EVALUATING DISCRIMINATION OF AI AND ALGORITHMIC LENDING DECISIONS WHEN RACE DATA ARE UNAVAILABLE

Shastri Sandy[†],
Joe Chance[††],
Christine Polek[†††],
and Daniel Wang[††††]

Abstract

Algorithms, big data, and artificial intelligence ("AI") models have become increasingly prevalent in lending practices and are recognized for their potential to reduce subjective bias and promote efficiency in credit risk assessment. However, they typically operate like "black boxes" and pose challenges for legal scrutiny and compliance. A key issue lies in the inadvertent use of variables that act as proxies for race, potentially leading to discriminatory outcomes. In traditional disparate impact regulations, if a proxy for race serves a legitimate business purpose, its use is not considered a violation. In contrast, black-box models pose a unique challenge: any finding that race or its proxy is being picked up by the model could constitute a disparate impact violation due to the opacity of the underlying parameters. We present a framework for evaluating disparate impact in "black-box" models when race is unobserved, using regression analysis to predict credit scores based on proxies for race, such as the racial composition of residential populations and mortgage borrowers. Additionally, we assess the robustness of our findings in race-neutral environments, offering a methodological approach to evaluate potential disparate impact in the absence of direct race data. Our work highlights the complexities of disparate impact analysis in the context of black-box algorithmic

† Shastri Sandy, Ph.D.; Principal, The Brattle Group. Dr. Sandy has experience in investigations and litigation involving income, credit, labor markets, and protection of consumers. He has worked with diverse and big data and reviewed complex computer programs and models.

†† Joe Chance, Ph.D.; Associate, The Brattle Group. Dr. Chance has experience consulting on antitrust cases involving collusion and market allocation in online advertising, media, and utilities. His work focuses on applied econometrics to perform statistical analysis of very large datasets and to model damages.

††† Christine Polek, Ph.D.; Senior Principal, Keystone Strategy. Dr. Polek has experience in investigations and litigation involving income, labor, and tax issues, including discrimination in employment and valuation.

†††† Daniel Wang; Senior Research Analyst, The Brattle Group.

models and underscores the importance of transparency and compliance in their deployment.

TABLE OF CONTENTS

INTRODUCTION

Algorithmic models that harness emerging machine learning and AI technologies are likely to become a key resource for financial services firms.[1] Lenders already rely on algorithmic models to make credit-related decisions.[2] While these models could be viewed as neutral in the sense that they do not rely on subjective assessment processes, they are usually proprietary "black-box" models that may nonetheless produce biased or inconsistent outcomes and violate regulatory or legal restrictions.[3]

However, it may not be straightforward to empirically assess whether lending outcomes are biased against race because nonmortgage lenders are

1. Charlotte Nan Jiang & Nadia Novik, *Leveraging Big Data and Machine Learning in Credit Reporting*, WORLD BANK (Aug. 10, 2021), https://blogs.worldbank.org/developmenttalk/leveraging-big-data-and-machine-learning-credit-reporting [https://perma.cc/QVD9-KGJA]; Lukas Ryll et al., *Transforming Paradigms: A Global AI in Financial Services Survey*, CAMBRIDGE CTR. FOR ALT. FIN., 1, 11 (Jan. 2020), https://ssrn.com/abstract=3532038 [https://perma.cc/MM53-L85F].

2. Michael Akinwumi et al., *An AI Fair Lending Policy Agenda for the Federal Financial Regulators*, BROOKINGS 1, 4 (2021), https://www.brookings.edu/wp-content/uploads/2021/12/Akinwumi_Merrill_Rice_Saleh_Yap_12-01-2021-1.pdf [https://perma.cc/X3HP-8ARY]; Mikella Hurley & Julius Adebayo, *Credit Scoring in the Era of Big Data*, 18 YALE J.L. & TECH. 148, 157 (2016); *see also* Raj Dash et al., *Designing Next-Generation Credit-Decisioning Models*, MCKINSEY & CO. (Dec. 2, 2021), https://www.mckinsey.com/capabilities/risk-and-resilience/our-insights/designing-next-generation-credit-decisioning-models [https://perma.cc/3T25-JHRL] (finding that AI-powered credit scoring models can be up to 10% more accurate than traditional models).

3. Holli Sargeant, *Algorithmic Decision-Making in Financial Services: Economic and Normative Outcomes in Consumer Credit*, 3 AI AND ETHICS 1295, 1303 (Nov. 21, 2022) (explaining that machine learning is referred to as "black box" because it obscures the rationale for a decision); Crystal S. Yang & Will Dobbie, *Equal Protection Under Algorithms: A New Statistical and Legal Framework*, 119 MICH. L. REV. 291, 301–02 (2020) (discussing predictive algorithms causing Equal Protection Clause concerns).

generally prohibited from collecting or identifying borrower race.[4] As non-mortgage lending such as auto-loans and credit cards have grown substantially to be a significant source of consumer lending,[5] these loans have also gained the attention of regulators.[6] This paper provides a case study to demonstrate how econometric techniques may be used to evaluate whether an algorithmic model results in disparate impact when race data for borrowers are not available.[7] More specifically, this paper illustrates how econometric analyses may be applied to detect and evaluate disparate impact of credit scoring, a key factor used to determine who gets credit and how that credit is priced.[8]

U.S. courts and regulatory agencies evaluate disparate impact claims using a structured framework that first examines whether a policy or practice disproportionately affects a protected class, such as race, color, or national origin.[9] If such an adverse effect is identified, the next step considers whether there is a substantial, legitimate justification for the policy.[10] Finally, the analysis assesses whether a less discriminatory alternative exists that would achieve the same objectives.[11] This paper focuses exclusively on the first stage—

4. Equal Credit Opportunity Act, 15 U.S.C. §1691 (restricting creditors from collecting race data in most credit transactions); Home Mortgage Disclosure, 12 C.F.R. § 1003 (2024) (requiring the collection and reporting of such data for mortgage lending to promote transparency and prevent discrimination*); REGStats*, U.S. DEPT. OF AGRIC., https://www.usda.gov/regstats [https://perma.cc/7ESB-EGRV] (last visited Feb. 13, 2024, 8:30 PM) (showing the USDA collection of voluntarily provided race data for targeted agricultural loan programs).

5. FED. RSRV. BANK OF N.Y. RSCH. AND STAT. GRP., QUARTERLY REPORT ON HOUSEHOLD DEBT AND CREDIT 2024: Q2 (2024), https://www.newyorkfed.org/medialibrary/interactives/householdcredit/data/pdf/HHDC_2024Q2 [https://perma.cc/U2ND-CXLJ] (noting that auto loans increased by $10 billion to reach $1.63 trillion and credit card balances increased by $27 billion to reach $1.14 trillion). *Household Debt Increased Moderately in Q2 2024; Auto and Credit Card Delinquency Rates Remain Elevated*, FED. RSRV. BANK OF N.Y. (Aug. 6, 2024), https://www.newyorkfed.org/newsevents/news/research/2024/20240806 [https://perma.cc/RY9R-EKF7].

6. CONSUMER FIN. PROT. BUREAU, SUPERVISORY HIGHLIGHTS: SERVICING AND COLLECTION OF CONSUMER DEBT ISSUE 34, 2 (2024) (discussing that the Consumer Financial Protection Bureau provided supervision on non-mortgage lending such as auto loans, student loans, and credit cards).

7. Griggs v. Duke Power Company, 401 U.S. 424, 426 (1971) (highlighting disparate impact as referred to a selection test or tool that disproportionately affects a protected class and is often unintentional); *Cf. Disparate Treatment*, CORNELL L. SCH.: LEGAL INFORMATION INST., https://www.law.cornell.edu/wex/disparate_treatment [https://perma.cc/F5EC-QJZN] (last visited Feb. 13, 2024, 8:45 PM) (defining disparate treatment as involves intentional actions designed to disadvantage a protected class); *see generally* Civil Rights Act of 1964, 42 U.S.C. § 2000e (prohibiting both disparate impact and disparate treatment in employment); Age Discrimination in Employment Act of 1967, 29 U.S.C. § 621, (prohibiting both disparate impact and disparate treatment based on age in employment); Fair Housing Act, 42 U.S.C. § 3604 (prohibiting both disparate impact and disparate treatment in housing rentals and sales).

8. Lisa Rice & Deidre Swesnik, *Discriminatory Effects of Credit Scoring on Communities of Color*, 46 SUFFOLK U. L. REV. 935, 936 (2013) (showing that credit scoring models can unintentionally embed socioeconomic and racial disparities, raising concerns about potential disparate impact); Robert B. Avery et al., *Does Credit Scoring Produce a Disparate Impact?*, 40 REAL EST. ECON. S65, S66 (2012) (examining the individual predictive factors included in credit scoring models and assessing whether including each of these factors in a credit scoring model results in a disparate impact by race or ethnicity, age or gender); *How Your Credit Score Impacts Your Financial Future*, FINRA, https://www.finra.org/investors/personal-finance/how-your-credit-score-impacts-your-financial-future [https://perma.cc/Z63Y-FT9Y] (last visited Feb. 14, 1:20 PM) ("The riskier you appear to the lender, the less likely you will be to get credit or, if you are approved, the more that credit will cost you. In other words, you will pay more to borrow money.").

9. *Title VI Legal Manual: Section VII- Proving Discrimination- Disparate Impact*, U.S. DEPT. OF JUST., https://www.justice.gov/crt/fcs/T6Manual7 [https://perma.cc/7VJQ-UDHZ] (last visited Mar. 2, 2025).

10. *Id.*

11. *Id.*

determining whether the algorithmic credit scoring model produces disparate outcomes[12]—in a setting where individual-level data on protected class status is unavailable.

We examine the credit scores as measured by a leading credit scoring model, VantageScore, for a sample of approximately 500,000 credit score users ("consumers") from North Carolina.[13] We use ZIP code-level racial shares from both Census data and Home Mortgage Disclosure Act (HMDA) mortgage applicant data to proxy for race.[14] By using HMDA racial composition data, our example provides a framework which takes into consideration how well the racial composition of the residential population reflects the racial composition of the sub-population in question—in this case, the credit applicant pool.[15] We examine the robustness of the results of the illustrative example by replicating our analyses in a race neutral environment.[16]

I. DISPARATE IMPACT IN CREDIT SCORING

Lending decisions are increasingly implemented with AI, "black-box" algorithms in which the inputs and outputs are known while the specific processes and calculations may not be well-known to the end users.[17] While these models promise enhanced accuracy and efficiency in evaluating borrowers' creditworthiness,[18] the opacity of AI models generally and the rise of their use has raised the concern of policymakers, exemplified, for example, by the White House's "Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence."[19]

Moreover, policymakers have warned and are concerned that the use of algorithmic models in decision making processes can lead to disparate impact.[20] According to the Consumer Financial Protection Bureau (CFPB), this occurs when:

> "A creditor employs facially neutral policies or practices that have an adverse effect or impact on a member of a protected class unless it

12. *Id.*

13. *See infra* Part II (discussing a credit score study).

14. *Mortgage Data (HMDA)*, CONSUMER FIN. PROT. BUREAU, https://www.consumerfinance.gov/data-research/hmda/ [https://perma.cc/SPV6-TPVR] (last visited Feb. 14, 2025, 1:36 PM) (explaining that the Home Mortgage Disclosure Act requires financial institutions to maintain, report, and publicly disclose loan-level information about mortgages including records on mortgage applications, denials, and originations).

15. Cherry v. Amoco Oil Co., 490 F. Supp. 1026, 1028–29, 1030 (N.D. Ga. 1980) (holding that plaintiff failed to make out a prima facie case after plaintiff presented a scattergram of percentage credit application acceptance rate versus percentage non-white for all ZIP codes in Atlanta with a low rating on Amoco's credit application scoring model as evidence showing negative correlation).

