National College of Ireland

AI in Loan Prediction

MSc Research Project
MSc in Finance Technology

Huy Hoang Nguyen
Student ID: 23429461

School of Computing
National College of Ireland

Supervisor: Faithful Onwuegbuche

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Huy Hoang Nguyen		
Student ID:	23429461		
Programme:	Msc in FinTech	**Year:**	2024-2025
Module:	MSc (Research) Practicum/Internship Part 2		
Supervisor:	Faithful Onwuegbuche		
Submission Due Date:	11/08/2025		
Project Title:	AI in Loan prediction		
Word Count:	5929	**Page Count: 21**	

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Huy Hoang Nguyen

Date: 11/08/2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

AI in Loan Prediction

Huy Hoang Nguyen
23429461

Abstract

Low loan approval rates and inefficient lending procedures have hindered access to essential financing for housing, education, small business expansion, and healthcare—factors that can significantly improve quality of life. This study investigates the effectiveness of machine learning (ML) models in predicting loan amounts based on individual financial history, addressing the need for a simple, accessible evaluation tool for borrowers with limited financial literacy. Unlike previous studies that focused on classification models for loan approval or default prediction, this research innovates by using a regression-based ML approach to estimate the loan amount. Five machine learning models were trained, with the XGBoost model achieving the best performance, yielding an R-squared score of 0.82 and an RMSE of 2,417.89 on the test data. These results demonstrate robust accuracy and the potential of ML applications in finance. In practice, the model provides a scalable tool for lending institutions to streamline loan evaluation and for borrowers to better understand their borrowing potential, thereby enhancing financial inclusion. However, challenges remain in addressing AI transparency and mitigating bias in training data, which can impact the fairness and reliability of model outcomes. Future research should focus on improving model interpretability and ensuring unbiased predictions to fully realize the potential of machine learning in transforming lending practices.

1 Introduction

The global lending market, valued at $10.4 trillion in 2023 and projected to reach $14.9 trillion by 2028, is foundational for economic development (Business Research Company, 2024), as loans are a critical tool for individuals to access financial stability for housing, vehicles, and education. However, low loan approval rates, which stood at just 14.5% at large U.S. banks in 2023, have limited financial inclusion, a problem exacerbated by tighter credit standards (IBS Intelligence, 2023; European Central Bank, 2023). While previous AI-based lending studies have focused on classification models to predict loan approval or default risk, they have largely neglected predicting the loan amount itself, a crucial aspect for both borrowers and lenders. This study addresses that gap by developing ML regression models to predict loan amounts based on a borrower's financial profile, with the goal of improving lending efficiency and accessibility.

The research question is: *How effective can ML models be in predicting loan amounts to improve loan approval rates and lending efficiency?*

The objectives were to train and compare five ML models, aiming for high predictive accuracy. The best-performing model, XGBoost, achieved an R-squared of 0.82 and an

RMSE of 2,417.89, demonstrating its potential for a scalable tool that benefits both lenders and borrowers. A key limitation is that the dataset was originally designed for loan approval, not amount prediction, and does not always contain the maximum possible loan amount, which may affect the model's ability to capture a borrower's full potential. The report's structure details the methodology, presents results, discusses findings in the context of existing research, and concludes with future directions for improving AI transparency and mitigating bias.

2 Related Work

Sravani and Mahaveerakannan (2023) explored loan status prediction by comparing Random Forest (RF) and Support Vector Machine (SVM) algorithms, using a dataset with financial and demographic features like income and credit history. The methodology involved training both models and applying an independent t-test to confirm RF's superior accuracy (85.30%) over SVM (75.10%), with a statistically significant difference ($p = 0.001$). The paper highlighted RF's robustness, demonstrated by minimal standard deviation, and provided a practical framework for credit scoring by analyzing feature influence. Strengths of this study include its rigorous statistical validation and clear methodology, offering a reproducible approach for financial institutions. However, its publication in the Journal of Survey in Fisheries Sciences, a non-relevant journal, raises concerns about the rigor and credibility of the peer-review process. The study was also limited in scope, focusing on only two algorithms and overlooking more advanced methods like XGBoost or Neural Networks. The paper also failed to provide detailed information on the dataset's origin, which hinders reproducibility and verification. This work contributes to understanding RF's effectiveness in binary loan prediction but suggests that future research should broaden algorithmic comparisons and enhance dataset transparency to improve generalizability in credit risk management (Sravani and Mahaveerakannan, 2023).

Manoj and Geetha (2024) developed a machine learning system to predict loan amounts, leveraging customer data such as income, spending, and credit scores to enhance lending efficiency. The study used regression models, Linear Regression, Decision Tree, Random Forest, and Bagging Regressor, on a pre-processed CSV dataset, employing the R-squared score to evaluate performance. Random Forest outperformed the other models, demonstrating strong predictive capabilities and scalability due to its ability to capture complex data relationships. The study's strengths lie in its comprehensive comparison of regression models and its practical implications for loan amount estimation, aligning with industry needs for automated decision-making. However, relying solely on R-squared as the only metric limited the depth of evaluation, as metrics like RMSE or MAE could have better captured prediction errors. Additionally, the black box nature of Random Forest limited interpretability, a crucial factor for financial applications (Manoj, K. M., & Geetha, M.,2024).

Bhatnagar, Chow, and Lai studied Lending Club peer-to-peer lending data from 2007–2015 to predict loan approval and interest rates, aiming to explore trends in lending criteria. They applied supervised learning (Random Forests, Support Vector Machines, Neural Networks)

and linear regression with PCA, achieving a 98% F-measure for approval prediction and an RMSE of <0.03 for interest rates. The study revealed relaxed approval standards (higher debt-to-income ratios) and identified stable predictive factors such as Loan Grade. Strengths of the study include its high predictive performance and insightful temporal analysis, providing a robust framework for understanding lending dynamics. This research makes a significant contribution to loan prediction by demonstrating the effectiveness of ensemble methods but suggests that future research should validate models on diverse platforms, incorporate proprietary data, and address implementation challenges to ensure real-world applicability (Pujun, Nick, & Max, 2016).

Odegua (2020), Lin (2024), Addo et al. (2018), Lai (2020), and Liang et al. (2019) have published several papers with a shared approach to enhancing loan risk management by applying machine learning to predict loan defaults, utilizing datasets like Kaggle and Lending Club. Odegua used XGBoost with 5-fold cross-validation, achieving 79% accuracy and identifying demographic features like location as important. Lin compared Logistic Regression, Random Forest, XGBoost, and AdaBoost, with XGBoost yielding 93.26% accuracy. Addo et al. found that tree-based models (Random Forest, Gradient Boosting) outperformed deep learning, highlighting the role of feature selection. Lai's Logistic Regression achieved 81.1% accuracy, focusing on comprehensive evaluation, while Liang et al. used multiple models, with LightGBM excelling on a large dataset. Noteworthy strengths include high accuracy and analysis of feature importance, showing how features interact and influence the target. However, the inconsistencies in algorithmic performance (Logistic Regression vs. XGBoost) stem from differences in dataset size and feature complexity. These studies highlight the superiority of ensemble methods but also underscore gaps in explainability and real-world implementation, suggesting that future research should prioritize interpretable models and robust validation on diverse datasets.

