Historical Redlining and Over Policing: A Working Paper

Carter Hanford
Independent Study
Saint Louis University - Fall 2020

I - Introduction

Established by the Federal Housing Administration in the early twentieth century, historical redlining segregated African American communities from white communities by denying them access to loans and lending needed to purchase homes. This form of legal segregation, or “de jure,” segregation, had long-lasting impacts on the communities it targeted even after the passing of the Fair House Act of 1968 which banned all forms of discrimination against race in the housing market. Although redlining practices seemingly ended in 1968, illegal segregation still took place in many of the communities affected by historical racial discrimination, and the effects of this discrimination in the housing market persisted in the form of “de facto segregation,” or willful segregation by local neighbourhoods and communities (Johnson 2020). In current social science research, an abundance of literature exists examining historical redlining and identifying the consequences it has on minority communities, indicating parallels between deteriorating and run-down areas of legacy cities and redlining maps from almost a century ago (Aaronson, Hartley, & Mazumder 2017, Berkovec et. al 1994, Tighe & Ganning 2015, Zenou & Bouccard 2000).

The result of a systemic denial of resources to African American communities due to redlining are communities that lack opportunities for families to experience upward mobility. Similarly, racially segregated neighbourhoods and communities with concentrated poverty experience higher levels of crime, both violent and non-violent crime, due to the neighbourhood’s access to resources and the ability to solve problems that foster crime (Hipp 2007, Kuhl, Krivo, & Peterson 2009). Thus, the fostering of crime in these communities perpetuates a false perception by police that African Americans are inherently more violent in nature, resulting in racial prejudice and the mismatch between feeling under-protected as victims and over-policed as suspects, regardless of geographical location (Palmer 2012). The racial prejudice towards minority communities in police forces manifests in the number of overall arrests for crime. According to the United States Bureau of Justice Statistics, in 2018 34% of the male prison population was black, whereas only 29% were white males even though the black population is only 13.4%. This is a direct result of the racial prejudice and over-policing that takes place in African American communities, especially those that experience the long-lasting effects of historical redlining.

The racial history of St. Louis plays a substantial role in the makeup of the city today. The effects of legal segregation and decades of predatory redlining practices disproportionately affected the black communities that make up North St. Louis and enclaves of South St. Louis. A majority of the black population in St. Louis are clustered into census tracts north of Delmar Boulevard, the infamous "Delmar Divide." The idea of this project and working paper is to examine the relationship between redlining and policing in these St. Louis communities. As a working paper, this project functions as an exploration of a few different datasets and techniques to examine a potential relationship between crime data in St. Louis and financial data on mortgage lending practices from financial institutions. The principal objective of this study is to eventually build a dataset for St. Louis that includes crime, race, and financial information that can be used for a machine learning classification analysis in order to predict the relationship between redlining and policing in St. Louis.

II(a) - Methodology for HMDA Data

The first of two methods of analysis in this project deals with the examination of financial data from the Home Mortgage Disclosure Act (HMDA), enacted by Congress in 1975 which requires financial institutions to disclose individual-level information about loan applications. This analysis uses 2017 data from the HMDA dataset to examine the relationship between an individual's race and their chances of defaulting on a home loan or having their loan application accepted by the final institution. By highlighting racial bias by financial institutions in the home mortgage market, one can theorize that this is a manifestation of the long-lasting effects of historical redlining.

The first analysis uses this data to build a logistic regression model using a machine learning algorithm from sklearn, or sci-kit learn, a machine learning library for python, which attempts to predict a loan acceptance or denial based on a set of chosen features from the dataset. The HMDA data structure consists of a column, or feature, representing character or numeric data, and its corresponding feature with one-hot encoding. Since all available HMDA data includes encoded columns, we eliminate the need for a label encoder step. However, for organization purposes, a label encoder from sklearn is still used to keep track of the features. The dataset comes with 25 features to choose from, many of which are unnecessary for this analysis, so we chose variables that could potentially influence the decision a financial institution has to make on the basis of accepting or rejecting a loan. This slims the dataset down to one label and four features in which the label is a binary indicator that represents class one, a financial institution accepting a loan application, and class zero, a financial institution denying a loan application.

