MERITRUST CREDIT UNION

Factors Influencing Loan Default Dataset

Report Draft
BSAN 885 - Capstone Project
Prof. Dr. Sue, Rosemary Radich
Barton School of Business

Team 6
Sudhanshu Sekhar Dash
Uma Maheswara Reddy Pammi
Sri Revanth Kandepi

Table of Content

List of Figures

1 - INTRODUCTION

The financial services industry plays a pivotal role in supporting economic stability and fostering national prosperity. Within this sector, Meritrust Credit Union serves as a vital community pillar, providing an alternative to traditional banking institutions (Yeboah, 2018) [1]. As the financial landscape undergoes evolution, Meritrust encounters multifaceted challenges, with loan defaults emerging as a critical concern (Foster & Zurada, 2013) [8]. Effectively managing loan defaults is imperative, given its profound impact on financial stability, institutional reputation, and community trust (Nigmonov & Shams, 2021) [7]. This report undertakes a comprehensive examination of Meritrust Credit Union's operations, with a particular focus on the intricacies of loan defaults. By exploring its historical journey, identifying pressing business challenges related to loan defaults, and conducting an in-depth exploration of pertinent topics, this study aims to furnish invaluable insights to assist Meritrust in overcoming existing challenges and ensuring its sustained success. The research endeavors to provide Meritrust with a strategic roadmap to navigate the intricate terrain of modern financial services, emphasizing its mission to serve the community effectively while maintaining economic sustainability and competitiveness amidst evolving challenges related to loan defaults.

A Deep Dive into Its History, Mission, and Core Values

Meritrust Credit Union, headquartered in Wichita, Kansas, was founded in 1935 as a not-for-profit financial institution. Since its inception, Meritrust has undergone remarkable growth, transitioning from a small credit union with a limited membership base into one of the largest credit unions in Kansas. Presently, Meritrust proudly serves a diverse membership exceeding 100,000 individuals and businesses, offering an extensive array of financial products and services encompassing savings accounts, loans, mortgages, and investment opportunities.

The credit union's core mission revolves around delivering financial solutions, promoting financial literacy, and enhancing the overall well-being of its members and the communities it serves. With assets surpassing $1.8 billion, Meritrust has firmly established itself as a prominent and influential player within the regional financial sector.

1.2 Problem Statement

Meritrust Credit Union, a prominent financial institution with a rich history of community engagement, is currently facing pressing challenges in managing loan defaults within its extensive member base (Yeboah, 2018) [1]. While the credit union has been a trusted financial partner to its members, it is confronted with a rising number of loan defaults that have adverse implications for its financial stability and long-term viability (Foster & Zurada, 2013) [8]. These loan defaults can not only lead to financial losses but also tarnish the reputation of Meritrust Credit Union, potentially resulting in the loss of valuable members (Nigmonov & Shams, 2021) [7]. To address these issues effectively, it is essential to gain a deep understanding of the factors contributing to loan defaults, develop strategies to mitigate these challenges, and protect the financial well-being of both the credit union and its members.

The research seeks to uncover patterns and trends associated with the current scenario of loan defaults. This involves a thorough analysis of data to identify commonalities, triggers, and recurring factors contributing to defaults. The goal is to extract actionable insights from these patterns to inform the development of effective mitigation strategies. The report focuses on proposing practical and proactive strategies for Meritrust Credit Union to address the challenges posed by loan defaults. This includes recommendations based on the identified causes and patterns, with the ultimate goal of safeguarding the financial stability of the credit union and ensuring the well-being of its members. The outcomes of this research will contribute valuable insights to the credit union and the broader financial sector, aiding in the development of effective strategies to manage and mitigate the impact of loan defaults.

1.3 Background Research:

Background research is pivotal for informed decision-making in the financial industry, especially for entities like Meritrust Credit Union (Yeboah, 2018; Isbister, 1992) [1],[2]. To address challenges, our research explores factors driving loan defaults (Financial Stability Board, 2022) [3]. Notably, credit scores, crucial in charge-offs, reveal an intricate relationship with default risks (Puri et al., 2017) [5]. This research arms Meritrust, and similar institutions, with actionable intelligence to proactively tackle loan default challenges (Mueller & Yannelis, 2017) [6].

Extensive research, encompassing studies from CUNA and insights from industry experts, stresses member-centric approaches and technology adoption (World Bank, 2021) [4]. Amidst a dynamic environment influenced by regulatory shifts, fintech advancements, and demographic changes, credit unions, including Meritrust, must learn from experiences, analyze feedback, and embrace technological trends (Serrano-Cinca et al., 2015) [9]. Research on loan defaults, exploring variables like credit scores, debt-to-income ratios, and financial education (Netzer et al., 2019) [10], guides institutions in tailoring practices. External factors such as regulatory changes, interest rates, and market conditions shape loan defaults (Appati et al., n.d.) [12]. Background research provides a comprehensive understanding, enabling institutions to develop robust risk management strategies, enhance lending practices, and ensure financial well-being (Mayer et al., 2009) [11].

In summary, the importance of background research lies in its ability to provide a comprehensive understanding of the intricate factors influencing loan defaults. By examining economic indicators, borrower characteristics, and external influences, financial institutions like Meritrust Credit Union can develop robust risk management strategies, enhance lending practices, and ensure the financial well-being of both the institution and its members (Mayer et al., 2009) [11].

METHODOLOGY

The following section describes the data set, how it was pre-processed, and the methodology used to build and determine the most accurate model.

2.1 Data Used

The dataset under analysis focuses on historical lending data, particularly loans issued in the year 2017. Its primary purpose is to predict loan charge-offs and understand the risk factors associated with them. Loan charge-offs are events where borrowers fail to repay their loans, resulting in financial losses for lending institutions. Effectively predicting and managing these charge-offs is crucial for an organization's financial health and sustainable growth.

The dataset utilized for this project encompasses information pertaining to loans issued in the year 2017, with tracking extending from 2017 to 2023. Critical for evaluating the performance and associated risk factors of loans within this timeframe, the dataset boasts a substantial size, comprising 15,482 rows and 33 distinct columns. This extensive dataset size facilitates a comprehensive analysis of lending patterns and occurrences of loan charge-offs. To ensure data quality, thorough checks for duplicates were conducted, and various preprocessing steps were implemented. These steps included handling missing values and removing columns with limited relevance to the analysis. The rigorous data cleaning procedures contribute to the overall high quality of the dataset, laying a robust foundation for the ensuing analysis and findings of the project.

2.2 Data Processing

In the initial phase of data processing, a thorough examination of the dataset was conducted to identify and address potential issues such as missing values, duplicates, outliers, and irrelevant features. Fortunately, no duplicate records were found in the dataset. To optimize the dataset for subsequent analysis and modeling, certain features were intentionally excluded based on their lack of relevance to the project's objectives. These excluded features included unique account identifiers, collateral information, and detailed banking account specifics. As part of this process, the 'Total Deposits,' 'Savings, Checking Account,' and 'Auto, Boat, Monetary, Real Estate Collateral' columns were specifically removed due to their limited correlation with the dependent variable. This comprehensive preprocessing approach ensures that the dataset is refined, optimized, and well-prepared for subsequent analytical and modeling tasks in line with the overarching project goals.

