Exploring Loan Discrimination with Big Data

Cody Gilbert
NYU Computer Science
New York, USA
cjg507@nyu.edu

Fang Han
NYU Computer Science
New York, USA
fh643@nyu.edu

Jeremy Lao
NYU Computer Science
Washington, USA
jjl359@nyu.edu

Abstract

The objective of this analysis is to draw actionable and insightful analysis regarding US loan approval bias using the Home Mortgage Disclosure Act (HMDA) data between 2007-2017, and to create an application for determining biases in lender practices. The tool was constructed using the Apache Spark framework to analyze and process data, with the MLLib package used to construct a model for predicting loan approval given loan applicant demographics. The model will produce an output that feeds a user interface (UI) to display a summary of what lenders are most likely to approve a loan given the user's input demographics. The final model used a Naive Bayes model with a poor predictive quality of 80

Index Terms

mortgage, HMDA, Naive Bayes, fair lending, logistic regression, loans

1. Introduction

In the ongoing struggle and debate over civil rights within the United States, the notion of equality in the delivery of essential services to persons of all races, genders, and ethnicities is considered an essential element to a free and productive nation. Housing, and the services to own and lease property, are critically important elements of these essential services that have had a tumultuous recorded history of discrimination within the US. To combat the loan inequality that hurt minority populations, US Congress enacted the 1975 Home Mortgage Disclosure Act (HMDA) that required lending institutions to supply loan data to the US Federal Government with the goal of identifying possible discriminatory lending patterns among lenders. To put it in plainly, if someone applies for a mortgage related loan, that data point and the anonymized metadata of the applicant are reported to the Consumer Financial Protection Bureau (CFPB). The objective of this analysis is use this HMDA data to create a user-interactive tool for discovering biases and trends in US lending practices. This tool will allow users to select a set of demographics to see how differences in race, ethnicity, and gender impact the chances of having a loan approved, how those chances have changed over time, and what institutions show the highest levels of bias. The goal of this project to provide potential lenders information on what lending institutions best fit their needs, and contribute an additional resource to the ongoing discussion of inequality within the United States.

2. Motivation

The Home Mortgage Disclosure Act (HMDA) dataset is one of the largest publicly available datasets in banking with application and demographic level data. While the banking regulators have used the dataset to monitor the mortgage market and unfair lending practices for many years, it is also important for the public to understand the dataset. The application that we are proposing is just one tool that can help the public utilize the publicly available information in choosing a lender when applying for a loan. The machine learning methods can help a loan applicant identify a lender that has the highest probability of lending to the applicant using the applicants metadata.

3. Related Work

Previous research on this topic are primarily based on the regulators perspective. For example, in [1] the banking regulators outlined the process by which they conduct investigations into unfair lending practices using the HMDA dataset as a starting point. The Federal Reserve and the FDIC both use the dataset to monitor and enforce both the community reinvestment act and fair lending practices. The Federal Reserve

publishes analysis performed on the HMDA dataset, and have highlighted both the rich analysis that can be obtained from the data along with the shortcomings of the dataset. For example, in [2] Federal Reserve economists highlight the importance of modeling and controlling for various loan application characteristics when interpreting the denial rate of a loan application by race and gender. Current regulatory research on the topic shows that the HMDA dataset contains valuable insight that can be analyzed and modeled to reveal patterns of discriminatory behavior. It is this insight that motivates the construction of the tool created in this project.

Based on the challenges of using geographic data in [2] due to the 2010 census update to Metropolitan Statistical Areas in 2014, we focused on analyzing data at the state and county level using FIPS (Federal Information Processing Standard) data. The authors at the Fed [2] also raised the issue of applications filed under HMDA identificaiton number of whatever entity owns the lender (i.e. regulator), if there is an acquisition in the middle of the year, the applications are filed under one lender. In addition, we found that additional external information is required to find the "high holder" of a lender from the panel data. This drove our decision to focus on the largest lenders in our application.

We found in [7] that the decrease in number of loan applications per annum between 2007 and 2010 is likely due to the mortgage meltdown of 2008. While we visualize denial rates from 2007-2017, our analytic & machine learning model looks at mortgage application data after 2010.

