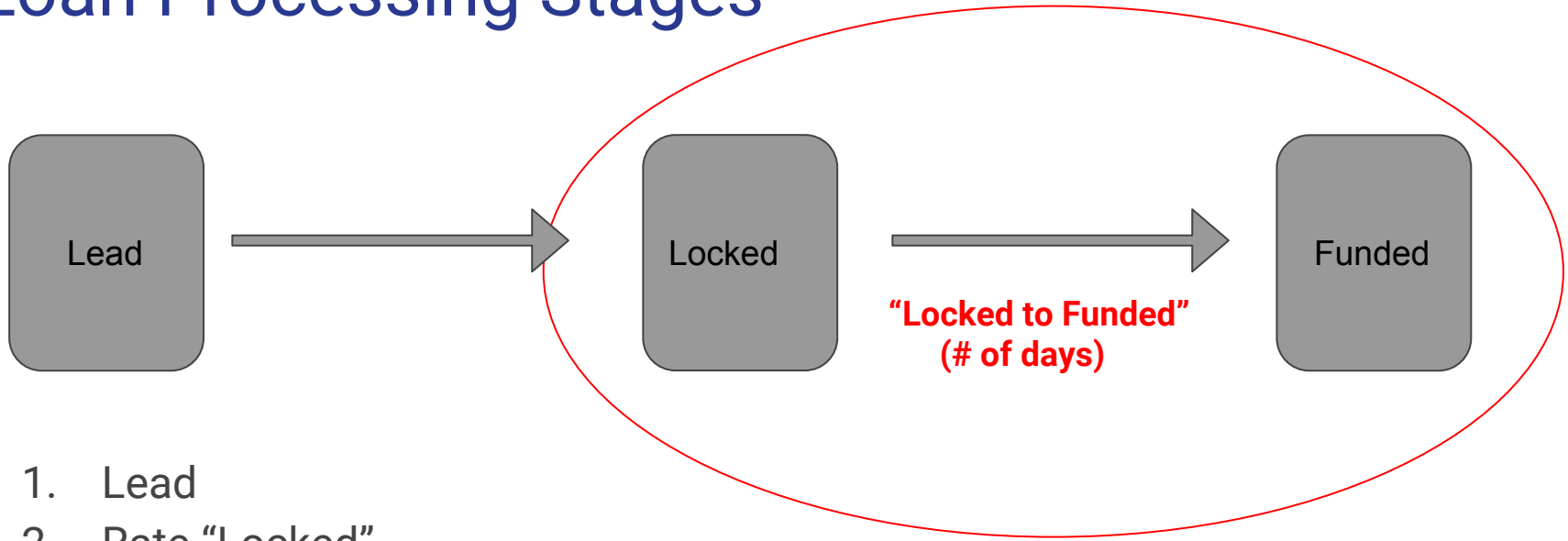
Mortgage Data Analysis

A Capstone Project for a mortgage lender

Shige Tajima, Ph.D.

