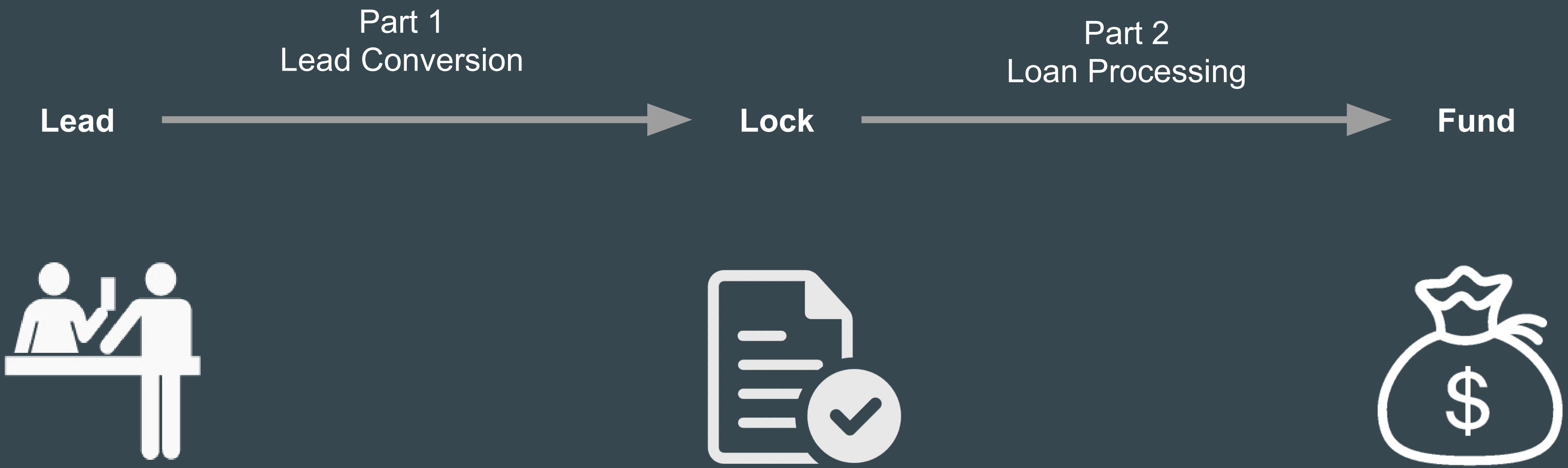
Project Home-a-loan

Wei Cheung
July 20, 2017

The Challenge



Part 1
Lead Conversion

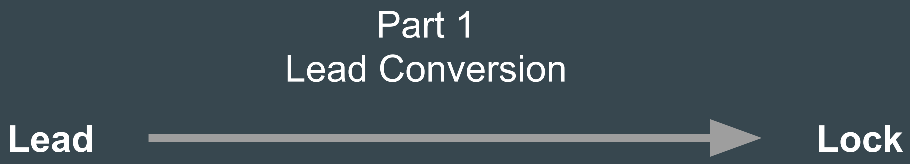
Part 2
Loan Processing

Lead

Lock

Fund

The Challenge



Part 1
Lead Conversion

Lead **Lock**

Mission:

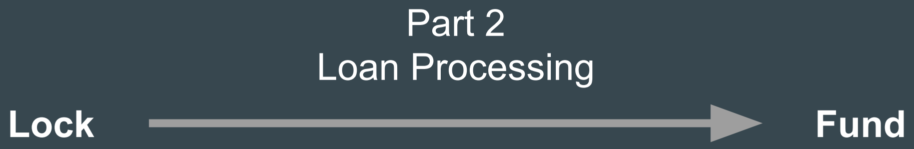
- Predict whether a lead will convert to “lock”

Values for the Company:

- Know the potential of each customer - know the ones to focus on

- Marketing - based on Profile of “Ideal Customers”

The Challenge



Part 2
Loan Processing

Lock → **Fund**

Mission:

- Predict “locked-to-funded” time (efficiency)

Values for the Company:

- Give expected time -
  improve customer experience
- Know areas of improvement for efficiency

The data

- ~9800 cases
- ~420 features
- Cleaning / pre-processing:

Data Type	Examples	Processing
Numerical	Borrower income, loan amount	None
Categorical	Type of home, Education level, City of property, Gender	Dumification
Text	Goal of refinancing, Unqualified reason note	Tfidf Vectorization (limiting stop words)
Datetime	Created time, Last modified	Categorize (year, quarter, month, dow), Calculate Period (difference of dates/times), Calculate Cohort (quarter/month since initiation)

