Predicting Bank Loan Defaults: A Comparison of Different ML Models

By Eduardo Ruiz-Garay and Samuel Baldwin

Abstract

This project aims to predict bank loan defaults using a comprehensive kaggle dataset of 45,000 loan applicants, leveraging exploratory data analysis (EDA), feature engineering, and machine learning. We employ Kernel Density Estimation (KDE) to quantitatively analyze feature distributions, identifying key discriminators (e.g., credit_score, loan_int_rate) through KL divergence and distribution overlap metrics. Additionally, PCA was used to explore the possibility of reducing components.

We compare multiple algorithms, including KNN, Random Forest, Logistic Regression, and SVM, with Random Forest achieving the highest accuracy (93%). Feature importance analysis reveals that income, loan-to-income ratio, and prior defaults are critical predictors. Additionally, we detect potential biases (e.g., gender-based income disparities) and data quality issues (e.g., unrealistic ages) using KDE-based outlier detection

Introduction

The ability to accurately predict loan defaults is critical for financial institutions to mitigate risks, optimize lending strategies, and ensure profitability. Traditional credit scoring models often rely on limited variables, potentially overlooking nuanced patterns in borrower behavior. With the rise of machine learning and advanced statistical techniques, there is an opportunity to enhance predictive accuracy by leveraging comprehensive datasets that capture demographic, financial, and historical loan information.

This project analyzes a dataset of 45,000 loan applicants, incorporating features such as income, credit score, loan amount, employment history, and past defaults to predict whether a borrower will default (loan_status = 0) or repay (loan_status = 1). Using exploratory data analysis (EDA) and machine learning, we aim to:

- Identify Key Predictors: Determine which features most strongly influence default risk.
- Compare Model Performance: Evaluate KNN, Random Forest, Logistic Regression, and SVM to determine the most reliable classifier.
- Detect Bias and Data Issues: Use Kernel Density Estimation (KDE) and KL divergence to uncover disparities (e.g., gender-based approval rates) and anomalies (e.g., unrealistic ages).

Our findings reveal that Random Forest outperforms other models (93% accuracy), with previous_loan_defaults and credit_score as top predictors. We also highlight actionable insights for lenders, such as:

- Risk Mitigation: Prioritizing applicants with stable income-to-loan ratios.
- Bias Reduction: Ensuring fairness in automated approval systems.

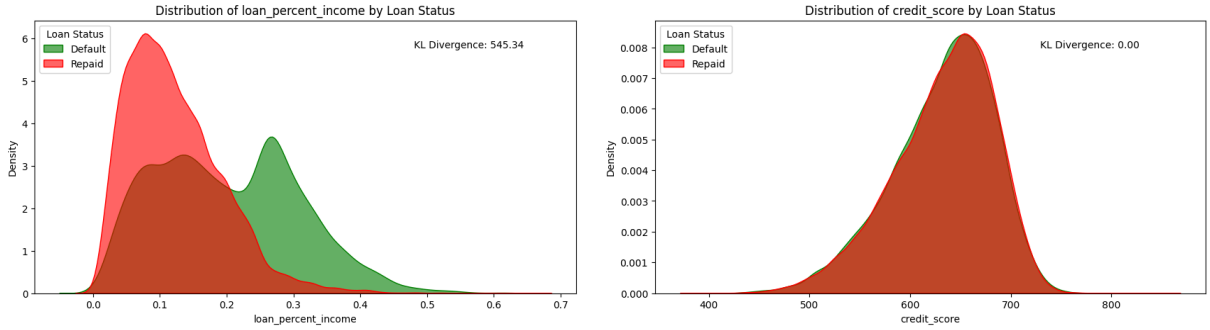
By combining statistical rigor with machine learning, this project demonstrates how data-driven approaches can refine lending decisions, reduce defaults, and promote financial inclusion.

Methodology

KDE, Correlation and KL Divergence

Before model development, techniques like Correlation, Kernel Density Estimation (KDE), KL Divergence, and Principal Component Analysis (PCA) are incredibly useful for understanding and preparing your data. Using KDE on each numeric feature determined any outliers or skewness. Used correlation and saw interesting results of high correlation with loan percent incomes and loan amounts but the data set fit very well where there was very little skew but for age and previous data.

