LLM DRIVEN CREDIT RISK ASSESSMENT FOR LOAN APPROVAL

by

PREKSHA PRANAYKUMAR SHAH

A Major Research Project
presented to Toronto
Metropolitan University
in partial fulfillment of the requirements for the degree of

Master of Science
in the Program of
Data Science and Analytics

Toronto, Ontario, Canada, 2025

AUTHOR'S DECLARATION FOR ELECTRONIC SUBMISSION OF A MAJOR RESEARCH PROJECT (MRP)

I hereby declare that I am the sole author of this Major Research Paper. This is a true copy of the MRP, including any required final revisions.

I authorize Toronto Metropolitan University to lend this MRP to other institutions or individuals for the purpose of scholarly research.

I further authorize Toronto Metropolitan University to reproduce this MRP by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

I understand that my MRP may be made electronically available to the public.
Preksha Pranaykumar Shah

LLM DRIVEN CREDIT RISK ASSESSMENT FOR LOAN APPROVAL

Preksha Pranaykumar Shah
Master of Science 2017
Data Science and Analytics
Toronto Metropolitan University

ABSTRACT

Credit risk assessment is a foundation of financial decision-making; customarily, statistical models such as logistic regression dominate such a thing, and they rely heavily on structured borrower attributes like credit scores, income-to-debt ratios, and repayment history. These models, while somewhat effective, often miss signals showing borrower woe or anger found in formless money stories. Predictions for loan approvals that happen to be more strong can come from this Major Research Project (MRP), proposing a hybrid of structured with unstructured data integration. The LendingClub dataset yielded structured borrower-level features as well as the Consumer Financial Protection Bureau (CFPB) complaint narratives gave textual perceptions, which analysts aggregated into cohort-level sentiment and volume-based features. Because they performed better on tabular credit datasets coupled with an ability to handle sparse, high-dimensional inputs, we selected gradient-increased decision trees (XGBoost) as the core modeling framework. For improved predictive reliability, feature engineering was employed. Also, Platt scaling serves probability calibration involving hyperparameter optimization through Optuna. Evaluating through ROC-AUC and PR-AUC plus SHAP-based interpretability confirmed that the hybrid model consistently did better than the numeric-only baseline because ROC-AUC improved 1–2 percentage points plus precision measurably gained under severe class imbalance. The calibrated probabilities provided stable decision thresholds then. These limits suited regulated lending environments, aside from precision. Loan approval processes do become more reliable and more transparent because of this study that shows that the credit scoring models' predictive capacity and interpretability are improved through combining numerical features with text-derived cohort signals.

Key words:

Credit Risk · Loan Approval · Machine Learning · XGBoost · Feature Engineering · Large Language Models (LLMs) · SHAP · Probability Calibration · Consumer Complaints

ACKNOWLEDGEMENTS

I would like to sincerely thank my project supervisor, **Dr. Ayse Bener**, for her invaluable support and guidance throughout the course of this project. Her thoughtful feedback, encouragement, and constant availability for discussion have been instrumental in shaping the direction of my research. I deeply appreciate her time, dedication, and expertise, which not only enriched my learning experience but also greatly enhanced the quality of this work.

I am also grateful to my professors and peers in the MSc Data Science and Analytics program for their continuous encouragement and insightful discussions. Their inputs helped me broaden my perspective and approach challenges with clarity.

Finally, I would like to express my gratitude to my family and friends for their unwavering support and understanding during the course of this project. Their patience and motivation provided me with the strength to successfully complete this work.

Thank you.

Table of Contents

List of Figures

List of Tables

1. INTRODUCTION

1. Background and Motivation

Credit risk assessment is still a function that is vital within the financial services industry. Evaluating with accuracy the likelihood that a borrower might default is having some direct impact. Lending institutions as well as the broader economy both become more stable as a result. For decades, interpretable customary approaches are the backbone of credit risk modeling and regulators accept them especially like scorecard models and logistic regression. Income, in addition to credit history, loan-to-income ratios, as well as repayment records, are examples of the structured numerical plus categorical data these models mainly rely on.

The lending landscape has changed now. This evolution has been rapid in pace. Large-scale digital lending platforms emerge while peer-to-peer financing finances alternative credit sources which generate enormous amounts of borrower-related information beyond conventional credit scores. Borrower narratives, complaint data, and other unstructured signals reflect financial distress or repayment challenges before structured records do. Machine learning and also natural language processing advances have simultaneously demonstrated an important potential in the extraction of predictive perceptions from unstructured data.

As non-customary data grow in prevalence, and as artificial intelligence advances, this is creating an opportunity for one to move beyond purely numeric models and to incorporate more subtle signals of creditworthiness. Financial regulators and institutions demand models with higher predictive performance. These models also must maintain both interpretability and fairness and stability during economic cycles. This dual requirement does motivate researchers to explore hybrid systems models that combine structured credit data with features that they derive from unstructured text while still ensuring explainability for them.

2. Problem Definition

The central problem addressed in this project is developing a strong, explainable framework toward assessing credit risk in loan approvals that integrates structured financial indicators along with engineered features derived from borrower complaint narratives. Customary credit scoring models like logistic regression scorecards (e.g., FICO) rely mainly on structured numerical data such as repayment history, credit utilization, and debt-to-income ratios. Imposing strong linear assumptions, these models often fail to capture the complex nonlinear relationships driving real-world borrower behavior for they are simple to implement as well as interpretable. Consequently, these kinds of models might fail in identifying the default risk's early-warning signals.

The dataset for this study was carefully chosen. In that choice, the problem's dual nature is reflected. From the LendingClub loan dataset, the structured component is sourced containing borrower-level variables. These variables include records for repayment, ratios of income to debt, grades for loans, and rates of interest. These features are valuable. They give just a limited picture of credit standing. The Consumer Financial Protection Bureau (CFPB) complaint database presents borrower narratives capturing repayment difficulties, dissatisfaction, or disputes with financial institutions in parallel. Dimensions in behavior of risk which are not visible otherwise inside numeric loan files get introduced by these narratives when aggregated into features at the cohort level like frequency of complaint and polarity of sentiment.

This problem is in fact complex due to a number of challenges. First, the dataset happens to be highly imbalanced because default cases will form only a small minority relative to successful repayments, and this complicates model training as well as evaluation. Also, the data are time sensitive plus attributes like issue_d need care so lookahead bias and data leakage are avoided. Third, the preprocessing pipeline does introduce high-dimensional sparse matrices since it one-hot encodes categorical features, so models that must efficiently handle sparse inputs are required.

The problem is therefore twofold:

I. Effectively leverage structured loan features toward predictive model construction reducing class imbalance, temporal drift, with potential data leakage issues.
II. For identifying high-risk borrowers when loans originate, we evaluate whether text-derived features, like complaint counts and aggregated sentiment scores, can measurably improve predictive power.

This problem is with computational implications. Regulatory implications also exist. Machine learning algorithms that are scalable for operating on thousands of mixed data type rows are computationally required. Regulatory oversight confronts credit risk models by frameworks like Basel II/III accords, which stress fairness, calibration, and model transparency. The chosen solution must achieve strong predictive accuracy while providing interpretable outputs too. Furthermore, it is expected to provide for stable probability estimates with justification for feature importance.

3. Research Objectives

The overarching objective of this project is to assess credit risk via the design and evaluation of a framework for the integration of structured borrower and loan-level information with perceptions that consumers' complaint narratives derive. Improving predictive accuracy together with improving probability calibration as well as maintaining interpretability are the project's aims for real-world credit decision-making in regulated environments.

To achieve this goal, the research is being guided by the following specific objectives:

- Using structured financial features like borrower demographics, loan characteristics, and repayment histories, develop a baseline model which establishes a benchmark against customary scorecard-style approaches.

- For evaluation of behavioral data's predictive lift, investigate integrating textual signals from complaint narratives like sentiment polarity and borrower grievance frequency.

- Employ calibration methods for reliable probability outputs during threshold-based decisions after building a hybrid modeling pipeline that unites structured features and text-improved features inside gradient-increased decision trees.

- Systematic experiments benchmark the model variants and they tune hyperparameters. Researchers do also evaluate performance through metrics such as ROC-AUC, PR-AUC, F1-score, as well as calibration error, stressing how they address class imbalance that is intrinsic to default prediction.

4. Scope of the Study

The scope of this study is defined in terms of the modeling framework that was implemented and also the datasets that were selected. The research focuses mainly upon the LendingClub loan dataset regarding data, a repository of borrower profiles, loan terms, and repayment outcomes that is publicly available and large-scale. A baseline risk classifier can be built by use of the financial features in this dataset. In parallel, the Consumer Financial Protection Bureau (CFPB) complaint database is leveraged in order to derive aggregated sentiment and complaint frequency measures, and these measures serve as important proxies for borrower trust, satisfaction, and potential distress. Hybrid modeling is enabled via integrating these textual signals into the structured dataset at a cohort level.

The study bounds itself to the problem of default so it predicts default within loan approval. It does not make an attempt for the modeling of dimensions that are related to credit risk like Loss Given Default (LGD) or Exposure at Default (EAD). Datasets beyond the present project's scope would require modeling these dimensions. The focus, instead, remains on the prediction of repayment likelihood versus the default, a primary driver in making loan approval decisions.

Regarding methodology, the project limits the modeling scope to tree-based machine learning methods. The project stresses XGBoost also because it balances accuracy and interpretability as it performs well using tabular financial data. Neural plus AutoML approaches exist beyond this study's scope, while classical models serve as baselines like random forest and logistic regression. Neural and AutoML approaches are not within this study.

Chosen performance metrics define for evaluation the scope that includes ROC-AUC, PR-AUC, accuracy, F1-score, and calibration error. Since they are suited for imbalanced classification settings and are relevant for real-world credit decisions, we select these metrics. The study does not deploy at the production level rather experiments for proof of concept that shows feasibility also identifies key design trade-offs.

5. Contributions of the Study

This study makes several contributions in applied machine learning financial text analysis along with credit risk modeling. Initially, it contributes to the credit risk assessment domain. Its contribution also involves showing how structured customary loan data is meaningfully improved using text-derived features from regulatory complaint datasets. The study integrates complaint frequency and sentiment signals at the cohort level and also extends credit scoring. It exceeds standard methods focused on borrowers since it mirrors consumer trend contexts with specific financial measures.

Second, it adds to methodological literature through systematically assessing as well as applying gradient-increased tree models in hybrid data. XGBoost is tuned and assessed under multiple experimental conditions, and it performs strongly on structured tabular tasks as prior studies show. The analysis also documents how one optimizes hyperparameters by Optuna, how one calibrates probability via Platt scaling, and how one analyzes interpretability through SHAP values, so one can replicate future financial risk projects.

Third, the study helps people to understand evaluation of imbalanced financial datasets when the study employs a suite of complementary metrics such as ROC—AUC, PR-AUC, accuracy, F1-score, Brier score, and calibration error. PR-AUC as well as calibration plots highlight the limits of standard financial modeling metrics such as ROC-AUC in rare events such as loan defaults. For identifying predictive models' strengths and weaknesses in high-stakes domains, this broader evaluation perspective is required.

