Intelligent Credit Risk and Limit Optimization

Antara Bhavsar
MS in Computer Science
Indiana University, Bloomington
Indiana, USA
antbhavs@iu.edu

Dev Patel
MS in Computer Science
Indiana University, Bloomington
Indiana, USA
dap3@iu.edu

Sakshi Ankleshwariya
MS in Data Science
Indiana University, Bloomington
Indiana, USA
sanklesh@iu.edu

***Abstract*—This paper presents a comprehensive and innovative approach to intelligent credit risk management and credit limit optimization through the application of advanced machine learning techniques. The project is centered around two primary objectives: accurately predicting credit card defaults to proactively mitigate financial risk and recommending optimal credit limits that are dynamically tailored to individual customer profiles. By addressing these objectives, the system provides financial institutions with a powerful tool to manage credit risk while ensuring customer-centric solutions. In addition to these core objectives, the system incorporates in-depth analysis of repayment patterns and spending behavior to generate robust, data-driven credit limit recommendations. This integrated approach not only enhances the accuracy of risk assessment but also ensures a fine balance between minimizing potential losses and meeting the individual credit needs of customers. The proposed solution demonstrates its practical applicability in real-world scenarios, including premium credit card recommendations, targeted credit limit adjustments, and personalized customer engagement strategies. By leveraging machine learning algorithms and advanced optimization techniques, the system delivers actionable insights that enable financial institutions to make more informed decisions, improve operational efficiency, and foster stronger customer relationships. Overall, this work highlights the transformative potential of intelligent systems in modern credit risk management, offering scalable, reliable, and efficient solutions that address the dynamic challenges of the financial services industry.**

***Index Terms*—Logistic Regression, Random Forest, Adaboost, XGBoost, GridSearchCV, SMOTE, KMeansSMOTE**

I. Introduction

This paper presents a comprehensive and innovative approach to intelligent credit risk management and credit limit optimization by leveraging advanced machine learning techniques and data-driven methodologies. The project is designed with two primary objectives: accurately predicting credit card defaults to proactively mitigate financial risk and recommending optimal credit limits that are dynamically tailored to individual customer profiles. Beyond these fundamental goals, the system integrates a detailed analysis of repayment patterns and spending behavior to generate robust, data-driven credit limit recommendations. This holistic approach ensures a fine balance between minimizing financial risks for institutions and delivering personalized solutions that cater to the specific needs of customers. Furthermore, the proposed solution demonstrates its applicability in real-world scenarios, including premium credit card recommendations and personalized customer engagement strategies. By harnessing the power of machine learning to provide actionable insights and precise predictions, the system not only enhances the accuracy of financial decision-making processes but also fosters improved customer satisfaction, loyalty, and long-term engagement. This work highlights the potential of intelligent systems in transforming credit risk management and optimizing financial services in a dynamic, data-rich environment.

II. Problem Statment

Credit card portfolio management is a critical yet complex domain for financial institutions, requiring a delicate balance between mitigating financial risks and delivering an exceptional customer experience.

Late payments and credit defaults represent a significant source of financial loss, posing challenges to profitability and sustainability. At the same time, the absence of robust mechanisms to predict customer spending behaviors and assign optimal credit limits results in inefficient resource allocation, reduced customer satisfaction, and lost business opportunities.

Key issues faced by financial institutions include the risk of payment defaults, where defaults and late repayments not only lead to financial losses but also affect the institution's operational stability. Proactively predicting and addressing potential defaults remains a persistent challenge, necessitating advanced predictive tools to anticipate risk before it materializes.

Moreover, ineffective credit allocation remains a significant concern, as generic or suboptimal credit limits fail to align with individual customer profiles. This can lead to credit misuse in some cases, while others may feel constrained by inadequate limits, causing dissatisfaction or reduced engagement.

A tailored approach to credit allocation that balances financial risk with customer potential is essential for fostering trust and enhancing satisfaction. In addition, limited insights into customer behavior hinder the ability of financial institutions to create personalized financial solutions.