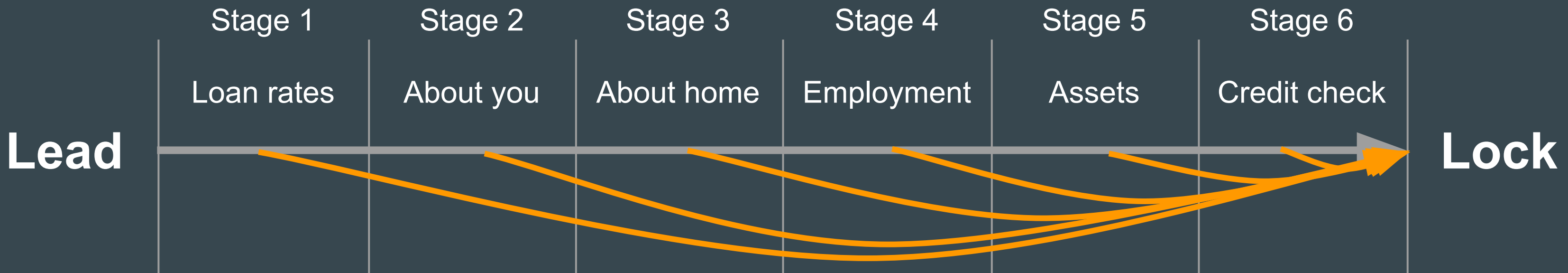
Part 1 - Leads Conversion

Part 1
Lead Conversion

Lead

Lock

Part 1 - Lead Conversion



Stage 1 Loan rates

Stage 2 About you

Stage 3 About home

Stage 4 Employment

Stage 5 Assets

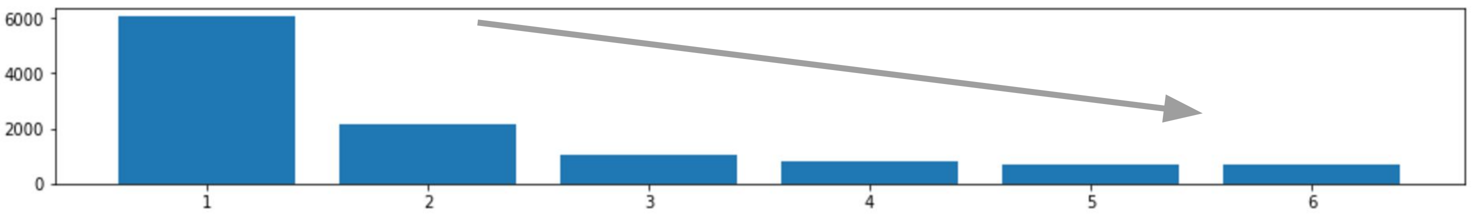
Stage 6 Credit check

Lead

Lock

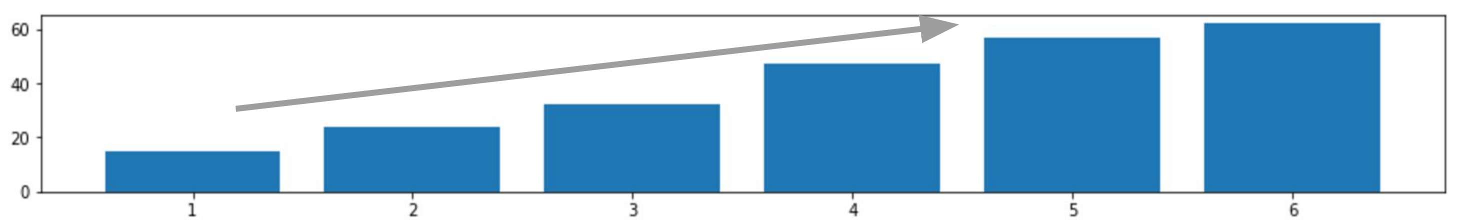
Sample Sizes

(Rows)



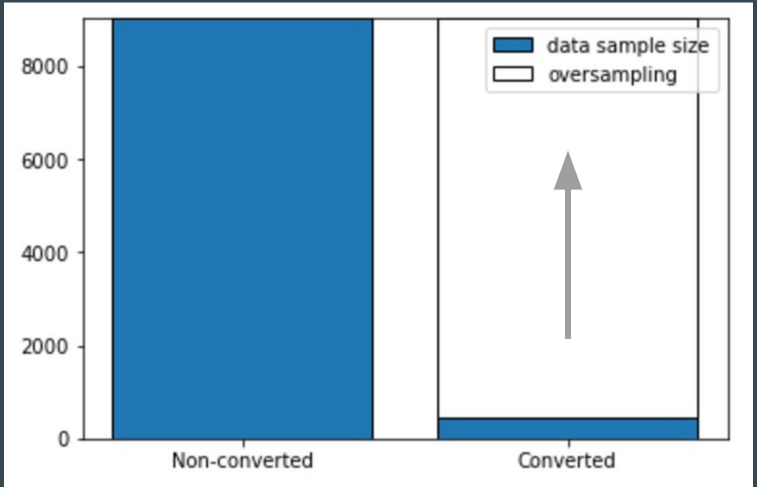
#Features

(Columns)

