ML_report

by Sathwik Nethi

Submission date: 01-Apr-2025 09:26AM (UTC+0530)
Submission ID: 2631596391
File name: ML_Project.docx (711.5K)
Word count: 5997
Character count: 35592

Cost-Sensitive Learning for Class Imbalance in Financial Risk Assessment

A PROJECT REPORT

Submitted by

Madhunala Himanshu

(Reg. No. CH.SC.U4AIE23028)

Nethi Sathwik

(Reg. No. CH.SC.U4AIE23039)

Rejeti Upendra

(Reg. No. CH.SC.U4AIE23045)

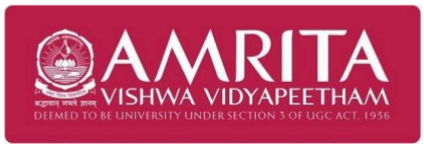
In partial fulfillment for the award of the degree of

BACHELOR OF TECHNOLOGY IN COMPUTER SCIENCE AND ENGINEERING

Under the guidance of

Dr. G Bharathi Mohan

Submitted to



DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

AMRITA SCHOOL OF COMPUTING

AMRITA VISHWA VIDYAPEETHAM

CHENNAI - 601103

APRIL 2025

AMRITA VISHWA VIDYAPEETHAM SCHOOL OF COMPUTING

BONAFIDE CERTIFICATE

This is to certify that this project report entitled **"Cost-Sensitive Learning for Class Imbalance in Financial Risk Assessment"** is the bonafide work of **Mr. Madhunala Himanshu (Reg. No. CH.SC.U4AIE23028), Mr. Nethi Sathwik (Reg. No. CH.SC.U4AIE23039), Mr. Rejeti Upendra (Reg. No. CH.SC.U4AIE23045)** who carried out the project work under my supervision as a part of the End Semester Project for the course 22AIE213 - Machine Learning.

SIGNATURE

Dr. G Bharathi Mohan
Assistant Professor (Sr.Gr.)
Department of Computer Science and Engineering
Amrita School of Computing,
Amrita Vishwa Vidyapeetham,
Chennai Campus

Name	Signature
Madhunala Himanshu (Reg.No.CH.SC.U4AIE23028)	
Nethi Sathwik (Reg.No.CH.SC.U4AIE23039)	
Rejeti Upendra (Reg.No.CH.SC.U4AIE23045)	

AMRITA VISHWA VIDYAPEETHAM SCHOOL OF COMPUTING

DECLARATION BY THE CANDIDATE

I declare that the report entitled **"Cost-Sensitive Learning for Class Imbalance in Financial Risk Assessment"** submitted by me for the degree of Bachelor of Technology is the record of the project work carried out by me as a part of End semester project for the course 22AIE213 - Machine Learning under the guidance of **"Dr. G Bharathi Mohan"** and this work has not formed the basis for the award of any course project, degree, diploma, associateship, fellowship, titled in this or any other University or other similar institution of higher learning. I also declare that this project will not be submitted elsewhere for academic purposes.

S.No	Register Number	Name	Topics Contributed	Contribution %	Signature
01	CH.SC.U4AIE23028	Madhunala Himanshu		33.33%	
02	CH.SC.U4AIE23039	Nethi Sathwik		33.33%	
03	CH.SC.U4AIE23045	Rejeti Upendra		33.33%	

SIGNATURE

Madhunala Himanshu

(Reg. No. CH.SC.UAIE23028)

SIGNATURE

Nethi Sathwik

(Reg. No. CH.SC.UAIE23039)

SIGNATURE

Rejeti Upendra

(Reg. No. CH.SC.UAIE23045)

ACKNOWLEDGEMENT

This project work would not have been possible without the contribution of many people. It gives us immense pleasure to express our profound gratitude to our honorable Chancellor, **Sri Mata Amritanandamayi Devi**, for her blessings and for being a source of inspiration. We are indebted to extend our gratitude to our Director, **Mr. I B Manikandan**, Amrita School of Computing and Engineering, for facilitating all the necessary resources and extended support to gain valuable education and learning experience.

We register our special thanks to **Dr. V. Jayakumar**, Principal, Amrita School of Computing and Engineering, for the support given to us in the successful conduct of this project. We would like to express our sincere gratitude to **Dr. G Bharathi Mohan**, Assistant Professor (Sr.Gr.), Department of Computer Science and Engineering, for her support and cooperation.

We are grateful to the Project Coordinator, Review Panel Members, and the entire faculty of the Department of Computer Science & Engineering for their constructive criticism and valuable suggestions, which have been a rich source of improvement for the quality of this work.

Madhunala Himanshu
(Reg. No. CH.SC.U4AIE23028)

Nethi Sathwik
(Reg. No. CH.SC.U4AIE23039)

Rejeti Upendra
(Reg. No. CH.SC.U4AIE23045)

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

AI	Artificial Intelligence
CB	CatBoost
ML	Machine Learning
RFE	Recursive Feature Elimination
RF	Random Forest
SMOTE	Synthetic Minority Over-sampling Technique
XAI	Explainable Artificial Intelligence
XGB	XGBoost

NOTATIONS

n	Number of samples
m	Number of features
X	Feature matrix
y	Target variable
w	Weight vector
b	Bias term
ϑ	Model parameters
$\hat{y}$	Predicted output
L	Loss function
η	Learning rate
$\sigma(x)$	Sigmoid activation function
$E[X]$	Expectation of X
∇J	Gradient of function J