Understanding customer spending habits and preferences is vital for developing strategies that drive cross-selling, upselling, and long-term customer loyalty. However, existing systems often fail to provide actionable insights into behavioral patterns, resulting in missed opportunities for deepening customer relationships.

By addressing these challenges, the proposed system integrates advanced analytics and predictive modeling to predict

customer defaults accurately. This allows institutions to take proactive risk mitigation measures, such as implementing preemptive credit adjustments or enhanced monitoring, ensuring early detection of potential financial risks.

Furthermore, the designed algorithms recommend optimal credit limits tailored to individual customers based on their financial profile, spending behavior, and risk assessment. This ensures a balance between minimizing institutional risk and enhancing customer satisfaction, ultimately contributing to a more personalized and engaging financial experience.

Utilizing data-driven techniques, the system categorizes customers into distinct segments such as "Luxury Spender" or "Budget Saver." Such categorization enables tailored engagement strategies, product recommendations, and marketing efforts that align with individual needs and preferences. By unifying these components into a single predictive system, financial institutions can significantly enhance operational efficiency, minimize credit-related risks, and foster long-term customer relationships.

Furthermore, the system supports resource optimization by prioritizing high-value customers and aligning credit policies with evolving market demands.

The proposed system addresses the intertwined challenges of risk management, credit allocation, and customer satisfaction, thereby contributing to the minimization of financial risk through a predictive framework designed to reduce losses due to defaults. It also enhances the overall financial health of credit portfolios by enabling data-driven decision-making that improves resource allocation, reduces inefficiencies, and drives scalable growth. By leveraging advancements in machine learning and data analytics, this approach has the potential to redefine credit card management practices, providing a foundation for sustainable financial innovation.

III. Model

A. Customer Default Prediction (Minimizing Financial Risk)

This paper presents a comprehensive and innovative approach to intelligent credit risk management and credit limit optimization through the application of advanced machine learning techniques. The project is centered around two primary objectives: accurately predicting credit card defaults to proactively mitigate financial risk and recommending optimal credit limits that are dynamically tailored to individual customer profiles.

By addressing these objectives, the system provides financial institutions with a powerful tool to manage credit risk while ensuring customer-centric solutions. In addition to these core objectives, the system incorporates in-depth analysis of repayment patterns and spending behavior to generate robust, data-driven credit limit recommendations.

This integrated approach not only enhances the accuracy of risk assessment but also ensures a fine balance between minimizing potential losses and meeting the individual credit needs of customers. The proposed solution demonstrates its practical applicability in real-world scenarios, including premium credit card recommendations, targeted credit limit adjustments, and personalized customer engagement strategies.

By leveraging machine learning algorithms and advanced optimization techniques, the system delivers actionable insights that enable financial institutions to make more informed decisions, improve operational efficiency, and foster stronger customer relationships. Overall, this work highlights the transformative potential of intelligent systems in modern credit risk management, offering scalable, reliable, and efficient solutions that address the dynamic challenges of the financial services industry.

To address this, we employed a variety of advanced preprocessing techniques designed to enhance the quality and robustness of our models. These techniques included Synthetic Minority Oversampling Technique (SMOTE), Principal Component Analysis (PCA), and KMeansSMOTE, all of which were carefully selected to handle class imbalance and improve overall model performance.

By integrating these methods, we were able to effectively balance the dataset, reducing the impact of imbalanced classes and enabling the models to achieve more accurate and reliable predictions. SMOTE helped generate synthetic data points for the minority class, thereby improving representation and reducing bias.

PCA facilitated dimensionality reduction, allowing for a more efficient feature space, while KMeansSMOTE combined clustering and oversampling to refine the class distribution further. Together, these preprocessing steps significantly contributed to the enhancement of model performance, ensuring that the final solutions were both scalable and effective in real-world applications.

