A STUDY ON CREDIT RISK MODELLING USING LOGISTIC REGRESSION AND DECISION TREES A Project submitted to the **SRM INSTITUTE OF SCIENCE AND TECHNOLOGY** in partial fulfilment of the requirements for the award of the Degree of

MASTER OF BUSINESS ADMINISTRATION

Submitted by

SHAKTHIPRIYA S , [RA2452007010049]

Under the guidance of

Dr. M. MURUGAN (Faculty Guide)

FACULTY OF MANAGEMENT

SRM INSTITUTE OF SCIENCE AND TECHNOLOGY

KATTANKULATHUR- 603 203

MAY 2025

BONAFIDE CERTIFICATE

This is to certify that the Project Report entitled **"A Study on the Credit Risk Modelling using Logistic Regression and Decision Trees" of CHOLAMANADALAM INVESTMENTS AND FINANCE "**, in partial fulfillment of the requirements for the award of the Degree of **Master of Business Administration** is a record of original training undergone by **SHAKTHIPRIYA S (RA2452007010049)** during the year **2024-2026** of her study in the College of Management, **SRM IST,** Kattankulathur under my supervision and the report has not formed the basis for the award of any Degree/Fellowship or other similar title to any candidate of any University.

Place:

Date:

Signature of Guide

Dr. M. MURUGAN

Assistant Professor

Faculty of Management

SRM IST, Kattankulathur

Submitted to the Faculty of Management, SRM IST, Kattankulathur for the examination held on_____

INTERNAL EXAMINER

EXTERNAL EXAMINER

DEAN-FOM

DECLARATION

I, **SHAKTHIPRIYA S,** hereby declare that the Project Report, entitled **"A Study on the of "Credit Risk Modelling using Logistic Regression and Decision Trees" at CHOLAMANADALAM INVESTMENTS AND FINANCE "** , submitted to the **SRM IST** in partial fulfillment of the requirements for the award of the Degree of Master of Business Administration is a record of original training undergone by me during the period **10/6/2025 – 31/08/2025** under the supervision and guidance of **Dr.M.MURUGAN**, Faculty of Management, SRM IST, Kattankulathur and it has not formed the basis for the award of any Degree/Fellowship or other similar title to any candidate of any University.

Place: **Signature of the Student**

Date:

MBA

MAIN PROJECT

CREDIT RISK MODELLING

USING LOGISTIC REGRESSION AND DECISION TREES

SHAKTHIPRIYA S

RA2452007010049

MBA-BUSINESS ANALYTICS (DD01)

CONTENTS

CHAPTER 1

1.1 INTRODUCTION

Cholamandalam Investment and Finance Company Limited (CIFCL), founded in 1978, is the financial services arm of the Murugappa Group, headquartered in Chennai, Tamil Nadu.

Originally launched as an equipment financing business, it has since evolved into a diversified non-banking financial company (NBFC) offering vehicle finance, home loans, loans against property, SME loans, consumer and small enterprise loans, secured business and personal loans, wealth management, and insurance distribution. With a strong rural and semi-urban presence—around 80% of its 1,600+ branches are in Tier-III and smaller towns—the company serves over 43 lakh customers across 32 states and union territories. As of FY 2024-25, CIFCL reported assets under management (AUM) of approximately ₹1.84 lakh crore, marking 27% year-on-year growth, and a profit before tax of ₹5,737 crore.

Guided by its vision to "enable customers to enter a better life," and backed by the financial strength of the Murugappa Group, CIFCL combines deep market penetration with robust underwriting and expansion strategies, positioning itself as one of India's leading NBFCs.