There were some methodological choices that needed to be made when dealing with the dependent variable. In the original dataset, the label had seven classes, some of which are not abundantly understandable from the documentation. For simplicity, the action taken by the financial institution, the outcome being measured, was converted to a binary variable. The reasoning behind this decision is that some of the actions taken by the financial institution are unclear and do not seem to be easily predictable by the data. For example, one of the classes involves the application being withdrawn by the applicant. There is no feature in the dataset that can predict why the application was withdrawn, so this action, as well as others, were excluded from the model. The features used to predict this label are as follows; property type, loan purpose type, the race of the applicant, and the sex of the applicant.

Another important methodological choice needed to be made regarding the races included in the applicant race feature in the dataset, which involves some context with a much larger research project. The 2017 HMDA data used in this project is part of a larger research project involving a professor in the Department of Sociology & Anthropology at Saint Louis University, in which the focus is on redlining and its effects on African American communities in St. Louis City. Historical racial discrimination in St. Louis has been primarily an issue of black and white, thus this project only includes black and white applicants even though there are many more listed in the dataset.

The methodology for the logistic regression model is two-fold; the first step is training the model on the training dataset to evaluate performance and the second step is to implement logistic regression feature importance to examine which features are influencing the model's performance. Feature importance is implemented on the logistic regression model to examine the coefficient properties for each input feature. A positive coefficient score indicates that the input feature predicts class one, whereas a negative score indicates that the input feature predicts class zero. This helps us examine which of the selected features has a higher impact on the prediction of the algorithm. In this case, since the model attempts to predict a yes or no decision from the financial institution, feature importance helps indicate which features influence the prediction of that decision.

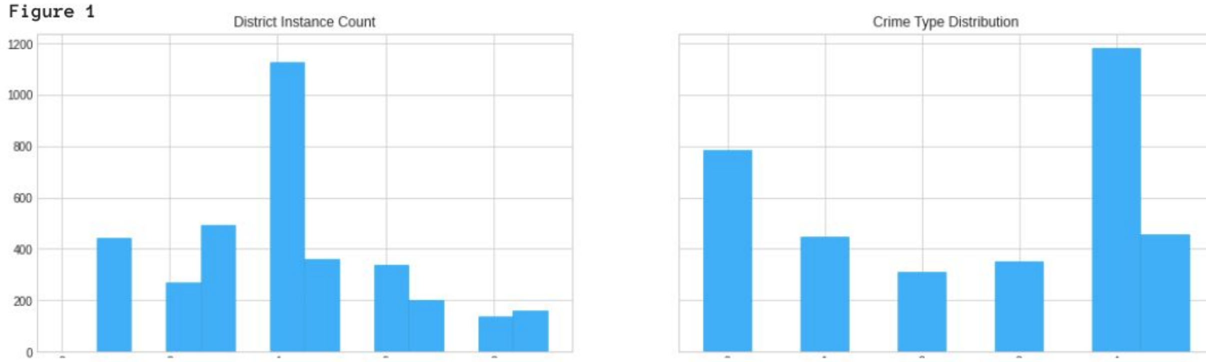
II(b) - Methodology for STLMPD Data

As mentioned in the introduction, the St. Louis Metropolitan Police Department's database of crime for St. Louis City was accessed by the R package, Compstatr, developed by Christopher Prener, PhD., Cree Foeller, & Taylor Braswell. The STLMPD dataset has many variables to choose from, but most do not have relevance for this analysis. The cleaning and organizing of this data was accomplished in an R notebook, resulting in a dataset from 2010-2018 with six variables; the year of the crime, the crime code, the district the crime took place in, the neighbourhood the crime took place in, the address of the crime, and the description of the crime. While this dataset includes all types of crime, the dataset only includes crime that is considered "non-violent" or "non-threatening." Thus, the goal of this dataset is to examine the relationship between the type of non-violent crime and where it took place in the city, looking for patterns of over-policing.