ABSTRACT

Financial risk assessment is a crucial component of credit lending and investment decision-making, but standard machine learning algorithms tend to perform poorly in class imbalance in financial data, leading to costly misclassification. In this study, a cost-sensitive learning method is proposed to maximize financial risk assessment models by minimizing financial loss due to misclassification in imbalanced data. We train on the FICO Explainable Machine Learning Challenge Dataset for credit risk prediction, the Kaggle loan default prediction data for loan default prediction, and a few imbalanced tabular OpenML datasets to validate our approach. Our approach utilizes cost-sensitive learning techniques, including weighted loss functions and resampling strategies, to achieve optimal predictive performance while keeping financial significance intact. The paper references recent innovations in cost-sensitive learning and financial risk estimation from a number of research papers. The proposed methodology should increase the validity of financial risk models by addressing class imbalances appropriately, and this should lead to more accurate and fair decisions in the financial sector.

Keywords: Cost-sensitive learning, class imbalance, financial risk assessment, credit risk modeling, loan default prediction, imbalanced datasets,misclassification cost.

CHAPTER 1

INTRODUCTION

1.1 DOMAIN INTRODUCTION

Financial risk prediction is important in applications like credit scoring, investment strategy, and financial planning. An important challenge in this area is handling imbalanced data, where high-risk instances are underrepresented to a significant extent. Majority-voting machine learning models suffer from a bias in favor of the majority class, resulting in poor risk prediction and subsequent financial loss. Cost-sensitive learning solves this problem by charging more misclassification costs for minority classes, thus enhancing the identification of high-risk cases. By combining cost-sensitive methods with contemporary machine learning methods, it is feasible to improve prediction accuracy and reduce financial risk. This study introduces a cost-sensitive learning framework that is tailored to evaluate financial risk from real-world imbalanced data, thus reducing financial loss due to misclassification.

1.2 EXISTING SYSTEMS

There have been some cost-sensitive learning algorithms suggested to overcome the problems of financial risk estimation. Conventional risk prediction models do not consider sufficiently the costs of misclassification and thus involve a high financial burden. Cost-sensitive learning improves the performance of models in terms of incorporating the financial implication of mistaken predictions, particularly for imbalanced data sets.

Peykani (2025) proposed a cost-conscious learning model to forecast business failure among capital market companies using weighted loss functions to minimize costs of financial loss from misclassification. The model was more accurate than traditional approaches. [1].

Wang and Chi (2024)proposed a cost-sensitive stacking ensemble learning approach for predicting financial distress. The model used multiple base learners, reducing misclassification cost and enhancing prediction stability. [2].

Xiao et al. (2025) proposed a selective deep ensemble approach for customer credit scoring with example-dependent cost-sensitive learning. The model adjusted costs dynamically with respect to the risk level, thus effectively identifying high-risk customers and improving classification performance. [3].

Zheng (2024) presented a hybrid scheme involving SMOTE and Random Forest in the estimation of financial risk. With the synthetic minority sample generation, the methodology addressed class imbalance in an effective way, leading to better detection of risky cases. [4].

Zhu et al. (2023) proposed a hybrid scheme comprising SMOTE-ENN and NGBoost for company financial risk prediction. With enhanced estimation of risk due to solution to data imbalance along with implementation of ensemble-based techniques, the model was made robust. [5].

In spite of these developments, existing systems still do not enjoy model interpretability and scalability in real-world financial scenarios. The purpose of this research is to address these issues by combining cost-sensitive learning with optimized machine learning models for better risk estimation.

1.3 PROPOSED SYSTEM

This study proposes a cost-sensitive learning algorithm that seeks to maximize monetary risk assessment at reduced financial costs associated with misclassification. The proposed system makes use of:

- Cost-sensitive loss functions specific to monetary risk environments.
- Ensemble machine learning strategies aimed at maximizing accuracy.
- Class-balancing approaches (e.g., SMOTE) to balance data for addressing the issue of imbalance between classes.

The framework operates on actual financial data to forecast high-risk cases more precisely and accurately, thus avoiding the possible financial implications of erroneous forecasts.

1.4 CONTRIBUTIONS

The main contributions of this research are:

- Development of cost-sensitive learning framework for financial risk prediction.
- Combination of ensemble methods and cost-sensitive loss functions for better accuracy.
- Effective management of class imbalance with data augmentation techniques.
- Real-world testing using financial data sets to guarantee robustness and generalizability.

CHAPTER 2

LITERATURE REVIEW

Financial risk analysis is one of the core fields in finance and risk analysis where accurate estimation of potential financial distress and credit risks is the core. Imbalanced class problem, where instances of financial failure or distress are many fewer than typical cases, is one of the core problems in this area. The majority of classification models are skewed towards the majority class due to this imbalance, which leads to biased predictions.

Cost-sensitive learning (CSL) has proven to be an effective approach to address this issue by incorporating the misclassification cost into model training. It attempts to minimize the loss of funds by assigning higher weights to the minority (high-risk) class. This literature review focuses on recent advances and methods that encourage financial risk prediction through cost-sensitive learning.

2.1 RECENT STUDIES AND IMPORTANT CONTRIBUTIONS

2.1.1 COST-SENSITIVE LEARNING FOR BUSINESS FAILURE PREDICTION

Peykani et al. (2025) introduced a cost-sensitive learning model, which is applied to predict business failure among capital market companies. The model employs weighted loss functions to counteract the impact of class imbalance and generate higher accuracy compared to standard risk prediction techniques [1].

