Mortgage Default Prediction

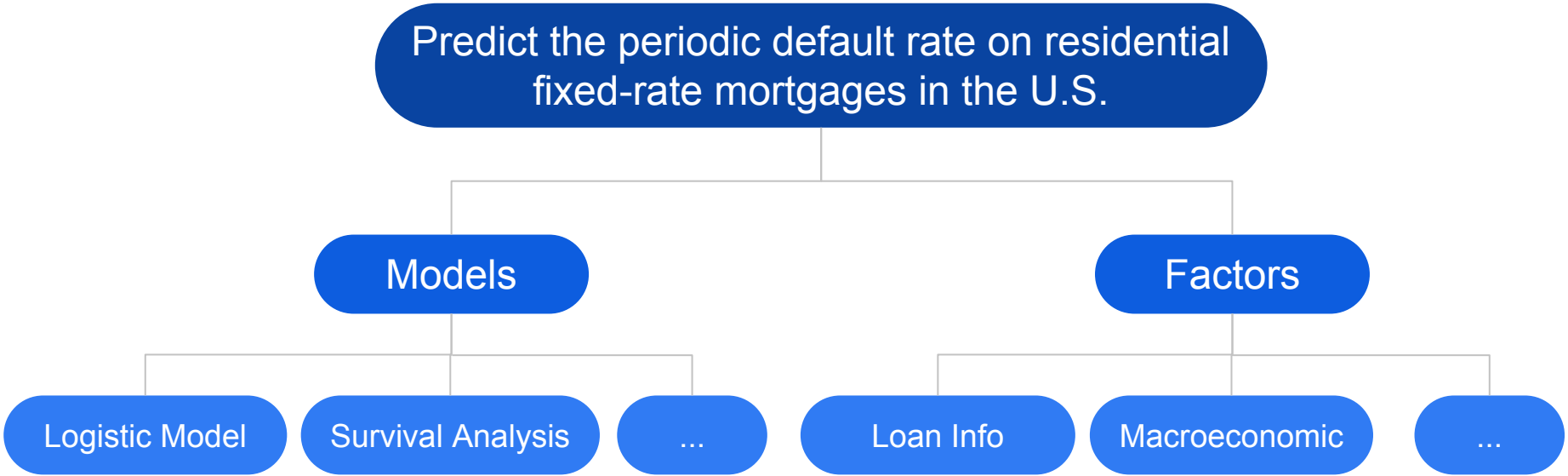
April 2, 2019

Tianxiao Chen, Jessica Chau, David Veitch and Shirly Wang

Agenda

1) Intro of the Problem/Criteria from CPPIB
2) Quick summary of research papers
3) Data collection and data
4) Exploratory data analysis
5) Model Approach
6) Model Evaluation
7) Results
8) Challenges of the project
9) Q&A

The Problem


Predict the periodic default rate on residential fixed-rate mortgages in the U.S.
Models
Factors
Logistic Model
Survival Analysis
...
Loan Info
Macroeconomic
...

Criteria from CPPIB

Expectations:

- Specification of the objectives
- Cross validation of third party vendor
- Interpretability of results

Requirements:

- Prefer to use interpretable models (no black-box models)
- Display the magnitude of the covariates' impact on the score
- Investigate microeconomic and macroeconomic factors

Literature Review

Candidate Factors:

- Loan and borrower variables
  - Loan-to-value ratio
  - Loan interest rate
  - Mortgage age
  - Credit quality
- Property variables
  - Year built
  - Property condition
- Market variables
  - House price appreciation
  - Growth in employment
  - Change in Treasury yield

Li, Mingxin. "Residential mortgage probability of default models and methods." *The Journal of Financial Institutions Commission, British Columbia* (2014).