Once the STLMPD data were collected and cleaned in R, the analysis moved back to python for visualization and further examination. The purpose of the exploratory analysis in python was to indicate whether or not the STLMPD crime data would be suitable for a classification analysis using machine learning. It is important to note that this step of the analysis is purely exploratory. The raw crime data from the police department does not have many features suitable for a proper machine learning classification analysis. For example, one of the columns is the address in the city where the crime took place. While this feature could be useful for a GIS visualization on a map, it is useless for a classification model because every address is a unique observation, making the model as complex as every row which leads to an insufficient model. Thus, when examining the STLMPD, the goal was to search for variables that have a "class," to them, or in other words, variables that can be used to predict an outcome.

The first step of the exploration in python was to visualize the data first. Figure one examines the relationship between the two prominent variables highlighted as potential indicators for a classification analysis, the district the crime took place in, and the description of the crime. On the left histogram, the highest instances of crime took place in district four which is located in North St. Louis City. This is a solid first step because it indicates that there is a clustering of crime taking place in areas of the city that have historical ties to redlining practices. The histogram on the right examines the distribution of non-violent crime types in the dataset. The highest instances come from encoding four, which represents “Public Order or Unspecified Public Order Violation,” while the second-highest instances come from encoding zero, which represents “Disorderly Conduct.” These two non-violent crime types are important because they could be vague enough to indicate over-policing if the correct data is available to predict this. It is not all that unreasonable to suggest that police may entrapment to instigate casual interactions into serious altercations that lead to “disorderly conduct,” and this idea comes to fruition in policies such as “stop and frisk” enacted by many police departments (Meares 2014).

Figure 1



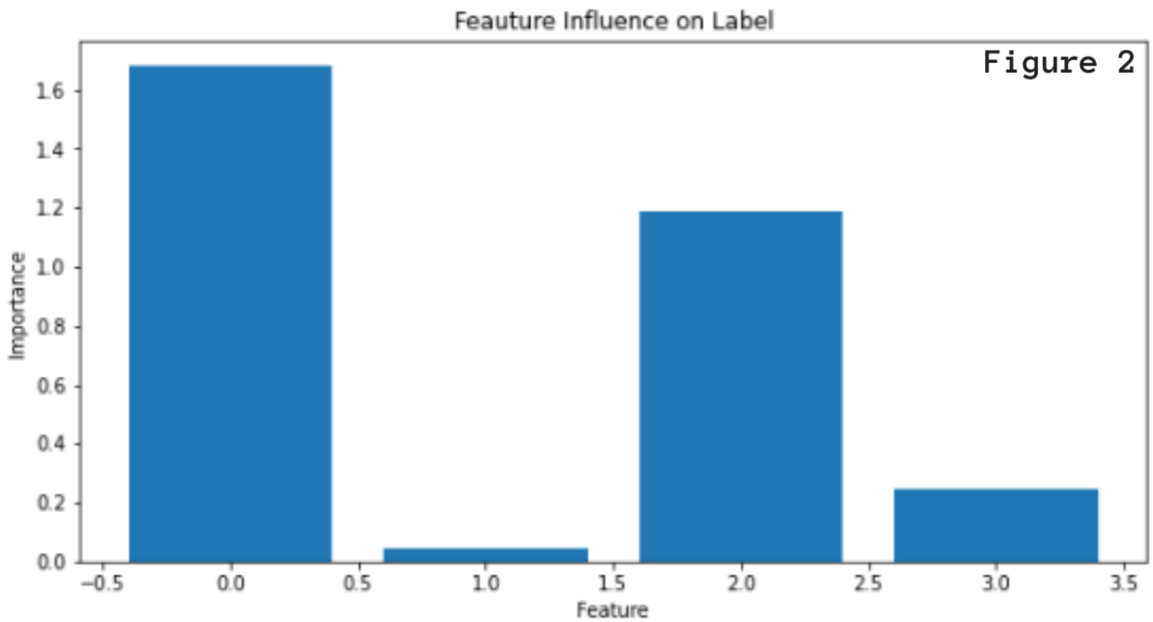
The second step in the STLMPD analysis is to explore whether a classification algorithm could potentially be utilized to predict a relationship between redlining and over-policing in St. Louis. As a result of the lack of predictability in the crime data mentioned previously, the only feature suitable for predicting is the district which the crime took place in. In theory, this feature could be used as a geographic location predictor that could indicate over-policing in certain areas of the city. Secondly, the label that we want to predict is the type of crime committed. Thus, the dataset is only one-dimensional, with one feature and one label. It is important to note that a model with this low of complexity will not achieve high classification accuracy on the training or testing dataset, but instead, this analysis serves as an exploratory tool to indicate whether or not this crime data could be utilized for a machine learning workflow.

