Assessment of Minorities Access to Finance

Emi N. Harry and Dr. Gary Weiss
Department of Computer and Information Science
Fordham University
113 W 60th Street,
Manhattan, New York 10023, USA
{eharry & gaweiss} @fordham.edu

***Abstract* – Access to finance is a fundamental aspect of an individual's social and economic growth. Despite the prohibition of discriminatory lending by financial institutions to minorities, there are still claims that such practices are still prevalent in the American society. By applying data mining techniques (exploratory and predictive analytics) to the Home Mortgage Disclosure Act data spanning a decade (from 2007 to 2016), insight is gained on the major determinants for mortgage approval or denial and how it has affected minorities over the past decade.**

Index Terms – Visualization, Machine Learning, Neural Network, Decision Trees, Data Mining.

I. INTRODUCTION

"The American Dream" is a phrase that transcends time in the American society; giving the average man and woman a goal to work towards. For some it still remains the idea of having that house that you can finally call your home, with a "white picket fence", and for others, the idea has evolved. Whatever the attainable, encapsulated in that dream, the constant themes are the need for socio-economic stability, success and freedom. Yet they can only be achieved through sufficient and unbiased **Access to Finance**.
In recent times, the following questions have become prevalent in public discussion forums;

1. If things have changed in the United States, why are there still few minority owned business?
2. Why do fewer minorities the homes they live in?
3. Why is there a stagnant growth of minorities in the STEM fields and college as a whole?
4. Why are there few minority owned financial institutions (Commercial Banks, Community Banks, and Micro-Lenders).
5. Why do minorities get higher interest rates on loans?

The list goes on. Yet to understand the issues with minorities in the United States, you cannot do so without talking about the issues African American have faced and continue to face in the United States. Hence, most of the references of the term "minorities" in this paper will be mostly referring to African-Americans.

Due to the lack of public access to good credit loan and student loan data, and Small Business data indicating the race of the business owner, the primary data being used in this research is the Home Mortgage Disclosure Act data, downloaded from the **Consumer Financial Protection Bureau** (**CFPB**); an agency of the United States government which was established in July 2011, with the goal of putting the consumer's needs first, they serve as a shield, protecting them from the financial industry.
The Home Mortgage Disclosure Act (HMDA) is a federal act was approved by congress in 1975. It requires mortgage lenders to keep records of vital pieces of information regarding their lending practices (in other words, information on income, family size, the loan amount, purpose of the loan, the race of the applicant, marital status etc), which they must submit to regulatory authorities. Furthermore, a publication by Investopedia also states that;

"Regulation C is an important component of the Home Mortgage Disclosure Act. It was created by the Federal Reserve to detail and explain the requirements of the Act, as well as to assign specific additional conditions that all established financial institutions (Banks) must follow. In general, the primary purposes of the Home Mortgage Disclosure Act and Regulation C, are to monitor the geographic targets of mortgage lenders, provide an identification mechanism for any predatory lending practices and to provide reporting statistics on the mortgage market to the government. The Home Mortgage Disclosure Act helps to support the community investment initiatives sponsored by government programs. In 2017, 6,762 lenders were required to report HMDA statistics with 16.3 million loan records reported." [1]

Features currently reported in the HMDA data includes: action taken; lien status; applicant ethnicity; applicant race; co-applicant ethnicity; co-applicant race; state; loan amount; applicant income; etc. but does not contain vital features like credit score, loan rate, payment spread, and income-to-debt-ratio, which are the type of features that can truly aid in the detection of discriminatory lending. In 2017, the new features that were approved for inclusion in the HMDA reporting includes: age of borrower; application channel; mortgage loan originator NMLS identification; credit score; combined loan-to-value (CLTV) ratio; borrower's debt-to-income (DTI) ratio; borrower-paid origination charges; points and fees; discount points; lender credits; loan term; prepayment penalties; non-amortizing loan features; interest rate; and rate spread for all loans. These new features will be seen in the 2018 HMDA data report that will be published in 2019.
Despite the approval of the Home Mortgage Disclosure Act over four decades ago, and the Consumer Financial

