Comparative Analysis of ML Algorithms for Personal Loan Prediction

Dai Vinh Quach
dept. of Faculty of Business
Humber College
Toronto, Canada
n01607919@humber.c

Krishna Prasad Tanneeru,
dept. of Faculty of Business
Humber College
Toronto, Canada
n01650146@humber.ca

Tenzin Yangzom
dept. of Faculty of Business
Humber College
Toronto, Canada
n01667419@humber.ca

Thi Thu Hien Nguyen
dept. of Faculty of Business
Humber College
Toronto, Canada
n01604562@humber.ca

Wenrui Shan
dept. of Faculty of Business
Humber College
Toronto, Canada
n01064064@humber.ca

***Abstract*— This study aims to enhance the accuracy of personal loan acceptance prediction using machine learning techniques, with a specific focus on Neural Networks (NNs) and model selection. The research utilizes the "Bank Personal Loan Modelling" dataset from Kaggle, which includes various customer attributes such as age, income, education level, and credit card usage. The personal loan acceptance problem is framed as a binary classification task, where the target variable indicates whether a customer accepted a loan (1) or not (0). In addition to conventional models—such as Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Naive Bayes, Decision Tree, and Logistic Regression—this study evaluates the performance of Neural Networks. The results indicate that Decision Tree Regression is the best model for Personal Loan Prediction.**

Keywords— Neural Networks (NN), Personal Loan Prediction, Bank Personal Loan, Machine Learning, Classification, Linear Regression (SVM), KNN, Naive Bayes, Decision Tree, Logistic Regression

I. Introduction

In the competitive retail banking landscape, accurately predicting customer acceptance of personal loan offers is crucial for optimizing marketing strategies, enhancing risk management, and improving customer engagement. Banks collect a wealth of customer data, including demographic information, financial behaviors, and account activities, which can be leveraged for informed, data-driven decisions. By employing machine learning techniques to analyze this structured data, financial institutions can more effectively identify potential loan applicants

The objective of this study is to develop a robust predictive model for personal loan acceptance using a real-world dataset. The dataset, sourced from Kaggle, contains information on 5,000 customers, including demographic details, account information, and a binary indicator of whether the customer accepted a personal loan. By systematically applying data preprocessing techniques, feature engineering, and machine learning algorithms, this project aims to identify key factors influencing personal loan acceptance and build a model that can accurately predict customer behavior.[1][2]

The project aims to uncover the underlying patterns and key features that influence customers' decisions through supervised learning techniques. A critical aspect of this research involves implementing data preprocessing steps and feature selection methods, which are crucial for minimizing noise and enhancing the model's performance. Techniques like forward stepwise selection have been employed to identify the most relevant variables, thus improving both model accuracy and computational efficiency. Several classification algorithms have been trained and evaluated, including Logistic Regression, K-Nearest Neighbors (KNN), Naive Bayes, Decision Tree, and Support Vector Machine (SVM), based on performance metrics such as accuracy, precision, recall, and F1-score. Additionally, a Neural Network model (Multi-Layer Perceptron Classifier) has been implemented to assess its capability in capturing complex, non-linear relationships within the data. This approach provides a comparative analysis of traditional algorithms alongside advanced deep learning techniques.

The latter part of this report discusses methodology, including data preprocessing, model training, and evaluation. A comparative analysis of different machine learning algorithms is conducted to assess their predictive performance and highlight the importance of feature selection and preprocessing in improving model accuracy.

II. Literature Review

With the continuous advancement of financial technology, loan approval prediction has become an important area of research for automating credit assessment. The availability of structured customer data—such as demographic information, financial behavior, and historical borrowing records—has provided a solid foundation for the application of machine learning models. In recent years, researchers have explored a variety of algorithms to improve prediction accuracy. Neural Networks (NNs), particularly Multi-Layer Perceptrons (MLPs), have gained popularity in financial prediction tasks due to their ability to model non-linear relationships and handle large feature spaces. It's the most commonly used with other conventional machine learning algorithms including logistic regression, decision trees, K-nearest neighbors, Naive Bayes, and support vector machines.

Logistic regression is widely applied in credit scoring due to its interpretability and simplicity, yet its reliance on linear relationships between variables may limit its performance in more complex datasets. In contrast, models like decision trees and K-nearest neighbors are capable of capturing non-linear patterns, though they may suffer from overfitting or computational inefficiency.[3] Naive Bayes assumes independence among features but still performs robustly in high-dimensional or imbalanced datasets. Support vector machines offer high accuracy in non-linear separable problems but are often computationally demanding and less interpretable.

