

Predict Mortgage Default

Strictly For Fun and Learn!!

Hemant Bundeale

05/19/2014

Questions!

Just ask me latter

or

ask Aaron right now!!

Agenda

- What is problem statement
- Dataset
- Model
- Predict
- Next Steps

Problem statement

- How can a borrower find out likelihood of default?
- How can bank price the risk of mortgage default?
 - Banks already know how to do, this is just another attempt.

Dataset to train model

- Last year Fannie Mae and Freddie Mac released loan level dataset for 25-35 year fixed mortgage loan level information
- This includes Acquisitions and monthly servicing information

Fannie Mae Loan level dataset

- It has 18.7 million mortgage loans.
- Fixed-Rate, fully amortizing mortgage loans
- Original Term greater than 25 and less than 35 years
- Full documentation mortgage loans
- Acquisition data file has 22 features in it.
- Monthly servicing has 15 features(including one target feature).

Fannie Mae Acquisition Fields

POSITION	FIELD NAME	TYPE
1	LOAN IDENTIFIER	ALPHA-NUMERIC
2	CHANNEL	ALPHA-NUMERIC
3	SELLER NAME	ALPHA-NUMERIC
4	ORIGINAL INTEREST RATE	NUMERIC
5	ORIGINAL UNPAID PRINCIPAL BALANCE (UPB)	NUMERIC
6	ORIGINAL LOAN TERM	NUMERIC
7	ORIGINATION DATE	DATE
8	FIRST PAYMENT DATE	DATE
9	ORIGINAL LOAN-TO-VALUE (LTV)	NUMERIC
10	ORIGINAL COMBINED LOAN-TO-VALUE (CLTV)	NUMERIC 14
11	NUMBER OF BORROWERS	NUMERIC
12	DEBT-TO-INCOME RATIO	NUMERIC
13	CREDIT SCORE	NUMERIC
14	FIRST-TIME HOME BUYER INDICATOR	ALPHA-NUMERIC
15	LOAN PURPOSE	ALPHA-NUMERIC
16	PROPERTY TYPE	ALPHA-NUMERIC
17	NUMBER OF UNITS	ALPHA-NUMERIC
18	OCCUPANCY STATUS	ALPHA-NUMERIC
19	PROPERTY STATE	ALPHA-NUMERIC
20	ZIP (3-DIGIT)	ALPHA-NUMERIC
21	MORTGAGE INSURANCE PERCENTAGE	NUMERIC
22	PRODUCT TYPE	ALPHA-NUMERIC

Fannie Mae servicing fields

POSITION	FIELD NAME	TYPE
1	LOAN IDENTIFIER	ALPHA-NUMERIC
2	MONTHLY REPORTING PERIOD	DATE
3	SERVICER NAME	ALPHA-NUMERIC
4	CURRENT INTEREST RATE	NUMERIC
5	CURRENT ACTUAL UNPAID PRINCIPAL BALANCE (UPB)	NUMERIC
6	LOAN AGE	NUMERIC
7	REMAINING MONTHS TO LEGAL MATURITY	NUMERIC
8	ADJUSTED REMAINING MONTHS TO MATURITY	NUMERIC
9	MATURITY DATE	DATE
10	METROPOLITAN STATISTICAL AREA	ALPHA-NUMERIC
11	CURRENT LOAN DELINQUENCY STATUS	ALPHA-NUMERIC
12	MODIFICATION FLAG	ALPHA-NUMERIC
13	ZERO BALANCE CODE	ALPHA-NUMERIC
14	ZERO BALANCE EFFECTIVE DATE	DATE
15	REPURCHASE INDICATOR	ALPHA-NUMERIC

Freddie Mac Loan level dataset

- It has 16.5 million mortgage loans.
- It has 719 million servicing records.
- 30-year fixed-rate mortgages that the company purchased or guaranteed from 1999 to 2012
- Acquisition data file has 25 features in it.
- Monthly servicing has 15 features(including one target feature).

