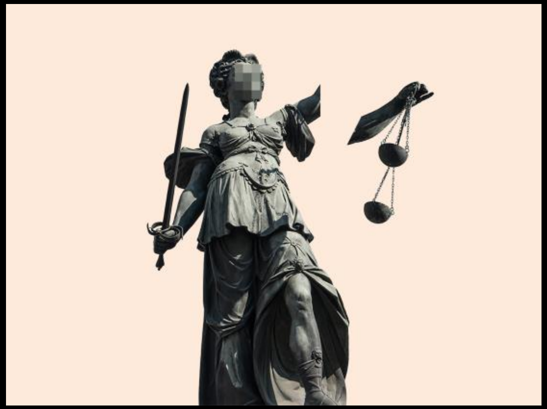
Deloitte.

Artificial Intelligence Guild

January 2023



Achieving Process Fairness through Automated Feature Selection

A Comparative Study by Aritra Nath

Agenda

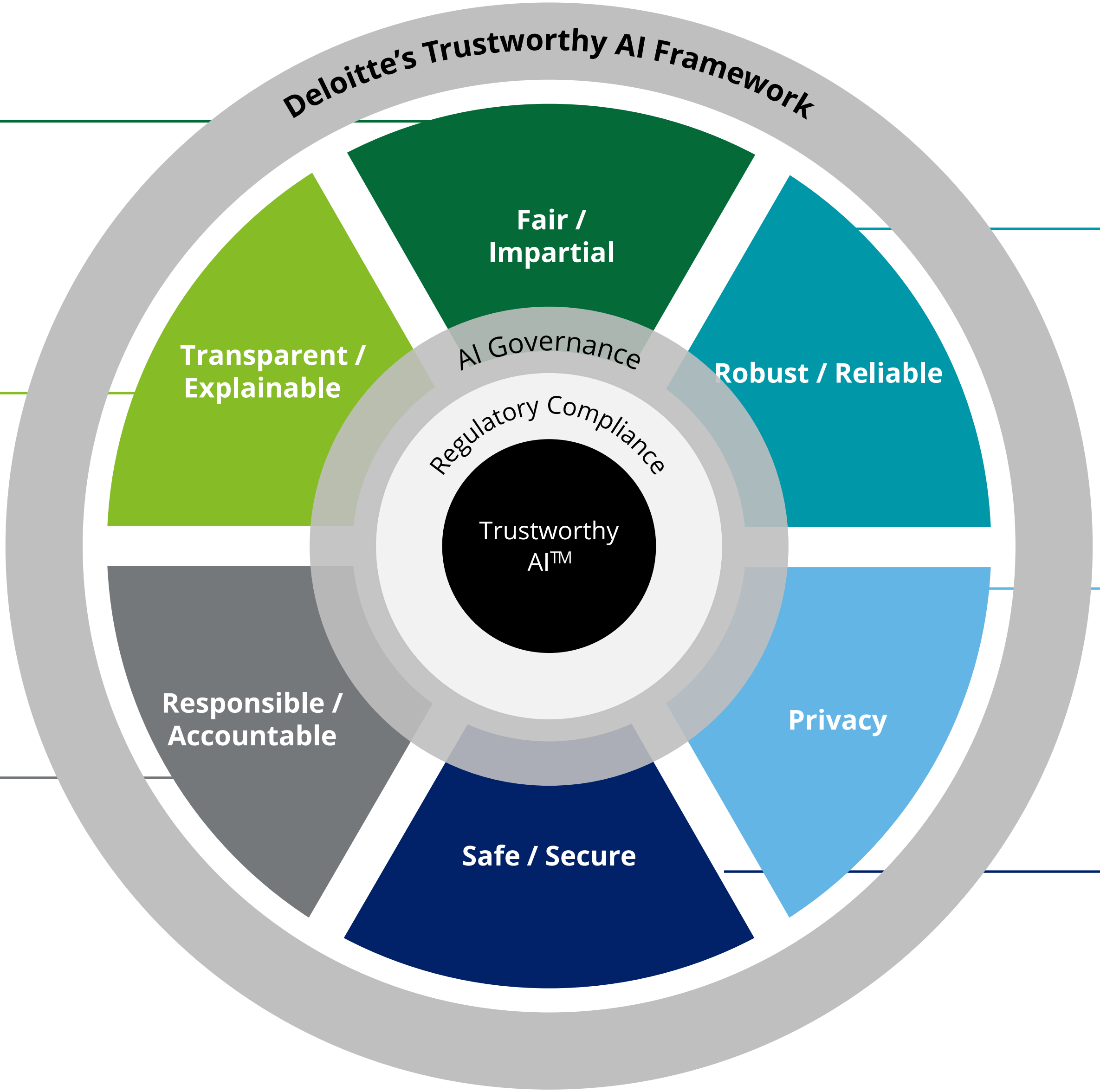
Motivation and Approach

Deloitte's **Trustworthy AI™** framework is an effective first step in having an approach to manage AI risks, which can be integrated into broader enterprise risk management.

Fair / Impartial
AI applications include automated and manual checks to help enable equitable application across all participants

Transparent / Explainable
All participants are able to understand how their data is being used and how AI systems make decisions; algorithms, attributes, and correlations are open to inspection

Responsible / Accountable
Policies are in place to determine who is held responsible for the output of AI system decisions

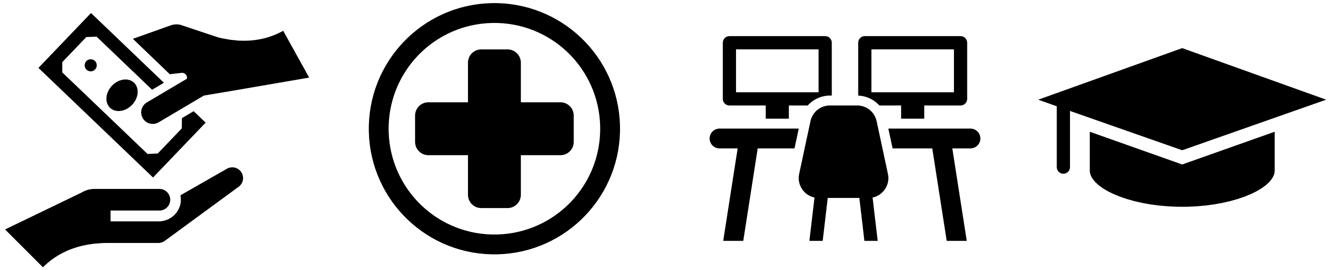


Robust / Reliable
AI systems have the ability to learn from humans and other systems and produce consistent and reliable outputs

Privacy
Consumer privacy is respected and customer data is not used beyond its intended and stated use; consumers are able to opt in/ out of sharing their data

Safe / Secure
AI systems can be protected from risks (including cyber risks) that may cause physical and/or digital harm

AI Should Work for All

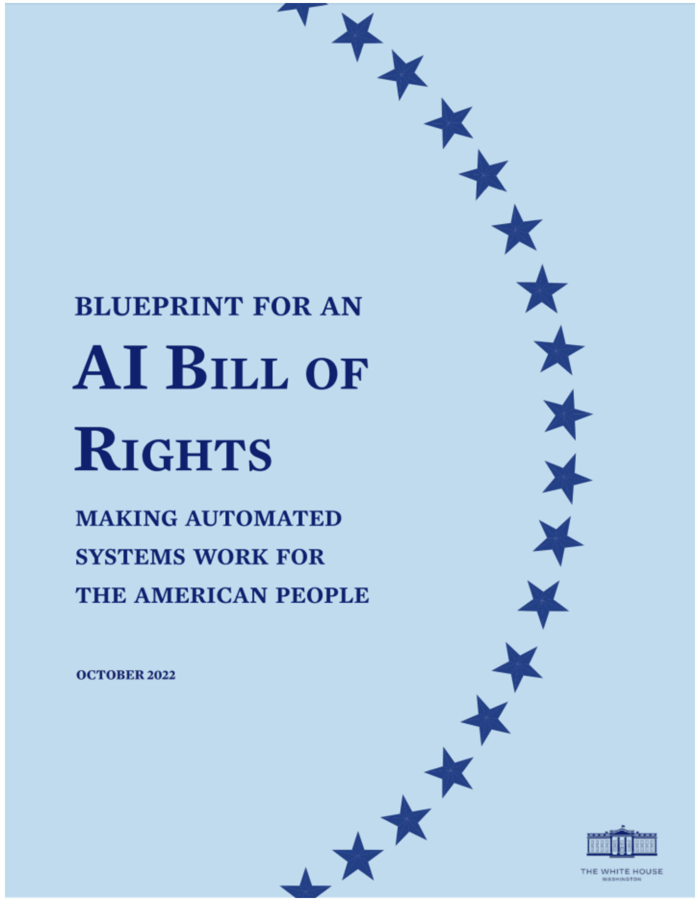


Procedural fairness

It requires the same set of transparent and non-discriminatory policies to be applied to everyone; basically, due process.

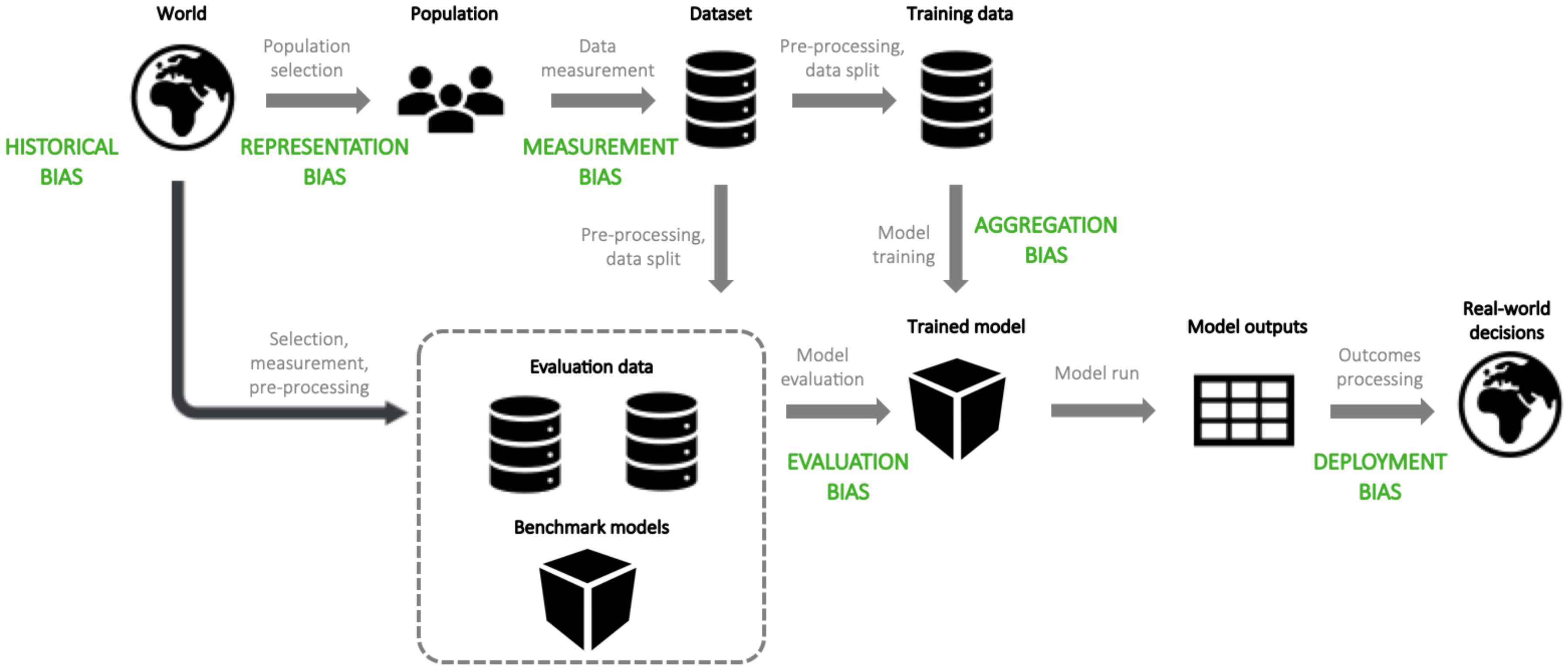
Source: https://pnigel.com/papers/belitz-AIES_2021_ACM.pdf

AI Regulations



Source: https://bluewatercredit.com/consumer-rights/
Source: https://www.assoc-law.com/blog/2021/05/05/a-guide-to-the-fair-housing-act-and-its-exemptions/