Moreover, the literature indicates that data preprocessing plays a crucial role in enhancing model performance. Techniques such as feature scaling, outlier treatment, and appropriate train-test splitting are essential for improving prediction outcomes. While some studies advocate for the use of ensemble algorithms and advanced feature selection methods, comparative analysis of common classifiers under standard preprocessing remains highly practical in many real-world scenarios.[6]

In addition to these traditional machine learning methods, feedforward neural networks (FNN), also known as multilayer perceptrons (MLP), have been increasingly explored for loan approval prediction due to their ability to model complex, non-linear relationships between features. For instance, Shinde and Aphale (2020) compared FNN with six other supervised learning algorithms—including logistic regression, decision trees, Naive Bayes, and support vector machines—on a structured credit dataset. The FNN model demonstrated competitive performance, achieving an accuracy of around 80%, although the study lacked detailed reporting on network architecture or training optimization techniques. This suggests that, while neural networks offer strong predictive capabilities, their performance depends heavily on proper tuning and may require more computational resources and interpretability considerations than simpler models. [7]

In summary, existing research suggests that both traditional machine learning models and neural networks exhibit distinct strengths under different data conditions. Shallow models such as logistic regression and decision trees remain popular due to their interpretability and efficiency, while feedforward neural networks offer improved capacity to model complex, non-linear relationships—albeit at the cost of increased computational complexity and reduced transparency. Therefore, selecting algorithms based not only on predictive performance but also on model interpretability, resource constraints, and the specific operational needs of financial institutions is key to building effective loan prediction systems. This study, in light of this context, systematically compares the performance of several commonly used classifiers, including a feedforward neural network, in loan approval prediction tasks.

III. Methodology

This project utilizes six machine learning models, including Neural Network (NN), Support Vector Machine (SVM) for linear regression, Logistic Regression (LR), K-Nearest Neighbors (KNN), Decision Trees (CART), and Naive Bayes (NB), all implemented using Python. The performance of these models is evaluated through comparisons of Accuracy rate and Recall rate (Sensitivity). The evaluation process encompasses several stages, including data collection, pre-processing, training models, applying algorithms to analyze and classify variables, and comparing performance.

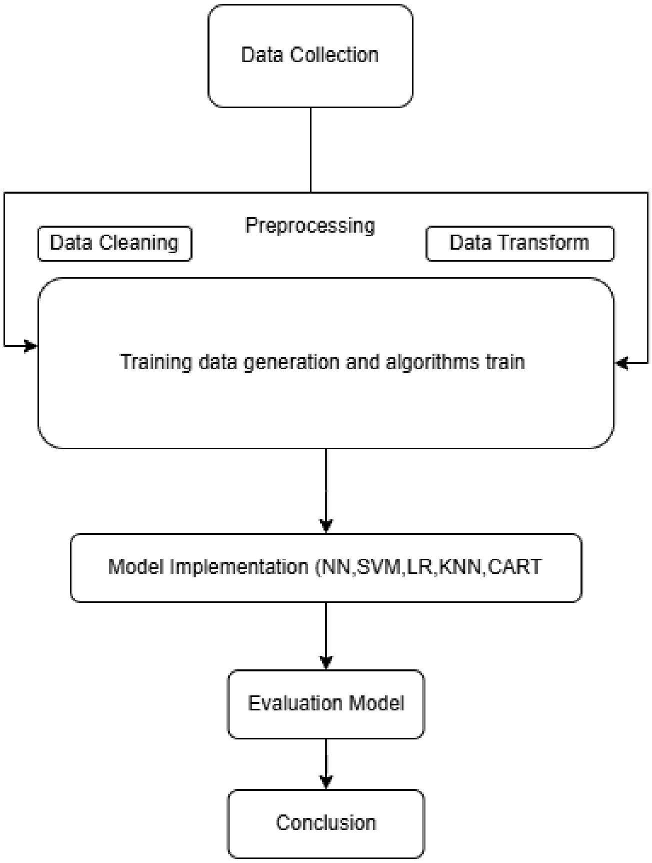
Data Collection
Preprocessing
Data Cleaning
Data Transform
Training data generation and algorithms train
Model Implementation (NN,SVM,LR,KNN,CART
Evaluation Model
Conclusion

Figure 1: Flow diagram of project work

The data set consists of two distinct classification values: a liability customer who buys personal loans is represented by the value (1), while a customer who does not buy a personal loan is represented by the value (0). This binary classification system categorizes each customer based on whether they choose to purchase a personal loan, allowing for a clearer analysis of loan perspectives and decision-making processes.