Freddie Mac Acquisition Fields

POSITION	FIELD NAME	TYPE
1	CREDIT SCORE	NUMERIC
2	FIRST PAYMENT DATE	DATE
3	FIRST-TIME HOME BUYER FLAG	ALPHA-NUMERIC
4	MATURITY DATE	DATE
5	METROPOLITAN STATISTICAL AREA (MSA) OR METROPOL	NUMERIC
6	MORTGAGE INSURANCE PERCENTAGE (MI %)	NUMERIC
7	NUMBER OF UNITS	NUMERIC
8	OCCUPANCY STATUS	ALPHA-NUMERIC
9	ORIGINAL COMBINED LOAN-TO-VALUE (CLTV)	NUMERIC 14
10	ORIGINAL DEBT-TO-INCOME (DTI) RATIO	NUMERIC
11	ORIGINAL UPB	NUMERIC 14
12	ORIGINAL LOAN-TO-VALUE (LTV)	NUMERIC
13	ORIGINAL INTEREST RATE	NUMERIC
14	CHANNEL	ALPHA-NUMERIC
15	PREPAYMENT PENALTY MORTGAGE (PPM) FLAG	ALPHA-NUMERIC
16	PRODUCT TYPE	ALPHA-NUMERIC
17	PROPERTY STATE	ALPHA-NUMERIC
18	PROPERTY TYPE	ALPHA-NUMERIC
19	POSTAL CODE	ALPHA-NUMERIC
20	LOAN SEQUENCE NUMBER -	ALPHA-NUMERIC
21	LOAN PURPOSE	ALPHA-NUMERIC
22	ORIGINAL LOAN TERM	NUMERIC
23	NUMBER OF BORROWERS	NUMERIC
24	SELLER NAME	ALPHA-NUMERIC
25	SERVICER NAME	ALPHA-NUMERIC

Freddie Mac servicing fields

POSITION	FIELD NAME	TYPE
1	LOAN SEQUENCE NUMBER	ALPHA-NUMERIC
2	MONTHLY REPORTING PERIOD	DATE
3	CURRENT ACTUAL UPB	NUMERIC
4	CURRENT LOAN DELINQUENCY STATUS	NUMERIC
5	LOAN AGE	NUMERIC
6	REMAINING MONTHS TO LEGAL MATURITY	NUMERIC
7	REPURCHASE FLAG	ALPHA-NUMERIC
8	MODIFICATION FLAG	NUMERIC
9	ZERO BALANCE CODE	NUMERIC
10	ZERO BALANCE EFFECTIVE DATE	DATE
11	CURRENT INTEREST RATE	NUMERIC

What are looking for

- Zero Balance Code == 97
 - Means Loan has been delinquent for more than 180 days(D180)
- Number of months since the first payment date for the loan and the default (D180)

Data Munging – Tidy data

- Load 200,00 loans - one quarter of data in HP Vertica
- Drop rows with empty values
- Create a big row – with initial acquisition data and last servicing record
- Get the number of months it took for loans to be 6 months delinquent

Training the model

- Logistic Regression
 - Could predict the default – low accuracy
 - Not the number of months to default
- Random Forest classification
 - can predict default and number of months to default
 - Much better accuracy than Logistic regression

Which were important features?

Based on current training dataset(Q1 2000):

- Original Unpaid Principal Balance
- Debt To Income Ratio
- Credit Score
- Zip code (First Three Digit)

Ask for these features from user and supply that to trained model for prediction.

What user is expected to get back?

- Whether mortgage will be 180 days delinquent in future
- After how many months mortgage will be delinquent



Predict if mortgage will default!

Logged in as John Doe [sign out](#)

Predict if mortgage will default by providing values for following!

Original Unpaid Principal Balance:	<input type="text" value="800000"/>
Debt To Income Ratio:	<input type="text" value="35"/>
Credit Score:	<input type="text" value="500"/>
Zip Code - Three Digit:	<input type="text" value="201"/>

Default or Not?:



This is currently work in progress!

Challenges

- Training Random forest on data of size roughly of 200 GB
- Understanding which feature has the most impact on prediction
- How to predict the month/year when the mortgage will default
- Deal with Problem of over fitting with Random Forest
 - With 200,000 loans accuracy is 97%+

Next steps

- Train on complete dataset
- Re-evaluate the important features
- Web App – that lets user predict default for themselves
- Merge with other public data – like Census

www.PredictMortgageDefault.com

Technology

- Data Munging
 - HP – Vertica – on local laptop
 - Planning to use
 - Redshift for preparing tidy data
 - EC2 – instance(32 core – 60 GB RAM)
- Classification
 - Python, scikit
- Web App
 - Grails/Groovy

Contact

Hemant@ibLaunchEnergy.com

703 862 0517

Thanks for listening.

References:

http://www.freddiemac.com/news/finance/sf_loanlevel_dataset.html

<http://www.fanniemae.com/portal/funding-the-market/data/loan-performance-data.html>