Predictive Analysis of Loan Defaults

Project Teammates

Name – UNI:

Xianglong Bai - xb2166

Haochu Yang - hy2812

Jiachen Tang - jt3453

Jiayu Chen - jc6116

Xinyang Chen - xc2713

Introduction

In this project, we aim to develop a predictive model that can determine the likelihood of loan defaults based on various borrower characteristics and loan details. The dataset provided has been meticulously pre-processed, scaled and normalized for this purpose, featuring 23,814 entries and 22 attributes per entry, including both borrower-specific information and loan-specific variables from Kaggle. The dataset includes:

index: Unique identifier for the row entries.

person_age: Normalized age of the person.

person_income: Normalized income of the person.

person_emp_length: Normalized employment length of the person.

loan_amnt: Normalized loan amount requested.

loan_int_rate: Normalized interest rate of the loan.

loan_percent_income: Loan amount as a percentage of the person's income, normalized.

cb_person_cred_hist_length: Normalized credit history length of the person.

person_home_ownership: One-hot encoded home ownership status with categories for MORTGAGE, OTHER, OWN, and RENT.

loan_intent: One-hot encoded loan intent with categories such as DEBTCONSOLIDATION, EDUCATION, HOMEIMPROVEMENT, MEDICAL, PERSONAL, and VENTURE.

cb_person_default_on_file: One-hot encoded indicator if the person has a default on file, with categories N and Y.

loan_grade: Categorical grade of the loan, normalized.

loan_status: Target variable indicating loan status (0 or 1).

The primary objective of this project is to develop a predictive model that can effectively determine the likelihood of loan defaults using historical data. By understanding which factors most significantly affect the likelihood of default, the project seeks to provide insights that could help financial institutions with risk management. Another key goal is to establish a risk assessment framework that dynamically adjusts loan terms based on the predicted risk of default, thereby optimizing financial operations, and reducing losses due to defaults.

In the contemporary landscape of financial services, the assessment and management of credit risk stand as pivotal tasks for both lenders and borrowers. With the proliferation of data and advancements in machine learning techniques, there exists an opportune moment to enhance credit risk modeling and analysis methodologies. This paper embarks on a journey to explore the efficacy of various models and methodologies in credit risk assessment, leveraging insights from both traditional statistical approaches and cutting-edge machine learning algorithms.

Our research delves into the realm of credit risk modeling and analysis, aiming to address critical questions surrounding the selection and evaluation of models, the integration of machine learning for enhanced assessment, and the predictive capabilities of these methodologies. We draw upon the findings of seminal works such as “Comparison of

Selected Models of Credit Risk by Tomáš Klieštik and Juraj Cúg, and “Evaluating Credit Risk Models” by Jose A. Lopez and Marc R. Saidenberg, which offer comprehensive examinations of established credit risk models and evaluation techniques.

Furthermore, we incorporate recent advancements highlighted in studies like “Machine Learning for Enhanced Credit Risk Assessment: An Empirical Approach” by Suhadolnik, Ueyama, and Da Silva, and “Credit Risk Prediction Based on Machine Learning Methods” by Y. Li. These works underscore the potential of machine learning algorithms to augment traditional approaches, providing deeper insights and more accurate predictions in credit risk assessment.

Moreover, our exploration extends to the realm of consumer credit risk models, as elucidated by Khandani, Kim, and Lo in their work “Consumer Credit Risk Models via Machine-Learning Algorithms.” By examining consumer-centric perspectives, we aim to tailor our analysis to the intricacies of individual credit assessment, thereby enriching our understanding and applicability in real-world scenarios.

In this paper, we synthesize insights from these diverse sources to inform our project goals of developing and evaluating credit risk models. Through a comparative analysis of selected methodologies, we seek to contribute to the ongoing discourse surrounding effective credit risk management practices, ultimately striving to enhance decision-making processes for lenders and borrowers alike.

Literature Review

Comparison of Selected Models of Credit Risk

Credit risk involves the potential for a company to suffer financial losses if a business partner fails to meet their obligations, leading to default. This failure can stem from various sources such as credit agreements, trade deals, investments, payment systems, or trade

settlements. Modeling credit risk is complex due to the rarity and unpredictability of defaults, yet they can result in significant losses for creditors, which are difficult to anticipate. Many studies and publications delve into understanding and quantifying credit risk, each with its own methodology.

This study aims to dissect and compare four main approaches for characterizing and measuring credit risk: CreditRisk+, Credit Metrics, Merton model, and Credit Grades. The comparison will consider factors like computational efficiency, suitability for different company types (public or privately traded), volatility of credit events, correlation among these events, required input data, data currency, and more. The findings will provide insights and recommendations for applying these approaches in specific scenarios.

Merton Model: The Merton method for assessing credit risk provides versatility by utilizing stock market data, making it applicable to any publicly traded company. However, practical implementation has shown weaknesses, like its tendency to underestimate credit spreads compared to actual values. Extensions like the KMV model aim to address these issues.

Credit Metrics: CreditMetrics, a component of JP Morgan's Credit Manager software, focuses on valuing bond prices and employs an empirical approach known as Value at Risk (VaR). It emphasizes historical transition matrices and forward prices, assessing potential funding loss in worst-case scenarios. While it incorporates elements of both structural and reduced-form models, it primarily relies on empirical market data.

CreditRisk+: Developed by Credit Suisse First Boston, CreditRisk+ applies actuarial methods to directly model credit risk, specifically focusing on default probabilities using Bernoulli's distribution. It offers concise loss probability results for loan portfolios and individual borrowers without assuming specific causes of default. However, it overlooks migration or market risks of borrowers, limiting its realism.

Credit Grades: Introduced by RiskMetrics, Credit Grades aims to provide a transparent, standardized model for evaluating credit risk of publicly traded firms. It quantifies credit risk based on credit spreads rather than default probabilities, utilizing market observatory data to capture market dynamics. The model assumes a stochastic process for asset value evolution, defining default as the point where asset value falls below a predetermined threshold.

We will surely try some of these models for our project, because we prefer to focus on model selection for a better model performance.

Reference

Tomáš Klieštik, and Juraj Cúg. “Comparison of Selected Models of Credit Risk.” *Procedia Economics and Finance*, vol. 23, 2015, pp. 356 – 361.

