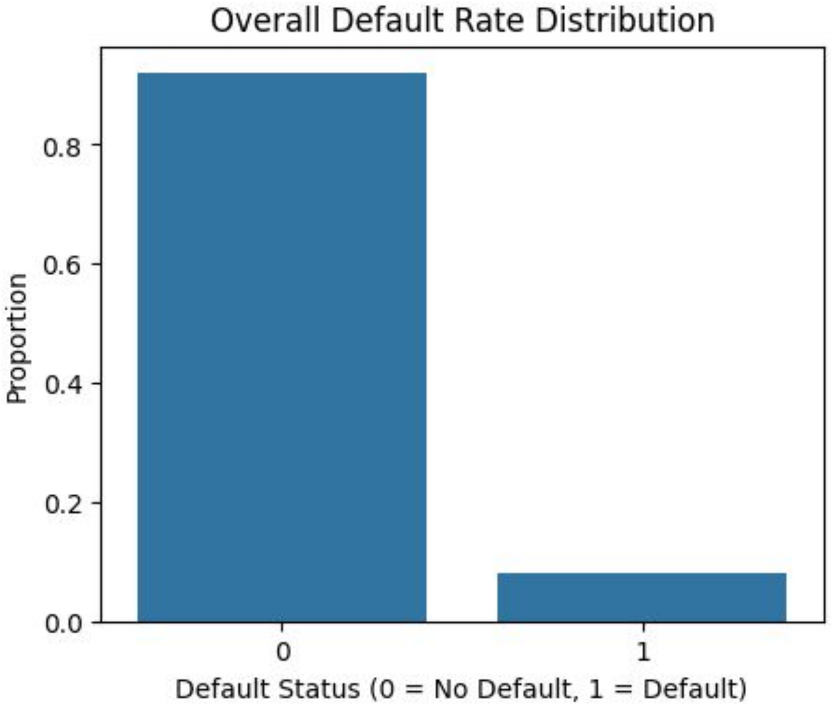
# EDA Findings

Jasmine Zhang, Ishaan Banerjee, Zien Zhu

# Overview and Baseline

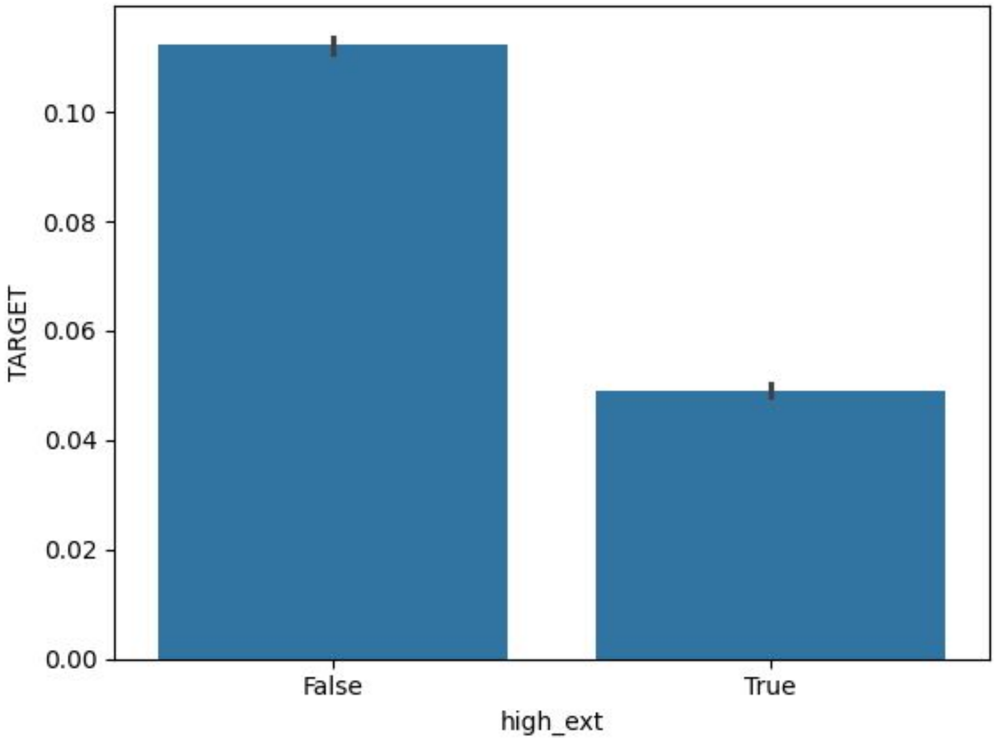


The overall default rate is approximately 8%, indicating a highly imbalanced dataset where most applicants do not default.

This imbalance is typical in credit risk datasets and highlights the importance of identifying strong predictive variables that can distinguish the relatively small group of likely defaulters from the majority of safe applicants.