Protection Bureau, **"Redlining"** is still pervasive in the American society. A major part of this, is the data fields (such as loan interest rate, credit score, lender credits, loan term, prepayment penalties etc) being required as clear indicators of this illegal lending practice were not being disclosed. Thus, making it very difficult for external parties to identify, which in turn makes it nearly impossible to hold these institutions accountable for their actions. An article on CNN Money by Tami Luhby, which was published in 2014 stated that the average black household had amassed less than one-tenth of the wealth of a typical white one. It also predicted and warned that the financial gap was getting worse. In the article, Tami further stated that in a research which had been carried out by Brandeis University, they discovered that the financial/wealth gap between blacks and whites had nearly tripled as seen in figure 1 below. [2]

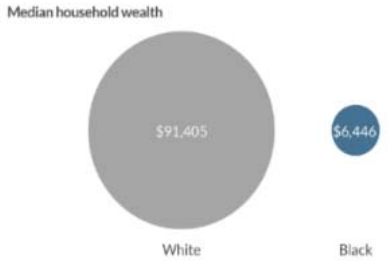

Fig 1. Showing the median wealth distribution by the white and black in the US as at 2014.
Image gotten from money.cnn.com

Luhby added that the greatest factor in this wealth gap is due to the fact that home ownership among blacks is so much lower as seen in figure 2 below. With the explanation that in America, housing is often the greatest asset and a major component of overall wealth. [2]

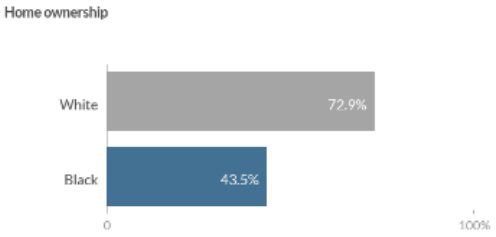

Fig 2. Showing the percentage distribution of home ownership by the white and black race in the US as at 2014.
Image gotten from money.cnn.com

Several publications including the NAACP 2009 report, have cited the following as prevalent consequences of the discriminatory lending practices that have impeded the accumulation of wealth in African American communities:

1. Disparities in lending were "particularly worrisome for African Americans" with respect to very high-cost loans covered by the Home Ownership and Equity Protection Act (HOEPA). In 2005, African Americans were the only racial group to receive a substantially higher percentage of very high-cost loans than market-rate loans. [3, 4]
2. In 2007, African Americans paid an average of 128 basis points more for loans than did their white counterparts; and in the subprime market, the difference was 275 basis points more than their white counterparts. [3, 5]
3. Even when income and credit risk are equal, African Americans are up to 34 percent more likely to receive higher-rate and subprime loans with a prepayment penalty than are their similarly situated white counterparts. [3, 6]
4. African Americans are 15 to 16 percent more likely to receive a higher-rate ARM purchase loan than if they were white. [3, 7]
5. Lending discrimination placed at least one million African Americans and other people of colour at great risk of loss of wealth; an estimated loss of at least $164 billion. [3, 8]

Hypothesis

In this research, the hypothesis is that race plays a huge role in the decision-making process of home mortgage approvals or denials, and it has not improved since the last major report on it in 2014. While a lot of financial institutions now claim to be racially blind in the decision-making process, there have been numerous publications which claim otherwise. Hence the need for this experiment.

The intent of this research is to see how accurately the decision of a loan application can be predicted, taking race into consideration, in order to determine if the hypothesis is true. Two tests were carried out; the first test was to check if the outcome of the application would have been different if race was not used as a predictive feature, by creating two models (one with racial features and the other without racial features) and comparing their results. The second test was to predict the race of applicants based on the application decision and other features, excluding highly correlated racial features like applicant and co-applicant ethnicity. Bearing in mind that with vital features like credit score and borrowers debt-to-income-ratio missing from the existing data, it might be difficult to detect discriminatory lending.

