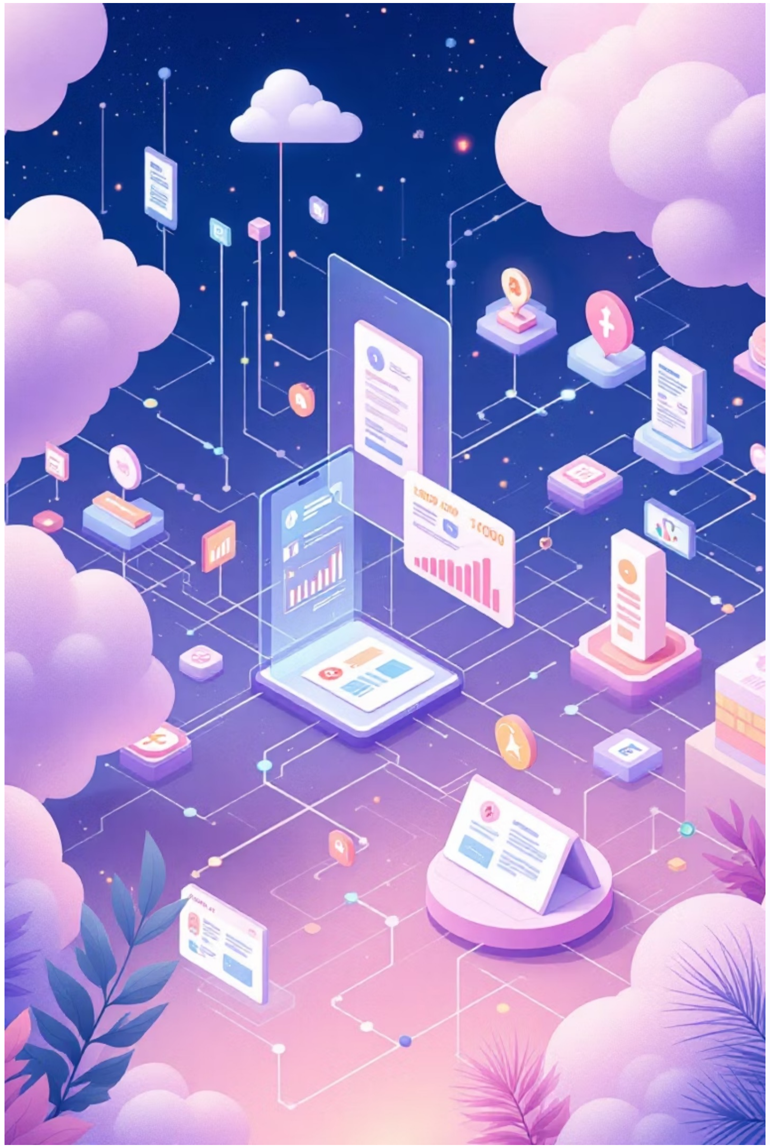


Evaluating Algorithmic Fairness: A Comparative Audit of Machine Learning Models in Loan Approval System

Esha Agarwal

GH1031345

Introduction & Motivation

The financial sector is increasingly adopting AI. However, "black box" algorithms pose ethical risks regarding protected features like gender.

Protected Features/Sensitive Attributes: The feature that may induce bias towards human and/or algorithms
Black Box Algorithm: A system that produces an output from an input without revealing how it reached that conclusion

Management Problem

Financial institutions pose a risk of regulatory non-compliance & reputational damage if, their *loan approval algorithms* exhibit ***disparate impact***.

Literature Review

This literature review conducted over key academic contributions relevant to algorithmic fairness in machine learning models, particularly within financial decision-making processes.

Author & Year	Theme	Relevance
Mehrabi et al. (2021)	Bias Origins	Algorithms often inherit "historical bias" from training data. If past loan officers discriminated, the model learns to do the same.
Dutta et al. (2020)	The Trade-off	There is often an inherent tension between maximizing Accuracy and maximizing Fairness. Increasing one frequently degrades the other.
Zafar et al. (2017)	Fairness Metrics	Defined "Disparate Mistreatment" (error rate differences) as a critical metric for decision boundaries in sensitive domains.
Sudhakar et al. (2020)	Model Comparison	Random Forest consistently outperforms Logistic Regression in pure accuracy but suffers from "Black Box" opacity, making it harder to audit for bias.

Fig: Relevant Literature

Research Objectives & Questions

Research Gap

1. The direct, comparative audit of three specific models (SVM, RF, and LR) focused on their performance on financial data across “Disparate Impact” & “Disparate Mistreatment”.