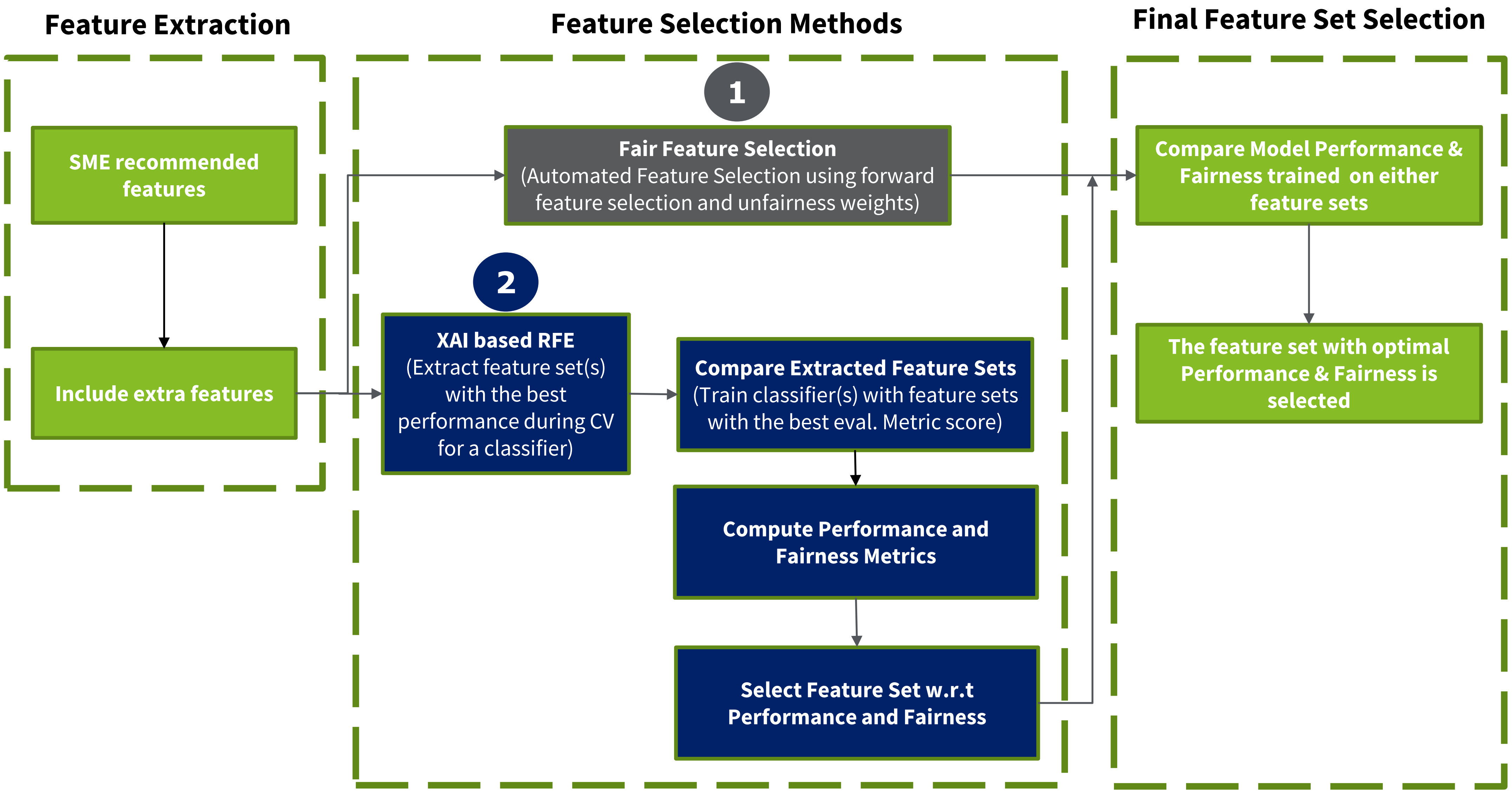
The Sources of Unfairness



What are "potentially" unfair features?

- Are those where one group is likely to benefit more than another from their inclusion. Ex. Gender, Race, Ethnicity, Age, Pregnancy Status
- Features that are correlated with sensitive features though apparently naïve could create unfairness as well. Ex. ZIP code as a proxy for race .

Approach Outline



Use Case

Exploring the HMDA Dataset

HMDA 2019	
Size	2 MN loans
Dataset Size Used	200K loans
No. of features	99
No. of features used	22 (predictors) + 1 (response)
Target	Action Taken (Application Outcome)

Data Dictionary:
https://ffiec.cfpb.gov/documentation/2019/lar-data-fields/

All Features

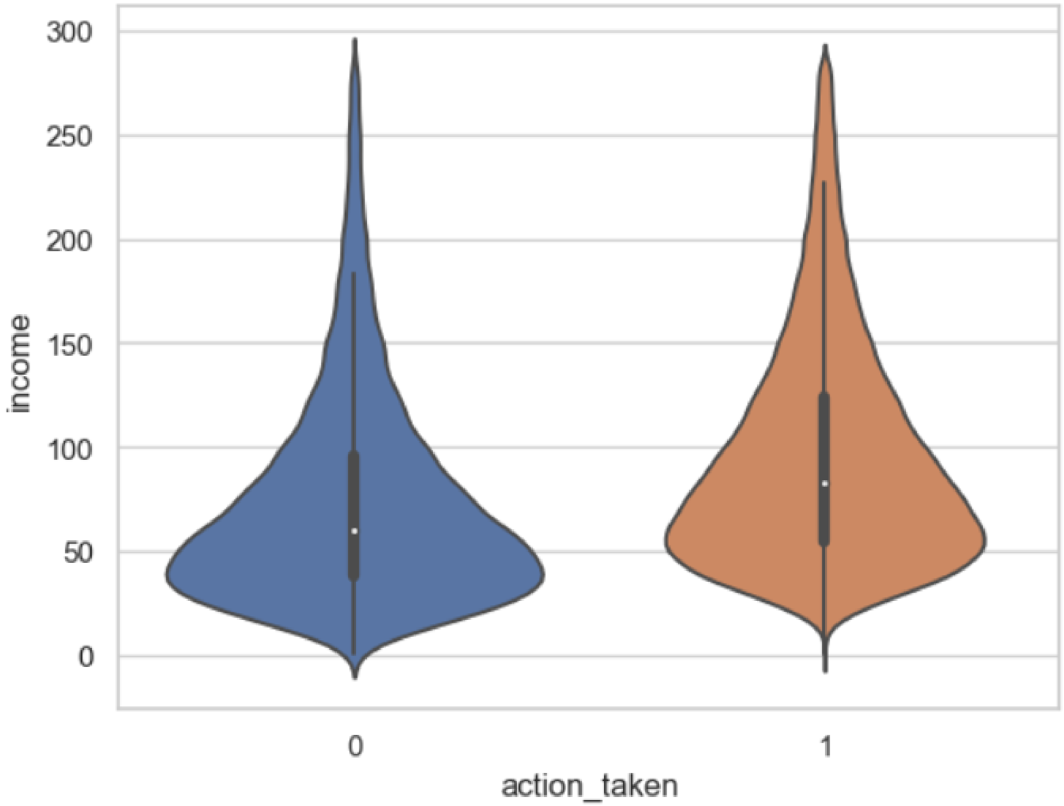
- *activity_year*
- *lei*
- *derived_msa-md*
- *state_code*
- *county_code*
- *census_tract*
- *derived_loan_product_type*
- *derived_dwelling_category*
- *conforming_loan_limit*
- *derived_ethnicity*
- *derived_race*
- *derived_sex*
- *action_taken*
- *purchaser_type*
- *preapproval*
- *loan_type*
- *loan_purpose*
- *lien_status*
- *reverse_mortgage*
- *open-end_line_of_credit*
- *business_or_commercial_purpose*
- *loan_amount*
- *combined_loan_to_value_ratio*
- *interest_rate*
- *rate_spread*
- *hoepa_status*
- *total_loan_costs*
- *total_points_and_fees*
- *origination_charges*
- *discount_points*
- *lender_credits*
- *loan_term*
- *prepayment_penalty_term*
- *intro_rate_period*
- *negative_amortization*
- *interest_only_payment*

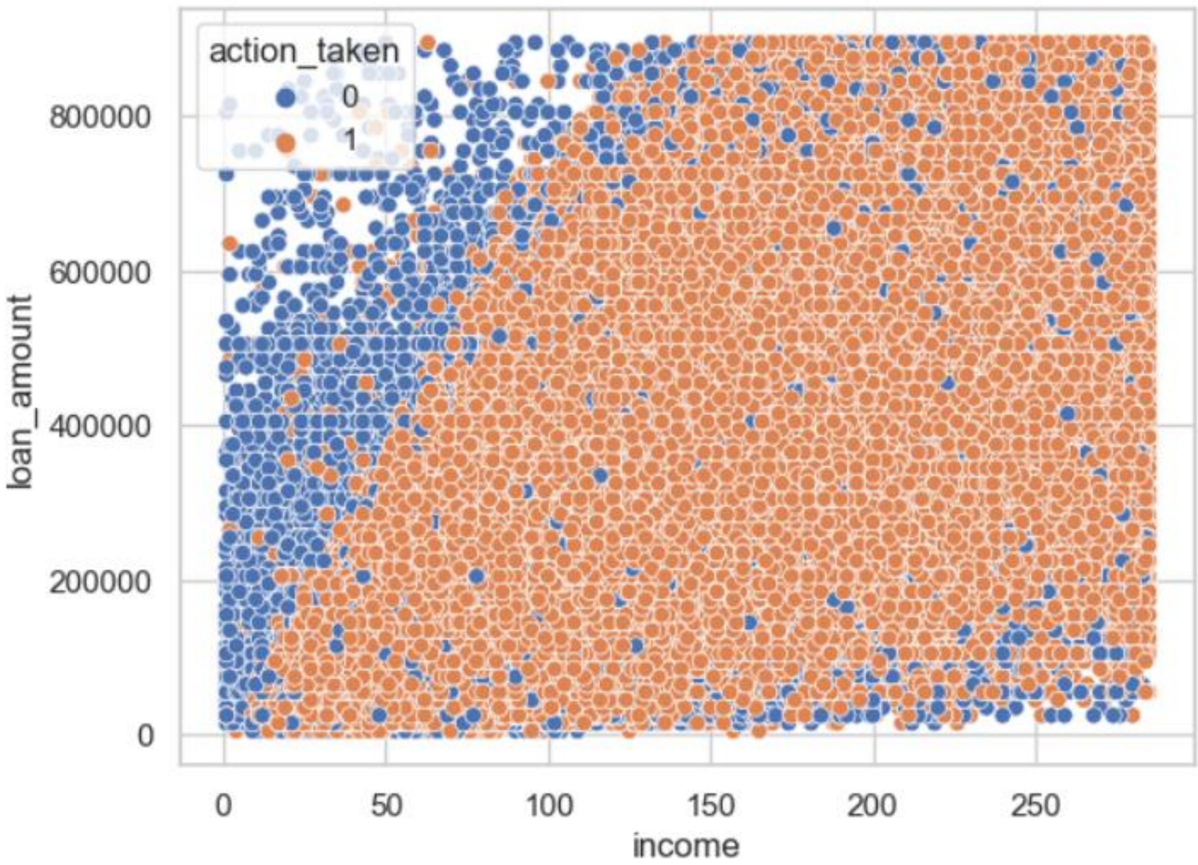
- *balloon_payment*
- *other_nonamortizing_features*
- *property_value*
- *construction_method*
- *occupancy_type*
- *manufactured_home_secured_property_type*
- *manufactured_home_land_property_interest*
- *total_units*
- *ageapplicant*
- *multifamily_affordable_units*
- *income*
- *debt_to_income_ratio*
- *applicant_credit_score_type*
- *co-applicant_credit_score_type*
- *applicant_ethnicity-1*
- *applicant_ethnicity-2*
- *applicant_ethnicity-3*
- *applicant_ethnicity-4*
- *applicant_ethnicity-5*
- *co-applicant_ethnicity-1*
- *co-applicant_ethnicity-2*
- *co-applicant_ethnicity-3*
- *co-applicant_ethnicity-4*
- *co-applicant_ethnicity-5*
- *applicant_ethnicity_observed*
- *co-applicant_ethnicity_observed*
- *applicant_race-1*
- *applicant_race-2*
- *applicant_race-3*
- *applicant_race-4*
- *applicant_race-5*
- *co-applicant_race-1*
- *co-applicant_race-2*
- *co-applicant_race-3*

- *co-applicant_race-4*
- *co-applicant_race-5*
- *applicant_race_observed*
- *co-applicant_race_observed*
- *applicant_sex*
- *co-applicant_sex*
- *applicant_sex_observed*
- *co-applicant_sex_observed*
- *co-applicant_age*
- *applicant_age_above_62*
- *co-applicant_age_above_62*
- *submission_of_application*
- *initially_payable_to_institution*
- *aus-1*
- *aus-2*
- *aus-3*
- *aus-4*
- *aus-5*
- *denial_reason-1*
- *denial_reason-2*
- *denial_reason-3*
- *denial_reason-4*
- *tract_population*
- *tract_minority_population_percent*
- *ffiec_msa_md_median_family_income*
- *tract_to_msa_income_percentage*
- *tract_owner_occupied_units*
- *tract_one_to_four_family_homes*
- *tract_median_age_of_housing_units*

