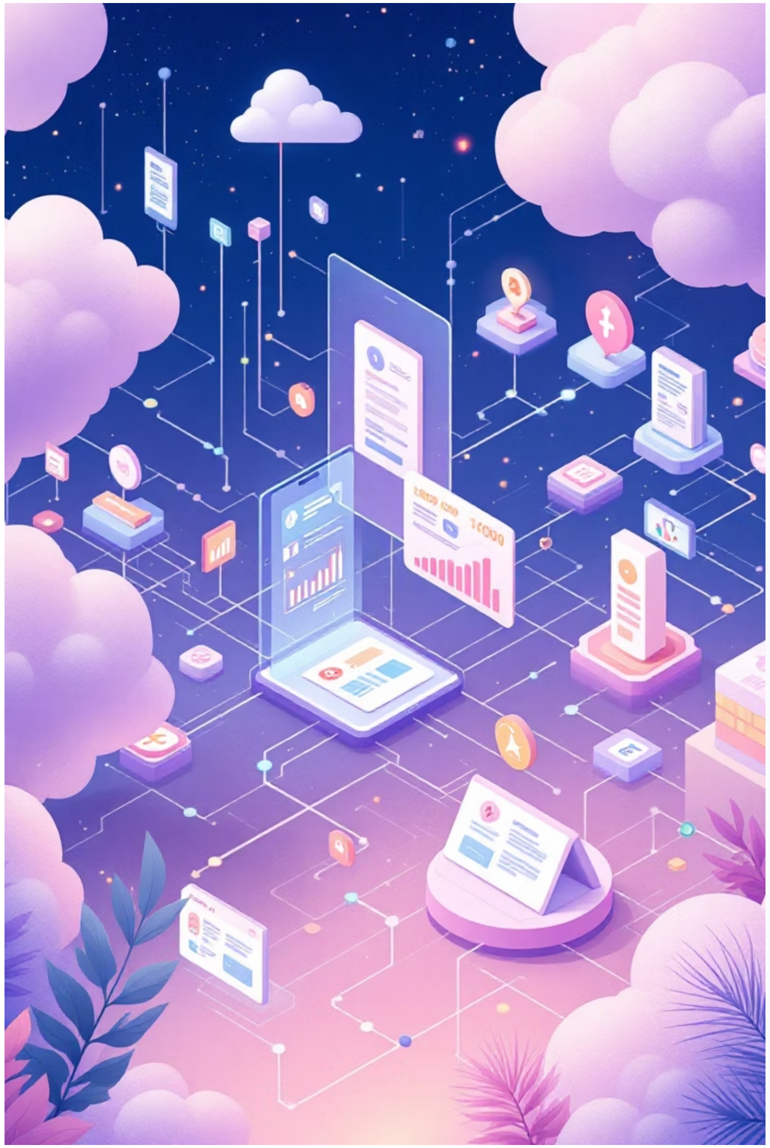


Evaluating Algorithmic Fairness: A Comparative Audit of Machine Learning Models in Loan Approval System

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 like finance.
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

Questions

1. To what extent does historical loan data contain gender bias?
2. Which classification algorithm (SVM, RF, or LR) minimizes Disparate Mistreatment while maintaining predictive accuracy?

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

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).

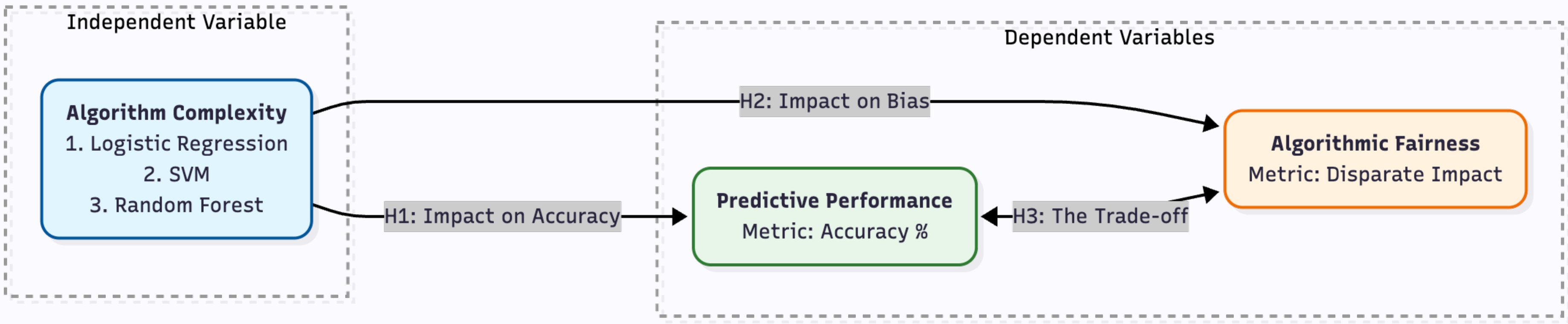
Metric 2: Disparate Mistreatment (Predictive Equality)

