Estimating Key Drivers of Peer to Peer Loan Interest Rates

Brodie Deb, Ethan Moody, and Eugene Oon | August 2022

Introduction

It is difficult to understate the size and importance of the personal loan industry. According to Federal Reserve Economic Data (FRED), the current balance of consumer loans in the U.S. stands at over $1.5 trillion – a staggering number which translates to roughly $4,500 in loans per individual.[1] Although personal loan debt slumped during the height of the COVID crisis in 2020, the world saw a resurgence in unsecured debt demand in 2021 which has continued into this year.[2] As a result, now may be an opportune time for major lending companies to reevaluate their loan product underwriting practices and interest rate offerings to attract new loan applicants.

In the following study, we assume the persona of a team that has been approached by a well-known financial services company – e.g., Discover, Inc. ("DI") – with an opportunity to help them in this way. The context is that DI wants to enter the peer-to-peer lending market with a focus on personal medium-term microloans (i.e., loan amounts of ~$10,000 over a 3-year term). They have asked us to perform a competitive research study that focuses on how LendingClub ("LC") – a key competitor and pioneer in the fintech industry – sets interest rates for their own peer-to-peer loan product. Given current economic conditions, DI is keen to mitigate the risk of loan default. They are specifically looking for a way to offer *higher quality* (lower risk) loans to applicants as part of their new peer-to-peer offering. DI is interested to know how a lower (more restrictive) debt-to-income ratio requirement would reduce the loan interest rate offered to a new customer. This requirement is a *controllable feature within their product underwriting criteria* (i.e., something they could adjust to optimize their new peer-to-peer loan product), and they consider interest rate to be a measure of loan risk.

Data and Methodology

For this study, we applied a set of regression models to show how DI could target higher quality/lower risk loans. To do so, we used a historical record of loans issued by LC in Q1 2018. The choice of timeframe for this data was intentional, as we wanted to avoid any noise from the unique economic impacts of the COVID pandemic throughout 2020-2021. The dataset contains 10,000 observations on 55 variables detailing a variety of loan characteristics (e.g., issue month, principal balance, etc.) and loan applicant attributes (e.g., location state, employment length, etc.). It was compiled and made publicly available on OpenIntro.[3] Each record represents a loan made between January to March 2018. We performed all exploration and model building on a ~30% subsample of the data and used the remaining ~70% to generate the statistics in this report. Since DI's focus is on medium-term microloans, we cleaned out unavailable ("NA") values and limited our analysis to just those records associated with LC's most-equivalent loan option: a $10,000 loan amount with a 36-month (3-year) term. This left us with *223 observations for exploration and model building* and *458 observations for model confirmation.*

From the dataset documentation, we knew there were potential clustering effects at-play within the data. For example, LC typically issues loans only to those who are U.S. citizens or permanent residents, are 18+ years of age, and have a minimum Fair Isaac Cooperation ("FICO") credit-risk model score (which presupposes *some* credit history). The loan holders represented in the data appear mostly geographically independent from each other – being distributed across all 50 states – though we saw a high share of records originating from populous states like CA (13.3%), NY (7.9%), and FL (7.3%) where residents might share some similar characteristics. We regarded the period of loans within the data as sufficiently short (i.e., first three months in 2018) and representative of a

[1] Consumer loans, all commercial banks. FRED. (2022, July 15). Retrieved July 17, 2022, from https://fred.stlouisfed.org/series/CONSUMER.

[2] Schulz, M. (2022, June 22). Personal loan statistics: 2022. LendingTree. Retrieved July 17, 2022, from https://www.lendingtree.com/personal/personal-loans-statistics/.

[3] OpenIntro. (2018). Loan data from Lending Club. OpenIntro Data Sets. Retrieved July 19, 2022, from https://www.openintro.org/data/csv/loans_full_schema.csv.

non-economically turbulent time, which allowed us to treat external forces as essentially constant. Additionally, we recognize that while the filters placed on loan amount and term limited dataset size, they also provided us with enough data for a large-sample regression analysis and helped us control for potential confounds and reverse causal pathways between interest rate and loan amount/term. Given these considerations, we thought the final "cleaned" version of the dataset was reasonable and relevant for our analysis.

