Asymmetric information in the Brazilian credit market: testing adverse selection predictions

Gustavo Araujo Everton Gomes Felipe Iachan Flavio Moraes*

The views expressed in this paper are those of the authors and do not necessarily reflect those of the Banco Central do Brasil

March 16, 2024

Abstract

Exploring firm-level data we empirically study the degree of asymmetric information in the Brazilian credit market, and test whether the usual predictions of theoretical models regarding asymmetric information are observable in our data. Considering that the credit decision is based on credit risk models, we have investigated how private data, which can only be accessed by some lenders (usually incumbents) can improve the models and produce advantages for competitors granted access to said data. Our findings suggest that private information substantially increases the accuracy of credit models, enabling competitors who are granted access the opportunity to increase their potential client portfolio by more than 100%, without facing a higher default rate, which suggests that the degree of asymmetric information is highly significant. In addition, our results provide evidence that new competitors face adverse selection conditions, since riskier clients are 27% more likely to migrate from their original bank, and subsequently have a 29% higher default rate. Nevertheless, by exploring the value of private information in different clusters of clients, we have found evidence that the usual data available in the Brazilian market are useful for more easily identifying high-risk clients, but insufficient for identifying very low-risk clients. Finally, we have tested a random forest model to investigate whether using a more modern modeling approach could replace reliance on private information but have not found evidence to support this.

Keywords: Credit Market, Asymmetric Information; Information Sharing

*Araujo: Banco Central do Brasil (gustavo.araujo@bcb.gov.br); Gomes: Verde Asset Management (everton.gomes@verdeasset.com.br); Iachan: FGV EPGE, Brazilian School of Economics and Finance (felipe.iachan@fgv.br); Moraes: COPPEAD Graduate Business School, Federal University of Rio de Janeiro (flavio.moraes@coppead.ufrj.br).

I Introduction

Money-lending is probably one of the businesses that requires the most information and predictive capabilities, considering that the return on each loan depends on the probability of the borrower defaulting, which is an unknown variable. In addition, the depth of information about this probability varies among players in this market: borrowers are better informed about themselves and their projects than lenders; and lenders who are more familiar with their borrowers (for instance, incumbents who have longstanding relationships with them) are better informed than new lenders trying to attract these same borrowers as new clients. Hence, credit markets are rich fields for exploring information-related issues, which is what we aim to do in this paper. Specifically, we empirically study the degree of asymmetric information in the Brazilian credit market, and test whether the usual predictions of theoretical models regarding asymmetric information, applied to credit markets, are observable in our data, especially with respect to adverse selection. In addition, we have divided our sample into clusters, to evaluate whether information is more useful in specific groups, such as riskier clients.

Considering that the credit decision is based on credit risk models, which are used to predict the likelihood of default, we have investigated how private data, which can only be accessed by some lenders (usually incumbents) can improve the models and produce advantages for competitors granted access to said data. Afterwards we look for signs of adverse selection, by testing whether the propensity to migrate from an incumbent bank to a new lender is correlated to the credit risk; whether the clients who migrate have a higher default rate in comparison to other clients; and whether new clients are offered inferior credit terms (higher interest rates and lower credit limits). Our findings suggest that private information substantially increases the accuracy of credit models, enabling competitors who are granted access the opportunity to increase their potential client portfolio more than 100%, without facing a higher default rate, which suggests that the degree of asymmetric information is highly significant. In addition, we have found evidence that the usual results expected in this environment of asymmetric information are present in the Brazilian market: riskier clients are 20% more likely to switch bank, and thus clients who move to a new bank have higher default rates and worse credit terms[1].

Our Brazilian dataset is interesting to study for two reasons: i) the country's high spreads and banking concentration, which suggest a significant impact from asymmetric information; and ii) researchers authorized by the Brazilian Central Bank are allowed to access private information, which produces advantages for major banks (because banks are required to report detailed information about their credit transactions to the Central Bank). Hence, we have sufficient information to evaluate the extent of the advantage enjoyed by incumbent banks and test the predictions from models in this context. We

[1]Stroebel (2016), for instance, has developed a theoretical model for such predictions

also analyze the type of information and its utility for predicting default rates by different groups of clients. Studying this information makes a significant contribution to this paper, considering that one of the problems in evaluating the effects of asymmetric information is that private information is not usually available to researchers. Thus, our data allow us to test predictions considering different datasets that lenders can observe about each firm.

We have used econometric models that are typically applied for credit decisions, in which the dependent variable is the default, simulating availability of different datasets. Thus, we have estimated a model that we have named 'complete model', whose explanatory variables include private and public information, and another model called the 'incomplete model', whose inputs originate solely from public information. By comparing the performance of these two models, we gain a measure of the level of information asymmetry, since incumbent lenders can develop models like the complete one, while new competitors must rely on models like the incomplete one. To quantify this situation, we have used indicators that are common to the credit business: Kolmogorov–Smirnov (KS) test results and measures of default rates X approval rates (better models are able to approve credit for more clients while keeping the default rate constant or the approval rate constant with a lower default rate). We have found that the Kolmogorov–Smirnov test result for the incomplete model was 14 percentage points lower than for the complete model, suggesting that the discriminating power of the incomplete model is significantly lower.

Having evaluated the degree of asymmetric information, as described above, we have followed Stroebel (2016) [1] and Foley et al. (2018) [2], testing the main predictions from asymmetric information models as they apply to credit markets:

- Borrowers that migrate from incumbent banks to new ones have higher default rates;
- Consequently, lenders establish more stringent credit terms for new clients that migrate from other banks;
- Observable high-risk borrowers are more likely to move from incumbent banks to other lenders, since the private information advantage is less pronounced once it is easier to identify a very high-risk client. In this prediction, our test is more precise than usual: by accessing both public and private information, we can test whether clients that present low risk according to public information (incomplete model) and high risk according to private information (complete model) are more likely to change from the incumbent bank. This shift is expected because in this situation, only lenders that can access private information know this borrower is high risk, so new competitor will evaluate the same borrower as having lower risk and may offer relatively better terms.

All expected results listed above were observed in our data and were statistically significant. In addition, when investigating the value of private information in subsamples of our data, we found that banks are able to separate clients with very high propensity to default from clients with low propensity, but are not able to differentiate between clients that have a low propensity to default from those with a very low propensity, due to the kind of variables that are available with significant information value – most of these variable relate to short payment delays, which is not observable on the databases held by credit bureaus (which mainly identify borrowers longer than 60 days overdue).

Finally, we tested an alternative modeling method, to investigate whether a more modern approach could replace reliance on private information (using random forest instead of logit models), but we did not find evidence in that direction, so it seems that more information has much greater importance.

The remainder of the paper is structured as follows: the next section describes the related literature, section three the database, section four specifies the empirical approach, section five presents our results, and the last section contains our concluding remarks.