Exploratory Data Analysis

Applicant Income

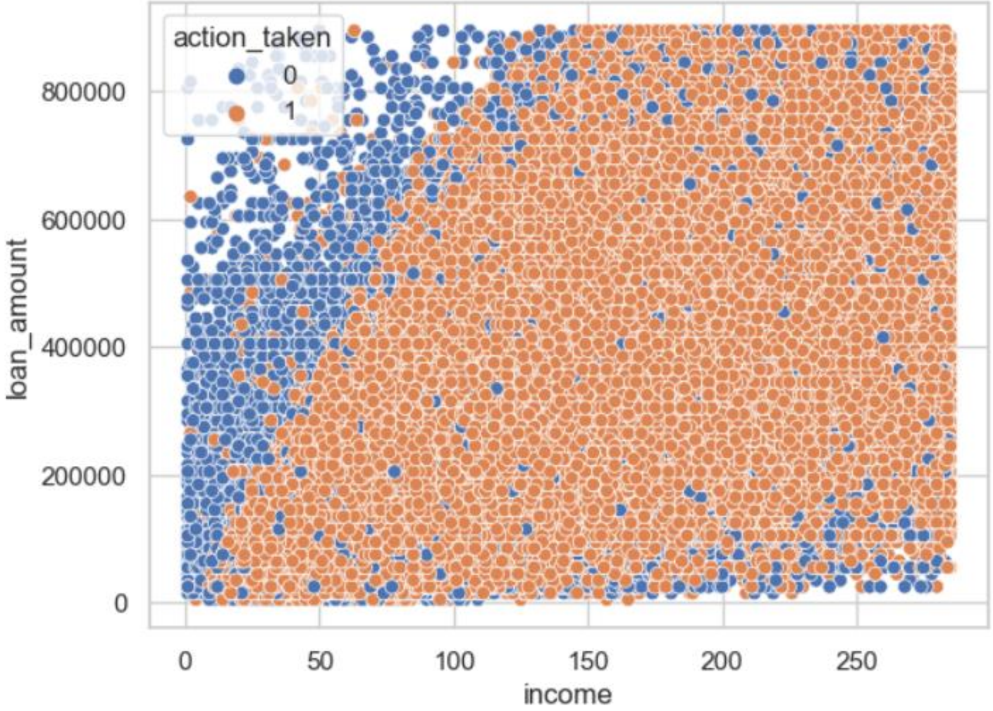




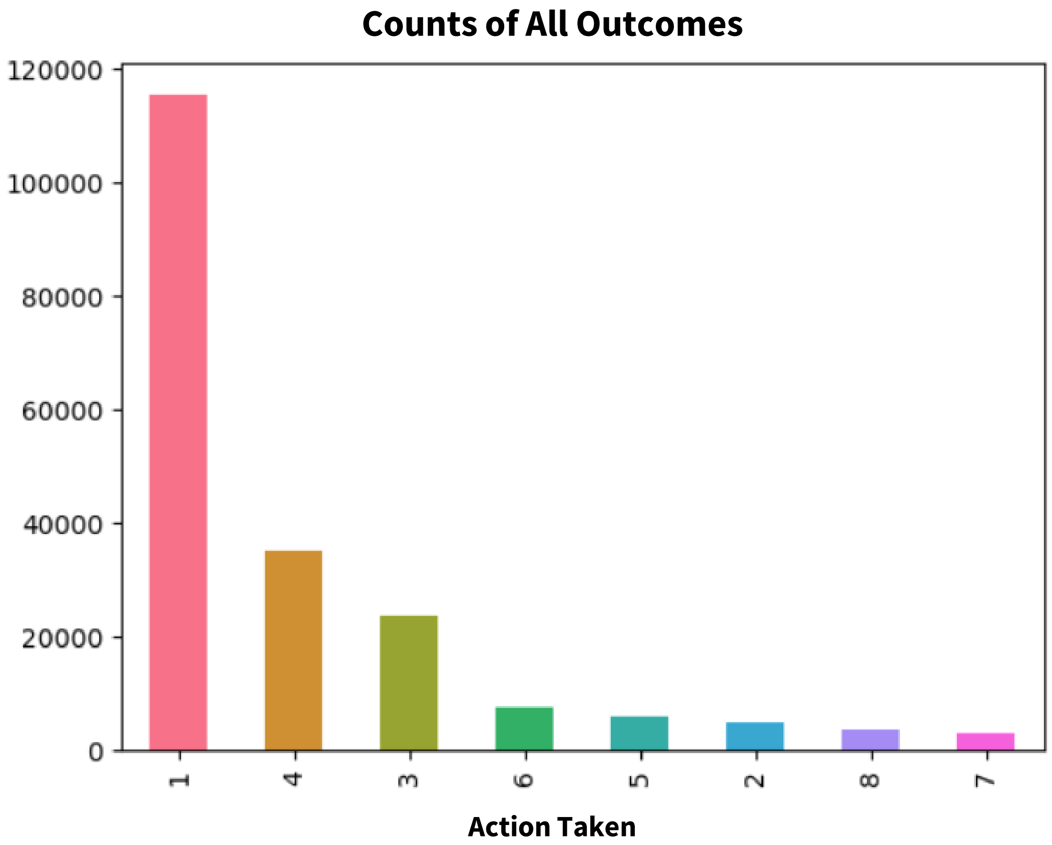
Exploring the HMDA Dataset

Correlation Heat Map

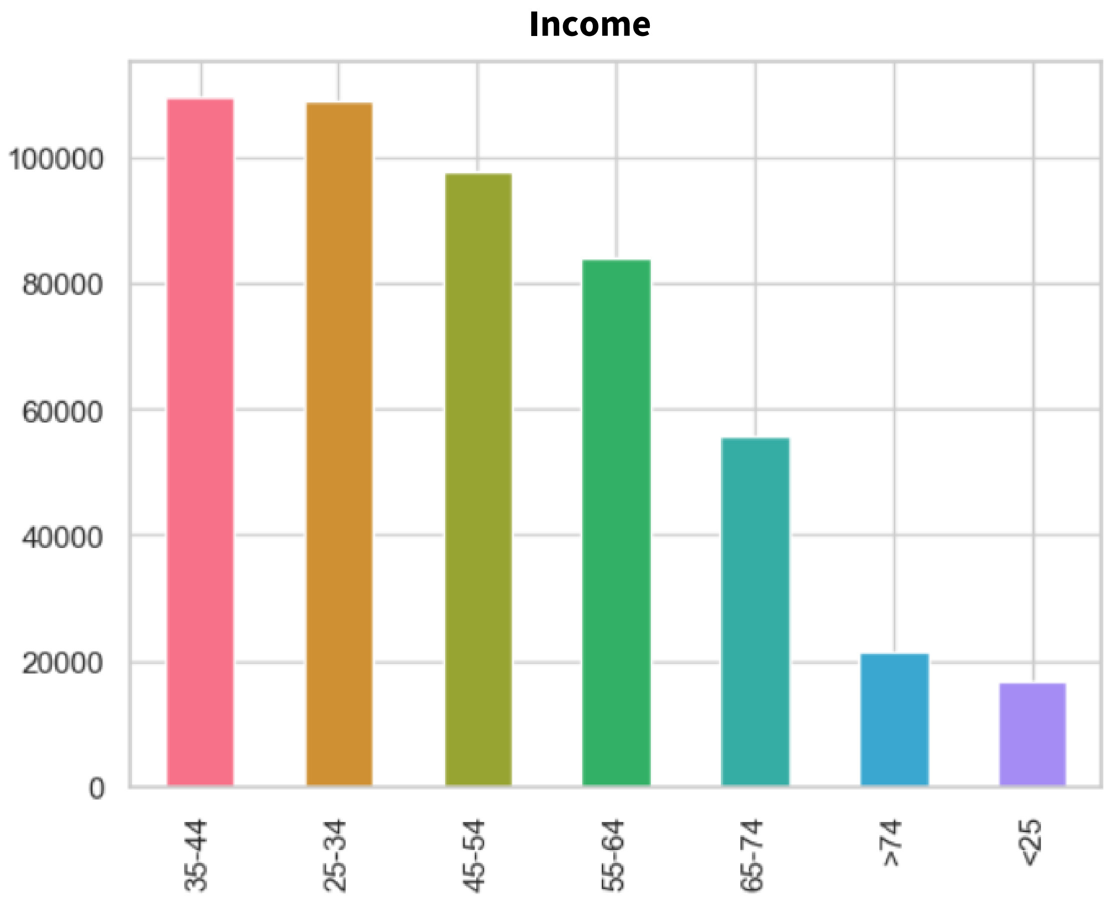
	income	loan_amount	derived_ethnicity	derived_race	derived_sex	preapproval	loan_type	loan_purpose	interest_only_payment	balloon_payment	action_taken	county_code	tract_minority_population_percent	tract_population	tract_to_msa_income_percentage	tract_median_age_of_housing_units	ffiec_msa_md_median_family_income
income	1.00	0.57	0.03	-0.04	-0.33	-0.04	-0.16	0.14	-0.03	0.00	0.15	0.01	-0.07	0.06	0.33	-0.12	0.22
loan_amount	0.57	1.00	-0.04	-0.07	-0.17	0.07	-0.01	0.10	-0.02	0.04	0.13	-0.10	0.10	0.09	0.25	-0.16	0.37
derived_ethnicity	0.03	-0.04	1.00	0.00	0.06	-0.05	-0.04	0.03	-0.02	-0.02	0.03	0.14	-0.30	-0.05	0.09	-0.00	0.02
derived_race	-0.04	-0.07	0.00	1.00	0.03	-0.04	-0.05	0.01	-0.01	-0.02	0.03	0.05	-0.19	-0.03	0.04	0.01	-0.06
derived_sex	-0.33	-0.17	0.06	0.03	1.00	0.03	0.00	-0.05	0.03	0.01	-0.06	-0.04	0.10	-0.04	-0.11	0.08	-0.01
preapproval	-0.04	0.07	-0.05	-0.04	0.03	1.00	0.10	0.24	0.03	-0.02	-0.10	-0.10	0.13	0.03	-0.05	-0.06	0.04
loan_type	-0.16	-0.01	-0.04	-0.05	0.00	0.10	1.00	-0.18	0.04	0.03	0.01	-0.07	0.05	0.02	-0.14	-0.02	-0.09
loan_purpose	0.14	0.10	0.03	0.01	-0.05	0.24	-0.18	1.00	0.04	0.01	-0.06	-0.07	0.01	-0.00	0.11	-0.02	0.10
interest_only_payment	-0.03	-0.02	-0.02	-0.01	0.03	0.03	0.04	0.04	1.00	0.52	-0.01	-0.08	0.05	0.02	-0.01	-0.00	-0.00
balloon_payment	0.00	0.04	-0.02	-0.02	0.01	-0.02	0.03	0.01	0.52	1.00	-0.00	-0.06	0.04	0.02	0.00	-0.02	0.00
action_taken	0.15	0.13	0.03	0.03	-0.06	-0.10	0.01	-0.06	-0.01	-0.00	1.00	0.01	-0.04	0.02	0.05	-0.03	0.06
county_code	0.01	-0.10	0.14	0.05	-0.04	-0.10	-0.07	-0.07	-0.08	-0.06	0.01	1.00	-0.29	-0.04	-0.00	0.08	0.09
tract_minority_population_percent	-0.07	0.10	-0.30	-0.19	0.10	0.13	0.05	0.01	0.05	0.04	-0.04	-0.29	1.00	0.12	-0.34	0.05	0.08
tract_population	0.06	0.09	-0.05	-0.03	-0.04	0.03	0.02	-0.00	0.02	0.02	0.02	-0.04	0.12	1.00	0.14	-0.40	-0.01
tract_to_msa_income_percentage	0.33	0.25	0.09	0.04	-0.11	-0.05	-0.14	0.11	-0.01	0.00	0.05	-0.00	-0.34	0.14	1.00	-0.32	-0.07
tract_median_age_of_housing_units	-0.12	-0.16	-0.00	0.01	0.08	-0.06	-0.02	-0.02	-0.00	-0.02	-0.03	0.08	0.05	-0.40	-0.32	1.00	0.08
ffiec_msa_md_median_family_income	0.22	0.37	0.02	-0.06	-0.01	0.04	-0.09	0.10	-0.00	0.00	0.06	0.09	0.08	-0.01	-0.07	0.08	1.00



Feature Distribution



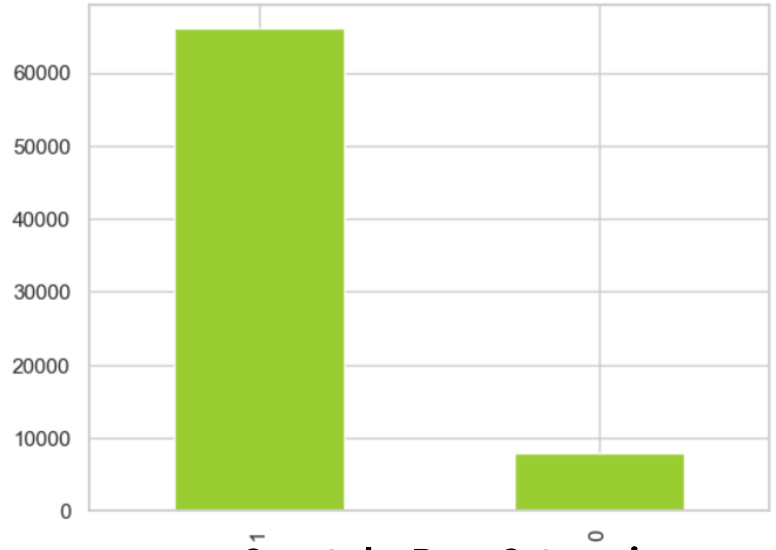
1 - Loan originated
2 - Application approved but not accepted
3 - Application denied
4 - Application withdrawn by applicant
5 - File closed for incompleteness
6 - Purchased loan
7 - Preapproval request denied
8 - Preapproval request approved but not accepted



*The income feature is skewed to the right, with more of the dataset having lower income levels.

