Loan Default Prediction Using Machine Learning

Zachary Lowe

*Abstract*— This project investigates the application of machine learning techniques to predict loan defaults using LendingClub’s real-world financial data. Three classification models—Logistic Regression, Support Vector Machine with RBF kernel, and Multi-Layer Perceptron—are developed and evaluated. The dataset is preprocessed through cleaning, encoding, and scaling to enhance model performance. Evaluation metrics include accuracy, F1 score, and ROC AUC to address class imbalance and predictive reliability. Results indicate that Logistic Regression achieves the best balance of accuracy and interpretability, making it suitable for practical credit risk assessment.

Keywords— Loan default prediction, machine learning, classification, credit risk, logistic regression, support vector machine, neural networks, data preprocessing, imbalanced datasets, ROC AUC.

# Introduction

Loan default prediction is a critical aspect of financial risk management that significantly impacts the profitability and sustainability of lending institutions. Accurately identifying borrowers who are likely to default allows lenders to optimize credit approval processes, minimize financial losses, and comply with regulatory requirements. Traditional credit scoring approaches, which often rely on heuristic rules and limited financial indicators, may fail to capture complex borrower behaviors and subtle risk factors.

Recent advancements in machine learning provide powerful tools for analyzing large, high-dimensional datasets to extract meaningful patterns that can enhance credit risk assessment. Machine learning models can learn non-linear relationships and interactions between borrower characteristics and loan outcomes, offering improved predictive accuracy over conventional methods.

This project investigates the application of supervised machine learning algorithms to predict loan defaults using historical loan data from LendingClub, a peer-to-peer lending platform. Specifically, we evaluate three classification models: Logistic Regression, Support Vector Machine (SVM) with a radial basis function (RBF) kernel, and a Multi-Layer Perceptron (MLP) neural network. These models were selected based on their complementary strengths and coverage in the course curriculum.

The main objective is to compare these models in terms of their predictive performance, handling of class imbalance, computational efficiency, and interpretability. The project emphasizes the importance of using appropriate evaluation metrics—such as F1 score and ROC AUC—to account for the imbalanced nature of loan default data.

# Dataset overview

## Data Source

The dataset utilized in this project is sourced from LendingClub, a leading peer-to-peer lending platform that connects borrowers with individual and institutional investors. LendingClub has pioneered the democratization of credit by providing detailed, publicly accessible loan data. The version used in this study was obtained from the Kaggle platform, where it is maintained as a large, curated repository for research and benchmarking.

The dataset spans loan applications submitted from 2007 through 2018, capturing a wide range of loan and borrower attributes along with the eventual loan outcomes. With over two million records and more than 150 columns, the dataset presents a comprehensive view of the lending process, encompassing borrower demographics, financial status, loan characteristics, and repayment history.

## Feature Description

Features in the dataset fall into several broad categories. Borrower demographic information includes employment length, home ownership status, and annual income. Financial indicators include credit scores, debt-to-income ratios, revolving credit balances, and delinquency counts. Loan-specific features such as loan amount, interest rate, loan term, and purpose provide context on the credit extended.

The dataset also contains LendingClub-specific metrics like internal credit grades and subgrades, which reflect proprietary risk assessments. Together, these features offer a rich and heterogeneous data environment conducive to machine learning applications.

## Target Variable

The target variable for this study was derived from the loan\_status field, which indicates the repayment status of each loan. For the binary classification task, loans labeled “Fully Paid” were categorized as non-default (0), while loans labeled “Charged Off” or “Default” were categorized as default (1). Other statuses—such as “Current,” “Late,” or “In Grace Period”—were excluded from the modeling dataset to ensure that only loans with definitive outcomes were included. This approach avoids introducing noise or ambiguity into the target labels.

## Data Quality & Missing Values

An initial assessment of the dataset revealed numerous columns with significant proportions of missing data, as well as features with inconsistent or redundant information. Columns with more than 50% missing values were dropped to maintain data integrity and reduce noise. The remaining features underwent further cleaning to remove irrelevant identifiers, textual fields, and variables unavailable at loan issuance to prevent data leakage.

