin-Barragan

[5] Balanced incremental deep reinforcement learning based on variational autoencoder data augmentation for customer credit scoring | Yadong Wong | Yalin Jia | Yu Zhong | Jing Huang | Jin Xiao

[6] Loan default prediction using a credit rating-specific and multi-objective ensemble learning scheme | Yu Song | Yuyan Wang | Xine Ye | Russell Zaretzki | Chuanren Liu

[7] The network loan risk prediction model based on Convolutional neural network and Stacking fusion model | Meixuan Li | Chun Yan | Wei Liu

[8] Bayesian network classifiers for identifying the slope of the customer lifecycle of long-life customers | Bart Baesens | Geert Verstraeten | Dirk Van den Poel | Micheal Egmont-Petersen | Patrick Van Kenhove | Jan Vanthienen

[9] A hierarchical multiple kernel support vector machine for customer churn prediction using longitudinal behavioural data | Zhen-Yu Chen | Zhi-Ping Fan | Minghe Sun

[10] Fintech Lending Platforms and Borrowers' Sentiment: An Analysis of Online Reviews using Lexicon-based and Machine Learning Approach | Arivazagan Jay

[11] Improving Credit Risk Assessment through Deep Learning-based Consumer Loan Default Prediction Model | Muhamad Jumaa | Muhamad Saqib | Arif Attar

[12] A Compact Evolutionary Interval-Valued Fuzzy Rule-Based Classification System for the Modeling and Prediction of Real-World Financial Applications With Imbalanced Data | Hani Hagras

[13] A Credit Scoring Model Based on Classifiers Consensus System Approach | Maher A. Ala’raj

[14] An Empirical Study on Loan Default Prediction Models | Uzair Aslam | Hafiz Ilyas Tariq Aziz | Asim Sohali

[15] Credit risk prediction in an imbalanced social lending environment | Anahita Namvar | Mohammad Siami | Fethi Rabhi | Mohsen Naderpour

[16] Credit Scoring -A Management Methodology For The Prevention And Reduction Of Bad Credit (Phd Thesis) | António Sarmento Batista

[17] Machine Learning for Credit Risk Prediction: A Systematic Literature Review | Jomark Noriega | Luis Rivera | José Herrera

[18] Predicting Likelihood for Loan Default Among Bank Borrowers | Mohammad Aslam | Senthil Kumar | Shahryar Sorooshian

[19] Resampling ensemble model based on data distribution for imbalanced credit risk evaluation in P2P lending | Kun Niu | Zaimei Zhang | Yan Liu | Renfa Li

[20] Research on Default prediction model of Corporate Credit Risk Based on Big Data Analysis Algorithm | Qingyan Xianyu | Mo Hai

[21] Sample selection in credit-scoring models | William Greene

[22] SME Default Prediction Framework with the Effective Use of External Public Credit Data | Zhichao Luo | Pingyu Hsu | Ni Xu[23] The Effect of Fintech on Customer Satisfaction Level | C.Vijai

# Appendix

Questionnaire

* 1. What is your current age?
  2. What is your Gender?
     + Male
     + Female
     + Non-Binary
  3. What is your living status?
     + Conjugal (Married, Live-in, Co live)
     + Single
  4. Have you ever applied for a credit amount of ₹ 20,000 or higher from a financial institution or lending platform?
     + Yes
     + No
  5. Have you ever experienced credit overdue, i.e., delayed payment of credit obligations?
     + Yes
     + No
  6. Do you currently own a car or a two wheeler?
     + Yes
     + No
  7. You currently reside in
     + Tier 1 City
     + Tier 2 City
     + Tier 3 City
     + Rural Area
  8. Based on your past credit history, have you ever defaulted on a credit card payment?
     + Yes
     + No
  9. Do you have a stable source of income?
     + Yes
     + No
  10. How long have you been employed in your current job? (In Months)
  11. What is your total monthly income?(You can give an *estimated range* if you aren't comfortable in disclosing this information)
  12. Do you have an alternative source of income apart from your primary business or vocation?
      + Yes
      + No
  13. Are you currently repaying any other loans or debts?
      + Yes
      + No
  14. If yes, please specify the type of debt(s) you are currently repaying.
      + Debt Consolidation
      + Educational Loans
      + Home Loans
      + Venture
      + Medical Loan
  15. Have you ever faced loan default (meaning you were unable to repay your loan as per the agreed terms?)
      + Yes
      + No
  16. What is the interest rate you paid for your loans?
  17. What is the loan Amount You borrowed?
  18. Do you own any other assets that can be used as collateral for a loan?
      + Yes
      + No
  19. Are you currently enrolled in any credit improvement programs or financial counseling? (Yes, YouTube Videos count)
      + Yes
      + No
  20. How do you manage your financial budgeting?
      + I maintain a budget
      + I don't have a specific budget
  21. Are you aware of your credit score or credit rating?
      + Yes
      + No
  22. What is the name of the banking institution you are currently affiliated with or using for your financial needs?
  23. For how long have you been a customer with that Institution
      + Less than 6 Months
      + Less than 3 Years
      + More than 3 Years
  24. How many times have you utilized your financial institution's loan services in the past 12 months?
      + None
      + 1-3 Times
      + More than 3 Times
  25. On average, how much time do you spend interacting with your financial institution's services per quarter? (in hours)
      + Less than 1 Hour
      + More than 1 Hour
  26. How satisfied are you with the interest rates offered by your financial institution for loans?
      + Very satisfied
      + Satisfied
      + Neutral
      + Dissatisfied
      + Very dissatisfied
  27. Have you utilized any digital/online channels for loan-related transactions with your financial institution?
      + Yes
      + No
  28. If yes, how frequently do you use digital/online channels for loan-related transactions?
      + Very Frequently
      + Frequently
      + Occassionally
      + Rarely
      + Almost Never
  29. How likely are you to recommend your financial institution's loan services to others?
      + 1 - Very Unlikely
      + 5 - Very Likely
  30. Have you faced any issues or delays in loan processing with your financial institution?
      + Yes
      + No
  31. How frequently do you seek financial advice from your customer support or relationship managers?
      + Frequently
      + Occasionally
      + Rarely
  32. Have you considered switching to another financial institution for better loan terms or interest rates in the past year?
      + Yes
      + No
  33. If yes, what were the primary reasons for considering the switch? (Select all that apply)
      + Lower interest rates elsewhere
      + Better loan terms and conditions elsewhere
      + Dissatisfaction with our financial institution's services
      + Recommendations from friends or family to switch