UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

Colegio de Posgrados

Optimizing Credit Risk Assessment through Data-Driven Scorecards

Proyecto de Titulación

Martín Emilio López Suárez

Israel Pineda, Ph.D.

Director de Trabajo de Titulación

Trabajo de titulación de posgrado presentado como requisito para la obtención del título de Magíster en Ciencia de Datos

Quito, 02 de diciembre de 2024

UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

COLEGIO DE POSGRADOS

HOJA DE APROBACIÓN DE TRABAJO DE TITULACIÓN

Optimizing Credit Risk Assessment through Data-Driven Scorecards

Martín Emilio López Suárez

Nombre del Director del Programa:	Felipe Grijalva
Título académico:	Ph.D. en Ingeniería Eléctrica
Director del programa de:	Ciencia de Datos
Nombre del Decano del colegio Académico:	Eduardo Alba
Título académico:	Doctor en Ciencias Matemáticas
Decano del Colegio:	Ciencias e Ingenierías
Nombre del Decano del Colegio de Posgrados:	Dario Niebieskikwiat
Título académico:	Doctor en Física

Quito, diciembre 2024

© DERECHOS DE AUTOR

Nombre del estudiante:	Martín Emilio López Súarez
Código de estudiante:	00338689
C.I.:	171668818-7
Lugar y fecha:	Quito, 02 de Diciembre de 2024.

ACLARACIÓN PARA PUBLICACIÓN

UNPUBLISHED DOCUMENT

RESUMEN

La puntuación de riesgo crediticio es fundamental para las instituciones financieras, ya que permite una gestión efectiva del riesgo y decisiones de préstamo informadas al evaluar la probabilidad de incumplimiento de préstamos, que plantea desafíos financieros y regulatorios significativos. Este estudio propone un marco basado en datos para la construcción de modelos de puntuación de riesgo crediticio adaptados a solicitantes comerciales, combinando técnicas estadísticas como la regresión logística con juicio experto para mejorar la precisión predictiva, la interpretabilidad y la aplicabilidad práctica. Aprovechando diversas fuentes de datos, incluidos datos de burós de crédito, solicitudes y desempeño, la metodología integra técnicas avanzadas como el Peso de la Evidencia (WoE) y el Valor de la Información (IV) para una construcción robusta de modelos de puntuación, con un fuerte enfoque en el cumplimiento normativo y la alineación empresarial. La evaluación del rendimiento mediante métricas como AUC-ROC, coeficiente de Gini y la estadística Kolmogorov-Smirnov resalta la eficacia del marco para predecir probabilidades de incumplimiento y apoyar la gestión proactiva del riesgo. Este enfoque híbrido combina rigor estadístico con conocimientos del mundo real, proporcionando a las instituciones financieras herramientas accionables para fortalecer la resiliencia y fomentar prácticas crediticias sostenibles.

Palabras clave: Riesgo crediticio - Scorecards - Regresión Logística - Criterio Experto -Cumplimiento Normativo

ABSTRACT

Credit risk scoring is pivotal for financial institutions, enabling effective risk management and informed lending decisions by assessing the likelihood of loan defaults, which pose significant financial and regulatory challenges. This study proposes a data-driven framework for constructing credit risk scorecards tailored to commercial applicants, combining statistical techniques like logistic regression with expert judgment to enhance predictive accuracy, interpretability, and practical applicability. Leveraging diverse data sources, including credit bureau, application, and performance data, the methodology integrates advanced techniques such as Weight of Evidence (WoE) and Information Value (IV) for robust scorecard construction, with a strong focus on regulatory compliance and business alignment. Performance evaluation using metrics like AUC-ROC, Gini coefficient, and Kolmogorov-Smirnov statistic highlights the framework's efficacy in predicting default probabilities and supporting proactive risk management. This hybrid approach bridges statistical rigor with real-world insights, equipping financial institutions with actionable tools to strengthen resilience and foster sustainable credit practices.

Key words: Credit Risk - Scorecards - Logistic Regression - Expert Judgment - Regulatory Compliance

TABLA DE CONTENIDO

LISTA DE TABLAS

LISTA DE FIGURAS

Optimizing Credit Risk Assessment through Data-Driven Scorecards

Martín López, Israel Pineda

***Abstract*—Credit risk scoring is pivotal for financial institutions, enabling effective risk management and informed lending decisions by assessing the likelihood of loan defaults, which pose significant financial and regulatory challenges. This study proposes a data-driven framework for constructing credit risk scorecards tailored to commercial applicants, combining statistical techniques like logistic regression with expert judgment to enhance predictive accuracy, interpretability, and practical applicability. Leveraging diverse data sources, including credit bureau, application, and performance data, the methodology integrates advanced techniques such as Weight of Evidence (WoE) and Information Value (IV) for robust scorecard construction, with a strong focus on regulatory compliance and business alignment. Performance evaluation using metrics like AUC-ROC, Gini coefficient, and Kolmogorov-Smirnov statistic highlights the framework's efficacy in predicting default probabilities and supporting proactive risk management. This hybrid approach bridges statistical rigor with real-world insights, equipping financial institutions with actionable tools to strengthen resilience and foster sustainable credit practices.**

I. Introduction

CREDIT scoring has emerged as a critical tool for financial institutions to evaluate the likelihood of loan defaults, enabling them to mitigate risks and optimize lending decisions. Loan defaults pose significant challenges for lenders, leading to financial losses, reduced liquidity, and increased regulatory scrutiny. The ability to predict and manage these risks is essential for maintaining the stability of financial institutions and ensuring sustainable credit practices. Credit scoring frameworks provide lenders with a systematic method to quantify an applicant's creditworthiness based on statistical probabilities derived from historical data. These frameworks are indispensable to address the complexities of modern lending environments [1].

The foundation of credit risk scoring lies in its ability to aggregate diverse borrower characteristics into a structured model, typically represented as a scorecard. Attributes such as income levels, past repayment behaviors, and credit bureau data are evaluated for their predictive power, contributing to an applicant's overall credit score. These scorecards are effective not only in evaluating new loan applications, but also in monitoring existing accounts, facilitating proactive risk management strategies such as credit limit adjustments and tailored repayment plans [2].