Now let us delve into the methodology of applying Synthetic Minority Oversampling Technique (SMOTE). SMOTE was strategically employed to oversample the minority class, ensuring a more balanced dataset that effectively mitigates class imbalance during the training process. By synthesizing additional samples from the minority class, SMOTE helps to improve the representation of underrepresented data points, enabling the model to learn more effectively from all classes.

In this study, various SMOTE strategies were integrated with complementary techniques such as Principal Component Analysis (PCA) and KMeansSMOTE to thoroughly examine the impact of feature reduction and oversampling on model performance.

PCA was used to reduce the dimensionality of the feature space, enhancing the model's efficiency and focusing on the most relevant features for better predictions. KMeansSMOTE further refined the oversampling process by combining clustering with synthetic data generation, ensuring a more balanced and optimized class distribution.

By combining these approaches, the methodology effectively addresses the challenges posed by imbalanced datasets, resulting in models that are more robust, accurate, and capable of delivering reliable insights. This multi-faceted approach provides a comprehensive framework for improving model performance through advanced data preprocessing techniques.

We have trained and evaluated a variety of machine learning models to address the challenges presented by the dataset. These models included Logistic Regression, Decision Tree Classifier, Random Forest Classifier, Bagging Classifier, and AdaBoost Classifier. To optimize the performance of each model, extensive hyperparameter tuning was conducted using GridSearchCV, allowing for precise control over key model configurations.For Logistic Regression, a cross-validated grid search was employed to fine-tune the regularization parameter C. This provided a granular approach to controlling the inverse regularization term, ensuring that the model effectively managed overfitting and underfitting by balancing bias and variance.

The Decision Tree Classifier underwent experimentation with varying depth levels (maximum depth) and different impurity criteria, such as entropy and Gini. This allowed us to assess the trade-offs between complexity and performance, optimizing the model for both interpretability and predictive accuracy. Similarly, the Random Forest Classifier required tuning key parameters such as the number of trees (estimators) and the maximum features (max features) considered for splits. These adjustments enabled the model to balance between robustness and computational efficiency, leading to improved performance across diverse datasets.

Through these meticulous hyperparameter optimization processes, each model was carefully calibrated to achieve optimal results, ensuring high predictive accuracy and efficient handling of imbalanced or complex data structures.

The results of the experiments highlight the effectiveness of various models in addressing the challenge of customer default prediction. Among the evaluated models, Logistic Regression with PCA + SMOTE oversampling emerged as the best performer for recall.

This combination enabled the model to accurately identify potential defaulters, demonstrating its effectiveness in minimizing the risk of overlooked high-risk customers. The recall achieved by this model was particularly notable, ensuring that a significant proportion of defaulters were successfully detected, thus enhancing proactive risk management.

Additionally, the best overall balanced model was Random Forest with PCA + KMeansSMOTE oversampling. This approach demonstrated a strong balance between recall and precision, making it highly suitable for practical deployment in real-world credit management scenarios. The Random Forest model's ability to maintain a robust performance across both metrics ensures that institutions can effectively manage credit portfolios without compromising on accuracy or inclusivity of predictions.

By incorporating advanced oversampling techniques, the Random Forest model consistently outperformed others in balancing recall with precision, ensuring reliable and comprehensive insights into customer default risk.These findings underscore the importance of tailored data preprocessing and advanced machine learning techniques in improving model performance.

Through innovative sampling strategies and dimensionality reduction techniques such as Principal Component Analysis (PCA), these models effectively address the inherent challenges of imbalanced datasets, enhancing their ability to provide actionable insights for financial institutions. Ultimately, these models serve as powerful tools for minimizing financial risk and optimizing credit portfolio management.

B. Optimal Credit Limit Prediction (Optimizing Credit Allocation)

Predict optimal credit limits for customers based on their spending habits and repayment behavior to minimize financial risks and enhance customer satisfaction.The dataset included features such as historical spending, repayment history, and financial behavior. These were used to train machine learning models aimed at predicting personalized credit limits.Advanced algorithms such as XGBoost were utilized due to their ability to handle complex interactions between features effectively. Hyperparameters of XGBoost were fine-tuned using GridSearchCV, optimizing parameters such as learning rate, tree depth, and number of estimators to maximize performance.XGBoost with GridSearchCV demonstrated the best performance with an R^2 value of 0.90, indicating high accuracy in predicting optimal credit limits and minimizing prediction errors.