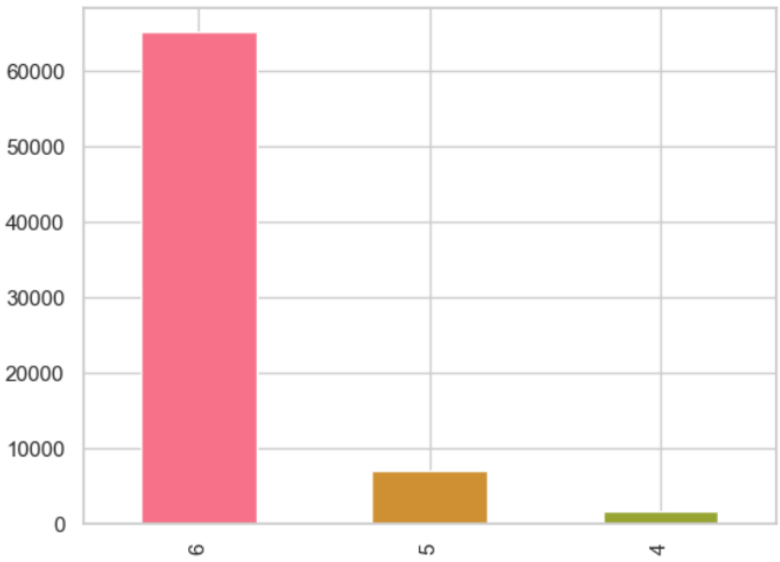
Feature Distribution

Counts by Binary Outcome



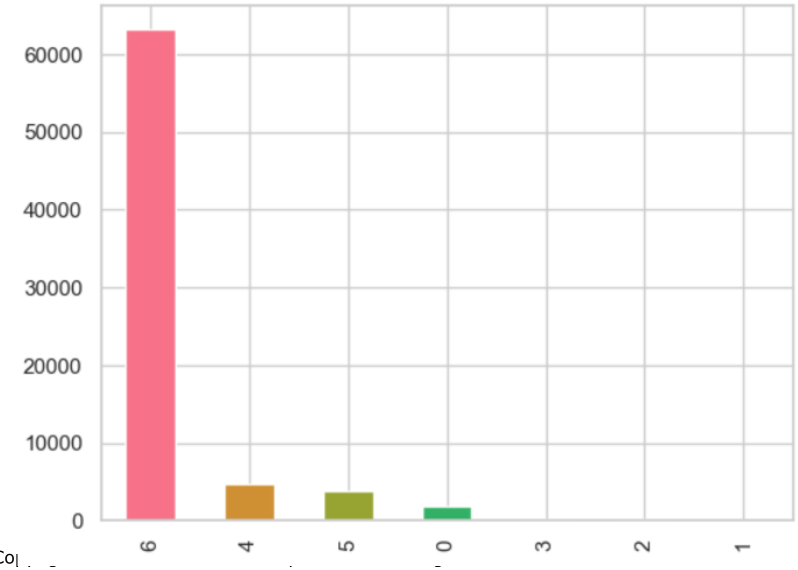
Action Taken:
1- Loan Granted
0- Loan Denied

Count by Ethnicity

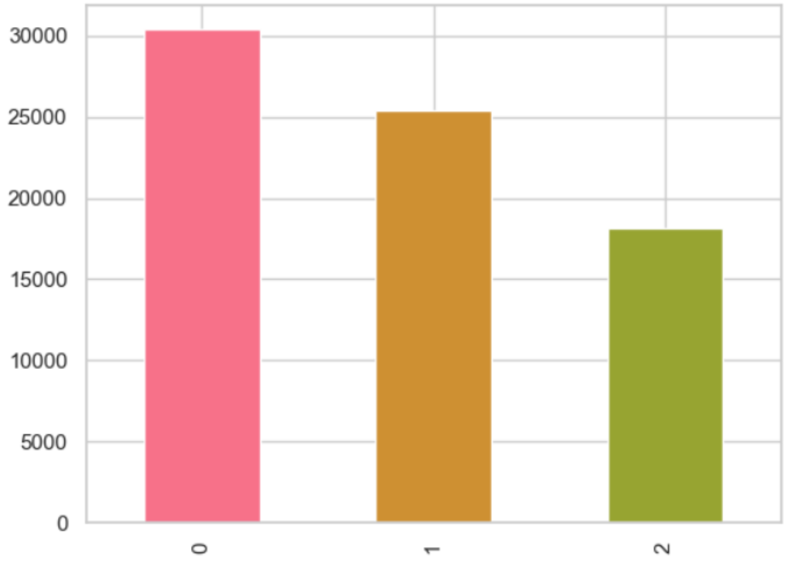


Ethnicity:
6- Not Hispanic or Latino
5- Hispanic or Latino
4- Joint

Counts by Race Categories



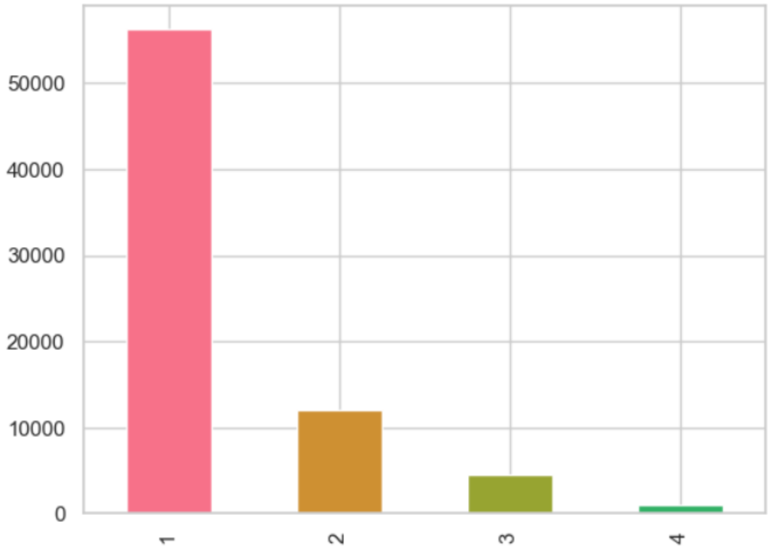
Race Categories:
6- White
5- Black or African American
4- American Indian or Alaska Native
3- Asian
2- Native Hawaiian or Other Pacific Islander
1- 2 or more minority races
0- Joint



Sex:
2- Female
1- Male
0- Joint

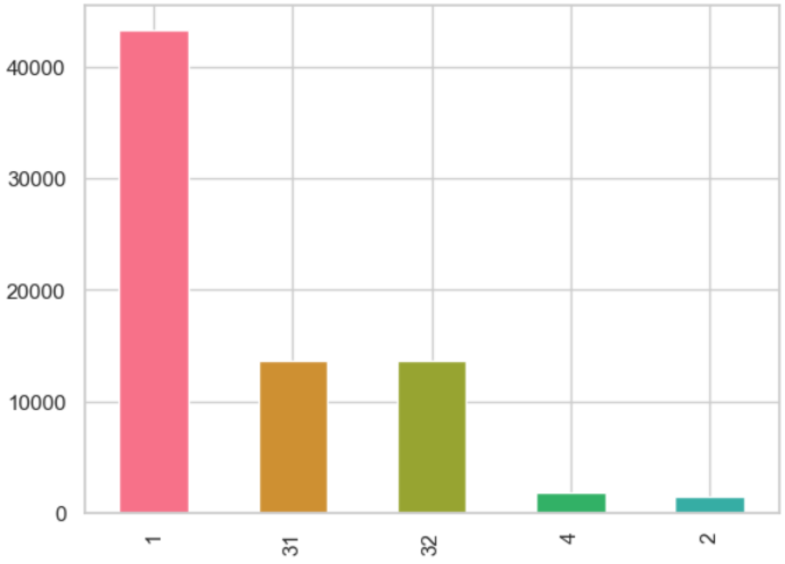
Feature Distribution

Counts by Loan Type



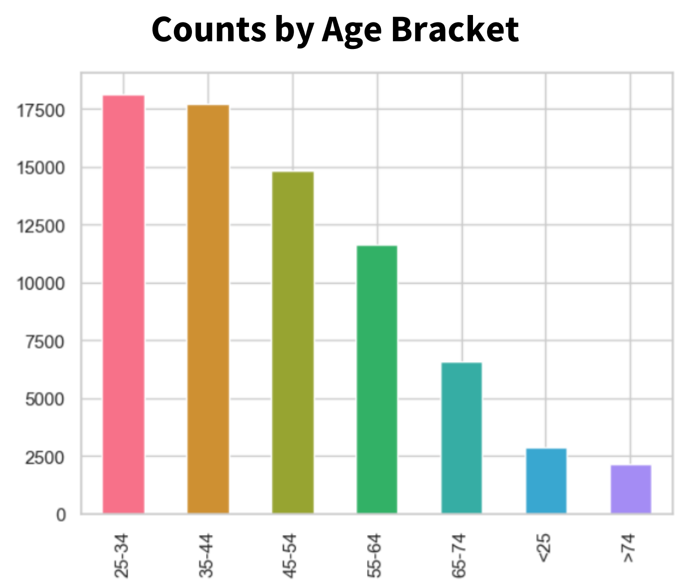
Loan Type:
1- Conventional (not insured or guaranteed by FHA, VA, RHS, or FSA)
2- Federal Housing Administration insured (FHA)
3- Veterans Affairs guaranteed (VA)
4- USDA Rural Housing Service or Farm Service Agency guaranteed (RHS or FSA)

