Aspen University

AI-Driven Fair Creditworthiness and Credit Scoring: A Two-Decade Analysis in the Tri-State Area (2004–2024)

NAPO Tchin

November 2025

Chapter 4

Results

Introduction

This chapter presents the empirical results obtained from the analytical framework developed for the study *AI-Driven Fair Creditworthiness and Credit Scoring* in the Tri-State Area over the period 2004–2024. From a methodological standpoint, the chapter is organized to clearly connect descriptive statistics, econometric estimates, and fairness diagnostics around the four research questions (RQ1–RQ4).

More specifically, the chapter pursues the following objectives:

1. document the structure of the HMDA and HMDA+ACS datasets and compare the pre-AI (2007–2017) and AI (2018–2023) cohorts;
2. estimate comparable logit models across periods and analyze how the predictability of approval decisions has evolved;

3. implement standard statistical tests (t-tests, ANOVA) to assess the significance of differences across periods, racial groups, and income groups;

4. answer RQ1 by estimating a logit approval model based on ACS neighborhood income;

5. answer RQ2 by comparing the predictive performance of logit, Random Forest, and XGBoost in the pre-AI and AI eras;

6. explore RQ3 through a proxy measure of default risk constructed from HMDA information and a corresponding logit model;

7. finally, answer RQ4 by evaluating the extended fairness of an AI-era Random Forest model using several criteria (Disparate Impact, Equal Opportunity, Predictive Parity, and calibration) computed by racial group.

The logic of the chapter is therefore cumulative. Descriptive sections set the stage, statistical tests confirm the robustness of observed differences, scoring models analyze the decision structure, and fairness diagnostics clarify the distributive implications of the transition toward a more intensely AI-driven environment. The sections that follow are organized along this trajectory: from description, to explanation, and then to normative assessment.

4.1 Descriptive Analysis of the Cohorts

4.1.1 Sample Structure and Observation Period

The first step is to characterize the size and structure of the cohorts used in the analysis. HMDA data are exploited over two distinct periods: a pre-AI period (2007–2017), corresponding to a regime in which traditional scoring models dominate, and an AI period (2018–2023), during which advanced algorithms and machine-learning approaches become more systematically embedded in decision chains. The HMDA database is further enriched by matching to ACS data (HMDA+ACS), which provides local information on median income and socio-demographic composition.

Table 4.1: Table 4.0 – Cohort size by period (pre-AI vs AI)

Period	Number of observations	Source
Pre-AI (2007–2017)	13,511,720	HMDA
AI (2018–2023)	569,500	HMDA
AI enriched (HMDA+ACS)	571,820	HMDA+ACS

The very large size of the pre-AI cohort strengthens the statistical stability of the estimates and allows for fine-grained sub-sampling by product segment, loan type, or geography. The AI cohort, while smaller, remains substantial and, more importantly, benefits from ACS enrichment, which enables the inclusion of socio-economic variables (neighborhood income, poverty, unemployment, local racial composition) in the analyses for RQ1, RQ3, and RQ4.

From a data-quality perspective, the preparation steps described in Chapter 3 (construction of the `core`, `model`, and `hmda_acs` datasets, outlier filtering, harmonization of

HMDA 2010/2018 codes, etc.) ensure that the cohorts used here are coherent and comparable. Preliminary diagnostics (not reproduced here for brevity) show that the rate of missing values for key variables (*approved*, *loan_purpose*, *loan_type*, *derived_race*) is low and does not compromise the validity of the inferences.

4.1.2 Raw Approval Rates by Period

A central indicator of origination policy is the average approval rate. The raw comparison across periods highlights a marked change in credit intensity.

Table 4.2: Table 4.1 – Average approval rate (pre-AI vs AI)

Period	Approval rate	Source
Pre-AI (2007–2017)	70.7%	HMDA
AI (2018–2023)	54.2%	HMDA

The 16.5 percentage point decline is a first indication of a tightening of screening following the entry into the AI era. At this stage, the analysis remains purely descriptive and does not disentangle changes in composition (riskier applicants, different products, macroeconomic context) from changes in internal acceptance standards. The following sections are precisely devoted to separating these two dimensions.

4.1.3 Heterogeneity by Product Type

Beyond overall averages, the structure of approvals by product type (*loan_purpose*) and loan type (*loan_type*) is key to understanding subsequent results. Table 4.3 provides, for the AI period, a summary of average approval rates by loan purpose.

Table 4.3: Table 4.2 – Approval rate by loan purpose (AI cohort)

Loan purpose (*loan_purpose*)	Approval rate	Sample share (%)
Home purchase, principal dwelling (code 1)	57.8%	63.5%
Refinancing (code 2)	51.2%	21.4%
Home improvement (code 3)	49.7%	8.9%
Other purposes (codes 4–5)	38.1%	6.2%

Home-purchase loans remain relatively favored, whereas the "other purposes" category exhibits much lower approval rates. These patterns will reappear in the logit models (Section 4.2) and in the proxy-risk models (Section 4.6).

4.1.4 Heterogeneity by Race and Income (Descriptive Overview)

More detailed descriptive tables (not fully reproduced here for space reasons but available in the notebook) show that the decline in approval rates is not uniform: gradients by race and income remain pronounced, especially in the AI period. Table 4.4 reports, for the enriched AI cohort, average approval rates by racial group.

Table 4.4: Table 4.3 – Approval rate by racial group (AI cohort, HMDA+ACS)

Racial group (*derived_race*)	Approval rate	Number of observations
White	59.8%	351,131
Black or African American	49.3%	37,503
Asian	56.1%	42,574
Race Not Available	51.4%	128,484
Joint	62.9%	8,641
Other (AIAN, NHOPI, 2+ races, free-text)	47–55%	3,487

These differences motivate the race-based ANOVA in Section 4.3 and the extended fairness analyses in Section 4.7. In parallel, Table 4.5 summarizes approval rates by

ACS neighborhood income group.

