Better Together? Estimating the Effect of Multiple Borrowers on Mortgage Default Risk Using Double Machine Learning

Andre Oviedo Mendoza

2024-12-12

Abstract

[PRELIMINARY] This final project write-up investigates the effect of multiple borrowers on mortgage default risk using double machine learning. By using this approach, we can account for complex relationships between variables while producing consistent and asymptotically normal estimates of treatment effects. We find that having multiple borrowers at origination is associated with a lower probability of default of around -1.6 percentage points. This is consistent with the findings of previous research that suggests that multiple borrowers can create a collective risk mitigation mechanism that enhances overall loan repayment probability. This result is robust to different models and control variables.

1 Introduction

The landscape of mortgage lending has evolved significantly in recent decades, with multi-borrower arrangements becoming increasingly prevalent in residential real estate financing. These arrangements, typically involving co-borrowers or multiple individuals sharing responsibility for a mortgage, present a complex interplay of financial dynamics that critically impact default risk assessment.

The fundamental premise of multi-borrower mortgages challenges traditional single-borrower models by introducing additional layers of financial complexity. Research by Deng and Quigley (2007) suggests that multiple borrowers can fundamentally alter the risk profile of a mortgage through diversified income streams and shared financial responsibility.

Multiple borrowers typically provide more stable income sources, potentially reducing the likelihood of default. A seminal study by Ghent and Kudlyak (2011) demonstrated that households with multiple income earners exhibit more resilient mortgage repayment capa-

bilities. The presence of multiple borrowers creates a collective risk mitigation mechanism. According to research by Lusardi et al. (2011), co-borrowers often develop informal risk-sharing strategies that enhance overall loan repayment probability.

The nature of relationships between co-borrowers significantly impacts default risk. Research by Mian and Sufi (2009) highlighted that familial or close personal relationships among co-borrowers can create additional psychological incentives for loan repayment.

This research document aims to estimate the causal effect of multiple borrowers on mortgage default risk using double machine learning methods. We leverage a comprehensive dataset of mortgage loans Our analysis focuses on identifying whether and to what extent having multiple borrowers affects the probability of mortgage default.

The use of double machine learning allows us to control for a rich set of covariates while maintaining valid statistical inference. This methodology, developed by Chernozhukov et al. (2024), combines the flexibility of machine learning with the rigorous causal inference framework of econometrics. By using this approach, we can account for complex relationships between borrower characteristics, loan features, and economic conditions that might influence both the decision to have multiple borrowers and the likelihood of default.

Our findings contribute in principle to practical policy discussions on a simple to understand topic: the effect of multiple borrowers on mortgage default risk. Understanding how multiple borrowers affect default risk can inform lending practices, risk assessment models, and regulatory policies in the mortgage market. Additionally, our methodological approach demonstrates the value of modern machine learning techniques in addressing traditional economic questions.

The remainder of this paper is organized as follows. Section 2 describes our data and key variables. Section 3 outlines our empirical methodology, including the double machine learning framework and identification strategy. Section 4 presents our main results and robustness checks. Finally, Section 5 concludes with policy implications and directions for future research.

2 Data

Our analysis uses data from Freddie Mac's Single-Family Loan Performance dataset, which provides detailed information on mortgage loans acquired by Freddie Mac from 2000 to 2023. This comprehensive dataset includes both loan-level characteristics at origination and monthly performance data, making it ideal for studying mortgage default patterns over time.

For computational reasons, we focus on the mortgages loans originated between 2011 and 2016 for all loans originated in a Metropolitan Statistical Area (MSA) which had information for the rate of unemployment and authorization permits for these years. This dataset contains approximately 900,000 unique mortgage loans, with each loan tracked monthly from acquisition until it is either paid off, enters default, or is still active at the end of our observation period. For each loan, we observe a rich set of characteristics including:

Table 1: Table: Variables in the dataset

Category	Variable	Description
Borrower	credit_score	Borrower's credit rating
Borrower	original_dti_ratio	Debt-to-income ratio
Borrower	first_time_homebuyer_flag	Whether this is their first home purchase
Loan	original_cltv	Combined loan-to-value ratio
Loan	original_interest_rate	Initial interest rate on the loan
Loan	occupancy_P	Property occupancy type (Primary)
Loan	occupancy_S	Property occupancy type (Secondary)
Loan	loan_purpose_N	Loan purpose (No Cash-Out Refi)
Loan	loan_purpose_P	Loan purpose (Purchase)
Property	property_state	State where property is located
Property	property_type_PU	Property type (Planned Unit Development)
Property	property_type_SF	Property type (Single Family)
Property	number_of_units	Number of units in the property
MSA	unemployment_rate	Local unemployment rate
MSA	building_permits	Number of building permits issued
MSA	house_price_index	Local house price index
MSA	median_income	Median income in MSA

We clean our dataset of any missing value in these variables. To ensure consistency in our analysis, we focus on conventional, 30-year fixed-rate mortgages originated between 2010

and 2024, allowing us to observe post-origination performance for at least three years for all loans in our sample.