II. BACKGROUND

The term "Redlining" has been used over the years to describe a perpetual form of discrimination; where financial institutions refused to provide their services to neighbourhoods they termed as "African-American" neighbourhoods. According to a publication by Brent Gaspaire in BlackPast.org, the origin of the term stems from the policies developed by the Home Owners Loan Corporation (HOLC) created in 1933 by the Franklin Roosevelt Administration to reduce home foreclosures during the Depression and then institutionalized by the 1937 U.S. Housing Act which established the Federal Housing Association (FHA). Federal housing agencies including the HOLC and the FHA determined whether areas were deemed unfit for investment by banks, insurance companies, savings and loan associations, and other financial services companies. The areas were physically demarcated with red shading on a map. In contrast, zones which were to receive preferential lending status were marked in green shading and intermediate areas in blue shading. Often these decisions were arbitrarily

based on the area's racial composition rather than income levels. [9]

Despite the abolishment of "redlining" in the 1968 Civil Acts law that was passed, it still continued. Neighbourhoods that local/community banks deemed unfit for investment were left underdeveloped or in disrepair. [9] The ripple effects are very much evident in modern day America.

III. METHODOLOGY

A. *Data Collection and Preprocessing*

The 2007 to 2016 data which is about 100 gigabytes was downloaded in two parts from the Consumer Financial Protection Bureau. Through the CFPB API, the public has the option of downloading the data. As such, the data came with labels as well. Hence, because of the size of the data, only the first five thousand rows were read in initially. The only actions relevant to this study are: 1-- Loan originated, which is a way of saying that the loan was approved by the institution and accepted by the applicant; 2 -- Application approved but not accepted; 3 -- Application denied by financial institution. The application approved but not accepted, represented by code 2, is replaced to code 1, thus ensuring that all loan approvals have the same identifier. While the application denied by financial institution, represented by code 3 is replaced with code 0.

The processed data was then exported for the exploratory visualization in Tableau.

B. *The Story Within the Data*

The processed data was then exported for the exploratory visualization in Tableau. Since the recession occurred in 2009, the primary focus of the exploratory analysis will be on the data from 2012 to 2016.

Question 1: What is the overall ratio of loan approvals to denials from 2012 to 2016?

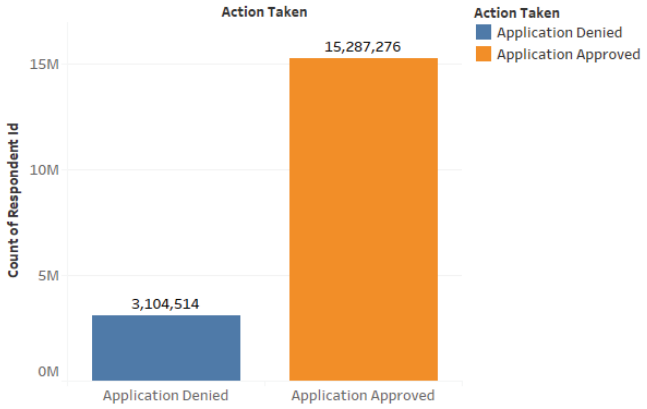

Fig 3. Showing the total number of home mortgage approvals and denials from 2012 to 2016.

From figure 3, it is clear that the ratio of approvals to denials is approximately a 5 to 1 ratio.

Question 2: What is the trend in approvals and denials by race from 2012 to 2016?

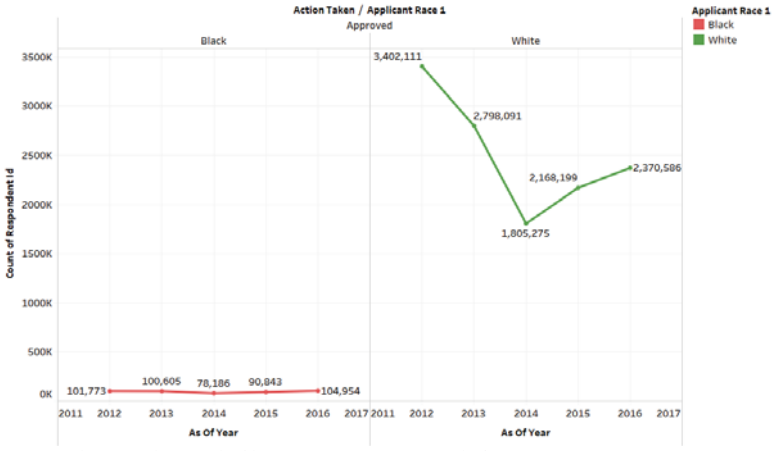

Fig 4. Showing the trend of home mortgage approvals from 2012 to 2016 by race.

