Home Mortgage Approval

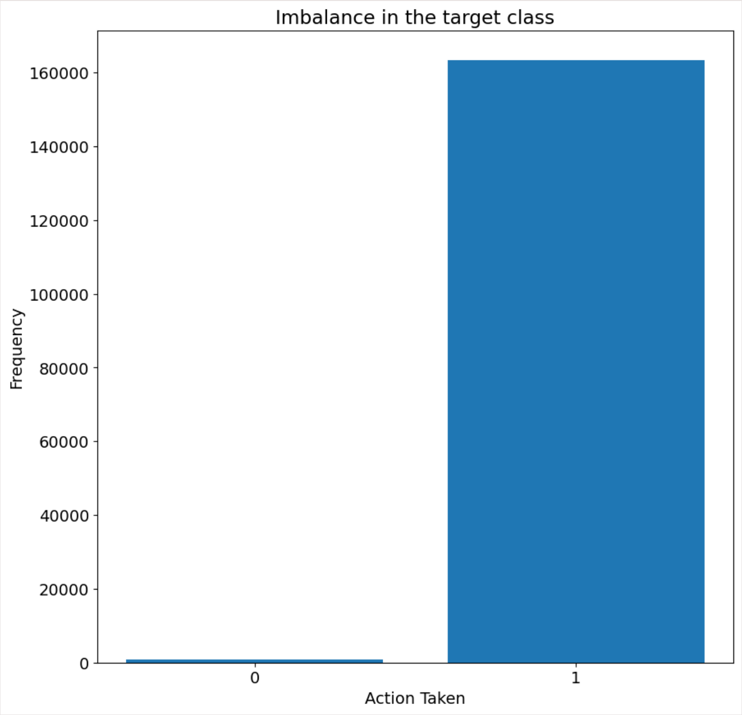
Ankit Mistry and Shivam Mistry

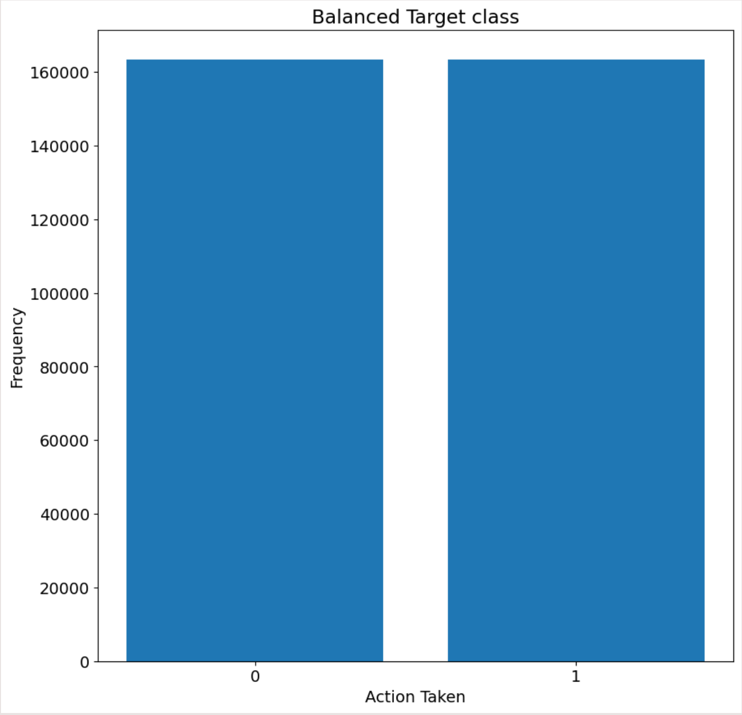
Introduction and Motivation

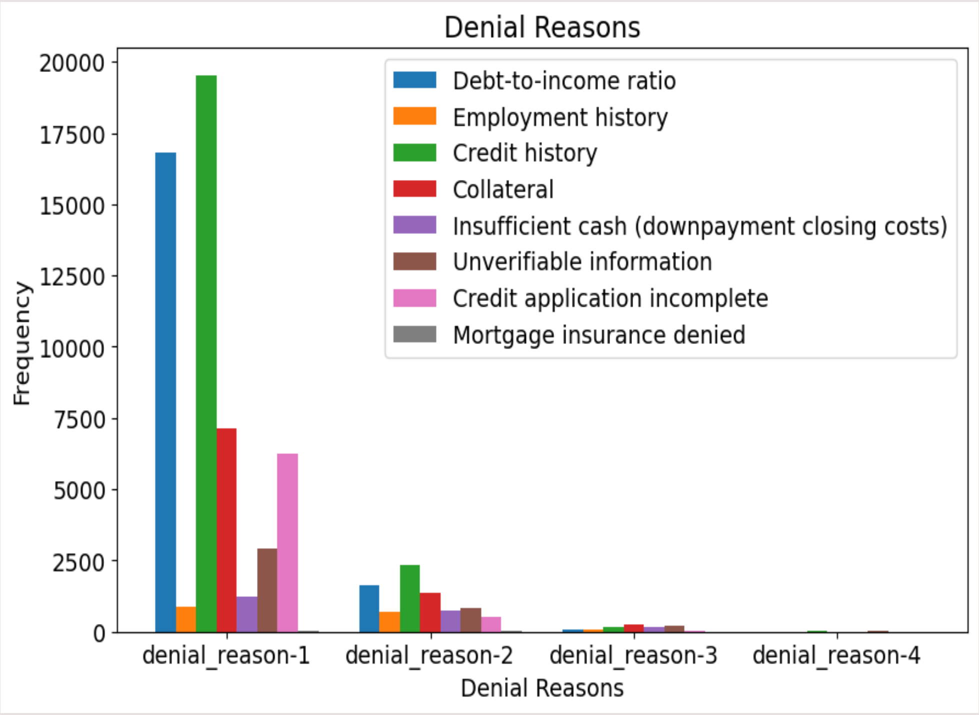
- Our primary objective was to explore which features are most significant when determining eligibility for mortgage.
- Also evaluating the suitability of machine learning techniques in this context and discovering dataset biases.