M. López and I. Pineda are with Universidad San Francisco de Quito USFQ

Technological advances and regulatory developments have further refined credit scoring methodologies. Frameworks such as Basel II have emphasized the importance of robust risk management practices, which require models that are transparent, repeatable, and compliant with regulatory standards [3]. Logistic regression remains a widely used technique due to its simplicity and interpretability, although there is increasing interest in advanced machine learning methods that can integrate alternative data sources for enhanced predictive accuracy [1]. However, the implementation of complex algorithms must be approached cautiously to preserve the transparency and reliability of credit risk assessments, especially considering lessons learned during the 2008 financial crisis, where the overreliance on opaque models contributed to systemic failures [4].

The prediction of loan defaults is particularly crucial in the current economic landscape, characterized by volatile markets and increased regulatory expectations. By accurately estimating the probability of default, lenders can establish appropriate credit terms, reduce exposure to high-risk applicants, and improve portfolio performance. This capability not only strengthens institutional resilience, but also ensures responsible lending practices that benefit the broader financial ecosystem [3].

This thesis focuses on developing a comprehensive, data-driven framework for constructing scorecards to evaluate the risk of loan defaults among commercial applicants. The proposed methodology combines the robustness of statistical models, specifically logistic regression, with the contextual insights provided by expert judgment. While the framework is predominantly data-driven, the incorporation of domain expertise ensures that the model reflects practical realities and captures nuances that may not be evident in the data alone. This hybrid approach aims to improve the predictive power, interpretability, and applicability of credit scoring models, enabling financial institutions to make well-informed, risk-sensitive lending decisions.

II. Prior works

Credit risk assessment has long relied on scoring models as an essential tool in financial decision-making. Among these, logistic regression remains a cornerstone due to its simplicity, transparency, and alignment with regulatory requirements [1]. These models use structured data, such as demographic details and credit history, to predict the probability of loan defaults, offering reliable and interpretable insights for risk management [2].

In recent years, advances in technology have expanded the scope of credit scoring methodologies. For example, Li et al. developed a dynamic network-based model that incorporates borrower interactions to improve default predictions [3]. Similarly, Zheng proposed a federated learning approach that allows multiple institutions to collaboratively develop models while maintaining data privacy and compliance with regulatory standards [4].

These developments underscore the growing emphasis on balancing predictive performance with interpretability and ethical considerations [5]. Despite these advances, challenges persist. Complex machine learning models, while powerful, can lack the transparency needed for effective communication with regulators and stakeholders. Furthermore, the inclusion of alternative data raises questions about privacy and potential biases, which require careful governance [2], [6].

This thesis aims to contribute to this evolving field by presenting a data-driven framework for constructing credit scorecards. The approach emphasizes methodological rigor using logistic regression, a well-established technique known for its robustness and interpretability. The framework is further refined by incorporating contextual insights from expert judgment, ensuring alignment with business realities, and improving the practical applicability of the scorecards. This integration complements the data-driven methodology, resulting in tools that are both predictive and actionable, while maintaining transparency and compliance with regulatory standards.

III. METHODOLOGY

The development of credit risk scorecards begins with Data Exploration, where the data set is analyzed for patterns, missing values, and outliers to ensure precision and readiness for modeling [1]. The next stage is to perform a default definition analysis; in this section, it is studied how to define the default that maximizes the bad behavior. This definition is crucial because it represents the supervision variable in which the model is trained and tested [2]. This is followed by Correlation Analysis, which evaluates the relationships between variables and the dependent variable. This is done to eliminate redundancy and ensure independence between predictors, improving the reliability of the model [3].

The next step is to conduct a multicharacteristic analysis that identifies the most predictive variables as a group of the dependent variable [4]. Once we have the variables that are going to be used in logistic regression, the following step is to build structural buckets that segment the default in an optimized way, which is crucial because these buckets in the probability of default are going to be the points assigned to each group, using metrics such as Information Value (IV) and Weight of Evidence (WOE), ensuring that the constructed buckets are both statistically significant and aligned with business objectives [1], [5]. Finally, in the Development of the Preliminary Scorecard, these variables are integrated into a logistic regression model, creating a predictive and operationally feasible scorecard framework [2].

Each of these steps forms the foundation of the scorecard methodology and will be explored in more detail in later sections, focusing on advanced techniques such as validation, scaling, and refinement to optimize performance [3].

A. Datasets

This project leverages four distinct datasets to develop and evaluate credit risk scorecards. The first dataset, the Application Dataset, encompasses all applications submitted for funding decisions. It includes critical variables such as Time in Business, Number of Guarantors, and other preliminary details essential for assessing initial creditworthiness. This data set provides the foundational information required to decide whether to approve or reject funding for an applicant.

The second dataset, the Performance Dataset, focuses on accounts that have been approved and funded (referred to as "booked" status). This back-office data set exclusively tracks the performance of funded clients, capturing key variables such as Days Past Due, Charge-Off Dates, and Amounts, and other performance indicators that reveal repayment behavior and financial health. Complementing these are the Asset Dataset, which details the funded assets (e.g., Model, Cost, New or Used status) and the Bureau Dataset, which offers external financial insights such as Days Past Due, Revolving Accounts, Inquiries, and Bankruptcies with other lenders. Together, these datasets provide a comprehensive view of client behavior from application through post-funding performance, ensuring a robust foundation for scorecard development.

B. *Data Cleaning and Preparations*

The development of scorecards is based on a structured data preparation framework that aligns technical methodologies with strategic business objectives. Defining clear problem statements and deliverables, such as identifying high-risk customers and optimizing loan approvals, ensures that data-driven insights address key business goals. Cross-disciplinary teams consisting of domain experts, data scientists, and IT specialists play a critical role in addressing diverse challenges such as data integration and regulatory compliance. For example, in highly regulated industries such as banking, privacy concerns and legal restrictions must be addressed during the project planning phase [1], [2]. Effective collaboration, supported by well-defined glossaries and iterative review processes, facilitates communication and mitigates risks associated with data inconsistencies or misaligned objectives [3].