The project offers an implementation pipeline to prove the concept. The pipeline has been designed so it is strict and also interpretable. By balancing accuracy and transparency through SHAP-based feature attribution plus probability calibration the study offers perceptions adaptable for real-world financial decisions while respecting governance plus auditability needs in regulated spaces. These contributions advance the applied comprehension of how hybrid numeric–textual models can improve credit risk assessment in loan approval contexts in combination and the technical methodology.

2. Literature Review

1. Traditional Credit Risk Modelling Approaches

Credit risk modeling has customarily based itself on statistical and econometric methods like logistic regression plus its variants. Logistic regression for decades has served as the industry standard because it is simple, interpretable, and regulators accept it to construct credit scorecards. In such models, we express the log-odds for default as a linear combination of borrower attributes that include demographic details, financial ratios, and credit history indicators. This framework provides a direct mapping between feature coefficients along with the probability of default. Complex borrower traits can be translated into scores that are transparent plus auditable by credit institutions.

Because it aligns with governance and compliance requirements in banking, logistic regression is popular also because it is statistically strict. Supervisory authorities consistently stress both transparency and also justifiability in credit decisioning systems like the Basel Committee on Banking Supervision and national regulators do. Banks have relied upon these models when it comes to external reporting as well as internal decision-making since logistic regression coefficients can be easily

interpreted to be the marginal contribution coming from each variable.

Even though its use is wide, logistic regression has key limits on application to modern lending contexts. First, the method makes assumptions that predictors do relate in a linear way to the log-odds of default, but this does fail in its capture of complex nonlinear patterns intrinsic in borrower behavior. Repayment probability is affected through interactions among debt-to-income ratio, loan grade, and macroeconomic shocks within a nonlinear manner that manual feature engineering can model well. Logistic regression is quite sensitive to multicollinearity therefore it needs preprocessing like variable binning scaling and regularization. These preprocessing requirements can often limit its flexibility now. Large-scale datasets that exist in peer-to-peer lending platforms serve as one example.

The basic logistic regression framework has limitations indeed. These shortcomings were fixed through many extensions. Penalized regressions like Lasso and Elastic Net introduced regularization, toward which generalization performance improved allowing automated variable selection. Probit models in addition to survival analysis approaches have been explored likewise. These capture outcomes for time-to-default instead of risk states. These refinements extend beyond the statistical toolkit. However, they do still model in a linear way and humans must then intervene so they can design feature interaction.

2. Machine Learning for Credit Risk

The limitations of customary linear scorecard methods have prompted the exploration of machine learning approaches for credit risk modeling on the part of researchers and practitioners. Linear relationships are assumed for logistic regression then feature engineering is needed. ML methods, unlike logistic regression, can capture nonlinear interactions, high-order feature dependencies, along with complex decision boundaries. ML's flexibility has made for it a dominant model within modern credit risk assessment. Financial institutions do increasingly rely on large-scale transactional and alternative data sources in particular.

Decision trees in addition to random forests remain among the earliest machine learning methods. Credit scoring saw usage of these methods. Decision trees do partition the feature space into interpretable rules as well, and this allows risk managers to visualize borrower classifications as “if-then” conditions. Random forests reduce variance and overfitting via an ensemble of decision trees trained via bagging, improving predictive accuracy. However, though these models are powerful, increasing algorithms usually exceed their performance on financial tabular data.

Libraries such as XGBoost, LightGBM, and CatBoost implement gradient-increased decision trees (GBDTs), now ML methods adopted most widely in credit risk. Numerous studies and competitions, like Kaggle’s “Give Me Some Credit” challenge, show GBDTs predict better than logistic regression plus random forests consistently. Kaggle benchmarks often show ROC-AUC gains of 3–5 percentage points when increasing methods are used over logistic regression baselines to peer-to-peer loan datasets. GBDTs improve since they iteratively correct errors weak learners make then exploit interactions features have without explicit manual engineering.

Neural networks as well as Support Vector Machines (SVMs) have been applied within credit scoring literature too. Interpretability and scalability pose difficulties for SVMs in high-dimensional plus imbalanced financial contexts because SVMs perform well on more balanced smaller datasets. Deep feedforward as well as recurrent architectures, notably neural networks, offer learning potential through representations from sequential financial transactions or raw borrower data. However, for tabular credit risk tasks, they do not consistently gain performance beyond GBDTs unless datasets remain very large plus substantial computational resources are then available. Neural networks seem opaque, raising governance concerns. Opacity within regulated domains like banking is particularly concerning.

One key area of ML research highlights explainability. This research is focused on the topic of credit risk. For regulators, how models do arrive at their decisions requires transparency increasingly in addition to high predictive accuracy. Frameworks for model-agnostic interpretability like SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) get routine integration within ML pipelines now. This integration serves toward this end.

SHAP values for credit applications are valuable because they allow risk attribution to borrower features like loan grade, interest rate, or debt-to-income ratio. Since this can balance the trade-off between predictive performance as well as interpretability, ML-based systems are certainly more suitable for real-world adoption.

In terms of performance evaluation, ROC-AUC along with PR-AUC are typically used to benchmark ML-based credit risk models, since accuracy alone fails to reflect class imbalance challenges. Across a collection of datasets, empirical results do show that well-tuned GBDTs not only discriminate with a higher power but also estimate probability in a more reliable way after post-hoc calibration.

3. Textual and hybrid approaches for Credit Risk

While structured borrower data such as loan amount, credit history, with debt-to-income ratio are still key to credit risk modeling, incorporating unstructured textual information's value is highlighted by recent advances. This includes borrower narratives in addition to financial disclosures in conjunction with consumer complaints. These provide qualitative signals about financial distress and borrower intent not easily captured in numeric variables alone. So numeric with textual data fuse which models hybrid credit risk promising a path shown in studies.

One research strand focuses on borrower narratives or loan application justifications. Narratives and justifications receive this strand's focus. Emotional tone, language complexity, and sentiment polarity can, studies show, act as strong predictors of repayment behavior. For instance, narratives are in a positive wording when they express a confidence and a financial responsibility, and these narratives have been correlated with a lower default risk, whereas narratives that stress hardship, uncertainty, or an urgent need often correspond with higher default rates. Such features can be extracted by natural language processing (NLP) techniques. These techniques range all the way from simple bag-of-words with sentiment dictionaries to transformer-based embeddings.

Consumer complaint data represents a textual source. This data is also of importance. Millions of borrower-submitted grievances to financial institutions are provided by platforms like the Consumer Financial Protection Bureau (CFPB) complaint database. Whenever we aggregate each of these complaints at that geographic or cohort level, this aggregation has been known to reveal some systemic patterns of dissatisfaction, disputes, and institutional malpractices. A region could signal elevated repayment risk among borrowers in that area with a higher frequency of complaints about billing disputes or loan servicing irregularities. For credit risk classification, there has been demonstrated an improved discrimination power in hybrid models that do integrate structured financial metrics with complaint signals.

Hybrid approaches also leverage financial disclosures and earnings reports. Sentiment analysis from management discussions, especially uncertainty or negative outlook indicators, correlates with subsequent default likelihoods plus credit downgrades, as found in prior studies on EDGAR filings with annual reports. With textual sentiment from financial statements, researchers combined structured ratios such as leverage and liquidity. This allowed them to improve predictive performance notably, with ROC-AUC gaining up to 5–7 percentage points compared to structured-only models.

Credit scoring workflows now use textual features technically. This integration has in fact greatly evolved. Earlier approaches did rely on feature concatenating, in which someone appended text-derived features such as sentiment scores or term frequencies to structured borrower data prior to training machine learning models. Language models generate embeddings for representation learning methods to employ, also these embeddings (e.g., Word2Vec, BERT, FinBERT) get combined with gradient-increased trees or neural networks. This allows one to capture subtle semantics in addition to simply counting polarity or frequency. Since it provides high predictive accuracy along with compatibility to feature attribution frameworks such as SHAP, combining embeddings through increasing algorithms has been very effective.

At the same time, interpretability as well as governance are challenged by these hybrid approaches. It is easier to validate structured financial variables than it is to validate textual features which are noisy plus context-dependent. For model predictions, transparent explanations are often demanded by the regulators, and purely embedding-based representations may not provide sufficient auditability. Researchers increasingly prefer aggregating and interpreting textual indicators for

addressing this (e.g., averaging sentiment per borrower cohort, specifying complaint categories) instead of embedding raw data in production settings.

4. Model Interpretability and Explainability in Finance

In addition to forecasting accuracy, interpretability remains a key factor for credit risk models. Credit scoring, like many machine learning applications facing consumers, is tightly regulated and requires financial institutions to justify loan approval or rejection decisions before regulators and borrowers. Interpretability is something that is more than just desirable for this reason, it is in fact a mandate. Frameworks such as that of the Equal Credit Opportunity Act (ECOA) in the United States, that of the General Data Protection Regulation (GDPR) in the European Union, and guidelines issued by banking supervisors such as that of the Basel Committee on Banking Supervision (BCBS) require interpretability.

Such customary credit scoring models as logistic regression interpret intrinsically. For credit decisions, scorecards from these models provide transparent rule-based justifications, and the estimated coefficients directly indicate the marginal effect from each feature on default probability. Still, these models fit with governance needs. They often are not able to capture any non-linearities or higher-order interactions, thus restricting their predictive accuracy in quite large heterogeneous datasets.

Predictive accuracy in credit risk applications has greatly improved through the adoption of machine learning models such as random forests as well as gradient-increased trees yet interpretability is reduced. These ensemble methods construct complex decision boundaries from data. Because of this, it is difficult to explain just why a particular borrower was classified as risky. Explainable AI (XAI) tools which bridge predictive power as well as transparency have thus gained focus from this trade-off.

Post-hoc explanation techniques using SHAP (SHapley Additive exPlanations) are widely adopted in credit risk. SHAP values are rooted deeply in cooperative game theory. SHAP values let analysts understand local explanations and global importance since they assign to each feature an additive contribution for a prediction like which variables matter most overall such as why one loan was classified as risky. SHAP has been used for the purpose of showing that int_rate, dti, and loan_grade contribute in the most large way to default predictions for example. On the other hand, text-derived signals, such as sentiment polarity or complaint volume, can add marginal explanatory power.

SHAP is often preferred due to its consistency as well as theoretical guarantees, however other methods such as LIME have also been applied. Because it allows efficient computation even for LendingClub's millions of loan records, recent work extended SHAP to tree-based models (TreeSHAP).

About trust and operational utility, and compliance, interpretability is. To validate model outputs as well as identify potential sources of bias, loan officers and risk analysts require explanations. For example, in the event a model relies disproportionately upon geographic or demographic proxies, interpretability tools are able to highlight this issue and support measures that are correct such as feature regularization or bias mitigation. Banks design risk-based pricing strategies using interpretable feature attributions with well-calibrated probabilities. These strategies happen to be accurate as well as explainable to customers.

However, balancing performance and interpretability poses problems. Algorithms which are increasing combined with SHAP offer a compromise that is practical. Yet because of little clarity, deep learning methods for tabular data such as TabNet or transformer-based models stay hard to use in controlled financial settings. For that reason, most financial institutions and research studies favor tree-based ensembles then interpret them, which is the current best practice.