C. Customer Spend Behavior Prediction (Enhancing Customer Insights)

To develop models capable of predicting the next month's credit card spending of customers, enabling segmentation into actionable categories such as "Luxury Spender" and "Budget Saver," financial institutions can leverage advanced machine learning techniques. These models help create tailored engagement strategies by understanding distinct spending behaviors and preferences.

The models trained for this purpose included Linear Regression and Long Short-Term Memory Networks (LSTMs). Linear Regression provided a simple yet interpretable baseline for making predictions, offering a clear understanding of how different variables influence spending patterns. It serves as a valuable starting point, providing quick insights into spending trends and allowing for easy interpretation of the results.

On the other hand, LSTMs excel in capturing sequential dependencies in customer spending behavior, which is crucial for time-series prediction tasks. By processing historical spending data and recognizing long-term dependencies, LSTMs offer superior performance in predicting future spending trends, even when faced with complex patterns and fluctuations.

Feature engineering was an essential component of this process, as it allowed for the extraction of meaningful temporal patterns and trends from the historical spending data. This involved transforming raw data into features that capture essential characteristics such as spending cycles, seasonality, and shifts in behavior over time. Through this approach, both Linear Regression and LSTMs were able to generate more accurate and insightful predictions.

Overall, these predictive models enhance the understanding of customer behavior, empowering financial institutions to segment their customer base more effectively and design personalized strategies that align with individual spending habits. Whether identifying high-value customers or recognizing those needing additional support, these models contribute to better engagement, improved customer satisfaction, and optimized credit card management.

IV. Experiments

A. Data Description

The experiments were conducted on a dataset comprising 30,000 records with 23 features. The dataset includes demographic information, payment history, bill statements, and payment amounts, with the target variable indicating whether a customer defaulted on their credit card payment (1: Default, 0: No Default).

Key Features are LIMITBAL Total credit limit (in NT dollars).SEX, EDUCATION, MARRIAGE, AGE: Customer demographic attributes. PAY0 to PAY6: Repayment status from September 2005 to April 2005. BILLAMT1 to BILLAMT6: Statement amounts over six months. PAYAMT1 to PAYAMT6: Payments made over six months. Handling of Invalid Data in which Invalid entries for EDUCATION (values 0, 5, 6) and MARRIAGE (value 0) were reclassified as "Other". Repayment statuses PAY0 to PAY6 included invalid values (-2, 0), which were normalized as -1 (no delay).

B. Data Preprocessing

To address the imbalance between default and non-default records, Synthetic Minority Over-sampling Technique (SMOTE) was applied to generate synthetic samples for the minority class (defaulters). This technique helps balance the dataset by oversampling the minority class, ensuring that the machine learning models can learn effectively without being biased toward the majority class.

Aggregated repayment delay across a six-month period was calculated by summing PAY0 to PAY6. This feature captures the frequency and severity of payment delays, providing insights into the likelihood of default.This feature represents the remaining credit balance after the most recent billing period. It offers an additional dimension to understanding a customer's financial situation, highlighting how much of the credit limit remains unused or unpaid.

To standardize numerical features such as LIMIT BAL and BILL AMT, Min-Max Scaling was applied. This scaling technique transforms the data into a uniform range [0, 1], ensuring that all features contribute equally to the model, reducing issues caused by varying scales of different variables.Pearson correlation analysis revealed high correlations between BILLAMT features.