Existing research has demonstrated the strengths of ML models in predicting complex loan outcomes by analyzing surrounding factors. These studies also highlight the importance of factors like income and credit scoring in the lending process. However, weaknesses such as limited dataset transparency and poor model interpretability reduce the widespread adoption of these models. The primary focus on binary outcomes (default or approval prediction) has optimized bank efficiency but has not addressed high loan rejection rates, overlooking borrower-centric solutions. This leaves a gap that needs to be addressed. This paper addresses these gaps by developing regression models to **predict the loan amount** based on credit history, with the aim of reducing rejection rates and supporting borrowers. Predicting the loan amount not only assists lenders in evaluating customers to optimize the lending process but also helps borrowers assess their own credit profile and keep a budget in mind before applying for a loan

3 Research Methodology

The research methodology includes data collection, exploratory data analysis (EDA), data preprocessing, model training, and evaluation, ensuring feasibility and transparency for validation by other researchers

3.1 Data collection

The dataset was sourced from a public Kaggle dataset, consisting of 45,000 loan records, of which 10,000 were approved (loan_status=1). This dataset has 14 features: 9 numerical columns (person_age, person_income, person_emp_exp, loan_amnt, loan_int_rate, loan_percent_income, cb_person_cred_hist_length, credit_score, loan_status) and 5 categorical columns (person_gender, person_education, person_home_ownership, loan_intent, previous_loan_defaults_on_file). The target variable is loan_amnt, a continuous variable representing the approved loan amount. This dataset, originally designed for loan approval prediction, captures relevant loan, demographic, and financial information, which could be a potential limitation for predicting the loan amount.

Data source: https://www.kaggle.com/datasets/taweilo/loan-approval-classification-data

3.2 Materials and Tools

The analysis was performed using Python (version 3.10) on Google Colab with a T4 GPU, which accelerated computational speed for model training. The following libraries were used: pandas and numpy for data manipulation; matplotlib.pyplot and seaborn for visualization; IPython.display for interactive outputs; category_encoders for encoding categorical variables; sklearn.preprocessing (RobustScaler, PolynomialFeatures) for feature scaling; sklearn.model_selection (train_test_split, cross_val_score) for data splitting and validation; sklearn.metrics (mean_squared_error, mean_absolute_error, r2_score, median_absolute_error) for evaluation; sklearn.linear_model (LinearRegression), sklearn.ensemble (RandomForestRegressor), sklearn.tree (DecisionTreeRegressor), sklearn.svm (SVR), and xgboost (XGBRegressor) for model implementation; sklearn.neighbors (KNeighborsRegressor) for missing value imputation; sklearn.utils (resample) for data handling; and time for measuring training duration. The warnings library was used to suppress non-critical warnings. Jupyter Notebook in Google Colab supported interactive development and visualization.

3.3 Data Preprocessing

Sample Preparation

The dataset was split into a training set (80%, 36,000 records) and a testing set (20%, 9,000 records) using train_test_split with stratified random sampling to maintain the distribution of the loan_amnt variable. Missing values were imputed using the median for numerical features and the mode for categorical features via pandas. Unapproved loan amounts (35,000 records, loan_status=0) were estimated using a K-Nearest Neighbors (KNN) regression model (KNeighborsRegressor, k=5) trained on the approved loan data to ensure all records had a valid loan_amnt value.

Outlier Handling

Outliers in numerical columns (person_age, person_income) were identified using the interquartile range (IQR) method via pandas and capped at the 1st and 99th percentiles to reduce skewness while preserving data integrity.

Feature Encoding

Categorical columns (person_gender, person_education, person_home_ownership, loan_intent, previous_loan_defaults_on_file) were encoded using category_encoders. LabelEncoder to convert them into numerical representations. Label encoding was chosen to minimize dimensionality, given the moderate number of categories.

Feature Scaling

Numerical features were scaled using RobustScaler from sklearn.preprocessing to mitigate the impact of outliers, transforming features by subtracting the median and scaling to the IQR. This ensured model stability against skewed financial data.

3.4 Model Development

Five machine learning regression models were trained: Linear Regression (LinearRegression), Decision Tree (DecisionTreeRegressor), Random Forest (RandomForestRegressor), XGBoost (XGBRegressor), and Support Vector Regression (SVR). Hyperparameter tuning was performed using grid search with 5-fold cross-validation via sklearn.model_selection.

3.5 Metrics and Calculations

The models predicted a continuous loan amount (loan_amnt). Performance was evaluated using the following metrics, implemented through sklearn.metrics:

- Training R^2 score
- Test R^2 score
- Mean Squared Error (MSE)
- Root Mean Squared Error (RMSE)
- Mean Absolute Error (MAE)
- Median Absolute Error (MedAE)
- Cross-validation R^2
- Training time

3.6 Statistical Techniques

Model performance was compared using the mean values of the Training R^2, Test R^2, RMSE, MSE, MAE, and MedAE scores across the training and testing sets. Cross-validation R^2, calculated using cross_val_score, ensured the model's robustness against overfitting. Paired t-tests (α=0.05), performed using scipy.stats, assessed the statistical significance of performance differences between XGBoost and other models. Feature importance was analyzed using XGBoost's gain-based scores to identify key predictors

4 Design Specification

This section outlines the techniques, architecture, and frameworks that underpin the final implementation of the loan amount prediction system. It focuses on the models developed, data transformations, feature selection, and tools used. The system predicts continuous loan amounts based on a borrower's financial, demographic, and loan-related information, aiming to enhance the efficiency and accessibility of loans. A detailed analysis of each technique provides clarity on its function and contribution to the implementation. The implementation uses a supervised machine learning regression framework that integrates a preprocessing

pipeline, feature selection, and five regression models: Linear Regression, Decision Tree, Random Forest, XGBoost, and Support Vector Regression (SVR).

4.1 Exploratory Data Analysis (EDA)

EDA examines the dataset to explore its features and derive insights for training a machine learning regression model to predict loan amounts. The analysis includes eight charts and one heatmap, organized to systematically explore loan status, loan amount, age, loan intent, gender, previous loan defaults, and correlations.