Evaluating credit risk models

The article “Evaluating Credit Risk Models” by Jose A. Lopez and Marc R. Saidenberg, published in the Journal of Banking & Finance in 2000, discusses the burgeoning interest and development in credit risk models by commercial banks over the past decade. These models aim to quantify financial risks and determine the necessary economic capital more accurately. Their development has been encouraged by bank regulators, such as the 1997 Market Risk Amendment (MRA) to the Basel Capital Accord, which incorporated banks’ internal market risk models into regulatory capital calculations.

Credit risk models focus on the variability in debt instruments and derivatives due to changes in borrowers and counterparties’ credit quality. The article highlights the challenges in validating these models, primarily due to the long forecasting horizons that limit the number of available forecasts, and the scarcity of historical data on credit losses across different economic cycles. This makes it difficult to test the models’ forecast accuracy effectively.

To address these challenges, Lopez and Saidenberg propose using a panel data approach that involves cross-sectional resampling techniques, allowing for a broader evaluation of models not only over time but also across simulated credit portfolios at a given time. By generating forecasts for these portfolios, one can apply various statistical methods to evaluate the models’ performance. This approach helps in overcoming the limited historical data on credit defaults and rating migrations, providing quantifiable measures of forecast accuracy that are crucial for model validation and for making informed decisions regarding credit portfolio management.

The article also discusses general issues in credit risk modeling, including the rapid development of the field and the different assumptions and methodologies used by various models. It emphasizes the importance of accurately forecasting the probability distribution of

losses from a bank's credit portfolio, which is essential for effective risk management and regulatory purposes.

In conclusion, Lopez and Saidenberg's work presents a methodological framework for evaluating credit risk models using simulated credit portfolios, addressing a significant challenge in the field. Their approach not only facilitates the quantitative assessment of model performance but also offers insights into improving risk management practices and regulatory standards related to credit risk.

Reference

Lopez, Jose A., and Marc R. Saidenberg. "Evaluating Credit Risk Models." *Journal of Banking & Finance*, vol. 24, 2000, pp. 151-165.

Machine Learning for Enhanced Credit Risk Assessment

In this article the authors systematically approach the selection, comparison, and analysis of machine learning models to effectively evaluate credit risk. They begin this empirical journey by reviewing relevant literature to identify a diverse set of machine learning algorithms previously applied in similar contexts. Their selection includes both traditional models like Logistic Regression and Decision Trees, as well as advanced ensemble techniques such as Random Forests, Gradient Boosting, and XGBoost. This selection aims to include a broad range of methodologies, from conventional to state-of-the-art, ensuring a comprehensive comparison.

The authors tackle the imbalance of data by employing resampling techniques like RandomUnderSampler to balance the dataset. The comparative analysis of the models is based on a robust evaluation framework employing 5-fold cross-validation. This technique enhances the reliability of the performance metrics, which include accuracy, precision, recall, F1 score, and the Area Under the Curve (AUC). These metrics provide a multi-dimensional view of model performance, essential for a nuanced understanding of each model's capabilities and limitations.

By Comparing the performance of different machine learning algorithm, the author reveals that ensemble models, particularly XGBoost, exhibit superior performance in credit risk classification tasks. The authors attribute this to the ensemble methods' ability to aggregate insights from multiple decision trees, capturing complex patterns in the data that might elude simpler models. This finding highlights the potential of advanced machine learning techniques in extracting nuanced insights from complex datasets.

Beyond mere model performance, the study delves into the practical implications of their findings. The superior performance of ensemble models in credit risk assessment tasks suggests their potential for real-world applications. However, the authors also caution about

the interpretability challenges associated with some advanced models. They advocate for a balanced approach that considers not only predictive accuracy but also model transparency and computational efficiency.

In summary, the study presents a comprehensive approach to selecting, evaluating, and analyzing machine learning models for credit risk assessment. Through meticulous data preparation, rigorous model comparison, and thoughtful consideration of practical implications, the authors contribute valuable insights to the field of financial risk management. Their findings underscore the effectiveness of ensemble methods in credit classification and offer guidance for leveraging machine learning in financial decision-making processes.

Reference

Suhadolnik, N., Ueyama, J., & Da Silva, S. (2023). Machine Learning for Enhanced Credit Risk Assessment: An Empirical Approach. *Journal of Risk Financial Management*, 16(12), 496. https://doi.org/10.3390/jrfm16120496

Credit Risk Prediction Based on Machine Learning Methods

Credit risk prediction has always been an important aspect of a financial institution's business, with the aim of assessing the likelihood of a borrower defaulting on a loan. Traditionally, this assessment has been conducted using subjective methods that rely heavily on the personal judgment of credit analysts. However, with the advent of statistical methods and advances in machine learning techniques, the approach to predicting credit risk has evolved. The article "Credit Risk Prediction Based on Machine Learning Methods" by Yu Li provides a comprehensive overview of this transition, focusing on the application of machine learning methods, in particular the XGBoost algorithm, to this area. The results show that the XGBoost algorithm is more accurate than traditional logistic regression models in predicting credit risk.

The paper outlines the progression from subjective assessment methods to statistical models like logistic regression, which has been a staple in credit scoring since the 1960s. Logistic regression offered a more objective and quantitative approach to credit risk assessment but came with limitations, including the requirement for data to meet certain assumptions and a cap on its discriminative power. Over the past four decades, as highlighted in the paper, machine learning methods have rapidly developed, offering more sophisticated tools for credit risk prediction, such as neural networks, decision trees, and ensemble methods like XGBoost.

Yu Li's research conducts a detailed comparative study between logistic regression and XGBoost in the context of credit risk prediction. The study leverages a dataset with both "good" and "bad" credit samples, applying feature engineering to enhance predictive modeling. The findings demonstrate XGBoost's superior performance over logistic regression in terms of model discrimination (as measured by AUC and KS statistics) and stability (evaluated using PSI values). This aligns with broader literature indicating the

effectiveness of ensemble methods and advanced machine learning algorithms in handling complex, nonlinear relationships within data.