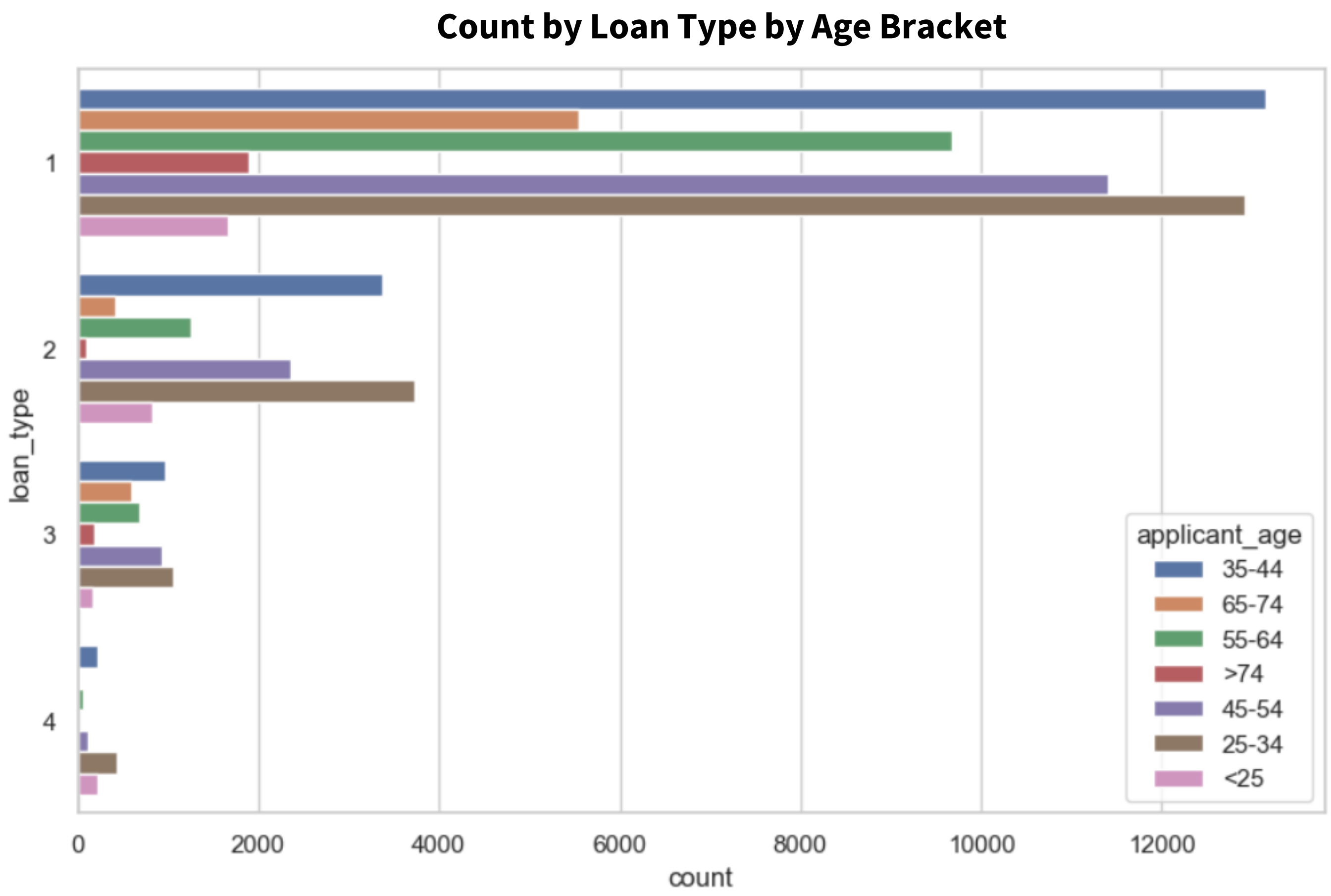
Count by Loan Purpose



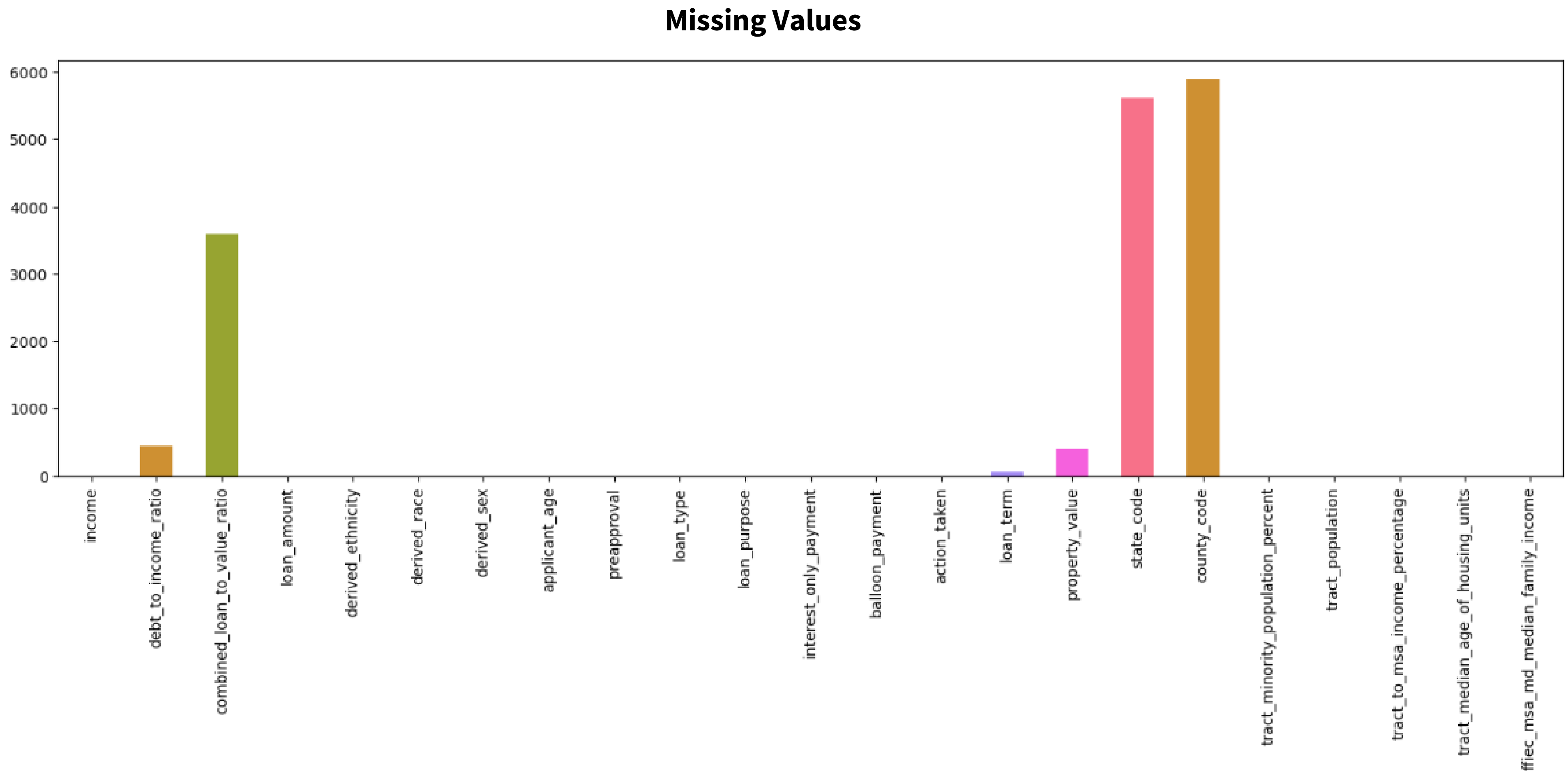
Loan Purpose:
1 - Home purchase
2 - Home improvement
31 - Refinancing
32 - Cash-out refinancing
4 - Other purpose
5 - Not applicable