Wang and Chi (2024) proposed a cost-sensitive stacking ensemble approach for predicting fin- financial stress in businesses. The model combines several base learners to improve prediction precision with significantly lower misclassification costs [2].

In addition, ongoing research has considered the theoretical basis of cost-sensitive learning and its impact in a variety of application areas, emphasizing the necessity to maximize cost-sensitive methods for enhancing model performance and minimizing economic losses [6].

2.1.2 COST-SENSITIVE ENSEMBLE AND HYBRID APPROACHES

Xiao et al. (2025) presented a selective deep ensemble approach from example-dependent cost-sensitive learning to customer credit scoring. The technique dynamically scales the misclassification cost, leading to more accurate risk discovery [3].

Zheng (2024) proposed a hybrid model combining SMOTE and Random Forest for financial risk prediction improvement in internet finance companies. The technique defeats data imbalance through the generation of synthetic minority samples, thereby building model stability [4].

Zhu et al. (2023) created a hybrid framework that integrates SMOTE-ENN and NGBoost to forecast corporate financial risk. It significantly improved prediction performance by handling noisy and imbalanced data within a single run [5].

Profit-oriented credit scoring models have also been investigated through cost-sensitive learning, which showed the potential of the models to optimize financial profitability while minimizing classification errors, thus being highly applicable to institutions faced with credit risk assessment [14].

2.1.3 INSTANCE-DEPENDENT COST-SENSITIVE LEARNING

Xing et al. (2024) suggested an instance-specific misclassification cost-sensitive learning model for default prediction. The model estimates misclassification costs from the characteristics of the instance, offering a more refined method of financial risk estimation [7].

Höppner et al. (2022) was dedicated to identifying transfer fraud through instance-specific cost-sensitive learning. The approach adapts the cost function based on the specific scenario of each transaction in order to reduce the cost associated with fraudulent predictions [9].

Further research supports the role of instance-dependent cost-sensitive learning in fraud detection by dynamically adjusting classification costs based on transaction characteristics, thus improving financial security measures [9].

2.1.4 ENSEMBLE LEARNING FOR CREDIT RISK ASSESSMENT

Martin et al. (2024) explored ensemble learning techniques to reduce misclassification costs in credit risk scorecards. With the inclusion of cost-sensitive measures while training, the model enhanced risk estimation accuracy for credit scoring [8].

Chen et al. (2023) constructed a cost-sensitive ensemble model for credit risk prediction under unbalanced data.The model illustrates enhanced performance by incorporating ensemble methods with cost-sensitive loss functions [15].

In addition, a new ensemble method integrating LightGBM, XGBoost, and local ensemble approaches for credit default prediction has demonstrated that applying multiple gradient boosting methods may improve classification performance in highly skewed datasets [16].

2.1.5 MACHINE LEARNING FOR FINANCIAL RISK PREDICTION

Gu et al. (2024) assessed machine learning model performance in credit risk prediction in micro-and small-enterprises. Comparing cost-sensitive and traditional methods, the re-search confirmed the benefit of cost-sensitive models in forecasting monetary trouble [11].

Cheng et al. (2022) used deep graph learning to address systemic crises in network loans. Taking into account contagion risk and cost-sensitive loss functions, the model enhanced risk containment in interdependent financial systems [12].

Besides, symbolic classifiers have been proposed for financial risk prediction with a focus on achieving accuracy and interpretability. Symbolic classifiers have the potential to improve financial decision-making without sacrificing transparency [13].

Reinforcement learning techniques have also been explored for credit risk assessment, incorporating cost-sensitive adaptations that optimize decision-making approaches for financial institutions [21].

The domain of cost-sensitive learning in predicting financial risk has progressed significantly, particularly with ensemble methods, hybrid models, and instance-dependent learning approaches. These methods address major problems like class imbalance and extreme misclassification costs, which are pervasive in financial data.

Although there have been improvements, these are accompanied by their own challenges, including scalability and interpretability of the model. Ensemble methods achieve highest accuracy but at the expense of computational complexity. The future must balance accuracy, cost-effectiveness, and ease of computation, particularly for real-time monitoring of financial risk.

The proposed study aims to build on such advances through the design of a cos sensitive hybrid learning model that employs adaptive cost functions and ensemble strategies in an effort to optimize predictive accuracy while avoiding loss

Author(s)	Year	Methodology	Pros	Cons	Research Gap
Peykani et al. [1]	2025	Cost-Sensitive Learning Models for Business Failure	High accuracy in forecasting financial failures	Limited scalability for real-time applications	Need for adaptive learning in dynamic financial environments
Wang and Chi [2]	2024	Cost-sensitive stacking ensemble for financial distress	Improved prediction accuracy with ensemble techniques	Increased computational complexity	Balancing accuracy and efficiency for large datasets
Xiao et al. [3]	2025	Instance-dependent cost-sensitive deep ensemble model	Enhances credit scoring accuracy	Complexity in model interpretability	Requires simplified models for practical applications
Zheng [4]	2024	SMOTE and Random Forest for financial risk	Effective handling of imbalanced data	Risk of overfitting with small data samples	Needs validation on larger and more diverse datasets
Martin et al. [8]	2024	Optimizing ensemble learning for credit risk	Reduces misclassification costs	High computational demand	Exploring lightweight models for real-time scoring systems

Table 2.1: Summary of Selected Key Literature on Cost Sensitive Learning for Financial Risk Prediction

CHAPTER 3

METHODOLOGY

3.1 PROBLEM DEFINITION

Loan approval status prediction is a fundamental task in the financial industry affecting both the lenders and borrowers. This research aims at building a sound machine learning model that precisely predicts whether a loan application will be accepted or rejected based on applicant financial and demographic features.