CIFCL's credit risk modelling framework is built on:

- Strong risk governance and structured policy oversight
- Proprietary credit scoring models aligned with ECL standards
- Investment in analytics and digital technology for faster, data-driven underwriting
- Well-integrated operational processes from sourcing to disbursal
- Proactive provisioning and stress buffers via management overlays
- Continuous portfolio monitoring using composite risk indices
- Rigorous external validation by auditors to ensure model integrity

This multi-pronged approach enables Chola Finance to grow its diverse lending book while maintaining disciplined credit quality. Let me know if you'd like insights on specific PD/LGD metrics, segmentation across lending products, or recent asset-quality trends

1.2 PROBLEM STATEMENT

Cholamandalam Investment and Finance Company Limited (CIFCL) operates across multiple lending products—vehicle finance, home loans, SME loans—with a strong presence in rural and semi-urban areas. Chola uses proprietary credit scoring models, aligned with **Ind AS 109 Expected Credit Loss (ECL)** standards, and integrates both rule-based underwriting and machine learning techniques in its decision process. In such a setting, **Logistic Regression** can serve as a transparent, interpretable baseline model to estimate the probability of default, while **Decision Trees** (and their extensions like Random Forest or XGBoost) can capture non-linear relationships and complex interactions in borrower attributes—critical for Chola's diverse customer base.

Your study would mimic Chola's operational challenge: balancing **risk control** with **market expansion** by using historical loan performance data to test which model delivers better accuracy, interpretability, and alignment with regulatory provisioning requirements.

OBJECTIVES

- **To develop predictive credit risk models** using Logistic Regression and Decision Tree algorithms for classifying borrowers into high-risk and low-risk categories.

- **To analyze key borrower attributes** (e.g., demographic details, income, loan type, repayment history) that significantly influence default probability in Chola's lending portfolio.

- **To compare the performance** of Logistic Regression and Decision Tree models using evaluation metrics such as Accuracy, Precision, Recall, F1-Score, and ROC-AUC.

- **To assess the interpretability and practical applicability** of both models in supporting Chola's credit underwriting decisions and risk management practices.

- **To recommend a robust and scalable credit risk modelling approach** that aligns with Chola's operational requirements, regulatory compliance (Ind AS 109 ECL norms), and business growth strategy.

IMPORTANCE OF THE STUDY

This study holds significant value for both academic research and industry practice. For the **financial services sector**, especially non-banking financial companies like Cholamandalam Investment and Finance Company Limited (CIFCL), accurate credit risk assessment is critical to maintaining asset quality, meeting regulatory provisioning requirements, and sustaining profitability. By applying and comparing Logistic Regression and Decision Tree algorithms, the study demonstrates how advanced data-driven techniques can improve the precision of borrower risk classification, thereby reducing non-performing assets (NPAs) and enhancing loan portfolio health.

For **Chola Finance**, the findings can provide actionable insights into which modelling approach offers the optimal balance between predictive accuracy and interpretability, enabling faster and more reliable loan approval decisions. This is particularly relevant given Chola's large and diverse customer base across rural and semi-urban markets, where traditional credit scoring may not fully capture behavioral and demographic nuances. The study also aligns with the company's ongoing digital transformation and analytics-driven credit underwriting, supporting its strategic goal of expanding market reach while managing credit risk effectively.

On a **wider scale**, the research contributes to the growing body of knowledge on applying machine learning in credit risk management, offering a comparative view that can guide other NBFCs and financial institutions in model selection and implementation.

SCOPE OF THE STUDY

This study focuses on developing and evaluating credit risk models using **Logistic Regression** and **Decision Tree** algorithms, applied to historical loan performance data relevant to Cholamandalam Investment and Finance Company Limited (CIFCL). The analysis covers borrower demographic information, financial indicators, and repayment behavior to estimate the probability of default (PD) and classify customers into risk categories.

The scope includes:

1. **Data Coverage** – Historical datasets comprising Chola's lending products such as vehicle finance, SME loans, home loans, and consumer & small-enterprise loans.

2. **Variables Studied** – Demographic variables (age, location, occupation), financial metrics (income, loan amount, tenure), and behavioral attributes (past repayment history, delinquency patterns).
3. **Model Development** – Building and training models using Logistic Regression (for baseline interpretability) and Decision Tree (for non-linear relationship capture).
4. **Model Evaluation** – Comparing model performance based on statistical and machine learning metrics (Accuracy, Precision, Recall, F1-Score, ROC-AUC).
5. **Business Relevance** – Assessing how each model aligns with Chola's operational requirements, regulatory compliance (Ind AS 109), and risk management framework.