The implementation of machine learning algorithms applied to data before and after the normalization process. The normalization process involves transforming the data features to a common scale, typically by standardizing them to have a mean of zero and a standard deviation of one. The purpose of normalization is to enhance the models' ability to learn more effectively and efficiently, as it helps mitigate issues related to differing scales of measurement among various features.

In the Logistic regression models, we undertook additional feature selection techniques, including forward selection and grid search, to enhance the model's performance.

The evaluation criteria for machine learning models are clearly defined and involve a comprehensive analysis of their performance metrics. This consists of assessing key indicators, including accuracy rate and recall rate. These rates were calculated through the confusion matrix as follows: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) [5].

Accuracy is the key metric that measures the proportion of correctly predicted transactions among all predictions made. [5]

$$Accuracy\ =\ \frac{TP + TN}{TP + FP + TN + FN}$$

Recall (Sensitivity) refers to the ratio of actual positive values that are accurately predicted as positive. It is also known as the True Positive Rate (TPR) [5]

$$Recall\ =\ \frac{TP}{TP + FN}$$

It is also essential to understand how normalization affects the models' predictive capabilities. By comparing these metrics, we can gain insights into the importance of selecting the right model to enhance the effectiveness and robustness of predictions.

IV. Collection of the dataset and pre-processing

The dataset, obtained from Kaggle, includes information on 5,000 customers along with the variables listed in the table below.

Table 1: Dataset Features

No	Variable	Description
1	ID	Customer ID
2	Age	Customer's age in completed years
3	Experience	Number of years of professional experience
4	Income	Annual income of the customer ($000)
5	ZIP Code	Home Address ZIP code
6	Family	Family size of the customer
7	CCAvg	Average spending on credit cards per month ($000)

8	Education	Education Level
		1: Undergrad;
		2: Graduate;
		3: Advanced/Professional

9	Mortgage	Value of house mortgage if any ($000)
10	Securities Account	Does the customer have a securities account with the bank?
11	CD Account	Does the customer have a certificate of deposit (CD) account with the bank?
12	Online	Does the customer use internet banking facilities?
13	Credit card	Does the customer use a credit card issued by this Bank?
14	Personal Loan	Did this customer accept the personal loan offered in the last campaign?

Data Processing

During the Processing stage, the dataset underwent a thorough cleansing to ensure its integrity and quality. Utilizing Python, we meticulously identified and addressed missing values, which brought to light several crucial insights. Among these were the presence of duplicated entries and the identification of erroneous data, such as negative numbers where they were not applicable.

To enhance the accuracy of our analysis, we categorized the variables into two distinct groups: numeric and categorical. This classification allowed us to implement more tailored handling strategies for each type. Furthermore, we employed standardization techniques to normalize the dataset, ensuring that different scales did not skew our results. Additionally, we applied principal component analysis (PCA) to effectively transform correlated variables into independent principal components, thereby simplifying the complexity of our data while retaining its essential patterns and information.

A correlation analysis was conducted to examine the relationships among variables. This analysis revealed a robust negative correlation (-0.99) between Age and Experience. The correlation matrix, which presents correlations among the various features, is illustrated in the Figure below.

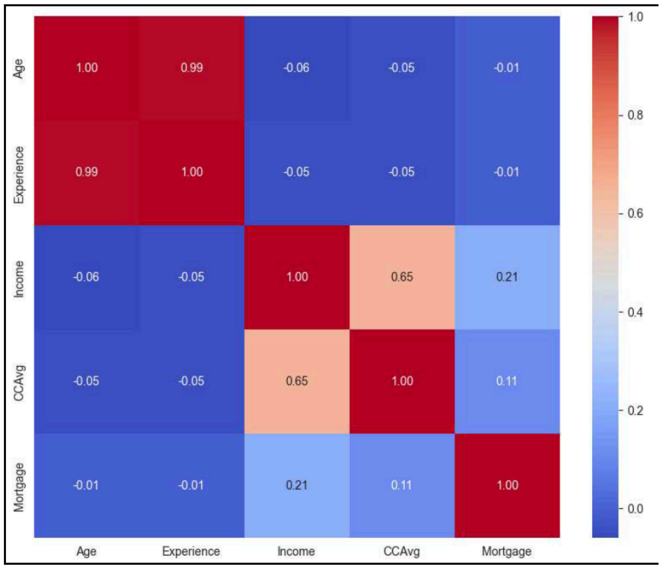

Figure 2: The relationships between all variables

After applying PCA, 2 new variables consist of PCA1 representing the Age&Experience relationship, and PCA2, representing Anual income.

