Predicting Loan Default and Borrower Risk Segmentation

Author:
Pramod Timalsina

Date: June 8, 2025

Table of Contents

1. Executive Summary

This report outlines the development of a logistic regression model to predict the probability of loan default and segment borrowers based on risk. The model shows strong predictive performance, with an AUC of 0.879 and an accuracy of 86.7%. Key predictors of default include loan grade, income, home ownership status, and the loan-to-income ratio. While the model is highly precise in identifying defaults, it shows a trade-off in recall—highlighting an opportunity to adjust the classification threshold to reduce the number of missed default cases. These insights support effective customer segmentation, enabling financial institutions to tailor lending strategies, manage risk more efficiently, and improve profitability. Overall, the model serves as a powerful, data-driven tool for making informed decisions in credit risk assessment.

2. Introduction

In the financial sector, accurately assessing credit risk is essential for sustainable lending practices. This project focused on building a reliable predictive model to:

- Estimate the likelihood of a loan default based on a borrower's financial profile.
- Segment borrowers into high-risk and low-risk groups, offering actionable insights to support loan approval decisions and risk management strategies.

Logistic regression was used as the primary analytical method, given its effectiveness and wide acceptance for modelling binary outcomes like default versus non-default.

3. Data Description and Preparation

The analysis utilized a credit risk dataset containing various borrower and loan-specific attributes.

3.1 Features of the Dataset:

Original Feature	Description
person_age	Age of the borrower
person_income	Borrower's annual income
person_home_ownership	Type of home ownership (e.g., RENT, OWN)
person_emp_length	Employment length (in years)
loan_intent	Purpose of the loan (e.g., EDUCATION, MEDICAL)
loan_grade	Loan's assigned risk grade (A, B, C, etc.)
loan_amnt	Total loan amount
loan_int_rate	Interest rate of the loan
loan_status	Loan status (0 = Non-Default, 1 = Default)
loan_percent_income	Loan amount as a percentage of annual income
cb_person_default_on_file	Indicator of historical default
cb_person_cred_hist_length	Length of credit history

3.2 Data Cleaning and Preprocessing:

Several steps were taken to prepare the dataset for analysis:

- Unrealistic Values: Records with implausible values such as age over 100 or employment length exceeding 50 years were removed, as these were likely data entry errors.
- Missing Values: Missing data in the person_emp_length and loan_int_rate columns were handled using median imputation. This approach replaced missing entries with the median of the respective column, preserving the larger dataset size.
- Duplicate Records: Duplicate rows were identified and removed, resulting in the deletion of 172 entries.
- Categorical Variables: Categorical fields such as person_home_ownership, loan_intent, and loan_grade were converted into numerical dummy variables. For example, "Rent" was used as the reference category for home ownership, with other categories like "Own", "Mortgage", and "Other" compared against it. Similarly, "Grade A" was the reference for loan grades, and "Personal" for loan intent.
- Variable Transformation: The person_income variable was log-transformed to reduce the effect of extreme values and improve model performance. Some column names were also updated for clarity for instance, person_age was renamed to Age.

4. Model Building: Logistic Regression Analysis

This section highlights the logistic regression model developed using EViews, serving as a representative example of the relationships uncovered through the analysis. In this model, the dependent variable is Loan_Status, where 1 indicates a default and 0 indicates a non-default. The final dataset used for the model contains 32,409 observations.

The logistic regression model follows the typical form:

$$\ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 A_i + \beta_2 B_i + \beta_3 C_i + \ \ldots\ldots\ldots + \varepsilon_i$$

Where:

- P represents the probability of loan default.
- $\ln\left(\frac{P}{1-P}\right)$ is the log-odds (or logit) of loan default.
- β0 is the intercept. It represents the log-odds of default when all independent variables are zero or at their reference categories.
- β1, β2, β3 …. are the coefficients of the independent variables (A, B, C….). In this model independents variables are the predictors of loan default such as Age, Income, Loan Grade etc. Each coefficient (βi) represents the change in the log-odds of default for a one-unit increase in the corresponding predictor, holding all other predictors constant.