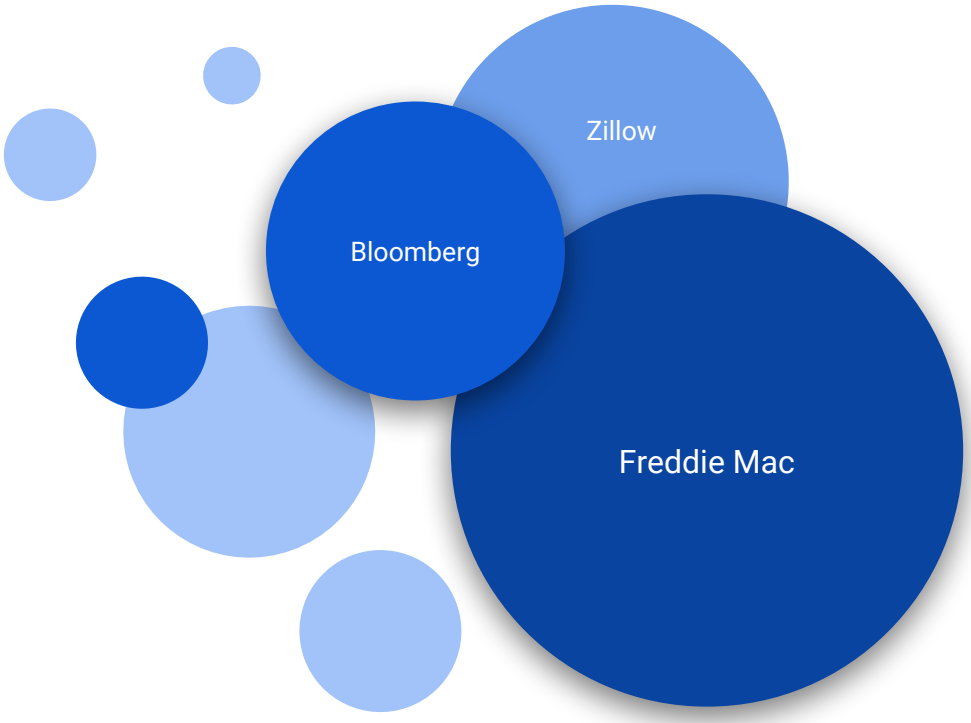
Literature Review

Candidate Models:

- Multinomial logistic models
  - Predict probability of default over a fixed time horizon
- Survival models
  - Estimate the relationship between default probability over time
  - Model fitting is tricky with time covariates
- Optimization models
  - Maximize wealth and utility of the borrower
  - Need assumptions on how house prices and interest rates evolve over time

Li, Mingxin. "Residential mortgage probability of default models and methods." *The Journal of Financial Institutions Commission, British Columbia* (2014).

Data Collection and Data



Data Collection and Data

Freddie Mac:

- Extract loan-level data from 1999 to 2017, including credit score, origination interest rate, first time home buyer, house location etc.
- Define default as 90-day delinquency
- Use stratified sampling to shrink data size due to the limitation of computing power