To operationalize *loan interest rate* (the "outcome variable" in our study), we used the `interest_rate` column from the original dataset, which represents the interest rate of the loan an applicant received. To operationalize *debt-to-income ratio* (the primary "predictor" or variable of interest in our study), we used the `debt_to_income` column from the original dataset, which is calculated as an applicant's total debt divided by their annual income. Figure 1 provides a basic representation of how debt-to-income ratio and interest rate are related (though, presently, without controlling for other variables). While the plot shows a weak positive semi-linear relationship between these two variables, we also see a wide range of interest rates offered to applicants with the same debt-to-income ratio. This result encouraged us to explore several additional predictor variables within our study.

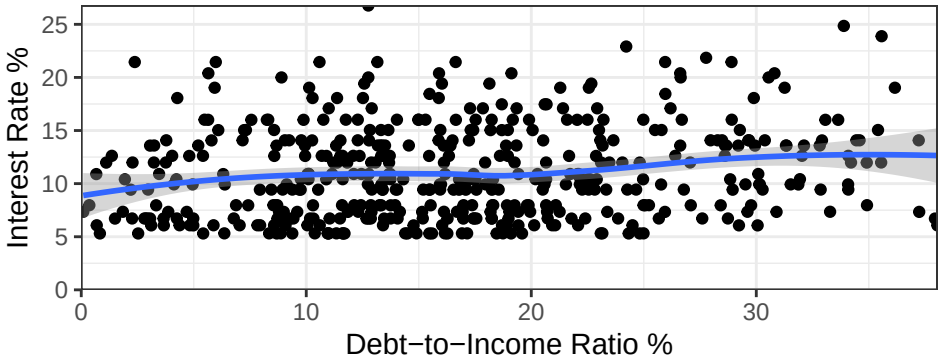

Figure 1: Interest Rate as a Function of Debt-to-Income Ratio

Exploratory analysis revealed that both *interest rate* and *debt-to-income ratio* have heavily right-skewed distributions. As indicated by a Box-Cox transformation analysis, we applied a natural logarithm to *interest rate* to transform it to $ln(interest\ rate)$ and a square root to *debt-to-income ratio* to transform it to $\sqrt{debt\ to\ income\ ratio}$, normalizing both distributions for linear modeling. We then fit regressions of the form,

$$\widehat{ln(interest\ rate)} = \beta_0 + \beta_1 \cdot \sqrt{debt\ to\ income\ ratio} + \mathbf{Z}\gamma$$

where β_0 is a constant, β_1 represents the change to interest rate driven by each unit of increase in the square root of debt-to-income ratio, $\mathbf{Z}$ is a row vector of additional covariates, and γ is a column vector of coefficients.

We evaluated three models of varying complexity. As we sought to improve model performance, we incorporated a number of loan applicant factors that were mentioned in external research and that a lender (like DI) would include in their product/underwriting guidelines.[4] These factors were (a) percentage of total credit utilized, (b) number of credit inquiries over last 12 months, (c) number of satisfactory accounts, (d) loan purpose, (e) homeownership status, and (f) credit history. We reviewed exploratory distributions of these variables and evaluated Box-Cox transformations as appropriate to support linear modeling. We also chose to exclude the following variables from our models: (g) annual income (since it was already accounted for in the debt-to-income ratio), (h) employment length (since it did not add any explanatory power or come up in external research), and (i) grade/sub-grade of loan (since it was so closely tied to our outcome variable and influences the interest rate for a loan). Finally, external research revealed that FICO scores are commonly used by lenders to assess potential borrowers' creditworthiness. However, no credit score data was present in our original dataset. This led us to a supplementary data source containing Q1 2018 FICO score ranges for applicants by their loan grade.[5] We used this data to calculate the average FICO credit score across each loan grade, and then mapped these averages to each record in the original LC loan dataset based on the `grade` column. We incorporated a variable for (z) credit score into our third model.

[4]Renton, P. (2011, November 7). "How Lending Club and Prosper Set Interest Rates." Fintech Nexus News. Retrieved July 30, 2022, from https://news.fintechnexus.com/how-lending-club-and-prosper-set-interest-rates/.

[5]George, N. (2019, April 10). All Lending Club loan data. Kaggle. Retrieved July 29, 2022, from https://www.kaggle.com/datasets/wordsforthewise/lending-club?select=accepted_2007_to_2018Q4.csv.gz.