Feature Distribution





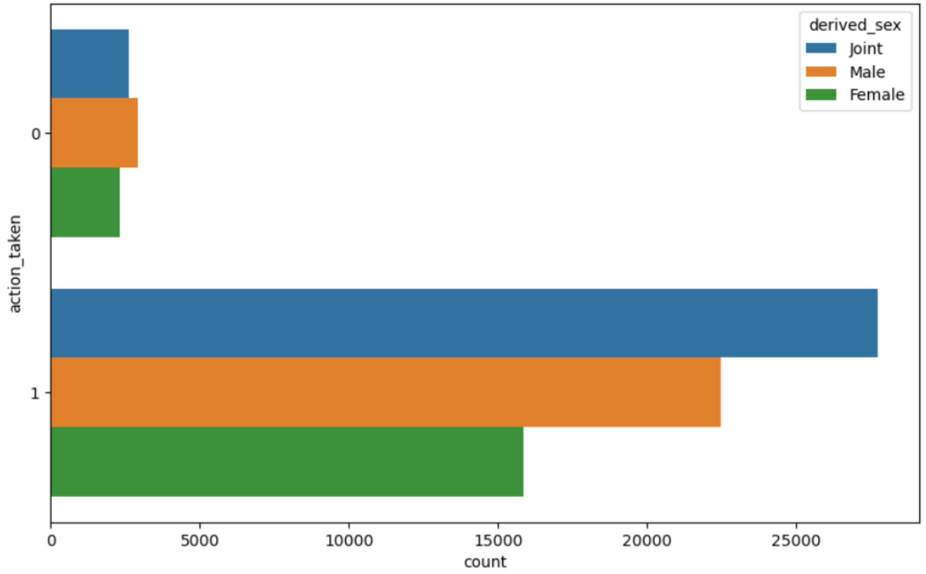
Missing Value Analysis



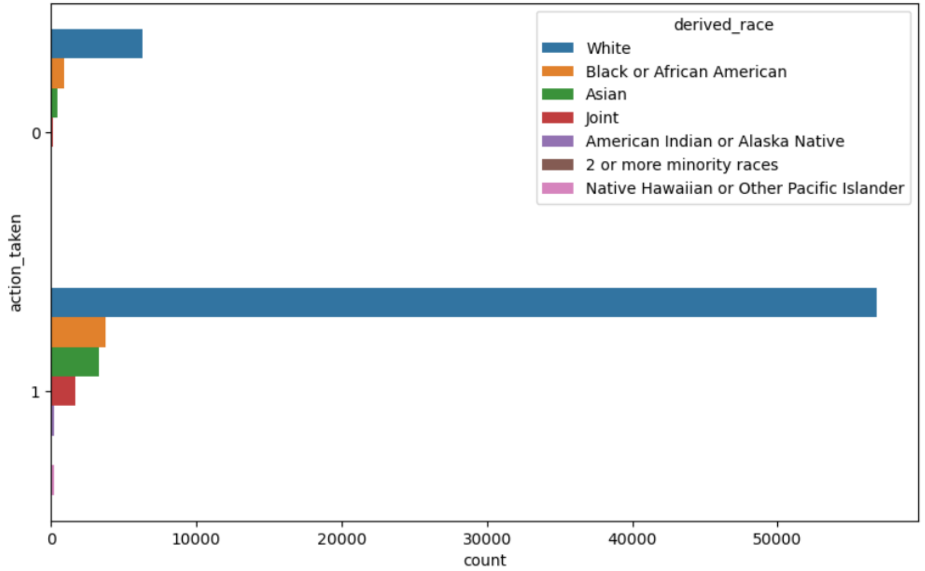
Number of Rows retained: 73947

Applicant Sensitive Feature vs Outcome

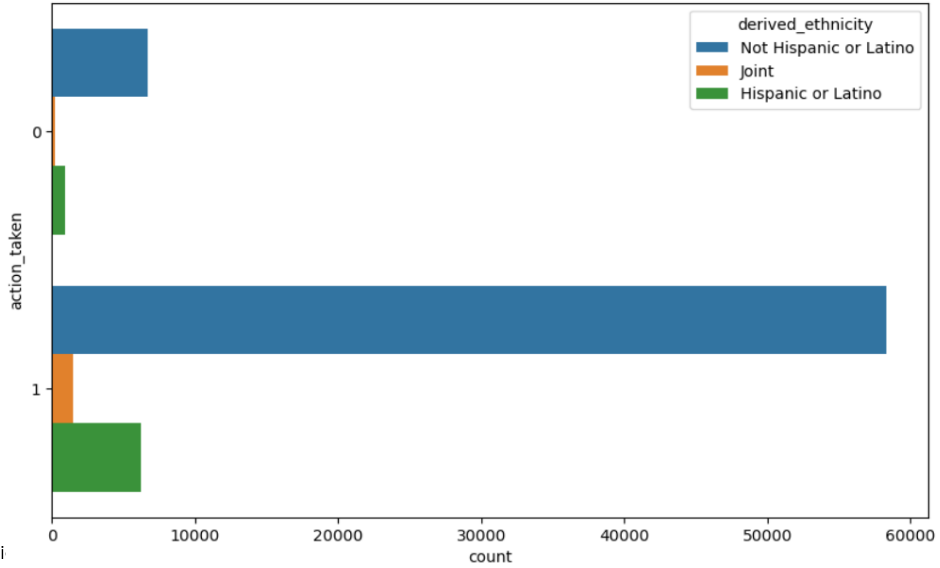
Sex



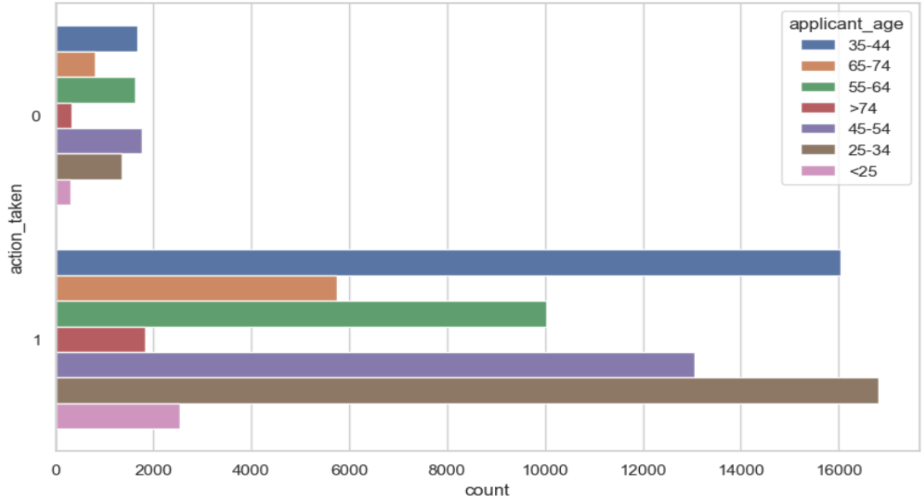
Race



Ethnicity

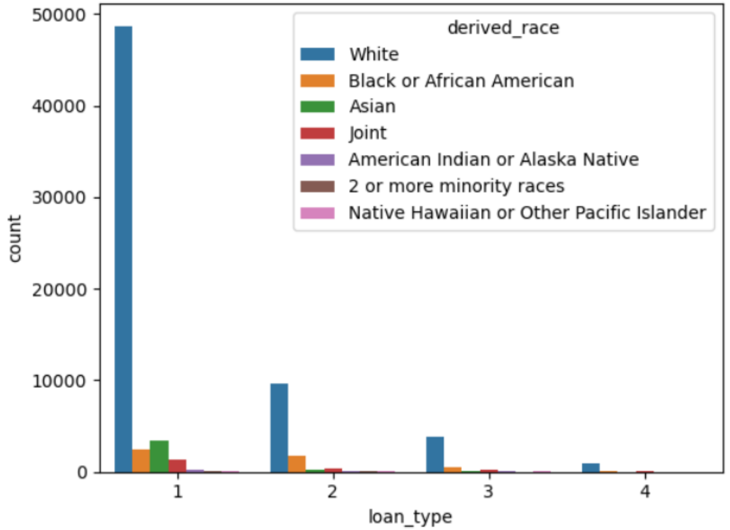


Age

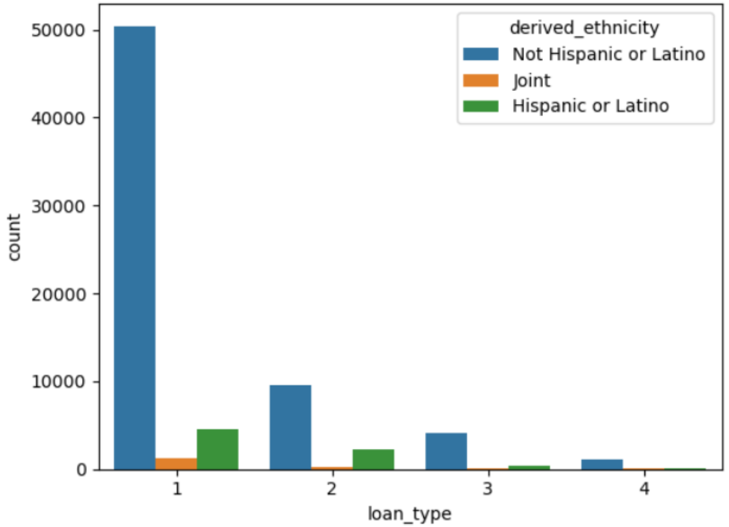


Applicant Sensitive Feature vs Loan Type

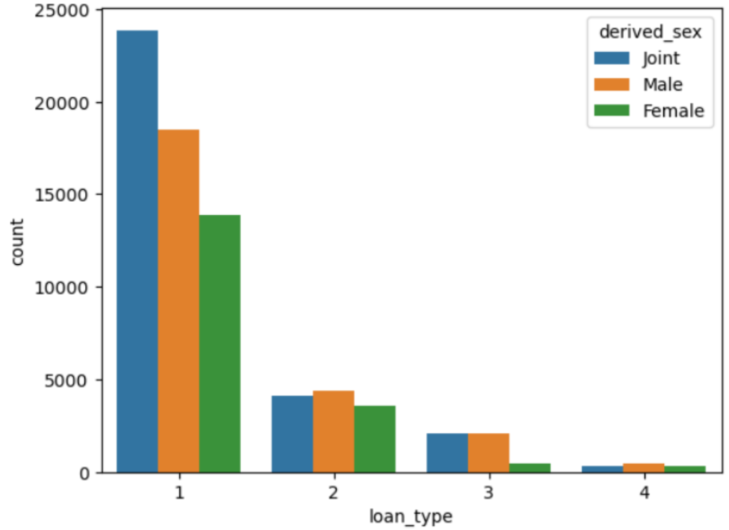
Race



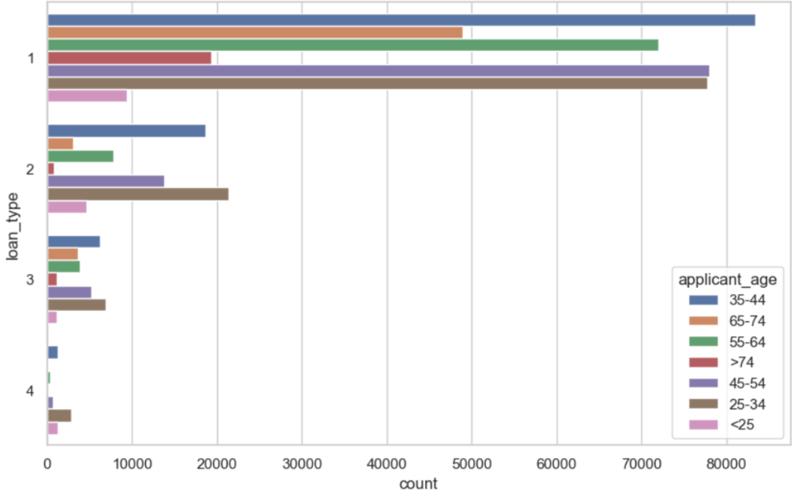
Ethnicity



Sex



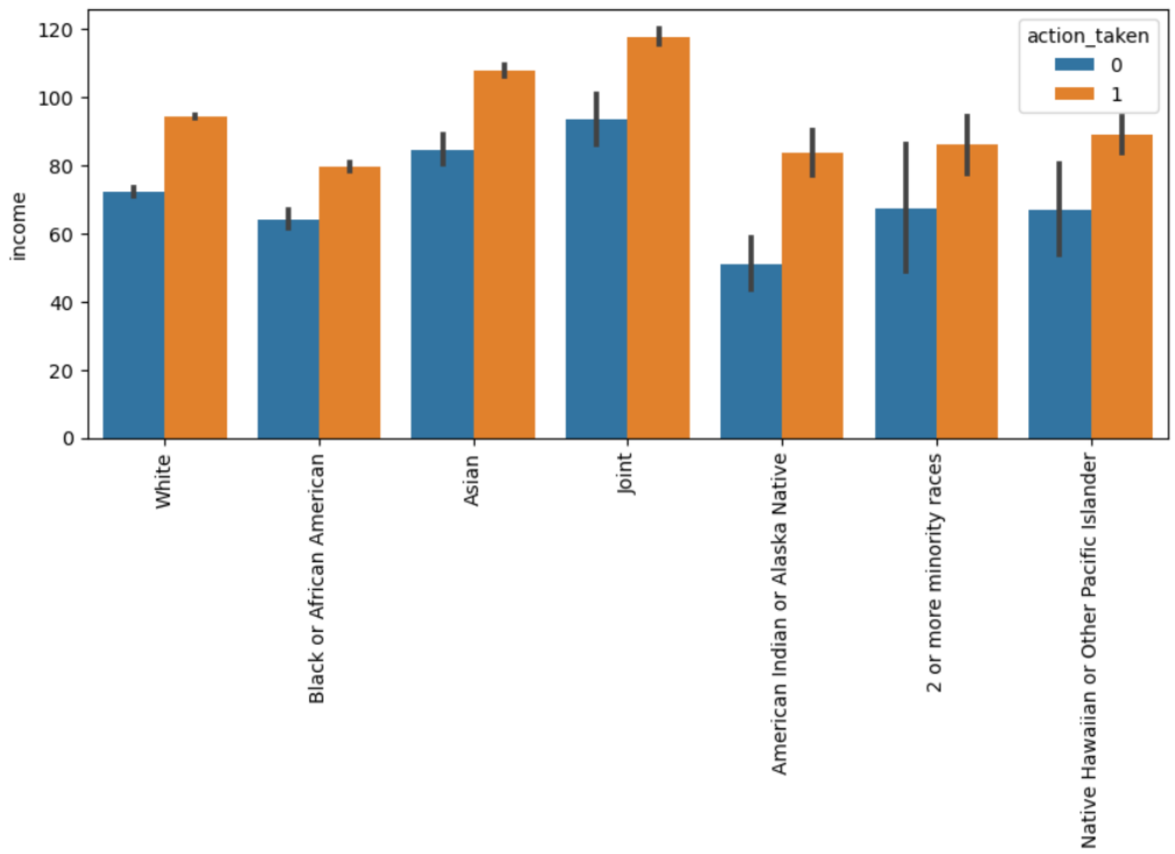
Age



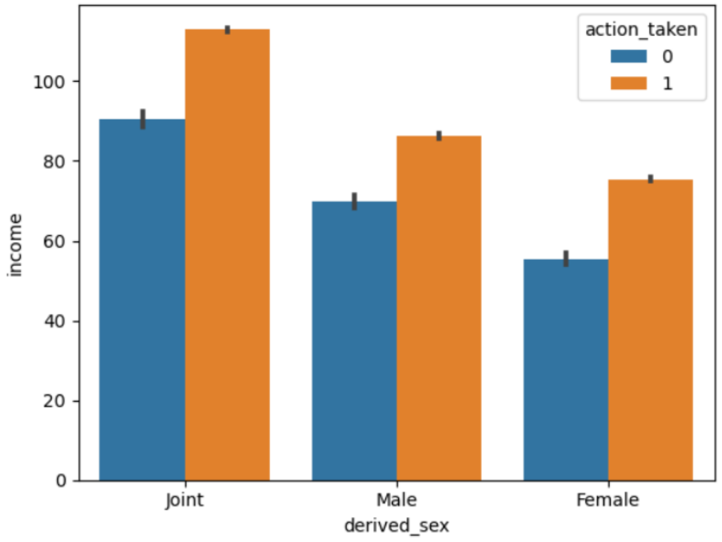
Loan Type:
1- Conventional (not insured or guaranteed by FHA, VA, RHS, or FSA)
2- Federal Housing Administration insured (FHA)
3-Veterans Affairs guaranteed (VA)
4- USDA Rural Housing Service or Farm Service Agency guaranteed (RHS or FSA)

Applicant Sensitive Feature vs Income & Outcome

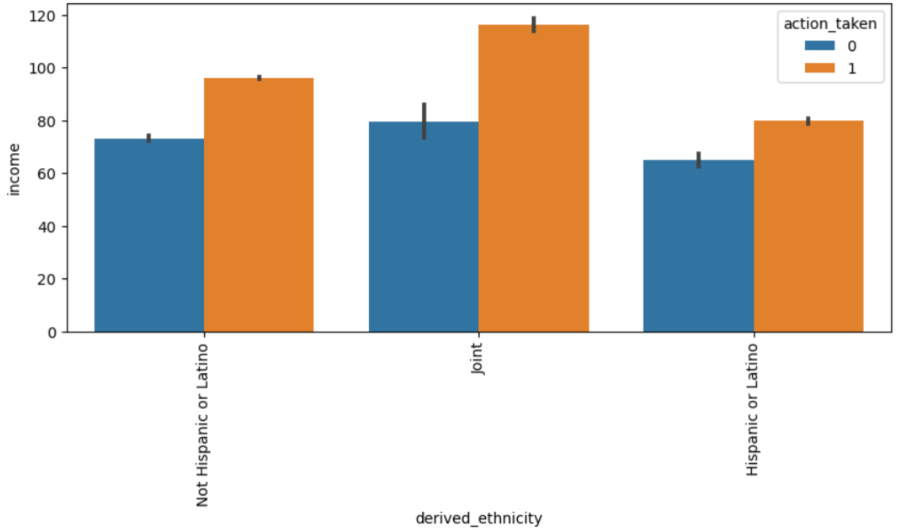
Race



Sex

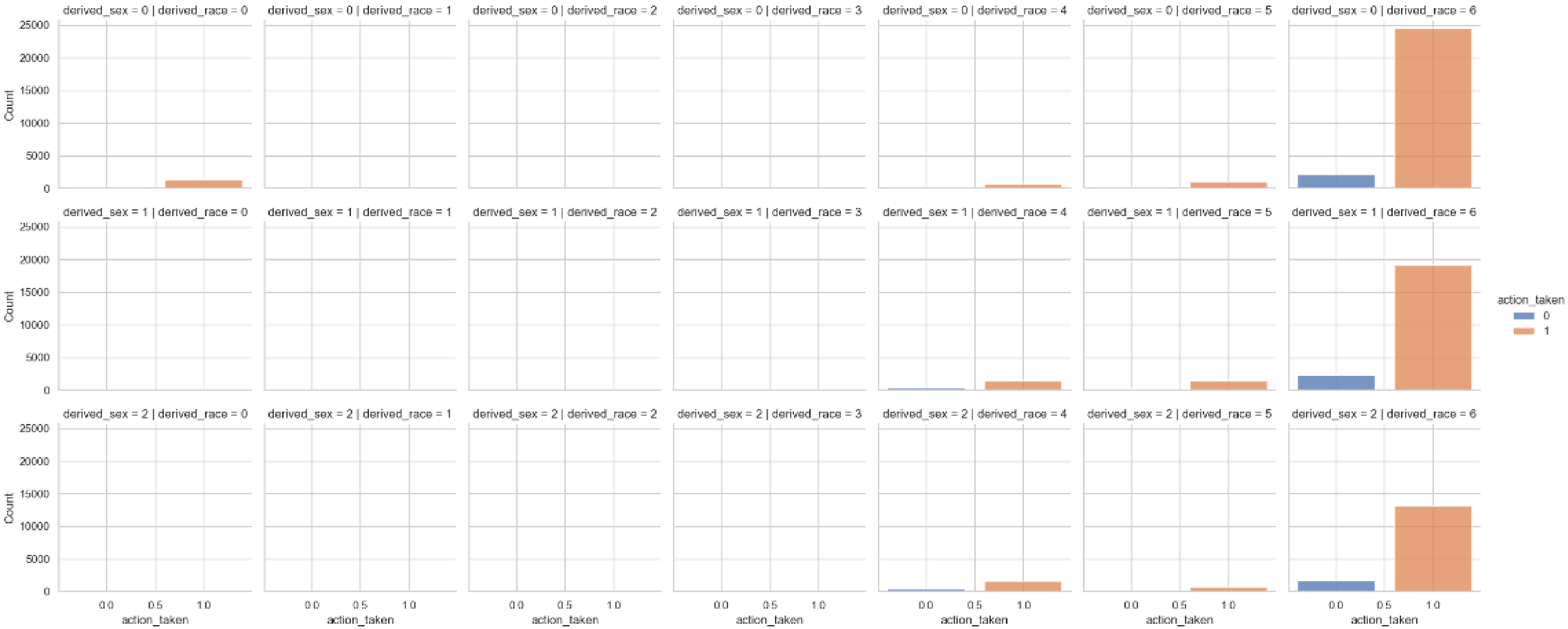


Ethnicity



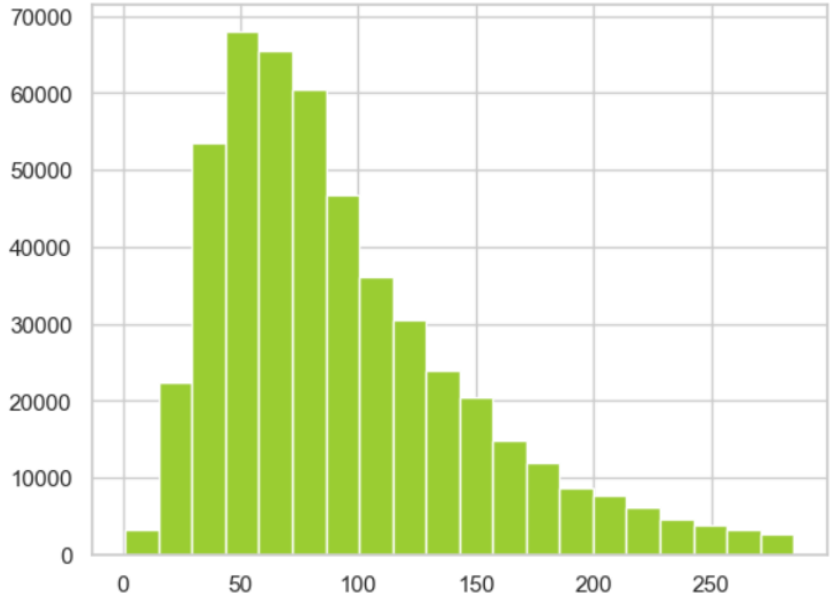
Intersectional Analysis of Outcome vs Sensitive Features

- Sex
- Race



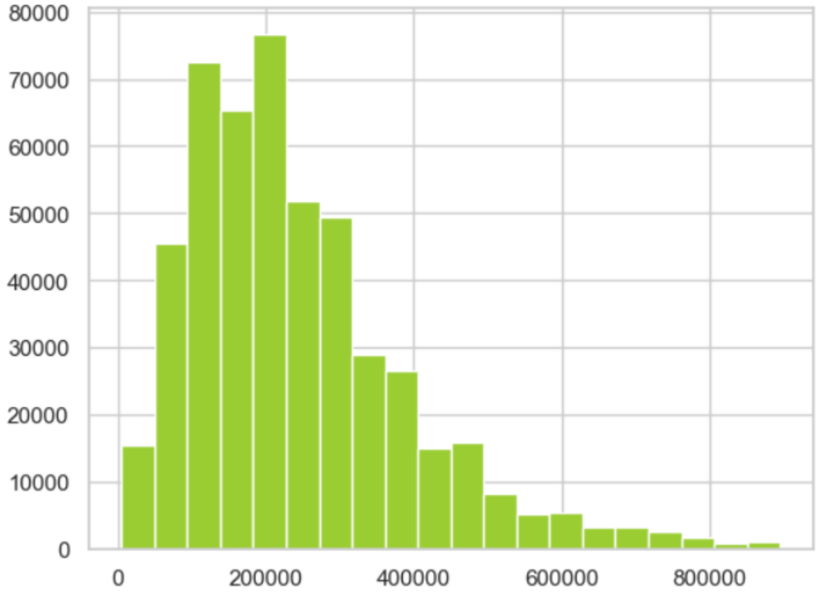
Feature Distribution

Income



The income feature is skewed to the right, with more of the dataset having lower income levels.

