Chapter 10

Algorithmic Bias in Traditional Credit Scoring Models

Jason Dietrich
March 2026

I. Introduction

The financial industry and its regulators have recently shown significant and growing interest in the use of machine learning (ML) classifiers to build scoring models for credit decision-making. Since ML classifiers excel at identifying small but meaningful correlations and patterns in data, one interesting implication of the adoption of these classifiers has been an increase in the number of variables in scoring models. Historically, scoring models built using traditional econometric estimators included a small number of variables, often less than 20. It is not unusual for models built using ML classifiers to have hundreds, if not thousands, of variables, especially if we count interactions. One potentially significant benefit of transitioning to ML scoring models with larger numbers of variables is improved accuracy, which could increase access to credit for some consumers, especially marginal consumers. Improved accuracy could also impact pricing, although the direction of this impact would likely vary across consumers. At the same time, this transition to ML classifiers also raises several usage and compliance challenges that regulators and industry need to consider. Requiring data on a larger number of variables to score applicants increases data integrity challenges when lenders use these scoring models. In addition, identifying the main drivers of scores for Adverse Action Notice purposes can also be more challenging, since with larger numbers of variables each individual variable typically has only a small effect on overall model predictiveness.

In addition to these usage and compliance challenges, larger numbers of variables in scoring models also have two important fair lending implications. First, including more variables in scoring models likely reduces variation in prediction errors across demographic groups. Anecdotal evidence from supervisory exams suggests that scoring models built using traditional econometric estimators and with smaller numbers of variables often benefit Black and Hispanic borrowers by under-predicting their true default or delinquency risk, and often harm Asian and

White borrowers by over-predicting their true default or delinquency risk. Recent research by Robb and Robinson (2018), Aggarwal et al. (2022), and Hebert-Johnson et al. (2018) provide evidence consistent with these findings. Expanding the number of variables included in scoring models will generally reduce the prediction errors for every demographic group and therefore reduce the variation in prediction errors across groups. Second, adding more variables to scoring model creates the risk that combinations of these variables will become proxies for prohibited bases groups. As a result, model developers may end up indirectly including prohibited bases terms into scoring models, which would create a significant concern of disparate impact or possibly disparate treatment.

In this report we demonstrate how both of these fair lending implications can plausibly occur, and how they tend to occur, with traditional econometric approaches to model development. We first develop an analytical framework showing how traditional credit scoring models with smaller numbers of variables tend to under-predict risk for some demographic groups and over-predict risk for others, and that including additional variables increases the likelihood that those variables in combination become proxies for race.[1] Throughout the analytical framework we define bias in a scoring model specifically as the difference between the predicted bad rate and actual bad rate for a given demographic group. To develop our main analytical findings, we leverage off of several key aspects of typical development data, standard model development processes, and logistic estimators including,

- The underlying premise behind credit scoring models is that past performance accurately predicts future performance.

[1] There are a variety of purposes of scoring models, such as to predict attrition, response, fraud, productivity, and revenue, among others, and a variety of applications of scoring models, such as in credit markets, for health care, in education, by courts, and in labor markets, among others. Throughout this report we focus specifically on credit scoring models used to predict credit risk, which lenders use to make underwriting and pricing decisions on credit applications. We focus specifically on credit decisions on mortgage applications, but the concepts and results apply to underwriting and pricing decisions on other credit product more broadly.

- Some minority groups (primarily Black borrowers) often have higher bad rates (worse performance on past loans) than Asian and White borrowers.
- Model development is typically a sequential process beginning with a parsimonious model and then adding variables to improve predictiveness.
- The logistic estimator is well-calibrated and preserves marginal probability.
- There is typically a stopping criterion limiting the number of variables in a scoring model.

After developing this analytical framework, we then use data from the National Mortgage Database (NMDB) to build credit scoring models for a variety of products to explore how treatment of borrowers from different demographic groups changes as we include additional variables to the models. This empirical evidence using real-world data provides additional support to the plausibility and likelihood of each of the two fair lending implications. To be clear, our goal is not to show that every traditional credit scoring model with smaller numbers of variables under-predicts risk for Black borrowers and over-predicts risk for Asian and White borrowers in every instance, or that adding additional variables to the model will always result in combinations of those variables proxying for race. Our goals are only to show how both results can occur, and that both results are plausible and tend to occur under traditional model development approaches.

The main takeaways in this report are:

- An analytical framework based on standard credit score development processes, as well as empirical evidence based on the NMDB, suggest that traditional scoring models with smaller numbers of variables tend to under-predict risk for Black borrowers and over-predict risk for Asian and White borrowers. This finding provides a general benchmark for comparison when assessing bias in ML scoring models, which is important for understanding how transitioning from traditional scoring models to ML scoring models might impact access to credit and pricing for different demographic groups.

- When there is variation across demographic groups in the bad rates in the data used to develop a scoring model, including more variables to the model naturally leads to a risk of combinations of those variables becoming ever stronger proxies for race. This finding has important implications for disparate impact analyses of scoring models as the transition to ML scoring models with increased numbers of variables may increase the likelihood that combinations of these variables proxy for race.

- Variation across demographic groups in the bad rates in the data used to develop scoring models is a main driver of differences in average score values across demographic groups and also impacts the bias in scoring models across demographic groups. This highlights one potential benefit of using alt data as a source of information on accounts, such as rental payments or utility payments, which may have less variation in performance measures across demographic groups.

II. Relevant Literature

There is a relatively recent stream of theoretical and empirical literature focused on prediction errors across demographic groups in scoring models. On the theoretical side, Hebert-Johnson et al. (2018) provides a theoretical foundation for how credit scoring models can under- and over-predict risk across demographic groups, including several theorems and proofs about the properties of these scoring models. Kleinberg and Mullainathan (2018) show that simplistic models can breed unfairness and that, by increasing accuracy, more flexible algorithms can actually reduce demographic biases. Finally, Ramabachan et al. (2020) discusses the concept of reverse bias in lending markets. If lenders discriminate against Black and Hispanic mortgage applicants by applying a higher underwriting standard, Black and Hispanic consumers with loans would have lower credit risks and be less likely to default, on average. Therefore, bias in underwriting decisions could lead to less bias in credit scoring models.

In addition to these theoretical findings, a small number of recent papers have shown empirical evidence of variation in prediction errors in scoring models across demographic groups. Aggarwal et al. (2022) shows that Asian patients have higher incidence of diabetes than

White patients with comparable age and BMI. As a result, race-blind models systematically under-estimate risk for Asian patients and systematically over-estimate risk for White patients. In the small business lending space, Robb and Robinson (2018) finds that credit scoring models under-predict the rate of payment delinquency among minority-owned businesses. Fuster et al. (2020) shows that transitioning from traditional to ML scoring models disproportionately benefits non-Hispanic White borrowers relative to Black and Hispanic borrowers. These results focus on the impact of different estimators/classifiers, but the results are consistent with the results in this report which stem from the number of variables included in scoring models. Rothblun and Yona (2022) focuses on potential threshold-based solutions when scoring models yield systematically different predictions for different demographic groups. Finally, Obermeyer et al. (2019) finds that conditional on a health risk score, Black borrowers have higher actual illness rates than White borrowers, suggesting that the scoring model under-predicts risk for Black borrowers. One caveat for this research is that the choice of the dependent variable used to develop the scoring model drives the under-prediction of health risk for Black borrowers, which is a slightly different focus than in our report.

To a large extent, these empirical papers do not focus specifically on variation in prediction errors across demographic groups, but instead show evidence of this variation as one result of research focused on different questions. Our report therefore contributes to this empirical literature in three ways by: 1) providing a specific mechanism for how variation in prediction errors across demographic groups can operationally occur for credit scoring models, 2) demonstrating the inverse relationship between variation in prediction errors across demographic groups and the likelihood that combinations of variables in scoring models proxy for race, and 3) generating an additional set of empirical evidence of these prediction errors and tradeoffs using real-world data on credit markets.

In addition to the above research, our report is also related to several streams of literature in both computer science and economics. There is an extensive literature on algorithmic fairness and bias in computer science dating back to Friedman and Nissenbaum (1996), which was one of the first comprehensive assessments of bias in computer systems. Mitchell et al. (2018), Vanhoof et al. (2018), and Caliskan et al. (2017) summarize various sources of algorithmic bias. Zliobaite (2017), Corbet-Davies et al. (2018), Mitchell et al. (2018), Barocas et al. (2018), and Liu et al. (2019) discuss possible definitions of, and tests for, discrimination and fairness, and a related literature argues that it is not possible to simultaneously achieve multiple fairness definitions (see Chouldechova (2017) and Barocas et al. (2017)). A recently emerging area of research has begun to explore approaches to reducing algorithmic bias. That research explores the potential impact of imposing fairness criteria on decisions (see Hardt et al. (2016)) and the effectiveness of debiasing techniques (see Cowgill (2018)). Romei and Ruggieri (2013) adds to this work by providing a summary of possible solutions to algorithmic bias across different disciplines (law, economics, statistics, and computer science). Specific to mortgage markets, which is the focus of our report, Fuster et al. (2020) and Bartlett et al. (2019) analyze algorithmic bias in mortgage markets in the U.S. using HMDA data, and Bono et al. (2021) analyzes algorithmic bias in mortgage markets in the U.K.

The economics literature on algorithmic bias is not as extensive as the computer science literature, but it has shown significant growth in recent years. Cowgill and Tucker (2019) summarizes the theoretical and empirical economics research on algorithmic bias. The authors include a summary of definitions of bias in economics, economic theories of bias, policy options for reducing bias, four general sources of algorithmic bias (unrepresentative development samples, mis-labeling outcomes in the development data, feedback loops, and modeler bias), behavioral economics considerations, and policy implications. There is also a growing line of

economic research focused on fairness definitions (statistical parity, equal odds, predictive parity, and overall accuracy, (see for example Hurlin et al. (2022)).

III. Analytical Framework

The analytical framework we use to develop credit scoring models and conduct the empirical analyses below consists of three primary components: 1) bias source and measure, 2) development of credit scoring models, and 3) definition of race proxies. We discuss each of these in turn.

Bias Source and Measure

People often cite algorithmic bias as a reason for differences in average credit scores across demographic groups. Behind this general statement is a significant amount of nuance and complexity since algorithmic bias encompasses several specific sources and potential measures of bias. The first component of our analytical framework details the specific source of bias we focus on and the specific approach we use to measure bias in credit scoring models. Focusing on different sources and measures of bias may lead to different results and conclusions.

In general, there are three primary sources of bias in credit scoring models: data, modeler decisions, and the model development process. For this report, we focus only on bias that the model development process creates. By focusing on just one general source of bias, we are implicitly assuming that the modeler injects no bias into the scoring model and that all results and findings are conditional on the available data. For the specific measure of bias, we considered two options that were particularly applicable to the model development process. The first option is based on the Becker test, which states that an economic decision-making process is biased if it produces unequal productivity across groups at the margin. In a credit scoring

scenario, this approach compares the loan performance of marginal applicants from different demographic groups. In real-world scenarios, the score threshold identifying marginal applicants will be specific to the lender, the scoring model, and the particular application of the scoring model. Since lenders know their appetite for risk and choose score thresholds accordingly, the Becker test is well-suited for measuring bias in credit scoring models in real-world scenarios. It is more difficult to apply this approach here, however, since it is not clear which threshold to use and the results will likely vary by threshold. As a result, it would be difficult to generalize any findings. The second option is to segment the score distribution (into deciles for example) and compare, separately for each demographic group, predicted bad rates from the scoring model by segment to actual bad rates in the development dataset by segment. For a given demographic group, predicted bad rates that are systematically lower (higher) than actual bad rates across score segments suggests the scoring model under-predicts (over-predicts) risk for that group.[2] [3] For this report, we use a slightly simplified version of this second option, which is similar to the concept of a biased predictor from Hebert-Johnson (2018).[4] Instead of comparing the distributions of predicted bad rates and actual bad rates for each demographic group, we compare the overall predicted bad rate to the overall actual bad rate for each group. Although we lose useful information about variation in bad rates across the score distribution by focusing on overall bad rates instead of distributions of bad rates, this approach is easier to convey and

[2] This approach is similar to the sufficiency fairness criteria in Bono et al. (2021) and the calibration fairness criteria in Hebert-Johnson (2018).

[3] Another common and related fairness measure found in the ML literature is separation, which tests whether the score distribution conditional on the outcome is the same for every subgroup. If separation is achieved, then for any cutoff used in a lending decision, the true positive and false positive rates for each subgroup will be the same (see Bono et al. (2021)).

[4] This measure of fairness is also called classification parity and requires that error rates are equal across demographic groups (see Hurlin (2022) and Hardt et al. (2016)).

explain both graphically and analytically without losing the substance needed to demonstrate our two fair lending implications of interest.

Development of Credit Scoring Models

The second component of the analytical framework is the process to develop a credit scoring model. The main premise underlying the development of credit scoring models is that past performance on loans accurately predicts performance on future loans. Therefore, the first step of building a credit scoring model is to obtain a development dataset containing a large number of originated loans from the past; a measure of performance on these loans, specifically whether each loan had "good" or "bad" performance; and a large number of variables that might be predictive of performance. Since denied applications have no performance data, modelers often use a reject inference approach to infer performance for these applications to avoid excluding them from the development dataset. The next step is to use an econometric estimator or ML classifier to identify the variables that are most predictive of bad performance, and to quantify the relationships between these variables and the likelihood of bad performance subject to several conditions, such as ensuring expected relationships between the variables and the likelihood of a bad outcome, fair lending concerns, and acceptable complexity, among others. An important part of this step is to transform all potentially predictive variables into sets of 0/1 flags in such a way as to optimize their predictiveness of bad performance.[5] Once modelers have finalized the scoring model, they typically calibrate the model so that a given increase in score

[5] For example, three 0/1 flags indicating debt-to-income (DTI) values less than 36%, values between 36% and 50%, and values greater than 50%, may best capture non-linear relationships between DTI and the likelihood of bad performance on a loan.

results in a specific reduction in the risk of default or delinquency. A common last step is to apply various modifications to the scoring model so that it meets all relevant business objectives.

For the purposes of this report, we would ideally want to develop credit scoring models using a comprehensive and detailed approach similar to what model developers use, and then explore the two fair lending implications of interest for models with different numbers of variables. That level of granularity is beyond the scope of this report, however. Instead, we incorporate just the general aspects of the development process most relevant to our purposes and exclude the more nuanced components that likely have less impact on our objectives. For example, similar to actual credit scoring models, prior to estimation, we transform all potentially predictive continuous variables into sets of 0/1 flags. Unlike actual credit scoring models, we use a general approach to construct based on deciles, quartiles, and medians, and not an approach that optimizes their predictiveness of performance. As a second example, we focus on the predicted probabilities of loans being bad directly from the logistic models and do not calibrate these probabilities so that a given increase in score results in a specific reduction in the risk of default or delinquency. Since calibration is just a monotonic transformation of predicted probabilities, the predicted probabilities from our models rank-order risk similar to actual credit scoring models, just not in the same specific way. Therefore, although we do not meet the specific, rigorous standards model developers use when building actual scoring models, we do incorporate the general aspects of credit scoring models that are most important for our purposes. Exploring whether a more rigorous and detailed model development approach affects the results presented in this report is one area for future research.

As noted above, we use a logistic estimator to estimate all models. In addition, to make the results nationally representative, we utilize the loan-level sampling weight in the NMDB for all regressions. The logistic estimator is a good estimator to use for several reasons. First, it is a

standard estimator that developers of scoring models have used for many years. Second, since it is both an econometric estimator to Economists and an ML classifier to Data Scientists, it provides a nice bridge between traditional scoring models and ML scoring models. Finally, it has two properties that are instrumental in analyzing the two fair lending implications of interest in this report, it is well-calibrated and it preserves marginal probability. Well-calibrated means it produces predictions that are, on aggregate, closely aligned with actual outcomes. Specifically, the average of the predicted probabilities of loans being bad will equal the overall bad rate in the development dataset. Preserving marginal probability means that, for each binary variable in the model, if we subset the data to just the observations where a given binary variable is on (i.e., equal to 1), the sum of the predicted probabilities will equal the sum of the outcome variable (i.e. the 0/1 flag indicating bad performance). Extending this slightly, for a model including a 0/1 flag for a demographic group, or a combination of variables that proxy for that demographic group, the average of the predicted probabilities of loans being bad from the model for that group will equal the bad rate for that group in the development dataset. Appendix A provides additional details and further explanation of both of these properties.