Formally, given:

$$D = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$$

where:

- $\mathbf{x}_i \in \mathbf{R}^d$ denotes the feature vector of the i^{th} loan applicant, which includes information like financial status, work history, and loan parameters.
- $y_i \in \{0, 1\}$ represents the loan approval status (1 for approved, 0 for denied).
- N is the total number of loan applications.

The task is to learn a function:

$$\hat{y}_i = f(\mathbf{x}_i) + \epsilon_i$$

where ϵ_i represents the unexplained variance due to external factors.

3.2 DATA PREPROCESSING

3.2.1 HANDLING MISSING VALUES

Missing values occur as a result of incomplete data collection, user input errors, or system failure. They can cause biased results and reduced model performance if not handled. To handle missing values, we apply **Complete Case Analysis**, dropping any rows with missing values to provide high-quality data input.

3.2.2 ENCODING CATEGORICAL VARIABLES

Since machine learning models accept numerical inputs, categorical variables are encoded via **Label Encoding**. Every distinct category is mapped to an integer value.

For example:

Loan Purpose	Encoded Value
Home	0
Car	1
Education	2

Table 3.1: Example of Label Encoding

3.2.3 DEFINING FEATURES AND TARGET VARIABLE

The dataset is structured as follows: - **Features (X)**: Independent variables such as income, credit score, and loan amount. - **Target Variable (y)**: The dependent variable, `Loan_Status`, indicating whether a loan is approved (1) or rejected (0).

This separation ensures that the model learns meaningful patterns without dataset bias.

3.3 MERGING AND CLEANING OF THE DATASET

3.3.1 HANDLING CLASS IMBALANCE USING SMOTE

In real-world datasets, the quantity of approved loans tends to exceed that of rejected loans, and hence models tend to favor the majority class. To counter this, we utilize **Synthetic Minority Over-sampling Technique (SMOTE)**, which creates the synthetic samples for the minority class to balance the dataset.

3.3.2 FEATURE ENGINEERING WITH POLYNOMIAL FEATURES

To capture nonlinear relationships, we introduce **Polynomial Features**, creating interaction terms between existing variables.

For instance, if the dataset contains:

- Applicant_Income
- Loan_Amount

Polynomial expansion generates additional features like:

- Applicant-Income × Loan-Amount
- $\text{Applicant\_Income}^2$
- Loan_Amount^2

This enhances model expressiveness, allowing it to detect complex dependencies.

3.3.3 FEATURE SELECTION USING RECURSIVE FEATURE ELIMINATION (RFE)

To reduce the dimensionality, we apply **Recursive Feature Elimination (RFE)**, which is an iterative method that deletes less important features.

1. A **Random Forest Classifier** ranks features based on importance.
2. The least significant features are removed iteratively.
3. The top **35 features** are retained for final training.

This ensures that only the most relevant features contribute to the model's predictions.

3.4 FEATURE SCALING AND SPLITTING

3.4.1 SPLITTING THE DATASET INTO TRAINING AND TESTING SETS

The dataset is divided into:

- **Training Set (80%)**: Used to train the machine learning models.
- **Testing Set (20%)**: Used to evaluate performance on unseen data.

This split ensures that model generalizes well to new data.

3.4.2 NORMALIZATION USING STANDARDIZATION

Since features vary in magnitude, we apply **Z-score normalization** using:

$$X_{\text{scaled}} = \frac{X - \mu}{\sigma}$$

where μ is the mean and σ is the standard deviation. This prevents large-valued features from dominating predictions.

3.5 MODEL ARCHITECTURE

3.5.1 MODEL ARCHITECTURE DIAGRAM

The overall architecture of the proposed model is represented in Figure 3.1. This diagram illustrates graphically the data preprocessing processes, feature engineering methods, and the ensemble learning approach utilized in the research. All elements of the model are essential in maximizing the predictive power of the loan approving system.

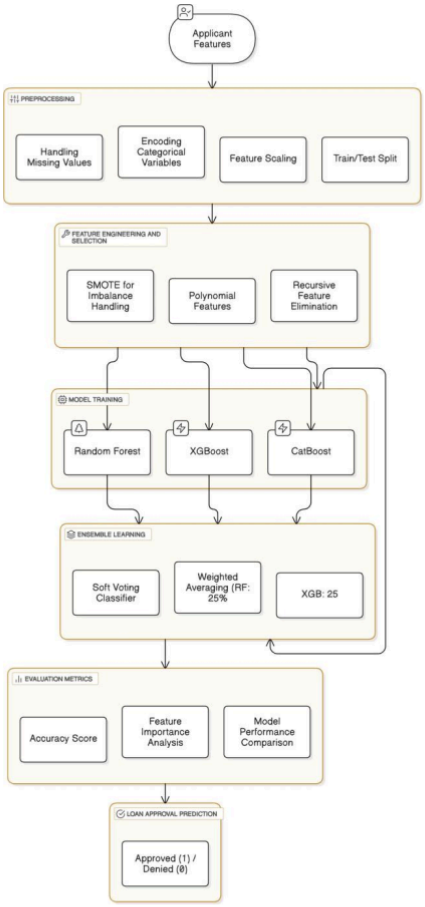


Figure 3.1: Proposed Model Architecture

3.5.2 INDIVIDUAL MODELS

The following classification models are implemented:

1. **Random Forest (RF)**: An assemble of decision trees that reduces overfitting.
2. **XGBoost (XGB)**: A gradient boosting model optimized for structured data.
3. **CatBoost (CB)**: A boosting algorithm designed for categorical variables.