Dataset

- The dataset used for this project was sourced from the Home Mortgage Disclosure Act (HMDA) for the year 2022, specifically focused on the state of Tennessee.
- We chose the dataset from HMDA considering its credibility and diversity.



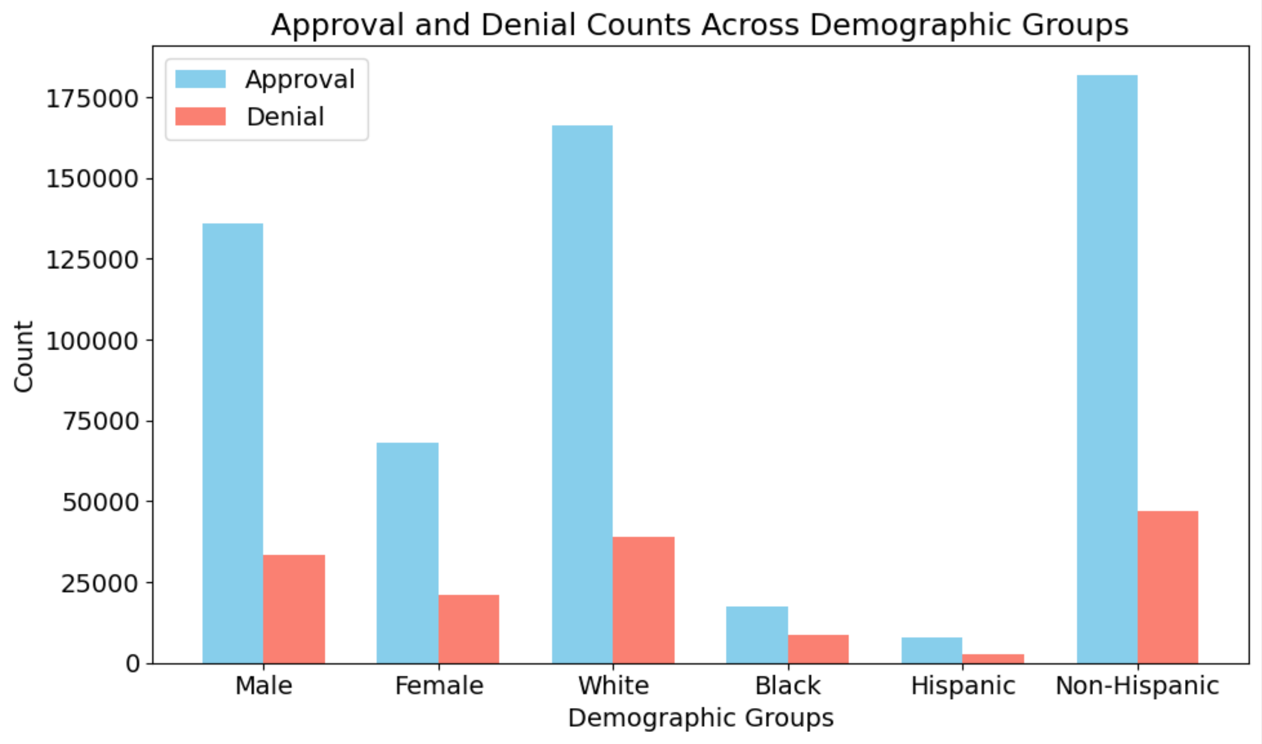




Dataset: Biases

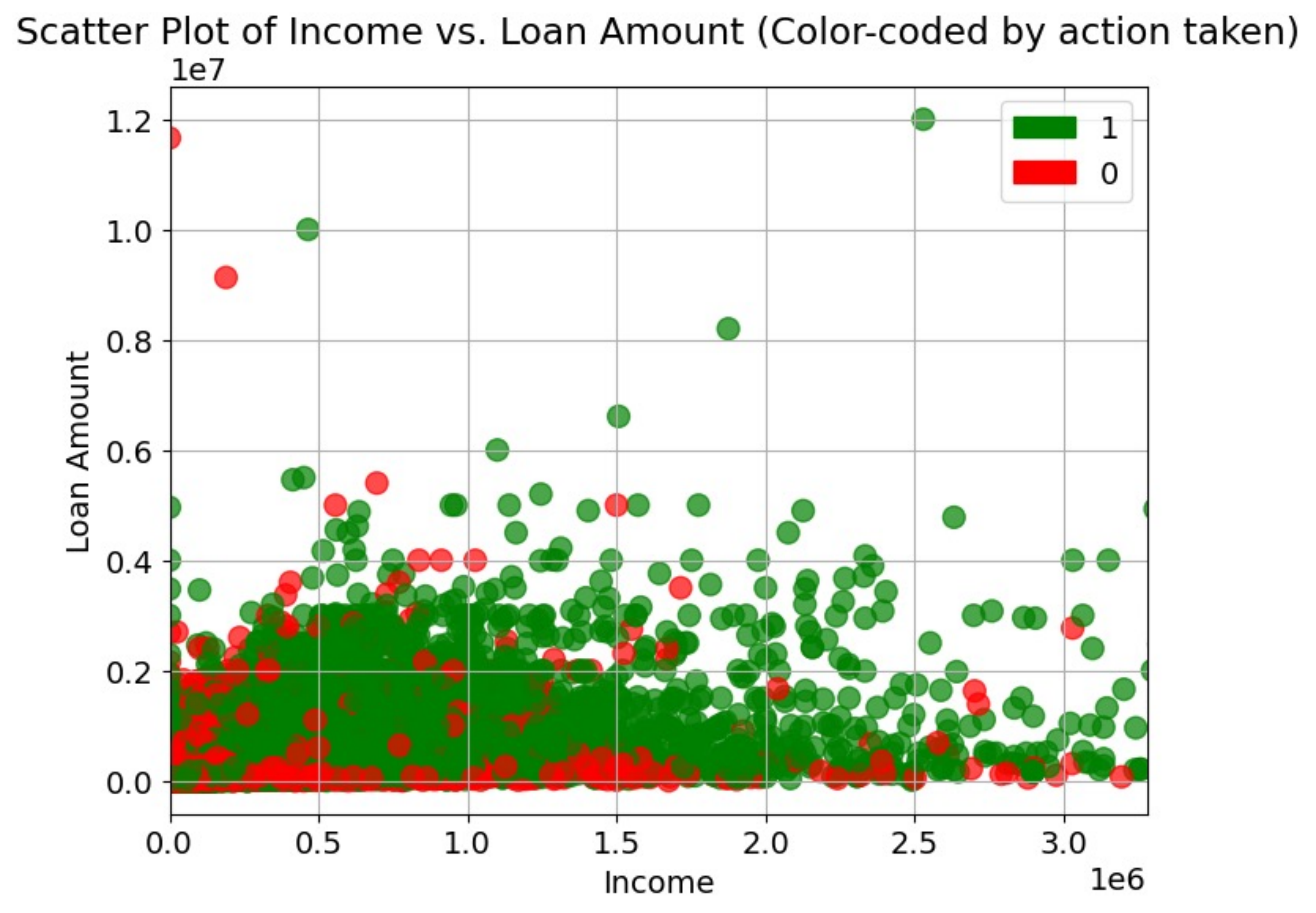
- Since it is well known that biases exists when making credit decisions, we did the following analysis:

Groups	Count	Income	Loan Amount	Interest Rate
Male	126102	178477.1	280438.38	4.84
Female	64466	101403.08	232323.4	4.92
White	154297	160196.45	263447.67	4.87
Black	16448	95395.91	220867.58	4.84
Hispanic	7567	109010.31	262646.36	5.14
Non Hispanic	169163	155873.35	261915.28	4.85



