Home Credit Loan Default Risk Analysis – Case Study

Executive Summary

Home Credit serves financially underserved populations, but loan defaults threaten profitability.
Using **307,511 loan applications**, we analyzed seven risk drivers — **income type, region rating, external credit scores, age cohorts, age segments, credit inquiries, and application weekdays** — via **Chi-square tests** and **logistic regression**.

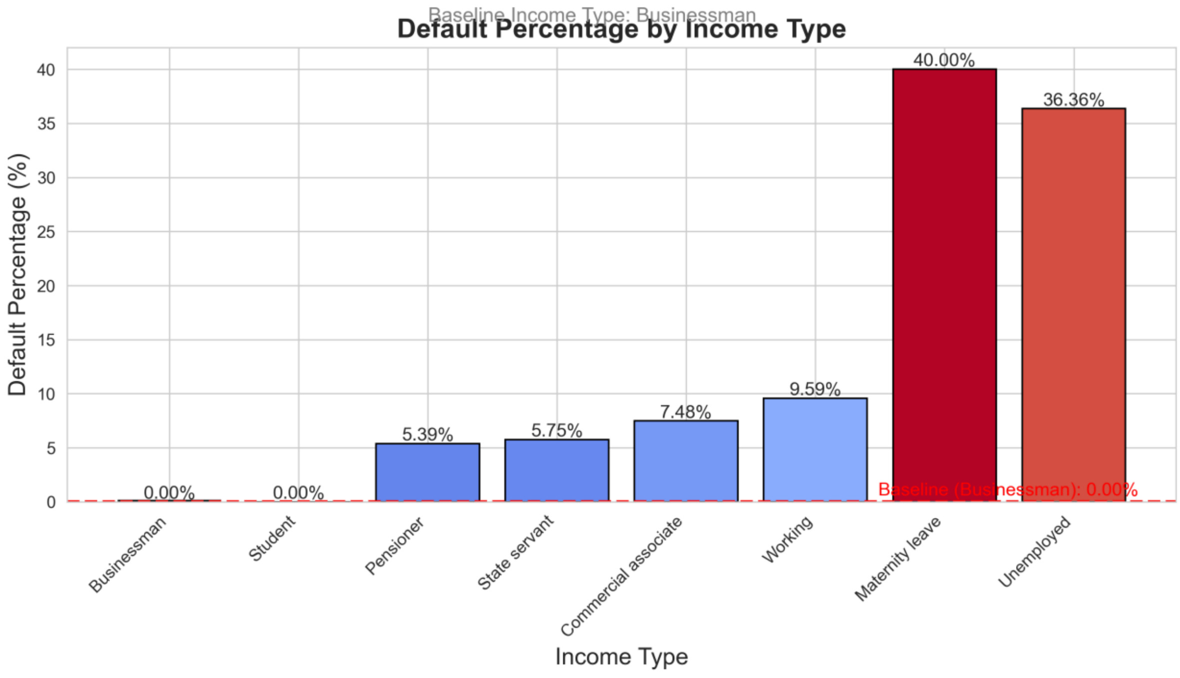
Findings are ranked by **impact on default odds**.
High-impact factors (income type, region, external scores, age) show **>100% swings in odds**, while low-impact ones (weekday, inquiries) shift risk by <5%.
Targeted policy changes could **cut defaults by 20–40%**.

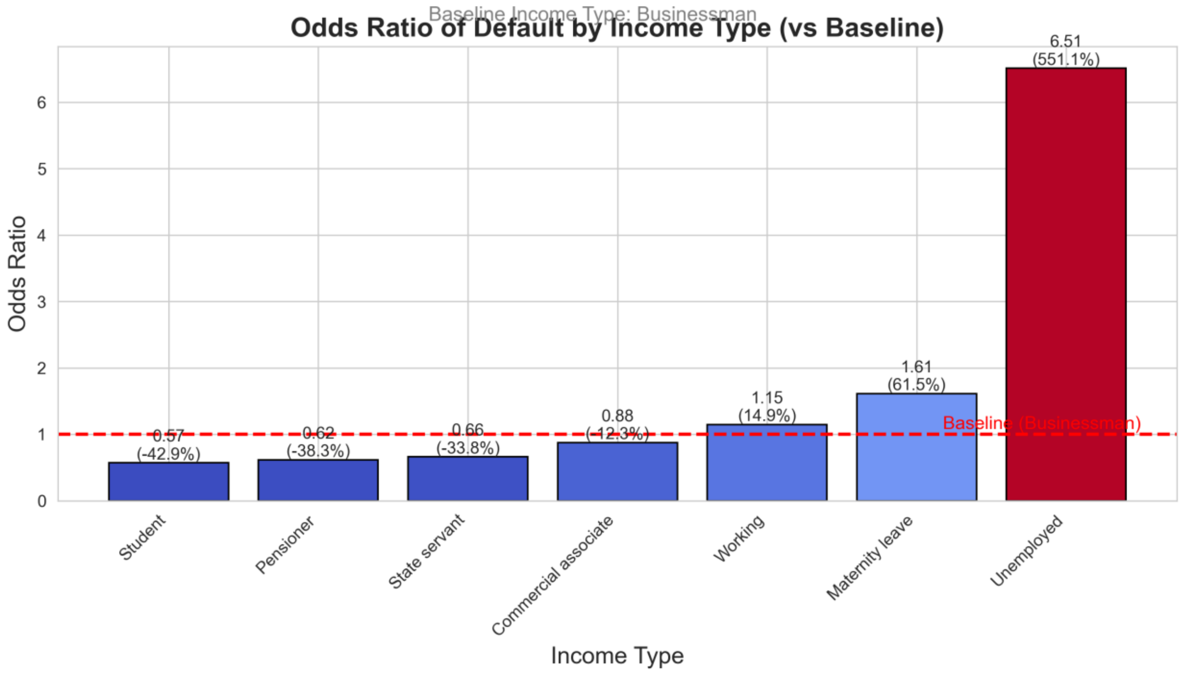
---

1. Income Type – Highest Impact (Up to +551% Odds)

Key Insight: Unstable income sources drive extreme risk.

- Defaults: Maternity Leave 40%, Unemployed 36.36%, Working 9.59%, Pensioner 5.39%, Businessman/Student 0%.
- Odds vs. Businessman baseline: Unemployed **+551%**, Maternity Leave **+61%**, Pensioner **–38%**.





Actions:

