



Delivered By-

Anirudh Chandnani

Rohith Bommagani

Akshita Sharma

Harish Konduru

Srutej Kodiganti

Dharav Shah

**Table of contents**

|  |  |  |
| --- | --- | --- |
| **S. No**. | **Topics** | **Page No.** |
| 1 | Executive Summary | 3 |
| 2 | Data | 4 |
| 3 | Research Questions | 6 |
| 4 | Analytical Methods | 12 |
| 5 | Figures and Tables | 19 |
| 6 | References | 21 |

**Executive Summary**

This report delivers a comprehensive analysis of the Lending Club Loan Data, with the objective of enhancing financial decision-making and investor satisfaction. Our research leverages advanced analytics, predictive modeling, and data-driven insights to provide strategic recommendations for investment decisions.

We anticipate uncovering significant insights in our forthcoming study of loan performance that will substantially impact our investment strategies. Through meticulous data examination, we aim to identify the primary factors influencing loan defaults and paybacks, such as creditworthiness, loan terms, borrower behavior, and economic trends. We intend to develop precise predictive models that will facilitate proactive investment decisions and risk assessment using state-of-the-art predictive modeling techniques, including decision trees and regression analysis. Our goal is to conduct a thorough root cause analysis to ascertain the origins of loan defaults. The findings of this analysis are expected to inform policy changes, such as improving investor communication, optimizing loan offerings during peak demand periods, and enhancing risk management. By adopting these data-driven insights, we strive to minimize defaults, boost financial efficacy, and ultimately secure a competitive edge for investors in the peer-to-peer lending market.

Our project bridges the gap between data science and operational excellence, positioning our organization for success in a dynamic financial ecosystem.

**Data**

**Data Collection**

The Lending Club Loan data was collected from Kaggle (Nathan George). The glossary was obtained from Data World (Josh Devlin).

**Data Legal Privacy**

Our dataset, meticulously gathered from reputable sources such as Kaggle and Data World, alleviates any apprehensions regarding legal or privacy implications. The process entails simply creating an account on these esteemed platforms, thereby granting users unfettered access to the data for professional utilization.

**Data Format**

The data was obtained in comma-separated value (CSV) format and would be serialized in pickle format for faster processing via Python.

**Integrated Datasets**

In alignment with our research objectives, we have identified a singular primary dataset for our project. We are also referring to an additional data frame serving as an exhaustive glossary, elucidating the significance of each feature within our primary dataset. There is no integration needed as we are only referring to the data frame for the glossary. This strategic approach ensures a comprehensive understanding and analysis of the data, thereby elevating the professionalism and depth of our project.

**Data Cleaning**

The cleaning of the Lending Club dataset is imperative due to numerous missing values and several categorical variables that necessitate encoding. Furthermore, a notable data imbalance is observed, as this project pertains to fraud detection. To address this imbalance between good loans vs bad loans instances effectively, oversampling (SMOTE) (Cory M.) or undersampling methodologies will be implemented. Python will serve as our primary tool for data analysis, ensuring a robust and comprehensive approach to the task at hand.

**Dataset Attributes/ Variables**

Our dataset comprises 155 variables that can influence a customer's financial decision-making. These variables encompass diverse factors such as loan amount, employee title, employee tenure, home ownership status, as well as intricate banking details including account type, opening date, last payment amount, and last payment date, among others. The variable “Loan\_Status” is our response variable which has 9 unique values amongst which we are treating "Charged Off", "Default", "Does not meet the credit policy. Status: Charged Off", "In Grace Period", "Late (16-30 days)", "Late (31-120 days)" as bad loans whereas “Fully Paid”, “Current” and “Does not meet the credit policy. Status: Fully Paid” are considered amongst good loans.

**Final Dataset Size (Number of variables, etc.)**

The final raw dataset has a total of 2260701 rows and 155 variables with 1957089 good loans and 303612 bad loans.

**Research Questions**

**The following are the six research questions we would like to answer by analyzing the dataset:**

**Research Question #1:** What factors contribute to borrowers' loan repayment rates? We'll examine elements such as their employment tenure, income levels, demographic attributes, and the purposes for their loans. How do these factors interact to influence loan performance, and do we notice any noteworthy trends or correlations among different borrower groups and loan intentions?