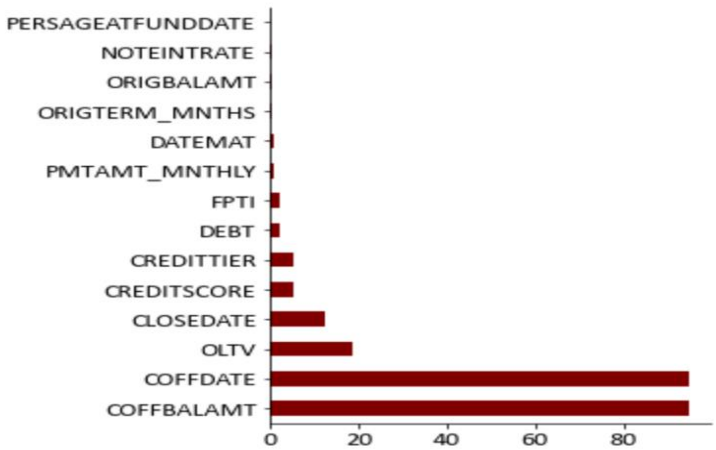

Figure 1. Percentage of Missing values

Additionally, to enhance the dataset's suitability for further analysis, feature engineering techniques were applied to three key variables: Credit Score, Loan Amount, and Loan Type.

2.3 Correlation Analysis:

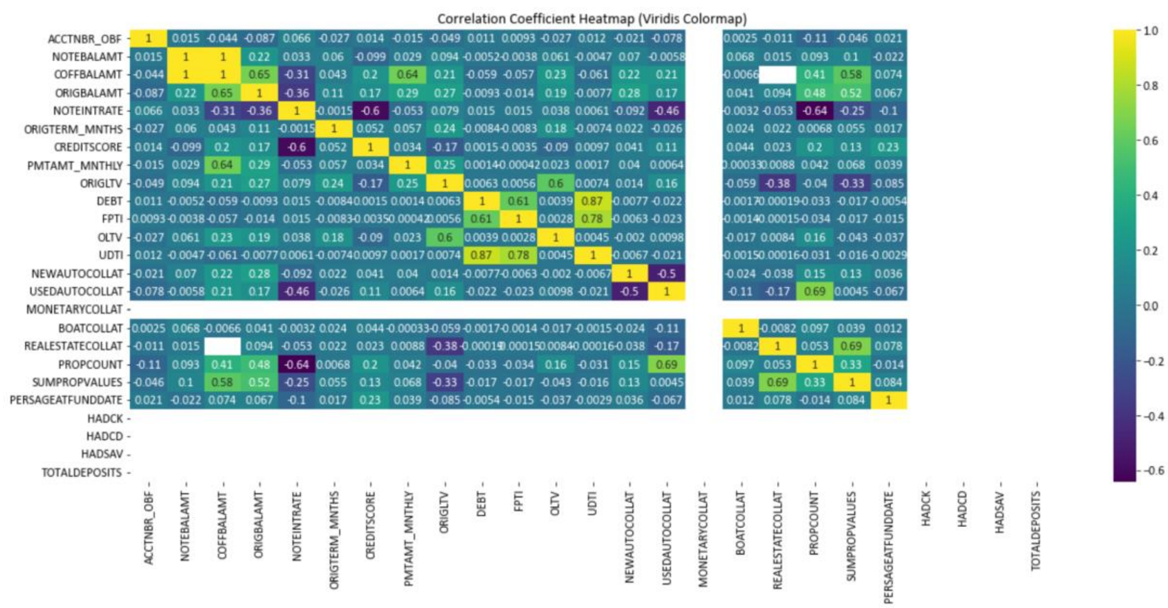

Figure 2. Correlation Matrix

The correlation matrix analysis plays a pivotal role in our project, offering insights into the intricate relationships between various features in the loan default dataset. This heatmap visualization reveals both strong and weak correlations, highlighting significant connections among variables. Notably, it uncovers the influence of factors like 'Original Balance Amount' on 'Total Deposits,' emphasizing that higher original balances tend to be associated with more substantial total deposits. Additionally, the matrix showcases negative correlations, such as the impact of the 'Debt-to-Income Ratio' on 'Credit Scores,' where increased debt-to-income ratios coincide with lower credit scores. These findings are instrumental for financial institutions seeking to make data-driven decisions and devise strategies to manage risk and enhance lending practices.

The correlation matrix acts as a foundation for predictive modeling, aiding in the identification of key variables that impact loan defaults. It serves as a valuable reference for crafting strategies to minimize charge-offs. By recognizing these interrelationships, financial institutions gain the insight required to optimize lending practices, ultimately contributing to reduced charge-offs and improved financial health.

2.4 Data Distribution - Distribution of Independent Variables:

The analysis of credit scores emerges as a pivotal factor in understanding the dynamics of loan repayment outcomes, unraveling a distinct dichotomy between successful repayments and loan defaults. A noteworthy revelation from the data is the discernible difference in average credit scores, with successful repayments clustering in the range of 680 to 780, while loan defaults align with credit scores ranging from 600 to 700. This stark contrast accentuates the significance of creditworthiness, as manifested through credit scores, in shaping the trajectory of loan repayments. It implies that individuals below the 700 credit score threshold may encounter an elevated risk of defaulting on their loans. These findings constitute a cornerstone for risk assessment and underwriting strategies, equipping financial institutions with actionable insights to enhance lending decisions. The implication is clear: a more nuanced evaluation of credit scores can empower institutions like Meritrust Credit Union to tailor their approaches, effectively mitigating the inherent risks associated with loan defaults and fostering a more resilient and sustainable lending portfolio.

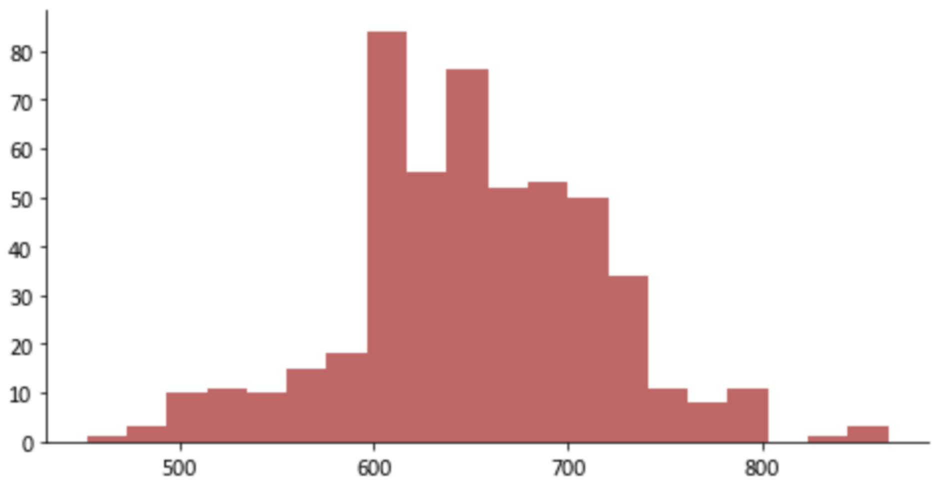

Figure 3. Distribution 'Credit Score'

The examination of interest rates within the dataset unveils a nuanced landscape of borrower preferences and potential predictors of loan default. A clear trend emerges, indicating that the majority of borrowers express a preference for interest rates falling within the 0.18 to 0.25 range, reflecting a comfort zone for loan seekers. This underscores the pivotal role of competitive and affordable interest rates in the lending industry. However, a distinctive pattern surfaces when analyzing the subset of loan defaulters, revealing a mean interest rate clustering between 0.25 and 0.30. This intriguing observation suggests a plausible correlation between higher interest rates and an increased likelihood of loan default. The inference drawn is that individuals accepting loans with elevated interest rates may encounter heightened financial stress or constraints, elevating their susceptibility to default. Integrating these insights into risk assessment models and lending strategies can empower financial institutions, like Meritrust Credit Union, to craft tailored interest rate structures. Such personalized approaches not only mitigate the risk of loan defaults but also optimize overall portfolio performance, ensuring a balanced and resilient lending portfolio.

