PulseCheck
A Streamlit App for Instant, Explainable Credit Risk Scoring

Lina El-Husseini

Meghna Ganesh Kumar

Murughanandam Sivasubramanian

Pravin Kumar Rajak

Sarfraz Nawaz

MS_DSP 422 Practical Machine Learning

Northwestern University

Saturday, 30 August 2025

Executive Summary **2**
1. Introduction **4**
2. Literature Survey **5**
2.1 Traditional Approaches 5
2.2 Emergence of Machine Learning 5
2.3 Explainability and Regulatory Compliance 6
2.4 Real-Time Credit Scoring Systems 6
3. Research Objective(s) **7**
4. Methodology and Tooling **7**
The Dataset **7**
Tooling 8
5. Exploratory Data Analysis (EDA) **9**
Data Analysis, Visualizations, Data Mining Techniques 9
Clarity in Underlying Assumptions, Unit of Analysis, Feature Interactions 9
6. Data Preparation & Feature Engineering **10**
Handling of Features, Feature Extraction and Engineering 10
Variable Transformations, Data Scaling, Assumptions, and Tests 10
7. Models & Training **11**
Hyperparameter Tuning 14
Model Performance 14
Model Evaluation 15
8. Lessons Learned and Additional Methods **17**
Addressing Fairness and Algorithmic Bias 18
9. References **18**

Executive Summary

The credit scoring process is evolving as digital lending expands and regulators demand faster, more transparent decisions. Traditional models, such as logistic regression, though simple and interpretable, cannot capture the complex, non-linear patterns in borrower behavior. This project develops a machine learning based credit risk scoring system with real time deployment through Streamlit, delivering instant predictions and interpretable explanations that improve customer experience, reduce risk, and ensure compliance.

Using a Kaggle dataset of 32,500 loan applications covering demographics, loan details, and credit history, the data was cleaned, transformed, and engineered to capture repayment capacity. Among several models, a stacking model that combined XGBoost and LightGBM, with a Logistic Regression Metamodel performed best. It surpassed other models both in speed and goodness of fit, with an ROC-AUC of 0.9451, precision of 0.94, recall of 0.73, F-1 score of 0.83. Recall and F-1 were emphasized because defaults represented only 22 percent of the data. In lending,the financial loss from approving a loan to a customer who later defaults is typically much higher than the opportunity cost of denying a loan to a customer who would have repaid it. Accordingly, recall and a balanced F-1 score are the most relevant measures of success in this context.

The model's business impact is significant. 5 to 8 percent reduction in defaults could save $500,000 annually, with larger gains from improved recall. It also enables risk-based pricing while SHAP explanations ensure transparency and compliance.

Overall, this project shows how advanced machine learning and explainable AI provide predictive strength and strategic value for responsible, profitable lending.

1. Introduction

The credit scoring process, particularly in the domain of unsecured lending, is undergoing a transformative shift due to the rise of digital financial services and the increasing demand for faster, more reliable risk assessments. Financial institutions today must not only make accurate decisions but also do so in real time to enhance customer experience, minimize default risk, and ensure compliance with evolving regulatory standards. Traditional scoring systems, which rely heavily on fixed rules or limited statistical models, often fail to capture the nuanced, dynamic nature of individual credit behavior in a digitized ecosystem (Hand and Henley 1997).

This project aims to develop a machine learning-powered, real-time credit risk scoring system. The scoring system will also be integrated with an interactive Streamlit application to ensure that deployment is seamless and decision support is provided. The solution will combine a robust ML pipeline for predictive modelling with an intuitive, user-friendly web interface, enabling stakeholders to input applicant data, receive instant credit risk assessments, and access interpretable explanations for each decision.

A successful well-built model has direct, real-world applications in the banking industry, with opportunities to be integrated into credit decision engines. Financial institutions are under increasing regulatory pressure to practice responsible lending, and this system can support safer, data-driven credit decisions in the best interest of both banks and their customers.