linkedin.com/in/stajima
github.com/shigetajima

Loan Processing Stages



1. Lead
2. Rate “Locked”
3. Money “Funded”

Motivations on “Locked to Funded” study

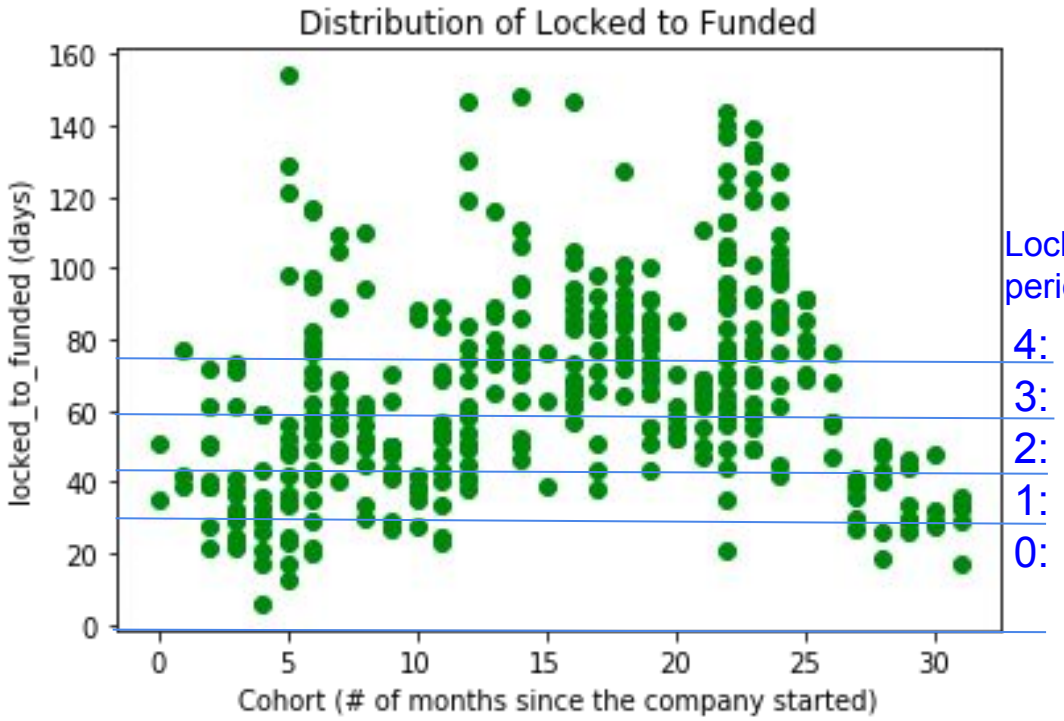
- Prediction of “Locked to Funded” time
- How to reduce “locked to funded”
- Identify most important features
- Estimation of the lock period

* Prediction of “Locked to Funded” and reducing it will be beneficial to both the company and customers

Data set

- Data is from a mortgage lender in the Bay Area.
- Many features (~500) including 200 categorical features
  - Fees/costs
  - Status of tasks
  - Milestone dates
  - Info on customers and company staff
- Loan processing procedures changed a few times since the company started

Locked to Funded (# of days)



Lock period

4: > 75 days

3: 60-74

2: 45-59

1: 30-44

0: < 30 days

EDA / Feature Engineering

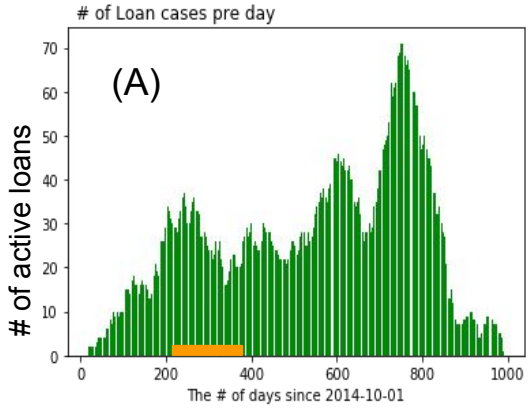
New features added:

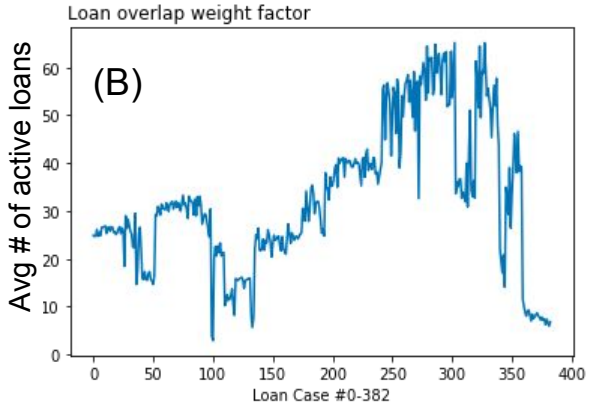
A = # of active loans per day

B = avg # of active loans for each loan

C = # of available staff for each day

D = Loan Weight = B / C





EDA / Feature Engineering

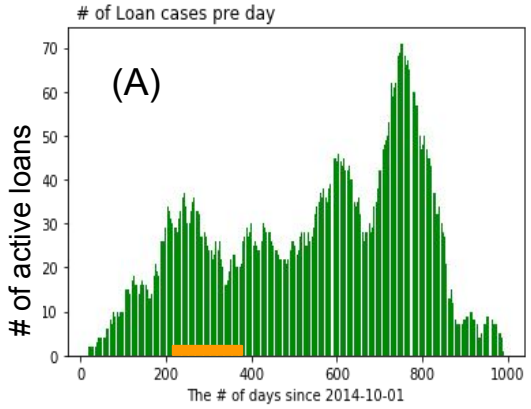
New features added:

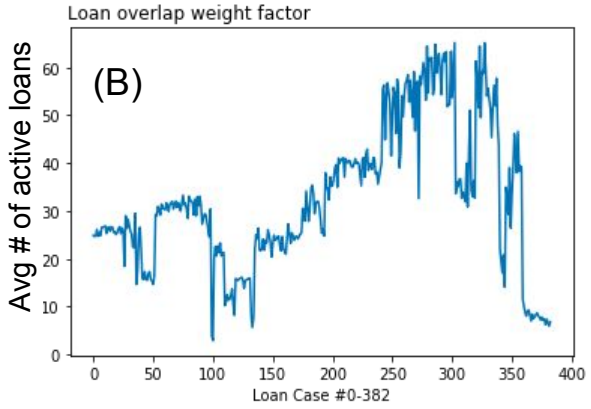
A = # of active loans per day

B = avg # of active loans for each loan

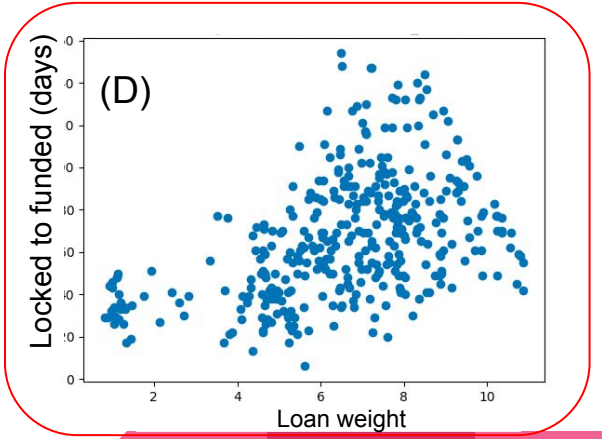
C = # of available staff for each day

D = Loan Weight = B / C





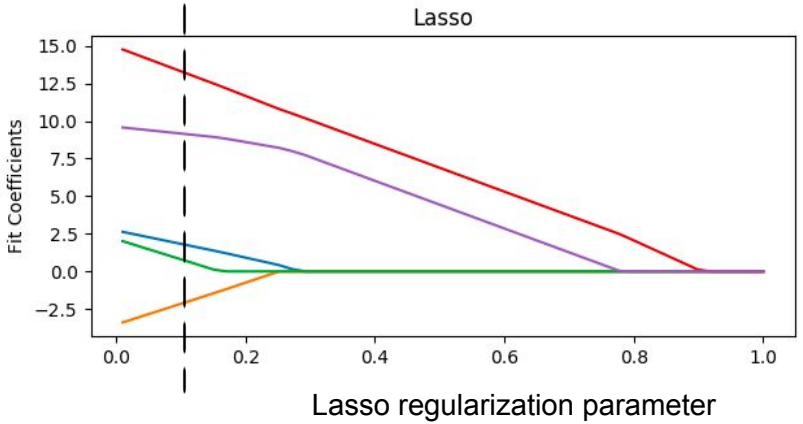
"Loan Weight" (x) shows a correlation with "Locked To Funded" (y)



Analysis #1 -- Regression models

- Prediction of “locked to funded”
- Lasso: found two dominant features:
  - Fees for extending the lock period
  - Loan weight
- Models and R2 scores

models	Cross validation R2 score
OLS	0.57
Random forest	0.60
Ada Boost	0.62
Gradient Descent Boosting	0.66



Analysis #2 -- Classification models

Prediction of Locked to Funded “period”