## Sampling Strategy

Given the extensive size of the full dataset and computational resource limitations, it was necessary to reduce the dataset to a manageable size for model training. A stratified random sampling method was employed to select approximately 100,000 records, ensuring that the proportion of default and non-default loans remained consistent with the full dataset. Stratification preserved class distribution and enabled robust model training while addressing the constraints of available memory and processing power.

The sampled dataset retained sufficient diversity and complexity, representing a wide range of borrower profiles and loan types, to support meaningful machine learning model development and evaluation.

# Data Preprocessing

## Handling Missing Values

The LendingClub dataset contained many features with missing or incomplete entries, which can negatively affect machine learning model training and performance. To address this, all columns with more than 50% missing values were removed from the dataset. This step ensured that only features with sufficient data coverage were retained, reducing noise and potential biases introduced by imputation on sparse data.

For the remaining features, rows containing missing values were dropped during the final data cleaning stage. Although this reduced the total number of samples, it guaranteed that the models would be trained on complete data without requiring complex imputation strategies that might introduce inaccuracies.

## Feature Selection and Removal ofIirrelevant Colums

Several columns were excluded based on their irrelevance, redundancy, or potential to leak future information into the model. For example, identifiers such as id and member\_id do not provide predictive value and were removed. Additionally, textual fields such as emp\_title and url were excluded due to high cardinality and lack of structured information. Features that reflect loan outcomes or collections occurring after loan origination were also removed to prevent information leakage that could artificially inflate model performance.

This step focused the dataset on features that are available at loan issuance and that provide meaningful predictive signals.

## Encoding Categorial Variables

Many features in the dataset were categorical, including loan grades, home ownership status, and loan purposes. These variables were transformed using one-hot encoding, converting each category into a binary indicator variable. One-hot encoding was chosen for its simplicity and compatibility with all chosen machine learning models. It avoids imposing ordinal relationships on nominal variables, preserving the integrity of categorical information.

This encoding significantly increased the feature space, raising the number of columns from around 100 to approximately 150.

## Feature Scaling

Continuous numerical features were standardized using Scikit-learn’s StandardScaler, which rescales variables to have zero mean and unit variance. Scaling was critical for models sensitive to feature magnitude and distance metrics, such as Support Vector Machines and Neural Networks. It also improved training stability and convergence speed for gradient-based optimization used in the MLP model.

## Cleaning & Synchronization

After encoding and scaling, any residual missing values were identified and removed. The final dataset contained only complete cases, ensuring compatibility with all machine learning algorithms used. The target variable was isolated from the features, resulting in a clean input matrix and corresponding output vector for modeling.

## Train-Test Split

The cleaned and transformed dataset was split into training and testing sets using an 80/20 ratio. Stratification was applied during splitting to preserve the proportion of defaulted and non-defaulted loans in both subsets. This ensured that model evaluation would reflect realistic class distributions and prevented bias toward the majority class. The final feature matrix contained approximately 150 columns and 100,000 rows, ready for input into classification models.

# Methodology & Model Formulation

## Problem Framing & Objective

This project addresses loan default prediction as a supervised binary classification problem. The input consists of borrower and loan characteristics, and the output is a binary label indicating whether the loan was fully paid (0) or defaulted (1). Predicting defaults accurately is vital for financial institutions to manage credit risk, allocate capital effectively, and minimize losses.

The dataset’s inherent class imbalance—where non-defaulted loans greatly outnumber defaults—adds complexity to the problem. Careful consideration was given to selecting models and training strategies that mitigate bias toward the majority class, ensuring reliable identification of risky loans.

## Models Selected

Three models were chosen for their complementary strengths and practical relevance:

1. **Logistic Regression**: A simple, interpretable linear model widely used in credit scoring. Its probabilistic output and coefficient interpretability make it a strong baseline and a valuable tool for regulatory compliance and decision support.
2. **Support Vector Machine (SVM) with RBF Kernel**: SVMs excel at classification with clear margin maximization principles. The RBF kernel allows the model to capture complex, nonlinear patterns by projecting data into higher-dimensional space, potentially improving classification accuracy on nuanced financial data.
3. **Multi-Layer Perceptron (MLP) Neural Network**: Neural networks provide flexible architectures capable of modeling complex feature interactions. The MLP used in this project consists of a single hidden layer with 100 neurons, utilizing ReLU activation to introduce nonlinearity. It is well-suited for capturing nonlinear dependencies common in borrower behavior and loan performance.