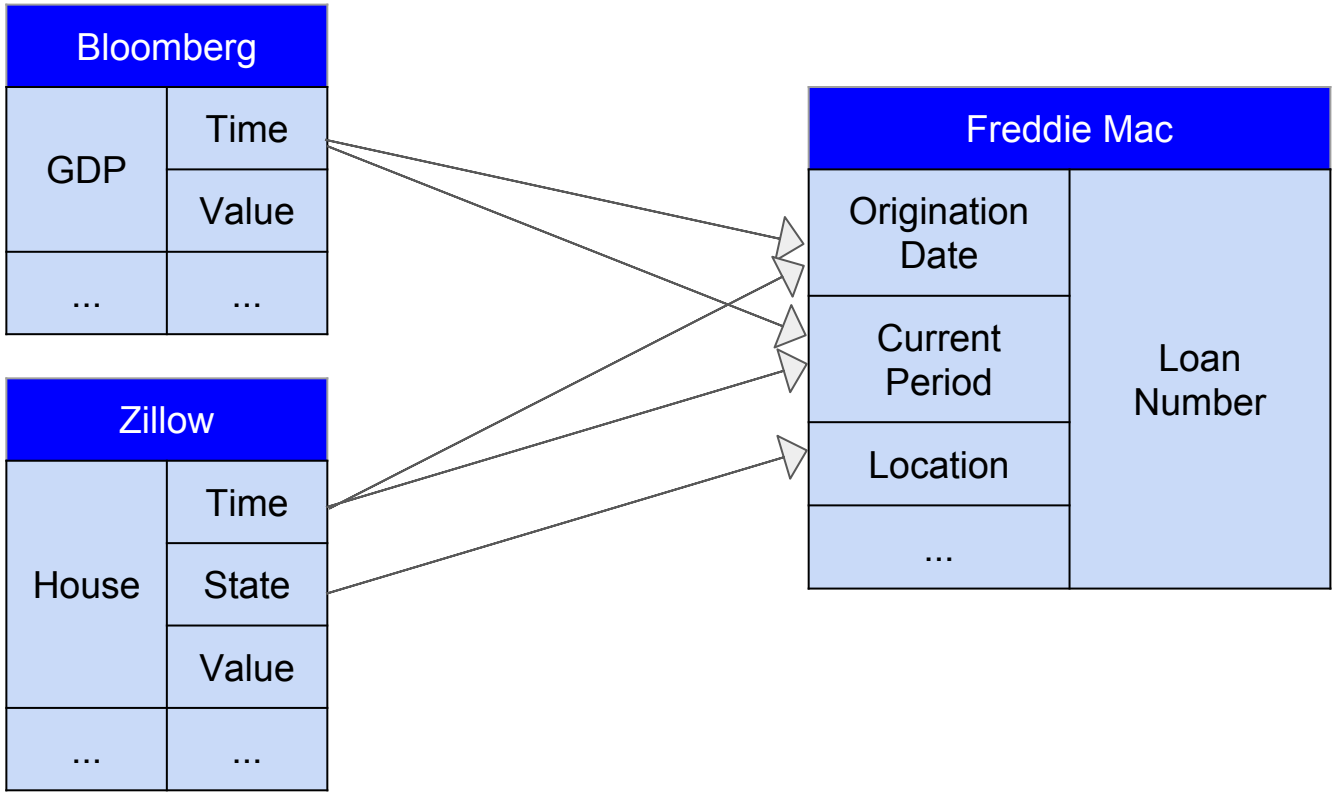
Bloomberg:

- Collect macroeconomic factors and real estate indicators that may affect default risk, including GDP, treasury yields, CPI, unemployment rate etc.

Zillow:

- Obtain U.S. state level housing price

Data Collection and Data



Data Collection and Data

loan seq num	period	default	prepay	credit score	curr upb	upb_prev	loan age	first time homebuyer flag	Housing Starts	...	New Home Sales	New Home Sales_orig	State Unemployment Rate	State Unemployment Rate_orig
F100Q1000066	2000-03-01	0	0	756.0	130000.00	130000.00	0.0	N	1737.0	...	856.0	900.0	3.1	3.1
F100Q1000066	2000-06-01	0	0	756.0	129000.00	130000.00	3.0	N	1575.0	...	857.0	900.0	3.1	3.1
F100Q1000066	2000-09-01	0	0	756.0	129285.64	129000.00	6.0	N	1541.0	...	848.0	900.0	3.0	3.1
F100Q1000066	2000-12-01	0	0	756.0	129007.80	129285.64	9.0	N	1551.0	...	880.0	900.0	2.9	3.1
F100Q1000066	2001-03-01	0	0	756.0	128724.38	129007.80	12.0	N	1625.0	...	963.0	900.0	3.1	3.1

Exploratory Analysis

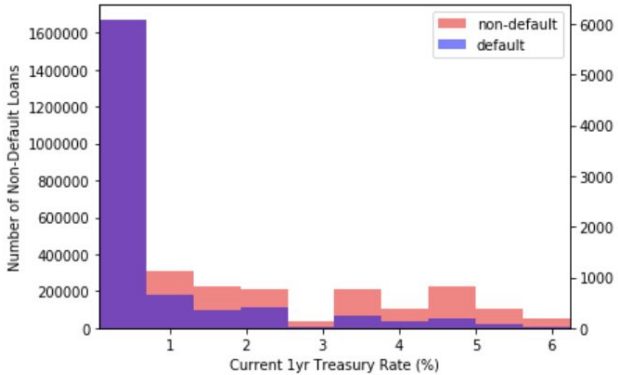
Used for Initial Feature Selection

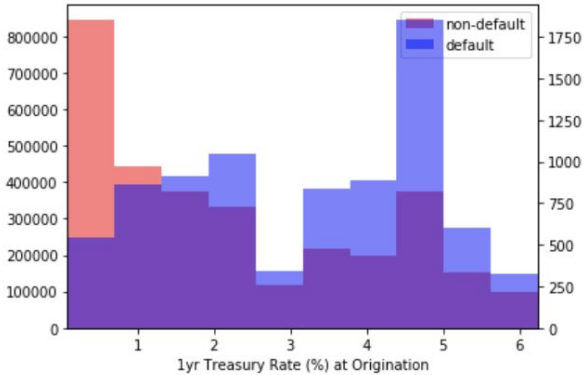
Methods:

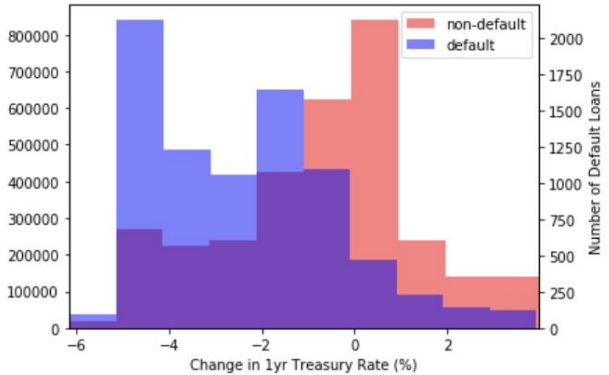
- Histograms with default loans vs non-default loans and prepaid loans vs non-prepaid loans
- For time series variables, consider values as of origination date, current date, and the difference between the two dates

Exploratory Analysis

1 Year Treasury Rate

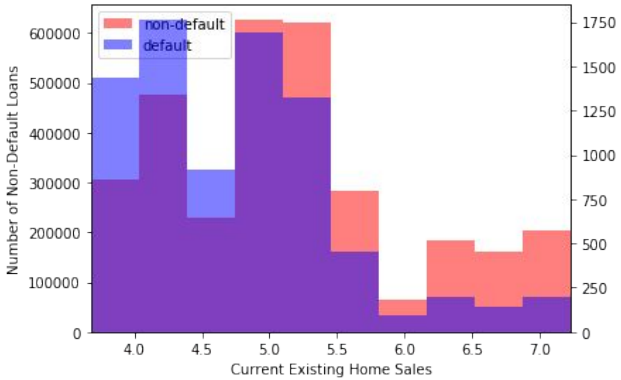


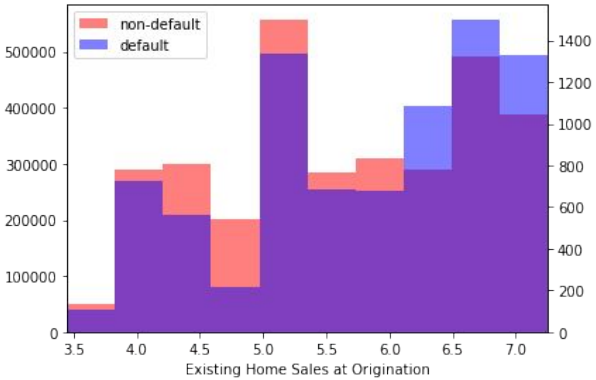


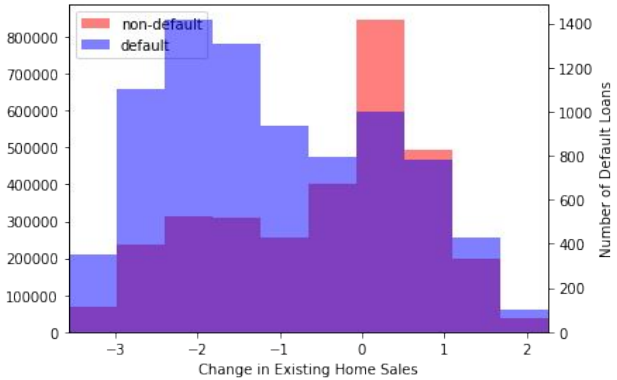


Exploratory Analysis

Existing Home Sales

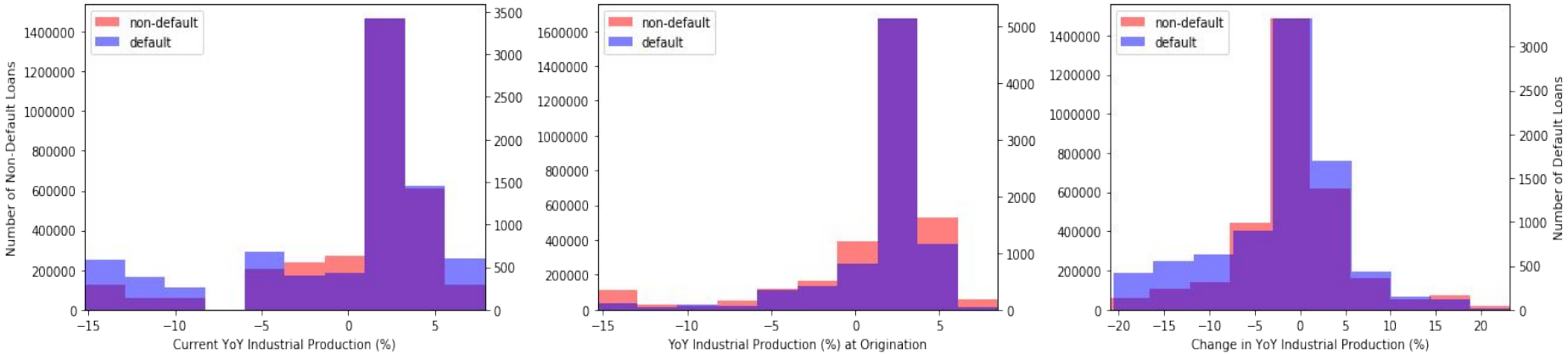






Exploratory Analysis

YoY Industrial Production



Model Approach - Key Features

Input | Model | Output

Loan Number: *F199Q1185054*
Quarter: *2001-03-01*

Covariates
Number of Borrowers: 2
Loan Age: 25 months
Loan-to-Value: 50%
House Price Change: 1.23x
State Unemployment: 3.7%
...

P(Default)	0.51%
P(Prepay)	4.32%