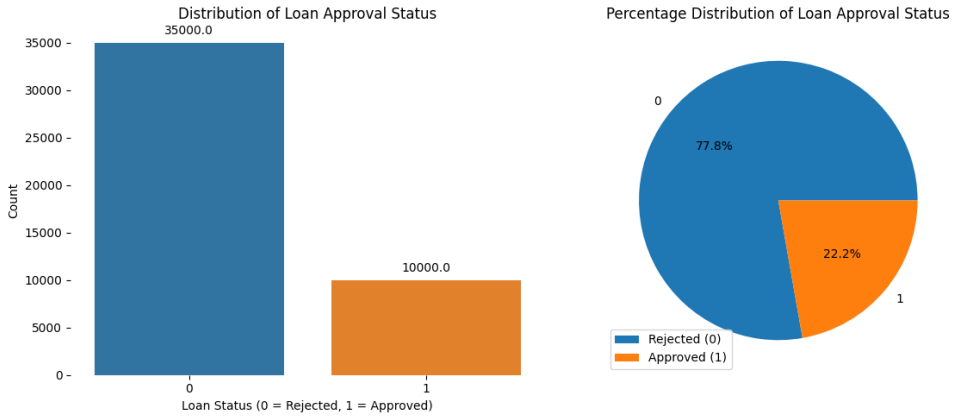

Figure 1: Distribution of loan status

The bar chart shows the count of loan statuses, with approximately 35,000 rejected loans and 10,000 approved loans. The pie chart complements this by indicating that 77.8% of loans are rejected, while 22.2% are approved.

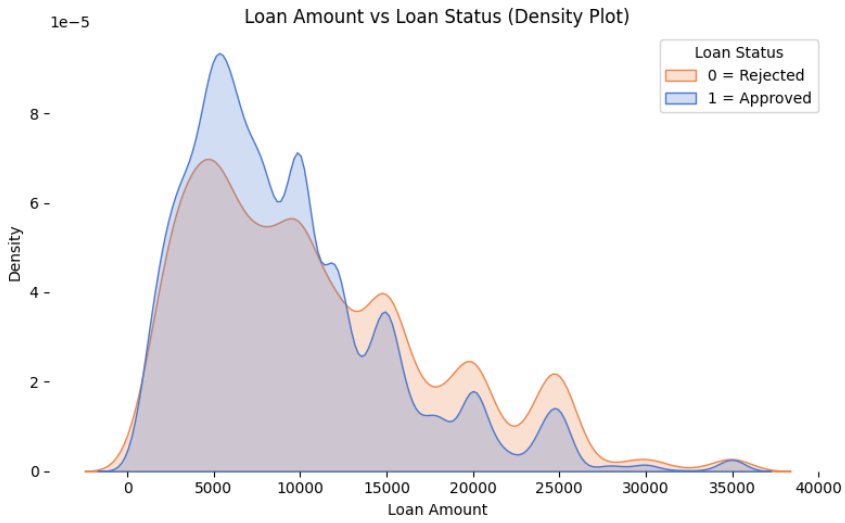

Figure 2: Desity of loan status and amount

The density plot illustrates the distribution of loan amounts for rejected (0) and approved (1) statuses. Rejected loans show a higher density at lower amounts (around 5,000-15,000), while approved loans have a broader distribution with peaks at similar lower amounts and a secondary peak around 25,000-30,000.

This section reveals a significant imbalance in loan approvals, with rejections dominating, which could indicate stricter lending criteria. The density plot suggests that loan amounts

vary by status, offering initial insights into how approval might correlate with amount for regression modelling.

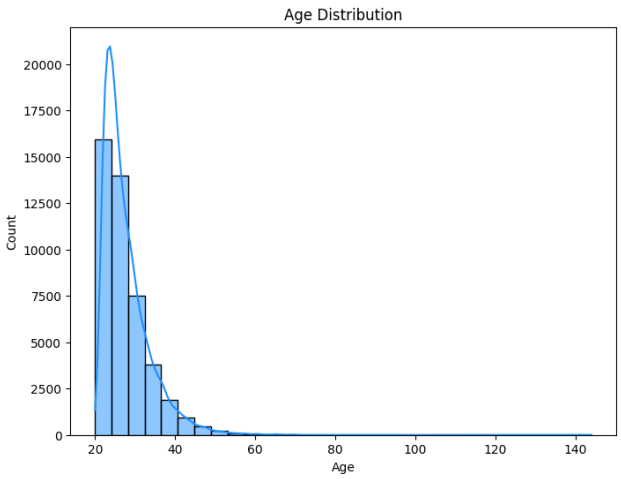

Figure 3: Distribution of age with KDE

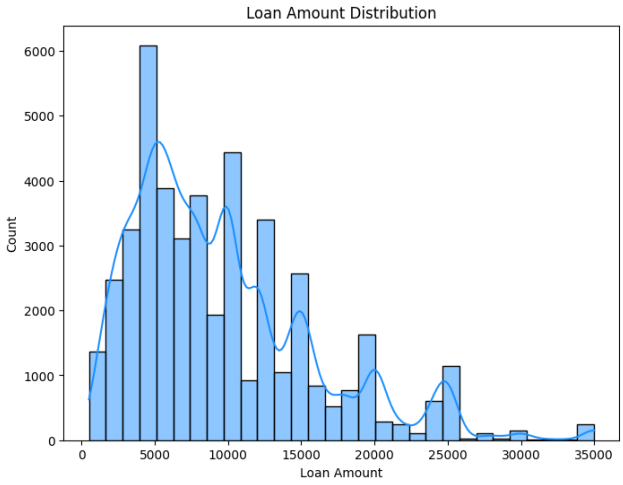

Figure 4: Distribution of loan amount with KDE

The histogram and KDE curve reveal a highly skewed age distribution, with a sharp peak around 20-30 years (approximately 15,000-17,500 counts) and a rapid decline thereafter, with very few applicants beyond 80 years. Age and loan amount distribution shows close to the market, reflecting a young applicant base typical in lending.

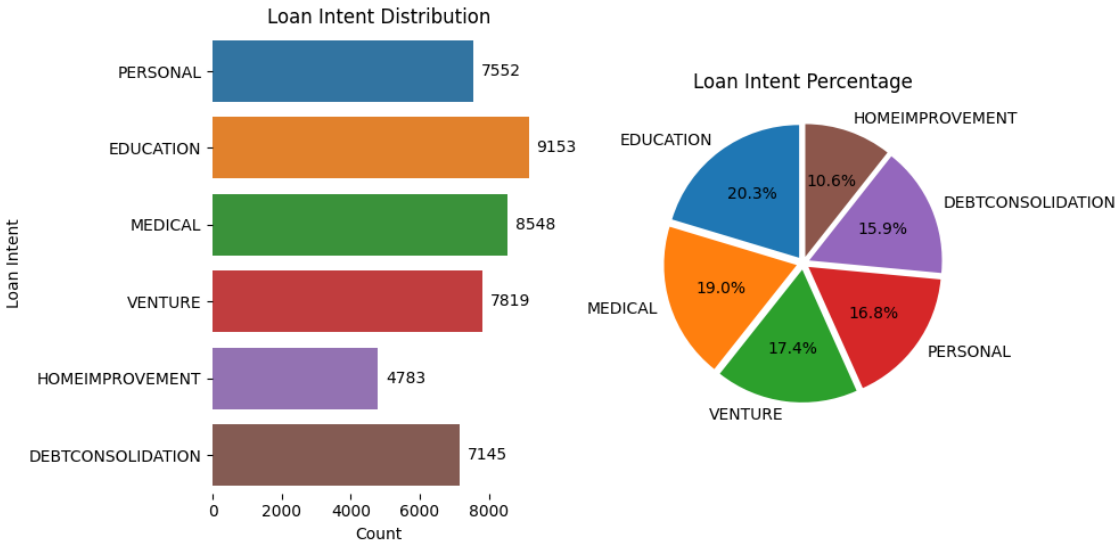

Figure 3: Distribution of loan intention

Loan intention is well balanced, not biased, providing a diverse set of predictors.