**Attributes/Variables:**emp\_length, annual\_inc, purpose, addr\_state, loan\_status, loan\_amnt, int\_rate, grade, sub\_grade, term, issue\_id, dti, home\_ownership, verification\_status, delinq\_2yrs, inq\_last\_6mths, open\_acc, pub\_rec, revol\_bal, revol\_util, total\_acc, initial\_list\_status, recoveries, collection\_recovery\_fee, last\_pymnt\_d, last\_pymnt\_amnt, next\_pymnt\_d, pub\_rec\_bankruptcies, tax\_liens, total\_bal\_ex\_mort, total\_bc\_limit, total\_il\_high\_credit\_limit, acc\_open\_past\_24mths, num\_tl\_op\_past\_12m, pct\_tl\_nvr\_dlq, fico\_range\_low, fico\_range\_high, last\_fico\_range\_low, last\_fico\_range\_high, revol\_bal, revol\_util, term, int\_rate, installment, total\_pymnt, total\_rec\_prncp, total\_rec\_int, pub\_rec.

**Managerial Decision Making:**

By answering these questions managers can make more informed decisions across their operations if they have insights into the factors that affect borrowers' capacity to repay loans. They can adjust how they evaluate risks and grant loans, emphasizing elements that point to borrowers' propensity to repay their debts. Managers can recruit ideal loan candidates by customizing their marketing strategies and improving their understanding of various borrower categories. They can even design brand-new loan choices or modify already-existing ones to meet client's requirements better. Additionally, managers can reduce risks by altering interest rates or providing financial guidance to individuals who may find it difficult. Managers can fine-tune their strategy and ensure compliance with regulations while providing loans to individuals in need by monitoring loan performance and analyzing data.

**Research Question #2:**

What patterns can be observed in a borrower's financial behavior in relation to loan repayment rates, and how do these behaviors correlate with economic indicators?

**Variables/Attributes:**

Emp\_length, annual\_inc, purpose, addr\_state, loan\_status, loan\_amnt, int\_rate, grade, sub\_grade, term, issue\_d, dti, home\_ownership, verification\_status, delinq\_2yrs, inq\_last\_6mths, open\_acc, pub\_rec, revol\_bal, revol\_util, total\_acc, initial\_list\_status, recoveries, collection\_recovery\_fee, last\_pymnt\_d, last\_pymnt\_amnt, next\_pymnt\_d, pub\_rec\_bankruptcies, tax\_liens, total\_bal\_ex\_mort, total\_bc\_limit, total\_il\_high\_credit\_limit, acc\_open\_past\_24mths, num\_tl\_op\_past\_12m, pct\_tl\_nvr\_dlq, fico\_range\_low, fico\_range\_high, last\_fico\_range\_low, last\_fico\_range\_high, revol\_bal, revol\_util.

**Managerial Decision-making:**

In-depth pattern analysis of borrowers' financial behaviors connected to loan repayments and its correlation with economic indicators is pivotal. For managerial staff, this transcends beyond mere data interpretation; it's about crafting a predictive, responsive, and dynamic credit strategy. With over 155 individual attributes potentially influencing loan repayments, the challenge lies in identifying the most predictive factors. By leveraging data insights, managers can proactively adjust credit policies, aligning with the evolving economic landscape. This proactive stance not only enhances the borrower's experience but also safeguards the lender's portfolio. Anticipating shifts in borrower behavior due to economic trends can shape strategic initiatives, from developing new financial products to adjusting risk mitigation plans. Such preemptive strategies are integral for maintaining portfolio health and achieving customer satisfaction.

**Research Question #3:**

Which loan characteristics most significantly impact default rates and what is their relation to recovery rates post-default?