For our main analysis, we define default as a loan becoming 90 or more days delinquent, which is a standard definition in the mortgage literature. We construct our treatment variable based on whether a loan has multiple borrowers at origination, and we create a rich set of control variables including borrower characteristics, loan features, and local economic conditions.

2.1 Identification strategy

To identify a causal effect of multiple borrowers on mortgage default risk, we create an identifying variable in the form of a dummy variable that takes the value of 1 if the loan has multiple borrowers at origination and 0 otherwise. This coupled with a rich set of control variables helps us isolate the causal effect of multiple borrowers from other confounding factors that might influence both the decision to have multiple borrowers and default risk.

The default variable is defined as a loan becoming 90 or more days delinquent, which is a standard definition in the mortgage literature, and it is observed for all loans in our sample for the whole period of each loan's observation window.

3 Methodology

3.1 Double machine learning

Our empirical strategy employs a double machine learning (DML) approach to estimate the causal effect of multiple borrowers on mortgage default risk. This methodology, introduced by Chernozhukov et al. (2024), allows us to handle high-dimensional controls while maintaining valid statistical inference. The key advantage of DML is that it can accommodate complex relationships between variables while producing consistent and asymptotically normal estimates of treatment effects.

The DML framework proceeds in three main steps:

1. First Stage: We estimate the relationship between our treatment variable (multiple borrowers) and control variables using multiple machine learning methods. This produces residualized treatment values that are orthogonal to the control variables.

2. Second Stage: We estimate the relationship between our outcome variable (default) and control variables, again using multiple machine learning methods. This produces

residualized outcome values.

3. Final Stage: We regress the residualized outcome on the residualized treatment to obtain our treatment effect estimate.

To implement this approach, we use a combination of methods including Random Forests, Boosted Trees and Neural Networks. We employ cross-fitting to avoid overfitting concerns, splitting our data into K folds and ensuring that the sample used for prediction is separate from the sample used for estimation.

Our main specification can be represented as the following the partially linear model:

$$default_i = \beta D_i + g(Z_i) + \epsilon_{j,t}$$

Where D_i is the treatment indicator for multiple borrowers, Z_i is our vector of control variables (both at the borrower, loan and MSA level) and $\epsilon_{j,t}$ is the error term.

We implement sample splitting using 5-fold cross-validation to ensure robustness of our results. For inference, we use clustered standard errors at the state level to account for potential spatial correlation in default patterns and fixed effects to control for different vintages of loans (origination month and year).

4 Results and further research

From analysis of the data, we find that having multiple borrowers at origination is associated with a lower probability of default. This is consistent with the findings of previous research that suggests that multiple borrowers can create a collective risk mitigation mechanism that enhances overall loan repayment probability.

Table 2: Effect of Multiple Borrowers on Default Risk Across Different Models

Model	Estimate	Std. Error	Lower CI	Upper CI	RMSE Y	RMSE D
No Controls	-0.018161	0.001598	-0.021293	-0.015029	0.166927	0.499150
Basic Controls	-0.016652	0.001244	-0.019090	-0.014213	0.165138	0.491929
Random Forest	-0.015743	0.001308	-0.018308	-0.013179	0.169374	0.486639

Model	Estimate	Std. Error	Lower CI	Upper CI	RMSE Y	RMSE D
Boosted Trees	-0.016000	0.001295	-0.018538	-0.013463	0.162378	0.485752
Logit NN	-0.015981	0.001298	-0.018526	-0.013436	0.169374	0.485868

The following figures visualize our main results. Figure 1 shows the coefficient estimates and their corresponding 95% confidence intervals across different models, while also comparing the Root Mean Square Error (RMSE) values for both the outcome (Y) and treatment (D) predictions.