P(Nothing)	95.17%
Actual	**Nothing**

- Multinomial Logistic Regression Model with Regularization Penalty
  - GLM -> high interpretability and produces coefficients with confidence intervals
  - Able to model multiple events

Model Approach - Specification

Learning Model Parameters:

$$\ln \frac{\Pr(Y_{m;t} = \text{Def})}{\Pr(Y_{m;t} = \text{Nothing})} = \beta_0 + \beta_1 x_{m;t;1} + \cdots + \beta_K x_{m;t;K}$$

$$\ln \frac{\Pr(Y_{m;t} = \text{Pre})}{\Pr(Y_{m;t} = \text{Nothing})} = \gamma_0 + \gamma_1 x_{m;t;1} + \cdots + \gamma_K x_{m;t;K}$$

$$\Pr(Y_{m;t} = \text{Nothing}) = 1 - \Pr(Y_{m;t} = \text{Def}) - \Pr(Y_{m;t} = \text{Pre})$$

Model Approach - Regularization

Regularization Penalty (L1):

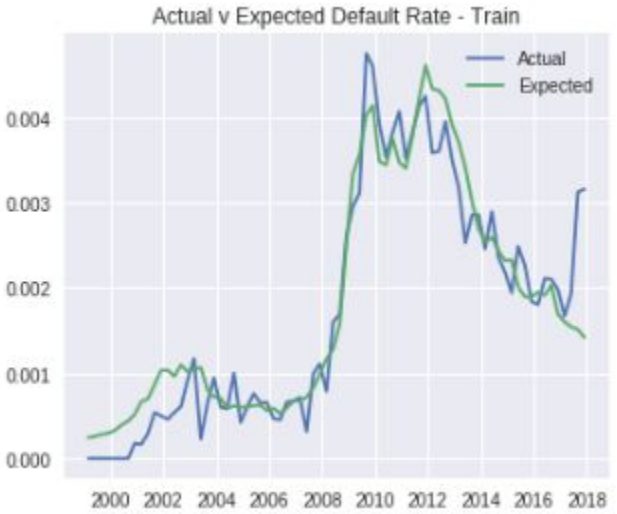
$$\text{Loss} = -(\text{Log-Likelihood}) + \text{Regularization Penalty}$$

