Credit Risk Management
Using
ScoreCards

Prepared by: Joan Ngugi

Reference Book

- **Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring** by Naeem Siddiqi (Author)

Strathmore
UNIVERSITY

Credit Scorecard Development

Introduction

Unlike most data science use cases, the development of scorecards has its unique considerations. It requires a holistic iterative process that integrates domain and technical expertise, regulatory compliance, and ethical considerations to build reliable, and transparent models that support responsible lending practices.

As a result, most companies will adhere to the following guidelines when building scorecards

- ✓ The decisions made based on credit scoring should be explainable, transparent, and ethical. Eg not use features such as religion or race in the decision process.
- ✓ Data privacy and consumer protection is adhered to in the decision process.
- ✓ Adherence to a governance framework through regulatory requirements and industry best practices.

Data Sources

Scorecards can be built from one rich data source or a variety of two or more reliable data sources such as payment and credit history behaviour, demographic data, etc.

Data Exploration & Cleaning

Thorough data exploration is essential to evaluate data quality, completeness, and imbalance, as well as to identify trends crucial for informing model and scorecard development. Successful data cleaning relies on a deep understanding of the business context to rectify anomalies and inconsistencies accurately.

Feature Generation

Feature generation is done to enhance the predictive power of the model. This is essentially the same process as most data science use cases.

Definition of Good and Bad Loans

Defining good and bad loans is a critical step in scorecard development, serving as the target variable for model evaluation and training. In real-world scenarios, loans may have varied statuses like current, defaulted, paid, in grace period etc. Collaborating with domain experts is crucial to accurately categorize loans as good or bad based on their status.

Sample Window

The next crucial step is identifying the sample window. This time frame defines when you’ll gather your data sample for constructing the scorecard. It needs to encompass a substantial number of both good and bad loans. Additionally, it should strike a balance between not being too outdated or too recent.

Binning

Binning is the process of grouping a continuous variable into distinct bins or categories. **A primary goal in scorecard development is to maximize information summarization**. This is achieved by binning continuous variables and subsequently constructing a model that removes weak variables or those not aligned with sound business logic. Variables that are already categorical or binary do not require binning because they naturally represent distinct groups or states. Binning categorical variables may lead to loss of information or introduce unnecessary complexity into the modeling process.

Some common binning techniques include ChiMerge Algorithm, Decision Tree-Based Binning, Equal-Width binning, Equal-Size binning, Manual Binning, etc.

It is usually a very detailed process that will need to be iterated. Regardless of the binning technique used, a good bin should have the following characteristics:

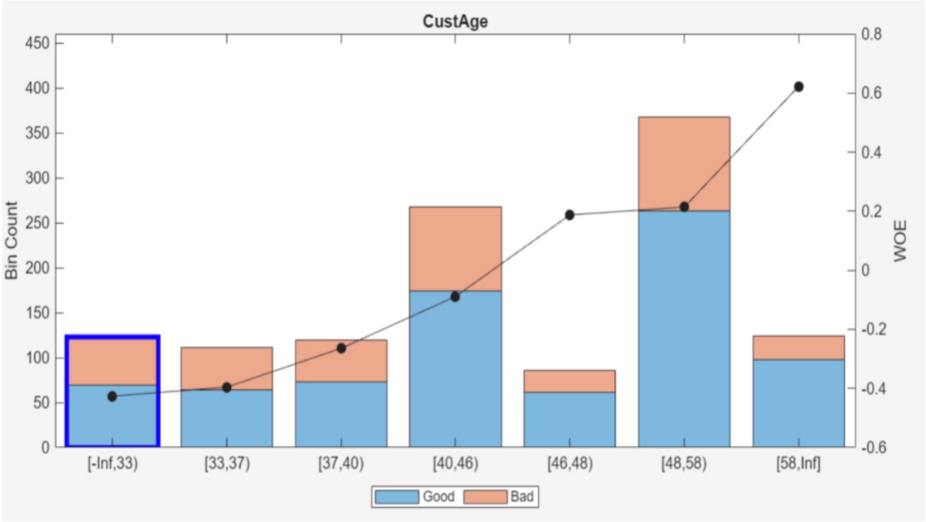
□ **Each bin should contain at least 5% of observations i.e. a minimum of 5 percent in each bin/bucket.**

□ **Missing values are binned separately.**

□ **There are no groups with 0 counts for good or bad.**

□ **The Weight of Evidence (WoE) for non-missing values adheres to a coherent distribution, progressing consistently from negative to positive values without any reversals. This alignment with business logic is confirmed by the pattern observed.**

□ **The disparity between the bad rate and Weight of Evidence (WoE) across different groups is significant. This indicates that the grouping has been structured to optimize the distinction between good and bad outcomes, enhancing differentiation from one group to the next.**



Binning converts continuous variables into discrete intervals (or bins), and it serves several important purposes:

- **Enhancing Interpretability:**

  By grouping continuous values into categories, you can more easily interpret the relationship between a variable and the target. For example, you might see that risk increases dramatically once a variable exceeds a certain threshold.

- **Reducing Noise and Handling Outliers:**

  Binning can smooth out small fluctuations and reduce the impact of extreme values, which might otherwise distort the model's understanding of the relationship.

