Fair Lending Finder

Cody Gilbert, Fang Han, Jeremy Lao

NYU Courant, Computer Science

August 8, 2019

Introduction

- How would you like to know your chances of obtaining a mortgage related loan?
- We leverage Home Mortgage Disclosure Act (HMDA) data to tell you just that

Example

Give your application your details and it will return the lender with the highest approval probability: income, loan amount, race, gender, and state

How do we predict? Machine learning, duh

$$
\begin{aligned}
P(y = k) \quad = \quad & \beta_0 loanAmt_{obs} \\
& + \beta_1 applicantIncome_{obs} \\
& + \delta_0 race \\
& + \delta_1 gender \\
& + \delta_2 lender
\end{aligned}
$$

Where δ are dummy variables for the categorical variables and β are coefficients. The outcome (k), approve or deny, $k \in 0, 1$

Hmm, what is HMDA?

- Home Mortgage Disclosure Act
- Lenders are required to collect and report information about housing-related loans to the Consumer Financial Protection Bureau (CFPB)
- Data are shared in an anonnymised manner
- The CFPB and FDIC both monitor the data to ensure community reinvestment and fair lending

Our Application

We want to use the data to help you find the lender that will originate your loan

Three Data Sets

- Each data set supports and complements the other through offering data that helps provide more information for the analytic

HMDA Data Set

- $>$ 125 *GB*
- 2007-2017

Geo Data

- Geospatial
- $>$50 MB

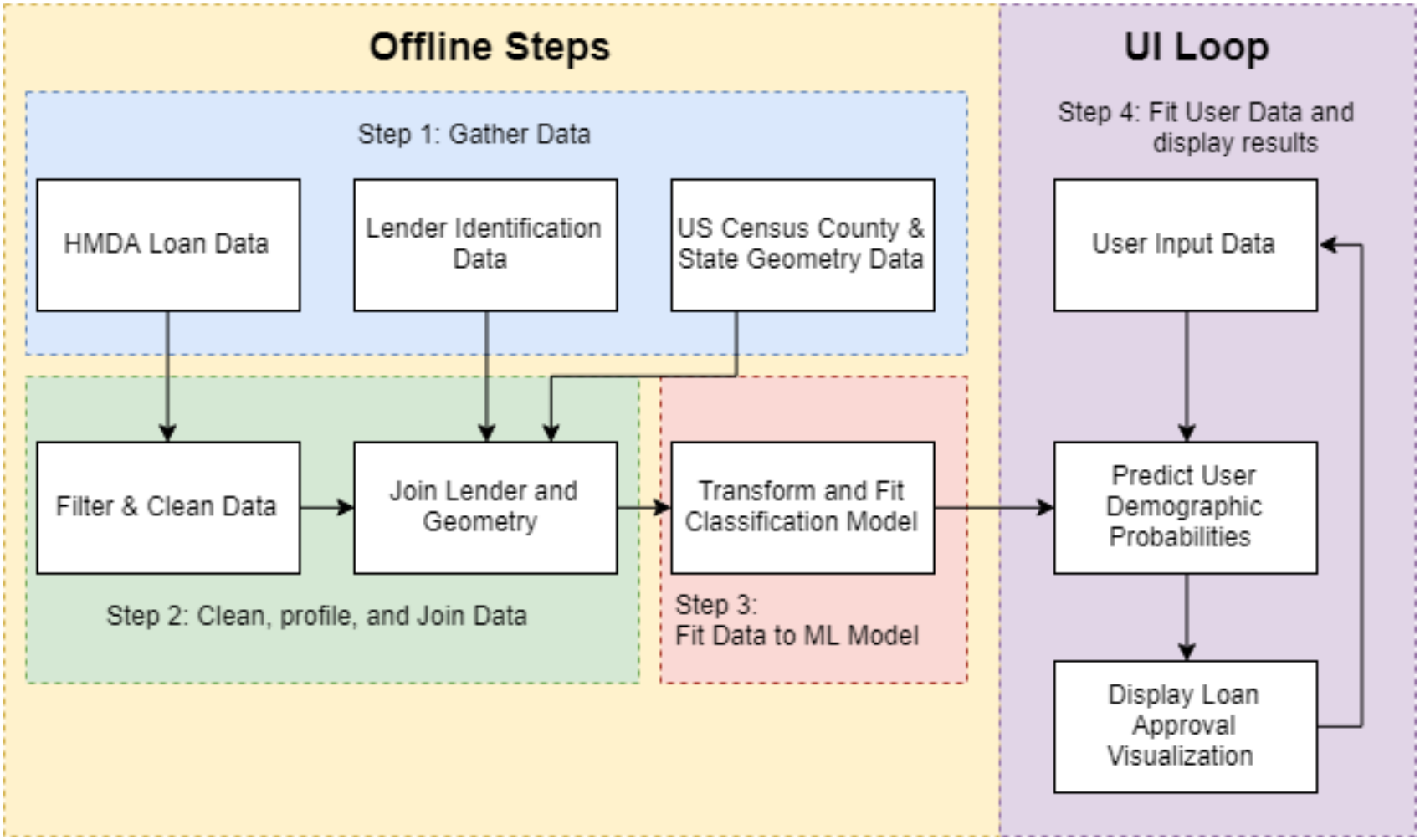
Panel Data

- Reporters
- $>$200 MB

Environments and Tools Used

- Scala Spark & Maven
- Packages: MLLib, Spark Context-RDD, Spark SQL-Dataframes, GeoSpark
- Python Plotly

Design Diagram



MLLib - Models

MLModel	Training/Test	AUC
Logistic Regression	80/20	60%
SVM	80/20	59%
Naive Bayes	80/20	79%

Table 1: Model Evaluation

MLLib - Models

MLModel	Training/Test	AUC
Logistic Regression	80/20	60%
SVM	80/20	59%
Naive Bayes	80/20	79%

- Naive Bayes also had the fastest model performance
- We chose Naive Bayes for the ML Model