16. *See infra* Part IV (analyzing results in a race neutral environment).

17. Christine Polek & Shastri Sandy, *The Disparate Impact of Artificial Intelligence and Machine Learning*, 21 COLO. TECH. L.J. 85, 96 (2023).

18. *See* Sargeant, *supra* note 3, at 1298 (discussing banks' uses of machine learning to evaluate consumers' credit worthiness).

19. Exec. Order No. 14110, 88 Fed. Reg. 210, 75191 (Oct. 30, 2023) (rescinded 2025).

20. *Artificial Intelligence 2023 Legislation*, NAT'L. CONF. OF STATE LEGISLATORS (Jan. 12, 2024), https://www.ncsl.org/technology-and-communication/artificial-intelligence-2023-legislation [https://perma.cc/GGM2-PGS8] (adopting resolutions to assess the use of AI among Connecticut's state agencies to ensure no disparate impact occurs).

> meets a legitimate business need that cannot reasonably be achieved by means that are less disparate in their impact."[21]

While policies such as the Equal Credit Opportunity Act and the Fair Housing Act explicitly prohibit nonmortgage lenders from considering an applicant's race,[22] [23] algorithmic models that rely on metrics that appear neutral or relevant for business purposes may function as proxies for a protected class and lead to disproportionately adverse effects.[24] For example, one recent study identified several variables relating to an individuals' digital profile outperforms traditional credit score models in predicting who would pay back a loan.[25] While none of these variables (type of computer, structure of email name, etc.) explicitly identify an individual's race, gender, or age, each is correlated with one or more of these protected classes.[26] If the correlation is significant enough, the model may perform similarly to a model that uses race, gender, or age directly.[27] As the FDIC has said, "[t]he system could include a prohibited basis as one of the variables, or, if not a prohibited basis itself, a factor that is so highly correlated with a prohibited basis that it serves as a proxy for the basis."[28]

U.S. courts have generally ruled that if facially neutral policy variables lead to disparate outcomes across protected groups, they must be prohibited unless they have a legitimate business justification.[29] In the landmark disparate impact ruling, *Griggs v. Duke Power Company*, the Supreme Court ruled that if a facially neutral employment criterium—in this case a high school diploma—has a disparate impact on a protected class, the employment criteria must be justified by business necessity to be permissible.[30] In 2019, the Department of Justice (DOJ) brought a case against Meta Platforms alleging that its ad delivery algorithms caused a disparate impact by disproportionately steering housing

---

21. *CFPB Consumer Laws and Regulations: Equal Credit Opportunity Act (ECOA)*, CONSUMER FIN. PROT. BUREAU 1, 1 (June, 2013), https://files.consumerfinance.gov/f/201306_cfpb_laws-and-regulations_ecoa-combined-june-2013.pdf [https://perma.cc/UC2T-LJC3]; *see also Comment for 1002.6 – Rules Concerning Evaluation of Applications*, CONSUMER FIN. PROT. BUREAU, https://www.consumerfinance.gov/rules-policy/regulations/1002/interp-6/ [https://perma.cc/XN5X-EDQZ] (last visited Feb. 16, 2025, 9:17 AM) (discussing the effects test that established the level of scrutiny for creditor actions that cause a disproportionally negative impact on a prohibited basis).

22. *Fair Lending: Race and Gender Data Are Limited for Nonmortgage Lending*, U.S. GOV'T. ACCOUNTABILITY OFF. (July 17, 2008), https://www.gao.gov/products/gao-08-1023t [https://perma.cc/94D8-EKGH]; *see also Using Publicly Available Information to Proxy for Unidentified Race and Ethnicity*, CONSUMER FIN. PROT. BUREAU 1, 4 (2014) https://files.consumerfinance.gov/f/201409_cfpb_report_proxy-methodology.pdf [https://perma.cc/86EN-RMZM] (discussing the proxy for race and ethnicity constructed by the CFPB to avoid violating federal laws).

23. Rules Concerning Request for Information, 12 C.F.R. § 1002.5(b); Equal Credit Opportunity Act, 15 U.S.C. § 1691(a)(1).

24. Anya Prince & Daniel B. Schwarcz, *Proxy Discrimination in the Age of Artificial Intelligence and Big Data*, 105 IOWA L. REV. 1257, 1260–64 (2020); *see also* Polek & Sandy, *supra* note 17, at 91.

25. *See* Prince & Schwarcz, *supra* note 24, at 1277.

26. *Id.* at 1263.

27. *Id.* at 1263–64.

28. *Fair Lending Implications of Credit Scoring Systems*, FDIC, 23, 25 (2005), https://www.fdic.gov/regulations/examinations/supervisory/insights/sisum05/sisummer05-article3.pdf [https://perma.cc/5PGN-CRLV].

29. Griggs v. Duke Power Co., 401 U.S. 424 (1971); *see also* Tex. Dep't of Hous. & Cmty. Affairs v. Inclusive Cmtys. Project, Inc., 576 U.S. 519 (2015) (applying the disparate impact test to housing discrimination under the Fair Housing Act).

30. *Griggs*, 401 U.S. at 424.

advertisements away from users in protected classes, thereby limiting their access to housing opportunities in violation of the Fair Housing Act.[31] In 2023, the DOJ and Department of Housing and Urban Development (HUD) filed a Statement of Interest in explaining the Fair Housing Act's application to algorithm-based tenant screening systems.[32] The Statement of Interest was filed in a lawsuit alleging that defendants' use of an algorithm-based scoring system to screen tenants discriminates against Black and Hispanic rental applicants in violation of the Fair Housing Act.[33]

Credit scoring models are typically "black-box" algorithms that rely on inputs that are directly linked to consumer credit worthiness.[34] However, some of the factors that the algorithms incorporate, such as the frequency of address changes, may lack a clear business purpose but still have predictive value.[35] These variables may inadvertently act as proxies for demographic characteristics including race or ethnicity, and may contribute to disparate impacts on protected classes.[36] It is relatively straightforward to evaluate whether the frequency a consumer's address changes functions as a proxy for race and causes an adverse impact when the consumer's actual race is known, using methods that quantify

31. United States v. Meta Platforms, Inc., No. 23-MC-80249-PHK, 2023 WL 8438579 (N.D. Cal. Dec. 5, 2023); *Justice Department Secures Groundbreaking Settlement Agreement with Meta Platforms, Formerly Known as Facebook, to Resolve Allegations of Discriminatory Advertising*, U.S. DEP'T OF JUST.: ARCHIVES (June 21, 2022), https://www.justice.gov/opa/pr/justice-department-secures-groundbreaking-settlement-agreement-meta-platforms-formerly-known [https://perma.cc/BBR3-LFJH].

32. Statement of Interest of the United States, Louis v. Saferent Sols., LLC, 685 F. Supp. 3d 19 (D. Mass. 2023) (No. 22cv10800-AK).

33. *Id*.

34. Tamon Tananilgul, *Achieving Fairness in Machine Learning-Powered Credit Scoring Models*, MEDIUM (Oct. 26, 2024), https://michikotananilgul.medium.com/achieving-fairness-in-machine-learning-powered-credit-scoring-models-a1df4c9001e0 [https://perma.cc/6WHS-KBGH].

35. *Does Changing My Address or Moving Often Affect My Credit Score?*, LOQBOX (Nov. 25, 2024), https://www.loqbox.com/en-gb/blog/does-changing-my-address-or-moving-often-affect-my-credit-score [https://perma.cc/BYU4-ZFAS] ("Changing your address regularly won't change your score but it could affect your ability to get credit.").

36. *See* Prince & Schwarcz, *supra* note 24, at 1260. (2020) ("Proxy discrimination is a particularly pernicious subset of disparate impact. Like all forms of disparate impact, it involves a facially neutral practice that disproportionately harms members of a protected class. But a practice producing a disparate impact only amounts to proxy discrimination when the usefulness to the discriminator of the facially neutral practice derives, at least in part, from the very fact that it produces a disparate impact."); *see also* Robert Bartlett et al., *Algorithmic Discrimination and Input Accountability under the Civil Rights Acts*, 36 BERKELEY TECH. L.J. 675, 686 (2021), "[T]he use of these proxy variables can result in members of a protected class experiencing disparate outcomes that are not justified by business necessity. The problem arises from what researchers call 'redundant encodings'-the fact that a proxy variable can be predictive of a legitimate target variable and membership in a protected group. Relying on these proxy variables, therefore, risks penalizing members of the protected group who are otherwise qualified in the legitimate target variable."); *see also* Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 CALIF. L. REV. 671, 673 (2016) ("If data miners are not careful, the process can result in disproportionately adverse outcomes concentrated within historically disadvantaged groups in ways that look a lot like discrimination."); *see also* Hurley & Adebayo, *supra* note 2 ("The machine-learning and feature-selection process may also produce models that perpetuate implicit forms of bias and that inadvertently factor in sensitive characteristics such as race. As we will discuss in further detail, longstanding Federal law prohibits lenders from directly taking characteristics such as race or sex into account when making lending decisions. When a credit scorer has thousands of data points to work with, however, the machine-learning process may indirectly consider sensitive characteristics, such as race, even when those characteristics are not directly designated as input values.").

disparities.[37] However, when consumer race data is unavailable, racial demographics of consumer ZIP codes can be reasonably used to infer consumer race.[38] By linking ZIP code demographics to borrowers, analysts can determine whether credit scoring models using facially neutral proxies such as change of address frequency produce disparate impacts that adversely affect specific groups.[39] Below, we present an illustrative analysis of a 'black-box' credit scoring model to demonstrate an approach for evaluating whether such models—along with broader applications of AI—exhibit disparate impacts when using known proxies for race.[40]

II. CASE STUDY: CREDIT SCORES OF CREDIT CONSUMERS IN NORTH CAROLINA

To provide an illustrative example of how statistical analysis could be used to detect and evaluate disparate impact of an algorithmic model when consumer race data is unavailable, we examine the underlying credit attributes and credit scores of credit line holders in North Carolina in March 2010.[41] In particular, we examine credit scores that were determined by a major credit scoring company, VantageScore, and proxy for race by using both Census data and HMDA mortgage data.[42] Because VantageScore does not report their modeling algorithm but does report the set of final attributes that generate the consumer credit score, we approximate the scoring model using regression analysis and identified credit attribute information using Equifax data.[43]

A. Reconstructing Borrower Credit Scores and Credit Characteristics

1. About Credit Scores

Credit scores modeled by VantageScore and its competitors inform lending decisions, including whether to grant credit, what terms to offer, and what

37. Joshua Grossman et al., *Reconciling Legal and Empirical Conceptions of Disparate Impact: An Analysis of Police Stops Across California*, 1 J. L. & EMPIRICAL ANALYSIS 118 (2024); *see also* Rice & Swesnik, *supra* note 8, at 952–53 (highlighting how seemingly neutral factors serve as proxies for race and exacerbate structural inequalities, necessitating targeted reforms to address adverse impacts on communities of color).

38. Kevin Fiscella & Allen M. Fremont, *Use of Geocoding and Surname Analysis to Estimate Race and Ethnicity*, NAT'L LIBR. OF MED. 1, 3 (May 5, 2006), https://pmc.ncbi.nlm.nih.gov/articles/PMC1797082/pdf/hesr0041-1482.pdf [https://perma.cc/J9VE-S6Z2].

39. *Id.* Algorithms based on machine learning may rely on specific data inputs to generate predictions or outcomes; however, these inputs are often not explicitly known or easily interpretable. This lack of transparency poses challenges in identifying the precise variables that may contribute to potential disparate impacts or other unintended biases. In such cases, establishing a business justification for the variables may become impractical.

40. *See infra* Part II (discussing analysis of VantageScore credit scores).