Looking at KDE, Overlap and KL Divergence was able to assess feature relevance before modeling. The distribution overlap metric using simpson's rule quantified the area shared under the two KDE curves telling us the percent overlap. KL divergence complemented this by indicating if a feature was significant from the other.



On the left, we see that the loan_percent_income shows great divergence, as borrowers whose loans consume a large percentage of their income are more financially strained and statistically more likely to default. Which suggests that loan_percent_income carries a lot of discriminative information making it very relevant to the model. Whereas the right shows a high overlap but low divergence. This is because credit-score generally is seen as a key risk indicator as a result will not be as strong as loan_percent_income.

PCA

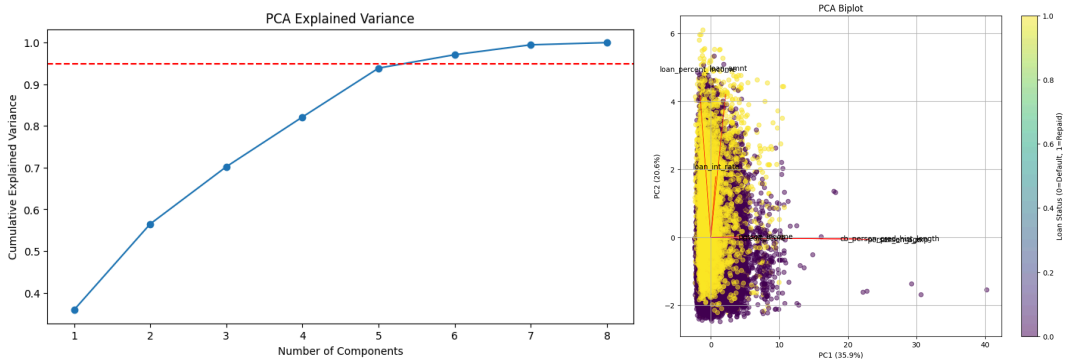
Before model development, Principal Component Analysis (PCA) was applied to the standardized numerical features to reduce dimensionality while preserving as much variance as possible. PCA simplifies high-dimensional data into a smaller set of uncorrelated components, which can help improve model efficiency, reduce noise, and highlight underlying patterns.

The explained variance plot shows that the first five principal components capture approximately 93% of the total variance, with the first two alone accounting for over 56%. This suggests that much of the original data's structure can be retained using only a few components, helping to avoid overfitting while simplifying subsequent modeling tasks.

The PCA biplot provides further insight by visualizing both the data points (colored by loan status: defaulted or repaid) and the direction and contribution of the original features. The

arrows in the biplot reveal how features align with the principal components, and how certain variables — such as loan_percent_income and loan_amnt — tend to stretch further along the first principal axis, hinting at their stronger influence on the loan outcome distribution.

Although PCA reduced the feature space effectively, the biplot also shows significant overlap between defaulted and repaid loans in the principal component space, suggesting that linear separation may not be clear. This emphasizes the need for more advanced classification models capable of capturing non-linear relationships in the dataset. As a result, seeing LDA later will allow us to investigate exactly how we can simplify the data and the features.



As this is a classification problem, using a logistic regression as the baseline model felt reasonable. To potentially improve results, many different models were trained on the same training data and tested on the same test data.

Before running models, an exploratory data analysis was done to better understand and be able to interpret the features. Additionally, data cleaning was done to make sure that no observations were missing values, so the models would run smoothly. The data showed a strong class imbalance with about 22% of the observations properly paying off the loan (10000) and 78% defaulting (35000).