Questions

1. Does historical loan data already contain gender bias?
2. Which classification algorithm (SVM, RF, or LR) has the least unfair errors: “Disparate Impact” & “Disparate Mistreatment”?

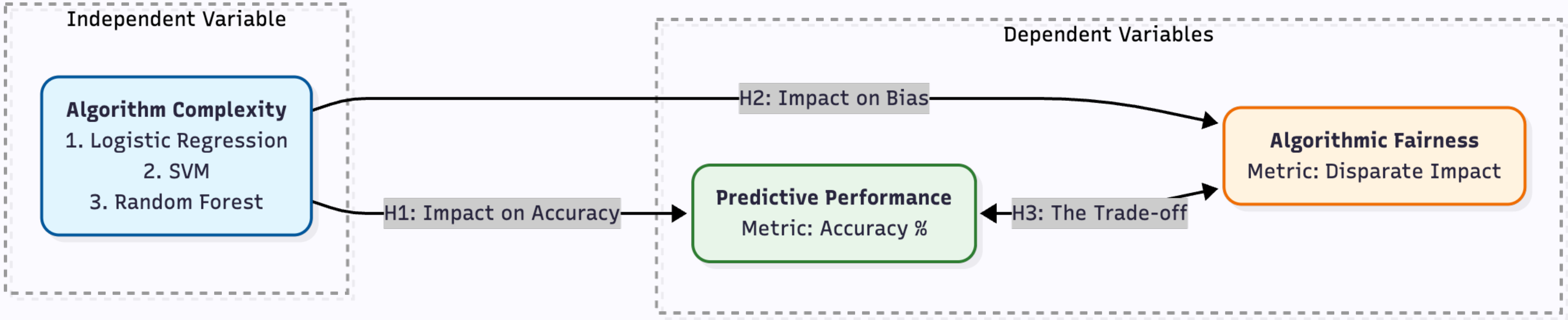
Objectives

1. To audit 3 distinct models (SVM, RF, LR) for fairness.

Definitions

- **SVM** - Support Vector Machine, Steinwart & Christmann, 2008.
- **RF** - Random Forest, Breiman, L. Random Forests. Machine Learning 45, 5–32 (2001).
- **LR** - Logistic Regression, King and Zeng, 200

Conceptual Framework



Hypotheses:

- H1 (Impact on Accuracy):
  - How does the model choice **impact its accuracy**?
- H2 (Impact on Bias):
  - How does the model choice **impact its fairness**?
- H3 (The Trade-off):
  - Represents the tension between maximizing **accuracy** and maximizing **fairness**.

Methodology

Quantitative Experimental Design

Comparison of 3 Classification Machine Learning Models

Data Source

Secondary Data Available openly on Kaggle

Fairness Metrics

"Disparate Impact" & "Disparate Mistreatment" (accuracy gaps)

Fairness Metrics Details

Metric 1: Disparate Impact (Statistical Parity)

- **Definition:** The ratio of the probability of a positive outcome (Approval) for the protected group vs. the unprotected group.
- **Threshold:** The "80% Rule" (legal standard in US employment/lending).

Metric 2: Disparate Mistreatment (Predictive Equality)

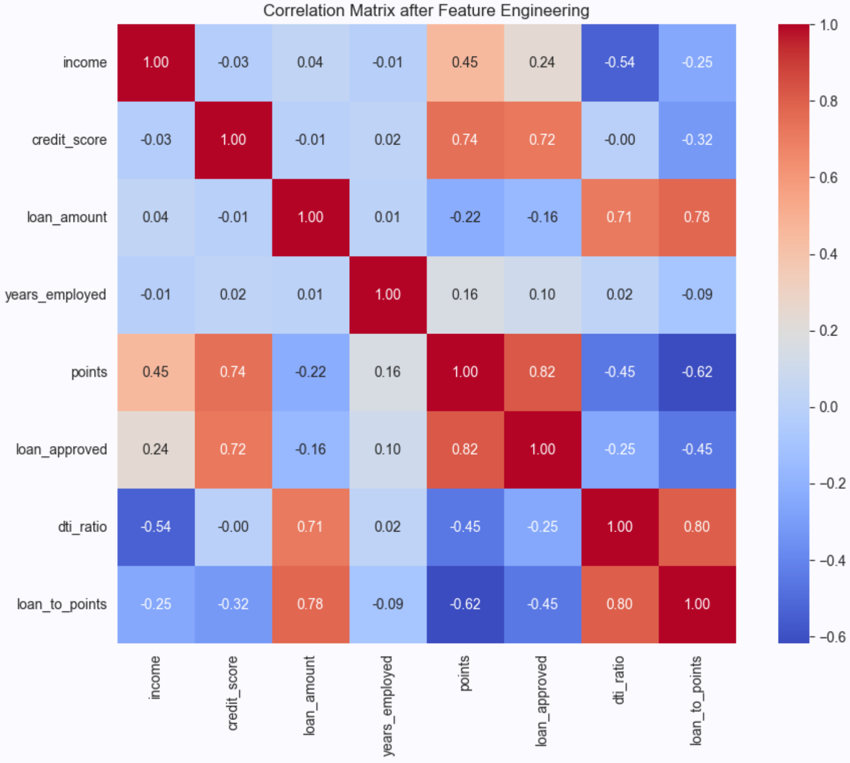
- **Definition:** The difference in Accuracy between groups.
- **Relevance:** Ensuring one group is not "wrongly denied" more often than another.