Results

Regression model results are shown in Table 1 below. Our first model regresses *interest rate* only on *debt-to-income ratio*, our second model incorporates covariates (a) through (f) above, and our third model adds *credit score* (FICO proxy) as a final covariate. We found that the coefficient for *debt-to-income ratio* was only statistically significant ($p < 0.05$) for our first two models, with point estimates ranging from 0.04 to 0.06. This result suggests that if a loan applicant's annual income was $100,000 and they suddenly took on $20,000 of debt, the interest rate they would be offered on a medium-term microloan would increase by 0.02 to 0.03 points, holding all else equal.

With our second model, we found that the coefficients for *credit utilization rate*, *credit inquiries in last 12 months*, *loan purpose* (specifically for *debt consolidation* and *other - excluding credit card*), and *age of earliest credit line* were all statistically significant ($p < 0.05$). The F statistic to gauge overall/joint significance of the variables was also statistically significant ($F = 9.69, p < 0.001$). This finding suggests that our second model improves upon the first by incorporating variables which increase the regression's goodness of fit ($R^2_{model1} = 0.03, R^2_{model2} = 0.16$).

With our third model, we found that only the coefficients for *credit score (FICO proxy)* and *credit inquiries in last 12 months* were statistically significant ($p < 0.05$). The negative point estimate of -0.02 indicates that *interest rate* increases as *credit score* decreases, which aligns with prevailing industry knowledge. The lack of significance found across all other variables – combined with a significant F statistic ($F = 384.16, p < 0.001$) and high coefficient of determination ($R^2_{model3} = 0.90$) – suggests that *credit score* by itself is enough to explain most of the variation in *interest rate* across loan applicants.

Table 1: Estimated Regressions

	Output Variable: Interest Rate (ln-transformed)		
	(1)	(2)	(3)
(a) Debt-to-Income Ratio (sqrt-transformed)	0.06^{***} (0.02)	0.04^{*} (0.02)	0.005 (0.01)
(b) Credit Utilization Rate (sqrt-transformed)		0.44^{***} (0.08)	0.01 (0.03)
(c) Credit Inquiries in Last 12 Months		0.03^{***} (0.01)	0.01^{*} (0.003)
(d) Number of Satisfactory Accounts		−0.004 (0.003)	0.001 (0.001)
(e.1) Loan Purpose: Debt Consolidation		0.15^{***} (0.04)	0.03 (0.01)
(e.2) Loan Purpose: Home Improvement		0.12 (0.07)	0.04 (0.03)
(e.3) Loan Purpose: Other - Exc. Credit Card		0.14^{**} (0.05)	0.03 (0.02)
(f) Homeownership Status: Own Home		0.01 (0.04)	0.02 (0.02)
(g) Age of Earliest Credit Line		-0.01^{*} (0.002)	−0.001 (0.001)
(z) Credit Score (FICO proxy)			-0.02^{***} (0.0004)
Constant	2.09^{***} (0.06)	1.87^{***} (0.08)	15.10^{***} (0.27)
Observations	458	458	458
R^2	0.03	0.16	0.90
Adjusted R^2	0.03	0.15	0.89
Residual Std. Error	0.39 (df = 456)	0.36 (df = 448)	0.13 (df = 447)
F Statistic	14.53^{***} (df = 1; 456)	9.69^{***} (df = 9; 448)	384.16^{***} (df = 10; 447)
Note:	*p<0.05; **p<0.01; ***p<0.001 Robust standard errors in parentheses (a) Debt-to-Income Ratio is total debt divided by annual income (b) Credit Utilization Rate is credit utilized divided by credit limit (e.1-3) Loan Purpose also encompasses Credit Card loans, though this category is excluded from models to prevent a closed system (g) Age of Earliest Credit Line is 2018 minus year of first credit line (z) Credit Score is a proxy for FICO score based on the average score associated with a loan grade		

Discussion and Limitations

These results lead to two interesting conclusions. First, they show that a lower debt-to-income ratio *does* yield a lower interest rate (holding all else constant). Second, they show that other variables might be even more predictive of interest rates than debt-to-income ratio. We plotted term-by-term R^2 contributions to determine which variables

drove the largest impact to our modeled R^2 results, and we saw that *credit utilization rate* and our proxy variable for *credit score* had the highest values. This means that DI might want to consider simultaneously adjusting *multiple* controllable features in their loan product/underwriting requirements – such as the credit utilization rate requirement and debt-to-income ratio requirement – to better target higher quality/lower risk loans. They could also consider adjustments to credit score requirements for even larger impact – though, as we discuss below, credit score may not always be seen as an *ethical or equitable* variable to include in loan design or underwriting criteria.