II Literature Review

Theoretical relationship between asymmetric information and credit market failures is deeply investigated since Stiglitz and Weiss (1981), that formalized a model which results suggests that imperfect information could lead to a disequilibrium in which prices are not able to drive the supply to match the demand. The problem is that even if marginal lenders would accept to pay higher interest rates to compensate higher risks, setting such rates could cause an additional increase in the risk, due to adverse selection (worsening the pool of borrowers) or even moral hazard (inducing the lenders to take more risks in his projects, increasing its expected return to pay for that rates). The effect described by Stiglitz and Weiss (1981) is a consequence of the asymmetric information between the lenders and the borrowers, once the last are more informed about his condition to repay the loan than the formers.

Another source of asymmetric information is the unequal level of information among different competitors about the same client. In general, the incumbent lender has the best information, developed while observing its client's repayment behavior, and may appropriate to itself some monopoly profit from that. The challenger, that offer some credit to potential lenders that are incumbent's clients will probably get only the riskier ones, facing higher non repayment rates and, consequently, lower profitability. Some papers, for instance Dell'Ariccia, Friedman, and Marquez (1999)[3], model this situation highlighting how the asymmetric information act as a barrier to new entrants. Some others, like Stroebel (2016) and Foley et al (2018) [1], focus on testing the predictions of this kind of model, usually producing evidence on the direction that less informed competitor faces

higher non repayment rates and set worse credit conditions to its borrowers, i.e. lower credit limits and higher interest rates.

In the light of this adverse effect of the asymmetric information in the credit markets, researchers investigated the potential impact of sharing information (Vives (1990), Foley et al (2018)) [4], how effective it could be to mitigate that effect, in which situations lenders could spontaneously agree to share information (Pagano and Jappelli (1993), Vives (1990)) [5], or when regulators actions are necessary. Interesting to note that there is several evidence that sharing information reduces interest rates, reducing monopoly power of the incumbent bank, nevertheless, there is no evidence that sharing information increase loans. Actually, theory predicts that exchange information cause increase in the credit amount only if in the equilibrium before the sharing there was low risky clients who were restricted from credit that could compensate the likely exclusion of high risk clients from the market. Hence, there is a trade-off between more competition (lower interest rates) against credit exclusion of higher risk borrowers.

Most of these results, or at least its intensity, depends on local markets structure, the degree of asymmetry in the local information, and even local consumers behavior (for instance more mobile clients tends to reduce market power originated from private information). These conditions are very different among countries, in the United Sates, for example, there is a very rich bureau of information (FICO) that reduces Asymmetric information. Brazilian case is intermediate, considering that there are some bureaus, sharing important information that any company can access (paying for that), however, big banks have better information about their clients.

Most of the papers described above, when testing predictions of theoretical models, produce evidence of the model's consequences, such as a less informed player facing a higher default rate. In this paper, we also investigate the private data that cause a disparity in the quantity and quality of information, while addressing the measurement of asymmetric information, and then its consequences.

III Data

The primary source of data is the Annual List of Social Information (Relação Anual de Informações Sociais - RAIS), generated by the Ministry of Labor and Employment (MTE), from which we selected our sample of firms and also collected information about the number of employees. The RAIS data have been collected annually since 1985 and contain detailed information on the employer-employee relationship in the formal sector. The RAIS database contains information on firms (such as sector of activity, year of incorporation and location), workers (gender, date of birth, educational level etc.) and employment terms (such as wages, occupation type, hiring/dismissal dates and reason for layoffs).

We selected a sample of 50,000 firms in 2016, excluding firms with fewer than three employees and those incorporated in 2016. The main goal with these filters was to eliminate newly established firms with insufficient historical information and very small that behave more like consumers than business entities.

The second source of data was Serasa Experian, the largest credit bureau in Brazil. It provided the credit score of its statistical models for our sample of 50,000 firms in June 2016 and June 2017. Its models make use of many records about firms, as well as different kinds of complaints from lenders about delinquency. Its scores provide the main source of information to potential lenders that are looking to attract new clients. This information is paid, but is available to any player in the market, so we considered it to be public information in this study.

The third source of data was the Credit Information System (Sistema de Informações de Credito - SCR) of the Central Bank of Brazil. This dataset includes virtually all loans made by financial institutions operating in Brazil since January 2003. Information on each loan is transmitted monthly and includes: type of loan, debt value (total and delinquent), interest rate, maturity etc. Only the Central Bank and the lending bank can access that information: when a firm opens a current account with a bank, it automatically authorizes this bank to access that information, but other banks cannot access it, so it is private information that winds up producing an advantage for banks with large numbers of clients. Accessing this information allowed us to analyze the reasons for advantages of incumbents, essential to the contribution of this paper.

We matched our sample of 50,000 firms to SCR data in June 2016 and June 2017. From these firms, 23,641 were found in June 2016 and 23,529 were found in June 2017. For the matched firms of each period, we merged the data again to SCR data to obtain all credit information up to one year beforehand - from June 2015 to May 2016 for the sample referenced in June 2016, and from June 2016 to May 2017 for the sample referenced in June 2017 – and one year ahead - from July 2016 to June 2017 and from July 2017 to June 2018 for the sample references, respectively, of June 2016 and June 2017. The "one year before" information is necessary to simulate the database that lenders could access to analyze whether or not to offer credit to a client, and the "one year ahead" information is necessary do construct the performance variable, which means verifying if that firm defaulted in the following 12 months.

Hence, using our first reference as an example, which is June 2016, we built our explained variable as a dummy that assumes 1 if the analyzed firm reached 90 days overdue in any loan from July 2016 to June 2017, and 0 otherwise. This default definition is an international standard, defined by credit regulators. The explanatory variables were built using the information from June 1015 to May 2016, and these variables were basically related to credit characteristics (like type of loan and duration of the debt), and credit history, like number of installments that were not paid on time, maximum share of debt

overdue, maximum days overdue, etc. Both to defined default and arrears, we used some materiality filters, considering only values higher than R$ 1000.

Another important definition was migration. According to our criteria, the main bank of the firm was the one holding the highest proportion of its debt. Consequently, we considered that a firm migrated from its incumbent bank to another one if the highest proportion of its credit was held by one bank in June 2016 and another one in June 2017.

Finally, since the behavior of borrowers at the start of a relationship with a new lender is especially important for this study, considering this group of borrowers is the potential result of adverse selection, and accounted for a small share of a random sample, we added a new sample only with "new" borrowers (those having up to six months of relationship with any financial institution). We did not have all the information about these clients, only the information of the Central Bank, because it is not possible to extract data from the Central Bank to merge with external data. Nevertheless, it was useful to increase our sample in analyses where external data were not necessary, for instance to compare performance of the average new client with other clients.

In the final Sample we have 73,640 firms. From this, 22,530 were in the first six months of the relationship with at least one financial institution ("new clients"). Table 1 compares some statistics of the new clients with other firms ("old clients"). It shows that percentage of new clients with working capital debt is lower, its average (and median) outstanding is lower as well, but interest rate they pay are higher. The only variable that we can direct compare with the market (central bank monthly report of credit) is the average interest rate, which is higher on our sample. This difference, seems acceptable, because it is not so large, and can be explained by the fact that in our sample new clients are overrepresented (because they are important to the adverse selection analysis), and because we have a random sample of firms, which may be different of a random sample of users of credit (that would be more representative of the central bank population).