## Model Training Details

## All models were implemented using the Scikit-learn Python library, ensuring consistency and reproducibility. Logistic Regression was configured with an increased iteration limit (max\_iter=1000) to guarantee convergence. The SVM model employed default regularization parameters with the RBF kernel, and probability estimates were enabled to facilitate ROC curve generation. The MLP was trained with the Adam optimizer, early stopping, and L2 regularization to prevent overfitting, using a maximum of 300 training iterations.

## Addressing Class Imbalance

Given the dataset’s skewed class distribution, all models incorporated class weighting (via class\_weight='balanced') to adjust the loss function and penalize misclassification of the minority default class more heavily. This strategy helps improve recall on defaults, reducing the risk of false negatives that could have severe financial consequences.

## Training & Validation

## The dataset was split into training and testing sets with an 80/20 stratified split to maintain class proportions. Models were trained exclusively on the training data and evaluated on the hold-out test set. No cross-validation was performed due to resource constraints, but stratification ensured representative performance metrics.

## Performance Metrics

Model performance was assessed using accuracy, F1 score, and ROC AUC on the test set. These metrics were chosen to balance the need for overall correctness and sensitivity to the minority default class, providing a comprehensive view of each model’s effectiveness.

# Evaluation Metrics & Justification

Evaluating the effectiveness of classification models in the context of loan default prediction requires selecting metrics that capture both overall accuracy and the ability to correctly identify minority class instances—defaults. This is especially important given the significant class imbalance in the dataset, where non-defaulted loans far outnumber defaulted ones.

Accuracy is a commonly used metric that reflects the proportion of correctly predicted loans, combining both defaults and non-defaults. While accuracy is straightforward and intuitive, it can be misleading in imbalanced datasets, as a model that predicts all loans as non-default could achieve high accuracy despite failing to identify any actual defaults.

To address this limitation, the F1 score was employed, which balances two critical aspects: precision and recall. Precision measures the proportion of predicted defaults that are true defaults, reflecting the model’s ability to avoid false alarms. Recall measures the proportion of actual defaults correctly identified, highlighting the model’s ability to catch risky loans. The F1 score combines these into a single value, emphasizing performance on the minority class and providing a more realistic assessment of model effectiveness in risk management.

Additionally, the Receiver Operating Characteristic Area Under the Curve (ROC AUC) was used to evaluate the models’ discrimination ability across various decision thresholds. The ROC AUC measures how well a model ranks defaulted loans higher in risk compared to non-defaulted loans, independent of any fixed classification threshold. This metric is robust to class imbalance and provides a comprehensive view of model performance beyond fixed point estimates.

To complement these metrics, confusion matrices were analyzed to visualize the trade-offs between true and false positives and negatives, aiding in detailed error analysis. These insights helped identify how each model balanced correctly flagging defaults against minimizing false alarms, a critical consideration in financial applications.

Overall, the combined use of these metrics ensured a thorough and balanced evaluation of model performance, with particular attention to detecting defaulted loans accurately while controlling for false positives.

# Experimental Results

The three machine learning models—Logistic Regression, Support Vector Machine (SVM) with RBF kernel, and Multi-Layer Perceptron (MLP)—were trained and evaluated on the preprocessed LendingClub dataset. The evaluation metrics described in the previous section were computed on the held-out test set, which comprised 20% of the sampled data, stratified to preserve class distribution.

## Performance Comparison

Table I summarizes the key performance metrics for each model. Logistic Regression achieved the highest accuracy of 99%, with an F1 score of 0.99 on the default class and an ROC AUC of 0.998. The MLP closely followed, with an accuracy of 99%, F1 score of 0.98, and ROC AUC of 0.997. The SVM model demonstrated slightly lower performance, with 97% accuracy, 0.93 F1 score, and ROC AUC of 0.998.

