Examining Bias in Loan Approval Processes with Machine Learning: Lending Club Dataset

Jared Watson, Shiv Desai

Emory University

Abstract

Bias in loan approval decisions is an ongoing challenge, often influenced by historical biases in lending practices that has grown over the past decade. The potential for machine learning to inadvertently replicate or exacerbate these biases in financial services is a serious concern for the future of loan approval processes. Our project aims to examine the fairness of machine learning models in the context of loan approval decisions using the Lending Club Loan dataset, which contains over 890,000 loan applications. This study will assess the fairness of these models by analyzing their ability to provide unbiased loan approval decisions across datasets with bias and ones without. Through data preprocessing, we will prepare the dataset by removing irrelevant features, encoding categorical variables, and normalizing numerical ones. We will then train KNN, decision tree, and logistic regression models to predict loan approval, evaluating their fairness by measuring the accuracy and precision scores across their performance. This project will identify which models make the most accurate unbiased loan decisions, contributing valuable insights into mitigating bias in automated financial decision-making

1 - Introduction

Bias in loan approval processes has historically led to discrimination against certain demographic groups. A 2018 UC Berkeley study found that Latinx and African American borrowers paid interest rates that were 6-9 basis points higher than others, leading to additional costs of $250-$500 million per year in extra mortgage interest (Brotcke L. 2022). Even with the transition to automated lending decisions, biases persist, as machine learning systems risk perpetuating historical biases by replicating patterns present in historical data.

Current approaches to mitigating bias in loan approval processes often fall short. While more sophisticated statistical testing procedures were developed in the 1990s, their application is limited due to insufficient data in many banks, leading to the use of less rigorous qualitative techniques. The inherent complexity of machine learning models makes it harder to detect individual features acting as proxies for prohibited biases. Furthermore, machine learning models' capacity to handle massive data volumes increases the risk of biases going unnoticed (Ward-Foxton, S. 2019). This puts further emphasis on a standard evaluation of the fairness of machine learning models to ensure unbiased decision-making in loan approvals.

Unbiased loan approval processes are essential for equitable access to financial services, as bias can have significant socioeconomic impacts on disadvantaged communities. The true potential of machine learning in lending decisions lies in its ability to make unbiased decisions if effectively controlled. Effective efforts to confront this problem at the outset will repay handsomely, allowing the true potential of machine learning to be realized most efficiently (Baer, T., & Kamalnath, V). Therefore, the purpose of this study is to evaluate the fairness of different machine learning models in loan approval decisions, focusing on their ability to avoid perpetuating existing biases. The experiment will train three different models with a dataset of loan approvals that includes demographic related features and one that does not (no existing bias). Based on the accuracy and precision scores of these models, we will evaluate whether a certain model is more capable of classification without bias data, or if the removal of bias features bottlenecks machine learning models' ability to accurately classify loan approvals.

2 - Background

Bias in loan approval processes has disproportionately affected protected groups, such as different races and ethnicities. Various factors contribute to this bias, including data imbalance, inherent bias in training data, and interactions between protected attributes and other variables (Jui, T. D., & Rivas, P. 2024). These issues often reinforce discrimination in lending. A study of over two million mortgage applications showed that lenders were 40% more likely to deny loans to Latino applicants, 50% more likely to deny Asian/Pacific Islander applicants, and 70% more likely to deny Native American applicants compared to similar White applicants. Black applicants faced the highest disparity, with an 80% higher rejection rate than similar White applicants. These disparities highlight the ubiquitous bias in current lending practices, and the need for action. (Martinez, E., & Kirchner, L. 2021).

Traditional statistical approaches have not successfully removed these biases, increasing the demand of advanced techniques like machine learning. However, machine learning models must be carefully designed and calibrated to avoid perpetuating existing biases present in historical data. Addressing these biases is critical to ensure fair access to credit and financial resources for all applicants.