Table I
Main Statistics of the sample

	New Clients	Other Clients	Total Market
Average Outstanding (R$)	65,186	152,775	-
Median Outstanding (R$)	19,407	68,780	-
Average Interest Rate (working capital line % p.y.)	43.98	34.29	26.7
% using working capital loan	18.0	39.0	-

Table I presents the main descriptive statistics of our sample, comparing clients in the first 6 months of relationship with a financial institution with other clients. Among new clients, the percentage that use credit is lower, while the interest rate is higher.

In Figure I we exhibit the correlation of the Serasa credit scoring (incomplete model) with one of the variables of the Credit Information System (SCR), the debt in arrears. It illustrates that Serasa model has a high correlation with that variable, once the higher the

credit scoring (lower risk), the lower the percentage of firms with some debt in arrears. Nevertheless, it suggests that information of the Credit Information System (which is not public) can add predictive power to Serasa Model, once even among the 20% best classified clients according to this model, 6% have some debt in arrears, which may be too much. Probably, if Serasa model could use SCR information[2] , the scoring of this clients with debt in arrears would be worse. Such hypothesis illustrates part of our investigation, which is measure the degree of asymmetry caused by the fact that only part of market players can access SCR information (each bank can access only the information of its current clients).

Figure I

% FIRMS WITH SOME DEBT IN ARREARS X QUINTILE OF SERASA CREDIT SCORING

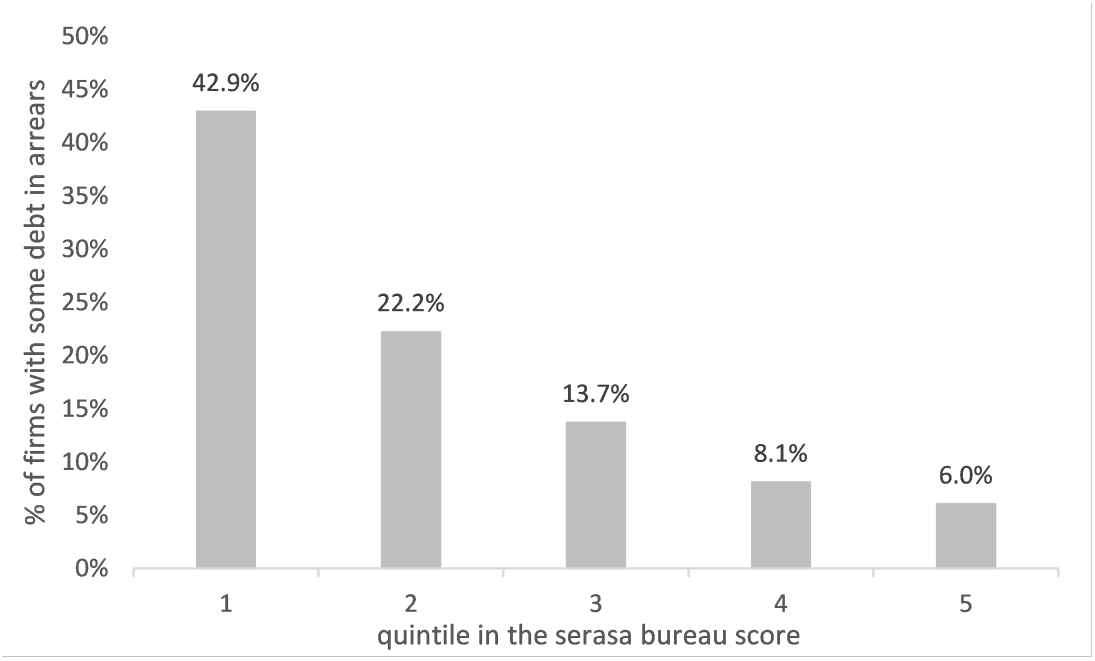

Figure I shows the percentage of firms with loan in arrears by quintile of the Serasa credit score,which is considered the incomplete model in this paper, once it considers only public information. The clients with the best scores are less prone to be in arrears, showing that the model of Serasa has an important correlation with variable of the risk system of the BCB. However, even among the clients with lower risk according to the risk model of Serasa, the percentage of firms with a loan in arrears is relevant at 6% of the sample. This suggests it is possible that having access to information from the credit system of the BCB improves the model of Serasa, which is part of our investigation here

[2]During the writing of this paper, the positive list mechanism proposed by the Central Bank was approved, and its implementation is ongoing. When it is fully implemented, Serasa and other bureaus will be able to access that information, improving their credit scoring

IV Empirical Analysis

IV.A Approach

Our main goal was to demonstrate that there is asymmetric information among lenders competing for borrowers in the Brazilian credit market, and test whether it causes adverse selection effects. There is no doubt that incumbent banks have more information about their clients than new competitors wanting to attract these clients. Nevertheless, our analysis aimed to verify if this extra information is informative enough to produce some advantage to incumbents, in other words, if it allows better-informed lenders to make better credit decisions than new competitors, which can only access public information. Hence, for our approach, to demonstrate the existence of information asymmetry it was necessary to demonstrate that this extra information improves credit decisions.

Therefore, to study the relationship of asymmetric information, adverse selection, and credit conditions, we proceeded in three steps. First, we replicated a credit rating system for incumbents and challengers considering the information that each of them can access. The second step consisted of showing that incumbents' credit rating performs better, giving them an advantage derived from the asymmetric information. Finally, we tested predictions of theoretical models for this environment with asymmetric information, like the existence of adverse selection, and its correlation with the credit systems described here.

- **Step 1. Replicating the credit rating system**: Financial institutions classify their clients by attributing a credit score to each of them. The quality of this credit scoring depends on the quality and availability of information. To simulate models that are used by incumbent lenders in their credit decisions, we used a logit model combining credit bureau information, which is public, with private information (which only incumbent banks can access), to estimate the probability of default (PD) of each firm. We called this the complete model. This model replicates the credit scoring that the better informed lenders use in their credit decisions.

$$Y_i = \beta_0.CreditBureau_i + \beta.Xprivate_i + \mu_i \tag{1}$$

  Where: Y is the performance variable, which is a dummy that equals 1 if the firm does not repay the loan or 0 otherwise; Xprivate is a vector of private information (such as amount and type of credit that the borrower obtained in the past, duration of debt, history of punctuality of installments).

  Competing lenders can access only partial information, obtained from the credit bureau (public information), so we denoted the model using this information as the incomplete model. Credit bureaus usually apply logit models as described above,

but only with parameters based on public information as explanatory variables.

We ranked all firms of our sample based on both models, so all of them had two credit scores: complete and incomplete. We expected the complete model to perform better than the incomplete one, and the measure of the difference between them would reflect the extent of information asymmetry.