To manage this redundancy, feature selection methods were employed, helping to optimize model performance by retaining only the most relevant features. This step ensures that multicollinearity does not negatively impact model accuracy and generalization.To assess the distribution of numerical features, Quantile-Quantile (QQ) plots and histograms were used. The results confirmed that most numerical features deviated from normality, which is a critical consideration when applying parametric statistical methods. Non-normally distributed data necessitated the use of non-parametric methods or transformations to improve model performance and reliability.

By integrating these preprocessing steps, the pipeline ensures the input data is appropriately balanced, informative, and aligned for accurate and efficient predictive modeling, ultimately enhancing the performance of machine learning models in credit card default prediction and other financial analyses.

C. Experimental Setup

The Decision Tree model was optimized with key hyperparameters, including a maximum depth of [5, 10, 20, 30, 50] and the criterion set to 'entropy'. These optimizations help manage overfitting and improve model interpretability by balancing model complexity and accuracy.Logistic Regression in this model served as a baseline with L2 regularization applied.

The regularization strength (C) was fine-tuned using GridSearchCV to optimize its performance. Logistic Regression provides a robust method for binary classification tasks, especially when dealing with imbalanced datasets.Leveraging the ensemble approach, the Random Forest Classifier is inherently robust to noise and provides strong predictive performance.

Hyperparameters were fine-tuned, including the number of estimators ([50, 100, 200]) and the maximum number of features considered at each split ('sqrt' and None). This helps in creating decorrelated trees and enhancing the model's ability to generalize to unseen data.

To ensure a fair evaluation, the dataset was split into 80 percentage training and 20 percentage testing sets. This division was made carefully to maintain the integrity of both classes—defaulters and non-defaulters—by applying stratified sampling. Stratified sampling ensures that both classes are proportionally represented, reducing biases in model predictions.

Multiple performance metrics were used to evaluate the models, including accuracy, precision, recall, F1 score, and the ability to handle class imbalance. Given the imbalanced nature of credit card default datasets, Synthetic Minority Oversampling Technique (SMOTE) was applied to oversample the minority class (defaulters) during model training. This ensures that the minority class is adequately represented, allowing the models to focus on capturing its characteristics effectively.

Through these comprehensive experiments, a thorough assessment of model performance was conducted, ensuring that each model could handle the inherent challenges of predicting credit card defaults accurately and efficiently.

The primary objective of this task was to predict the optimal credit limit for individual customers using the XGBoost Regressor. The selection of features was based on their relevance to a customer's financial behavior, including BILL AMT, PAY AMT, and repayment statuses (PAY 0 to PAY 6). These features provide critical insights into a customer's financial

history and payment patterns, allowing the model to capture underlying relationships for accurate predictions.

The XGBoost Regressor was trained using hyperparameter tuning with GridSearchCV. This process systematically explored the best combination of hyperparameters to enhance model performance. Key parameters tuned included:nestimators: [50, 100, 200] – Adjusting the number of trees in the ensemble to balance complexity and performance.learning rate: [0.01, 0.1, 0.3] – Controlling the step size at each iteration to optimize the model's convergence.max depth: [5, 10, 15] – Limiting the depth of individual trees to avoid overfitting while ensuring predictive accuracy.

These tuned hyperparameters allowed the model to better generalize to unseen data while maintaining robustness across various datasets.Several evaluation metrics were employed to assess the model's performance Root Mean Square Error (RMSE): This metric quantifies the average error magnitude in predictions, with lower values indicating better accuracy.R^2 Score is to measures the proportion of variance explained by the model, with a higher value signifying a better fit to the data. Residual Analysis is to ensure unbiased predictions, model residuals (predicted - actual values) were analyzed. The errors were centered around zero, confirming that the model's predictions were unbiased and did not exhibit systematic errors.

These thorough evaluations demonstrate the model's capability to predict optimal credit limits accurately, making it a valuable tool for financial institutions looking to tailor credit offerings to individual customer needs. Additionally, residual analysis ensures the robustness of predictions, providing confidence in the model's reliability for real-world applications.