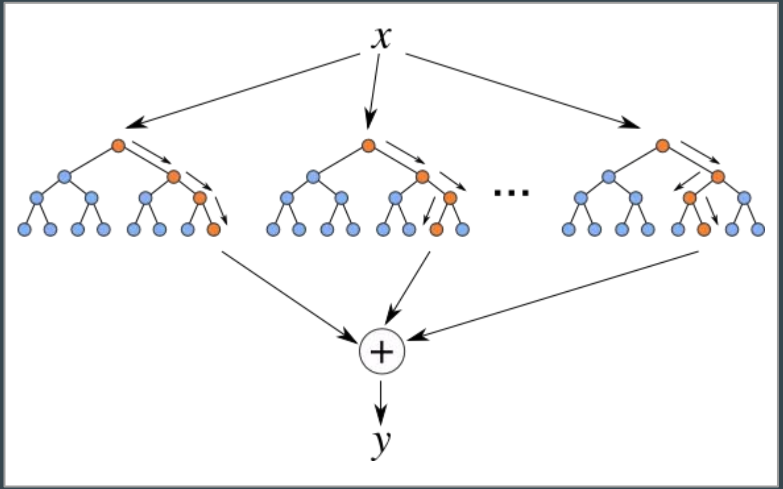


Part 1 - Methodology

Random Oversampling for Imbalance Classes

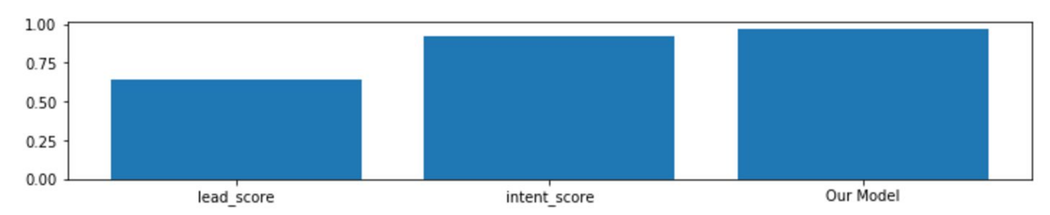


Random Forest Classifier

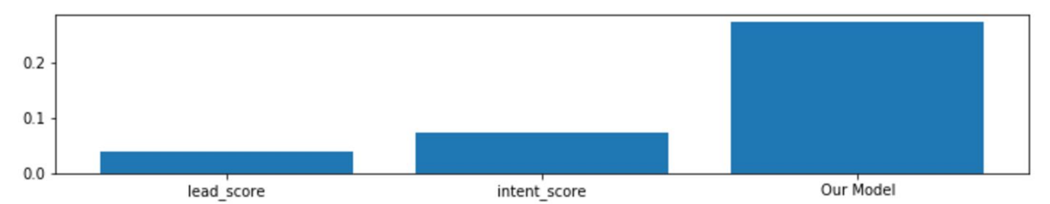


Part 1 - Results

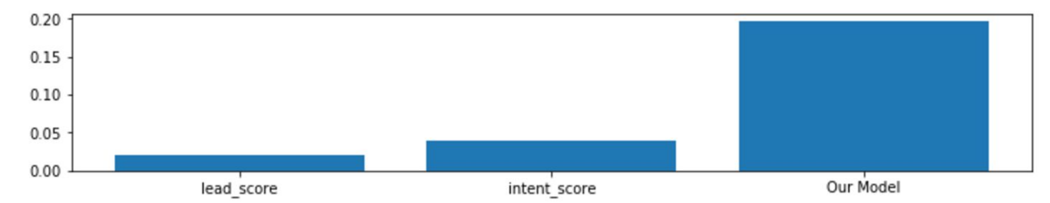
Accuracy



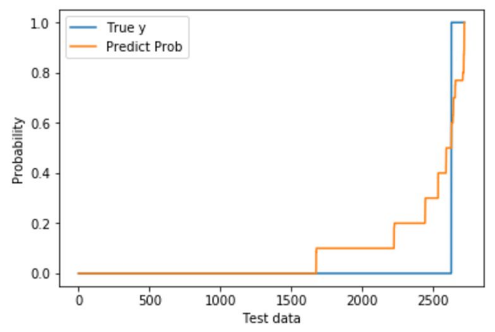