- **Step 2. Comparing models to measure asymmetric information**: The advantage of the incumbent banks comes from their additional information. However, this is only translated into a real advantage if this information is useful to improve estimation of probabilities of default of each firm, i.e., only if the models used by incumbents perform better. To investigate that aspect, we used two metrics:

  3.1: Kolmogorov–Smirnov test (KS). This is a nonparametric indicator, commonly used in evaluating the quality of credit scoring models. After ranking the sample with the score under evaluation, it compares the cumulative distribution of good borrowers (who do not default) with the cumulative distribution of bad borrowers, then calculates the maximum difference of these two distributions. The better the model that generates this score, the more bad clients (which defaulted) will be concentrated in the high-risk category and the more good clients will be concentrated in the low-risk category. Thus, the difference of this distribution will be high, so the better the model is, the higher will be the KS statistic calculated with the scoring generated by this model.

  3.2: Tradeoff between risk and approval rate. By comparing two models with different quality applied to the same sample of clients, the better model can allow the financial institution lend money to more clients than the worse model and obtain the same default rate, or to lend money to the same share of clients while achieving a lower probability of default. Hence, a possible measure of quality is the approval rate (share of clients that received loans) for each level of default probability. Complementary measures is the expect default rates of each level of approval rate.

  These models were developed in the first reference (in June 2016) and applied to the validation sample, which was that with our second reference date (the year 2017), to avoid overfitting problems. Therefore, we considered that incumbents are better informed, generating asymmetric information, if the model developed including private information (complete model) performed better, according to criteria described above, in the validation sample. Finally, we split our sample into clusters to analyze if the information works better in some group of clients.

- **Step 3. Testing theoretical predictions**: We looked for evidence that incumbent lenders keep lower risk borrowers, while challengers get riskier ones, implying a

higher default rate and worse credit conditions for clients who migrate from the incumbent bank to another one (and for new clients in general). Our definition of main bank (incumbent bank) was the one to which the firm owed the most money. Hence, firms that migrated from their incumbent banks to challengers were those that owed the most to one bank in the first reference period (June 2016) and to another lender one year later.

We tested if the propensity of this migration was correlated with the credit score (from the complete and incomplete models), whether the clients that migrated had a higher default rate compared to other clients, and if new clients receive worse credit conditions (higher interest rates and lower credit limits). To check statistical significance, we tested these predictions with a two sample t-test with unequal variances.

IV.B Results

As mentioned in the previous section, our first step was to run a logit model including all available information, private and public (Appendix1, equation 5), replicating models used by incumbent banks. Hence, we ranked all firms of the sample with a credit score generated by this model. Figure II illustrates the performance of this model in the validation sample (June of 2017). This rank can significantly differentiate good clients (who do not default) from bad clients (who default), since defaulters are concentrated in worse credit score categories while good clients are concentrated in better credit scores (KS of 53.6). In addition, as expected, the better the credit score, the higher the default was: in the 20% worst ranked firms, the observed default rate was 18%, while in the 20% best ranked firms the rate was 3%.

Figure II
COMPLETE MODEL DISCRIMINATION POWER

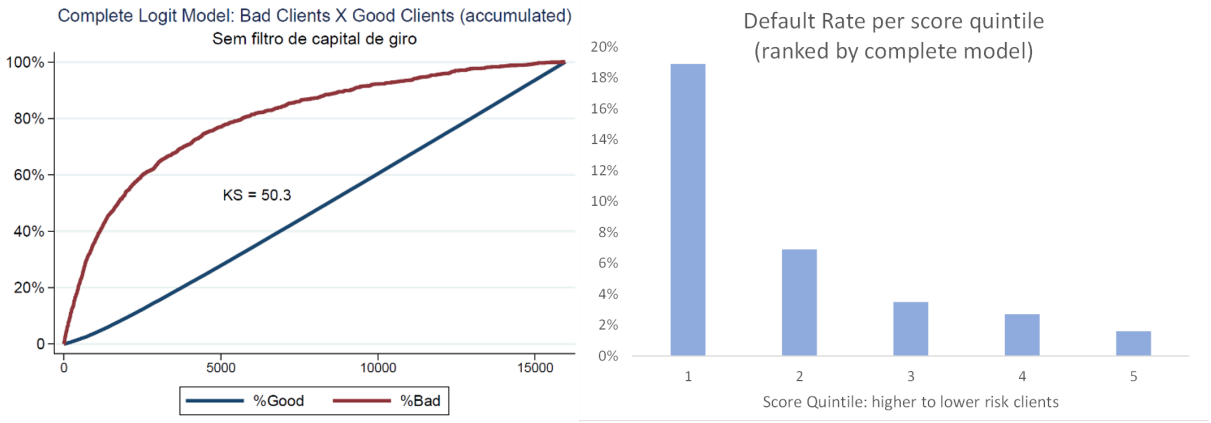

Figure II depicts the discrimination power of our complete model, which incorporates, besides credit score information from Serasa, information from the BCB's risk system. Figure II-a compares the cumulative distribution of clients that were in default in the 12 months after the reference date with the cumulative distribution of clients that were not in default (in order of the credit score generated by this model). The curve of defaulting clients (bad) has a strong concentration in the lower ratings (worse classifications), while curve of the good clients is more evenly distributed. The greatest distance between the curves is 50.3 (KS indicator). Figure II-b shows the percentage of clients in default by quintile of this same complete model. The default is higher among borrowers with worse ratings, and the default rate declines as the rating improves, showing the good performance of the model to rank clients.

The next step was to compare the quality of this complete model that incorporates private information with the incomplete model that considers only public information (from the Serasa credit bureau). The Kolmogorov–Smirnov test result of the incomplete model (35.9) is 18 points lower than that of the complete model, suggesting the discrimination power of the incomplete model is significantly lower (Figure III-a). Figure III-b illustrate the consequence of this better discrimination power in terms of default rate per credit scoring: clients that are in the 20% worse credit score according to complete model have higher default rate than clients that are in the 20% worse credit score according to incomplete model, while, clients that are in the 20% better credit score according to complete model have lower default rate than clients that are in the 20% lower credit score according to incomplete model.

Figure III

(A) Incomplete Model KS, (B)% Default (complete X incomplete)