Data cleaning is a cornerstone of credit scorecard development, ensuring that raw data is transformed into a reliable foundation for analysis. Addressing variable inconsistencies due to human error, which can mislead conclusions and bias the model, is essential to maintaining its future performance. To solve this issue, advanced ETL (Extract, Transform, Load) processes are required [4]. Numeric variables with erroneous values or textual variables with inconsistent labeling, such as varying representations of a loan type, must be standardized to eliminate inconsistencies. For example, credit limits reported in multiple currencies should be converted into a unified scale to allow meaningful comparisons and improve the reliability of subsequent analyzes [5]. In addition, geographical variables, such as state, must be validated, as they may be outside the underwriting area and affect the scope of the model.

Addressing missing values, by either removal or imputation, is another critical step in preventing biases that could compromise model outcomes. These cleaning processes not only reduce noise, but also ensure that the models reflect true patterns in customer behavior, thus enhancing predictive accuracy [6].

Once the data is clean, transformations are applied to prepare it for analysis and uncover actionable insights. Variables with wide ranges, such as annual incomes or outstanding balances, are normalized using methods such as min-max scaling or z-score transformations, ensuring that no single variable disproportionately influences model predictions [7]. Categorical attributes, such as loan types or payment statuses, are converted into numerical formats using techniques such as one-hot encoding or ordinal scaling, facilitating seamless integration into machine learning algorithms [8]. Feature engineering further enriches the data set by creating derived metrics, such as debt-to-income ratios, that provide deeper insights into customer behavior [9].

C. *Default Definition*

The default definition is a critical component in the development of credit risk scorecards, forming the basis for distinguishing between good and bad accounts. This definition not only shapes the predictive accuracy of the model, but also ensures alignment with regulatory frameworks, such as Basel II, which mandates a standardized approach to ensure consistency between financial institutions and jurisdictions [5], [12]. A robust default definition must balance operational relevance, regulatory compliance, and statistical rigor, ensuring that it captures meaningful indicators of credit risk while avoiding over or underclassification of accounts [8].

One of the core elements of a default definition is the delinquency threshold, which sets the number of days past due for an account to be considered in default. Under Basel II, this threshold is typically established 90 days past due, with the option for national regulators to extend it to 180 days for specific portfolios, such as retail loans secured by real estate [3], [9]. The threshold reflects a "point of no return" where delinquency becomes an irreversible indicator of credit risk, marking accounts unlikely to recover. Although this threshold is widely used, it must be carefully calibrated to account for differences between financial products, ensuring relevance to the specific characteristics of each portfolio [11].

Beyond delinquency, qualitative creditworthiness assessments further refine the definition of default. The "unlikely to pay" criterion captures accounts that, while not meeting the delinquency threshold, exhibit other high-risk indicators such as bankruptcy, distressed restructuring, or specific credit risk adjustments [7], [14]. This assessment relies on both internal data sources, such as loan exercise processes, and external inputs, including credit bureau reports or legal records [6]. Incorporating this criterion broadens the scope of the default definition, ensuring that it identifies potential risks that may not be immediately evident from payment behavior alone [10].

Materiality thresholds are another vital aspect of the default definition, determining the significance of arrears required to classify an account as defaulted. These thresholds may be expressed as a fixed monetary amount (for example, USD 2000) or as a percentage of total exposure (e.g. 2%), and ensure that defaults represent meaningful financial risks rather than minor discrepancies [2]. Materiality thresholds prevent overclassification of defaults due to negligible arrears, preserving the statistical integrity of default rates, particularly in portfolios with high volumes of low-value accounts [15].

The level of application for default definitions varies depending on the portfolio. In retail portfolios, defaults are generally assessed at the account level. However, events such as bankruptcy can elevate the default status to the customer level, identifying all associated accounts as defaulted to reflect the greater financial risk posed by the borrower [4]. In non-retail portfolios, default is inherently assessed at the customer level, with any default in an individual account resulting in a default classification for the entire customer relationship. This approach ensures that the definition captures the full extent of the credit risk associated with a borrower, particularly in cases involving significant exposures [16].

The operationalization of default definitions requires careful selection of performance and sample windows. The performance window, often set at 12 months as mandated by Basel II, represents the period during which the behavior of the account is monitored for evidence of default [1]. The sample window, on the other hand, specifies the historical time frame from which data are drawn for analysis [13]. Together, these windows ensure that the data set captures a wide range of borrower behaviors while mitigating biases caused by seasonality or economic cycles [17]. This structured approach ensures that the default definition remains relevant, consistent, and statistically sound [18].

To validate the robustness of the default definitions, analytical techniques such as roll rate analysis and delinquency comparisons are used. Roll-rate analysis tracks the progression of accounts to higher stages of delinquency, identifying thresholds where delinquency becomes irreversible [19]. Delinquency comparisons evaluate the differences between current and worst delinquency statuses, helping to confirm that thresholds align with operational realities and predictive objectives. These methods ensure that the default definitions are operationally practical and statistically valid [20].

In summary, the definition of default is a main point in the development of risk scorecards since this variable will be the one that supervises the process of learning. It is also the starting point for the feature selection process, encapsulating the default behavior that is desired to be predicted [12].

D. Variable Selection

In the development of credit risk scorecards, the selection of variables is an important step following the preparation of a clean and validated data set. This phase involves filtering and evaluating potential predictors to identify those most relevant and robust for modeling purposes. A rigorous selection process ensures that the resulting model is both statistically sound and operationally practical while adhering to regulatory and business constraints [1], [2].

The variable selection process begins with the elimination of variables that lack predictive power or are not suitable for inclusion. This includes variables such as names, dates, or phone numbers, which inherently do not contribute to risk prediction. Variables restricted by regulatory requirements, such as those that indicate age or sex, are also excluded at this stage [3]. In addition, variables that are specific to unique opportunities and lack generalizability are removed from consideration. This initial filtering ensures that the data set is streamlined for statistical evaluation and aligned with ethical and legal considerations [4].

The initial step in the variable selection process involves analyzing the fill rate and variability of the characteristics. For the filling rate, variables are retained only if they account for at least 20 percent of the total population. This threshold ensures that the variable contributes meaningful information to the final model, positively impacting its overall performance [5]. Feature variability focuses on categorical variables. These variables are expected to exhibit diverse classes. Variables with only one class provide limited information, similar to variables with low filling rates. To assess variability, the standard deviation is analyzed. If the standard deviation is close to zero, the variable is considered to have low variability and is excluded from the model. Instead, variables with a standard deviation closer to one are preferred, as they indicate sufficient variability to enhance the predictive performance of the model [6].