4. Datasets

4.1. Home Mortgage Disclosure Act

he Home Mortgage Disclosure Act (HMDA) was originally enacted by Congress in 1975 and is implemented by Regulation C. This regulation applies to certain financial institutions, including banks, savings associations, credit unions, and other mortgage lending institutions. HMDA requires many financial institutions to maintain, report, and publicly disclose loan-level information about mortgages. These data help show whether lenders are serving the housing needs of their communities. They give public officials information that helps them make decisions and policies, and they shed light on lending patterns that could be discriminatory. The public data are modified to protect applicant and borrower privacy. The data available for public use is summarized in Table 1.

Year	Activity Year	No. Reporters	No. Records (mm)
2017	2016	6,762	16.3
2016	2015	6,913	14.3
2015	2014	7,062	11.9
2014	2013	7,190	17
2013	2012	7,400	18.7
2012	2011	7,632	14.7
2011	2010	7,923	16.3
2010	2009	8,124	19.5
2009	2008	8,388	17.4
2008	2007	8,610	26.6
2007	2006	8,886	34.1

Table 1: HMDA Data Summary

4.1.1. HMDA Data Schema. This is the HMDA data schema:

Column of Interest	Range of Values
as_of_year	2007 to 2017
respondent_id	1-9 digit ID RSSD or 9 digit tax identifier (xx-xxxxxxx)
agency_code	one to nine representing different agencies
loan_type	there four distinct values (Conventional / FH / VA / FSA) however in some cases NA is submitted
property_type	there are only three types of property types but NA is sometimes submitted
loan_purpose	there are only three choices but NA if nothing is indicated on the application
owner_occupancy	there are three options (Owner occupied / not owner occupied / not applicable)
loan_amount_000s	the range is from 1000 dollars to 99 million dollars and there are some NA or blank values. However it appears that the loan amount is normally distributed between 1000 dollars to 500000 dollars
preapproval	three options
action_taken	there are 8 distinct actions that can be taken (such as loan originated or application denied)
state_code	string or integer of the abbreviations or FIPS number of the state (Federal Information Processing Standard)
county_code	the name of the county or a three digit FIPS number
applicant_sex	male or female are the options
applicant_income_000s	the range of incomes are 1000 to 99 million and the average income over the 11 year period is 3.3 million / however the weighted average income by frequency of applications per income listed is 98 thousand

Table 2

4.2. Nationwide Institution Data

In close relation to the HMDA data, the nationwide institution data is a series of fixed format flat files

Column of Interest	Data Type
as_of_year	String or Integer
respondent_id	String or Integer
agency_code	string or integer (string if NA is submitted)
loan_type	string
property_type	string or integer
loan_purpose	string or integer
owner_occupancy	string or integer
loan_amount_000s	integer unless NA
preapproval	integer or string
action_taken	string or integer
state_code	string or integer
county_code	string or integer
applicant_sex	string or integer
applicant_income_000s	string or integer

Table 3

of all lending institutions that reported HMDA data from 2007 to 2017. These files come in two different schemas, offering information regarding the lending institution, its parent institution and its top holder, including ID, name, and state.

4.3. Mapping Data

Data was acquired to validate the geographic distribution of HMDA data as well as provide enhanced visualizations within the output tool. The data used to map the geography of the loan data was taken from two source files for US County [3] and US State [4] geometry data. The geometry data was compiled by the US Census Bureau as a collection of TIGER/Line shapefiles and related database files. These files are extracts of selected geographic and cartographic information from the U.S. Census Bureau's Master Address File / Topologically Integrated Geographic Encoding and Referencing (MAF/TIGER) Database (MTDB).

5. Description of Analytic

The core principle of operation of this analysis and resulting application is the use of machine learning to create a supervised machine learning classification model to determine the probability of loan application approval. The user of the tool will enter their demographic information (e.g. US State, income, Race, etc.) which will be processed by the model to produce a probability of loan approval by historical year and lender. This model will effectively condense the 100+ million rows of HMDA and allow users to query acceptance rates for their particular cohort.