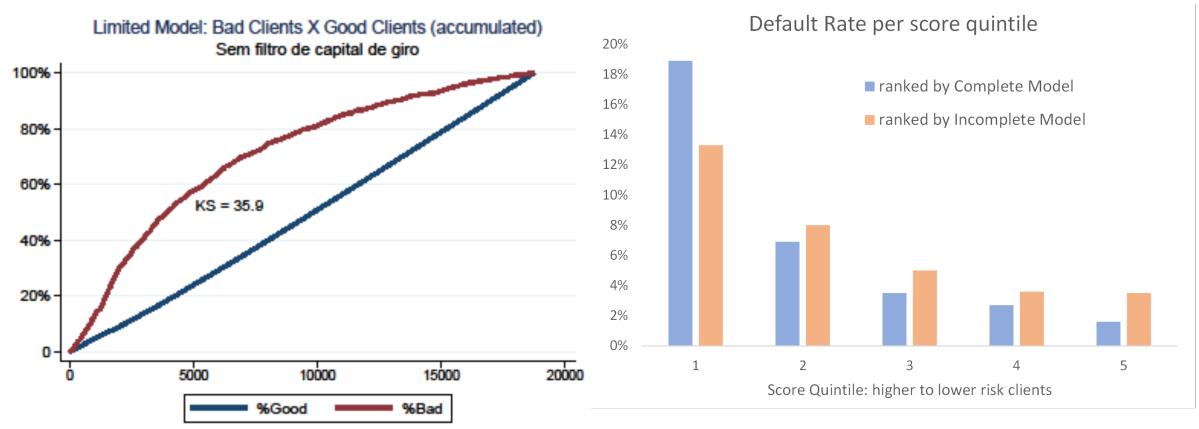

Figure III depicts the discrimination power of our incomplete model, and compares it with that of the complete model. The KS (greatest distance between the cumulative distributions of good and bad clients, as described in Figure II is 35.9, worse than that exhibited by the complete model in Figure II. Besides this, Figure III-b shows that clients classified in the highest quintiles of the incomplete model have a higher default rate than those classified in the best groups of the complete model, while those classified in the lowest quintiles according to the incomplete model have a lower default rate. This provides complementary evidence that the complete model ranks clients better regarding risk (i.e., the private information that only it considers adds value).

To translate this better performance of the complete model in terms of economic indicators, by measuring the advantage of the incumbents, we computed some comparative statistics, simulating different credit policies and evaluating the number of companies that would receive loans depending on the model (incomplete or complete). To do that, we first calculate the accumulated default by each percentage of total borrowers, for the complete and incomplete model, to estimate the tradeoff between higher approval rate and lower default rate for each model. Figure IV presents the resulting curves, illustrating that, for any percentage of clients in the sample, the accumulated default rate is lower if clients are ranked by the complete model than if they are ranked by incomplete level[3] . That means that, for any chosen target default rate, the lender that is using the complete model can reach more clients than lender using incomplete model, or, in the other point of view of the same graph, for a given approval rate, using the better model allows to reach lower default rates. In addition, Note that at the start the curves behave erratically, not always rising, which illustrates the difficulty of classifying clients with very low default risk.

Table II illustrates some options in the tradeoff between approval rate and default rate. If the credit policy, for instance, is set to reach 3% expected default rate, 12.9% of the clients would be approved under the restricted model, and 66.6% would be approved under the complete model. This means that if a new competitor could access the complete model

[3]Except for the 100% of borrowers point in the curve, because since the sample is the same, if we consider 100% of the sample, the default rate is the same, no matter the model.

instead of only using public information, the subsample of clients that could be eligible for credit (considering a 3% default target) would be 5 folds higher. The lower the default target, the greater the improvement (using the 1% default target, the sample would be 16.7 folds higher). This result illustrates that if the private information was shared with new competitors, this could increase the number of borrowers without increasing the risk, indicating the value of this extra information, or the degree of information asymmetry. Table II-b illustrate the other possibility of measuring the same gain with the better model, which is the potential reduction in the default rate for the same approval rate: if the target is to approve 60% of the potential clients, for instance, using the complete model would reduce 40% the default rate.

Figure IV

Accumulates default rate (complete X incomplete model)

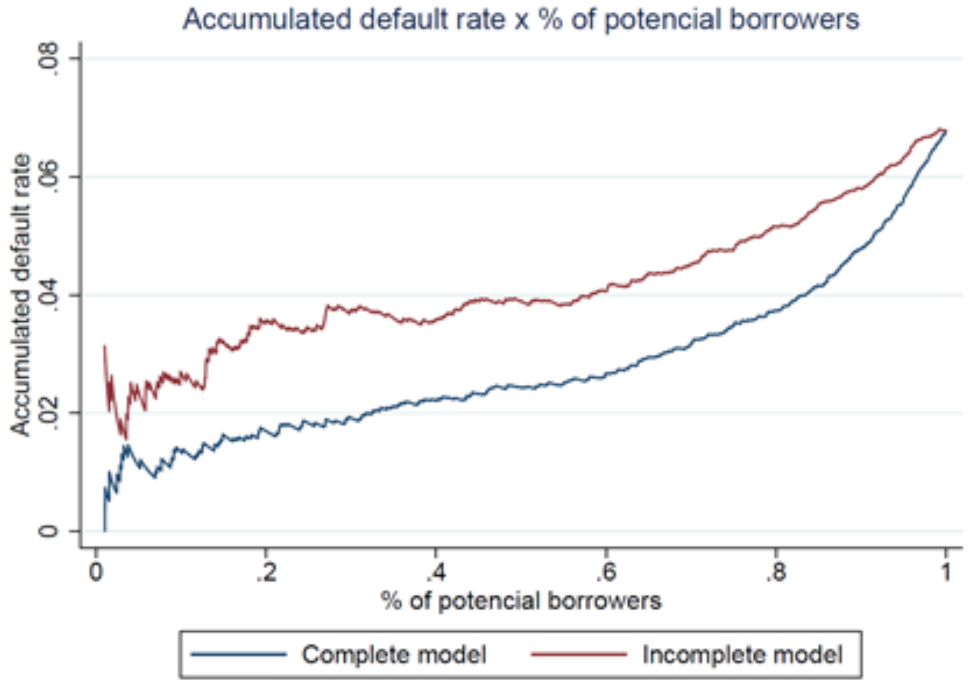

Figure IV depicts the cumulative default curves according to the percentage of clients ordered by the two models, complete and incomplete The cumulative default curve of the clients ranked by the complete model is always below that of the incomplete model, because by ranking the clients better, for any percentage of the best clients, the default rate of the complete model is lower, except when considering 100% of the sample, in which case the curves meet, coinciding with the mean default rate of the sample. Note that at the start the curves behave erratically, not always rising, which illustrates the difficulty of classifying clients with very low default risk.

Table II

DEFAULT RATE X % BORROWERS (A) AND % BORROWERS X DEFAULT RATE (B)

accumulated	Share of borrowers			Share of	Accumulated Default Rate		
default rate	(a) Incomplete	(b) Complete	(b) / (a)	borrowers	(a) Incomplete	(b) Complete	(b) / (a)
2%	1,9%	32,2%	16,8	**20%**	3,5%	1,6%	0,5
3%	12,9%	66,6%	5,2	**40%**	3,6%	2,2%	0,6
4%	58,0%	83,4%	1,4	**60%**	4,0%	2,6%	0,6
5%	78,1%	91,4%	1,2	**80%**	5,0%	3,7%	0,7
6%	91,7%	96,5%	1,1	**100%**	6,7%	6,7%	1,0

Table II illustrates the results of a comparative static exercise, showing the outcomes of different hypotheses about credit policies defined by the financial institution in relation to the target for the default rate or the rate of credit approval of clients. Table II-a shows the percentage of clients approved for different levels of the default rate target. Table II-b, in turn, shows the complementary view, namely the default rate for different approval rate targets. The comparison of the results between the complete and incomplete model shows that using the model that contains more information allows increasing the approval rate without increasing the default rate, or reducing the default rate without diminishing the approval rate.