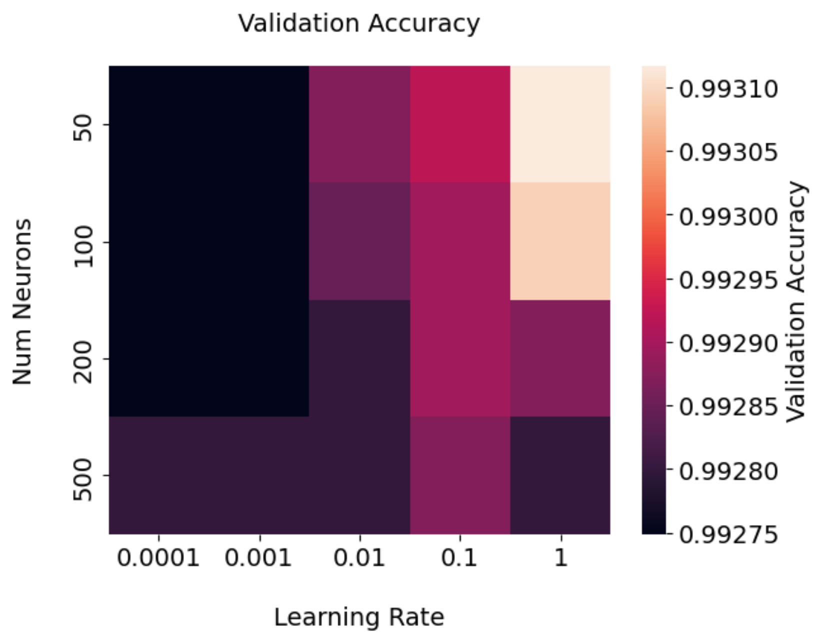
Dataset: Feature Selection

- We chose the following features:
  - *action_taken*
  - *purchaser_type*
  - *preapproval*
  - *loan_type*
  - *loan_purpose*
  - *business_or_commercial_purpose*
  - *loan_amount*
  - *loan_to_value_ratio*
  - *interest_rate*
  - *hoepa_status*
  - *property_value*
  - *occupancy_type*
  - *income*
  - *debt_to_income_ratio*
  - *applicant_credit_score_type*



ML Methods: MLP Classifier

- Hyperparameters: learning_rate_init, hidden_layer_sizes
- learning_rate_init = [0.0001, 0.001, 0.01, 0.1, 1]
- hidden_layer_sizes = [(100), (100, 100), (100, 100, 100), (200), (200,100), (200,100,100), (500), (500,200), (500,200,100)]



Network Combinations	Accuracy
100	99.30%
(100, 100)	99.21%
(100, 100, 100)	97.83%
200	99.28%
(200, 100)	97.89%
(200, 100, 100)	97.83%
500	99.27%
(500, 200)	99.27%
(500, 200, 100)	99.04%

ML Methods: Random Forests

- Hyperparameters: max_depth
- Our first approach was not oversampling, but it gave us skewed results, so we oversampled and tested the hyperparameters.

Depth	Training Accuracy	Testing Accuracy
2	0.979238	0.980452
4	0.992525	0.992847
6	0.992767	0.992896
8	0.993022	0.993141
10	0.994023	0.993949
12	0.994429	0.994292
14	0.995405	0.995052
16	0.996595	0.996203
18	0.997613	0.997134
20	0.998541	0.997624
22	0.999220	0.998212
24	0.999685	0.998530
26	0.999864	0.998628
28	0.999892	0.998555
30	0.999895	0.998579
32	0.999895	0.998530
34	0.999895	0.998555
36	0.999895	0.998555
38	0.999895	0.998555

ML Methods: AdaBoost

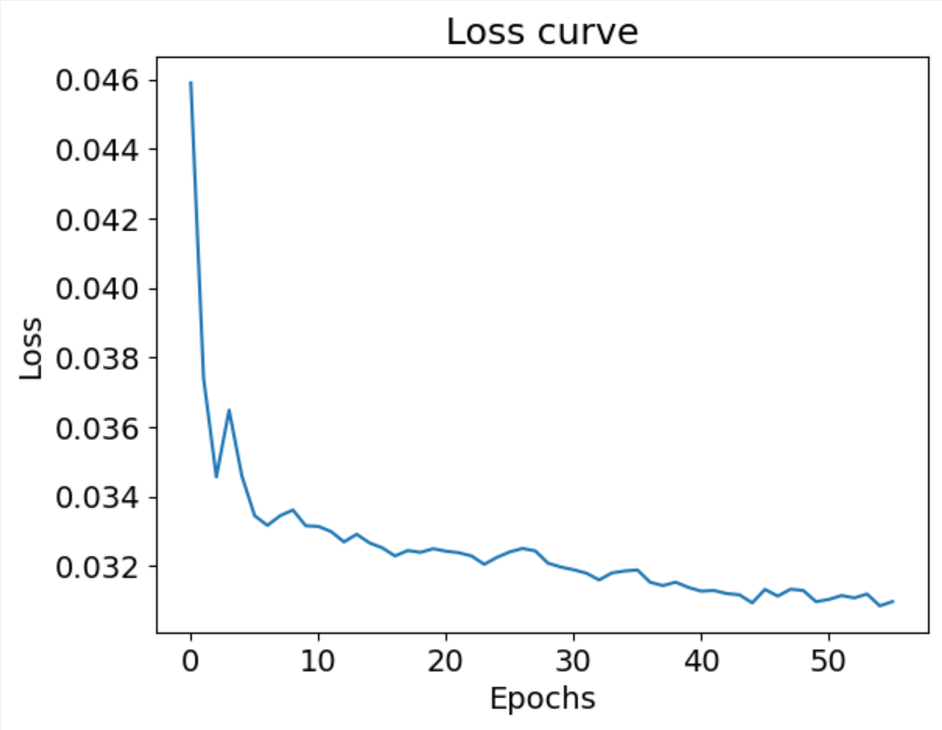
- Hyperparameters: n_estimators
- n_estimators = [50, 100, 150, 200, 250]