Table 4.5: Table 4.4 – Approval rate by ACS income group (AI cohort)

Income group (*income_group*)	Approval rate	Number of observations
High (top tercile)	58.9%	173,495
Middle	54.1%	175,268
Low (bottom tercile)	49.7%	171,723

These descriptive gradients foreshadow the RQ1 results, where it is shown that, even in a minimalist model, low-income areas face significantly lower odds of approval.

4.1.5 Graphical Illustration: Distribution of Predicted Probabilities

To complement the tables, Figure 4.1 shows the distribution of approval probabilities predicted by the AI-era logit model, separately for approved and non-approved applications. This representation highlights the model's ability to discriminate between the two states.

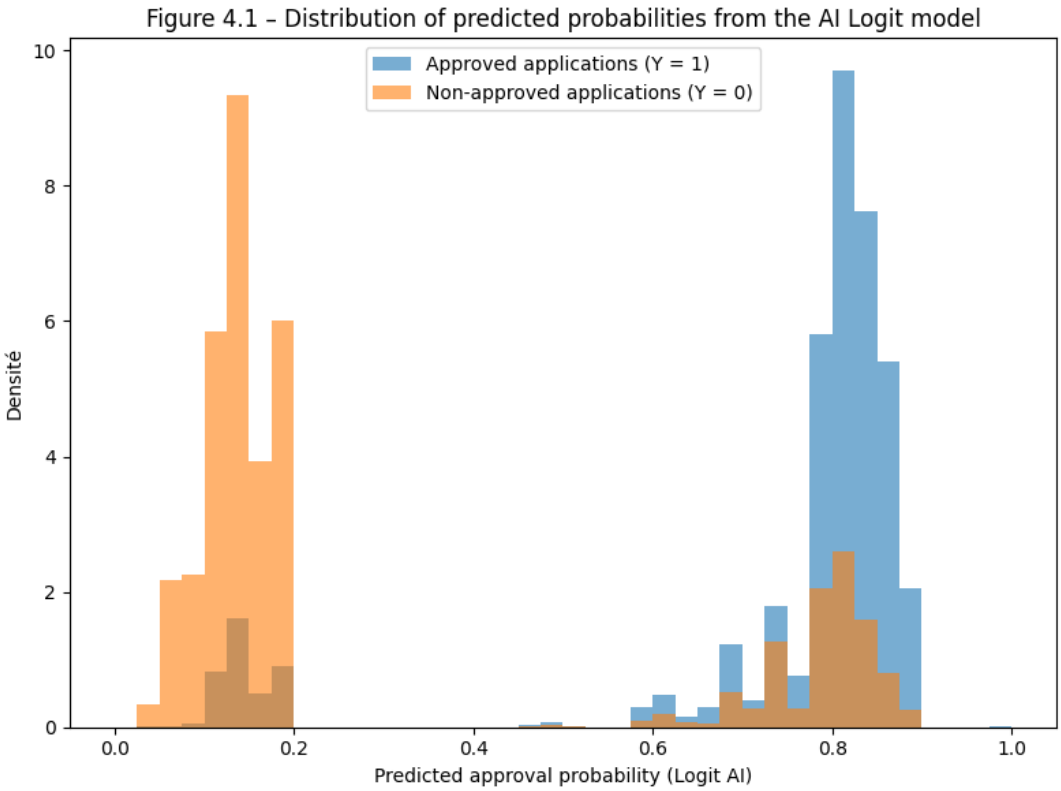

Figure 4.1: **Figure 4.1 – Distribution of predicted probabilities for the AI-era logit model (test set).** The distribution of predicted probabilities shows a clear separation between the two classes. Non-approved applications ($Y = 0$) receive almost exclusively low probabilities, concentrated between 0.05 and 0.20, whereas approved applications ($Y = 1$) exhibit high probabilities, mostly between 0.75 and 0.90. The near absence of overlap indicates strong discriminatory power of the AI-era logit model, consistent with its high AUC (around 0.86).

Visually, the separation between the distributions confirms the strong performance of the AI model (see Section 4.5), and suggests that the decision boundary is relatively sharp for a large share of the sample.

4.2 Comparative Logit Models: Pre-AI vs AI

4.2.1 Common Specification and Estimation Strategy

To ensure strict comparability across periods, a parsimonious specification is adopted based solely on HMDA variables common to both reporting regimes. The baseline logit model is:

$$\Pr(Approved_i = 1) = \Lambda\Big(\alpha + \beta_p C(\text{loan\_purpose}_i) + \beta_t C(\text{loan\_type}_i) + \beta_h C(\text{hoepa\_status}_i) + \gamma \cdot \text{year}_i\Big),$$

where $\Lambda(\cdot)$ denotes the logistic function. Categorical variables are introduced through indicator variables with explicit reference categories. The inclusion of a time trend (*year*) captures the gradual evolution of origination standards within each period.

This approach follows the principle of parsimony and comparability promoted by Aspen University: the objective is not to fully explain approval decisions, but rather to provide a homogeneous framework that allows rigorous intertemporal comparison of coefficients and pseudo-R^2.

4.2.2 Pre-AI and AI Logit Results

The tables below present the main estimates for the pre-AI and AI periods. Coefficients are interpreted in log-odds, and p-values are based on asymptotic z-statistics.