In this project, the dataset is divided into two subsets: a training set and a testing set, utilizing a ratio of 70:30. This approach ensures that a significant portion of the data is dedicated to training the models while reserving enough data for an unbiased evaluation of their performance.

V. Prediction Model Results

Neural Network

Neural network (NN) models were used to predict whether a customer would accept a personal loan, based on features such as age, income, education, and account status. To compare different neural network models, we experimented with several key parameters. First, we tested architectures with varying numbers of hidden neurons (4 and 7) in a single hidden layer. Second, we examined the effect of data preprocessing by training models on both unstandardized and standardized datasets. Third, we explored different training configurations, including activation functions (logistic and relu) and solvers (lbfgs and adam). All models were trained using the default maximum number of iterations (200 epochs), and early convergence was observed depending on the solver. These experiments demonstrated that proper data preprocessing and parameter tuning—especially with respect to activation function, optimizer, and hidden layer size—significantly improved model performance, particularly in terms of accuracy and recall.

Table 2: Model without standardized data

Metric	**Value (4)**	**Value (7)**
Accuracy	0.95	0.95
Recall	0.63	0.62
Precision	0.85	0.84

Table 3: Model standardized data

Metric	**Value (4)**	**Value (7)**
Accuracy	0.98	0.98
Recall	0.88	0.83
Precision	0.87	0.91

Interpretation

Based on the results, the 4-node neural network exhibited a higher recall (0.88), indicating better sensitivity in identifying actual loan acceptors. On the other hand, the 7-node model achieved higher precision (0.91), suggesting fewer false positives. Given that both models had the same overall accuracy (98%), the choice between them depends on business priorities: if minimizing missed opportunities is critical, the 4-node model is preferable; if maximizing the accuracy of positive predictions is more important, the 7-node model is more suitable.

Support Vector Machine

Support Vector Machine (SVM) models were used to classify whether a customer would accept a personal loan, based on features such as age, income, education, and account status. Four kernel types—Linear, RBF, Polynomial, and Sigmoid—were tested using 5-fold cross-validation on both raw and standardized data. Standardizing the features significantly improved model performance, especially for non-linear kernels. The RBF kernel on standardized data achieved the best accuracy (96.44%) and was selected as the final model.

Table 4: Model without standardized data

Kernel	**Mean Accuracy**
Linear	0.9477
RBF	0.9086
Polynomial	0.9118
Sigmoid	0.8730

Table 5: Model with standardized data

Kernel	**Mean Accuracy**
Linear	0.9483
RBF	0.9644
Polynomial	0.9578
Sigmoid	0.8848

Interpretation

The SVM model's accuracy improved from 96% to 98%, and recall increased from 0.58 to 0.81 after standardizing the features. This indicates the model became more effective at identifying actual loan acceptors (true positives) with data scaling. The results highlight the importance of preprocessing for SVM models and demonstrate that the RBF kernel with standardized data offers the most reliable and well-balanced performance for predicting personal loan acceptance.

KNN

The KNN Model is a data-driven classification approach that does not require the model to be trained beforehand. To begin, the best k value is identified by selecting the one that provides the highest classification accuracy through validation of the training set.

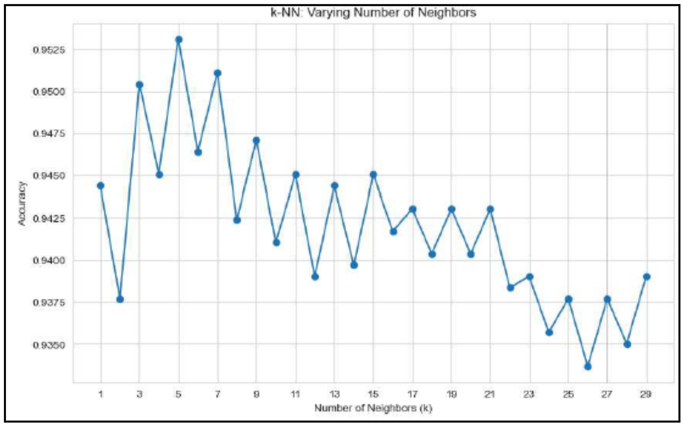

Figure 3: K value vs accuracy

As shown in Figure 1, K=5 yields the highest accuracy and is therefore selected as the optimal parameter. The model's performance using K=5 on the validation set is shown below.