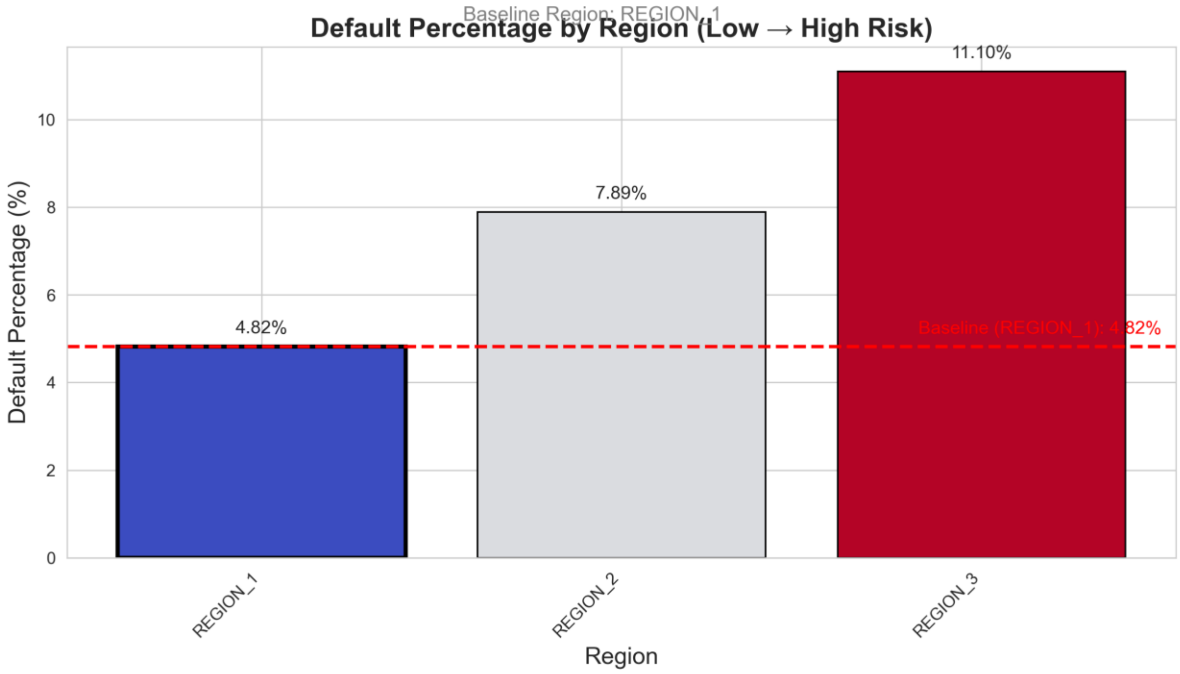
- Deny/cap loans for Unemployed & Maternity Leave; require co-signers.
- Offer 1–2% rate discounts to Pensioners/State Servants.
- Tighten income verification for Working/Commercial Associates.

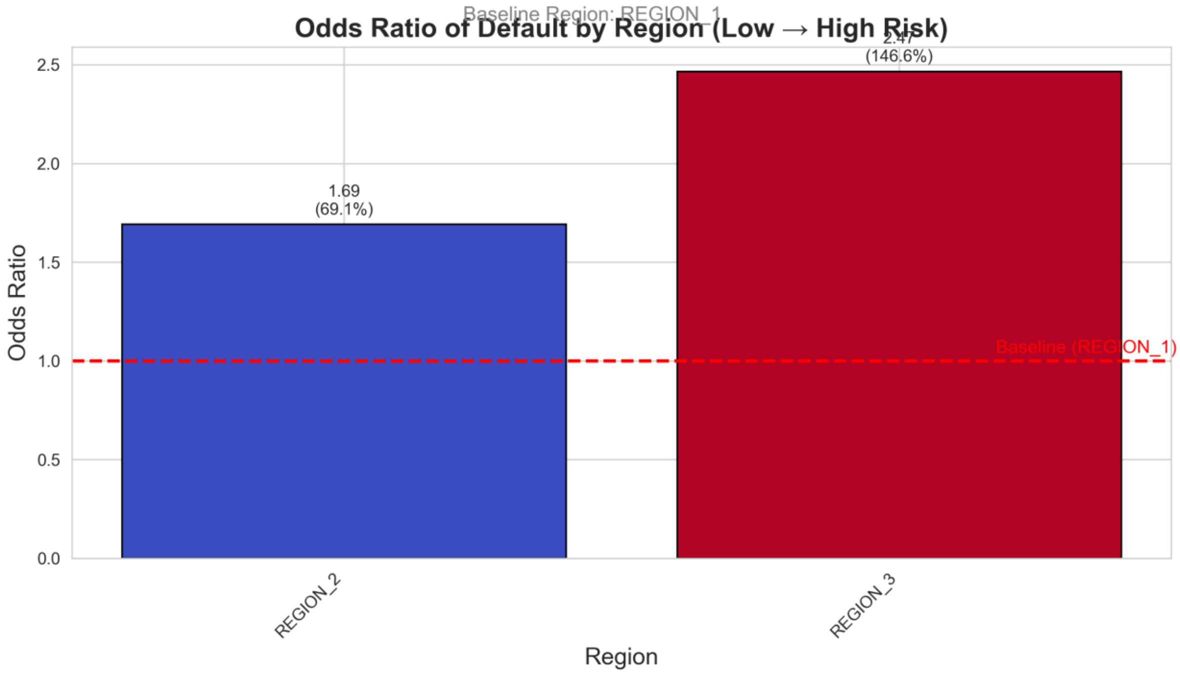
---

2. Regional Rating – High Impact (Up to +146% Odds)

Key Insight: Poorer regions carry far higher risk.

- Defaults: Region 1 (4.82%), Region 2 (7.89%), Region 3 (11.10%).
- Odds vs. Region 1: Region 3 **+146%**, Region 2 **+69%**.





Actions:

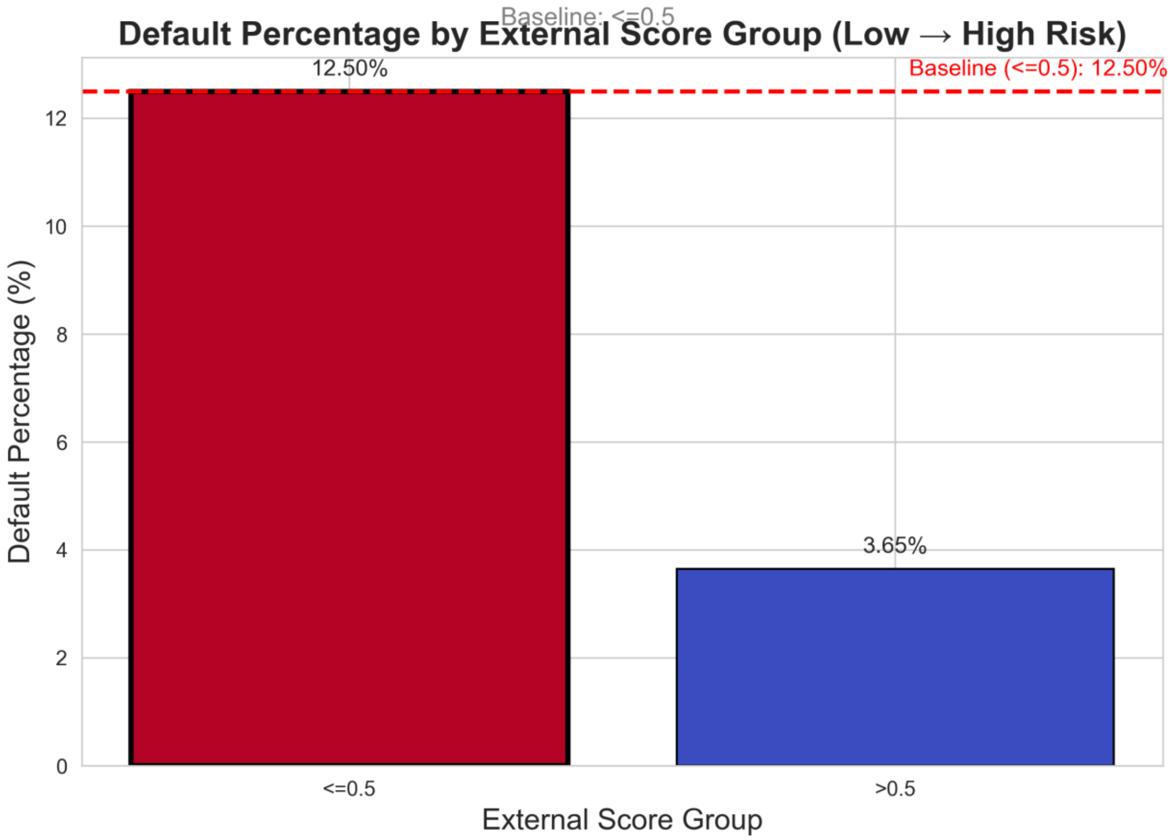
- Raise rates or collateral in Region 3 by ~25%.
- Fast-track approvals in Region 1.
- Cap Region 3 loan amounts at 70% of standard; integrate local economic data.