This paper underscores the potential of advanced machine learning methods to improve credit risk prediction. The good performance of the XGBoost algorithm not only illustrates the advancements in machine learning but also encourages further exploration into other sophisticated ML techniques for credit risk assessment. Additionally, the research highlights the importance of feature engineering in enhancing model performance, a critical aspect that future studies could delve deeper into.

Currently, researchers have begun to test the application of complex machine learning models in predicting credit risk. Yu Li's work provides insights into the effectiveness of machine learning (specifically the XGBoost algorithm) in predicting credit risk, validates the transition toward using more advanced analytics in financial risk assessment, and lays the groundwork for future research to explore the untapped potential.

Reference

Y. Li, "Credit Risk Prediction Based on Machine Learning Methods," 2019 14th International Conference on Computer Science & Education (ICCSE), Toronto, ON, Canada, 2019, pp. 1011-1013, doi: 10.1109/ICCSE.2019.8845444.

Consumer Credit-Risk Models via Machine-Learning Algorithms

The article “Consumer Credit-Risk Models via Machine-Learning Algorithms” from the Journal of Banking & Finance by Amir E. Khandani, Adlar J. Kim, and Andrew W. Lo presents an innovative approach to predicting consumer credit risk. This comprehensive study, conducted using data from 2005 to 2009 from a major commercial bank, utilizes machine-learning techniques to develop advanced, non-linear forecasting models for consumer credit risk, showcasing a notable improvement over traditional method.

The research addresses the pivotal role of consumer spending in macroeconomic conditions and systemic risks. It emphasizes the substantial impact of consumer lending practices on the economy, underscoring the necessity for accurate and robust credit risk assessment tools. A combination of transaction-level data, credit bureau information, and account balance data is employed, adhering to strict data security and confidentiality protocols. The comprehensive dataset allows for an in-depth analysis of consumer behavior and credit risk.

The study focuses on identifying key variables that indicate credit risk, such as balance-to-income ratios and sudden income drops. By employing classification and regression trees (CART), the study effectively handles high-dimensional data and uncovers non-linear relationships, which traditional credit scoring models might overlook. The research demonstrates that machine-learning models substantially outperform conventional credit scoring systems like the CScore, particularly in accurately forecasting credit risk during periods of economic uncertainty.

The study’s findings are instrumental in identifying high-risk customers, enabling banks to make informed decisions about credit line adjustments. The research also highlights the potential of using aggregated consumer credit risk forecasts as a tool for macroprudential risk management. The paper concludes that machine-learning models present a significant

advancement in consumer credit risk analysis. These models offer more accurate, real-time assessments, which are crucial for effective risk management in the financial sector.

In summary, this article makes a substantial contribution to the field of consumer credit risk assessment. It not only introduces a novel application of machine-learning algorithms but also illustrates their superiority in predicting credit risk, thus opening new avenues for risk management in banking and finance.

Reference

Khandani, A. E., Kim, A. J., & Lo, A. W. (2010). Consumer credit risk models via machine-learning algorithms. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.1568864

Bibliography Including Our Chosen Papers

Tomáš Klieštik, and Juraj Cúg. "Comparison of Selected Models of Credit Risk." *Procedia Economics and Finance,* vol. 23, 2015, pp. 356 – 361.

Lopez, Jose A., and Marc R. Saidenberg. "Evaluating Credit Risk Models." *Journal of Banking & Finance*, vol. 24, 2000, pp. 151-165.

Suhadolnik, N., Ueyama, J., & Da Silva, S. (2023). Machine Learning for Enhanced Credit Risk Assessment: An Empirical Approach. *Journal of Risk Financial Management*, 16(12), 496. https://doi.org/10.3390/jrfm16120496

Y. Li, "Credit Risk Prediction Based on Machine Learning Methods," 2019 14th International Conference on Computer Science & Education (ICCSE), Toronto, ON, Canada, 2019, pp. 1011-1013, doi: 10.1109/ICCSE.2019.8845444.

Khandani, A. E., Kim, A. J., & Lo, A. W. (2010). Consumer credit risk models via machine-learning algorithms. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.1568864

Ashofteh, A. (2023). Big data for credit risk analysis: Efficient Machine Learning models using PySpark. *Contributions to Statistics*, 245–265. https://doi.org/10.1007/978-3-031-40055-1_14

Hand, D. J., & Henley, W. E. (1997). Statistical Classification Methods in Consumer Credit scoring: A Review. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 160(3), 523–541. https://doi.org/10.1111/j.1467-985x.1997.00078.x

Goldberg, L. R., Kercheval, A. N., & Lee, K. (2005). t-statistics for weighted means in credit risk modeling. *The Journal of Risk Finance*, 6(4), 349–365. https://doi.org/10.1108/15265940510613688

Methodologies

In terms of methodologies, the project will employ a range of statistical and machine learning techniques to build and validate predictive models. Initial data processing steps include normalization and one-hot encoding to prepare the dataset for analysis. We plan to implement several machine learning models to evaluate their performance and suitability for predicting loan defaults. These models include:

- **Logistic Regression**: A robust statistical model that estimates probabilities using a logistic function, widely used for binary classification tasks like predicting loan defaults.
- **Decision Trees**: Useful for their ease of interpretation and implementation. They help in identifying critical decision rules from data.
- **K-Nearest Neighbors (KNN)**: A simple yet powerful model that makes predictions based on the proximity to the nearest data points.
- **Support Vector Machines (SVM)**: Effective in high-dimensional spaces, which is ideal for our dataset with many variables.
- **Random Forest**: An ensemble method that improves predictive accuracy by reducing overfitting through averaging multiple decision trees.
- **XGBoost**: An implementation of gradient boosted decision trees designed for speed and performance, which is particularly effective for large datasets and complex features.