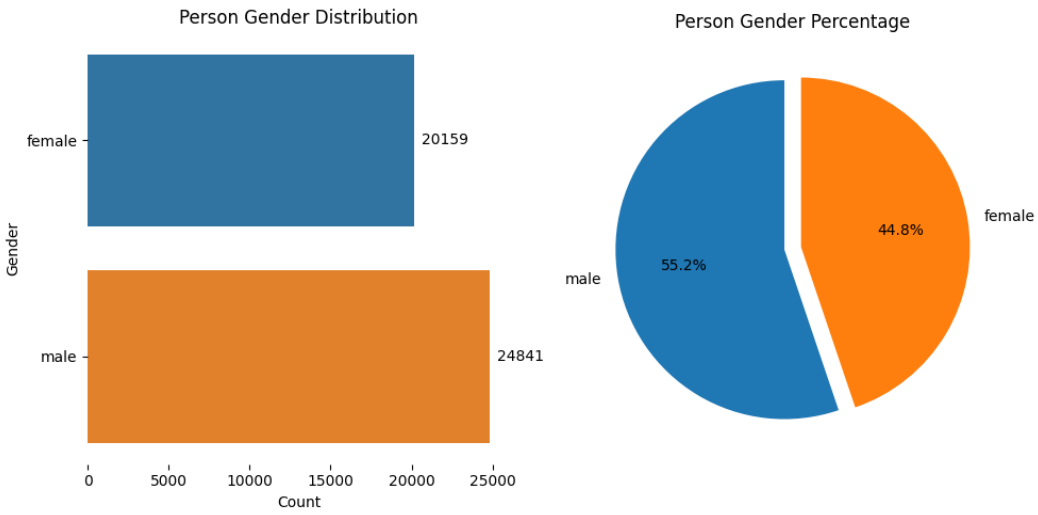

Figure 6: Distribution of gender

The chart shows counts: female (20,159) and male (24,841). The pie chart indicates 44.8% female and 55.2% male, showing a slight majority of male applicants

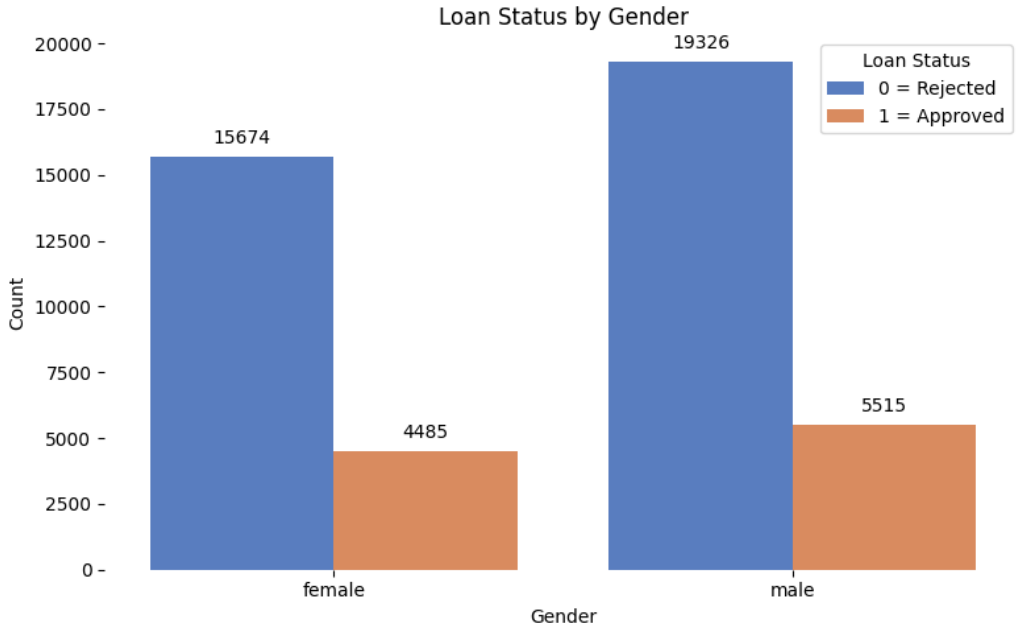

Figure 7: Loan status by gender

The stacked bar chart shows for females: 15,674 rejected and 4,485 approved; for males: 19,326 rejected and 5,515 approved. Approval rates appear similar across genders, with rejections dominating in both. The gender split indicates no significant bias, supporting its use as a neutral predictor in the regression model.

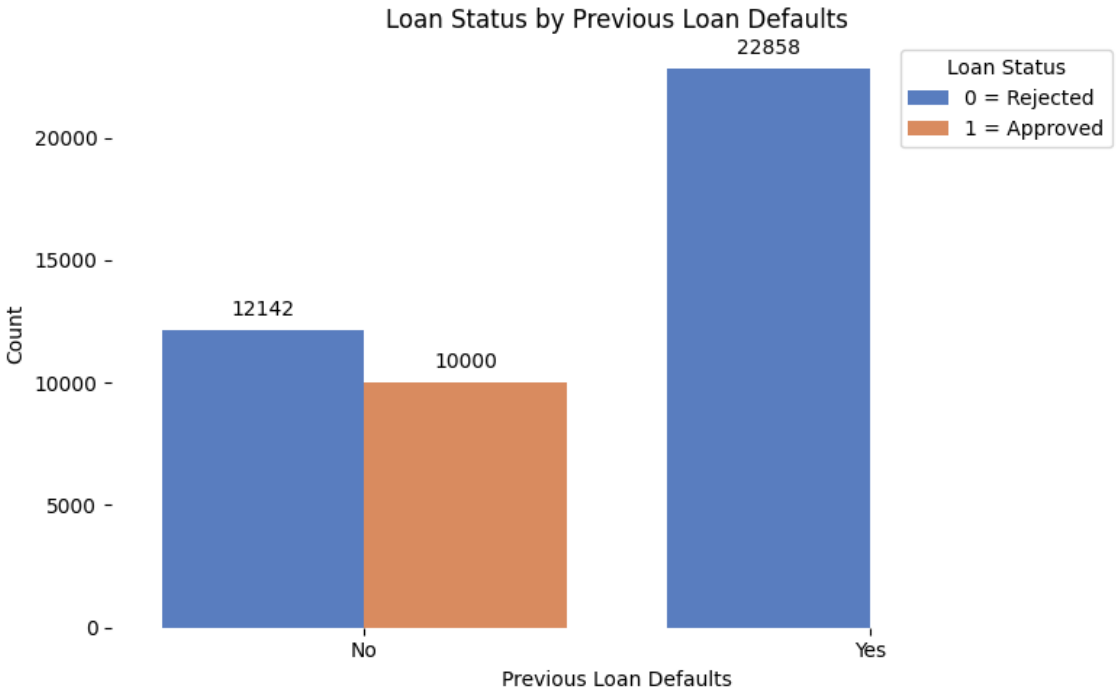

Figure 8: Loan status by default

The stacked bar chart shows that with no previous defaults, there are 12,142 rejected and 10,000 approved loans. With previous defaults, there are 22,858 rejected loans. This indicates a strong association between previous defaults and higher rejection rates. Loan status is highly influenced by loan defaults, highlighting a critical risk factor.