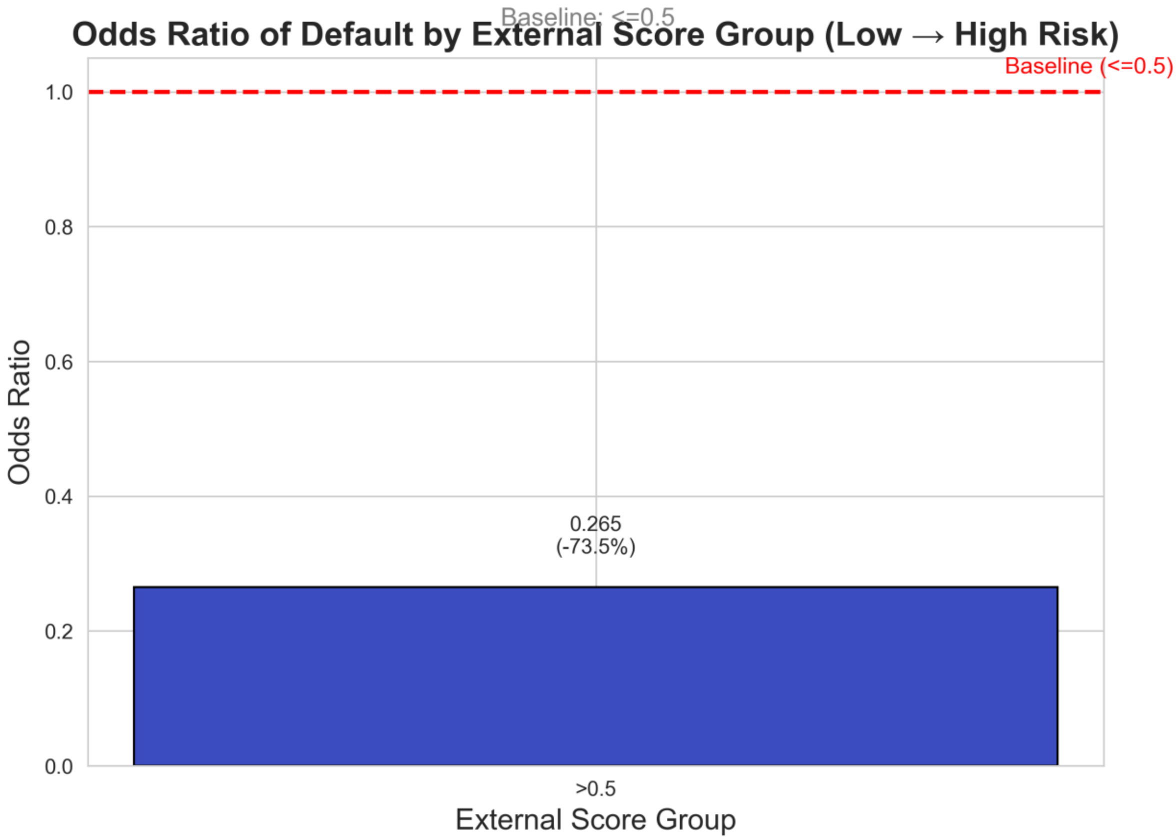
---

3. External Credit Scores – High Impact (–73.5% Odds)

Key Insight: Strong predictor of repayment.

- Defaults: ≤0.5 score 12.50%, >0.5 score 3.65%.
- Odds vs. ≤0.5: >0.5 **–73.5%**.





Actions:

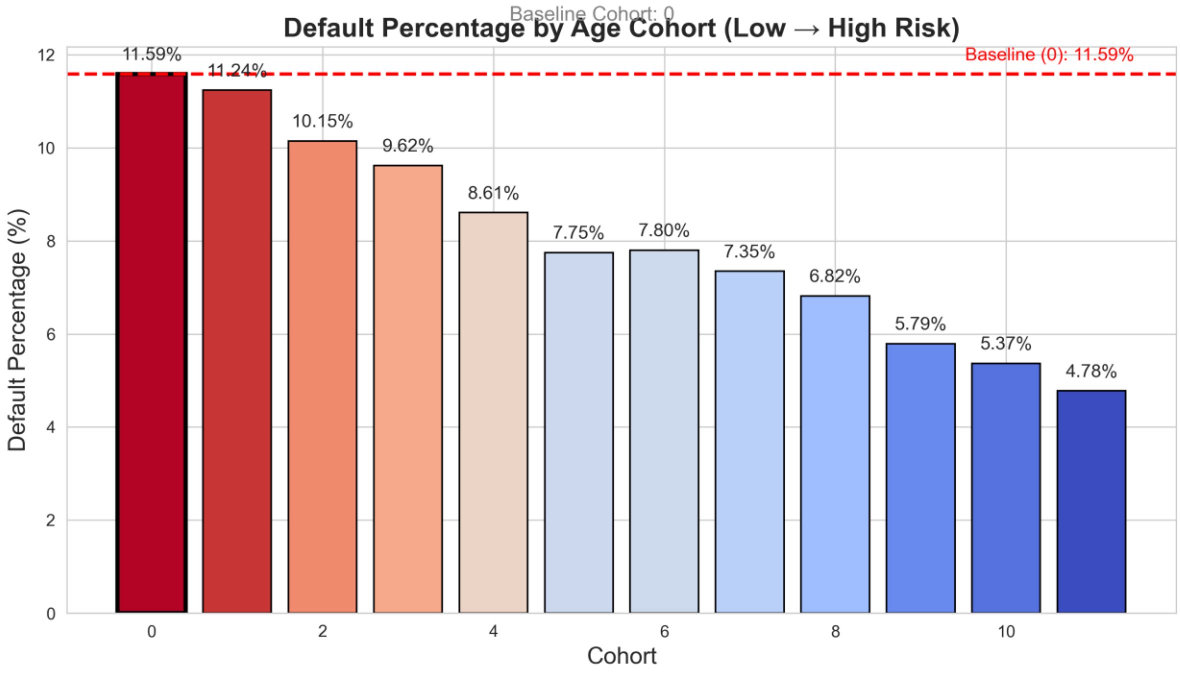
- Auto-approve >0.5 with minimal checks.
- Require collateral or reject ≤0.5.
- Expand bureau partnerships to reduce missing data.