Logistic Regression

The logistic regression model is well-suited for the task of credit risk assessment due to its ability to model binary outcomes, provide probabilistic outputs, and offer interpretable results. To fully evaluate the model's performance, reviewing specific metrics like accuracy,

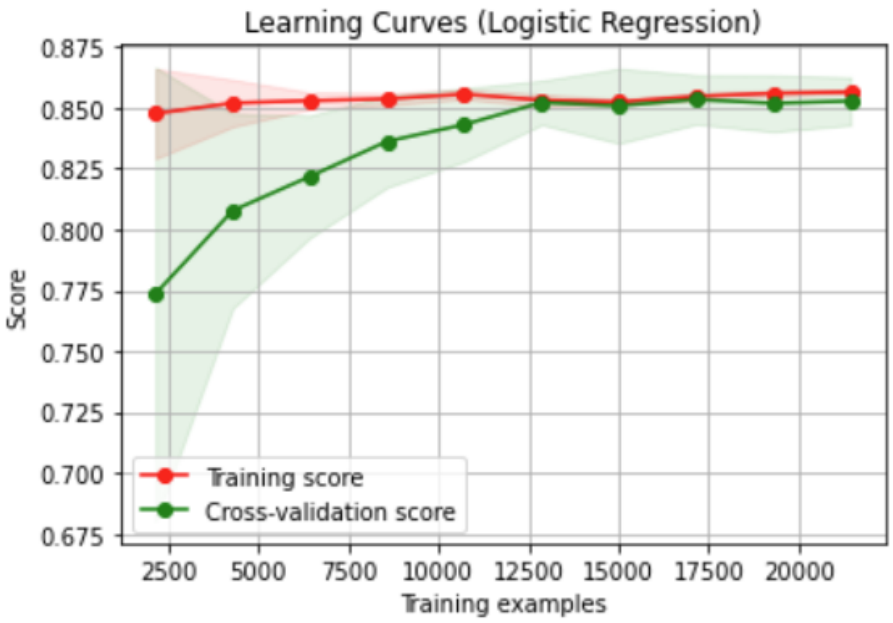
precision, recall, and the AUC for the ROC curve is essential.

The model achieves an accuracy of approximately 86%. This means that the model correctly predicts whether a loan will be paid off or defaulted on 86% of the time.

	precision	recall	f1-score	support
0	0.88	0.95	0.92	3842.00
1	0.70	0.48	0.57	921.00
accuracy	0.86	0.86	0.86	0.86
macro avg	0.79	0.72	0.74	4763.00
weighted avg	0.85	0.86	0.85	4763.00

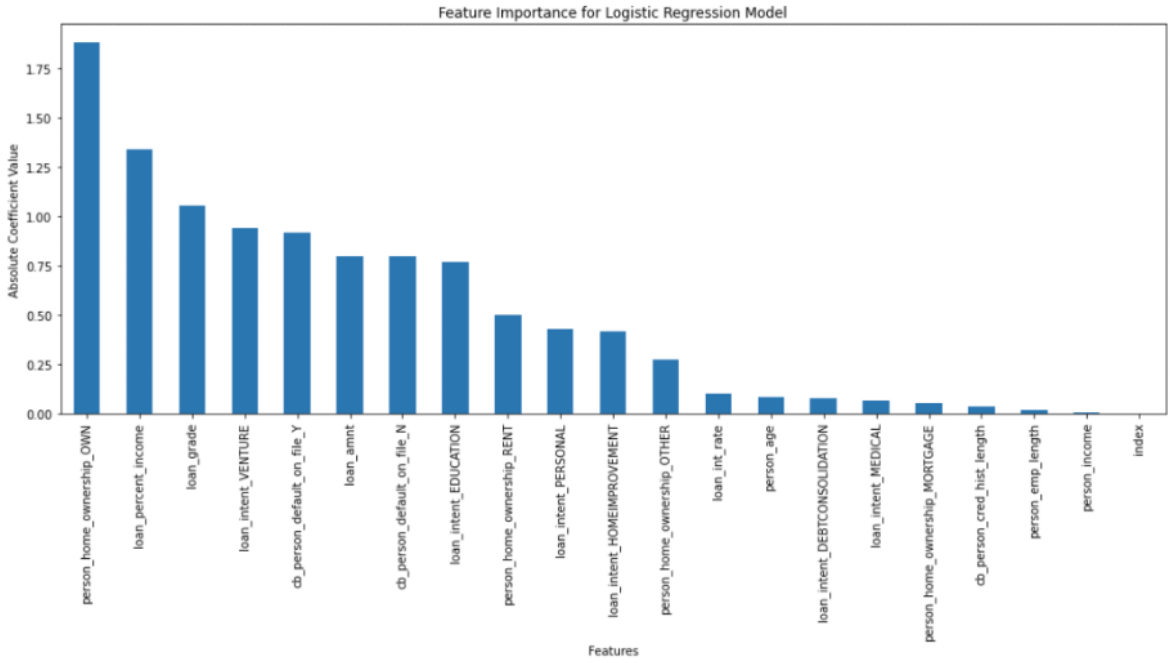
The logistic regression model for credit risk performs well in predicting non-default cases (88% precision, 95% recall, 92% f1-score) but is less effective in correctly identifying default cases, as seen in the lower recall (48%) and F1-score (57%) for the default class.

```
Fitting 10 folds for each of 14 candidates, totalling 140 fits
Best accruacy: 0.855178 using {'C': 100, 'penalty': 'l1'}
cross validation score is: ((0.8529857312173655, 0.00971812040855695), (0.7173395693211383, 0.07352365455693545), (0.45501
559114211726, 0.046451228117362145), (0.5525452803828379, 0.028360927847189375))
```



The grid search for hyperparameter tuning resulted in a logistic regression model with a regularization strength (‘C’) of 100 and ‘l1’ penalty as the best parameters, achieving an accuracy of approximately 85.52%. suggests that the model does not require much

regularization, which implies the model is not overfitting significantly with the provided features. The learning curve shows the training and cross-validation scores converging as the number of training examples increases. This is indicative of a well-fitting model that is generalizing well to unseen data.



The top3 importance features is person_home_ownership, loan_percent_income and loan_grade.