# External Credit Indicators

# External Credit Score (EXT_SOURCE_2)

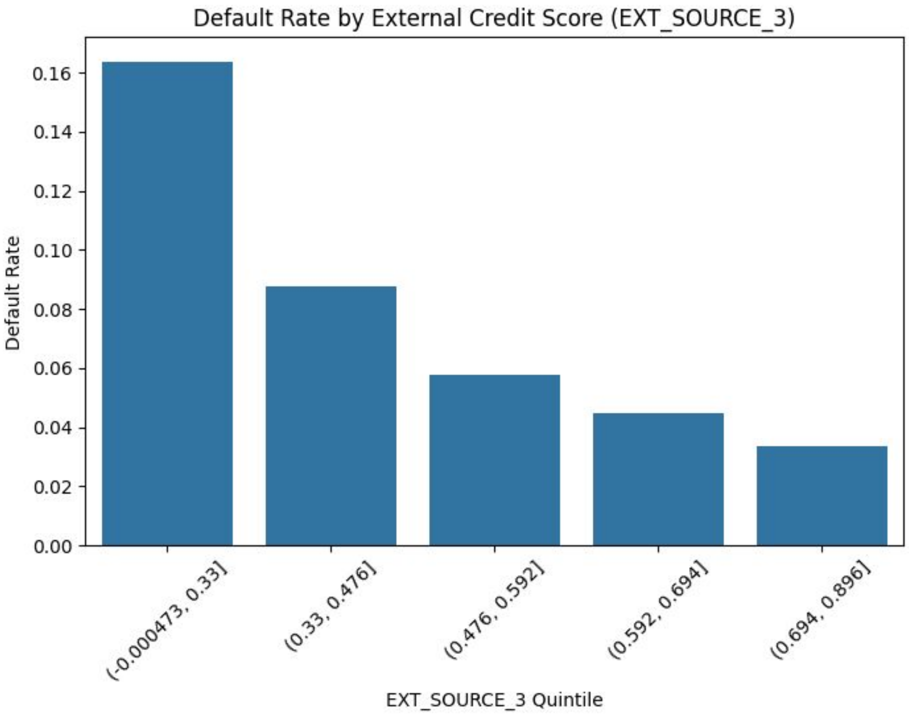


Strong relationship between the external credit score and default probability.

Applicants with below-median external credit scores have a default rate of 11.2%, compared to above-median credit scores of only 4.9%.

This is more than double increase in default risk, so external credit score the strongest predictor identified in my analysis.

# External Credit Score (EXT_SOURCE_3)



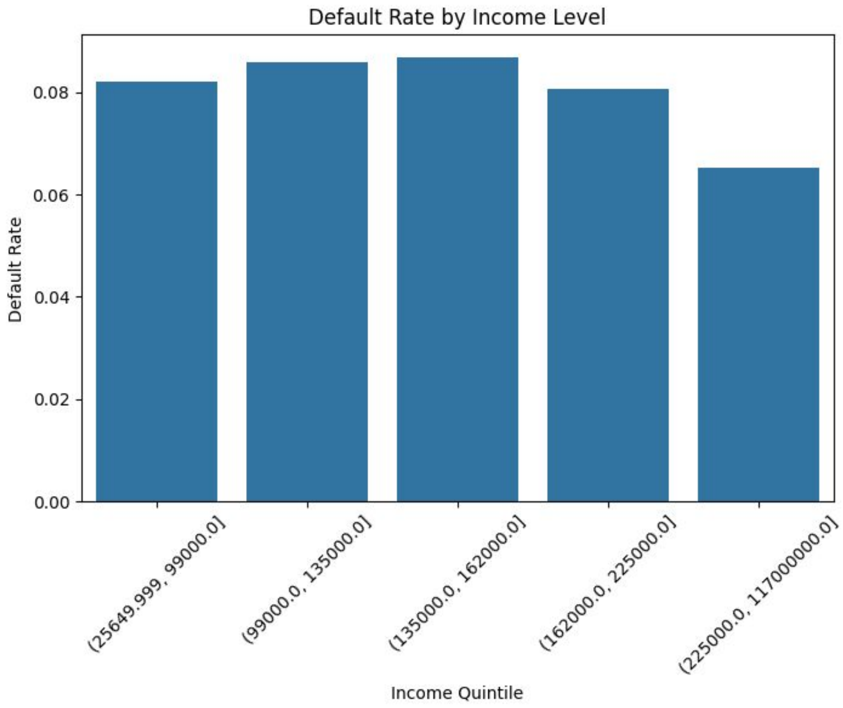
A strong and nearly monotonic decreasing relationship is observed between EXT_SOURCE_3 and default rate.

Applicants with lower external credit scores exhibit significantly higher default probabilities.

This variable demonstrates one of the strongest predictive signals in the dataset and is highly valuable for identifying high-risk applicants during loan approval.

