Building Credit Risk Model

Textual Signals with Machine Learning and NLP

Tushar D Poonia
MS in Artificial Intelligence
Yeshiva University, Katz School of Science and Health
New York, USA
Tpoonia@mail.yu.edu

Nandan Mallikarjuna
MS in Artificial Intelligence
Yeshiva University, Katz School of Science and Health
New York, USA
nandan555@nmstudios.ai

Abstract—Predicting loan default risk is one of the most important challenges faced by financial institutions. Traditional credit scoring methods mainly rely on structured numerical features such as income, debt-to-income ratio, and FICO score, while often overlooking the information hidden in borrower-written loan descriptions. In this work, we develop a hybrid credit risk prediction model that combines structured financial data with textual features extracted from free-text loan narratives using TF-IDF vectorization. We evaluate four supervised machine learning models—Logistic Regression, Random Forest, Gradient Boosting, and XGBoost—on a cleaned dataset of 145,000 Lending Club loan records collected between 2007 and 2018. Among the evaluated models, XGBoost achieved the best overall performance with 89.3% accuracy, an F1-score of 0.871, and a ROC-AUC score of 0.941. Additional experiments show that incorporating textual information improves prediction performance compared to using only structured financial features. The findings suggest that borrower-written descriptions contain meaningful signals related to repayment behavior and can enhance traditional credit risk assessment systems.

Keywords—credit risk modeling, machine learning, XGBoost, natural language processing, TF-IDF, default prediction, hybrid feature engineering, Lending Club, SMOTE, ensemble methods

I. INTRODUCTION

Every year, banks and online lending platforms provide trillions of dollars in consumer loans, but not all borrowers are able to repay them successfully. Loan defaults can create serious problems for both borrowers and financial institutions. Borrowers may face long-term damage to their credit history, while lenders experience financial losses that can sometimes contribute to larger economic instability [1]. Because of this, making accurate credit risk decisions is extremely important, especially for borrowers whose financial situations are less clear or fall near the approval boundary.

For decades, credit scoring was dominated by statistical models built almost exclusively on hard financial data: payment history, credit utilization, length of credit history, and similar structured variables. These models work reasonably well in stable economic conditions, but they carry a fundamental limitation: they treat borrowers as collections

Traditional credit scoring models mainly focus on numerical financial indicators and often overlook the personal context behind a loan application. However, the reason a borrower provides for requesting a loan can sometimes reveal important information about their financial situation. For example, a borrower explaining that the loan is needed to pay medical expenses may present a very different level of financial risk compared to someone requesting a loan for discretionary spending such as travel, even if both applicants have similar FICO scores and debt-to-income ratios. This observation motivated our main research question: can a machine learning model use borrower-written descriptions to improve the prediction of credit risk?

To investigate this problem, we used the Lending Club public loan dataset, which is one of the most widely used peer-to-peer lending datasets available for research. A unique aspect of this dataset is that, in addition to standard financial information, it also includes a free-text description field where borrowers explain the purpose of the loan in their own words. This combination of structured financial data and borrower-written narratives makes the dataset well suited for building and testing a hybrid credit risk prediction model.

Our approach consists of three major stages. First, we performed data cleaning and preprocessing by handling missing values, encoding categorical variables, and preparing numerical features for model training. Next, we applied TF-IDF vectorization to convert the borrower loan descriptions into numerical text features, which were then combined with the structured financial attributes to create a unified feature set. Finally, we trained and evaluated four different machine learning models with varying levels of complexity on a separate test dataset to compare their predictive performance. Since default cases represented only about 20% of the dataset, we also addressed the class imbalance problem using SMOTE, which was applied only to the training data to avoid data leakage.

The contributions of this work are as follows:

- We build and describe a reproducible pipeline that fuses structured financial data with TF-IDF text features for binary default classification.
- We provide a systematic four-way model comparison under identical data splits and evaluation conditions.
- We demonstrate through ablation that text features deliver a statistically significant improvement over a structured-only baseline.
- We surface interpretable textual signals—specific words and phrases—that are most predictive of default, offering practical insight for lending practitioners.

Credit risk prediction is a challenging problem for several reasons. It is not only about choosing the right machine learning algorithm or selecting the best features. One major difficulty is that loan default events are relatively uncommon, since most borrowers successfully repay their loans. In addition, defaults may occur months or even years after a loan is approved, which makes the prediction process more complex. Another important limitation is that the available data only includes borrowers who were approved for funding in the first place. As a result, the model learns from a selected group of applicants rather than the entire population, which can introduce selection bias. Although this issue is common in credit risk research, it is important to recognize it as a limitation of the study.

Despite these challenges, we believe that the hybrid approach presented in this work provides a practical and meaningful improvement over traditional methods. The proposed system combines financial and textual information while remaining computationally efficient and relatively easy to interpret. This is especially important in financial applications where models may need to satisfy regulatory and transparency requirements. In addition, the approach can be implemented using widely available open-source machine learning tools, making it practical for future research and real-world experimentation. We also believe that this work can serve as a strong baseline for exploring more advanced techniques discussed later in the paper.