Once the initial filtering is complete, statistical techniques are employed to evaluate the predictive strength of each remaining variable individually. Transformations are applied to numerical variables to uncover non-linear relationships or interactions with the default indicator. These transformations may include polynomial terms, logarithmic adjustments, and indicators of missing data. For categorical variables, dummy variables are generated to examine the predictive contributions of individual categories. Each variable is then analyzed in isolation using logistic regression, with predictive power assessed through metrics such as the GINI coefficient [7], [8].

Variables demonstrating strong predictive power and logical relationships with the target variable are retained for further consideration [9]. To ensure that the selected variables are not redundant, correlation analysis is performed. This involves calculating a correlation matrix to identify highly correlated variables, which may introduce multicollinearity and reduce the model's interpretability. When correlated variables are identified, the variable with the greater predictive power is typically retained [10]. However, business relevance and operational feasibility also influence these decisions, ensuring that the final selection supports both statistical rigor and practical application [11].

The process of selecting variables goes beyond statistical metrics, incorporating business reasoning to ensure that the model is aligned with operational objectives. Variables prone to manipulation or subjective interpretation, such as self-reported income, are excluded unless robust validation mechanisms are in place. Similarly, data elements with inconsistent or optional collection practices are avoided to ensure reliability and completeness [12]. Variables are also assessed for their interpretability, as complex or ambiguous characteristics can undermine the usability and effectiveness of the model [13]. Characteristics that are highly dependent on human judgment are considered only when supported by expert evaluations or extensive credit risk experience [14].

Business justification plays a crucial role in creating ratios or derived variables. For example, comparisons between short- and long-term credit utilization can provide valuable information on borrower behavior, making them strong predictors of risk [15]. Such variables are included only when their predictive value is supported by sound business logic and their construction is relevant to the specific context of the scorecard [16].

In addition, industry trends and changes in the competitive environment are considered to ensure that the selected variables remain relevant and predictive over time. This forward-looking approach ensures that the scorecard adapts to evolving market dynamics without compromising its effectiveness [17].

The final data set is carefully curated to balance statistical robustness with business relevance and compliance with regulatory standards. This ensures that the variables selected for the modeling are not only predictive, but also operationally viable and aligned with the institution's risk management framework [18]. By refining the data set to include only the most relevant and reliable variables, the variable selection process establishes a strong foundation for the development of a credit risk scorecard that meets both analytical and practical requirements [19], [20].

E. Variable Binning

Optimal binning is a sophisticated data preprocessing technique essential for credit scoring, designed to convert continuous or ordinal variables into discrete bins. Unlike arbitrary or fixed-interval binning methods, optimal binning uses statistical algorithms to determine bin boundaries that maximize the predictive power of a variable. This approach ensures that bins effectively differentiate between good and bad credit outcomes, providing better predictions of credit risk and default probabilities. By tailoring bin edges to the characteristics of the data, optimal binning improves both the interpretability of credit risk models and their compliance with regulatory standards [1], [7].

Several methodologies underpin optimal binning in credit scoring, each tailored to specific data characteristics and modeling goals. Supervised binning determines the boundaries of the bin based on the relationship between predictor variables and credit default results, often using decision trees or Information Value (IV) analysis to identify the most informative splits [3]. Chi-square binning focuses on statistical significance, merging adjacent bins to ensure meaningful distinctions in default risk [6]. Additionally, monotonic binning maintains a consistent directional relationship between predictors (e.g. income or credit utilization) and default likelihood, aligning with credit risk management practices. These approaches ensure that the binning not only captures relevant patterns in the data but also aligns with business logic and risk assessment requirements [9].

The validation of optimal binning is based on several key metrics to ensure its effectiveness and statistical soundness. The Information Value (IV) measures the discriminatory power of a binned variable, with values greater than 0.3 indicating strong predictive performance. IV combines the proportion of good and bad outcomes in each bin with Weight of Evidence (WoE), which quantifies the strength of the relationship between predictor bins and the target variable [10]. Statistical tests, such as Pearson's chi-square test, further validate the differences between bins, ensuring that they reflect genuine differences in default risk [5]. Homogeneity metrics, such as the Herfindahl-Hirschman Index (HHI), evaluate the evenness of the data distribution across bins, with lower HHI values indicating balanced and robust binning. These validations confirm that optimal binning contributes significantly to model precision and reliability [8].

The Weight of Evidence (WoE) is integral to optimal binning, providing a standardized numerical representation of the predictive strength of each bin. WoE is computed as the logarithm of the ratio between the proportions of good and bad outcomes within a bin, allowing credit risk models to interpret the relative risk associated with different bins [4]. By transforming variables into WoE values, logistic regression models can directly compare predictors on the same scale, improving the interpretability of the model and reducing bias. WoE also supports advanced modeling techniques, such as Marginal Stepwise Logistic Regression, by highlighting the most significant predictors and ensuring a consistent relationship between variables and outcomes [12].

Optimal binning offers substantial advantages in credit risk modeling, including improved predictive accuracy, enhanced interpretability, and effective handling of outliers and missing values. By discretizing continuous variables such as income or credit utilization into strategically defined bins, the relationship between predictors and the probability of credit default is clarified. Furthermore, the use of metrics such as IV and HHI ensures that binning strategies contribute meaningfully to the model while maintaining statistical integrity [13]. These transformations are not only vital for building transparent and compliant credit scoring models but also for effectively communicating risk assessments to stakeholders and regulatory bodies. Optimal binning remains a cornerstone of modern credit risk management, allowing financial institutions to make data-driven decisions with confidence [11].

IV. Scorecard Construction

A. Dataset Splitting and Population Stability

The first step in building a scorecard is to divide the data set into training and validation sets to ensure robust model development and evaluation. Typically, 70–80 percent of the sample is used for model training (development set), while the remaining 20–30 percent serves as the validation set [3], [7]. This split ensures that the model is trained in one data subset and tested in another, reducing the risk of overfitting and allowing an unbiased evaluation of its predictive performance [10]. In cases with small datasets, the entire sample can be used for training, followed by validation in randomly selected subsets of 50–80 percent [11]. Training and validation sets should maintain the same distribution of key variables, such as goods and bads, to ensure the stability and reliability of the scorecard to predict outcomes [13].