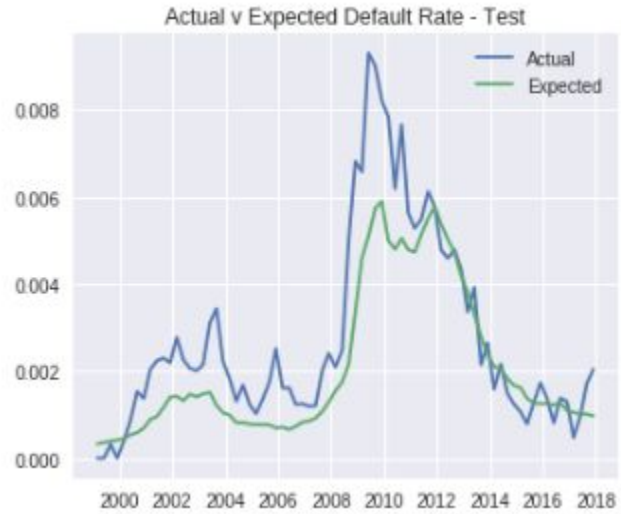
$$= -(\text{Log-Likelihood}) + \sum_{k=1}^{K} \alpha_k |\beta_k| + \nu_k |\gamma_k|$$

Initial Model	Variable Selection	Final Model
Fit model with all variables and a regularization penalty to help variable selection.	Discard all statistically insignificant variables.	Fit model with smaller subset of variables with regularization penalty to help generalization.

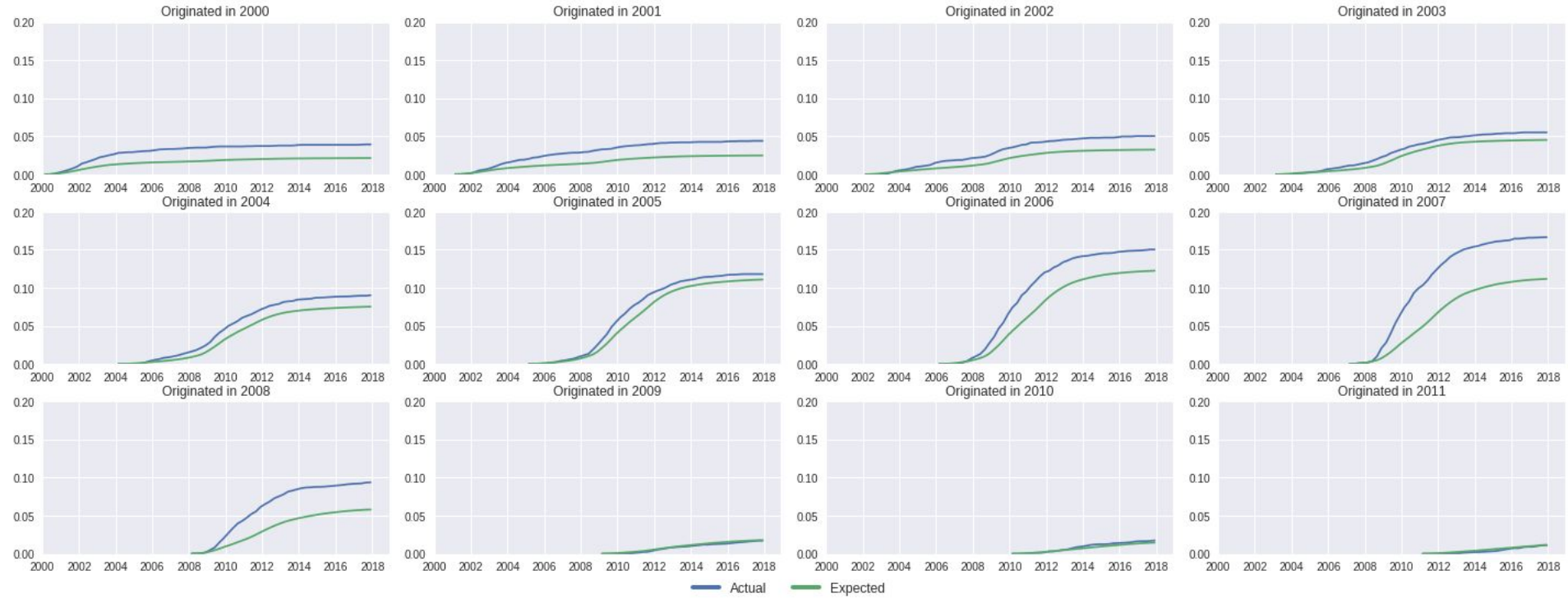
Model Evaluation - Aggregate

- Mean absolute difference of expected vs. actual default rate on test set of 0.08% (mean actual default rate ~0.3%)