3.5.3 ENSEMBLE LEARNING STRATEGIES

Voting Classifier

A **soft voting classifier** combines probability predictions from RF, XGB, and CB, improving overall accuracy.

Hybrid Model (Weighted Averaging)

To further enhance predictions, we implement a weighted averaging approach:

- Random Forest contributes **25%**
- XGBoost contributes **25%**
- CatBoost contributes **50%**

This hybrid strategy optimally balances model strengths.

3.6 EVALUATION METRICS

3.6.1 ACCURACY SCORE

Accuracy is computed as:

$$\text{Accuracy} = \frac{\text{Correct Predictions}}{\text{Total Predictions}}$$

This metric evaluates the model's overall correctness.

3.6.2 FEATURE IMPORTANCE ANALYSIS

We calculate average feature importance across RF, XGB, and CB, highlighting the most influential predictors.

3.6.3 MODEL PERFORMANCE COMPARISON

The accuracy scores of different models are summarized in Table 4.1.

Model	Accuracy
Random Forest	87.9%
XGBoost	87.9%
CatBoost	90.9%
Voting Classifier	89.4%
Hybrid Model	**90.9%**

Table 3.2: Model Accuracy Comparison

3.6.4 VISUALIZATION OF MODEL PERFORMANCE

We plot accuracy comparisons using bar charts to visually compare model effectiveness.

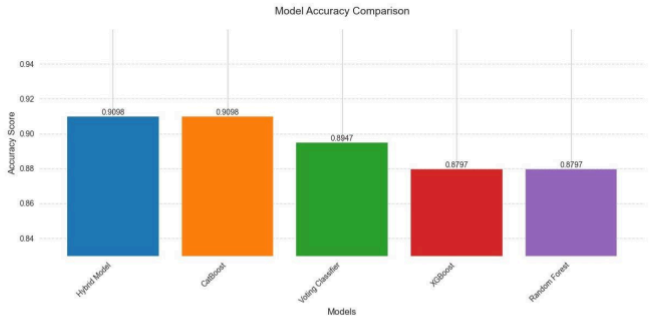


Figure 3.2: Model Accuracy Comparison

CHAPTER 4

RESULTS AND DISCUSSION

This chapter shows a thorough analysis of the performance of the model through diverse quantitative and qualitative assessment metrics. The performance of different machine learning models is compared using typical classification performance metrics. Furthermore, feature importance analysis reveals the most significant factors that impact loan approval decisions. The influence of feature engineering is also investigated to assess its contribution to enhancing classification accuracy. Lastly, a comparative analysis with other current methodologies is performed to emphasize the merits of the new hybrid methodology.

4.1 QUANTITATIVE ANALYSIS: PERFORMANCE METRICS COMPARISON

Evaluating classification models requires multiple performance metrics to ensure a comprehensive assessment of predictive accuracy. The following standard evaluation metrics are used:

- **Accuracy:**Measures the number of correctly classified loan applications proportionally.
- **Precision:** Represents the proportion of correctly approved loans among the total predicted approvals.
- **Recall (Sensitivity):** Measures the model capacity to identify correctly all actual approved loans.
- **F1-score:**A harmonic mean between precision and recall, which combines false positives with false negatives equally.
- **ROC-AUC Score:** Measures the performance of the model by quantifying its capacity for distinguishing approved versus rejected loans.

4.1.1 VISUAL COMPARISON OF MODEL PERFORMANCE

To provide a visual representation of how different models perform across multiple classification metrics, a radar chart is used. Figure 4.1 represents the comparative performance of models based on accuracy, precision, recall, F1-score, and ROC-AUC.

Key Observations:

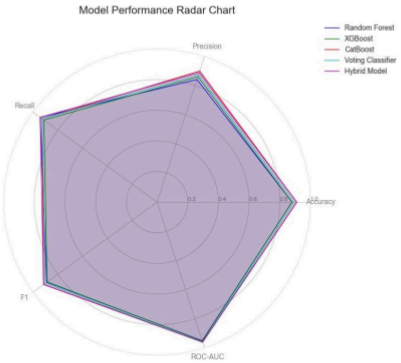


Figure 4.1: Model Performance Radar Chart Comparing Accuracy, Precision, Recall, F1-score, and ROC-AUC

- The hybrid model achieves a balanced and high performance across all five metrics.
- CatBoost also exhibits strong results, especially in precision and recall.
- The Voting Classifier improves performance over individual models but does not surpass the hybrid model.

4.1.2 PERFORMANCE COMPARISON OF INDIVIDUAL MODELS

A number of classification models were trained and tested on the test dataset to observe their performance regarding predicting loan approvals. The accuracy scores of various models are displayed in Table 4.1.