The input features for the model were chosen based on features both present in the HMDA dataset and considered relevant for loan application approval prediction. The following features are included in the model:

1) Applicant Race
2) Applicant Ethnicity
3) Applicant Gender
4) Applicant Income
5) Loan Amount
6) Respondent ID (Lender ID)
7) Year of Application
8) State of application

The classification outcome was a binary Approved or Denied status. The output used for the final visualization will be the models probability of approval, which is typically bucketed into a binary classification by some threshold value. The model was fit to the HMDA data based on these features to create a predictive model for use in the application.

Users of the application will enter their demographic and loan application information, which will be fed into the features described above. A pre-populated table of lenders and application years will be created and the cartesian product of lender, year, and user provided information will be input to the model. The model output will result in a probability of approval for each year and lender for the given demographics.

The table of loan approval probabilities will be presented to the user using visualization tools within a Flask web application. The user will then be able to repeat the submission process for different values to compare loan approval distributions.

6. Application Design

The design process used to generate the output loan bias visualization tool followed the data flow shown in Figure 1.

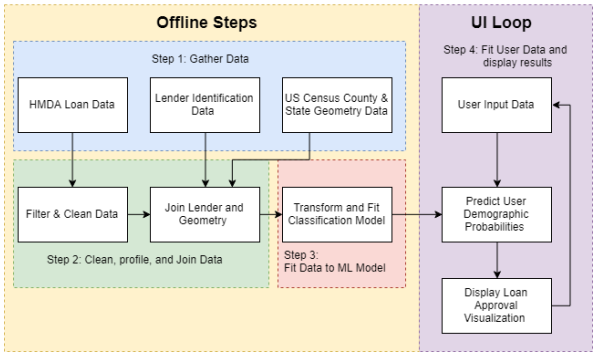

Figure 1: Design Data Flow Diagram

The design supports the following data flow steps:

1) HMDA data, lender identification data, and US Census geography data were downloaded and stored within the NYU Dumbo Hadoop cluster.
2) Spark was used to clean, profile, and filter the data sets. Each data set was joined to create a table containing all pertinent user demographic data.
3) HMDA features were selected, transformed, and fit to a logistic regression model. The model was saved to file for fast online access.
4) An application was created to support the loan bias visualization tool flow loop:
    a) User inputs demographic and loan application information
    b) Input information is transformed and the saved model is used to calculate a dataframe of loan approval probabilities
    c) The predicted data is shown to the user, who can input new information to iterate the loop

7. Actuation or Remediation

The end result of this project is a tool intended for use by potential loan applicants, equal housing advocates, and US financial regulatory policymakers to discern potential discriminatory patterns in lending and take appropriate actions. A potential loan applicant may enter their information and decide what lenders should be pursued or avoided for their demographic. Advocates and policymakers can iterate upon the results of the tool to determine the differences in lender behavior related to protected classes and take further regulatory action.

8. Analysis

8.1. HMDA Data

The HMDA data, the primary data source for this analysis, was uploaded to the NYU Dumbo Hadoop cluster and placed within HDFS. The data was extensively profiled in order to understand data quality, underlying distribution of some continuous variables, and understand each data fields values to determine which fields should be used in the machine learning model. The time period that this dataset covers is from 2007-2017.

HMDA data is collected by the Consumer Financial Protection Bureau (CFPB) and made available to the public on their website [6]. The data were ingested into NYUs scratch folder using a bash script of curl commands that downloaded the zip file and subsequently unzipped the file into scratch and subsequently loaded to HDFS. The HMDA dataset comes in two versions. The first version is a raw dataset that represents the respondents (lending institutions) categorical (i.e. textual information) as categorical variables and continuous variables as a string of integers. For example, the respondent represents the gender of the loan applicant as 1 for male and 2 for female. The state and county are represented by their FIPS code (Federal Information Processing Standard).

The second version is a complete textual representation of the loan applicants metadata. For example, gender is represented as either Male or Female and the state and county are represented by the abbreviation and county name.