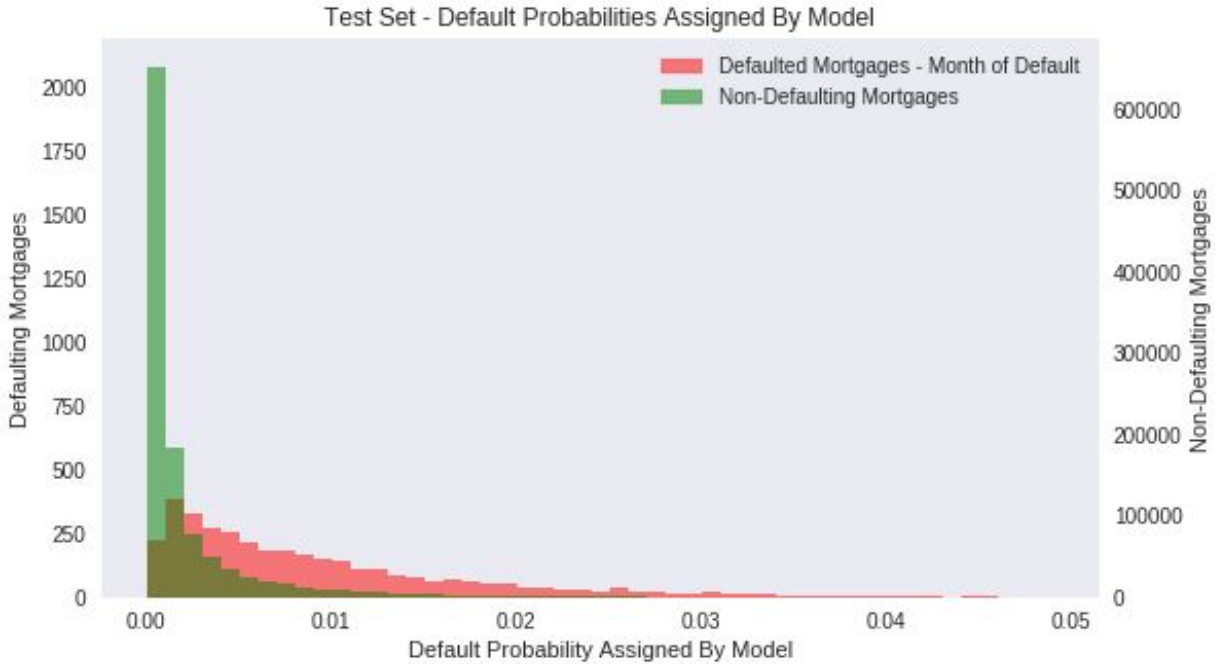


Model Evaluation - Default Rate By Origination Year



Model Evaluation - Loan Level Probabilities

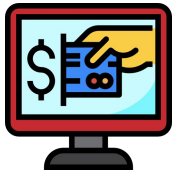
- Model learns to assign high default probabilities to mortgages that default



Model Results - Default

Default Model	Coefficient
constant	-6.924
1yr_diff	-0.104
2_borrowers	-0.148
30yr_diff	0.110
chan_broker	0.059
chan_not_specified	0.042
credit score	-0.403
dti	0.184
loan age	1.627
loan age squared	-1.038
ltv	0.415
orig interest rate	0.336
purpose_cashout	0.117
purpose_purchase	-0.088
Real GDP YoY	-0.075
State Unemployment Rate_diff	0.266
Treasury 30yr	-0.269
ZillowHouseChg	-0.329
ZillowHouseValue	0.132