# Applicant Financial Profile

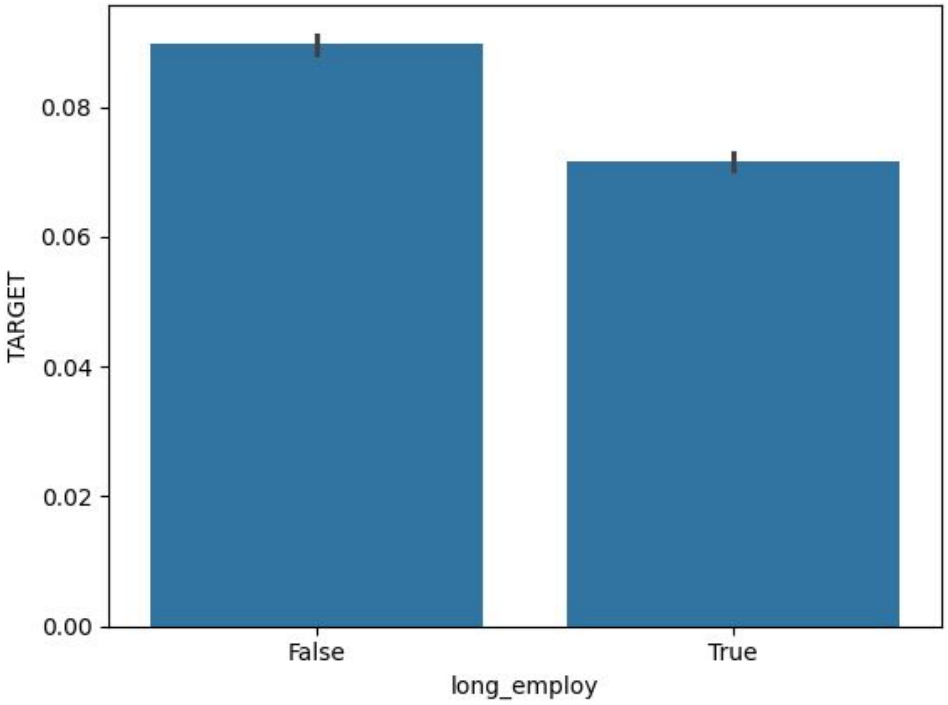
# Income Level (AMT_INCOME_TOTAL)



Higher income groups generally exhibit lower default rates, indicating that stronger financial capacity improves repayment ability.

Although the relationship is not perfectly monotonic, income remains an important factor for distinguishing lower-risk applicants from higher-risk ones.

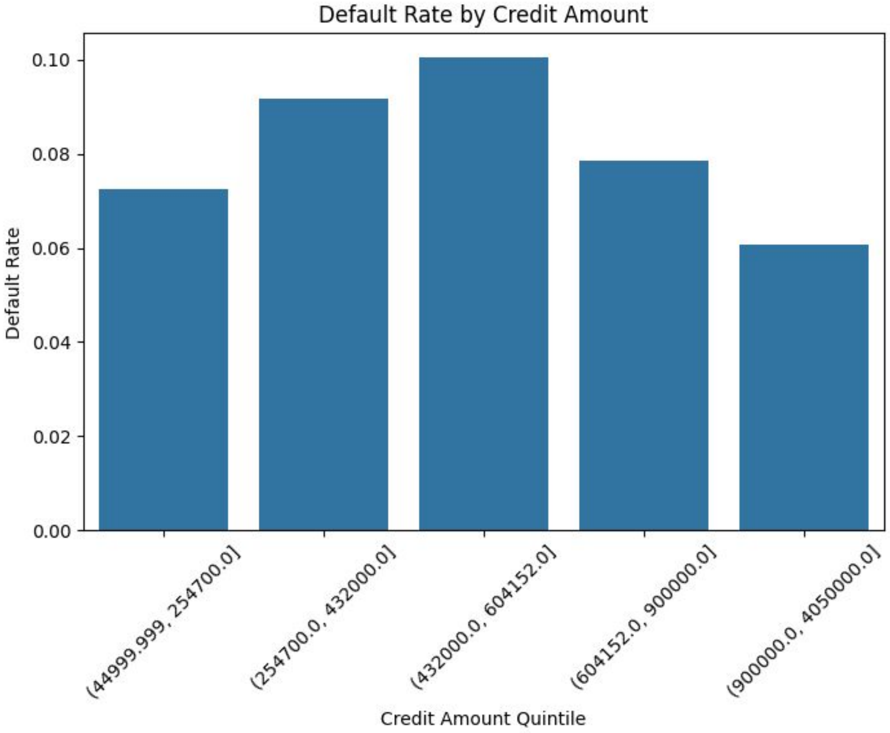
# Income Type (NAME_INCOME_TYPE)



For large groups, Working hsa a 9.6% default rate, Commercial associate has a 7.5% default rate, State servant has a 5.8% default rate, Pensioner has a 5.4% default rate.

Extremely high default rate for Maternity and Unemployed, but so few data entries that just a few people defaulting artificially skews the data very high, so this is not statistically significant. These groups are too small to be reliable.

# Credit Amount (AMT_CREDIT)

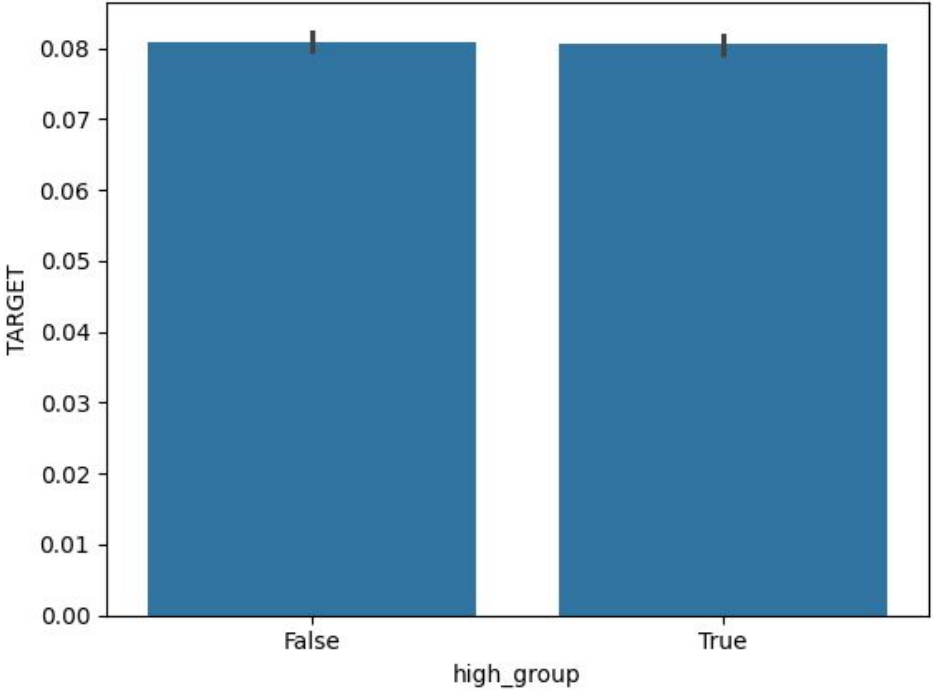


The relationship between credit amount and default risk is moderate.