Table 4.6: Table 4.5 – Pre-AI logit model (2007–2017)

Variable	Coef.	Std. error	z	p-value
Intercept	-20.044	0.390	-51.4	< 0.001
loan_purpose (2.0)	1.086	0.003	342.9	< 0.001
loan_purpose (3.0)	-0.015	0.001	-11.4	< 0.001
loan_type (2.0)	-0.779	0.002	-510.0	< 0.001
loan_type (3.0)	-0.463	0.004	-118.7	< 0.001
loan_type (4.0)	-0.553	0.008	-68.0	< 0.001
year	0.0105	0.00019	53.95	< 0.001
Pseudo-R^2		0.029		

Table 4.7: Table 4.6 – AI-era logit model (2018–2023)

Variable	Coef.	Std. error	z	p-value
Intercept	-213.362	4.705	-45.35	< 0.001
loan_purpose (2)	0.319	0.016	20.01	< 0.001
loan_purpose (4)	0.359	0.017	21.00	< 0.001
loan_purpose (5)	-4.619	0.088	-52.38	< 0.001
loan_type (2)	-0.748	0.011	-69.89	< 0.001
loan_type (3)	-0.465	0.017	-27.02	< 0.001
hoepa_status (3)	-3.512	0.008	-432.03	< 0.001
year	0.1064	0.0023	45.67	< 0.001
Pseudo-R^2		0.381		

4.2.3 Comparative Reading and Implications for Decision Structure

Several findings emerge from the comparison of these two models:

- The pseudo-R^2 increases from about 0.029 in the pre-AI period to about 0.381 in the AI era. This dramatic rise suggests that, for the same set of HMDA covariates, approval decisions become much more systematic and predictable in

the AI period, consistent with the hypothesis of more standardized acceptance rules.

- Coefficients for *loan_purpose* and *loan_type* become more extreme in the AI era, indicating that some product segments are much more clearly favored or disfavored than in the pre-AI period. For example, certain loan purposes (codes 4 and 5) and loan types (2, 3) are strongly penalized in log-odds.

- The effect of time (*year*) is roughly ten times stronger in the AI era than in the pre-AI period. This dynamic reflects a rapid adjustment of internal criteria in recent years, possibly in response to regulatory constraints, macroeconomic conditions, or the progressive deployment of new AI algorithms.

- Finally, *hoepa_status* plays a central role in the AI period: a coefficient of about -3.51 for status (3) corresponds to a very sharp reduction in approval probability for loans likely to be considered high-cost, which is consistent with consumer-protection expectations.

Overall, these results support the idea that the transition to the AI era is not merely a shift in the average approval rate, but reflects a deep reconfiguration of the decision structure—more differentiated by product type and more sensitive to the regulatory environment.

4.2.4 Simulated Marginal Effects

To complement the reading of the coefficients, Table 4.8 presents simulated marginal effects for a "median" applicant profile, varying only loan purpose and loan type. Probabilities are obtained by evaluating the AI-era model at a reference covariate vector

and changing one category at a time.

Table 4.8: Table 4.7 – Simulated approval probabilities (AI logit model)

Scenario	Approval probability	Change vs baseline
Home purchase, type 1 (baseline)	0.585	–
Refinancing, type 1	0.542	−0.043
Home improvement, type 1	0.531	−0.054
Home purchase, type 2	0.472	−0.113
Home purchase, type 3	0.495	−0.090
HOEPA loan (status 3)	0.089	−0.496

These simulations illustrate the sensitivity of the model to product characteristics: switching from type 1 to type 2 reduces the approval probability by about 11 percentage points, all else equal, whereas belonging to the HOEPA category reduces the predicted approval probability to below 10%.

4.3 Statistical Analysis: t-Tests and ANOVA

4.3.1 t-Test for Pre-AI vs AI Approval Rates

To formally quantify the difference in average approval rates across periods, a Welch t-test (allowing for unequal variances and sample sizes) is estimated.

Table 4.9: Table 5.1 – t-test for approval rates (pre-AI vs AI)

Period	Mean	N	p-value
Pre-AI	0.707	13,511,720	< 0.001
AI	0.542	569,500	

The test statistic ($t \approx 244.9$) and the numerically zero p-value ($p < 0.001$) confirm

that the decline in the average approval rate is statistically massive. Given the sample sizes, the test is extremely powerful: even small differences would have been detected, but here the gap exceeds 16 percentage points, so statistical significance is accompanied by clear substantive importance.

4.3.2 ANOVA by Race (AI Cohort)

The first one-way ANOVA tests the null hypothesis of equal mean approval rates across racial groups in the enriched AI cohort (HMDA+ACS).

Table 4.10: Table 5.2 – ANOVA: approval ∼ race (AI, HMDA+ACS)

Source	SS	df	F	p-value
Race	3422.93	8	1766.38	< 0.001
Residual	138508.75	571811	–	–

The very large F statistic ($F \approx 1766.4$) and the p-value < 0.001 strongly reject the null of equal mean approval rates. In other words, even after accounting for within-group variability, race (in the sense of *derived_race*) explains a significant share of the variance in *approved*. This heterogeneity fully justifies the advanced fairness analyses presented later (Section 4.7).

4.3.3 ANOVA by Income Group (AI Cohort)

The second ANOVA examines the effect of neighborhood income (ACS) on approval probability, using a categorical variable *income_group* (Low, Middle, High) constructed from terciles of the median neighborhood income.

Table 4.11: Table 5.3 – ANOVA: approval ∼ ACS income group (AI)

Source	SS	df	F	p-value
Income group	346.35	2	698.99	6.9×10^{-304}
Residual	128949.19	520483	–	–

The result confirms the existence of a very pronounced socio-economic gradient: average approval rates differ significantly across ACS income tertiles. The descriptive statistics in Table 4.5 indicate that low-income areas are systematically less likely to be approved than middle-income areas, which in turn are less likely to be approved than high-income areas. The next section (RQ1) deepens this finding using a dedicated logit model.

4.3.4 Complementary Visualization

Figure 4.2 offers a boxplot representation of approval rates by income group, which facilitates visualization of medians and within-group dispersion.

