# 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

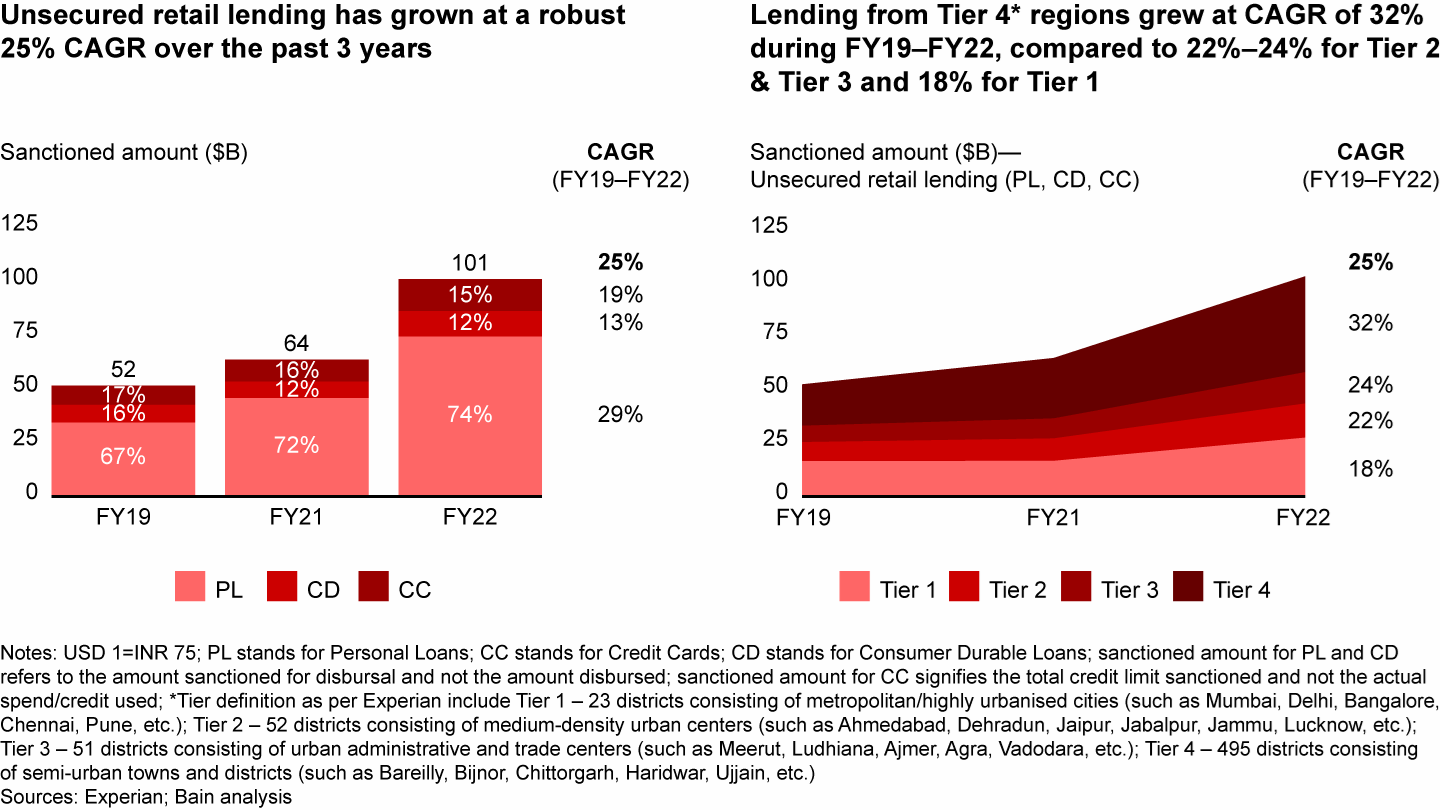
Meixuan Li, Chun Yu, Wei Liu (2021)