41. *See infra* Figure 1: VantageScore Credit Scores for Sample of Credit Consumers in North Carolina by ZIP code (illustrating VantageScore credit scores in North Carolina).

42. *Id.*

43. *See infra* Figure 2: VantageScore 3.0 Credit Attributes Matched to Equifax Data [hereinafter Figure 2] (matching Equifax data to VantageScore credit attributes)

interest rate to charge.[44] VantageScore uses a proprietary algorithm to determine credit scores for approximately 35 million individuals across the United States.[45] VantageScore credit scores are used by a wide range of creditors as an indicator of the likelihood that a person will fall at least 90 days behind on a bill within the next 24 months.[46]

Like most credit scores, the VantageScore credit score is a three-digit number, ranging from 300 to 850.[47] While the VantageScore credit score model details are not made publicly available, the company does disclose information about the inputs used to model credit scores.[48] These inputs include a number of traditional credit quality indicators, such as payment history and length of credit history.[49]

2. Credit Score Sample Selection

We use a sample of VantageScore 3.0 credit scores and associated credit characteristics from approximately 500,000 consumers in North Carolina to demonstrate a generalized approach for assessing disparate impact in algorithmic "black-box" models.[50] We focus our analysis on the state of North Carolina, given its representativeness of national characteristics in race, income, and credit scores.[51] In the 2010 Census for North Carolina, the population is 68.5% White, 21.5% Black, and 8.4% Hispanic or Latino with the median household income in North Carolina being $43,830.[52] Nationally in the 2010 Census, the population is 72.4% White, 12.6% Black, and 16.3% Hispanic or

44. Jennifer Streaks, *What Is a VantageScore?*, BUSINESS INSIDER, https://www.businessinsider.com/personal-finance/credit-score/what-is-vantagescore [https://perma.cc/X9NF-KB97] (last updated Nov. 25, 2024, 1:28 PM).

45. PYMNTS, *The New World of Credit Scoring*, PYMNTS (June 24, 2015), https://www.pymnts.com/company-spotlight/2015/the-new-world-of-credit-scoring/ [https://perma.cc/LN2B-K6R9] ("Among VantageScore's conclusions: As many as 30 million-35 million American consumers, flying under the radar of traditional scoring metrics can have an accurate credit score and are now able to become visible and attractive customers to lenders.").

46. VANTAGESCORE, *Credit Score Basics, Part 1: What's Behind Credit Scores?* (2011), https://www.vantagescore.com/wp-content/uploads/2022/02/VantageScore_CreditScoreBasics-Part_1_9-29-11.pdf [https://perma.cc/K6CS-6CNM].

47. VANTAGESCORE, VANTAGESCORE 3.0: BETTER PREDICTIVE ABILITY AMONG SOUGHT-AFTER BORROWERS 1 (2013) (illustrating how VantageScore credit scores are distinguished from credit scores calculated in other models, such as FICO, because it considers entire credit history instead of past 24 months, which allows scores to be modeled for consumer with less than six months of credit history on their credit file).

48. *The Science Behind VantageScore*, VANTAGESCORE, https://www.vantagescore.com/lenders/why-vantagescore/the-science/ [https://perma.cc/J6GH-9S89] (last visited Feb. 17, 2025) (discussing how VantageScore's model is periodically refined and updated to provide more accuracy and reliability with recent advancements reflecting the use of machine learning in the assessment of credit scoring). Our analysis relies on VantgageScore credit scores generated with the VantageScore Version 3.0 model.

49. *See* VANTAGESCORE, *supra* note 47. The VantageScore credit score is calculated by the model. VantageScore credit scores are not subjectively adjusted.

50. *Id*. We acquired data on non-mortgage borrowers VantageScore credit scores and credit characteristics from Equifax. As observations in the Equifax data are at the person-account level, we collapse the data to the person-level, using time-invariant characteristics of credit holders in March 2010.

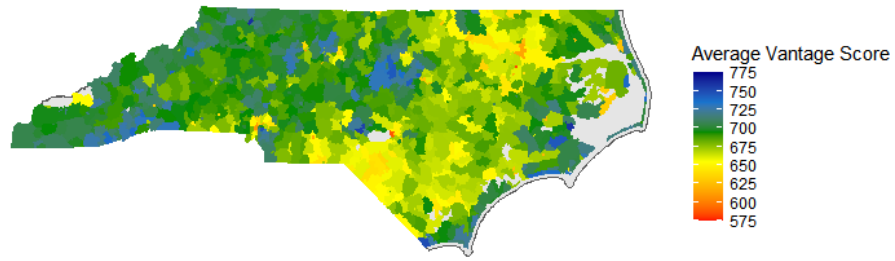
51. *2010 Census: North Carolina Profile*, U.S. CENSUS BUREAU. (2010), https://www2.census.gov/library/publications/decennial/2010/cph-2/cph-2-35.pdf [https://perma.cc/DEY9-H5NY]; *see also* Figure 1: VantageScore Credit Scores for Sample of Credit Consumers in North Carolina by ZIP code (illustrating Vantage Credit Scores in North Carolina).

52. *See North Carolina: 2010*, *supra* note 51.

Latino with a median household income of $49,280.[53] Thus, in 2010, North Carolina provides a reasonable representation of the demographic make-up of the U.S. population.[54]

The average credit score for our sample of borrowers is 690, and the distribution of average credit scores, for March 2010, by ZIP code is depicted in Figure 1 below.[55] Figure 1 below shows the **Western half** of the state and the state's **East coast** as having higher credit scores in comparison to the **Eastern inland half** of the state.[56]

Figure 1: VantageScore Credit Scores for Sample of Credit Consumers in North Carolina by ZIP code[57]



Note: Gray area is water.[58]

3. Identifying Credit Characteristics

VantageScore lists the following categories and relative weights of attributes as inputs in the 3.0 model: payment history (40%), depth of credit (21%), credit utilization (20%), balances (11%), recent credit activity (5%) and available credit (3%).[59] Because VantageScore does not provide the underlying inputs to the final credit score result, we rely on Equifax, one of the credit reporting agencies, for underlying measures of credit qualities used by

53. Karen R. Humes et al., *Overview of Race and Hispanic Origin: 2010*, U.S. CENSUS BUREAU (Mar. 2011), https://www.census.gov/content/dam/Census/library/publications/2011/dec/c2010br-02.pdf [https://perma.cc/G8ZJ-XTXK]; *Median Household Income in the United States*, FED. RSRV. BANK OF ST. LOUIS, https://fred.stlouisfed.org/series/MEHOINUSA646N [https://perma.cc/83L4-GMXB] (last updated Sept. 11, 2024, 9:45 AM).

54. *See supra* notes 51–53 and accompanying text (discussing 2010 Censuses of North Carolina and the U.S.).

55. *See supra* text accompanying note 50 (describing the method by which the sample credit scores were collected and collapsed together); *see also infra* note 60 (describing how Equifax data from 2010 was used to calculate the credit scores of 500,000 North Carolina consumers in March of 2010).

56. *Id.*

57. *Id.*

58. Jerad D. Bales et al., *North Carolina District Science Plan: Science Goals for 2003–2008*, U.S. GEOLOGICAL SURV., https://nc.water.usgs.gov/reports/ofr041025/report.html [https://perma.cc/CU9S-23JT] (last updated Dec. 6, 2016).

59. *See* VANTAGESCORE, *supra* note 47, at 7.

VantageScore to determine a borrower credit score.[60] Equifax helped create the VantageScore model, and reports customer credit attribute data, as well as VantageScore.[61]

Because Equifax and VantageScore track and categorize credit characteristics differently from one another, we construct categories of attributes using Equifax to approximate the categories of credit attributes used as inputs in the VantageScore model.[62] Furthermore, because observations in Equifax are available at the consumer-account level and VantageScore credit scores are consumer-level, we use an anonymized consumer identification number to collapse the Equifax data across accounts to the consumer-level to capture time-invariant characteristics of credit holders as of March 2010.[63]

The VantageScore categories and corresponding Equifax credit attributes we associate with the categories are:

- **Payment history** reflects repayment behavior over a consumer's entire history.[64] We measure payment history category with two constructed variables: the account satisfaction rate and the account past due rate.[65] The account satisfaction rate is the fraction of all accounts on credit file that is always paid as agreed (i.e., never delinquent).[66] The account past

60. *What Are the Benefits of Knowing Your VantageScore 3.0 Credit Score*, EQUIFAX, https://www.equifax.com/personal/education/credit/score/articles/-/learn/benefits-of-knowing-vantagescore/ [https://perma.cc/SH82-LJY4] (explaining that VantageScore uses underlying data from each credit bureau) (last visited Feb. 13, 2025). More specifically, we use snapshot data from Equifax in March 2010 to examine the credit qualities used to generate VantageScore credit scores for the sample of approximately 500,000 North Carolina consumers.

61. Louis DeNicola, *What is a VantageScore Credit Score?*, EXPERIAN (June 25, 2024), https://www.experian.com/blogs/ask-experian/what-is-a-vantagescore-credit-score/ [https://perma.cc/YD38-S8VC] ("The three credit reporting agencies, Experian, TransUnion and Equifax, created VantageScore as an independently managed joint venture in 2006.").

62. *What is an Equifax Credit Report?*, CAPITALONE (Aug. 22, 2024), https://www.capitalone.com/learn-grow/money-management/equifax-credit-report [https://perma.cc/LX4R-XC4W] (describing how the analytic dataset from Equifax contains tradeline - and consumer-level characteristics); *see also* VANTAGESCORE, *supra* note 47, at 1–7 (describing how those characteristics are used as inputs for the credit attribute categories used by VantageScore). For example, the Equifax data provides tradeline current and delinquent balances, while VantageScore uses a credit attribute "Balances" that is described as the total amount of recently reported current and delinquent balances. Both the current balance and delinquent balance data from a tradeline in the Equifax data can be used together as the balances credit attribute in VantageScore.

63. *Id.* (explaining that Vantage score and credit agency reports, like Equifax, are for individual consumers). Most of the characteristics are at the consumer level rather than the tradeline account level. For the characteristics in our analysis that are at tradeline we took the sum across all accounts for the total past due amount, total balance, and high variables. We took the maximum and minimum, and mean of account age across all accounts for the age of oldest account, age of newest account, and average age of account variables respectively. Finally, for the account mix, we looked across all accounts of a consumer to determine whether they had a revolving account, installment account, both, or neither. Descriptions of these variables are provided in the following paragraph.

64. *See* VANTAGESCORE, *supra* note 47, at 7 ("Payment history . . . Repayment behavior . . ."); *see also* Louis DeNicola, *What is a Good Credit Score*, EXPERIAN, https://www.experian.com/blogs/ask-experian/credit-education/score-basics/what-is-a-good-credit-score/ [https://perma.cc/4JZT-WGYE] (last visited Feb. 16, 2025) (describing how a more satisfactory repayment behavior yields a higher score and more delinquent or derogatory repayment leads to a lower score).

65. *See* DeNicola, *supra* note 61 ("[T]o improve your credit: Pay your bills on time. . . If you miss a debt payment by 30 days or more, it could damage your credit.").