Ultimately the size of the second version of the dataset is nearly twice as large as the raw version. While it is easier to deal with the full textual representation of the data, there was a considerable performance issue with ingesting and processing the dataset during the profiling stage. The data set is provided as a CSV file by the CFPB. However, in the textual representation of the HMDA data, some fields are long descriptions that include embedded commas. Therefore, the only way that we found to process that dataset in Spark was by using a SQL context and ingesting the data as a dataframe. However, with 150 GB of data, the data profiling process would require six hours to complete.

Due to the slow performance of data profiling the textual HMDA information the team decided to process the underlying raw dataset using the spark context resilient distributed dataset (RDD) abstraction. While there was a significant amount of coding required to filter out empty strings and NA representations, along with considerable work to translate the categorical variable to the textual representation, the data profiling portion of the project was much quicker by a factor of 4x.

We profiled the continuous variable, loan application amount for the 2013 dataset, and we found that the frequency of loan application amounts look somewhat normally distributed between $10,000 and $500,000. There is a very fat tail to the right (larger loan amounts). In deciding which data we should model, we decided to filter for loan applications between $50,000 and $500,000 in order to have a normally distributed continuous variable. Similarly we filtered for applicant income between $25,000 and $100,000 assuming there would be normal distribution of income frequencies in that range.

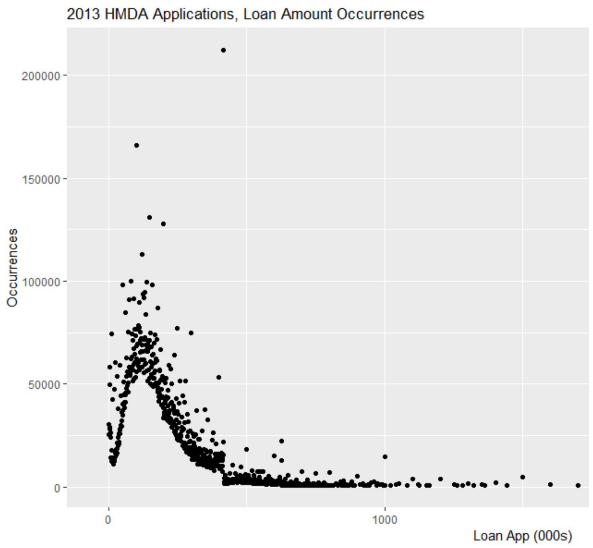

Figure 2: Loan App Frequency Distribution

8.2. Nationwide Institution Data

nder the Home Mortgage Disclosure Act (HMDA), financial institutions are required to report data about mortgages to the public, whose own information is in turn compiled into csv files and available from 2007 to 2017 by year. Compared to the millions of records in the HMDA data, the institution dataset is quite small, totaling no more than 30MB. Challenges, nevertheless, are still present from primarily these two aspects:

1) Data recorded before and after 2010 used formats that are dramatically different, namely about a third of non-trivial fields were not recorded at all before 2010.
2) We were hoping that Respondent ID could serve as the primary key to query into the dataset, in order to facilitate project actuation. However it turns out that not only are these IDs not unique, theres no pairwise bijection relationship within the set of Respondent ID, Respondent Name, Parent ID".

In response to these challenges, we tested several alternative logistics including:

1) Only accept institutions reported in 2017 into the database. One strong reasoning for it is if a previously reported institution isnt active anymore we shouldnt include it in our prediction.
2) It seems promising that the combination of Respondent ID and Respondent State will guarantee record uniqueness, which if proved will be beneficial to us because both these fields are included in our regression model.

The first alternative was rejected because the earlier the record, the more institutions are excluded, with proportion as high as 56% (see Table 2). Without further analyzing the distribution of HMDA records, such exclusion would be imprudent.