Data Collection

- Secondary data sourcing from a public repository (Kaggle), utilising 2,000 loan application records
- 7 features and 1 binary target variable

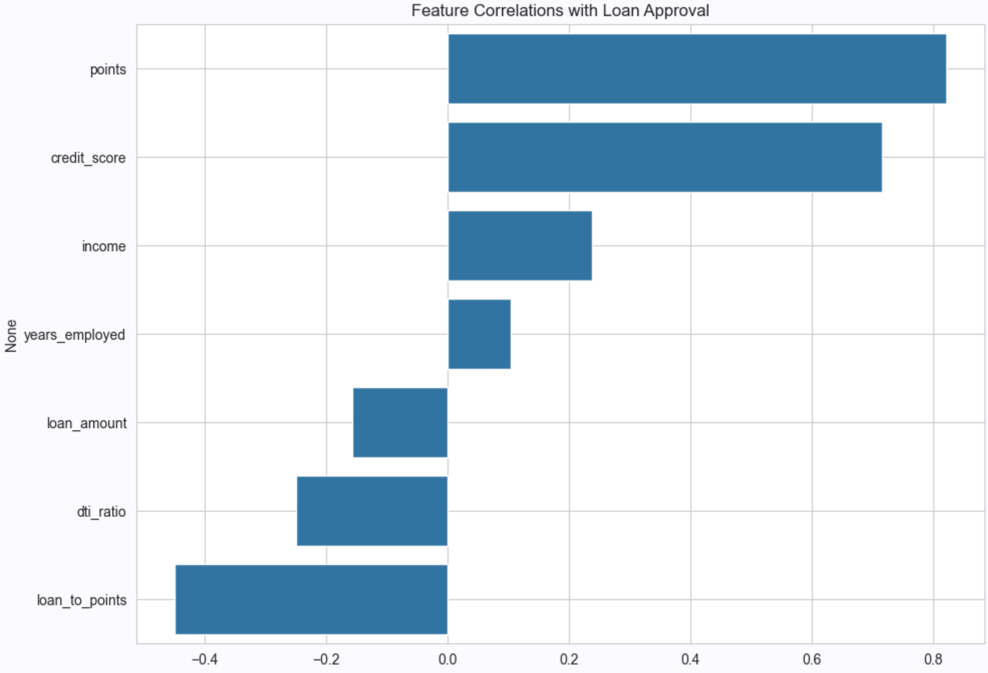
Results

Preliminary Test – Correlation Matrix



Correlation Matrix after Feature Engineering

	income	credit_score	loan_amount	years_employed	points	loan_approved	dti_ratio	loan_to_points
income	1.00	-0.03	0.04	-0.01	0.45	0.24	-0.54	-0.25
credit_score	-0.03	1.00	-0.01	0.02	0.74	0.72	-0.00	-0.32
loan_amount	0.04	-0.01	1.00	0.01	-0.22	-0.16	0.71	0.78
years_employed	-0.01	0.02	0.01	1.00	0.16	0.10	0.02	-0.09
points	0.45	0.74	-0.22	0.16	1.00	0.82	-0.45	-0.62
loan_approved	0.24	0.72	-0.16	0.10	0.82	1.00	-0.25	-0.45
dti_ratio	-0.54	-0.00	0.71	0.02	-0.45	-0.25	1.00	0.80
loan_to_points	-0.25	-0.32	0.78	-0.09	-0.62	-0.45	0.80	1.00