- **Facilitating Calculation of WOE and IV:**

  In credit scoring, binning is a crucial step before calculating the Weight of Evidence (WOE) and Information Value (IV). These metrics help assess the predictive power of each variable and ensure a monotonic relationship with the target variable.

- **Improving Model Stability:**

  By discretizing continuous data, binning can help prevent overfitting, especially in models like logistic regression where the relationship between features and the target needs to be stable across different data samples.

- **Aligning with Business Logic:**

  Bins can be created based on domain knowledge or business rules, making the model more actionable and understandable to stakeholders.

WOE (Weight of Evidence)

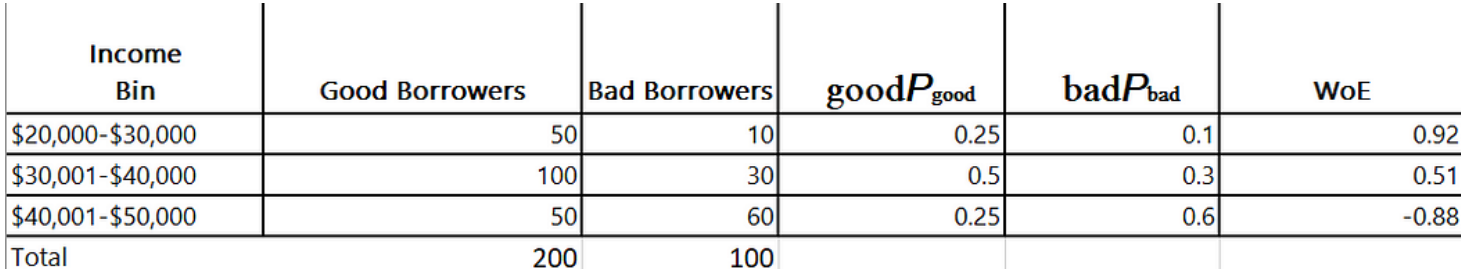
The WOE is a statistical technique that measures the strength of each attribute, or grouped attributes, in separating good and bad accounts.

$$\text{WoE} = \ln\left(\frac{\text{Proportion of Good}}{\text{Proportion of Bad}}\right)$$

In credit scoring, the WoE indicates the predictive power of a particular bin of a variable. Negative WoE values signify that the proportion of “bads” (e.g., defaulters) outweighs the proportion of “goods” (e.g., non-defaulters) within a category or bin. Conversely, positive WoE values indicate that the proportion of “goods” outweighs the proportion of “bads.” Categories with high WoE values are considered to have strong predictive power and vice versa.

Positive WOE: More favorable outcomes (lower risk).

Negative WOE: More unfavorable outcomes (higher risk).



Income Bin	Good Borrowers	Bad Borrowers	goodP_{good}	badP_{bad}	WoE
\$20,000-\$30,000	50	10	0.25	0.1	0.92
\$30,001-\$40,000	100	30	0.5	0.3	0.51
\$40,001-\$50,000	50	60	0.25	0.6	-0.88
Total	200	100			

The Weight of Evidence (WoE) should **exhibit a monotonic trend**, either increasing or decreasing across the bins. This characteristic enhances model stability and facilitates interpretability, strengthening the relationship between the target and predictor variables.

The Weight of Evidence (WoE) plays a crucial role in shaping the final scorecard outcome. When WoE values are closely clustered together, the corresponding points assigned in the scorecard will also exhibit minimal variation. Where the trend of the WoE, is illogical from a business perspective, the bins can be manually adjusted to reflect actual behavior.

IV (Information Value)

This statistical technique measures a variable's predictive power in distinguishing between good and bad credit risk. It helps in selecting and ranking the most significant variables for use in a credit scoring model.

The IV value can be interpreted using the following general guidelines:

- IV < 0.02: Not Predictive
- 0.02 ≤ IV < 0.1: Weak Predictive Power
- 0.1 ≤ IV < 0.3: Medium Predictive Power
- 0.3 ≤ IV < 0.5: Strong Predictive Power
- IV ≥ 0.5: Suspicious or Overfitting

Calculation of IV

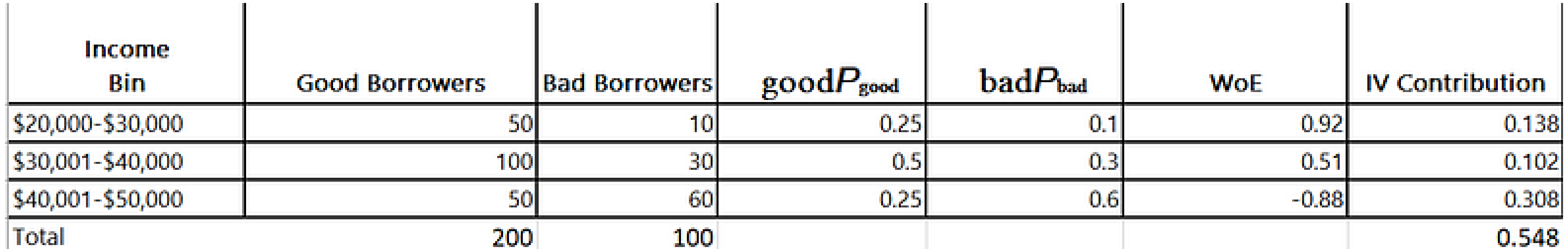
1.Binning: Divide the predictor variable into bins or categories.