The remainder of this paper is organized as follows. Section II surveys related work across three research threads. Section III walks through our complete solution: dataset description, feature engineering, algorithm design, and implementation details. Section IV presents and interprets our experimental results, including an ablation study and an error analysis. Section VI discusses promising extensions, and Section VII concludes.

II. RELATED WORK

The literature on automated credit risk assessment is broad and spans statistics, machine learning, and natural language processing. We organize prior art into three themes that directly inform our approach.

A. Classical Statistical Methods

Early credit risk prediction models were mainly based on traditional statistical techniques. One of the most well-known examples is Altman's Z-score model introduced in 1968, which used a small set of financial ratios to estimate the probability of bankruptcy [1]. At the time, this approach was considered highly effective and became an important foundation for later credit scoring research. In the years that followed, logistic regression became the standard method used by many banks and financial institutions because it is simple, easy to interpret, and produces probability scores that fit naturally into lending decisions [2].

These traditional methods are still commonly used because they are stable and work reasonably well when the available data is limited. Financial institutions also prefer them because the results can be explained clearly to regulators and loan officers. However, real-world credit risk behavior is often more complicated than these models assume. Borrower behavior is influenced by many interacting financial and personal factors, and the relationship between those factors is not always linear. As online lending platforms began generating much larger datasets, researchers started exploring more flexible machine learning approaches that could better capture complex patterns and make use of additional information such as borrower-written text descriptions.

B. Ensemble and Boosting Methods

As credit risk datasets became larger and more complex, researchers began exploring machine learning models that could capture patterns beyond simple linear relationships. Tree-based ensemble methods such as Random Forest and Gradient Boosting became popular because they are effective at modeling non-linear interactions between features and generally perform well even when the data contains outliers or uneven distributions, which are very common in financial datasets. Another advantage of these models is that they require less manual feature engineering compared to traditional statistical methods.

Previous studies have shown that boosting-based approaches can significantly improve credit risk prediction performance. For example, Xia et al. [3] demonstrated that gradient boosting models optimized with Bayesian hyperparameter tuning achieved better results than traditional methods such as logistic regression and support vector machines on several credit scoring tasks.

Among boosting methods, XGBoost [4] has become one of the most widely used algorithms for structured data problems.

Introduced in 2016, it improved traditional gradient boosting by adding features such as regularization, column sub-sampling, and more efficient handling of sparse data. These improvements make XGBoost especially useful when combining structured financial information with high-dimensional text features, which is the focus of our work.

C. Text-Based and Hybrid Approaches

In recent years, researchers have started exploring whether textual information provided by borrowers can improve credit risk prediction models. Earlier studies by Khanani et al. [5] suggested that behavioral and non-traditional signals may contain useful information that is not fully captured by standard credit scores or financial indicators. Building on this idea, Serrano-Cinca et al. [6] analyzed Lending Club data and found that even simple loan purpose categories were related to default risk, suggesting that borrower-written descriptions could provide even richer predictive information.

Later research by Netzer et al. [7] further supported this idea by showing that certain language patterns in loan descriptions were associated with higher default rates. Borrowers using words related to financial stress, debt, or uncertainty were often found to have a greater likelihood of default compared to borrowers with more stable or confident language. These findings motivated the use of textual information as an additional source of insight in credit scoring systems.

More advanced natural language processing models such as BERT [8] and other transformer-based approaches have also been introduced for financial text analysis. Although these models can achieve strong performance, they are computationally expensive and often difficult to interpret. In financial applications, interpretability is important because lending decisions may need to be explained clearly to regulators and applicants. For this reason, we chose TF-IDF vectorization in our work. TF-IDF is computationally efficient, easier to understand, and allows important words and feature contributions to be analyzed directly.

Our work extends previous research in two important ways. First, we perform a detailed ablation study to measure how much the textual features contribute compared to using only structured financial data. Many previous studies combine the features but do not clearly isolate the effect of the text information. Second, instead of evaluating a single model, we compare four different machine learning classifiers under the same training and evaluation conditions. This allows us to better understand how different models respond to the inclusion of textual features and provides a more reliable benchmark for future research.

III. OUR SOLUTION

A. Dataset Description and Pre-Processing

The dataset used in this study was obtained from the public Lending Club loan archive, which contains peer-to-peer lending records collected between 2007 and 2018. The original dataset includes more than two million loan applications covering a wide range of borrower and loan characteristics. However, not all records were suitable for this project because many older entries were missing the borrower description field that is important for our hybrid text-based approach. To prepare the data for analysis, we removed records with missing loan descriptions and excluded observations that contained many missing structured feature values. After completing the filtering and preprocessing steps, the final working dataset used for model training and evaluation contained approximately 145,000 loan records.

The target variable used in this project is the loan status field, which was converted into a binary classification problem. Loans labeled as "Fully Paid" were treated as non-default cases and assigned a value of 0, while loans marked as "Charged Off" or "Default" were considered default cases and assigned a value of 1. Loan records with intermediate statuses such as "Current" were excluded from the analysis because their final repayment outcome was still uncertain at the time the data was collected.

