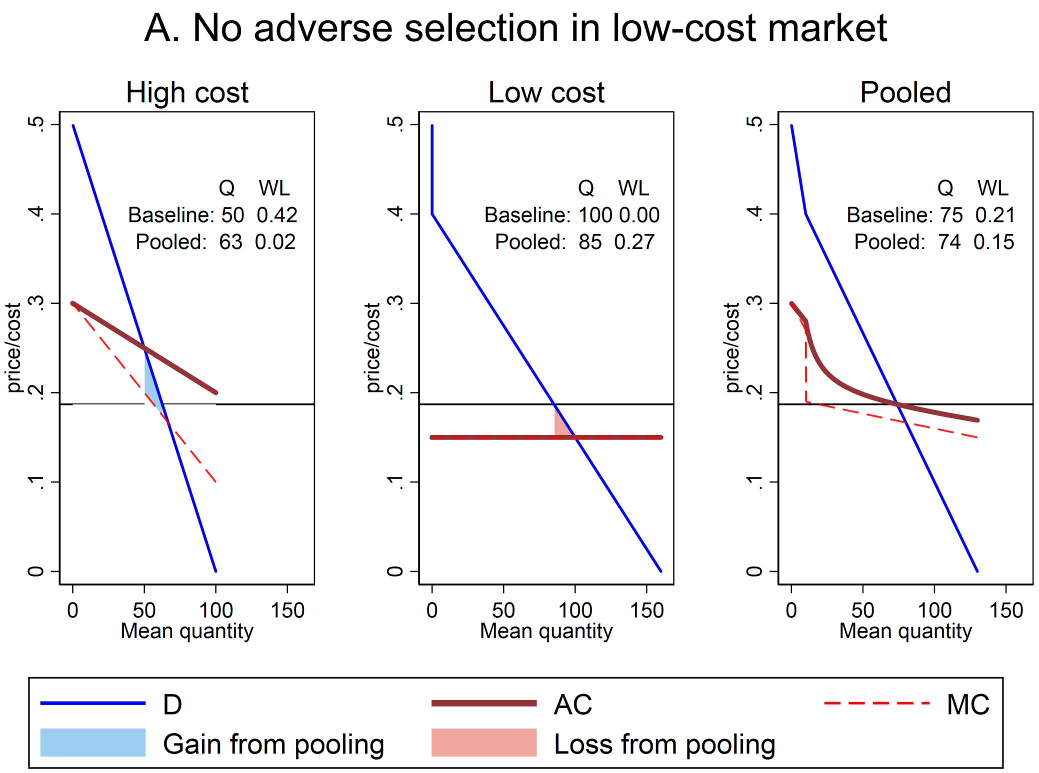
Internet Appendix

A Additional Results

Figure A1: Simulated separating and pooling equilibria


A. No adverse selection in low-cost market

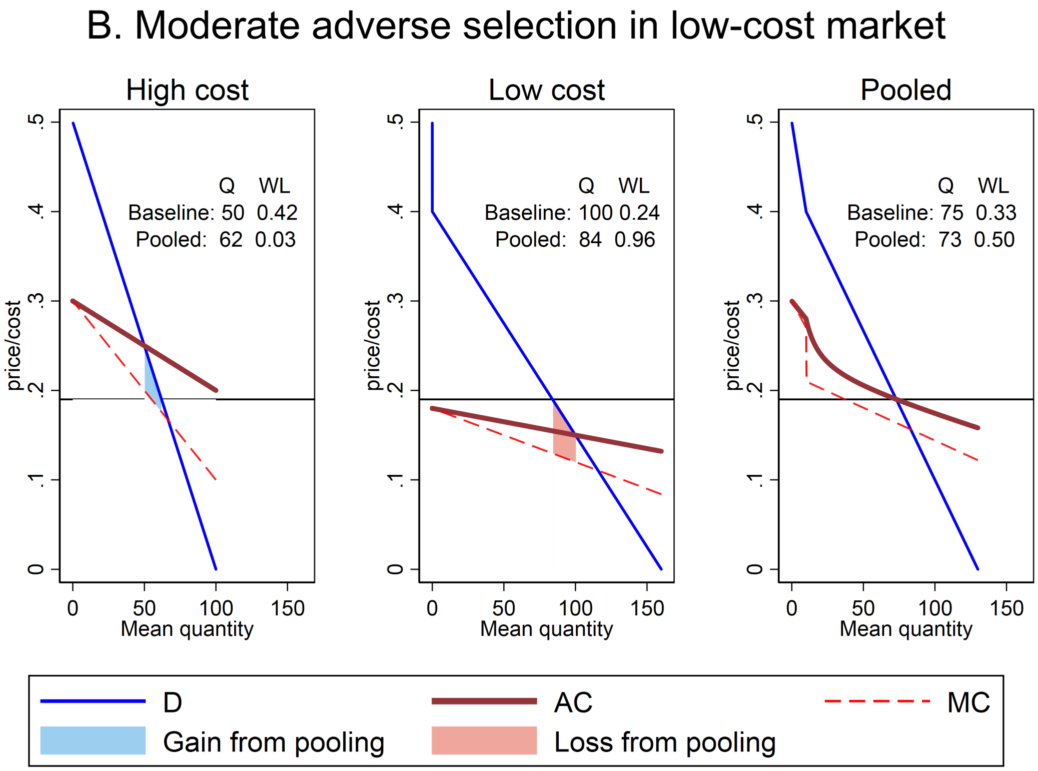
High cost

Low cost

Pooled

Each panel shows simulated separate-market (left two panels) and pooled market (right panel) equilibria under different assumptions about market sizes and slopes of average cost and demand curves in the high-cost and low-cost market. Text in each panel displays aggregate market quantities transacted ("Q" column) and welfare loss relative to the efficient quantity ("WL" column) under the separate ("baseline") equilibrium and the "pooled" equlibrium. To see changes in aggregate welfare from pooling compare the "pooled" and "baseline" welfare loss columns in the rightmost