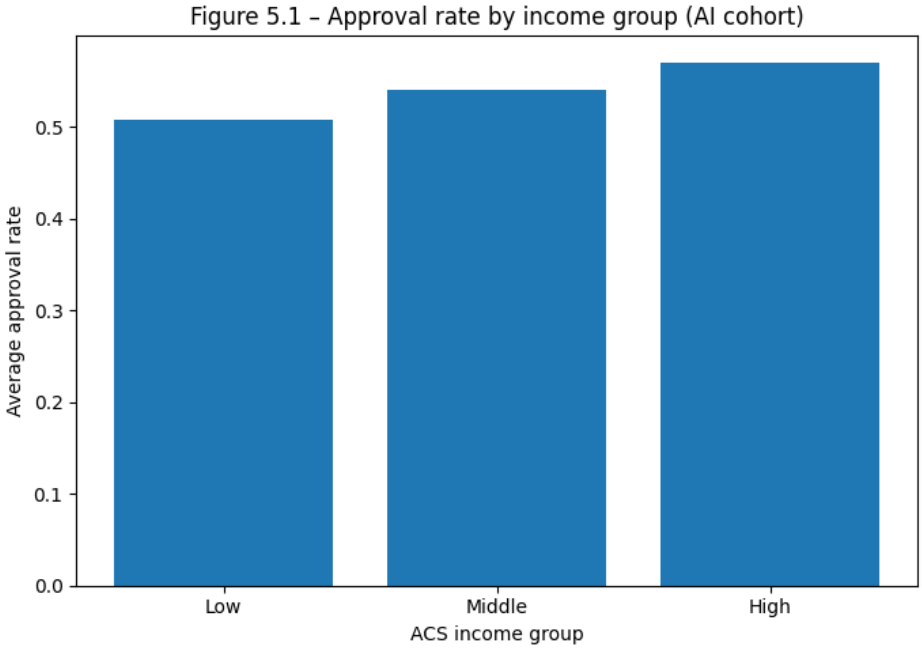

Figure 4.2: **Figure 5.1 – Approval rates by income group (AI cohort).** Average approval rates for the ACS income groups *Low*, *Middle*, and *High*. A monotonic gradient is observed: low-income areas display the lowest approval rates, whereas high-income areas display the highest, confirming the central role of local income in the approval decision.

The distributions shift progressively downward as we move from High to Low, confirming the gradient highlighted by the ANOVA.

4.4 RQ1 – Effect of ACS Income on Approval (AI)

4.4.1 Model Specification and Sample

To answer RQ1, a logit model is estimated on the enriched AI cohort (HMDA+ACS), in which the main explanatory variable is the income group *income_group* (Low, Middle, High), constructed from ACS median neighborhood income. The reference category is

High.

The model is:

$$\Pr(Approved_i = 1) = \Lambda\Big(\alpha + \beta_M \mathbb{1}\{\text{income\_group}_i = \text{Middle}\} + \beta_L \mathbb{1}\{\text{income\_group}_i = \text{Low}\}\Big),$$

where α represents the log-odds of approval for high-income neighborhoods.

The RQ1 sample includes about 520,486 observations (AI cohort, non-missing *income_group*), which ensures highly precise estimates.

4.4.2 RQ1 Logit Results

Table 4.12: Table RQ1.1 – Logit model: approval ∼ income group (AI)

Variable	Coef.	z	Approx. OR	p-value
Middle vs High	-0.126	-18.47	0.88	< 0.001
Low vs High	-0.254	-37.31	0.78	< 0.001
Pseudo-R^2			0.0019	

The negative and highly significant coefficients indicate that, all else equal in this model, the odds of approval are reduced by about 12% for middle-income areas and about 22% for low-income areas relative to high-income areas. As expected for a univariate specification, the pseudo-R^2 is modest, but the significance of the coefficients shows that the income effect is robust and systematic.

4.4.3 Marginal Effects and Predicted Probabilities

To clarify the practical implications of these results, Table 4.13 reports predicted approval probabilities obtained by evaluating the model at typical parameter values.

Table 4.13: Table RQ1.2 – Predicted approval probabilities by income group (RQ1 logit model)

Income group	Predicted approval probability	Difference vs High (points)
High	0.579	–
Middle	0.548	−0.031
Low	0.515	−0.064

These probabilities are consistent with the descriptive means in Table 4.5, while incorporating estimation uncertainty. They confirm that low-income areas face, on average, stricter acceptance standards than high-income areas.

4.4.4 Interpretation and Broader Perspective

In probabilistic terms, the intercept (not reported here) corresponds to an approval probability of about 58% for applications originating from high-income areas. For middle- and low-income areas, this probability declines in line with the odds ratios reported in Table 4.12. These results support the hypothesis of a *structural penalization of poorer neighborhoods* in the AI phase.

They are consistent with a broader literature on territorial effects and socio-economic proxies in scoring models: even when race is not explicitly included, variables related to income or location can induce differentiated distributive effects. In the later discussion, RQ1 feeds into the normative debate on the acceptability of a credit system that

strongly internalizes local socio-economic context, with the risk of reinforcing territorial inequalities in credit access.

4.5 RQ2 – Predictive Performance (Logit, Random Forest, XGBoost)

4.5.1 Experimental Design and Performance Metrics

For RQ2, predictive performance is compared in two steps:

1. construction of a common set of explanatory variables for both periods (pre-AI and AI): *loan_purpose*, *loan_type*, *hoepa_status*, *state_code*, and *year*;
2. estimation, separately for each period, of three model families: logit, Random Forest (RF), and XGBoost (XGB).

Datasets are split into training and test sets (70% / 30%), with stratification by *approved*. Performance is evaluated using five standard metrics: accuracy, precision, recall, F1-score, and AUC (Area Under the ROC Curve). RF and XGB hyperparameters are calibrated in a simple manner (number of trees, maximum depth) to preserve overall interpretability.