Nevertheless, new competitors cannot use this private information, so they are restricted to using the incomplete model. Hence, due to this asymmetric information, they are expected to face adverse selection. Indeed, we found that clients that are more likely to leave their incumbent banks for challengers are riskier, especially the ones that are classified as high risk in the complete model and low risk in the restricted model [4] – incumbent bank knows they are high risk while challenger banks do not (and may offer relatively good credit conditions). The share of clients with those characteristics that change their main bank was 28%, while 22% of other clients would do so. Additionally, the default rate of the clients that change was higher than for other borrowers (12.5% X 9.7%). These differences were tested with a two-sample t-test with unequal variances, reported in Appendix 2, and were statically significant.

[4]Defined as clients in the group of 25% with the worst credit score according to the complete model and 75% best credit scores in the incomplete model.

Figure V
MIGRATION X CREDIT RATING (A) MIGRATION X DEFAULT RATE (B)

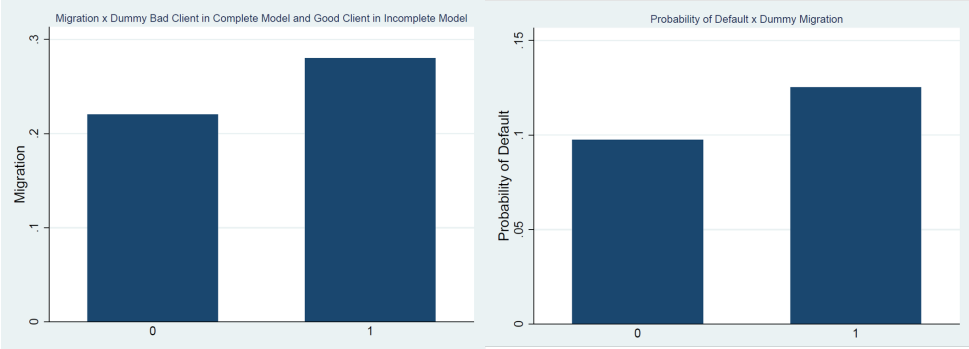

Figure V-a indicates that the clients that are better ranked by the incomplete model (public information) but poorly classified by the complete model (public 1 in figure V-a) are more likely to change their main bank. Figure V-b shows that clients that migrate from one bank to another (public 1 in figure V-b) have a higher default rate than the other banks in the sample. The two results are indications of the presence of adverse selection.

Furthermore, our results suggest that new clients (in the first 6 months of the relationship with the bank) are more likely to default, which is consistent with the fact that worse clients move more, opening more new accounts[5] . Clients in this period were 100% more likely to default than older clients, a difference which was statically significant according to the two-sample t-test.

[5]Evaluating the performance of new clients is complementary to evaluating the performance of clients that migrate from their incumbent bank, because the migration definition requires a definition of the main (incumbent) bank, which we define as bank to which the borrower owes the most. So, if we looked only on this criterion of migration, we could be missing clients that migrated but did not have credit from their former bank. Analyzing all clients at the beginning of a relationship with a financial institution acts as a robustness check of the theory that migrating clients are riskier.

Figure VI

DEFAULT RATE - CLIENTS UP TO 6 MONTHS RELATIONSHIP X OTHERS, (A): IN 2016 ; (B): IN 2017

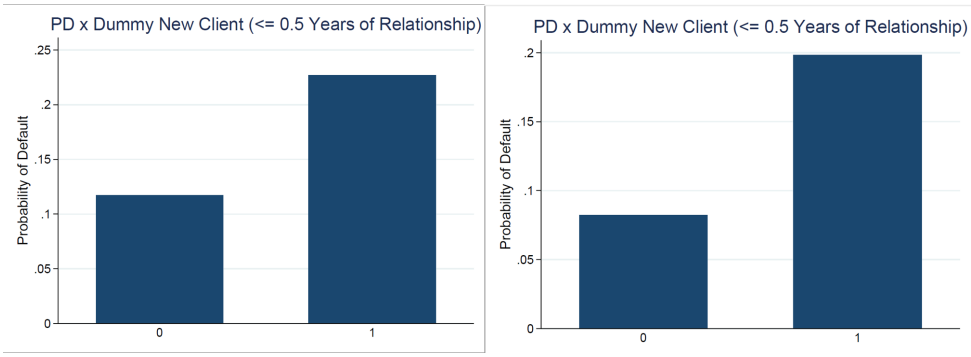

Figure VI shows that new clients (those with relationship up to six months with their main financial institution) present a significantly higher default rate: in the two references database analyzed, the default rate of new clients is more than twice that of longstanding clients. This finding is also consistent with the existence of adverse selection in the process of attracting clients.

The third finding, which is a consequence of the second, is that new clients have worse credit conditions compared to old clients. Figure VII shows that limits are higher and interest lower for older clients. Once more this result is statically significant according to two-sample t-test.

Figure VII

DISTRIBUTION OF CREDIT CONDITIONS FOR NEW CLIENT X OTHERS, (A): CREDIT LIMIT ; (B): WORKING CAPITAL INTEREST RATE (%P.Y.)

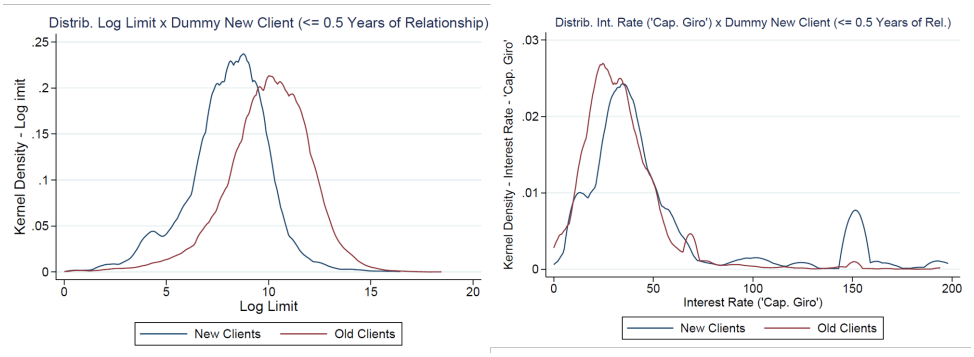

Figure VII shows the distribution of credit conditions to new clients compared to old ones. The graphs indicate that the new clients receive lower credit limits (VII-a) and have to pay higher interest rates (VII-b).

In addition, to better understand information that improves the credit risk model, giving advantage to the incumbent, we ordered the main variables according to the information value criteria [6] . Table III shows that the top 5 variables according to information

[6]

$$\sum \left(\%default_i - \%nondefault_i\right).ln\left(\frac{\%oftotaldefault_i}{\%oftotalnondefault_i}\right) \quad (2)$$

value are related to past negative behavior (except credit bureau record, which is the variable with the highest information value), suggesting they are more important than others, such as characteristics of the transaction (for instance duration). Indeed, simulating an intermediate model using only the public information and these top five variables (Figure VIII, Intermediate 1), we found a KS value close to that of the complete model, while when applying another intermediate model excluding these variables and including the complementary group of variables, the estimated KS value was close to that of the incomplete model (Figure VIII, Intermediate 3).