66. *Id.* (describing the importance of payment history, thus requiring a special method to calculate in this case).

due rate is similarly calculated as the fraction of the total number of accounts with a past due amount greater than $0.[67]

- **Depth of credit** or "age and type credit," which collectively comprises 21% of the VantageScore.[68] Lengthier credit histories combined with a varied mix of credit types positively impact the score.[69] For measures that approximate history and depth of credit, we use the age of the oldest line of credit, the age of the newest line of credit, the average age of a consumer's credit lines, and the number of distinct types of credit lines held by a consumer.[70]
- **Credit utilization** reflects the total credit used relative to credit limit.[71] A lower ratio of current credit used or owed on an account to the credit limit on the account is rewarded with a higher VantageScore.[72] We proxy for the percent of credit limit used category with two measures from Equifax: the number of high utilization accounts and the number of high utilization bank card accounts.[73] While, ideally, one would calculate the percent of credit limited used by taking the ratio of the total tradeline balance to the total tradeline credit limit for a customer, Equifax provides a mix of high credit and credit limit values that cannot be used to universally calculate a total credit limit.[74] Instead, we use measures of the number of accounts with high percent of credit limit used.[75] The two metrics we use are the number of revolving accounts with a balance at least 50% of the credit limit and the number of open bankcard accounts with a balance at least 75% of credit.[76]

---

67. *Id.*

68. *See* VANTAGESCORE, *supra* note 47, at 7.

69. *See* DeNicola, *supra* note 61 ("[T]o improve your credit: . . . Keep old credit cards open. Keeping credit cards open gives you more available credit . . . . Use different types of credit. Having open revolving and installment accounts can give you a diverse credit mix and increase your scores.").

70. *See Why Your Credit Scores May Drop After Paying Off Debt*, EQUIFAX, https://www.equifax.com/personal/education/credit/score/articles/-/learn/why-credit-scores-may-drop-after-paying-off-debt/ [https://perma.cc/7G5V-DL4G] (listing length of credit history, newer lines of credit, and credit mix as factors in determining credit scores) (last visited Feb. 13, 2025). The two main types of credit considered by VantageScore are revolving and installment debt.

71. *See* VANTAGESCORE, *supra* note 47, at 7 ("% of credit limit used . . . . [is the] [p]roportion of credit amount used/owned on accounts.").

72. *See* DeNicola, *supra* note 61 ("[T]o improve your credit: . . . . Keep your credit card balances low. . . . A lower utilization rate is best.").

73. *Id.* (describing the importance of low utilization).

74. *The Only Source Representing Total Outstanding Credit*, EQUIFAX, www.equifax.com/resource/-/asset/product-sheet/creditmix-product-sheet/ [https://perma.cc/89X3-ETWQ] (last visited Feb. 16, 2025) ("CreditMix can provide greater visibility into your target markets. . . . Its benefits include: Determining market size and share of total outstanding credit behavior through detailed credit categories."); *see also* Holly D. Johnson, *What is High Credit on a Credit Report?*, BANKRATE (May 16, 2024), https://www.bankrate.com/personal-finance/credit/dispute-high-balance-on-credit-reports/ [https://perma.cc/3FFW-H2LL] (referring to "the highest monthly balance or highest amount of credit you have owed on a specific credit card account or loan during a particular period of time").

75. *See The Only Source Representing Total Outstanding Credit*, *supra* note 74 (describing how the Equifax data cannot be used to calculate a total credit limit, thus requiring a specialized method to calculate a total credit limit).

76. *Id.*

- **Balances** reflect the total amount owed on various credit accounts, referred to as balances.[77] Lower balances, both in current accounts and in delinquent accounts, positively impact the creditworthiness evaluation.[78] We use the total tradeline balance and the total amount past due across accounts on a consumer's credit file.[79]
- **Recent credit activity** reflects a customer's propensity to open new lines of credit.[80] Multiple new accounts in a short timeframe may be viewed as a risk factor and negatively influence the score.[81] To measure recent credit activity, we use the number of accounts on a consumer's credit file that were opened in the last twelve months, as well as credit inquiries in the last twelve months.[82]
- **Available credit** gauges the aggregate credit available to an individual.[83] To measure available credit, we use the sum of credit limit or high credit across all of a consumer's accounts.[84]

The Equifax data that we match to each credit attribute identified by VantageScore is summarized in [Figure 2] below.[85]

77. *See* VANTAGESCORE, *supra* note 47, at 7 ("Balances . . . [is the] [t]otal amount of recently reported balances (current and delinquent).").

78. *See* DeNicola, *supra* note 61 ("[T]o improve your credit: . . . . Keep your credit card balances low. . . . Avoid taking on too much debt.").

79. *Id.* (describing the importance of balances on credit scores).

80. *See* VANTAGESCORE, *supra* note 47, at 7 ("Recent credit . . . [is the] [n]umber of recently opened credit accounts and credit inquires.").

81. *See* DeNicola, *supra* note 61 ("[T]o improve your credit: . . . . Avoid frequent credit applications. Apply for and opening new credit accounts can hurt your credit.").

82. *See infra* Figure 2 (measuring recent credit activity).

83. *Id.* (factors such as the number of accounts and changes in credit limits influence this category, with more available credit overall positively impacting the score).

84. *Id.* (showing the results of using the sum of credit limits).

85. *Id.* (displaying the VantageScore summaries).

Figure 2: VantageScore 3.0 Credit Attributes Matched to Equifax Data[86]

VantageScore Attributes	Matched Equifax Data	Equifax Description
Payment History	Number of accounts	Total number of accounts on credit file regardless of whether account is open, closed, or inactive.
	Number of accounts that are always satisfactory	Number of accounts on credit file always paid as agreed (i.e., never delinquent).
	Number of all accounts that are past due	Number of accounts on credit file with any of the following ever occurring: charge off, foreclosure, bankruptcy, internal collection, defaulted student loan.
Depth of Credit (Age and Type of Credit)	Age of oldest account	The number of months since the oldest account was opened or established on the consumer's credit file.
	Average Age of account	Average number of months since an account on credit file was established.
	Age of newest account	Number of months since the newest account on credit file was established.
	Number of distinct credit line categories	Number of distinct product categories.
Credit Utilization	Number of high utilization accounts	Number of open revolving accounts on credit file updated within the last 3 months with balance greater than or equals to 50% of credit limit/high credit.
	Number of high utilization bank cards	Number of open bankcard accounts on credit file updated within the last 3 months with balance greater than or equals to 75% of credit limit/high credit.
Recent Credit	New Accounts in last 12 months	Number of new accounts opened in last 12 months.
	Credit inquiries in last 12 months	Number of inquiries by lenders that result from a consumer's application for credit within the last 12 months.

86. *See* VANTAGESCORE, *supra* note 47, at 7; *see also supra* Section II.A.3.

Balances	Total balance	The tradeline balance at the archive date.[87]
	Total past due amount	Sum of past due amount for all accounts (open & closed) on credit file.
Available Credit	High credit	The tradeline High Credit/Credit Limit, whichever is reported by the lender.

B. ZIP Code Racial Share Data

VantageScore is prohibited from and does not collect and report consumer race.[88] Therefore, we use the racial composition of each borrower's ZIP code as a proxy measure for consumer race.[89] VantageScore similarly uses Census ZIP code level racial shares in its own analysis of disparate impact.[90] We also consider race proxied by the racial composition of mortgage applicants, as measured by Home Mortgage Disclosure Act (HMDA) data.[91] Mortgage applicants are a subpopulation of the Census that may provide an alternative representation of the racial composition of the relevant population.[92] More specifically, because credit scores are prerequisites for mortgage applications, this metric excludes non-credit seeking individuals that may be present in the Census data.[93]

We use data from the 2010 Census via the National Historical Geographic Information System to estimate total racial composition, and HMDA mortgage data provided by the CFPB to estimate mortgage applicant racial composition.[94]

87. *See supra* Figure 2 (A tradeline is another name for an account listed on a consumer's credit reports).

88. Julia Kagan, *VantageScore: Meaning, Model, Components*, INVESTOPEDIA (June 14, 2023), https://www.investopedia.com/terms/v/vantagescore.asp [https://perma.cc/G4M2-TRR8].

89. Equal Credit Opportunity Act (Regulation B), 12 C.F.R. § 1002.6 (2024) (The Equal Credit Opportunity Act (ECOA) prohibits credit discrimination on race and eight additional criteria. These ECOA prohibitions are implemented in 12 C.F.R. Part 1002 (Regulation B). Regulation B includes a requirement that credit scores be "empirically derived" and not include race as a variable.); *See also* Equal Credit Opportunity Act (Regulation B), 12 C.F.R. § 1002.2 (2024) ("Effects test and disparate treatment: An empirically derived, demonstrably and statistically sound, credit scoring system may include age as a predictive factor. . . . Besides age, no other prohibited basis may be used as a variable.").

90. VANTAGESCORE, TESTING METHODOLOGIES FOR CREDIT SCORE MODELS TO IDENTIFY STATISTICAL BIAS TOWARD PROTECTED CLASSES 1, 3 (2014) ("To measure ethnicity, a consumer's ZIP Code also was appended to the file for look-up purposes based on the US Census Bureau's database.").

91. *See infra* Figure 3: North Carolina Summary Statistics, March 2010 [hereinafter Figure 3] (describing the racial composition of mortgage applicants).

92. Ben Horowitz et al., *Lender-Reported Reasons for Mortgage Denials Don't Explain Racial Disparities*, FED. RSRV. BANK OF MINNEAPOLIS (Jan. 18, 2024), https://www.minneapolisfed.org/article/2024/lender-reported-reasons-for-mortgage-denials-dont-explain-racial-disparities [https://perma.cc/BZ52-AUC4] ("[T]he reasons lenders give for denying mortgages to people of color differ from the reasons they give for denying mortgages to White applicants.").

93. Ben Luthi, *What Credit Score Is Needed for a Personal Loan?*, EXPERIAN (June 4, 2024), https://www.experian.com/blogs/ask-experian/what-credit-score-is-needed-for-a-personal-loan/ [https://perma.cc/V6RP-XECS] ("You generally need a credit score of 580 or higher to qualify for a personal loan.").

94. Steven Manson et al., *IPUMS NHGIS: Version 17.0*, IPUMS (2022), https://www.ipums.org/projects/ipums-nhgis/d050.V17.0 [https://perma.cc/D4AR-3PS7]; *see also Mortgage Data (HMDA)*, *supra* note 94.

We merge the Census and HMDA mortgage data to the credit characteristics data using the consumer's ZIP code.[95]

Figure 3 **Error! Reference source not found.Error! Reference source not found.**below summarizes consumer credit attribute data as well as the racial composition data in the population of North Carolina consumers.[96] The figure shows that on average, consumers in our sample had an average of about $63,700 of debt across lines of credit and held balances around $5,440 past due on their accounts on average.[97] Notably, Amount Past Due, Tradeline Balance, High Credit, and Number of Credit Inquiries in LTM have larger outlier values.[98] Because of the potential for outlier values in these variables to skew coefficient estimates, we winsorize these variables to the 99th percentile in the credit data in all our regression analyses.[99]

Figure 3: North Carolina Summary Statistics, March 2010[100]

	Mean	Standard Deviation	Minumum	Median	Maximum
A: VantageScore & Credit Characteristics					
VantageScore	690	106	301	707	839
Payment History Satisfaction Rate	0.85	0.23	0.00	1.00	1.00
Payment History Past-Due Rate	0.04	0.12	0.00	0.00	1.00
Amount Past Due (000s $)	5.4	77.6	0.0	0.0	29,259.8
Age of Newest Account (mos.)	22.5	31.8	0.0	13.0	666.0
Age of Oldest Account (mos.)	207.4	122.6	0.0	187.0	866.0
Average Age of Account (mos.)	72.6	65.1	0.0	54.6	710.0
Account Mix	1.5	0.5	0.0	1.0	2.0
Number of High Utilization Accounts	0.8	1.5	0.0	0.0	28.0
Number of High Utilization Bankcard Accounts	0.3	0.8	0.0	0.0	18.0
Tradeline Balance (000s $)	63.7	132.0	0.0	12.9	7,580.6
Number of New Accounts in LTM	0.9	1.3	0.0	0.0	35.0
Number of Credit Inquiries in LTM	1.9	2.7	0.0	1.0	66.0
High Credit (000s $)	91.4	157.6	0.0	32.6	8,645.0
B: Racial Composition					
White Population Proportion	0.67	0.20	0.00	0.72	0.99
Black Population Proportion	0.21	0.17	0.00	0.16	0.87
Hispanic Population Proportion	0.08	0.05	0.00	0.07	0.47
Asian Population Proportion	0.03	0.03	0.00	0.02	0.28
Other Population Proportion	0.06	0.07	0.00	0.05	0.84
HMDA White Proportion	0.80	0.16	0.16	0.85	1.00
HMDA Black Proportion	0.14	0.14	0.00	0.09	0.82
HMDA Hispanic Proportion	0.03	0.02	0.00	0.03	0.33
HMDA Asian Proportion	0.02	0.03	0.00	0.01	0.27
HMDA Other Proportion	0.01	0.04	0.00	0.01	0.78
Observations	497,949				

95. *See infra* Figure 3 (HMDA uses Census tracts instead of ZIP codes. To account for this, we use data of Census tract and ZIP code overlaps from Housing and Urban Development to calculate racial composition by taking a weighted average of racial compositions of all Census tracts within a ZIP code).