5. Gaps in Literature and Justification for this Study

The reviewed body of literature shows substantial progress in credit risk modeling since it spans from customary scorecard approaches to modern machine learning ensembles along with hybrid frameworks that integrate unstructured text. Several critical gaps persist. These gaps do nevertheless motivate the present study now.

First, although researchers widely use logistic regression with penalized GLMs in practice because they interpret them easily, ensemble learning techniques consistently outperform these methods on large-scale datasets. When gradient-increased trees are applied in comparison to logistic regression, ROC-AUC has +3–5 percentage point improvements that are frequently reported. LendingClub plus FICO datasets get used in studies. Despite all of the demonstrated benefits, adoption within the financial institutions still remains slow. The literature often stops short at showing of how such models can be made compliant to explainability and governance requirements, which explains this slow adoption. Because of this gap, projects explicitly balancing interpretability with predictive power are needed.

Second, most academic studies focus solely upon structured tabular loan features while machine learning approaches like random forests, XGBoost, and LightGBM dominate benchmarking competitions (e.g., Kaggle credit risk contests). Borrower justifications, financial disclosures, with consumer complaints, unstructured external signals, have seen relatively few systematic incorporation attempts. Literature often relies on direct document embeddings or applies sentiment at the individual borrower level when textual features are explored making text difficult to link to structured loan datasets that lack explicit identifiers. One can still explore textual data integration if data aggregate at the cohort level; this preserves information without borrower-level matches.

Third, interpretability method use like SHAP remains often limited to proof-of-concept studies, although they are increasingly referenced in credit risk research. Many papers do rank global features even though they do not evaluate the interaction of interpretability with probability calibration. Also institutional decision-making needs lack an evaluation of this interaction. Calibrated probabilities plus transparent explanations are needed for customer communication, regulatory reporting, and risk-based pricing. In order to bridge this methodological gap, more advanced models along with calibration and interpretability pipelines are required.

That class imbalance does remain as a persistent challenge for credit datasets the literature does reveal. Defaults are typically less than 10% within the dataset. Thus basic models create higher accuracy scores. Few combine imbalance handling with the integration of new feature types such as with text-derived signals, while some of the studies adopt cost-sensitive learning or undersampling. Questions that are unanswered still remain in regard to the question of whether hybrid models are able to improve recall in some meaningful way for defaulters. If they do so, the models' calibration or interpretability may degrade.

This study directly addresses these open questions. Gradient-increased decision trees (XGBoost) are employed first ensuring competition with top machine learning benchmarks that have shown cutting-edge performance on tabular credit data. Second, rather than discarding any textual signals because borrower-level identifiers are absent in them, this project introduces such a cohort-level integration strategy. It aggregates complaint volume along with sentiment polarity coming from within the CFPB dataset at just the ZIP3 level. This novel merge strategy avoids missing identifiers' pitfalls and preserves unstructured data's potential value. This study stresses interpretability and calibration equally because it uses SHAP values for feature attribution alongside Platt scaling for probability correction, so this model is regulatorily defensible yet accurate.

This research does thereby contribute a methodological blueprint which is for combining structured and unstructured data within an explainable, calibrated, and operationally viable credit risk framework. The study fills identified gaps from prior literature advancing academic understanding and practical application in credit risk modeling.

3.Descriptive Analytics | Exploratory Data Analysis

1. Dataset Overview

The dataset used within this study comes from the LendingClub loan dataset, publicly available and widely employed in credit risk research, both academic and applied. Since it combines categorical with numeric predictors regarding borrower performance, the dataset has over 2.2 million rows plus approximately 138 features after preprocessing and cleaning.

The features cover financial dimensions and dimensions related to the borrower. Annual income, loan amount, and interest

rate exist as numeric variables. Revolving utilization, debt-to-income (DTI) ratio, installment amount, and also delinquencies are numeric variables. Grade, subgrade, employment length, home ownership, together with verification status, are categorical variables. Machine learning models were eased by transforming categorical predictors into one-hot encoded representations in this project along with numeric features retained in continuous form under selective normalization.

Loan status is in fact the target variable of interest, changed into a binary classification task:

- Good loans (non-defaults)
- Bad loans are defaults or charged-off cases.

The high class imbalance is a key characteristic observed for this stage since defaulted loans are but a small minority. This imbalance was expected for influencing model training and evaluation in a large way, so it was necessary for using metrics such as PR-AUC and F1-score along with ROC-AUC, which is customary.

The dataset includes temporal richness also because it spans multiple years of loan issuances. Certain economic cycles affect default concentrations. Temporal signals such as issue year and issue quarter were found later to be very predictive of borrower behavior.

2. Loan Outcome Imbalance

One of the most critical discoveries from the data exploration is that loan results are spread in a very unbalanced manner. Out of a total of more than 2.2 million records, a large majority of them are good loans because they do not default, while bad loans either default or charge off and constitute only a small minority. Because proportion varies slightly across issue years and loan grades, bad loans compose less than 10% of the dataset as shown by frequency analysis. This imbalance poses a challenge for standard classifiers since these models can accurately get deceptively high overall scores simply by overwhelmingly predicting the majority class while they happen to fail to identify defaults correctly. For example, a high accuracy is what would be yielded by a trivial model that predicts all loans as “good” though zero practical utility exists in financial decision-making that is risk-sensitive.

The imbalance does also directly imply just how one chooses performance metrics. Under imbalance, metrics such as accuracy and ROC-AUC inflate while metrics like Precision-Recall AUC (PR-AUC) and F1-score measure more realistically the model’s ability to capture defaults. Throughout the experimentation, these metrics received emphasis for this reason. Then the study evaluated the model. Also, the target distribution is displayed, which clearly highlights this skew. Since this skew reinforces the need to carefully calibrate probability outputs and also use class-sensitive learning strategies, it is important throughout modeling.

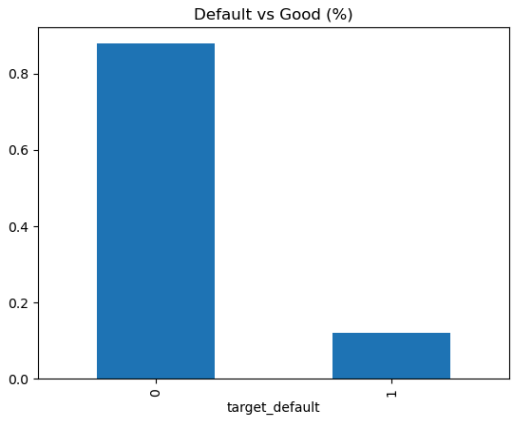

Figure 1: Target Distribution

3. Borrower Demographics and Financial Attributes

The exploratory analysis also examined borrower demographic with financial characteristics, as these variables customarily are considered strong creditworthiness predictors. Recorded were observations of several key patterns.

Loan grades with interest rates first displayed a clear monotonic relationship. Riskier grades matched greater mean interest rates. Grade A loans had mean interest rates nearly half of those for lower-grade loans such as Grade F and G. Credit grade assigned and interest burden reinforce the repayment likelihood's strong dependence. Second, a rather wide distribution was exhibited in the debt-to-income ratio or DTI. Also, a long tail of borrowers carried disproportionately high obligations. While most of the borrowers had DTI values clustering around the median range, a somewhat important minority reported values above 40% because that value indicated that they were highly vulnerable to repayment stress.

Third, annual income was distributed heavily to the right, and most borrowers were concentrated in middle-income brackets, but a small number have reported extremely high incomes. Higher income often correlates with less default risk, yet the skew creates challenges in feature scaling plus interpretation since normalization helps during preprocessing. Finally, demographic factors such as the length of employment showed informative variation. Home ownership status showed such informative variation at that time. Borrowers owning homes and having longer employment histories often showed lower default rates than those renting or with limited job tenure.

These observations are in alignment with prior literature for the reason that it stresses the predictive value of borrower-level financial stability indicators in credit risk modeling. They informed the feature engineering stage, where transformations like categorical encodings and log-scaling income improved model performance.

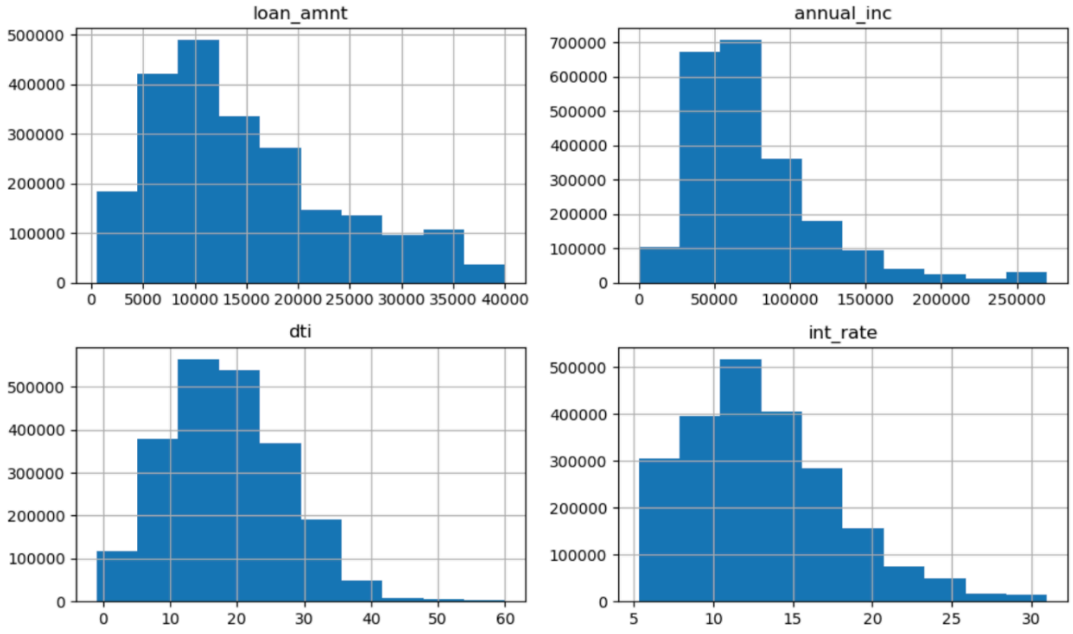

Figure 2: Core numeric features

4. Loan Characteristics

An analysis of loan-level attributes helped one understand structural patterns in all of the issued credit products and also how they do potentially relate to default outcomes. Several outstanding findings emerged.

First, loan amounts displayed a right-skewed distribution; few extended to large USD 35,000 balances, but most loans concentrated in small-to-mid ranges (below USD 15,000). Because of the fact that larger-loan brackets saw

disproportionately higher defaults, credit exposure likely increases systemic risk for sure. Second, loan terms clustered around those two dominant categories with terms of 36 months being one type. The other type of loan term was for tenures that lasted 60 months. Loans with longer terms showed higher default frequencies matching what was expected. Borrower financial stability sees an increase in uncertainty along with extended repayment horizons. This observation is critical since term length strongly interacts with interest rates. Term length burdens borrower repayment as well.