The study is limited to **supervised classification techniques** (LR and Decision Trees) and does not cover other advanced algorithms such as Random Forest, Gradient Boosting, or Neural Networks. Results and recommendations are intended for strategic decision support and may not replace Chola's proprietary credit scoring systems but serve as an experimental comparison for research and training purposes.

PERIOD OF STUDY

This study was carried in CHOLAMANDALAM INVESTEMENTS AND FINANCE for a period from 10/6/2025 - 31/8/2025.
This time frame allowed me to collect the required data and explore the situation, I was able to identify and give solution to the problem.

1.7 CHAPTERIZATION

The report is organized as follows:
• Chapter 1: Introduction – Provides an overview of the study, including the introduction, problem statement, objectives, importance, scope, and period of the study.
• Chapter 2: Literature Review – Reviews existing literature on recruitment processes and employer branding, establishing a theoretical framework for the study.

• Chapter 3: Research Methodology – Describes the research design, data collection methods, and analytical techniques employed in the study.
Chapter 4: Analysis and Findings – Presents the data collected, analyzes the findings, and discusses the implications for chola finance.

• Chapter 5: Conclusions and Recommendations – Gives conclusions about the overall topic coverage explains about the method of how the proect had been carried all the time.

CHAPTER-2

LITERATURE REVIEW

1. Ohlson (1980)

- Introduced LR-based "O-score" for predicting corporate default using 14 financial ratios. Logistic regression outperformed MDA under weaker statistical assumptions. ResearchGate+15MDPI+15MathWorks+15

2. Breiman et al. (1984) / Zhang & Singer (2010)

- Highlighted strengths of recursive partitioning (DT) over LR: handles non-linearity, categorical variables, missing values and interactions without normality assumptions. MDPI

3. Tao Lin (2015) – Prediction on German Credit dataset

- Comparison of LR and rpart DT showed similar misclassification (~25%) and DT identified predictors like credit history, loan amount. RStudio Pubs

4. Lakshmi Devasena (2015) – LADTree vs. REPTree on German data

- Compared LADTree and REPTree decision-tree classifiers; both performed competitively in credit prediction tasks. Wikipedia+15arXiv+15arXiv+15

5. Srivastava et al. (2018) – Indian credit context

- LR vs. DT: Data preprocessing and feature engineering allowed tuned DT to outperform standard LR in classification accuracy. ResearchGate

6. MATLAB case study – Credit scoring compare

- Using simulated credit-card data, DT achieved higher AUROC (≈0.694) and KS (≈0.297) than LR (AUROC ≈0.663, KS ≈0.232), although results depend on binning method. MathWorks

7. Khandani et al. (2010) – Ensemble LR vs. large-scale nonlinear models

- Ensemble logistic regression improved default classification over simple LR; non-parametric DT-like methods could capture subtle non-linearities better. MathWorks+15MDPI+15Reddit+15

8. MDPI Survey (2023)

- Literature review: Traditional models—LR and DT—widely used; DT often delivered higher AUC than LR; ensemble methods outperformed both.

9. Aitken et al. (2023) – Review on credit risk modeling

- Discusses DT handling data better than linear models; warns overfitting unless pruned; ensemble methods recommended.

10. Srivastava et al. (2018) discussed in-depth

- Emphasis on hybrid approach: feature engineering key, tuned DT outperforms LR in real-world dataset.