1. Model performance on test set

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Accuracy** | **F1 Score** | **ROC AUC** |
| Logistic Regression | 0.99 | 0.99 | 0.998 |
| SVM (RBF Kernel) | 0.97 | 0.93 | 0.998 |
| MLP Neural Network | 0.99 | 0.98 | 0.997 |

## Confusion Matrix Analysis

Confusion matrices provided deeper insights into each model’s classification tendencies. Logistic Regression exhibited the fewest misclassifications, with very low false negative and false positive rates, indicating robust identification of defaulted loans. The MLP model had slightly higher false negatives but maintained strong precision. The SVM model, while maintaining a good ROC AUC, had increased false negatives, indicating some defaults were missed, which could have financial consequences in practical applications.

This

## ROC Curve Observations

ROC curves (see Figure 1) show that all models perform well at distinguishing defaults from non-defaults. Logistic Regression’s curve was marginally steeper, indicating better discrimination at most thresholds. The MLP’s curve closely overlapped with Logistic Regression, suggesting comparable ranking ability. The SVM curve, while high, was slightly lower in true positive rate at specific false positive rates.

This

## Computational Efficiency

Training time varied significantly among models. Logistic Regression completed training in under a minute, demonstrating high efficiency and scalability. The MLP required longer training time due to backpropagation computations, but it remained practical for the dataset size. The SVM was the most computationally expensive, with training times exceeding several minutes, which may limit its use in real-time or large-scale environments.

## Summary

Overall, Logistic Regression provided the best balance of accuracy, interpretability, and computational efficiency, making it the most practical choice for loan default prediction in this setting. The MLP offered promising performance with more flexible modeling capacity, suitable for further exploration with tuning and larger datasets. The SVM, while conceptually powerful, was less competitive due to longer training times and lower recall.

# Interpretation & Analysis

The comparative results from the three models provide several insights into both their predictive capabilities and practical suitability for loan default prediction. While all models achieved strong overall performance, there are key differences in how each performed under the constraints and objectives of this project.

Logistic Regression stood out as the most balanced model across all evaluation criteria. Despite its linear nature, it performed exceptionally well with an accuracy of 99% and an F1 score of 0.99. Its ability to consistently and accurately identify default cases while maintaining a low false positive rate makes it highly effective in this financial context. One notable advantage of this model is its interpretability; coefficients can be examined to understand the influence of each feature on the likelihood of default. This transparency is valuable in finance, where decisions often require regulatory justification and model clarity.

The MLP Neural Network also performed at a high level, with metrics closely trailing those of Logistic Regression. Its architecture allowed it to capture more complex relationships between features, resulting in an F1 score of 0.98. However, this came at the cost of increased training time and reduced interpretability. While the MLP model may generalize well on larger or more complex datasets, the additional computational cost and black-box nature may make it less suitable for institutions that require explainable models for auditing and compliance.

The SVM with RBF kernel showed relatively lower performance in terms of F1 score and had the highest number of false negatives among the three models. Although its ROC AUC remained high, indicating strong ranking ability, its recall was comparatively lower, suggesting that more defaulted loans were missed. In a lending context, this is particularly problematic, as failing to identify risky borrowers can have significant financial implications. Additionally, the SVM model required the longest training time, which may not scale well for larger datasets or real-time deployment.

From a business perspective, the cost of false negatives (approving a risky loan) generally outweighs the cost of false positives (rejecting a low-risk applicant). Therefore, models that offer high recall and F1 scores—such as Logistic Regression and MLP—are more aligned with real-world risk management priorities. However, the trade-off between interpretability and complexity remains a critical consideration. While neural networks offer higher modeling capacity, their lack of transparency can limit their usability in regulated environments.

Ultimately, the analysis suggests that Logistic Regression is the most appropriate choice for this particular task, balancing prediction quality, speed, and interpretability. The MLP model is a viable alternative, especially in scenarios where marginal performance gains justify additional computational cost and complexity. The SVM, while theoretically appealing, was less practical given the size of the dataset and the nature of the problem.

# Challenges & Limitations

Although the project successfully developed and evaluated several machine learning models for loan default prediction, a number of challenges and limitations were encountered throughout the process. These affected model design, performance evaluation, and overall project scope.