Table III
Information Value of Main Variables

Main Variables	Information Value
Serasa Credit Bureau	1.33
Max % overdue (last 12 months)	1.24
Current % outstanding overdue	1.08
numbers of months with overdue payments (last 12 months)	1.06
max overdue period (last 12 months)	1.01
current overdue	1.00

Notes: Information value (IV) is a univariate measure of the discrimination power of the variable in question in relation to an event, in this case default. The higher the IV, the more informative the variable will be in relation to the event in question. Table III reports the variables with the highest IV. The table's results show that all the main predictive variables, except the Serasa score, are related with a negative credit history (current position or past history of belated loan installments).

Each I is a percentile of the original continuous variable.

Figure VIII

KS Value of Models with Different Groups of Variables

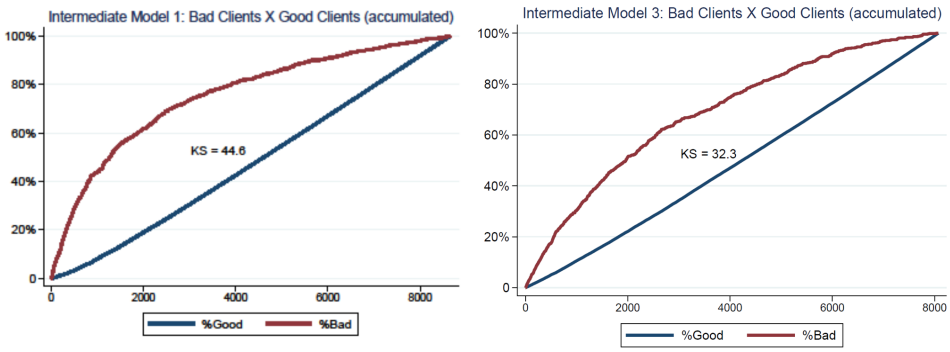

Figure VIII shows the discrimination power of the model by the KS criterion, considering different subgroups of the available variables. The Serasa credit score is included in both because it is public, and our intention here is to illustrate which subset of private variables increases the predictive power. In Figure VIII-a, the variables added to the model are those related to the track record of arrears, while those added to the model in Figure VIII-b are the others (basically related to the characteristics of the loan transactions and firms). The results suggest that the variables related to arrears are responsible for most of the gain, while the set of other variables adds little value, because the KS of the model including them is equally low as the model that only contains public information (which is basically the Serasa credit score).

In the light of that result, we divided our sample into groups to evaluate those in which that data are more informative, and estimated the KS statistic for each subgroup. The results (Table IV) suggested that both models, complete and incomplete, have poor performance when only considering clients without negative information, which always pay installments on time. This is consistent with the fact that the most important variables in our logit equation are those related to past negative behavior. In addition,, the performance when including private information was better in the samples including high-risk clients. Once again, this is consistent with the type of variable that is most significant for the model. The highest difference in the performance was in the subsample with clients that are borrowers (credit outstanding > 0), considering low and high-risk clients

Table IV

KS Indicator for Subsamples of Clients

	outstanding > 0	outstanding ≥ 0	no arrears and outstanding > 0	no arrears and outstanding ≥ 0
Complete	44.8	50.3	23.8	29.3
Incomplete	27.6	35.9	21.9	28.2
Difference	17.2	14.4	1.9	1.1

The table IV shows the discrimination power, measured by the KS criterion, of the complete and incomplete models for different samples. The first column only includes clients that obtained credit, the second column covers the complete sample, the third includes all clients without any history of late payment, and the fourth contains clients with an outstanding balance but without a history of credit denial. The results indicate that when considering only clients without history of arrears, the performance of the model declines significantly, strengthening the hypothesis that a good part of the model's discriminatory power comes from differentiating clients with a record of some type of arrears from the rest.

We finally checked if changing the approach from logit to random forest could improve the results, but we did not find any evidence in that sense. Indeed, despite the very high KS statistic in the developing sample (June 2016), our random forest model's performance declined substantially in the validation sample (Figures IX-a and IX-b), illustrating the risk of wrong conclusions due to overfitting. Indeed, considering only public information (Figure X), the performance of the random forest was very poor (KS 18), the worst one among models we evaluate.

Figure IX

Indicator Using Random Forest Model (including private information)

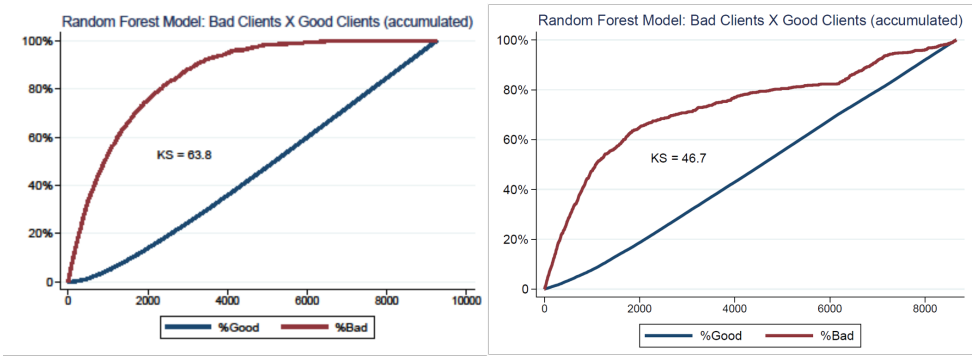

Figure IX illustrates the discriminatory power, measured by the KS indicator, of a model estimated by the random forest method, using the complete database, including private information. Figure 9a depicts the result in the developing sample, where the model performs very well. However, in the developing sample, the performance drops substantially, illustrating the risk of overfitting with this type of approach.

Figure X
KS Indicator Using Random Forest Model (only public information)