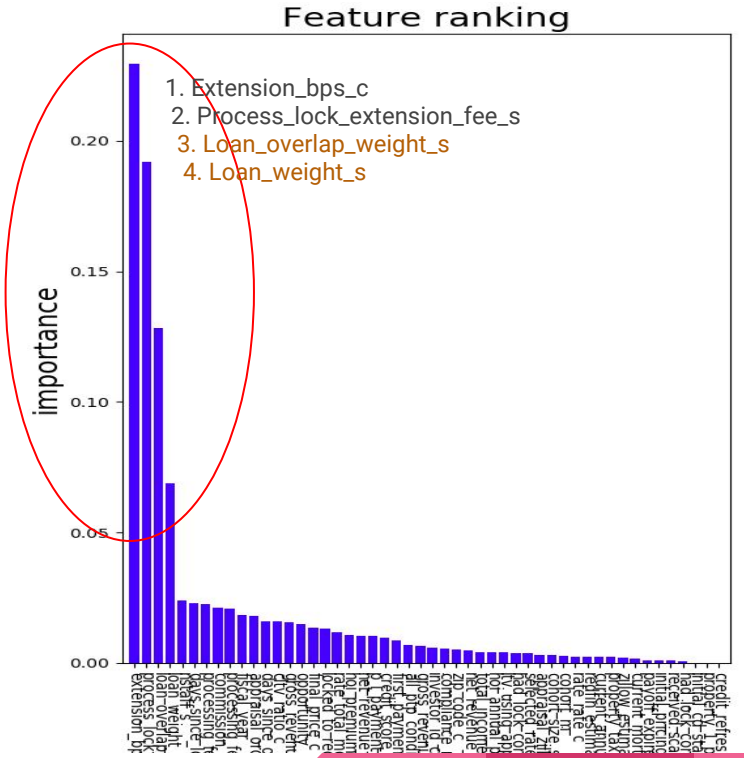
- ”Locked to funded” data were divided into 5 bins
- Used Random Forest Classifier

- Use top 50 features and do Grid Search to optimize params (n_estimators, max_features, max_depth, etc)

- Random Forest: cross validation accuracy = 0.54

Top features are related to:

- Fees for extending the lock period
- Loan weight



Future work

- Staff / pair performance analysis
- Estimation of the loan processing cost
- Estimation of # of days from “Locked” to “Clear to Close” (a major milestone before money is ‘funded’)

Thank you!

Shige Tajima, Ph.D.

stjm05@gmail.com
linkedin.com/in/stajima
github.com/shigetajima

Backup slides

Analysis -- OLS

Feature selection criteria:

1. Correlation with ‘y’ (locked_to_funded) (> 0.2)
2. Small VIF (Variance Inflation Factor) (< 5.0)
3. Large Statistical Significance: (p_value < 0.05)

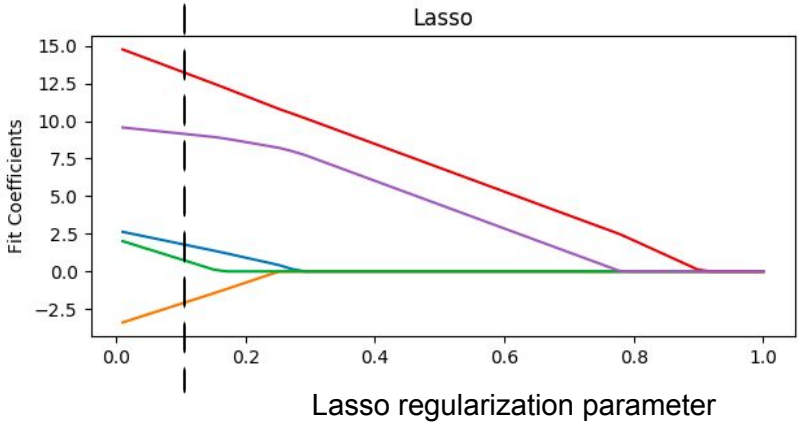
Iterations

Lasso regularization was used to identify most important features

→ select 5 best features, and use them in OLS:

OLS Results

- 5-fold cross validation score = 0.57
- Dominant features found
  - process_lock_extension_fee_s,
  - loan_overlap_weight_s,
  - appraisal_ordered_to_received_c,
  - commission_due_c (negative)



Random Forest Regressor-- Feature Importances

5-fold cross_val_score = 0.60 (avg)

1. process_lock_extension_fee_s (0.280300)
2. Extension_bps_c (0.221298)
3. loan_overlap_weight_s (0.145513)
4. loan_weight_s (0.023385)
5. appraisal_ordered_to_received_c (0.012643)
6. nstaff_s (0.012452)
7. commission_due_c (0.007477)
8. gross_revenue_c (0.006962)
9. rate_closing_costs_total_c (0.006785)
10. days_since_opp_created_c (0.006541)