These regression estimates rest upon the large-sample linear model assumptions: independent and identically-distributed (IID) observations and the existence of a unique best linear predictor (BLP). We made a few observations about potential threats to IID in our discussion of the original dataset (e.g., sources of geographic clustering/dependence across loan applicants) but consider them negligible. We also acknowledge the 3-month timeframe of the data could subject it to risk from external/economic confounds and present a source of serial correlation across samples. However, given the strictness of this timeframe, we consider this risk to be minimal. Additionally, we controlled for a major potential threat to the "identically-distributed" aspect of IID by limiting the scope of our study to a specific loan amount ($10,000) and term (36-months). This limit allowed us to avoid making conclusions across samples drawn from markedly different populations (microloan vs. macroloan applicants). To support the BLP assumption, we applied power transformations to our outcome variable, key predictor variable, and other covariates as appropriate to normalize their heavy tailed distributions. We also computed variance inflation factors (VIFs) for our model predictors and found no evidence of perfect collinearity between terms.

Nevertheless, we note the following key limitations of our study: 1) the specific loan criteria we imposed on the data limits how much our findings can be generalized; 2) there is no guarantee that *post-COVID* loan data would necessarily support the same models or conclusions drawn from the *pre-COVID* data we used; and 3) the use of credit scores in product offering, pricing, and underwriting decisions is becoming increasingly controversial, attracting social and ethical critiques from public and private entities.[6] The limited availability of FICO scores across some regions or populations – coupled with potential discriminatory concerns surrounding credit scoring – presents a good reason for DI to consider adjusting other aspects of their loan requirements or application process (e.g., credit utilization rate, debt-to-income ratio, etc.) *before* making adjustments to credit score requirements if they want to offer higher quality loans to the broadest array of new applicants.

We also acknowledge two additional limitations surrounding the structure of our models: 1) the transformations we applied to variables lessens the explanatory power of our models; and 2) the lack of *actual* FICO credit scores included in LC's loan data introduces omitted variable bias. This first limitation plays an important role in practical significance, as models of increasing complexity can be harder to communicate or operationalize. However, given the needs of our primary stakeholder (DI), we are comfortable exchanging some explanatory power for predictive power. The second limitation suggests that the hypothesis tests for our models may be overconfident. We suspect the absence of actual FICO scores produces a *positive* omitted variable bias on model coefficients (driving the effect away from zero) due to assumed negative correlations between A) FICO scores and debt-to-income ratios and B) FICO scores and interest rates. This limitation is something we attempted to address using average FICO scores by grade as a "proxy." While we achieved better performance with our third model using this proxy, we recognize our approach does not provide a *perfect* stand-in for the actual FICO score.

Conclusion

Our study evaluated three models estimating the interest rate (risk) for a medium-term microloan. While LC data suggests that *debt-to-income ratio* is a significant predictor of loan *interest rate*, we believe additional applicant and loan characteristics play a key role in determining this rate as well. This analysis should encourage DI to consider adjustments to debt-to-income ratio requirements (and other underwriting requirements/product features) as one way to target higher quality/lower risk loans as they enter the peer-to-peer lending space.

Future research might build upon our analysis by testing these models across different loan amounts, terms, and timeframes. As global economies continue to reopen and the U.S. Federal Reserve considers additional rate hikes, lending practices and demand for credit could change. It would be interesting to assess what these changes might mean for loan competitiveness, lender interest revenue, and the fate of peer-to-peer lending in the future. We hope this study not only adds value for DI, but also contributes to a conversation around product changes that other loan providers could make to mitigate risk during the post-pandemic era.

[6]Chopra, S. (2021). Current Regulatory Challenges in Consumer Credit Scoring Using Alternative Data-Driven Methodologies. Vanderbilt Journal of Entertainment and Technology Law, 23(3), 625–648. https://scholarship.law.vanderbilt.edu/jetlaw/vol23/iss3/4.