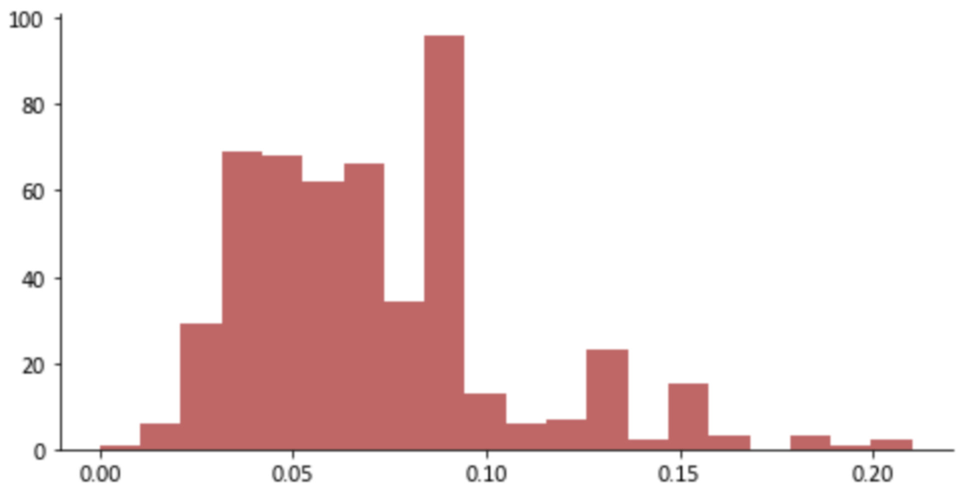

Figure 4. Distribution 'Interest Rate'

Distribution of Dependent Variable:

In this analysis, our primary focus centers on the dependent variable 'ACCTSTATCD,' a pivotal element signifying the account status of loans. A key performance indicator (KPI) guiding our assessment is the "Account Status." This metric is crucial as it distinctly indicates whether a loan has been "Charged Off (CO)" or not. The primary goal is to gauge and enhance predictive accuracy, facilitating effective risk management practices. The "default rate" emerges as a pivotal measure in financial analysis and this indicator serves as a direct reflection of the level of risk associated with our loan portfolio. A higher default rate signals increased riskiness in our loan ventures, emphasizing the critical importance of our analytical efforts. In our dataset, the "Charged Off (CO)" status holds immense significance as it reflects how well or poorly accounts are performing. We simplify this complex data into a binary system where "0" denotes "Not Charged Off," indicating sound account performance, while "1" designates "Charged Off," providing insights into instances where accounts may be underperforming. This strategic simplification is foundational, enabling clearer and more informed decision-making based on detailed account status assessments.

Visual Reference: The distribution of 'ACCTSTATCD' (Account status) shows an imbalanced dataset with most records labeled as 0 (ACT = Active, CO = Charged Off, NPFM = non-performing) and a smaller number labeled as 1 (Charged Off). Visual plots are available for reference.

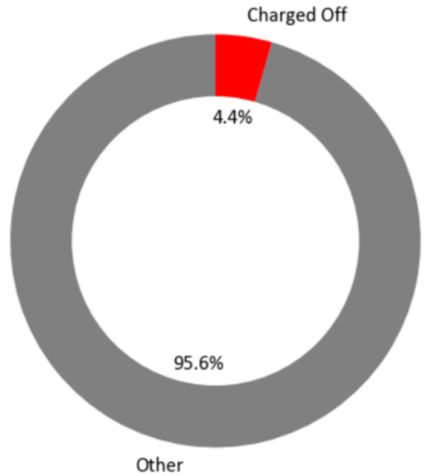

Figure 5. Main KPI - Account Status

Seasonality:

The 'Fund Issued Date' was meticulously analyzed to unveil insights into loan charge-offs, with a focus on understanding temporal patterns rather than traditional feature engineering. The analysis, employing advanced techniques, exposed a distinctive trend in loan issuance dynamics. March exhibited a peak in loan issuances, while December showed a comparatively lower volume, a pattern consistent even in charged-off cases. This observation underscores the impact of temporal factors on loan performance, revealing seasonal variations. The findings serve as a crucial foundation for comprehending how the timing of loan issuance correlates with default rates, empowering stakeholders to proactively address these seasonal fluctuations and optimize risk management strategies for enhanced decision-making.

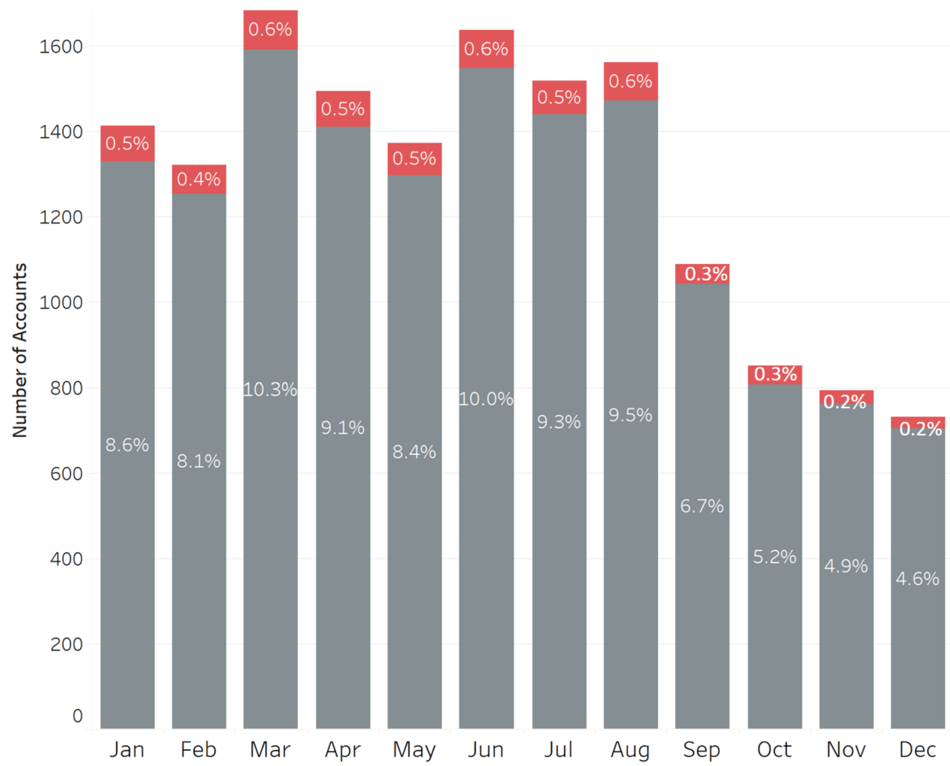

Figure 6. Seasonality on 'Fund Issued Date'

'Credit Score':

The 'Credit Score' underwent a targeted feature engineering process involving grouping to discern intervals associated with elevated loan charge-off rates. Specifically, the majority of loans with charge-off occurrences were concentrated within the credit score interval of 600-699. This strategic grouping sheds light on the critical relationship between credit scores and loan defaults, emphasizing the significance of this feature in predicting and understanding default patterns. The insights gained from this feature engineering enhance our ability to make informed decisions regarding credit risk and aid in developing tailored strategies to mitigate the risk of loans within specific credit score ranges.