The two classification methods tested are two of the simplest classifiers in machine learning, a decision tree and a random forest. A decision tree works like a flowchart that stops at each feature node, each branch represents an outcome, and each leaf node represents a label. In this case, there is only one branch since our data has only one feature, but there are six leaf nodes since there are six labels. Random forests are ensemble learning methods for classification that work by essentially generating multiple decision trees during training and outputting the class that is the mode of the classes. Both of these models are simple enough to use as exploration models to examine the St. Louis crime data.

III - Results

The logistic regression model tested on the HMDA data predicted the labels of the training data with an accuracy of 82% and predicted the labels of the testing set with an accuracy of 82% as well. While the results from the accuracy score seem in favor of a good model, many different performance metrics were used to evaluate the model's overall performance. A classification report indicates an f1 score of 0.74 which suggests that the model performs better than average. These performance measures seem to indicate that the model is performing well at classifying the labels correctly.

Checking feature importance from the HMDA data model indicates that two features from the dataset account for most of the variance in the label predictions, property type and applicant race. A positive feature analysis score indicates that the feature influences class one, which in this case is the financial institution accepting the loan. Both property type and applicant race have positive scores, suggesting that these features have a major influence on the financial institution accepting the loan from the applicant. Interestingly, property type has a higher score than applicant race, suggesting that the property type an applicant is requesting a loan for has a significant influence on whether or not the financial institution accepts the loan application. Recall that the features in the analysis are as follows; property type, loan purpose, applicant race, and application sex. The feature importance measurement indicating that an applicant's race significantly affects their chances of their loan application being accepted or denied is worth noting, since the other two features, property type and loan purpose, are financial parameters that a financial institution should consider when evaluating loan applications. An applicant’s race, one could argue, is not a proper measure for evaluating loan applications because it introduces a racial bias in the home-loan market. Figure two visualizes the feature importance measurement, with the first bar representing property type, the second bar representing loan type, the third bar representing applicant race, and the fourth bar representing applicant sex.



The classification exploration results from the STLMPD crime data seem to indicate that this dataset could be used for a proper classification analysis if the correct features were available to use for prediction. In a one-dimensional model with one feature and one label, both the decision tree and random forest classifiers were able to predict the crime based on the district 34% of the time. At face value, this would indicate that the model does not perform well, however, recall that these models were simply used to explore whether the crime data could be used for a classification analysis or not. With more relevant features built into the crime dataset, both these simple classification models could potentially achieve a high classification accuracy and be successful at highlighting a potential relationship between redlining and over-policing.

IV - Discussion

Overall, this analysis serves as a significant first step in highlighting the relationship between the historical effects of redlining and over-policing in St. Louis City. The logistic regression model performed on the HMDA financial data indicated a significant racial bias in the feature importance measurement in which the race of an applicant has a substantial influence on the financial institution's decision to deny or accept a loan application. However, more specific data and analysis would be needed to solidify the assertion that this disproportionately impacts African Americans within the dataset. Based on current social science research, it can be implied that this racial bias present in the HMDA data has negative consequences on black applicants (see Pager, Bonikowski, & Western 2009), but it cannot be directly proven based on the analysis itself.