Classification algorithms will be trained on historical credit data as part of the machine learning pipeline and will then be optimized using cross-validation and hyperparameter tuning techniques. Another key aspect of the project is model explainability, to ensure transparency and regulatory compliance (consumer protection regulations). This will be done

using SHAP (SHapley Additive exPlanations) values and feature importance visualizations. Streamlit will serve as the front-end framework to deliver real-time risk predictions and clear, accessible insights into model behavior.

The project impact is substantial: a reduction in loan defaults by 5-8%, can lead to an estimated annual saving of $500,000 - assuming a Credit Loss line of $5M. This scalable solution can be seamlessly integrated into existing credit decision engines or decision systems.

2. Literature Survey

2.1 Traditional Approaches

Credit scoring has historically relied on statistical models such as logistic regression and discriminant analysis. These models, while interpretable and computationally efficient, assume linear relationships between features and the target variable, and often struggle with high-dimensional or non-linear datasets. According to Hand and Henley (1997), logistic regression has been the industry standard due to its balance between simplicity and performance. However, the advent of big data has exposed the limitations of these methods in dealing with complex consumer profiles.

2.2 Emergence of Machine Learning

Advances in machine learning (ML) have significantly improved the predictive power of credit risk models. Algorithms such as Random Forests, Gradient Boosting Machines (e.g., XGBoost, LightGBM), and Support Vector Machines offer superior accuracy by capturing non-linear relationships and high-order feature interactions. Lessmann et al. (2015)

benchmarked many classification algorithms and demonstrated that ML-based scoring systems consistently outperform traditional models across various datasets. These models can process large volumes of structured and semi-structured data, making them ideal for modern financial datasets. Despite their high performance, these models are often criticized for being "black boxes," which has led to a growing emphasis on model interpretability.

2.3 Explainability and Regulatory Compliance

To address concerns around transparency, researchers and practitioners have adopted explainable AI (XAI) frameworks. One of the most widely used methods is SHAP, which is based on cooperative game theory and assigns each feature a contribution value for individual predictions (Lundberg and Lee, 2017). SHAP not only enhances model interpretability but also supports compliance with legal mandates that require clarity in algorithmic decision-making.

Additionally, regulators such as the European Commission and the U.S. Consumer Financial Protection Bureau (CFPB) have emphasized the importance of explainability in credit decisions, citing the need for fairness and accountability in automated systems. Projects that integrate SHAP and other visualization tools help stakeholders understand the rationale behind each decision, thereby improving trust and user adoption.

2.4 Real-Time Credit Scoring Systems

The need for real-time decision-making has propelled the integration of ML models into web applications. Tools like Streamlit offer a lightweight, Python-based framework for deploying interactive data science applications. It facilitates rapid development cycles and enhances user engagement by offering dynamic, visual feedback. When integrated with robust ML

backends, these platforms can serve as full-fledged decision support systems that offer instant predictions and insights.

3. Research Objective(s)

1. Replace the conventional scoring techniques - logistic regression is most often used- with a machine learning-based credit scoring model; create, implement, and assess it
2. Integrate SHAP to guarantee transparency and offer practical insights into model predictions or model explain-ability.
3. Use Streamlit to create an intuitive, user-friendly, and real-time online application that supports decision-making and model deployment.

4. Methodology and Tooling

The Dataset

The model training and evaluation were performed on a publicly available Credit Risk Dataset sourced from Kaggle.

Source: https://www.kaggle.com/datasets/laotse/credit-risk-dataset?resource=download

Dataset Structure: The dataset is made up of c.32,500 records and 12 key features. The target variable is loan_status, a binary classification for a good vs. bad loan (0 = non-default, 1 = default).

The features are categorized as shown below, all of which help to provide a comprehensive view of the borrower:

Borrower Information (Demographics & Employment)	Age Income Home Ownership Length of Employment

Credit History	Previous Defaults Credit History Length
Loan Characteristics	Loan Amount Loan Interest Rate Loan % of Income (Debt Burden Ratio) Loan Intent/Purpose Loan Grade

The above features are examined in more detail in the EDA section. The dataset provides valuable insights into the borrower's financial profile, past behavior, and the specifics of the lending request—each of which plays a critical role in accurate risk assessment.