96. *Id.* (detailing consumer credit attribute and racial composition data).

97. *Id.* (lines of credit include categories such as mortgages, auto loans, student loans, and bankcard accounts).

98. *Id.* (showing large outlier values for past due, tradeline balance, high credit, and credit inquiries).

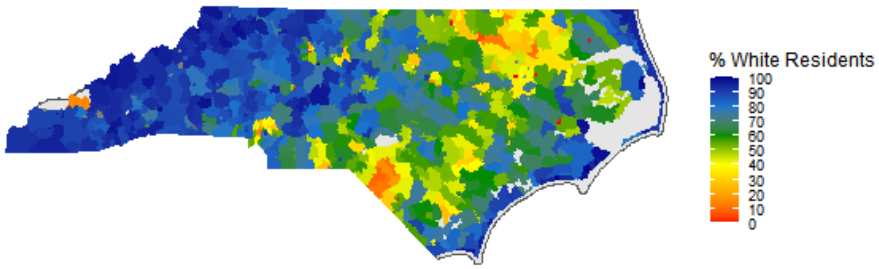
99. *Id.* (showing the winsorized data).

100. *See* VANTAGESCORE, *supra* note 47, at 7.

Notes: Hispanic includes both white and non-white Hispanics.[101] White is defined as the white non-Hispanic population in the data.[102]

Figure 4 below presents North Carolina's ZIP codes and their population racial composition using Census data, for March 2010.[103] The population racial composition map shows higher White population shares in the **Western half** of the state and along the **East coast**.[104] Lower White population shares are seen in the **inland Eastern half** of the state.[105]

Figure 4: White Population Share in North Carolina by ZIP Code[106]



Note: Gray area is water.[107]

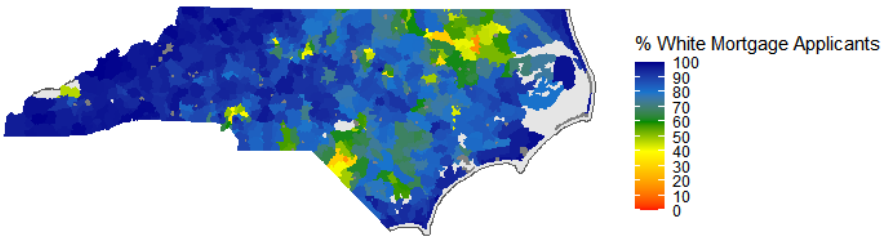
Figure 5 below shows the percentage of White mortgage applicants across North Carolina by ZIP code, for March 2010.[108] By mapping mortgage applicants, this allows us to examine geographic variation in racial shares among a population that may be likelier to represent the subpopulation of consumer credit score holders.[109] A visual comparison of Figure 4 and Figure 5: indicates that the White share of mortgage applicants is substantially higher than the share

101. *See supra* Figure 3 (including both white and non-white Hispanics).
102. *Id.* (providing a definition of white citizens).
103. *See infra* Figure 4: White Population Share in North Carolina by ZIP Code (presenting North Carolina ZIP codes and racial composition).
104. *Id.* (showing higher White population shares in the Western half of the state and along the East coast).
105. *Id.* (showing lower White population shares in the inland Eastern half of the state).
106. *2010 Census of Population and Housing*, U.S. CENSUS BUREAU, https://www.census.gov/quickfacts/fact/table/NC/POP010220#POP010220 [https://perma.cc/8Z55-KGAT] (last visited Feb. 13, 2025).
107. *See supra* Figure 4: White Population Share in North Carolina by ZIP Code (describing the map).
108. *See infra* Figure 5: White Mortgage Applicant Share by ZIP Code (showing the percentage of White mortgage applicants across North Carolina by ZIP code, for March 2010).
109. Amalie Zinn & Liam Reynolds, *How Local Differences in Race and Place Affect Mortgage Lending*, URBAN INST. (Nov. 15, 2022), https://www.urban.org/urban-wire/how-local-differences-race-and-place-affect-mortgage-lending [https://perma.cc/4976-D6GR] ("To ensure Black and Latino borrowers in majority-white areas aren't overlooked, decisionmakers seeking to close the racial homeownership gaps should consult data to assess where and when to use place as a stand-in for race.").

of population in ZIP codes throughout the state.[110] Overall, we observe higher percentage of White Mortgage applicants in the **Western half** of the state and along the **East coast**, and we observe lower White population shares in the **inland Eastern half** of the state.[111]

Figure 5: White Mortgage Applicant Share by ZIP Code[112]

.



Note: Gray area is water.[113]

A visual comparison of Figure 1, Figure 4, and Figure 5: suggests a positive geographic correlation between higher White shares and higher credit score.[114] We observe higher average VantageScores and higher white population and white mortgage applicant shares in the **Western half** of the state and along the **East coast** compared to the **inland Eastern half.**[115] However, the visual correlation between race and credit score across ZIP codes may be an artifact of correlation of race and consumer credit attributes.[116] Furthermore, the visual correlation between race and credit score may be misleading if Census racial variation differs from subset of residents that are consumers (i.e., seek to open lines of credit lines).[117] In order to assess whether consumers' geographically proxied race impacts their credit scores, we condition the VantageScore credit

110. *See supra* Figures 4: White Population Share in North Carolina by ZIP Code & 5: White Mortgage Applicant Share by ZIP Code (comparing two maps reflecting a contrast in White population and White mortgage application rates).

111. *See infra* Figure 5: White Mortgage Applicant Share by ZIP Code (presenting a higher percentage of White mortgage applicants in the Western half and along the East Coast and lower rates of White mortgage applicants in the inland Eastern half of North Carolina).

112. *See Mortgage Data (HMDA)*, *supra* note 94.

113. *Water Bodies*, NC ONEMAP, https://www.nconemap.gov/datasets/nconemap::water-bodies-western-nc-local-res/about [https://perma.cc/GXW6-ZUHE] (last visited Feb. 17, 2025).

114. *See supra* Figures 1, 4: White Population Share in North Carolina by ZIP Code, & 5: White Mortgage Applicant Share by ZIP Code (explaining an observed correlation between White population shares and higher credit scores).

115. *Id.*

116. Taz George et al., *The Geography of Subprime Credit*, FED. RSRV. BANK CHI. (2019), https://www.chicagofed.org/publications/profitwise-news-and-views/2019/the-geography-of-subprime-credit [https://perma.cc/PTS3-JFXQ] ("On average, we also find a number of indicators associated with less economic opportunity in places with lower credit scores . . .").

117. Laura Blattner & Scott Nelson, *How Costly is Noise? Data and Disparities in Consumer Credit*, ARXIV (May 5, 2021), https://arxiv.org/pdf/2105.07554 [https://perma.cc/N8NW-PTSJ].

scores on credit attributes, and evaluate the use of racial share as a proxy for consumer race.[118]

We present summary statistics for credit characteristics by ZIP code in the lowest, middle, and highest deciles of White share of mortgage applicants in Figure 6.[119] This allows us to observe the range of credit characteristics based on the percentile ranked by the proportion of White mortgage applicants in a ZIP code. For the bottom 10 percentile of ZIP codes, the proportion of mortgage applicants that are considered White is 43%, while the proportion of mortgage applicants that are considered White in the highest 10th percentile of ZIP codes is 97%.[120] For each credit attribute we present the mean in the 10th percentile for ZIP code, the 45th–55th percentile of ZIP code, and the 90th percentile of ZIP code. Across the different ZIP code percentiles, the Payment History Past Due Rate, Amount Past Due, and LTM Credit Inquiries are the credit characteristics with the largest variation, with the average of the lowest 10 percent White ZIP codes being 133%, 52%, and 67% higher respectively compared to the average of the highest 10 percent White ZIP codes.[121]

118. *See* Fiscella & Fremont, *supra* note 38.

119. *See infra* Figure 6 North Carolina Summary Statistics By ZIP Code Percentile Ranked BY Ratio of White Mortgage Applicants [hereinafter Figure 6] (displaying a table showing North Carolina statistics by ZIP code percentiles ranked by ratio of White mortgage applicants among various values).

120. *See infra* Figure 6 (explaining section B of Figure 6 showing White proportions).

121. *See infra* Figure 6 (demonstrating the contrast of key financial data points between the lowest 10 percent of White ZIP codes compared to the highest 10 percent).

Figure 6: North Carolina Summary Statistics By ZIP Code Percentile Ranked BY Ratio of White Mortgage Applicants[122]

Grouped by Proportion of White Mortgage Applicants	Bottom 10% HMDA White	45th - 55th Percentile HMDA White	Top 10% HMDA White
A: VantageScore & Credit Characteristics			
VantageScore	648	697	708
Payment History Satisfaction Rate	0.77	0.86	0.88
Payment History Past-Due Rate	0.07	0.04	0.03
Amount Past Due (000s $)	7.80	6.24	5.14
Age of Newest Account (mos.)	22	22	24
Age of Oldest Account (mos.)	180	212	224
Average Age of Account (mos.)	64	74	79
Account Mix	1.5	1.5	1.4
Number of High Utilization Accounts	0.9	0.8	0.7
Number of High Utilization Bankcard Accounts	0.4	0.3	0.3
Tradeline Balance (000s $)	51	70	60
Number of New Accounts in LTM	0.9	0.9	0.8
Number of Credit Inquiries in LTM	2.5	1.9	1.5
High Credit (000s $)	71	99	89
B: Racial Composition	Bottom 10% ZIP Code	45th - 55th Percentile ZIP Code	Top 10% ZIP Code
HMDA White Proportion	0.43	0.85	0.97
HMDA Black Proportion	0.45	0.10	0.01
HMDA Hispanic Proportion	0.06	0.03	0.01
HMDA Asian Proportion	0.03	0.02	0.00
HMDA Other Proportion	0.04	0.01	0.00
White Population Proportion	0.28	0.72	0.91
Black Population Proportion	0.53	0.17	0.03
Hispanic Population Proportion	0.12	0.08	0.04
Asian Population Proportion	0.03	0.02	0.00
Other Population Proportion	0.12	0.05	0.02
Observations	49,794	49,794	49,795

Notes: Hispanic includes both white and non-white Hispanics.[123] White is defined as the white non-Hispanic population in the data.[124]

C. Regression of Credit Score on Credit Attributes

We begin our statistical analysis by first showing that, for our illustrative example, an ordinary least squares ("OLS") regression using consumer-level credit characteristics as explanatory variables can be used to predict

122. *See Mortgage Data (HMDA)*, *supra* note 94.

123. *Updates to Race/Ethnicity Standards for Our Nation*, U.S. CENSUS BUREAU (Dec. 20, 2024), https://www.census.gov/about/our-research/race-ethnicity/standards-updates.html [https://perma.cc/B372-GERU] (explaining the U.S. Census designation of racial demographics, including Hispanics as a separate category).