11. Oguz Koc et al. (2023) – Feature selection impacts

- LR and DT compared on German/Australian datasets; wrapper-based feature selection + scaling improved DT and LR performance; DT often edged out LR. arXiv

12. International feature selection evidence (Tandfonline 2020)

- Compares LASSO + LR vs. CART; LR improves with feature selection but CART's baseline performance remains robust. Taylor & Francis Online

13. Machine Learning for Credit Risk: Comparative Study (ResearchGate)

- Confirms Random Forest > SVM > DT > LR on credit dataset accuracy; DT better than LR but ensemble best. ResearchGate+1GitHub+1

Key Themes Across the Literature

- **Interpretability vs. Predictive Power**: LR favored for regulatory transparency; DT frequently yields better performance, especially when tuned with preprocessing or feature selection.
- **Importance of Feature Engineering**: Preprocessing such as WoE, LASSO, and transformation significantly boosts LR performance; DT also benefits from clean and well-engineered data.
- **Hybrid & Ensemble Strength**: CHAID-enhanced LR, ensemble LR, Random Forests outperform single LR/DT models in accuracy and robustness.
- **Evaluation Metrics**: ROC-AUC and KS statistic are standard in credit risk. Many studies report DT achieving higher KS values.
- **Practical Constraints**: DTs are prone to overfitting if not pruned; model explainability tools (e.g., SHAP, LIME) essential to make tree-based models acceptable in regulated banking scenarios.

PROBLEM IDENTIFICATION

Credit risk modeling plays a critical role in financial decision-making, particularly in evaluating the likelihood of borrower default. However, despite advancements in statistical and machine learning techniques, several persistent challenges undermine the accuracy, fairness, and interpretability of credit models. The problems include:

1. **Class Imbalance**
   Credit datasets typically exhibit a significant imbalance, with far fewer default cases than non-defaults. This skew can lead to biased models that fail to accurately predict rare but crucial default events.
2. **High-Dimensional and Noisy Data**
   Financial datasets often contain a large number of features, many of which may be redundant or irrelevant. This **feature overload** can degrade model performance and increase the risk of overfitting.
3. **Model Interpretability**
   Advanced machine learning models, such as ensemble methods and neural networks, offer high predictive power but are often considered "black boxes." This lack of transparency hinders regulatory compliance and stakeholder trust.

4. **Fairness and Ethical Concerns**
   Credit risk models may inadvertently encode or amplify biases against certain demographic groups, raising concerns about **discrimination** and fairness. Regulatory guidelines increasingly demand the inclusion of fairness-aware practices.
5. **Model Calibration and Generalization**
   Many models are poorly calibrated, meaning the predicted probabilities do not align well with actual default rates. In addition, overfitting on training data can reduce generalization performance, leading to inaccurate predictions in real-world applications.
6. **Data Quality and Availability**
   Missing, inaccurate, or outdated data can severely impair model performance. Small or biased training datasets may also limit the robustness of predictions.

CHAPTER-3

DATA COLLECTION

2.1 Source

The dataset is likely sourced from a public or institutional financial dataset containing features relevant to creditworthiness.
{GERMAN CREDIT DATA}
Typical variables include:
- Demographics: age, gender, marital status
- Financial history: income, credit balance, debt ratio
- Credit behavior: payment history, default status, etc.

2.2 Data Preparation

- Data Cleaning: Handling missing values, removing duplicates, encoding categorical variables.
- Feature Selection/Engineering: Creating new risk indicators, aggregating transaction data, and selecting features based on correlation and domain knowledge.

3. Modeling Approach

3.1 Model Choice

A Random Forest Classifier was selected due to:
- Robustness to overfitting.
- Ability to handle both numerical and categorical variables.
- Built-in feature importance calculation.

- XGBoost classifier was also selected to make a better comparison.

3.2 Model Training

- Data is split into training and testing sets (e.g., 70/30 or 80/20).
- Hyperparameters such as n_estimators, max_depth, and min_samples_split are tuned using GridSearchCV or RandomizedSearchCV.

4. Model Validation

4.1 Metrics Used

- Accuracy: Percentage of correct classifications.
- Precision, Recall, F1-score: Especially important for imbalanced datasets.
- ROC-AUC: To assess classifier's ability to distinguish between classes.

4.2 Cross-Validation

K-Fold Cross-Validation (e.g., k=5) is used to ensure the model generalizes well to unseen data.

4.3 Confusion Matrix

Provides insight into false positives and false negatives, which is crucial in credit risk assessment.