panel of each row. Pooling causes welfare to rise in Panel A, fall in Panel B, and rise in Panel C. The number of individuals in high- and low-cost market are normalized to one. Separate equilibrium price and quantity (p, q) are the same in each panel, with $(p, q) = (0.25, 50)$ in the high market and $(p, q) = (0.15, 100)$ in the low-cost market. $\frac{dAC_0}{dq}$ and $\frac{AC_1}{dq}$ are the slopes of average cost curves in the low- and high-cost markets, respectively, with analogous definitions for demand curves. Slopes parameters vary across rows as follows. Panel A: $(0, -0.001, -400, -200)$ Panel B: $(-0.0003, -0.001, -400, -200)$. Panel C: $(0, -0.001, -400, -350)$. See text for model details.

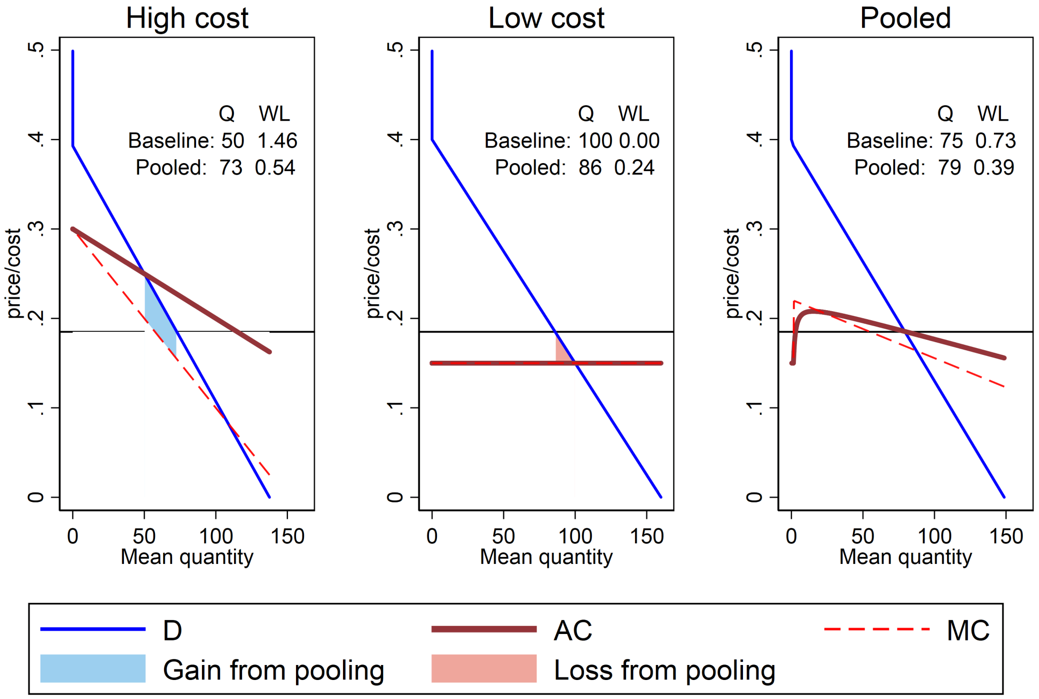
Figure A1: (Cont'd) Simulated separating and pooling equilibria