2.Calculate Distribution: For each bin, calculate the proportion of good and bad borrowers.

3.Weight of Evidence (WoE): Calculate the WoE for each bin.

4.IV Calculation: Sum up the contributions of each bin to get the total IV.

$$\mathrm{IV} = \sum_i (P_{i,\mathrm{good}} - P_{i,\mathrm{bad}}) \times \mathrm{WoE}_i$$



Income Bin	Good Borrowers	Bad Borrowers	goodP_{good}	badP_{bad}	WoE	IV Contribution
$20,000-$30,000	50	10	0.25	0.1	0.92	0.138
$30,001-$40,000	100	30	0.5	0.3	0.51	0.102
$40,001-$50,000	50	60	0.25	0.6	-0.88	0.308
Total	200	100				0.548

IV=0.138+0.102+0.308=0.548. This IV for instance indicates a very strong predictive power of the variable in distinguishing between good and bad borrowers.

PSI(Population Stability Index)

Population Stability Index (PSI) is a metric used to quantify how much the distribution of a variable has shifted over time or between different datasets (e.g., between a model's training data and new data). i.e. It answers the question: if the base year is 2020 and the comparison year is 2021, how did the distribution of credit scores behave? Did it change or remain the same?

Purpose:

- **Model Monitoring:** Helps detect changes or drifts in the population that could affect model performance.
- **Risk Management:** Alerts when significant changes occur, prompting a review or recalibration of the model.

Expected vs. Actual:

- **Expected Data:** This is the baseline or reference dataset (often the training or development dataset) where the model was originally built.
- **Actual Data:** This is the new or current dataset that you’re comparing against the baseline to see if there have been any shifts.

PSI is primarily used after the scorecard has been deployed to monitor its performance over time. It is calculated periodically (e.g., monthly, quarterly) to compare the distribution of scores in the current period with the distribution in a baseline period (usually the development or initial validation period**). However, it can be useful to calculate PSI between the training and validation datasets to ensure they are similar in distribution.** This helps to check if the model built on the training data is applicable to the validation data.

PSI Calculation

Calculation Steps:

1.Binning: Divide the variable into several bins (using quantiles, fixed intervals, etc.).

2.Compute Proportions: For each bin, calculate:

   1. Expected % – the percentage of observations in the baseline data.
   2. Actual % – the percentage of observations in the new data.

3. Formula: For each bin

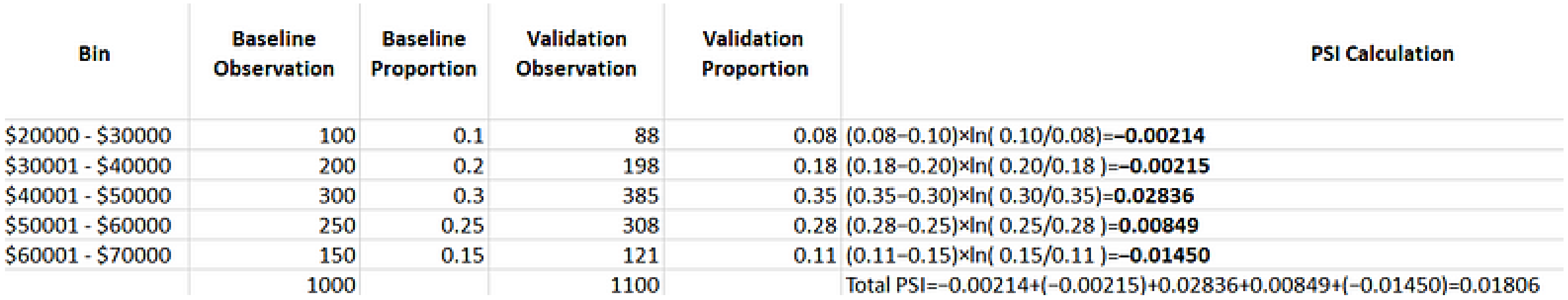
$$\mathrm{PSI}_i = (\mathrm{Actual\%}_i - \mathrm{Expected\%}_i) \times \ln\left(\frac{\mathrm{Actual\%}_i}{\mathrm{Expected\%}_i}\right)$$

4. Total PSI: Sum the contributions across all bins

$$\text{Total PSI} = \sum_i \mathrm{PSI}_i$$

PSI Calculation

PSI=(Proportion in Comparison−Proportion in Baseline)×ln(Proportion in Baseline/Proportion in Comparison)



Bin	Baseline Observation	Baseline Proportion	Validation Observation	Validation Proportion	PSI Calculation
\$20000 - \$30000	100	0.1	88	0.08	(0.08−0.10)×ln(0.10/0.08)=**−0.00214**
\$30001 - \$40000	200	0.2	198	0.18	(0.18−0.20)×ln(0.20/0.18)=**−0.00215**
\$40001 - \$50000	300	0.3	385	0.35	(0.35−0.30)×ln(0.30/0.35)=**0.02836**
\$50001 - \$60000	250	0.25	308	0.28	(0.28−0.25)×ln(0.25/0.28)=**0.00849**
\$60001 - \$70000	150	0.15	121	0.11	(0.11−0.15)×ln(0.15/0.11)=**−0.01450**
	1000		1100		Total PSI=−0.00214+(−0.00215)+0.02836+0.00849+(−0.01450)=0.01806

Interpretation of PSI Values:

•PSI < 0.1:

- Little or no change; the population is considered stable.