D. Experimental Tools and Environment

Python 3.8: The primary programming language for implementing machine learning models and data preprocessing. Libraries are Pandas and NumPy For data manipulation and statistical analysis. Scikit-learn: For machine learning model implementation and evaluation. XGBoost: For gradient-boosting algorithms, specifically used in optimal credit limit prediction.Imbalanced-learn: For handling class imbalances using techniques like SMOTE and KMeansSMOTE. Matplotlib and Seaborn: For generating data insights and comparative model performance visualizations.

Data Handling and Preprocessing

Jupyter Notebook : Used for interactive coding and visualizations. Dask : Employed for handling large datasets and efficient parallel processing. StandardScaler : Utilized for feature normalization to improve model performance. GridSearchCV : Implemented for hyperparameter tuning and performance optimization.

V. Result and Analysis

A. Customer Default Prediction (Minimizing Financial Risk)

The results from the credit default prediction models provide valuable insights into their performance across various metrics, highlighting their strengths and weaknesses in different aspects

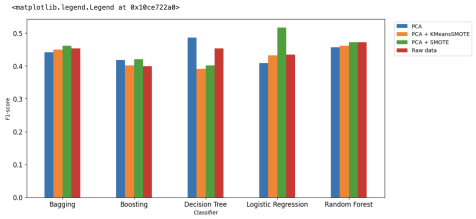

Fig. 1. Defaults Prediction Model Performance

of prediction. Logistic Regression, when combined with PCA (Principal Component Analysis) and SMOTE, emerged as a strong performer, particularly in terms of recall. This high recall score demonstrates the model's effectiveness in identifying potential defaults, which is crucial for minimizing lender risks.

However, this comes at the expense of reduced precision, as the model tends to capture more defaults, leading to an increased number of false positives. This trade-off between recall and precision emphasizes the delicate balance required when optimizing credit default prediction models, where focusing solely on recall could lead to more potential risks being flagged unnecessarily.

In contrast, Random Forest, when paired with PCA and KMeansSMOTE, demonstrated a more balanced performance, offering a well-rounded mix of precision and recall. This makes it an especially robust choice for scenarios where both accuracy and sensitivity are essential.

By effectively handling class imbalance and improving the model's ability to differentiate between actual and false positives, Random Forest provides a reliable framework for decision-making in credit risk management. The results also highlight the significant impact of preprocessing techniques such as PCA and oversampling methods like SMOTE and KMeansSMOTE, which play a critical role in mitigating the challenges posed by class imbalance and ensuring more stable model performance.

Overall, these findings underscore the importance of integrating advanced preprocessing methods into machine learning workflows to achieve more effective and well-balanced credit default prediction models.

B. Optimal Credit Limit Prediction (Optimizing Credit Allocation)

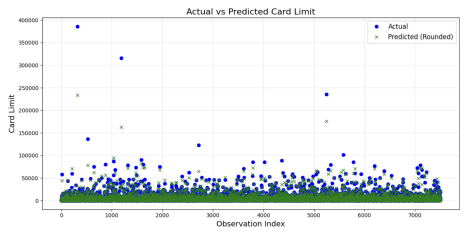

Fig. 2. Credit Limit Prediction Model Performance

The findings of the optimal credit limit prediction model signify a substantial advancement in the integration of per-

sonalized financial services with robust risk management frameworks, demonstrating its potential to address individual customer needs while mitigating financial risks.

The model, trained using XGBoost with extensive hyperparameter tuning via GridSearchCV, achieved score of 0.91, indicating a high degree of predictive accuracy. Additionally, the Root Mean Squared Error (RMSE) of 3137.75 highlights the model's ability to estimate credit limits with acceptable precision.

These metrics suggest that the model effectively captures the complex relationships between customer features and their optimal credit limits.