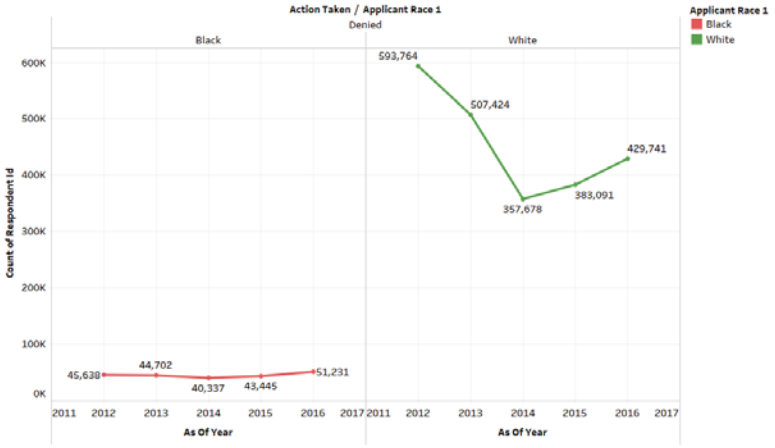

Fig 5. Showing the trend of home mortgage denials from 2012 to 2016 by race.

From figure 4 and 5, approximately 85% of all white applicants get approved for a home loan, which is better than all black applicants from which only approximately 68% of the applicants get approved.

C. *Building the Prediction Models*

- *Test 1 – with Action Take as Class*

The data set is divided into two parts, based on the action taken variable. The first being a subset of all the approved applications, and the second a subset of all the denied applications. These two subsets are broken down further by race and year. For the subset of approved applications, fourteen thousand records are randomly selected for each combination of application decision, year, and race. In like manner, for the subset of denied applications, fourteen thousand records are selected on the combination of application decision, year, and race. All the randomly selected subsets are then merged into one data frame. The goal was to have a balanced class in the data set on which all models would be trained and tested. The models selected were: C5.0 and NNET. The method of train control to avoid overfitting of all the models is cross validation, set to repeat three times. The second subset for the training and testing data is partitioned eighty percent to twenty percent respectively.

C5.0 – In both models, '1 – application approved" is used as the negative class while "0 – application denied is used as the positive class. The ranking of the features from the most important to the least important, for the model with racial features and the model without racial features can be seen in figures 6 and 7 respectively.

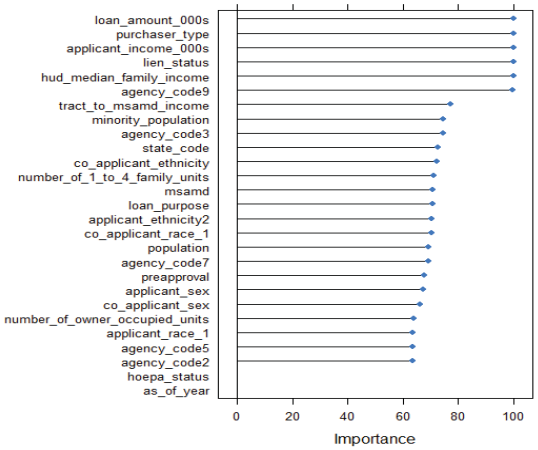

Fig 6. Showing a plot of the features in the order of their importance to the C5.0 model with racial features present.

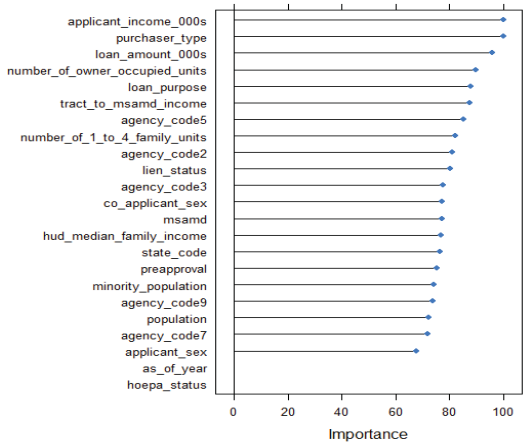

Fig 7. Showing a plot of the features in the order of their importance to the C5.0 model without racial features present.

NNET – The ranking of the features from the most important to the least important, for the model with racial features and the model without racial features can be seen in figures 8 and 9 respectively.

