Detecting bias against minorities in home loan applications

Team Name: Loan Sharks

Members: Ananta Iyengar (A20388360), Bhuvnesh Tejwani(A20444878), Omkar Pawar(A20448802), Virat Joshi (A20417850), Xiaoman Shen(A20449626)

Background

There have been a number of reports like the following:

- NPR Gap between white and black homeownership in Baltimore
- UrbanWire. What explains the gap between black and white young adults
- CNN. Racial bias cost
- Why does a homeownership gap exist between whites and minorities

The overarching theme remains the same. Minorities such as African Americans, Hispanics, Latinos, Asians, etc continue to struggle to obtain financing for purchasing/refinancing homes at a higher rate when compared with whites.

Homeownership lies at the heart of the American dream. In the US, wealth and financial stability are linked with homeownership. This is for good reasons. As per this article, homeownership is still financially better than renting. In general, owning a home for a reasonable period typically more than 10 years has historically outperformed the stock market and provided better returns for homeowners. As per the survey of consumer finances, the average homeowner has a net worth of 231K while the average renter has a net worth of 5K. Homeownership helps build communities by boosting investments in housing, construction, which in turn helps drive local economies. This, in turn, helps intangibles like reducing crime, etc.

In the US and Canada redlining is the term which basically means systematic denial of various services like loans and other forms of financing to specific communities on the basis of race for the most part. The term redlining specifically for financing means marking off certain zip codes/areas, etc as high risk. Banks and other financial lending institutions would make it extremely difficult for home loan borrowers to get loans for homes in these neighbourhoods.

The Federal fair housing act and the Community Reinvestment act were laws passed by Congress, the latter in 1977 to encourage banks and other institutions to reduce discriminatory practices against low income and minority neighbourhoods.

Problem Statement

The goal of this project is to analyze if there are discriminatory policies followed in home loan lending by banks and other financial institutions. We plan to analyze conventional home loans for this project. A conventional home loan is a loan which is not backed by FHA/VA, etc. We will build models for various zip codes which include African American and minority neighbourhoods and white neighbourhoods using data provided by the Consumer Finance Protection Bureau. These models will allow us to predict

the probability of an individual obtaining a conventional loan in a particular minority neighbourhood vs a similar loan application in white and other wealthier neighbourhoods.

Data Sources

We will be using the HMDA data provided by the Consumer Protection Financial Bureau for the years 2007, 2008, 2015, 2016, 2017 for one of the states or more from Illinois, Pennsylvania, Michigan, Florida or Iowa. The years 2007-2008 have been used since it was the time when the housing recession was at its peak and the trends during this display will help us perform a more accurate analysis. Data for 2007 and 2008 will be used for exploratory data analysis and comparison with the other periods.

Link to the dataset - https://www.consumerfinance.gov/data-research/hmda/historic-data/
The data here describes if the consumer got the mortgage—look for applications that were "originated"—or if the consumer was denied, didn't complete the application, or something else happened. For each record, it describes the loan, the property characteristics, the applicant demographics, and the lender.

We might also consider additional data sources such as the unemployment rates and bankruptcy rates from the Bureau of Labour Services and the Census datasets depending on the results of the exploratory analysis performed on the HMDA dataset.

BLS unemployment dataset - https://www.bls.gov/lau/#tables
Census dataset - https://www.census.gov/data/datasets/2010/dec/summary-file-1.html

Project Outline

1. Prepare HMDA dataset for a given state and perform exploratory Analysis for the same.
2. Analyse additional datasets based on our preliminary EDA.
3. Compare approval for different zip codes and different races which includes whites and other minorities.
4. Build prediction models which include explanatory variables like race, ethnicity, etc.
5. Verify if there is bias in the lending process. For the most part this would involve comparing predictions from the model for loan applications coming from minority groups with similar finances as white applicants.
6. Eliminate parameters inducing bias if any and build a model that functions without these parameters with similar performance.
7. Compare the results from two models and draw a conclusion.

Statistical Modeling Techniques

Logistic Regression: Create a model to classify a record as if it gets approved for a loan or not.
Decision Tree Classification: Even this algorithm will be used to classify the records based on the variables into two classes viz. Loan approved or not. We will compare these two models and deploy the model which has higher success metrics.

KPIs

Assumptions:

- Cost of False Positive: Medium (If the model predicts a borrower as safe when they aren't. This could technically result in a situation like the 2007 housing crisis. For now, we assume that this is a one-off)
- Cost of False Negative: High (If the model predicts a borrower as risky when they aren't. This results in a long term wealth gap between whites and minorities)

1. Recall(R): Number of true positives predicted/ total number of actual positives
2. Precision(P): Number of true positives predicted/ total number of positives predicted
3. F1 score: It can be interpreted as a weighted average of the Precision and Recall:
4. ROC Curve: Receiver Operating Characteristic to find the different cutoffs for the diagnostic test.

Success Metrics

Considering the above assumption, we need Recall to be high and Precision to be moderately high as well. Therefore, this setting would result in a high value of F1 score.
For example, in summary:

KPI	Threshold
Recall	0.90
Precision	0.80
F1	0.85

Deliverables

1. Exploratory Data Analysis and inferences
2. Logistic Regression Models predicting the probability of obtaining loans in various zip codes
3. Classification Models based on the available data to predict how likely a customer can get the loan approved.
4. Determine which parameters can be eliminated to give the applicant an equal chance as everyone else.
5. To come up with models which eliminate variables determined in above and build the model completely based on the financial situation of the applicant.