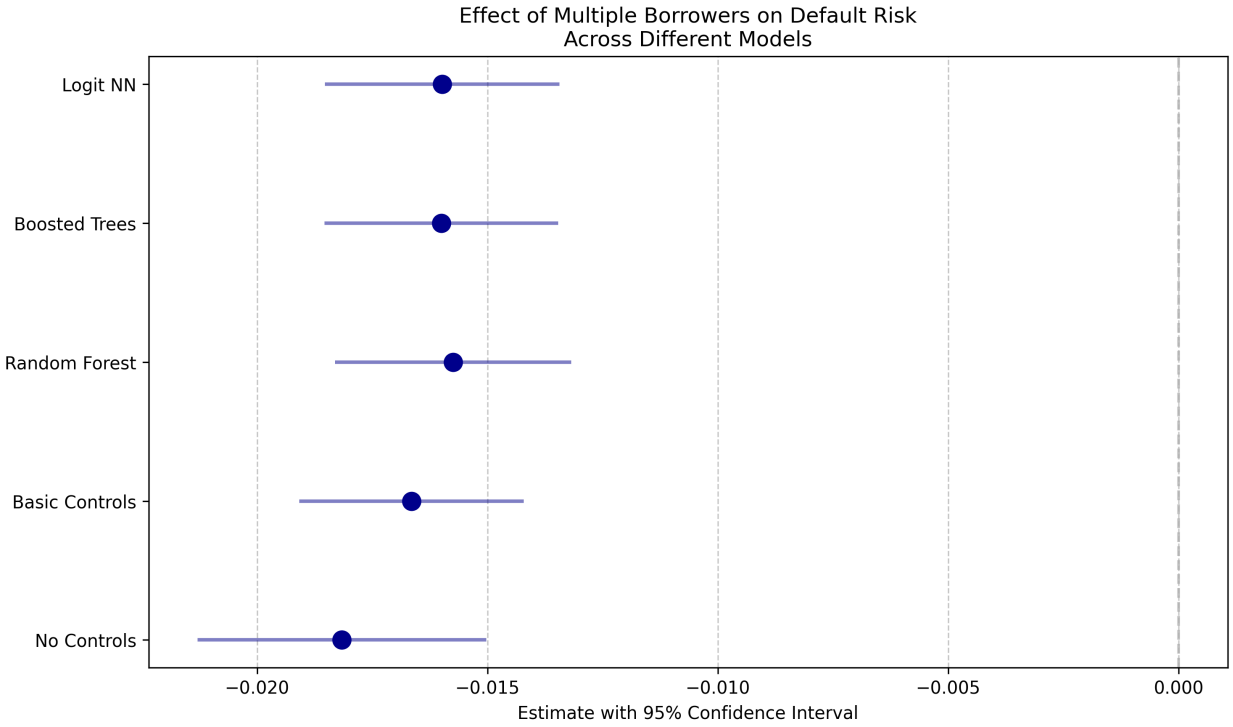

Figure 1: Coefficient estimates and confidence intervals across different models

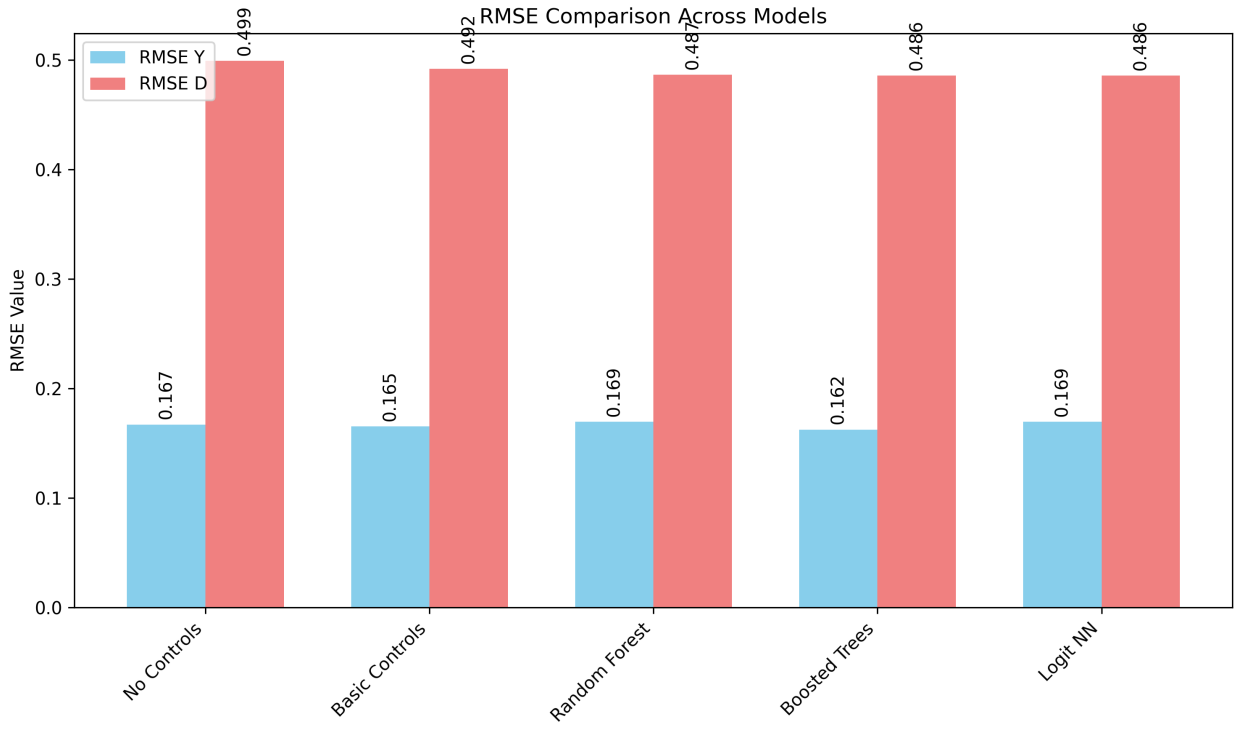

Figure 2: RMSE comparison for outcome and treatment predictions across models

As we can observe from the figures, all models consistently estimate a negative effect of multiple borrowers on default risk, with the magnitude ranging from approximately -1.8 to -1.6 percentage points. The confidence intervals are relatively tight and do not cross zero, indicating statistical significance across all specifications. The RMSE values show that our models achieve similar prediction accuracy, with slightly better performance in predicting the outcome variable compared to the treatment variable.

4.1 Further research

One interesting extension of this research would be to estimate the effect of multiple borrowers on default risk using a difference-in-differences approach. This would allow us to isolate the causal effect of multiple borrowers from other confounding factors that might influence both the decision to have multiple borrowers and default risk.

Given the long series of data, we could also estimate the effect of multiple borrowers on default risk using a panel data approach. This would allow us to control for unobserved heterogeneity and time-invariant factors that might influence both the decision to have multiple borrowers and default risk. There has been some regulation work on the Freddie Mac loan origination process which has removed the requirement to distinguish between single and

multiple borrower mortgages. This would allow us to estimate the effect of multiple borrowers on default risk using a difference-in-differences approach accounting for this change in regulation, but the data is not available for this analysis and would be needed to be analyzed in the upcoming years as the mortgages start maturing.

5 References