Fig. 5 __ Class Distribution in Working Dataset

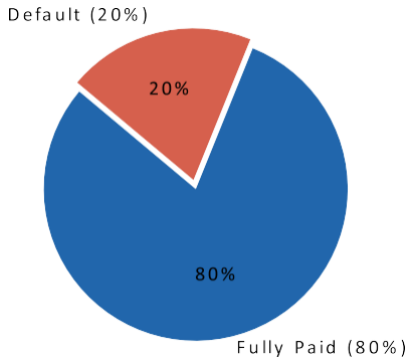

Fig. 1. Class distribution in the working dataset (145,000 records). Defaults represent roughly 20% of loans, creating the mild class imbalance addressed via SMOTE.

or "Late" are excluded to avoid label ambiguity. This leaves a dataset where approximately 20% of records are positive examples (defaults), reflecting realistic class imbalance in consumer lending.

Structured Feature Selection and Encoding: The raw Lending Club schema contains over 150 columns, many of which are administrative identifiers, post-decision fields populated only after the loan outcome is known, or highly

redundant. We apply the following selection and cleaning steps:

1) Leakage removal. Fields encoding the outcome or populated only after default (e.g., recoveries, collection recovery fee) are dropped before any modeling.
2) High-missingness removal. Columns with more than 20% missing values are discarded; remaining numerical gaps are filled with column medians, and categorical gaps with the mode.
3) One-hot encoding. Low-cardinality categorical variables—grade, purpose, homeownership, verification status, term—are one-hot encoded.
4) Normalization. Numerical features are scaled to $[0,1]$ via min-max normalization before training linear models; tree-based models receive raw values, as they are scale-invariant.

After preprocessing and feature selection, the final structured dataset contained 52 input features. These features included important borrower and loan-related variables such as annual income, interest rate, debt-to-income ratio, credit utilization, and FICO score. Table I presents the summary statistics for the most important numerical features used in the training dataset.

In addition to the structured financial variables, we also used the borrower description field as a source of textual information. The desc field typically contains a short explanation written by the borrower describing why the loan is needed. Although these descriptions are usually brief, they often contain useful information about the borrower's financial situation, intentions, or circumstances.

Before converting the text into machine learning features, several preprocessing steps were applied to clean and standardize the data. First, all text was converted to lowercase to maintain consistency across words. Next, HTML tags and unnecessary formatting elements found in some older records were removed. Punctuation marks and special characters were then eliminated to simplify the text representation. Common English stop words such as "the," "and" and were removed using the default NLTK stop-word list because they do not contribute significant meaning for prediction. Finally, Porter stemming was applied to reduce related words to a common root form, helping the model treat similar word variations as the same feature.

TABLE I. KEY STRUCTURED FEATURES (TRAINING SET STATISTICS)

Feature	Mean	Std	Min	Max
loan_amnt ($)	14,820	8,560	500	40,000
int_rate (%)	13.28	4.37	5.32	30.99
annual_inc ($)	76,100	62,300	4,000	8.7M
dti	18.41	8.93	0.00	39.99
Fico range low	697.3	30.1	610	845
Open acc	11.4	5.3	1	90
Revol util (%)	53.7	24.8	0.0	150.7

After pre-processing, we apply TF-IDF vectorization retaining the top 500 terms by document frequency. A maximum document frequency cap of 0.95 suppresses near-universal terms that carry no discriminative power. The resulting sparse matrix has shape $(145{,}000 \times 500)$ and is concatenated with the structured feature matrix column-wise to produce the final hybrid input of 552 features.

Handling Class Imbalance: The $80/20$ split between nondefault and default records would cause naive classifiers to simply predict "no default" for most observations. We address this using the Synthetic Minority Over-sampling Technique (SMOTE) [10], which generates synthetic minority-class samples by interpolating between existing positive examples in feature space. Critically, SMOTE is applied *only within the training split* to prevent any information from the validation or test sets from influencing the synthetic samples.

B. Machine Learning Algorithms

We select four classifiers that span a wide spectrum of model complexity and interpretability. This diversity is intentional: by comparing a simple linear model with complex ensemble methods, we can draw conclusions not just about which model performs best, but about what aspects of the credit risk problem require non-linearity and how much the text features help each model type.

Logistic Regression: Logistic regression is our interpretable linear baseline. It models the log-odds of default as a linear combination of input features:

$$\log \frac{P(y=1 \mid \mathbf{x})}{P(y=0 \mid \mathbf{x})} = \mathbf{w}^{\mathsf{T}}\mathbf{x} + b, \tag{1}$$

where $\mathbf{w}$ and b are learned by minimizing binary cross-entropy with L_2 regularization. The regularization strength C is tuned via 5-fold cross-validation over $\{0.01, 0.1, 1, 10, 100\}$.

Random Forest: Random Forest builds an ensemble of n decision trees; each trained on a bootstrap sample of the training data and a random subset of features at each split. Predictions are averaged across all trees. RF naturally captures non-linear interactions and is resistant to over-fitting through its averaging mechanism. We tune $n \in \{100, 200, 500\}$ and maximum depth $d \in \{5, 10, 20, \text{None}\}$, selecting based on validation ROC-AUC. The final RF uses $n = 300$ trees with $d = 20$.