Similarly, the exploration of the STLMPD crime data seems to indicate that it may be a potential dataset to perform classification analysis on if the dataset could be expanded to add additional features. Mentioned previously, building the STLMPD data into a new dataset with features corresponding to race and financial status indicators potentially from the HMDA data could provide for a rich and unique analysis to examine the relationship between redlining and crime in St. Louis City. The addition of more predictable class features to increase model complexity could benefit the classification model and increase the accuracy to a threshold where the analysis becomes meaningful and novel in nature. Many research papers and analyses exist documenting the effects of historical redlining on minority communities as well as the effects of racial prejudices and over-policing in police forces across the United States. However, there does not seem to be the many instances of social science research linking these interconnected phenomena.

V - Conclusion

Further analysis for this working paper includes building on the STLMPD crime dataset with features that could be used to predict crime outcomes in St. Louis, while also examining how the HMDA data could be utilized as well. The exploratory nature of this analysis produced a pathway for a potential novel and unique research project that describes the relationship between historical redlining and over-policing in St. Louis, one of the cities that has experienced these issues the most in history and in present-day. The logistic regression model mapped to the data from the Home Mortgage Disclosure Act produced a result that signified racial bias within financial institutions and how they make decisions on loan applications. Feature importance showed that the race of an applicant significantly influenced whether or not a financial institution accepted or denied a loan application. Crime data from the St. Louis Metropolitan Police Department indicated a cluster of non-violent crime taking place in North St. Louis City, an area of the city affected by historical redlining and segregation. And finally, an exploration of the dataset using simple classification methods indicated that this dataset has the potential to use machine learning algorithms to predict crime in St. Louis. Overall, this working paper provided a solid basis for a novel and unique social science research project about redlining and over-policing in St. Louis.

Works Cited

Aaronson, D., Hartley, D., and B. Mazumder. 2017. “The Effects of the 1930s HOLC Redlining Maps.” *EconStor* 12.

Berkovec, J. A. et al. 1994. “Race, Redlining, and Residential Mortgage Loan Performance.” *Journal of Real Estate Finance and Economics* 9: 263-294.

Hipp, J. R. 2007. “Income Inequality, Race, and Place: Does the Distribution of Race and Class within Neighborhoods Affect Crime Rates?” *Criminology* 45(3): 665-97.

Johnson, Walter. 2020. *The Broken Heart of America: St. Louis and the Violent History of The United States.* New York, NY: Basic Books.

Kuhl, D. C., Krivo, L. J., and R. D. Peterson. 2009. “Segregation, Racial Structure, and Neighborhood Violent Crime.” *American Journal of Sociology* 114(6): 1765–802.

Pager, D., Bonikowski, B., & B. Western. 2009. “Discrimination in a Low-Wage Labor Market: A Field Experiment.” *American Sociological Review* 74(5).

Pager, D., & H. Shepherd. 2008. “The Sociology of Discrimination: Racial Discrimination in Employment, Housing, Credit, and Consumer Markets.” *Annual Review of Sociology* 34: 181-209.

Meares, T. L. 2014. “The Law and Social Science of Stop and Frisk.” *Annual Review of Law and Social Science* 10: 335-352.

Palmer, S. 2012. “‘Dutty Babylon’: policing Black communities and the Politics of Resistance.” *Criminal Justice Matters* 87(1): 26-27.

Tighe, J. R., & J. P. Ganning. 2015. “The Divergent City: Unequal and Uneven Development in St. Louis.” *Urban Geography* 36(5): 654-673.

Zenou, Y., & N. Boccard. 2000. “Racial Discrimination and Redlining in Cities.” *Journal of*

Urban Economics 48(2): 260-285.