Correlation Analysis Between Credit Scores and Number of Missing Payments

Avinash Bunga

Master of Science in Information Systems and Business Analytics

Park University

CIS625HOS2P2025 Machine Learning for Business

Professor: Abdelmonaem Jornaz

March 23, 2025

Correlation Analysis Between Credit Scores and Number of Missing Payments

Introduction

Credit scores serve as a cornerstone metric in the lending industry, often influencing decisions on loan approvals and interest rates. This analysis investigates the relationship between credit scores and the frequency of missed payments. The outcome of this examination aims to provide actionable insights that can enhance lending strategies while maintaining strict adherence to ethical standards and regulatory frameworks (Consumer Financial Protection Bureau, 2024).

Strategic Framework for Analysis

Comprehensive Data Acquisition: The analysis begins with the procurement of a well-structured, anonymized dataset. Essential variables include credit scores, historical records of missed payments, loan amounts, borrower income levels, loan terms, and relevant demographic factors. Adherence to data privacy mandates such as the General Data Protection Regulation (GDPR) and California Consumer Privacy Act (CCPA) ensures compliance and protection of customer information (Thoropass, n.d.).

Hypothetical Example – Real-World Scenario: Consider a hypothetical case study: "Metro Finance," a regional lender, observes that a subset of borrowers with credit scores between 650-700 have an unexpectedly high rate of missed payments. On deeper investigation, it becomes evident that many of these borrowers recently experienced job layoffs due to a local factory closure. This highlights how external economic shocks, despite stable credit histories, can temporarily disrupt repayment behaviors. Such examples stress the importance of considering socioeconomic factors in tandem with credit data.

Data Integrity and Preprocessing: Data will undergo rigorous cleaning to eliminate inaccuracies, outliers, and incomplete records. Standardization processes will be applied to normalize numerical variables, ensuring comparability across different data points (Fatima, 2024).

Visual and Descriptive Exploratory Analysis: Descriptive statistical methods will summarize key metrics, including the mean, median, and standard deviation of credit scores and missed payments. Visual tools such as scatter plots and box plots will provide an intuitive understanding of data distributions and potential patterns (Donnelly, 2024).

Correlation Metrics and Relationship Quantification: The Pearson correlation coefficient will be used to measure the linear association between credit scores and missed payments. In cases where the data reveals non-linear tendencies, Spearman’s rank correlation will offer a more robust alternative. To further deepen understanding, a simple linear regression model may be constructed, enabling the quantification of how changes in credit scores potentially influence the likelihood of missed payments (Sereno, 2024).

Findings Interpretation and Communication: Clear, concise interpretation of the analytical outcomes will be prioritized. Emphasis will be placed on the strength and direction of the correlation, alongside an evaluation of the regression model's predictive accuracy (Sharma, n.d.).

Ethical Implications and Industry Concerns

Fairness and Socioeconomic Bias: Credit scoring models may unintentionally reflect systemic inequities, adversely affecting underrepresented or economically disadvantaged groups. Recognizing and mitigating such biases is essential to uphold principles of fairness (Campisi, 2021).

Data Confidentiality and Privacy Risks: Handling sensitive financial data introduces significant privacy concerns. Ensuring that personal identifiers are removed and implementing robust encryption methods are critical steps (Devane, 2022).

Transparency and Accountability: Misinterpretation or misuse of analytical findings can result in discriminatory lending practices. Ensuring transparency in methodology and results is vital to prevent such outcomes (Asurity, 2024).

Analytical Challenges and Considerations

Data Quality Assurance: High-quality, reliable data is the foundation of sound analysis. Gaps or inaccuracies in credit histories can compromise the validity of results (Atlan, 2023).

Distinguishing Correlation from Causation: While correlations may indicate relationships, they do not establish causal links. Other factors such as income variability, employment status, or economic conditions might confound the observed relationships (JMP, n.d.).

Multicollinearity and Confounding Variables: Variables such as income levels and loan amounts may display interdependencies that obscure the isolated impact of credit scores on missed payments (Kim, 2019).

Mitigation Strategies for Ethical and Analytical Risks

Robust Anonymization Techniques: All datasets will be stripped of personally identifiable information (PII), and encryption protocols will be implemented to ensure data security (Lidsky, 2024).