For the modeling process, we follow a sequential approach that is standard for developing credit scoring models. The initial model includes only a constant and no other variables. In this parsimonious scoring model, every application receives the same score, which is equal to the overall bad rate in the development dataset since the logistic estimator is well-calibrated. If no other information is available, the overall bad rate in the development dataset is a good estimator to use since it is an unbiased estimator of the true population bad rate. It may not be the best unbiased estimator in terms of minimum variance, however. As we will show graphically below, for demographic groups with bad rates greater than the overall bad rate in the development dataset, this parsimonious model under-predicts bad rates, and for demographic groups with bad

rates less than the overall bad rate, this model over-predicts bad rates. Since some minority groups often have higher bad rates, this parsimonious model would benefit these groups. This result is one of the keys behind the first fair lending implication in this report.

From the parsimonious model we then add variables to the model in a sequential manner to reduce the overall prediction error. Each time we add another variable we assess how the prediction errors change for each demographic group. When bad rates in the development dataset vary by demographic group, there will be subsets of loans (i.e., each demographic group) with systematically different levels of performance. The modeling process, which focuses on minimizing prediction error, will therefore find predictive, variables that take on systematically different values for different demographic groups. In other words, if bad rates vary across demographic groups, variables that are correlated with demographic groups will be predictive of performance.[6] The most direct and efficient way to reduce prediction error for each demographic group in the parsimonious model is to include minority flags directly into the model.[7] By doing so, the model will over- and under-predict the probability of bad outcomes for individual loans in the development dataset, but there will be no systematic over- or under-prediction for any demographic groups as the average predicted probabilities for each demographic group will exactly equal the actual bad rate for that group in the development dataset since the logistic estimator preserves marginal probabilities. Of course, directly including race as a variable in

[6] Throughout this report we assume that race and ethnicity do not directly impact performance, so that any differences in performance by race or ethnicity indirectly reflect correlations between race and ethnicity and the true observable (such as income and wealth) and unobservable (willingness to repay debts) variables that determine performance.

[7] Following Hebert-Johnson et al. (2018) a credit scoring model including demographic flags directly would be multi-calibrated, which means the scoring model's predictions are unbiased for each demographic group. The authors then show that forcing a scoring model to be multi-calibrated has only a small, if any, impact on overall accuracy.

scoring models is not appropriate since it violates the Equal Credit Opportunity Act (ECOA) and is illegal.

As an alternative to reducing prediction errors in the parsimonious model by directly including minority flags into the model, we instead explore variables available in the NMDB that might be predictive of bad performance. Table B1 in Appendix B lists the 36 available we analyze, sorted from most to least predictive based on KS-statistic values from a logistic model of the probability of a bad outcome regressed on just a constant and the given variable. This is the variable ordering we use to add variables to each scoring model we develop in this report. As we demonstrate in detail in section V below, adding the first variable (Vantage score) reduces the overall prediction error in the parsimonious model because it is predictive of bad performance. It also reduces prediction errors for each demographic group as well, but by only a portion of the reduction that including minority flags directly would create, since the reduction will come from the correlation between Vantage score and race. Adding the next variable (loan type) further reduces the prediction errors for each demographic group, but again only partially based on correlations with race. As long as systematic prediction errors by demographic group continue to exist, including additional variables will generally reduce these prediction errors based on their correlations with race.[8] Since each added variable will be correlated with race to some extent,

[8] Whether adding a given variable reduces or increases the in-sample prediction error for a given demographic group will depend on all of the underlying correlations between the given variable, each of the demographic groups, the other variables already in the model, and the outcome variable. In some instances, adding a given variable will reduce overall in-sample prediction error while at the same time increase the in-sample prediction error for a given demographic group. The overall trend, however, will be for in-sample prediction errors for each demographic group to move toward zero as additional variables are added to the model. In the extreme, if we continue adding variables, the model will eventually become fully specified with the outcome for each loan being perfectly predicted and no in-sample prediction error for any demographic group. The sequential model development process therefore moves from a parsimonious model with in-sample prediction errors for each demographic group to a fully-specified model with no in-sample prediction for any demographic group. Stated differently, including additional variables will reduce the in-sample prediction errors for each demographic group overall, even though any given variable might increase the in-sample prediction error for a given group.

combinations of all variables included in the model will naturally move toward becoming stronger and stronger proxies for each demographic group as more variables are added. If this sequential process is continued long enough, combinations of variables will eventually become perfect proxies for each demographic group and the average predicted probability for each group will equal the bad rate for that group in the development dataset, since the logistic estimator preserves marginal probability for combinations of variables that define a subset of loans. Therefore, adding variables sequentially will eventually eliminate all bias, which is the same result as if we added a set of minority flags directly to the parsimonious model.[9] We note here that the accuracy of a model built using the sequential approach will be higher since including more variables will capture variation in outcomes that adding just a set of minority flags will not. This is important since improved accuracy is one of the primary benefits of using ML classifiers to build credit scoring models.

A key question for this report is how many variables we need to add to the parsimonious model to reverse or eliminate the original prediction errors for each demographic group. The answer to this will be case specific and depend on the specific data, bad rates, correlations, and modeling process used. However, we show evidence below using the NMDB that the prediction errors for demographic groups tend not to reverse themselves and tend not to fully dissipate even after adding a large number of variables. Further, the results suggest that it takes longer for prediction errors to fully dissipate when differences in bad rates across demographic groups are larger and when the correlations between race and the available variables that are potentially

[9] We are assuming that group membership does not inherently impact variation in bad rates, all-else-equal, which means that any prediction errors in estimating bad rates across demographic groups is due solely to omitted variable bias. Once we account for these omitted variables (i.e., the financial and credit characteristics that truly explain bad rates), then all variation in prediction errors across demographic groups should disappear. A different assumption will lead to different interpretations of the results.

predictive of bad performance are smaller. Again, these findings will not hold for every credit scoring model, but they do provide additional supporting evidence that it is plausible that traditional scoring models with smaller numbers of variables tend to under-predict risk for Black borrowers and over-predict risk for Asian and White borrowers.

Definition of Race Proxies

The final component of our analytical framework is the strategy for determining whether a combination of variables is a proxy for race. The metric we use is the percentage of applicants from a given group that the combination of variables uniquely identifies.[10] As a simple example, suppose we have a sample of 100 Black applicants and 576 non-Hispanic White applicants. Further suppose that we want to assess whether the combination of loan purpose, which takes on 3 values (1, 2, and 3), and loan type, which takes on 4 values (1, 2, 3, and 4) proxy for race. We first partition all Black and non-Hispanic White applicants into segments defined by the unique combinations of the values of loan purpose and loan type. With this approach, each applicant can belong to only one segment. For a given race, we consider an applicant uniquely identified if it is in a segment containing only applicants of that race. For example, Table 1 shows all unique combinations of loan purpose and type and the number of Black and non-Hispanic White applicants in each segment. Segment "13" has 2 Black applicants and 0 non-Hispanic White applicants, which suggests that applications with loan purpose equal to 1 and loan type equal to 3 uniquely identifies 2 Black applicants. Aggregating across all segments, Table 1 shows that the combination of loan purpose and loan type uniquely identifies 5 of the 100 Black applicants, or 5

[10] Another common strategy for estimating the degree with which combinations of variables proxy for race is to estimate a logistic regression of the likelihood of belonging to a given race. A measure like the AUC then provides a measure of the degree to which the variables in the model proxy for race (see Bono et al. (2021)).

Table 1: Proxy Example

Loan Purpose, Loan Type Combinations	# Blacks	# non-Hispanic Whites	# Blacks uniquely identified by purpose and type
11	15	103	0
12	19	58	0
13	2	0	2
14	5	10	0
21	1	13	0
22	4	17	0
23	3	0	3
24	8	15	0
31	21	256	0
32	12	96	0
33	4	6	0
34	5	2	0
Total	100	576	5 (5% of Blacks)

percent, in this example. A higher aggregate percentage for a given demographic group suggests that the combination of variables is a stronger proxy for that group. This is the approach we apply below to assess whether combinations of variables in the scoring models are proxies for race.

There is one important caveat to the metric we use to measure whether a combination of variables is a proxy for race. In the previous section we noted that for a model including a combination of variables that proxy for a demographic group, the average of the predicted probabilities of loans being bad from the model for that group will equal the bad rate for that group in the development dataset. Even if our measure of whether a combination of variables is a proxy for race is near 100%, this result only holds if the model includes interactions of the variables in the model.[11] To explain this caveat, we extend the example from Table 1. Suppose

[11] There is one exception here, which is when each applicant for a given race is uniquely identified by one or more values of just one variable in the scoring model instead of values for two or more variables. This is a highly unlikely scenario.

we have a very simple scoring model including only loan purpose and loan type and that the loan purpose / loan type combination "13" uniquely identifies 100% of Black borrowers. With no interactions, the scoring model would include separate 0/1 flags for loan purposes 1 and 2 and loan types 1, 2, and 3. For Black borrowers, the 0/1 flag for loan purpose 1 and the 0/1 flag for loan type 3 would both contribute to the average of the predicted probabilities of loans being bad. These impacts are both main effects that reflect the overall impact of loan purpose 1 for loans of every loan type (including loan type 3), and the overall impact of loan type 3 for loans of every loan purpose (including loan purpose 1). Given that Black borrowers are uniquely identified by segment "13," we want the contribution to the average of the predicted probabilities of a bad outcome to come only from the impact of the subset of loans with loan purpose 1 and loan type 3. These contributions are specific to just Black applicants and are generated by including variables in the scoring model that interact loan purpose and loan type.

IV. Data

We use the NMDB data for all analyses. The NMDB is a de-identified, loan-level database of closed-end, first-lien, residential mortgages. The core data is a 1-in-20 random sample of all closed-end, first-lien mortgages in the files of Experian between 1998 and 2012. The database has been updated each quarter since 2012 with an additional 1-in-20 random sample of mortgages newly reported to Experian. As of June 2023, the dataset includes data on 14,312,597 loans and 22,124,244 total borrowers from 1998 through March 31st, 2023. For each loan, credit records are collected from at least 12 months prior to origination and until at least one year after termination. The NMDB includes variables characterizing loans, properties securing loans, and borrowers. Most importantly for our purposes, it includes performance data

on each loan for each month since origination, several credit bureau and application variables that might explain performance, and demographic data on each borrower on each loan.

Our development dataset includes only loans originated between January 1, 2015 and December 31, 2017. This time period allows for a sufficient volume of loans for developing credit scoring models. It also avoids any impact of the pandemic since the two-year performance window for loans originated in the last month (December 2017) ends on December 31, 2019. To make the development dataset more homogenous we exclude balloon, interest-only, negative amortization, HOEPA, HARP, and HAMP loans. In addition, we include only home purchase and refinance loans for owner-occupied, site-built properties.

We define a bad outcome as delinquent, in foreclosure, or in bankruptcy at any point during the first 24 months starting in the month after origination. Among the set of loans with no bad outcome during the first 24 months, we exclude those that were closed prior to 24 months, and those with missing or suppressed performance information for at least one of the 24 months, since we cannot determine with certainty if these loans were ever bad for any of the first 24 months after origination. Finally, among the remaining loans, we exclude those that were never current during any of the first 24 months after origination. After applying these filters, 600,482 total loans remained for analysis.

Table 2 presents the total number of loans, the number of loans with bad outcomes, and the bad rate by demographic group for the entire sample of 600,482 loans. The overall bad rate is 6.58%. Similar to many development datasets, the bad rates for Black and Hispanic borrowers are both higher than the overall bad rate, and the bad rates for Asian and White borrowers are both lower. Appendix C, which presents bad rates by loan purpose and type, shows similar patterns. These specific bad rate patterns are a key driver of the findings in this report.

Table 2: Summary Statistics for Loans Originated in 2015-2017			
Demographic Group	# Loans*	# Ever Bad in 2 Years After Origination	Bad Rate
Total	600,482	39,536	6.58%
Asian	41,470	1,246	3.00%
Black	34,540	5,856	16.95%
White	516,283	31,776	6.15%
Hispanic	62,496	5,984	9.58%
Non-Hispanic	537,986	33,552	6.24%
* For race, the sum of the counts for Asian, Black, and White do not sum to the total, because the total includes additional, small-volume racial groups, such as Native Americans. The counts do sum to the total for ethnicity.			

Table B1 in Appendix B lists the 36 variables available in the NMDB that we consider as possible predictors of bad performance. Table B2 provides the corresponding summary statistics for each of these variables. All of these potential predictors are measured either as of the quarter or month prior to origination, or at time of origination. For example, total dollar balances of auto loans/leases is measured quarterly, so a loan originated in July 2018 would use the borrower's total auto balances as of Q2 2018, and a loan originated in December 2019 would use the borrower's total auto balances as of Q3 2019. Prior to model development, we transformed each continuous variable into a categorical variable based on the percentage of missing values and the distribution of non-missing values.[12] [13] We then generated 0/1 flags for each value of each categorical variable. As an example, total dollar balances of open-ended credit cards is a continuous variable with 11.8 percent of loans containing missing values. We first set the categorical variable to 0 for all loans where total dollar balances of open-end credit cards were missing. We then used quartile values of all non-missing values to set the categorical value to 1 if

[12] For most variables, this transformation is based on deciles, quartiles, and medians. The exceptions to this approach are CLTV, DTI, PTI, income and Vantage score. We transform the continuous CLTV into 10-point buckets, DTI into 5-point buckets, PTI into 5-point buckets, income into $10,000 buckets, and Vantage score into 20-point buckets.
[13] Transforming continuous variables into sets of 0/1 flags is a common approach for building credit scoring models. In addition, doing so mitigates impacts of the extreme values that occur for many variables as shown in Table B2.

balances were between $0 and $1,229; 2 if balances were between $1,230 and $3,795; 3 if balances were between $3,796 and $9,749; and 4 if balances were greater than $9,749. We then generated five 0/1 flags for total dollar balances of open-ended credit cards, which we used in model development. The order we used to include variables in the model depends on KS-statistics from a regression of the probability of a bad outcome on each variable separately using the entire sample of 600,482 loans. We sorted the variables in Appendix B by the KS-statistic values, from most to least predictive.