Chernozhukov, V., Hansen, C., Kallus, N., Spindler, M., & Syrgkanis, V. (2024). Applied causal inference powered by ml and ai. https://doi.org/10.48550/arXiv.2403.02467

Deng, Y., & Quigley, J. M. (2007). Irrational borrowers and the pricing of residential mortgages.

Ghent, A. C., & Kudlyak, M. (2011). Recourse and residential mortgage default: Evidence from u.s. states.

Lusardi, A., Schneider, D. J., & Tufano, P. (2011). Nber working paper series.

Mian, A., & Sufi, A. (2009). The consequences of mortgage credit expansion: Evidence from the u.s. mortgage default crisis*. *Quarterly Journal of Economics*, *124*(4), 1449–1496. https://doi.org/10.1162/qjec.2009.124.4.1449

6 Appendix: Exploratory analysis

Unnamed: 0	count	mean	std	min	25%	50%	75%	max
credit_score	1522893.000	767.074	42.620	572.000	728.000	767.000	791.000	846.000
original_dti_ratio	1522893.000	32.527	9.738	1.000	25.000	33.000	41.000	59.000
original_cltv	1522893.000	73.567	16.064	6.000	66.000	78.000	80.000	164.000
original_interest_rate	1522893.000	3.927	0.554	2.250	3.625	3.875	4.250	6.375
first_time_homebuyer_flag	1522893.000	0.145	0.352	0.000	0.000	0.000	0.000	1.000
occupancy_P	1522893.000	0.891	0.311	0.000	1.000	1.000	1.000	1.000
occupancy_S	1522893.000	0.033	0.180	0.000	0.000	0.000	0.000	1.000
loan_purpose_N	1522893.000	0.337	0.473	0.000	0.000	0.000	1.000	1.000
loan_purpose_P	1522893.000	0.454	0.498	0.000	0.000	0.000	1.000	1.000
unemployment_rate	1522893.000	-11.648	8.138	-45.312	-16.667	-12.088	-7.500	30.769
private_housing_authorization	1522893.000	0.015	62.744	-98.095	-9.091	18.462	53.136	2050.000
defaulted	1522893.000	0.029	0.167	0.000	0.000	0.000	0.000	1.000
metropolitan_statistical_area	1522893.000	32.244	10681.196	12060.000	17140.000	26420.000	38060.000	47220.000
first_payment_date	1522893.000	201398.638	168.368	201102.000	201301.000	201407.000	201512.000	201701.000
more_than_one_borrower	1522893.000	0.529	0.499	0.000	0.000	1.000	1.000	1.000