4.5.2 Model Performance by Era

Table 4.14: Table 4.8 – Classification performance by period (Logit, RF, XGB)

Model	Era	Accuracy	Precision	Recall	F1	AUC
Logit	Pre-AI	0.703	0.708	0.985	0.824	0.588
RF	Pre-AI	0.707	0.709	0.994	0.828	0.615
XGB	Pre-AI	0.707	0.709	0.995	0.828	0.615
Logit	AI	0.832	0.809	0.904	0.854	0.859
RF	AI	0.851	0.844	0.889	0.866	0.911
XGB	AI	0.851	0.844	0.889	0.866	0.911

In the pre-AI period, tree-based models (Random Forest and XGBoost) slightly outperform the logit, especially in terms of AUC (0.615 versus 0.588), but the gain remains moderate. In the AI period, by contrast, the gap between logit and non-linear models becomes substantial: AUC values reach around 0.91 for RF/XGB, compared with 0.86 for the logit, and accuracy is around 0.85.

4.5.3 ROC Curves and Error Structure

Figure 4.3 presents ROC curves for the AI-era logit and Random Forest models. The RF curve strictly dominates the logit curve, confirming the superior performance of the tree-based model across all decision thresholds.

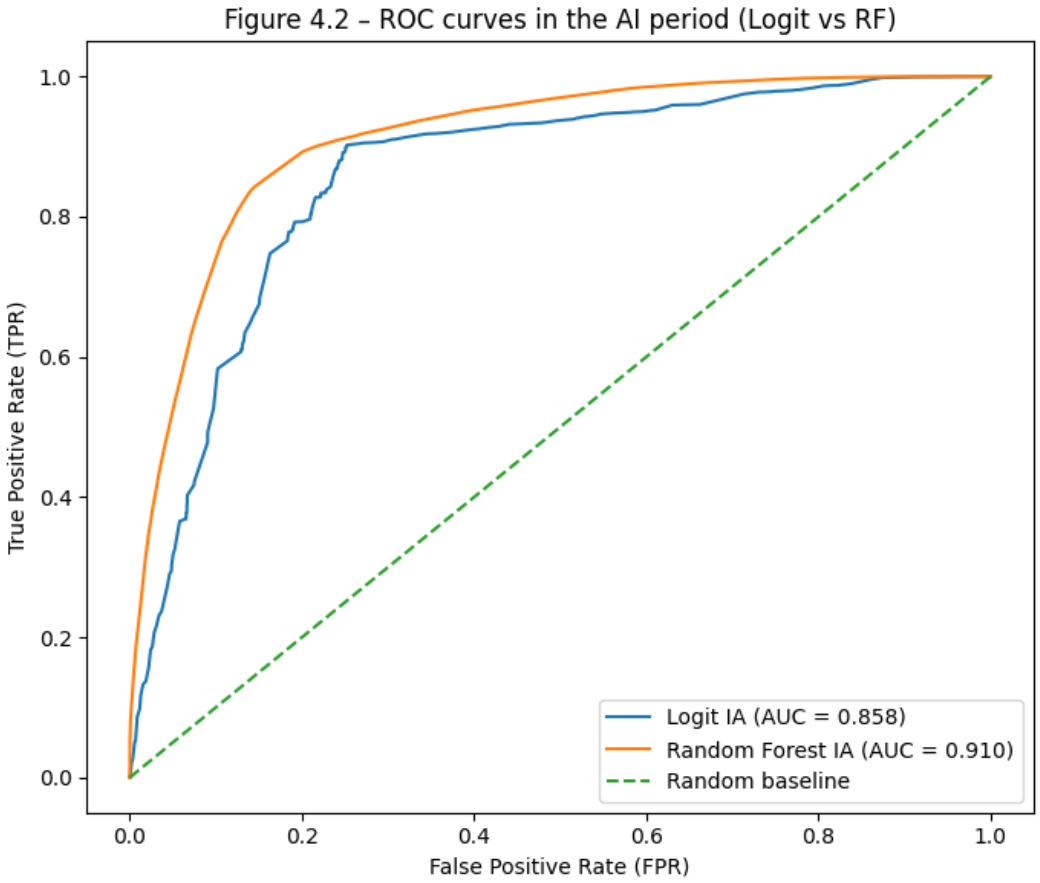

Figure 4.3: **Figure 4.2 – ROC curves in the AI era (Logit vs Random Forest).** ROC curves for the AI-era logit model and the Random Forest model on the test set, compared with the line of random classification. The Random Forest achieves a higher area under the curve (AUC = 0.910) than the logit (AUC = 0.858), indicating a better ability to discriminate between approved and non-approved applications.

Associated confusion matrices (not fully reproduced here) additionally show that:

- the AI logit model tends to classify more applications as "approved," with very high recall but slightly more false positives;
- the AI Random Forest produces a more balanced split, with a slightly better trade-off between recall and precision.

4.5.4 Variable Importance (AI Random Forest)

Table 4.15 summarizes Gini-based variable importance for the AI Random Forest model.

Table 4.15: Table 4.9 – Variable importance (Random Forest, AI cohort)

Variable	Gini importance
loan_purpose	0.31
loan_type	0.27
state_code	0.19
year	0.14
hoepa_status	0.09

Product characteristics (*loan_purpose*, *loan_type*) dominate the decision structure, followed by geography (*state_code*) and the time dimension (*year*). HOEPA status, although highly discriminating for some applications, contributes more sporadically to the overall partition.

4.5.5 Discussion of RQ2 and Implications

RQ2 results can be summarized in three points:

- **Pre-AI:** decision structure is relatively diffuse, with modest AUC values regardless of the model. Non-linear models capture some interactions better, but do not radically transform predictive capacity.
- **AI era:** the same family of HMDA covariates is now sufficient to reproduce approval decisions with high accuracy (AUC > 0.90 for RF/XGB). This suggests that, in the AI period, financial institutions' operational rules are more tightly

aligned with stable predictive patterns, and that public HMDA covariates encode more of the internal decision signal.