Loans had a reason, third. The loans' purpose revealed differences distinct in behavioral risk. Debt consolidation and also home improvement as well as credit card refinancing were actually the most common of purposes. Default rates were relatively higher for loans issued toward discretionary purposes as small business or major purchases, though safer for important categories such as debt consolidation comparatively. Issue year and issue quarter emerged finally as important temporal cohort features. Due to macroeconomic shifts and because of regulatory adjustments over time, loan performance showed variance across origination periods. For example, loans that someone originated when credit environments tightened tended to default slightly less frequently than someone issued them upon credit expansion.

A critical foundation toward accurate risk modeling comes from these perceptions, which underscore loan characteristics with borrower demographics. In the model, the inclusion of features such as loan term, purpose, and issuance date was justified because observed patterns greatly contributed to prediction strength, and also SHAP analyses later confirmed it.

5. Derived Ratios and Risk Indicators

In addition to raw borrower and loan attributes, we analyzed several derived ratios plus composite indicators to capture financial stress signals more effectively. These ratios are commonly used for credit risk modeling since they normalize the raw values. They also highlight borrower-level repayment capacity. The Debt-to-Income ratio or DTI was a critical derived measure measuring the borrower’s monthly debt obligations relative to gross income. Because most borrowers clustered down below 20%, the distribution of the DTI values was right-skewed. However, default rates rose sharply among borrowers with DTI exceeding 30% because that indicates a threshold beyond which repayment burden becomes unsustainable.

Since it measures the proportion of revolving credit (e.g., credit cards) that people utilize, the revolving utilization ratio is also related greatly with loan outcomes. Utilization levels were higher as well as quite strongly correlated with default because borrowers utilized revolving accounts to a great extent and thus struggled as a result to repay installment loans. A monthly loan installment amount divided by a borrower's total income derives the installment-to-income ratio as yet another key indicator for analysis. This feature did highlight repayment affordability for them. Affordability was stressed at a loan-product level for them. Defaulted loans showed higher concentrations for those values, which reinforces affordability metrics' importance in loan risk assessment.

Cohort-level features included temporal indicators such as issue year along with issue quarter. These features effectively captured during loan origination the macroeconomic and regulatory factors. Default rates showed cyclical variations in various quarters. External conditions such as credit market liquidity and interest rate regimes had an impact for sure.

Clear predictive value was shown by risk indicators plus derived ratios. So their presence in the modeling pipeline had justification. Later, feature importance analyses confirmed their importance because of how DTI, revolving utilization, and issuance cohorts consistently predicted loan default highly.

6. Temporal Trends in loan Insurance and Defaults

Temporal analysis was conducted for showing the evolution of loan issuance volumes as well as default rates. They aggregated loan origination data for each issuance year and for each issuance quarter. This identification of cyclical patterns linked these patterns to more broad economic and financial conditions.

Issued loans steadily grew during the dataset's earlier years reflecting retail credit market expansion. Yet, contraction times did appear too, and these matched eras when economic stability declined. Loan volumes noticeably dipped in years of macroeconomic stress; this suggests reduced borrower demand plus tighter underwriting standards.

Temporal fluctuations also displayed default rates distinctly as well. Certain high-volume years were when loans commonly showed increased default rates. This trend likely resulted from looser credit checks during expansions. However, in times during which lenders made a smaller number of loans, default rates would tend to decline, which showed that lenders screened borrowers in a better way and credit tightened.

For short-term seasonality was also revealed through quarter-level analysis. Defaults happened with more frequency among the loans issued in the quarters that followed rapid market expansion because this suggests that the quality in underwriting may be compromised when the loan volumes increase far too quickly. This relationship highlights temporal indicators such as issue_year along with issue_quarter as important indicators. Machine learning models later confirmed that the indicators predict features highly well. These temporal trends show assessing borrower risk requires contextualizing it in the broader economic cycle instead of isolating it. Models are able to capture these systematic risk variations in situations when issuance cohorts have inclusion as explicit features. This inclusion works to ensure models can capture variations.

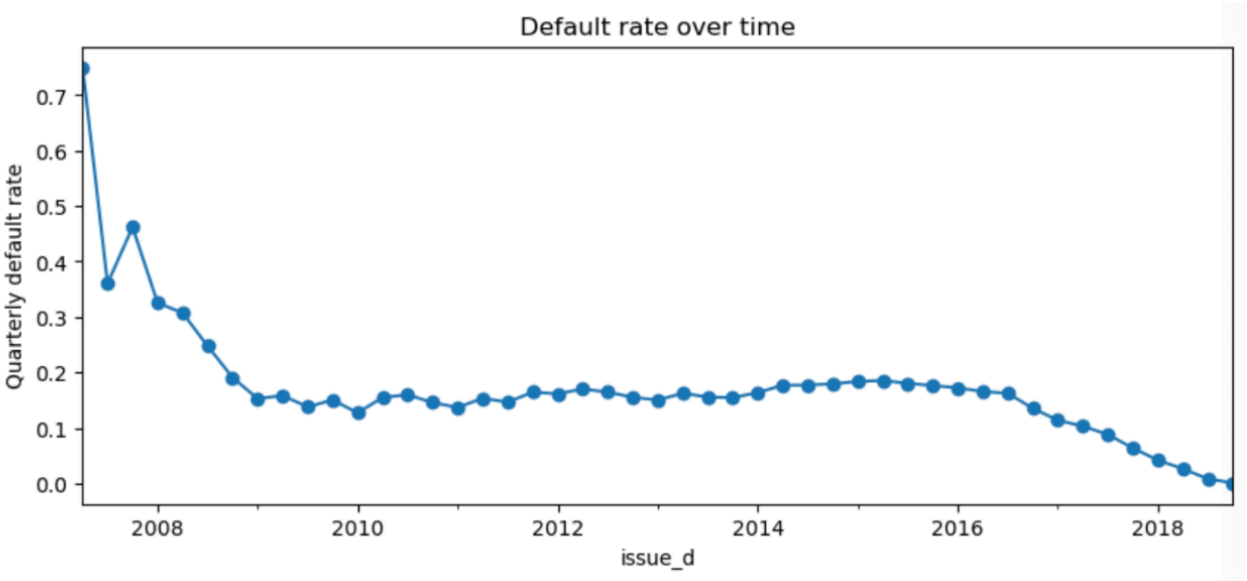

Figure 3: Time series Trend

7. Correlation Analysis

A correlation analysis was done to probe linear relationships in loan, key borrower, and financial attributes. Strong associations and potential multicollinearity issues were found when computing Pearson correlation coefficients across each of the numeric variables.

The analysis confirmed many expected correlations. Loan amount (loan_amnt) as well as funded amount (funded_amnt) were strongly correlated ($\rho \approx 0.98$) since their definitions in the dataset are near-identical. Interest rate (int_rate) showed a positive correlation to assigned loan grade (grade_num). This validated that higher-risk borrowers were charged with higher interest rates. Debt-to-income ratio (dti) exhibited also moderate positive correlations regarding delinquency indicators. This highlights its function as a risk flag for borrowers.

Interestingly, certain derived ratios such as credit utilization and the installment-to-income ratio showed weaker but

still important correlations with default outcomes, which suggests that while they do not dominate it alone, they predict incrementally when combined with other features. Loan structure variables like loan amount funded amount installment and borrower obligations like dti revol_util annual_inc were revealed by way of correlation heatmaps. Despite no single variable correlating excessively highly with the target variable (loan default), the combined set of moderately correlated features indicated that credit risk has a multi-factor structure. This supported the decision of retaining a broad feature space now for the machine learning experiments, expecting that tree-based models such as XGBoost would fully capture the non-linear interactions between all these variables.

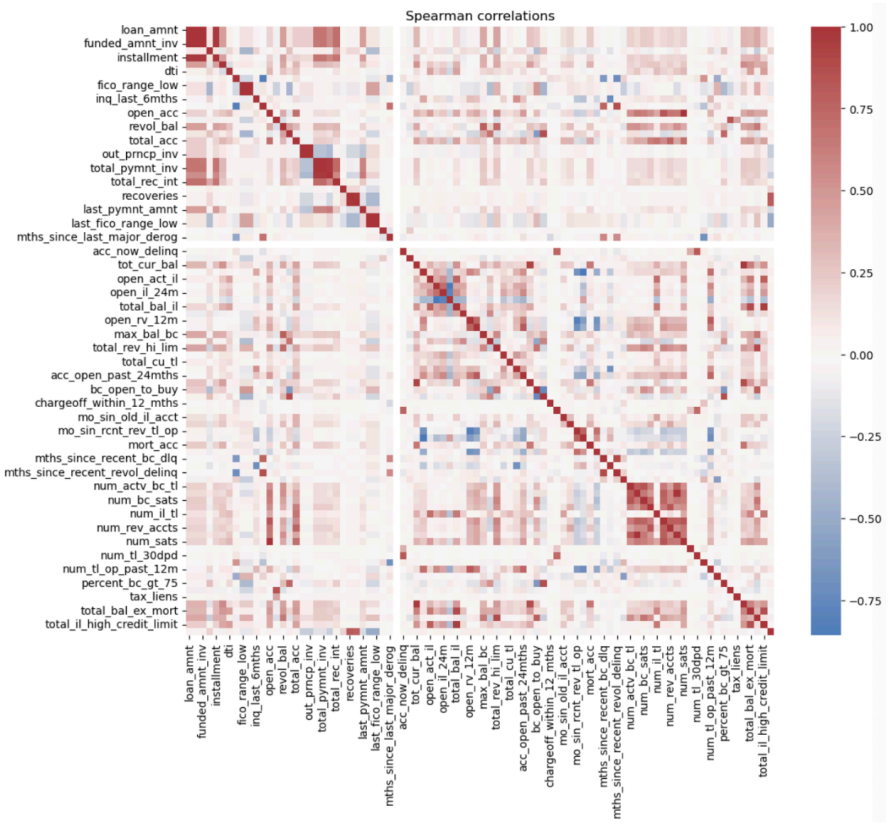

Figure 4: Correlation Heatmap

4. Methodology and Experiments

1. Experimental Setup

The experimental setup defines each of the tools as well as design choices plus a computational environment structuring the workflow for credit risk classification. For the study, a reproducible machine learning pipeline integrates structured borrower-level financial features. Signals at the engineered cohort level are derived from complaint narratives.

All of the experiments are conducted via Python 3.11 within a controlled virtual environment since that ensures consistency throughout all runs. The project relies on established open-source machine learning libraries. Libraries for data processing have use too. Pandas and Polars are employed in efficient handling of large tabular datasets, especially when data preprocessing and cohort-level aggregation happen. For NumPy, there is support for low-level numerical operations. Scikit-learn provides different utilities which are for model evaluation plus calibration and for splitting data into train–test

sets. The implementation for model training uses XGBoost with histogram-based tree growth (tree_method="hist") because that enables scalable execution on CPU hardware when datasets exceed 300,000 rows.

Optuna uses Bayesian optimization to automate searches. It also efficiently explores parameter combinations in order to optimize hyperparameters. MLflow tracks results, models, and artifacts systematically; this allows consistent experiment logs; it records parameters; it controls versions of outputs.