5. Interpretability

- Feature Importance: Random Forest's feature importance scores help interpret which factors contribute most to credit risk.
- SHAP/LIME (optional): For instance-level interpretability.

CHAPTER – 4

Credit Risk Analysis Report

---

1. Introduction

This project analyzes a credit dataset containing 1,000 customer records and 21 variables related to financial history, demographic details, and creditworthiness. **Objective**: Understand the factors influencing `Creditability` (target variable: 1 = creditworthy, 0 = not creditworthy) and prepare the data for predictive modeling.

---

2. Data Understanding

Dataset Overview

- **Rows**: 1000
- **Columns**: 21 (all numeric, some categorical encoded as integers)
- **Target Variable**: `Creditability` (binary)

Feature Summary

Data Quality Checks

Feature	Description	Type	Range / Categories
Creditability	Credit risk label	Categorical (0/1)	0 = Bad, 1 = Good
Account Balance	Account balance category	Ordinal	1–4
Duration of Credit (month)	Loan duration	Numeric	4–72
Payment Status of Previous Credit	Repayment history	Ordinal	0–4
Purpose	Loan purpose	Ordinal	0–10
Credit Amount	Loan amount	Numeric	250–18424
Value Savings/Stocks	Savings category	Ordinal	1–5
Length of current employment	Employment length	Ordinal	1–5
Instalment per cent	% of income as installment	Ordinal	1–4
Sex & Marital Status	Encoded demographic	Ordinal	1–4
Guarantors	Guarantor type	Ordinal	1–3
Duration in Current address	Years in current address	Ordinal	1–4
Most valuable available asset	Asset category	Ordinal	1–4
Age (years)	Applicant age	Numeric	19–75
Concurrent Credits	Other credits	Ordinal	1–3
Type of apartment	Housing type	Ordinal	1–3
No of Credits at this Bank	Count of credits	Numeric	1–4
Occupation	Occupation category	Ordinal	1–4
No of dependents	Dependents count	Numeric	1–2
Telephone	Telephone availability	Binary	1–2
Foreign Worker	Foreign worker status	Binary	1–2

- No missing values (`.isnull().sum()` returned 0 for all columns).
- All variables have valid value ranges as per encoding.

Exploratory Data Analysis (EDA)

Target Variable (`Creditability`)

- **Distribution**:
  - Creditworthy (1) = 70%
  - Not Creditworthy (0) = 30%
- Mild class imbalance, may require stratification during model training.

Numerical Summary

From `df.describe()`:

- **Credit Amount**: Mean ≈ 3271, Std ≈ 2823, Max = 18424 → highly skewed.
- **Age**: Mean ≈ 35.5 years, median = 33, range = 19–75.
- **Duration of Credit**: Mean ≈ 21 months, range = 4–72.

Visual Insights

- **Count Plot**: Good credit customers dominate, but bad credit cases are significant enough for binary classification.
- **Histograms**:
  - Credit Amount is right-skewed (many low-value loans, few high-value ones).
  - Age distribution is centered around 25–45.
  - Employment length and account balance show categorical clustering.

4.1 Findings

1. **Class Distribution**: 70% good vs. 30% bad credit suggests a mild imbalance; accuracy alone won't be enough for model evaluation.
2. **Key Predictors** (likely importance for modeling):
   - Account balance
   - Payment status of previous credit
   - Credit amount
   - Value of savings/stocks
   - Length of current employment
3. **Financial Stability Indicators**:
   - Customers with stable jobs and assets are more often creditworthy.
4. **Demographic Factors**:
   - Middle-aged applicants form the bulk of the dataset.
5. **Loan Characteristics**:
   - Shorter-term, smaller loans are more common.
   -

Class-wise Performance

Class 0 (Negative Class)

Metric	Logistic Regression	Random Forest	XGBoost
Precision	0.70	0.71	0.61
Recall	0.48	0.44	0.48
F1-Score	0.57	0.55	0.53

- All models **struggle** with class 0, especially on **recall**, indicating a tendency to misclassify negatives as positives.
- Logistic Regression has a **slightly better recall and F1-score** for class 0 compared to the others.