## Dataset Size & Resource Constraints

## The original LendingClub dataset contained over two million records and 150+ features. However, due to memory limitations within the Google Colab environment, it was not feasible to process or train models on the full dataset. To mitigate this, a stratified random sample of 100,000 records was selected. While this reduced training time and prevented system crashes, it may have limited the generalizability of the findings. Certain patterns that only appear in less frequent segments of the data may not be adequately captured in the sample.

## Imbalanced Classes

The target variable in the dataset was heavily imbalanced, with the majority of loans labeled as non-default. This created a risk that models would become biased toward the majority class, reducing sensitivity to defaults—the minority class of interest. Although class weighting and careful metric selection were used to address this issue, imbalance remains a structural limitation of the dataset. Oversampling, undersampling, or synthetic techniques like SMOTE could be explored in future work to further mitigate this.

## Limited Feature Engineering

This project relied primarily on the raw features provided in the dataset, with minimal domain-specific feature engineering. Certain engineered features—such as credit utilization ratios, time-based trends, or interaction terms—could have improved model performance. Additionally, temporal dynamics were not incorporated; all loans were treated as static entries, even though macroeconomic conditions and borrower behavior may change over time.

## Model Tuning & Exploration

## Due to time and resource constraints, only basic hyperparameter tuning was performed. Models were run using primarily default settings, with slight adjustments to iteration limits and activation functions. Further improvements might be achieved through grid search or randomized search across a broader parameter space, particularly for the MLP and SVM models. Moreover, ensemble approaches like Random Forest or Gradient Boosting were not explored, even though they are known to perform well in similar financial prediction tasks.

## Interpretability vs. Complexity

A trade-off was observed between model interpretability and performance. While Logistic Regression offered transparent, explainable coefficients, models like MLP and SVM introduced more complexity and made interpretation difficult. In high-stakes financial applications, where decisions must often be audited or justified, interpretability can be just as important as raw predictive power.

## Tools & Platform Limitations

Working within a cloud-based notebook environment like Google Colab offered accessibility but also introduced limitations in runtime duration, memory usage, and parallel processing capabilities. Training time for the SVM model, in particular, was significantly affected by these constraints. This restricted the extent to which deeper architectures or more computationally demanding algorithms could be explored.

# Challenges & Limitations

This project set out to evaluate the effectiveness of three supervised machine learning models—Logistic Regression, Support Vector Machine (SVM) with an RBF kernel, and Multi-Layer Perceptron (MLP)—in predicting loan default risk using real-world data from LendingClub. The goal was to determine which model offered the best combination of predictive performance, computational efficiency, and practical applicability in a highly imbalanced classification setting.

The dataset, consisting of over 2 million records, was cleaned, preprocessed, and sampled down to 100,000 representative examples due to memory constraints. Significant effort was devoted to preparing the data: dropping columns with high proportions of missing values, removing leakage-prone fields, encoding categorical variables, and scaling numerical ones. After preprocessing, the dataset included approximately 150 features, offering a rich and diverse set of inputs for training the models.

Each model was trained using stratified sampling to preserve class distribution, and all models were evaluated using accuracy, F1 score, ROC AUC, and confusion matrices. These metrics provided a balanced view of both overall performance and the ability to identify the minority class—defaulted loans.

Logistic Regression delivered the strongest and most balanced performance overall. It achieved a 99% accuracy, a 0.99 F1 score for the default class, and the highest ROC AUC score of 0.998. It required minimal training time and provided interpretable coefficients that allow institutions to understand how specific features affect the likelihood of default. Its linear nature did not limit its effectiveness, as the cleaned and transformed features provided enough signal to allow strong separation between defaulted and non-defaulted loans.

The MLP neural network also demonstrated high accuracy and an impressive F1 score of 0.98, suggesting it was also highly effective at identifying defaults. However, the added computational complexity and reduced interpretability of the model may present challenges in real-world applications, particularly in regulated industries where understanding and explaining decisions is critical. Still, its ability to learn non-linear relationships makes it a viable option in environments where performance outweighs clarity.