Using all selected predictors, the logistic regression model was estimated in EViews. The results of this model are presented below:

Dependent Variable: LOAN_STATUS
Method: ML - Binary Logit (Newton-Raphson / Marquardt steps)
Date: 06/08/25 Time: 09:45
Sample: 1 32409
Included observations: 32409
Convergence achieved after 7 iterations
Coefficient covariance computed using observed Hessian

Variable	Coefficient	Std. Error	z-Statistic	Prob.
C	8.628411	0.803634	10.73674	0.0000
AGE	-0.003876	0.005859	-0.661627	0.5082
CREDIT_HISTORY_YEARS	0.001857	0.008929	0.207953	0.8353
DEBTCONSOLIDATION_LOAN	0.636492	0.058433	10.89272	0.0000
HOMEIMPROVEMENT_LOAN	0.716803	0.066617	10.76008	0.0000
EDUCATION_LOAN	-0.220993	0.059687	-3.702551	0.0002
MEDICAL_LOAN	0.428737	0.056981	7.524263	0.0000
VENTURE_LOAN	-0.470340	0.063973	-7.352176	0.0000
DEFAULT_HISTORY	0.001576	0.050407	0.031272	0.9751
EMPLOYMENT_PERIOD_YEARS_	-0.007078	0.004855	-1.457920	0.1449
HOME_MORTGAGE_	-0.767911	0.040542	-18.94107	0.0000
HOME_OTHER	-0.253001	0.285385	-0.886526	0.3753
HOME_OWN_	-2.476649	0.094896	-26.09865	0.0000
INTEREST_RATE	0.056336	0.012968	4.344142	0.0000
LOAN_AMOUNT	-4.70E-06	7.06E-06	-0.665734	0.5056
LOAN_GRADE_B_	0.206943	0.063738	3.246799	0.0012
LOAN_GRADE_C_	0.437611	0.090887	4.814876	0.0000
LOAN_GRADE_D_	2.536680	0.111867	22.67585	0.0000
LOAN_GRADE_E_	2.785579	0.145620	19.12903	0.0000
LOAN_GRADE_F_	3.117987	0.217908	14.30872	0.0000
LOAN_GRADE_G_	6.882053	1.052882	6.536394	0.0000
LOAN_INCOME_RATIO	8.162879	0.364030	22.42364	0.0000
LOG_INCOME	-1.155963	0.073447	-15.73881	0.0000

McFadden R-squared	0.360557	Mean dependent var	0.218705
S.D. dependent var	0.413374	S.E. of regression	0.317860
Akaike info criterion	0.673172	Sum squared resid	3272.129
Schwarz criterion	0.679123	Log likelihood	-10885.41
Hannan-Quinn criter.	0.675074	Deviance	21770.82
Restr. deviance	34046.54	Restr. log likelihood	-17023.27
LR statistic	12275.72	Avg. log likelihood	-0.335876
Prob(LR statistic)	0.000000		

Obs with Dep=0	25321	Total obs	32409
Obs with Dep=1	7088		

The initial regression results indicated that several variables, namely Age, Credit History (Years), Default History, Employment Period (Years), and Loan Amount, were statistically insignificant. As a result, these variables were excluded from the set of predictors, and the model was re-estimated using only the significant variables. In the final model, the dummy variable Home_Ownership_Other remained statistically insignificant. However, it was retained to ensure the complete representation of the person_home_ownership categorical variable, preserving the integrity of the dummy variable structure.

The results of the final model estimation are presented below:

Dependent Variable: LOAN_STATUS
Method: ML - Binary Logit (Newton-Raphson / Marquardt steps)
Date: 06/08/25 Time: 09:52
Sample: 1 32409
Included observations: 32409
Convergence achieved after 6 iterations
Coefficient covariance computed using observed Hessian

Variable	Coefficient	Std. Error	z-Statistic	Prob.
C	9.082665	0.416824	21.79015	0.0000
DEBTCONSOLIDATION_LOAN	0.636764	0.058406	10.90240	0.0000
HOMEIMPROVEMENT LOAN	0.712381	0.066514	10.71018	0.0000
EDUCATION_LOAN	-0.215859	0.059550	-3.624848	0.0003
MEDICAL_LOAN	0.426174	0.056933	7.485530	0.0000
VENTURE LOAN	-0.468460	0.063922	-7.328583	0.0000
HOME_MORTGAGE_	-0.778541	0.039987	-19.47002	0.0000
HOME_OTHER_	-0.241477	0.284714	-0.848137	0.3964
HOME OWN	-2.477551	0.094662	-26.17251	0.0000
INTEREST_RATE	0.056112	0.012960	4.329648	0.0000
LOAN_GRADE_B_	0.207192	0.063698	3.252747	0.0011
LOAN GRADE C	0.440191	0.087309	5.041762	0.0000
LOAN_GRADE_D_	2.534735	0.108717	23.31497	0.0000
LOAN_GRADE_E_	2.784676	0.143328	19.42866	0.0000
LOAN GRADE F	3.116138	0.216673	14.38174	0.0000
LOAN_GRADE_G_	6.869108	1.052595	6.525878	0.0000
LOAN_INCOME_RATIO	7.934469	0.169150	46.90795	0.0000
LOG INCOME	-1.209579	0.037567	-32.19832	0.0000