Higher debt-to-income (DTI);
Higher (loan to value) LTV



Credit Score, Loan age, interest rates, unemployment rate

Brokers and unspecified channels have a higher default

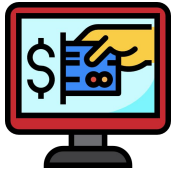
Number of borrowers, house value

Prepay Model	Coefficient
constant	-4.236
1yr_diff	-0.051
2_borrowers	0.130
30yr Fixed Rate Mortgage Average	-0.477
30yr_diff	-0.136
Building Permits	0.619
Building Permits_diff	-0.194
chan_broker	0.019
credit score	0.156
dti	0.016
Existing Home Sales	-0.360
Existing Home Sales_diff	0.275
Housing Starts	-0.177
loan age	2.628
loan age squared	-2.839
loan_age_cubed	0.715
ltv	0.013
New Home Sales	0.267
New Home Sales_diff	-0.124
OECD Leading Indicator YoY	0.364
orig interest rate	0.097
purpose_cashout	-0.034
purpose_purchase	0.028
Real GDP YoY	-0.074
SP500 YoY	-0.367
State Unemployment Rate	0.210
State Unemployment Rate Squared	-0.309
State Unemployment Rate_diff	0.083
Treasury 1yr	-0.341
Treasury 30yr	0.111
ZillowHouseChg	-0.021
ZillowHouseValue	0.056

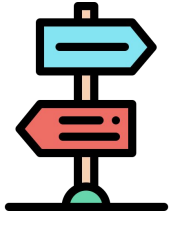
Model Results - Prepay

Loan age, more than 1 borrower

Number of Building Permits, broker vs, retailer



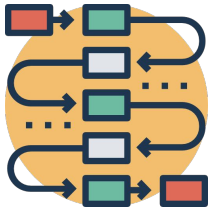
Credit Score, DTI, LTV



OECD Indicator, unemployment rate

Challenges of the Project

External Data Collection



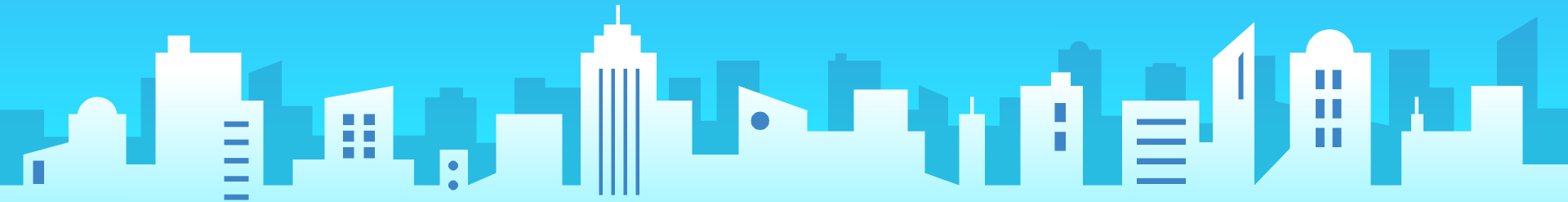
Processing power

Covariate Selection

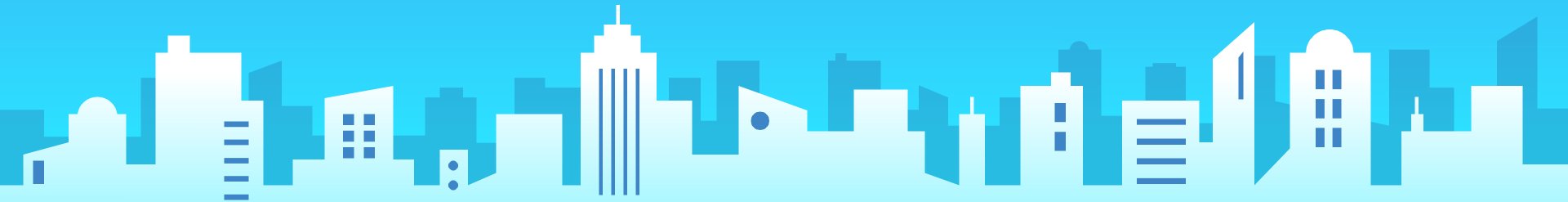
Discussions

Information Sharing

Q&A



Appendix



Appendix - Prepay