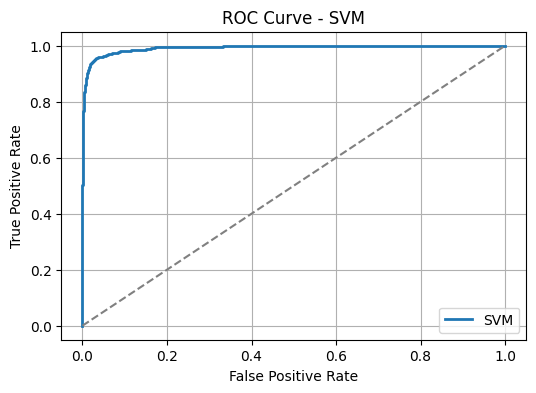
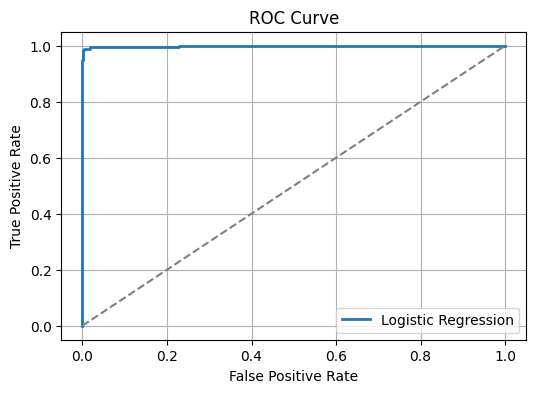
The SVM model, while achieving strong ROC AUC and reasonable accuracy, fell short in identifying defaulted loans, reflected in its lower F1 score. Additionally, it required significantly more training time than the other models, which makes it less scalable or practical for very large datasets. The relatively lower recall suggested that it was less reliable in detecting risky loans, which is a major consideration in financial applications.

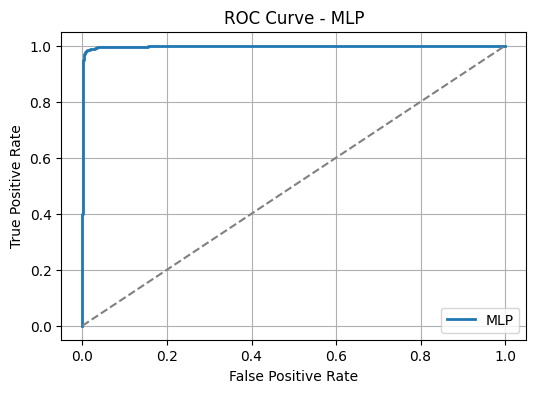
Overall, the results of this project affirm that even relatively simple models like Logistic Regression can perform exceptionally well in financial classification tasks when combined with thorough preprocessing and thoughtful metric selection. The success of this model also underscores the importance of interpretability and efficiency in practical deployment scenarios. While more complex models like MLP offer strong performance, they must be weighed against the cost of reduced transparency. SVM, despite its theoretical strengths, proved less practical for this dataset and problem context.

The work completed here demonstrates that machine learning can meaningfully support loan risk analysis. A strong foundation in data cleaning, balanced evaluation, and model interpretability proved to be just as important as algorithm selection itself. These findings reinforce the value of combining domain knowledge with machine learning to deliver actionable, trustworthy solutions in financial risk management.

# Appendix

1. ROC Curvres for Logistic Regression, SVM, and MLP





##### References

1. LendingClub, “Loan Data on Lending Club,” Kaggle, 2023. [Online]. Available: https://www.kaggle.com/datasets/wordsforthewise/lending-club
2. F. Pedregosa et al., “Scikit-learn: Machine Learning in Python,” Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.
3. J. Brownlee, Imbalanced Classification with Python: Better Metrics, Balance Skewed Classes, and Improve Model Performance, Machine Learning Mastery, 2020.
4. G. E. Batista, R. C. Prati, and M. C. Monard, “A study of the behavior of several methods for balancing machine learning training data,” SIGKDD Explorations, vol. 6, no. 1, pp. 20–29, 2004.
5. Scikit-learn documentation. [Online]. Available: https://scikit-learn.org/stable