3 - Methods

I. Dataset Analysis

The Lending Club dataset contains data related to all loans issued from the organization during 2007-2015. The dataset contains approximately 890,000 observations and 75 variables regarding each loan's status and the borrower's latest payment details. Some of the key features include:

1. Credit scores: for assessing the risk profile of borrowers
2. Number of Finance Inquiries: counts of recent financial inquiries conducted on the borrower
3. Address information: including zip codes and states to help identify trends
4. Collections: identifies borrowers who have missed one or more payments and are in the process of debt recovery.

The raw dataset available for download contains 145 numerical and categorical features.

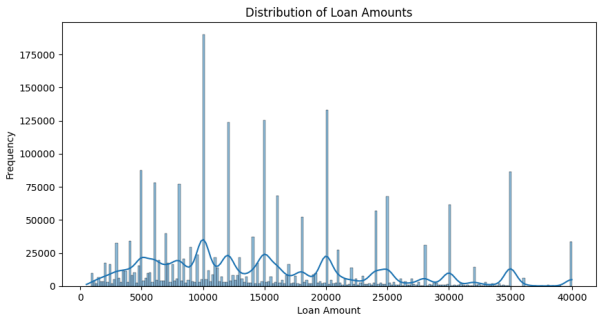

Fig 1: Distribution of Loan Amounts [1]

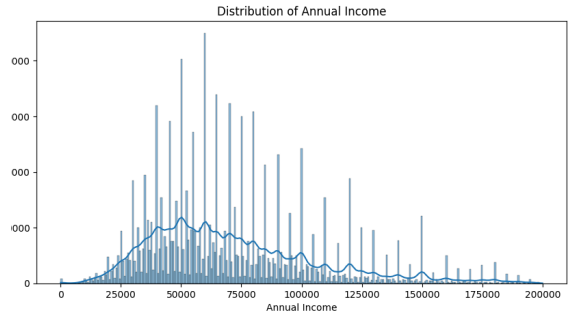

Fig 2: Distribution of Annual Income [2]

II. Data Preprocessing

1. Loading and Sampling data:
    a. Only 50% of the original dataset is sampled randomly to reduce the computational load of processing the data and training the models.
2. Split data
    a. Split data set into training data (80%) and testing data (20%), to ensure data integrity and utilize during the training period for our models
3. Normalizing Numerical Features:
    a. all numerical columns are normalized using `MinMaxScaler` keeping values between 0 and 1

[1] This plot shows the distribution of requested loan amounts, revealing that most loans are clustered around $10,000 to $15,000, with several peaks reflecting common borrowing thresholds across higher amounts.
[2] The second plot illustrates the distribution of borrowers' annual incomes, showing a left skewed distribution where most borrowers earn between $50,000 and $75,000 annually, with diminishing frequencies at higher income ranges.

4. Removing Columns with Missing Values:
    a. Columns which contain more than 50% missing values are removed to reduce noise
5. Encoding Categorical Variables:
    a. The target column (loan_status) is remapped to number 0-5 representing their outcome
        i. "Current": 0, "Fully Paid": 1, "Charged Off": 3, "Late": 4, "In Grade Period": 5
    b. All other categorical variables are converted to numerical representations using factorization

Note: the `late` value of loan_status represents all statuses deemed late. Originally all late values include the number of days late the borrower was (e.g. Late (16-30 days)). All these values are considered `late` for this project.

III. Feature Selection

1. Correlation Analysis
    a. The correlation between numerical features is analyzed through a heatmap (*Fig. 3*) to identify highly correlated features
2. Feature Importance
    a. Using a Random Forest classifier, feature scores are computed and ranked
3. Univariate Statistical Tests:
    a. The `SelectKBest` method is used to identify features using statistical scores
4. Saving Feature Statistics
    a. Feature importance and selection scores are saved into a JSON file (`data/feature_stats.json`)
5. Removing Bias:
    a. The models are trained twice, one with a dataset that contains sensitive demographic information and one that does not. After analyzing the features, we have collected a list of 11 demographic related features to remove to train models with no bias.
        1. **zip_code**
        2. **addr_state**
        3. **emp_title**
        4. **home_ownership**
        5. **emp_length**
        6. **annual_inc**
        7. **verification_status**
        8. **mths_since_recent_inq**
        9. **inq_last_6mths**
        10. **inq_last_12m**
        11. **loan_amnt**