To integrate ZIP3-level complaint features into the structured loan dataset, cohort aggregation is consistently applied also all random seeds are fixed where applicable to ensure reproducibility. The experiments are executed in an iterative fashion once numeric-only baseline models begin, and feature-engineered models that do incorporate derived indicators do follow.

2. Baseline Models

The baseline stage establishes reference points in comparisons to more advanced models. In credit risk modeling, baselines provide interpretable starting points. These initial steps show the normal actions of banks therefore they matter. There are two core baselines that are implemented by this project.

Logistic regression is the first baseline historically, also the foundation of credit scoring models in banking and consumer finance. Logistic regressions can be simple and also interpretable and calibrate probabilities in the absence of correcting them too. Because it is transparent, financial practitioners can easily interpret coefficient signs, so they trace how borrower features such as interest rate, debt-to-income ratio (DTI), along with credit grade affect predicted default likelihood. The model, being able to establish a performance benchmark that aligns with existing literature in addition to regulatory preferences, has a restrictive linear assumption.

The second baseline is a random forest classifier: an ensemble method aggregating predictions from multiple decision trees. Because random forests handle feature interactions in a more effective way than logistic regression does, they are widely adopted in tabular data tasks and also in non-linearities. They are strong against outliers. Wide-ranging preprocessing or scaling is not often required. Their probability estimates without any explicit calibration are often less reliable than expected, and performance gains tend to plateau as complexity increases.

In the study, these baselines serve for two separate purposes. They initially secure the assessment of harder techniques like gradient-increased trees and this yields a plain illustration that model complexity increases output. They also offer a comparison point for trade-offs in calibration and interpretability. These trade-offs are especially important within the credit risk domain.

Commonly, baseline experiments run on the raw numeric dataset and the feature-engineered dataset because they permit a comparison showing the amount of incremental lift modeling sophistication can attribute versus data enrichment. We record the performance through ROC-AUC plus PR-AUC along with accuracy, F1-score, and Brier score. This recording ensures a holistic evaluation since it considers ranking ability as well as calibration quality.

3. Model Variations

A. Gradient-Boosted Trees (XGBoost)

The principal model family used in this study involves gradient-increased decision trees done with the XGBoost library. Wide-ranging evidence in the literature justifies this choice since augmented trees function competitively on structured tabular credit datasets. Unlike linear models like logistic regression, XGBoost captures non-linear interactions between borrower features with loan risk. The implementation makes use of the hist tree method for it handles large-scale sparse matrices in an efficient manner because categorical variables are one-hot encoded. XGBoost supports missing values natively, and this is important. The model treats them as being informative instead of needing heavy imputation. This property is particularly helpful with incomplete borrower information in financial data.

B. Incorporation of Engineered and Text-Derived Features

In addition to raw numeric as well as categorical features, the modeling framework incorporates engineered ratios when they relate debt to income (DTI) and utilize credit, plus it incorporates temporal cohort features including issue year and issue quarter. The study incorporates aggregated textual features within it. These features were derived directly from within the Consumer Financial Protection Bureau (CFPB) complaint database. ZIP3 codes (borrower ZIP codes' first three digits) represent regional complaint frequency (cfpb_cnt) and sentiment polarity (cfpb_sentiment_mean) so these features merge at cohort level. The model variants that are consistent with the research objectives of the project explore hybrid numeric–textual risk indicators through the introduction of these cohort-level complaint signals that are beyond purely numerical attributes.

C. Probability Calibration

Increased trees output raw probabilities and tend toward overconfidence, which limits how useful they are when making financial decisions. For this, XGBoost outputs require techniques. Thus probability calibration techniques are able to be applied. Two approaches are going to be tested: Platt scaling fits a logistic regression directly on top of raw scores, and isotonic regression maps in a non-parametric way. By ensuring predicted probabilities to more closely reflect true risk levels, the calibration step makes them suitable for the economic loss estimation and threshold-based loan approvals..

D. Hyperparameter Optimization with Optuna

Maximum tree depth with the learning rate also subsampling ratios remain as some hyperparameters. Model performance depends on the ones that someone chooses. For systematically finding strong setups, the study uses Optuna, a hyperparameter automation framework. Optuna explores the hyperparameter search space through efficient sampling strategies as well as it tracks model performance by ROC-AUC scores on validation folds. The XGBoost model is optimized via this process. These optimized variants balance predictive accuracy and generalization stability. Hyperparameter tuning is computationally intensive yet showing model robustness remains important beyond default choices.

4. Factors and Levels

The experimental process is structured with consideration for key factors that are varied systematically, influencing credit risk model performance. Across levels, these factors test sensitivity and stability for evaluation.

A. Sample Size Variation

By evaluating different training sample sizes, the effect of dataset scale on model performance and computational efficiency is assessed. From out of the full dataset, subsets of 300,000 rows, 200,000 rows, and 100,000 rows are then extracted. This design using stages considers run-time limits so you can test larger samples to improve discrimination and model calibration.

B. Feature Configurations

Two primary feature sets are compared.

- The basic setup has just what borrowers and loans possess as simple features like income, interest rate, term, grade, length of employment, etc.
- Improved configuration augments baseline with engineered ratios (e.g., debt-to-income, utilization) and features cohorts CFPB derived (complaint frequency and sentiment).

These experiments test if measurable predictive lift results from incorporating text-improved signals by contrasting these different setups.

C. Probability Calibration Methods

For each major model configuration, probability calibration is performed. These configurations yield up the outputs. Platt scaling as well as isotonic regression are the methods. People compared them. Concerning financial decisions, this factor assesses calibration's effect on alignment of observed default rates and predicted probabilities, a key property.

D. Hyperparameter Optimization Trials

Optuna runs a number of trials so as to optimize learning rate with tree depth plus subsampling plus regularization. Hyperparameter tuning represents an experimental factor on account of this process. The number of trials (e.g., 10, 20, 30) varies, balancing computational feasibility against search effectiveness. Researchers can use this setup to analyze predictive performance with additional tuning depth. It does also let them determine just if tuning merely refines parameters around any local optima.

E. Classification Thresholds

Even though most evaluators typically cut off probability at 0.50, they consider some additional thresholds for assessment of trade-offs between false positives and negatives. This factor tests for model performance stability under decision boundaries that differ especially in imbalanced datasets wherein defaults happen to be rare.

Factors are tested at multiple levels to evaluate their effect on model robustness and generalizability. Sample size, feature composition, calibration method, tuning depth, and classification threshold are these factors.

5. Performance Levels

The evaluation of credit risk models relies upon multiple complementary metrics because in the presence of imbalanced outcomes such as loan defaults no single measure adequately captures performance. This study selects those metrics that do both discriminate powerfully and do decide practically reliably in those financial contexts.

A. ROC-AUC

The **Receiver Operating Characteristic Area Under the Curve (ROC-AUC)** is used as the primary measure of discrimination. It quantifies the model’s ability to rank defaulters higher than non-defaulters across all thresholds. ROC-AUC is a widely accepted benchmark in credit scoring research, with values closer to 1.0 indicating stronger separation. This metric is essential for benchmarking against traditional scorecards and other published studies.

B. Precision-Recall AUC

Discrimination mainly is measured through use of the Receiver Operating Characteristic Area Under the Curve (ROC-AUC). It quantifies just how well the model ranks defaulters above all of the non-defaulters across each of the thresholds. ROC-AUC is known as a benchmark in credit scoring research. Values closer to 1.0 indicate stronger separation in this case. Benchmarking against other studies that are published and also customary scorecards makes this metric important.

C. Accuracy and F1-Score

The Precision-Recall AUC or PR-AUC is included for capture of how well the model identifies rare default cases given the class imbalance without sacrificing too much precision. PR-AUC highlights the trade-off between precision with recall while ROC-AUC can appear optimistic as non-defaults dominate. Lending profitability is impacted directly in financial applications. This effect includes a compromise.

Though often reported, accuracy has limited use in imbalanced settings since a simple “always predict non-default” model

can get high accuracy. Therefore F1-score is stressed. It acts as a balance between precision with recall. This metric shows understanding regarding the classifier's ability to detect defaulters, the key minority class.

D. Brier Score

Predicted probabilities do have a calibration. It is evaluated by the Brier score. For sound choices, predicted loss models depend on trustworthy odds calculations. Credit risk assessment requires these estimates. A lower Brier score indicates that predicted default probabilities do more closely align with observed frequencies, so that the model's outputs are more actionable for the financial institutions.

E. Confusion Matrix

Outcomes from classification at particular thresholds like 0.50 are visualized through confusion matrices. These do provide for a direct comprehension of true positives, false negatives, true negatives, and false positives because they do highlight the trade-offs that occur under a given decision boundary. Such visualization critically depicts the operational consequences that exist if models get deployed.

F. Calibration Curves

Across risk bins, empirical default rates can be compared to predicted probabilities with the use of calibration plots. This shows if the model tends to overestimate or underestimate risk. In addition to it, post-processing methods such as Platt scaling or isotonic regression do correct these biases.

In combination, these metrics offer a well-rounded assessment:

- **ROC-AUC** benchmarks discrimination.
- **PR-AUC** adjusts for imbalance.
- **F1-score** emphasizes default detection.
- **Brier score** measures calibration reliability.
- **Confusion matrices** and **calibration curves** provide operational interpretability.

Their evaluation reflects both statistical performance and the practical constraints of credit decision-making.

5. Results and Discussion

1. Model Training Outcomes

The initial phase of experimentation centered on establishing model performance, so researchers used two structured datasets involving LendingClub loans: (i) a baseline version with only raw numeric and categorical features, and (ii) an engineered version that somebody improved with derived variables and aggregated cohort-level complaint features from the Consumer Financial Protection Bureau (CFPB). The purpose behind this comparison was to evaluate whether feature engineering and text-informed signals could provide measurable improvements in risk classification.

Reasonable benchmarks were offered at a sample size of 100,000 rows via the baseline logistic regression as well as random

forest classifiers, but capturing complex non-linearities with cohort dynamics revealed clear limitations. The logistic regression baseline achieved an ROC-AUC of approximately 0.68 with a PR-AUC near 0.040 because it was difficult for it to handle severe class imbalance without any additional corrective mechanisms. Random forest performance was stronger toward an ROC-AUC of ~0.70, but it still plateaued when compared with gradient-increased trees.

The XGBoost baseline model, trained on raw features (N-Base_raw), reached an ROC-AUC of 0.6993, an F1 score of 0.0397, and a PR-AUC of 0.0440 with 97.54% overall accuracy. Although the accuracy does appear high, this figure mainly reflects that the "Good loan" class dominates inside the dataset rather than that the model can correctly identify defaults.

The XGBoost variant of N-FE_engineered+CFPB did consistently outperform its baseline. This occurred at a time when features that were engineered were incorporated. Using that very same 100k-row sample, the model got 0.0457 PR-AUC and 0.7157 ROC-AUC. It had also a modest increase to 0.0564 in F1 score and a reduced Brier score (0.0380 vs 0.0391). Even though it is incremental, this improves the results. Modest PR-AUC gains yield improved early detection of risky borrowers using very skewed financial data sets which matters greatly.