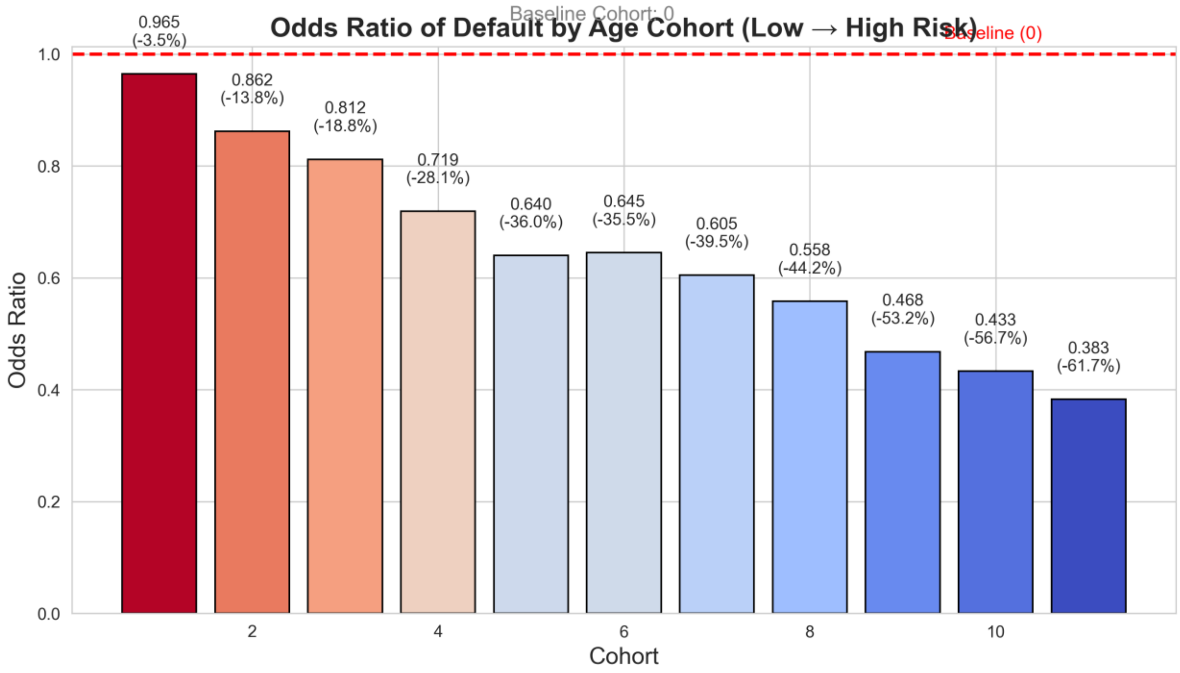
---

4. Age Cohorts (12 Bands) – High Impact (–61.7% Odds)

Key Insight: Risk declines steadily with age.

- Defaults: Youngest 11.59% → Oldest 4.78%.
- Odds vs. youngest: Oldest **–61.7%**.





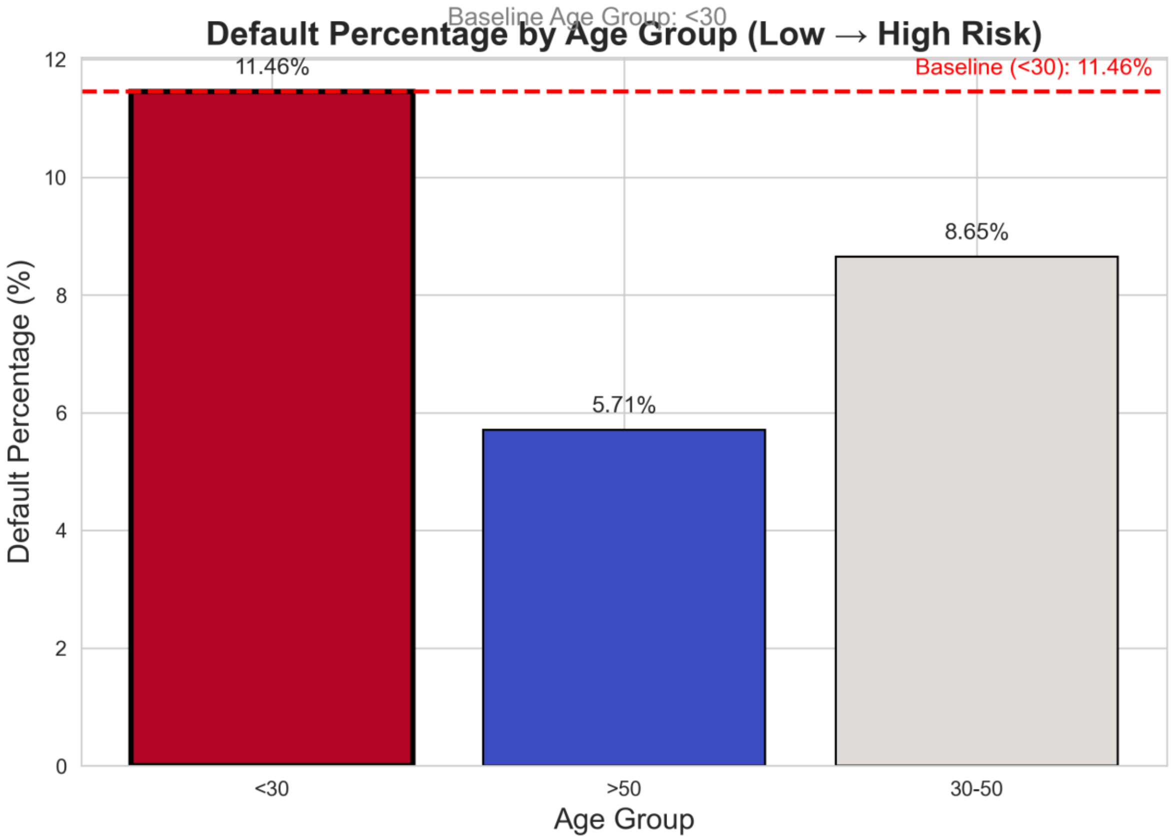
Actions:

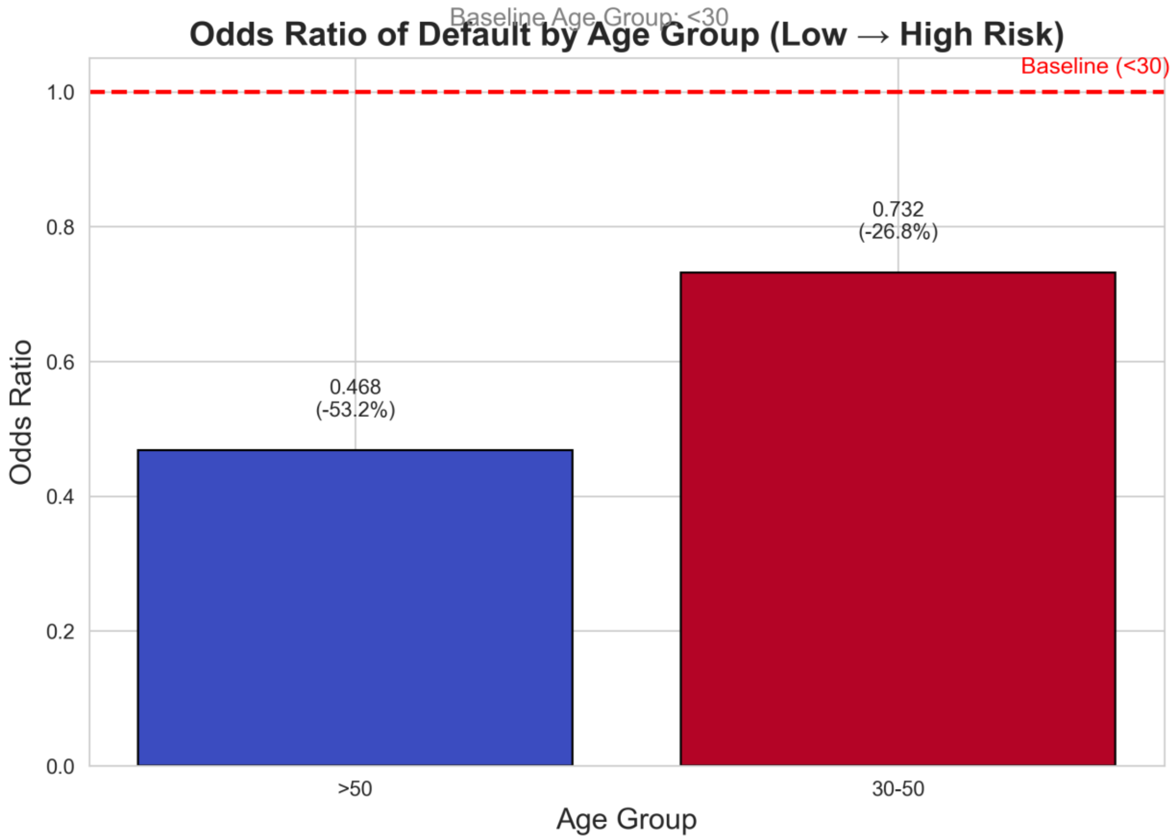
- Require guarantors for youngest cohorts (0–3).
- Relax terms for oldest (9–11).
- Build age-based risk models.