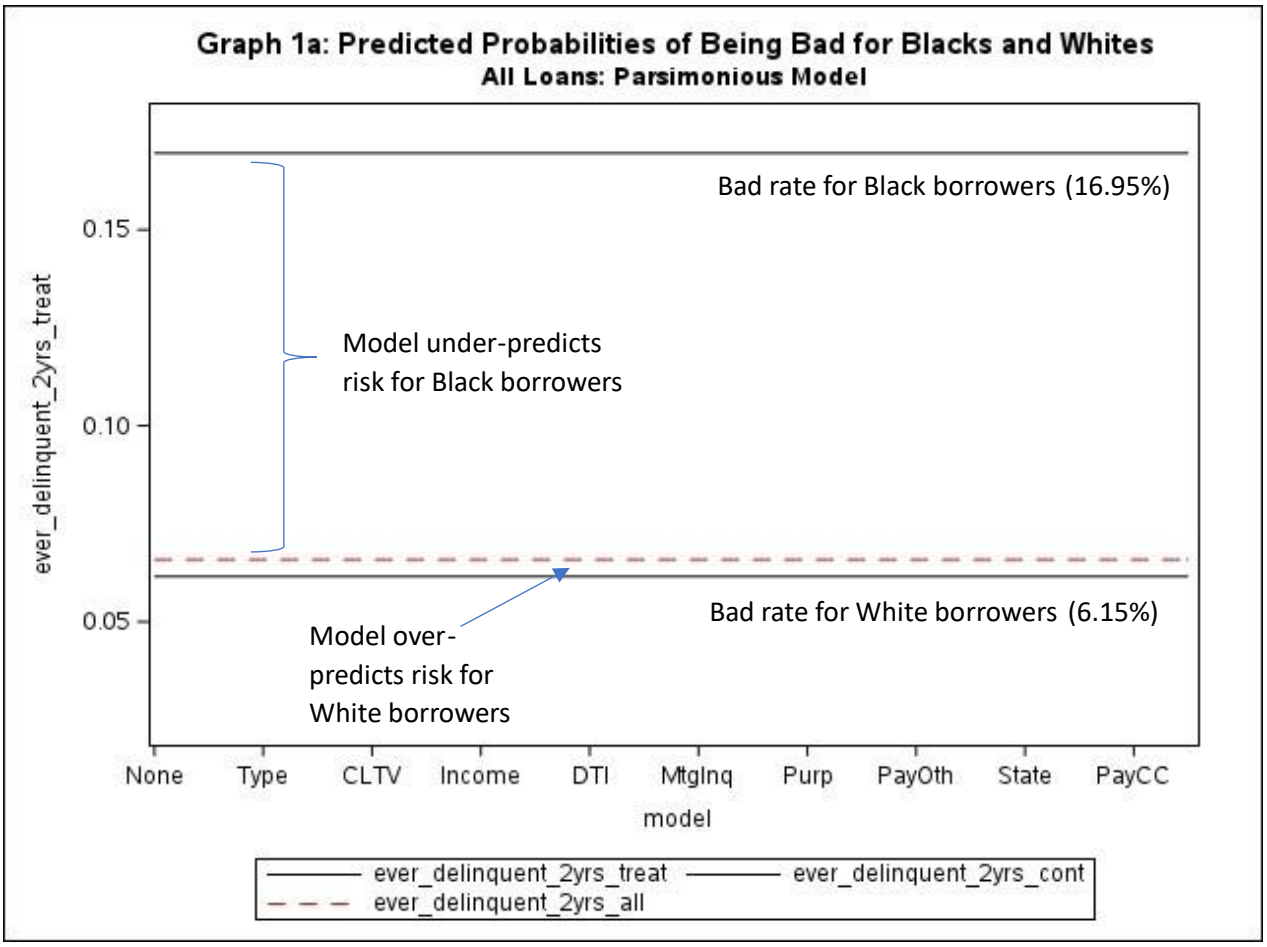
Throughout the report, we focus on results for Black, Asian, White, Hispanic, and non-Hispanic borrowers. We define Black as "Black only," Asian as "Asian only" and White as "White only" based on the primary applicant race data in the NMDB. Similarly, we define Hispanic and non-Hispanic using the ethnicity variable in the NMDB. Although we focus on just these five groups, we include loans from all groups when developing each scoring model and do not filter on race or ethnicity.

V. Analysis and Results

Using the analytical framework and data described above we now build credit scoring models to assess how including additional variables impacts measures of bias across demographic groups, and whether combinations of variables in the models proxy for race and ethnicity. Our primary scoring model is an aggregate model using all loans. As additional evidence of the robustness of the results, we also develop separate scoring models for conventional home purchase loans, conventional refinance loans, FHA-insured home purchase loans, and FHA-insured refinance loans. We focus most of the discussion of the modeling and results here on just the aggregate scoring model, leaving results for the remaining four products to Appendix C.

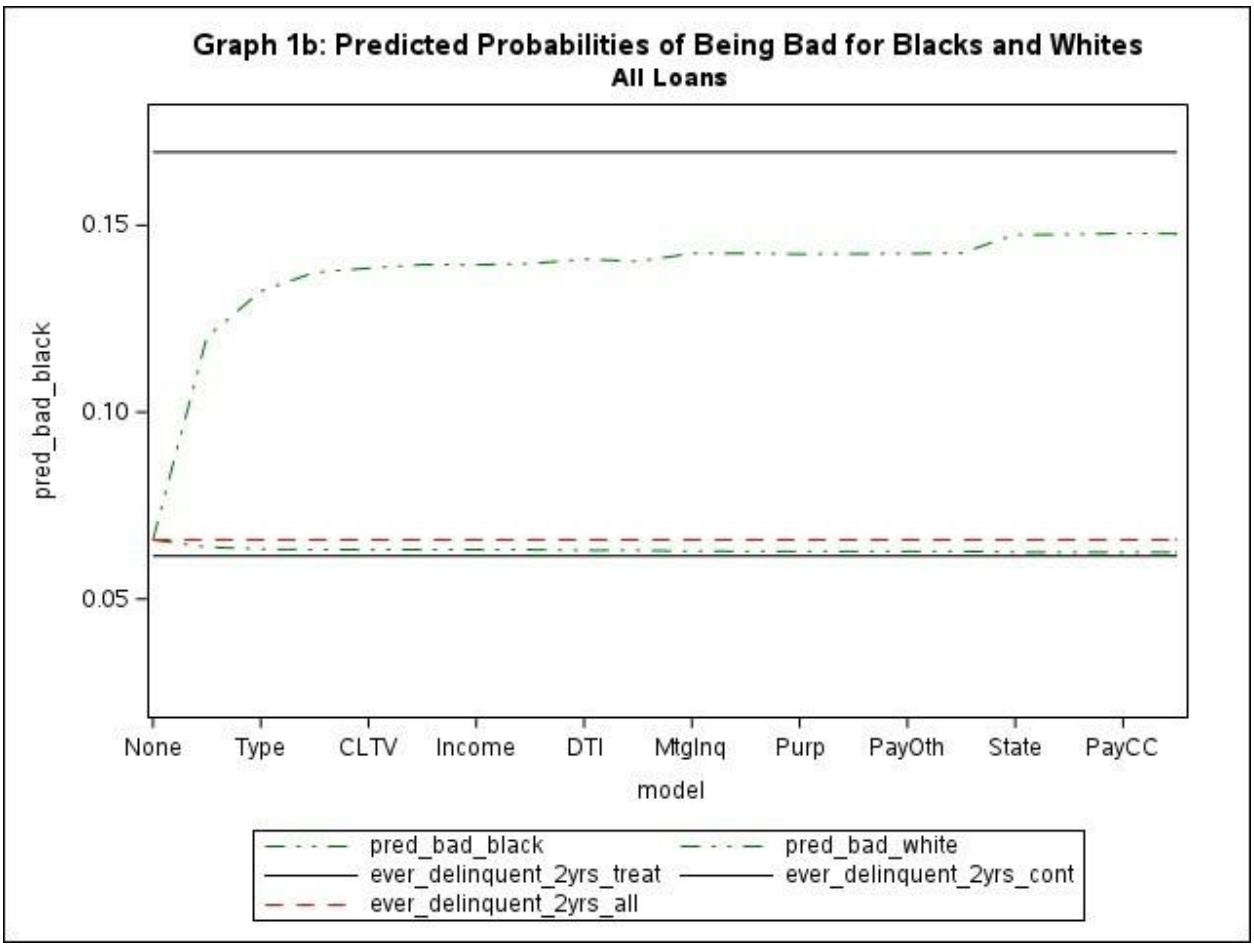
First Fair Lending Implication: Over- and Under-Prediction Errors by Demographic Group

We begin our analysis with the most parsimonious model and then build upon that model in a sequential manner to explore the impact of including additional variables. The bad rate for Black borrowers is 16.95% and the bad rate for White borrowers is 6.15%, as denoted by the two solid horizontal lines in Graph 1a. We can determine these two horizontal lines by calculating the bad rate for each group in the development dataset (see Table 2) or by estimating a logit regression with just a constant and a 0/1 flag identifying Black borrowers and then computing the average of the predicted probabilities of loans being bad separately for Black and White borrowers. If for some reason there are absolutely no variables available during model



development to predict bad loan performance, then one of the best predictors of bad loan performance is the overall bad rate in the development dataset, which is 6.58%. In this very parsimonious scoring model, every applicant receives the same score of 6.58%. As a reminder, we do not translate the predicted probabilities from the logistic regression into a formal score, so the term score here refers to the predicted probabilities. The red dashed line at 6.58% in Graph 1a shows the score that each applicant receives with this parsimonious scoring model. As the graph shows, this score is systematically lower than the bad rate on past loans for Black borrowers and systematically higher than the bad rate on past loans for White borrowers. Therefore, using our measure of prediction error across demographic groups, this parsimonious scoring model under-predicts risk for Black borrowers and over-predicts risk for White borrowers. In other words, Black applicants benefit from this scoring model.

We now add variables to the model sequentially to assess how the average of the predicted probabilities of loans being bad changes for each demographic group. We add variables in order of predictiveness as shown in Appendix B. After adding each variable, we update the average predicted probabilities of a bad outcome for each race. Graph 1b shows the results. The horizontal axis includes labels only for every other variable added to improve readability. Given the initial disparity in bad rates between Black and White borrowers in the development dataset there is a subset of loans (Black borrowers) that have a systematically higher likelihood of bad performance and a subset of loans (White borrowers) that have a systematically lower likelihood of bad performance. Since the modeling process focuses on minimizing prediction error, it will like variables that take on systematically different values for these two subsets of borrowers. Specifically, variables that are correlated with race will be predictive of bad outcomes. Adding Vantage score to the model (not labeled in Graph 1b) increases the average predicted probability of a bad outcome for Black borrowers and reduces the average predicted probability of a bad



outcome for White borrowers. In other words, Vantage score is correlated with race and the likelihood of a bad outcome. The model with just the Vantage score still under-predicts bad outcomes for Black borrowers and over-predicts bad outcomes for White borrowers but less so. Adding loan type to the model again increases the average predicted probability of a bad outcome for Black borrowers and reduces the average predicted probability of a bad outcome for White borrowers. In other words, loan type is correlated with race and the likelihood of a bad outcome. The model with just the Vantage score and loan type still under-predicts bad outcomes for Black borrowers and over-predicts bad outcomes for White borrowers but less so. Continuing this sequential process with 18 additional variables extends the dashed green lines to the right on

the graph.[14] As the graph shows, even after accounting for 20 variables, the model still under-predicts risk for Black borrowers and over-predicts risk for White borrowers, but again less so than for the parsimonious model.[15] As detailed above, continuing to add variables to the scoring model would eventually eliminate all prediction errors for both Black and White borrowers if either the variables in combination become proxies for race or when we have included so many variables that the model becomes perfectly specified.[16]

There are two main takeaways from this exercise. First, when the bad rate in the development dataset differs across demographic groups, and the number of variables included in the scoring model is relatively small, there is a tendency for the final scoring model to under-predict bad rates for the group with the higher bad rate, and over-predict bad rates for the group with the lower bad rate. Given that Black borrowers often have higher bad rates on past loans, this suggests that these types of credit scoring models benefit Black applicants. Second, since the modeling process minimizes prediction error, and Black and White borrowers are two distinct subsets of borrowers with different bad rates on past loans, the modeling process will continue to like additional variables that are correlated with race. Therefore, including additional variables will continue to reduce the prediction error for both groups until both are equal and near 0. At that point, the average predicted bad rates for Black and White borrowers will be equal to the bad rates for Black and White borrowers in the development dataset. We can achieve the exact same result by simply including a 0/1 Black flag as the only variable in the model as shown by Graph

[14] Including the remaining 16 variables listed in Table B1 had only a minor impact on the graphical results, so we stopped at 20 to make the graph more readable, while still maintaining the main takeaways of the results.

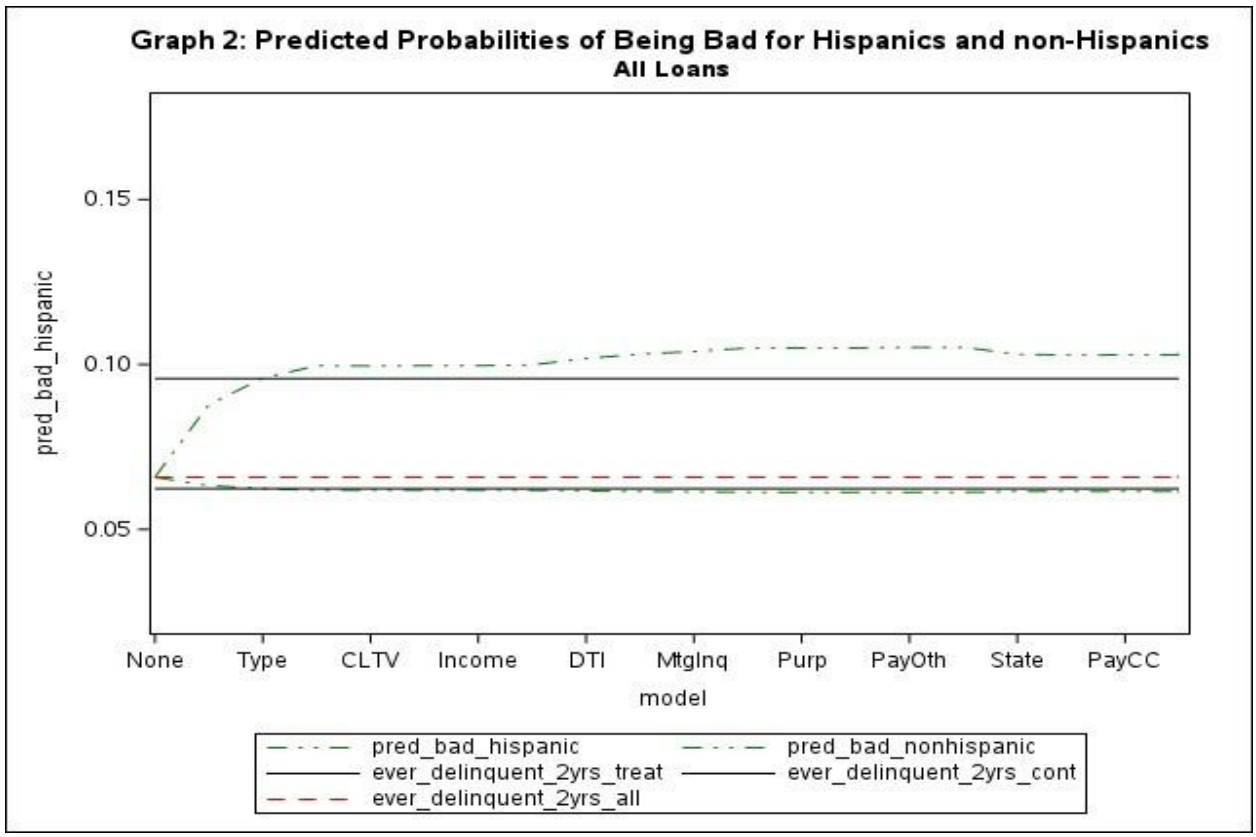
[15] We note that had we included interactions of various variables, similar to what is done for ML scoring models, the prediction errors would likely decrease at a faster rate.