Applicants requesting extremely large credit amounts tend to show slightly higher default rates, likely due to increased repayment burden.

Compared to external credit score and income, credit amount alone is a weaker predictor, but it still contributes useful information when combined with other financial indicators.

# Loan Size (Loan-to-Income Ratio)

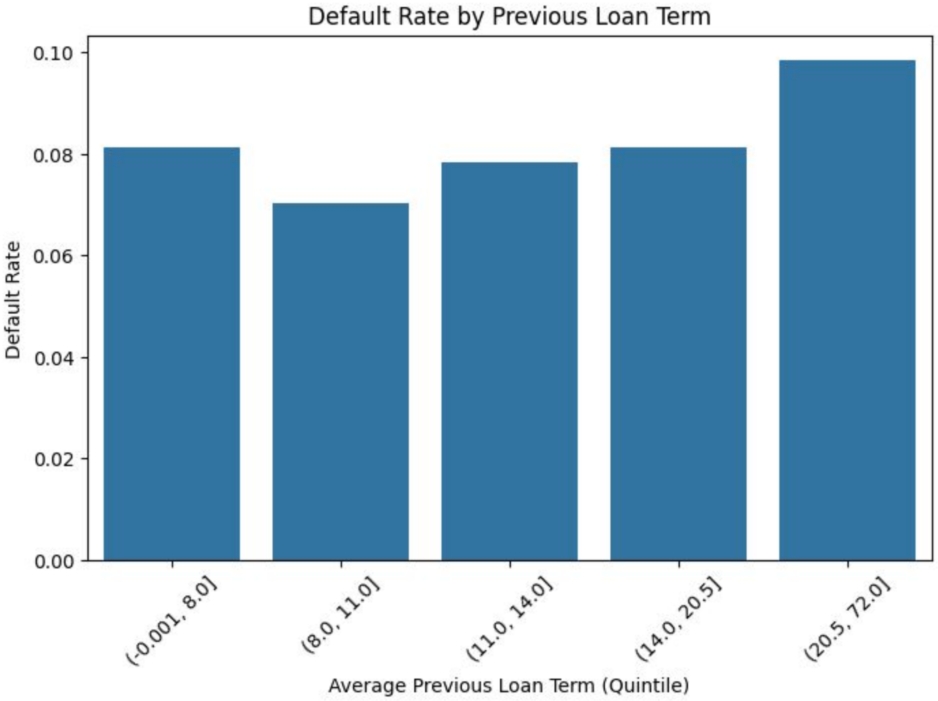


Borrowers split into two groups by median

Difference was negligible, suggests loan size relative to portion of borrower's income does not determine default risk.

# Historical Loan Behavior

# Previous Loan Term (CNT_PAYMENT)

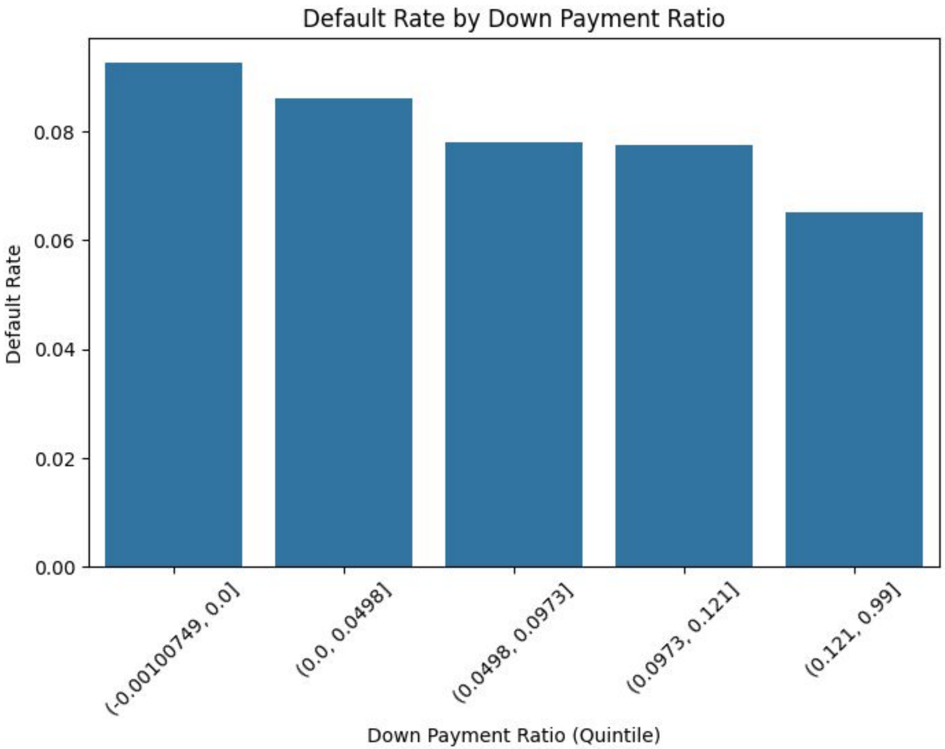


Clients with longer previous loan terms tend to exhibit higher default probabilities.