Table 6: Performance of Validation Set

Metric	**Value**
Accuracy	0.97
Recall	0.71
Precision	0.97

Confusion Matrix	Predicted: No Loan	Predicted: Loan
Actual: No Loan (n=1,353)	1347	6
Actual: Loan (n=140)	64	76

Interpretation

The KNN Model demonstrated exceptional performance on the validation set, with a very high accuracy of 95%, and balanced precision (93%) and recall (54%) However, due to the low recall, the model correctly flags most of No Loan cases, but fails to detect a significant number of actual Loan cases.

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

Figure 4: Euclidean distance measure formula [4]

Logistic regression (LR)

The initial phase of model development involved training the model using all available predictors. Subsequently, we implemented a grid search to optimize the model's parameters, ensuring that we identified the best configuration for our analysis. The performance of the algorithm remains unchanged after data standardization.

Table 7: Performance of Validation Set

Metric	**Value**
Accuracy	0.95
Recall	0.6
Precision	0.84

Confusion Matrix	Predicted: No Loan	Predicted: Loan
Actual: No Loan (n=1,353)	1337	16

Actual: Loan (n=140)	56	84

After applying algorithms to dataset before and after normalization, we derived the equation with the lowest AIC value (-435.49):

$$\begin{aligned} Y = &-4.48 + 0.12*\text{Experience} + 2.43*\text{Income} \\ &+ 0.78*\text{Family} + 0.14*\text{CCAvg} + + 1.29*\text{Education} \\ &+ 0.04*\text{Mortgage} - 0.28*\text{Security\_Account} \\ &+ 0.86*\text{CD\_account} - 0.25*\text{Online} - 0.51*\text{Credit\_card} \\ &- 0.02*\text{PCA2} - 0.04*\text{PCA2} \end{aligned}$$

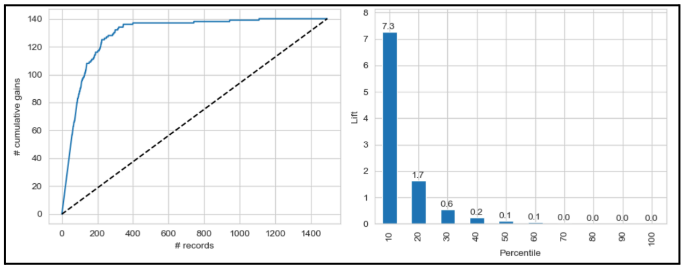

Figure 5: Cumulative Gains chart and Lift chart

Interpretation

The logistic regression (LR) model exhibited exceptional performance metrics, achieving an accuracy rate of 95%, which underscores its reliability in classifying outcomes correctly. It also recorded a recall of 60%, indicating its ability to effectively identify positive cases among the total actual positives.

Analyzing the performance through the Cumulative Gains chart reveals that the model can pinpoint a substantial proportion of positive instances early in the ranking process, significantly outperforming random selection strategies. This early identification is crucial for timely interventions and decision-making.

Moreover, the model achieved an impressive lift of 7.3 within the top 10th percentile of predicted probabilities, illustrating its strong predictive power in accurately identifying high-value cases within this segment. Such a lift indicates that the model is proficiently concentrating on the most relevant cases, thereby enhancing its practical application in real-world scenarios.

Naive Bayes

We implemented a Naive Bayes classification model using the MultinomialNB algorithm from scikit-learn. Prior to model training, all predictor variables were scaled using MinMaxScaler to meet the assumptions of the multinomial distribution. The model was trained on the training set and evaluated on the validation set (n = 1,493).

Performance on Validation Set

Table 8: Classification report of Naive Bayes

Metric	**Value**
Accuracy	0.91
Recall	0.01
Precision	0.0

Confusion Matrix	Predicted: No Loan	Predicted: Loan
Actual: No Loan (n=1353)	1353	0
Actual: Loan (n=140)	140	0

Interpretation

The Naive Bayes model achieved a high overall accuracy of 91%, largely due to the dominance of the “No Loan” class in the dataset. However, the model performed extremely poorly in identifying the minority “Loan Approved” class, with a recall of only 0%, indicating it rarely identifies positive cases Although the precision for class 1 was 100%, this is misleading, as the model predicted only one positive case and happened to get it right, which is likely by chance.

These results suggest that the Naive Bayes model is not well-suited for highly imbalanced classification tasks where the ability to identify the minority class is critical—such as predicting loan approvals.

