How Magic a Bullet Is Machine Learning for Credit Analysis? An Exploration with FinTech Lending Data

J. Christina Wang* and Charles B. Perkins

** Disclaimer: views expressed here are ours only, not necessarily those of anyone else in the Federal Reserve System.*

CEAR/CenFIS Conference, Oct 30, 2019

Motivation: If/How Are ML Methods Better at Predicting Default in FinTech Lending?

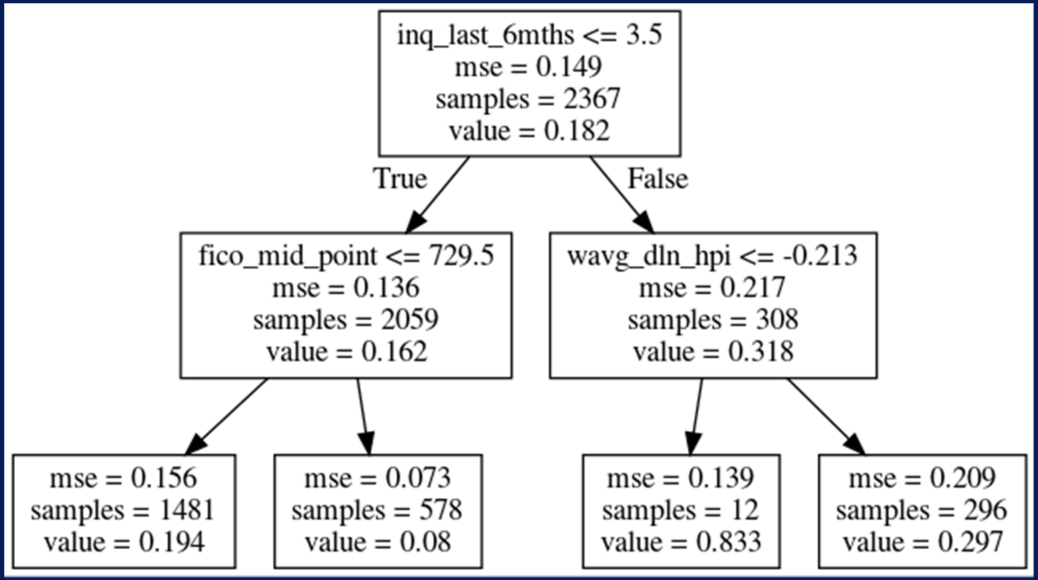
- How much do machine learning methods improve prediction accuracy?
- Which covariates are important in the ML models, and how do they compare with logistic regressions?
  - Any notable interactive effects across covariates?
- How much do more data help ML models (relative to logistic model)? How much do more input variables help, and how does it depend on the type of inputs?
- Do ML models predict more accurate and/or better default probabilities for subgroups of consumers?

Key Findings: ML Models Predict More Accurately, Help Uncover Complex Relation…

- Tree-based ML models improve prediction accuracy
  - Excel more in ranking than exact probability estimate
- List of important inputs (features) similar across ML models and logistic regressions
  - But ML models uncover notable interactive effects
- More observations help ML models relatively more, but only up to a point (~ 5,000 obs.)
- More predictors, esp. local conditions, help too
- Two tree-based ML models predict better default prob. for different subgroups of consumers, but neither is more accurate for any subgroup
  - Intrinsic algorithm matters…

Trees: Recursive Partition of Data Space → Nonparametric Flex. Approximation of Functions

- **Allow different relationships in different parts of the sample space**



- **Interactive effects a natural result**
- **Used for feature selection: only those inputs used for splits**

- **Competitive in classification problems (e.g., binary responses), even though trees using Gini gains for splits (e.g., CART) are subject to inherent biases**

Random Forests: Ensemble of Trees ➔ Low Variance

- Individual tree: low bias but high variance
- ➢ To reduce variance: Average predictions over many trees, and reduce correlation across trees

- Each tree: trained on bootstrapped random subsample, and random subset of covariates
  - Subsetting of inputs: efficient for input selection in high-dimension problems
  - Can be interpreted as adaptive nearest neighbors (NN)

- Easy to apply: fewer hyperparameters to tune than boosting and faster coverage; competitive performance
- But still subject to intrinsic CART bias

Gradient Boosting with Tree Base Learners: Stagewise Additive Modeling, Low Bias & Var.

- **Boosting: sequence of simple models (base learners), each successive step fits last step's residual or increases weights on obs. with wrong predictions**
  - **Gradient boosting: fits last step's residual to achieve largest descent in gradient of the loss function**
- **Final prediction: weighted average over all the steps**

- **Reduce both bias & variance; low risk of overfitting**
- **Boosting with trees as base learners found to excel in classification problems**

- **Feature Importance: an input's contribution to reducing the loss function**
  - **Defined on relative basis; normalize the sum to 100**

Misc. Additional Procedures to Implement ML Models: CV, Discretize Data,…

- K-fold Cross validation: set aside 1/K data for validating model, and train model on the rest (1 – 1/K) of data, then rotate
  - These ML models lack formal inference, so use CV to quantify nonparametrically the uncertainty regarding predictions
  - If suspect data drift, train + CV using loans made in period t, compare with error rate of tests on future loans (t + h)
- Hyperparameter tuning:
  - Tree depth (degree of interactive effects), min. terminal node size, % of input subset; learning rate in boosting (low rate ⇔ many trees)
- Discretize input variables to minimize the impact of intrinsic bias in CART
- Also try LASSO and Ridge regressions: L1 and L2 regularization
  - LASSO: L1 penalty leads to feature selection naturally

