US Mortgage Loan Data Analysis

Alyssa, Chris, Gerard, Seung

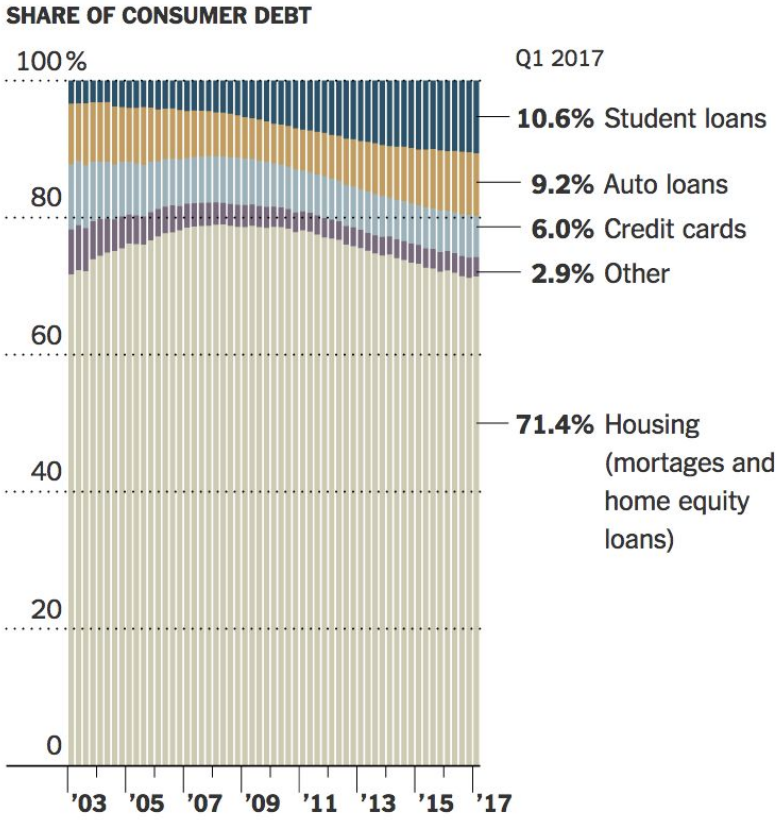
Problem & Motivation

Housing mortgages: primary component of ~$13 billion consumer debt in the United States.

Basis for an important class of financial asset; residential mortgage-backed securities (RMBS).

Predicting their future performance is a key task for securities analysts, especially the major ratings agencies.

STANDARD &POOR'S Moody's FitchRatings



Data: Freddie Mac/Fannie Mae

Acquisition Data

FICO Score
Loan-to-Value
Debt-to-Income
Freddie: 27 fields
Fannie: 25 fields

Monthly Performance Data

Remaining Balance
Sale
Recovery
Freddie: 23 fields
Fannie: 31 fields

Frequency

Quarterly update
Corrections possible

Coverage

Freddie: 1999 Q1-2016 Q3
Fannie: 2000 Q1-2016 Q2

Size

Freddie - 24 million loans (12GB)
Fannie - 35 million loans (23GB)
~500MB increase every quarter

Data: Housing Prices

Micro-Scale Data

Zillow API:
- 1000/day sampling
- Price record tracking

Redfin API:
- List-to-Sale spread
- % contract in 2 wks

Macro-Scale Data

US FRED:
- Avg Price by financing
- New Sales by price bucket

US Census, State & Urban: -
- New construction
- Vacancies

Size

US FRED: <10 KB
US Census: <10 KB
Zillow/Redfin: low, variable depending upon number of variables collected. Limit of 1000 calls per day.

Frequency

Micro: Monthly - Quarterly
Macro: Quarterly

Coverage

Redfin: Jan 2012-Present
Zillow: dependent upon unit
FRED:Sales buckets: 2002-Present
- Price by finance: 1988- Present

US Census: 2014-Present

Architecture Design

Acquire, store, and transform raw data: use Hadoop File System and Hive

Download the data

Load into HDFS

Load into Hive

Transform and clean data

Serving layer: Use SparkSQL/Pyspark and a Visualization tool

Use SparkSQL/Pyspark to allow open-ended querying of entire dataset

Create visualizations of particular insights (potentially Tableau or Data Studio)

THANKS!