2009 Home Mortgage Disclosure Act (HMDA) Loan Application Register (LAR) Data

Jonathan Howton

Project Purpose

- Dataset:
  - 2009 home mortgage loan application register data reported by certain banks, credit unions, savings associations, and non-depository institutions pursuant to the Home Mortgage Disclosure Act (HMDA)
- Goals:
  - Identify most important attributes for approval/denial
  - Test different models for predicting approval/denial (Not quite finished)

Dataset

- Originally over 19 million entries
- Over 12 million entries after preprocessing
- 45 Features

HMDA Loan Application Register Format

Fields	Length	Type
As of Year	4	Numeric
Respondent ID	10	Alphanumeric
Agency Code	1	Alphanumeric
Loan Type	1	Numeric
Property Type	1	Alphanumeric
Loan Purpose	1	Numeric
Occupancy	1	Numeric
Loan Amount (000s)	5	Numeric
Preapproval	1	Alphanumeric
Action Type	1	Numeric
MSA/MD	5	Alphanumeric
State Code	2	Alphanumeric
County Code	3	Alphanumeric
Census Tract Number	7	Alphanumeric
Applicant Ethnicity	1	Alphanumeric
Co Applicant Ethnicity	1	Alphanumeric
Applicant Race 1	1	Alphanumeric
Applicant Race 2	1	Alphanumeric
Applicant Race 3	1	Alphanumeric
Applicant Race 4	1	Alphanumeric
Applicant Race 5	1	Alphanumeric
Co Applicant Race 1	1	Alphanumeric
Co Applicant Race 2	1	Alphanumeric
Co Applicant Race 3	1	Alphanumeric
Co Applicant Race 4	1	Alphanumeric
Co Applicant Race 5	1	Alphanumeric
Applicant Sex	1	Numeric
Co Applicant Sex	1	Numeric
Applicant Income (000s)	4	Alphanumeric
Purchaser Type	1	Alphanumeric
Denial Reason 1	1	Alphanumeric
Denial Reason 2	1	Alphanumeric
Denial Reason 3	1	Alphanumeric
Rate Spread	5	Alphanumeric
HOEPA Status	1	Alphanumeric
Lien Status	1	Alphanumeric
Edit Status	1	Alphanumeric
Sequence Number	7	Alphanumeric
Population	8	Alphanumeric
Minority Population %	6	Alphanumeric
HUD Median Family Income	8	Alphanumeric
Tract to MSA/MD Income %	6	Alphanumeric
Number of Owner-occupied units	8	Alphanumeric
Number of 1-to 4-Family units	8	Alphanumeric
Application Date Indicator	1	Numeric

Preprocessing Steps

- Only entries with accept or deny status (to 12 million +)
- Dropped several columns
- Dropped all rows with missing values (Still 12 million +)
- Scaled using MinMaxScaler
- Binary Encoding using pandas getdummies
    - 45 Features

Random Forest Based Importance

Tried ExtraTrees, but used too much memory

Random Forest

- Bootstrap samples
- Randomly select features
- Build ensemble of decision trees