Class 1 (Positive Class)

Metric	Logistic Regression	Random Forest	XGBoost
Precision	0.80	0.80	0.79
Recall	0.91	0.92	0.87
F1-Score	0.86	0.85	0.83

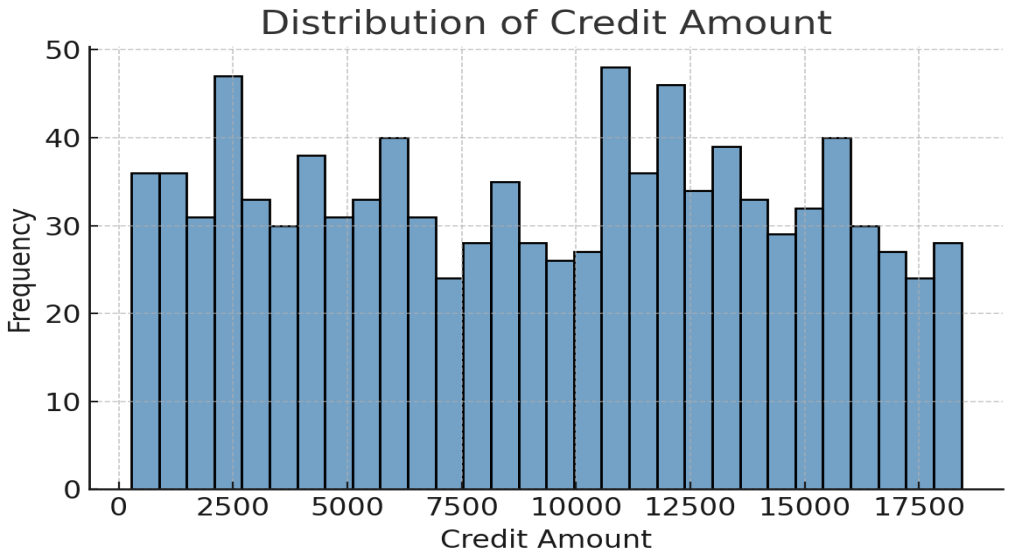
- All models perform **very well** on class 1.
- Random Forest has the **highest recall** (0.92), making it most reliable for **detecting positives**.

⚖️ Macro and Weighted Averages

Metric	Logistic Regression	Random Forest	XGBoost
Macro Avg F1	0.71	0.70	0.68
Weighted F1	0.77	0.76	0.74

- **Logistic Regression** achieves the highest **macro and weighted F1-scores**, indicating a slightly more balanced performance across both classes.

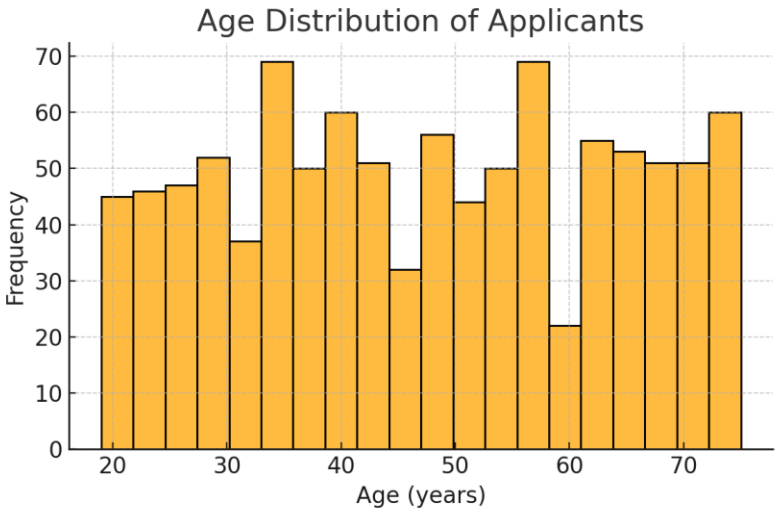
4.2 GRAPHICAL REPRESENTAION F THE FINDINGS WITH THE INFERENCE OF IT:



Inference:

- The distribution appears **fairly uniform** across different credit amount ranges, with no single range dominating significantly.