**Variables/Attributes:**

Loan\_amnt, term, int\_rate, installment, grade, sub\_grade, total\_pymnt, total\_rec\_prncp, total\_rec\_int, recoveries, collection\_recovery\_fee, loan\_status, dti, total\_bal\_il, total\_bal\_ex\_mort, total\_bc\_limit, total\_il\_high\_credit\_limit, total\_acc, total\_rec\_late\_fee, delinq\_2yrs, mths\_since\_last\_delinq, pub\_rec, mths\_since\_last\_record, mths\_since\_last\_major\_derog.

**Managerial Decision-making:**

Loan characteristics and their contribution to default rates is a complex area that demands careful consideration from lending institutions. For managerial teams, discerning these characteristics is not just a risk mitigation exercise; it's an opportunity to redefine the loan origination process. A deep dive into default influencers such as loan amount, term, and borrower's creditworthiness informs more than just risk assessment—it shapes the very products offered to consumers. By correlating default rates with recovery post-default, management can craft strategies that optimize recovery processes, improving financial outcomes even in default scenarios. As managers translate these insights into action, they balance risk with growth, securing the lender's position in the marketplace. This understanding becomes particularly crucial when entering new markets or launching novel financial products, ensuring that risk frameworks are both robust and adaptable.

**Research Question #4:**

Can machine learning models predict loan defaults and which features contribute most to the predictiveness of these models?

**Variables/Attributes:**

Loan\_status, grade, sub\_grade, loan\_amnt, term, int\_rate, dti, total\_pymnt, total\_rec\_prncp, total\_rec\_int, total\_rec\_late\_fee, recoveries, collection\_recovery\_fee, annual\_inc, emp\_length, home\_ownership, verification\_status, purpose, addr\_state, fico\_range\_high, fico\_range\_low, inq\_last\_6mths, mths\_since\_last\_delinq, mths\_since\_last\_record, open\_acc, pub\_rec, revol\_bal, revol\_util, total\_acc, num\_tl\_90g\_dpd\_24m, num\_tl\_op\_past\_12m, pct\_tl\_nvr\_dlq.

**Managerial Decision-making:**

Machine learning models hold transformative potential for predicting loan defaults, representing a paradigm shift in risk management practices. Through these models, managers can peer into the future of loan performance, steering clear of potential defaults before they crystallize. This predictive capability allows for the tailoring of intervention strategies, personalized borrower communication, and the recalibration of financial products. Moreover, identifying the most predictive features aids in the optimization of these models, ensuring that loan officers and underwriters have access to state-of-the-art tools for decision-making. The operational benefits extend beyond risk aversion, fostering a culture of innovation and continuous improvement within the lending institution. As these analytical tools become embedded in the lending process, they pave the way for more informed, data-driven managerial decisions, cementing a competitive edge in the financial services industry.

**Research Question #5:**

How does the ratio of income to loan amount affect the likelihood of a borrower defaulting, and what implications does this have for loan structuring?

**Variables/Attributes:**

Annual\_inc, loan\_amnt, dti, loan\_status, funded\_amnt, funded\_amnt\_inv, total\_pymnt, total\_pymnt\_inv, total\_rec\_prncp, installment.

**Managerial Decision-making:**

The ratio of a borrower's income to the loan amount is a vital gauge of financial stability and risk. For managers, insights derived from this ratio can drive the structuring of loans to better align with repayment capacities, fostering financial health across the borrower spectrum. Implementing strategies based on these insights can revolutionize loan offerings, leading to customized financial solutions that cater to the unique circumstances of each borrower. By doing so, lenders not only act in their financial interest but also embody a customer-centric approach that can enhance trust and loyalty. Adjustments in loan terms based on these findings can mitigate default risks, ensuring sustainability and profitability. This strategic approach underpins not just a singular loan decision but informs the overarching philosophy of lending, emphasizing responsible and empathetic lending practices.

**Research Question #6:**

Research Question: What is the relationship between the borrower's credit utilization ratio (revol\_util) and their likelihood of default (loan\_status), controlling for other financial and demographic variables?