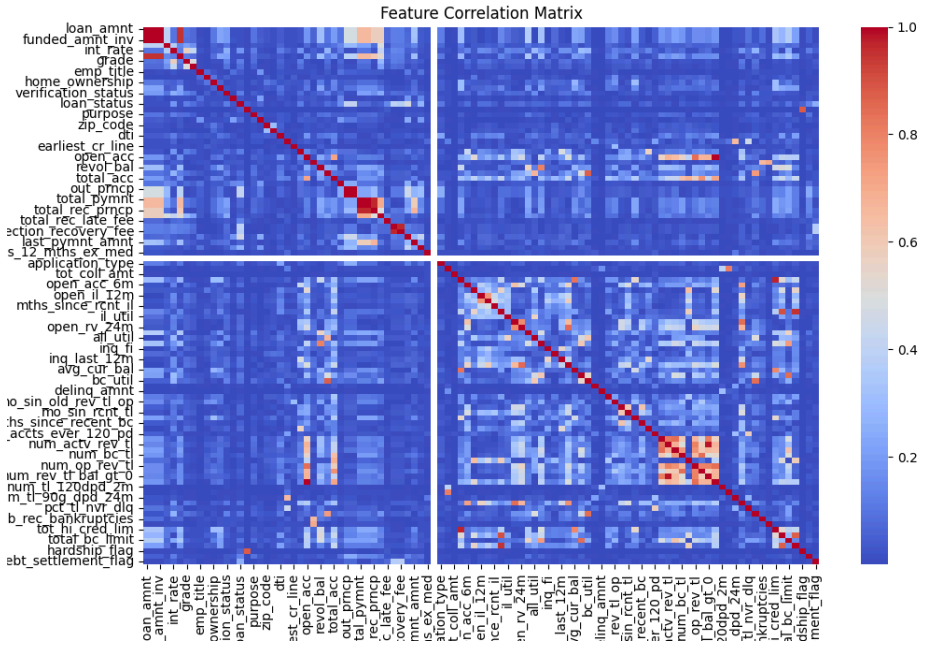

Fig 3: Heatmap representing correlation between each feature

IV. Models and Parameter Tuning

Since the target column (`loan_status`) for this dataset is categorical, our model selection focussed on classifiers suitable for this problem. After initial model selection, the parameters were optimized using k-fold cross validation to evaluate different configurations. Each model parameters were selected to minimize the misclassification error to maximize prediction accuracy. The models were trained twice, one with demographic related features, and one without to determine if the models were susceptible to these sensitive data types.

1. K-Nearest Neighbors (KNN):
    a. This classifier predicts the label based on the l-nearest neighbors to the data point. The predicted label is determined using a majority vote across the nearest neighbors' labels
    b. Tested k values of `[3, 5, 7, 9]` with both manhattan and euclidean distances
    c. Best parameters: `"n_neighbors": 7, "metric": "manhattan"`
2. Logistic Regression:
    a. Logistic regression models the probability that an instance belongs to a particular class.
    b. Tested C values of `[0.01, 0.1, 1, 10, 100]` each with a penalty of `"l2"`
    c. Best Parameters: `"C": 100, "penalty": "l2"`
3. Decision Tree:
    a. Decision trees build a hierarchical structure that splits the data into nodes by selecting the optimal split point based on a certain criteria

b. Tested criterion of `["gini", "entropy"]`, max_depth of `[5, 10, 20, 40, 80]`, min_samples_leaf of `[5, 10, 20, 40, 80]`, and min_samples_split of `[2, 4, 8, 16, 32]`
c. Best Parameters: `"criterion": "gini", "max_depth": 20, "min_samples_leaf": 10, "min_samples_split": 2`

4 - Experiment Results

Model	**Model Performance without Bias Features**					
	Accuracy	Precision (Weighted)	Recall (Sensitivity, Weighted)	Specificity	F1 Score (Weighted)	ROC AUC Score
Decision Tree	0.98	0.98	0.98	0.99	0.98	1.0
KNN	0.89	0.86	0.87	0.95	0.86	1.0
Logistic Regression	0.96	0.97	0.95	0.99	0.96	1.0