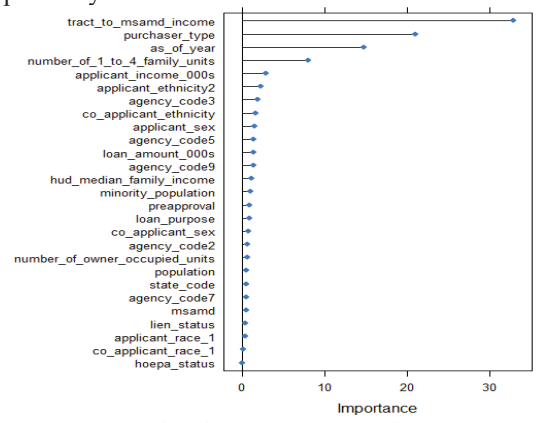

Fig 8. Showing a plot of the features in the order of their importance to the NNET -Neural Network model with racial features present.

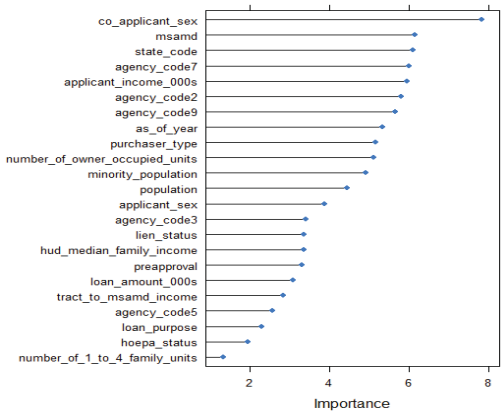

Fig 9. Showing a plot of the features in the order of their importance to the NNET -Neural Network model without racial features present.

- *Test 2 – with Race as Class*

Similar to the previous process, the data set is divided into two parts, based on the *"applicant_race_1"* variable. The first being a subset of all the approved applications, and the second a subset of all the denied applications. These two subsets are broken down further by sex (gender) and year. For the subset of approved applications, fourteen thousand records are randomly selected for each combination of year and sex (gender).

Figure 10 below shows the ranking of the features from the most important to the least important in the C5.0 model.

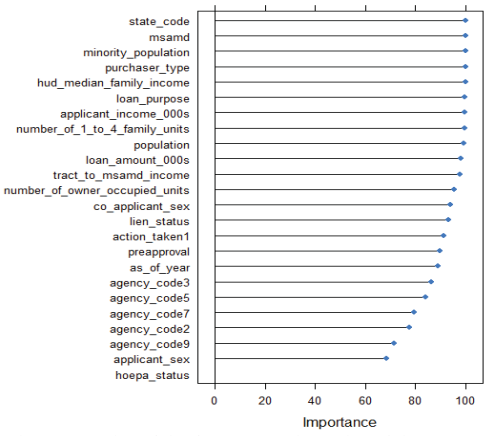

Fig 10. Showing a plot of the features in the order of their importance to the C5.0 model with race as the class.

Figure 11 is an image of a part of the decision tree for the prediction of race.

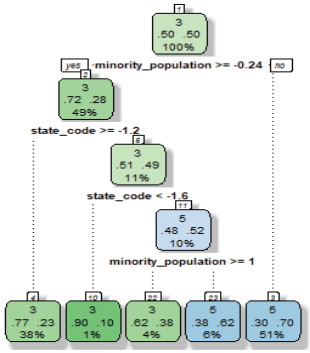

Fig 11. A partial plot the C5.0 decision tree with race as the class

Figure 12 below shows the ranking of the features from the most important to the least important in the NNET model.

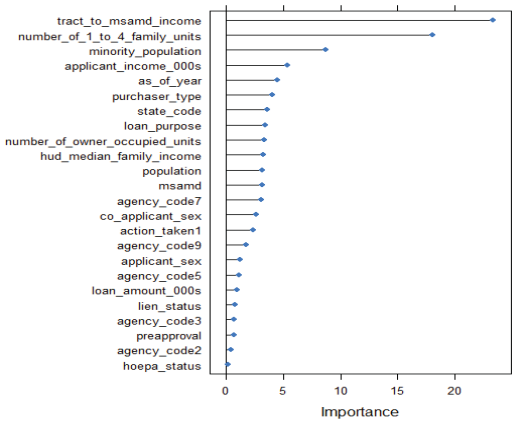

Fig 12. Showing a plot of the features in the order of their importance to the neural network model with race as the class.