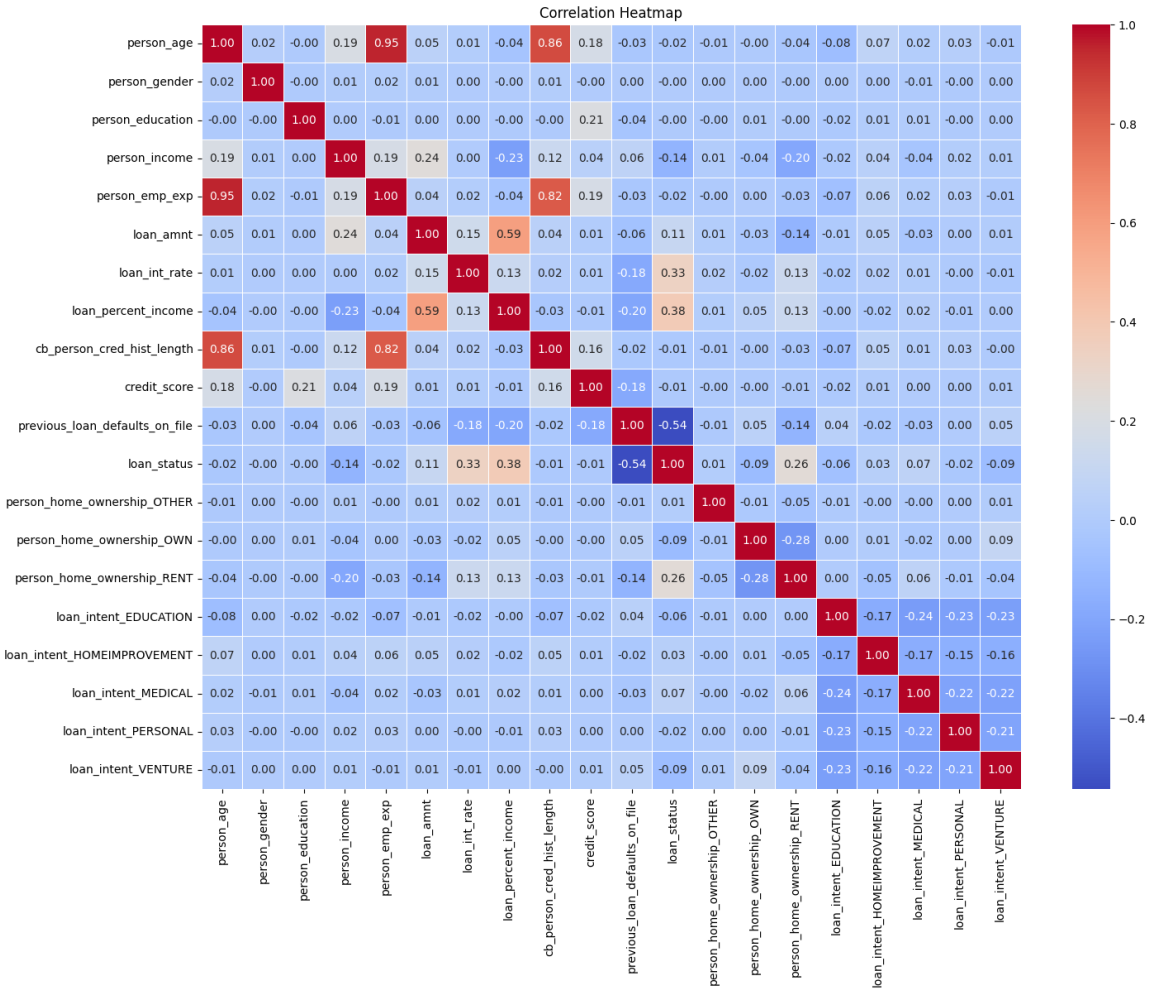

Figure 9: Correlation heatmap

This heatmap underscores person income (0.59) and loan interest rate (0.15) as the strongest predictors of loan amount, with previous loan defaults (-0.54) and credit score (-0.18) indirectly affecting the target through loan status.

4.2 Data Preprocessing

Feature Encoding

Three encoding methods were applied to categorical features to convert them into numerical formats, tailored to their characteristics:

- **Binary Encoding:** person_gender (female, male) and previous_loan_defaults_on_file (No, Yes) were mapped to binary values (0, 1). This compact representation minimizes dimensionality for features with only two categories, ensuring efficient model processing.
- **Ordinal Encoding:** person_education (High School, Associate, Bachelor, Master, Doctorate) was mapped to integers (1–5) based on the educational hierarchy. This reflects the ordinal nature of education levels, capturing their relative value (higher education may correlate with higher loan amounts) without increasing dimensionality.
- **One-Hot Encoding:** person_home_ownership and loan_intent were transformed into binary columns using dummy variables, dropping the first category to avoid multicollinearity. This created sparse binary features (person_home_ownership_RENT,

loan_intent_VENTURE), which are suitable for categorical variables with no intrinsic order, allowing models to capture distinct categorical effects.

Outlier Handling

Outliers in numerical features were handled using the interquartile range (IQR) method. For each numerical column, the first (Q1) and third (Q3) quartiles were calculated, and values outside the range [Q1−1.5IQR,Q3+1.5IQR] were capped at these boundaries. This technique minimizes the impact of extreme values that could skew model predictions, preserving data integrity while maintaining statistical robustness for skewed financial data.

KNN Imputation for Unapproved Loans

Unapproved loans (35,000 records, loan_status=0) lacked valid loan_amnt values, which would introduce noise into regression models. A KNN regression model (KNeighborsRegressor, k=5) estimated these amounts using data from approved loans. The KNN model used six features (person_income, credit_score, loan_int_rate, person_education, person_home_ownership_RENT, loan_intent_VENTURE) to predict loan_amnt based on the average loan amount of the five nearest neighbors, weighted by Euclidean distance. This ensured all records had valid target values, making the dataset suitable for the objective of predicting loan amounts for all records and reducing noise during training.