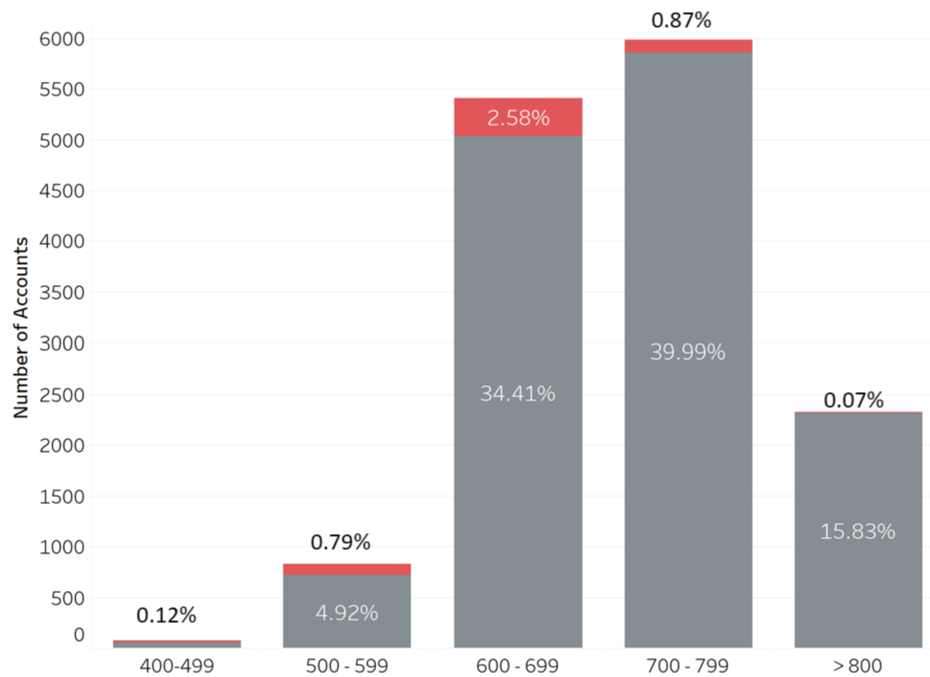

Figure 7. Feature Engineering on 'Credit Score'

'Original Balance Amount':

In a strategic feature engineering initiative, the variable 'Original Balance Amount' underwent meticulous partitioning to enhance the identification of charged-off loans within specific ranges. The implemented code proficiently categorized this parameter into well-defined intervals, shedding light on crucial patterns related to loan distribution. Remarkably, a substantial 50% of the total loans issued clustered within the 10k-20k range, signifying a noteworthy concentration within this particular interval. This meticulous segmentation not only streamlines data analysis but also enriches the interpretability of loan portfolio dynamics, providing actionable insights for decision-makers. The outcomes of this feature engineering process serve as a valuable asset, offering a refined understanding of how original balance amounts influence loan default rates and, consequently, supporting the development of targeted strategies for risk mitigation and effective loan portfolio management.

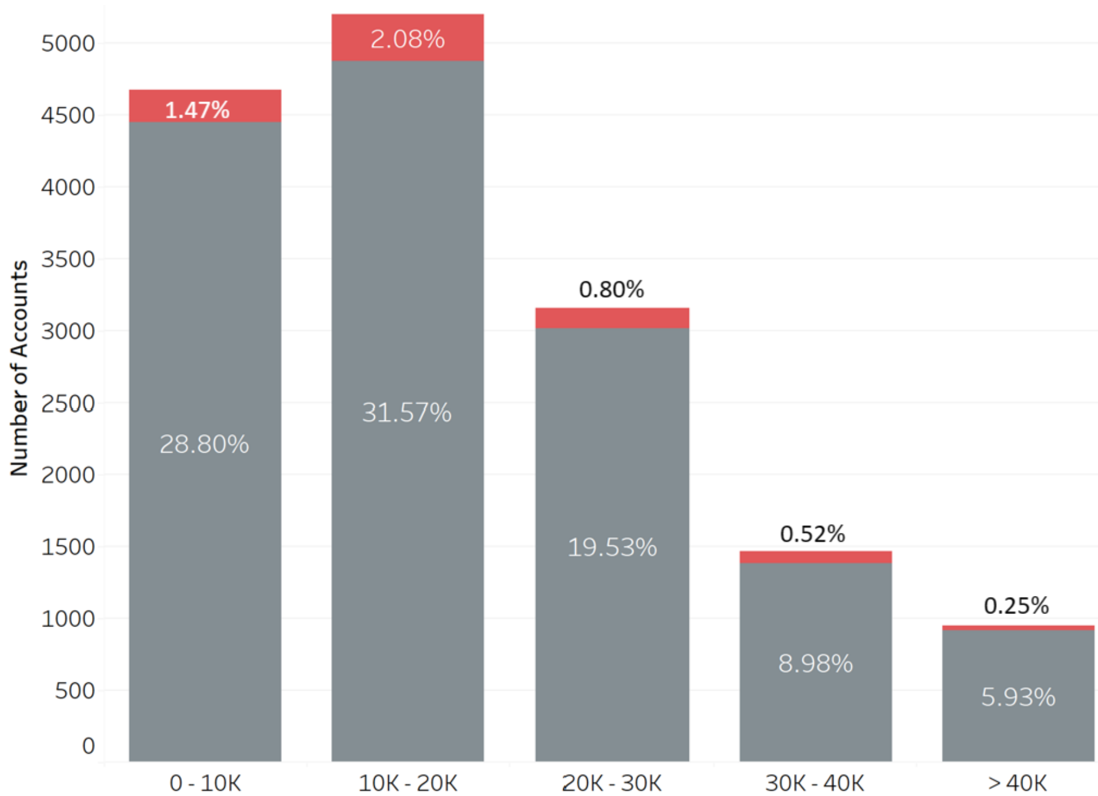

Figure 8. Feature Engineering on 'Original Balance Amount'

'Loan Type':

In the final phase of feature engineering, a thoughtful approach was taken to enhance the 'Loan Type' feature's analytical value. The code implemented a strategic division, categorizing the loans into two distinct types: indirect loans, involving third-party facilitation, and direct loans. This binary classification aimed to provide a nuanced analysis of charge-off rates associated with each loan type. The dataset analysis yielded a notable prevalence of indirect loans compared to direct ones, and intriguingly, the charged-off pattern aligned distinctly with this distribution.

This feature engineering not only refines the categorical representation of 'Loan Type' but also illuminates subtle differences in default rates between loans acquired directly and those facilitated by third-party intermediaries. The insights gained from this approach offer a more profound understanding of how the origination channel of a loan impacts default patterns. This information is invaluable for stakeholders, providing crucial insights for risk assessment and informing strategic decision-making processes within the lending institution, such as Meritrust Credit Union.