•PSI 0.1 to 0.25:

- Moderate change; suggests a shift that should be monitored.

•PSI > 0.25:

- Significant change; indicates a substantial shift that could negatively impact model performance and may require model review or recalibration.

Practical Implications:

- **Regular Monitoring:** PSI should be computed regularly to track if your model's input variables are shifting over time.
- **Actionable Insights:** A high PSI might lead to actions such as further investigation, model recalibration, or even rebuilding the model if the drift is severe.
- **Model Performance:** Significant population shifts (high PSI) often lead to model degradation, making this metric critical for maintaining model reliability.

Step Wise Regression

Stepwise regression is an automated, iterative procedure for selecting variables to include in a regression model. It works by either adding or removing predictors one at a time based on statistical criteria.

Types of Stepwise Methods:

1. **Forward Selection:**

   Starts with no predictors, then adds variables one at a time, choosing the one that most improves the model until no significant improvement is seen.

2. **Backward Elimination:**

   Begins with all candidate predictors in the model and removes the least significant one at each step until only statistically significant predictors remain.

3. **Bidirectional (Stepwise) Selection:**

   A combination of forward selection and backward elimination. It adds variables like forward selection but also checks and removes variables that become non-significant as new predictors are added.

Criteria for Adding/Removing Variables:

Common criteria include p-values, Akaike's Information Criterion (AIC), Bayesian Information Criterion (BIC), or adjusted R^2.

Advantages:

- **Automation:** Helps in handling a large number of candidate predictors.
- **Simplification:** Can result in a more interpretable model by selecting only the most significant variables

Score Card Development

Transition to a Scorecard:

□ **Step 1: Model Estimation**

Develop a logistic regression model using the WOE-transformed variables.

□ **Step 2: Parameter Translation**

Convert the logistic regression coefficients into score contributions for each variable.

□ **Step 3: Score Scaling**

Use the PDO to determine the Factor, ensuring that a predefined change in score reflects a doubling of the odds.

□ **Step 4: Setting the Base Score**

Choose an Offset so that a borrower with “average” risk gets a score that aligns with business requirements.

□ **Step 5: Final Score Calculation**

The final score is computed as the sum of the contributions from each variable plus the base score:

$$\mathrm{Score}_i = -\left(\beta_i \times \mathrm{WoE}_i + \frac{\alpha}{n}\right) \times \mathrm{Factor} + \frac{\mathrm{Offset}}{n}$$

Transition to a Scorecard:

The objective of a credit score is to indicate the odds of an application being a good or bad risk.

1. Understanding Scores and Odds

Odds- The likelihood of a borrower defaulting (not repaying the loan) compared to the likelihood of them repaying the loan. Ex **Odds of 50:1** , means that for every 50 good customers, there is 1 bad customer

Score — A score is a number that tells lenders how likely you are to pay back what you borrow. Higher scores indicate lower risks and vice-versa.

2. Basic Formula for Score Calculation

$Score=Offset+Factor\times ln(odds)$

3. Understanding Offset from the formula

Offset- This is a constant value added to the final score to ensure that the score falls within a desired range and meets specific business objectives or industry standards.

4. Understanding Factor from the formula

Factor — This is a multiplier applied to the logistic regression coefficients to convert them into points. It helps in precision adjustment. A more precise scorecard can better differentiate between **high risk and low risk applicants.** It is a number that helps us figure out how much the credit score should change when the odds of a person repaying a loan change.

The **Factor** is calculated using:

$$\text{Factor} = \frac{\text{pdo}}{\ln(2)}$$

PDO (Points to Double Odds) is a way of deciding **how much a credit score should change** when a person’s chances of repaying a loan (the **odds**) **double**.The Points to Double the Odds is the most widely used scaling formula in the credit risk industry. It creates a consistent relationship between scores and the odds of default.

We use ln(2)(natural logarithm of 2, approximately 0.693) because it represents the change in score required to double the odds. This is the basis for the Points to Double the Odds (PDO) system. It ensures that for instance with every increase of pdo points in the borrower’s score, the odds of defaulting double.

➢ For example, **50:1 odds** mean they are 50 times more likely to repay than default.

➢ When the odds **double**, they become **100:1**, meaning the person is now seen as less risky.