Each panel shows simulated separate-market (left two panels) and pooled market (right panel) equilibria under different assumptions about market sizes and slopes of average cost and demand curves in the high-cost and low-cost market. Text in each panel displays aggregate market quantities transacted ("Q" column) and welfare loss relative to the efficient quantity ("WL" column) under the separate ("baseline") equilibrium and the "pooled" equlibrium. To see changes in aggregate welfare from pooling compare the "pooled" and "baseline" welfare loss columns in the rightmost panel of each row. Pooling causes welfare to rise in Panel A, fall in Panel B, and rise in Panel C. The number of individuals in high- and low-cost market are normalized to one. Separate equilibrium price and quantity (p, q) are the same in each panel, with $(p, q) = (0.25, 50)$ in the high market and $(p, q) = (0.15, 100)$ in the low-cost market. $\frac{dAC_0}{dq}$ and $\frac{AC_1}{dq}$ are the slopes of average cost curves in the low- and high-cost markets, respectively, with analogous definitions for demand curves. Slopes parameters vary across rows as follows. Panel A: $(0, -0.001, -400, -200)$ Panel B: $(-0.0003, -0.001, -400, -200)$. Panel C: $(0, -0.001, -400, -350)$. See text for model details.

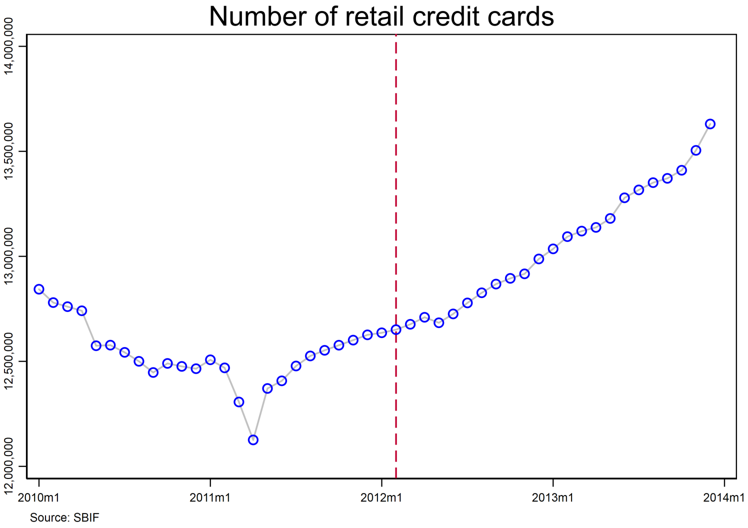
Figure A1: (Cont'd) Simulated separating and pooling equilibria

C. No adverse selection in low-cost market, less elastic demand



Each panel shows simulated separate-market (left two panels) and pooled market (right panel) equilibria under different assumptions about market sizes and slopes of average cost and demand curves in the high-cost and low-cost market. Text in each panel displays aggregate market quantities transacted ("Q" column) and welfare loss relative to the efficient quantity ("WL" column) under the separate ("baseline") equilibrium and the "pooled" equlibrium. To see changes in aggregate welfare from pooling compare the "pooled" and "baseline" welfare loss columns in the rightmost panel of each row. Pooling causes welfare to rise in Panel A, fall in Panel B, and rise in Panel C. The number of individuals in high- and low-cost market are normalized to one. Separate equilibrium price and quantity (p, q) are the same in each panel, with $(p, q) = (0.25, 50)$ in the high market and $(p, q) = (0.15, 100)$ in the low-cost market. $\frac{dAC_0}{dq}$ and $\frac{AC_1}{dq}$ are the slopes of average cost curves in the low- and high-cost markets, respectively, with analogous definitions for demand curves. Slopes parameters vary across rows as follows. Panel A: $(0, -0.001, -400, -200)$ Panel B: $(-0.0003, -0.001, -400, -200)$. Panel C: $(0, -0.001, -400, -350)$. See text for model details.