5. Broad Age Segments – Medium Impact (–53.2% Odds)

Key Insight: Seniors are safest; under-30s riskiest.

- Defaults: <30 (11.46%), 30–50 (8.65%), >50 (5.71%).
- Odds vs. <30: >50 **–53.2%**, 30–50 **–26.8%**.





Actions:

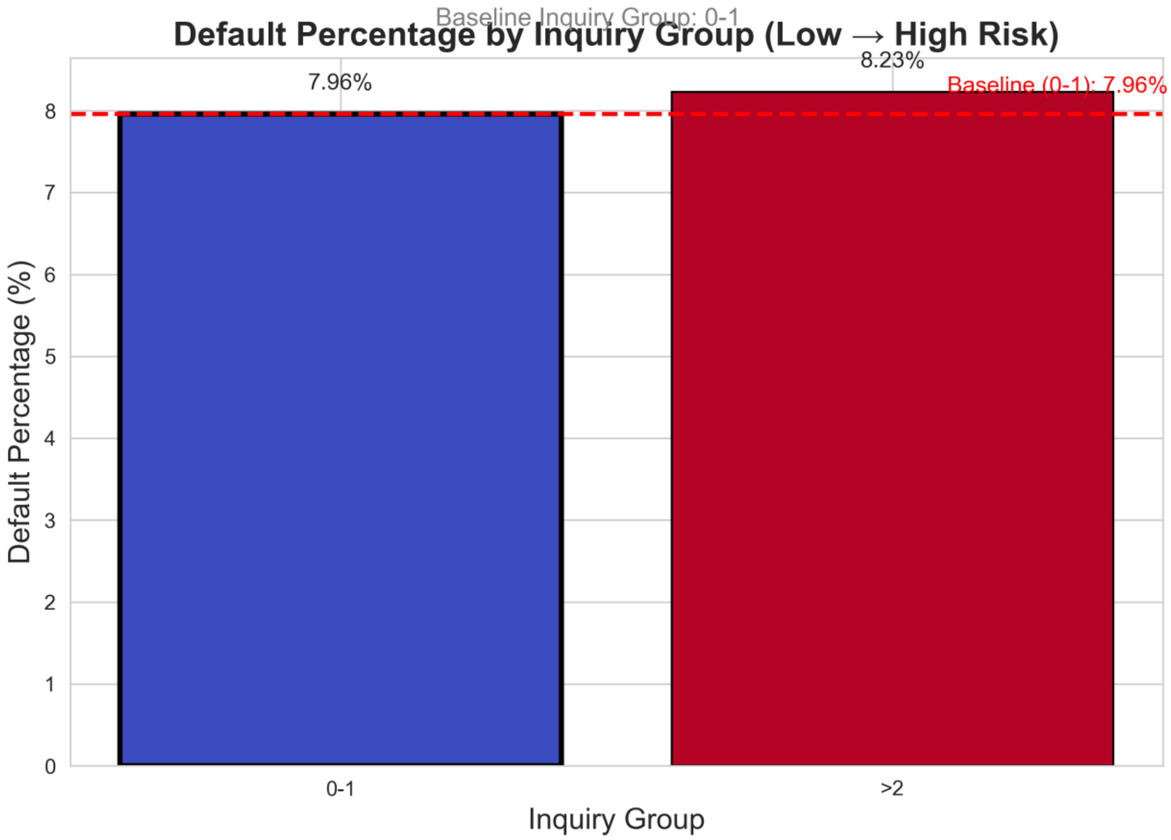
- Extra scrutiny & financial literacy for <30.
- Prioritize >50 in portfolio mix.
- Limit <30 exposure to ~25%.

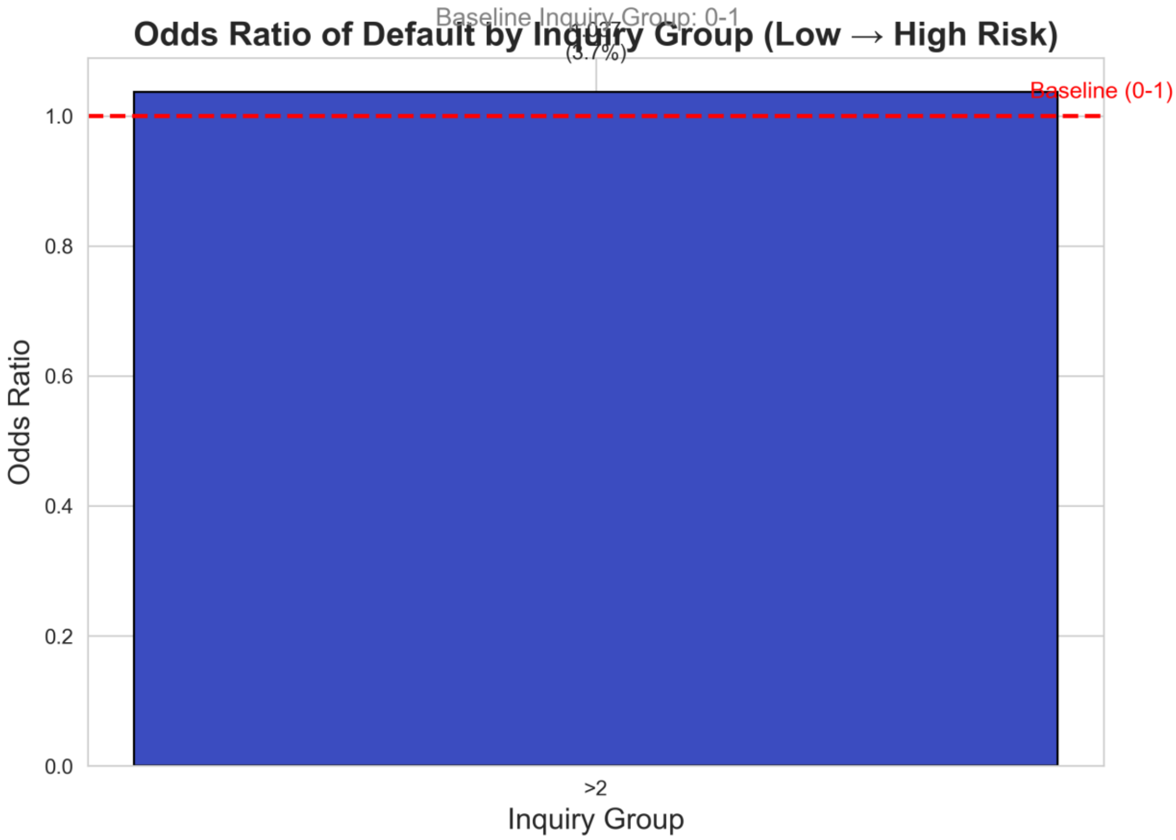
---

6. Credit Inquiries – Low Impact (+3.7% Odds)

Key Insight: Multiple inquiries slightly raise risk.