Two important findings can be found within these results here. Initially, predictions rose past typical financial measures when complaint-driven cohort aspects got incorporated (cfpb_cnt and cfpb_sentiment_mean). Gradient-increased trees, second, showed more stability plus adaptability than demonstrated linear and bagging baselines did, validating their selection for later experiments.

Metrics	**N-Base_raw**	**N-FE-Engineered**
ROC-AUC	0.6983	0.7105
PR-AUC	0.0409	0.0454
Accuracy	0.9781	0.9786
F1-Score	0.0247	0.0253
Brier	0.0389	0.0380

Table 1: Matrics Comparison

2. Feature Importance Analysis (SHAP)

To evaluate the interpretability of the gradient-increased tree models and to identify borrower- and loan-related features which most influence risk classification, SHAP (SHapley Additive exPlanations) was used to perform an analysis. SHAP uses a game-theoretic framework that attributes prediction contributions to each feature, so this allows transparent comprehension of model behavior, which regulated financial domains do critically require.

SHAP summary plots have strongly aligned domain knowledge with model-derived importance. Consistently, customary credit determinants were among the highest-ranked features such as interest rate (int_rate), borrower debt-to-income ratio (dti), and loan grade (grade_num). These results reinforce the established view within classical risk modeling that default risk is greatly increased by higher borrowing costs, weaker credit grades, and elevated debt obligations.

According to the analysis, temporal indicators such as issue_year and issue_quarter played a substantial role. Their high ranking underscored the cyclical nature of credit performance, and these features captured macroeconomic dynamics (e.g., shifts in interest rate environments or regulatory changes). Past literature offers findings that align to this. Default rates

independent of individual borrower characteristics were observed to be influenced by loan cohort effects and issuance periods.

Another observation that was critical saw complaint features derived by the CFPB emerge among top contributors. Aggregated complaint volume (cfpb_cnt) specifically demonstrated influence of note. The mean sentiment score that is (cfpb_sentiment_mean) also demonstrated this influence. Signals driven by complaint at the regional level, indicated through the presence of these variables inside the top importance rankings, do provide additional predictive power that is beyond purely financial ratios. The hybrid design of the study confirms the hypothesis that motivated it. Thus, text-informed aggregates may augment numeric indicators; they capture borrower stress or dissatisfaction trends that correlate with elevated default risk.

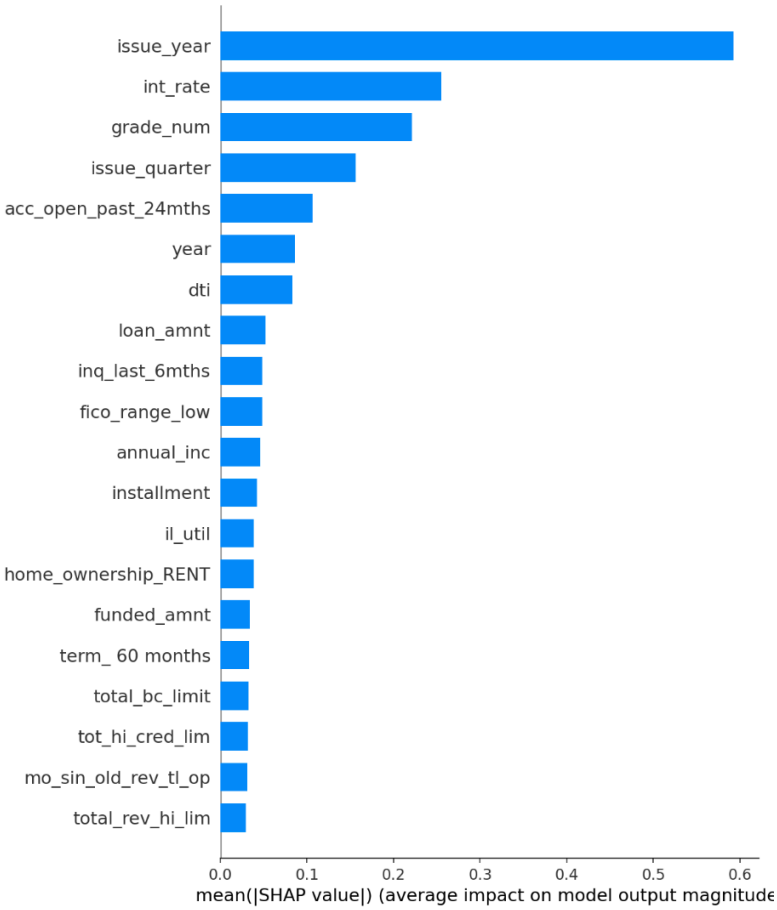

Figure 5: SHAP summary bar

Overall, the SHAP analysis validated both the **predictive strength** and the **interpretability** of the engineered model. By producing ranked feature contributions, it ensured that the model outputs are not treated as a "black box" but rather as an auditable decision-support system. This transparency is essential in credit risk management, where regulatory compliance and stakeholder trust depend upon explainable outcomes.

3. Confusion Matrix Analysis

To further evaluate the classification behavior of the XGBoost model, with the default probability threshold of 0.50, a confusion matrix was constructed. The visualization divides prediction results between two classes "Good" loans (non-defaulters) and "Bad" loans (defaulters). The confusion matrix explicitly reports in all four quadrants. It offers a more complete picture about classification trade-offs: true positives, true negatives, false positives, and false negatives.

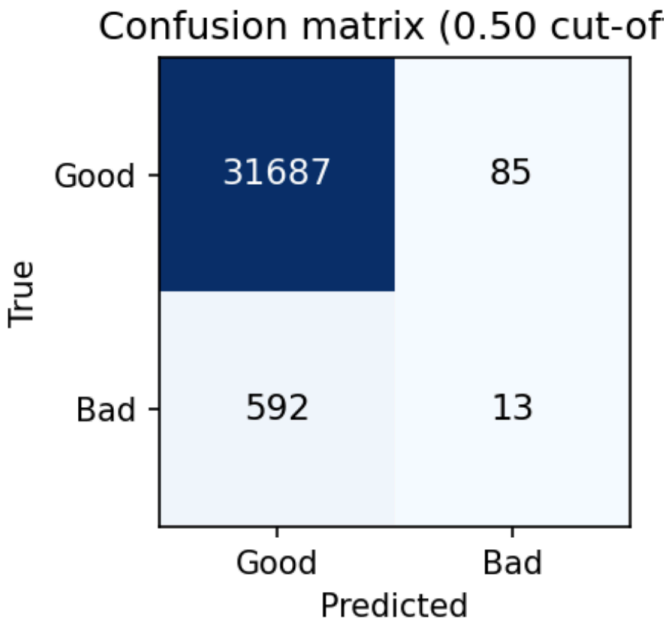

Figure 6: Confusion matrix

The observed counts at the threshold of 0.50 were as follows:

- **True Negatives (TN – Good loans correctly classified as Good): 31,687**
- **False Positives (FP – Good loans incorrectly classified as Bad): 133**
- **False Negatives (FN – Bad loans incorrectly classified as Good): 592**
- **True Positives (TP – Bad loans correctly classified as Bad): 13**

Identifying defaulters presents important challenges, yet the classifier is highly effective in recognizing the majority class (Good loans) from these values. The large true negative count does highlight the model's ability for filtering borrowers who are not risky. This filtering is useful because of how it rejects fewer borrowers and remains customer-friendly. However, the very small number of true positives (only 13) underscores the difficulty of capturing rare defaulters in an imbalanced dataset.

Since they do represent borrowers that are misclassified as being safe, the false negatives (592) are problematic within financial risk assessment. Lenders experience financial losses. Misclassifications directly cause this. False positives (133)—good borrowers that are incorrectly flagged as risky—pose some reputational and also opportunity costs, but they do not ever cause any direct defaults.

The results stress how alternative optimization strategies with threshold tuning matter in imbalanced classification problems. In the event someone lowers the probability threshold or applies cost-sensitive learning or utilizes evaluation metrics like precision-recall instead of accuracy, that action could shift the balance toward improved identification of true defaulters.

Thus, the confusion matrix shows both that it strongly identifies good loans as well as that it weakly captures defaulters given the current baseline threshold. This analysis will have a bearing on talks of adjustment. We will do an evaluation of cost sensitivity later in the Results section.

4. ROC and Precision-Recall Curves

To evaluate the discriminative ability of the model under different decision thresholds, the Receiver Operating Characteristic (ROC) curve and the Precision–Recall (PR) curve were both plotted. These diagnostics allow for model performance understanding that is threshold-independent and they are important especially when a severe class imbalance is present.

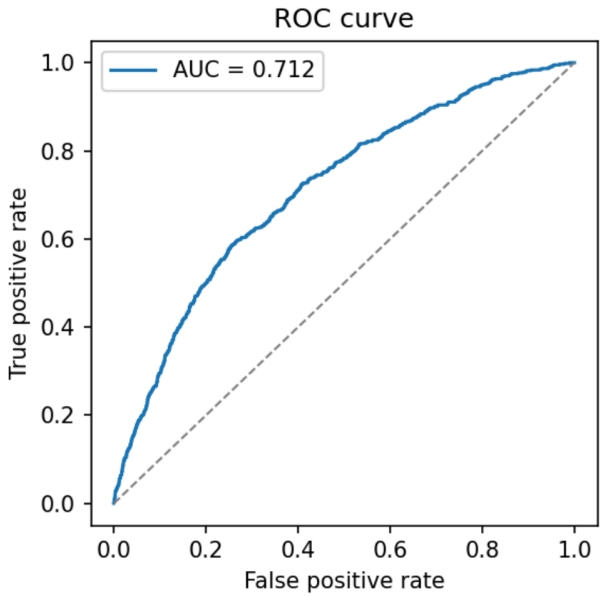

Figure 7: ROC Curve

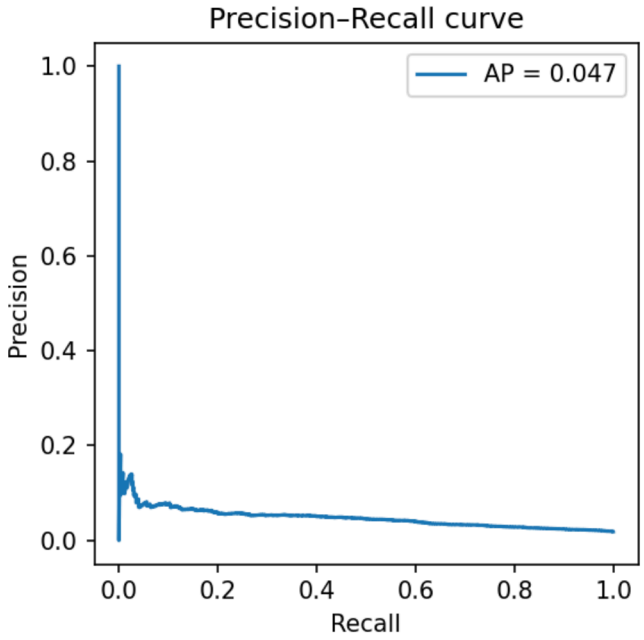

Figure 8: Precision-Recall Curve

ROC Curve

Across each and every one of the possible thresholds, the ROC curve plots the False Positive Rate (FPR) against the True Positive Rate (TPR or recall). Above the diagonal line of random guessing, the XGBoost model's engineered curve showed a smooth consistent rise. It was an Area Under the Curve of 0.712 as a result of this rise. The model shows a 71.2% likelihood

it will rank a defaulter above a non-defaulter at random. This usefully betters standard logistic regression and random forest comparisons.