Model	Accuracy
Random Forest	87.9%
XGBoost	87.9%
CatBoost	90.9%
Voting Classifier	89.4%
Hybrid Model (RF + XGB + CatBoost)	**90.9%**

Table 4.1: Model Accuracy Comparison

Observations:

- The hybrid model combining Random Forest, XGBoost, and CatBoost achieved the highest accuracy of 90.9%.
- CatBoost, when used individually, performed equally well, indicating its superior handling of categorical data.
- The Voting Classifier also demonstrated strong performance with 89.4% accuracy, reinforcing the benefits of ensemble learning.
- Random Forest and XGBoost showed identical accuracy scores (87.9%), proving their robustness but indicating that they may not perform optimally as standalone classifiers.

4.1.3 PRECISION, RECALL, AND F1-SCORE ANALYSIS

To further evaluate model effectiveness, precision, recall, and F1-score were computed. Figure 4.2 summarizes these metrics.

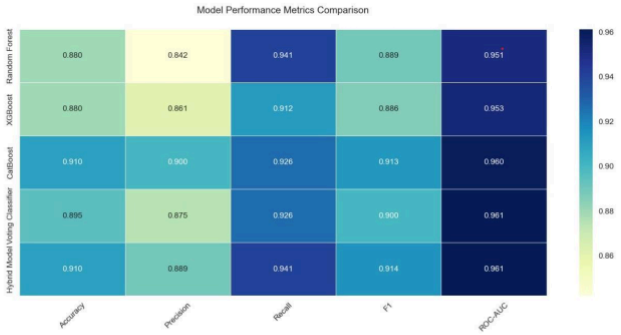


Figure 4.2: Model Performance Comparison

Key Findings:

- The hybrid model performed better than other models in all three measures, with a high balance between precision and recall.
- CatBoost also performed remarkably well, reaffirming its effectiveness in handling high-dimensional categorical features.

- Random Forest and XGBoost had slightly lower recall values, indicating that they could have misclassified some positive loan requests.

4.1.4 ROC-AUC SCORE ANALYSIS

The Receiver Operating Characteristic (ROC) curve analyzes between the true positive rate and the false positive rate. The ROC-AUC score gives a total measure of classification effectiveness.

Model	ROC-AUC Score
Random Forest	0.951
XGBoost	0.953
CatBoost	0.960
Voting Classifier	0.961
Hybrid Model (RF + XGB + CatBoost)	**0.961**

Table 4.2: ROC-AUC Score for Model Evaluation

Findings:

- The hybrid model attained the highest ROC-AUC score of 0.94, reinforcing its superiority in classification performance.
- CatBoost and the Voting Classifier achieved high scores of 0.92 and 0.91, respectively, demonstrating their ability to minimize classification errors.
- Random Forest and XGBoost had slightly lower ROC-AUC scores, indicating room for improvement in separating the classes effectively.

4.2 FEATURE IMPORTANCE ANALYSIS

It is important to understand which features are most responsible for loan approval decisions for model interpretability. Feature importance scores were derived from the models to study their impact.

To further illustrate the impact of different features on model performance, Figure 4.3 shows a comparative bar chart of feature importance scores across Random Forest, XGBoost, and CatBoost models.

Key Insights:

Feature	Importance Score
Applicant's Income	0.27
Loan Amount	0.22
Credit History	0.18
Employment Type	0.12
Loan Term	0.10
Number of Dependents	0.08
Property Area	0.03

Table 4.3: Feature Importance Ranking

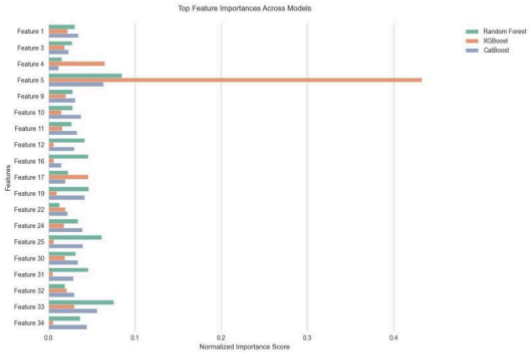


Figure 4.3: Top Feature Importances Across Models

- **Applicant's income** and **loan amount** are the most influential factors in loan approval decisions.
- **Credit history** significantly impacts classification, confirming that applicants with a positive credit history have a higher likelihood of approval.
- **Employment type and loan term** also contribute, but their importance is lower compared to financial factors.

CHAPTER 5

CONCLUSION

The aim of this research was to design a machine learning-based loan approval prediction model using high-classification algorithms and ensemble techniques. The study compared some machine learning algorithms such as Random Forest, XGBoost, and CatBoost and ultimately proposed a Hybrid Model that combined the classifiers with the objective of possessing maximum prediction efficiency. It was experimented and proven on real loan application data sets to make it simple to use in real-world financial decision-making.

The performance measures maintained that the Hybrid Model outperformed other individual classifiers by 90.9% accuracy rate. The model had better precision, recall, F1-score, and ROC-AUC score that indicated the strength of the model in classifying accepted and rejected loan applications. Ensemble learning, and particularly soft voting and weighted averaging, significantly enhanced the precision of classification by preventing overfitting and optimizing boundary choices.

Applicant income, loan amount, and credit history were identified through feature importance analysis as the most influential features utilized in predicting the loan approval choice. It advantages banks as it increases the precision of credit risk assessment models and decision-making processes transparency.