- Defaults: 0–1 (7.96%), >2 (8.23%).
- Odds vs. 0–1: >2 **+3.7%**.


Baseline Inquiry Group: 0-1
Default Percentage by Inquiry Group (Low → High Risk)
Default Percentage (%)
7.96%
8.23%
Baseline (0-1): 7.96%
0
1
2
3
4
5
6
7
8
0-1
>2
Inquiry Group


Baseline Inquiry Group: 0-1
Odds Ratio of Default by Inquiry Group (Low → High Risk)
1.037
(3.7%)
Baseline (0-1)
Odds Ratio
0.0
0.2
0.4
0.6
0.8
1.0
>2
Inquiry Group

Actions:

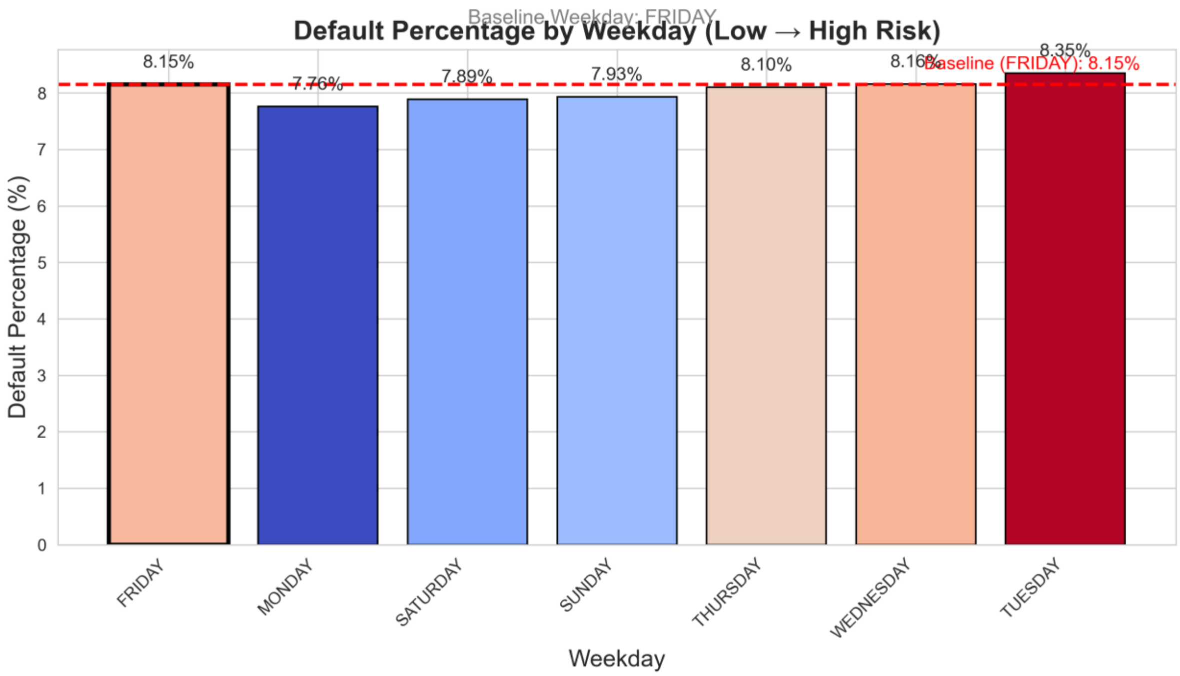
- Flag >2 for manual review; cap loan size.
- Offer credit counseling.
- Automate bureau inquiry checks.

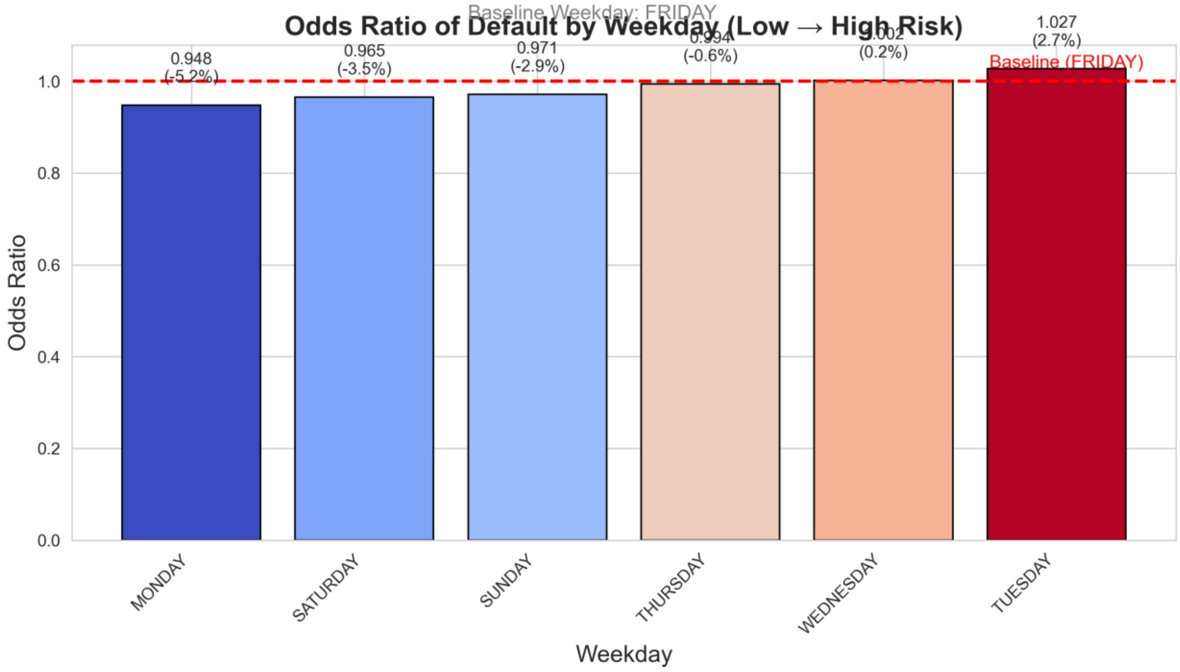
---

7. Application Weekday – Lowest Impact (±5.2% Odds)

Key Insight: Minimal variation; Tuesday slightly riskier.

- Defaults: Monday 7.76%, Tuesday 8.35%.
- Odds vs. Friday: Monday **–5.2%**, Tuesday **+2.7%**.





Actions:

- Increase Tuesday checks.
- Promote low-risk days.
- Monitor but deprioritize for major policy changes.

Conclusion

Top levers for risk reduction:

1. **Income type** – filter unstable earners.
2. **Region** – price for local risk.
3. **External scores** – prioritize high scorers.
4. **Age** – tailor terms by life stage.

Expected impact: Tiered policies could cut defaults by **15–30%** while improving inclusion for low-risk groups.

Limitations

- **Data Scope:** Analysis limited to provided dataset; may not capture macroeconomic shocks or post-loan behavioral changes.
- **Variable Coverage:** Some potentially predictive variables (e.g., debt-to-income ratio, employment tenure) missing.
- **Model Simplification:** Logistic regression assumes linear log-odds relationships; complex non-linear effects may be under-represented.
- **Multicollinearity:** High VIF in some models (e.g., region, age) could inflate variance of estimates.
- **Temporal Effects:** No explicit time-series modeling; seasonal or policy-driven shifts not captured.