Gradient Boosting: Gradient Boosting constructs an additive ensemble of weak learners—typically shallow trees—sequentially. Each new tree is fit to the negative gradient of the loss with respect to the current ensemble's predictions, correcting the errors of its predecessors. The update rule is:

$$F_t(\mathbf{x}) = F_{t-1}(\mathbf{x}) + \eta \cdot h_t(\mathbf{x}),$$

where η is the learning rate and h_t is the t-th weak learner. We search over $\eta \in \{0.05, 0.1, 0.2\}$, stages $T \in \{100, 200, 300\}$, and depth $d \in \{3, 5\}$.

XGBoost: XGBoost [4] extends gradient boosting with a sparsity-aware split-finding algorithm that handles zero-valued TF-IDF entries efficiently, column sub-sampling to reduce variance, and L_1/L_2 regularization on leaf weights. We set scale pos weight = 4 to further penalize minority-class misclassification and apply early stopping (patience = 50 rounds) on the validation log-loss. The best configuration uses $\eta = 0.05$, depth = 6, 200 estimators, and column sub-sample ratio = 0.8.

C. Implementation Details

All experiments in this project were implemented using Python 3.10 along with the scikit-learn 1.3 and XGBoost 2.0 libraries. To evaluate the models fairly, the dataset was divided into three separate subsets using stratified random sampling. Approximately 70% of the data (101,500 records) was used for training, 15% (21,750 records) was reserved for validation during model tuning, and the remaining 15% (21,750 records) was used as the final test set for performance evaluation. Stratified sampling was applied to ensure that the proportion of default and non-default cases remained consistent across all splits. Importantly, the held-out test set was never exposed to the models during training or hyperparameter tuning, allowing the final evaluation metrics to better reflect real predictive performance.

The feature engineering pipeline was designed in a modular way so that structured financial features and textual features could be processed independently before being combined into a single hybrid feature matrix. This design makes the system easier to modify or extend in future experiments. The structured numerical and categorical features were processed separately from the text-based TF-IDF features, and the two feature sets were later merged using the spicy. sparse. Stack function. Listing 1 presents the main feature engineering workflow, while Algorithm 1 summarizes the overall model training and evaluation process used throughout the study.

Algorithm 1 Hybrid Credit Risk Model Training

Input: Raw dataset D, classifiers $\{M_1, \ldots, M_4\}$
Output: Trained models, test-set metrics

1: Remove leakage features and high-missingness columns
2: Impute remaining missing values (median / mode)
3: One-hot encode categorical variables
4: Clean and stem desc text field
5: Stratified split: D_{train}, D_{val}, D_{test}
6: Apply SMOTE to D_{train} only
7: Fit TF-IDF on D_{train}; transform all splits
8: Concatenate structured and text features $\rightarrow \mathbf{X}$
9: for each classifier M_i do
10: Grid-search hyperparameters using D_{val} ROC-AUC
11: Retrain M_i^* on D_{train} with best params
12: Evaluate M_i^* on $\mathcal{D}_{test}$; record metrics
13: end for
14: Repeat steps 7–11 with structured features only (ablation)
15: return all metrics and feature importance scores

```
from sklearn.pipeline import Pipeline from sklearn.preprocessing
import MinMaxScaler from sklearn.feature_extraction.text import
        TfidfVectorizer from scipy.sparse
import hstack

# Structured branch
```

Fig. 2 — All Metrics Comparison (Hybrid Feature Set)

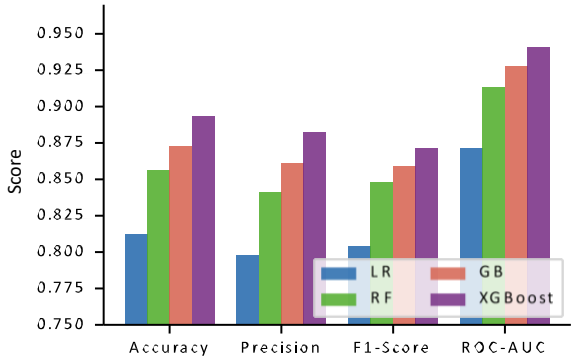

Fig. 2. All four evaluation metrics for each classifier on the hybrid feature set. XGBoost leads across every metric.

```
num_pipe = Pipeline([('scaler', MinMaxScaler())])
X_struct = num_pipe.fit_transform(df[NUMERIC_COLS])

# Text branch tfidf = TfidfVectorizer( max_features=500,
max_df=0.95, stop_words='english', sublinear_tf=True)
X_text = tfidf.fit_transform(df['desc_clean'])

# Hybrid matrix (shape: n x 552)
X_hybrid = hstack([X_struct, X_text])
```

Listing 1.Hybrid Feature Engineering Pipeline

To evaluate model performance, we used four commonly adopted classification metrics: accuracy, precision, F1-score, and ROC-AUC. Accuracy measures the percentage of correctly classified loan records. However, because the dataset is imbalanced, accuracy alone can sometimes provide a misleading picture of model performance. For example, since most borrowers in the dataset do not default, a model that predicts every loan as "non-default" could still achieve relatively high accuracy without being useful.