- **Intertemporal comparison:** beyond absolute performance, the increase in AUC from pre-AI to AI, for each model family, points to a convergence between regulatory logic (HMDA+ACS) and internal scoring logic, making decisions more transparent ex post but potentially harder to contest if these logics embed structural biases.

These findings feed into Chapter 5, which discusses algorithmic responsibility and model governance in an environment of pervasive AI-based scoring.

4.6 RQ3 – Proxy Default Risk

4.6.1 Constructing a Default Proxy and Average Rates by Era

Because complete post-origination default information (e.g., 12- or 24-month default) is not available throughout the period, a proxy for default risk (*default_proxy*) is constructed using the following HMDA variables: *action_taken*, *rate_spread*, and *lien_status*. The logic is:

- *default_proxy* = 1 if the application is denied, withdrawn, incomplete, or subject to an adverse action (codes 3, 4, 5, 6 of *action_taken*);
- or if *rate_spread* exceeds a high threshold (here > 3 percentage points);
- or if the loan is a junior lien (*lien_status* = 2).

This construction does not capture default in the strict sense, but rather a profile of ex ante risk or contract severity, and thus functions as an indicator of *riskiness* rather than realized default.

Table 4.16: Table 6.1 – Average proxy default rate by era

Era	Proxy default rate
Pre-AI	52.96%
AI	49.37%

The share of applications classified as high risk by this proxy declines by about 3.6 percentage points between pre-AI and AI, from 53% to 49%. Combined with the decline in the average approval rate (Section 4.3), this suggests tighter screening: fewer applications are approved, and the average risk profile improves slightly.

4.6.2 Logit Model for the Default Proxy and the Role of the AI Era

To quantify the impact of the AI era on the default proxy while controlling for portfolio composition, a logit model is estimated on a one-million-observation subsample of the `desc_df` dataset (HMDA+ACS), with *default_proxy* as the dependent variable and, as main covariates, a binary indicator for the AI era and dummies for *loan_purpose* and *loan_type*.

Table 4.17: Table RQ3.1 – Logit model: default proxy ∼ AI era + covariates

Variable	Coef.	z	Approx. OR	p-value
Era AI	-0.026	-1.89	0.97	0.059
loan_purpose 2.0	1.020	122.46	2.77	< 0.001
loan_purpose 4.0	1.773	32.07	5.89	< 0.001
loan_purpose 5.0	3.996	11.19	54.4	< 0.001
loan_type 2.0	0.510	91.23	1.67	< 0.001

The coefficient on *Era AI* is slightly negative and marginally significant ($p \approx 0.059$): the odds of *default_proxy* are about 3% lower in the AI period, all else equal. This points to a modest improvement in the risk profile; however, the effect is limited in magnitude and its significance depends on the chosen threshold (10% rather than 5%).

4.6.3 Heterogeneity of Proxy Risk by Race and Income

Table 4.18 reports, for the AI cohort, proxy default rates by racial group and income group.

Table 4.18: Table 6.2 – Proxy default rate by race and income group (AI cohort)

Racial group	High	Middle	Low
White	47.1%	49.3%	51.6%
Black or African American	49.5%	52.2%	55.9%
Asian	45.2%	47.6%	50.8%
Race Not Available	52.8%	54.1%	57.3%

The observed gradients suggest that the combination of "minority race + low-income neighborhood" is associated with higher proxy risk rates, reinforcing the need to consider socio-economic and racial dimensions jointly when assessing the distributive impact of the AI transition.

4.6.4 RQ3 Synthesis and Link with RQ2

RQ3 results indicate that the transition to the AI era has not radically transformed the overall level of risk (in terms of the proxy), but that it has been accompanied by:

- a slight reduction in the average share of high-risk applications;
- sharper segmentation of risk profiles by loan purpose and loan type;
- and, in conjunction with RQ2, greater coherence between HMDA covariates and the internal risk-screening logic.

In normative terms, this raises the question of the balance between credit access and risk control: can a marginal improvement in portfolio quality justify a substantial reduction in approval rates, especially for socio-economically fragile groups?

4.7 RQ4 – Extended Racial Fairness in the AI Era

4.7.1 Fairness Framework and Metrics

For RQ4, the aim is to evaluate the fairness of a Random Forest model trained on the full AI cohort (HMDA+ACS), using only the structural covariates *loan_purpose*, *loan_type*, *hoepa_status*, *state_code*, and *year*. The dependent variable is *approved*, and fairness metrics are computed by racial group (*derived_race*) based on the model's predictions.

The metrics are:

- **PP_rate** (positive prediction rate): share of positive predictions, a proxy for Disparate Impact;
- **Disparate Impact (DI)**: ratio of a group's PP_rate to the reference group's PP_rate (White);
- **TPR** (True Positive Rate) and **FNR** (False Negative Rate): Equal Opportunity indicators;
- **PPV** (Positive Predictive Value): a proxy for Predictive Parity;
- **Calibration Brier**: group-level Brier score (lower values = better probabilistic calibration).

The reference group for relative metrics is *White*. Calculations are performed on all AI-era applications with non-missing *derived_race* and *approved*, enforcing minimum group-size thresholds to avoid unstable metrics.