Decision Tree

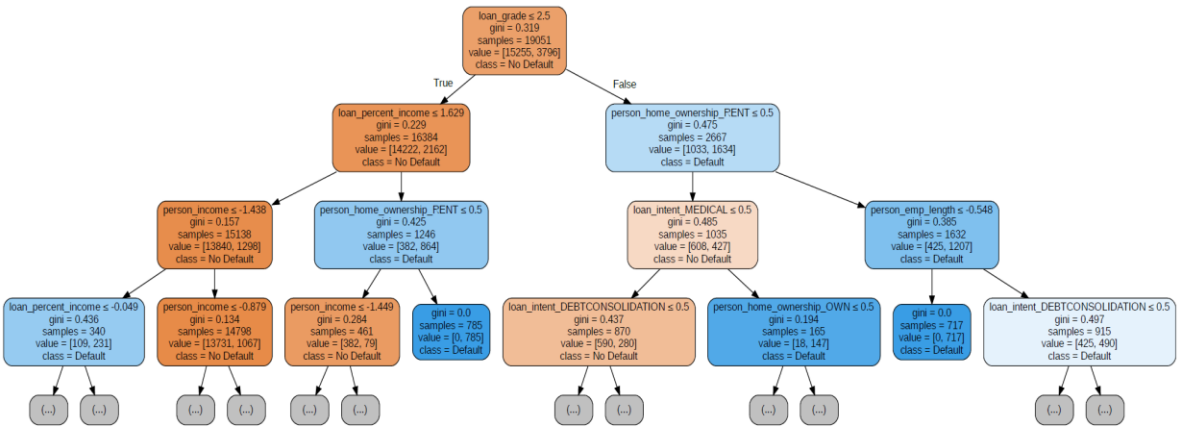
The initial model, without cross-validation, suggested a high accuracy of approximately 89.2% in predicting loan defaults. This model was particularly adept at identifying non-defaulting loans, with precision and recall both at 93%. However, for defaulting loans, precision and recall were lower, at 73% and 74%, respectively, indicating a higher instance of false positives and false negatives in comparison.

Cross-validation revealed substantial variability in model accuracy, ranging from around 20.7% to 86.3%, with an average best score of approximately 77.9%. Such variability, especially the anomalously low score, may indicate potential overfitting or inconsistencies in

the data, where the model performs well on some subsets of the data but poorly on others. The best parameters identified via hyperparameter tuning were a “max_depth” of 10, “min_samples_leaf” of 1, and “min_samples_split” of 10. This parameter setting indicates a preference for a more constrained tree, potentially reducing overfitting and improving the model’s generalization to new data.

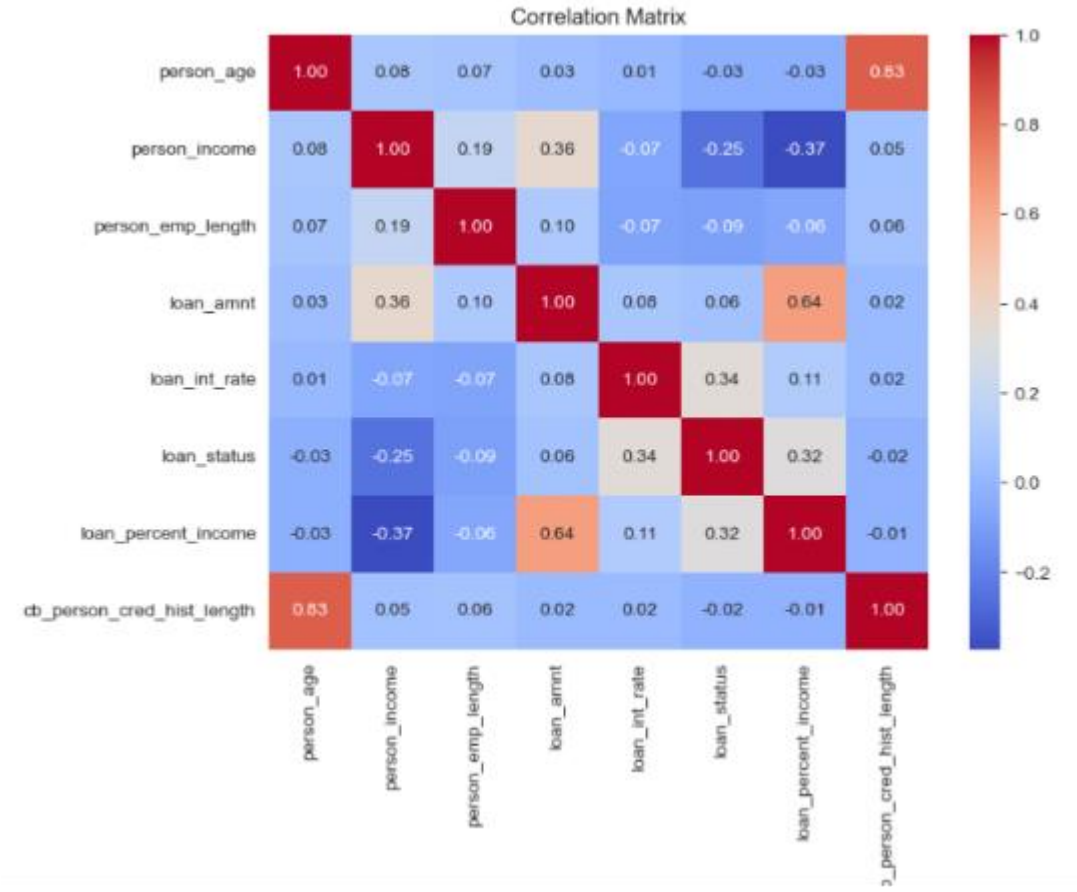
```
Accuracy: 0.8922947722023935
Confusion Matrix:
 [[3526  264]
 [ 249  724]]
Classification Report:
              precision    recall  f1-score   support

           0       0.93      0.93      0.93      3790
           1       0.73      0.74      0.74       973

    accuracy                           0.89      4763
   macro avg       0.83      0.84      0.84      4763
weighted avg       0.89      0.89      0.89      4763
```

In conclusion, the decision tree’s initial impressive performance metrics need to be understood in the context of the cross-validation results. The significant spread in cross-validation scores urges a cautious interpretation of the model’s predictive reliability. While the tuned model parameters suggest a more stable and generalizable model, the extreme variation in cross-validation scores points to a need for further investigation into data quality, feature selection, and possibly model choice. It’s crucial to delve deeper into the underlying causes of this variation to improve model stability and ensure consistent performance across different data segments.