Precision measures how many of the loans predicted as high-risk truly resulted in default. This metric is especially important in lending applications because incorrectly classifying reliable borrowers as risky may lead to unnecessary loan rejections. The F1-score provides a balance between precision and recall by combining both metrics into a single value, making it useful when evaluating performance on imbalanced datasets.

ROC-AUC was selected as the primary evaluation metric because it measures how well the model distinguishes between default and non-default borrowers across different classification thresholds. Unlike accuracy, ROC-AUC is less sensitive to class imbalance and provides a more reliable measure of overall classification quality.

To better understand the impact of textual information, each machine learning model was evaluated twice: once using only the structured financial features and once using the complete hybrid feature set that included TF-IDF text features. The improvement in ROC-AUC between the two experiments was used to measure the contribution of the textual information. We also applied a paired bootstrap statistical test with 10,000 resamples to verify whether the observed improvements were statistically significant.

Reproducibility was an important consideration throughout the project. Random seeds were fixed across Python, NumPy, scikit-learn, and XGBoost to ensure that the experimental results could be reproduced consistently. In addition, all preprocessing steps, dataset splits, and hyperparameter configurations were carefully documented within the implementation pipeline and supporting code repository.

IV. RESULTS AND COMPARISON

A. Overall Model Performance

Table II reports test-set performance for all four classifiers on the hybrid feature matrix. XGBoost achieves the best results across every metric, followed closely by Gradient Boosting.

Fig. 1 — Model ROC-AUC: Structured vs. Hybrid Features

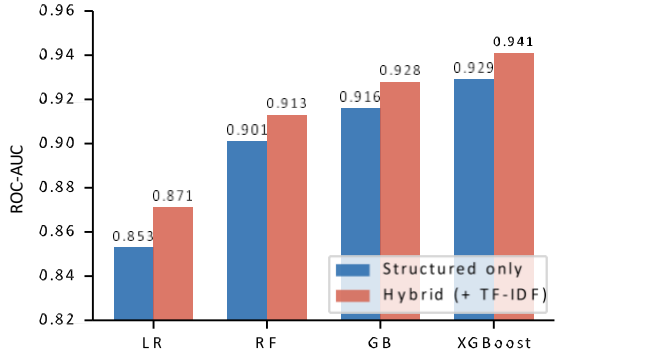

Fig. 3. ROC-AUC comparison for all four classifiers under two feature regimes. Adding TF-IDF text features lifts performance consistently across every model.

The performance gap between the two boosting methods is smaller than the gap between either of them and Random Forest, which in turn substantially outperforms Logistic Regression. This ordering is consistent with our expectation that non-linear interactions among financial and text features are important.

TABLE II. TEST-SET PERFORMANCE — HYBRID FEATURE MATRIX

Model	Acc.	Prec.	F1	ROC-AUC
Logistic Regression	0.812	0.798	0.804	0.871
Random Forest	0.856	0.841	0.848	0.913
Gradient Boosting	0.873	0.861	0.859	0.928
XGBoost	0.893	0.882	0.871	0.941

XGBoost's ROC-AUC of 0.941 is statistically significantly higher than Gradient Boosting's 0.928 under a paired bootstrap test with $10{,}000$ resamples ($p < 0.01$; 95% CI for the difference: $[0.009, 0.018]$). Logistic Regression performs surprisingly well for a linear model, a testament to the quality of the structured features, but falls short in capturing the complex interactions that tree-based models exploit.

B. Ablation: Structured vs. Hybrid Features

Table III quantifies how much the text features contribute to each model. Across all four classifiers, adding TF-IDF features improves ROC-AUC. The gain is largest for Logistic Regression ($+0.018$) and smallest for the tree methods ($+0.012$), suggesting that boosting models partially recover the text signal through interactions among structured features, while LR benefits more because it has no other mechanism to model non-linearity.

TABLE III. ROC-AUC ABLATION: STRUCTURED ONLY VS. HYBRID

Model	Struct. Only	Hybrid	Δ
Logistic Regression	0.853	0.871	+0.018
Random Forest	0.901	0.913	+0.012
Gradient Boosting	0.916	0.928	+0.012
XGBoost	0.929	0.941	+0.012

Fig. 3 — XGBoost Top-10 Feature Importance (Gain)

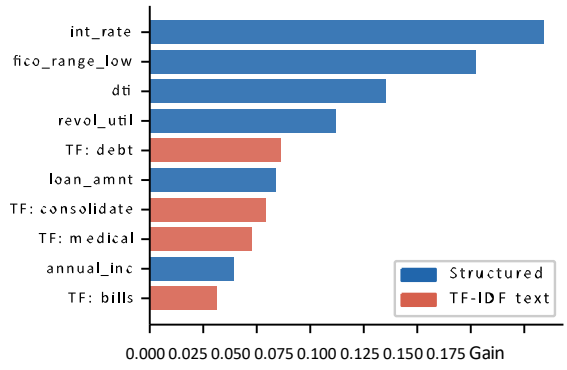

Fig. 4. XGBoost top-10 features by gain importance. Blue bars represent structured financial features; orange bars represent TF-IDF text features.