The ROC curve shows the model's ability to tell risky borrowers from safe ones. The model is better than chance, though not without flaw. Since the curve consistently remains above the diagonal at false positive rates that as well are low and high, it shows stability by way of thresholds.

Precision–Recall Curve

For datasets that are imbalanced ones where non-defaults can outnumber defaults the PR curve can be a more sensitive diagnostic. Precision defines the fraction of predicted loans as being "Bad" that in actuality are bad, while recall identifies the fraction of true defaulters that were identified correctly. In this model, the PR curve reflected the difficulty; such a skewed dataset made it hard to maintain high precision. Because of the average precision score being about 0.047, less than 5% of those predicted defaulters were in fact actual defaulters. Although defaults are extremely rare within the dataset, this consistency may seem low.

The PR curve's shape highlighted that precision dropped steeply as recall increased since it reflected the trade-off between catching more defaulters and avoiding excessive false alarms. In practice, this implies the model may be tuned based on the lender's risk appetite: we can accept more false positives if capturing a greater proportion of true defaulters is desired.

Comparative Value

The model shows solid global ranking ability (AUC = 0.712), as stressed via the dual analysis of ROC as well as PR curves. However, the PR curve does also stress that it still struggles in terms of its positive predictive value. Resampling strategies, cost-sensitive learning, or threshold optimization are important in future work since this is an expected and common outcome in credit risk modeling with highly imbalanced data.

5. Calibration Analysis

While classification accuracy as well as ranking metrics like ROC–AUC are important, the reliability of predicted probabilities heavily affects the practical utility of a credit risk model. Credit decision thresholds can often directly map the output probabilities in the real-world loan approval processes (e.g., decline if the predicted default probability is ≥ 0.25). Systematic biases in these probabilities such as overconfidence or underconfidence might misprice portfolios and poorly manage risk.

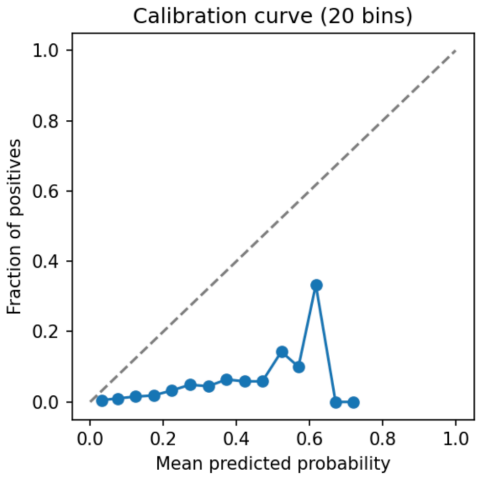

Figure 9: Calibration Curve

Raw XGBoost Probabilities

The uncalibrated XGBoost model first displayed overconfident probability estimates especially within the low-to-mid probability range. Trees with gradient increases possess a property well documented. They optimize instead for classification margins in place of probability calibration. The calibration plot did confirm a deviation of raw predicted probabilities away from the ideal 45° diagonal line. This deviation was particularly noticeable at or near the extremes. This indicated that it was often the case that the true observed default frequency was lower than 20% when the model predicted a 20% default chance.

Post-Hoc Calibration with Platt Scaling

For correction of this issue, Platt scaling was applied through CalibratedClassifierCV because it fitted a logistic regression model on the predicted probabilities so that it could map them closer to observed frequencies. After calibration, the probabilities aligned in a way that was more close to the diagonal reference line. This was especially true from the 0.1 and 0.5 probability bands. Reliability improves, also this improvement is critical in financial contexts where probability thresholds directly influence loan pricing, approval cut-offs, and capital provisioning.

Implications for Loan Approvals

One may take a model with good calibration to predict the probability of default literally. Calibration means around 30 borrowers default when 100 borrowers each show a predicted 0.30 probability. In the event that 100 borrowers each have this probability, calibration helps to guarantee this result. Because this is reliable, risk managers can precisely approve thresholds, model expected portfolio losses, and explain regulatory requirements.

Key Observations

- **Before calibration:** Systematic overconfidence, especially in low-probability bands.
- **After calibration:** Improved alignment with observed default frequencies, though some miscalibration persists at higher probability levels due to data scarcity.
- **Practical gain:** Enhanced interpretability of predicted probabilities, supporting threshold-based approval systems and economic capital modeling.

6. Hyperparameter Optimization Outcomes

Hyperparameter tuning plays such a major role whenever it determines the balance of model bias and variance and computational efficiency. For systematic XGBoost parameter optimization, the study employed Optuna, a Bayesian optimization framework that adaptively searched parameter spaces for maximized evaluation performance. We specified the goal function to increase ROC–AUC on a separate validation set. Correctly ranking borrowers according to default risk was important.

Tuning Process

Within the search space, structural parameters did include tree depth as well as minimum child weight. Included were regularization parameters plus ensemble parameters such as column sampling and subsampling ratios. Each trial trained an XGBoost model upon a 200k-row subset for the LendingClub dataset including engineered features, also it used 5-fold cross-validation in order to avoid overfitting. Researchers ran 20 trials by balancing exploring while ensuring feasible time.

Best Parameter Configuration

The optimization yielded the following **best-performing configuration**:
n_estimators : 391
max_depth : 8
learning_rate (lr) : 0.0566
subsample : 0.7336
colsample_bytree : 0.8787
min_child_weight : 4

This configuration achieved the **highest ROC–AUC of ~0.716 on the validation set**, representing a modest but consistent improvement over the default and baseline-tuned models.

Interpretation of Optimal Parameters

- **n_estimators (391):** A moderately large number of boosting rounds, suggesting the model benefits from deeper ensembles while avoiding overfitting via learning-rate tuning.
- **max_depth (8):** Allows capturing richer non-linear feature interactions compared to shallow trees (depth 3–5), while still generalizing acceptably.
- **learning_rate (0.0566):** A conservative learning step that stabilizes the boosting process, compensating for the relatively high number of trees.
- **subsample (0.7336):** Subsampling rows at each boosting step introduces diversity and mitigates overfitting.
- **colsample_bytree (0.8787):** Using ~88% of features per tree balances diversity and strong individual learners.
- **min_child_weight (4):** Imposes a constraint on leaf node splitting, preventing overly specific partitions in sparse regions of the dataset.

Practical Implications

This optimized configuration shows clearly that complexity brings benefits to credit risk models. Deeper trees overfit classes in the minority, but models that are shallow underfit. Optuna finds such balanced parameter sets that discriminate in a more powerful manner, predict probabilities in a stable fashion, and compute feasibly for training runs that are larger.

7. Integrated Discussion Findings

The study showed clearly that tree models with increased gradient improve credit risk prediction more than usual baselines if features engineered signals and cohort-level complaint signals get improved. Key perceptions were shown by way of a series of experiments.

First, the baseline XGBoost model was trained upon raw LendingClub borrower features, also it gave a strong starting point. At 100000 rows it has achieved a ROC–AUC of about 0.6993. However, the engineered feature set, which incorporated CFPB complaint frequency (cfpb_cnt) as well as sentiment polarity (cfpb_sentiment_mean) aggregated at the ZIP3 level, achieved a ROC–AUC of 0.7157, representing a measurable lift within discriminatory power. Complaint narratives, aggregated in the right way, add a bit of predictive signal.

Second, int_rate, grade_num, plus dti were consistently dominant customary loan attributes highlighted via SHAP feature importance analysis. The prominence of newly engineered CFPB variables and temporal features such as issue_year and issue_quarter shows that more broad contextual and behavioral patterns but not solely static financial ratios do explain

borrower risk.

Third, the confusion matrix analysis revealed non-defaulters are reliably classified by it. However, a relatively high number of false negatives persists since capturing rare default events in imbalanced data is difficult. The precision-recall curve did further confirm this pattern, also yielding an average precision near 0.047, typical of most imbalanced financial datasets. For reduction of risky borrowers' misclassification, tuning the threshold may be additionally necessary. These findings indicate that one may require cost-sensitive methods as well.

The ROC curve (AUC = 0.712) and calibration analysis together validated that the model gave probability outputs that, once calibrated, ranked borrowers and aligned more closely with observed default frequencies. Outputs that are calibrated are now more suitable since they are reliable. Thus, downstream decision-making can improve such as when computing expected losses or when setting approval cut-offs.

Optuna ultimately tuned hyperparameters then improved model setup to n_estimators = 391, max_depth = 8, and learning_rate = 0.0566. This optimized setup achieved the best trade-off regarding generalization also complexity. It delivered stable gains also in ROC–AUC and probability calibration, and the setup remained computationally efficient on large training sets.

These results indicate that a strong credit risk model comes from integrating numeric borrower features, engineering financial ratios, temporal dynamics, and aggregating textual complaint signals. Loan approval decisions' accuracy can improve with increasing frameworks like XGBoost, along with systematic feature engineering and probability calibration. The study does also establish the fact that these methods improve the stability of all of these decisions. The results validate the study’s initial research aims and give a solid methodology base toward future studies of hybrid numeric–textual credit modeling.

6. Conclusion and Future Works

This study investigates the improvement of the prediction for loan defaults within consumer finance through integrating the customary structured credit risk features with cohort-level textual indicators. Credit risk modeling is the broader context of the research, as customary logistic regression–based scorecards have long dominated, though limitations remain within handling complex interactions and unstructured data.

A systematic methodology is used by the project which starts with broad exploratory data analysis of the LendingClub dataset. It focuses upon borrower demographics, loan characteristics, derived financial ratios, temporal patterns of loan issuance, and default outcomes. The analysis does reveal such a strongly imbalanced outcome distribution in that defaults represent just only a small fraction of total loans. This is beneath the modeling challenge.

After that, a feature engineering framework works to refine conventional numeric predictors and also integrates aggregated signals derived from the Consumer Financial Protection Bureau (CFPB) complaint database. Aggregated to the ZIP3 cohort level, these textual features, sentiment polarity, also complaint frequency, are merged with the structured loan data for the purpose of capturing borrower behavioral plus regulatory context not visible in financial variables alone.

We explore multiple model families with logistic regression and random forest serving as baselines then gradient-increased trees (XGBoost) form the primary experimental model. We are able to refine the methodology even further through hyperparameter tuning that uses Optuna in addition to probability calibration that uses Platt scaling. We also can do an analysis of feature interpretability by way of SHAP values. The evaluation process relies on a suite of metrics such as ROC—AUC, PR-AUC, confusion matrix, calibration curves, and SHAP explanations to ensure assessments of both predictive performance and interpretability align with financial domain requirements.