F1-score

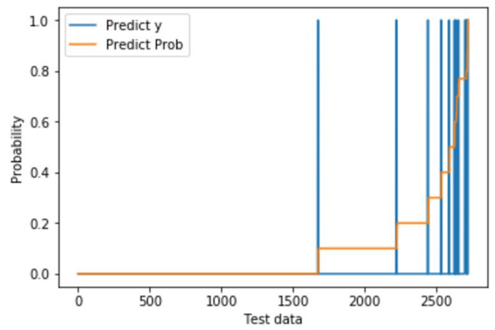


Precision



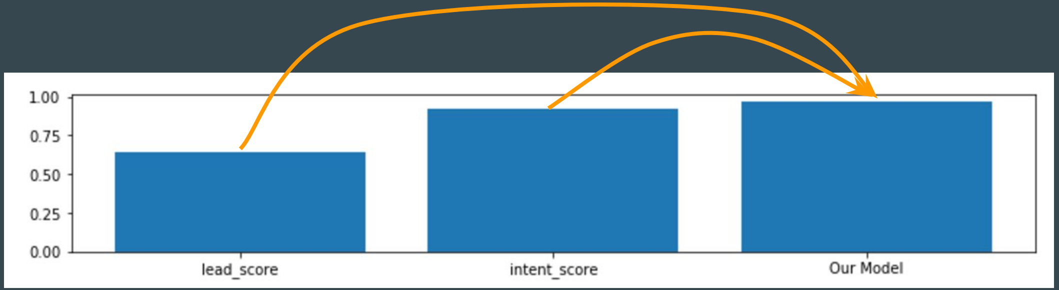
Predicted Probability v.s. True Value



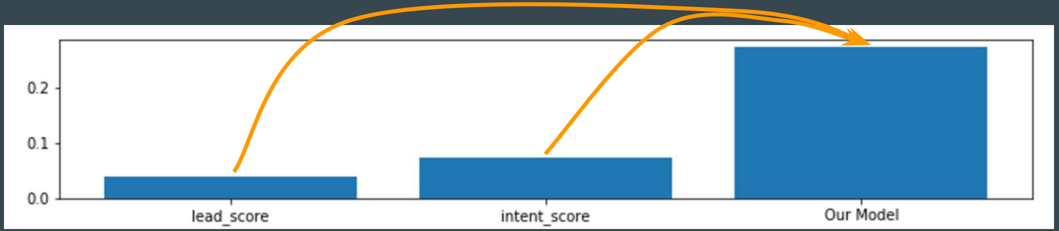


Part 1 - Results

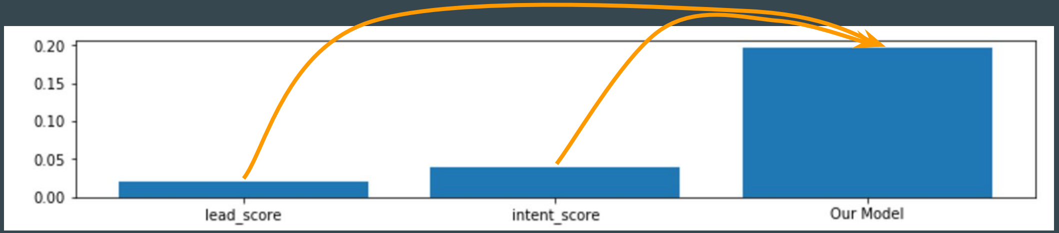
Accuracy



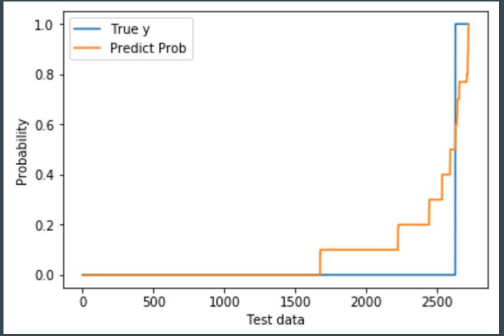
F1-score