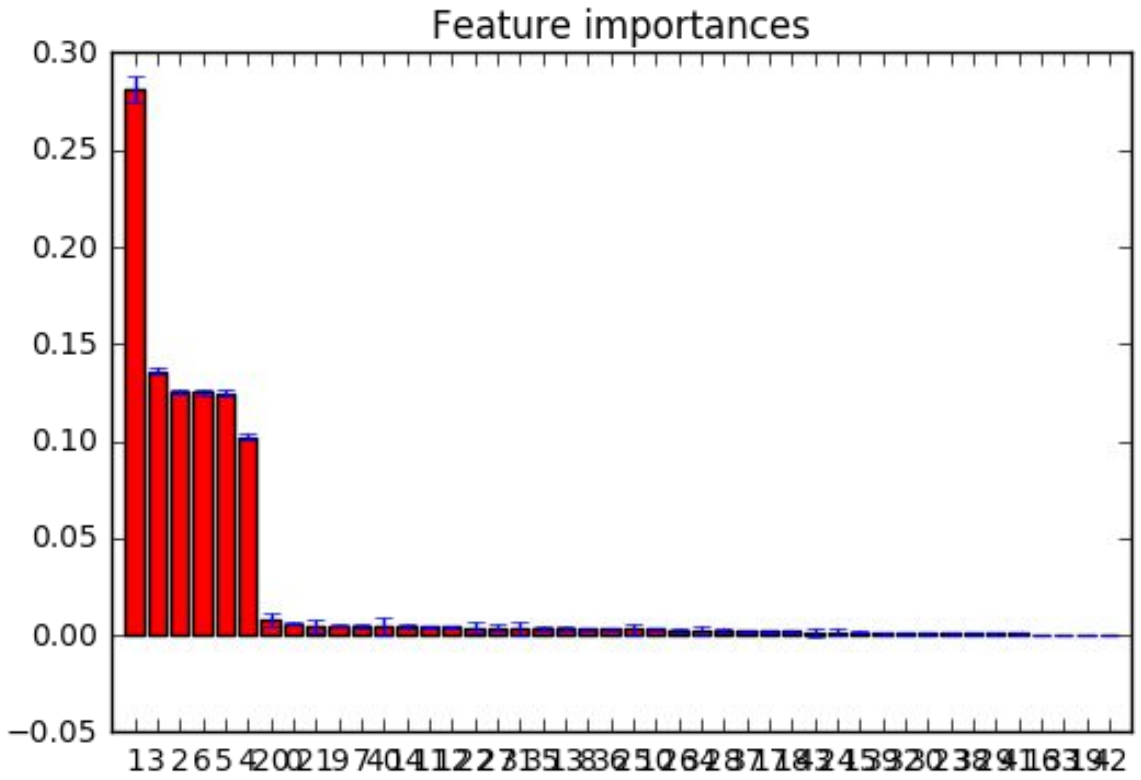
My Random Forest

- 100 Trees (250 tree = too much memory)
- ~70% Accuracy

Random Forest Based Importance

Top 6

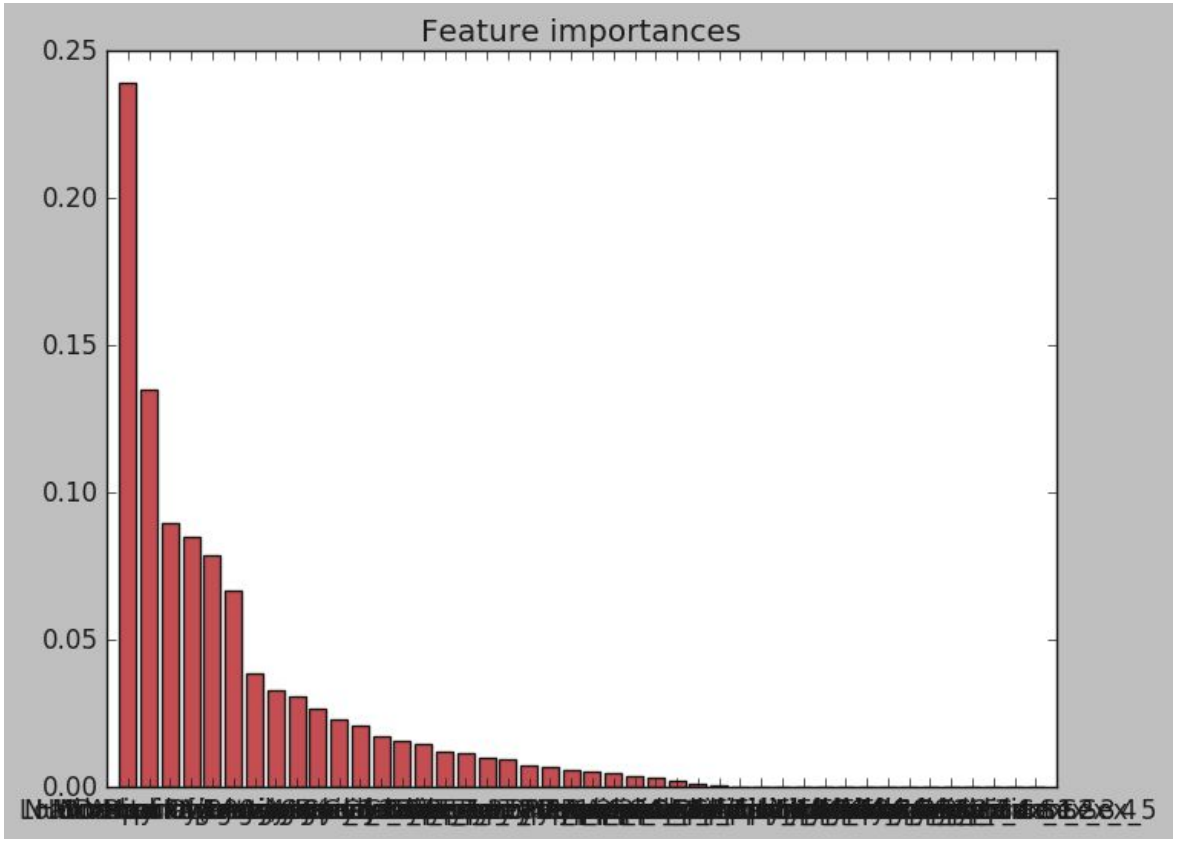
- Loan Amount (000s) (0.281669)
- Minority Population % (0.136070)
- Population (0.125451)
- Number of 1- to 4-Family units (0.125167)
- Number of Owner-occupied units (0.124825)
- HUD Median Family Income (0.102179)



Principal Component Analysis

Above 0.05

- Occupancy (0.239321)
- Loan Amount (000s) (0.135050)
- Population (0.089684)
- Minority Population % (0.085135)
- HUD Median Family Income (0.078909)
- Number of Owner-occupied units (0.066519)



Recursive Factor Elimination Using Logistic Regression

- Minority Population %
- HUD Median Family Income
- Number of Owner-occupied units
- Number of 1- to 4-Family units
- Loan Type_3 (VA)
- Property Type_2 (Manufactured Housing)
- Loan Purpose_1 (Home Purchase)
- Loan Purpose_2 (Home Improvement)
- Preapproval_1 (Request for Preapproval)
- Applicant Sex_4 (Not Applicable (?))