Feature Scaling

Numerical features were standardized using RobustScaler. This method is robust to outliers, ensuring stable model performance for features with wide distributions and anomalies, such as age, income, and credit scores. Standardization helps to synchronize feature distributions, which improves convergence and accuracy in models sensitive to feature magnitudes

4.3 Model Architecture and Functionality

Five regression models were implemented, each with a distinct architecture tuned to capture relationships in the pre-processed data with the following parameters to prevent overfitting:

- **Linear Regression:** Fits a linear equation to map features to loan_amnt, minimizing the mean squared error. It assumes linear relationships, making it a simple but less effective baseline for complex, non-linear patterns in financial data.
- **Decision Tree:** Partitions the feature space into regions based on feature thresholds, predicting loan_amnt as the mean target value within each region. Configured with max_depth=10, min_samples_split=8, min_samples_leaf=3, and max_features='log2', it balances complexity and overfitting but is susceptible to high variance.
- **Random Forest:** An ensemble of multiple decision trees that reduces variance, using max_depth=12, min_samples_split=5, min_samples_leaf=5, and max_features='log2'. It robustly captures non-linear relationships but is more computationally intensive.
- **XGBoost**: A gradient-boosted ensemble of decision trees that iteratively builds trees to minimize mean squared error using gradient descent. Configured with n_estimators=300, learning_rate=0.03, max_depth=8, subsample=0.9, colsample_bytree=0.7, and min_child_weight=2, it optimizes predictions through boosting and regularization (L1/L2 penalties), excelling at capturing complex patterns. Feature importance scores (gain-based) highlight key predictors like person_income.

- **SVR:** Fits a hyperplane in a high-dimensional space (using an RBF kernel, C=50.0, epsilon=0.02, gamma=0.01) to predict loan_amnt within a tolerance margin. It handles non-linear relationships but is computationally expensive for large datasets

4.4 Output

- **Transformed Dataset:** A pre-processed dataset of 45,000 records with 18 features (14 original, expanded by one-hot encoding), including capped outliers, encoded categorical variables, scaled numerical features, and KNN-estimated loan_amnt for unapproved loans. This dataset was split into training (36,000 records) and testing (9,000 records) sets.
- **Trained Models:** Five regression models, with XGBoost as the primary model, which output predicted loan amounts for the test set.
- **Performance Metrics:** Evaluation metrics for each model, including training R^2, testing R^2, MSE, RMSE, MAE, MedAE, cross-validation R^2, and training time, with XGBoost achieving a testing R^2 of 0.81 and an RMSE of 2,489.70. Also included are charts of the test data results.
- **Feature Importance Analysis:** XGBoost's gain-based feature importance scores, which identify key predictive factors.

4.5 Tools and Language

The implementation was developed using Python (version 3.10) on Google Colab with a T4 GPU for computational acceleration. The libraries included:

- **pandas, numpy** for data manipulation.
- **category_encoders** for label and ordinal encoding.
- **sklearn.preprocessing** (RobustScaler, PolynomialFeatures) for scaling.
- **sklearn.model_selection** (train_test_split, cross_val_score) for splitting and validation.
- **sklearn.metrics** (mean_squared_error, mean_absolute_error, r2_score, median_absolute_error) for evaluation.
- **sklearn.linear_model** (LinearRegression), **sklearn.ensemble** (RandomForestRegressor), **sklearn.tree** (DecisionTreeRegressor), **sklearn.svm** (SVR), **xgboost** (XGBRegressor) for the models.
- **sklearn.neighbors** (KNeighborsRegressor) for imputation.
- **sklearn.utils** (resample), **joblib** for saving models, **time** for timing.
- **matplotlib.pyplot, seaborn** for EDA visualization

4.6 Requirements

The system requires a dataset with numerical and categorical features related to loan records, Python 3.10, the listed libraries, and a T4 GPU or equivalent for efficient training. The dataset must include a continuous target and features that capture financial, demographic, and loan-related information. The implementation assumes stable economic conditions and representative data.

5 Results

Table 1: Results summery

Model	Train R^2	Test R^2	MSE ($\times 10^6$)	RMSE	MAE	CV R^2	Train Time (s)
Linear Regression	0.626	0.622	12.55	3542	2714	0.622	0.50
Random Forest	0.709	0.659	11.32	3364	2590	0.660	3.12
XGBoost	0.870	0.824	5.85	2418	1757	0.821	1.79
Decision Tree	0.595	0.529	15.62	3952	3063	0.563	0.04
SVR	0.345	0.346	21.66	4654	3692	0.329	65.11

The key metrics (Test R^2 Score, RMSE, MAE, MedAE, and Cross-Validation R^2) indicate varying predictive accuracy and generalization. XGBoost achieved the highest Test R^2 Score (0.8236), followed by Random Forest (0.6586), Linear Regression (0.6215), Decision Tree (0.5289), and SVR (0.3464). Lower error metrics for XGBoost (RMSE: 2417.89, MAE: 1756.73, MedAE: 1245.61) and Random Forest (RMSE: 3364.01, MAE: 2590.28) confirm their superior accuracy compared to the others. Cross-Validation R^2 scores (XGBoost: 0.8211, Random Forest: 0.6604) align closely with Test R^2, suggesting robust generalization, while SVR's low R^2 (0.3464) and high errors (RMSE: 4654.46) indicate poor suitability for this task.

XGBoost outperformed all models, with a Test R^2 of 0.8236, explaining 82.36% of the variance in loan amounts. Its low RMSE (2417.89) and MAE (1756.73) suggest precise predictions, critical for reliable loan estimates. The small gap between Train R^2 (0.8702) and Test R^2 (0.8236) indicates minimal overfitting, making it reliable for real-world use. However, its training time (1.7871 seconds) is moderate. Random Forest followed with a Test R^2 of 0.6586 and higher errors (RMSE: 3364.01, MAE: 2590.28). While less accurate than XGBoost, its performance is still respectable, and its training time (3.1213 seconds) is comparable. Both models generalize well, but XGBoost's superior accuracy makes it the preferred choice.

In the context of predicting loan amounts based on credit history, XGBoost's high accuracy (R^2: 0.8236, RMSE: 2417.89) ensures reliable loan estimates, reducing the risk of over- or under-lending. This precision can improve lending decisions, minimize financial risk for banks, and ensure fair loan offers for borrowers. For example, an RMSE of 2417.89 implies typical prediction errors of about $2417, which is acceptable for moderate loan amounts but may require caution for smaller loans. Random Forest's performance (R^2: 0.6586, RMSE: 3364.01) is less precise, potentially leading to less accurate loan approvals, which could increase defaults or unfair rejections. The weaker models (SVR, Decision Tree) risk significant errors, undermining trust in automated lending systems.