While the feature ranking in the various C5.0 models are quite similar, it can be observed that the neural network rankings are diverse.

IV. RESULTS

Results of Predictions with "Decision" as Class

The confusion matrices and statistical measures of the C5.0 model that had racial features, with the "*action_taken*" as the class are as follows:

TABLE 1
METRICS OF C5.0 MODEL WITH RACIAL FEATURES

Accuracy : 0.8776
Sensitivity : 0.9731
Specificity : 0.7821

In table 1 above, the accuracy of the racially biased C5.0 model in predicting the action taken is approximately 88% which is relatively high. It shows that the model is pretty good at predicting the present outcome of mortgage applications based on the existing data. The sensitivity or recall of the model, is approximately 97% which will be very good if the existing lending system was not purported to be biased. While the specificity, which is the rate at which the model accurately identifies applicants that will be approved is approximately 78%.

TABLE 2
CONFUSION MATRIX OF C5.0 MODEL WITH RACIAL FEARURES FOR BLACK APPLICANTS

	Denied	Approved	Total Predictions
Denied	1525	344	1869
Approved	27	1109	1136
Total Actuals	1552	1453	

TABLE 3
CONFUSION MATRIX OF C5.0 MODEL WITH RACIAL FEARURES FOR WHITE APPLICANTS

	Denied	Approved	Total Predictions
Denied	1412	343	1755
Approved	42	1192	1234
Total Actuals	1454	1535	

In table 2, the rate of misclassifying the application decision for black applicants with respect to approvals as denials is approximately 24%, while the rate of misclassifying their application decision with respect to denials as approvals is approximately 2%. On the other hand, in table 3, the rate of misclassifying the application decision for white applicants with respect to approvals as denials is approximately 22%, while the rate of misclassifying their application decision with respect to denials as approvals is approximately 3%.

For comparison, the confusion matrices and statistical measures of the C5.0 model that had racial features removed, with the "*action_taken*" as the class are as follows:

TABLE 4
METRICS OF C5.0 MODEL WITHOUT RACIAL FEATURES

Accuracy : 0.8664
Sensitivity : 0.9523
Specificity : 0.7806

In table 4 above, the accuracy of the racially blind model in predicting the action taken is approximately 87% which is just one-point shy of its predecessor. With a sensitivity or recall of approximately 95%, and a specificity 78%, while there was no change in the specificity.

TABLE 5
CONFUSION MATRIX OF C5.0 MODEL WITHOUT RACIAL FEARURES FOR BLACK APPLICANTS

	Denied	Approved	Total Predictions
Denied	1486	341	1827
Approved	66	1112	1178
Total Actuals	1552	1453	

TABLE 6
CONFUSION MATRIX OF C5.0 MODEL WITHOUT RACIAL FEARURES FOR WHITE APPLICANTS

	Denied	Approved	Total Predictions
Denied	1372	326	1698
Approved	82	1209	1291
Total Actuals	1454	1535	

In table 5, the rate of misclassifying the application decision for black applicants with respect to approvals as denials is approximately 24%, while the rate of misclassifying their application decision with respect to denials as approvals is approximately 4%. On the other hand, in table 6, the rate of misclassifying the application decision for white applicants with respect to approvals as denials is approximately 21%, while the rate of misclassifying their application decision with respect to denials as approvals is approximately 6%.

In comparing the approval rate of black and white applicants in both C5.0 models, it can be observed that in the first model, their approval rates are approximately 76% and 78% respectively, while in the second model,

their approval rates are approximately 77% and 79% respectively.

The statistical measures and confusion matrices of the NNET model that had racial features, with the "*action_taken*" as the class are as follows:

TABLE 7
METRICS OF NNET MODEL WITH RACIAL FEATURES

Accuracy : 0.8702
Sensitivity : 0.9897
Specificity : 0.7507

In table 7 above, the accuracy of the racially biased NNET model in predicting the action taken is approximately 87% which is one-point lower than the racially biased C5.0 model. Similarly, this means that the model is ok at predicting the present outcome of mortgage applications based on the existing data. Yet, the sensitivity or recall of this neural network model, is approximately 99%. This implies that unlike the C5.0 models, it is nearly perfect at detecting applicants who will be denied which will be extremely good in a fair lending system. On the other hand, the specificity, which is the rate at which the model accurately identifies applicants that will be approved is approximately 75%.