**Variables/Attributes:**

Loan\_status, revol\_util, annual\_inc, dti, int\_rate, loan\_amnt, purpose, total\_acc, verification\_status, home\_ownership, sub\_grade, num\_tl\_90g\_dpd\_24m, pub\_rec, pub\_rec\_bankruptcies

**Managerial Decision-making:**

Analyzing the impact of credit utilization on default rates can inform lenders about the importance of this factor in assessing borrower creditworthiness. Lenders can use this information to refine their risk assessment models, adjust interest rates, and set appropriate credit limits based on borrowers' credit utilization patterns. Additionally, understanding the relationship between credit utilization and default can guide borrowers in managing their credit wisely to improve their chances of loan approval and lower default risk.

**Analytical Methods**

**Research Question #1:**

What factors contribute to borrowers' loan repayment rates? We'll examine elements such as their employment tenure, income levels, demographic attributes, and the purposes for their loans. How do these factors influence loan performance, and do we notice any noteworthy trends or correlations among different borrower groups and loan intentions?

**Methods**:

Several analytical techniques are used to identify the factors impacting borrowers' loan payback rates. To gain a basic grasp of borrower characteristics, descriptive statistics first offers insights into the main elements of the data, such as the mean, median, and frequency distributions. Finding the direction and degree of links between variables such as income levels and loan performance is made easier using correlation analysis, which also offers important insights into other factors that may have an impact. Regression analysis provides a greater knowledge of predictive factors by exploring the relationship between independent variables such as income levels, employment duration, and loan repayment rates. The Chi-Square Test of Independence provides insight into borrowing patterns by examining possible connections between categorical variables, such as loan reasons and demographic traits. Furthermore, by identifying borrower segments according to common attributes, Cluster Analysis makes it possible to distinguish different borrower groups that exhibit comparable loan repayment tendencies. Decision Trees and Random Forests employ machine-learning approaches to identify important factors influencing loan payback rates and reveal complex correlations between borrower qualities.

**Reason for choosing method**:

Descriptive Statistics gives us a clear picture of who our borrowers are, helping us understand their demographics and characteristics. Correlation Analysis guides us in identifying potential links between borrower attributes and loan performance, prompting deeper exploration into what drives repayment behavior. Regression Analysis offers a detailed view of how various factors impact loan repayment rates, aiding in making well-informed lending decisions. The Chi-Square Test of Independence and Cluster Analysis provides valuable insights into borrowing trends and segmentation, allowing us to tailor our lending strategies accordingly. Finally, Decision Trees and Random Forests delve into the intricacies of borrower data, revealing complex relationships and enabling us to make more accurate predictions about loan repayment behavior. These methods empower us to navigate the lending landscape effectively and better serve our borrowers' needs.

**Software Programs/ Applications for analyzing**:

1. R, Python (using libraries like NumPy, Pandas, StatsModels, scikit-learn, SciPy, Packages: Cluster), Jupiter.
2. Excel, Tableau or Power-Bi, Python (Matplotlib, Seaborn, scikit-learn), R (ggplot2, corrplot, ggfortify, vcd, factoextra, rpart.plot, randomForestExplainer).

**Acquiring Software**:

The software programs are easily available through the CSUF Division of Information Technology student portal and open-source platforms.

**Research Question #2:** What patterns can be observed in borrowers' financial behavior with loan repayment rates, and how do these behaviors correlate with economic indicators?

**Methods:**

Descriptive Statistics and Multivariate Analysis will be utilized to discern patterns and correlations in financial behavior and loan repayment. Predictive Modeling will forecast the likelihood of loan repayment.

**Reason for Chosen Method(s):**

The chosen method of descriptive statistics allows us to gain a foundational understanding of the distribution and characteristics of each variable individually, enabling us to identify any outliers or trends. Multivariate analysis, on the other hand, delves deeper into the interrelationships between various variables, helping uncover complex patterns that may not be apparent through univariate analysis alone. Finally, predictive modeling is essential for forecasting borrower behavior and loan repayment rates, providing actionable insights for developing strategies to mitigate risks and optimize lending practices in response to economic indicators.

**Software Programs / Applications for Analyzation:**

R (RStudio for statistical analysis), Python (pandas and scikit-learn for data manipulation and predictive modelling), Excel (for initial data exploration and visualization), and Tableau (for advanced visual analytics).