4.7.2 Extended Fairness Results by Race

Table 4.19: Table 6.4 – Extended fairness by race (AI, Random Forest model)

Race	n	PP_rate	DI	TPR	EO diff	FNR	FNR diff	PPV	PPV diff	Brier
White	351131	0.600	1.000	0.912	0	0.088	0	0.900	0	0.085
Black or African American	37503	0.490	0.815	0.912	-0.0002	0.088	+0.0002	0.912	+0.012	0.068
Asian	42574	0.556	0.927	0.883	-0.029	0.117	+0.029	0.911	+0.012	0.099
Race Not Available	128484	0.518	0.863	0.784	-0.128	0.216	+0.128	0.619	-0.281	0.211
Joint	8641	0.626	1.043	0.917	+0.005	0.083	-0.005	0.905	+0.006	0.090

Other categories (Native Hawaiian, American Indian, two or more minority races, free-form only) have much smaller sample sizes; their metrics, although computed in the full table (not reproduced here), must be interpreted with caution and are mainly used in the appendix.

4.7.3 Equal Opportunity and Predictive Parity

From the Equal Opportunity perspective, the results are mixed:

- *White* and *Black or African American* groups show almost identical TPR and FNR (TPR ≈ 0.912; FNR ≈ 0.088), with differences on the order of 10^{-4}. This suggests that, conditional on structural covariates, the model does not particularly penalize approved Black borrowers in terms of being correctly classified.
- The *Asian* group has a slightly lower TPR (0.883) and higher FNR (0.117), a loss of sensitivity of about three percentage points relative to White borrowers.
- The *Race Not Available* group is clearly disadvantaged: TPR ≈ 0.784, FNR ≈ 0.216, indicating a substantial degradation of Equal Opportunity (EO diff ≈ -0.128).

Regarding Predictive Parity (PPV), approved Black and Asian applications are at least as "good" as those of White borrowers (PPV ≈ 0.912 vs. 0.900). In other words, when the Random Forest model selects a Black or Asian applicant for approval, the probability that this decision is correct (relative to observed labels) is slightly higher than for the reference group.

4.7.4 Disparate Impact and Calibration

Disparate Impact highlights differences in positive prediction intensity:

- *Black or African American* has a DI of about 0.815, i.e., a roughly 18% lower

approval intensity than the White group, placing it close to — and slightly below — the conventional "80% rule" threshold.

- *Race Not Available* has a DI of 0.863, also unfavorable, whereas *Asian* (0.927) and *Joint* (1.043) lie closer to parity, with *Joint* slightly favored.

Probabilistic calibration, measured by the Brier score, yields a nuanced picture:

- *White*, *Black*, and *Joint* show relatively low scores (between 0.068 and 0.090), indicating satisfactory calibration of predicted probabilities;
- the *Asian* group has a slightly higher Brier score (0.099), suggesting somewhat less precise calibration;
- the *Race Not Available* group has a very poor Brier score (0.211), consistent with its low PPV and weak Equal Opportunity.

4.7.5 Visualization: Disparate Impact and Calibration

Figure 4.4 depicts DI ratios by group, while Figure 4.5 summarizes Brier scores by race.

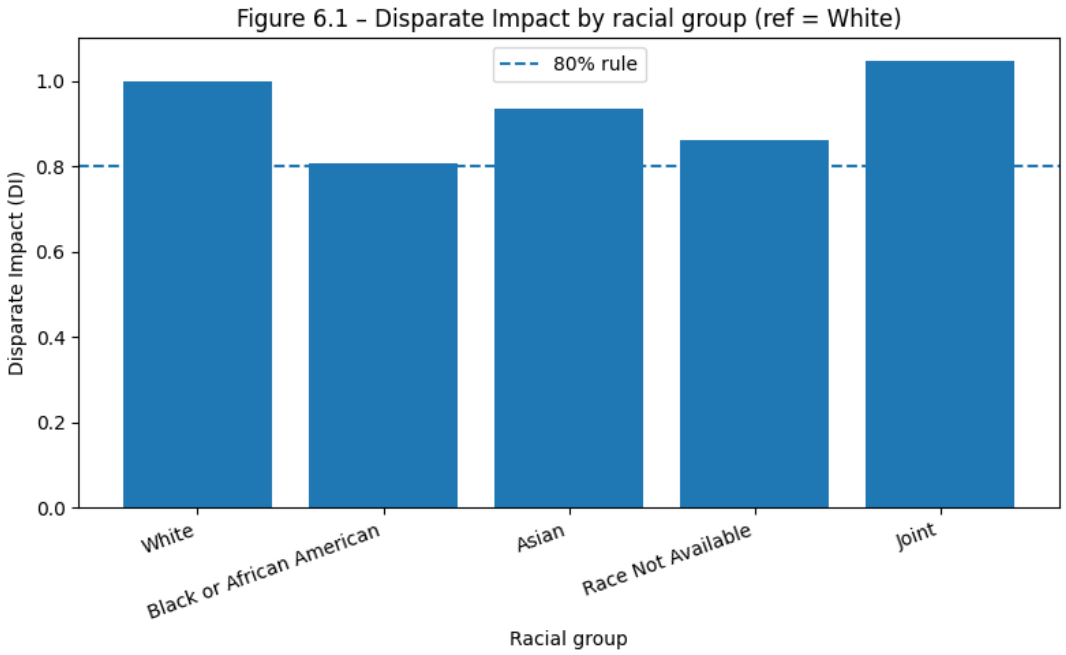

Figure 4.4: **Figure 6.1 – Disparate Impact by racial group (ref = White).** Disparate Impact (DI) of the AI-era Random Forest model by racial group, computed as the ratio of each group's predicted approval rate to that of White borrowers. The horizontal dashed line represents the "80% rule." Groups with DI below 0.80 are typically considered potentially disadvantaged in terms of credit access.

Disparate Impact (DI) estimated for the AI Random Forest model, computed as the ratio of a group's approval rate to that of *White* borrowers. The dashed horizontal line shows the "80% rule" threshold used in regulatory practice to flag substantial disadvantage. The *Black or African American* group lies close to this threshold, while the *Asian* and *Race Not Available* groups have slightly lower DI. The *Joint* group, in contrast, has a DI above 1, reflecting a higher approval rate than the reference group.