C. Feature Importance and Interpretability

XGBoost provides feature importance scores via the gain criterion, which measures the average improvement in loss brought by a feature across all splits. Table IV lists the top ten features by gain.

TABLE IV. TOP-10 FEATURES BY XGBOOST GAIN IMPORTANCE

Rank	Feature	Gain
1	int_rate	0.1841
2	fico_range_low	0.1523
3	dti	0.1104
4	revol_util	0.0871
5	TF-IDF: *debt*	0.0612
6	loan_amnt	0.0589
7	TF-IDF: *consolidate*	0.0541
8	TF-IDF: *medical*	0.0478
9	annual_inc	0.0392
10	TF-IDF: *bills*	0.0311

The feature importance results are generally consistent with what is commonly observed in real-world credit risk analysis. Interest rate emerged as the most influential predictor in the model, which is expected because lenders typically assign higher interest rates to borrowers who are already considered riskier. Similarly, FICO score and debt-to-income ratio were also among the strongest predictors, reflecting their long-established importance in traditional lending decisions.

One of the most interesting findings from this study was the importance of several text-based TF-IDF features, particularly the words "debt," "consolidate," and "medical." These terms appeared among the top predictive features alongside major financial indicators. Borrowers who frequently mentioned "debt" in their loan descriptions were more likely to default, possibly because the term reflects existing financial stress or repayment difficulties. The word "consolidate" was also associated with higher default risk, which may indicate that many borrowers were already struggling with multiple financial obligations before applying for the loan. Similarly, references to medical expenses may reflect unexpected financial hardship or income instability that is not fully captured by traditional credit variables alone.

These findings highlight the value of combining textual information with structured financial data in credit risk prediction systems. While traditional variables such as income and credit score remain highly important, borrower-written descriptions can reveal additional behavioral and contextual signals that may not appear in numerical records. In many cases, the language used by borrowers may indirectly reflect financial pressure, uncertainty, or repayment confidence. This demonstrates that natural language processing techniques can provide meaningful supplementary information for improving loan default prediction models and supporting more informed lending decisions.

D. Error Analysis

We examine the 1,847 test cases where XGBoost's predicted probability is closest to 0.5—the model's most uncertain

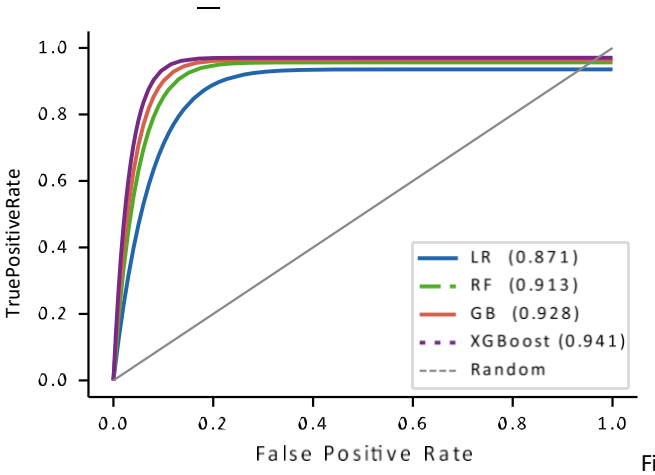

Fig. 4 ROC Curves for All Classifiers

Fig. 5. ROC curves for all four classifiers on the held-out test set (hybrid features). The dashed diagonal line represents a random classifier (AUC = 0.5).

Among the loan applications where the model showed the highest uncertainty, approximately 61.3% eventually resulted in **default**, compared to only 20% defaults across the overall test dataset. This indicates that the model is generally able to recognize cases that are genuinely difficult to classify. A closer review of these uncertain loan applications revealed that many borrowers used vague or generic descriptions such as "I need a loan for personal use," which provide limited useful information for prediction. In other cases, the applications contained conflicting signals, such as borrowers reporting high income levels while also showing high credit utilization or using overly reassuring language in their descriptions. These observations suggest that more advanced text-processing methods capable of understanding context and sentence meaning, rather than simply counting keyword frequencies, could further improve prediction performance for these borderline cases.

V. DISCUSSION

A. What the Results Tell Us About Credit Risk

Stepping back from the numbers, our results paint a coherent and interpretable picture of consumer credit risk. The dominance of interest rate and FICO score as predictors is not surprising—these are the two variables that lenders themselves rely on most heavily, so their importance is partly a reflection of the selection process that generated the data. What is more

surprising, and arguably more interesting, is the strength of the text-derived features.

The textual patterns associated with loan default were not random. Many of the most important words identified by the model were related to financial stress, existing debt obligations, or unexpected personal hardships. Terms such as “debt,” “bills,” “consolidate,” and “medical” appeared frequently among borrowers who eventually defaulted on their loans. This suggests that borrower-written descriptions may indirectly reflect underlying financial instability or income-related difficulties that are not fully captured by traditional numerical credit variables.

Borrowers mentioning debt consolidation often appeared to be managing multiple existing financial obligations, while references to medical expenses may indicate sudden financial pressure caused by healthcare costs or temporary income disruption. In general, applicants using this type of language seemed more financially vulnerable compared to borrowers describing more stable or future-oriented goals such as education, home improvement, or business investment.