TABLE 10
CONFUSION MATRIX OF NNET MODEL WITH RACIAL FEARURES FOR BLACK APPLICANTS

	Denied	Approved	Total Predictions
Denied	1543	380	1923
Approved	9	1073	1082
Total Actuals	1552	1453	

TABLE 11
CONFUSION MATRIX OF NNET MODEL WITH RACIAL FEARURES FOR WHITE APPLICANTS

	Denied	Approved	Total Predictions
Denied	1432	367	1799
Approved	22	1168	1190
Total Actuals	1454	1535	

In table 10, the rate of misclassifying the application decision for black applicants with respect to approvals as denials is approximately 26%, while the rate of misclassifying their application decision with respect to denials as approvals is approximately 1%. On the other hand, in table 11, the rate of misclassifying the application decision for white applicants with respect to approvals as denials is approximately 24%, while the rate of misclassifying their application decision with respect to denials as approvals is approximately 2%.

For comparison, the confusion matrices and statistical measures of the NNET model that had racial features removed, with the "*action_taken*" as the class are as follows:

TABLE 12
METRICS OF NNET MODEL WITHOUT RACIAL FEATURES

Accuracy : 0.8674
Sensitivity : 0.9960
Specificity : 0.7389

In table 12 above, the accuracy of the racially blind model in predicting the action taken is approximately 87% which is the same as the racially biased NNET model and the racially blind C5.0 model. And just like the its predecessors, the model is pretty good at predicting the present outcome of mortgage applications based on the existing data. With a sensitivity or recall of approximately 100%, it is perfect at detecting applicants who will be denied which. But it has a specificity of approximately 74%, making it average for determining which applicants will be approved for a home loan, within the current lending system.

TABLE 13
CONFUSION MATRIX OF NNET MODEL WITHOUT RACIAL FEARURES FOR BLACK APPLICANTS

	Denied	Approved	Total Predictions
Denied	1550	394	1944
Approved	2	1059	1061
Total Actuals	1552	1453	

TABLE 14
CONFUSION MATRIX OF NNET MODEL WITHOUT RACIAL FEARURES FOR WHITE APPLICANTS

	Denied	Approved	Total Predictions
Denied	1447	407	1854
Approved	7	1128	1135
Total Actuals	1454	1535	

In table 13, the rate of misclassifying the application decision for black applicants with respect to approvals as denials is approximately 27%, while the rate of misclassifying their application decision with respect to denials as approvals is approximately 0.001%. On the other hand, in table 14, the rate of misclassifying the application decision for white applicants with respect to approvals as denials is approximately 27%, while the rate of misclassifying their application decision with respect to denials as approvals is approximately 0.01%. **In comparing the approval rate of black and white applicants in both NNET models, it can be observed that in the first model, their approval rates are approximately 74% and 76% respectively, while in the second model, their approval rates are approximately 73% and 75% respectively.**

Results of Predictions with "Race" as Class

The confusion matrices and statistical measures of the C5.0 model, with the "*applicant-race-1*" as the class are as follows:

TABLE 15
METRICS OF C5.0 MODEL WITH RACE AS CLASS

Accuracy : 0.767
Sensitivity : 0.7514
Specificity : 0.7826

In table 15 above, the accuracy of the C5.0 model in predicting the race of the applicant based on the action taken and other recorded features is approximately 77%, with a sensitivity or recall of approximately 75%, and a specificity 78%.

TABLE 16
CONFUSION MATRIX OF C5.0 MODEL WITH RACE AS CLASS BY APPROVALS

	Black	White	Total Predictions
Black	6401	2057	8458
White	2955	9453	12408
Total Actuals	9356	11510	

TABLE 17
CONFUSION MATRIX OF C5.0 MODEL WITH RACE AS CLASS BY DENIALS

	Black	White	Total Predictions
Black	4118	987	5105
White	526	1503	2029
Total Actuals	4644	2490	

In table 16 and 17 above, approximately 32% of the black applicants who were approved for a home loan were misclassified as white, while only 18% of the white applicants who were approved for a home loan were misclassified as black. In the case of the denials, approximately 11% of the black applicants who were denied a home loan were misclassified as white, while approximately 40% of the white applicants who were denied a home loan were misclassified as black.