McFadden R-squared	0.360439	Mean dependent var	0.218705
S.D. dependent var	0.413374	S.E. of regression	0.317897
Akaike info criterion	0.672987	Sum squared resid	3273.387
Schwarz criterion	0.677645	Log likelihood	-10887.42
Hannan-Quinn criter.	0.674476	Deviance	21774.84
Restr. deviance	34046.54	Restr. log likelihood	-17023.27
LR statistic	12271.70	Avg. log likelihood	-0.335938
Prob(LR statistic)	0.000000		

Obs with Dep=0	25321	Total obs	32409
Obs with Dep=1	7088		

Overall Model Fit: The value of McFadden R-squared is 0.3604. This suggests the model explains approximately 36% of the variability in loan default. Furthermore, Prob(LR statistic) is 0.0000. The model is highly statistically significant, indicating that its predictors collectively provide strong predictive power.

5. Interpretation of Key Predictors:

The table below presents selected predictors, their coefficients (log-odds), and their corresponding Odds Ratios (e^coefficient). An Odds Ratio greater than 1 suggests an increased likelihood of default, while less than 1 suggests a decreased likelihood, holding other factors constant.

Variable	Coefficient	Odds Ratio	Interpretation

Intercept	9.0827	—	This is the log odds of default when all other predictor variables are zero. In practical terms, this represents a baseline log odd of default under hypothetical "zero" conditions for all other factors.
Debt Consolidation Loan	0.6368	1.89	Increases default likelihood by 89% relative to Personal Loans.
Home Improvement Loan	0.7124	2.04	Home improvement loan (compared to a personal loan) increases the odds of default by approximately 2.04 times.
Education Loan	-0.2159	0.81	Reduces default odds by 19% relative to Personal Loans. This suggests education loans are significantly less risky (Prob. 0.0003).
Venture Loan	-0.4685	0.63	Venture loan (compared to a personal loan) decreases the odds of default by approximately 37%. This suggests venture loans are significantly less risky (Prob. 0.0000)
Medical Loan	0.4262	1.53	Medical loan (compared to a personal loan) increases the odds of default by approximately 53%.
Loan Grade B	0.2072	1.23	Compared to a Loan Grade A, a loan with Grade B increases the odds of default by approximately 23%.
Loan Grade C	1.4402	1.55	Compared to a Loan Grade A, a loan with Grade C increases the odds of default by approximately 55%.
Loan Grade D	2.5347	12.61	Compared to a Loan Grade A, a loan with Grade D substantially increases the odds of default by approximately 12.61 times.
Loan Grade E	2.7845	16.20	Compared to a Loan Grade A, a loan with Grade E substantially increases the odds of default by approximately 16.20 times.
Loan Grade F	3.1161	22.56	Compared to a Loan Grade A, a loan with Grade F drastically increases the odds of default by approximately 2156%.
Loan Grade G	6.8691	962.19	Compared to a Loan Grade A, a loan with Grade G has an extremely large increase in the odds of default by approximately 96,119%. This strongly suggests that Grade G loans carry an exceptionally high risk of default (Prob. 0.0000).

Home Ownership: Mortgage	-0.7785	0.46	Having a home with a mortgage (compared to renting) decreases the odds of default by approximately 54%. This is highly significant and intuitive, as homeowners with mortgages generally exhibit greater financial stability (Prob. 0.0000).
Home Ownership: Own	-2.4776	0.08	Owning a home outright (compared to renting) decreases the odds of default by a substantial 92%. This is highly significant and strongly indicates that outright homeowners are much less likely to default (Prob. 0.0000).
Home Ownership: Other	-0.2415	0.79	Having "other" home ownership status (compared to renting) decreases the odds of default by about 21%. However, this finding is not statistically significant (Prob. = 0.3964), meaning we cannot confidently conclude a real difference in default risk from renting based on this category.
Interest Rate	0.0561	1.06	For every one-unit increase in the INTEREST_RATE, the odds of default increase by approximately 6%. This is statistically significant and intuitive: higher interest rates often correlate with higher perceived risk or higher payment burdens (Prob. 0.0000).
Loan-Income Ratio	7.9345	2792.8	For every one-unit increase in the LOAN_INCOME_RATIO, the odds of default increase by an extremely large factor of approximately 2792.8 times, or 279,180%. This is highly significant and indicates that as the loan amount becomes a larger proportion of income, the default risk rises dramatically (Prob. 0.0000).
Log_Income	-1.2096	0.30	For every one-unit increase in LOG_INCOME (representing a multiplicative increase in actual income), the odds of default decrease by approximately 70%. This is highly significant and intuitive: higher income significantly reduces the likelihood of loan default (Prob. 0.0000).