Data Sources

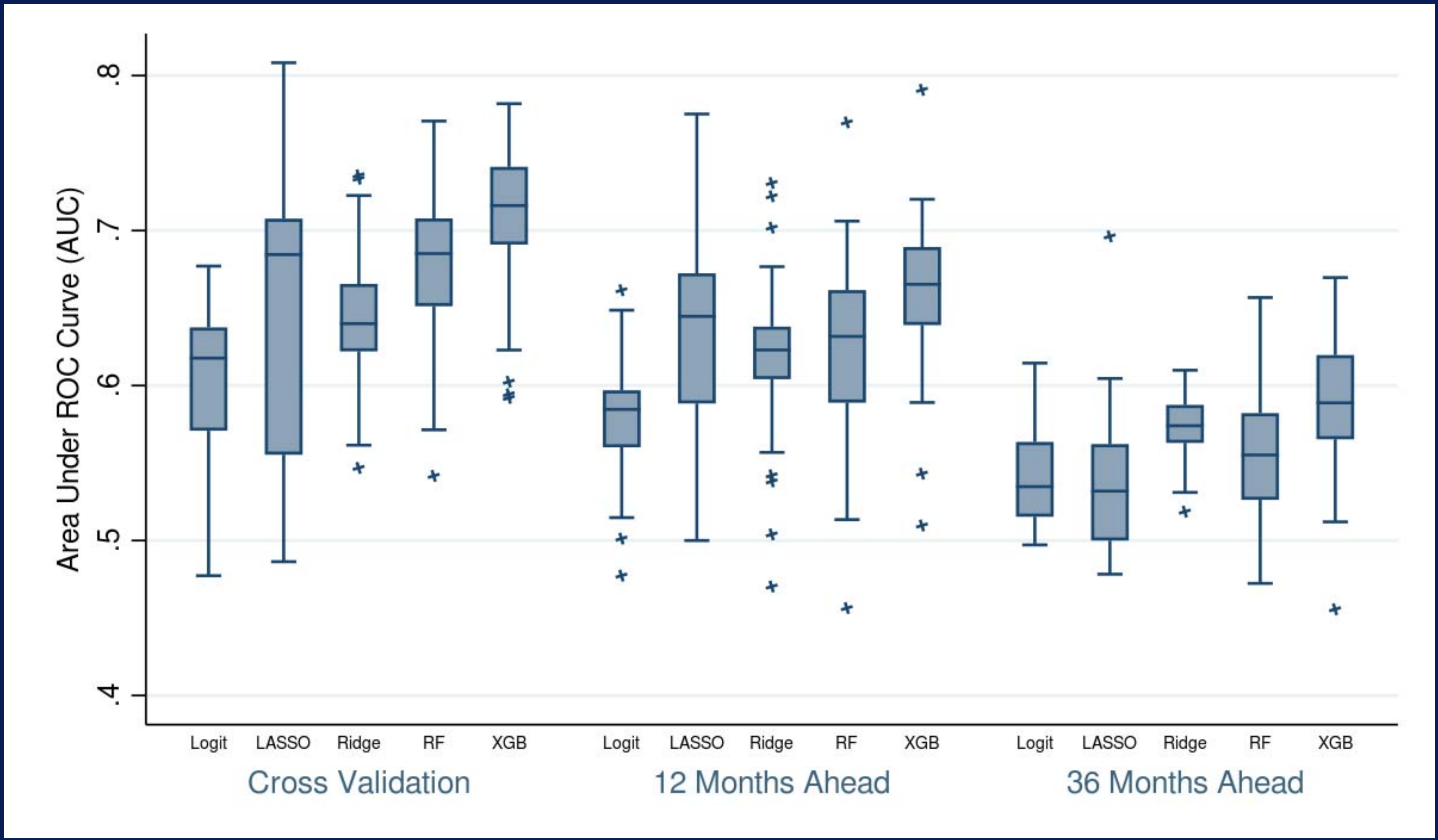
- LendingClub:
  - Borrower attributes: mainly credit bureau data (FICO score, DTI, # of inquiries last 6 months, etc.)
  - Loan outcome (3-year loans only for max. data)
- Census Bureau (ACS): prime-age population, poverty share, share with college degrees, etc.
- BLS unemployment rate (US, by county → by 3-digit zip code)
- FHFA HPI (by 3-digit zip code)
- Equifax (CCP) data by 3-digit zip: avg. balance on credit card, student loan and other non-mortgage debt
- Banking market conditions: wt. avg. NPL of CRE and RRE of banks in the zip area, CET1 cap ratio, deposit HHI
- BEA and BLS: major NIPA indicators (GDP and PCE growth, deflator inflation)