Long repayment horizons may reflect higher financial stress or structural repayment risk.

This pattern suggests that loan duration history is a meaningful indicator of default risk.

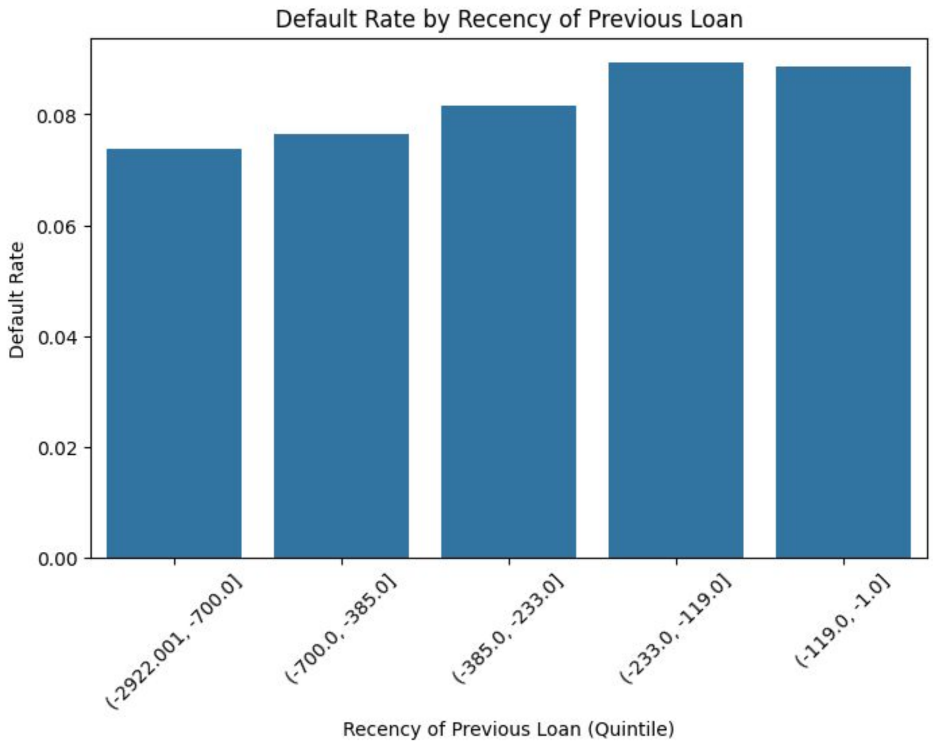
# Down Payment Ratio (RATE_DOWN_PAYMENT)



A clear decreasing trend is observed: higher down payment ratios are associated with lower default rates.

Clients who contribute more upfront capital are likely to be financially stronger and less risky, making down payment ratio a strong protective indicator against default.

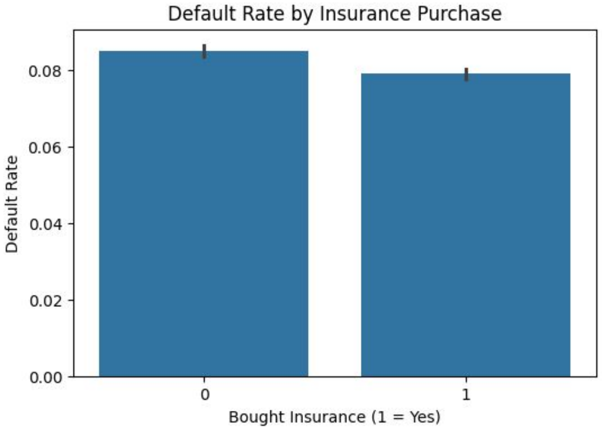
# Recency of Previous Loan (DAYS_DECISION)

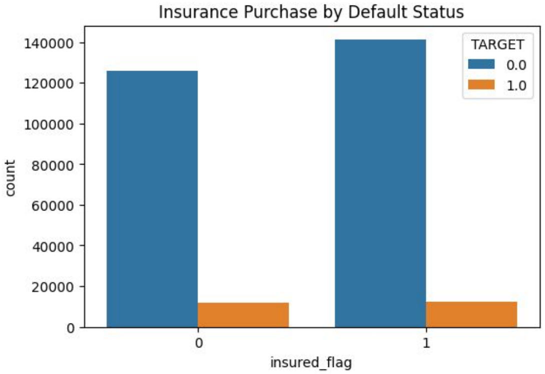


Clients who borrowed more recently tend to show higher default rates.

This suggests that frequent or recent borrowing activity may reflect ongoing financial pressure or increased debt burden.