Based on the comprehensive interpretation, following key conclusions is drawn about the factors influencing loan default:

1. Borrower Characteristics are Paramount (Income & Home Ownership):
    - Higher income significantly reduces default risk: As confirmed by the LOG_INCOME coefficient, borrowers with higher incomes are substantially less likely to default.
    - Homeownership is a strong protective factor: Owning a home outright or having a mortgage drastically decreases the odds of default compared to renting, indicating greater financial stability and asset backing.
2. Loan Quality (Grade) is a Dominant Predictor of Risk:
    - Worse loan grades correlate with drastically higher default rates: There is a clear and escalating relationship where lower loan grades (from B to G, relative to A) correspond to progressively and significantly higher odds of default. Loan Grade G, in particular, indicates an extremely high probability of default. This suggests that internal or external risk assessments (captured by loan grade) are highly effective in predicting default.
3. Loan Purpose and Structure Influence Risk:
    - High-risk loan purposes: Loans for Debt Consolidation, Home Improvement, and Medical expenses are associated with a significantly higher likelihood of default compared to a 'Personal Loan'. This suggests these purposes might indicate existing financial strain or a less stable financial position for the borrower.
    - Lower-risk loan purposes: Education and Venture loans, conversely, show a significantly lower likelihood of default compared to 'Personal Loans'. This could imply that borrowers for these purposes are either more financially disciplined or that these loan types have built-in structures (e.g., career benefits from education, higher scrutiny for venture capital) that mitigate risk.
    - Loan-to-Income Ratio is a critical risk amplifier: A higher LOAN_INCOME_RATIO dramatically increases the odds of default, highlighting that the burden of the loan relative to a borrower's earning capacity is a primary driver of default risk.
    - Interest Rate indicates risk: A higher INTEREST_RATE is associated with increased default risk, which is logical as it reflects either higher borrowing costs or the lender's initial assessment of higher risk for that loan.

Overall, the model reveals that loan default risk is primarily driven by the borrower's fundamental financial health (income, asset ownership), the inherent risk profile assigned to the loan (loan grade), and the specific purpose and burden of the loan. Lenders can use these insights to assess creditworthiness more accurately and manage their portfolios effectively.

5 Customer Segmentation:

5.1 Customer Segment Analysis

The results from logistic regression model provide a rich basis for loan applicant or borrower segmentation based on their default risk profile. Segmentation is a powerful tool for lenders to tailor their strategies, whether for loan approvals, interest rates, marketing, or collections.

Here are some practical segmentations based on the findings, categorizing them by the key variables:

1. Loan Grade-Based Segmentation (Primary Risk Indicator)

This is the most obvious and powerful segmentation.

- Segment 1: Very Low Risk (Loan Grade A)
    - Characteristics: Borrowers who would be in the omitted reference category (A) or are classified as Loan_Grade_A.
    - Risk Profile: Lowest odds of default.
    - Lending Strategy: Ideal customers. Offer competitive rates, streamlined approval processes.
- Segment 2: Moderate Risk (Loan Grades B & C)
    - Characteristics: Borrowers with Loan_Grade_B or Loan_Grade_C.
    - Risk Profile: Significantly higher odds of default than Grade A (23% and 55% higher, respectively).
    - Lending Strategy: Still potentially desirable customers but require careful consideration. May warrant slightly higher interest rates, more stringent income/debt-to-income checks, or smaller loan amounts.
- Segment 3: High Risk (Loan Grades D, E, F)
    - Characteristics: Borrowers with Loan_Grade_D, Loan_Grade_E, or Loan_Grade_F.
    - Risk Profile: Drastically higher odds of default (1161% to 2156% higher than Grade A).
    - Lending Strategy: Very cautious approach. Likely require very high interest rates, significant collateral, strict repayment terms, or might be declined depending on the lender's risk appetite.
- Segment 4: Extremely High Risk (Loan Grade G)
    - Characteristics: Borrowers with Loan_Grade_G.
    - Risk Profile: Exceedingly high odds of default (96,119% higher than Grade A).

    - Lending Strategy: Likely should not be approved for loans under standard conditions. If approved, it would be under highly specialized, secured, and extremely high-interest terms, possibly with significant upfront payments.

2. Home Ownership-Based Segmentation (Financial Stability Indicator)

This provides another strong indicator of financial stability.