Training vs. Testing Samples, Metrics for Model Comparison

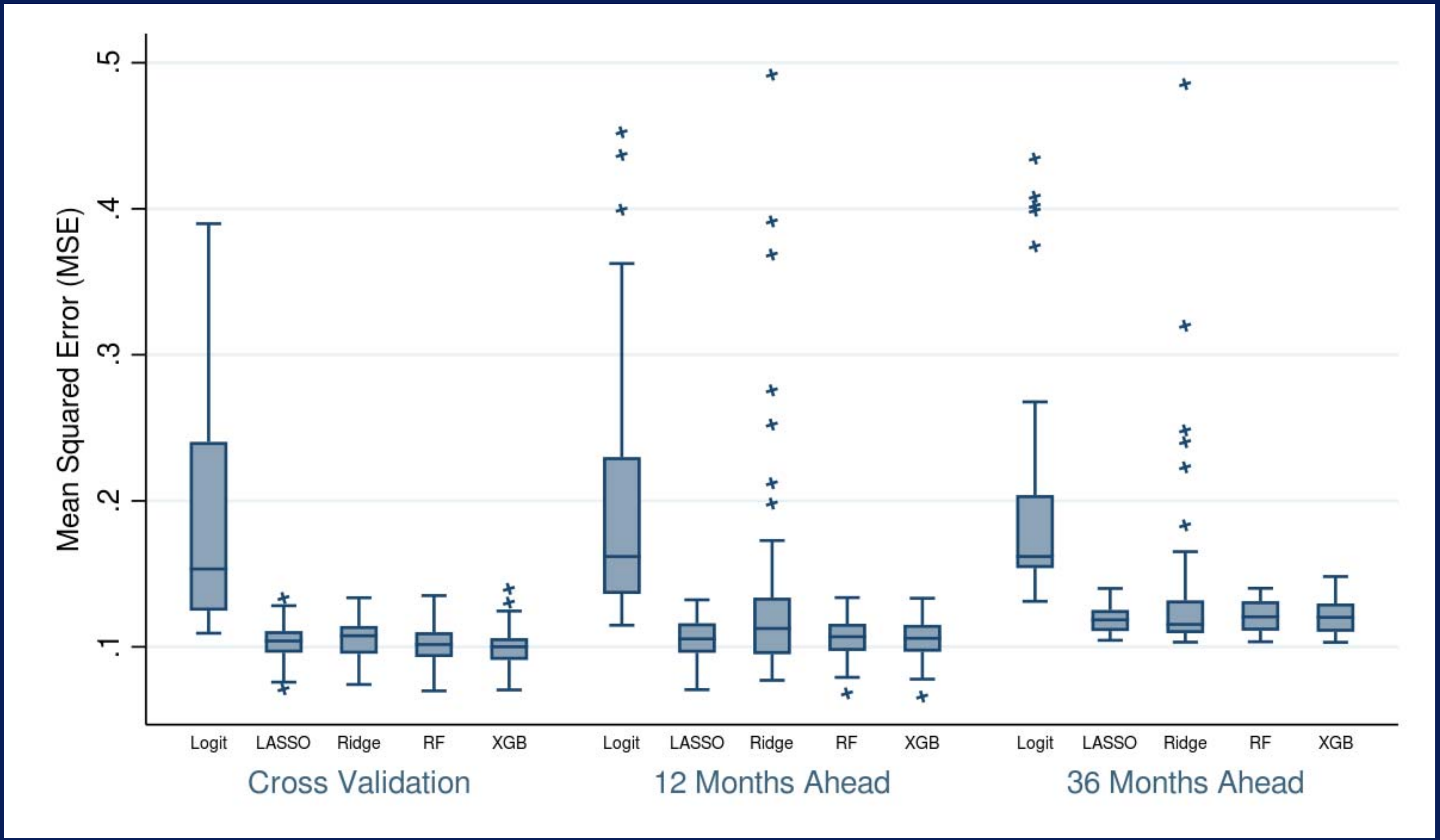
- Training: by monthly loan cohort 2009:M1 – 2014:M2
- Testing: loans made in the same month (CV test subsample) vs. in future months
  - Train models on cohort t, test on cohorts $t + h$, $h \geq 0$
  - In real time, need data released $\geq t + 36$ to train models on cohort t; need data released $\geq t + h + 36$ to test cohort $t + h$
  - Last fully matured loan cohort 2015:M8

- Metrics for model comparison:
  - Mean squared error (MSE): exact value of predicted PD matters
  - Area under ROC curve (AUC):
    probability of ranking a random obs. with $y = 1$ higher than a random obs. with $y = 0$
    ➔ ranking matters, but not exact value of predicted PD
    - AUC = 0.5 for a completely uninformative model