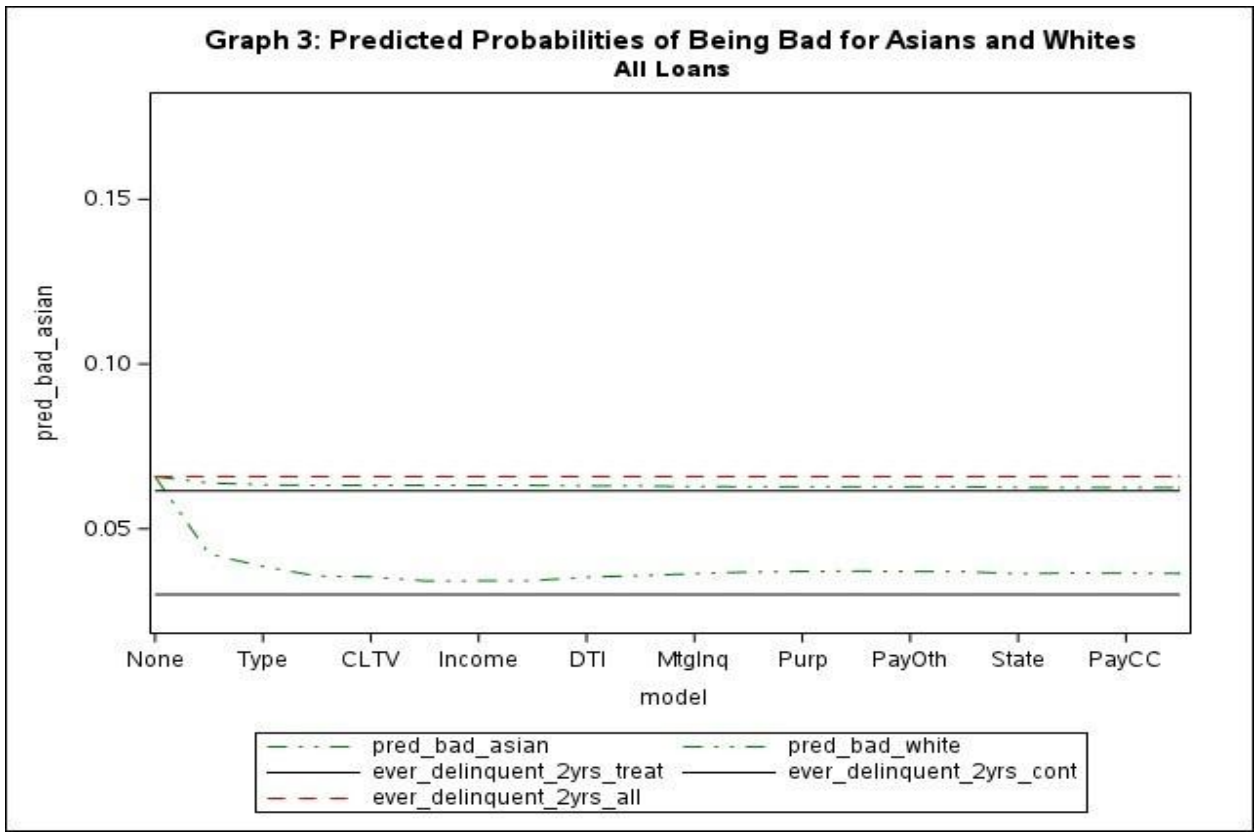
[16] It should be noted that overfitting becomes an increasing challenge as more variables are added to the scoring model, which would erode the overall purpose of the scoring model, which is to predict risk out-of-sample.

1a. In other words, adding additional variables reduces the prediction error by race, but the combinations of variables included in the scoring model become ever closer proxies for race.

Graphs 2 and 3 present the results for Hispanics and Asians, respectively. The story for Hispanic borrowers is slightly different than for Black borrowers. Similar to Black borrowers, Hispanic borrowers have a higher bad rate on past loans as compared to non-Hispanics (9.58% to 6.24%) as indicated by the two solid horizontal lines in Graph 2. However, the disparity in bad rates on past loans is almost twice as large for Black borrowers (16.95-6.15=10.80 percentage points (pps) to 9.58-6.24=3.34 pps). Given the smaller disparity in bad rates on past loans, the initial prediction error in the parsimonious model disappears, and actually reverses itself, after adding just three variables (Vantage score, loan type, and total credit limit on credit cards), and then remains close to zero for all additional models. The different patterns in the results for Black and Hispanic borrowers highlight the importance of the magnitude of differences in bad rates in the development dataset used to develop scoring models. The story is also slightly different for Asian borrowers as well. Asian borrowers had the lowest bad rate on past loans among all races (3.00%), and the bad rates for both Asian and White borrowers were below the overall bad rate. As a result, the initial parsimonious model over-predicted bad rates for both Asian and White borrowers. Including additional variables to the model reduced the prediction errors for both groups. However, even after adding 20 variables, the model still slightly over-predicted bad rates for both Asian and White borrowers. These results show how credit scoring models with variation in bad rates on past loans and small numbers of variables can over-predict risk for both Asian and White borrowers.

Appendix C provides similar sets of results separately for conventional home purchase loans, conventional refinance loans, FHA-insured home purchase loans, and FHA-insured refinance loans. In general, the patterns of results and main takeaways for these four products are


Graph 2: Predicted Probabilities of Being Bad for Hispanics and non-Hispanics
All Loans
pred_bad_hispanic
0.15
0.10
0.05
None
Type
CLTV
Income
DTI
MtgInq
Purp
PayOth
State
PayCC
model
pred_bad_hispanic
pred_bad_nonhispanic
ever_delinquent_2yrs_treat
ever_delinquent_2yrs_cont
ever_delinquent_2yrs_all


Graph 3: Predicted Probabilities of Being Bad for Asians and Whites
All Loans
pred_bad_asian
0.15
0.10
0.05
None
Type
CLTV
Income
DTI
MtgInq
Purp
PayOth
State
PayCC
model
pred_bad_asian
pred_bad_white
ever_delinquent_2yrs_treat
ever_delinquent_2yrs_cont
ever_delinquent_2yrs_all

similar to the results for the aggregate scoring model in Graphs 1-3, adding additional support for the plausibility of the first fair lending implication in this report. There is one important item of note, however. The prediction errors for Black borrowers for both conventional refinance and FHA-insured refinance products converge more quickly to 0 (by model 14), which is similar to including a flag for Black borrowers directly in the model. The main reason for this different result is that, given the smaller sample sizes as compared to the aggregate scoring model, it is more likely that smaller numbers of variables in combination will be proxies for race. If smaller numbers of variables in combination are proxies for race, prediction errors will converge more quickly to 0, since the logistic estimator preserves marginal probabilities. The next section provides additional details on these proxies.

Second Fair Lending Implication: Combinations of Variables Proxying for Race

For all of the model development we conducted and described above we focused primarily on variables that were predictive of performance, and on prediction errors by demographic group. In this section, we analyze the extent to which combinations of the variables added during the modeling process above act as proxies for race. Table 3 presents the results of the proxy analysis for Black, Asian, and Hispanic borrowers. The first column shows the order of variables we added in our sequential modelling process. The second column shows the actual number of 0/1 flags we added for each subsequent model. For each variable, the number of flags we added depends on whether the variable is categorical or continuous and the strategy we used to transform continuous variables into mutually exclusive categories as discussed in Section IV. The third column shows the total number of variables included in each model. The final three columns show the percent of Black, Asian, and Hispanic borrowers uniquely identified

Table 3: Do Combinations of Variables Proxy for Race?

	# variables added	# variables total	% Blacks IDed	% Asians IDed	% Hispanics IDed
Model 1: Just a constant	1	1	0.00%	0.00%	0.00%
Model 2: Model 1 + Vantage Score	11	12	0.00%	0.00%	0.00%
Model 3: Model 2 + Loan Type	1	13	0.00%	0.00%	0.00%
Model 4: Model 3 + Total Credit Limit on Cards	4	17	0.00%	0.00%	0.00%
Model 5: Model 4 + CLTV	6	23	0.00%	0.00%	0.00%
Model 6: Model 5 + Balances on Other Closed-end Loans	3	26	0.06%	0.02%	0.04%
Model 7: Model 6 + Income	20	46	2.37%	1.18%	1.76%
Model 8: Model 7 + Delinquencies on Other Loans	1	47	4.59%	2.20%	3.76%
Model 9: Model 8 + DTI	6	53	17.68%	10.32%	14.77%
Model 10: Model 9 + # non-Mortgage Inquiries	2	55	30.63%	20.72%	26.70%
Model 11: Model 10 + # Mortgage Inquiries	3	58	48.60%	40.95%	45.84%
Model 12: Model 11 + PTI	5	63	71.19%	69.30%	69.83%
Model 13: Model 12 + Loan Purpose	2	65	78.44%	79.01%	77.34%
Model 14: Model 13 + Balance on Open-end Cards	3	68	85.54%	88.71%	85.22%
Model 15: Model 14 + Monthly Payment on Other Loans	3	71	91.63%	93.08%	91.74%
Model 16: Model 15 + Payoff Amount on 1st Liens	3	74	92.03%	93.88%	92.20%
Model 17: Model 16 + State	51	125	99.27%	99.37%	99.34%
Model 18: Model 17 + Credit Limit on Other Loans	4	129	99.35%	99.43%	99.41%
Model 19: Model 18 + Monthly Payment on Cards	4	133	99.53%	99.68%	99.60%
Model 20: Model 19 + Balance on Open-end Other Loans	3	136	99.55%	99.69%	99.62%

by the combinations of variables included in each model. The main result in Table 3 is how, with just a small number of variables, and very broad binning, we can uniquely identify almost every Black, Asian, and Hispanic borrower with combinations of the variables included in the models. With even slightly more granular binning, it is easy to uniquely identify 100 percent of all borrowers of each demographic group.

Overall, similar results hold for the four specific products in Appendix C (results not shown). The one major difference is that the risk of combinations of variables proxying for race occurs more quickly and with smaller numbers of variables. This is due primarily to the smaller sample sizes for these four products compared to the aggregate model. For example, using the aggregate dataset and model, the 58 variables in model 11 uniquely identify 48.60 percent of Black borrowers, 40.95 percent of Asian borrowers, and 45.84 percent of Hispanic borrowers (see Table 3). Using the conventional refinance sample, the 58 variables in model 11 uniquely identify 82.93 percent of Black borrowers, 77.32 percent of Asian borrowers, and 82.65 percent of Hispanic borrowers (results not provided in this report).

One final result of note in these proxy results is how, after adding 20 variables to the model, the percentage of Black borrowers uniquely identified by our measure is nearly 100%, yet the under-prediction error for Black borrowers in Graph 1b is still fairly large, at nearly 2 pps. This result seemingly contradicts one argument made above that the preserving marginal probabilities property of the logistic estimator ensures that for a model including a combination of variables that proxy for a given race, the average of the predicted probabilities for that race will equal the bad rate for that race in the development dataset. Following the discussion at the end of Section III, this result does not exactly hold here because the variables that uniquely identify Black borrowers are not mutually exclusive. Therefore, even though our proxy measure is near 100%, to achieve the result of no over- or under-prediction, we need to include the variables, as well as interactions of the variables, in the model. This

point is particularly relevant for credit scoring models built using ML classifiers, since these models commonly include large numbers of interaction variables. The inclusion of interaction variables in ML models will further increase the likelihood that combinations of variables will proxy for race, and further eliminate prediction errors by demographic group.

VI. Conclusion

Anecdotal evidence from fair lending reviews suggests that traditional credit scoring models with smaller numbers of variables often under-predict risk for Black borrowers and over-predict risk for Asian and White borrowers. In this report we have used standard model development strategies, along with properties of the logistic estimator, to demonstrate the mechanism behind this anecdotal evidence. In addition, using the NMDB, we have developed credit scoring models for several products, providing empirical evidence of these relationships as well.

These results have several implications as industry and regulators consider the potential impacts of the current transition from credit scoring models developed using traditional econometric estimators, which typically have smaller numbers of variables, to credit scoring models developed using ML classifiers, which typically have larger numbers of variables. First, this report adds to the growing base of research on algorithmic bias that provides the foundation for understanding the impacts of future ML innovations in this space. A better understanding of the drivers of differences in average scores and bias across demographic groups in traditional scoring models will help regulators and industry assess the impacts on bias, subsequent access to credit, and pricing from moving to ML scoring models. Very specifically, the analysis and results in this report suggest that transitioning from traditional scoring models to ML scoring models might reduce some bias favorable to Black borrowers and some bias harmful to Asian and White borrowers. Second, the results demonstrating that including larger numbers of variables into

scoring models may naturally lead to combinations of those variables becoming ever stronger proxies for race have important policy and regulatory implications regarding disparate impact. Specifically, transitioning from traditional scoring models to ML scoring models likely increases the risk that combinations of variables in the scoring model proxy for race. Finally, this report underscores the important role that variation across demographic groups in the bad rates used to develop scoring models has on both differences in average scores and algorithmic bias across groups. This result highlights the potential value of using alt data as a source of information on accounts, such as rental payments or utility payments, which may have less variation in performance measures across demographic groups, and might more accurately reflect true bad rates across demographic groups. With increasing availability of alt data containing a much broader set of performance metrics, modelers may be able to develop scoring models using metrics with less variation across groups, which will result in less variation across groups in average scores and potentially in algorithmic bias.

We conclude with several important caveats. First, throughout, we have taken the development data as given and focused only on bias that the modeling process generates. We emphasize that this is only one source of potential algorithmic bias, and that it is important to consider, explore, and address all potential sources of bias. This includes how a specific scoring model is used, such as what products and decisions it impacts, what thresholds are applied, what other variables the scorecard is combined with, what other variables impact the given decision, and whether there are exceptions or overlays to the use of the scoring model, among others. Second, our results are specific to our analytical choices and assumptions. For example, we have used a specific definition of bias that is closely aligned with the well-calibrated and preserves marginal probabilities properties of the logistic estimator. As Corbet-Davies et al. (2018) and others have pointed out, there are several definitions of fairness and bias in the literature, and

different definitions may lead to different results. Relatedly, as Cowgill and Tucker (2019) point out, “Multiple theoretical papers demonstrate the mathematical impossibility of satisfying all fairness criteria, particularly simultaneously.” We have also used a specific dataset, the NMDB, with a specific measure of bad performance and available predictors. Different data may lead to different results as well. Third, whether a given scoring model over- or under-predicts risk for a demographic group will be model-specific and depend on a large number of factors including, the development data, the analytical approach used to develop the model, the objective of the model, and the assumptions and choices the modeler makes, among others. The prediction errors discussed and shown here will therefore not occur for every model. Fourth, we focused solely on prediction errors and did not consider offsetting impacts on accuracy, and the potential to increase access to credit and reduce pricing, which are some of the key benefits of ML models. As Blattner and Nelson (2021) show, credit scores are statistically noisier indicators of default risk for historically under-served groups, so equalizing the precision of credit scores across groups can reduce disparities in approval rates for disadvantaged groups.

As noted throughout, our goal was to provide a reasonable analytical framework, along with corresponding empirical results, showing the feasibility of algorithmic bias patterns found anecdotally during fair lending reviews. We leave to future research, formalizing and expanding on the results in this paper, and exploring each of the caveats above.

References

Aggarwal, Rahul, Kirsten Bibbins-Domingo, Robert W. Yeh, Yang Song, Nicholas Chiu, Rishi K. Wadhera, Changyu Shen, and Dhruv S. Kazi. (2022). "Diabetes Screening by Race and Ethnicity in the United States: Equivalent Body Mass Index and Age Thresholds," Annals of Internal Medicine, 175(6), pp. 765–773.

Barocas, S., and A. D. Selbst. (2016). "Big Data's Disparate Impact," 104 California Law Review 671, doi: 10.2139/ssrn.2477899.

Barocas, S., Moritz Hardt, and Arvind Narayanan. (2017). "Fairness in Machine Learning," in Conference on Neural Information Processing Systems, Long Beach, CA.

Barocas, S., M. Hardt, and A. Narayanan. (2018). Fairness and Machine Learning, fairmlbook.org, http:// www.fairmlbook.org.

Bartlett, R. P., A. Morse, R. Stanton, and N. Wallace. (2019). "Consumer-lending Discrimination in the Fintech Era," NBER Working Paper No. w25943.

Bono, Teresa, Karen Croxson, and Adam Giles. (2021). "Algorithmic Fairness in Credit Scoring," Oxford Review of Economic Policy, Volume 37, Number 3, pp. 585-617.

Caliskan, A., J. J. Bryson, and A. Narayanan. (2017). "Semantics Derived Automatically from Language Corpora Contain Human-like Biases," Science, 356(6334), pp. 183–186.

Chouldechova, A. (2017). "Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments," Big Data, 5.

Corbett-Davies, Sam, Johann D. Gabler, Hamed Nilforoshan, Ravi Shroff, and Sharad Goel. (2018). "The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning," arXiv preprint arXiv:1808.00023.

Cowgill, Bo, and Catherine Tucker. (2019). "Economics, Fairness and Algorithmic Bias," Columbia Business School Research Paper, doi: 10.2139/ssrn.3361280.