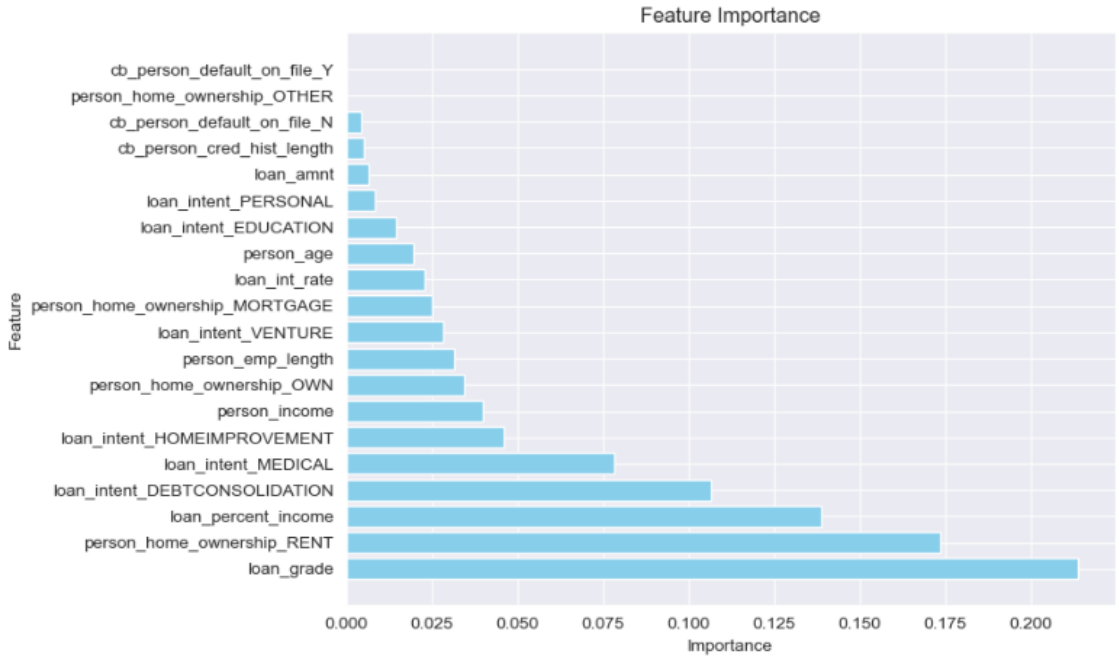
K-Nearest Neighbors

Data preprocessing is essential to optimize the input data for better model performance. There are a couple of key steps undertaken. We began by loading the dataset using Pandas' read_csv function. Missing values were addressed through imputation strategies, ensuring no data loss that could affect model accuracy. We engineered new features that are crucial for predicting credit risk, such as debt-to-income ratio and number of past delinquencies. Numerical data were normalized using StandardScaler to bring all features to a similar scale, avoiding bias toward variables with higher magnitude. Categorical variables were transformed using OrdinalEncoder, facilitating their use in our mathematical models. A correlation matrix in credit risk modeling provides insight into how different variables relate to one another and their influence on credit risk. Each entry in the matrix shows the correlation coefficient between two variables, which ranges from -1 to 1.



K-Nearest Neighbors (KNN) is a simple, versatile, and widely used machine learning

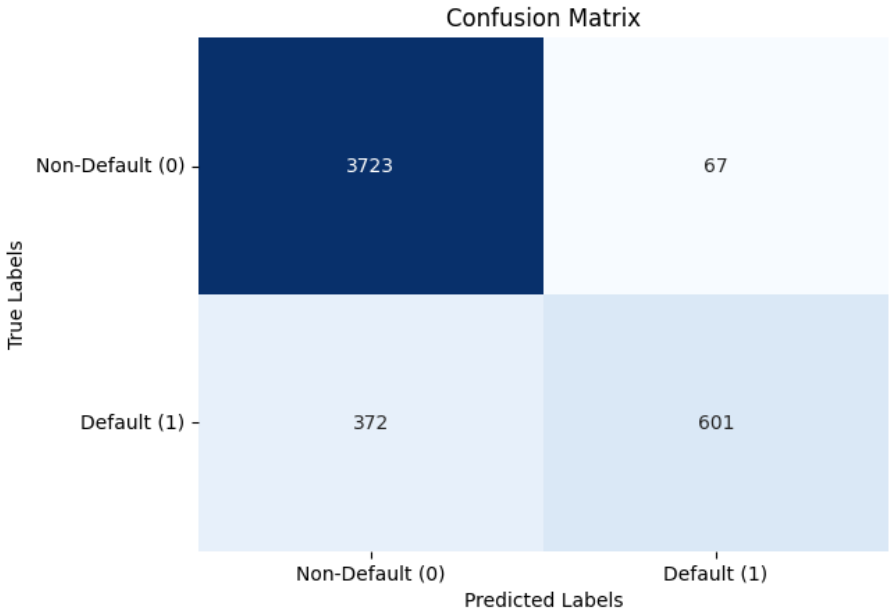
algorithm. It belongs to the family of instance-based, non-parametric learning models. KNN is primarily used for classification tasks, though it can also be applied to regression. KNN operates on a very straightforward principle: it classifies a new data point based on the majority vote of its ‘k’ nearest neighbors. The data point is assigned to the class most common among its nearest neighbors as measured by a distance metric, typically Euclidean distance. The choice of ‘k’ is crucial as it determines how well the model will generalize to new data. A smaller ‘k’ can make the model sensitive to noise in the dataset, whereas a larger ‘k’ makes it computationally expensive and may include features too far away, potentially leading to misclassification. Euclidean distance is the most common metric, but others like Manhattan, Minkowski, or Hamming distance can be used depending on the type of data. Calculate the distance from the new point to all points in the training set and select the ‘k’ smallest distances. For classification, use the most frequent class label among the ‘k’ nearest neighbors. For regression, the average or median of the dependent variable of the nearest neighbors is used. KNN is very intuitive and simple to implement. It makes no assumptions about the underlying data distribution, which is an advantage over models like linear regression that require the data to follow a certain distribution. It can be used for both classification and regression tasks. KNN can be very slow and inefficient when dealing with large datasets because it requires computing the distance of each query instance to all training samples. Its performance degrades as the number of features (dimensions) increases, unless the number of samples is sufficient to cover the multidimensional space. KNN is sensitive to noise in the dataset, especially if ‘k’ is set too low. Our KNN score is 0.9 which is good. The most important feature is loan grade.