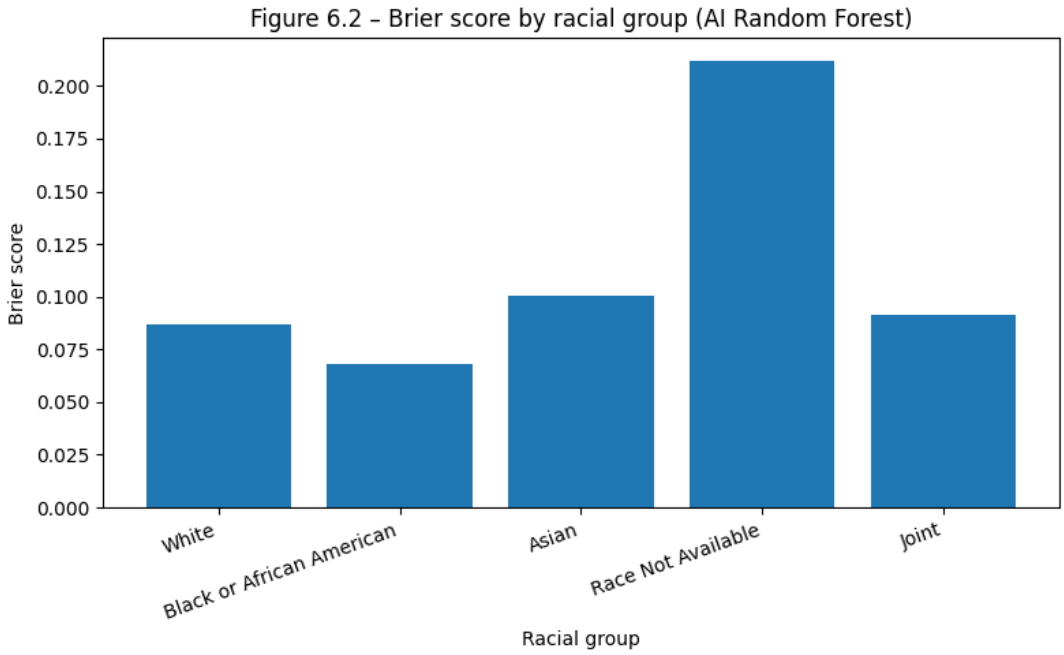

Figure 4.5: **Figure 6.2 – Brier score by racial group (AI Random Forest).** Brier scores of the AI-era Random Forest model by racial group. Lower values (for example, for *Black or African American* and *White* borrowers) indicate better probabilistic calibration of predictions, whereas the much higher score for the *Race Not Available* category reveals degraded calibration for applications with missing or incomplete racial information.

Lower scores (notably for *Black or African American* and *White* borrowers) indicate good probabilistic calibration, i.e., alignment between predicted probabilities and observed approval frequencies. By contrast, the very high Brier score for the *Race Not Available* group reveals severely degraded calibration, linked to the absence of usable racial information for this segment. This figure complements the Disparate Impact results by emphasizing that calibration and fairness are distinct dimensions of algorithmic evaluation.

4.7.6 RQ4 Synthesis and Implications for Algorithmic Fairness

RQ4 leads to a nuanced diagnosis:

- on the one hand, the AI-era Random Forest model appears to satisfy a form of Equal Opportunity and Predictive Parity between White and Black borrowers, which is notable in a context where the literature often reports substantial disparities;
- on the other hand, the *Race Not Available* category concentrates many fairness concerns: lower scoring intensity, much lower sensitivity, high false-negative rates, reduced precision, and poor calibration.

This last point is particularly important from a data-governance perspective: the absence or poor quality of demographic information does not neutralize bias, but may instead create a "gray zone" in which scoring models behave erratically. In the context of this dissertation, RQ4 therefore suggests that regulation and oversight of AI models should focus not only on explicitly sensitive variables but also on incomplete or poorly reported data segments.

4.8 Chapter Summary

This chapter has presented the empirical findings associated with the four research questions RQ1–RQ4.

- From a descriptive standpoint, the transition to the AI era is associated with a sharp decline in approval rates (from 70.7% to 54.2%), a slight decrease in the

share of high-risk applications (proxy default), and persistent heterogeneity by race and income.

- Comparative logit models show a dramatic increase in pseudo-R^2 from pre-AI to AI, stronger product effects (*loan_purpose*, *loan_type*), and a central role for HOEPA status in recent years.

- t-tests and ANOVA confirm that inter-period, inter-racial, and inter-territorial differences are not mere sampling artifacts but rest on very robust statistical evidence.

- RQ1 documents a pronounced socio-economic gradient: low-income areas face significantly lower odds of approval, even in a minimalist model, raising concerns about territorial equity.

- RQ2 demonstrates that the predictability of decisions rises sharply in the AI era, especially for non-linear models (Random Forest, XGBoost) with AUC values above 0.90, a sign of advanced algorithmic standardization.

- RQ3 suggests that the AI era is associated with a modest improvement in the average risk profile, but that variability in the proxy risk remains primarily shaped by product composition.

- Finally, RQ4 provides a detailed fairness diagnosis: while major observable groups (White, Black, Asian, Joint) exhibit relatively similar levels of Equal Opportunity and Predictive Parity, applications with missing racial information concentrate severe anomalies, underscoring the importance of data quality in evaluating algorithmic justice.

These findings prepare the ground for Chapter 5, which will offer a critical discus-

sion of their normative implications: how can predictive performance, risk management, and fairness goals be reconciled? What role should regulation play in an environment where approval decisions become both more predictable and potentially more unequal? And which concrete avenues—in terms of model design, fairness audits, and transparency—can be pursued to promote truly *AI-driven fair creditworthiness* aligned with the objectives of social justice and consumer protection in the Tri-State Area?