Regular Bias Assessments: Periodic fairness audits will be conducted to evaluate and correct any biases within models and datasets (Semrad, 2023).

Fairness-Optimized Modeling Techniques: Advanced fairness-aware algorithms will be integrated to adjust predictions and reduce potential discriminatory impacts (Caton & Haas, 2024).

Transparent Documentation: Methodologies, assumptions, and limitations will be explicitly documented, enabling stakeholders to interpret findings responsibly (Kibuacha, 2023).

Cross-Validation and Reliability Testing: Robust statistical techniques, including cross-validation, will be employed to confirm the reliability and generalizability of results (Chawla, 2024).

Conclusion

The examination of the correlation between credit scores and missed payments offers valuable insights into borrower behavior patterns. By incorporating ethical safeguards, ensuring data accuracy, and emphasizing transparency, the resulting analysis can support informed, equitable lending practices.

References

Asurity. (2024, March 5). *Credit algorithms, disparate impact, and the search for less discriminatory alternatives*. https://www.asurity.com/blogs/credit-algorithms-disparate-impact-and-the-search-for-less-discriminatory-alternatives/

Atlan. (2023, September 29). *The importance of data quality in financial services: 5 reasons!* https://atlan.com/importance-of-data-quality-in-financial-services/

Campisi, N. (2021, February 26). *From inherent racial bias to incorrect data—The problems with current credit scoring models*. Forbes. https://www.forbes.com/advisor/credit-cards/from-inherent-racial-bias-to-incorrect-data-the-problems-with-current-credit-scoring-models/

Caton, S., & Haas, C. (2024). Fairness in machine learning: A survey. *ACM Computing Surveys, 56*(7), Article 166, 1–38. https://doi.org/10.1145/3616865

Chawla, A. (2024, July 1). 5 cross validation techniques explained visually. *Daily Dose of Data Science*. https://blog.dailydoseofds.com/p/5-cross-validation-techniques-explained

Consumer Financial Protection Bureau. (2024, December 31). *Does my credit score affect my ability to get a mortgage loan or the mortgage rate I pay?* https://www.consumerfinance.gov/ask-cfpb/does-my-credit-score-affect-my-ability-to-get-a-mortgage-loan-or-the-mortgage-rate-i-pay-en-319/

Devane, H. (2022, May 12). *What are the top data anonymization techniques?* Immuta. https://www.immuta.com/blog/data-anonymization-techniques/

Donnelly, S. (2024, July 24). *16 of the best financial charts and graphs for data storytelling*. Finance Alliance. https://www.financealliance.io/financial-charts-and-graphs/

Fatima, N. (2024, June 26). *Understanding standardization in data preprocessing*. Medium. https://medium.com/%40noorfatimaafzalbutt/understanding-standardization-in-data-preprocessing-6c7760b5790a

JMP. (n.d.). *Correlation vs. causation*. https://www.jmp.com/en/statistics-knowledge-portal/what-is-correlation/correlation-vs-causation

Kim, J. H. (2019). Multicollinearity and misleading statistical results. *Korean Journal of Anesthesiology, 72*(6), 558–569. https://doi.org/10.4097/kja.19087

Kibuacha, F. (2023, August 29). The path to confident decisions: Data transparency in research reporting. *GeoPoll*. https://www.geopoll.com/blog/data-transparency/

Lidsky, O. (2024, November 22). Understanding data encryption: Types, algorithms and security. *Forbes*. https://www.forbes.com/councils/forbestechcouncil/2024/11/22/understanding-data-encryption-types-algorithms-and-security/

Sharma, S. (n.d.). *Correlation and regression analysis: Exploring relationships in data*. Data Science Salon. https://roundtable.datascience.salon/correlation-and-regression-analysis-exploring-relationships-in-data?

Semrad, M. (2023, January 5). The fast and effective way to audit ML for fairness. *KDnuggets*. https://www.kdnuggets.com/2023/01/fast-effective-way-audit-ml-fairness.html

Sereno. (2024, December 10). *Comparison of Pearson vs Spearman correlation coefficients*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/03/comparison-of-pearson-and-spearman-correlation-coefficients/

Thoropass. (n.d.). *GDPR vs CCPA: A thorough breakdown of data protection laws*. Thoropass. https://thoropass.com/blog/compliance/gdpr-vs-ccpa/