- Segment 1: Lowest Default Risk (Home_Own)
    - Characteristics: Borrowers who own their home outright.
    - Risk Profile: 92% lower odds of default compared to renters.
    - Lending Strategy: Highly desirable customers, often qualify for the best rates and terms.
- Segment 2: Low Default Risk (Home_Mortgage)
    - Characteristics: Borrowers with a home mortgage.
    - Risk Profile: 54% lower odds of default compared to renters.
    - Lending Strategy: Also desirable, but slightly higher risk than outright owners. Still good candidates for competitive offers.
- Segment 3: Baseline/Higher Risk (Rent)
    - Characteristics: Borrowers who rent.
    - Risk Profile: This is your baseline, so other categories are compared to it. They have higher default odds than homeowners.
    - Lending Strategy: Standard assessment, potentially higher scrutiny than homeowners.
- Segment 4: Unclear/Variable Risk (Home_Other)
    - Characteristics: Borrowers with "other" home ownership status.
    - Risk Profile: No statistically significant difference from renters in terms of default risk.
    - Lending Strategy: May require more individual assessment or be grouped with the "Rent" segment if no clear pattern emerges with more data.

3. Loan Purpose (Intent) Based Segmentation

This helps understand which types of loans are inherently riskier for a given borrower profile.

- Segment 1: Higher Default Risk Purposes:
    - Characteristics: Borrowers seeking Debt Consolidation, Home Improvement, or Medical loans.
    - Risk Profile: Significantly higher odds of default (53% to 104% higher) compared to 'Personal Loans'.

    - Lending Strategy: Requires stricter underwriting, potentially higher rates, or smaller loan amounts due to the higher inherent risk associated with these loan intents.

- Segment 2: Lower Default Risk Purposes:
    - Characteristics: Borrowers seeking Education or Venture loans.
    - Risk Profile: Significantly lower odds of default (19% to 37% lower) compared to 'Personal Loans'.
    - Lending Strategy: May be offered more favorable terms, even if other risk factors are present, as the purpose itself is mitigating.
- Segment 3: Baseline Risk Purpose:
    - Characteristics: Borrowers seeking Personal Loans.
    - Risk Profile: This is the baseline.
    - Lending Strategy: Standard assessment.

4. Financial Health (Income & Ratio) Based Segmentation

These are continuous variables, so segmentation would involve defining thresholds.

- Segment 1: High Income, Low Loan Burden:
    - Characteristics: Borrowers with high LOG_INCOME and low LOAN_INCOME_RATIO.
    - Risk Profile: Very low default risk.
    - Lending Strategy: Prime candidates for the best loan products.
- Segment 2: Moderate Income, Moderate Loan Burden:
    - Characteristics: Borrowers in the middle ranges for LOG_INCOME and LOAN_INCOME_RATIO.
    - Risk Profile: Moderate default risk.
    - Lending Strategy: Standard offers, balanced assessment of other factors.
- Segment 3: Low Income, High Loan Burden:
    - Characteristics: Borrowers with low LOG_INCOME and high LOAN_INCOME_RATIO.
    - Risk Profile: Very high default risk. The LOAN_INCOME_RATIO has an exceptionally strong positive impact on default odds.
    - Lending Strategy: Likely to be declined or offered highly restrictive, secured loans with very high rates. This segment requires extreme caution.

Combining Segmentations

The most powerful segmentation comes from combining these factors. For example:

- "Gold Tier" Applicants: Home_Own + LOG_INCOME (high) + Loan_Grade_A + Education_Loan or Venture_Loan. These would be the lowest risk and receive the best offers.
- "High-Risk Alert" Applicants: Rent + Loan_Grade_G + LOAN_INCOME_RATIO (high) + DebtConsolidation_Loan. These would be very likely to default and warrant almost certain denial.
- "Borderline" Applicants: These might have a mix of good and bad factors (e.g., Home_Mortgage + Loan_Grade_D + average LOG_INCOME). For these, the lender would need to weigh the competing factors and decide on appropriate terms or further review.

By creating these segments, a lending institution can develop targeted strategies for underwriting, pricing, risk management, and even collection efforts, leading to more profitable and sustainable lending practices. The results provide a strong basis for customer segmentation, especially for a lending institution. By segmenting your customer base based on these significant risk factors, you can tailor your lending strategies, marketing efforts, risk assessment, and even product offerings.