Table 1

Model	**Model Performance with Bias Features**					
	Accuracy	Precision (Weighted)	Recall (Sensitivity, Weighted)	Specificity	F1 Score (Weighted)	ROC AUC Score
Decision Tree	0.98	0.98	0.98	0.99	0.98	1.0
KNN	0.86	0.86	0.86	0.95	0.85	1.0
Logistic Regression	0.97	0.97	0.97	0.99	0.97	1.0

Table 2

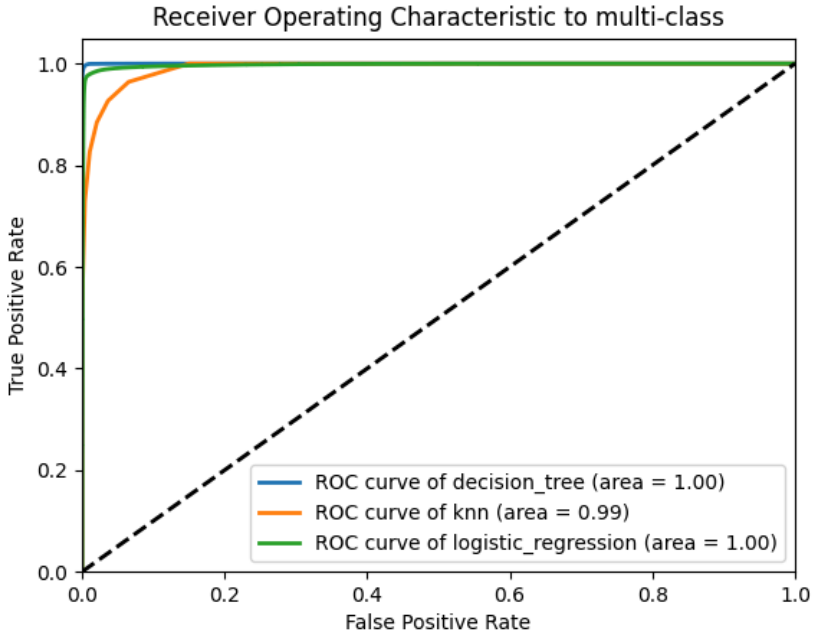

Fig 4: ROC Curve for model w/ bias

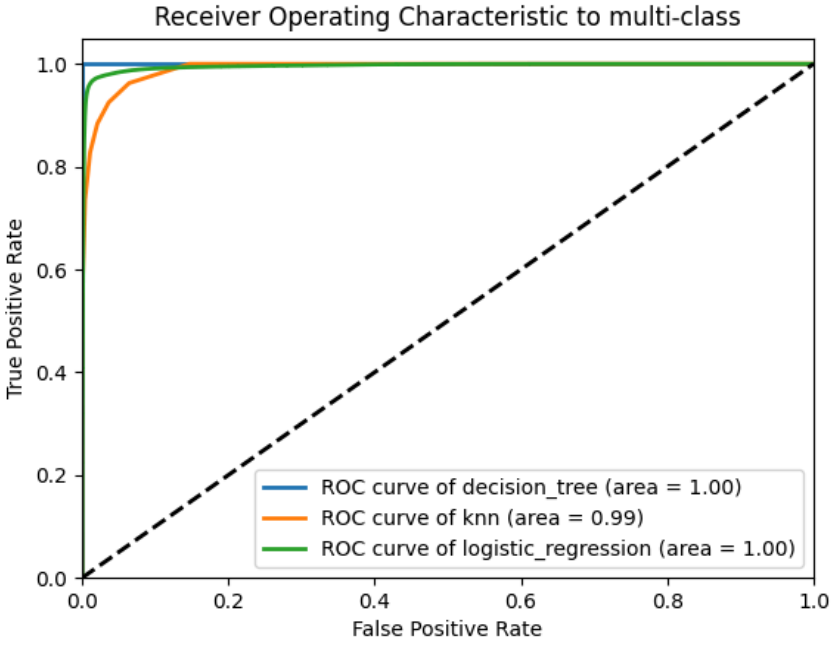

Fig 5: ROC Curve for model w/o bias

1. **Decision Tree**: The decision tree model achieved an accuracy of 98% which is similar to the logistic regression model. Its precision, recall and F1 scores are also high at 98% showing that the model was able to handle performance across all classes for classification. The specificity is 99%

which means it is effectively distinguishing between the various classes minimizing false positives.