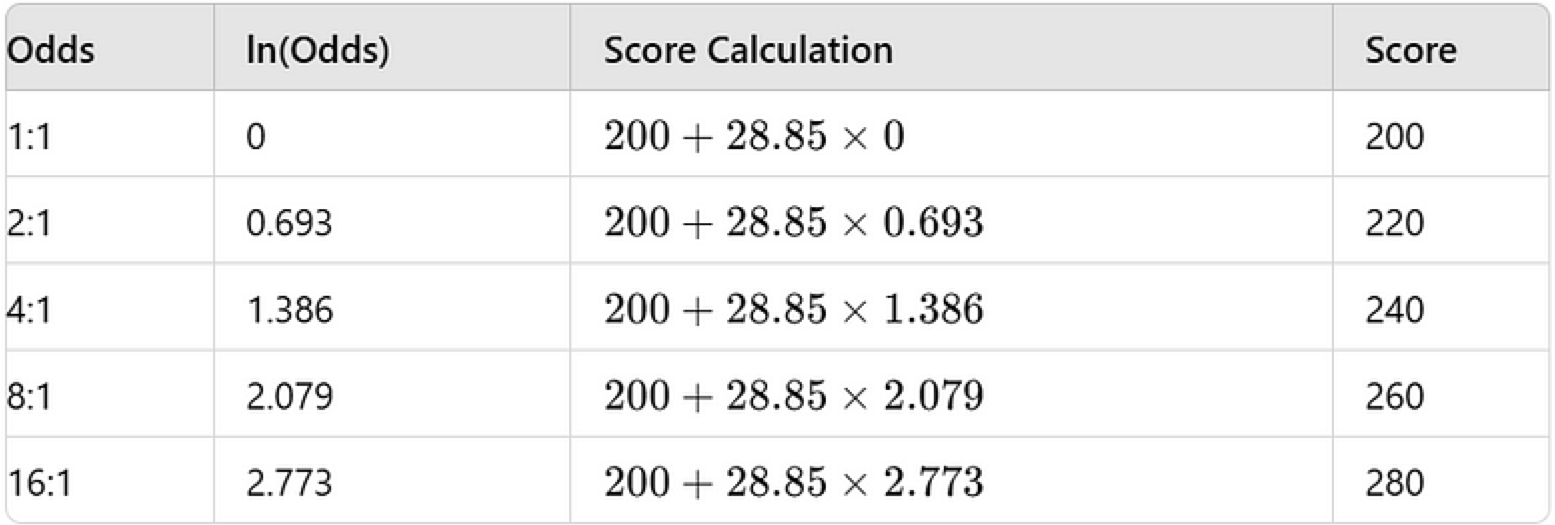
❑ A lower pdo value (e.g., pdo = 10 or 15) means smaller score increments are needed to double the odds.This finer granularity allows for more precise differentiation between different levels of risk.

❑ A higher pdo value (e.g., pdo = 20 or 25) means larger score increments are needed to double the odds. This coarser granularity might simplify the scorecard but could reduce the model’s sensitivity to smaller variations in risk.

- If the **PDO is 20**, it means:

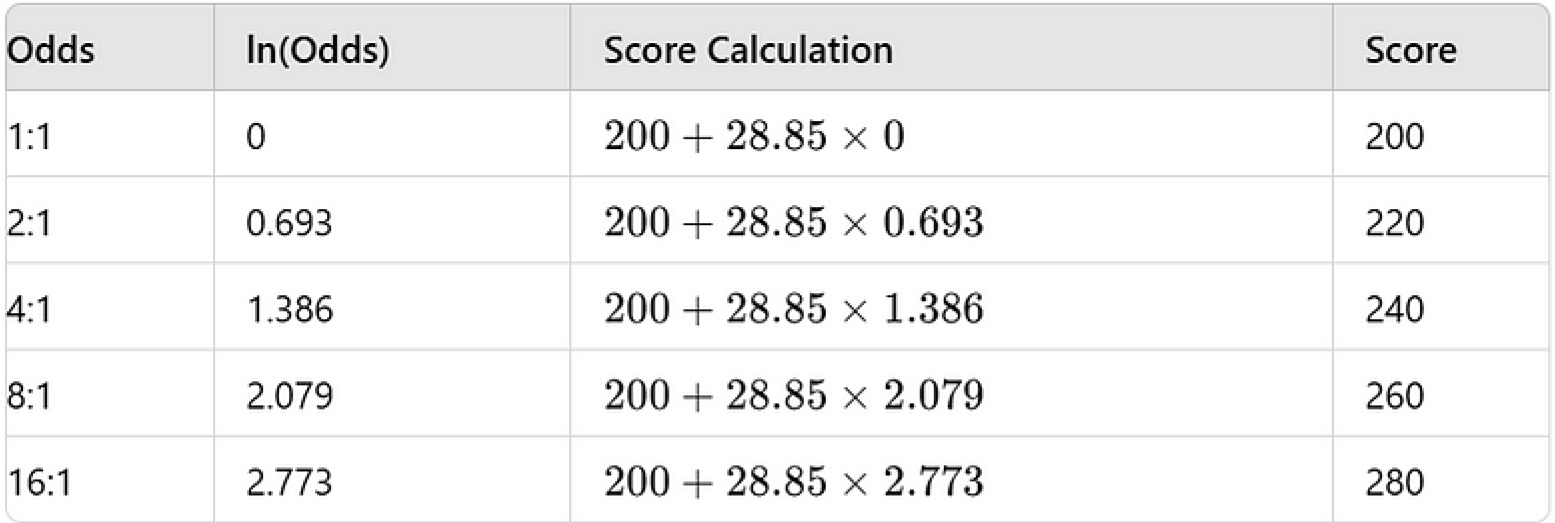
☞ Every time the odds of repayment **double**, we **add 20 points** to the credit score.