C. Customer Spend Behavior Prediction (Enhancing Customer Insights)

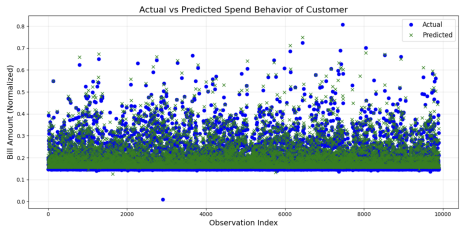

Fig. 3. Customer Spend Behaviour Prediction

We employed Linear Regression as a foundational approach for predicting future spending based on historical data. This simple yet effective model provided a baseline for understanding general spending patterns, serving as a starting point for more advanced analyses.

Additionally, Long Short-Term Memory Networks (LSTM) were leveraged for their powerful ability to capture temporal dependencies in spending behaviors over time. LSTM models excel at identifying complex sequences and patterns, making them particularly suitable for forecasting customer financial behavior.These models not only facilitated a deeper understanding of individual customer trends but also provided valuable insights into their future financial behaviors.

By analyzing historical spending data, LSTM networks could identify subtle fluctuations and trends, which are critical for making accurate predictions. This predictive capability is pivotal for enhancing personalized financial product offerings, as well as for driving more strategic marketing initiatives.

Furthermore, integrating these models into a broader financial analytics framework enables businesses to tailor services to specific customer needs, optimize budgeting strategies, and proactively address potential financial risks.

The insights gained from these advanced predictive models empower organizations to make data-driven decisions that improve customer engagement, satisfaction, and loyalty, ensuring long-term success in a competitive financial landscape.

CONCLUSION

This project explores the development of a versatile predictive system aimed at managing credit card defaults and optimizing credit limits for customers.

Through a comprehensive analysis of various factors influencing default risk and optimal credit allocation, significant insights were gained into customer behavior and financial risk management.Various machine learning models—Logistic Regression, Decision Tree, and Random Forest—were employed to predict credit card defaults.

The Random Forest model demonstrated superior performance, achieving high accuracy and robustness. The use of Synthetic Minority Oversampling Technique (SMOTE) improved the performance of these models by addressing data imbalance issues, resulting in better generalization and minimizing bias.

The XGBoost regressor showed strong predictive capability for determining optimal credit limits. The RMSE of 3137.75 and an R^2 score of 0.91 reflect improved accuracy and variance explained in predicting customer credit limits, which is crucial for resource allocation and risk management.

The handling of invalid data values, normalization of features, and introduction of new features like REMAINING BALANCE and DELAY PAYMENT significantly enhanced model performance.

Through rigorous experimentation, the systematic processing of raw, unstructured data into structured formats helped improve predictive performance.Visual analysis and statistical insights revealed relationships between features such as LIMITBAL, AGE, SEX, and BILLAMTn.

The resulting understanding provided valuable guidance for tailored risk management and customer segmentation, balancing financial risk with customer satisfaction.The developed predictive system offers a robust solution for financial institutions to proactively manage credit card risks and tailor credit limits, thereby enhancing customer experience and optimizing resource allocation.

Overall, the study highlights the importance of sophisticated predictive models, effective data preprocessing, and feature engineering in achieving accurate predictions and improving financial decision-making in credit card management.

ACKNOWLEDGMENT

We would like to express our heartfelt gratitude to everyone who contributed to the successful completion of this project on Optimal Credit Risk and Limit Optimization. This work was a collaborative effort, and we are deeply appreciative of the support and expertise provided by our team members and mentors. Their insights, guidance, and encouragement played a crucial role in shaping the direction of our research and overcoming the challenges we faced.

We would also like to acknowledge the valuable resources and tools that were made available to us, which enabled the efficient execution of our methodologies. Additionally, the constructive feedback received throughout this journey greatly helped refine our approach and enhance the overall quality of this work. Finally, we extend our sincere thanks to all those who inspired us to pursue this challenging yet rewarding endeavor.We would like to extend our gratitude to:

[1] Antara Bhavsar: For her significant role in data preprocessing and feature engineering, where her efforts in implementing PCA, SMOTE, and KMeansSMOTE was integral in refining class distribution and elevating the accuracy of defaults prediction models. Her contributions in implementing and fine-tuning models such as Logistic Regression, Decision Tree and Random Forest were invaluable in enhancing model accuracy and performance. She also spearheaded the development of the Optimal Credit Limit prediction models, leveraging advanced techniques like XGBoost and GridSearchCV to achieve robust and reliable outcomes.