124. *Id.*

VantageScore 3.0 credit scores with high explanatory power.[125] OLS regression is a tractable approach for this analysis due to its simplicity, interpretability, and accepted use in litigation.[126] While we do not assume that the relationship between credit attributes and VantageScore is inherently linear, OLS provides a straightforward method for estimating statistical associations and understanding the direction and relative magnitude of these effects.[127] Although actual relationships may involve non-linear dynamics, the OLS framework serves as a practical starting point for identifying patterns and detecting systematic biases within the model.[128] The coefficients, which reveal the marginal impact of each variable on the predicted score (holding other factors constant), are interpretable.[129] Furthermore, an OLS analysis allows us to evaluate whether race, directly or through proxies, disproportionately influences credit scoring outcomes, even after accounting for legitimate credit factors, thereby facilitating the identification of potential disparate impacts.[130]

We regress consumer credit scores on each of the credit attributes described as inputs (see Figure 2) in the credit scoring algorithm.[131] We use the entire sample of our consumer data for North Carolina for March 2010.[132] The coefficient estimates for each credit attribute are presented in Figure 7.[133] Each coefficient is statistically significant and has a sign that is consistent with how VantageScore describes the impact of the attributes on scores.[134] For example, payment history satisfactory rate has a large and positive coefficient estimate while payment history past due rate has a negative coefficient estimate.[135]

The overall adjusted R^2 for the regression is presented at the bottom of Figure 7 and has a value of 0.748, which indicates that the variation in the VantageScore credit attributes in the OLS model explain 74.8% of the variation

125. Although VantageScore 4.0 was launched in fall 2017, we use VantageScore 3.0 given its availability during the March 2010 period for which credit score and credit attribute data is available.

126. David H. Kaye & David A. Freedman, *Reference Guide on Statistics*, *in* REFERENCE MANUAL ON SCI. EVIDENCE 337 (3d ed. 2011) ("Most statisticians use the least squares regression technique because of its simplicity and its desirable statistical properties. As a result, it also is used frequently in legal proceedings.").

127. *Id.* at 342 ("Under appropriate assumptions, the least squares estimators provide 'best' determinations of the true underlying parameters. In fact, least squares has several desirable properties.").

128. We rely on the OLS adjusted R^2in evaluating whether our specification captures a reasonable relationship between the dependent variables described by VantageScore and predicted credit scores. Adjusted R^2 indicates the share of variation in the outcome variable of VantageScores that is explained by the included credit attribute variables and includes a penalty for unnecessary explanatory variables.

129. *See* Kaye & Freedman, *supra* note 126, at 227 ("Given independence, the correlation coefficient (infra Section V.B) between repeated measurements can be used as a measure of reliability.").

130. *Id.* at 260 ("Such [regression] models have been offered in court to prove disparate impact in discrimination cases, to estimate damages in antitrust actions, and for many other purposes.").

131. *See supra* Figure 2 (listing the credit attributes: payment history, depth of credit, credit utilization, recent credit, balances, and available credit).

132. *See Mortgage Data (HMDA)*, *supra* note 94.

133. *See infra* Figure 7: North Carolina Regression of Vantage Score Variables [hereinafter Figure 7] (displaying the coefficient estimates used for each credit attribute).

134. *VantageScore 4.0 Model Attributes Made Available for the First Time to Sophisticated Lenders Developing Custom Credit Scoring*, VANTAGESCORE (Nov. 14, 2024), https://www.vantagescore.com/press_releases/vantagescore-4-0-model-attributes-made-available-for-the-first-time-to-sophisticated-lenders-developing-custom-credit-scoring/ [https://perma.cc/E8PV-B99E].

135. *Id.* ("Credit score model attributes are model components reflective of a person's credit behavior or financial history. They help assess someone's likelihood of repaying a loan by considering factors like payment history, credit usage, and length of credit history.").

in the VantageScore values in the sample.[136] In the context of real-world data and complex social phenomena, this is typically considered high explanatory power, particularly when studying the outcome of black-box models such as credit scores, which are influenced by numerous observable and unobservable factors.[137] Specific to our results, the remaining 25% of unexplained variation is a result of the OLS linear model's inability to perfectly replicate the "black-box" algorithmic process from VantageScore.[138]

The second column presents the partial-R^2 of each set of VantageScore components.[139] These partial-R^2's generally correlate with the relative weights VantageScore details for the VantageScore components.[140] VantageScore describes their weights as the amount a credit attribute category contributes to a credit score, and the partial-R^2 is the proportion of variance explained in the credit score by the specific independent variables in a credit attribute category, while holding other independent variables constant.[141] Thus, one can interpret the partial-R^2 as comparable to the assigned weight.[142] The intuitively signed coefficient estimates, the correlation between partial-R^2 and the weighting, and the high overall R^2 all provide evidence that the OLS regression model here reasonably predicts the impact of each credit attribute on the VantageScore.[143]

136. *See infra* Figure 7 (describing a table showing a 0.748 value indicating the variation in the VantageScore values).

137. *See* Kaye & Freedman, *supra* note 126, at 316 ("In general, the more complete the explained relationship between the included explanatory variables and the dependent variable, the more precise the results.").

138. *See Infra* Figure 7; *See generally* Cynthia Rudin & Joanna Radin, *Why Are We Using Black Box Models in AI When We Don't Need To? A Lesson From an Explainable AI Competition*, HARV. DATA SCI. REV. (Nov. 22, 2019) https://hdsr.mitpress.mit.edu/pub/f9kuryi8/release/8 [https://perma.cc/ZB6U-68V4].

139. *See Infra* Figure 7.

140. *Id*.

141. *The Complete Guide to Your VantageScore*, VANTAGESCORE (Oct. 11, 2019), https://www.vantagescore.com/the-complete-guide-to-your-vantagescore/ [https://perma.cc/Q42R-3J78]; Julian Wang, *How to Calculate the R Squared Score for Each Individual Coefficient*, MEDIUM (Aug. 26, 2024), https://medium.com/@j.wang.mlds/how-to-calculate-the-r-squared-score-for-each-individual-coefficient-df805b6ed9ec [https://perma.cc/W959-ZDDZ].

142. *Id*.

143. *See Supra* notes 133–142 and accompanying text (discussing Figure 7's coefficient as statistically significant and detailing the correlation between partial-R^2 and the VantageScore weighting).

Figure 7: North Carolina Regression of Vantage Score Variables[144]

Factor		Coefficients	Partial R Squared	VantageScore Weighting
Intercept	Intercept	469.36***		
Payment History	Payment History Satisfaction Rate	227.78***	0.51	40%
	Payment History Past-Due Rate	-115.30***		
Depth of Credit	Age of Oldest Account (mos.)	0.17***	0.17	21%
	Age of Newest Account (mos.)	-0.11***		
	Average Age of Account (mos.)	0.14***		
	Account Mix	8.50***		
Utilization	Number of High Utilization Accounts	-12.25***	0.18	20%
	Number of High Utilization Bankcard Accounts	-11.81***		
Balances	Tradeline Balance (000s $)	-0.38***	0.09	11%
	Amount Past Due (000s $)	-0.87***		
Recent Credit	Number of New Accounts in LTM	-1.96***	0.05	5%
	Number of Inquiries in LTM	-4.93***		
Available Credit	High Credit (000s $)	0.32***	0.04	3%
Observations		497,949		
Adj R^2			0.748	

Notes: Hispanic includes both white and non-white Hispanics.[145] White is defined as the white non-Hispanic population in the data.[146] Standard errors are clustered at the ZIP code level. Due to the presence of large outlier values in Amount Past Due, Tradeline Balance High Credit, and Number of Credit Inquiries in LTM, we winsorize these four variables to the 99th percentile in the credit data.[147] The 99th percentile is 100.7 thousand dollars, 550.3 thousand dollars, and 677.3 thousand dollars respectively, 12 inquiries, respectively.[148]

Thus, while our OLS model for predicting VantageScore credit scores retains some unexplained variation, this does not diminish the importance of the explanatory variables included in the model, which VantageScore has identified as critical drivers of credit scores.[149]

144. *HMDA Maps*, FFIEC, https://ffiec.cfpb.gov/data-browser/maps/2023?geography=state [https://perma.cc/9WBR-5KTT] (last visited Feb. 16, 2025).

145. *Hispanic or Latino Origin*, U.S. CENSUS BUREAU, https://www.census.gov/acs/www/about/why-we-ask-each-question/ethnicity/ [https://perma.cc/3RY5-RH8G] (last visited Feb. 16, 2025) ("OMB defines "Hispanic or Latino" as a person of Cuban, Mexican, Puerto Rican, South or Central American, or other Spanish culture or origin *regardless of race*.") (emphasis added).

146. *New Estimates Highlight Differences in Growth Between the U.S. Hispanic and Non-Hispanic Populations*, U.S. CENSUS BUREAU (June 27, 2024), https://www.census.gov/newsroom/press-releases/2024/population-estimates-characteristics.html [https://perma.cc/K6N8-2RMQ] (describing the white population as the "non-Hispanic White population").

147. *See Supra* Figure 6; see also *Supra* Figure 7; s*ee also* Zach Bobbitt, *How to Winsorize Data: Definition & Examples*, STATOLOGY (Jan. 22, 2021), https://www.statology.org/winsorize/ [https://perma.cc/TVD9-8JRE] ("To winsorize data means to set extreme outliers equal to a specified percentile of the data.").

148. *Id*.

149. *Supra* note 147 (detailing the outliers to be winsorized); *The Complete Guide to Your VantageScore*, *supra* note 141 (enumerating the explanatory variables included in the model).

III. REGRESSION ANALYSIS WITH GEOGRAPHIC RACIAL PROXY VARIABLES

Next, we build on the baseline credit scoring model[150] by including racial shares of consumers' ZIP codes to test for whether the scoring practice disproportionately and adversely affects protected groups. We add explanatory variables for the racial makeup of the consumer's ZIP code to the OLS regression (see Figure 7) of VantageScore on credit attributes.[151] If the racial makeup of a consumer's ZIP code is found to be statistically significant, it could be reasonably considered to impact their VantageScore, separately from their credit characteristics.[152] A statistically significant coefficient on the racial makeup of a consumer's ZIP code therefore would suggest statistical evidence that in addition to the main credit attributes, the racial makeup of a ZIP code also correlates with the level of their VantageScore and suggests that there may be a structural relationship between race and credit score.[153]

We describe an approach to test whether race as proxied by population-level racial shares impact VantageScore credit score outcomes differently than race as proxied by racial shares of the credit consumer market.[154] To refine our sample of credit score holders to be more closely represented by the racial composition of the mortgage applicants, we also evaluate the regression using HMDA mortgage application data.[155] The HMDA data provides applicant racial demographics for consumers who applied for mortgages, a subset of the population with credit scores.[156] This approach allows us to examine whether evidence of discrimination differ when using the racial makeup of a ZIP code's overall population versus that of a credit-filtered applicant pool.[157]

---

150. *See Infra* notes 151–57 and accompanying text (detailing the methods to be used to determine how racial makeup impacts the model).

151. *See Supra* Figure 7; *see also infra* Figure 8: Credit Score Regression Results with ZIP Code Racial Shares [hereinafter Figure 8].

152. *See Infra* notes 162–63 and accompanying text (arguing that because the coefficient estimate of Black ZIP Share is statistically significant and negative, this indicates that black consumers have disproportionately lower credit scores). *See* Pritha Bhandari, *An Easy Introduction to Statistical Significance (With Examples)*, SCRIBBR (June 22, 2023), https://www.scribbr.com/statistics/statistical-significance/ [https://perma.cc/6399-NWR4] ("If a result is statistically significant, that means it's unlikely to be explained solely by chance or random factors. In other words, a statistically significant result has a very low chance of occurring if there were no true effect in a research study.").