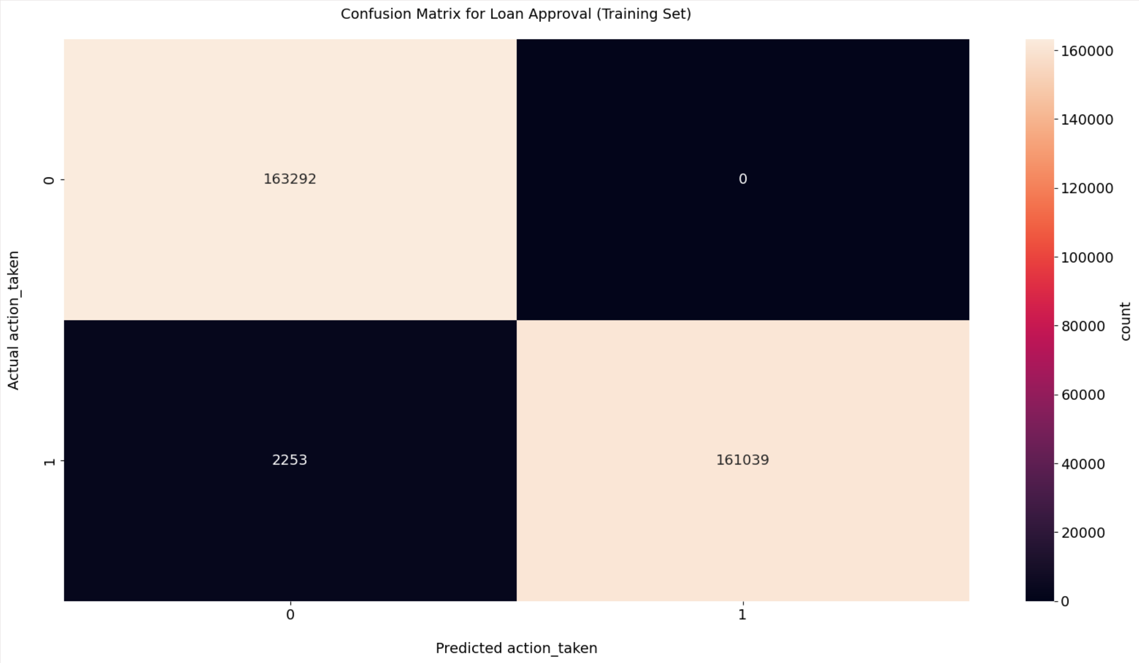
Number of Estimators	Training Accuracy	Testing Accuracy
50	99.33%	99.32%
100	99.34%	99.33%
150	99.35%	99.33%
200	99.34%	99.35%
250	99.34%	99.35%

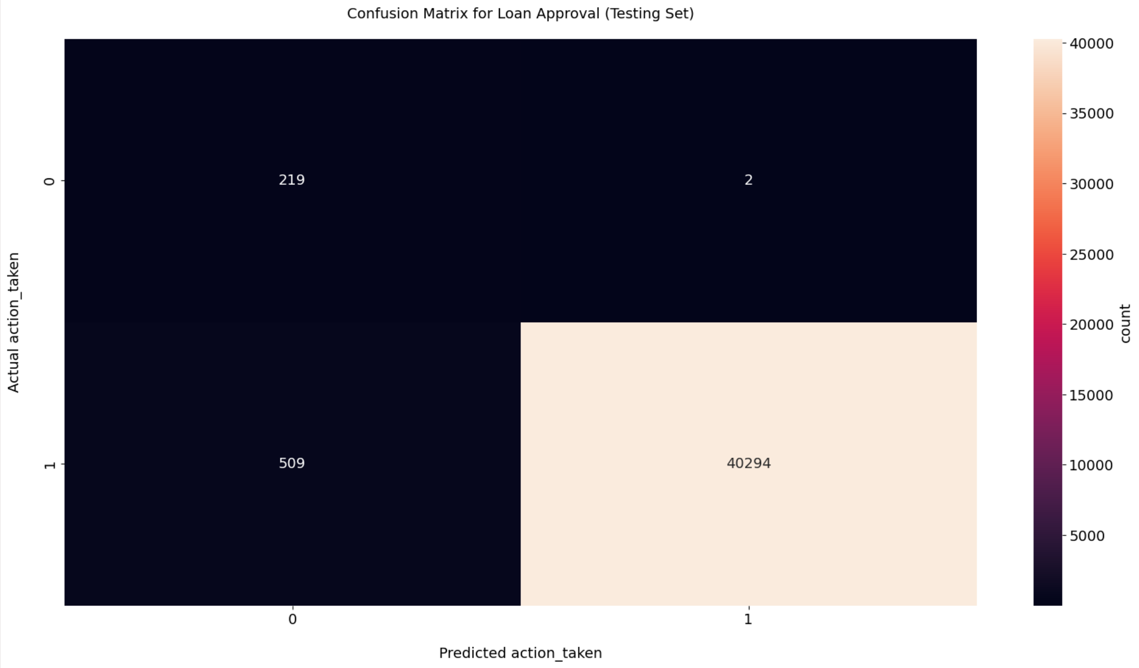
ML Methods: Summary

	MLP Classifier	Random Forest	AdaBoost
Training Accuracy	99.3%	99.9%	99.3%
Testing Accuracy	98.8%	99.6%	98.8%
F1 Score	99.4	99.8	99.4