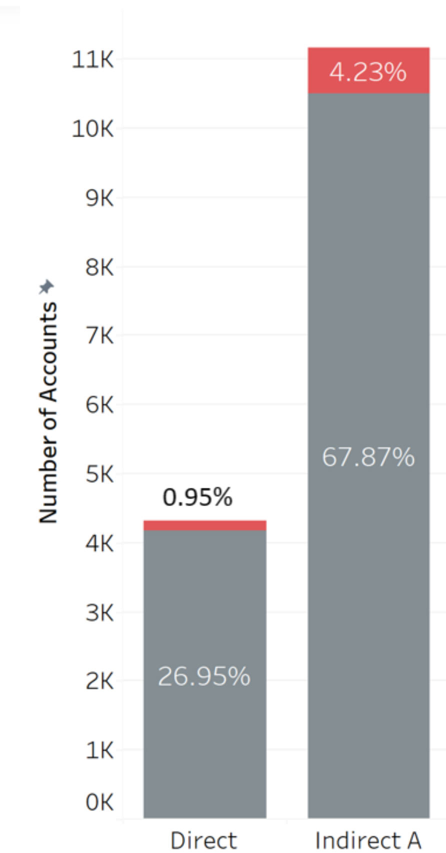

Figure 9. Engineering on 'Loan Type'

In summary, feature engineering of 'Original Balance Amount,' 'Credit Score,' and 'Loan Type' has transformed these variables, making them more interpretable and insightful in the context of the loan default dataset. These engineered features provide valuable information to financial institutions for risk mitigation and data-driven lending decisions.

2.6 Modeling:

In our modeling approach, we meticulously selected key fields based on their substantial influence on loan defaults. These fields Credit Score, Debt-to-Income Ratio, Loan Type, Interest

Rate, and Loan Amount underwent thorough examination to unravel their respective impacts. Lower Credit Scores correlated with higher default rates, emphasizing the critical role of creditworthiness assessment. The Debt-to-Income Ratio provided insights into financial stability, indicating the proportion of income allocated to debt payments. Loan Type distinctions offered nuanced insights for tailored risk management. Higher Interest Rates were potential contributors to financial stress, influencing default risks. The Loan Amount's size impacted default risks, requiring strategic management. This comprehensive examination formed the foundation for our modeling process, contributing to a nuanced understanding of factors influencing loan defaults at Meritrust Credit Union. The implementation of two models - Random Forest and Logistic Regression is crucial to our methodology. An early knowledge of the link between variables and loan defaults is provided by logistic regression, which is well-known for its interpretability and simplicity.

RESULTS

Our project has yielded promising results with a focus on predicting loan defaults and understanding the associated risk factors. Through rigorous data analysis, feature engineering, and modeling, we have achieved notable accuracy rates using Logistic Regression and Random Forest as our primary modeling techniques.

3.1 Model Accuracy & Significance:

The model utilized sophisticated techniques, such as logistic regression analysis and Random Forests Classifier, to comprehensively examine various features influencing loan defaults within Meritrust Credit Union.

We evaluate our models through a series of robust measures, including accuracy, precision, recall, and F1-score. These evaluation metrics allow us to assess the models' performance in detail, understanding not only the overall accuracy but also their ability to correctly classify charged-off and non-charged-off loans. By considering multiple evaluation criteria, we ensure a comprehensive understanding of our models' performance.

Within our methodology, the significance of implementing two models, Logistic Regression and Random Forest, is paramount. Logistic Regression, known for its simplicity and interpretability, provides an initial understanding of the relationship between variables and loan defaults. Achieving an accuracy of 87.2% underscores its potential as a valuable tool. Random Forest, an ensemble technique, further refines predictions by combining multiple decision trees. With an accuracy of 88.9%, it proves its worth in identifying significant features for risk assessment and decision-making. These models are critical for financial institutions to predict defaults and make informed lending decisions.

This analysis focuses on several fields related to loan defaults, including 'Credit Score,' 'Debt-to-Income Ratio,' 'Loan Type,' 'Interest Rate,' and 'Original Balance Amount.' These fields were chosen for examination due to their perceived significance in influencing loan default rates. 'Credit Score' and 'Debt-to-Income Ratio' are crucial indicators of borrowers' creditworthiness and financial stability, respectively. 'Loan Type' provides insights into different default patterns associated with various loan origination channels. 'Interest Rate' and 'Original Balance Amount' offer perspectives on the impact of loan terms and size on default rates. Statistical techniques such as Logistic Regression and Random Forest were employed to model and predict loan defaults. Logistic Regression, known for its simplicity and interpretability, was used for its ability to establish fundamental relationships between variables. Random Forest, as an ensemble technique, provided a more complex and powerful approach to capture non-linearities within the dataset, offering a comprehensive understanding of factors contributing to loan defaults. These techniques enable accurate predictions and identification of significant features crucial for proactive risk management in financial institutions.

Logistic Regression:

Our initial modeling approach with Logistic Regression has provided an accuracy rate of approximately 87.2%. This accuracy reflects the model's capability to correctly classify charged-off and non-charged-off loans. With simplicity and interpretability, Logistic Regression allows us to establish a fundamental understanding of the relationships between various independent variables and the likelihood of loan defaults.

Random Forest:

The Random Forest technique offers a more complex and powerful approach to understanding and predicting loan defaults. With an accuracy rate of approximately 88.9 %, Random Forest harnesses the strength of multiple decision trees. This ensemble method captures the intricacies and non-linearities within the dataset, providing a more comprehensive understanding of the underlying factors contributing to loan defaults. Achieving these accuracy rates underscores the effectiveness of our modeling techniques in addressing the critical task of predicting loan defaults. These results are particularly significant for financial institutions seeking to manage risk proactively and make informed lending decisions. With these models, we can confidently identify significant features that impact loan defaults and use these insights to mitigate risk effectively, ultimately contributing to the organization's financial health and reputation.

Model	Train Accuracy	Test Accuracy
Logistic Regression	87.2	86.9
Random Forests Classifier	88.9	88.7

Evaluation Metrics	Logistic Regression	Random Forests Classifier
F1 Score	0.86	0.88
Precision	0.91	0.91
Recall	0.82	0.85
Error	0.13	0.11

Feature 10. Accuracy and Model Scores of Logistic Regression and Random Forest

The Results section assesses model accuracy through performance metrics like precision, recall, and the F1 score, providing a nuanced understanding of its predictive capabilities. **Random Forests Classifier outperformed Logistic Regression, achieving an accuracy of 88.9% in training and 88.7% in testing, compared to Logistic Regression's 87.2% and 86.9%, respectively.**

3.2 Feature Importance and Significance:

The analysis delves into the importance of features, shedding light on how 'Credit Score,' 'Interest Rate,' 'Loan Type,' and 'Debt-to-Income Ratio' contribute to the model's predictions. The feature engineering process, especially applied to 'Credit Score' and 'Original Balance Amount,' strategically groups and segments data to identify intervals associated with higher loan charge-off rates. This careful feature engineering enhances the dataset's interpretability and analytical capabilities. The subsequent use of Logistic Regression and Random Forest as modeling techniques emphasizes the significance of these features in predicting loan defaults. The interpretation is reinforced by the claim that these key variables profoundly impact the model's predictions, providing a nuanced understanding of the complex relationships between borrower characteristics, loan terms, and the likelihood of defaults. This information empowers financial institutions, like Meritrust Credit Union, with actionable insights to customize their approaches and effectively manage risks in lending decisions. Notably, credit scores in the range of 700-799 are considered most significant, credit scores in the range of 800-899 are identified as the second most important, and Debt category 76-100 is recognized as the least important feature among the 14 features analyzed.