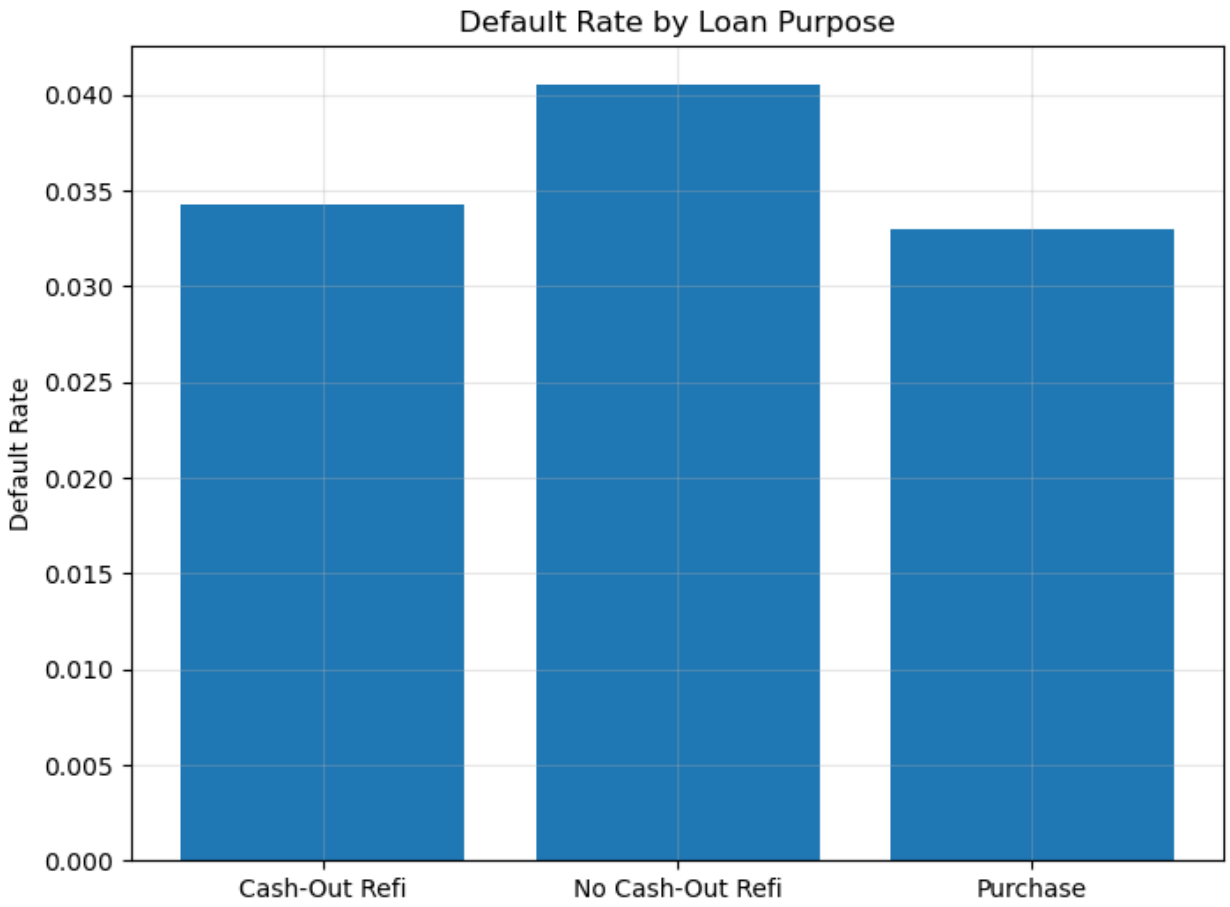

Figure 3: Default Rate by Loan Purpose

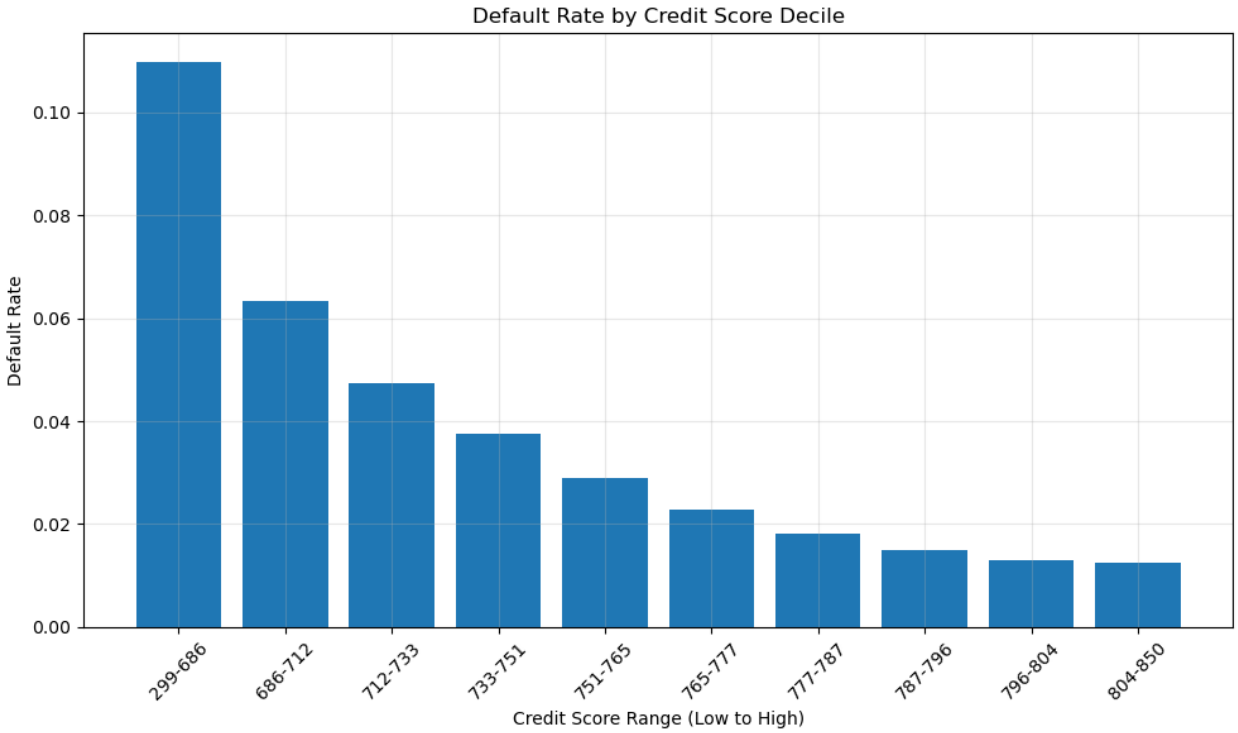