- **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

Conceptual Framework

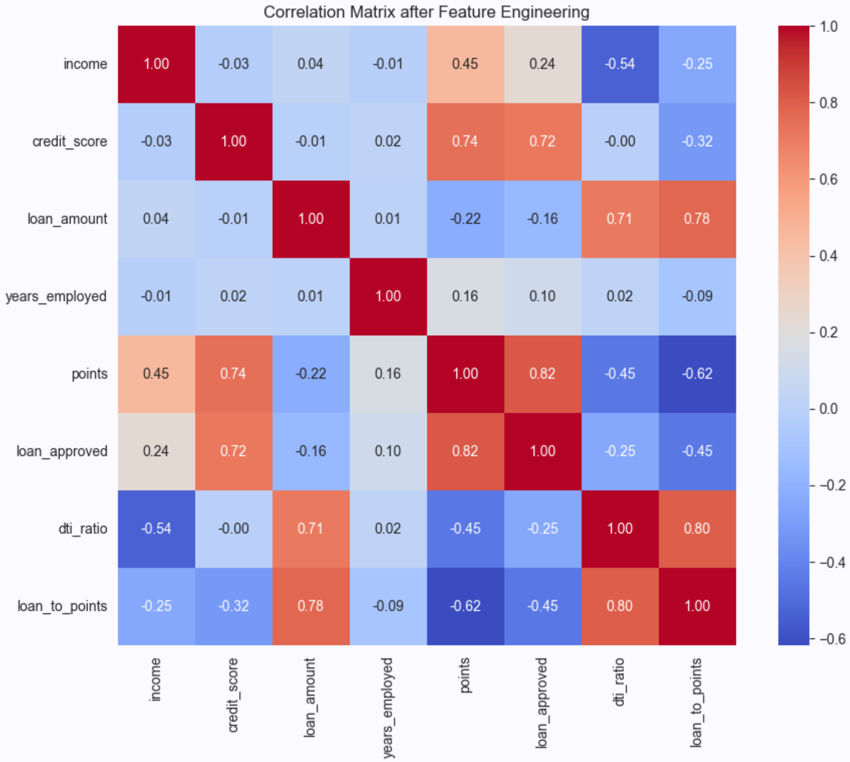


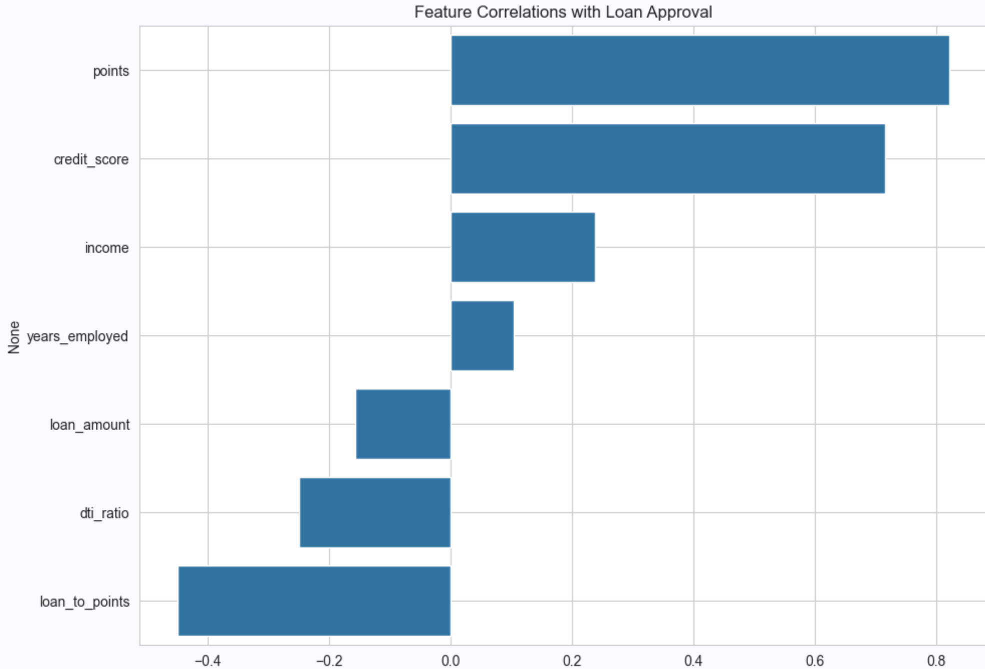
Hypotheses:

- H1 (Impact on Accuracy):
  - Tests how the model type changes predictive power.
- H2 (Impact on Bias):
  - Tests how the model type changes fairness.
- H3 (The Trade-off):
  - Represents the theoretical tension between accuracy and fairness.

Results

Preliminary Test – Correlation Matrix

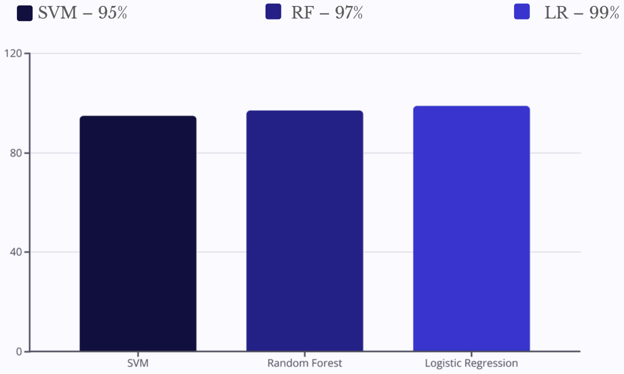




Results

Predictive Findings

Overall Accuracy Comparison



Fairness Audit

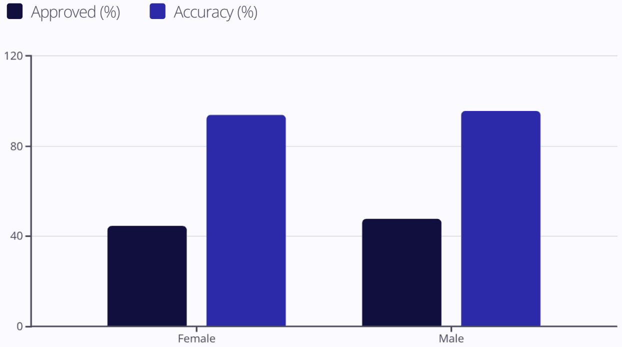
The "Disparate Impact" tables show approval rate differences across gender.
Logistic Regression demonstrated the smallest error gap, indicating superior fairness performance alongside its highest accuracy.

Discussion

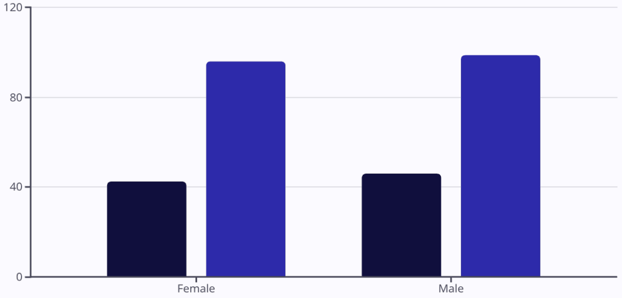
The "Proxy Variable" issue emerged during analysis. The loan_to_points feature drove decisions and the Random Forest Classifier "learned" bias more than the simple Logistic Regression, highlighting how complex models can inadvertently amplify historical biases present in training data.

Disparity Mistreatment on Accuracy

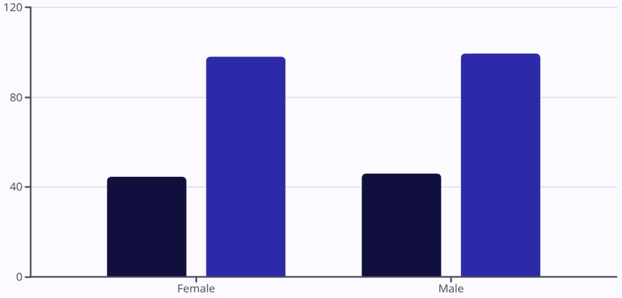
SVM



Random Forest



Logistic Regression



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.