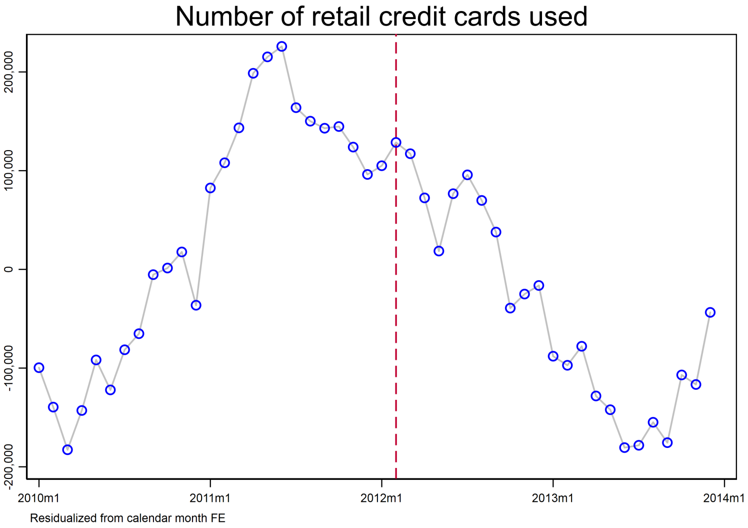
Figure A2: Stock of retail credit cards over time



Source: SBIF

Stock of retail credit cards by month. Time of deletion policy noted with vertical line.

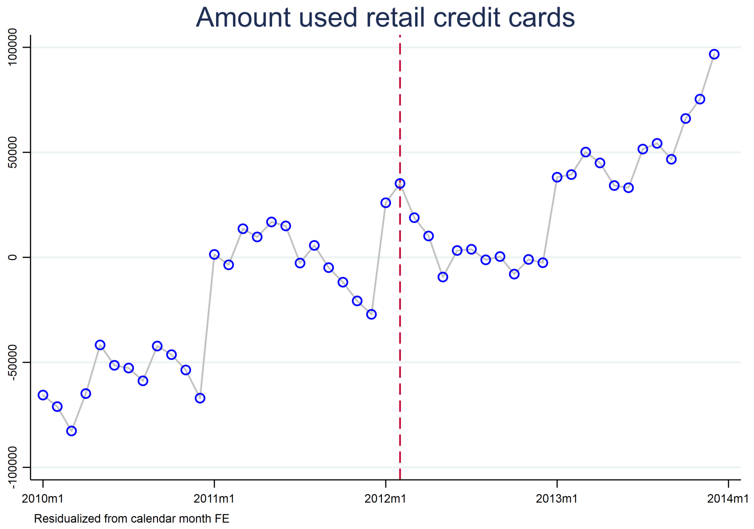
Figure A3: Retail credit cards in use over time



Residualized from calendar month FE

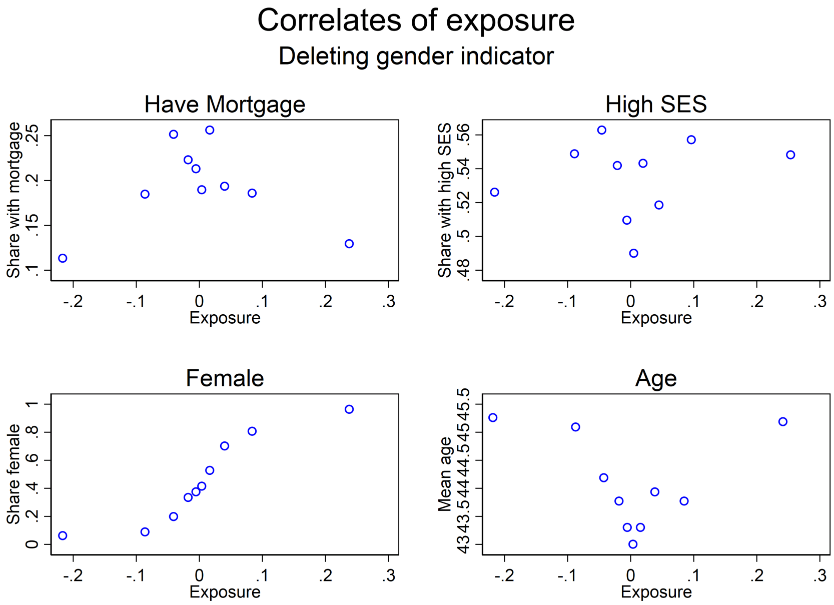
Number of retail credit cards used by month. Time of deletion policy noted with vertical line. Source: SBIF.