Cowgill, Bo. (2018). "Bias and Productivity in Humans and Algorithms," Working Paper.

Friedler, S. A., C. Scheidegger, and S. Venkatasubramanian. (2016). "On the (Im)possibility of Fairness," arXiv:1609.07236.

Friedman, Batya and Helen Nissenbaum. (1996). "Bias in Computer Systems," ACM Transactions on Information Systems (TOIS), 14 (3), pp. 330–347.

Fuster, A., P. Goldsmith-Pinkham, T. Ramadorai, and A. Walther. (2020). "Predictably Unequal? The Effects of Machine Learning on Credit Markets," Working Paper, doi: 10.2139/ssrn.3072038.

Hardt, M., E. Price, and N. Srebro. (2016). "Equality of Opportunity in Supervised Learning," Advances in Neural Information Processing Systems 29 (NIPS).

Hebert-Johnson, Ursula, Michael Kim, Omer Reingold, and Guy Rothblum. (2018). "Multicalibration: Calibration for the (Computationally-identifiable) Masses," In International Conference on Machine Learning, pp. 1939–1948. PMLR.

Hurlin, Christophe, Christophe Perignon, and Sebastien Saurin. (2022). "The Fairness of Credit Scoring Models," arXiv:2205.10200v1.

Kleinberg, J. M., and S. Mullainathan. (2018). "Simplicity Creates Inequity: Implications for Fairness, Stereotypes, and Interpretability," Computing Research Repository, abs/1809.04578.

Lee, M. S. A., and L. Floridi. (2020). "Algorithmic Fairness in Mortgage Lending: From Absolute Conditions to Relational Trade-offs," Minds and Machines, doi: 10.2139/ssrn.3559407.

Liu, L. T., M. Simchowitz, and M. Hardt. (2019). "The Implicit Fairness Criterion of Unconstrained Learning," arXiv:1808.10013.

Ludwig, Jens, Jon Kleinberg, Sendhil Mullainathan, and Ashesh Rambachan. (2018). "Should Algorithms Be Regulated?"

Mitchell, S., E. Potash, and S. Barocas. (2018). "Prediction-based Decisions and Fairness: A Catalogue of Choices, Assumptions, and Definitions," ArXiV: 1811.07867v2.

Obermeyer, Ziad, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. (2019). "Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations," Science 366, pp. 447-453.

Pedreshi, D., S. Ruggieri, and F. Turini. (2008). "Discrimination-aware Data Mining," in Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 560–568, New York, NY.

Rambachan, Ashesh, Jon Kleinberg, Jens Ludwig, and Sendhil Mullainathan. (2020). "An Economic Perspective on Algorithmic Fairness," AEA Papers and Proceedings, May 2020, Vol. 110, pp. 91-95.

Robb, Alicia, and David T. Robinson. (2018). "Testing for Racial Bias in Business Credit Scores," Small Business Economics, Vol. 50, No. 3, Special Issue: Minority Entrepreneurship in 21st Century America, pp. 429-443.

Romei, Andrea, and Salvatore Ruggieri. (2013). "A Multidisciplinary Survey on Discrimination Analysis," The Knowledge Engineering Review, Vol. 29, pp. 582-638.

Rothblum, Guy N. and Gal Yona. (2022). "Decision-making Under Miscalibration," arXiv preprint arXiv:2203.09852.

Vanhoof, M., F. Reis, T. Ploetz, and Z. Smoreda. (2018). “Assessing the Quality of Home Detection from Mobile Phone Data for Official Statistics,” Journal of Official Statistics, 34(4), doi: 10.2478/ jos-2018-0046.

Zliobaite, Indre. (2017). “Measuring Discrimination in Algorithmic Decision Making,” Data Mining and Knowledge Discovery, pp. 1–30.

Appendix A: Technical Appendix

This appendix provides a detailed explanation of the well-calibrated and preserves marginal probability properties of the logistic estimator. We start with a general likelihood function for a limited dependent variable regression,

$$L(\beta) = \prod_{i=1}^{n} p_i^{y_i} \ (1 - p_i)^{1-y_i} \tag{A1}$$

where i indexes individual loans from i = 1 to n and,

β = the parameters on the variables in the scoring model
p_i = the unobserved probability of bad performance for loan i
y_i = 1 if the observed performance on loan i was bad; 0 otherwise

For computational purposes, prior to the optimization process that determines the estimates of β, the likelihood function, which is a product of probabilities, is typically transformed into a negative log-likelihood function, which is a sum of probabilities,

$$l(\beta) = \sum_i^n -y_i \log p_i - \ (1 \ - \ y_i) \ \log(1 \ - \ p_i) \tag{A2}$$

For the logistic estimator, the probabilities take the form,

$$p_i = \frac{1}{1 + e^{-x\beta}} \tag{A3}$$

$$(1 - p_i) \ = \ \frac{1+e^{-x\beta}}{1+e^{-x\beta}} - \frac{1}{1+e^{-x\beta}} = \frac{e^{-x\beta}}{1+e^{-x\beta}} \tag{A4}$$

where x is the set of variables in the scoring model. Substituting these probabilities into (A2) and simplifying yields the negative log-likelihood function for the logistic estimator,

$$l(\beta) = \sum_i^n -y_i \log\left(\frac{1}{1+e^{-x\beta}}\right) - (1 - y_i) \log\left(\frac{e^{-x\beta}}{1+e^{-x\beta}}\right) \tag{A5}$$

$$l(\beta) = \sum_i^n -y_i\left[\log(1) \ - \ \log\left(1 + e^{-x\beta}\right)\right] - (1 - y_i)\left[\log\left(e^{-x\beta}\right) - \log\left(1 + e^{-x\beta}\right)\right] \tag{A6}$$

$$l(\beta) = \sum_i^n y_i \log\left(1 + e^{-x\beta}\right) - (1 - y_i)\left[(-x\beta) - \log\left(1 + e^{-x\beta}\right)\right] \tag{A7}$$

To identify the β values that minimize the negative log likelihood function we take the derivative of (A7) with respect to β_j, where j denotes the jth variable in the model, and then set this derivative to 0.

$$\frac{\partial l(\beta)}{\partial \beta_j} = \sum_i^n y_i \left[\frac{-x_{ji}e^{-x\beta}}{1+e^{-x\beta}}\right] - (1-y_i)\left[(-x_{ji}) - \frac{-x_{ji}e^{-x\beta}}{1+e^{-x\beta}}\right] = 0 \quad \text{(A8)}$$

$$\frac{\partial l(\beta)}{\partial \beta_j} = \sum_{i=1}^n y_i \left[\frac{-x_{ji}e^{-x\beta}}{1+e^{-x\beta}}\right] + (1-y_i)(x_{ji}) + (1-y_i)\left[\frac{-x_{ji}e^{-x\beta}}{1+e^{-x\beta}}\right] = 0 \quad \text{(A9)}$$

$$\frac{\partial l(\beta)}{\partial \beta_j} = \sum_{i=1}^n y_i \left[\frac{-x_{ji}e^{-x\beta}}{1+e^{-x\beta}}\right] + (1-y_i)(x_{ji}) + \left[\frac{-x_{ji}e^{-x\beta}}{1+e^{-x\beta}}\right] - y_i\left[\frac{-x_{ji}e^{-x\beta}}{1+e^{-x\beta}}\right] = 0 \quad \text{(A10)}$$

$$\frac{\partial l(\beta)}{\partial \beta_j} = \sum_{i=1}^n (1-y_i)(x_{ji}) + \left[\frac{-x_{ji}e^{-x\beta}}{1+e^{-x\beta}}\right] = 0 \quad \text{(A11)}$$

$$\frac{\partial l(\beta)}{\partial \beta_j} = \sum_{i=1}^n (1-y_i)(x_{ji}) + [-x_{ji}(1-p_i)] = 0 \quad \text{(A12)}$$

$$\frac{\partial l(\beta)}{\partial \beta_j} = \sum_{i=1}^n \left(x_{ji} - (y_i x_{ji})\right) - x_{ji} + x_{ji}p_i = 0 \quad \text{(A13)}$$

$$\frac{\partial l(\beta)}{\partial \beta_j} = \sum_{i=1}^n (x_{ji}p_i - y_i x_{ji}) = 0 \quad \text{(A14)}$$

$$\sum_{i=1}^n x_{ji}\, p_i = \sum_{i=1}^n y_i\, x_{ji} \quad \text{(A15)}$$

Since we took the derivative of the negative log likelihood function with respect to β_j, the general equation (A15) holds for each of the j variables in the model. This set of j equations is called the set of calibration equations. In the special case of a model with only a constant, x_{1i} equals 1 for all i, so the sum of predicted probabilities equals the sum of the outcome variable. Dividing each side by n, we get that the average of the predicted probabilities equals the average of the outcome variable, which is the well-calibrated property of the logistic estimator.

For a model that includes a minority flag, the x_{ji} variable corresponding to this flag will contain 1's and 0's. In this model, the above equation simplifies to the sum of the predicted

probabilities for minorities equal to the sum of the outcome variable for minorities. Again, dividing each side by n, we get that the average of the predicted probabilities for minorities equals the average of the outcome variable for minorities. If instead of a minority flag, the model includes several variables that in combination are a perfect proxy for minority status, the same result holds. Specifically, in this model, equation (A15) simplifies to the sum of the predicted probabilities for minorities equal to the sum of the outcome variable for minorities. Again, dividing each side by n, we get that the average of the predicted probabilities for minorities equals the average of the outcome variable for minorities. So, for each binary variable in our regression, or combinations of binary variables that together perfectly proxy for minority status, if we subset down to the observations where this binary variable (or the combination of binary variables) is on, the sum of the predicted probabilities for these observations equals the sum of the response. This is what is meant by "preserves marginal probability".

Appendix B: Background Information on Potential Predictors in NMDB

Table B1: Variable Definitions, Frequency, and Predictiveness

Variable Name	Description	Frequency	KS-Statistic
Score_orig_1	Vantage 3.0 score at origination date	Q	0.12434
Loan Type	Conventional vs FHA-Insured	Q	0.09970
limtcardqQ	Total credit limit for credit cards	Q	0.08371
Cltv_all	Combined loan-to-value at origination	Q	0.06699
CloseothrqQ	Total $ balances of closed-end other loans	Q	0.05372
Income	Income relied upon in underwriting	Q	0.05213
DelqothrqQ	Highest delinquency for other loans	Q	0.05181
DTI	Debt-to-income ratio at origination	Q	0.04409
Inqn	# of non-mortgage inquiries in 12 mths prior to orig	M	0.04359
Inqm	# of mortgage inquiries in 12 mths prior to orig	M	0.03899
Pti	Payment-to-income ratio at origination	Q	0.03592
Loan Purpose	Home Purchase vs Refinance	Q	0.03264
opencardqQ	Total $ balances of open-ended credit cards	Q	0.03175
PaymothrqQ	Total $ amt of monthly payments on other loans	Q	0.03008
Payoff_first	Payoff amt of 1st liens at origination	Q	0.03001
Geo2021_st	State where the property is located	Q	0.02788
limtothrqQ	Total credit limit for open-ended other loans	Q	0.02237
PaymcardqQ	Total $ amt of monthly payments on credit cards	Q	0.02106
openothrqQ	Total $ balances of open-ended other loans	Q	0.01981
Paym2ndqQ	Total $ amt of monthly payments on 2nd liens	Q	0.01758
Open2ndqQ	Total $ balances of open-ended 2nd liens	Q	0.01489
Limt2ndqQ	Total credit limit for open-ended 2nd liens	Q	0.01489
acctautoqQ	Total $ balances of auto loans/leases	Q	0.01200
Payoff_sub	Payoff amt of subordinate liens at origination	Q	0.01035
PaymautoqQ	Total $ amt of monthly payments on auto loans	Q	0.01031
acctstudqQ	Total $ balances of student loans	Q	0.00967
Open_month	Month mortgage was opened	Q	0.00933
DelqcardqQ	Highest delinquency for credit cards	Q	0.00868
PaymstudqQ	Total $ amt of monthly payments on student loans	Q	0.00776
DelqautoqQ	Highest delinquency for auto loans	Q	0.00622

Close2ndqQ	Total $ balances of closed-end 2nd liens	Q	0.00313
DelqstudqQ	Highest delinquency for student loans	Q	0.00249
Sub_open	Sum of open-ended piggyback subordinate liens	Q	0.00237
Sub_close	Sum of close-ended piggyback subordinate liens	Q	0.00189
Units	Number of units in property	Q	0.00043
Delq2ndqQ	Highest delinquency for 2nd liens or HELOCS	Q	0.00029

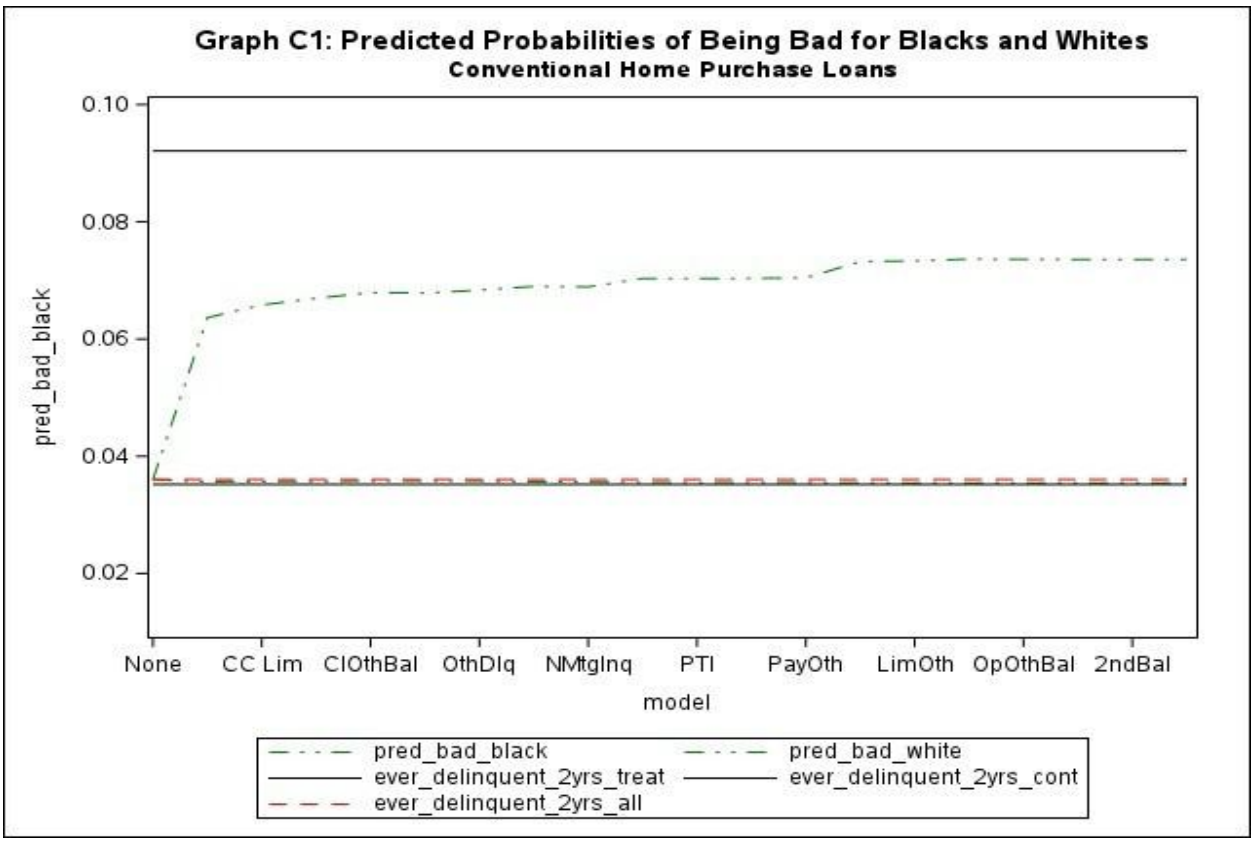
Table B2: Summary Statistics of Potential Predictors in NMDB