Feature Correlations with Loan Approval

Pearson Correlation by Karl Pearson, 1895:

$$r = \frac{n\sum XY - \sum X \sum Y}{\sqrt{(n\sum X^2 - (\sum X)^2)(n\sum Y^2 - (\sum Y)^2)}}$$

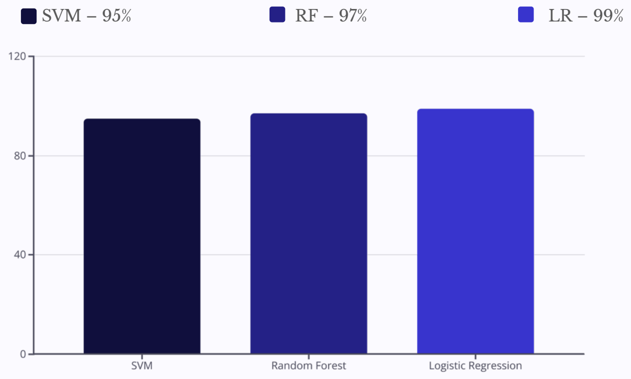
Challenge-1:

Potential Data Leakage: The feature 'points' is a cause for potential data leakage because it is not present at the time of approval but rather assigned after loan process.

Results

Predictive Findings

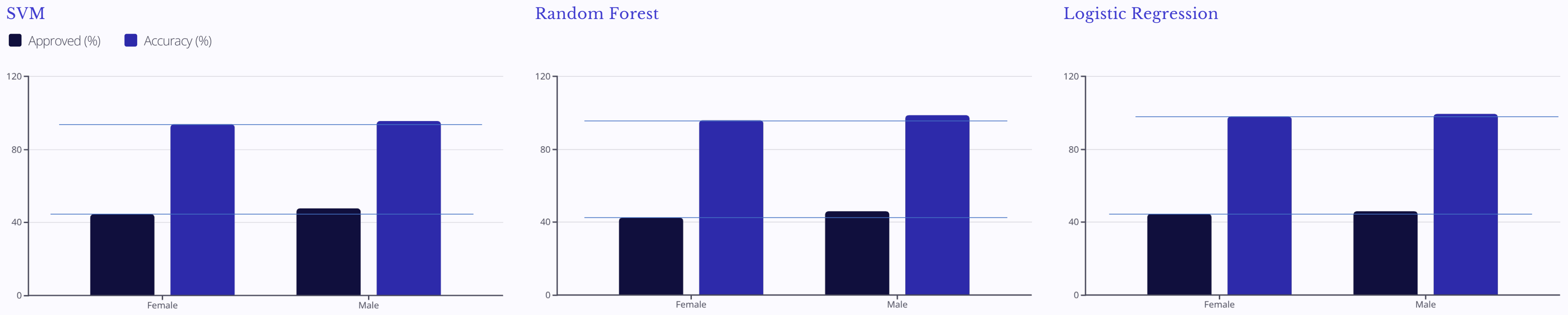
Overall Accuracy Comparison



Fairness Audit

- The "Disparate Mistreatment on Accuracy" is the difference in Accuracy between Males & Females.
- The "Disparate Impact" is the difference in Approved % between Males & Females.
- **Logistic Regression demonstrated the smallest gap in both cases.**

Disparate Impact & Disparate Mistreatment on Accuracy


SVM

Random Forest

Logistic Regression

Research Findings

Questions

1. Does historical loan data already contain gender bias?

A. Yes, historical loan data does contain a minimal gender bias in this dataset.

2. Which classification algorithm (SVM, RF, or LR) has the least unfair errors: “Disparate Impact” & “Disparate Mistreatment”?

A. LR outperforms RF and SVM in terms of fairness and overall accuracy because the relation of features and target was simple. Which is different than Sudhakar et al. (2020) who found RF to outperform LR but harder to interpret.

Conclusion & Recommendations

Based on this complete analysis, the **Logistic Regression** model is the clear winner for this business problem.

Performance

It achieved the highest overall accuracy at 99%.

Fairness

It demonstrated the lowest bias across both fairness metrics, showing the smallest gap in both approval rates (Disparate Impact) and error rates (Disparate Mistreatment).

Recommend Logistic Regression for this specific use case but advise human oversight for the ***1.59%*** error gap.

Further work could involve bias mitigation techniques, such as re-weighting the samples or using a different fairness-aware algorithm.