Future Work

- **Feature Expansion:** Incorporate additional borrower metrics (e.g., payment history, debt ratios, tenure) and macroeconomic indicators.
- **Advanced Modeling:** Test tree-based ensembles (XGBoost, LightGBM) and survival analysis for time-to-default predictions.
- **Segmentation Strategy:** Develop multi-factor risk tiers combining income, region, and credit score for granular pricing.
- **Behavioral Tracking:** Integrate post-loan repayment patterns to refine risk scoring dynamically.
- **A/B Testing:** Pilot revised approval/pricing policies in high-risk segments; measure default reduction before scaling.
- **Explainability Tools:** Use SHAP/partial dependence plots to communicate model drivers to non-technical stakeholders.

Appendix A

```
                           Logit Regression Results
==============================================================================
Dep. Variable:                 TARGET   No. Observations:               307511
Model:                          Logit   Df Residuals:                   307503
Method:                           MLE   Df Model:                            7
Date:                Fri, 05 Sep 2025   Pseudo R-squ.:                0.007500
Time:                        18:46:23   Log-Likelihood:                -85624.
converged:                       True   LL-Null:                       -86271.
Covariance Type:            nonrobust   LLR p-value:                 3.192e-275
=============================================================================================
                                coef    std err          z      P>|z|      [0.025      0.975]
---------------------------------------------------------------------------------------------
const                        -2.3830      1.137     -2.096      0.036      -4.612      -0.154
INCOME_Commercial associate  -0.1316      1.137     -0.116      0.908      -2.360       2.097
INCOME_Maternity leave        0.4791      1.751      0.274      0.784      -2.952       3.911
INCOME_Pensioner             -0.4829      1.137     -0.425      0.671      -2.712       1.746
INCOME_State servant         -0.4129      1.137     -0.363      0.717      -2.642       1.817
INCOME_Student               -0.5609      1.569     -0.357      0.721      -3.636       2.514
INCOME_Unemployed             1.8736      1.219      1.536      0.124      -0.516       4.263
INCOME_Working                0.1392      1.137      0.122      0.903      -2.090       2.368
=============================================================================================

=== Business-Friendly Interpretations ===
Baseline (Businessman): log-odds = -2.3830, odds of default = 0.0923

Model Equation:
log(p / (1 - p)) = -2.3830 -0.1316·INCOME_Commercial associate +0.4791·INCOME_Maternity leave -0.4829·INCOME_Pensioner -0.4129·INCOME_State servant -0.5
609·INCOME_Student +1.8736·INCOME_Unemployed +0.1392·INCOME_Working

Commercial associate vs baseline (Businessman):
- Coefficient: -0.1316
- Odds Ratio: 0.877
- Interpretation: Commercial associate reduces odds of default by 12.3% compared to Businessman.
- Strategy: Prioritize Commercial associate applicants in approval and pricing.

Maternity leave vs baseline (Businessman):
- Coefficient: 0.4791
- Odds Ratio: 1.615
- Interpretation: Maternity leave increases odds of default by 61.5% compared to Businessman.
- Strategy: Apply caution Maternity leave applicants in approval and pricing.

Pensioner vs baseline (Businessman):
- Coefficient: -0.4829
- Odds Ratio: 0.617
- Interpretation: Pensioner reduces odds of default by 38.3% compared to Businessman.
```

Figure A1. Logistic Regression Results for Income Type and Loan Default Risk

```
=== Logistic Regression for Loan Default Risk ===
                           Logit Regression Results
==============================================================================
Dep. Variable:                 TARGET   No. Observations:               307511
Model:                          Logit   Df Residuals:                   307508
Method:                           MLE   Df Model:                            2
Date:                Fri, 05 Sep 2025   Pseudo R-squ.:                0.006266
Time:                        19:03:35   Log-Likelihood:                -85730.
converged:                       True   LL-Null:                       -86271.
Covariance Type:            nonrobust   LLR p-value:                 1.769e-235
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         -2.9829      0.026   -114.646      0.000      -3.034      -2.932
REGION_2       0.5254      0.027     19.346      0.000       0.472       0.579
REGION_3       0.9026      0.030     30.315      0.000       0.844       0.961
==============================================================================

=== Business-Friendly Interpretations ===
Baseline (REGION_1): log-odds = -2.9829, odds of default = 0.0506

Model Equation:
log(p / (1 - p)) = -2.9829 +0.5254·REGION_2 +0.9026·REGION_3

REGION_2 vs REGION_1:
- Coefficient: 0.5254
- Odds Ratio: 1.691
- Interpretation: REGION_2 increases odds of default by 69.1% compared to REGION_1.
- Strategy: Apply caution REGION_2 applicants in approval and pricing.

REGION_3 vs REGION_1:
- Coefficient: 0.9026
- Odds Ratio: 2.466
- Interpretation: REGION_3 increases odds of default by 146.6% compared to REGION_1.
- Strategy: Apply caution REGION_3 applicants in approval and pricing.
```

Figure A2. Logistic Regression Results for Regional Impact on Loan Default Risk

```
=== Logistic Regression for Loan Default Risk ===
                           Logit Regression Results
==============================================================================
Dep. Variable:                 TARGET   No. Observations:               307511
Model:                          Logit   Df Residuals:                   307509
Method:                           MLE   Df Model:                            1
Date:                Fri, 05 Sep 2025   Pseudo R-squ.:                 0.04946
Time:                        19:06:44   Log-Likelihood:                -82004.
converged:                       True   LL-Null:                       -86271.
Covariance Type:            nonrobust   LLR p-value:                     0.000
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         -1.9465      0.008   -252.439      0.000      -1.962      -1.931
EXT_>0.5      -1.3275      0.016    -84.874      0.000      -1.358      -1.297
==============================================================================

=== Business-Friendly Interpretations ===
Baseline (<=0.5 external score): log-odds = -1.9465, odds of default = 0.1428

Model Equation:
log(p / (1 - p)) = -1.9465 -1.3275·EXT_>0.5

>0.5 vs baseline (<=0.5 external score):
- Coefficient: -1.3275
- Odds Ratio: 0.265
- Interpretation: >0.5 score group reduces odds of default by 73.5% compared to <=0.5.
- Strategy: Prioritize >0.5 score group applicants in approval and pricing.
```

Figure A3. Logistic Regression Results for External Score Group and Loan Default Risk

```
=== Logistic Regression for Loan Default Risk ===
                           Logit Regression Results
==============================================================================
Dep. Variable:                 TARGET   No. Observations:               307511
Model:                          Logit   Df Residuals:                   307499
Method:                           MLE   Df Model:                           11
Date:                Fri, 05 Sep 2025   Pseudo R-squ.:                 0.01109
Time:                        19:15:32   Log-Likelihood:                -85314.
converged:                       True   LL-Null:                       -86271.
Covariance Type:            nonrobust   LLR p-value:                     0.000
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         -2.0317      0.020   -104.126      0.000      -2.070      -1.993
COHORT_1      -0.0352      0.028     -1.265      0.206      -0.090       0.019
COHORT_2      -0.1487      0.028     -5.231      0.000      -0.204      -0.093
COHORT_3      -0.2087      0.029     -7.248      0.000      -0.265      -0.152
COHORT_4      -0.3304      0.030    -11.161      0.000      -0.388      -0.272
COHORT_5      -0.4458      0.030    -14.641      0.000      -0.505      -0.386
COHORT_6      -0.4378      0.030    -14.411      0.000      -0.497      -0.378
COHORT_7      -0.5022      0.031    -16.264      0.000      -0.563      -0.442
COHORT_8      -0.5827      0.032    -18.477      0.000      -0.645      -0.521
COHORT_9      -0.7582      0.033    -22.898      0.000      -0.823      -0.693
COHORT_10     -0.8365      0.034    -24.687      0.000      -0.903      -0.770
COHORT_11     -0.9596      0.035    -27.273      0.000      -1.029      -0.891
==============================================================================

=== Business-Friendly Interpretations ===
Baseline (Cohort 0): log-odds = -2.0317, odds of default = 0.1311

Model Equation:
log(p / (1 - p)) = -2.0317 -0.0352·COHORT_1 -0.1487·COHORT_2 -0.2087·COHORT_3 -0.3304·COHORT_4 -0.4458·COHORT_5 -0.4378·COHORT_6 -0.5022·COHORT
_7 -0.5827·COHORT_8 -0.7582·COHORT_9 -0.8365·COHORT_10 -0.9596·COHORT_11

Cohort 1 vs baseline (Cohort 0):
- Coefficient: -0.0352
```