Loan Amount



The loan to value ratio feature is skewed to the right, with more of the dataset having lower debt to income ratios.

Feature Engineering

Feature Engineering

- **Select outcomes that resulted in Loan Origination/ Rejection:**
  1-loan originations
  2-Application approved but not accepted
  3-applications denied
  7-Preapproval request denied
  8-Preapproval request approved but not accepted

- **Categorize outcomes into Approved or Denied Loan:**
  - **Loan Approved (1)**
  1 - Loan originated
  2 - Application approved but not accepted
  8 - Preapproval request approved but not accepted

  - **Loan Denied (0)**
  3 - Application denied
  7 - Preapproval request denied

- **Treating outliers (SME advised):**
  - Remove the top 3% and the bottom 1% of income values as they are outliers

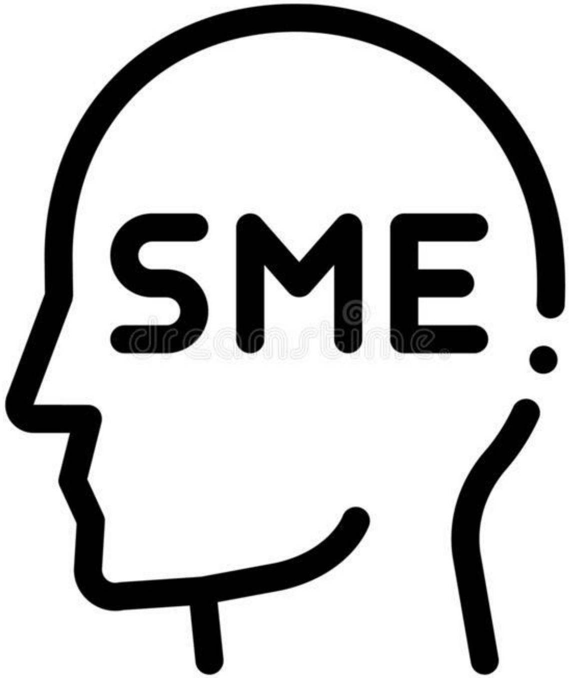
- **Data Cleansing:**
  - Filter on singlefamily 1-4 units
  - Select occupancy type = 1 - principal residences
  - Select on lien_status = 1 - first mortgage loans
  - Eliminate records that don't have a derived_sex
  - Remove values that arent useful
  - Remove “Not Applicable” values for Loan Purpose
  - Remove exempt values for interest_only_payment
  - Dropped records with blanks

- **Group protected/sensitive feature groups:**
  1- Majority
  2,3,4,5,6 - Minority

Feature Extraction

SME Recommended Features

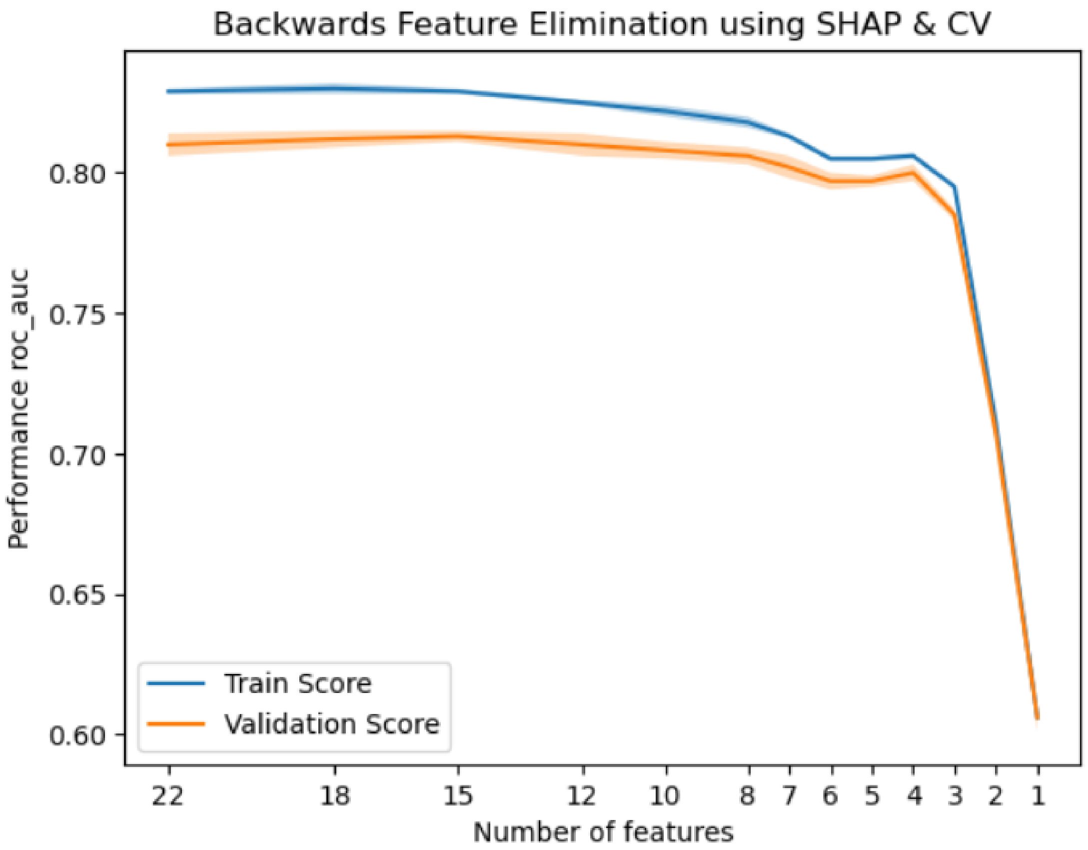
SME Recommended Training Data Features
• Income
• debt_to_income_ratio
• combined_loan_to_value_ratio
• loan_amount
• derived_ethnicity
• derived_race
• derived_sex
• applicant_age
• preapproval
• loan_type
• loan_purpose
• interest_only_payment
• balloon_payment
• action_taken
• loan_term
• property_value
• state_code
• county_code
• tract_minority_population_percent
• tract_population
• tract_to_msa_income_percentage
• tract_median_age_of_housing_units
• ffiec_msa_md_median_family_income



XAI Based Fair Feature Selection with Probatus

PROBATUS

SHAP Based Feature Selection: RandomForest



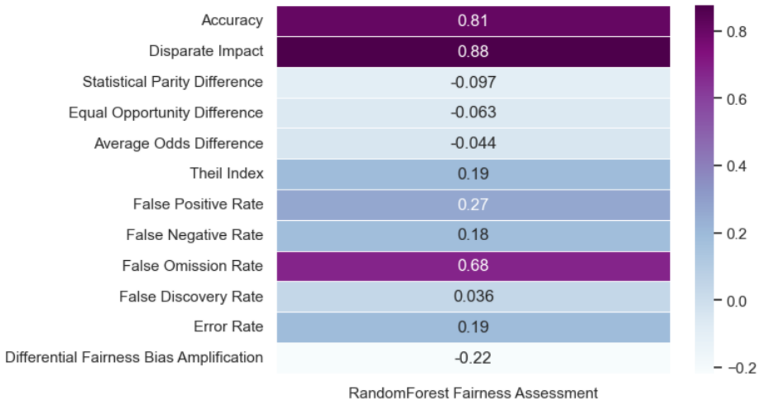
	num_features	features_set	val_metric_mean
1	22	[income, debt_to_income_ratio, combined_loan_t...	0.810
2	18	[loan_type, tract_minority_population_percent,...	0.812
3	15	[tract_minority_population_percent, loan_type,...	0.813
4	12	[derived_race, debt_to_income_ratio, property_...	0.810
5	10	[debt_to_income_ratio, property_value, combine...	0.808
6	8	[debt_to_income_ratio, property_value, combine...	0.806
7	7	[debt_to_income_ratio, property_value, county_...	0.802
8	6	[debt_to_income_ratio, property_value, preappr...	0.797
9	5	[debt_to_income_ratio, property_value, preappr...	0.797
10	4	[preapproval, debt_to_income_ratio, property_v...	0.800
11	3	[debt_to_income_ratio, property_value, loan_pu...	0.785
12	2	[property_value, loan_purpose]	0.707
13	1	[loan_purpose]	0.606

SHAP Based Feature Selection: RandomForest

Training Parameters	Evaluation Metric Value (ROC_AUC)
• **Hyperparameters:** {'reg_lambda': 1.0, 'reg_alpha': 0.1, 'min_child_weight': 0.5, 'max_depth': 5, 'colsample_bylevel': 0.5} • **No. of Features:** 7 • **Features:** ['debt_to_income_ratio', 'property_value', 'combined_loan_to_value_ratio', 'county_code', 'preapproval', 'loan_purpose', 'state_code']	0.802
• **Hyperparameters:** {'reg_lambda': 1.0, 'reg_alpha': 0.1, 'min_child_weight': 0.5, 'max_depth': 5, 'colsample_bylevel': 0.5} • **No. of Features:** 8 • **Features:** ['debt_to_income_ratio', 'property_value', 'combined_loan_to_value_ratio', 'county_code', 'preapproval', 'loan_purpose', 'loan_amount', 'state_code']	0.806
• **Hyperparameters:** {'reg_lambda': 1.0, 'reg_alpha': 0.1, 'min_child_weight': 0.5, 'max_depth': 5, 'colsample_bylevel': 0.5} • **No. of Features:** 10 • **Features:** ['debt_to_income_ratio', 'loan_type', 'property_value', 'state_code', 'combined_loan_to_value_ratio', 'county_code', 'preapproval', 'loan_amount', 'loan_purpose', 'income']	0.808
• **Hyperparameters:** {'reg_lambda': 1.0, 'reg_alpha': 0.1, 'min_child_weight': 0.5, 'max_depth': 5, 'colsample_bylevel': 0.5} • **No. of Features:** 12 • **Features:** ['derived_race', 'debt_to_income_ratio', 'loan_type', 'property_value', 'income', 'combined_loan_to_value_ratio', 'county_code', 'preapproval', 'loan_purpose', 'loan_amount', 'applicant_age', 'state_code']	0.810
• **Hyperparameters:** {'reg_lambda': 1.0, 'reg_alpha': 0.1, 'min_child_weight': 0.5, 'max_depth': 5, 'colsample_bylevel': 0.5} • **No. of Features:** 15 • **Features:** ['tract_minority_population_percent', 'loan_type', 'income', 'county_code', 'loan_term', 'derived_race', 'ffiec_msa_md_median_family_income', 'loan_amount', 'applicant_age', 'debt_to_income_ratio', 'property_value', 'combined_loan_to_value_ratio', 'preapproval', 'loan_purpose', 'state_code']	0.813
• **Hyperparameters:** {'reg_lambda': 1.0, 'reg_alpha': 0.1, 'min_child_weight': 0.5, 'max_depth': 5, 'colsample_bylevel': 0.5} • **No. of Features:** 18 • **Features:** ['loan_type', 'tract_minority_population_percent', 'tract_median_age_of_housing_units', 'income', 'county_code', 'loan_term', 'tract_population', 'loan_amount', 'derived_race', 'tract_to_msa_income_percentage', 'ffiec_msa_md_median_family_income', 'applicant_age', 'debt_to_income_ratio', 'property_value', 'combined_loan_to_value_ratio', 'preapproval', 'loan_purpose', 'state_code']	0.812
• **Hyperparameters:** {'reg_lambda': 1.0, 'reg_alpha': 0.1, 'min_child_weight': 0.5, 'max_depth': 5, 'colsample_bylevel': 0.5} • **No. of Features:** 22 • **Features:** ['income', 'debt_to_income_ratio', 'combined_loan_to_value_ratio', 'loan_amount', 'derived_ethnicity', 'derived_race', 'derived_sex', 'applicant_age', 'preapproval', 'loan_type', 'loan_purpose', 'interest_only_payment', 'balloon_payment', 'loan_term', 'property_value', 'state_code', 'county_code', 'tract_minority_population_percent', 'tract_population', 'tract_to_msa_income_percentage', 'tract_median_age_of_housing_units', 'ffiec_msa_md_median_family_income']	0.810