# Insurance Purchase Behavior (NFLAG_INSURED_ON_APPROVAL)

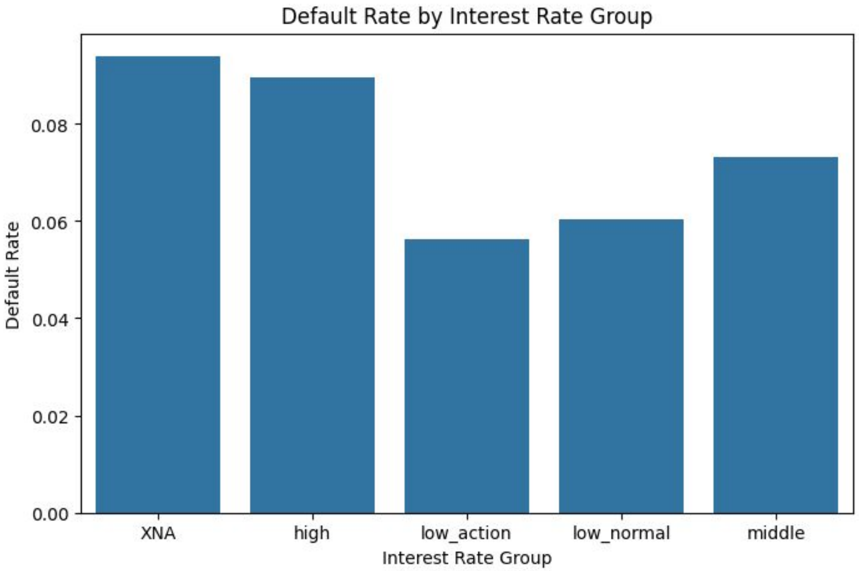




Default rates differ between clients who purchased insurance and those who did not.

Insurance purchase behavior may reflect underlying risk characteristics and contains additional predictive information.

# Interest Rate Group (NAME_YIELD_GROUP)

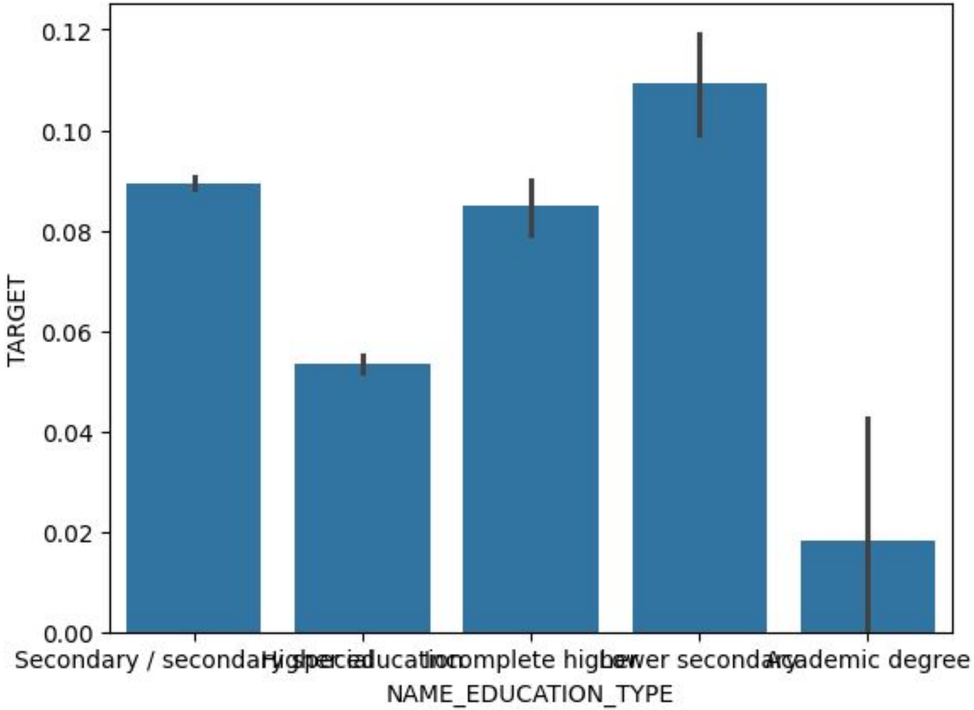


Higher interest rate groups exhibit higher default rates, which aligns with risk-based pricing principles.

This confirms that interest rate assignment captures meaningful information about borrower risk and default likelihood.

# Personal Demographics

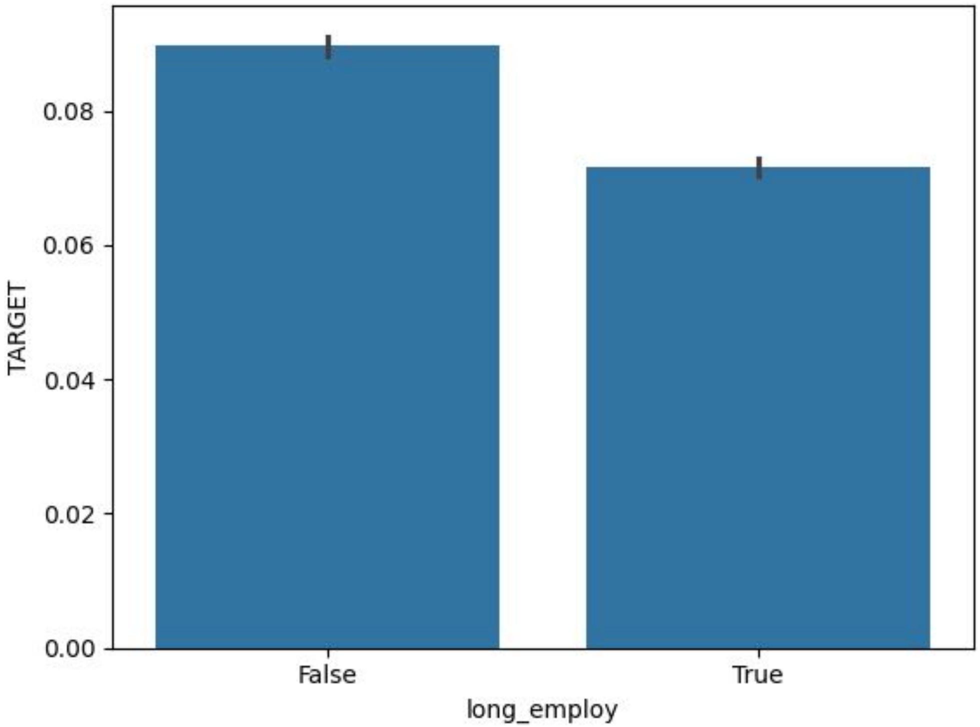
# Education Level (NAME_EDUCATION_TYPE)