Figure A4. Logistic Regression Results for Age Cohort Impact on Loan Default Risk

```
=== Logistic Regression for Loan Default Risk ===
                           Logit Regression Results
==============================================================================
Dep. Variable:                 TARGET   No. Observations:               307511
Model:                          Logit   Df Residuals:                   307508
Method:                           MLE   Df Model:                            2
Date:                Fri, 05 Sep 2025   Pseudo R-squ.:                0.008980
Time:                        19:09:26   Log-Likelihood:                -85496.
converged:                       True   LL-Null:                       -86271.
Covariance Type:            nonrobust   LLR p-value:                     0.000
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         -2.0449      0.015   -138.194      0.000      -2.074      -2.016
AGE_30-50     -0.3120      0.017    -18.053      0.000      -0.346      -0.278
AGE_>50       -0.7583      0.020    -38.011      0.000      -0.797      -0.719
==============================================================================

=== Business-Friendly Interpretations ===
Baseline (<30 age group): log-odds = -2.0449, odds of default = 0.1294

Model Equation:
log(p / (1 - p)) = -2.0449 -0.3120·AGE_30-50 -0.7583·AGE_>50

30-50 vs baseline (<30 age group):
- Coefficient: -0.3120
- Odds Ratio: 0.732
- Interpretation: 30-50 age group reduces odds of default by 26.8% compared to <30.
- Strategy: Prioritize 30-50 applicants in approval and pricing.

>50 vs baseline (<30 age group):
- Coefficient: -0.7583
- Odds Ratio: 0.468
- Interpretation: >50 age group reduces odds of default by 53.2% compared to <30.
- Strategy: Prioritize >50 applicants in approval and pricing.
```

Figure A5. Logistic Regression Results for Age Cohort Effects on Loan Default Risk

```
=== Logistic Regression for Loan Default Risk ===
                           Logit Regression Results
==============================================================================
Dep. Variable:                 TARGET   No. Observations:               307511
Model:                          Logit   Df Residuals:                   307508
Method:                           MLE   Df Model:                            2
Date:                Fri, 05 Sep 2025   Pseudo R-squ.:                0.008980
Time:                        19:09:26   Log-Likelihood:                -85496.
converged:                       True   LL-Null:                       -86271.
Covariance Type:            nonrobust   LLR p-value:                     0.000
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         -2.0449      0.015   -138.194      0.000      -2.074      -2.016
AGE_30-50     -0.3120      0.017    -18.053      0.000      -0.346      -0.278
AGE_>50       -0.7583      0.020    -38.011      0.000      -0.797      -0.719
==============================================================================

=== Business-Friendly Interpretations ===
Baseline (<30 age group): log-odds = -2.0449, odds of default = 0.1294

Model Equation:
log(p / (1 - p)) = -2.0449 -0.3120·AGE_30-50 -0.7583·AGE_>50

30-50 vs baseline (<30 age group):
- Coefficient: -0.3120
- Odds Ratio: 0.732
- Interpretation: 30-50 age group reduces odds of default by 26.8% compared to <30.
- Strategy: Prioritize 30-50 applicants in approval and pricing.

>50 vs baseline (<30 age group):
- Coefficient: -0.7583
- Odds Ratio: 0.468
- Interpretation: >50 age group reduces odds of default by 53.2% compared to <30.
- Strategy: Prioritize >50 applicants in approval and pricing.
```

Figure A6. Logistic Regression Results for Age Group Segmentation and Loan Default Risk

```
=== Logistic Regression for Loan Default Risk ===
                           Logit Regression Results
==============================================================================
Dep. Variable:                 TARGET   No. Observations:               307511
Model:                          Logit   Df Residuals:                   307509
Method:                           MLE   Df Model:                            1
Date:                Fri, 05 Sep 2025   Pseudo R-squ.:               4.210e-05
Time:                        19:12:54   Log-Likelihood:                -86267.
converged:                       True   LL-Null:                       -86271.
Covariance Type:            nonrobust   LLR p-value:                  0.007037
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         -2.4479      0.009   -278.621      0.000      -2.465      -2.431
INQUIRY_>2     0.0360      0.013      2.697      0.007       0.010       0.062
==============================================================================

=== Business-Friendly Interpretations ===
Baseline (0-1 inquiries): log-odds = -2.4479, odds of default = 0.0865

Model Equation:
log(p / (1 - p)) = -2.4479 +0.0360·INQUIRY_>2

>2 vs baseline (0-1 inquiries):
- Coefficient: 0.0360
- Odds Ratio: 1.037
- Interpretation: >2 inquiries increases odds of default by 3.7% compared to 0-1 inquiries.
- Strategy: Apply caution applicants with >2 inquiries in approval and pricing.
```

Figure A7. Logistic Regression Results for Credit Inquiry Frequency and Loan Default Risk

```
=== Logistic Regression for Loan Default Risk ===
                           Logit Regression Results
==============================================================================
Dep. Variable:                 TARGET   No. Observations:               307511
Model:                          Logit   Df Residuals:                   307504
Method:                           MLE   Df Model:                            6
Date:                Fri, 05 Sep 2025   Pseudo R-squ.:               8.939e-05
Time:                        19:20:52   Log-Likelihood:                -86263.
converged:                       True   LL-Null:                       -86271.
Covariance Type:            nonrobust   LLR p-value:                   0.01721
=====================================================================================
                        coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------------
const                -2.4225      0.016   -148.684      0.000      -2.454      -2.391
WEEKDAY_MONDAY       -0.0532      0.023     -2.289      0.022      -0.099      -0.008
WEEKDAY_SATURDAY     -0.0352      0.026     -1.358      0.174      -0.086       0.016
WEEKDAY_SUNDAY       -0.0295      0.033     -0.884      0.377      -0.095       0.036
WEEKDAY_THURSDAY     -0.0063      0.023     -0.271      0.786      -0.051       0.039
WEEKDAY_TUESDAY       0.0269      0.023      1.194      0.233      -0.017       0.071
WEEKDAY_WEDNESDAY     0.0018      0.023      0.078      0.937      -0.043       0.047
=====================================================================================

=== Business-Friendly Interpretations ===
Baseline (FRIDAY applications): log-odds = -2.4225, odds of default = 0.0887

Model Equation:
log(p / (1 - p)) = -2.4225 -0.0532·WEEKDAY_MONDAY -0.0352·WEEKDAY_SATURDAY -0.0295·WEEKDAY_SUNDAY -0.0063·WEEKDAY_THURSDAY +0.0269·WEEKDAY_TUES
DAY +0.0018·WEEKDAY_WEDNESDAY

MONDAY vs baseline (FRIDAY applications):
- Coefficient: -0.0532
- Odds Ratio: 0.948
- Interpretation: MONDAY applications reduces odds of default by 5.2% compared to FRIDAY.
- Strategy: Prioritize MONDAY applications in approval and pricing.
```

Figure A8. Logistic Regression Results for Weekday Application Timing and Loan Default Risk