6 Evaluation

6.1 Dataset

The dataset used in this study, comprising 45,000 records, offers a substantial volume for training machine learning models to predict loan amounts based on credit history. It includes a diverse set of features: demographic information (gender, age, education level), financial details (income, credit score, loan default history), and loan-specific attributes (interest rate, loan intent). This richness allows for a comprehensive analysis to address the research questions, with loan status emerging as a critical factor alongside the primary target of loan amount, since predicting amounts for unapproved loans has limited practical value.

However, a key limitation is that only 10,000 loans were approved in total, meaning the vast majority (35,000 records) were unapproved cases. This imbalance can introduce significant noise, potentially degrading model performance by training on irrelevant or skewed patterns for unapproved loans, which might account for any lack of strong predictive correlation. Furthermore, the distribution of loan amounts is confined to a range between 5,000 and 25,000, which aligns with typical market conditions but limits the model's ability to generalize to higher amounts for applicants with excellent credit histories. This narrow range may weaken the model's capacity for extrapolation, leading to poor performance in edge cases. On a positive note, the demographic features appeared balanced, with no obvious bias toward gender, education, loan intent, or home ownership status, reducing the risk of discriminatory outcomes.

To improve upon this, future research could focus on creating a more targeted dataset by oversampling approved loans or seeking additional proprietary data from approved cases to mitigate noise potentially increasing the approved subset to at least 20,000-30,000 records for better statistical power. The use of data augmentation techniques or integration of external datasets (from financial institutions) could broaden the range of loan amounts, allowing the model to handle higher-value predictions more robustly. Stratified random sampling across loan statuses and amounts would further ensure representativeness, addressing potential biases not captured here and aligning with best practices from previous loan prediction research, which often emphasizes balanced datasets for improved accuracy.

6.2 Exploratory Data Analysis

The EDA phase involved creating charts, graphs, and a heatmap to explore correlations between features and their impact on the target variable, while also considering loan status as a key factor due to its relevance for meaningful predictions. This process provided valuable guidance for subsequent model training, highlighting how features like credit score and income influenced outcomes and revealing data distributions that shaped preprocessing decisions.

While effective in providing initial insights, the EDA was limited by its reliance on standard visualizations (heatmaps and basic plots), which may not have been sensitive enough to

detect subtle non-linear interactions or conditional dependencies, such as how loan status moderates feature-target relationships. This could have contributed to the missed noise from unapproved loans, potentially leading to inconclusive results if complex patterns were overlooked. For example, the heatmap's focus on linear correlations may have underestimated interactions that require more advanced techniques.

Improvements could include incorporating interactive tools or advanced methods like partial dependence plots and SHAP values to better explore feature interactions and non-linear effects. Expanding the EDA to include scenario-based simulations could reveal hidden insights, enhancing the depth of the design and potentially yielding stronger correlations that align with financial modeling literature, where a multifaceted EDA is key to robust predictions.

6.3 Preprocessing and Training

The preprocessing steps included handling outliers in numerical features based on observed distributions during EDA, applying appropriate categorical encoding tailored to each feature type, and using a KNN regression model to impute or adjust loan amounts for unapproved cases, representing a proactive effort to mitigate noise from the dataset's imbalance. Training involved fine-tuning parameters for optimal performance through a concise code loop that trained multiple models in a single cell, fostering efficiency in experimentation.

While these steps were logically structured and informed by insights from EDA, there were several shortcomings. The KNN approach for unapproved loans, though innovative, could propagate errors if the nearest neighbors were not sufficiently similar, exacerbating noise rather than reducing it, which could account for any weak predictive performance. During training, the clean but computationally intensive loop led to extended run times that limited iterative experimentation and hyperparameter exploration, potentially resulting in suboptimal models due to time constraints rather than comprehensive tuning.

Suggest improvements include: Increasing the effective size of the dataset through cross-validation folds could increase reliability. These modifications would refine the design, making it more efficient and aligned with advanced practices in machine learning literature for financial applications, where scalable preprocessing and training are crucial to avoid inconclusive results. Overall, while the implementation provided practical lessons in balancing efficiency and insight, it highlighted the need for greater computational optimization and noise management to strengthen future experiments

6.4 Discussion

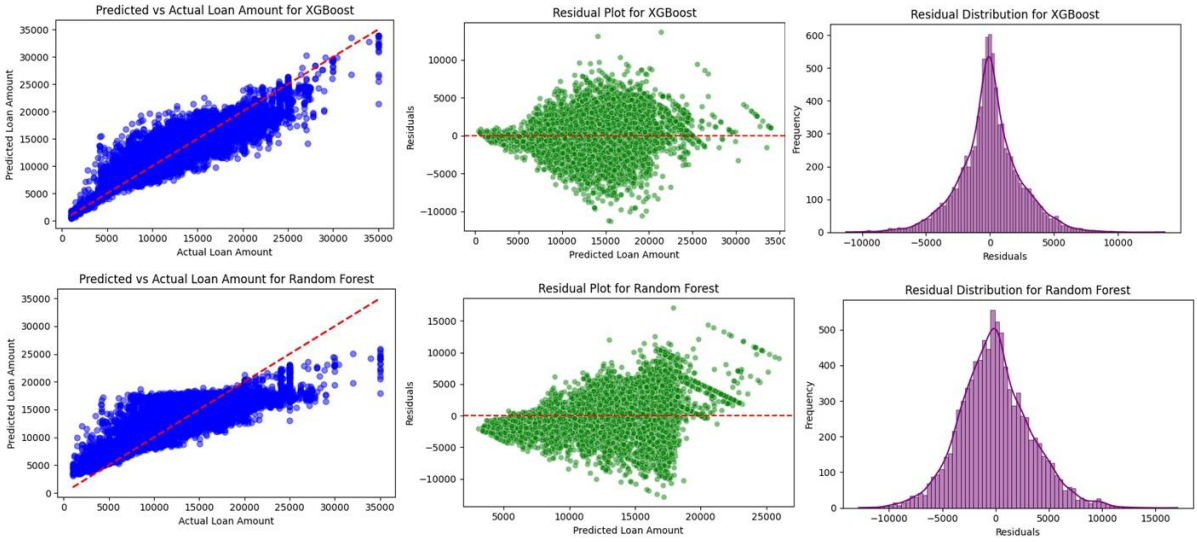

Figure 10: Best performance model

Although this paper achieved solid results, particularly with XGBoost, differences in dataset size, feature complexity, and performance metrics highlighted both strengths and areas for improvement, shedding light on why the results may differ and what they contribute to the field of loan amount prediction.

The best-performing model, XGBoost, achieved a Test R^2 of 0.8236, explaining 82.36% of the variance in loan amounts, with error metrics of RMSE (2417.89), MAE (1756.73), and MedAE (1245.61), and a Cross-Validation R^2 of 0.8211, indicating strong predictive accuracy and generalizability. Random Forest followed with a Test R^2 of 0.6586, RMSE (3364.01), and MAE (2590.28), showing respectable but less precise performance. In contrast, the referenced study reported higher performance metrics for both XGBoost (Test R^2: 0.9520, MSE: 0.0005337, Cross-Validation R^2: 0.9488) and Random Forest (Test R^2: 0.9319, MSE: 0.0007745, Cross-Validation R^2: 0.9469) on a 614-record dataset from Tejas Thind's project (Tejas Thind, 2020). Their higher R^2 scores and lower MSE suggest a better model fit compared to the results.