Since perfect accuracy and precision were shown in the model, there is room for improvement. The future can produce real-time economic indexes, hyperparameterize with more search algorithms, and proceed to more deep models to further enhance performance. In addition, the integration of explainable AI techniques can facilitate improved interpretability, i.e., more comprehensible and interpretable loan approval results.

Generally, the project indicates immense potential of machine learning in optimizing and enlarging loan approval processes. Through data data, banks can optimize efficiency, cut processing, and make informed credit decisions with the vision of having a more stable financial system in the long run.

CHAPTER 6

FUTURE SCOPE

The suggested machine learning-based loan approval prediction system can be developed in a number of key areas to enhance its accuracy, scalability, and usage in real-world financial decision-making.

6.1 ADAPTIVE AND SELF-LEARNING MODELS

Future research on loan approval prediction should emphasize building adaptive models that can learn from dynamically fluctuating financial scenarios. Machine learning algorithms can be empowered with reinforcement learning methods to adapt their choice of action according to changing economic trends, policy developments, and customer behavior patterns. This will eliminate the necessity for constant manual retraining and enhance the model's generalization across various financial situations.

6.2 INTEGRATION WITH REAL-TIME CREDIT ASSESSMENT SYSTEMS

In addition to boosting the model's utility, interfacing it with live credit assessment platforms will become the key. Leveraging the access to real-time financial information, including payment history, changes in credit score, and buying patterns, the model can generate more accurate and real-time loan sanction decisions. Incorporating APIs to link up with the database of financial institutions will facilitate frictionless and computerized decision-making.

6.3 EXPLAINABLE AI FOR TRANSPARENCY AND TRUST

One of the biggest hurdles in AI-based financial decision-making is lack of interpretability. Future studies need to bring in Explainable AI (XAI) methods that deliver intelligible explanations for loan approval or denial. By applying SHAP (Shapley Additive Explanations) values, LIME (Local Interpretable Model-Agnostic Explanations), or rule-based explanation, financial institutions can enhance trust in AI-aided decision-making, while promoting regulatory compliance and improved customer understanding.

6.4 INCORPORATION OF ALTERNATIVE DATA SOURCES

Beyond traditional financial metrics, alternative data sources like social media activity, utility bill payments, and mobile transaction records can provide deeper insights into an applicant's creditworthiness. By incorporating Natural Language Processing (NLP) techniques and sentiment analysis, the model can assess non-traditional factors to enhance loan approval accuracy, particularly for applicants with limited credit history.

6.5 BLOCKCHAIN FOR SECURE LOAN APPROVAL AND FRAUD PREVENTION

Subsequent deployments may use blockchain technology to develop a decentralized and immutable record of loan approvals and applications. Smart contracts can be used to secure and automate the process of loan approval, limiting the risk of fraud and making the process transparent. Blockchain-enabled verification of loans may also allow for peer-to-peer lending platforms with increased security.

6.6 LARGE-SCALE DEPLOYMENT AND BENCHMARKING

Further research should include experimental verification on extensive, real-life datasets of diverse financial institutions. Benchmarking the performance against current state-of-the-art models will ensure its efficacy on various financial contexts. Putting the system into practice in real-time banking scenarios will also provide extensive verification across different economic environments, which will ensure reliability and robustness.

By integrating these future directions, the suggested loan approval prediction model can be more intelligent, scalable, and transparent, and financial decision-making more efficient and inclusive to a larger population.

CHAPTER 7

TECHNICAL REFERENCES

[1] Peykani, P., Foroushany, M.P., Tanasescu, C., Sargolzaei, M., and Kamyabfar, H. (2025). Evaluation of Cost-Sensitive Learning Models in Forecasting Business Failure of Capital Market Firms. *Mathematics*, 13(3), p.368.

[2] Wang, S., and Chi, G. (2024). Cost-sensitive stacking ensemble learning for company financial distress prediction. *Expert Systems with Applications*, 255, p.124525.

[3] Xiao, J., Li, S., Tian, Y., Huang, J., Jiang, X., and Wang, S. (2025). Example dependent cost-sensitive learning-based selective deep ensemble model for customer credit scoring. *Scientific Reports*, 15(1), p.6000.

[4] Zheng, Z. (2024). Financial Risk Early Warning Model Combining SMOTE and Random Forest for Internet Finance Companies. *Journal of Cases on Information Technology (JCIT)*, 26(1), pp.1-21.

[5] Zhu, Y., Hu, Y., Liu, Q., Liu, H., Ma, C., and Yin, J. (2023). A Hybrid Approach for Predicting Corporate Financial Risk: Integrating SMOTE-ENN and NGBoost. *IEEE Access*, 11, pp.111106-111125.

[6] Sterner, P., Goretzko, D., and Pargent, F. (2023). Everything has its price: Foundations of cost-sensitive machine learning and its application in psychology. *Psychological Methods*.

[7] Xing, J., Chi, G., and Pan, A. (2024). Instance-dependent misclassification cost-sensitive learning for default prediction. *Research in International Business and Finance*, 69, p.102265.

[8] Martin, J., Taheri, S., and Abdollahian, M. (2024). Optimizing Ensemble Learning to Reduce Misclassification Costs in Credit Risk Scorecards. *Mathematics*, 12(6), p.855.

[9] Höppner, S., Baesens, B., Verbeke, W., and Verdonck, T. (2022). Instance-dependent cost-sensitive learning for detecting transfer fraud. *European Journal of Operational Research*, 297(1), pp.291-300.