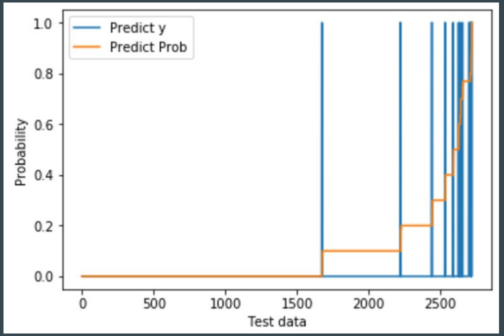


Precision

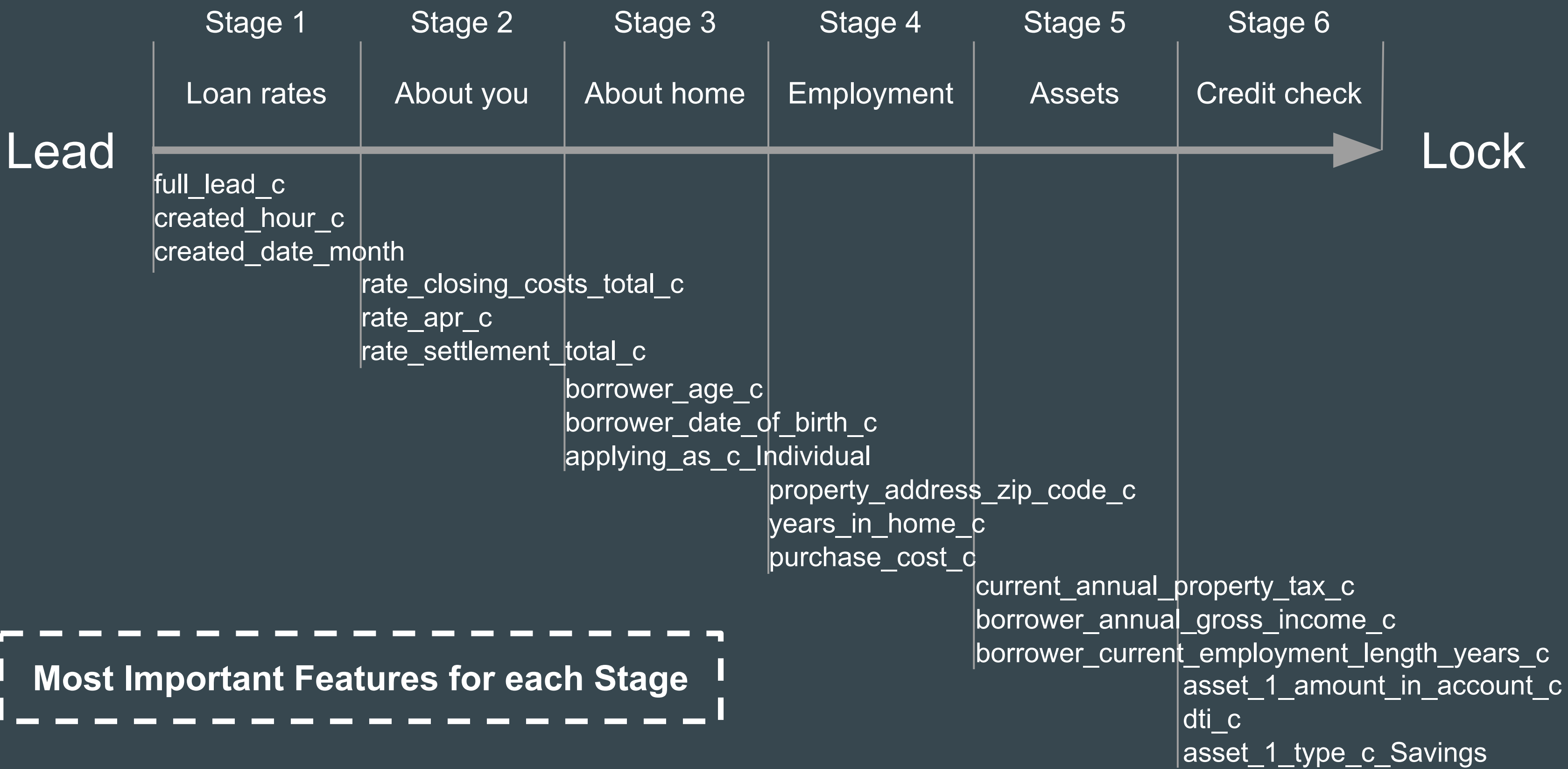


Predicted Probability v.s. True Value





Part 1 - Interpretation



Lead

Stage 1 - Loan rates
- full_lead_c
- created_hour_c
- created_date_month

Stage 2 - About you
- rate_closing_costs_total_c
- rate_apr_c
- rate_settlement_total_c