Support Vector Machines

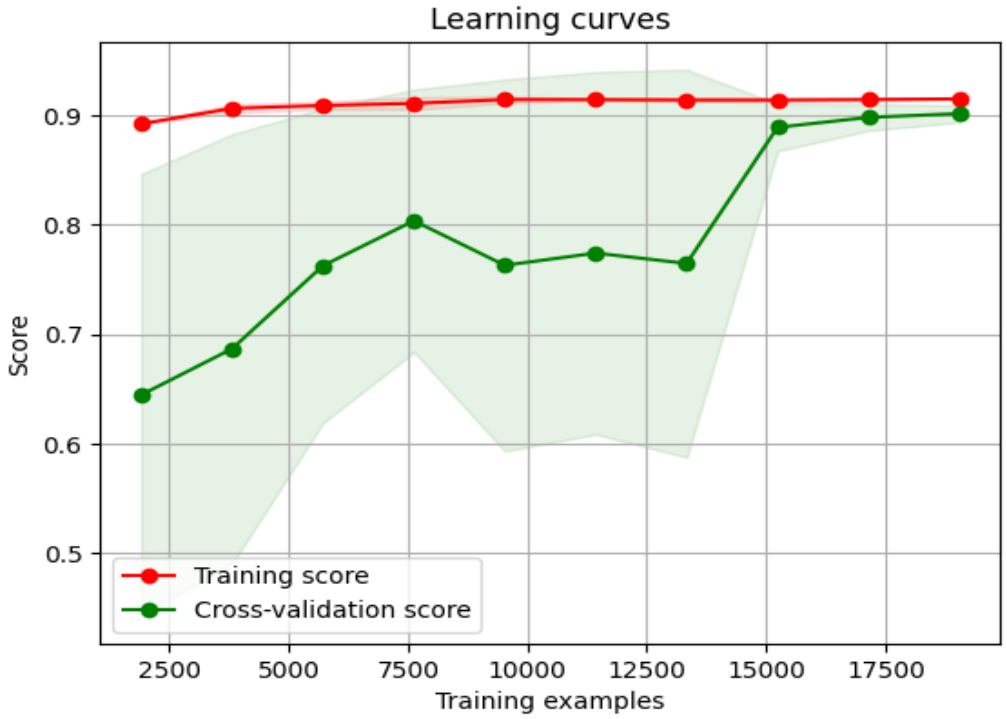
The SVM model achieved an accuracy of approximately 91.07% on the test set, which is a strong indicator of its overall predictive performance. The model achieved a precision of 91% for the negative class (Non-Default) and 90% for the positive class (Default), indicating a high likelihood that the predicted labels are correct. The recall was 98% for the negative class and 62% for the positive class, suggesting the model is very effective at identifying the negative class but less so for the positive class. The model scored an F1-score of 94% for the negative class and 73% for the positive class, which are harmonic means of precision and recall.

The confusion matrix showed 3723 true negatives and 601 true positives, which are correct predictions. There were 67 false positives and 372 false negatives, indicating instances where the model predicted the label incorrectly.

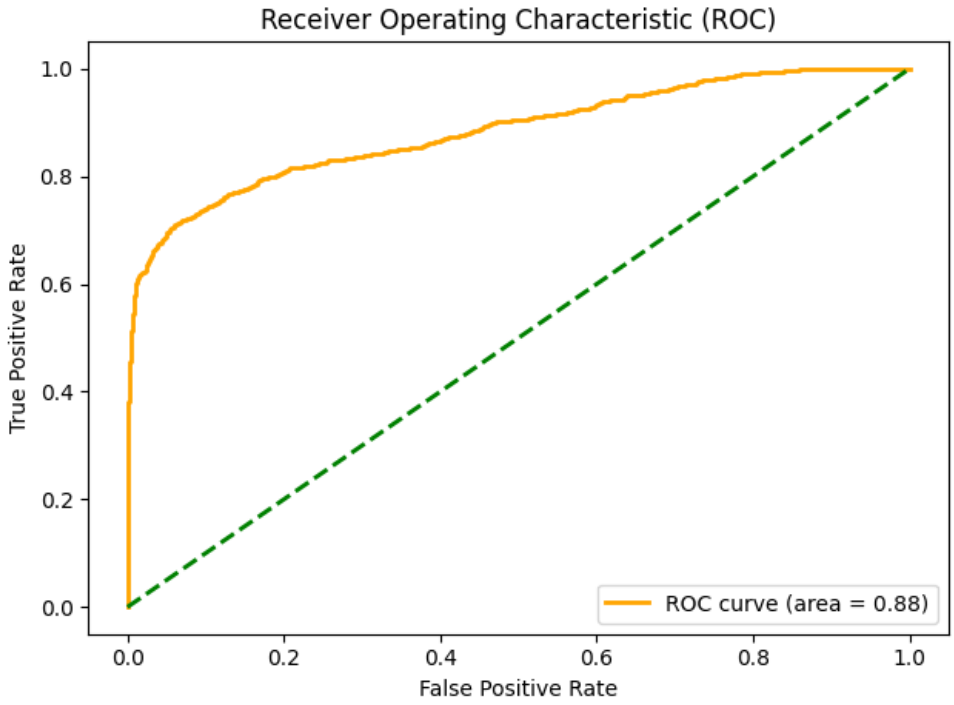


The model underwent 5-fold cross-validation, resulting in accuracy scores that were consistently high across all folds, with a mean accuracy of approximately 91.62% and a low standard deviation, indicating stable performance across different subsets of the data.

A learning curve plot would provide insight into how the SVM model's performance changes with the size of the training data, showing the scores for training and validation datasets. However, the specific output for the learning curve was not included in the snippets.



The ROC curve, which is a plot of the true positive rate against the false positive rate, was not directly shown in the snippets provided, but it is indicated that the model achieved a notable AUC of approximately 0.848, reflecting its good discriminative ability.