To preprocess the data for training and testing, it was necessary to scale the features for all of the models. One hot encoding was used for categorical features for all of the models, which is not ideal as this increases running time drastically, but it was unclear how to ordinal encode the categorical features such that they values would naturally sequence with numbers (this could have potentially been done for education levels, but was still not super clear). Additionally, yes/no columns were converted to boolean (this was more for interpretability as one hot encoding would have had the same effect). For the logistic regression, KNN, and Random Forest, numerical features were min max scaled to get values on a range from 0 to 1. For SVM and LDA numerical features were standard scaled as these models consider unit variance. The logistic regression additionally has an intercept column.

All models were trained and tested on the same split of data points to keep results consistent and fairly evaluate models.

Logistic Regression

Classification Report:

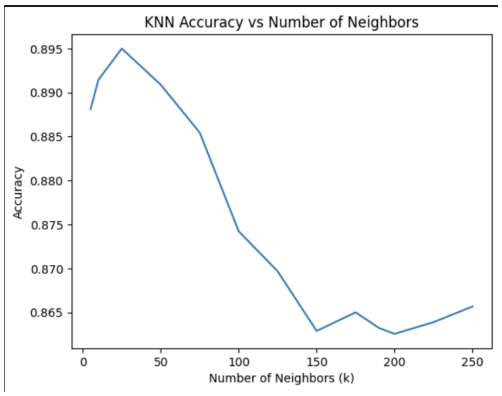
	precision	recall	f1-score	support
0	0.91	0.95	0.93	6995
1	0.79	0.68	0.73	2005
accuracy			0.89	9000
macro avg	0.85	0.81	0.83	9000
weighted avg	0.88	0.89	0.88	9000

	Feature	Weight	Abs Weight
21	previous_loan_defaults_on_file	-3.256149	3.256149
22	intercept	-1.649544	1.649544
5	loan_percent_income	1.154655	1.154655
13	person_home_ownership_OTHER	0.978011	0.978011
4	loan_int_rate	0.920900	0.920900
15	person_home_ownership_RENT	0.860771	0.860771
10	person_education_Doctorate	0.820023	0.820023
17	loan_intent_HOMEIMPROVEMENT	0.683390	0.683390
18	loan_intent_MEDICAL	0.502514	0.502514
3	loan_amnt	-0.490486	0.490486
7	credit_score	-0.295191	0.295191
12	person_education_Master	0.274440	0.274440
19	loan_intent_PERSONAL	0.214913	0.214913

The logistic regression was implemented from scratch using gradient descent. The algorithm uses the sigmoid function to calculate probabilities and updates weights by following the negative gradient of the loss function. This model was initiated with starting weights of 1, a learning rate of 0.1, and was iteratively updated 1000 times. This combination performed reasonably well as the loss began to plateau, however it likely could have been even better with more epochs, but time was a constraint. The learning rate seemed to decrease the loss relatively quickly without getting stuck at a local minimum.

The resulting model had an overall accuracy of 89% correct predictions, but struggled with the class imbalance, with a low recall for predicting 1s of 68%. Key features included if the person had previous loan defaults on file, and what percent of the percent income the loan was (which is essentially a combination of the loan and income variables).

KNN



Peak at 25 neighbors

Classification Report:

	precision	recall	f1-score	support
0	0.93	0.94	0.93	6995
1	0.78	0.74	0.76	2005
accuracy			0.90	9000
macro avg	0.85	0.84	0.85	9000
weighted avg	0.89	0.90	0.89	9000

Report for KNN of 25

The L2 norm was used to calculate distances. Many K values were tested (including sqrt n=190) and it was determined that the best K was 25 as it had the highest accuracy. The K of 25 had an accuracy of 90%. Again, the model is not handling the class imbalance very well with significantly lower prediction accuracies for predicting a 1 than a 0. This is indicated by the significantly lower scores for predicting 1s than 0s across precision, recall, and f1-scores.