# 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.

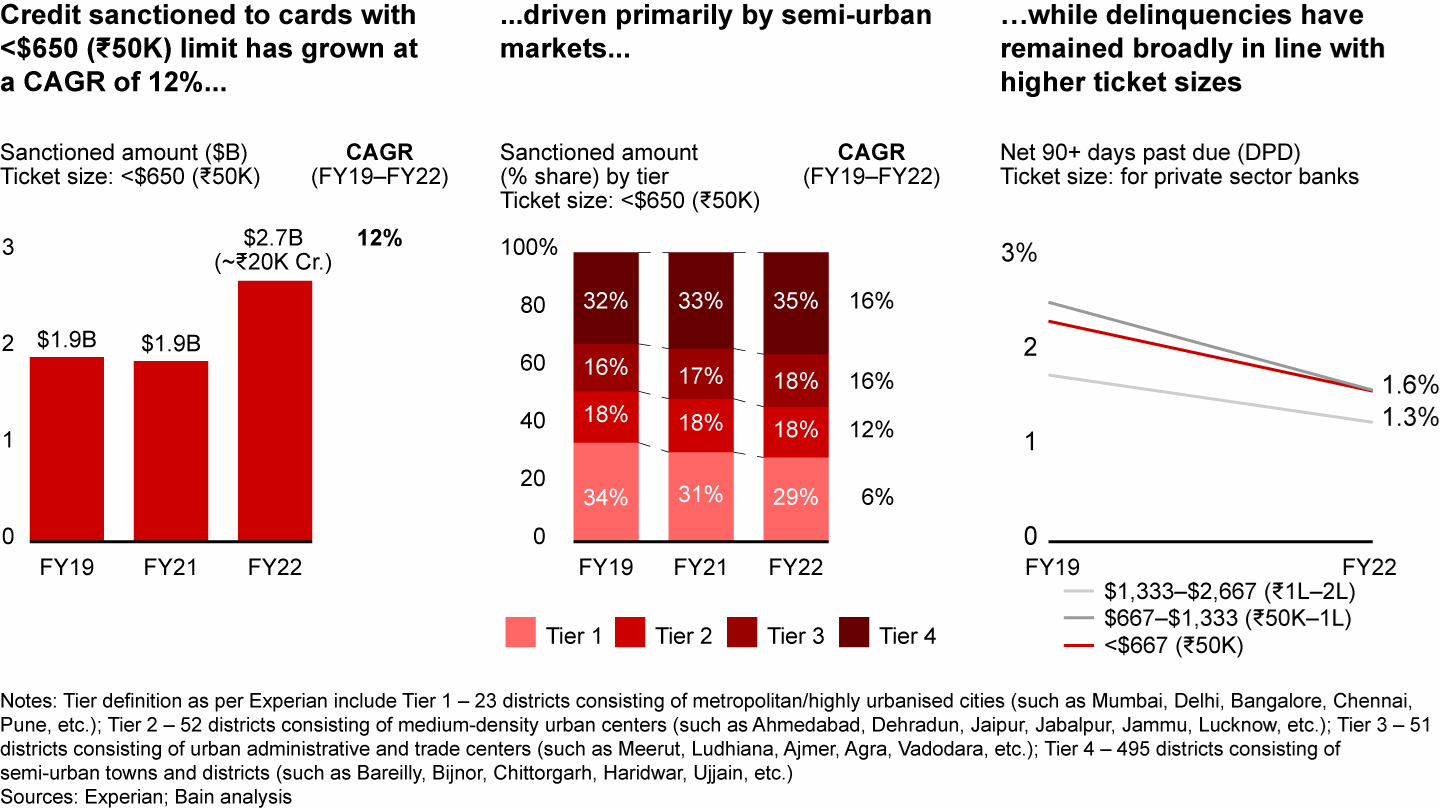
Despite the rapid growth of the online lending market in Bangalore, very little research has been conducted on credit risk assessment, specifically concerning the incorporation of various types of data such as demographic, financial, behavioural, and social media data. The lack of comprehensive studies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

