Exploring payment trends in single-family mortgage data

JEANA CURRO

DATA SCIENCE- FINAL PROJECT 1

The dataset:

- Released November 2016
- Approximately 22.5 million single-family residential mortgage loans originated between January 1, 1999 – September 30, 2015 that are currently guaranteed by Freddie Mac
- We will use a smaller **sample dataset** of 837,500 loans: 50,000 loans randomly selected from each origination year
- Data is organized in smaller data files: for each calendar year:
  - origination data file: includes loan characteristics at origination
  - monthly performance data file: includes credit performance and actual loss data which goes through March 31, 2016
- Useful links:
  - Dataset download: http://www.freddiemac.com/news/finance/sf_loanlevel_dataset.html
  - Dataset user guide: http://www.freddiemac.com/news/finance/pdf/user_guide.pdf

The dataset (continued)

ORIGINATION DATA FILE

Includes:

VARIABLE	TYPE
fico	integer
first_pay_date	date
first_time_homebuyer	categorical (1)
mature_date	date
msa	categorical (5)
mi_pct	float
units	integer
occupancy	categorical (1)
cltv	float
dti	float
orig_bal	float
ltv	float
init_rate	float

VARIABLE	TYPE
channel	categorical (1)
prepay	categorical (1)
product	categorical (5)
state	categorical (2)
prop_type	categorical (2)
prop_zip	categorical (5)
loan_id	categorical (12)
purpose	categorical (1)
orig_term	integer
num_borrowers	integer
seller	categorical (20)
servicer	categorical (20)
super_conforming_flag	categorical (1)

MONTHLY PERFORMANCE DATA FILE

Includes:

VARIABLE	TYPE
loan_id	categorical (12)
factor_date	date
act_bal	float
mba	categorical (3)
age	integer
rem_term	integer
repurch_flag	categorical (1)
mod_flag	categorical (1)
zero_bal_code	categorical (2)
last_date	date
cpn	float
defer_bal	float
next_due_date	date

VARIABLE	TYPE
mi_recov	float
net_sales_proceeds	categorical (14)
non_mi_recov	float
expenses	float
legal_costs	float
maint_preserve_costs	float
ti	float
misc_expenses	float
actual_loss	float

Full description of variables can be found here (starting p7): http://www.freddiemac.com/news/finance/pdf/user_guide.pdf

Potential project 1:

PROBLEM 1 :

Using the Freddie Mac loan level origination dataset which covers 837,500 randomly selected single-family mortgage loans spanning 1999 - 2015:

We will explore different origination trends pre and post the financial crisis (September 2008)

Hypothesis:

We hypothesize that loans originated prior to the crisis generally show weaker credit characteristics (e.g. lower credit scores, lower down-payments, and higher debt to income ratios) than loans originated post crisis.

Potential projects 2 & 3:

PROBLEM 2:

Using the Freddie Mac loan level performance dataset which covers the repayment and loss history of 837,500 single family mortgage loans from 1999 until Q1 2016:

We attempt to identify which characteristics correlate with a borrower voluntarily repaying his mortgage

We hypothesize that rate incentive will be the strongest driver of voluntary repayments but other factors such as larger loan balance and a higher credit score also play a part

PROBLEM 3:

Using the Freddie Mac loan level performance dataset which covers the repayment and loss history of 837,500 single family mortgage loans from 1999 until Q1 2016:

We attempt to identify which characteristics correlate with a borrower defaulting on his mortgage.

Bonus: we attempt to quantify probability of a borrower defaulting given the loan's origination characteristics

We hypothesize that loans with lower credit scores, lower-down payments, and higher debt to income ratios were the most likely to become delinquent.

Past research:

1. The Urban Institute, July 21 2015: *The Credit Shows Early Signs of Loosening*
2. Federal Reserve Bank of Kansas City, July 7 2014: *Tight Credit Conditions Continue to Constrain the Housing Recovery*
3. Federal Reserve Board Divisions of Research, Statistics and Monetary Affairs, November 2008: *The Rise in Mortgage Defaults*