5.2 Segmentation Strategies and Actions:

Primary Segmentation Axes (Strongest Predictors of Default):

1. Loan Grade (Most Impactful)
2. Home Ownership Status
3. Income Level (via LOG_INCOME)
4. Loan-to-Income Ratio
5. Loan Intent

Recommended segments and how institution could act on these segments are given below:

Segment 1: "High-Risk Borrowers"

- Characteristics:
    - Loan Grade: D, E, F, or especially G (very high default odds).
    - Home Ownership: Renting.
    - Loan Intent: Debt Consolidation, Home Improvement, Medical.
    - Loan-Income Ratio: High.
    - Income: Lower income levels.
- Implications: These borrowers are at an exceptionally high risk of default based on multiple strong indicators.
- Potential Actions:

    - Lending Decisions: High likelihood of loan rejection. If approved, require very high interest rates, strong collateral, or co-signers.
    - Risk Management: Close monitoring, stricter repayment schedules, immediate follow-up on missed payments.
    - Product Offering: Potentially offer alternative financial products like secured loans (if collateral is available) or financial counseling instead of traditional unsecured loans.
    - Marketing: Avoid targeted marketing for new loans.

Segment 2: "Moderate-Risk Borrowers"

- Characteristics:
    - Loan Grade: B or C.
    - Home Ownership: Could be mixed, but perhaps more renters or those with mortgages.
    - Loan Intent: Could be Personal, or mixed.
    - Loan-Income Ratio: Moderate.
    - Income: Moderate income levels.
    - *(Could also include individuals with some risk factors, but offset by others)*
- Implications: These borrowers represent a balanced risk. They are more likely to default than low-risk groups but not as severe as high-risk.
- Potential Actions:
    - Lending Decisions: Offer loans with competitive but appropriately risk-adjusted interest rates. May require more stringent income verification or slightly lower loan amounts.
    - Risk Management: Regular monitoring, perhaps automated alerts for early signs of distress.
    - Product Offering: Standard loan products, potentially with tiered pricing based on specific risk scores.
    - Marketing: Targeted marketing for specific loan products, potentially offering incentives for good repayment behavior.

Segment 3: "Low-Risk Borrowers"

- Characteristics:
    - Loan Grade: A (or unstated best grade).
    - Home Ownership: Own outright (HOME_OWN) or have a mortgage (HOME_MORTGAGE).
    - Loan Intent: Education, Venture.

    - Loan-Income Ratio: Low.
    - Income: Higher income levels.
- Implications: These borrowers have a significantly lower likelihood of default.
- Potential Actions:
    - Lending Decisions: Offer highly competitive interest rates and favorable loan terms (e.g., higher loan amounts, longer repayment periods). Fast-track approval process.
    - Risk Management: Less intensive monitoring, focus on relationship management.
    - Product Offering: Premium loan products, cross-selling other financial services (e.g., investments, credit cards with higher limits).
    - Marketing: Focus on retention, loyalty programs, and attracting new "prime" borrowers through appealing offers.

5.3 Implement of Segmentation:

1. Define Clear Rules/Scores: You can create internal scoring models based on the coefficients. For instance, assign points for different loan grades, home ownership statuses, and ranges of log income/loan-to-income ratio.
2. Combine Variables: Don't just segment on one variable. Combine the strongest predictors. For example, a "Grade G, Renter with High Loan-Income Ratio" vs. a "Grade A, Homeowner with Low Loan-Income Ratio."
3. Automate (if possible): Use the logistic regression model directly to predict the probability of default for each new applicant. Then, create probability thresholds to define your segments (e.g., Top 10% probability of default = High Risk, Next 30% = Moderate Risk, Bottom 60% = Low Risk).
4. Validate: Regularly review your segments to ensure they accurately reflect actual default rates and that your strategies for each segment are effective.

By implementing these results for segmentation, the institution can make more informed, data-driven decisions that optimize risk management, improve profitability, and enhance customer relationships.

6. Model Performance Evaluation:

The logistic regression model was trained and tested using a stratified shuffle split sampling technique, employing 20 random samples with 80% of the data, and targeting class 1('loan default').

6.1. Overall Performance Metrics (Scores Table)

Test and Score Mon Jun 09 25, 14:19:05

Settings

Sampling type: Stratified Shuffle split, 20 random samples with 80% data
Target class: 1

Scores

Model	AUC	CA	F1	Prec	Recall	MCC
Logistic Regression	0.879	0.867	0.647	0.772	0.556	0.579

Write a comment...

Confusion Matrix Mon Jun 09 25, 14:19:20

Confusion matrix for Logistic Regression (showing number of instances)

		Predicted 0	Predicted 1	Predicted Σ
Actual	0	96617	4663	101,280
	1	12580	15780	28,360
	Σ	109,197	20,443	129,640

Write a comment...