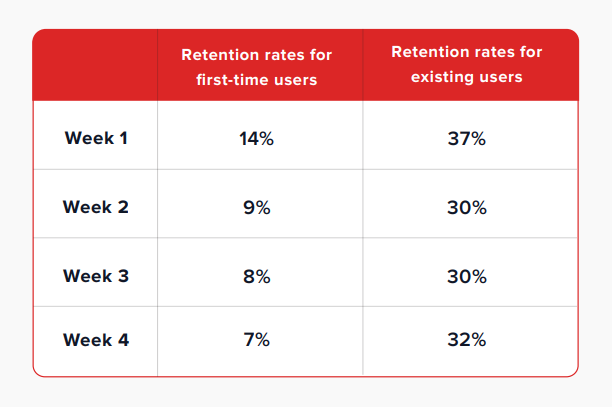
Bangalore, often referred to as the Silicon Valley of India, is home to a thriving fintech ecosystem. The customer churn rate in the Indian fintech sector averages a staggering 73%, a figure that has substantial implications for business sustainability and profitability. Furthermore, the credit default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



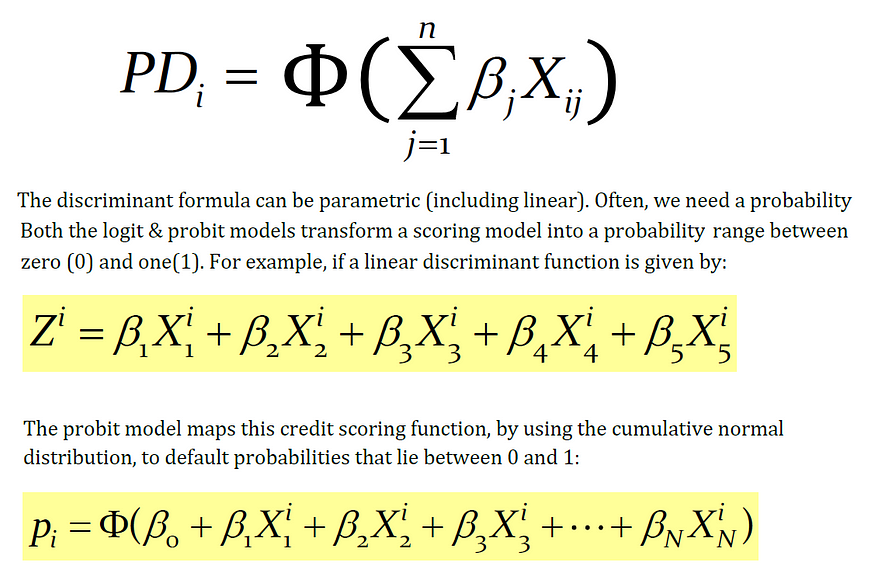
**Source : Clevertap** [**[7]**](https://brandequity.economictimes.indiatimes.com/files/cp/1139/cdoc-1658399014-Fintech_Retention_Guide.pdf)

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.

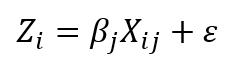
With a robust dataset in place, the construction of clear and testable research hypotheses ensues, encompassing both null and alternative stances. The investigative prowess is further amplified through the employment of Probit and Logit Regression methodologies. These techniques unfurl the intricacies 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% |

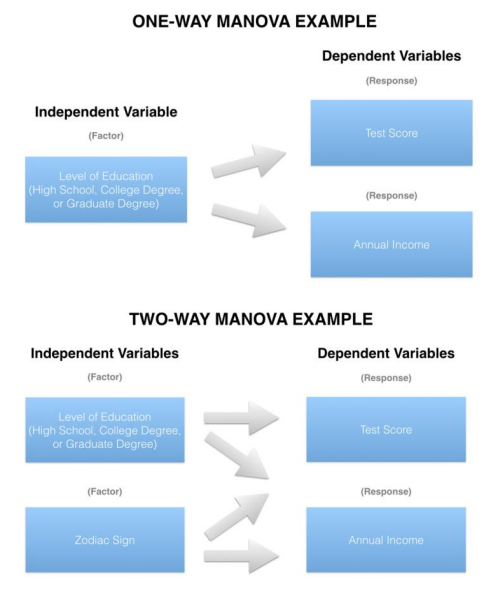
The second objective is to assess the creditworthiness of loan applicants in the instant loan app sectors employed through fintech platforms. The same questionnaire data will be used to build a comprehensive credit scoring model. Machine learning techniques will be applied to handle the data, including 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.