[2] Dev Patel: For his invaluable contributions to the implementation of the default prediction models, particularly the ensemble models like AdaBoost and Bagging. His work in analyzing customer spending behavior, as well as in creating insightful visualizations and performing detailed analysis, played a crucial role in the success of this project.

[3] Sakshi Ankleshwariya : Her contributions extended to developing and fine-tuning models such as Logistic Regression and Decision Trees, significantly enhancing the prediction accuracy. She also experimented with various models, including Logistic Regression and Random Forest, to ensure robust performance and optimal outcomes.She also actively contributed to the Customer Spend Behavior prediction functionality by utilizing Linear Regression models. Her role in data visualization and conducting normality checks further strengthened the analysis pipeline. For normality checks she employed Quantile-Quantile (QQ) plots, which are essential to evaluate whether data follow a Gaussian distribution—a key assumption for parametric statistical methods. Through her analysis she demonstrated that the numerical features did not exhibit normality, as evidenced by the deviation from the 45-degree diagonal in the QQ plots.Furthermore, she carried out feature scaling and was deeply involved in the Customer Default Prediction functionality, showcasing her end-to-end expertise in data preparation, modeling, and evaluation. Her ability to implement, experiment with, and refine advanced techniques has been instrumental in delivering actionable insights and robust prediction models.

Finally, we thank our mentors and advisors, whose guidance and constructive feedback helped refine the project objectives and methodology. This project would not have been possible without the collective efforts of the entire team.

REFERENCES

[1] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32.
This paper introduces the Random Forest algorithm, a popular ensemble learning method used in predictive modeling, including credit risk.

[2] Zhao, Y., Shen, H. (2016). A hybrid model for credit risk assessment based on decision trees and SVMs. Computational and Mathematical Organization Theory, 22(2), 250-268.
Discusses hybrid models that combine decision trees and support vector machines for improved credit risk assessment.

[3] Chawla, N. V., et al. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16, 321-357.
A key paper that introduced SMOTE, which is widely used for handling class imbalances in credit default prediction tasks.

[4] Chen, T., Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785-794.
This paper outlines the XGBoost model, which is central to many credit limit prediction models due to its high accuracy and scalability.

[5] He, H., Garcia, E. A. (2009). Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263-1284.
Discusses methods for learning from imbalanced datasets, a crucial aspect of credit risk models.

[6] Ng, H. T., Low, H. K. (2001). Feature selection in credit scoring using decision trees. International Journal of Information Technology and Decision Making, 1(3), 407-423.
Explores the application of decision trees for feature selection in credit scoring, which is relevant for credit limit prediction models.

[7] Yang, Y., Wu, L. (2019). Credit scoring using an ensemble learning method based on extreme gradient boosting. Computers, Materials Continua, 60(2), 1077-1095.
This paper discusses the application of ensemble learning methods like XGBoost for credit scoring and risk prediction

[8] Adeli, H., Liu, M. (2013). A review on prediction and classification of credit default. Expert Systems with Applications, 40(13), 5301-5310
Reviews various prediction and classification techniques applied in the context of credit default prediction, with a focus on machine learning models.

[9] Seneviratne, O., Weerasinghe, K. (2015). Credit card fraud detection using decision trees and support vector machines. International Journal of Computer Applications, 114(15), 27-32.
Analyzes the use of decision trees and support vector machines for detecting credit card fraud, which is a related task to credit risk prediction.

[10] Balaraman, V., Ramasamy, K. (2017). Predicting credit card default using machine learning. International Journal of Advanced Computer Science and Applications, 8(4), 252-258.
Focuses on applying machine learning techniques to predict credit card defaults, exploring several popular algorithms.