Decision tree (CART)

The Decision Tree model using the CART (Classification and Regression Tree) algorithm was evaluated as part of the classification modeling efforts. The model underwent training and was assessed using a validation set to ensure that the results generalized well beyond the training data. (n=1,353)

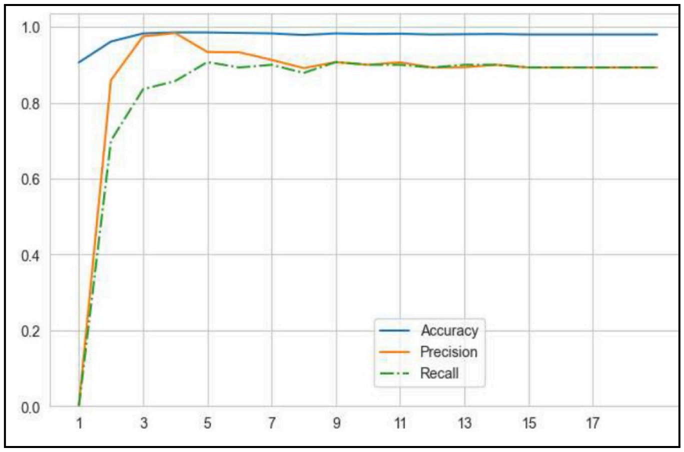

Figure 6: CART’s accuracy vs precision vs recall

Performance on Validation Set

Table 9: Classification report of Decision Tree (CART)

Metric	**Value**
Accuracy	0.98
Recall	0.93
Precision	0.91

Confusion Matrix	Predicted: No Loan	Predicted: Loan
Actual: No Loan (n=1,353)	1340	13
Actual: Loan (n=140)	10	130

Interpretation

The Decision Tree model demonstrated exceptional performance on the validation set, with a very high accuracy of 98%, and balanced precision (91%) and recall (93%), suggesting it is highly effective at distinguishing between loan and non-loan applicants. Additionally, the confusion matrix confirms minimal misclassification, making this model a strong candidate for deployment or further analysis.

VI. Results and Discussion

Without Data normalization

Table 10: Accuracy and Recall rates without normalization

Model	**Recall**	**Accuracy**
Neural network (NN)	0.63	0.95
Linear regression (SVM)	0.58	0.96
Decision tree (CART)	**0.93**	**0.98**
Logistic regression (LR)	0.60	0.95
Naı¨ve Bayes (NB)	0.00	0.91
K-nearest neighbour (KNN)	0.50	0.93

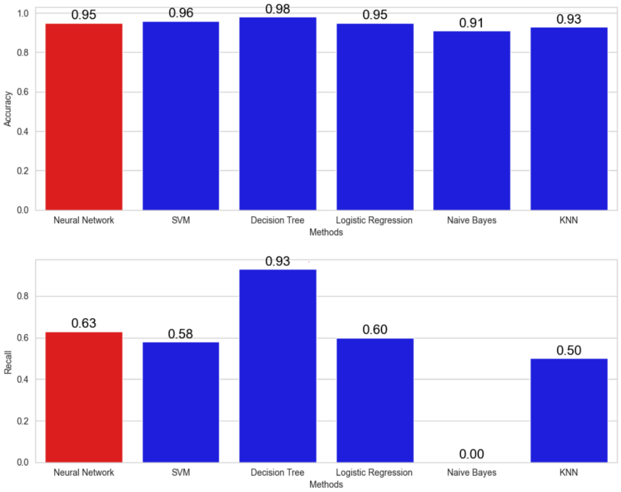

Figure 7: Accuracy and Recall without normalization

In the analysis of various classification models for identifying customers likely to accept personal loans, the Decision Tree model emerged as the most effective, achieving an impressive recall rate of 93% and an accuracy of 98%. This makes it the most reliable choice for this task. The Support Vector Machine (SVM) and Logistic Regression models also showed strong accuracy, recording values of 96% and 95%, respectively. However, their recall scores were lower, at 0.58 for SVM and 0.60 for Logistic Regression, indicating a moderate capability in capturing positive cases of loan acceptance. The K-Nearest Neighbor (KNN) model had an accuracy of 93%, but its recall rate was only 0.50, which implies that it missed a significant number of potential positive cases. Conversely, the Naïve Bayes model produced the least favorable results, resulting in a recall of 0.00 and an accuracy of 91%, indicating that it did not detect any instances of loan acceptance. Finally, the Neural Network (NN) model performed reasonably well, achieving a recall of 0.63 and an accuracy of 95%. This suggests a balanced classification ability, despite being applied without data normalization.