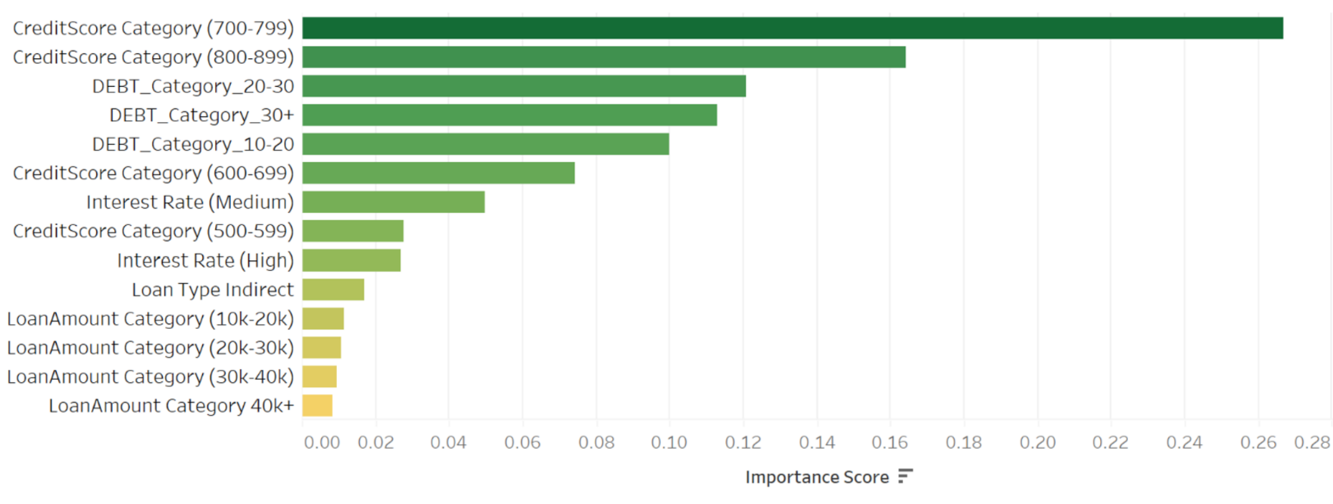

Figure 11. Feature Importance

The identification of significant features is a cornerstone of our project's success, providing valuable insights into the drivers behind loan defaults. Through our models, several key features have emerged as crucial contributors to understanding and predicting default rates. Among these, 'Credit Score' takes center stage, showcasing a strong influence on default rates and underscoring the importance of evaluating borrowers' creditworthiness thoroughly. The 'Debt-to-Income Ratio' proves to be another pivotal feature, shedding light on borrowers' financial stability. A higher ratio is indicative of an increased likelihood of default, emphasizing the role of financial well-being in loan repayment. The one-hot encoded 'Loan Type' adds a layer of sophistication to our analysis, allowing us to discern between direct and indirect loans. This distinction provides nuanced insights into default patterns associated with different loan types. Additionally, the 'Interest Rate' and 'Original Balance Amount' features offer crucial perspectives on the impact of loan terms and size on default rates.

In summary, our project's models, and significant features collectively contribute to proactive risk management, enabling financial institutions to make informed lending decisions and minimize charge-offs. These findings are pivotal in maintaining financial health, reputation, and market value, ensuring sustainable growth and profitability.

3.3 Crucial Independent Variable for Predicting the Dependent Variable:

In this comprehensive analysis, a range of crucial independent variables is scrutinized to evaluate their impact on the dependent variable. These independent variables encompass 'LOAN_TYPE,' which categorizes loans as direct or indirect, 'ORIGBALAMT,' representing the original loan balance and its potential risk-increasing effect, 'NOTEINTRATE,' reflecting the interest rate, 'ORIGTERM_MNTHS,' indicating the original loan term in months, 'CREDITSCORE,' which is the borrower's credit score, and 'FPTI,' denoting the ratio of payment amount to the borrower's income. These variables are of paramount importance due to their potential associations with the dependent variable, and the analysis aims to uncover the intricacies of their relationships and influences on loan performance.

Among these key variables, Credit Score stands out as a critical determinant, with lower credit scores often linked to higher default rates. Simultaneously, the Debt-to-Income Ratio is another pivotal factor, signaling a substantial commitment of the borrower's income to debt payments. Furthermore, Loan Type differentiation assumes significance as it unveils potential variations in default patterns between loans obtained indirectly through dealerships or third parties and those obtained directly. Moreover, the Interest Rate plays a central role, with higher rates increasing the likelihood of defaults. Lastly, the Original Balance Amount emerges as a risk factor, as it can heighten the risk of default when misaligned with the borrower's financial capacity and creditworthiness. These multifaceted variables are examined to reveal their correlations with the dependent variable, providing a comprehensive understanding of their impact on loan performance.

Furthermore, the accuracy in predicting the dependent variable is evaluated using metrics like the area under the ROC curve, offering insights into the model's discrimination between charged off and non-charged off cases. The comprehensive analysis of independent variables in the Results section highlights the distinctive role each variable plays in predicting the dependent variable of loan default. Credit scores demonstrate an inverse relationship with default likelihood, while loan amounts, debt-to-income ratios, interest rates, and loan types contribute unique insights. The temporal dynamics captured by the 'Fund Issued Date' feature unveil seasonal variations influencing default rates. The collaborative analysis enhances the model's predictive accuracy, providing a multifaceted understanding of risk factors associated with loan defaults.

3.4 Matrix Evaluation

Logistic Regression: Confusion Matrix

The Confusion Matrix provides a comprehensive overview of the Logistic Regression model's classification performance. It reveals the accuracy in predicting both non-default and default cases. The model excelled in accurately identifying loans labeled as "Not Charged Off," achieving an accuracy rate of 84.43%. This indicates a robust ability to correctly classify loans with low default risk. However, there is an opportunity for improvement in precision, as evidenced by the 15.57% false positive rate. Addressing this would enhance the model's capability to more accurately identify loans with a higher risk of default. Additionally, the Confusion Matrix highlights a 18.57% misclassification rate for cases labeled as "Not Charged Off," emphasizing the need for further optimization to minimize overlooking actual defaults.

