Dan Clayton

DSC-630 Predictive Analytics

**Final Project Milestone 3**

Note regarding milestone 2 vs milestone 3:

Milestone 3 is built on top of milestone 2, with changes from milestone 2 written in blue font to help differentiate old from new material.

Introduction

Discrimination has been a facet of money lending since before recorded history, and with good reason–if you lend money to someone you want to make as sure as possible that they will pay you back. While some discrimination in lending is not only permissible, but essential (you don’t want to lend to someone with a history of defaulting on loan repayments), other types of discrimination are non-sensical, hurtful, and since the passage of titles 8 and 9 of the Civil rights act of 1968 (AKA “The Fair Housing Act”) downright illegal. Of course, what I am referring to is racially motivated lending discrimination. If two people with equal incomes, debt levels, assets, and vocational prospects but different skin colors both apply for loans then both should have equal access to capital.

As one who works in finance, I find it difficult to believe that a company would fail to embrace borrowers of any skin color who proved creditworthy—effectively this would constitute companies refusing to do business (otherwise stated refusing to make money) because of the color of their potential client’s skin. This doesn’t make any sort of business sense. On the other hand, as one who grew up in the South, I can truthfully say that I have witnessed racism, both overtly, driven by nothing other than a person’s skin color, and in more subtle forms, targeting more the culture behind certain racial demographics. So I am willing to entertain the notion that perhaps racism continues to be a problem for borrowers today.

Racially motivated lending discrimination is supposedly still prevalent in the world today. If that is the case, then it means that businesses are not lending to qualified applicants, thus they are throwing away potential interest income. Social justice aside, it means that there is a business opportunity servicing clients of color who lack access to traditional loans because of their skin color but who are otherwise good lending candidates.

## Data Used

Home Mortgage Disclosure Act (HMDA) data (available [here](https://www.consumerfinance.gov/data-research/hmda/historic-data/)) will form the backbone of this analysis. To make this data set more manageable, only Illinois data from 2017 (the latest year available) will be analyzed. This smaller dataset still contains over a half a million records for mortgage applications, with 54 different data fields.

## Do I need to adjust the data and/or driving questions?

## Some additional calculations will need to be made using data that is yet a part of the data set. For example, property tax averages for each county could be used, along with average mortgage rates for 2017 to calculate the payments expected of each applicant. Looking at the expected payment as a percent of the applicant’s income could serve as a more meaningful measure of creditworthiness than simply comparing the applicant’s income to their perspective borrowing amount (particularly given that this data set is from Illinois, which has some of the highest property tax rates in the nation, making the tax information particularly relevant to their lending equation).

## If a useful proxy could be found for an applicant’s credit score, that would be very valuable in making this analysis more accurate. As of this writing, I have been unable to identify any such proxy that could be made from the available data. The Fair Isaac corporation guards FICO score data disclosures, making this analysis difficult.

## Research questions

1. Show lending information by race to see if there are differences in lending outcomes
2. Identify discriminatory lending examples where otherwise equivalent groups of borrowers receive different treatment
3. Try to size the business opportunity to sell loans to these otherwise neglected customers

## Will I be able to answer the questions I want to answer with the data I have?

1. Show lending information by race to see if there are differences in lending outcomes
   1. This question (restated, “is it possible to see differences in lending outcomes when viewed by race”) can be easily answered with the data provided.
2. Identify discriminatory lending examples where otherwise equivalent groups of borrowers receive different treatment
   1. This question is likely to present a challenge because the data that would justify a lending decision is only partially provided in this data set. For example, outstanding debt is not provided, and neither is the applicant’s credit score. Thus, trying to compare two candidates could be done based on their incomes, perspective mortgage amounts, and the location of the properties being purchased, which makes a complete comparison difficult.
3. Try to size the business opportunity to sell loans to these otherwise neglected customers
   1. This is an important question because it speaks to the underlying problem. While many might point to different lending outcomes for people of different races as proof that companies are being racist in how they extend credit, that argument doesn’t make economic sense. By not lending to people based on their race, a company would be systematically denying income to its shareholders—and any decent board of directors would not allow that to happen.
   2. Given the limited data set a subset of candidates who were denied loans and happen to adhere to certain ethnic groups will be identified, and the size of their loans summed. Net interest margin can then be calculated using today’s market rates (or those from 2017 to match the data set) which can be used to calculated the profit before tax that is being sacrificed by not lending to these individuals.

## Approach and Models Used

The goal is to establish if credit is being extended [or withheld] solely on the basis of race, because if it is, then a business opportunity exists within an under-served but valuable segment of potential customers.

This study assumes that systemic issues associated with race are out of scope. For example, it might be argued that credit scores are lower for black Americans than for white Americans, and while this may be true the reasons why are out of scope for this assignment. It is assumed that an objective measure of credit score, assets, liabilities, and earnings potential are all possible without regards to a person’s race.

Publicly available data on lending applications, approvals, and offers will be analyzed in an effort to identify discrimination.

Potential lenders from each of the racial groups need to be stratified by credit score, loan type, property type, loan purpose, occupancy type (if it will be a primary residence or investment property), and income, thus allowing us to make apples to apples comparisons across ethnic groups. If, for example, all Asian borrowers are compared to all Native American borrowers, there is likely to be a difference in the average income of the groups, which would have a very real impact on loan approvals. It is important to eliminate as much of this noise as possible before running a series of independent T-tests to determine if there are outcome differences for the comparable groups of different races.

Do I need to adjust my model/evaluation choices?

## It is entirely likely that my models will need to be adjusted as I start building them. Ultimately, I would like to see differences in lending outcomes that are associated only with the applicant’s race (Forgive the poor phrasing here—I don’t want to see this in real life, but this analysis seeks to identify any such occurrences). Clustering algorithms will likely proof valuable in this pursuit and should also build easy to interpret visuals that help to inform my conclusions.

## What visualizations are especially useful for explaining my data?

## Visualizations that show the population sizes of perspective borrowers in the data set grouped by race and lending outcome can be quickly generated and would help to answer the questions that are posed in this study.

## Clusters grouped by race and shown on an axis of percent of applications approved (approved applications / total applications) could also be a good visual to contrast the different lending outcomes by racial group.

## Risks

It is highly likely that the data set will not contain enough data to determine the full comparability of the borrowing applicants of different races. For example, the data set does not have applicant credit score information, not does it list the applicant’s outstanding debts, or monthly debt payment obligations, all of which would be relevant to making a lending decision. This could make it harder to truthfully compare applicants of different races. If no proxy can be found, then another analysis might focus on the likelihood of an applicant of any given race to be extended credit for a home in a predominantly minority neighborhood. This still fails to account for differences in credit scores and outstanding debt, but it might shed some light on the ability of the average applicant of any race to get a loan for a property in a minority neighborhood.

## Are my original expectations still reasonable?

I expect to be able to show differences in lending outcomes by race with my data. This correlation is well documented and should prove easy to observe. What will be harder to prove is that race was the causation motivating the lending decision (or even one of many causes). The underlying data used to make the lending decision is simply insufficient. If the federal government really wants to eliminate racism, sexism, or other biases or bigotries from lending decisions in this country, then the Fair Lending Act needs to be updated to require all of the data that is being used to make a lending decision. Given that the average loan today is being approved or denied based on stringent underwriting criteria in a computer system, and not by a banker, data is really the only way to prove that racism is still a factor in unfair lending—without that data, we are left to speculate.