Data normalization

Table 11: Accuracy and Recall rates with normalization

Model	Recall	Accuracy
Neural network (NN)	0.87	0.98
Linear regression (SVM)	0.81	0.98
Decision tree (CART)	**0.91**	**0.99**
Logistic regression (LR)	0.60	0.95
Naı¨ve Bayes (NB)	0.00	0.91
K-nearest neighbour (KNN)	0.71	0.97

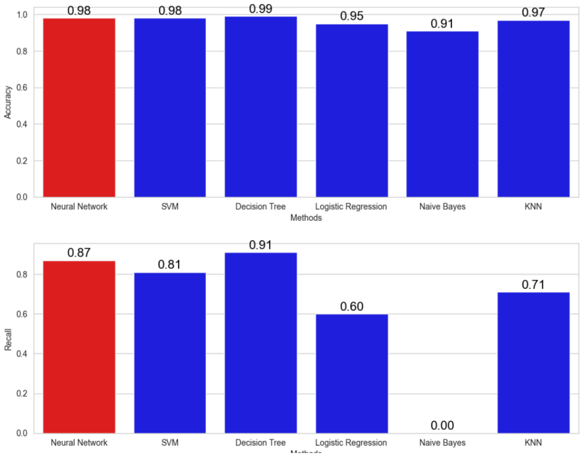

Figure 8: Accuracy and Recall with Normalization

The analysis of various machine learning models revealed how normalization affects their performance. The Decision Tree model emerged as the top performer, maintaining high metrics with a recall of 0.91 and accuracy of 0.99 after normalization. In contrast, the Logistic Regression model did not show any improvement with normalization, retaining its recall at 0.60 and accuracy at 0.95. This suggests that feature scaling does not provide significant benefits for this particular algorithm. The Support Vector Machine (SVM) model demonstrated notable improvements post-normalization, achieving a recall of 0.81 and an accuracy of 0.98, indicating its enhanced performance with scaled data. Similarly, the K-Nearest Neighbors (KNN) model also exhibited a marked improvement, achieving a recall of 0.71 and accuracy of 0.97. This highlights the sensitivity of distance-based algorithms to feature scaling.

On the other hand, the Naïve Bayes model exhibited persistently poor performance, with a recall of 0.00 and accuracy of 0.91, indicating a consistent tendency to misclassify the positive class even after normalization. The Neural Network (NN) model performed strongly with normalized data as well, achieving a recall of 0.87 and accuracy of 0.98. This underlines the advantages that deep learning models can achieve when trained on standardized input features.

The results underscore the significance of normalization in enhancing model performance for certain algorithms, such as SVM, KNN, and Neural Networks, while models like Logistic Regression and Decision Trees remain relatively stable regardless of normalization. This emphasizes the importance of tailoring preprocessing techniques to the specific characteristics of each algorithm when developing predictive models.

VII. Significance

This project offers valuable contributions to the financial services industry by promoting data-driven decision-making through the implementation of advanced machine learning techniques such as SVM, KNN, Naive Bayes, Decision Tree, Logistic Regression. Its significance can be summarized through the following key points:

Enhanced Decision-Making: The project introduces a reliable predictive framework using the CART algorithm to assess personal loan applications. This helps financial ınstitutions reduce the risk of loan default and streamline approval processes by enabling more accurate and informed decisions.

Improved Customer Experience: By accurately forecasting loan approvals, the model accelerates the loan application process. Utilizing demographic and financial data allows banks to offer personalized lending solutions, ultimately enhancing customer satisfaction.

Increased Operational Efficiency: Integrating predictive models into the loan approval workflow significantly reduces manual effort, shortens processing times, and cuts operational costs. Automation based on data insights contributes to greater efficiency and scalability within banking operations.

VIII. Challenges

Despite the promising outcomes of our model, the project encountered several practical challenges:

Limited Feature Diversity: The dataset provided a relatively narrow set of features, limiting the model’s ability to capture complex relationships between applicant attributes and loan approval outcomes. This constrained the scope for meaningful feature engineering.

Missing and Inconsistent Values: Several records in the dataset included missing or inconsistent values, particularly in fields like income, credit history, and employment status. This required thorough data cleaning and imputation strategies to maintain model accuracy.

Overfitting During Initial Model Training: Early iterations of the CART model showed signs of overfitting, performing well on training data but poorly on test data. We addressed this by pruning the decision tree and tuning hyperparameters to improve generalization.

IX. Conclusion

This project aims to forecast the likelihood that customers will purchase personal loans by employing five different machine learning models, both with and without data normalization. The performance of each model is evaluated based on two key metrics: recall and accuracy, which help in predicting customer behavior. The findings emphasize the importance of selecting the right model to achieve optimal predictive performance in financial decision-making.