Figure 4: Default Rate by Credit Score Decile

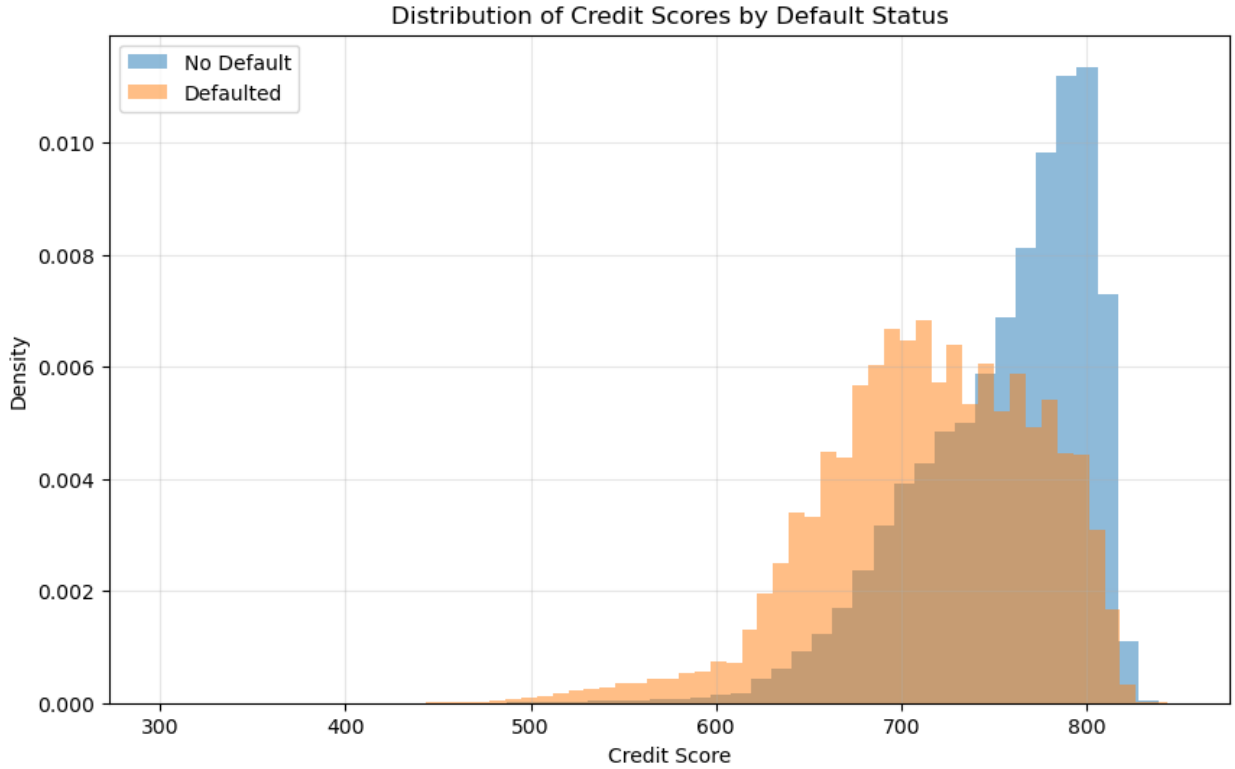

Figure 5: Distribution