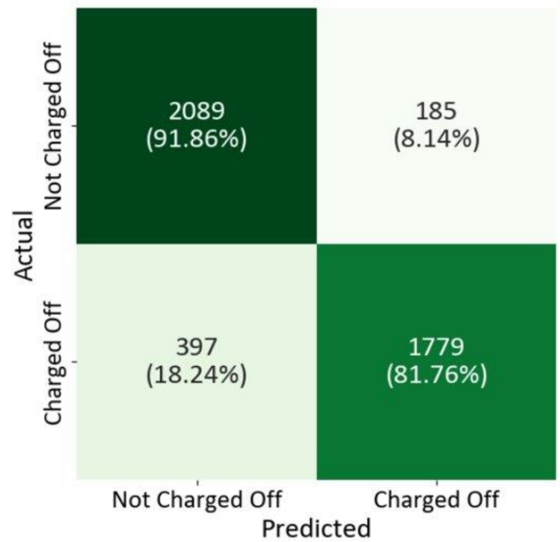

Figure 12. Logistic Regression – Confusion Matrix

ROC Curve Analysis:

The Receiver Operating Characteristic (ROC) Curve provides insights into the Logistic Regression model's discriminatory power. With an impressive Area Under the ROC Curve (AUC) score of 0.82, the model showcases a high ability to distinguish between 'Charged Off' and 'Not Charged Off' cases. The curve's proximity to the upper-left corner indicates the model's strong performance in minimizing false positives and false negatives. This robust discriminatory power reinforces the model's reliability and effectiveness in predicting loan outcomes. The high AUC score underscores the model's accuracy and precision, making it a valuable tool for financial institutions in assessing and managing the risk associated with loan portfolios.

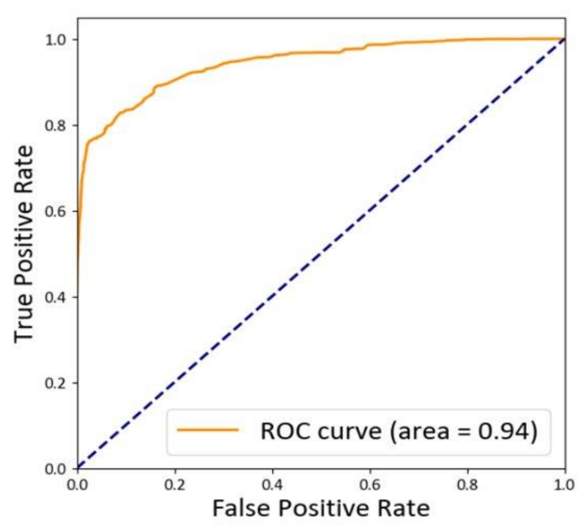

Figure 13. Logistic Regression - ROC

Random Forest: Confusion Matrix

The Confusion Matrix offers an in-depth assessment of the Random Forest model's classification performance. Demonstrating a commendable accuracy of 84.96% in predicting "Not Charged Off" loans, the model excels in identifying cases with a lower risk of default. With an accuracy rate of 83.87% in detecting "Charged Off" cases, the model showcases its proficiency in identifying loans at a higher risk of default. However, there is a precision enhancement opportunity, as seen in the 15.04% false positive rate. Improving precision would be valuable in reducing the misclassification of non-default loans as potential defaults. The 16.13% misclassification rate for cases labeled as "Not Charged Off" suggests an opportunity for further optimization to enhance the model's ability to accurately identify loans with a lower risk of default.

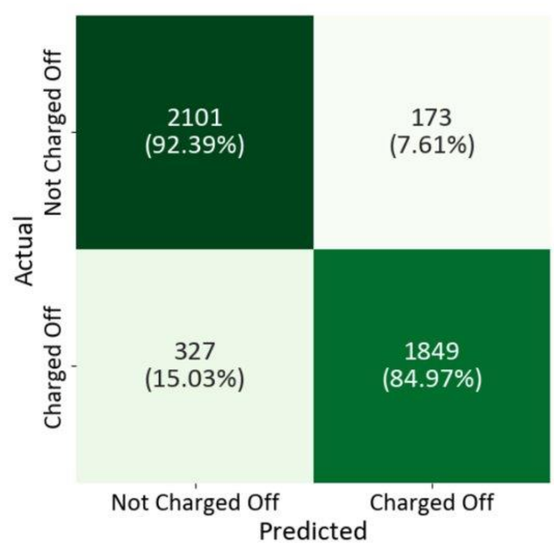

Figure 14. Random Forest Confusion Matrix

ROC Curve Analysis:

The Receiver Operating Characteristic (ROC) Curve is a crucial tool for evaluating the Random Forest model's discriminatory power. With an impressive Area Under the ROC Curve (AUC) score of 0.84, the model exhibits a robust ability to distinguish between 'Charged Off' and 'Not Charged Off' cases. The high AUC score signifies the model's accuracy in balancing true positive and false positive rates, highlighting its reliability in predicting loan outcomes. The ROC Curve's proximity to the upper-left corner underscores the model's effectiveness in minimizing both false positives and false negatives. Financial institutions can leverage the model's robust discriminatory power to make informed decisions in managing and mitigating the risk associated with loan portfolios.

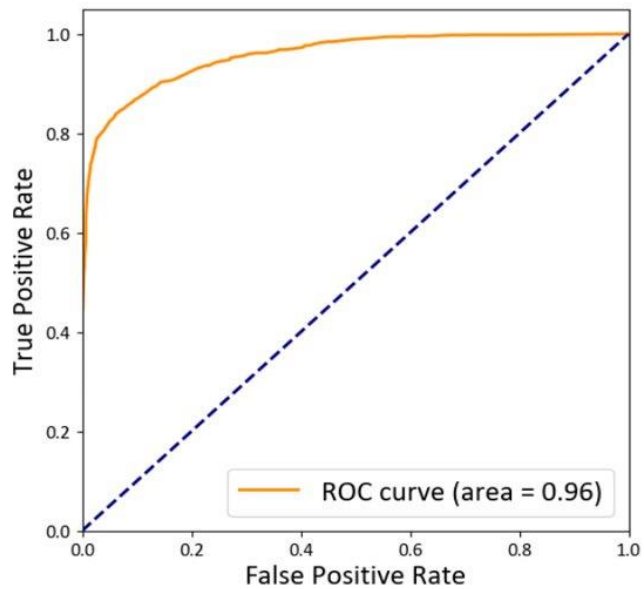

Figure 15. Random Forest – ROC

Additionally, it is important to note that the Appendix section will delve into Log-Odd Ratio Analysis, providing a detailed interpretation of the log-odd ratios resulting from the logistic regression analysis. This section will offer insights into the factors influencing loan default within Meritrust Credit Union, further enriching the discussion on the model's findings and enhancing the comprehensiveness of the overall analysis.

4. DISCUSSION:

Limitations of the Project:

The project, while meticulous, has certain limitations. The research could benefit from a deeper exploration of external factors impacting loan defaults, such as economic trends or regional economic stability. Additionally, the discussion on model evaluation metrics could be more nuanced, providing a detailed examination of false positives and false negatives. Explicit consideration of potential biases in the dataset and model outputs is necessary to ensure the findings are applicable across diverse member profiles. An acknowledgment of these limitations and potential areas for future research would enhance the overall depth of the project.

Recommendations for Future Research:

Future research endeavors could delve into external economic factors influencing loan defaults, providing a more holistic understanding of the financial landscape. A more nuanced discussion on model evaluation metrics, along with considerations for potential biases, would enhance the depth of future analyses. Exploring the impact of socio-economic and regional factors on loan performance could offer valuable insights for credit unions and financial institutions. Additionally, a deeper exploration of the implications of false positives and false negatives in the model outputs could contribute to a more comprehensive understanding.