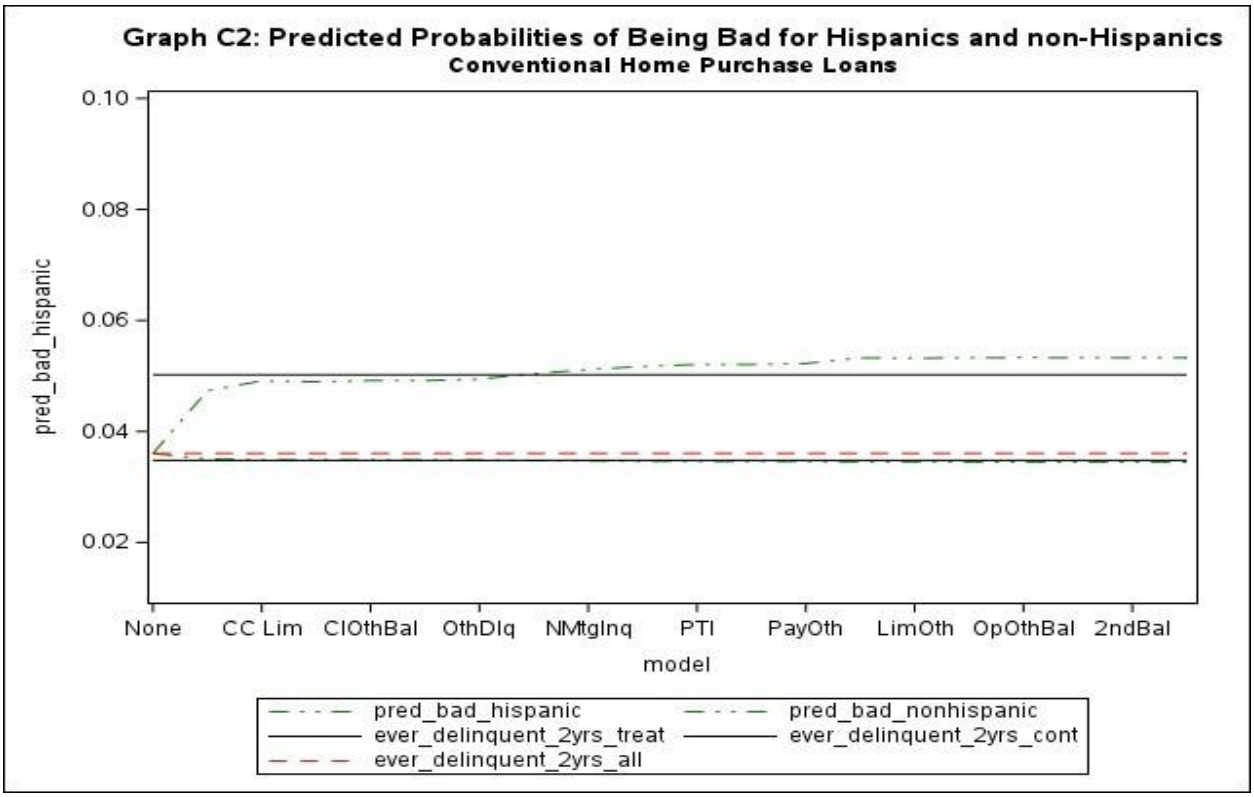
Variable	Label	N	Mean	Std Dev	Min	Max
score_orig_1	Vantage 3.0 score at origination date	600481	740.44	63.81	300.00	839.00
conv	Conventional	600482	0.79	0.41	0.00	1.00
fha	FHA-Insured	600482	0.21	0.41	0.00	1.00
limit_card	Total credit limit for cards	529786	26250.35	26003.04	50.00	468500.00
cltv_all	Combined loan-to-value at origination	600482	76.72	19.24	1.00	999.00
cl_oth_bal	$ balances of closed-end other loans	172636	14845.71	58757.52	1.00	6850160.00
income	Income relied upon in underwriting	600482	103550.59	95124.74	3000.00	7256000.00
del_oth	Highest delinquency for other loans	53590	8.64	1.62	1.00	9.00
dti	Debt-to-income ratio at origination	600482	35.22	10.52	2.00	99.00
N_nonmtg_inqs	# of non-mtg inquiries 12 mths pre-orig	600482	1.38	2.04	0.00	79.00
N_mtg_inqs	# of mtg inquiries in 12 mths pre-orig	600482	1.90	1.34	0.00	48.00
pti	Payment-to-income at origination	600482	21.74	9.82	1.00	99.00
hp	Home Purchase	600482	0.52	0.50	0.00	1.00
refi	Refinance	600482	0.47	0.50	0.00	1.00
op_card_bal	$ balances of open-ended credit cards	529786	8047.29	11856.65	1.00	295001.00
mth_pay_oth	$ amt of mthly payments on other loans	352818	288.19	6289.86	0.00	1100027.00
payoff_first	Payoff amt of 1st liens at origination	279417	199903.59	204791.79	0.00	12465238.00
limit_oth	$ credit limit open-ended other loans	278425	9190.88	41367.07	16.00	8000000.00
mth_pay_card	$ amt of mthly payments on cards	529786	175.90	285.49	0.00	51416.00
op_oth_bal	$ balances of open-ended other loans	278425	3758.63	30126.83	1.00	7878309.00
mth_pay_2nd	$ amt of mthly payments on 2nd liens	76636	451.12	4930.06	0.00	678278.00
op_2nd_bal	$ balances of open-ended 2nd liens	58987	66989.63	108681.14	1.00	9540797.00
limit_2nd	$ limit for open-ended 2nd liens	58987	91456.12	127586.98	35.00	9540797.00
auto_bal	$ balances of auto loans/leases	352213	19929.75	17663.58	1.00	1091919.00
payoff_sub	Payoff amt of sub liens at orig	279417	9022.36	43080.69	0.00	9577951.00
mth_pay_auto	$ amt of mthly payments on auto loans	352213	539.34	438.35	0.00	105903.00
student_bal	Total $ balances of student loans	160764	40837.01	56155.27	1.00	1048483.00
open_month	Month mortgage was opened	600482	223.19	10.06	205.00	240.00
del_card	Highest delinquency for credit cards	8216	6.34	3.61	1.00	9.00
mth_pay_stud	$ mthly payments on student loans	160764	299.28	659.68	0.00	202908.00
del_auto	Highest delinquency for auto loans	3460	6.25	3.70	1.00	9.00
cl_2nd_bal	$ balances of closed-end 2nd liens	19181	43704.88	70811.93	1.00	2982645.00
del_stud	Highest delinquency for student loans	1966	7.09	2.72	1.00	9.00
sub_open	Sum of open-ended piggyback sub liens	16320	51708.33	92578.03	0.00	2500000.00
sub_close	Sum of close-ended piggyback sub liens	16320	25513.62	81137.41	0.00	4000000.00
units	Number of units in property	600482	1.02	0.16	1.00	4.00
del_2nd	Highest delinq for 2nd liens or HELOCS	408	5.69	3.67	1.00	9.00

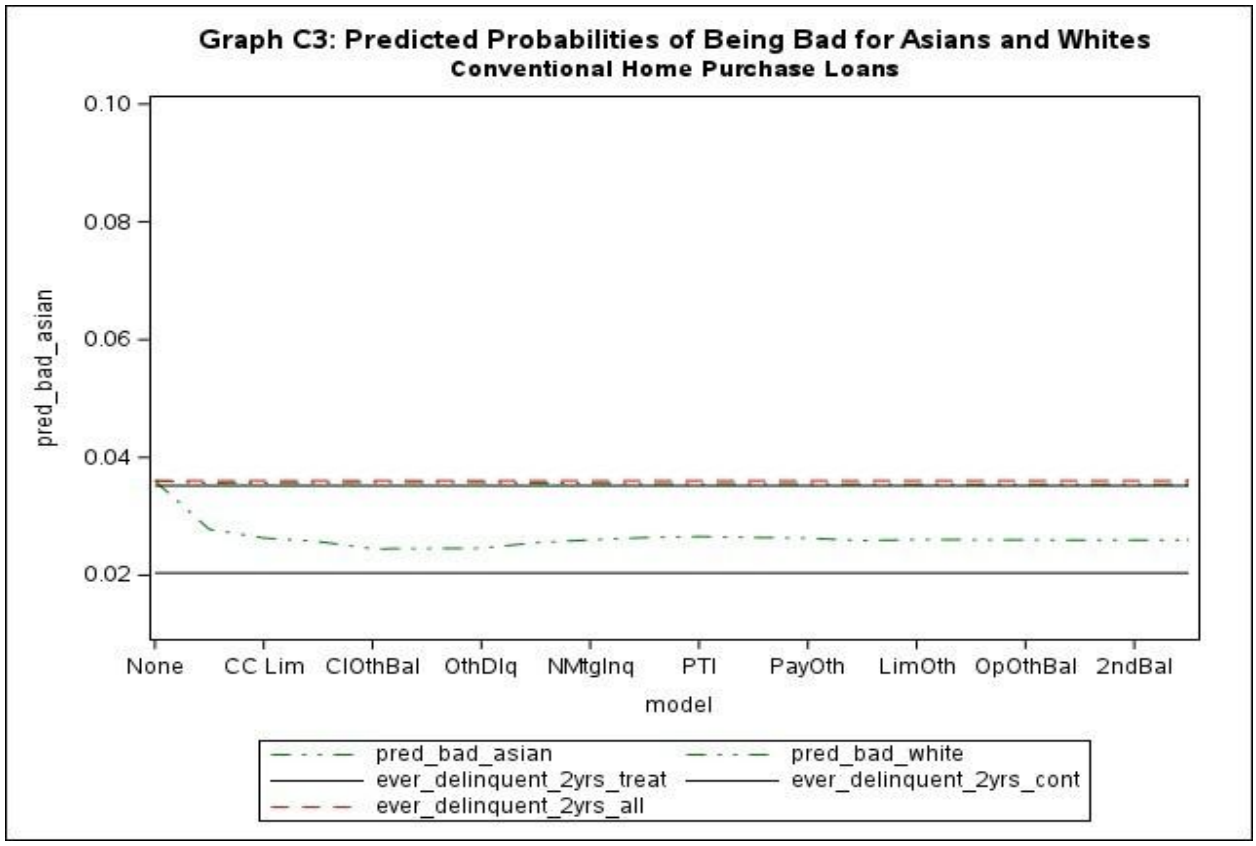
Appendix C: Results for Conventional Home Purchase, Conventional Refinance, FHA-Insured Home Purchase, and FHA-Insured Refinance Products

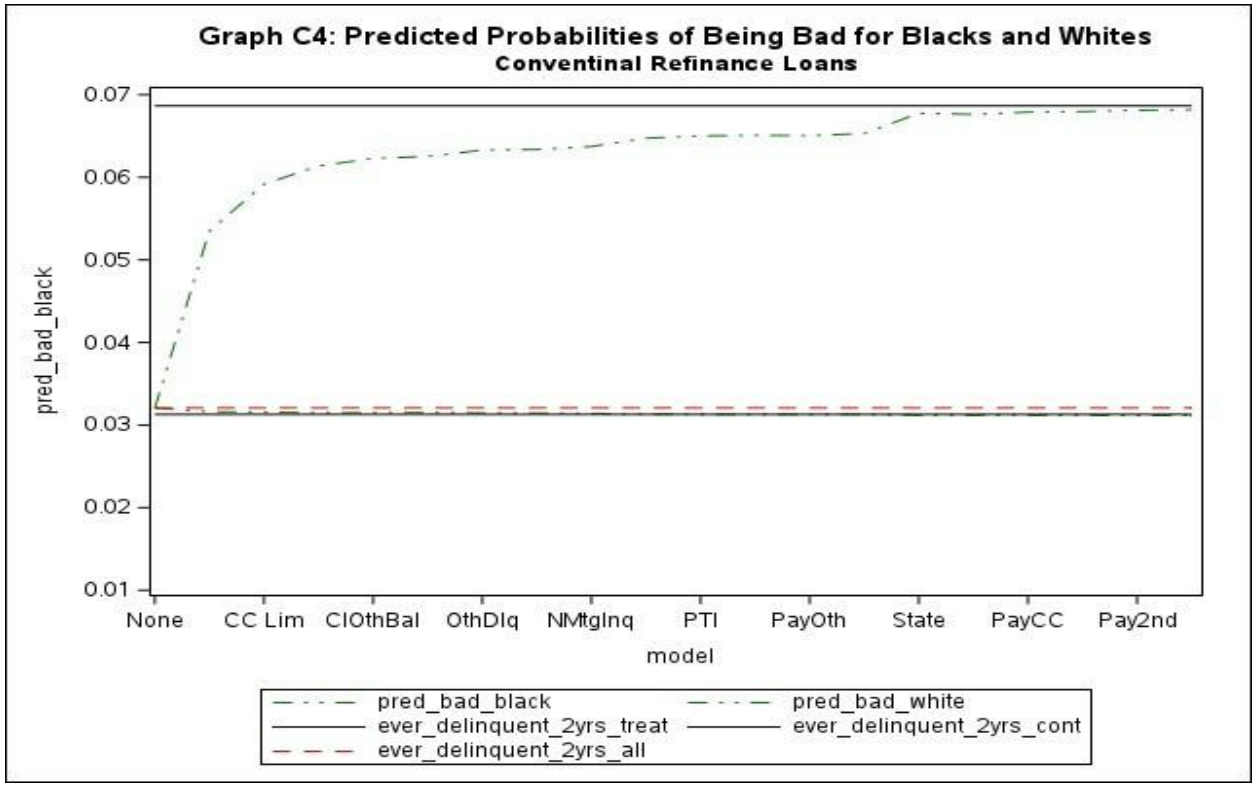
Table C1: Bad Rates by Product and Demographic Group from the Development Data