Tooling

- **Development Environment:** Jupyter Notebook was used end-to-end, from EDA to model training and evaluation.
- **Programming Language:** Python
- **Libraries:** Key libraries included:
    - pandas and numpy for data manipulation
    - scikit-learn for data preprocessing and model training
    - streamlit for building the interactive web application.
    - shap for model explainability
- **Hardware:** A local computer with a CPU was used for the initial development and training. The procedure would be moved to a cloud-based environment such as AWS SageMaker, which provides scalable and managed infrastructure with access to GPU/TPU resources for accelerated training, and for production-scale training on larger datasets.

5. Exploratory Data Analysis (EDA)

Data Analysis, Visualizations, Data Mining Techniques

The dataset contains 12 features across demographics (age, income, home ownership, employment length), loan characteristics (loan intent, loan grade, loan amount, interest rate), and credit history (loan percent of income, prior defaults, credit history length). Class imbalance is present with approximately 21.8 percent defaults versus 78.2 percent non-defaults, which motivates the use of F1-score and ROC-AUC over accuracy for model evaluation and feature selection.

Outlier inspection surfaced unrealistic values in age and employment length, which were cleaned or capped. Extreme values in interest rate and loan percent of income were retained because they reflect real-world high-risk borrower behavior. To mitigate scale distortion, income and loan amount were log-transformed. These steps preserve predictive signal while reducing undue influence from extreme magnitudes.

Correlation patterns indicated that repayment capacity measures dominate predictive power. Loan-to-income ratio and interest rate are the strongest drivers of default, while income is secondary. Age, employment length, and credit history length had weaker associations with default outcomes, suggesting lenders prioritize financial capacity features.

Clarity in Underlying Assumptions, Unit of Analysis, Feature Interactions

Unit of analysis: individual loan application. Assumptions: extreme ratios such as high loan percent of income indicate elevated default risk rather than noise; applicants act in line with repayment capacity constraints. Feature interactions were emphasized, particularly the relationship between loan amount and income, which captures borrower stress more

effectively than absolute income. Interactions between employment stability and prior defaults help distinguish long-term stability from sporadic repayment behavior.

6. Data Preparation & Feature Engineering

Handling of Features, Feature Extraction and Engineering

Missing numeric values (for example, employment length and interest rate) were imputed using the median to reduce bias under skewed distributions. Categorical features (loan grade, loan intent, home ownership, prior default flag, and binned employment length) were one-hot encoded with the first category dropped to reduce multicollinearity. Outliers in age and employment length were capped or corrected.

Key engineered features included an income-to-loan ratio capturing repayment capacity, and employment stability bins (0–2 years, 3–7 years, 7–20 years, 20+ years, Unknown) to reflect credit risk practice. Encoding of default history and loan purpose supported modeling of non-linear effects, with medical and debt restructuring intents associated with higher observed risk.

Variable Transformations, Data Scaling, Assumptions, and Tests

Log transformations were applied to income and loan amount to address skewness. Feature scaling ensured numeric features contributed to comparable ranges during model training. Outliers in interest rate and loan percent income were deliberately retained to preserve credit risk signals. Median imputation was assumed adequate given distribution skew. The overall preparation strategy balanced data quality improvements with the need to retain informative extremes characteristic of credit risk datasets.

7. Models & Training

Model selection, descriptions, evaluation approach and key decisions

A total of six (6) models were developed and evaluated using cross-fold validation to ensure robust performance metrics. The key evaluation metric we used was the AUC-ROC, which was the single measure we used to compare the models' abilities to distinguish between the two classes (good and bad loans) and is particularly useful for a dataset such as this one with a class imbalance (4:1 ratio). Given the class imbalance, a crucial trade-off was evaluated between precision and recall for the predicted class. While a high recall would identify more potential defaulters (reducing risk), it might also incorrectly flag many non-defaulters (low precision), leading to a high number of false positives. Conversely, a high precision would minimize false positives but might miss some true defaulters (low recall). Depending on the organization's priorities, the cutoff must be defined accordingly. By default, it is generally assumed that the cost of a default outweighs the cost of a false positive. The final model selection was based on finding a balance between these metrics that aligned with the business objective of minimizing default losses while maintaining an acceptable approval rate.