Figure A4: Number of retail credit card uses over time



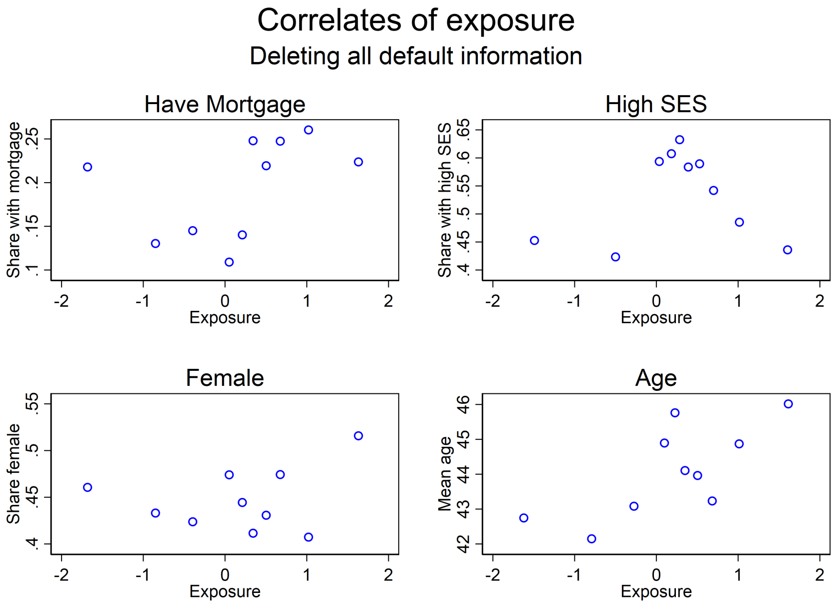
Amount of retail credit purchases by month. Time deletion policy noted with vertical line. Source: SBIF.

Figure A5: Correlates of exposure under counterfactual deletion policy – no gender



Binscatters of correlates of exposure under the counterfactual policy of deleting a gender indicator. See text for details.

Figure A6: Correlates of exposure under counterfactual deletion policy – all default information



Binscatters of correlates of exposure under the counterfactual policy of deleting all default information. See text for details.

Table A1: Difference-in-difference predictions using long run cost measures

	Low cost market			*High cost market*		
	Predicted Cost	Average Cost	New Borrowing	Predicted Cost	Average Cost	New Borrowing
Jun. 2010	0.01	0.00	-7.09^{*}	0.03	0.03	-5.68^{+}
	(0.02)	(0.02)	(3.05)	(0.05)	(0.05)	(3.23)
Dec. 2010	0.01	0.01	-2.11	0.02	0.01	0.30
	(0.02)	(0.02)	(3.52)	(0.05)	(0.05)	(3.25)
Jun. 2011	0.00	0.00	0.00	0.00	0.00	0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Dec. 2011	0.25^{***}	0.12^{***}	-13.28^{**}	-0.30^{***}	0.04	17.98^{***}
	(0.02)	(0.02)	(4.21)	(0.04)	(0.04)	(3.47)
Elasticity		0.48	-0.24		-0.12	-0.36
Dep. Var. Base Period Mean	0.08	0.08	214.70	0.14	0.14	165.09
N Clusters	307	307	307	299	299	300
N Obs.	2,929,133	4,961,674	13,163,613	1,486,567	2,519,339	8,117,207
N Individuals	1,844,615	2,394,399	4,373,700	1,104,246	1,571,258	3,422,263
N Exposed Individuals	452,132	765,941	1,967,865	79,572	134,306	589,628