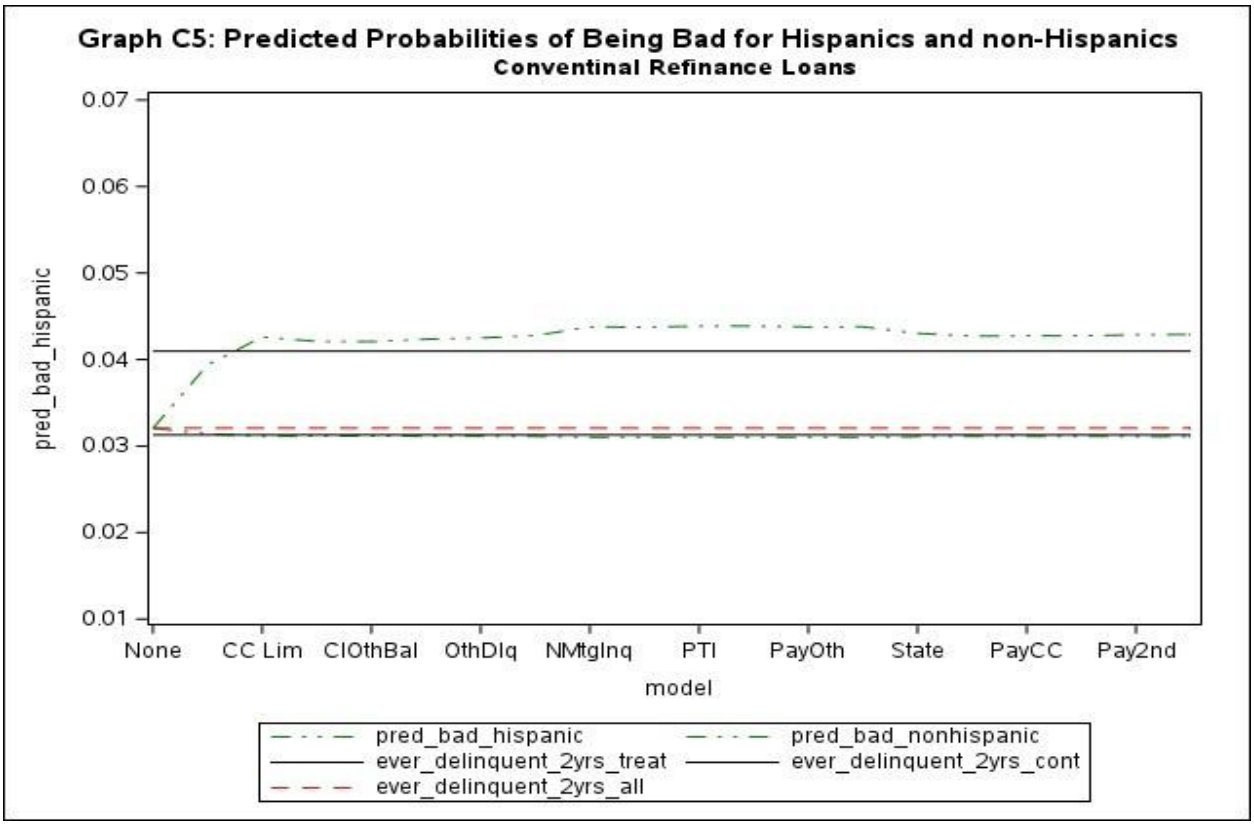
Product	Demographic Group	# Loans	# Ever Delinquent in 2 Years from Origination	Bad Rate
Conventional Home Purchase	Total	223,840	8,067	3.60%
Conventional Home Purchase	Asian	20,202	412	2.04%
Conventional Home Purchase	Black	8,080	744	9.21%
Conventional Home Purchase	White	192,724	6,786	3.52%
Conventional Home Purchase	Hispanic	18,103	909	5.02%
Conventional Home Purchase	Non-Hispanic	205,737	7,158	3.48%
Conventional Refinance	Total	243,740	7,820	3.21%
Conventional Refinance	Asian	16,831	321	1.91%
Conventional Refinance	Black	9,875	678	6.87%
Conventional Refinance	White	214,040	6,699	3.13%
Conventional Refinance	Hispanic	20,086	823	4.10%
Conventional Refinance	Non-Hispanic	223,654	6,997	3.13%
FHA-Insured Home Purchase	Total	85,903	16,521	19.23%
FHA-Insured Home Purchase	Asian	2,684	344	12.82%
FHA-Insured Home Purchase	Black	10,486	3,152	30.06%
FHA-Insured Home Purchase	White	71,180	12,750	17.91%
FHA-Insured Home Purchase	Hispanic	17,688	3,192	18.05%
FHA-Insured Home Purchase	Non-Hispanic	68,215	13,329	19.54%
FHA-Insured Refinance	Total	35,714	5,718	16.01%
FHA-Insured Refinance	Asian	1,051	134	12.75%
FHA-Insured Refinance	Black	4,455	868	19.48%
FHA-Insured Refinance	White	29,572	4,612	15.60%
FHA-Insured Refinance	Hispanic	5,167	883	17.09%
FHA-Insured Refinance	Non-Hispanic	30,547	4,835	15.83%

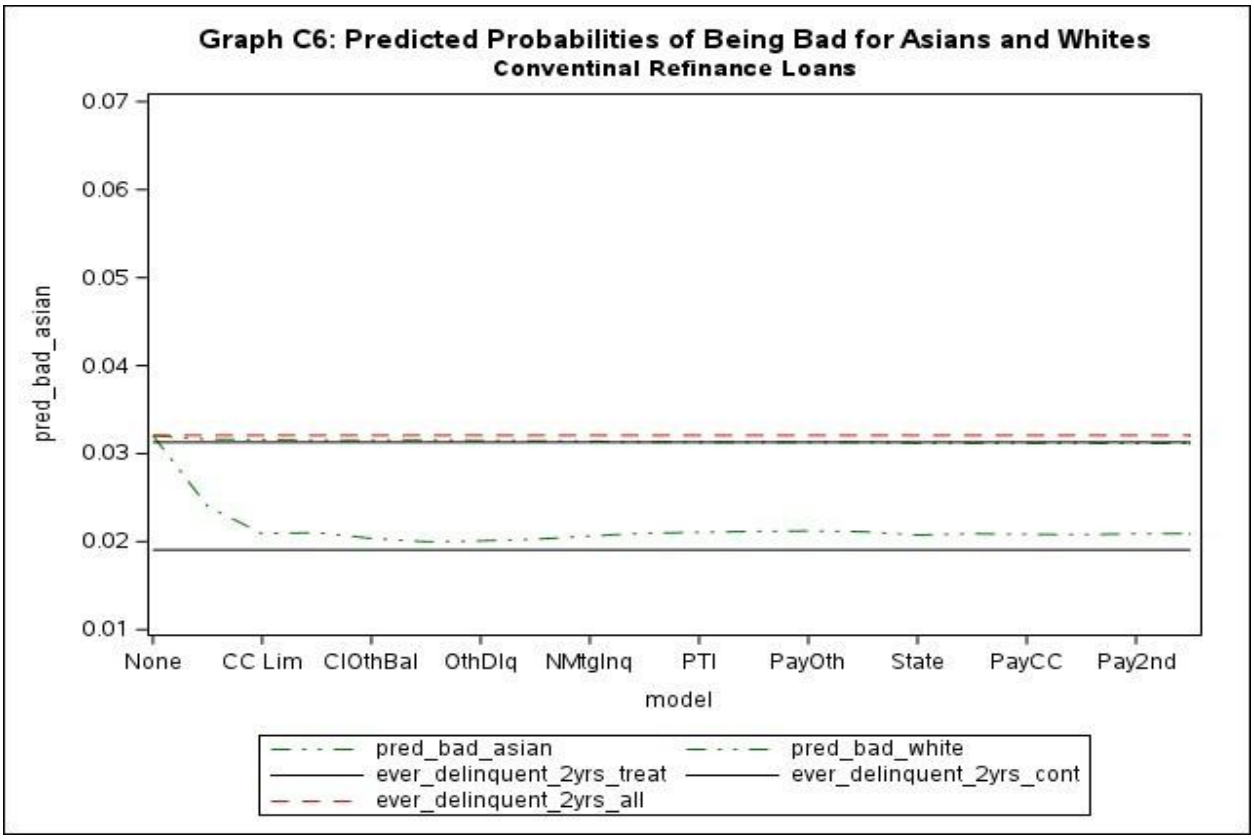

Graph C1: Predicted Probabilities of Being Bad for Blacks and Whites
Conventional Home Purchase Loans
pred_bad_black
0.10
0.08
0.06
0.04
0.02
None
CC Lim
ClOthBal
OthDlq
NMtglnq
PTI
PayOth
LimOth
OpOthBal
2ndBal
model
pred_bad_black
pred_bad_white
ever_delinquent_2yrs_treat
ever_delinquent_2yrs_cont
ever_delinquent_2yrs_all

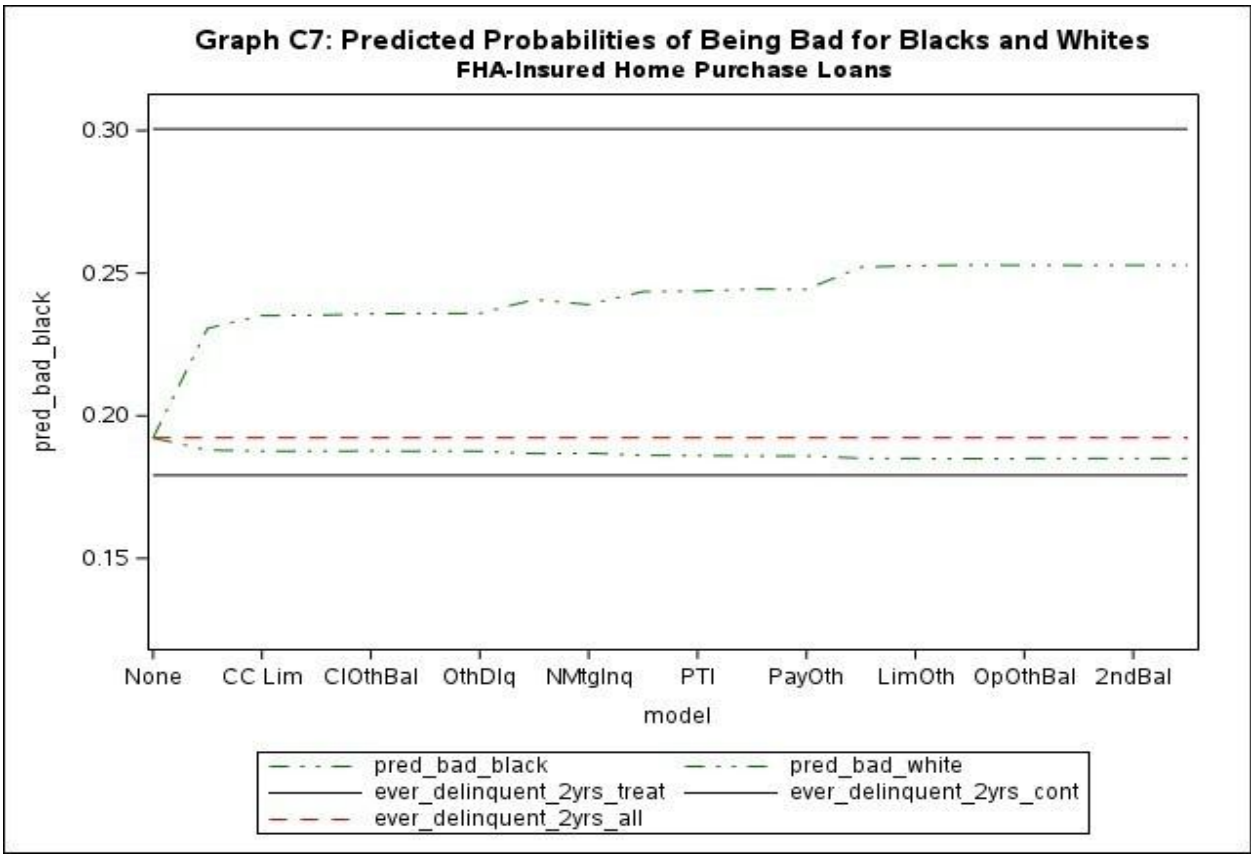

Graph C2: Predicted Probabilities of Being Bad for Hispanics and non-Hispanics
Conventional Home Purchase Loans
pred_bad_hispanic
0.10
0.08
0.06
0.04
0.02
None
CC Lim
ClOthBal
OthDlq
NMtglnq
PTI
PayOth
LimOth
OpOthBal
2ndBal
model
pred_bad_hispanic
pred_bad_nonhispanic
ever_delinquent_2yrs_treat
ever_delinquent_2yrs_cont
ever_delinquent_2yrs_all

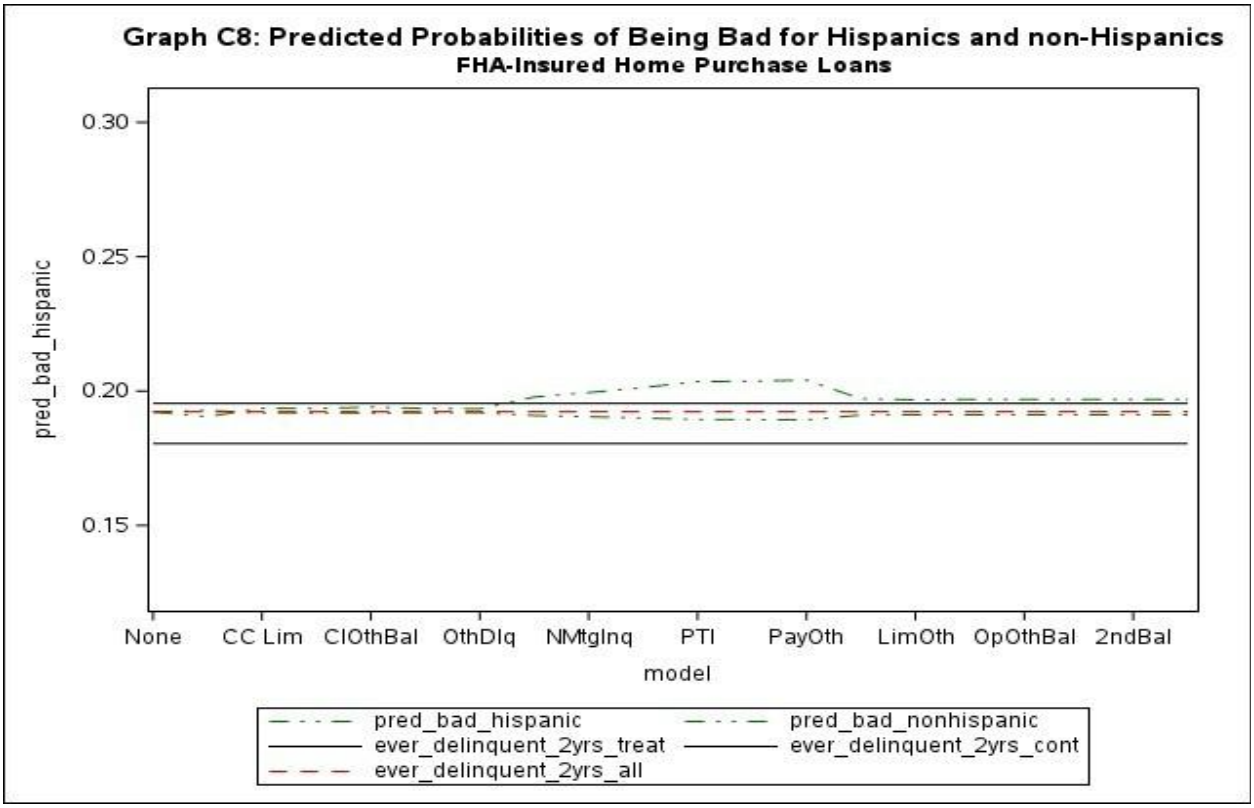

Graph C3: Predicted Probabilities of Being Bad for Asians and Whites
Conventional Home Purchase Loans
pred_bad_asian
0.10
0.08
0.06
0.04
0.02
None
CC Lim
ClOthBal
OthDlq
NMtgInq
PTI
PayOth
LimOth
OpOthBal
2ndBal
model
pred_bad_asian
pred_bad_white
ever_delinquent_2yrs_treat
ever_delinquent_2yrs_cont
ever_delinquent_2yrs_all