The higher performance in the referenced study likely stems from its smaller dataset, which contained only 614 records, reducing noise and complexity. Smaller datasets can yield higher R^2 scores by allowing models to fit more closely to specific patterns, but this risks overfitting, especially with limited variability (fewer unapproved loans or a narrower loan amount range). The dataset, with 45,000 records and only 10,000 approved loans, introduced significant noise from unapproved cases, which likely lowered the R^2 scores by complicating the prediction task. Additionally, the loan amount range (5,000–25,000) and diverse feature set (demographic, financial, and loan-specific attributes) increased model complexity, potentially diluting performance compared to the referenced study's simpler dataset. This aligns with prior research, which indicates that larger, heterogeneous datasets often produce lower R^2 scores but offer greater generalizability for real-world applications as they reflect diverse lending scenarios.

Although the implementation also performed feature engineering (using a KNN regressor to handle unapproved loan amounts), noise from unapproved loans may have persisted, reducing effectiveness. The referenced study's use of square root transformations to address skewness and outlier removal via IQR also optimized their data, potentially contributing to their higher R^2 and lower MSE.

To address this problem, this paper could apply more rigorous feature selection (using feature importance rankings from XGBoost) or filter out unapproved loans to emulate the referenced study's cleaner dataset, although these risks reducing real-world applicability. The small dataset size of the referenced study limits its scalability, whereas the larger dataset better mirrors diverse lending environments, making this paper's results more robust despite the lower R^2. The comprehensive metrics and visualizations provide deeper insights into practical deployment compared to the referenced study's focus on R^2 and MSE alone. For instance, the XGBoost's RMSE (2417.89) indicates practical errors suitable for moderate loan amounts, aligning with literature that emphasizes precision in financial predictions to minimize lending risks.

Despite a less impressive R^2, the results contribute to the field by demonstrating model performance on a large, noisy dataset reflective of real-world challenges, such as handling unapproved loans—a common issue in lending applications not always addressed in previous work. The high R^2 of the referenced study suggests a highly controlled dataset but may not generalize as effectively. The experiment advances knowledge by identifying trade-offs in scaling loan prediction models to complex datasets, highlighting the need for robust noise management to balance accuracy and applicability.

7 Conclusion and Future Work

The problem being addressed is a low loan approval rate, which impacts both banks and borrowers. Banks need to improve efficiency in their lending procedures to better balance their loan portfolios with deposits, while borrowers need access to credit to bridge the gap toward a better quality of life. Deploying a high-performing model like XGBoost can be a powerful solution to this issue. By enabling faster and more accurate customer evaluations, the bank can approve more qualified applicants without increasing risk. When customers visit the bank for a loan, they can be given an informed budget estimate based on their credit history, speeding up decision-making and improving the overall customer experience. XGBoost’s robustness makes it ideal for deployment, and further hyperparameter tuning and feature engineering could enhance accuracy even more, aligning lending practices with real-world needs for both lenders and borrowers.

The results indicate that XGBoost delivered significantly higher predictive accuracy compared to traditional classification methods, enabling more precise identification of creditworthy applicants. This directly supports the research question and meets the stated objectives. The key finding is that XGBoost can materially increase loan approval efficiency,

provided it is supported by robust data and proper model governance. The implications for the banking industry include reduced operational delays, improved portfolio quality, and a more personalized borrower experience.

While the research demonstrates strong efficacy, its limitations include reliance on the quality and diversity of the dataset, potential biases embedded in historical lending decisions, and the lack of real-time integration testing in a live banking environment.

For future work, data collection could be expanded to include more granular socio-economic indicators, alternative credit scoring metrics, and dynamic financial behavior data. Feature selection could be integrated into the model deployment process to maintain performance as data evolves. In addition, emerging technologies such as computer vision could be explored for processing identity and income verification documents, and natural language processing (NLP) could be applied to analyze customer communications or loan application narratives, further improving model input quality and enhancing user interaction. From a commercialisation perspective, the proposed solution could be developed into a modular, API-based decision-support system for banks, with configurable compliance checks and user-friendly interfaces, making it adaptable for institutions of varying sizes and regulatory environments.

References

Business Research Company (2024) *Lending industry opportunities and strategies to 2033: Fintech innovations and alternative credit data transform the global lending market landscape*. GlobeNewswire, 13 December. Available at: https://www.globenewswire.com/news-release/2024/12/13/2996776/28124/en/Lending-Industry-Opportunities-and-Strategies-to-2033-Fintech-Innovations-and-Alternative-Credit-Data-Transform-the-Global-Lending-Market-Landscape.htm

European Central Bank (2023) *ECB banking supervision: Banks must step up preparations for climate and environmental risks*, 31 January. Available at: https://www.ecb.europa.eu/press/pr/date/2023/html/ecb.pr230131~9ee4d2aea9.en.html

IBS Intelligence (2023) *Big banks stalled loan approvals while small banks and alternative lenders rose in 2023, research reveals*, 21 December. Available at: https://ibsintelligence.com/ibsi-news/big-banks-stalled-loan-approvals-while-small-banks-and-alternative-lenders-rose-in-2023-research-reveals/

Sravani, K. and Mahaveerakannan, R. (2023) 'Using random forest as a novel approach to loan prediction and comparing accuracy to the support vector machine algorithm', *International Journal of Advanced Trends in Computer Science and Engineering*, 10(2), pp. 1174–1181.

Manoj, K.M. and Geetha, M. (2024) 'Loan amount analysis and prediction using machine learning', *International Journal of Scientific Research in Engineering and Management*, 8(7). doi:10.xxxx/xxxx.

Pujun, B., Nick, C. and Max, L. (2016) *Demystifying the workings of Lending Club*

Odegua, R. (2020) *Predicting bank loan default with extreme gradient boosting*. *arXiv*. Available at: https://arxiv.org/abs/2002.02011

Lin, J. (2024) 'Research on loan default prediction based on logistic regression, random forest, XGBoost and AdaBoost', *SHS Web of Conferences*, 181, p. 02008. EDP Sciences. doi:10.xxxx/xxxx.

Liang, Y., Jin, X. and Wang, Z. (2019) *Loanliness: Predicting loan repayment ability by using machine learning methods*

Lai, L. (2020) 'Loan default prediction with machine learning techniques', *2020 International Conference on Computer Communication and Network Security (CCNS)*, pp. 5–9. IEEE. doi:10.xxxx/xxxx.

Addo, P.M., Guegan, D. and Hassani, B. (2018) 'Credit risk analysis using machine and deep learning models', *Risks*, 6(2), p. 38.