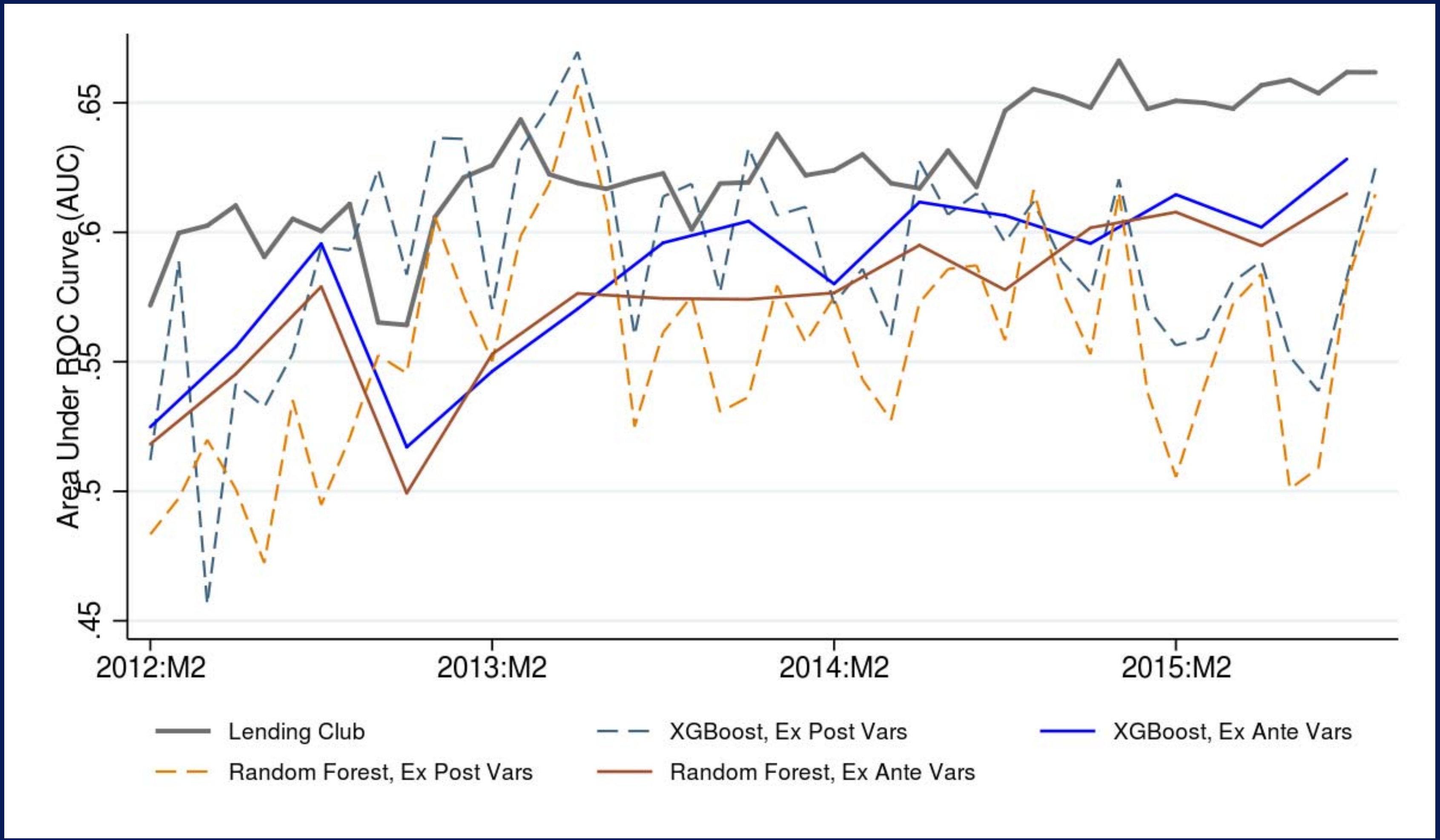
ML Models Rank Default Prob. More Accurately: AUC Comparison Across All Models



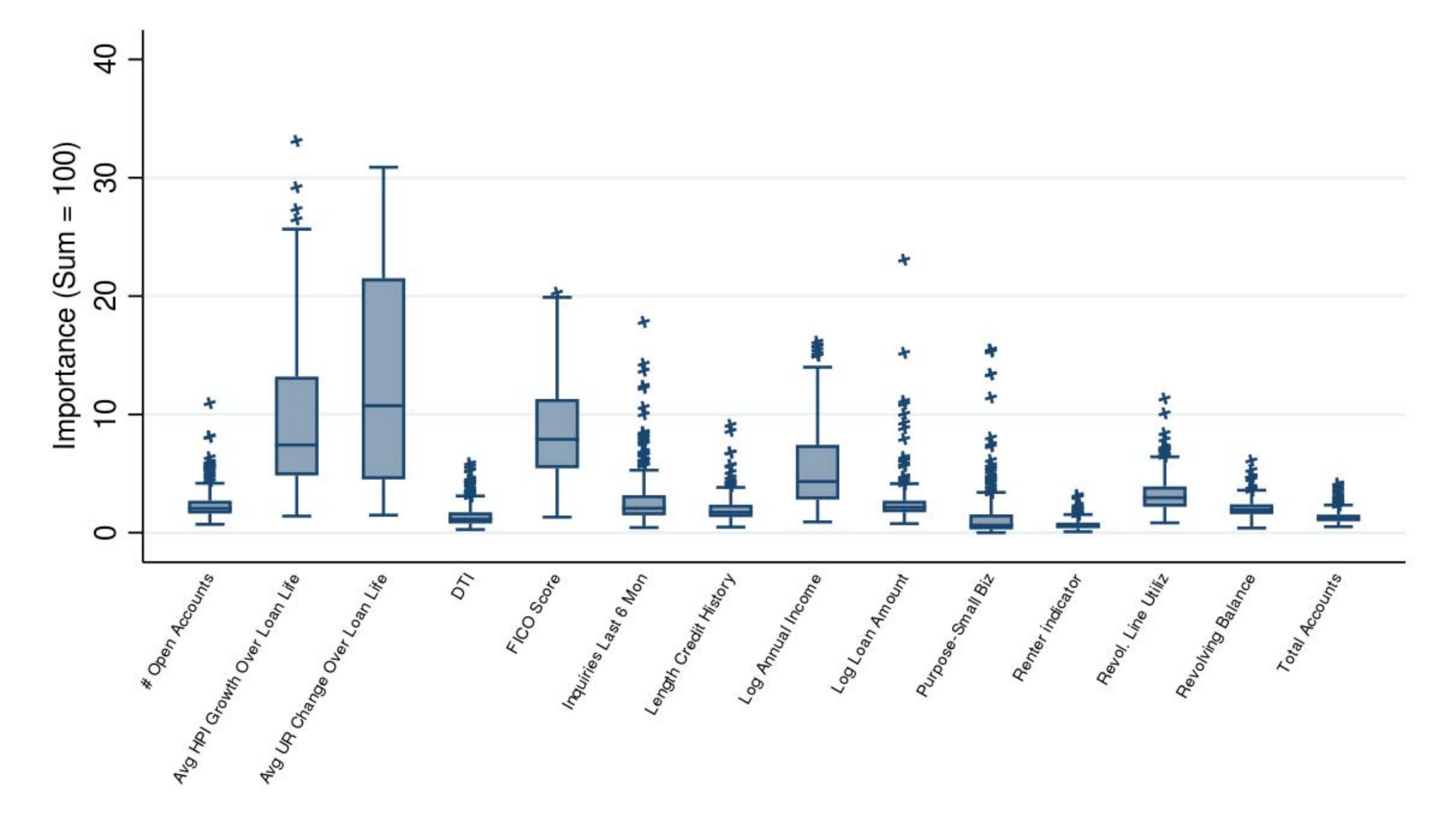
In contrast to AUC, ML Model MSEs Comparable to LASSO, Ridge → Regularization is Key