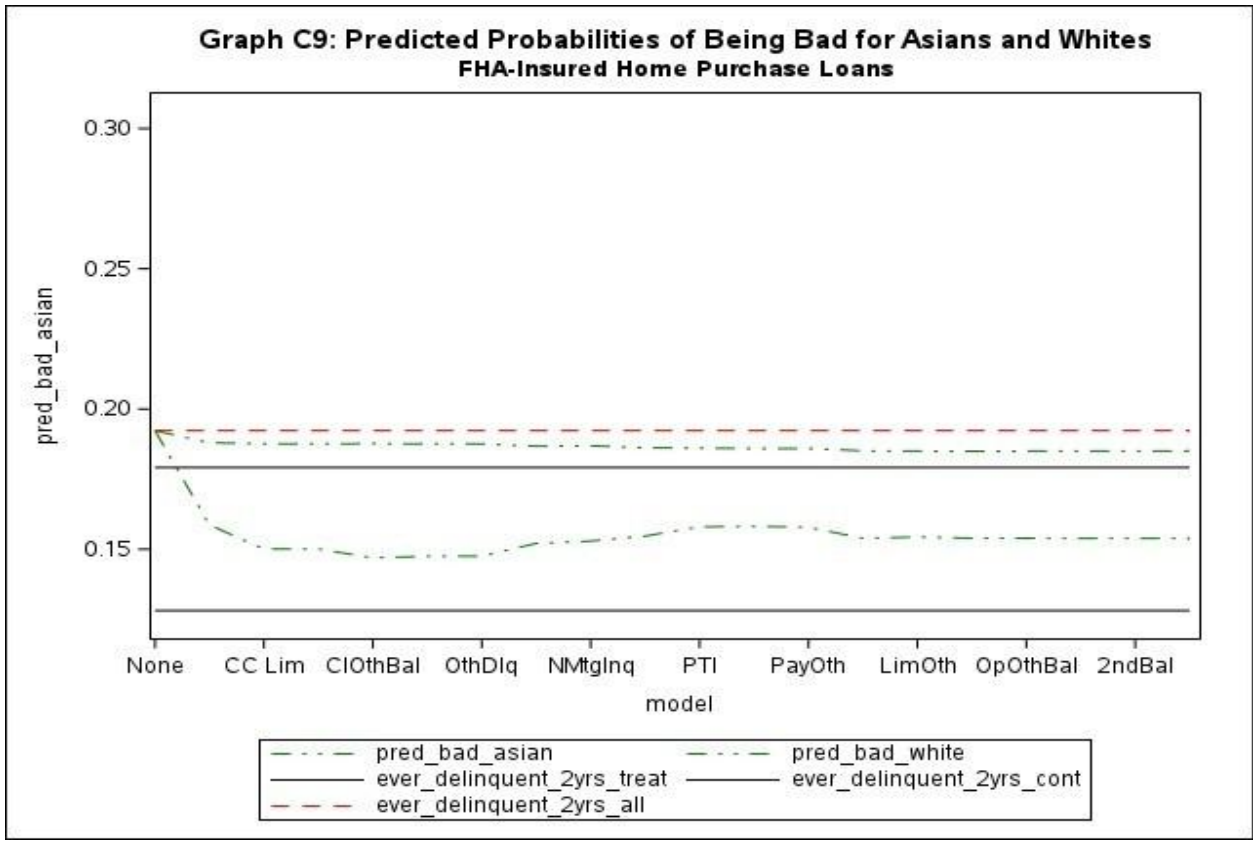

Graph C4: Predicted Probabilities of Being Bad for Blacks and Whites
Conventinal Refinance Loans
pred_bad_black
0.07
0.06
0.05
0.04
0.03
0.02
0.01
None
CC Lim
ClOthBal
OthDlq
NMtgInq
PTI
PayOth
State
PayCC
Pay2nd
model
pred_bad_black
pred_bad_white
ever_delinquent_2yrs_treat
ever_delinquent_2yrs_cont
ever_delinquent_2yrs_all

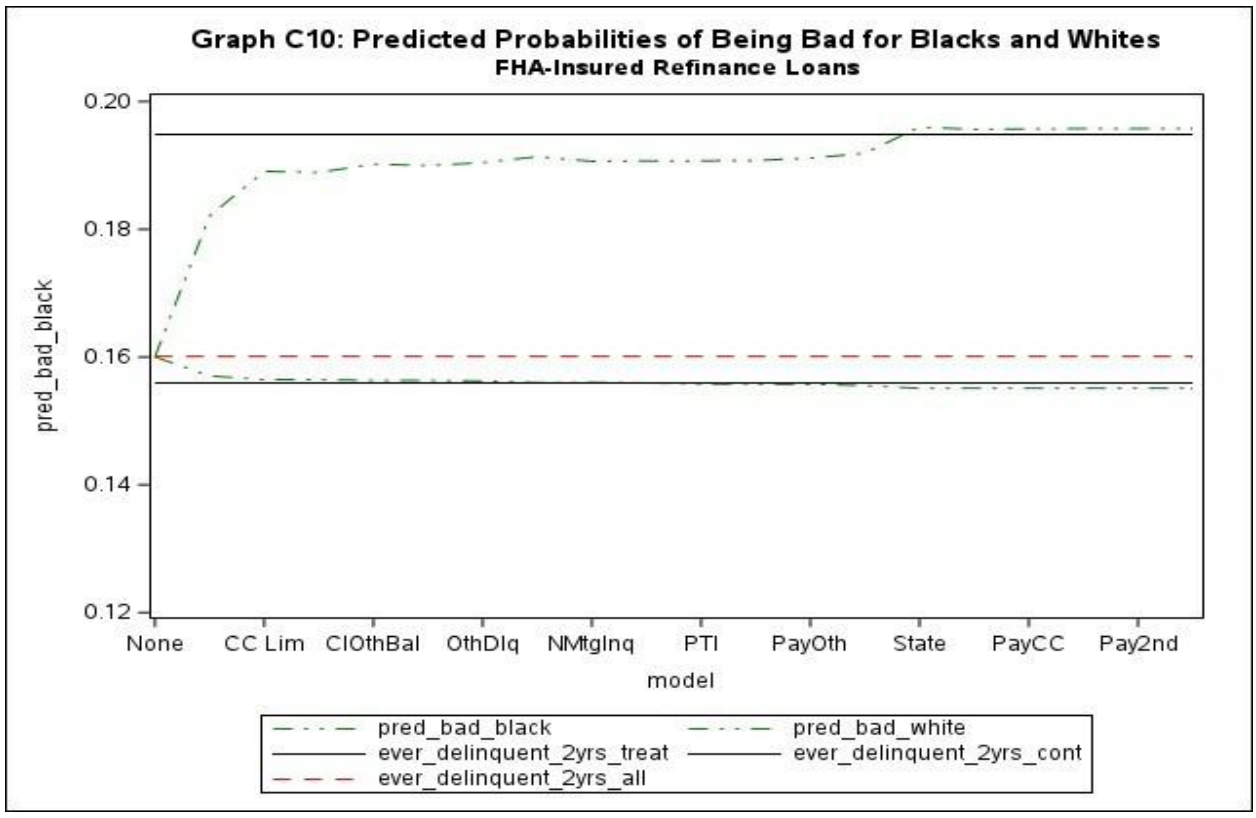

Graph C5: Predicted Probabilities of Being Bad for Hispanics and non-Hispanics
Conventinal Refinance Loans
pred_bad_hispanic
0.07
0.06
0.05
0.04
0.03
0.02
0.01
None
CC Lim
ClOthBal
OthDlq
NMtglnq
PTI
PayOth
State
PayCC
Pay2nd
model
pred_bad_hispanic
pred_bad_nonhispanic
ever_delinquent_2yrs_treat
ever_delinquent_2yrs_cont
ever_delinquent_2yrs_all

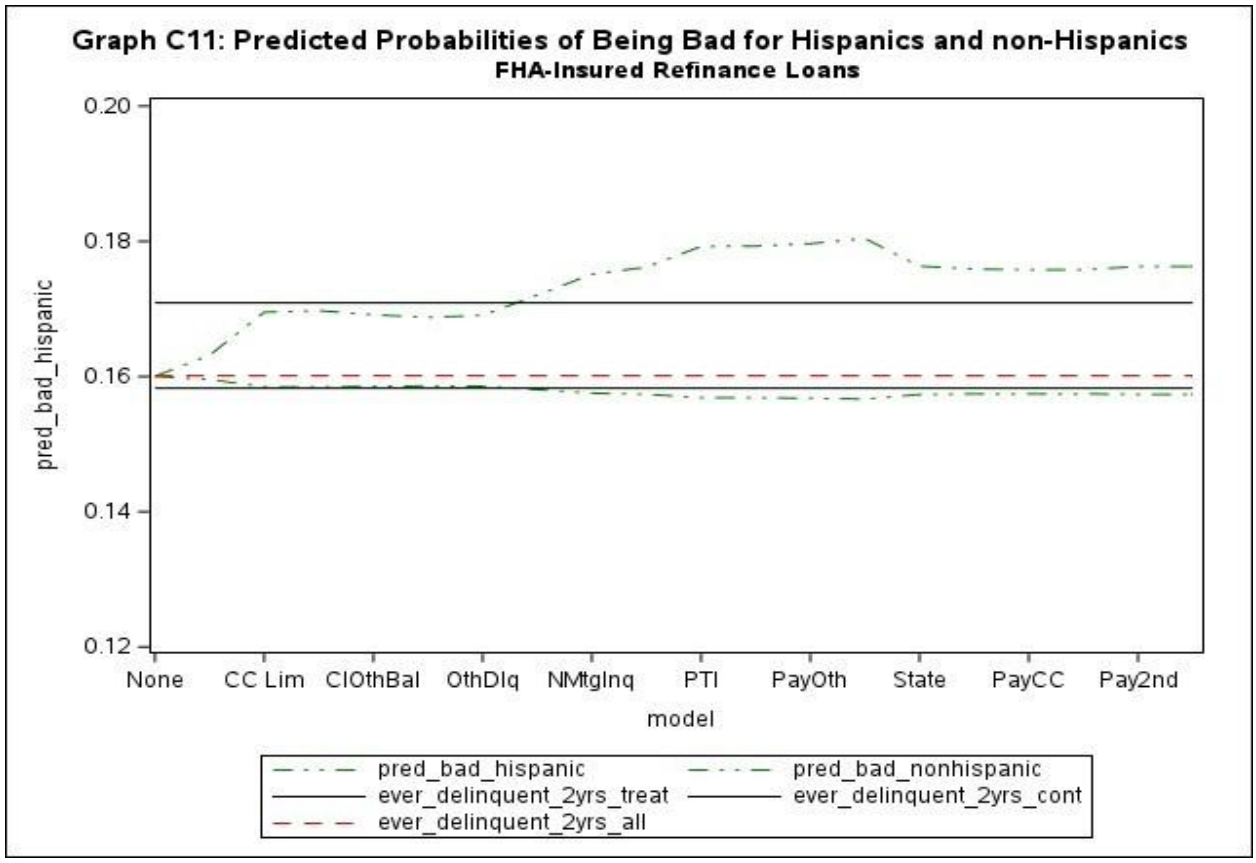

Graph C6: Predicted Probabilities of Being Bad for Asians and Whites
Conventinal Refinance Loans
pred_bad_asian
0.07
0.06
0.05
0.04
0.03
0.02
0.01
None
CC Lim
ClOthBal
OthDlq
NMtglnq
PTI
PayOth
State
PayCC
Pay2nd
model
pred_bad_asian
pred_bad_white
ever_delinquent_2yrs_treat
ever_delinquent_2yrs_cont
ever_delinquent_2yrs_all

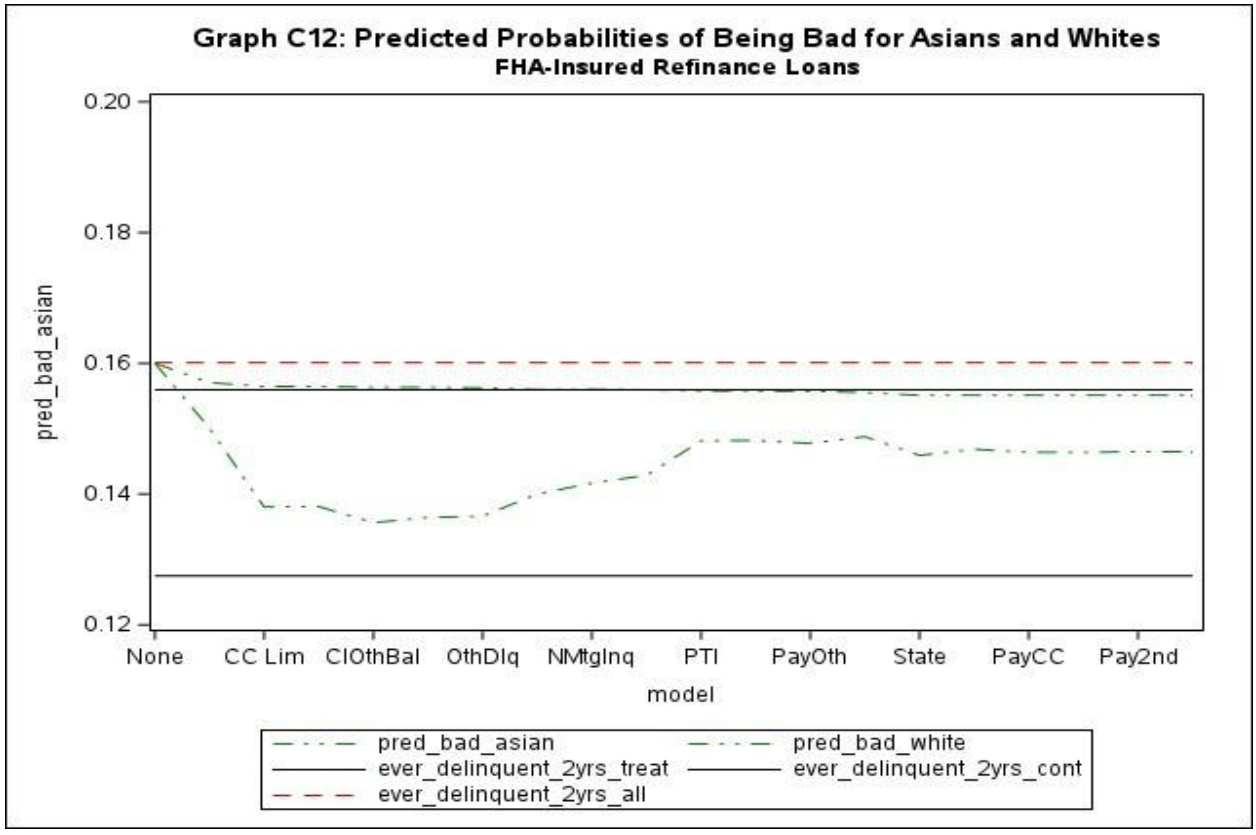

Graph C7: Predicted Probabilities of Being Bad for Blacks and Whites
FHA-Insured Home Purchase Loans
pred_bad_black
0.30
0.25
0.20
0.15
None
CC Lim
ClOthBal
OthDlq
NMtgInq
PTI
PayOth
LimOth
OpOthBal
2ndBal
model
pred_bad_black
pred_bad_white
ever_delinquent_2yrs_treat
ever_delinquent_2yrs_cont
ever_delinquent_2yrs_all


Graph C8: Predicted Probabilities of Being Bad for Hispanics and non-Hispanics
FHA-Insured Home Purchase Loans
pred_bad_hispanic
0.30
0.25
0.20
0.15
None
CC Lim
ClOthBal
OthDlq
NMtgInq
PTI
PayOth
LimOth
OpOthBal
2ndBal
model
pred_bad_hispanic
pred_bad_nonhispanic
ever_delinquent_2yrs_treat
ever_delinquent_2yrs_cont
ever_delinquent_2yrs_all


Graph C9: Predicted Probabilities of Being Bad for Asians and Whites
FHA-Insured Home Purchase Loans
pred_bad_asian
0.30
0.25
0.20
0.15
None
CC Lim
ClOthBal
OthDlq
NMtgInq
PTI
PayOth
LimOth
OpOthBal
2ndBal
model
pred_bad_asian
pred_bad_white
ever_delinquent_2yrs_treat
ever_delinquent_2yrs_cont
ever_delinquent_2yrs_all


Graph C10: Predicted Probabilities of Being Bad for Blacks and Whites
FHA-Insured Refinance Loans
pred_bad_black
0.20
0.18
0.16
0.14
0.12
None
CC Lim
ClOthBal
OthDlq
NMtgInq
PTI
PayOth
State
PayCC
Pay2nd
model
pred_bad_black
pred_bad_white
ever_delinquent_2yrs_treat
ever_delinquent_2yrs_cont
ever_delinquent_2yrs_all


Graph C11: Predicted Probabilities of Being Bad for Hispanics and non-Hispanics
FHA-Insured Refinance Loans
pred_bad_hispanic
0.20
0.18
0.16
0.14
0.12
None
CC Lim
ClOthBal
OthDlq
NMtgInq
PTI
PayOth
State
PayCC
Pay2nd
model
pred_bad_hispanic
pred_bad_nonhispanic
ever_delinquent_2yrs_treat
ever_delinquent_2yrs_cont
ever_delinquent_2yrs_all


Graph C12: Predicted Probabilities of Being Bad for Asians and Whites
FHA-Insured Refinance Loans
pred_bad_asian
0.20
0.18
0.16
0.14
0.12
None
CC Lim
ClOthBal
OthDlq
NMtgInq
PTI
PayOth
State
PayCC
Pay2nd
model
pred_bad_asian
pred_bad_white
ever_delinquent_2yrs_treat
ever_delinquent_2yrs_cont
ever_delinquent_2yrs_all