Year	Unique Respondent ID Count	Intersection w/2017	Unacceptance Rate
2017	5762	5762	0%
2016	6644	5496	17%
2015	6790	5347	21%
2014	6926	5159	26%
2013	7053	4954	30%
2012	7253	4824	33%
2011	7480	4677	37%
2010	7686	4222	45%
2009	7878	4045	49%
2008	8127	3855	53%
2007	8339	3687	56%

Table 4: Yearly counts of RespondentIDs appeared in the 2017 set and percentages not included in the intersection

With the unique combinations of Respondent ID and Respondent State, duplicate records are largely factored out. But collisions do happen at rare occasions. It turns out that including Agency Code solves the problem, as shown in Table 3. It can be argued that using the combination of these fields we preserve information embedded in the institution data to the largest extent, meanwhile guarantee the uniqueness of query results.

Year	Distinct Count Including All Fields	Distinct Combinations: RespondentID State	Distinct Combinations: RespondentID State AgencyCode
2017	5852	5851	5852
2010	7923	7919	7923
2007	8610	8603	8610

Table 5: Distinct records counts sampled when selecting specific fields

Supported by the above observations, we joined the yearly institution data on the set of columns: Respondent ID , Agency Code, Respondent State, and then with HMDA data on the same set of keys.

8.3. Mapping Data

In the initial design phase of this project, loan data was to be broken down by US county to allow users to analyze lender behavior on the neighborhood level. To enrich the interface supplied to the user, a visualization of the county map was to be displayed in a section of the UI. To create these displays, geometry data had to be collected for all counties and states.

The county and state geometries used for visualizations were contained in shapefiles created by the US Census Bureau and downloaded directly from the US government open data source Data.gov. These data files were downloaded as ZIP files on a local machine, passed to the NYU Dumbo Hadoop cluster via SCP, then uploaded to Hadoop HDFS.

Shapefile data are distributed over several different files, and require specialized tools to convert to Spark-readable text. The open-source Spark plugin GeoSpark contains a Scala method 'ShapefileReader', among other tools, that were used to translate the data into Spark DataFrames. A Scala script was used to convert the Shapefile data Spark DataFrames, which were then saved as JSON files where geometry data was stored as well-known-text (WKT) formatted strings. These JSON files could more easily be used clean and profile the data in the following steps.

Further Scala cleaning scripts imported the geometry JSON files, dropped unnecessary columns, and joined the county and state data on the common state-ID keys. The number of counties present in each US state was for analysis of county distribution over the data and to check for join errors. The number of records for each US Geographical region were calculated for analysis of regional county distribution. To better understand the geographical distribution of HMDA data, the combined state-county table was inner-joined to the HMDA data set by state and county name. The number of HMDA records per state and county were calculated and to a local machine. A Python script using the Plotly module was executed on the data to generate the choropleth of the number of HMDA records per county.

The results show that the overwhelming majority of counties have few records within the HMDA data set. Partitioning data by county would severely limit the accuracy of any lending model fit to such a small dataset, if the model could be fit at all. The decision was then made to eliminate partitions on the county level and partition instead by state. The state and county geometry data were retained for potential illustrative uses within the tool as future work, but would not be included in the final visualization for this project.

In the process of validating the HMDA to state and county join, an anti-join was created and profiled by a breakdown of county and state.The results show that approximately 2.5% of the data was not joined, due to either special characters within the county names or the state and county names being missing altogether. This amount of data was considered minor in the context of the greater data set and was dropped from further analysis.

8.4. Model Creation

The application returns to the loan applicant the lender with the loan applicant will have the highest probability of obtaining a loan given certain characteristics of a loan applicant.

$$\begin{aligned} P(y=k) \quad = \quad & \beta_0 loanAmt_{obs} + \beta_1 applicantIncome_{obs} \\ & + \delta_0 race + \delta_1 ethnicity + \delta_2 gender \\ & + \delta_3 lender + \delta_4 state + \delta_5 year \end{aligned}$$

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

After the above dataset were cleaned and filtered, a final dataset was created that joined the lender parent ID and name to the HMDA datas respondent ID field. This join created the final data frame that contains all the features used for the final model and UI. A variety of binary classification models were considered and evaluated. All models were taken from the Apache Spark MLLib library to support modeling on big data. The range of models considered were limited by the time constraints of this project, therefore model prediction accuracy will be less than could potentially be designed given more time to test a broader range of models.