**Acquiring Software:**

These tools are accessible through various academic and open-source platforms, with extensive support for data analytics processes.

**Research Question #3:** Which loan characteristics most significantly impact default rates and what is their relation to recovery rates post-default?

**Methods:**

Logistic Regression to identify the loan characteristics that most significantly affect default rates. Time-series analysis for understanding trends and patterns in recovery rates post-default.

**Reason for Chosen Method(s):**

Logistic Regression is well-suited for analyzing the impact of various loan characteristics on binary outcomes such as default rates. By employing this method, we can quantify the significance of each variable in predicting loan defaults and identify the most influential factors. Additionally, time-series analysis enables us to track recovery rates over time, offering insights into the effectiveness of recovery strategies and potential correlations with economic cycles or policy changes. Together, these methods provide a comprehensive understanding of the relationship between loan characteristics, default rates, and post-default recovery dynamics, essential for informing risk management strategies and decision-making in lending practices.

**Software Programs / Applications for Analyzation:**

R (RStudio for logistic regression and time-series modelling), Python (statsmodels for advanced statistical modelling), Excel (for trend analysis and preliminary visualizations), and Tableau (for dynamic and interactive visual representations of data trends).

**Acquiring Software:**

Both R and Python are available for free and are supported by a wide community for analytical tasks. Excel is part of the Microsoft Office suite, and Tableau offers academic licenses for learning purposes.

**Research Question #4:** Can machine learning models predict loan defaults and which features contribute most to the predictiveness of these models?

**Methods:**

A combination of Feature Selection techniques and Machine Learning algorithms like Decision Trees, Random Forest, and Logistic Regression will be used to predict loan defaults and identify the most influential features.

**Reason for Chosen Method(s):**

The chosen methods leverage the strengths of feature selection and machine learning algorithms to predict loan defaults effectively. Feature selection techniques allow us to identify the most influential variables in predicting defaults, enhancing model accuracy and interpretability. Decision Trees and Random Forests excel in providing clear insights into feature importance, aiding in understanding the underlying relationships between variables and default outcomes. Logistic Regression complements this by estimating probabilities, offering a nuanced understanding of the likelihood of default for individual loans based on their features. Together, these methods enable robust prediction of loan defaults while providing valuable insights into the key factors driving default risk.

**Software Programs / Applications for Analyzation:**

Python (SciPy, NumPy, pandas, scikit-learn for machine learning), R (RStudio, caret package for model training), Excel (for data preprocessing), and Tableau (for visualizing model performance and feature importance).

**Acquiring Software:**

Python and R are open-source and freely available. Excel is part of the Microsoft Office suite, and Tableau provides licenses for educational purposes or a free public version.

**Research Question #5:** How does the ratio of income to loan amount affect the likelihood of a borrower defaulting, and what implications does this have for loan structuring?

**Methods:**

Regression Analysis will be conducted to explore the relationship between the income-to-loan amount ratio and default likelihood. Sensitivity Analysis will also be performed to understand how changes in the income-to-loan amount ratio affect default probability.

**Reason for Chosen Method(s):**

Regression Analysis is chosen because it allows for a detailed examination of how changes in the income-to-loan ratio influence the likelihood of default. By quantifying this relationship, lenders can better understand the risk associated with different income levels relative to loan amounts. Additionally, Sensitivity Analysis is crucial for assessing the stability and reliability of this relationship across various economic scenarios, providing insights into potential vulnerabilities and informing loan structuring decisions to mitigate default risks effectively. Together, these methods offer a comprehensive approach to understanding the impact of income-to-loan ratios on default likelihood and its implications for loan structuring strategies.

**Software Programs / Applications for Analyzation:**

R (RStudio for statistical modeling), Python (pandas for data manipulation, statsmodels for regression analysis), Excel (for preliminary data analysis and visualizations), and Tableau (for advanced data visualization and sensitivity analysis).

**Acquiring Software:**

The software tools are widely accessible through educational licenses or open-source platforms, suitable for comprehensive data analysis.

**Research Question #6:**

Research Question: What is the relationship between the borrower's credit utilization ratio (revol\_util) and their likelihood of default (loan\_status), controlling for other financial and demographic variables?