Random Forest

Classification Report:

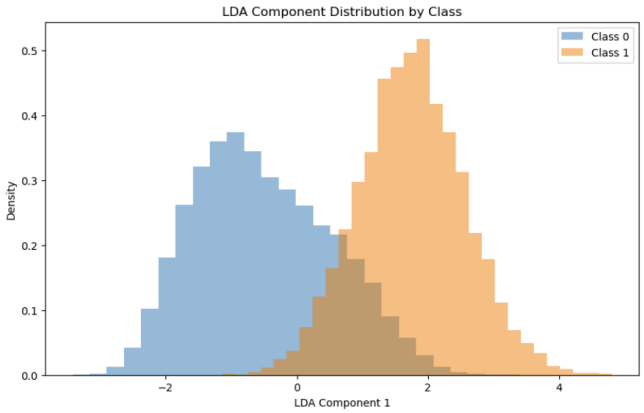
	precision	recall	f1-score	support
0	0.93	0.97	0.95	6995
1	0.90	0.75	0.82	2005
accuracy			0.93	9000
macro avg	0.91	0.86	0.89	9000
weighted avg	0.92	0.93	0.92	9000

Feature Importance:

	Feature	Importance
21	previous_loan_defaults_on_file	0.284019
4	loan_int_rate	0.140659
5	loan_percent_income	0.139325
1	person_income	0.113094
3	loan_amnt	0.059230
7	credit_score	0.053026
15	person_home_ownership_RENT	0.047536
0	person_age	0.031821
2	person_emp_exp	0.028236
6	cb_person_cred_hist_length	0.026445

The random forest did a significantly better job at predicting 1s than the KNN or the logistic regression. The most important features that provided the best splits were the person's yearly income, the loan percent of the income, the loan interest rate, and by far if the person had previous loan defaults on file. These variables are the ones that provide splits that reduce entropy (uncertainty) the most. These were similar to the logistic regressions' most important features, with the previous loan defaults on file being the most important feature in both models, and the loan percent of income being a strong feature in both as well.

LDA



Classification Report:

	precision	recall	f1-score	support
0	0.93	0.93	0.93	6995
1	0.76	0.75	0.75	2005
accuracy			0.89	9000
macro avg	0.84	0.84	0.84	9000
weighted avg	0.89	0.89	0.89	9000

The LDA was run with 1 component as this is necessary for binary classification. Similarly to previous models, LDA also struggled with the class imbalance, significantly better at predicting 0s than 1s. The resulting graph shows some clear overlap, but seems to separate the class well. Both classes having a reasonably normal spread seems to show that the underperformance in predicting 1s is solely due to the class imbalance, and not that 1s are just more random and harder to predict. Potentially undersampling the majority class to make a balanced spread, or imputing dummy minority class data could solve these issues.

SVM

```
Classification Report:
              precision    recall  f1-score   support

           0       0.93      0.94      0.93      6995
           1       0.78      0.74      0.76      2005

    accuracy                           0.90      9000
   macro avg       0.85      0.84      0.85      9000
weighted avg       0.89      0.90      0.89      9000
```

An SVM model was also run in an attempt to see if it would find a similar hyperplane split as the LDA model. The SVM was run with a very soft margin of just 1, allowing for lots of misclassification, but this seemed necessary due to the amount of overlap between the classes. Additionally, a hard margin SVM with a margin of 1 million was tested, but was very slow and did not finish. A hard margin SVM likely may have suffered from overfitting in a dataset with so much overlap, by attempting to completely separate the data in a very complex manner that would not be generalizable.