Default risk decreases steadily as education level increases. Lower secondary education has a 10.9% default rate, secondary education has a 8.9% default rate, higher education has a 5.4% default rate, academic degree has a 1.8% default rate.

Applicants with lower education have roughly double the default risk

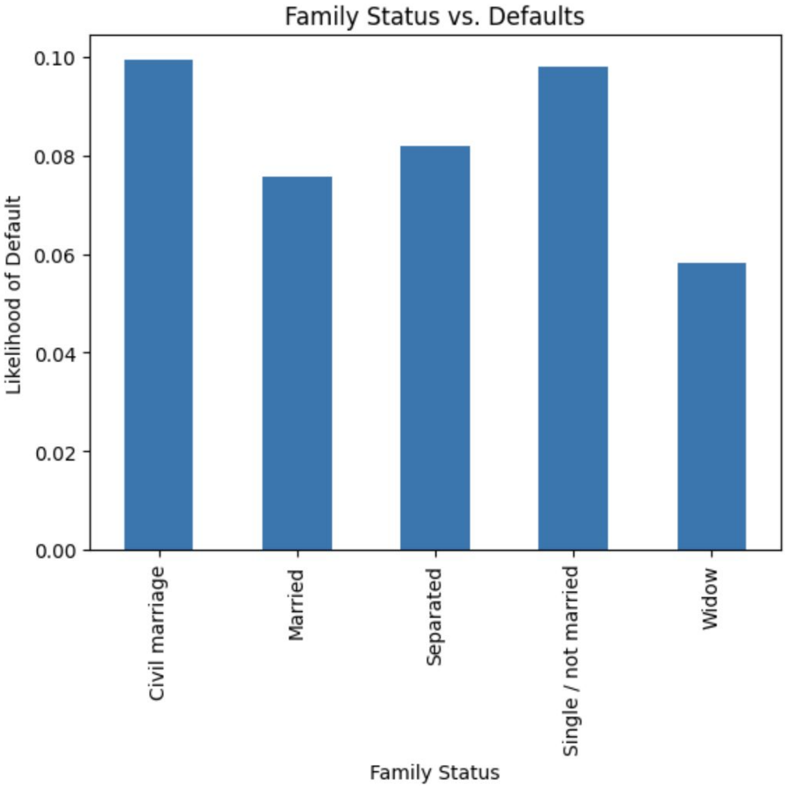
# Employment Length (DAYS_EMPLOYED)



Short employment had a 8.97% default rate and long employment had a 7.17% default rate.

Moderate increase of 15% default risk when going from long employment to short

# Family Status (NAME_FAMILY_STATUS)

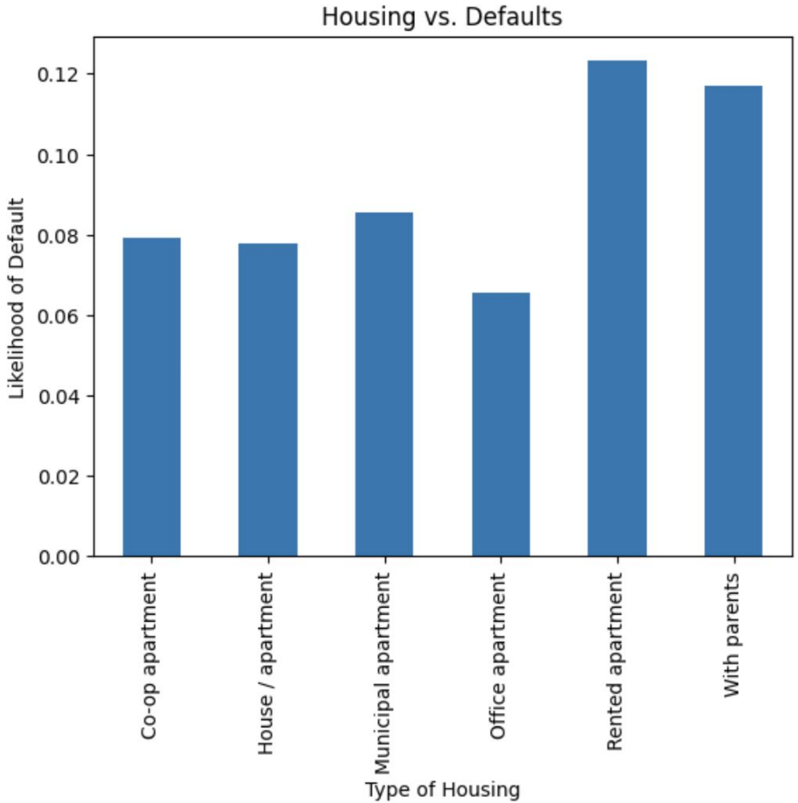


Individuals in a civil marriage or those who are single/unmarried face the highest likelihood of default, with both groups approaching a 10% rate.