The Population Stability Index (PSI) is a key metric to assess whether the distributions of variables in the training and validation datasets are consistent. PSI measures changes in the distribution of a variable, such as a credit score, between the baseline (training) and current (validation) datasets [8]. PSI is calculated by dividing the range of the variable into bins, comparing the proportion of observations in each bin between the datasets, and summing the differences using the formula.

$$PSI = \sum \left((p_{\text{current}} - p_{\text{baseline}}) \cdot \ln \frac{p_{\text{current}}}{p_{\text{baseline}}} \right)$$

A PSI below 0.1 indicates population stability, while values between 0.1 and 0.25 suggest moderate changes that may require investigation. Values above 0.25 indicate significant instability, which could necessitate a model recalibration. Monitoring PSI ensures that the training and validation data sets are comparable and helps detect changes that could affect the performance of the model [2], [5].

Logistic regression is a fundamental statistical method used for binary classification tasks, where the goal is to predict the probability of an outcome that belongs to one of two classes, such as "default" versus "non-default" in credit scoring. Unlike linear regression, which models a continuous dependent variable, logistic regression estimates the relationship between predictor variables $(X_1, X_2, \ldots, X_n)$ and the logarithmic odds of the binary dependent variable (Y)[7], [11] The logistic regression model begins with the logit function, which expresses the relationship between predictors and log-odds of the outcome [8]:

$$\text{Logit}(P) = \ln\left(\frac{P}{1-P}\right) = \beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_nX_n$$

Here:

- P is the probability of a positive outcome (e.g., the probability of default).
- $\ln\left(\frac{P}{1-P}\right)$ is the natural logarithm of the odds ratio, which transforms the bounded probability space $[0,1]$ into the real number line $((-\infty,+\infty))$.
- β_0 is the intercept, representing the logarithmic odds of the outcome when all predictors are zero.
- $\beta_1,\beta_2,\ldots,\beta_n$ are coefficients that quantify the impact of each predictor $X_1,X_2,\ldots,X_n$ on the logarithmic odds of the result.

B. *From Logit to Probability*

To convert logit ($\text{Logit}(P)$) back to a probability, the following logistic function is applied:

$$P = \frac{1}{1+e^{-(\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_nX_n)}}$$

This transformation ensures that the predicted probabilities P are restricted within the interval $[0,1]$.

Logistic regression makes several key assumptions:

1) **Linear Relationship:** The log-odds of the dependent variable are linearly related to the predictor variables.
2) **Independence of Errors:** Observations are independent, and errors are not auto-correlated.
3) **Multicollinearity:** Predictor variables should not be highly correlated with each other.
4) **Large Sample Size:** Sufficient data is needed to ensure the robustness of the maximum likelihood estimation (MLE) used to compute the coefficients.

C. *Estimation of Coefficients*

The coefficients $\beta_0,\beta_1,\ldots,\beta_n$ are estimated using the maximum likelihood estimate (MLE). This method finds the coefficient values that maximize the likelihood of observing the given data:

$$L(\beta) = \prod_{i=1}^{n} P_i^{y_i}(1-P_i)^{1-y_i}$$

Where P_i is the predicted probability of observation i, and y_i is the actual result (0 or 1). Taking the natural logarithm of $L(\beta)$, the log-likelihood function is maximized iteratively to obtain the estimates.

In the context of scorecard development, logistic regression coefficients (β) play a critical role in defining the relationship between predictor variables and outcome. A negative coefficient ($\beta < 0$) signifies an inverse relationship between the predictor variable and the likelihood of the positive outcome, such as the default in credit scoring [1]. This means that as the predictor value increases, the probability of default decreases. For example, if the predictor variable is "years with current employer," a negative coefficient indicates that a longer tenure with an employer reduces the likelihood of default, reflecting greater job stability and consequently lower credit risk [3]. This aligns with the intuitive and widely accepted principles of risk assessment, ensuring that the outcome of the scores remains logical and interpretable [6].

The presence of negative coefficients also serves as a critical indicator of data consistency and alignment with domain knowledge. When coefficients contradict expectations, such as a positive coefficient for "income," which would suggest that higher income increases credit risk, it could signal issues such as multicollinearity among predictors or inaccuracies in the data [8]. These cases require careful examination and potential adjustments to the model or data preprocessing to ensure that the relationships being captured are genuine and not artifacts caused by statistical noise or poor data quality [9].

Negative coefficients further influence the allocation of points within a scorecard. Credit scorecards are designed so that higher scores correspond to lower credit risk. A variable with a negative coefficient contributes to this design by assigning higher scores to higher attribute values, thereby lowering the risk assessed [11]. For example, a higher value of "years with current employer" translates into more points, reinforcing the trend that greater job stability is associated with a reduced likelihood of default. This ensures that the scorecard provides clear and logical risk classifications between different applicants, allowing better decision making [13].

In general, negative coefficients are essential to maintain the interpretability and reliability of a scorecard. They validate the logical trends expected in credit scoring, help detect potential model inconsistencies, and ensure that the allocation of points reflects sound risk assessment practices [15]. Their role is fundamental in aligning statistical results with business intuition and regulatory requirements, which makes them indispensable in the development of robust credit scoring systems [16].

D. Advantages of Logistic Regression

1) **Interpretability:** Each coefficient represents the change in the logarithmic odds of the outcome for a change of one unit in the predictor, ensuring that the other variables are constant.
2) **Probability Outputs:** Logistic regression directly predicts probabilities, making it ideal for risk assessment tasks such as credit scoring.
3) **Flexibility:** Can handle binary, ordinal, and multinomial classification problems (with extensions).

E. Applications in Credit Scoring

In credit risk modeling, logistic regression is the cornerstone of constructing scorecards. Predictor variables such as income, credit utilization, or payment history are transformed using **Weight of Evidence (WoE)** to ensure linearity and interpretability. The model estimates the probabilities of default (P), which are used to classify applicants by risk and calculate scores based on defined odds ratios.

F. Limitations

Although logistic regression is robust, it assumes linearity in logarithmic odds, which may not capture complex relationships in the data. For this reason, interactions between variables or transformations like polynomial terms are sometimes included to improve the fit.