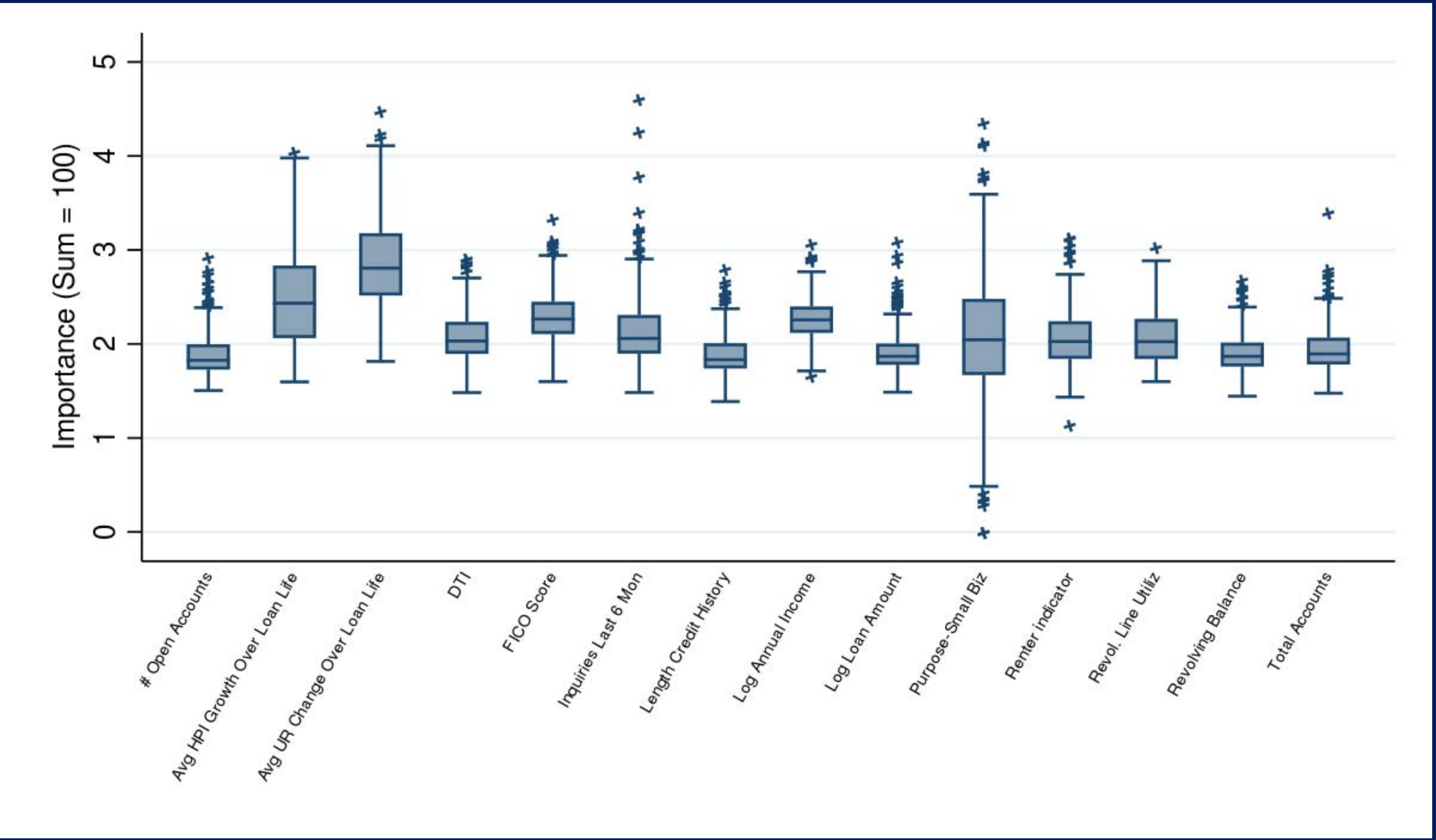
AUC comparison: LendingClub Credit Grades Rank Borrowers Most Accurately