Significance: $^{+}$ 0.10 * 0.05 ** 0.01 *** 0.001. Difference and difference estimates from equation 2. Table is identical to Table 4 but uses a one-year ahead measure of default to compute predicted costs. See section 5.5 for details. The first two columns report the difference-in-difference estimated effect of deletion on outcome variables listed in column headers, while the third and fourth estimate the dif-in-dif effect on the different exposure-defined markets. We take the log of 'Predicted cost' for estimation but report the base period mean in levels. 'Elasticity' is borrowing effect scaled by base period outcome mean and predicted cost effect. '*N* exposed individuals' reports the number of individuals not in the 0 group included in the regression sample in the treatment period. Since some individuals appear in multiple snapshots we report both individuals and observations. Standard errors clustered at market level.

Table A2: Distribution of deletion effects using long run cost measures

	Separate	Pooled	Difference
Low cost market			
Predicted cost	0.065	0.081	0.016
Average cost	0.065	0.073	0.008
New borrowing (1000s CLP)	234.779	222.246	−12.533
Welfare loss (1000s CLP)	1.711	2.138	0.427
Aggregate new borrowing (Bns CLP)	447	424	−24
Aggregate welfare loss (1000s CLP)	3,261,672	4,075,579	813,908
			24.95%
N individuals	1,905,946	1,905,946	1,905,946
High cost market			
Predicted cost	0.120	0.081	−0.039
Average cost	0.120	0.125	0.005
New borrowing (1000s CLP)	112.490	132.079	19.589
Welfare loss (1000s CLP)	0.140	1.128	0.988
Aggregate new borrowing (Bns CLP)	67	78	12
Aggregate welfare loss (1000s CLP)	83,086	668,656	585,570
			704.77%
N individuals	592,732	592,732	592,732
Combined			
Average price	0.072	0.081	0.008
Average cost	0.072	0.081	0.008
New borrowing (1000s CLP)	205.770	200.857	−4.913
Welfare loss (1000s CLP)	1.339	1.899	0.560
			41.84%
Aggregate new borrowing (Bns CLP)	514	502	−12
Aggregate welfare loss (1000s CLP)	3,344,758	4,744,236	1,399,478
			41.84%
N individuals	2,498,678	2,498,678	2,498,678

This table describes changes in key welfare metrics before and following deletion, with inputs to the theoretical framework using the long-run cost measure, assuming a 0% markup.

B Detail on the machine learning procedure

We generate cost predictions by regressing an indicator for new default against a large selection of features using a random forest algorithm. We create four sets of predictions trained on 10% of the data with new borrowing within each snapshot – approximately 8% of the overall data. Predictions are trained and predicted either within each 6-month post-December snapshot (AC^{post}), or only in the December 2009 snapshot (AC^{pre}). The random forests for each type are constructed with or without registry information. We use `python`'s `sklearn` package to perform our machine learning tasks (Pedregosa, Varoquaux, Gramfort, Michel, Thirion, Grisel, Blondel, Prettenhofer, Weiss, Dubourg, Vanderplas, Passos, Cournapeau, Brucher, Perrot and Duchesnay 2011).

Our random forest regression design constructs regression trees using a feature vector of the following observable characteristics of each observation: a gender indicator, and one and two period lags of innovations in borrowing, innovations in total debt, total borrowing, total debt, average costs, and credit line information. We additionally include the default history deleted from the credit registry in some of the trees. In total, these trees have either thirteen or fourteen predictor variables.

We scale our features by binning their nonzero values into quartiles. This reduces noise in the feature vector and creates parsimonious regression trees. In our dataset, we find that this additionally decreases the time necessary to construct a random forest. Finally, we subset over only new borrowers in each period so that our cost estimates reflect costs conditional on borrowing.

To generate our $\mathrm{AC}^{\mathrm{pre}}$ predictions, we train a model only using observations in the December 2009 snapshot. $\mathrm{AC}^{\mathrm{post}}$ predictions are generated using a training sample from each snapshot; these predictions are actually generated using a suite of models each tied to a particular snapshot.