The applicant gender, ethnicity, race, state, the application year, and lender were transformed into one-hot encoding vectors using the Apache Spark MLLib function OneHotEncoder. OneHotEncoder allows for the categorical variables to be treated as dummies in the analysis, which is useful for data analysis in the social sciences. The remaining numeric features were unchanged. The resulting sparse vectors were assembled using the Assembler and Pipeline transformers to create a transformation pipeline from the data source RDD to the final model to be fit.

Models were fit on a random sample of the HMDA data, as the full set of data can take on the order of days to fit a single model with cross-validation and tuning grid. All assessed models were fit on an 80% training set to 20% test set split, with an assessment criteria of AUC-ROC. After model assessment, the final model was trained on the full dataset, therefore the given AUC-ROC are conservatively bound. The following sections cover the binary classification models assessed with this process.

Logistic Regression - the Apache Spark MLLib binary logistic regression model was used with a parametric grid of tuning parameters fit with the CrossValidator function. The logistic regression was

fit to a linear combination of the input features, and higher-order combinations could not be explored due to time constraints. Tuning parameters included the elastic net parameter fit with values of 0.0, 0.5, and 1.0, and the regularization parameter with values of 0.0, 0.01, and 0.1. 3-fold cross validation was performed using the CrossValidator transformer included in the Pipeline. The final test set AUC-ROC was an abysmal 0.6, barely outperforming uniform random guessing. This low accuracy is attributed to the low-order linear combination of terms that possess high covariance.

Naive Bayes - model performed the best with an AUC-ROC of 79%. Based on already conducted research, the Naive Bayes algorithm is far less affected by sparse data compared to other commonly used machine learning algorithms. The features of our analysis included both continuous and categorical values as well as dummy variables for race and gender of a loan applicant. The underlying data in our analytic also exhibited some degree of covariance and the data was not evenly distributed across all the combinations of races, genders, and states that were analyzed. Therefore, we expect the feature matrix to be sparse, and the Naive Bayes algorithm performs better on sparse feature matrices [5].

Linear Support Vector Machine (L-SVM) - an L-SVM classifier was fit using the MLLib LinearSVC function. A single regularization parameter 0.1 was used as a previous profiling of the data indicated that the data is not linearly separable, and this model would likely perform poorly regardless of regularization. The AUC-ROC was 59%, confirming this assumption. Based on the results of the above models, the Naive Bayes model was chosen as the predictive model used in the loan assessment tool. It is noted that even though this model outperformed all other considered models, its accuracy is still below that of limits acceptable for mainstream predictive analytics. In the consideration of time and the overall uncertainty of the data, this model was considered to be adequate enough for this analysis. Future work on the subject must perform additional analysis of models to find a more accurate assessment of loan approval.

9. Conclusion

The conclusion goes here.

Acknowledgments

The authors would like to thank...

References

[1] S Frumkin. “HMDA Data: Identifying and Analyzing Outliers”. *Supervisory Insights, Winter 2007*. Federal Deposit Insurance Corporation.

[2] R. B. Avery, K. P. Brevoort, G. B. Canner. “Opportunities and Issues in Using HMDA Data”. *Journal of Real Estate Research, American Real Estate Society*. Vol. 29(4), pages 351-380. 2007.

[3] United States Census Bureau, Department of Commerce. TIGER/Line Shapefile, 2017, nation, U.S., Current County and Equivalent National Shapefile. Data.Gov. Updated June 2019. Retrieved July 2019.

[4] United States Census Bureau, Department of Commerce. TIGER/Line Shapefile, 2017, nation, U.S., Current State and Equivalent National. Data.Gov. Updated February 2019. Retrieved July 2019.

[5] Bissmark, Johan and Warnling, Oscar. The Sparse Data Problem Within Classification Algorithms: The Effect of Sparse Data on the Nave Bayes Algorithm. KTH, Datavetenskap, June 2017.

[6] HMDA: Consumer Financial Protection Bureau public website of HMDA data

[7] Darolia and Skanderson. Impact of the Mortgage Meltdown on the HMDA Data. Charles River Associates. January 2010.