XGBoost

The XGBoost model's hyperparameters were carefully selected for tuning to cater to the dataset's unique characteristics. The dataset was duly split into training and testing subsets, with 80% allocated for training and 20% for testing, ensuring a random seed for reproducibility. The chosen hyperparameters for optimization included the number of trees (`n_estimators`), the depth of the trees (`max_depth`), and the learning rate. The selection of these particular hyperparameters is guided by their direct impact on the model's complexity and its ability to learn from data without overfitting.

Utilizing a 5-fold cross-validation approach, the GridSearchCV method honed in on the most effective combination of hyperparameters. The optimal settings were determined to be 100 trees, a maximum depth of 5 for each tree, and a learning rate of 0.1. This combination

strikes a balance between model complexity and learning capability, ensuring that the model is both accurate and generalized well.

Upon evaluation, the fine-tuned XGBoost model demonstrated impressive predictive performance, achieving an accuracy of approximately 92.78% on the test set.

```
Best hyperparameters: {'learning_rate': 0.1, 'max_depth': 5, 'n_estimators': 100}
Accuracy on test set: 0.9277766113793827
```

Random Forest

The Random Forest classifier was fine-tuned using grid search cross-validation on a dataset, leading to the selection of hyperparameters that optimized its performance. The model was trained with the following optimal settings: 300 trees (n_estimators), a maximum depth of 20 (max_depth), a minimum of 5 samples required to split nodes (min_samples_split), and a minimum of 1 sample required at each leaf node (min_samples_leaf).

The accuracy of the Random Forest model on the test set was 92.65%. This high accuracy suggests that the model is effective at predicting the correct class labels for the given dataset.

```
Best hyperparameters: {'max_depth': 20, 'min_samples_leaf': 1, 'min_samples_split': 5, 'n_estimators': 300}
Accuracy on test set: 0.9265169011127441
```

Results

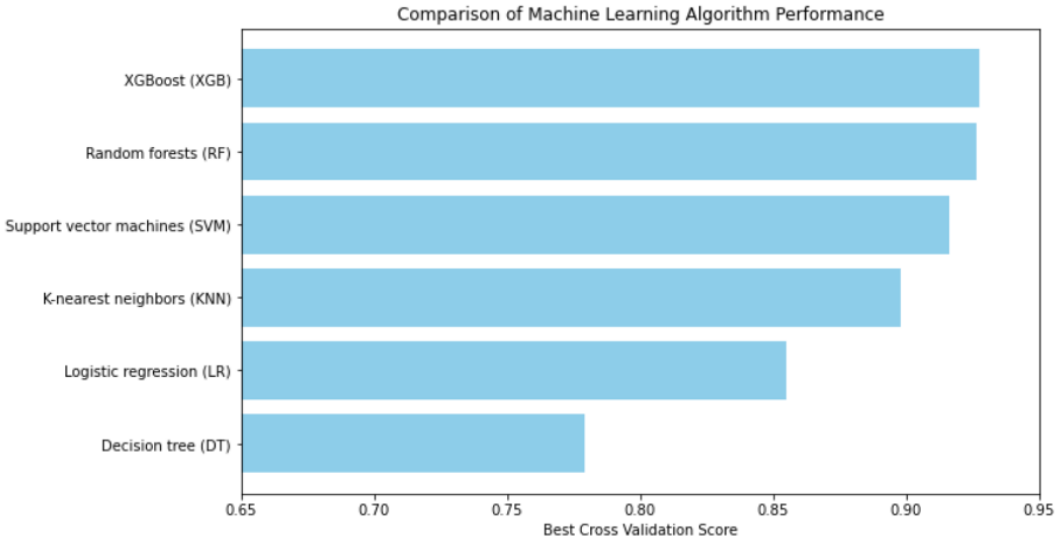
The section starts by outlining its design, which aims to assess the predictive performance of various algorithms selected for the classification of credit risk. The preliminary choice of these algorithms stemmed from important insights identified within literature review related to credit risk classification. Consequently, a total of six algorithms were selected for this purpose. To find the optimal model, we did hyperparameter tunning. Below are the simulated parameters that make the best prediction for each model.

	Algorithm	Hyperparameter
0	Logistic regression (LR)	class_weight="balanced", C=100
1	Decision tree (DT)	max_depth=10, min_samples_leaf=1, min_samples_...
2	K-nearest neighbors (KNN)	n_neighbors=12, class_weight="distance"
3	Support vector machines (SVM)	kernel="rbf", gamma="scale", C=10
4	Random forests (RF)	max_depth=20, min_samples_leaf=1, min_samples_...
5	XGBoost (XGB)	learning_rate=0.1, max_depth=5, n_estimators=100

In the simulations, the 10-fold cross-validation technique was employed, as it strikes a good balance between bias and variance. Initially, the pre-processed dataset was divided, with 90% of the observations forming the first subset and the remaining 10% constituting the second. This training subset was partitioned at random into ten equal-sized folds.

	Algorithm	Best Cross Validation Score
0	Logistic regression (LR)	0.8551
1	Decision tree (DT)	0.7791
2	K-nearest neighbors (KNN)	0.8979
3	Support vector machines (SVM)	0.9162
4	Random forests (RF)	0.9265
5	XGBoost (XGB)	0.9277

In summary, the machine learning algorithms are ranked in terms of their cross-validation score as follows (from highest to lowest score): XGBoost, Random Forests, Support Vector Machines, K-Nearest Neighbors, Logistic Regression, and Decision Tree. XGBoost achieves the highest score, indicating it may be the best performing model among those evaluated, while the Decision Tree appears to be the least effective for the given dataset based on the information provided. We also did data visualization with the score ranks.



The differing results among machine learning models in this study are likely due to a variety of factors. Ensemble tree models such as XGBoost and Random Forest can enhance predictive performance by combining multiple weak learners, and they show superior performance in handling complex datasets and capturing intricate data patterns, thus surpassing single models like Decision Trees. Support Vector Machines (SVM) typically perform well on high-dimensional datasets, especially when the classification boundaries are not clearly defined. If a dataset exhibits complex, non-linear relationships, models such as SVM and XGBoost may outperform logistic regression (LR) and K-Nearest Neighbors (KNN). Recognizing these differences is crucial in selecting the most appropriate model for credit risk scenarios.