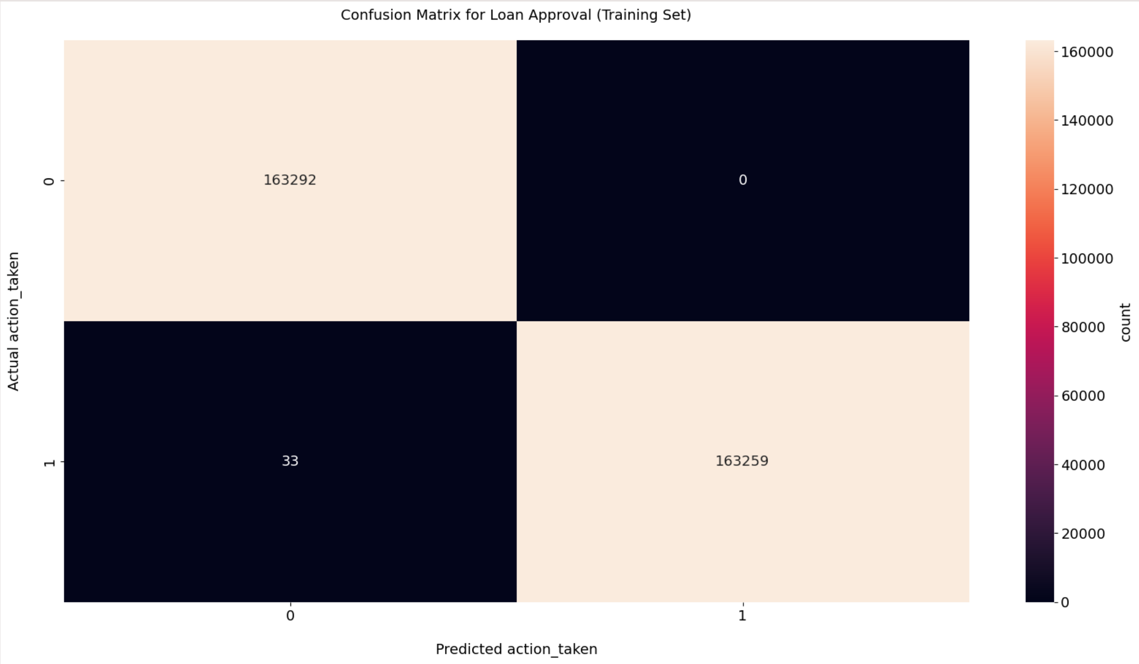
Results: MLP Classifier

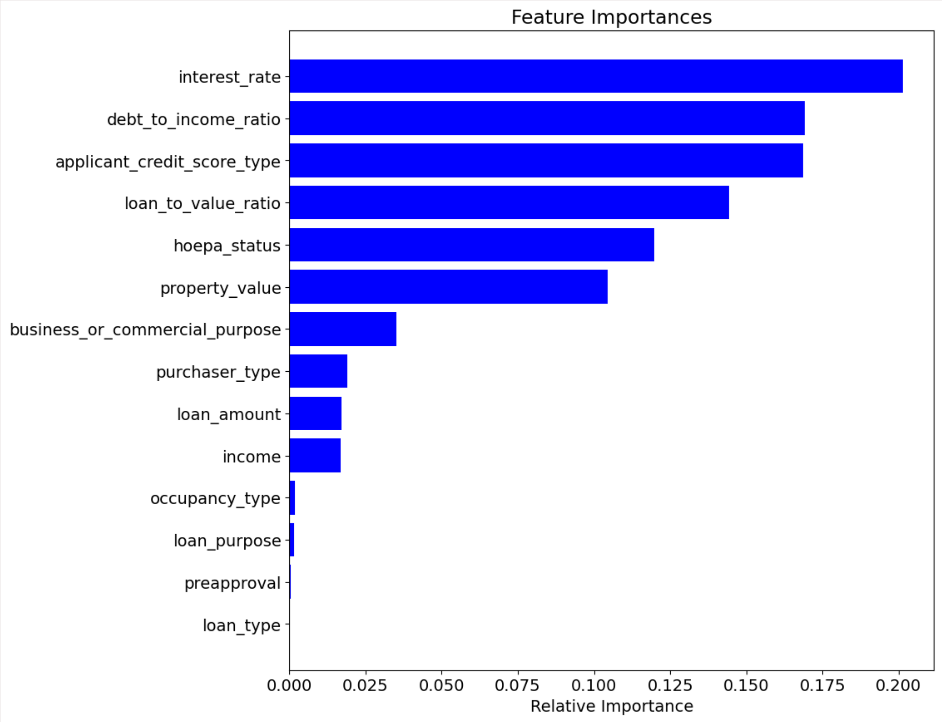


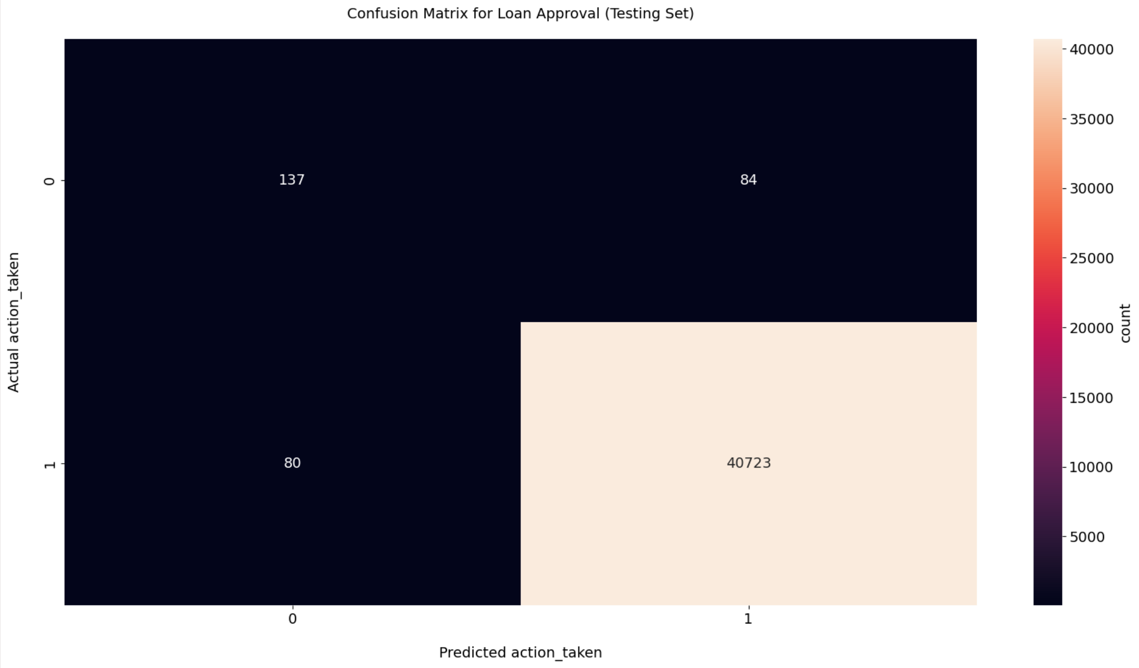




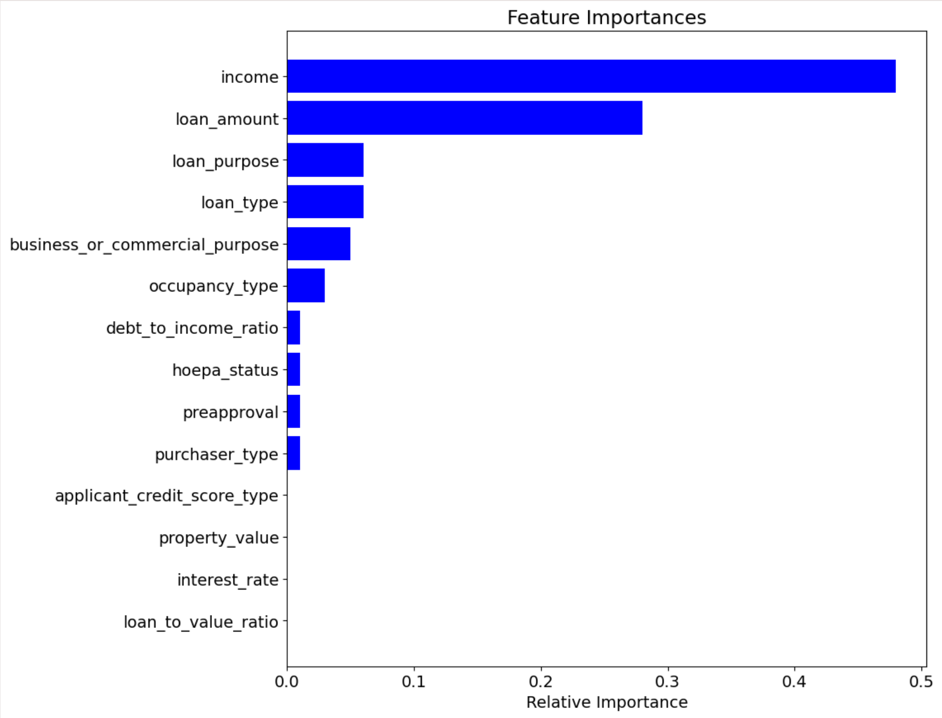
Results: Random Forest Classifier

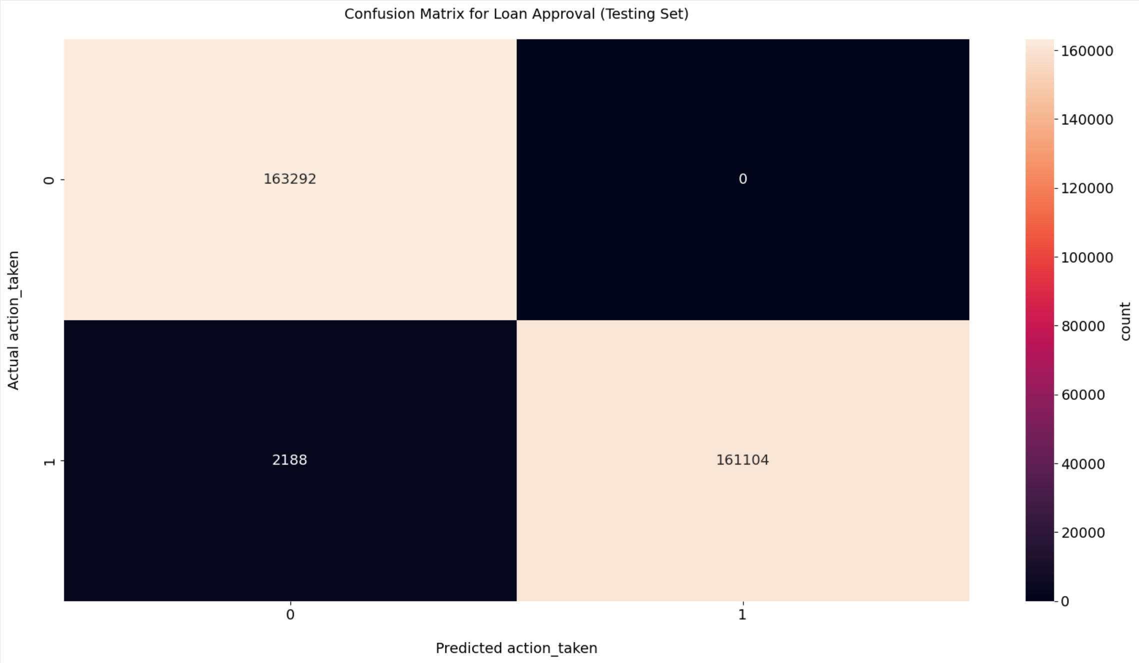