We use three-fold cross validation combined with a grid search to pick parameters for each model. The parameters over which we search are the minimum number of observations in a terminal node (*minleaf*) and the number of features over which each tree can sample. We set the number of trees in a forest to 150. Predictive power is not sensitive to choices in this range. See figures B1 and B2 to see outcomes from this procedure.

Constructing random forests is (generally) a supervised learning task. Breiman (2001) defines a random forest as a set of regression trees, $h_k = h(x, \Theta_k)$ where h is a tree and Θ_k is a random selection of observations and features from the training data, where each tree "votes" on the output given an observation. We pick splits in the data to reduce mean-squared error, as is common with regression tasks. We use this loss function and a regression task, despite our target variable existing only in $\{0,1\}$, to ensure that our outputs are continuous on $[0,1]$ and reflect probabilities. Our predictions are best thought of as a weighted average of default rate in pools of observations clustered together by similarity along a set of their covariates.

We additionally estimate a regression tree[16] to bin borrowers into smaller markets. We define a market as a set of observations M such that $h(x_i, \Theta)$ returns a prediction stemming from the same terminal node for all $i \in M$. We use this method to cluster borrowers into borrowers with similar features and default rates. These clusters therefore represent infered groups in the data at the level which we believe the treatment is applied and are analagous to the clusters defined in each tree in the forest.

Finally, we recreate the analysis above, exchanging the random forest algorithm for two other machine learning procedures that return classification probabilities. These are a naive Bayes classifier and a logistic LASSO. Our naive Bayes classifier first bins nonzero values along the feature vector into quartiles. Under the naive assumption of independence of features in the feature vector, the classifier constructs $P(\text{default}|X)$ using Bayes' formula under the assumption that $P(X|\text{default})$ is Gaussian, though this is functionally irrelevant due to binning.

For the logistic LASSO, we take the log of nonzero values of continuous features, dummying out zero values using indicator variables. We perform a logistic regression with a λ penalty term of the sum absolute value of the coefficients and use three-fold cross validation to pick λ for each model; see figure B3.

Finally, we classify observations' socioeconomic status by training a random forest classifier on observations for whom the bank defined socioeconomic status group. Our three-fold cross validation procedure indicates that we are able to do this with approximately 35% accuracy using a random forest composed of 100 trees and built on a feature vector consisting of continuous measures of consumer debt, mortgage amount, debt balance, credit line, bank default, average cose, age, total default amount, and indicators for gender, new borrowing, and having positive borrowing cap.

[16] We estimate CART-style regression trees that split using variance reduction (Breiman, Friedman, Stone and Olshen 1984).

Figure B1: Cross-validation output for AC$^{\text{pre}}$ random forest predictions