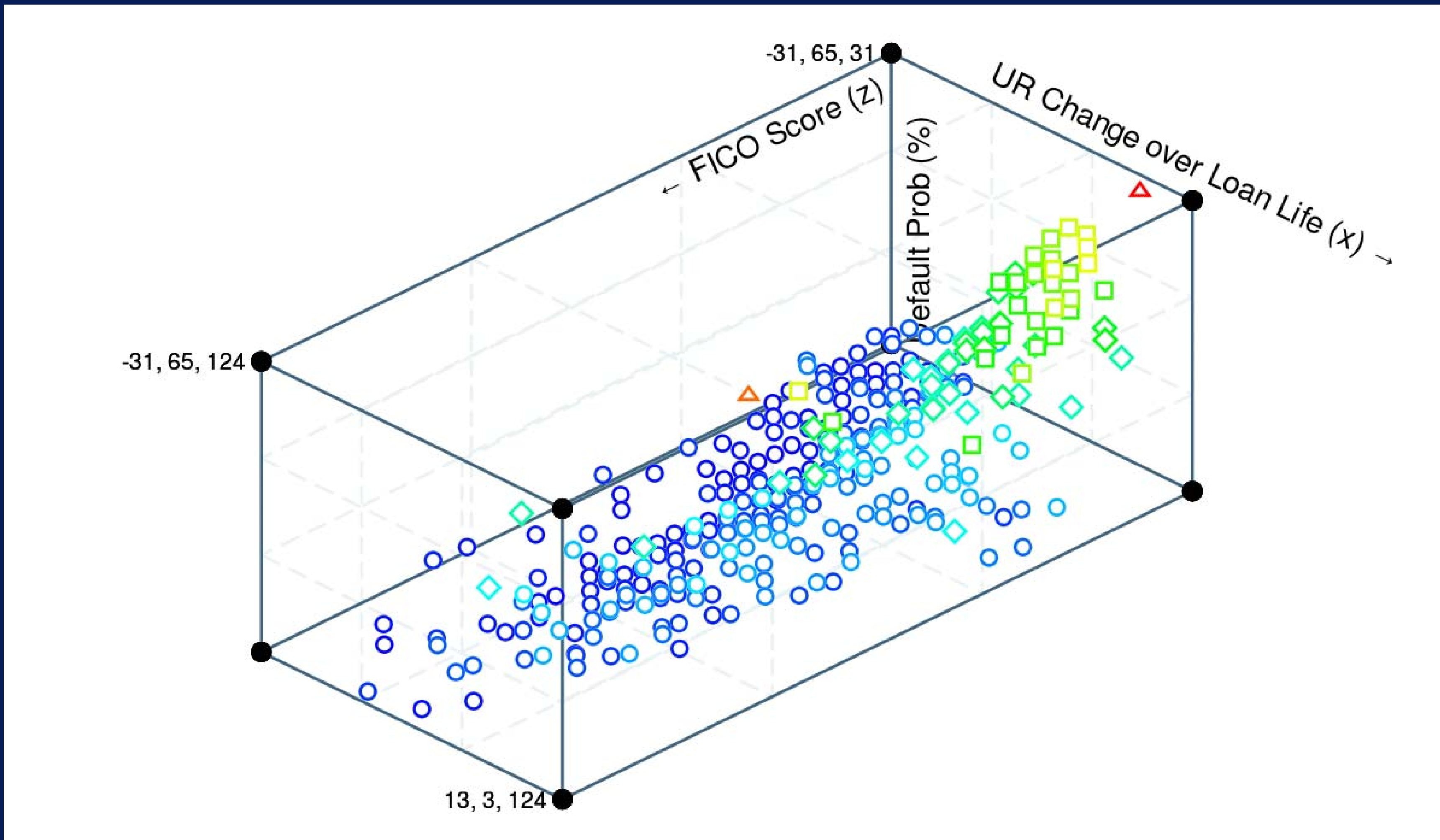
Random Forests Feature Importance: Similar to Logistic and LASSO Coeff. Significance Ranking



Boosted Trees Feature Importance: Similar Ordering but More Uniform across Inputs