Stage 3 - About home
- borrower_age_c
- borrower_date_of_birth_c
- applying_as_c_Individual

Stage 4 - Employment
- property_address_zip_code_c
- years_in_home_c
- purchase_cost_c

Stage 5 - Assets
- current_annual_property_tax_c
- borrower_annual_gross_income_c
- borrower_current_employment_length_years_c

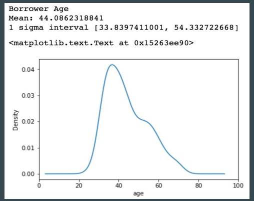
Stage 6 - Credit check
- asset_1_amount_in_account_c
- dti_c
- asset_1_type_c_Savings

Lock

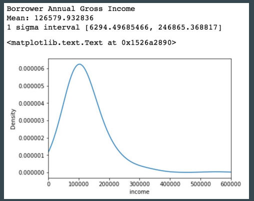
Most Important Features for each Stage

Part 1 - Interpretation

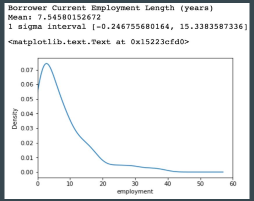
Examples of the Important Features



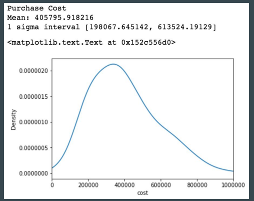
```
Borrower Age
Mean: 44.0862318841
1 sigma interval [33.8397411001, 54.332722668]

<matplotlib.text.Text at 0x15263ee90>
```



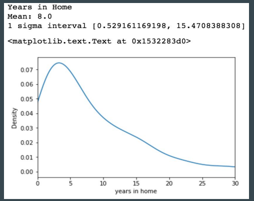
```
Borrower Annual Gross Income
Mean: 126579.932836
1 sigma interval [6294.49685466, 246865.368817]

<matplotlib.text.Text at 0x1526a2890>
```



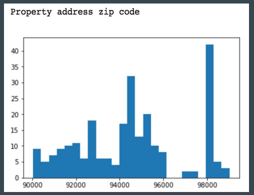
```
Borrower Current Employment Length (years)
Mean: 7.54580152672
1 sigma interval [-0.246755680164, 15.3383587336]

<matplotlib.text.Text at 0x15223cfd0>
```



```
Purchase Cost
Mean: 405795.918216
1 sigma interval [198067.645142, 613524.19129]

<matplotlib.text.Text at 0x152c556d0>
```



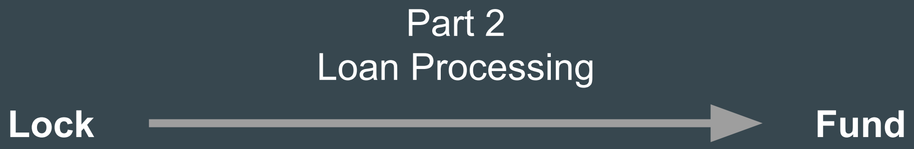
```
Years in Home
Mean: 8.0
1 sigma interval [0.529161169198, 15.4708388308]

<matplotlib.text.Text at 0x1532283d0>
```



```
Property address zip code
```

Feature	Borrower Age	Borrower's Annual Income	Borrower current employment	Home purchase cost	Years in home	Property zip code
Mean	44	127,000	7.5	406,000	8	98103
Common range	33 - 54	6,300 - 247,000	0 - 15	190,000 - 614,000	0.5 - 15	98103, 92691, 98125, 93003