The confusion matrices and statistical measures of the NNET model, with the "*applicant-race-1*" as the class are as follows:

TABLE 18
METRICS OF NNET MODEL WITH RACE AS CLASS

Accuracy : 0.7204
Sensitivity : 0.6896
Specificity : 0.7511

In table 18 above, the accuracy of the NNET model in predicting the race of the applicant based on the action taken and other recorded features is approximately 72%, with a sensitivity or recall of approximately 69%, and a specificity 75%.

TABLE 19
CONFUSION MATRIX OF NNET MODEL WITH RACE AS CLASS BY APPROVALS

	Black	White	Total Predictions
Black	5565	2244	7809
White	3791	9266	13057
Total Actuals	9356	11510	

TABLE 20
CONFUSION MATRIX OF C5.0 MODEL WITH RACE AS CLASS BY DENIALS

	Black	White	Total Predictions
Black	4090	1240	5330
White	554	1250	1804
Total Actuals	4644	2490	

In table 19 and 20 above, approximately 41% of the black applicants who were approve-d for a home loan were misclassified as white, while only 24% of the white applicants who were approved for a home loan were misclassified as black. In the case of the denials, approximately 12% of the black applicants who were denied a home loan were misclassified as white, while approximately 50% of the white applicants who were denied a home loan were misclassified as black.

V. CONCLUSIONS & FUTURE WORK

While it is difficult to clearly say that the results from the first sets of tests are indicative of discriminatory lending, the second set of tests that trained the C5.0 decision tree and NNET neural network models to predict the race of the applicant based on the application decision and other variables are more definitive. Given that in both models, the rate of misclassification of black applicants as white applicants, who had their loans approved, was approximately **twice the rate** of misclassification of white applicants as black applicants, who had their loans approved.

In addition, in both models, the rate of misclassification of white applicants as black applicants, who had their loans denied, was approximately **four times the rate** of misclassification of black applicants as white applicants, who had their loans denied.

Based on these results, it is conclusive that the hypothesis "*that race plays a huge role in the decision-making process of home mortgage approvals or denials, and it has not improved since the last major report on it in 2014"* **is true**.

Within the next five years, another research can be carried out using the new HMDA data, since it would have been updated to have important features like the applicant's credit score, loan interest rate and spread, debt-to-income-ratio, and prepayment penalties. Predictions can then be run to see if your race determines things like the loan interest rate and spread.

REFERENCES

[1] Staff, I. (2018, March 16). Home Mortgage Disclosure Act (HMDA). Retrieved from https://www.investopedia.com/terms/h/home-mortgage-disclosure-act-hmda.asp

[2] Luhby, T. (2014, August 21). 5 disturbing stats on black-white inequality Retrieved from http://money.cnn.com/2014/08/21/news/economy/black-white-inequality/index.html

[3] Morris, M. W. (2009). *Discrimination and Mortgage Lending in America* (pp. 2-3, Publication). Baltimore, MD: NAACP.

[4] National Community Reinvestment Coalition, *The 2005 Fair Lending Disparities: Stubborn and Persistent II* (Washington, DC: National Community Reinvestment Coalition, 2006), http://www.ncrc.org/images/stories/pdf/research/ncrc%202005%20hmda%20report.pdf

[5] Federal Home Loan Mortgage Corporation, *Annual Report* (McLean, VA: Federal Home Loan Mortgage Corporation, 2007).

[6] U.S. Department of Housing and Urban Development, *Unequal Burden*.

[7] Ibid.

[8] Center for Responsible Lending, http://www.responsiblelending.org, 2009. See also A. Rivera, B. Cotto-Escalara, A. Desai, J. Huezo, and D. Muhammad, *Foreclosed: State of the Dream* (Boston, MA: United for a Fair Economy, 2008).

[9] Gaspaire, Brent. "Redlining (1937-)." *BlackPast.org Remembered and Reclaimed*, www.blackpast.org/aah/redlining-1937.