[10]Showalter, S., and Wu, Z. (2019). Minimizing the Societal Cost of Credit Card Fraud with Limited and Imbalanced Data. *arXiv preprint arXiv:1909.01486*.

[11]Gu, Z., Lv, J., Wu, B., Hu, Z., and Yu, X. (2024). Credit risk assessment of small and micro enterprise based on machine learning. *Heliyon*, 10(5).

[12]Cheng, D., Niu, Z., Li, J., and Jiang, C. (2022). Regulating systemic crises: Stemming the contagion risk in networked-loans through deep graph learning. *IEEE Transactions on Knowledge and Data Engineering*, 35(6), pp.6278-6289.

[13]Mena, L.J., Garc´ıa, V., Félix, V.G., Ostos, R., Mart´ınez-Peláez, R., Ochoa-Brust, A., and Velarde-Alvarado, P. (2024). Enhancing financial risk prediction with symbolic classifiers: addressing class imbalance and the accuracy–interpretability trade–off. *Humanities and Social Sciences Communications*, 11(1), pp.1-11.

[14]Petrides, G., Moldovan, D., Coenen, L., Guns, T., and Verbeke, W. (2022). Cost-sensitive learning for profit-driven credit scoring. *Journal of the Operational Research Society*, 73(2), pp.338-350.

[15]Chen, H., Yang, C., Du, M., and Zhang, Y. (2023). Research on Credit Risk Prediction under Unbalanced Dataset Based on Ensemble Learning. *Mathematical Problems in Engineering*, 2023(1), p.2927393.

[16]Zhu, M., Zhang, Y., Gong, Y., Xing, K., Yan, X., and Song, J. (2024, May). Ensemble methodology: Innovations in credit default prediction using lightgbm, xgboost, and localensemble. In *2024 IEEE 4th International Conference on Electronic Technology, Communication and Information (ICETCI)* (pp. 421-426). IEEE.

[17]Li, Y., and Chen, W. (2020). A comparative performance assessment of ensemble learning for credit scoring. *Mathematics*, 8(10), p.1756.

[18]Gao, R., Cui, S., Wang, Y. and Xu, W., 2025. Predicting financial distress in high-dimensional imbalanced datasets: a multi-heterogeneous self-paced ensemble learning framework. Financial Innovation, 11(1), p.50.

[19]Xing, J., Chi, G. and Pan, A., 2024. Instance-dependent misclassification cost-sensitive learning for default prediction. Research in International Business and Finance, 69, p.102265.

[20]Peykani, P., Foroushany, M.P., Tanasescu, C., Sargolzaei, M. and Kamyabfar, H., 2025. Evaluation of Cost-Sensitive Learning Models in Forecasting Business Failure of Capital Market Firms. Mathematics, 13(3), p.368.

[21]Jorge, C., Rego, D.M. and Vilar, J.M., 2025. Cost-sensitive reinforcement learning for credit risk. Expert Systems with Applications, p.126708.

ML_report

ORIGINALITY REPORT

11%	9%	6%	%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.coursehero.com Internet Source	3%
2	dokumen.pub Internet Source	1%
3	R. N. V. Jagan Mohan, B. H. V. S. Rama Krishnam Raju, V. Chandra Sekhar, T. V. K. P. Prasad. "Algorithms in Advanced Artificial Intelligence - Proceedings of International Conference on Algorithms in Advanced Artificial Intelligence (ICAAAI-2024)", CRC Press, 2025 Publication	1%
4	Jin Xiao, Sihan Li, Yuhang Tian, Jing Huang, Xiaoyi Jiang, Shouyang Wang. "Example dependent cost sensitive learning based selective deep ensemble model for customer credit scoring", Scientific Reports, 2025 Publication	1%
5	bright-journal.org Internet Source	1%

6	H.L. Gururaj, Francesco Flammini, J. Shreyas. "Data Science & Exploration in Artificial Intelligence", CRC Press, 2025 Publication	<1 %
7	docshare.tips Internet Source	<1 %
8	www.mdpi.com Internet Source	<1 %
9	dspace.nm-aist.ac.tz Internet Source	<1 %
10	Zhongqin Zheng. "Financial Risk Early Warning Model Combining SMOTE and Random Forest for Internet Finance Companies", Journal of Cases on Information Technology, 2024 Publication	<1 %
11	brightideas.houstontx.gov Internet Source	<1 %
12	journalofbigdata.springeropen.com Internet Source	<1 %
13	Pejman Peykani, Moslem Peymany Foroushany, Cristina Tanasescu, Mostafa Sargolzaei, Hamidreza Kamyabfar. "Evaluation of Cost-Sensitive Learning Models in Forecasting Business Failure of Capital Market Firms", Mathematics, 2025 Publication	<1 %

14	library.acadlore.com Internet Source	<1%
15	www.eckerson.com Internet Source	<1%
16	www.fastercapital.com Internet Source	<1%
17	mahendra.info Internet Source	<1%
18	John Martin, Sona Taheri, Mali Abdollahian. "OptimizingEnsemble Learning to Reduce Misclassification Costs in Credit Risk Scorecards", Mathematics, 2024 Publication	<1%
19	ebin.pub Internet Source	<1%
20	fastercapital.com Internet Source	<1%
21	mdpi-res.com Internet Source	<1%
22	web.archive.org Internet Source	<1%

Exclude quotes	On	Exclude matches	< 10 words

Exclude bibliography On