Metric	Value	Interpretation
AUC	0.879	The Area Under the Receiver Operating Characteristic (ROC) curve is 0.879. This is a strong result, indicating that the model has a high ability to distinguish between the two classes (default vs. non-default). An AUC of 0.5 indicates a model no better than random guessing, while 1.0 indicates a perfect classifier. 0.879 suggests excellent discriminative power.
CA	0.867	Classification Accuracy is 0.867 (86.7%). This metric represents the proportion of total predictions that were correct (both true positives and true negatives). While generally good, accuracy can be misleading in imbalanced datasets (where one class is much more frequent than the other), as a model can achieve high accuracy by simply predicting the majority class.
F1	0.647	The F1-score is 0.647. This is the harmonic mean of precision and recall, providing a balanced measure of the model's performance, especially useful in cases of imbalanced classes. A higher F1-score indicates a better balance between precision and recall. For predicting defaults (a likely minority class), this is a crucial metric. Its value indicates a decent, but not outstanding, balance.
Prec	0.772	Precision (Positive Predictive Value) is 0.772 (77.2%). This means that when the model predicts a default (class 1), it is correct 77.2% of the time.

		High precision is important if the cost of false positives (incorrectly predicting default) is high.
Rec	0.556	Recall (Sensitivity or True Positive Rate) is 0.556 (55.6%). This indicates that the model correctly identifies 55.6% of all actual default cases. High recall is critical if the cost of false negatives (failing to predict an actual default) is high. The relatively lower recall compared to precision suggests the model is more conservative in predicting defaults, leading to more missed defaults.
MCC	0.579	The Matthews Correlation Coefficient (MCC) is 0.579. MCC is a more robust and informative metric than accuracy or F1-score, particularly for imbalanced datasets, as it takes into account all four values in the confusion matrix. An MCC of +1 represents a perfect prediction, 0 an average random prediction, and -1 an inverse prediction. An MCC of 0.579 indicates a moderately good prediction quality.

6.2. Confusion Matrix Analysis

The confusion matrix provides a detailed breakdown of the model's predictions versus actual outcomes (target class: 1 for default).

		Predicted		
		0 (Non-Default)	1 (Default)	Total Actual (Σ)
Actual	0 (Non-Default)	96,617 (TN)	4,663 (FP)	101,280
	1 (Default)	12,580 (FN)	15,780 (TP)	28,360
	Total Predicted (Σ)	109,197	20,443	129,640

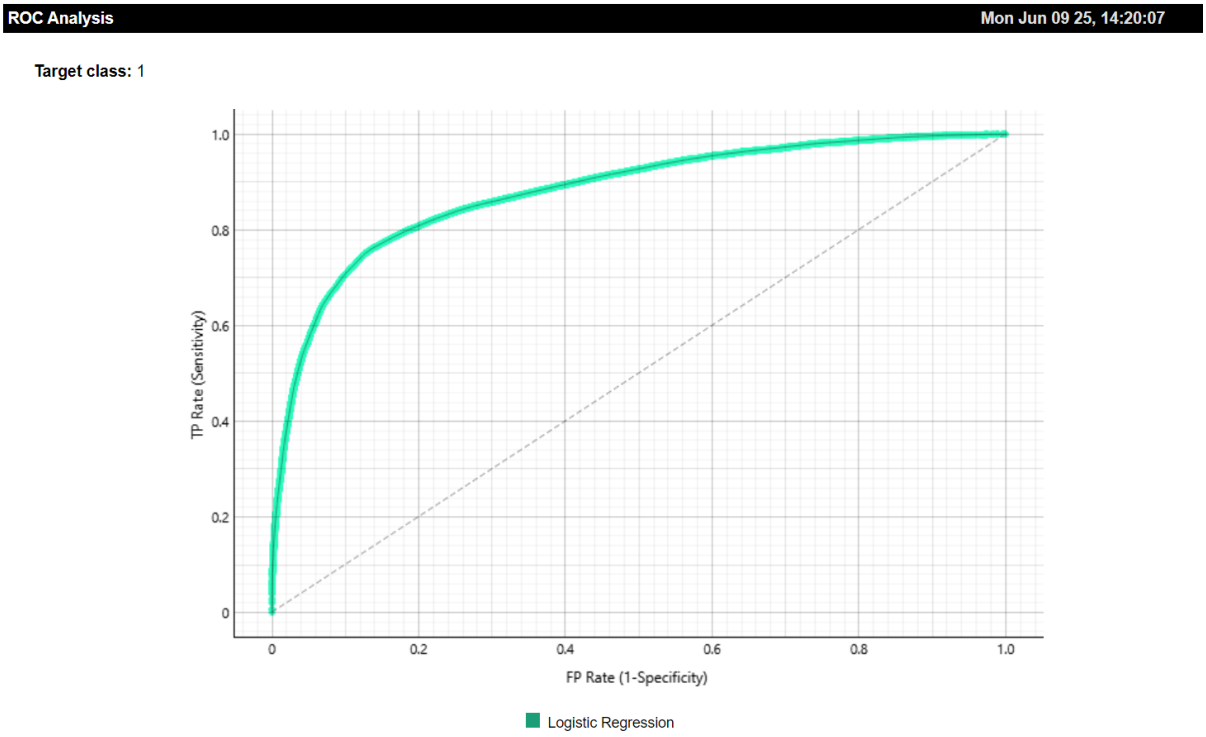
- True Negatives (TN): 96,617
    - The model correctly identified 96,617 instances where the loan did not default.
- False Positives (FP): 4,663
    - The model incorrectly predicted 4,663 instances as default when they were actually non-default. These are "Type I errors."
- False Negatives (FN): 12,580
    - The model incorrectly predicted 12,580 instances as non-default when they actually defaulted. These are "Type II errors" and are often more costly in a lending scenario (missed defaults).
- True Positives (TP): 15,780