of Credit Scores by Default Status

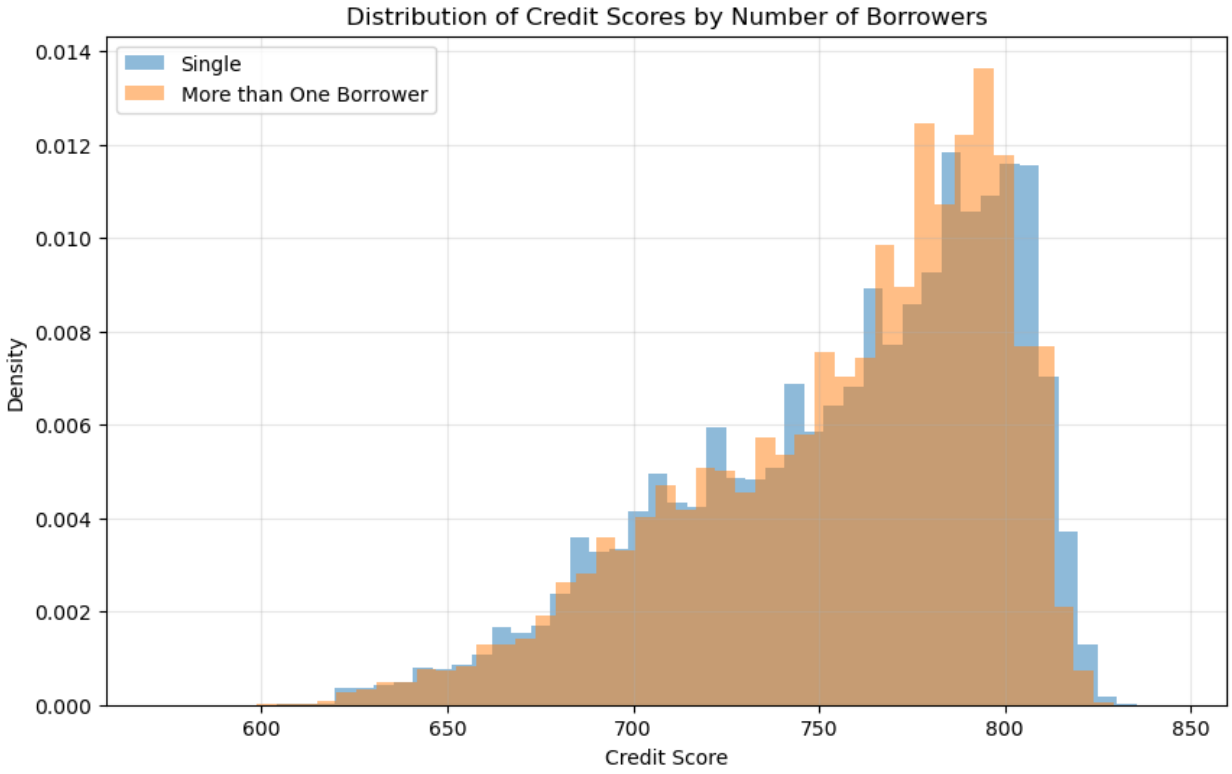

Figure 6: Distribution of Credit Scores