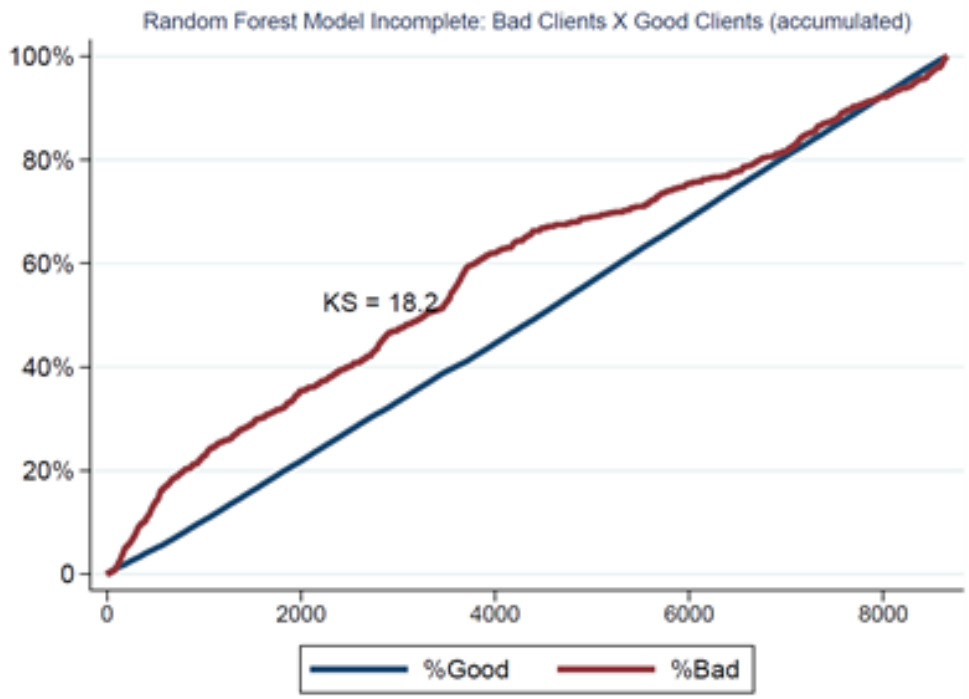

Figure X shows the discriminatory power of the model estimated via the random forest method only using public information. The performance is poor, indicating that the alternative modeling techniques do not serve to replace, or even partially offset, the restricted access to information.

V Conclusion

Our results produce evidence that private information that only incumbent banks can access in Brazil, gives an important advantage to them compared to new competitors, since credit models including that information worked substantially better than models using only public information. The KS indicator of the model including private information was 14 points higher than in the model excluding this information. According to our estimates, if this private data were shared with new competitors, they could increase in more than 100% its potential clients, without increasing their default rate. The benefits of using this private information, in terms of increasing the potential clients, are negatively correlated with the risk of the client (we estimate higher potential increases in low risk clients).

We found evidence that without accessing this information, new competitors face adverse selection problems because riskier clients were 27% more likely to migrate from their original main bank, and after migrating, presented a 29% higher default rate. As expected, consequently, new clients tended to obtain worse credit conditions, with higher interest rates and lower credit limits.

Furthermore, when exploring private information value in different clusters of clients, we found signs that the usual available dataset in the Brazilian market is useful to better discriminate high-risk clients, but not to identify clients with very low risk, once even the complete model has poor performance when only considering clients without negative information. Finally, we tested a random forest model to investigate if a more recent modeling approach could offset the lack of private information, but we did not find evidence in that direction, so it seems that more information is the most important factor.

It is important to note that we did not address the question of whether sharing information could increase total credit granted. Our results are related to the pool of potential clients available to new competitors, illustrating the potential increase in competition, but it is not possible to predict the net effect, considering that some high-risk clients may not receive credit anymore. In other words, low-risk clients would probably be benefited due to more competition, but high-risk clients might face more restricted credit conditions, but we cannot predict the net effect.

Finally, the Brazilian Central Bank is already addressing the information sharing issue, with the aim of increasing competition, by establishing a publicly available "positive credit list", which will include most of the data considered as private in this paper (for consumers, this process is more advanced), and by implementing open banking, which is a very ambitious project to share information. The success in implementing these projects will probably mitigate the effects we found in this study.

References

[1] JOHANNES STROEBEL. Asymmetric information about collateral values. *The Journal of Finance*, 71(3):1071–1112, 2016.

[2] C. Fritz Foley, A. Magdalena Hurtado, Andrés Occhipinti Liberman, and Alberto Sepúlveda. The effects of information on credit market competition: Evidence from credit cards. *Banking & Insurance eJournal*, 2020.

[3] Giovanni Dell'Ariccia, Ezra Friedman, and Robert Marquez. Adverse selection as a barrier to entry in the banking industry. *The RAND Journal of Economics*, 30(3):515–534, 1999.

[4] Xavier Vives. Trade association disclosure rules, incentives to share information, and welfare. *The RAND Journal of Economics*, 21(3):409–430, 1990.

[5] Marco Pagano and Tullio Jappelli. Information sharing in credit markets. *The Journal of Finance*, 48(5):1693–1718, 1993.

Appendices

Appendix Table 1 : Logit Models

Table V
LOGIT MODELS

t statistics in parentheses

	(1) Incomplete Model	(2) Intermediate Model 1	(3) Intermediate Model 2	(4) Intermediate Model 3	(5) Complete Model
serasa_score	-0.00467***	-0.00344***		-0.00442***	-0.00340***
	(-40.90)	(-28.11)		(-33.84)	(-24.50)
Months overdue		0.0192***			0.00729*
		(5.68)			(1.99)
max_days overdue		0.192***			0.164***
		(8.70)			(6.09)
max_%_overdue		-0.0282			0.369
		(-0.10)			(0.97)
Current % overdue		0.400			2.100*
		(0.81)			(2.52)
Current days overdue		0.537***			0.545***
		(12.67)			(10.36)
% debt shor term			-0.172	-0.0343	-0.00179
			(-1.46)	(-0.54)	(-0.05)
Duration			0.000482***	0.000456***	0.000362***
			(7.87)	(6.79)	(5.15)
overdraft / total debt			0.903***	0.311*	0.277
			(7.27)	(2.28)	(1.90)
Limit utilization			0.00000112	-0.00000518	-0.00000908
			(0.23)	(-0.79)	(-1.35)
Relationship period			-0.00661	-0.0000502	0.00165
			(-1.20)	(-0.01)	(0.27)
Bndes line / total debt			-1.120***	-0.489**	-0.366
			(-6.26)	(-2.63)	(-1.93)
Interest rate			-498.4	-1470.6	-1227.8
			(-0.68)	(-1.12)	(-0.83)
Number of Financial institutions			0.337***	0.270***	0.213***
			(22.16)	(16.25)	(11.98)
Age of the firm			-0.0219***	-0.00812	-0.0118**
			(-5.32)	(-1.93)	(-2.61)
_cons	-0.623***	-1.390***	-2.905***	-1.437***	-1.878***
	(-17.20)	(-28.22)	(-39.31)	(-18.00)	(-21.67)
N	19914	19800	17173	17173	17173
Pseudo R2	0.176	0.249	0.0675	0.204	0.256
LR chi2	2280.5	3214.1	719.9	2177.9	2733.6
Prob>chi2	0	0	3.57e-149	0	0

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

The table presents the main models estimated. The first equation only includes public information (basically the Serasa score), and the fifth equation specifies the complete model, incorporating all the private variables. The intermediate columns contain the equations that only include subgroups of private information, in an attempt to identify which variables in fact add explanatory power.

Table VI

Two Sample t-Tests with unequal variance

	Default Rate	Default Rate	Credit Limit (ln)	Interest Rate	Migration Rate
New Clients	0.227		8.02	51.95	
Old Clients	0.117		9.95	34.17	
Low Risk Incomplete / High Risk Complete					0.2801
Others					0.2203
Migrated Clients		0.0975			
Other Clients		0.1253			
Difference	0.110	0.0278	-1.93	17.78	0.0598
P (diff=0)	0.000	0.001	0.000	0.000	0.0181

The table presents the results for the two sample T-Test with unequal variance, illustrating that all mentioned means differences were statically significant