- The model correctly identified 15,780 instances where the loan actually defaulted.

The confusion matrix confirms that while the model has high overall accuracy, it sacrifices some recall to achieve its precision. It is better at avoiding false alarms (FP) than at catching all actual defaults (FN). Given the context of loan default, missing a default (False Negative) is typically more financially detrimental than a false alarm (False Positive), suggesting there might be room to adjust the classification threshold to improve recall if desired.

6.3. ROC Curve Analysis

The ROC curve plots the True Positive Rate (Sensitivity/Recall) against the False Positive Rate (1-Specificity) at various threshold settings.



- The curve shows a sharp upward trend from the origin, staying well above the diagonal dashed line (which represents a random classifier). This visually reinforces the high AUC score of 0.879, indicating strong discriminative power.
- The model's ability to achieve a high True Positive Rate while keeping the False Positive Rate relatively low in the initial segments of the curve is evident. For example, it reaches a True Positive Rate of approximately 0.8 with a False Positive Rate below 0.2.
- The curve's shape suggests that there is a good trade-off between sensitivity and specificity available. The "elbow" or steepest part of the curve indicates the optimal threshold for balancing these two rates if the costs of false positives and false negatives are considered equal. However, for loan default, the cost of a false negative is often higher, which might prompt a shift in the operating point (threshold) on this curve towards higher recall.

6.4 Model Evaluation

The Logistic Regression model demonstrates strong predictive capabilities for identifying loan defaults. Its high AUC (0.879) indicates excellent discrimination between defaulting and non-defaulting loans. The overall accuracy of 86.7% is commendable.

However, the model exhibits a trade-off between precision (0.772) and recall (0.556). While it is quite good at correctly identifying actual defaults among its positive predictions (high precision), it misses a notable proportion (approximately 44.4%) of actual defaults (lower recall). Given that the consequence of a missed default (False Negative) is often significant financial loss for a lender, the current operating point might not be ideal depending on the business's risk tolerance.

Recommendations:

1. Threshold Adjustment: If minimizing missed defaults (reducing False Negatives) is a higher priority, consider lowering the classification threshold. This would increase recall, but likely at the cost of reduced precision (more False Positives). The ROC curve can guide this decision.
2. Cost-Benefit Analysis: Perform a detailed cost-benefit analysis where the financial implications of False Positives and False Negatives are explicitly quantified. This will inform the optimal threshold selection for deployment.
3. Further Optimization: While strong, explore if other modeling techniques (e.g., Gradient Boosting, Random Forests) or feature engineering could further improve recall without excessively compromising precision, or overall MCC.
4. Imbalance Handling: The number of defaults (28,360) is significantly less than non-defaults (101,280), indicating class imbalance. Although Logistic Regression can handle some imbalance, advanced techniques like oversampling the minority class, undersampling the majority class, or using algorithms designed for imbalanced data might further boost performance, especially recall.

In essence, the model is a robust tool for default prediction, providing valuable insights, but its operational utility can be further enhanced by fine-tuning its classification threshold based on specific business objectives and the asymmetrical costs of misclassifications.

7. Conclusion:

the developed Logistic Regression model demonstrates strong overall capability in predicting loan defaults, evidenced by its high AUC of 0.879 and 86.7% accuracy. While the model achieves good precision (77.2% of predicted defaults are correct), its recall of 55.6% indicates it misses a significant portion of actual defaults. This trade-off suggests the model is conservative, and strategic adjustments to the classification threshold, informed by a cost-benefit analysis of misclassifications, could further optimize its practical utility for risk management. The model provides valuable insights for segmenting borrowers based on key factors like loan grade, home ownership, income, and loan purpose, enabling tailored lending strategies.