Conversely, those classified as a Widow exhibit the lowest risk of default at approximately 6%, while Married individuals show a moderate risk level near 7.5%.

Family structures involving singlehood or informal partnerships are associated with higher financial default rates compared to traditional marriage or widowhood.

# Housing Type (NAME_HOUSING_TYPE)

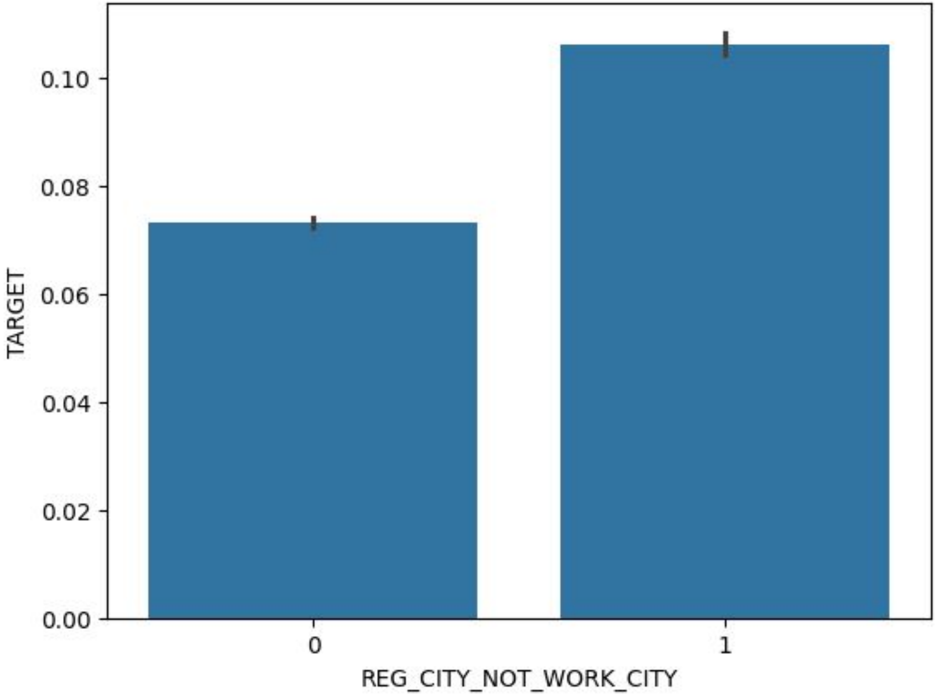


Individuals in more stable or ownership-oriented housing, such as office or co-op apartments, maintain the lowest default rates, ranging between 6.5% and 8.5%.

A significant upward trend in risk appears for those in rented apartments or living with parents, where the likelihood of default peaks at over 12%.

# Geographic and Contractual Risk

# City Registration Mismatch (REG_CITY_NOT_WORK_CITY)

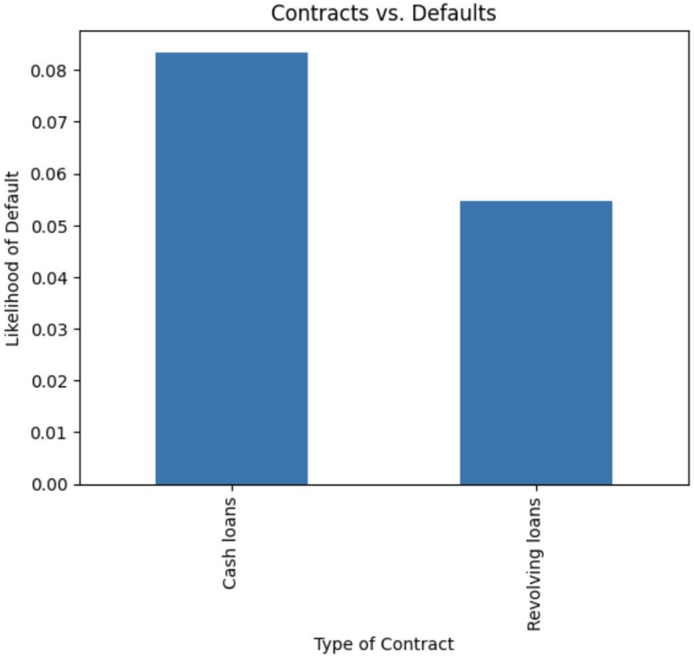


Borrowers whose registered city is different from their work city have slightly higher default rates.

Difference is moderate, registered city being different suggests some instability.

Default rate increases from 7.31% to 10.61% when registered city doesn't match work city, which is a 49% increase in default risk.

# Contract Type (NAME_CONTRACT_TYPE)



Individuals with cash loans have a notably higher default risk, with a likelihood of approximately 8.3%. Conversely, those with revolving loans demonstrate a significantly lower default rate, which is situated near 5.5%.

This disparity suggests that the type of financial contract is a strong indicator of repayment reliability, with revolving credit posing less risk than cash loans.