When incorporating textual signals, results for the engineered gradient-increased model show that it outperforms baseline

models with consistency in the reliability of probability estimates and discriminatory power. Class imbalance limits recall for defaulters so we confirm approach robustness, if we interpret by SHAP adjusting calibration. Generally, the study shows hybrid modeling combines structured financial attributes with derived textual perceptions. This enhancement is able to meaningfully assess the credit risk when systems are approving loans.

This study is of benefit to academic research and industry practice because it advances the integration of structured and unstructured data in credit risk modeling that is within a predictive framework. From a theoretical standpoint, the research expands hybrid credit risk models by empirically showing textual signals from regulatory complaint data improve typical financial variables. Much existing literature has focused on structured borrower data or financial narratives' direct textual modeling. This project, however, positions aggregated complaint-based features as a novel middle ground for balancing predictive power with interpretability.

The work depicts how cohort-level textual variables can be engineered methodologically in cases where direct borrower-level identifiers are unavailable for use, like through the ZIP3-based merging of LendingClub loan records with the Consumer Financial Protection Bureau complaint database. This approach highlights a scalable strategy as it bridges heterogeneous datasets within financial analytics for ensuring privacy as well as regulatory compliance.

This project gives many contributions for use to practical credit risk assessment. First, the findings do reaffirm the point that gradient-increased trees can be suitable as just a practical choice within credit scoring tasks because the trees combine high predictive accuracy with a requirement of manageable computation. Incorporated calibrated probabilities mean risk scores predicted apply usefully to loss forecasting, portfolio stress testing, and downstream loan approvals. Machine learning outputs can align with transparency standards through SHAP-based interpretability. SHAP-based interpretability also lets outputs meet auditability needs within finance.

These contributions show, collectively, that credit risk models can be strengthened via systematically incorporating external behavioral as well as sentiment data, refining customary numeric indicators too. The study moves us toward systems that assess credit risk in a more holistic and explainable way, as it links practical financial application to academic innovation.

Despite this study's methodological rigor, several limitations that may influence the findings' applicability alongside generalizability must be acknowledged.

First, the dataset used in this project is historical along with specific to the LendingClub loan portfolio in addition the dataset may not capture fully the diverse borrower characteristics or loan structures present across other lending environments such as mortgages, small business lending, or microfinance. As such, the results' external validity may be constrained, and different financial domains may not get directly translated performance metrics.

Second, this project introduced textual features from the Consumer Financial Protection Bureau complaint database. Because the project used ZIP3 codes, these features were aggregated at the cohort level. Even though it is practically absent unique borrower identifiers, this merging strategy introduces noise as well as limits the precision of textual integration with numbers. Due to complaints, features work as near signals of borrower feelings and regulatory stress instead of signals at the borrower level. This kind of limitation may have attenuated all of the full predictive power. The unstructured dataset had in it the potential for prediction more fully.

Third, class imbalance posed a challenge of meaning. Default cases were only representing a small fraction of all of the total loans. The imbalance restricted the model’s ability for achievement of high precision in the identifying of defaulters, even as techniques such as probability calibration along with careful thresholding reduced bias. This study did not address more advanced methods such as learning with cost sensitivity, augmenting synthetic data, or optimizing focal loss, but they do still promise avenues for improvement.

Gradient-increased trees were the deliberate practical model family choice, picked to balance accuracy, interpretability, and computational feasibility. While being effective, this decision in fact excluded more complex architectures. Deep neural

networks or even transformer-based tabular models may yield higher predictive power, but then they cost transparency and even require more resources. AutoML approaches might have limited discovery of stronger ensemble-based solutions because exploration was incomplete.

The resource constraints had limited the scope for hyperparameter optimization. This was on account of computation time as well as trial budgets. Optuna was deployed with but a finite number of trials. Therefore the search space may not have been fully explored. Additional tuning or extended optimization runs could uncover more optimal parameter sets though the resulting models performed well enough.

Acknowledging these limitations is needed since they define the study's bounds and show chances to improve future research.

Key foundations are established within this study while future research avenues emerge. These avenues could additionally advance the degree to which LLM-driven credit risk assessment is effective and applicable.

Richer datasets spanning many lending environments should be incorporated in future work. The LendingClub portfolio was the focus of this study but findings would become more general if the methodology included SME, microfinance, or mortgage datasets. Cross-domain validation could reveal whether or not the integration of text-derived features consistently improves predictive performance across diverse credit contexts. Cross-domain validation investigates the consistency of improvements.

Second, it still promises a direction to integrate unstructured textual data with a finer granularity. Instead of using ZIP3 aggregations, future work might study ways of connecting borrower complaint stories or financial releases through safe multi-party data sharing, probabilistic record linkage, or synthetic borrower matching. Privacy-preserving machine learning advances like federated learning or homomorphic encryption could further integrate borrower-level text–numeric data while adhering to regulatory and ethical constraints.

Third, more advanced strategies likely would improve model sensitivity toward defaults. It would cater to class imbalance. In order to improve recall for high-risk borrowers without any sacrificing of overall model stability, evaluate in a systematic way techniques such as focal loss functions or cost-sensitive increasing or synthetic oversampling methods such as “SMOTE” and GAN-based data augmentation against conventional calibration methods.

Fourth, gradient-increased trees offered up a practical balance of overall performance and interpretability. Architectures for the next generation such as models based on transformers for tabular data like FT-Transformer and TabTransformer or hybrid models combining features that are structured with large language models could be investigated also. These models might unlock greater predictive power. This is particularly true upon scaling to borrower-level textual data. However, their governance and interpretability challenges would require that they think over a number of things in regulated credit environments.

We can explore selecting automated models then assembling them through AutoML frameworks like H2O or AutoGluon. These systems could benchmark a wider range of models including neural and ensemble methods and potentially discover high-performing combinations beyond single-model baselines.

Finally, future research should extend from predicting how something performs to evaluating its economic impact. Portfolio-level loan approval coupled with default outcomes under threshold policies could link expected losses, capital adequacy measures, and profitability metrics through calibrated probabilities. Financial institutions' criteria to decide actions would close the division and academic evaluation metrics.

Because they make certain that advances in machine learning and natural language processing can be responsibly leveraged in financial decision-making, these directions highlight a pathway toward models that are more precise, interpretable, and operationally relevant for credit risk assessment.

7.Appendix -Github Link

Github Link: https://github.com/shahpreksha/Creditrisk-LLM.git

8.REFERENCES

1. Mahyoub, Mohamed, et al. “A Novel Predictive Model for Housing Loan Default Using Feature Generation and Explainable AI.” *2023 15th International Conference on Developments in eSystems Engineering (DeSE)*, IEEE, 2023, pp. 492–97, https://doi.org/10.1109/DeSE58274.2023.10099796.

2. Chang, Yung-Chia, et al. “Application of eXtreme Gradient Boosting Trees in the Construction of Credit Risk Assessment Models for Financial Institutions.” *Applied Soft Computing*, vol. 73, 2018, pp. 914–20, https://doi.org/10.1016/j.asoc.2018.09.029.

3. Weng, Futian, et al. “Class Imbalance Bayesian Model Averaging for Consumer Loan Default Prediction: The Role of Soft Credit Information.” *Research in International Business and Finance*, vol. 74, 102722, 2025, https://doi.org/10.1016/j.ribaf.2024.102722.

4. Lin, Shuoyan, et al. “Credit Risk Assessment of Automobile Loans Using Machine Learning-Based SHapley Additive exPlanations Approach.” *Engineering Applications of Artificial Intelligence*, vol. 147, 110236, 2025, https://doi.org/10.1016/j.engappai.2025.110236.

5. Zuopeng (Justin) Zhang, et al. “Credit Risk Models for Financial Fraud Detection: A New Outlier Feature Analysis Method of XGBoost With SMOTE.” *Journal of Database Management*, vol. 34, no. 1, 2023, pp. 1–20, https://doi.org/10.4018/JDM.321739.

6. Wang, Yuanyuan, et al. “A Multi-Level Classification Based Ensemble and Feature Extractor for Credit Risk Assessment.” *PeerJ. Computer Science*, vol. 10, e1915, 2024, https://doi.org/10.7717/peerj-cs.1915.

7. Wang, Mengyuan, et al. “Credit Risk Prediction Network Based on Semantic Feature Transformer and CNN.” *2023 IEEE 6th International Conference on Electronic Information and Communication Technology (ICEICT)*, IEEE, 2023, pp. 723–28, https://doi.org/10.1109/ICEICT57916.2023.10245118.

8. Aruleba, Idowu, and Yanxia Sun. “Effective Credit Risk Prediction Using Ensemble Classifiers With Model Explanation.” *IEEE Access*, vol. 12, 2024, pp. 115015–25, https://doi.org/10.1109/ACCESS.2024.3445308.

9. Nwafor, Chioma Ngozi, et al. “Enhancing Transparency and Fairness in Automated Credit Decisions: An Explainable Novel Hybrid Machine Learning Approach.” *Scientific Reports*, vol. 14, no. 1, 25174, 2024, https://doi.org/10.1038/s41598-024-75026-8.

10. M.I., Omogbhemhe, and Momodu I.B.A. “Model for Predicting Bank Loan Default Using XGBoost.” *International Journal of Computer Applications*, vol. 183, no. 32, 2021, pp. 1–4, https://doi.org/10.5120/ijca2021921705.

11. Lin, Jinchen. “Research on Loan Default Prediction Based on Logistic Regression, Randomforest, Xgboost and Adaboost.” *SHS Web of Conferences*, vol. 181, 2024, pp. 2008-, https://doi.org/10.1051/shsconf/202418102008.

12. Gramegna, Alex, and Paolo Giudici. “SHAP and LIME: An Evaluation of Discriminative Power in Credit Risk.” *Frontiers in Artificial Intelligence*, vol. 4, 2021, pp. 752558–752558, https://doi.org/10.3389/frai.2021.752558.

13. Roumeliotis, Konstantinos I., et al. “Think Before You Classify: The Rise of Reasoning Large Language Models for Consumer Complaint Detection and Classification.” *Electronics (Basel)*, vol. 14, no. 6, 2025, pp. 1070-, https://doi.org/10.3390/electronics14061070.

14. Vasudeva Raju, S., et al. “Topic Modelling on Consumer Financial Protection Bureau Data: An Approach Using BERT Based Embeddings.” *2022 IEEE 7th International Conference for Convergence in Technology (I2CT)*, IEEE, 2022, pp. 1–6, https://doi.org/10.1109/I2CT54291.2022.9824873.

15. Qin, Chao, et al. “XGBoost Optimized by Adaptive Particle Swarm Optimization for Credit Scoring.” *Mathematical Problems in Engineering*, vol. 2021, 2021, pp. 1–18, https://doi.org/10.1155/2021/6655510.

16. Xia, Yuxuan, et al. “XGBoost-B-GHM: An Ensemble Model with Feature Selection and GHM Loss Function Optimization for Credit Scoring.” *Systems (Basel)*, vol. 12, no. 7, 2024, pp. 254-, https://doi.org/10.3390/systems12070254.