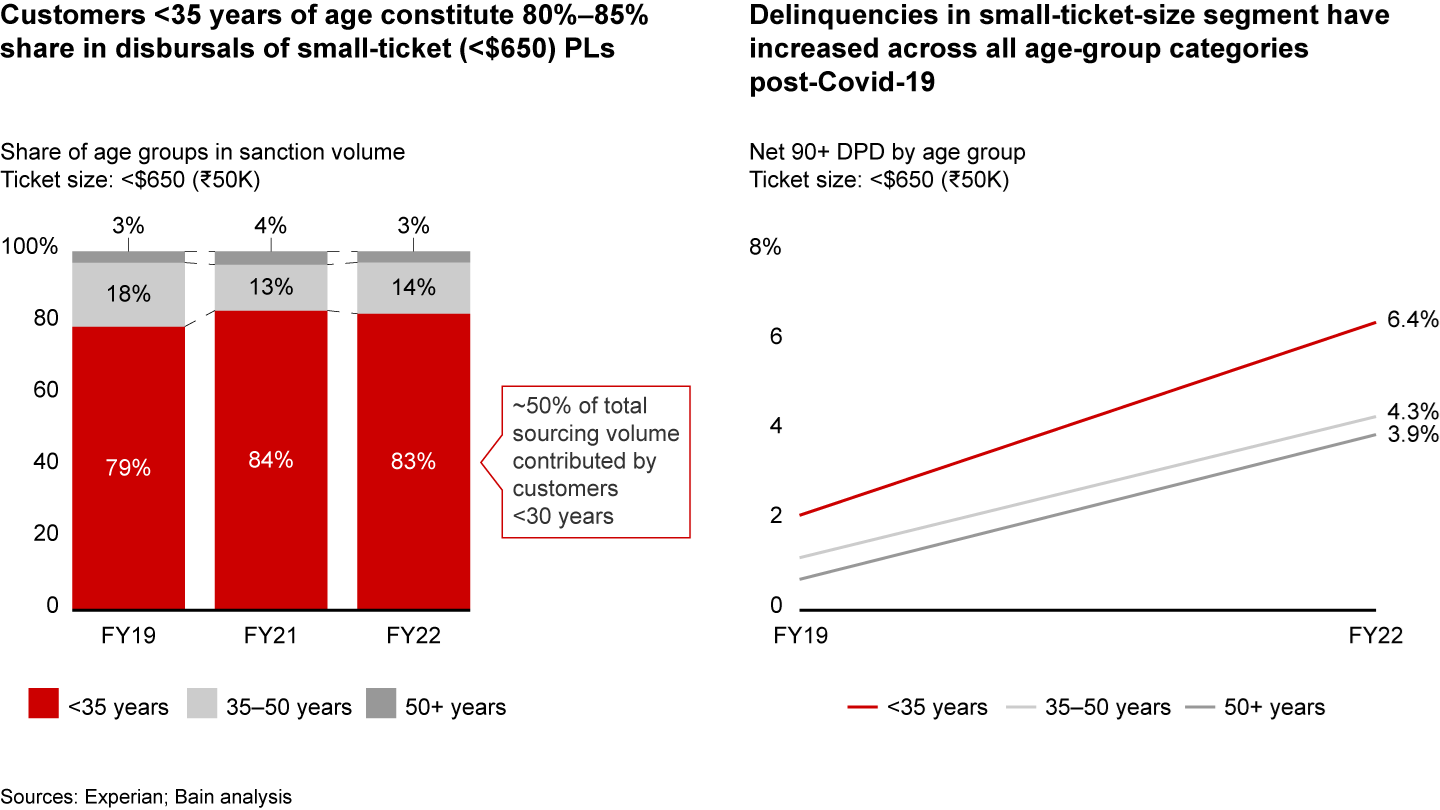


Source: Medium [[10]](https://medium.com/nerd-for-tech/everything-about-manova-and-mancova-4c1c237af464)

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.

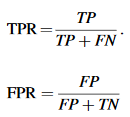
Evaluation indicators

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.



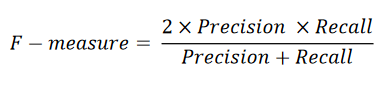
Accuracy is a widely employed metric to gauge the performance of a classification model. It quantifies how many samples the classifier has accurately predicted. To calculate accuracy, the model is employed to classify the test dataset, and the accuracy is determined by dividing the number of correctly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

**