Upon comprehensive evaluation of all models applied in the personal loan prediction analysis, it is evident that they exhibit a high level of accuracy. Accuracy is defined as the ratio of the total number of correct predictions—comprising both True Positives and True Negatives—to the total number of predictions made. While this metric provides a general overview of model effectiveness, our primary goal is to accurately identify customers likely to accept a personal loan offer. Therefore, a strong emphasis is placed on predicting positive cases correctly.

To enhance this effort, we prioritize the recall rate, also known as sensitivity, as a critical performance indicator for the five predictive models. Recall measures the proportion of actual positive cases identified accurately by the model, which is essential for effectively recognizing customers interested in purchasing a personal loan and optimizing outreach strategies. By focusing on the recall rate, we can better assess how well each model captures the attention of potential borrowers, ensuring that marketing efforts are directed toward the most appropriate audience.

Among the models tested, the Decision Tree model showcased the best overall performance, achieving an accuracy of 99% and a recall rate of 91%. This positions it as the most reliable choice for targeting customers likely to accept personal loan offers. Following closely are the Neural Network, which achieved a recall of 87% and an accuracy of 98%, and the Support Vector Machine (SVM), with a recall of 81% and an accuracy of 98%, particularly when standardized data was utilized.

The process of normalization significantly enhanced the performance of both the SVM and Logistic Regression models. Specifically, the SVM with an RBF kernel improved its accuracy from 96% to 98% post-normalization. Although Logistic Regression also displayed performance improvements, these were more modest in comparison. It is noteworthy that normalization did not have a significant impact on the K-Nearest Neighbors (KNN) model; its recall remained comparatively lower, with only minor fluctuations in accuracy.

According to the logic employed by the Decision Tree model, customers with an income exceeding 110.5 are likely to be classified based on their experience rather than their age—especially if their education level is 1.5 or lower and their income surpasses 126.5. In such cases, the model exhibits high confidence in classifying them according to experience. On the other hand, if a customer's income is 110.5 or lower and their credit card average (CCAvg) is below 2.95, they are usually categorized based on age. When the income further decreases below 106.5, the model is completely certain that they should be classified by age. Overall, income and CCAvg are identified as critical factors in the classification decisions made by the model. These insights can aid financial institutions in gaining a deeper understanding of customer profiles, thereby enabling the design of more targeted marketing strategies and personalized loan offerings.

Overall, income and CCAvg play a big role in deciding whether a customer is sorted by age or experience. This insight can help businesses better understand customer profiles and make more targeted offers or services.

X. References

[1] K. Walke, Bank_Personal_Loan_Modelling. (2020, April 17). Kaggle. https://www.kaggle.com/datasets/krantiswalke/bank-personal-loan-modelling/data [Accessed April 16, 2025]

[2] Arunmohan. (2021, June 22). Pruning decision trees - tutorial. Kaggle. https://www.kaggle.com/code/arunmohan003/pruning-decision-trees-tutorial [Accessed April 16, 2025]

[3] Li, Y. (2024). Analysis of USA national home prices based on different machine learning models. In Advances in economics, business and management research/Advances in Economics, Business and Management Research (pp. 100–109). https://doi.org/10.2991/978-94-6463-459-4_13 [Accessed April 16, 2025]

[4] Fiori, L. (2021, December 14). Distance metrics and K-Nearest Neighbor (KNN) - Luigi Fiori - Medium. Medium. https://medium.com/@luigi.fiori.lf0303/distance-metrics-and-k-nearest-neighbor-knn-1b840969c0f4 [Accessed April 16, 2025]

[5] Itoo, F., Meenakshi, & Singh, S. (2021). Comparison and analysis of logistic regression, Naïve Bayes and KNN machine learning algorithms for credit card fraud detection. International Journal of Information Technology, 13(4), 1503-1511.

[6] C. Madan, V. Sinap, C. Hongsanukulsant, D. V. Quach, T. Yangzom, T. T. H. Nguyen, and W. Shan, “A Comparative

Study of Loan Approval Prediction Using Machine Learning Methods," *ResearchGate*, 2023

[7] S. R. Shinde and A. S. Aphale, "Predict Loan Approval in Banking System Machine Learning Approach for Cooperative Banks Loan Approval," *International Journal of Engineering Research & Technology (IJERT)*, vol. 9, no. 5, pp. 1266–1270, 2020. [Online]. Available: https://www.ijert.org/predict-loan-approval-in-banking-system-machine-learning-approach-for-cooperative-banks-loan-approve