Model Summary

Group Type	**Model Name**	**Description**	**Strengths (Why it was used)**
Linear Model	Logistic Regression	A statistical model used as an interpretable	Simple, highly interpretable, fast to train, and provides

		baseline to classify	clear insights into feature relationships.
Ensemble Models	XGBoost (Extreme Gradient Boosting)	An efficient gradient boosting framework that corrects the errors of the previous models iteratively.	Superior performance, speed, and accuracy
	LightGBM (Light Gradient Boosting Machine)	A gradient boosting framework that uses a leaf-wise tree growth approach	Faster training speed and lower memory usage than XGBoost
	Stacking Ensemble	A hybrid model using XGBoost and LightGBM as base models, with a Logistic Regression meta-model	Combines the strength of multiple models. It was used to assess potential improvements in accuracy compared to standalone models.
Nonlinear Model	Support Vector Machine (SVM)	This model finds the optimal hyperplane	Effective in high-dimensional

		to separate data into classes (0 and 1)	spaces. Works well for scenarios with clear margins of separation.
Deep Learning Models	Artificial Neural Network (ANN) & 1D Convolutional Neural Network (1D CNN)	Modern neural network approaches that learn complex patterns and relationships in the data.	Can capture highly complex non-linear relationships

Each model was fine-tuned using grid search and cross-validation to find the optimal hyperparameters.

Hyperparameter Tuning

Model	Hyperparameter Tuned	Best Parameter Found
Logistic Regression	C (inverse of regularization strength), penalty (L1 or L2)	{C: 0.1, penalty: 'l1'}
Support Vector Machine (SVM)	C (regularization parameter), gamma, kernel	{C: 1, gamma: 'scale', kernel: 'rbf'}
LightGBM	n_estimators, max_depth, learning_rate, num_leaves	{learning_rate: 0.1, max_depth: 10, n_estimators: 200, num_leaves: 50}
XGBoost	n_estimators, max_depth, learning_rate	{learning_rate: 0.1, max_depth: 7, n_estimators: 200}
Artificial Neural Network (ANN)	Optimizer, Epochs, Dropout	{optimizer: 'Adam', epochs: 50, dropout: 0.2}
1D CNN	Optimizer, Epochs	{optimizer: 'Adam', epochs: 10}

For all models, the dataset was split into an 80% training set and a 20% test set.

Model Performance

Model	ROC-AUC	Precision (1)	Recall (1)	F1-Score (1)	Training Time
Stacking Ensemble	0.9451	0.94	0.73	0.83	4.75 seconds
XGBoost	0.9429	0.97	0.72	0.82	14.91 seconds

LightGBM	0.9434	0.95	0.72	0.82	113.72 seconds
ANN	0.8664	0.66	0.78	0.71	33.97 seconds
1D CNN	0.9113	0.94	0.62	0.74	32.44 seconds
SVM	0.8893	0.93	0.62	0.74	387.32 seconds
Linear Regression	0.8617	0.75	0.53	0.62	7.12 seconds

Model Evaluation

The Stacking Ensemble delivered the highest AUC-ROC, albeit marginally, and was selected as the final model for deployment due to its superior performance and robustness. It also had swift training time compared to the other models, reinforcing our choice to deploy in a production environment. The ANN model did show promise but required further tuning to perform better even though it achieved the best recall (1).

Classification Report - Stacking Ensemble

	Precision	**Recall**	**F1-Score**
0	0.93	0.99	0.96
1	0.94	0.73	0.83

Confusion Matrix - Stacking Ensemble

		True Label	
		0	1

Predicted Label	0	4433	330
	1	54	911

Validating Assumptions and Business Impact

The results clearly demonstrate that machine learning can effectively capture complex credit risk patterns. In particular, the stacking ensemble model shows strong potential to significantly reduce annual credit losses. Since credit losses directly influence the Risk-Adjusted Revenue (RAR) line, even small improvements yield meaningful impact.