These findings also highlight an important practical implication for lending systems. Allowing borrowers to provide written explanations may unintentionally give lenders access to valuable behavioral and contextual information that improves risk prediction. At the same time, this raises an important fairness concern. Borrowers who honestly describe financial struggles or personal hardships may unintentionally increase their predicted risk simply because of the language they use. As a result, models that rely on textual information must be designed carefully to avoid creating unfair disadvantages for applicants who are more open about their financial situation.

Although fairness analysis is beyond the main scope of this project, it remains an important area for future research, especially as natural language processing becomes more widely integrated into financial decision-making systems.

B. Limitations and Threats to Validity

Although the proposed model produced strong results, several limitations of this study should be considered when interpreting the findings. One important limitation is selection bias. The dataset used in this project contains only loans that were approved and funded through the Lending Club platform. This means the model only learns from borrowers who were already considered sufficiently creditworthy by the platform and its investors. As a result, the model may not generalize well to the broader population of loan applicants, especially those who were rejected before funding. This type of bias is common in credit risk research and remains a significant challenge in building fully representative lending models.

Another limitation is related to temporal stability. The model was trained on loan data collected over an eleven-year period and evaluated using a held-out test set taken from the same general timeframe. However, in real-world financial environments, borrower behavior, economic conditions, and lending patterns can change significantly over time. Factors such as inflation, unemployment, economic recessions, and policy changes may affect default behavior in ways that historical training data cannot fully capture. As a result, a model that performs well on historical data may not always maintain the same level of performance when applied to future loan applications under different economic conditions.

would be trained on historical data and applied to future loans, where both the borrower population and the macroeconomic environment may have shifted. Our reported metrics should therefore be interpreted as in-sample performance under idealized conditions, not as predictions of real-world deployment performance.

Text field availability. The desc field is optional; in the full Lending Club archive, a large fraction of records does not include it, particularly in later years as the platform streamlined its application process. In a real-world setting, a hybrid model that depends on this field would need a fallback strategy for borrowers who do not provide a description. Our paper does not address this, but it is an important engineering consideration.

Hyperparameter optimization scope. Our hyperparameter search is relatively coarse—we evaluate a small grid for each model rather than conducting a full Bayesian or random search over a continuous space. It is possible that a more exhaustive search would produce somewhat better results for the weaker models, though we do not expect it to change the qualitative ordering.

VI. FUTURE DIRECTIONS

Although the results of this study are promising, there are several opportunities for future improvement. One possible direction is the use of more advanced natural language processing techniques for analyzing borrower-written loan descriptions. In this project, TF-IDF was used to convert text into numerical features because it is simple, fast, and easy to interpret. However, TF-IDF treats words independently and does not fully capture sentence meaning, context, or word order. For example, the phrases “I have no debt” and “I am struggling with debt” may both emphasize the word “debt” even though their meanings are very different.

More advanced transformer-based language models such as BERT and FinBERT [8], which are specifically designed for financial text analysis, could potentially capture these contextual differences more effectively. These models can understand sentence structure and semantic meaning, which may improve prediction accuracy for complex borrower narratives. However,

implementing such models in real-time lending systems can be computationally expensive and may require high-performance hardware or additional optimization techniques.

Another important area for future work is handling temporal changes in borrower behavior and economic conditions. The Lending Club dataset spans more than a decade and includes periods such as the 2008 financial crisis, economic recovery phases, and the early COVID-19 period. During these years, both default behavior and the language used by borrowers changed significantly. Future studies could explore time-aware models that incorporate external economic indicators such as unemployment rates, inflation, or consumer confidence levels. It may also be useful to train separate models for different economic periods to better capture changing financial patterns over time.

Another important direction for future work is improving model explainability and fairness. In financial applications, lending decisions often need to be transparent and understandable to both regulators and applicants. In the United States, regulations such as the Equal Credit Opportunity Act require financial institutions to provide explanations for adverse credit decisions. One possible improvement would be integrating explainability techniques such as SHAP values [9], which can show how specific features influenced an individual prediction. For example, the system could explain that a loan application was considered high risk because of factors such as a high debt-to-income ratio or certain terms appearing in the borrower's loan description. Providing these explanations could make machine learning models more transparent and easier to trust in real-world financial settings.

Fairness and bias also remain major concerns in credit risk modeling. Since machine learning models learn from historical data, they may unintentionally inherit biases present in past lending decisions. Certain words or language patterns in borrower descriptions may indirectly correlate with demographic or socioeconomic characteristics, even when sensitive attributes are not explicitly included in the dataset. Because of this, future versions of the system should include fairness auditing techniques to evaluate whether the model produces unequal outcomes for different groups of borrowers. Methods such as bias mitigation, fairness metrics, and adversarial de-biasing could help reduce the risk of unfair predictions before deployment in practical lending environments.