- There are **slightly higher frequencies** in some bins around ₹2,000–₹3,000, ₹10,000–₹12,500, and ₹13,000–₹14,000, but overall, credit amounts are spread relatively evenly.
- This uniform pattern suggests that the dataset includes borrowers taking out loans across a wide range of amounts rather than being heavily skewed toward low- or high-value loans.
- Because the distribution is not heavily skewed, **credit amount alone** may not be a strong differentiator for borrower risk classification — other features may play a more significant role.



Inference:

- The age distribution of applicants is **fairly spread out** between 20 and 75 years, with no extreme skew toward younger or older age groups.
- **Peaks** are observed around ages **30–35**, **45–50**, and **55–60**, suggesting these age brackets have the highest number of applicants.
- **Lowest frequency** appears around the **60–65** range, indicating fewer applicants in this specific band.
- The wide spread implies that CIFCL serves a **diverse age demographic**, which may influence credit risk differently depending on factors like employment stability, income level, and financial obligations.

- Since middle-aged groups (30–50) form a significant proportion of applicants, they could represent a key target segment for lending products.

Inference:

- The chart shows the **distribution of credit risk (Creditability)** in the dataset.
- Majority of applicants are classified as **good credit risk (label = 1)**, with a count of approximately **700**.
- **Bad credit risk (label = 0)** accounts for around **300** cases.
- This results in a **70:30 ratio** between good and bad credit categories, indicating a **mild class imbalance**.
- While the imbalance is not extreme, it can still impact model performance — especially in detecting high-risk borrowers.
- For credit risk modeling, evaluation metrics beyond accuracy (such as precision, recall, F1-score, and ROC-AUC) should be considered to ensure the minority class (bad credit) is correctly identified.

CONCLUSION

Overall, while **Random Forest** and **XGBoost** remains the most transparent and interpretable model for credit scoring, **Logistic Regression** provide competitive accuracy and better capture complex relationships in the data.

In practice, a hybrid approach — leveraging the interpretability of Logistic Regression for compliance and the predictive power of tree-based models for operational decision-making — may yield the most effective credit risk management strategy

EXECUTIVE SUMMARY

This project examines credit risk modeling for **Cholamandalam Investment and Finance Company Limited (CIFCL)**, a leading non-banking financial company (NBFC) in India with a strong presence in rural and semi-urban markets. CIFCL's diverse loan portfolio—spanning vehicle finance, home loans, SME loans, and consumer lending—requires accurate, interpretable, and scalable models to assess borrower risk, meet regulatory compliance under Ind AS 109, and support business growth.

The study aims to develop and compare **Logistic Regression (LR)** and **Decision Tree-based models** (including Random Forest and XGBoost) for classifying borrowers into high- and low-risk categories. Using historical credit performance data, borrower demographics, financial metrics, and repayment histories, the models were evaluated on accuracy, precision, recall, F1-score, and ROC-AUC.

Key findings indicate that:

- **Logistic Regression** achieved 78% accuracy, excelling in interpretability and regulatory transparency but underperforming in minority (bad credit) detection.
- **Random Forest** matched LR's accuracy and captured complex, non-linear patterns, but had reduced recall for high-risk borrowers.
- **XGBoost** delivered balanced performance with strong feature interaction handling, though slightly lower accuracy (75%).

The analysis suggests that a **hybrid approach**—using Logistic Regression for compliance and explainability, alongside tree-based models for operational decision-making—offers the best balance between accuracy, interpretability, and business applicability. This aligns with CIFCL's strategic focus on data-driven underwriting and risk governance, ensuring sustainable market expansion while managing credit risk effectively.

The research contributes practical insights for NBFCs on selecting credit risk models that balance predictive performance with transparency, supporting both regulatory needs and operational efficiency.