For example, in a financial institution with $100M in annual acquisition volume and a 5% default rate (where defaults are defined as write-offs, not delinquency buckets at 30, 60, or 90 DPD), a 0.1% improvement in the default rate translates to a $10,000 increase in RAR. Extrapolating this effect, the model could enhance RAR by up to $3M annually, assuming all other revenue lines remain unchanged and frontline productivity is not adversely affected by lower approval rates.

What is next for the business

In the lending space, a risk score not only helps reduce total credit losses but can also be leveraged to implement risk-based pricing. This approach incorporates the inherent risk profile of each customer, enabling more accurate pricing and ultimately boosting the top-line revenue. A simplified banking Profit & Loss (P&L) statement illustrates the impact:

Gross Interest Margin (GIM)	8.50%	Improved through risk-based pricing
Cost of Funds (CoF)	3.50%	
Net Interest Margins (NIM)	5.00%	
Net Fee Income (NFI)	0.5%	

Revenue	5.50%	
Credit Losses	1.50%	Reduced through credit risk models
Risk Adjusted Revenue (RAR)	4.00%	

Beyond lending, banks can also apply these models in parallel domains such as fraud detection, as well as across different product lines, including mortgages, installment plans, and other credit products.

8. Lessons Learned and Additional Methods

Working with a relatively small dataset was intentional, as the primary objective was to compare model performance. However, data quality remains paramount. The richer and cleaner the dataset, the higher the likelihood of strong performance—often more so than searching for a so-called “magic” model.

Explainability is also not an afterthought. With governments and regulators placing increasing emphasis on consumer protection, model transparency has become critical. In fact, the lack of explainability has historically been one of the main barriers preventing banks from fully deploying credit risk models, leading many to instead rely on internal ML scores combined with rule-based cutoffs.

On agility, while the stacking ensemble delivered the best results here, other approaches could be explored. For instance, CatBoost—a gradient boosting model known for its strong handling of categorical variables—could be stacked with other gradient models to push accuracy even further.

In terms of features, many more possibilities remain untapped. Third-party datasets can enrich models with attributes such as employment type (salaried, self-employed, retired), job

title, and job category (white-collar, blue-collar, etc.). Credit bureau behavioral data is also highly valuable—for example, the number of credit inquiries made by a customer, since frequent inquiries can signal credit hunger and higher risk.

The opportunities are nearly limitless. With the right data and models capable of capturing complex relationships, driving credit losses down to their lowest possible levels is no longer a distant aspiration, but a realistic and achievable goal.

Addressing Fairness and Algorithmic Bias

The model depends on historical interest rates to evaluate how borrowers handled their finances in the past. The field shows past lending results but it is noted that algorithmic bias can persist through this measurement. For example, the interest rate a customer received in the past could have been affected by market conditions and outdated lending practices that did not accurately represent their actual risk level at that time and may have possibly contained discriminatory elements. The model risks learning to connect specific population groups with elevated risk through past pricing practices which might have been discriminatory. A person who received high interest rates in the past because of outdated factors will stay in a high-risk category even though their financial situation has improved. The evaluation of this feature's effects on various population groups should occur during future model development to identify alternative risk assessment features that directly measure current borrower risk through credit utilization and debt-to-income ratio calculations.

9. References

Hand, D. J., and W. E. Henley. 1997. "Statistical Classification Methods in Consumer Credit Scoring: A Review." Journal of the Royal Statistical Society: Series A (Statistics in Society) 160 (3): 523–541. https://academic.oup.com/jrsssa/article/160/3/523/7102381.

Lessmann, Stefan, Bart Baesens, Hsin-Vonn Seow, and Lyn C. Thomas. 2015. "Benchmarking State-of-the-Art Classification Algorithms for Credit Scoring: An Update of Research." European Journal of Operational Research 247 (1): 124–136. https://www.sciencedirect.com/science/article/abs/pii/S0377221715004208.

Lundberg, Scott M., and Su-In Lee. 2017. "A Unified Approach to Interpreting Model Predictions." Advances in Neural Information Processing Systems 30. https://arxiv.org/pdf/1705.07874.