153. *See infra* notes 162–63 and accompanying text (explaining that while a statistically significant coefficient on the racial makeup of a ZIP code may indicate a disparate impact, it does not necessarily imply intentional bias or a causal relationship. However, this analysis underscores the need for transparency and fairness reviews of proprietary or machine learning models. Without insight into the design and decision-making processes of such models, it is not feasible to attribute observed correlation between race and credit scores to an input with legitimate business justification).

154. *See Supra* notes 151–157 and accompanying text (explaining how adding explanatory variables for the racial makeup to the model can test whether race, in additional to main credit attributes, can impact the VantageScore).

155. *See supra* figure 3 and accompanying text ("We use data of Census tract and ZIP code overlaps from Housing and Urban Development to calculate racial composition"); *see also infra* Figure 8; Brian Beers, *Regression: Definition, Analysis, Calculation, and Example*, INVESTOPEDIA (July 31, 2024), https://www.investopedia.com/terms/r/regression.asp [https://perma.cc/K5E3-E2W9]. *See Data Browser API*, FFIEC, https://ffiec.cfpb.gov/documentation/api/data-browser/ [https://perma.cc/G7PC-FNRQ] (last visited Feb. 15, 2025) (allowing users to download HMDA data).

156. *See HMDA Maps*, *supra* note 144.

157. *Id.*

The regression results shown in Figure 8 below use the same specification as Figure 7, but with the racial shares of the consumers' ZIP code as explanatory variables.[158] The first two columns estimate the data in the full North Carolina sample of credit score holders, with the first column using population racial shares to proxy for race and the second using mortgage applicant racial shares to proxy for race.[159] In both columns, the coefficient estimate of Black ZIP Share is statistically significant and negative, while the coefficient for Hispanic ZIP Share is statistically insignificant.[160] This suggests that, for our example, there is some evidence of disproportionately lower in credit scores of Black consumers, while for Hispanics, Asians, and other racial or ethnic groups, we do not find any statistical evidence of disproportionately lower scores.[161] The magnitude of the Black ZIP Share coefficient estimate using Census Data varies from -0.11 to -0.22.[162] To place this result in context, using Census population shares and the results of column [1], this analysis suggests that a consumer at the 10th percentile of Black population share among consumers in the sample - corresponding to residing in a ZIP code where 3.5% of people are Black - moving to a ZIP code with the 90th percentile Black population share (45.2% Black population) would experience a credit score decline of 9.2 points or 0.087 standard deviations.[163] Similarly, when using mortgage applicant share (column [2]) and holding all other credit characteristics constant, increasing a consumer's Black mortgage applicant share from the 10th percentile value to that of the 90th percentile, a 33.6 percentage point increase, leads to a decrease in the predicted credit score by 8.7 points.[164] Furthermore, we also find that the adjusted R^2 does not increase in columns 1 and 2 of Figure 8 relative to Figure 7, indicating that the addition of geographic racial proxies does not meaningfully impact the goodness-of-fit of the linear regression model.[165]

158. *Id*.; see also *supra* Figure 7; *infra* Figure 8.

159. *See Infra* Figure 8.

160. *See id*. (noting, however, that given our large sample size, the statistical significance of the ZIP shares for these protected classes is not surprising even with the use of robust standard errors clustered at the ZIP code level. These standard errors correct for potential intra-ZIP correlation and heteroskedasticity. As the ZIP code is the level at which the racial proxies are constructed, this makes it a natural unit for clustering consumer-level observations).

161. *Id*.

162. *Id*.

163. Sarah Thomas, *Coefficient Regression: Definition, Formula, and Examples*, OUTLIER (May 13, 2023), https://articles.outlier.org/coefficient-regression [https://perma.cc/4VRQ-EJZL] (explaining coefficient regression); *see also* Courtney Taylor, *Statistics and How to Calculate It*, THOUGHTCO. (June 7, 2024), https://www.thoughtco.com/what-is-a-percentile-3126238 [https://perma.cc/Y2QJ-V8L6] ("The *n*th percentile of a set of data is the value at which *n* percent of the data is below it."); *See also* Marshall Hargrave, *Standard Deviation Formula and Uses vs. Variance*, INVESTOPEDIA (Aug. 5, 2024), https://www.investopedia.com/terms/s/standarddeviation.asp [https://perma.cc/C4XE-XQ86] ("Standard deviation is a statistical measurement that looks at how far individual points in a dataset are dispersed from the mean of that set.").

164. *Id*.

165. Jason Fernando, *R-Squared: Definition, Calculation, and Interpretation*, INVESTOPEDIA (Nov. 13, 2024), https://www.investopedia.com/terms/r/r-squared.asp [https://perma.cc/W46M-T3MF] ("R-squared (R^2) is defined as a number that tells you how well the independent variable(s) in a statistical model explains the variation in the dependent variable."); *see also Supra* Figure 7; *see also Infra* Figure 8.

The third and fourth columns restrict the sample to only credit score holders with a mortgage.[166] The fifth column uses racial shares of successful mortgage applicants and restricts the sample to credit score holders with a mortgage and matches the racial share of the consumers to the underlying sample of consumers in the regression.[167] In each of these columns, the Black share of the ZIP code is estimated to have a statistically significant effect, with credit scores dropping from 4.3 to 4.7 points if the consumer moves from the 10th percentile to 90th percentile of ZIP code based on Black (mortgage) population.[168]

While statistically significant, the economic magnitude of the drop in credit score for the Black consumer ranges from 0.04 to 0.09 standard deviations.[169] It is natural to ask whether this drop is economically significant. As mentioned earlier in the paper, credit scores tend to vary from 300 to 850.[170] In general, a decline of 9 points on a linear scale up to 850 may not be considered economically significant.[171] However, for lending purposes, credit scores around thresholds may also need to be considered.[172] Depending on the credit score, the consumer may qualify for a conventional mortgage loan, or be viewed as a subprime borrower.[173] This has implications for the borrowing interest rate charged to the consumer.[174] Additional analysis would need to be undertaken to determine whether the decline in credit score has economic implications to the borrower.[175]

It is also worth noting the consistently positive and statistically significant coefficients for the estimate of Asian ZIP share.[176] These coefficients indicate that a ZIP code having a higher Asian share is related to also having higher VantageScore.[177] The in-sample variation in Asian share is lower than that of

166. *See Infra* Figure 8.
167. *Id.*
168. *Id.*
169. *Id.*
170. *See* VANTAGESCORE, *supra* note 47, at 1.
171. *See generally id.* (presenting the wide range of possible credit scores).
172. *See* VANTAGESCORE, *supra* note 47, at 7 (identifying four credit tier ranges for VantageScore 3.0); *FAQs: VantageScore Credit Scores and the Mortgage Market*, VANTAGESCORE, 1, 1 (2018), https://www.vantagescore.com/wp-content/uploads/2022/02/FAQs-VantageScore-Credit-Scores-and-the-Mortgage-Market.pdf [perma.cc/H3RE-MB79] ("Both Fannie Mae and Freddie Mac use credit scores to determine product eligibility and pricing. . . . Eligibility is defined using a baseline score minimum, typically 620.").
173. *See* VANTAGESCORE, *supra* note 47, at 7 (labeling the lowest tier of credit scores as "subprime"); *see also Carol M. Kopp, What Is a Subprime Mortgage? Credit Scores, Interest Rates*, INVESTOPEDIA (Nov. 2, 2024), https://www.investopedia.com/terms/s/subprime_mortgage.asp [perma.cc/4V2A-JUML] (describing subprime mortgages).
174. Louis DeNicola, *Average Mortgage Rates by Credit Score*, EXPERIAN (Jan. 29, 2025), https://www.experian.com/blogs/ask-experian/average-mortgage-rates-by-credit-score/ [https://perma.cc/793A-ALHJ] (listing common mortgage interest rates for specific credit scores).
175. Steven Laufer & Andrew Paciorek, *The Effects of Mortgage Credit Availability: Evidence from Minimum Credit Score Lending Rules*, FIN. AND ECON. DISCUSSION SERIES 1, 8–9 (2016), https://www.federalreserve.gov/econresdata/feds/2016/files/2016098pap.pdf [https://perma.cc/WHH7-GRH8] (finding that over 2005 to 2012 between 20 to 70 percent as many mortgages were originated to borrowers just below minimum FICO credit score thresholds compared to those just above the threshold).
176. *See supra* text accompanying note 50; *see also supra* Figure 1; *See also infra* Figure 8.
177. *Id.*

Black and Hispanic shares, as the 90th percentile ZIP code in terms of Asian population share among consumers in the sample is 5.8%.[178] Moving from the 10th percentile (0.0%) to the 90th percentile ZIP code (5.8%) leads to an increase of 6.26 points.[179]

Overall, the regression results indicate that for this illustrative example in the setting of North Carolina in 2010 using a credit-population measure of racial share of ZIP code such as mortgage applicants leads to similar results as using the broader population racial shares.[180]

178. *Id.*
179. *Id.*
180. *See supra* text accompanying notes 160–79.

Figure 8: Credit Score Regression Results with ZIP Code Racial Shares[181]

	[1]	[2]	[3]	[4]	[5]
ZIP Black Share	-0.22***	-0.26***	-0.11***	-0.15***	-0.17***
	(0.01)	(0.02)	(0.01)	(0.02)	(0.02)
ZIP Hispanic Share	0.11***	0.06	0.02	0.04	0.07
	(0.04)	(0.10)	(0.03)	(0.10)	(0.10)
ZIP Asian Share	1.08***	0.99***	0.91***	0.80***	0.81***
	(0.13)	(0.12)	(0.15)	(0.13)	(0.14)
ZIP Other Share	-0.23***	-0.25***	-0.13***	-0.13***	-0.12***
	(0.03)	(0.03)	(0.04)	(0.05)	(0.04)
Payment History Satisfaction Rate	225.39***	225.58***	245.07***	245.18***	245.29***
	(1.03)	(1.03)	(1.44)	(1.44)	(1.44)
Payment History Past-Due Rate	-114.35***	-114.53***	-204.72***	-204.62***	-204.62***
	(1.70)	(1.69)	(4.17)	(4.17)	(4.17)
Amount Past Due (000s $)	-0.87***	-0.87***	-0.53***	-0.53***	-0.53***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Age of Oldest Account (mos.)	0.17***	0.17***	0.08***	0.08***	0.08***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Age of Newest Account (mos.)	-0.10***	-0.11***	-0.18***	-0.18***	-0.18***
	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)
Average Age of Account (mos.)	0.14***	0.14***	0.25***	0.25***	0.25***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Account Mix	8.85***	8.81***	16.27***	16.27***	16.27***
	(0.22)	(0.22)	(0.42)	(0.42)	(0.42)
Number of High Utilization Accounts	-12.07***	-12.09***	-13.56***	-13.57***	-13.58***
	(0.15)	(0.15)	(0.18)	(0.18)	(0.18)
Number of High Utilization Bankcard Accounts	-12.02***	-11.98***	-11.85***	-11.82***	-11.82***
	(0.17)	(0.17)	(0.24)	(0.24)	(0.24)
Tradeline Balance (000s $)	-0.38***	-0.38***	-0.26***	-0.26***	-0.26***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Number of New Accounts in LTM	-1.92***	-1.93***	1.33***	1.33***	1.33***
	(0.11)	(0.11)	(0.13)	(0.13)	(0.13)
Number of Credit Inquiries in LTM	-4.81***	-4.81***	-4.94***	-4.94***	-4.95***
	(0.06)	(0.06)	(0.07)	(0.07)	(0.07)
High Credit (000s $)	0.31***	0.32***	0.21***	0.21***	0.21***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Observations	497,949	497,949	169,757	169,757	169,757
Adj R^2	0.75	0.75	0.79	0.79	0.79
VantageScore Sample Restricted to Mortgage Holders?	No	No	Yes	Yes	Yes
Racial Composition Source	All Census	All HMDA	All Census	All HMDA	HMDA Mortgage Originations

Note: Hispanic includes both white and non-white Hispanics.[182] White is defined as the white non-Hispanic population in the data.[183] Robust standard errors are clustered at the ZIP code level. Due to the presence of large outlier values in Amount Past Due, Tradeline Balance High Credit, and Number of Credit Inquiries in LTM, we winsorize these four variables to the 99th percentile in the credit data.[184] The 99th percentile is 100.7 thousand

181. *See Supra* text accompanying note 50; *See also* Humes et al, *supra* note 53 (providing ZIP code data used as a racial heuristic).