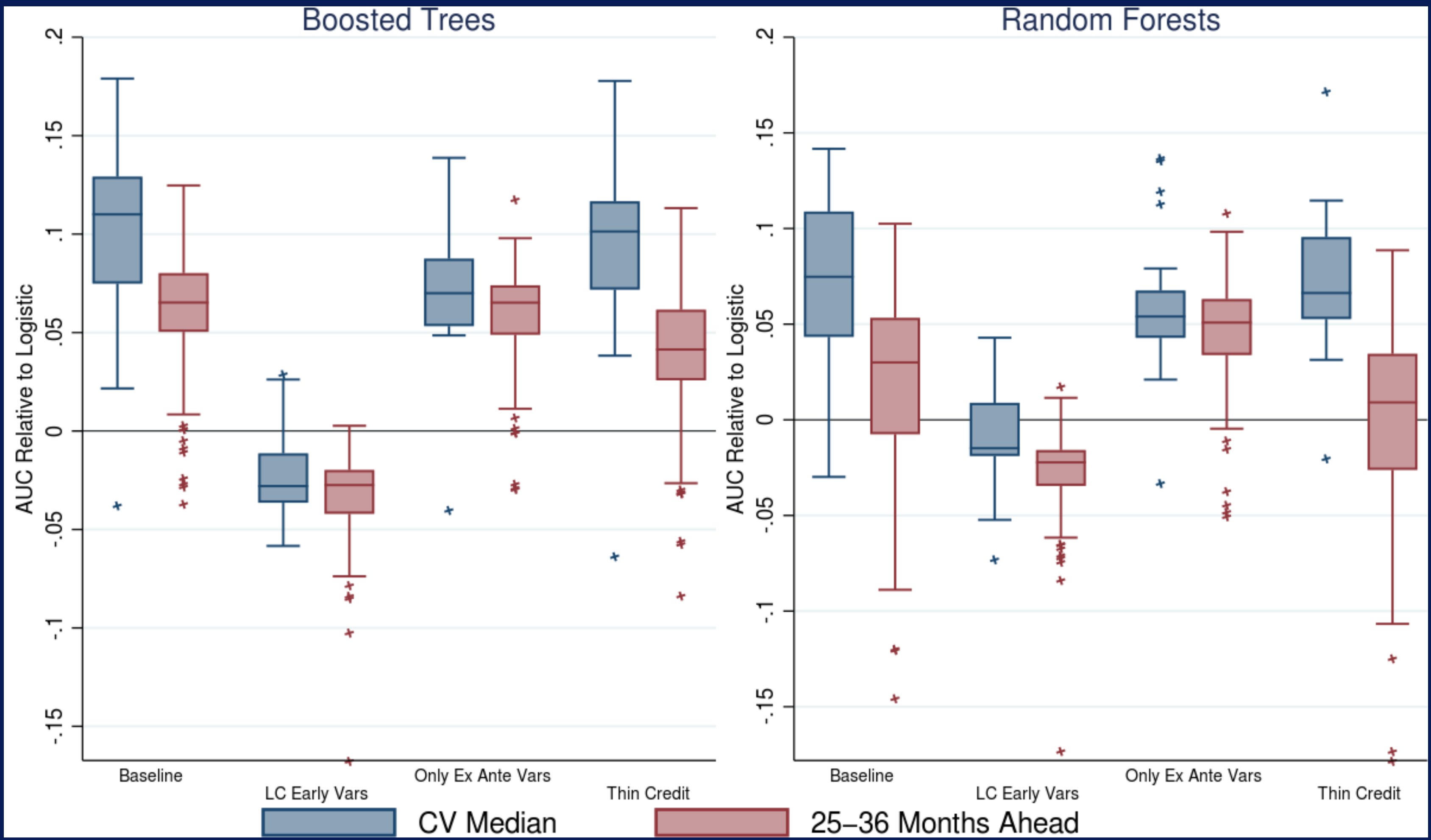
Partial Dependence—Interactive Effect betw. FICO & Unemployment Rate (Low FICO X High UR)



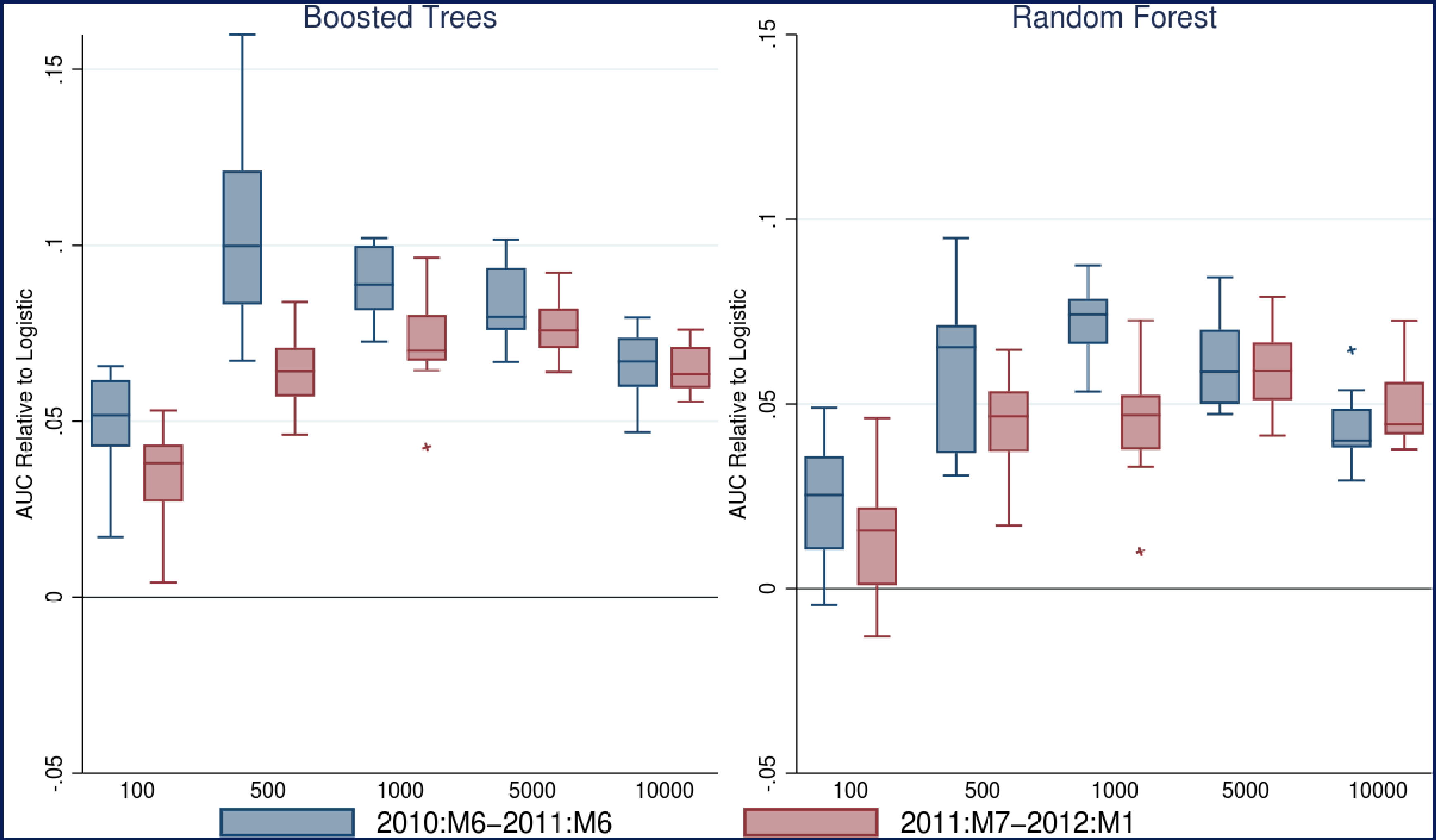
The More Covariates, the Better?