Based on the insights derived from the analysis, several strategic recommendations can be made to enhance risk management and optimize lending processes for Meritrust Credit Union:

- **Implement Historical Loan Monitoring:**

Leverage the developed model to retrospectively analyze loans granted from 2018 to the present. This will provide valuable insights into the historical performance of loans, enabling a comprehensive understanding of trends and patterns. By forecasting potential defaults in 2025, Meritrust can proactively identify risk areas and tailor strategies to mitigate future challenges.

- **Conduct Scenario Analysis:**

Utilize the model for scenario analysis by incorporating economic variables such as recession and growth. This will enable Meritrust to simulate the impact of various economic conditions on the loan portfolio. By preparing for different scenarios, the credit union can develop robust risk management strategies, ensuring resilience in the face of economic fluctuations.

Use Cases and Implications:

The findings hold practical implications for Meritrust Credit Union and the broader financial sector. The emphasis on 'Credit Score,' 'Debt-to-Income Ratio,' 'Loan Type,' 'Interest Rate,' and 'Original Balance Amount' provides actionable insights for risk mitigation. Robust modeling techniques like Logistic Regression and Random Forest offer a strategic roadmap for managing loan defaults. These insights empower financial institutions to proactively address challenges, fostering sustainable growth and positive community impact in an evolving financial landscape.

Enhance Credit Risk Assessment:

Integrate the model into the credit risk assessment process to improve the precision of identifying high-risk applicants. By incorporating key features like 'Credit Score,' 'Debt-to-Income Ratio,' and 'Loan Type,' Meritrust can strengthen its ability to assess creditworthiness accurately. This, in turn, minimizes the likelihood of approving loans that may turn non-performing.

5.CONCLUSION:

In conclusion, the analysis of Meritrust Credit Union's loan default challenges provides a comprehensive strategy for navigating the complex operational landscape. The study highlights the critical issue of managing loan defaults and its potential impact on financials, reputation, and the member base. By employing Logistic Regression and Random Forest models, the project delivers precise predictions of loan outcomes, emphasizing key features such as 'Credit Score,' 'Debt-to-Income Ratio,' 'Loan Type,' 'Interest Rate,' and 'Original Balance Amount' as crucial factors for risk mitigation. While insightful, there is room for further model refinement, and implementing the study's recommendations will empower Meritrust to proactively address loan default challenges, fostering sustainable growth and reinforcing its commitment to community service. In summary, this project provides a valuable guide for Meritrust Credit Union and contributes essential knowledge to the financial sector, laying the foundation for a data-driven approach to risk management and ensuring continued prosperity amidst evolving financial landscapes.

6. REFERENCES:

1. Yeboah, E. (2018). Determinants of Loan Defaults in Some Selected Credit Unions in Kumasi Metropolis of Ghana. Open Journal of Business and Management, 6(3), 59.

   https://doi:10.4236/ojbm.2018.63059

2. Isbister, J. (1992). The Lending Performance Of Community Development Credit Unions. https://ideas.repec.org/p/ags/ucdrrp/140050.html

3. Financial Stability Board. (2022). Financial Stability Report: Challenges and Developments. https://www.fsb.org/publications/reports/

4. World Bank. (2021). Global Economic Prospects: Pandemic, Poverty, and Policy. https://www.worldbank.org/en/publication/global-economic-prospects

5. Puri, M., Rocholl, J., & Steffen, S. (2017). What do a million observations have to say about loan defaults? Opening the black box of relationships. Journal of Financial Intermediation, 31, 1–15. https://doi.org/10.1016/j.jfi.2017.02.001

6. Mueller, H. M., & Yannelis, C. (31 May 2017). The rise in student loan defaults. Journal of Financial Economics. https://doi.org/10.1016/j.jfineco.2018.07.013

7. Nigmonov, A., & Shams, S. (2021, Dec 8). COVID-19 pandemic risk and probability of loan default: evidence from marketplace lending market. Risk Management. https://doi.org/10.1186/s40854-021-00300-x

8. Foster, B. P., & Zurada, J. (2013). Loan defaults and hazard models for bankruptcy prediction. Managerial Auditing Journal, 28(6), 516-541. https://doi.org/10.1108/02686901311329900

9. Serrano-Cinca, C., Gutiérrez-Nieto, B., & López-Palacios, L. (October 1, 2015). Determinants of Default in P2P Lending. PLOS ONE. https://doi.org/10.1371/journal.pone.0139427

10. Netzer, O., Lemaire, A., & Herzenstein, M. (August 15, 2019). When Words Sweat: Identifying Signals for Loan Default in the Text of Loan Applications. Journal of Marketing Research. https://doi.org/10.1177/0022243719852959

11. Mayer, C., Pence, K., & Sherlund, S. M. (2009). The Rise in Mortgage Defaults. Journal of Economic Perspectives, 23(1), 27-50. https://doi.org/10.1257/jep.23.1.27

12. Appati, J. K., Owusu, E., Quainoo, R., & Mensah, S. A Deep Learning Approach for Loan Default Prediction Using Imbalanced Dataset. International Journal of Intelligent Information Technologies.
https://doi.org/10.4018/IJIIT.318672

7.APPENDIX:

Feature	Class	Odds Ratio	Insights
Credit Score	500-599	0.021***	Higher credit scores significantly decrease the odds of loan default, highlighting the pivotal role of good credit in risk mitigation.
	600-699	0.016***	
	700-799	0.001***	
	800-900	0.000***	
Loan Amount	10k-20k	1.474***	Higher loan amounts are associated with increased odds of default, emphasizing the importance of managing risk in larger loan transactions.
	20k-30k	1.964***	
	30-40k	2.607***	
	40k+	3.282***	
Debt to Income Ratio	10-20	0.025***	A lower debt-to-income ratio significantly reduces the odds of loan default.
	20-30	0.037***	
	30+	0.072***	
Interest Rate	Moderate	0.302***	Higher interest rates are linked to higher odds of default, suggesting that low interest serves as a protective factor against loan default.
	High	0.770***	
Loan Type	Indirect	1.286***	Indicates an increased likelihood of default compared to other loan types.

Figure 16. Logs Odds Ratio

Credit Score Analysis:

Odds ratio of 0.021 (*) for scores between 500-599, emphasizing higher odds of loan default with lower credit scores.

Odds decrease with rising credit scores, reaching 0.000 () for scores between 800-899, highlighting the risk mitigation role of good credit.

Loan Amount Analysis:

Higher loan amounts correspond to increased default likelihood, with odds ratios of 1.474 (), 1.964 (), 2.607 (), and 3.282 () for amounts 10k-20k, 20k-30k, 30k-40k, and 40+, respectively.

Underscores the importance of meticulous risk assessment in larger loan transactions.

Debt-to-Income Ratio Analysis:

Higher ratios (e.g., 30+) exhibit higher odds of default (odds ratio of 0.072 ()), while lower ratios (e.g., 10-20) show a diminishing likelihood (odds ratio of 0.025 ()).

Highlights the correlation between balanced income-to-debt ratios and reduced likelihood of default.

Interest Rate Analysis:

Higher interest rates are associated with increased odds of default (0.302 () for moderate rates, 0.770 () for high rates). Suggests lower interest rates as a potential risk mitigation strategy.

Loan Type Analysis:

'Indirect' loans have an odds ratio of 1.286 (***), indicating an increased likelihood of default compared to other loan types.

Quantifies the comparative risk associated with this specific loan category.

Note:

Odds ratios are presented with corresponding significance levels: (): p < 0.001, (): p < 0.01.*=