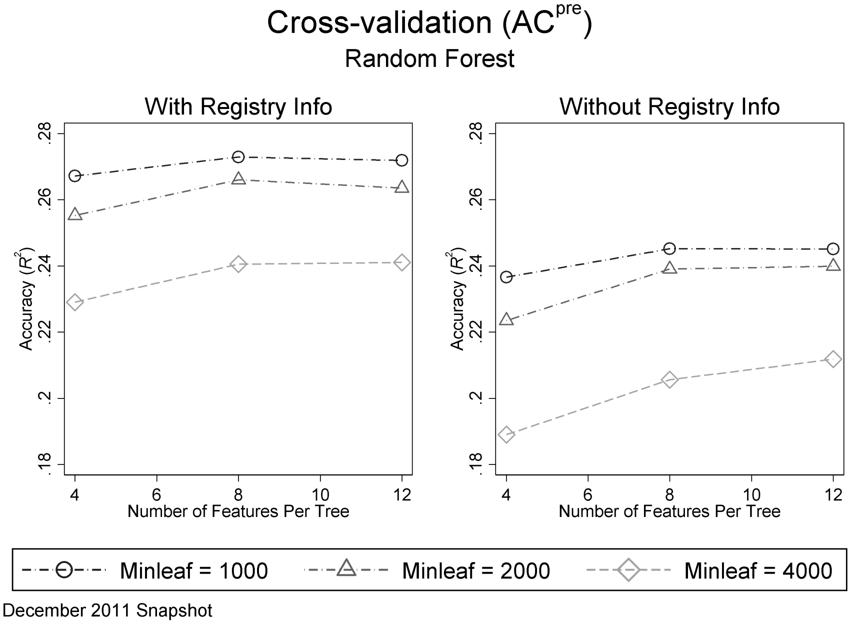


Figure B2: Cross-validation output for AC^{post} random forest predictions

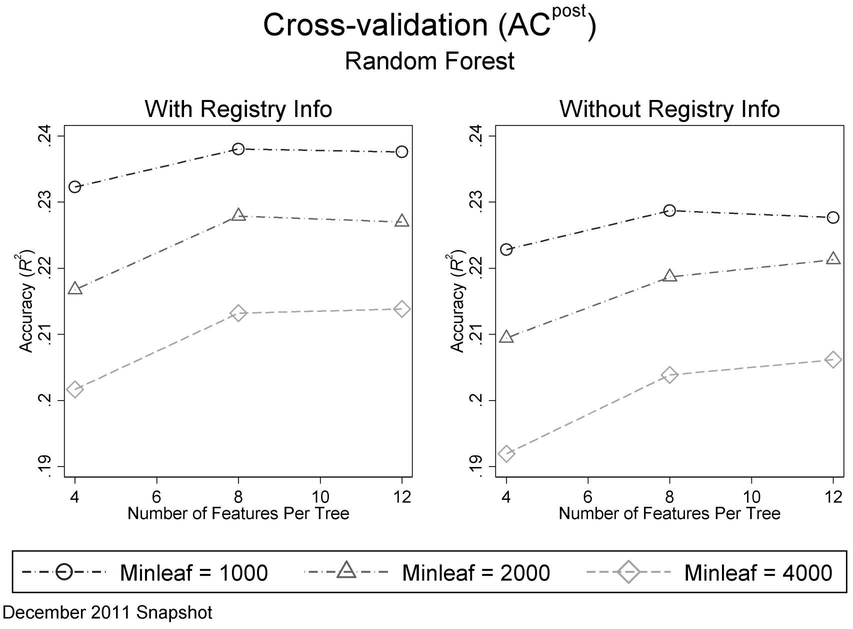
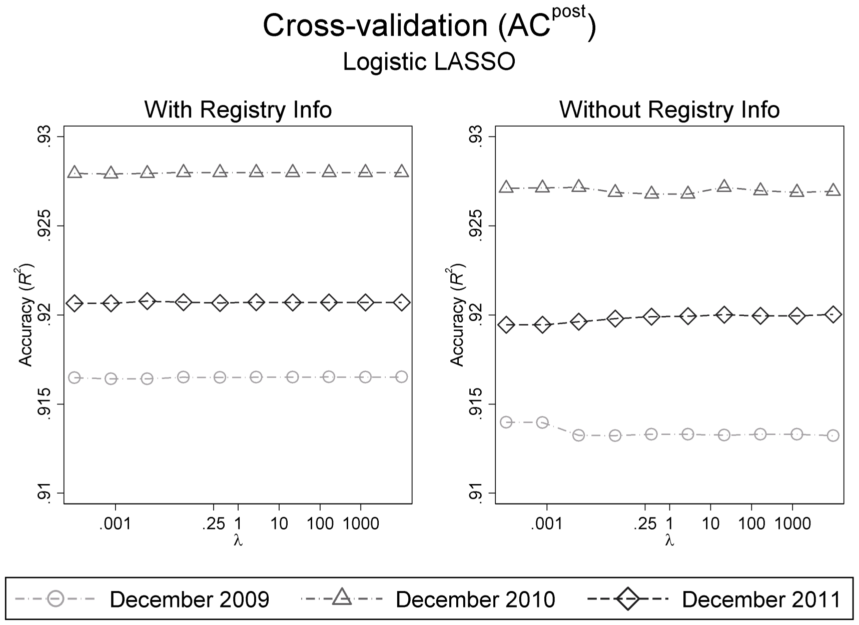


Figure B3: Cross-validation output for AC^{post} logistic LASSO predictions


Cross-validation (AC^post)
Logistic LASSO
With Registry Info
Without Registry Info
Accuracy (R²)
λ
December 2009
December 2010
December 2011