- Let’s say someone starts with a score of **200** at 50:1 odds.
- If their odds improve to 100:1, their score would become **220**.
- If their odds double again to 200:1, the score would go up to **240**, and so on.



Odds	ln(Odds)	Score Calculation	Score
1:1	0	$200 + 28.85 \times 0$	200
2:1	0.693	$200 + 28.85 \times 0.693$	220
4:1	1.386	$200 + 28.85 \times 1.386$	240
8:1	2.079	$200 + 28.85 \times 2.079$	260
16:1	2.773	$200 + 28.85 \times 2.773$	280

In the above example, you can see increasing the odds from say 1:1 to 2:1 increases the score by 20 points. 1:1 to 4:1 increases the score by 40 points.



Odds	ln(Odds)	Score Calculation	Score
1:1	0	$200 + 28.85 \times 0$	200
2:1	0.693	$200 + 28.85 \times 0.693$	220
4:1	1.386	$200 + 28.85 \times 1.386$	240
8:1	2.079	$200 + 28.85 \times 2.079$	260
16:1	2.773	$200 + 28.85 \times 2.773$	280

In the above example, you can see increasing the odds from say 1:1 to 2:1 increases the score by 20 points. 1:1 to 4:1 increases the score by 40 points.

5. Calculating Offset and Factor

Assume:

- ❑ Base Score = 600
- ❑ Odds at Base Score = 50:1
- ❑ pdo =20

Step a: Calculate factor

$$\text{Factor} = \frac{\text{pdo}}{\ln(2)} \quad \Rightarrow \quad \text{Factor} = \frac{20}{0.693} \approx 28.85$$

Step b: Calculate Offset

Offset = Score − {Factor ∗ ln (Odds)}

$$\text{Offset} = 600 - (28.85 \times \ln(50))$$
$$\text{Offset} = 600 - (28.85 \times 3.91) \approx 487.123$$

6. Incorporate Logistic Regression and Weight of Evidence (WoE)

Each feature would have a coefficient and each bin in that feature its corresponding WoE. To integrate the WoE

$$\text{Score}_i = -\left(\beta_i \times \text{WoE}_i + \frac{\alpha}{n}\right) \times \text{Factor} + \frac{\text{Offset}}{n}$$

Where:

- WoE = weight of evidence for each grouped feature
- βi = regression coefficient for each feature
- α is the Logistic Regression Intercept
- n = number of bins in the feature

Evaluation Methods

Kolmogorov–Smirnov (KS)

The **Kolmogorov-Smirnov (K-S) technique** is a statistical method used to compare two distributions and measure their difference. In credit scoring and model evaluation, it is commonly used to assess the discriminatory power of a model in distinguishing between "good" and "bad" customers.

Importance of K-S in Credit Scoring

❑ Measures how well a model separates "good" and "bad" applicants.

❑ A higher K-S value indicates better discrimination.

❑ Helps in assessing the predictive power of a scorecard.

K-S Interpretation in Credit Scoring

K-S Value (%)	Interpretation
< 15%	Poor separation. Scorecard might not be useful
15% - 20%	Fair separation. Potentially useful but should be carefully evaluated
20% - 28%	Poor separation but useful
28% -35%	Average separation, definitely useful
35%-45%	High separation for application scorecard
>45	Very high quality application scorecard

What Are We Trying to Measure?

The **goal of a scorecard** is to correctly rank customers by their credit risk. A good model should give **higher scores to "good" customers** and **lower scores to "bad" customers** (or vice versa, depending on the scoring logic).

The **K-S statistic** helps us measure **how well the scorecard separates the two groups**:

- "Good" customers (target = 0)
- "Bad" customers (target = 1)

It does this by checking **how different their cumulative distributions are across score buckets**.

Use of Cumulative Percentages

Imagine sorting customers by their **credit score** (high to low).

For each score threshold, we ask:

➢ **How many "good" customers have a score below this threshold.**

➢ **How many "bad" customers have a score below this threshold?**

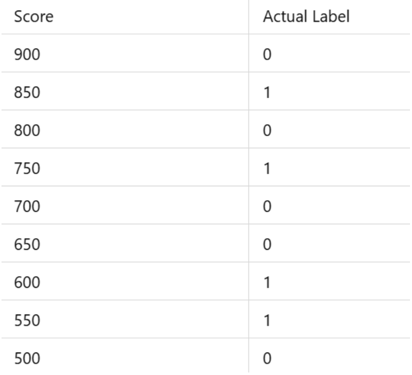
If the model is **perfect**, all **bad** customers should be on one end, and all **good** customers should be on the other.

If the model is **random**, the distributions will look very similar.

The **cumulative percentage** at each threshold helps us track this **separation**.