**Methods:**

Logistic regression is a commonly used method for analyzing the relationship between a binary outcome variable (default or not) and one or more predictor variables (such as revol\_util, income, age, etc.). In this case, you would include revol\_util along with other relevant variables as predictors of loan default.

**Reason for Chosen Method(s):**

Logistic regression excels in analyzing binary outcomes, such as loan defaults. Its utility lies in the interpretability of its output, which elucidates the relationship between variables like credit utilization and the likelihood of default. The method facilitates clear communication with stakeholders by producing precise coefficients that reflect the magnitude and direction of the predictors' impact. Additionally, logistic regression's multivariate capability allows for the adjustment of confounding factors, offering a robust analysis of default risk against a backdrop of various demographic and financial covariates.

**Software Programs / Applications for Analyzation:**

Python (SciPy, NumPy, pandas, scikit-learn for machine learning), R (RStudio, caret package for model training), Excel (for data preprocessing), and Power BI (for visualizing model performance and feature importance).

**Acquiring Software:**

The software tools are widely accessible through educational licenses or open-source platforms, suitable for comprehensive data analysis.

# **Figures and Tables**

# 

# **Types of Figures and Tables:**

1. Scatter Plot
2. Histogram
3. Descriptive Statistics
4. Bar Chart
5. Pie Chart
6. Decision Trees
7. Pivot Tables
8. Network Graph
9. Violin Plots

**Reasons for these specific figures/tables:**

1. Scatter Plot: This can be used to visualize the relationship between loan amounts and interest rates or any pair of variables, helping to identify patterns, correlations, or outliers in the data.
2. Histogram: Useful for showing the distribution of loans by interest rate or loan amount, which can indicate the most common rates or loan sizes.
3. Descriptive Statistics: Provides a summary of the dataset, including measures like the mean, median, and mode of loan amounts, as well as borrower credit scores.
4. Bar Chart: This can compare the number of loans by state or by loan grade, giving a visual representation of where and how the loans are distributed.
5. Pie Chart: Shows the proportion of loans by purpose (e.g., debt consolidation, home improvement), offering insights into the primary reasons borrowers are seeking loans.
6. Decision Trees: Helps in understanding the factors that lead to a loan being fully paid or charged off, by mapping out decision points based on borrower information.
7. Pivot Tables: Allows for complex analysis, such as examining the default rates by credit grade or the average loan amount by state.
8. Network Graphs: Can illustrate the connections between borrowers and lenders or show how peer-to-peer lending networks are structured.
9. Violin Plots: To illustrate the comparison between different grades, loan amounts, and loan status.

**Software programs that will be used to create the figures/tables:**

1. Excel
2. Power BI
3. Python
4. Jupyter IDE
5. R Studio

# 

# **References**

1. Cory Makin (May 14, 2022) Synthetic Minority Over-sampling TEchnique (SMOTE). Retrieved March 8, 2024, from <https://medium.com/@corymaklin/synthetic-minority-over-sampling-technique-smote-7d419696b88c>
2. A Logistic Regression Model for consumer default risk Eliana Costa e Silva, Isabel Cristina Lopes, Aldina Correia and Susana Fariac. (Published 2020 May 5). <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9041570/>
3. Nathan, G. (2019)., All Lending Club loan data. Retrieved March 8, 2024, from <https://www.kaggle.com/datasets/wordsforthewise/lending-club/data>
4. Josh, D (2017)., Lending Club Loan Data 2007-11. Retrieved March 8, 2024, from <https://data.world/jaypeedevlin/lending-club-loan-data-2007-11/workspace/file?filename=LCDataDictionary.csv>
5. Loan Default Rate and its Impact on Profitability in Financial Institutions by E. Ntiamoah, E. Oteng, Beatrice Opoku, Anthony Siaw (Published 2014): ​​<https://www.semanticscholar.org/paper/Loan-Default-Rate-and-its-Impact-on-Profitability-Ntiamoah-Oteng/fbd2a38f80fcbb256dd31e49886cbc7ccc961c8f>