by Number of Borrowers

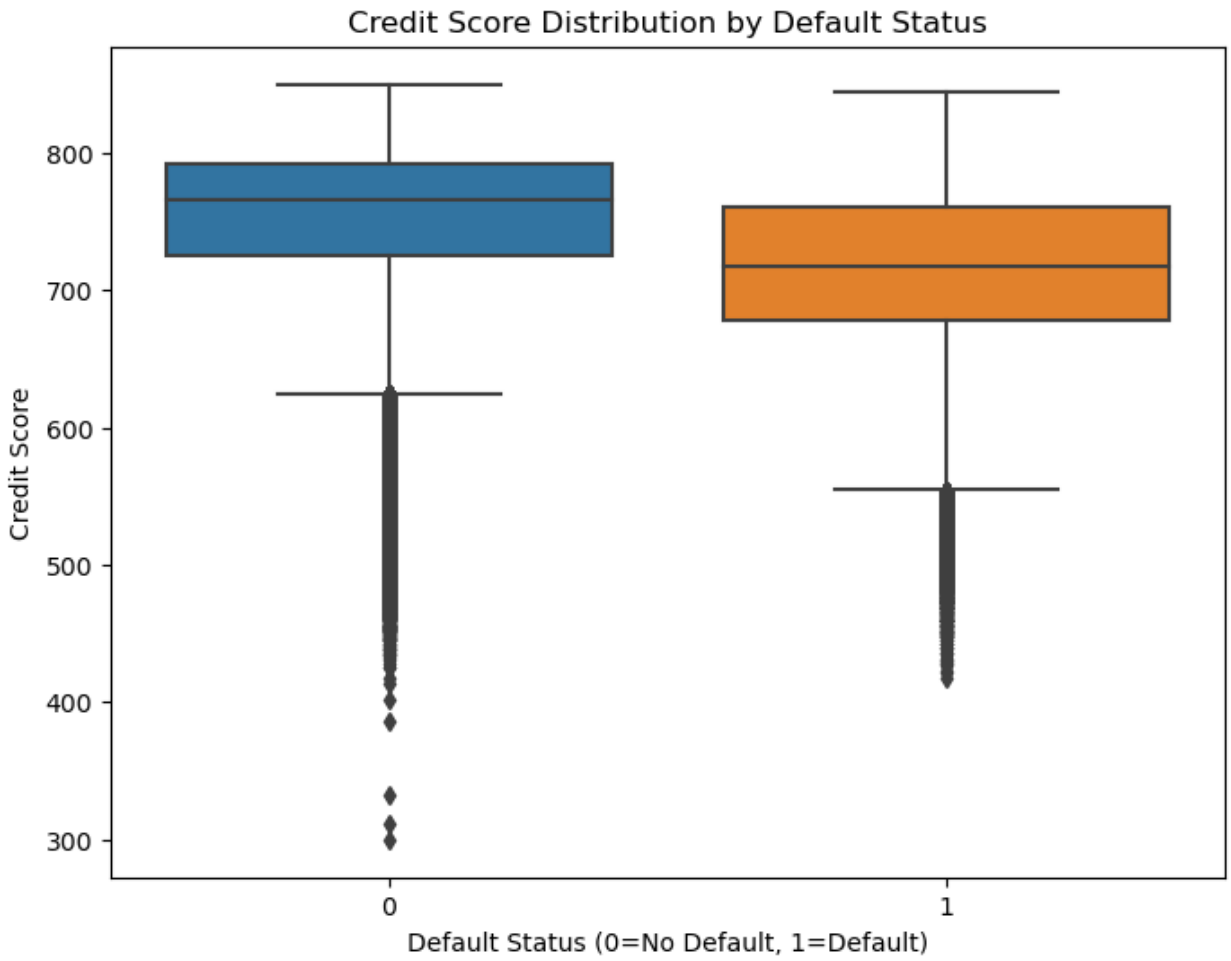

Figure 7: Credit Score Distribution by Default Status