182. *See generally QuickFacts*, U.S. Census Bureau, https://www.census.gov/quickfacts/fact/table/US/PST045224 [https://perma.cc/64SU-X6AE] (last visited Feb. 17, 2025) (listing Hispanic and non-Hispanic or Latino white people as separate demographics).

183. *Id.*

184. *See Supra* text accompanying note 50.

dollars, 550.3 thousand dollars, and 677.3 thousand dollars respectively, 12 inquiries, respectively.[185]

IV. RACE-NEUTRAL ANALYSIS

Next, we describe an approach to examine discrimination through a race neutral regression analysis.[186] This illustrative example seeks to analyze disparate impact by first re-estimating the baseline model (see Figure 7) in a demographically neutral environment.[187] Ideally, when the regression is estimated in a sample without racial variation, the coefficients are estimated with variation only in the credit attributes.[188] After estimating the regression in both the full sample and in a racially neutral subsample, the two regression models are then used to predict scores for samples restricted to specific protected classes.[189] If the predicted scores for protected class members are higher when using the race neutral model than the model estimated using the full sample, that indicates disparate impact negatively impacts the credit scores of protected classes.[190]

Because race data is unavailable for individual consumers, we proxy for race using the racial makeup of HMDA applicants in each ZIP code.[191] The average White mortgage applicant share in the sample is 79.7%, with averages of 13.8% and 3.3% for the Black and Hispanic shares.[192] To create a subsample that is near neutral in racial composition of White consumers, we restrict the data to ZIP codes that are in the 90th percentile White share or higher of HMDA applicants.[193] This results in ZIP codes with mortgage applicant pools that are at least 97.1% White.[194] While we would ideally estimate our regression in a subsample without any variation in the racial applicant shares, this 90th percentile threshold removes nearly all variation in the White mortgage applicant share while providing a large sample of 71 ZIP codes and 15,360 credit score consumers.[195]

When predicting credit scores in different subsamples of ZIP codes, we use the 90th percentile threshold and the regression model to predict VantageScores

185. *Id.*

186. *See generally* Avery et al., *supra* note 8 (utilizing the approach).

187. *Id.* (employing data linking scores and credit attributes to individuals' race data, which allows designation of subsamples specific to race. As we do not have race data for consumers, we use the top decile of ZIP codes in our data in terms of White percentage to approximate a subsample with near-neutral racial status).

188. *See generally id.* (attempting to control for racial attributes in some data analyses).

189. *See infra* Figure 9: Comparison of Mean Predicted Scores [hereinafter Figure 9] (showing predicted scores).

190. *See generally id.* (comparing predicted scores with race-neutral results).

191. *See supra* Figure 8 (showing that White racial share in HMDA is generally higher than the Census share, which better allows identification of a nearly-all White ZIP code in terms of HMDA consumers).

192. *See* Avery et al, *supra* note 8 (providing data); *see also infra* Figure 9 (comparing data).

193. *See* Avery et al, *supra* note 8 (providing data).

194. *See* Avery et al., *supra* note 8; *Download HMDA Data*, CONSUMER FIN. PROTECTION BUREAU, https://www.consumerfinance.gov/data-research/hmda/historic-data/ (last visited Feb. 12, 2025).

195. *Id.*

in separate subsamples of ZIP codes in the 90th percentiles of White, Black, Hispanic, and Asian, respectively.[196]

Figure 9 presents the different mean predicted scores when estimating either the baseline model or race neutral model in either the full sample or in subsamples in the 90th percentile of applicant racial shares for White, Black, Hispanic, and Asian.[197] The predicted scores are roughly similar for both the baseline and race neutral models in the ZIP codes highest in each racial group's applicant shares.[198] Using a model estimated with far less racial variation in the sample leads to similar predicted scores for consumers in the ZIP codes with the highest shares of Black and Hispanic consumers.[199] These results indicate that for this approach, credit characteristics in the baseline model are unlikely to proxy for racial shares for the data we examine under the given assumptions.[200]

Figure 9: Comparison of Mean Predicted Scores[201]

Sample	Observations	ZIP Codes	Baseline Model Predicted Score	Race Neutral Model Predicted Score
All ZIP Codes	497,949	730	690.2	691.2
90th Percentile White ZIP Codes	15,360	71	707.2	709.0
90th Percentile Black ZIP Codes	56,437	72	658.9	658.7
90th Percentile Hispanic ZIP Codes	93,827	73	667.8	668.0
90th Percentile Asian ZIP Codes	118,821	73	690.6	691.7

V. CONCLUSION

We match credit scores with data on consumers' credit histories and racial demographics at the ZIP code level to assess whether an algorithmic model may inadvertently use proxies for race in credit scoring.[202] We first show that an OLS regression of credit scores on known inputs to the credit model reliably reproduces most of the variation in credit scores, and then show that two alternative geographic proxies for race are also statistically significant.[203] Our results indicate that moving from a 10th percentile to 90th percentile ZIP code by share of the Black population would lead to a statistically significant decrease in credit scoring.[204] In practice, a significant correlation between race and outcomes in a high-stake setting demands attention to ensure equity and mitigate

196. *Id.*

197. *See generally* Kaye & Freedman, *supra* note 126, at 211 (explaining OLS regression analysis); *See also Infra* Figure 9.

198. *See infra* Part IV (providing the papers conclusion on race neutral analysis and consumer credit scoring).

199. *Id.*

200. *Id.*

201. *See Download HMDA Data*, *supra* note 194.

202. *See supra* Part IV (providing the papers conclusion on race neutral analysis and consumer credit scoring).

203. *See supra* Parts III–IV (describing race natural regression model and outlining conclusions.).

204. *See supra* Part IV (providing the papers conclusion on race neutral analysis and consumer credit scoring).

disparate impacts.[205] In our analysis, the impact of race as proxied by ZIP code on consumer credit scores is not large within the set of credit scores that we considered,[206] however more analysis would be needed to fully evaluate the economic effects.

We also present a model estimated using a race-neutral environment. For our data, time period, and assumptions, this approach does not lead to relatively more favorable scores for consumers in areas with higher shares of protected classes.[207]

The analysis described in this paper provides an illustrative example of an approach to assessing disparate impact when a decision-making algorithm is a "black box," and the researcher has reasonable proxy measures of protected classes for the data.[208] More specifically, we use two representative proxies for race to illustrate how an OLS model can be used to examine whether any of the input variables act as a proxy for a protected class.[209] In this example, similar results are produced when using either a general population-level measure of racial shares from the Census or a measure of racial shares that shares characteristics (having a credit score) with the underlying consumer population impacted by the credit-scoring algorithm.[210]

With the growth of AI in consumer finance, there is growing risk of disparate impact for credit consumers because of tools such as AI-driven underwriting.[211] Regulators are increasingly concerned by this risk.[212] This paper provides one method of testing for disparate impact with two key advantages for this trend.[213] First, it aligns with regulatory requirements that ensure new AI tools are generally prohibited from collecting or using data related to race of the consumers.[214] Second, our method accommodates the

205. *See* Rice & Swesnik, *supra* note 8, at 952–53; *see* generally, *Title VI Legal Manual*, *supra* note 9 (discussing how to prove claims of disparate impacts discrimination).

206. *See Supra* Part III (describing race natural regression model and outlining conclusions.).

207. *See Supra* Part IV (providing the papers conclusion on race neutral analysis and consumer credit scoring).

208. *See Supra* Part II.B, Part III (outlining research methodology)

209. *See Supra* Part III. (describing race natural regression model and outlining conclusions.).

210. *Id.*

211. Olga Mack, *Promoting AI Fairness: The Application of Disparate Impact Theory*, COMPUTATIONAL L. (Aug. 31, 2023, 2:29 PM), https://law.mit.edu/pub/promoting-ai-fairness-disparate-impact-theory/release/1 [https://perma.cc/25L7-MHYR] ("The use of AI in decision-making raises concerns about bias and fairness."); *See* generally El Bachir Boukherouaa et al., *Powering the Digital Economy: Opportunities and Risks of Artificial Intelligence in Finance*, IMF ELIBRARY (Oct. 22, 2021), https://www.elibrary.imf.org/view/journals/087/2021/024/article-A001-en.xml [https://perma.cc/KY5T-4MGH] (discussing "the impact of the rapid adoption of artificial intelligence (AI) and machine learning (ML) in the financial sector").

212. *Id.*; *See also* Akinwumi et al., *supra* note 2 ("In light of the growing adoption of AI/ML, federal regulators—including the Consumer Financial Protection Bureau (CFPB), Federal Trade Commission (FTC), the Department of Housing and Urban Development (HUD), Office of the Comptroller of the Currency (OCC), Board of Governors of the Federal Reserve (Federal Reserve), Federal Deposit Insurance Corporation (FDIC), and National Credit Union Administration (NCUA)—have been evaluating how existing laws, regulations, and guidance should be updated to account for the advent of AI in consumer finance.")

213. *See Supra* Part I, II (outlining disparate impact in credit scoring and providing case study of VantageScore)

214. Adi Robertson, *FTC Warns It Could Crack Down on Biased AI*, THE VERGE (Apr. 20, 2021, 11:35 AM), https://www.theverge.com/2021/4/20/22393873/ftc-ai-machine-learning-race-gender-bias-legal-violation

opaque nature of many AI models, allowing for disparate testing without information on the underpinnings of the AI model.[215] Our approach only requires a reasonable approximation of the economic factors driving the model, making it a practical solution for disparate impact in AI-driven credit decisions.[216]

It is important to acknowledge the limitations of this study. Primarily, the case study presented is limited to the state of North Carolina during the specific month of March 2010.[217] Consequently, the findings may not be readily applicable to other states or different timeframes.[218] Additionally, our analysis focuses solely on evaluating the potential disparate impact among Black, Hispanic, and Asian consumers, neglecting other demographic groups that may be affected differently.[219] Thus, while our study provides valuable insights into the dynamics of consumer behavior and potential disparities, its scope remains constrained to a particular region and demographic subset, warranting caution in generalizing the findings to broader contexts.[220] This paper presents an exercise to demonstrate an approach[221] and is not intended to prove or disprove disparate impact of credit scores.

[https://perma.cc/9Z9L-NHAY] ("[C]ompanies could be prosecuted under the Equal Credit Opportunity Act or the Fair Credit Reporting Act for biased and unfair AI-powered decisions, and unfair and deceptive practices could also fall under Section 5 of the FTC Act."); *See supra* Part II (discussing how VantageScore does not collect or report consumer race data, aligning with regulatory requirements that generally prohibit the use of race-related data in AI tools.)

215. *See Supra* Part II., III (describing race natural regression model and outlining conclusions.).
216. *See Supra* Part II.C, III. (describing race natural regression model and outlining conclusions.).
217. *See Supra* Part II.A. (outlining background of VantageScore data).
218. *Id.*
219. *Id.*
220. *Id.*
221. *Id.*