1. Baseline:
   - All individual + loan indicators
   - Local economic conditions: ex ante indicators, ex post unemployment rate & HPI growth rate
2. LendingClub early grade model variables:
   - 8 key borrower credit indicators
3. Ex ante economic variables only:
   - Baseline – ex post local conditions
4. Thin credit: to mimic cases with little credit history
   - # of inquiries in last 6 months, months since last inquiry, total current/high balance, credit history length, requested loan amount
   - All local economic conditions

More Predictors Increase ML Models' AUC More, esp. for Test Loans in the Same Month



More Observations Increase ML Models' AUC More, but peak around 500~5000 obs.



ML Models' Prediction Accuracy (Relative MSE) Hardly Differs by Borrower Risk, Income, etc.

	XGBoost	Random Forest
Risk Grade A	-1.270	-1.669
	(1.661)	(1.674)
Risk Grade B	-1.219	-1.907
	(1.343)	(1.362)
Risk Grade C	-0.968	-1.762
	(1.050)	(1.073)
Risk Grade D	-1.020	-1.711
	(0.897)	(0.922)
Risk Grade E	-0.739	-1.144
	(0.689)	(0.716)

ML Models' Prediction Accuracy (Relative MSE) Hardly Differs by Borrower Risk, Income, etc.

	XGBoost	Random Forest
FICO Score	0.00253	0.00732
	(0.00759)	(0.00750)
Debt-to-Income Ratio	-0.00211	-0.0167
	(0.0228)	(0.0223)
Log of Applicant Income	0.284	0.532
	(0.347)	(0.344)
Log of Loan Amount	0.0586	0.147*
	(0.0738)	(0.0727)
Log of 3-digit Zip Code Population	-0.0400	-0.154
	(1.373)	(1.443)
Unemploy. Rate Difference from US Rate	1.683**	1.847**
	(0.428)	(0.432)
HPI Growth Rate (t-1)	-0.0596*	-0.0602*
	(0.0290)	(0.0287)
Poverty Share (%)	0.529*	0.508
	(0.262)	(0.263)
Share with Card Utilization >= 85%	0.194	0.220
	(0.355)	(0.356)

Boosted Trees Predict Lower Prob. for Borrowers Already Deemed Safe, Random Forests Less So

	XGBoost	Random Forest
Risk Grade A	4.196**	-1.538
	(1.077)	(1.050)
Risk Grade B	2.459**	-1.728*
	(0.812)	(0.783)
Risk Grade C	0.976	-1.839**
	(0.651)	(0.631)
Risk Grade D	0.0492	-1.646**
	(0.585)	(0.571)
Risk Grade E	-0.111	-0.984*
	(0.469)	(0.460)

Boosted Trees Predict Lower Prob. for Borrowers Already Deemed Safe, Random Forests Less So

	XGBoost	Random Forest
FICO Score	0.0639**	0.0266**
	(0.00880)	(0.00853)
Debt-to-Income Ratio	-0.0867**	0.0622**
	(0.0204)	(0.0193)
Log of Applicant Income	2.137**	0.0928
	(0.392)	(0.397)
Log of Loan Amount	1.162**	1.094**
	(0.141)	(0.134)
Log of 3-digit Zip Code Population	0.932	1.218
	(1.722)	(1.865)
Unemploy. Rate Difference from US Rate	-0.637	-1.002
	(0.615)	(0.639)
HPI Growth Rate (t-1)	0.0232	0.0377
	(0.0468)	(0.0459)
Poverty Share (%)	-0.773*	-0.617
	(0.390)	(0.385)
Share with Card Utilization >= 85%	-0.484	-0.359
	(0.498)	(0.491)

Summary of Findings

- Tree-based ML models improve prediction accuracy
  - Excel more in ranking than exact probability estimate
- List of important inputs (features) similar across ML models and logistic regressions
  - But ML models uncover notable interactive effects
- More observations help ML models relatively more, but only up to a point (~ 5,000 obs)
- More predictors, esp. local conditions, help too
- Two tree-based ML models predict better default prob. for different subgroups of consumers, but neither is more accurate for any subgroup
  - Algorithm matters: averaging helps risky borrowers more