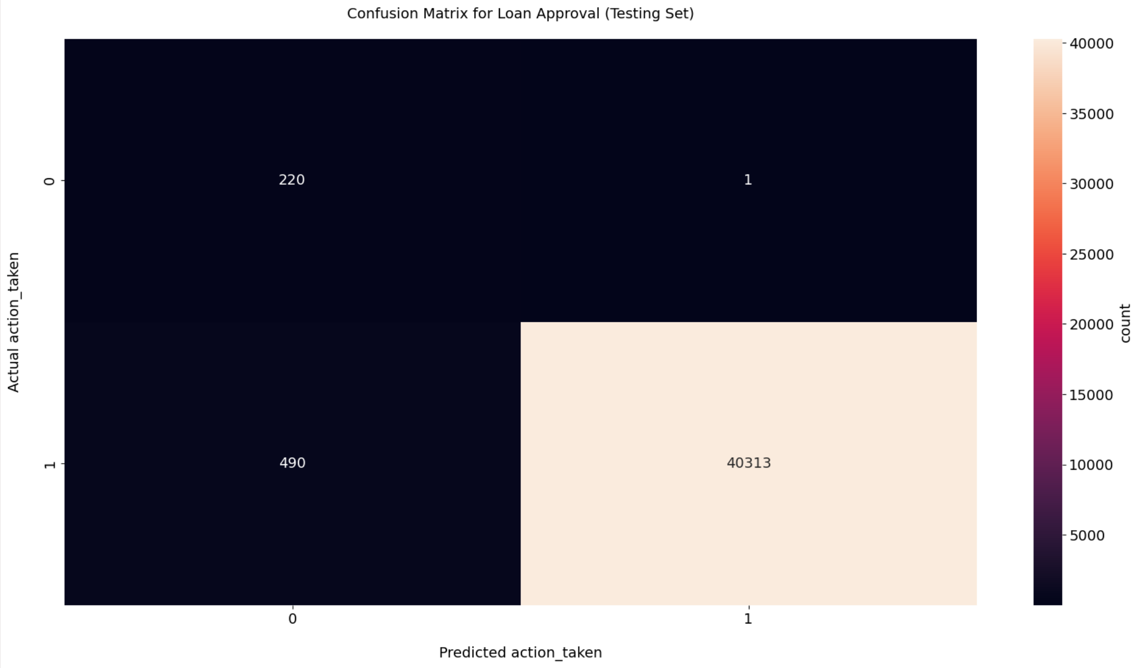




Results: AdaBoost Classifier







Future Work

- Looking ahead, we would like to get a preprocessed dataset directly from authoritative sources like HMDA or any other institution that specialized in gathering such data. This would help enhance data integrity and accuracy.
- Also, in future we would like to prioritize mitigation of biases ingrained within the dataset and strive for balanced samples of approved and denied applications across all demographic groups, fostering a more equitable dataset for robust analysis and modeling.