G. Converting Logistic Regression to Scorecard Points

In credit scorecard development, the logistic regression outputs, which describe the relationship between predictor variables and the log-odds of the outcome, are transformed into scorecard points. This transformation allows raw statistical results to be converted into an interpretable scoring system [1], [7]. The results of logistic regression are first used to calculate the Weight of Evidence (WoE) for the groups of a variable, and then these WoE values are assigned to the points using the relationship [9], [12]:

$$\text{Points} = \text{Factor} \cdot \text{WoE} + \text{Offset}$$

The **Factor** determines the scaling of points and is derived from the "Points to Double Odds" (PDO), which specifies the number of points associated with a doubling or halving of the odds. If the PDO is set to 20, the Factor is calculated as:

$$\text{Factor} = \frac{\text{PDO}}{\ln(2)} \approx 28.85$$

The **Offset** is a baseline adjustment that ensures the scores align with business goals, such as setting a specific score for a given odds ratio.

In the context of the variable `Equifax FICO Score`, which represents an applicant's FICO score, the scorecard assigns specific points to predefined buckets based on the range of FICO scores. These points are precalculated based on the results of the logistic regression and reflect the relative risk associated with each bucket. The buckets, ranges, and their assigned points are as follows:

- **Bucket 0**: FICO score between 0 and 598, assigned **0 points**.
- **Bucket 1**: FICO score between 598 and 677, assigned **22 points**.
- **Bucket 2**: FICO score between 677 and 693, assigned **37 points**.
- **Bucket 3**: FICO score between 693 and 758, assigned **44 points**.
- **Bucket 4**: FICO score greater than 758, assigned **71 points**.

For example, an applicant with a FICO score of 700 would fall into Bucket 3 (693–758) and receive 44 points. This reflects their lower likelihood of default compared to applicants in lower buckets. In contrast, an applicant with a FICO score of 550, which falls into bucket 0 (0-598), would receive 0 points, indicating a higher level of credit risk.

The logic of the scorecard ensures interpretability and alignment with intuitive risk assessments. Higher FICO scores correspond to higher points, which indicate lower risk. This scoring approach allows credit decision makers to easily rank and evaluate applicants based on their risk profile. In addition, precalculated points simplify operational processes by directly linking the scorecard to business decisions without requiring further mathematical transformations.

Using a well-structured bucket system and predetermined points, the scorecard provides an actionable framework that translates the statistical rigor of logistic regression into a practical tool for assessing credit risk. This system ensures consistency, transparency, and usability in real-world credit scoring applications.

Table I: Population Stability Index

Feature	PSI
Equifax FICO Score	0.0031
Paynet Inquiries 6 Month	0.0045
Worst Payment Status	0.0075
Paynet Average Days Past Due	0
Number of Bankruptcy's	0.0005
Number of Repossessions	0.0052
Total Past Due Amount	0.0065
SIC	0.0025
Equip Code	0.0018
State	0.0024

V. Results

A. Population Stability Index

B. Bucketization

For the numerical variables, the optimum cutoffs were calculated by maximizing the IV and WoE as shown in the following tables.

Equifax FICO Score:

Table II: Equifax FICO Score Binning

BIN	Population %	Event Rate (%)
(0 , 490)	2.55	22.77
[490 , 575)	19.99	15.74
[575 , 679)	49.35	11.22
[689 , 720)	19.56	7.69
[720 , 800]	8.55	3.12

The table highlights the relationship between credit scores, population distribution, and event rates, which is critical to the development of credit score cards. Shows an inverse correlation between FICO scores and event rates, where higher scores correspond to lower financial risk. Most of the population (49.35%) falls in the 575–679 range, indicating a key segment for risk modeling, as it balances moderate event rates (11.22%) with a significant population share. Low scores (0–490) show the highest event rate (22.77%), emphasizing the need for stricter credit policies, while high scores (720–800) have minimal risk (3.12%), making them ideal for less stringent lending criteria. These data serve as a foundation for segmenting populations, calibrating risk factors, and optimizing decision rules in credit scorecard models.

Paynet Inquiries 6 Month:

Table III: Paynet Inquiries 6 Month Binning

BIN	Population %	Event Rate (%)
(inf , 7)	12.25	17.25
[7 , 3)	55.25	13.01
[3 , 0)	13.50	8.28
[0 , -inf)	19.00	4.63

This table analyzes the relationship between the number of Paynet inquiries in the last six months, the population distribution, and the event rates, providing information on the behavior of the risk of the borrower. Most of the population (55.25%) falls into the (7, 3] bin with an event rate of 13. 01%, representing a moderate risk segment. The group (inf, 7], with 12. 25% of the population, exhibits the highest event rate of 17.25%, indicating that a higher number of recent inquiries correlates with greater risk. On the other hand, the bin [0, -inf), which comprises 19. 00% of the population, shows the lowest event rate of 4.63%, demonstrating that fewer or no recent inquiries are associated with reduced risk. The segment (3, 0], representing 13. 50% of the population, has an event rate of 8.28%, reflecting an intermediate level of risk. These insights reinforce the relevance of recent inquiry activity in credit scorecard development, as a higher frequency of inquiry can signal increased financial stress or credit-seeking behavior.

Paynet Average Days Past Due:

Table IV: Paynet Average Days Past Due Binning

BIN	Population %	Event Rate (%)
(inf , 13)	37.66	15.64
[13 , 4)	18.75	12.21
[4 , -inf)	43.59	5.31

The table shows the distribution of the population and event rates based on Paynet average days past due, which reflects the borrower's payment behavior. The largest segment, representing 43.59% of the population, falls in the 4 to infinity bin, with the lowest event rate of 5.31%, indicating strong credit performance and low risk. In contrast, the (-infinity, 13) group, which accounts for 37. 66% of the population, has the highest event rate of 15. 64%, which means elevated risk. The bin [13, 4] represents an intermediate segment, with 18.75% of the population and an event rate of 12.21%. This analysis highlights the importance of payment timeliness in predicting credit risk, with fewer days past due correlated to better credit outcomes.