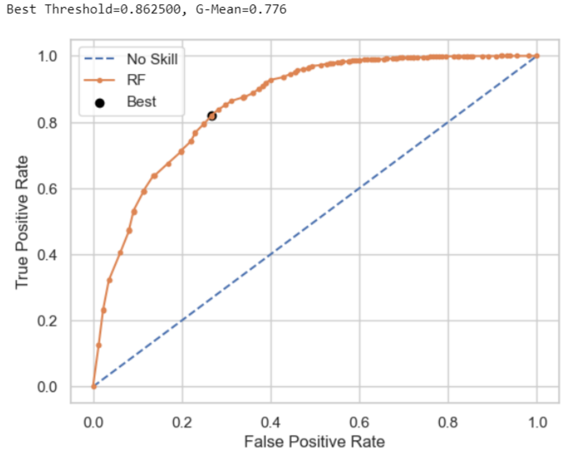
Fairness Assessment Comparison: Race

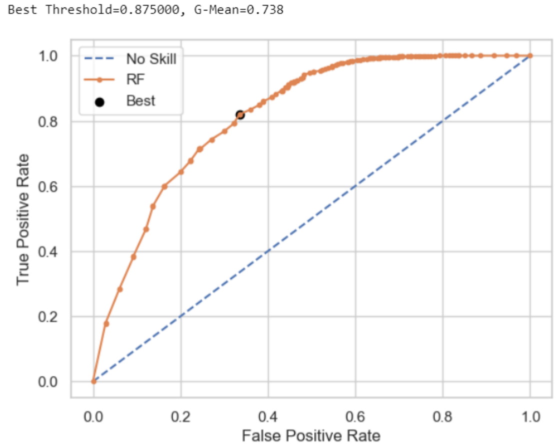
XAI Features: Including Race



XAI Features: Excluding Race

AUC/ROC Curve

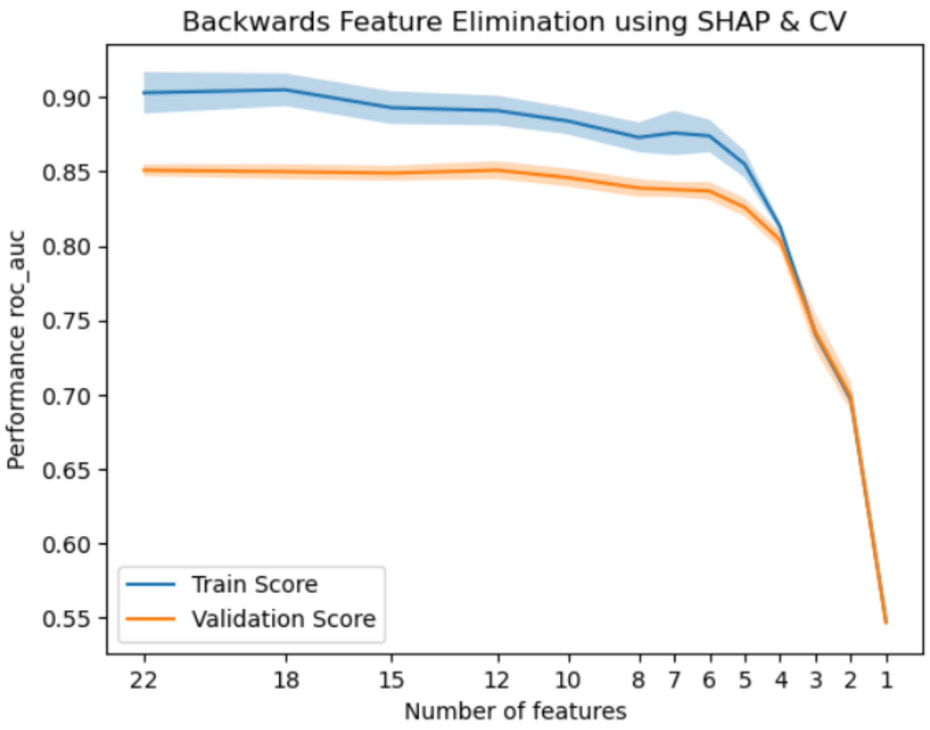




Features: ['tract_minority_population_percent', 'loan_type', 'income', 'county_code', 'loan_term', **'derived_race'**, 'ffiec_msa_md_median_family_income', 'loan_amount', 'applicant_age', 'debt_to_income_ratio', 'property_value', 'combined_loan_to_value_ratio', 'preapproval', 'loan_purpose', 'state_code']

Features:
['debt_to_income_ratio', 'loan_type', 'property_value', 'state_code', 'combined_loan_to_value_ratio', 'county_code', 'preapproval', 'loan_amount', 'loan_purpose', 'income']

SHAP Based Feature Selection: XGBoost



	num_features	features_set	val_metric_mean
1	22	[income, debt_to_income_ratio, combined_loan_t...	0.851
2	18	[loan_type, tract_minority_population_percent,...	0.850
3	15	[tract_minority_population_percent, loan_type,...	0.849
4	12	[derived_race, debt_to_income_ratio, loan_type...	0.851
5	10	[debt_to_income_ratio, loan_type, property_val...	0.846
6	8	[debt_to_income_ratio, property_value, combine...	0.839
7	7	[debt_to_income_ratio, property_value, combine...	0.838
8	6	[debt_to_income_ratio, property_value, combine...	0.837
9	5	[debt_to_income_ratio, property_value, county_...	0.826
10	4	[preapproval, debt_to_income_ratio, property_v...	0.804
11	3	[preapproval, debt_to_income_ratio, loan_purpose]	0.742
12	2	[preapproval, debt_to_income_ratio]	0.699
13	1	[preapproval]	0.547

Fair Feature Selection with FFS

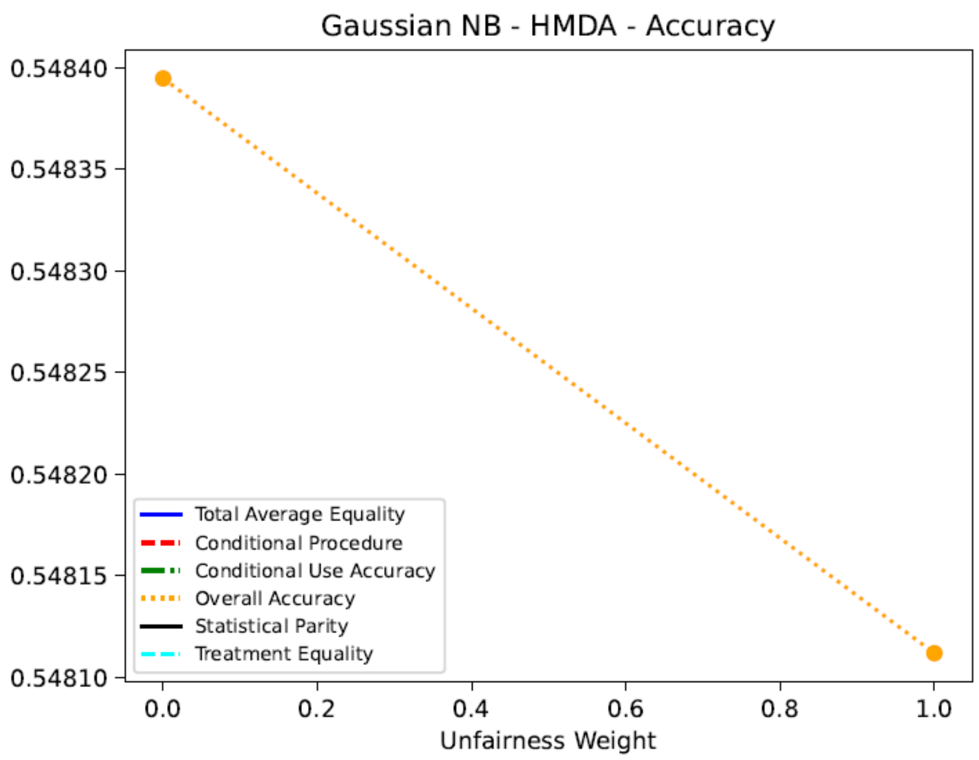
Automating Procedurally Fair Feature Selection in Machine Learning

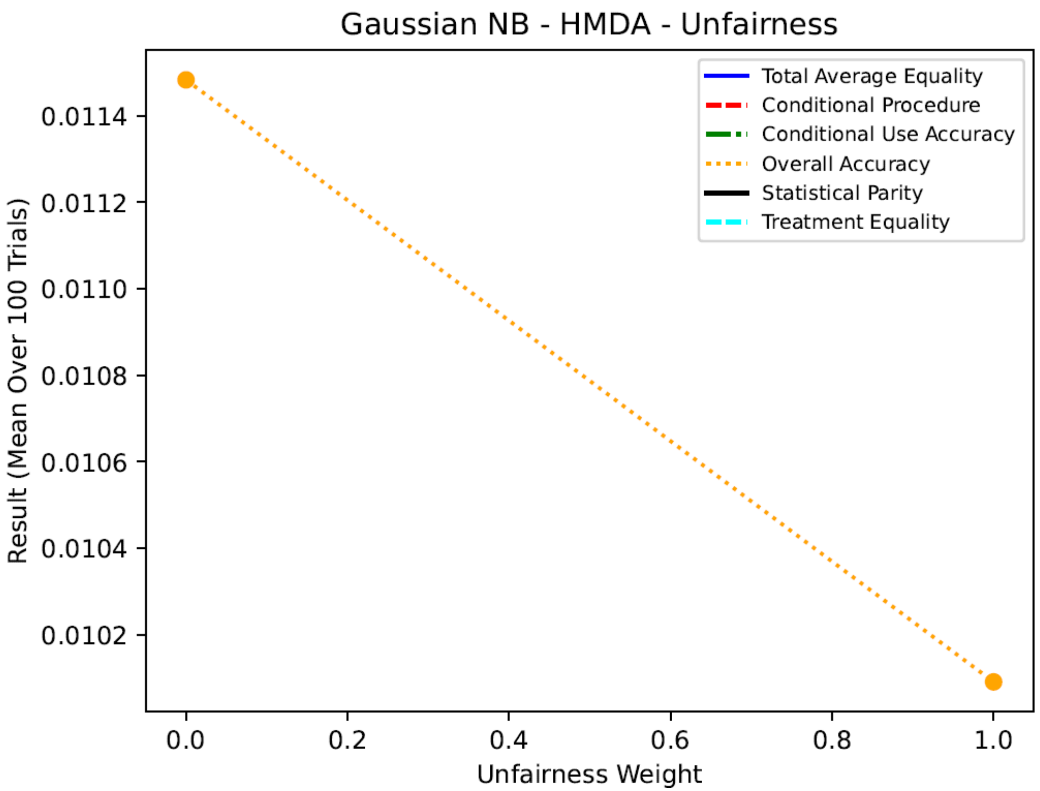
Clara Belitz
University of Illinois
Urbana–Champaign
Champaign, IL, USA
cbelitz2@illinois.edu

Lan Jiang
University of Illinois
Urbana–Champaign
Champaign, IL, USA
lanj3@illinois.edu

Nigel Bosch
University of Illinois
Urbana–Champaign
Champaign, IL, USA
pnb@illinois.edu

Effect of Unfairness Weight on Accuracy & Fairness





Unfairness Weights: 0,1,2,3,4

Effect of Unfairness Weight on Accuracy & Fairness

model	unfairness_metric	unfairness_weight	iteration	unfairness	auc	unfairness_scaled	protected_column_selected_prop
GaussianNB	overall_accuracy_equality	0	1	0.01110778	0.548255027	0.33949866	0.5
GaussianNB	overall_accuracy_equality	0	2	0.012673039	0.548475748	0.594346585	0.25
GaussianNB	overall_accuracy_equality	0	3	0.01033668	0.548312514	0.213951797	0.5
GaussianNB	overall_accuracy_equality	0	4	0.012376864	0.547820423	0.546124787	0.25
GaussianNB	overall_accuracy_equality	0	5	0.011125865	0.548211159	0.342443134	0.5
GaussianNB	overall_accuracy_equality	0	6	0.010054859	0.548288688	0.168067124	0.25
GaussianNB	overall_accuracy_equality	0	7	0.010825149	0.548547466	0.293482053	0.5
GaussianNB	overall_accuracy_equality	0	8	0.011934804	0.548180179	0.474150776	0.25
GaussianNB	overall_accuracy_equality	0	9	0.012532974	0.548522802	0.571541894	0.5
GaussianNB	overall_accuracy_equality	0	10	0.01145999	0.548558632	0.396843874	0.25
GaussianNB	overall_accuracy_equality	0	11	0.010331136	0.548325458	0.21304912	0.25
GaussianNB	overall_accuracy_equality	0	12	0.011257508	0.548410869	0.363876658	0.5
GaussianNB	overall_accuracy_equality	0	13	0.010946889	0.548358524	0.313303148	0.25
GaussianNB	overall_accuracy_equality	0	14	0.011330916	0.54840898	0.375828533	0
GaussianNB	overall_accuracy_equality	0	15	0.011626862	0.548605232	0.42401305	0.5
GaussianNB	overall_accuracy_equality	0	16	0.011964312	0.548593329	0.478955133	0.5
GaussianNB	overall_accuracy_equality	0	17	0.012766976	0.548546274	0.60964093	0.75
GaussianNB	overall_accuracy_equality	0	18	0.012170372	0.54846999	0.512504859	0.25

Conclusion

Study Conclusions

- AI bias can be diagnosed and mitigated in the training data before being amplified by the models.
- Feature selection plays a crucial role in not just the discriminatory power of models but also in discriminating among population groups.
- The results of the approach must be subject to SME scrutiny
- The methods used are resource intensive

Questions?