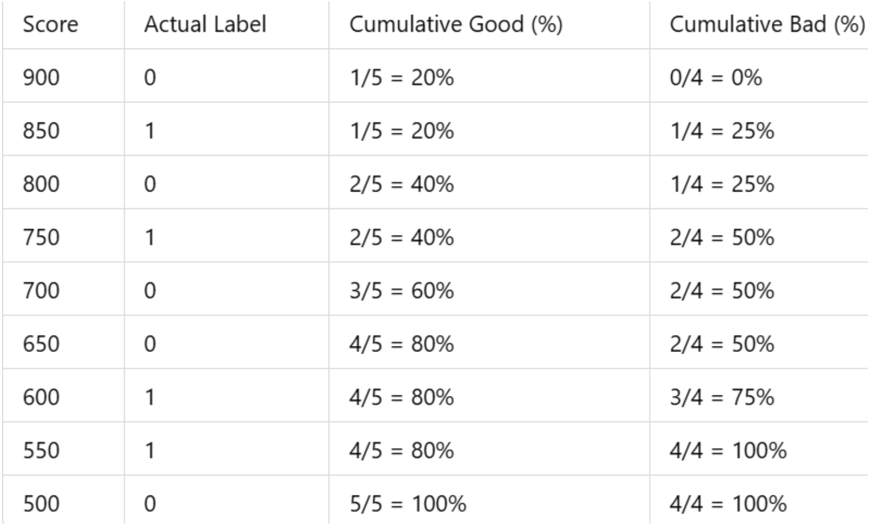
Steps to Calculate K-S

1.Sort the predicted scores in descending order.



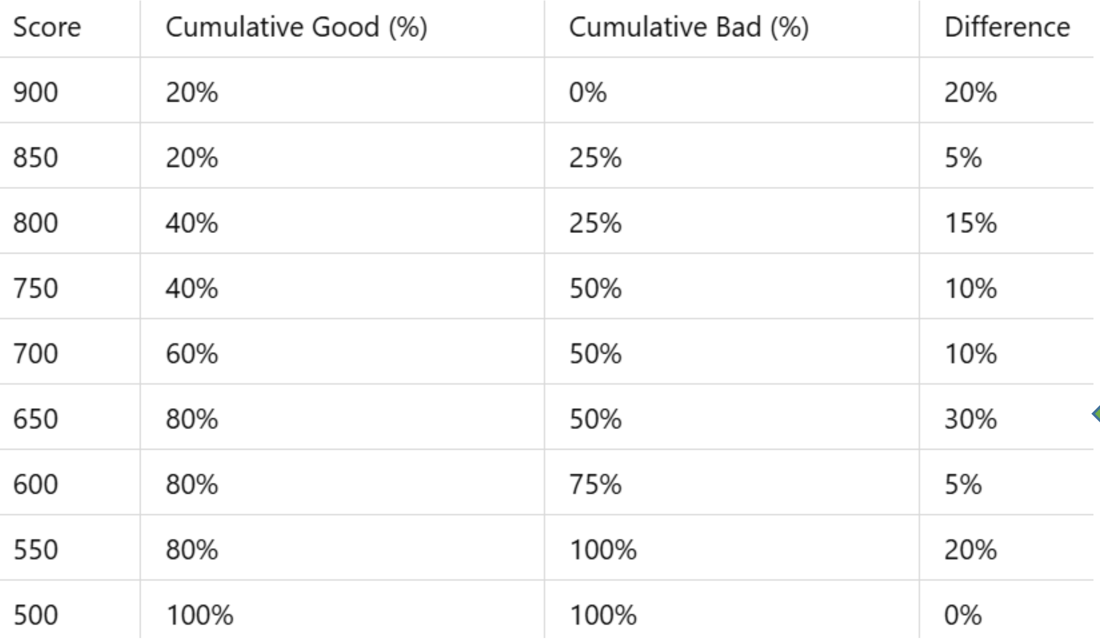
Score	Actual Label
900	0
850	1
800	0
750	1
700	0
650	0
600	1
550	1
500	0

2. Calculate the cumulative percentage of "goods" and "bads" at each score threshold.



Score	Actual Label	Cumulative Good (%)	Cumulative Bad (%)
900	0	1/5 = 20%	0/4 = 0%
850	1	1/5 = 20%	1/4 = 25%
800	0	2/5 = 40%	1/4 = 25%
750	1	2/5 = 40%	2/4 = 50%
700	0	3/5 = 60%	2/4 = 50%
650	0	4/5 = 80%	2/4 = 50%
600	1	4/5 = 80%	3/4 = 75%
550	1	4/5 = 80%	4/4 = 100%
500	0	5/5 = 100%	4/4 = 100%

3.Compute Absolute Differences



Score	Cumulative Good (%)	Cumulative Bad (%)	Difference
900	20%	0%	20%
850	20%	25%	5%
800	40%	25%	15%
750	40%	50%	10%
700	60%	50%	10%
650	80%	50%	30%
600	80%	75%	5%
550	80%	100%	20%
500	100%	100%	0%

4: Find Maximum Difference (K-S Statistic)

➢ **At score 650**, the difference is 30%, which is the **K-S Statistic**.

➢ This means the **best separation** between "good" and "bad" happens at this score.

❑ If good and bad customers are well-separated, their cumulative distribution curves will be **far apart** → **High K-S value** (Good Model).