Number of Bankruptcy's Max:

Table V: Number of Bankruptcy's Binning

BIN	Population %	Event Rate (%)
(inf , 3)	8.35	23.25
[3 , 1)	15.52	17.44
[1 , 0)	35.47	13.15
[0 , -inf)	40.66	6.56

This table analyzes the relationship between the maximum number of bankruptcies, population distribution, and event rates, highlighting the impact of previous bankruptcies on financial risk. The bin (inf, 3) representing 8. 35% of the population shows the highest event rate (23. 25%), indicating a significant risk associated with individuals with multiple bankruptcies. The segment (3, 1], comprising 15. 52% of the population, also exhibits an elevated risk with an event rate of 17.44%. Conversely, the [0, -inf) group, which makes up the largest share of the population (40.66%), has the lowest event rate (6.56%), reflecting low-risk behavior in individuals with no prior bankruptcies. The category [1, 0] accounts for 35. 47% of the population, with a moderate event rate of 13.15%. This analysis demonstrates that the number of past bankruptcies is a strong predictor of future financial risk and emphasizes its critical role in credit risk segmentation and scorecard development.

Number of Repossessions:

Table VI: Number of Repossessions Binning

BIN	Population %	Event Rate (%)
(inf , 2)	7.53	20.14
[2 , 1)	18.25	15.76
[1 , 0)	32.35	10.91
[0 , -inf)	41.87	4.98

This table evaluates the relationship between the number of repossessions, the population distribution, and the event rates, offering insight into how past repossessions correlate with financial risk. The bin (inf, 2), representing 7. 53% of the population, exhibits the highest event rate of 20. 14%, indicating a significant risk among individuals with multiple repossessions. The bin [2, 1], covering 18. 25% of the population, follows with a high event rate of 15.76%. In contrast, the [0, -inf) group, which makes up the largest proportion of the population (41.87%), has the lowest event rate of 4.98%, highlighting the low risk associated with no history of repossessions. The category [1, 0], which represents 32. 35% of the population, reflects moderate risk with an event rate of 10.91%. This data underscores the importance of the history of repossession as a key risk factor in the development of credit scorecards, with a higher number of repossessions correlated with an increased financial risk.

Total Past Due Amount:

Table VII: Total Past Due Amount Binning

BIN	Population %	Event Rate (%)
(inf , 3,650)	5.00	27.47
[3,650 , 1,245)	30.36	18.35
[1,245 , 0)	64.64	7.74

This table analyzes the relationship between the total amount due in the past, the distribution of the population, and the events, highlighting its relevance to the assessment of financial risk. The (inf, 3,650) bin, comprising 5.00% of the population, shows the highest event rate of 27.47%, indicating significant risk associated with high past-due amounts. The [3,650, 1,245] bin, representing 30.36% of the population, has a moderate event rate of 18.35%, reflecting a considerable level of financial stress. The majority of the population (64.64%) falls within the [1,245, 0] bin, with the lowest event rate of 7.74%, highlighting minimal risk among individuals with lower past-due amounts. These findings reinforce the critical role of past-due amounts in predicting financial risk, with higher amounts correlating strongly with elevated risk levels, making it a key variable for segmentation and scorecard calibration in credit risk models.

C. Performance Metrics

With the available dataset, the process involves building, optimizing and evaluating a predictive model for binary classification, specifically tailored for applications such as credit risk analysis. The initial step focuses on preparing the data set by scaling features to ensure consistency between variables, which is critical for algorithms sensitive to input magnitudes. This step ensures that all variables contribute equally to the model's learning process, improving the reliability of predictions.

To address potential imbalances in the dataset, class balancing techniques are applied, which is particularly important in scenarios where positive outcomes, such as default events, are relatively rare. The model is trained using a robust cross-validation strategy, such as repeated k-fold cross-validation, to validate its performance across multiple data splits, thereby enhancing generalizability and stability. The process includes systematic hyperparameter tuning to find the optimal configurations of the logistic regression model, focusing on regularization techniques to balance model complexity and predictive power.

The performance of the model is evaluated using key metrics: AUC-ROC, Gini coefficient and

Kolmogorov-Smirnov (KS) statistic. AUC-ROC measures the model's ability to discriminate between positive and negative outcomes, the Gini coefficient quantifies inequality in predictions derived from the AUC, and the KS statistic identifies the maximum separation between true positive and false positive rates. These metrics collectively provide a robust assessment of the model's ability to accurately predict outcomes and its potential for deployment in high-stakes decision-making environments such as credit risk management. As a result of this process the following metrics were obtained:

Table VIII: Model Performance Metrics

KPI	**AUC%**	**GINI%**	**KS%**
Train	0.7474	49.49	36.37
Test	0.7125	42.51	32.83

To better understand these results and further develop some interpretations, we also studied the following learning curves for each metric.