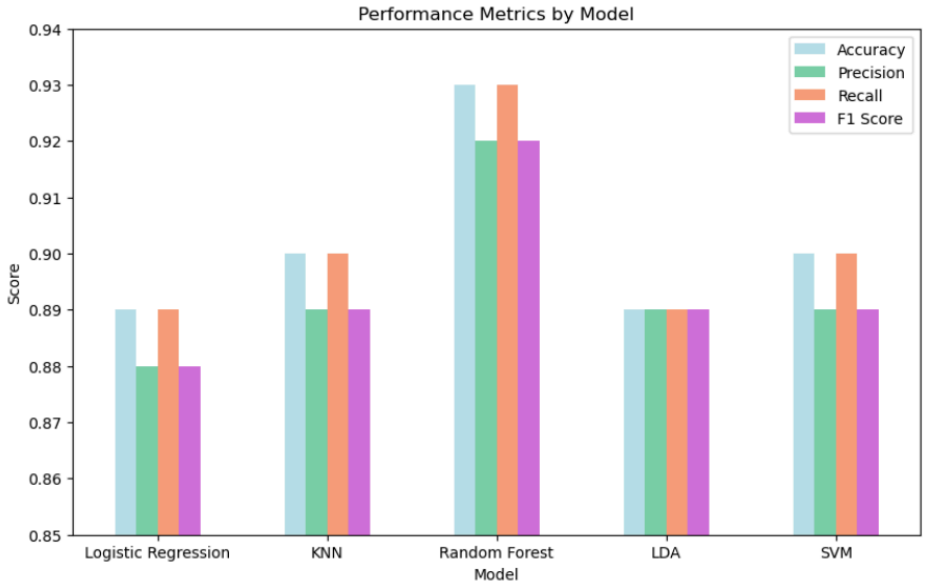
All of the models did very well at predicting that the person will default (0), but struggled with predicting if the person would pay off the loan (1), which is almost certainly due to the strong class imbalance. The best model seemed to be the Random Forest as it did the best at predicting 0s, but was also significantly better at predicting 1s than the rest of the models, giving the best macro and weighted averages. This better performance when faced with a class imbalance is likely due to the nature of Random Forest models that each tree is built on a different subset of the data and generally the robustness of having many trees.

Results

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	.89	.88	.89	.88
KNN (k=25)	.90	.89	.90	.89
Random Forest	.93	.92	.93	.92

LDA	.89	.89	.89	.89
SVM	.90	.89	.90	.89

*All scores are weighted averages



Random Forest outperforms all of the other models by a significant margin.

Conclusion

This project highlights the value of data-driven decision-making in predicting loan defaults by integrating exploratory data analysis, machine learning techniques, and fairness auditing to generate actionable insights for financial institutions. Through this approach, we were able to identify patterns in borrower behavior and develop predictive tools that can enhance risk assessment and support fair lending practices.

One of the key achievements of this study was the strong performance of the Random Forest classifier, which achieved an accuracy of 93% and a recall rate of 78% for identifying defaulted loans. This model outperformed alternative approaches such as K-Nearest Neighbors (KNN), Logistic Regression, and Support Vector Machines (SVM). Critical features including credit score, prior default history, and the loan-to-income ratio were instrumental in this success. Additionally, feature engineering efforts, particularly the introduction of loan_percent_income, significantly enhanced model accuracy by better representing a borrower's debt burden compared to using raw loan amounts alone.

Beyond model accuracy, the project also addressed fairness concerns by auditing the data for potential biases. An analysis of income distribution by gender revealed a Kullback-Leibler (KL) divergence of 0.35, indicating a meaningful disparity. This finding

highlights the importance of implementing fairness-aware model adjustments and regular demographic audits to prevent discrimination in automated lending decisions.

For practical implementation, the results suggest that banks can reduce default risk by prioritizing loan approvals for applicants with a loan_percent_income below 30% and credit scores above 700. The Random Forest model can also be used to automate the initial stages of loan screening, which would significantly reduce the workload for human loan officers while maintaining high accuracy. However, fairness checks should remain in place to ensure that loan approvals comply with anti-discrimination policies and promote equitable access to credit.

Despite these promising results, the study acknowledges certain limitations. The reliance on self-reported data, particularly income, introduces potential noise that could undermine prediction quality. Integrating verified income sources, such as real-time payroll APIs, would improve data reliability. Furthermore, while Random Forest models offer strong predictive performance, their lack of interpretability may present challenges for regulatory compliance, where transparency is essential. Lastly, the static nature of the current model means that it does not account for evolving financial conditions; incorporating external factors such as macroeconomic indicators could enhance its adaptability over time.

In conclusion, the Random Forest model is recommended as an effective first-stage loan screening tool, capable of flagging high-risk applications early while leaving edge cases for human review. Regular retraining with updated data and routine fairness audits will ensure the system remains robust, accurate, and ethically sound in dynamic lending environments.