❑ If good and bad customers are mixed across scores, their curves will be **close together** → **Low K-S value** (Weak Model).

Gini-Coefficient

The **Gini Coefficient** in scorecards measures **how well the scorecard separates good and bad customers** (discriminatory power).

What Does Gini Coefficient Test?

- It tests the **ranking ability** of the scorecard.
- It shows how **different** the score distributions are between good (non-defaulting) and bad (defaulting) customers.
- **Higher Gini = Better separation** (good customers have high scores, bad customers have low scores).

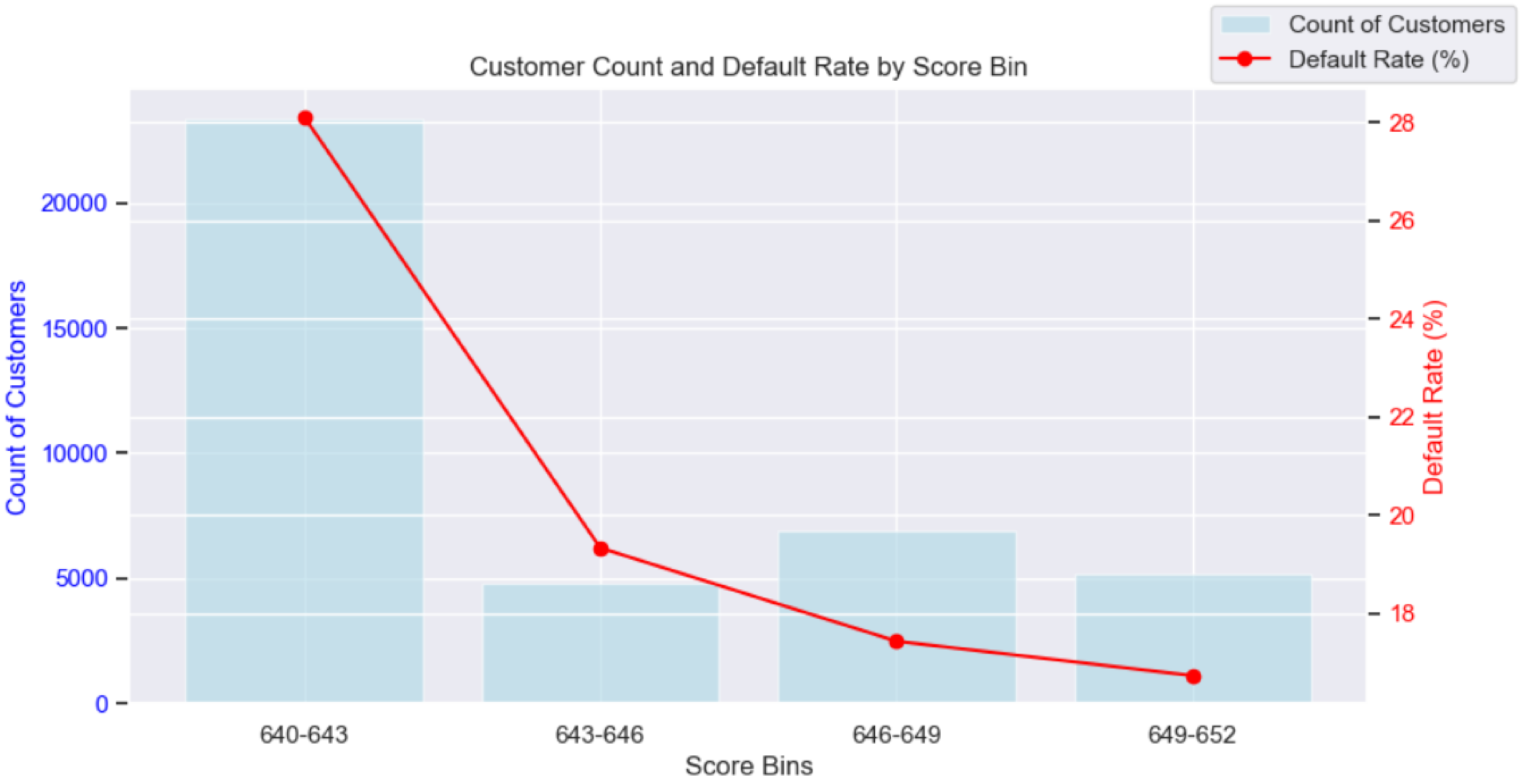
How Gini is Related to AUC (Area Under Curve)

$$Gini = 2 \times AUC - 1$$

- Constructing the Lorenz curve and extract Corrado Gini’s measure to get the Gini coefficient.**(Measures Inequality of distribution)**
- Constructing the ROC curve to extract the AUC and then compute the Gini coefficient.**(Measures discriminatory power)**

  - **AUC = 0.5 → Gini = 0 (random model, no discrimination)**
  - **AUC = 1 → Gini = 1 (perfect model, ideal separation)**
  - **Typical scorecards: Gini between 0.3 - 0.6** (higher is better)

Score Distribution on Test


Customer Count and Default Rate by Score Bin
Count of Customers
Default Rate (%)
Count of Customers
Default Rate (%)
Score Bins