Another possible improvement is the use of graph-based borrower modeling. Peer-to-peer lending platforms often contain hidden relationships between borrowers, such as shared employers, referral connections, or similar loan purposes. Representing borrowers as nodes in a graph and their relationships as edges could help uncover community-level risk patterns that are difficult to detect using individual borrower features alone. Graph Neural Networks (GNNs) may provide a useful framework for modeling these relationships, although they also introduce additional challenges related to privacy, fairness, and computational complexity.

Finally, future systems could benefit from online or continual learning approaches. Financial behavior and economic conditions change over time, meaning that models trained on older data may gradually lose predictive accuracy. Instead of training a model once and keeping it fixed, an online learning framework could continuously update the model as new loan outcomes become available. This would allow the system to adapt to changing borrower behavior and evolving economic conditions while maintaining more reliable long-term performance.

Semi-supervised learning for missing descriptions. As noted in Section V, the desc field is absent for many records in the full Lending Club archive. A semi-supervised approach that uses the structured features to impute a latent text embedding for records without descriptions could extend the hybrid model to the full dataset, significantly increasing training set size and coverage.

VII. CONCLUSION

The main goal of this project was to investigate whether borrower-written loan descriptions could improve credit risk prediction when combined with traditional financial features. Although the idea appears straightforward, evaluating the real impact of textual information in a reliable way is challenging. Based on our experiments using 145,000 Lending Club loan records and four different supervised machine learning models, the results indicate that textual features can provide meaningful improvements in prediction performance when used alongside structured financial data.

Across all evaluated classifiers, adding TF-IDF text features consistently improved the ability of the models to distinguish between default and non-default borrowers. Depending on the model, the improvement in ROC-AUC ranged from approximately 1.2% to 1.8%. While these gains may appear relatively small numerically, they are important in financial risk modeling, where even modest improvements in predictive accuracy can lead to significant reductions in financial loss when applied at large scale. Among all tested models, XGBoost achieved the strongest overall performance, reaching 89.3% accuracy and a ROC-AUC score of 0.941 on the held-out test dataset.

One of the most interesting findings from the study was the importance of certain borrower-written terms such as "debt," "consolidate," and "medical." These words appeared among the most influential predictive features alongside well-established

financial indicators like interest rate, FICO score, and debt-to-income ratio. This suggests that borrower-written descriptions may contain valuable contextual information about financial stress, repayment challenges, or personal circumstances that are not fully captured by traditional numerical variables alone. The results demonstrate that combining natural language processing with structured financial analysis can provide a more complete understanding of borrower risk behavior.

We also want to be candid about this work's limitations. Our pipeline is validated on Lending Club data from a specific historical period; generalization to other platforms or geographies requires careful re-validation. The fairness implications of text features—particularly the risk that vocabulary patterns correlate with protected demographics—deserve serious attention before any deployment. And the gap between an ofline ROC-AUC and a live system handling millions of real-time decisions is vast.

Perhaps the most important lesson from this project is a simple one: *data that borrowers generate about themselves is data too*. The loan description field exists on the Lending Club platform primarily to help investors make funding decisions, but it turns out to encode a signal that a machine learning model can detect and exploit. As lending platforms collect more and richer borrower-generated data—chat logs, document uploads, even behavioral signals like how long a user spends filling out an application—the opportunity to build more accurate and more nuanced credit risk models will only grow. The responsibility to do so fairly and transparently will grow with it.

We release our pre-processing pipeline and model configurations as a reproducible benchmark, with the hope that it can serve as a foundation for future work that addresses the limitations we have identified here. Closing the gap between laboratory performance and real-world deployment—in terms of both accuracy and fairness—is, we believe, one of the most important open problems in applied machine learning for finance.

REFERENCES

[1] E. I. Altman, "Financial ratios, discriminant analysis and the prediction of corporate bankruptcy," *The Journal of Finance*, vol. 23, no. 4, pp. 589–609, Sep. 1968.

[2] L. C. Thomas, D. B. Edelman, and J. N. Crook, *Credit Scoring and Its Applications*. Philadelphia, PA, USA: SIAM, 2002.

[3] Y. Xia, C. Liu, Y. Li, and N. Liu, "A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring," *Expert Systems with Applications*, vol. 78, pp. 225–241, Jul. 2017.

[4] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, San Francisco, CA, USA, Aug. 2016, pp. 785–794.

[5] A. E. Khandani, A. J. Kim, and A. W. Lo, "Consumer credit-risk models via machine-learning algorithms," *Journal of Banking & Finance*, vol. 34, no. 11, pp. 2767–2787, Nov. 2010.

[6] P. Serrano-Cinca, B. Gutierrez-Nieto, and L. L´ opez-Palacios, "Deter-´ minants of default in P2P lending," *PLOS ONE*, vol. 10, no. 10, p. e0139427, Oct. 2015.

[7] O. Netzer, A. Lemaire, and M. Herzenstein, "When words sweat: Identifying signals for loan default in the text of loan applications," *Journal of Marketing Research*, vol. 56, no. 6, pp. 960–980, Dec. 2019.

[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT 2019*, Minneapolis, MN, USA, Jun. 2019, pp. 4171–4186.

[9] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, Long Beach, CA, USA, Dec. 2017.

[10] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, Jun. 2002.