AUC-ROC

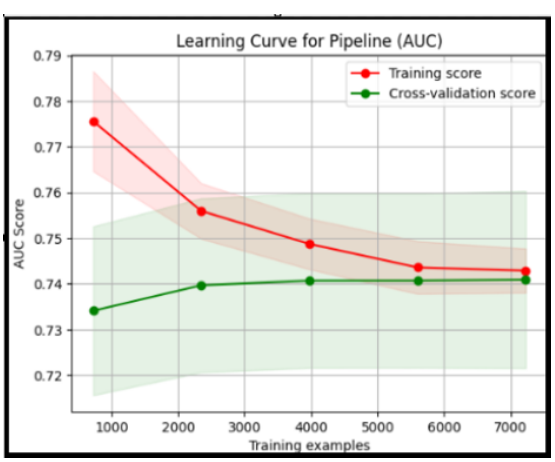

Figure 1: AUC-ROC Learning Curve

GINI

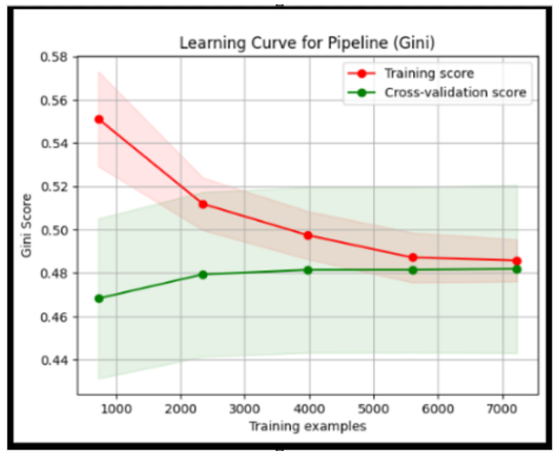

Figure 2: GINI Learning Curve

Kolmogorov-Smirnov (KS) statistic

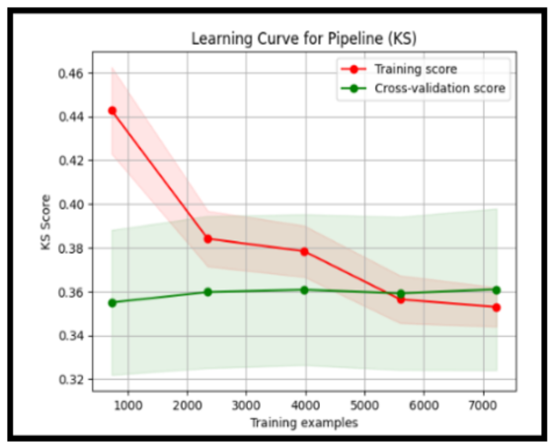

Figure 3: Kolmogorov-Smirnov (KS) Statistic Learning Curve

The performance evaluation of the logistic regression model indicates a moderate level of predictive power based on the results of the training and testing data sets. The model achieves an AUC-ROC of 0.7474 on the training set and 0.7125 on the test set, with corresponding Gini coefficients of 49.49% and 42.51%, and Kolmogorov-Smirnov (KS) statistics of 36.37% and 32.83%, respectively. The slight decline in performance on the test set highlights some generalization limitations, although the gap remains within acceptable bounds, indicating that the model is not overfitted.

The learning curves for AUC, Gini, and KS provide additional insights into the behavior of the model. Training scores start high, but decrease as the number of training examples increases, stabilizing at a reasonable level. Meanwhile, cross-validation scores gradually improve with more training data, reducing the gap between training and validation performance. This pattern suggests that the model benefits from additional data and is moving toward convergence, but it has not yet reached a point where further training data would provide diminishing returns.

The Gini coefficient and the KS statistical learning curves exhibit similar trends, with initial high training scores and increasing cross-validation scores as more training data are introduced. Both metrics stabilize at levels that align with the overall model performance, reinforcing the conclusion that the model generalizes reasonably well while maintaining sufficient discrimination power.

In summary, the logistic regression model shows solid performance, with acceptable discrimination and separation capabilities. However, the relatively modest metrics, particularly on the test set, suggest that while the model captures meaningful patterns in the data, there is room for improvement. Enhancing the model might involve exploring alternative feature engineering techniques, testing additional algorithms, or further optimizing hyperparameters. The learning curve trends confirm that adding more data could further improve generalization and overall model robustness.

VI. Conclusion

- The project's data cleaning and preprocessing steps were effective in ensuring model reliability and consistency. Standardization of features and the adjustment of class imbalances contributed to a stable learning process and interpretable outcomes. These efforts ensured that the model could effectively use the data set, minimizing potential biases or noise. The results validate the importance of these steps in building a solid foundation for predictive modeling.

- The careful selection of variables, including critical credit behavior metrics, and the use of bucketization strategies effectively captured meaningful patterns in the data. Bucketization helped simplify complex variables, making them easier to interpret and use in the model. This approach added to the model's explainability, which is crucial in credit risk assessment. Although there is a balance between granularity and interpretability, the bucketization applied here aligns well with the project objectives.

- The logistic regression model demonstrated robust performance, with reasonable AUC-ROC, Gini, and KS values for both training and test sets. The model effectively captured the underlying relationships in the data and generalized well to unseen examples. The learning curves further highlight the model's stability and potential for improvement with additional data, confirming the strength of the current approach.

- While the model performed well, improvements could focus on exploring alternative models, such as gradient boosting or ensemble methods, to enhance predictive power. Refinement of bucketization through data-driven techniques, such as optimal binning, could improve the representation of the variables. Adding advanced feature engineering techniques and expanding the data set could further enhance the discriminatory ability of the model. These steps, combined with continuous monitoring and updates, would strengthen the long-term performance and adaptability of the model in dynamic credit environments.

References

[1] N. Siddiqi, *Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring*, Hoboken, NJ: Wiley, 2012.

[2] Basel Committee on Banking Supervision, *International Convergence of Capital Measurement and Capital Standards: A Revised Framework*, Bank for International Settlements, 2006.

[3] R. Anderson, *The Credit Scoring Toolkit: Theory and Practice for Retail Credit Risk Management and Decision Automation*, Oxford, UK: Oxford University Press, 2007.

[4] E. Mays, *Credit Scoring for Risk Managers: The Handbook for Lenders*, New York, NY, USA: McGraw-Hill, 2004.

[5] L. C. Thomas, D. B. Edelman, and J. N. Crook, *Credit Scoring and Its Applications*, Philadelphia, PA: SIAM, 2002.

[6] D. Bensoussan, *Credit Risk Management: Basic Concepts*, Chichester, UK: Wiley, 2015.

[7] S. Finlay, *Credit Scoring, Response Modeling, and Insurance Rating: A Practical Guide to Forecasting Consumer Behavior*, Palgrave Macmillan, 2010.

[8] J. Li, J. Zhang, and X. Wang, "Dynamic network-based credit scoring model," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 50, no. 2, pp. 73–85, 2020.

[9] C. Zheng, Y. Xu, and Q. Wu, "Federated learning for credit risk assessment: A vertical approach," in *Proceedings of the 14th ACM Conference on Recommender Systems*, New York, NY: ACM, 2020, pp. 345–353.

[10] M. Crook, D. Edelman, and L. Thomas, "Recent developments in consumer credit risk assessment," *European Journal of Operational Research*, vol. 183, no. 3, pp. 1447–1465, Dec. 2007.

[11] S. Thomas, *Consumer Credit Models: Pricing, Profit, and Portfolios*, Oxford, UK: Oxford University Press, 2009.

[12] Project Management Institute, *A Guide to the Project Management Body of Knowledge (PMBOK Guides)*, 5th ed., Project Management Institute, 2013.

[13] T. Antony, *Design of Experiments for Engineers and Scientists*, Butterworth-Heinemann, Oxford, 2003.

[14] D. Montgomery, *Design and Analysis of Experiments*, 8th ed., Hoboken, NJ: Wiley, 2012.

[15] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed., Morgan Kaufmann Publishers, 2012.