Part 2 - Loan Processing

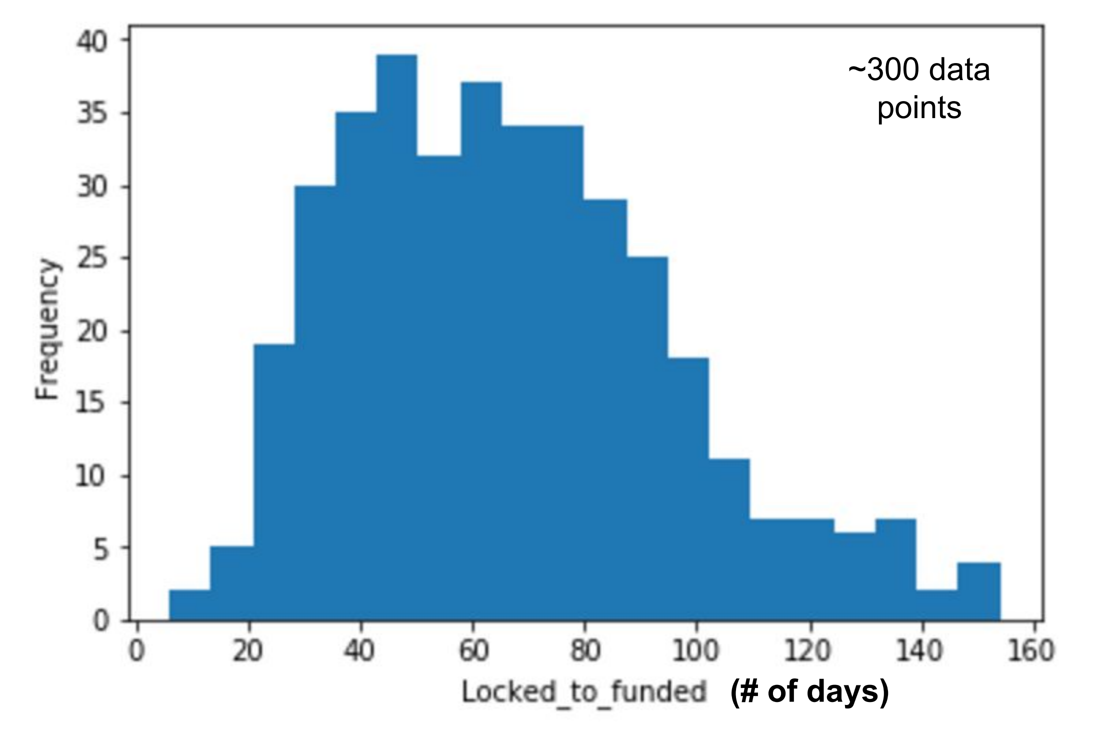


Part 2
Loan Processing

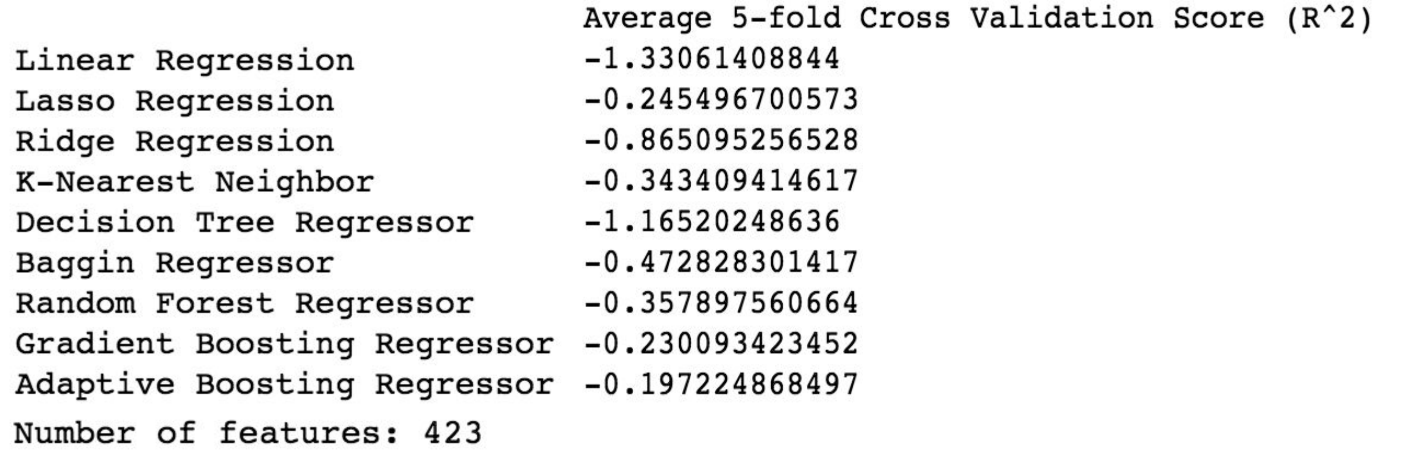
Lock

Fund

Part 2 - Results



Part 2 - Results

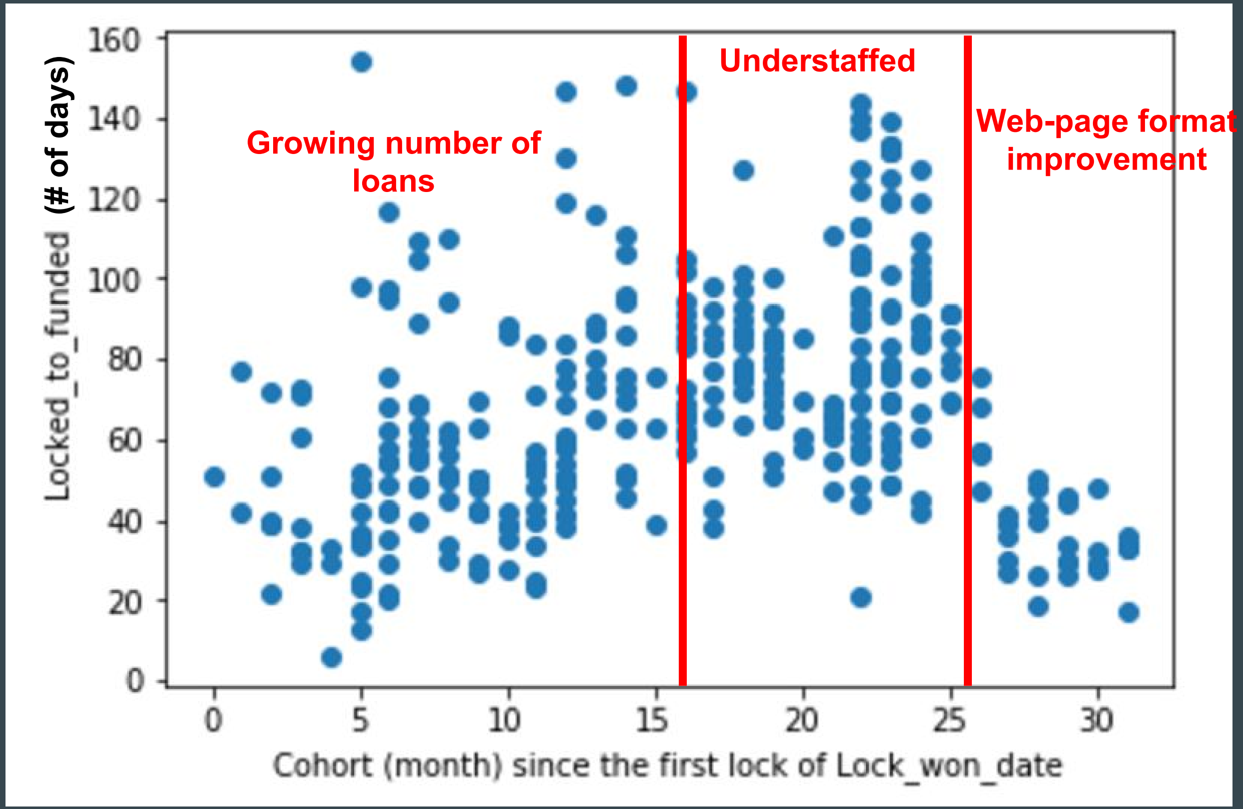


```
                            Average 5-fold Cross Validation Score (R^2)
Linear Regression           -1.33061408844
Lasso Regression            -0.245496700573
Ridge Regression            -0.865095256528
K-Nearest Neighbor          -0.343409414617
Decision Tree Regressor     -1.16520248636
Baggin Regressor            -0.472828301417
Random Forest Regressor     -0.357897560664
Gradient Boosting Regressor -0.230093423452
Adaptive Boosting Regressor -0.197224868497

Number of features: 423
```

Outcome:
- Experimented with different features, different algorithms
- No strong signal (near zero R2)

Potential reason:
- Data set too small
- Signals lie in external factors

Part 2 - Insights



Growing number of loans

Understaffed

Web-page format improvement

(the higher, the worse)

(or, months since the company launched)

Conclusions

Part 1

- Multistage model construction, Oversampling, Random Forest Classification
- Outperform existing models in Accuracy, F1-score, Precision
- Import feature extracted, Best Customer Profiling

Conclusions

Part 1

- Multistage model construction, Oversampling, Random Forest Classification
- Outperform existing models in Accuracy, F1-score, Precision
- Import feature extracted, Best Customer Profiling

Part 2

- Too few data points (~300) to detect signal for prediction
- Time cohort analysis indicated trends in outcome

Conclusions

Part 1

- Multistage model construction, Oversampling, Random Forest Classification
- Outperform existing models in Accuracy, F1-score, Precision
- Import feature extracted, Best Customer Profiling

Part 2

- Too few data points (~300) to detect signal for prediction
- Time cohort analysis indicated trends in outcome

Future work

- Implement prediction pipeline into existing workflow
- Include new features (from internal and external sources) into prediction models
- Try different models (neural network) that can take care of wide data sets

Thank you

Wei Cheung

cheungwz@gmail.com

github.com/weizhic