2. **kNN:** The kNN model is decently robust however it does not perform as well as the other models. The accuracy, precision, recall, and F1 scores fall around 85-86% which is overall lower performance than the decision tree and logistic regression models.
3. **Logistic Regression:** This model also is highly accurate at 97%, with precision, recall, and F1 scores all at 97%. Its specificity of 99% makes it similarly performing to the decision tree model minimizing false positives.

Both the decision tree and logistic regression models deliver similar performance while the kNN model is less accurate. The decision tree and logistic regression models have high precision, recall, and F1 scores while maintaining a specificity of 99%. Because both models have nearly identical performance either could be used, but the decision tree runs in less time making it the more efficient model.

Discussion

Our empirical results reveal that through appropriate pre-processing that encompasses scaling, one-hot encoding, and feature selection, the three selected models achieved remarkable accuracy percentages and a respectable AUROC score. Out of the three models, the selective best for the Lending Club data would be a tie between Decision Tree and Logistic Regression as both models, with bias and without, maintained accuracy and F1-Scores around 97%.

Our original research question aimed to reveal if any particular model would perform better when bias was removed from the training and test data set. The results reveal that after removing demographic related features, the trained models showed nearly minimum difference in training accuracy, precision, and AUCROC score. The models trained with bias performed remarkably well in regards to accuracy and predicting the loan status; however, when training the results without bias, the scores remained the same. In regards to the research question of this study, our results show that demographic features are irrelevant in terms of model evaluations, thus indicating existing loan data sets do not exhibit perpetual bias over years of data collection. This is a positive insight in terms of the future application of machine learning models for loan and mortgage approvals as this large data set does not contain any inherent bias from data collected over several years. The removal of bias in loan approvals statistically benefits the larger social-economic state of the United States. The economic wealth gap that persists and continues to expand in many countries will continue to grow unless low-income citizens are given an equal opportunity financially. If applicants are able to receive the loans they are qualified for, without taking in account ethnicity and background, we will see more wealth accumulate in the pockets of minority and immigrant groups, and more American dreams being made.

Contributions

Jared Watson:

Created scripts for preprocessing dataset, selecting/tuning parameters for each model, and created models for training on data without demographic features. Additionally, created a script to analyze performance of models in order to obtain which model performs the best.

Shiv Desai:

Improved preprocessing for dataset, researched relevant demographic related features and built a script to remove those from training and test data, trained all models for data set without bias, conducted relevant market research within this field of study, and structured github documentation.

Code

The code for this project can be found *here*. The repository contains all code files relevant to our project. The raw dataset is located in a dropbox folder *here*.

Citations

Baer, T., & Kamalnath, V. (n.d.). Controlling machine-learning algorithms and their biases. McKinsey & Company. https://www.mckinsey.com/capabilities/risk-and-resilience/our-insights/controlling-machine-learning-algorithms-and-their-biases

Brotcke, L. (2022, February 5). Time to Assess Bias in Machine Learning Models for Credit Decisions. MDPI. https://www.mdpi.com/1911-8074/15/4/165

Ward-Foxton, S. (2019, 4 30). Reducing Bias in AI Models for Credit and Loan Decisions. EE|Times. Reducing Bias in AI Models for Credit and Loan Decisions

Jui, T. D., & Rivas, P. (2024). Fairness issues, current approaches, and challenges in machine learning models. International Journal of Machine Learning and Cybernetics. https://doi.org/10.1007/s13042-023-02083-2

Martinez, E., & Kirchner, L. (2021, April 25). The Secret Bias Hidden in Mortgage-Approval Algorithms. The Markup. https://themarkup.org/denied/2021/08/25/the-secret-bias-hidden-in-mortgage-approval-algorithms