Machine Learning and Counterfactual Fairness Analysis to Detect and Counter Bias in Credit Analysis for Loan Granting Systems

University of Turku
Department of Computing
Master of Science (Tech) Thesis
Department of Computing
June 2025
Mishaal Naeem

UNIVERSITY OF TURKU
Department of Computing

Mishaal Naeem: Machine Learning and Counterfactual Fairness Analysis to Detect and Counter Bias in Credit Analysis for Loan Granting Systems

Master of Science (Tech) Thesis, 60 p.
Department of Computing
June 2025

Creditworthiness analysis and the loan granting process are core services of the finance sector. With the shifts towards Machine Learning and Artificial Intelligence based solutions, the finance sector has also incorporated them into their creditworthiness analysis and loan granting processes over the years. As the finance sector is a heavily regulated industry, machine learning solutions require transparency, interpretability, and fairness. In this research, the questions of how to detect bias in the dataset before training, how to train a fair model, and how to assess the fairness of a trained model are explored. The purpose of this thesis is to derive a modular approach that addresses inherited bias from the dataset in the model training itself. This research uses existing literature for defining bias, to determine the machine learning models that work best for creditworthiness analysis, and methodologies for determining bias in trained models, along with determining how bias can be tackled at the model training level. The core methodology that this thesis suggests and focuses on is a Hybrid SVM and Random Forest, which utilizes custom decision trees based on the leave-group-out concept. The dataset utilized in the thesis is the German Credit Dataset from the UCI repository. It has several features, including sensitive features like gender and foreign worker status. The thesis determines the importance of all features for quantifying the effect of sensitive features, analyzes the dataset, trains models including the proposed model for comparison of bias in the model outcome and accuracy. Evaluation metrics show the proposed model performs better than or as well as the other models. Counterfactual analysis and statistical tests for fairness show that the proposed solution significantly reduces bias and is much more fair in its predictions. The thesis lastly sheds light on the practical use of the proposed model.

Keywords: Machine Learning, Bias, Random Forest, Credit Analysis, Loan Granting, Finance, Fairness, Counterfactual Analysis, Creditworthiness, SVM, Feature Selection

Contents

List of Figures

List of Tables

1 Introduction

The loan granting process is a very specific and complex process, and it directly affects the stability of the financial markets. Creditworthiness analysis over the years has contributed to economic mistakes and crises. Oftentimes, loans are granted or rejected at face value, which can be termed as an implicit bias in the creditworthiness analysis, but because of the complexity of the process, the bias can be covered up. That is just one aspect. The complexity of the process entails risks due to simpler things, such as human error. Thus, an automated and data-driven solution to creditworthiness analysis can reduce bias and risks, leading to more secure and risk-free loan granting.

The purpose of this research is to identify using a credit dataset the fairness of loan granting and to develop a machine learning solution that can reduce bias. The idea behind the research is to compare how automation of the process would not only improve the loan granting process but also how decisions would be more justifiable.

The research in this thesis is a combination of literature review and experiments. The literature review in Section 3 presents the existence of bias in finance services and in existing machine learning solutions. Studies are used to derive an approach to answer the problem statement (defined in Section 2) by deriving a model based on existing literature for credit analysis in machine learning, defining and tackling bias in machine learning, and methodologies for finding existence of bias in trained

machine learning models. The datasets in the literature are also reviewed to pick datasets relevant to this study. The finalized approach is laid out in Section 4.

The chosen dataset is explored and analyzed in Sections 5 and 6. The trained models are evaluated on performance and fairness metrics, and counterfactual analysis is performed on them in Section 7.

The final discussions are carried out under Section 8, with conclusion and prospect in Section 9.

2 Problem Statement

Credit Analysis and Lending are at the heart of finance. These services have been provided for a long period of time. Bias in credit analysis has also existed since due to human involvement and causal impact of culture. Historically, racism, sexism, and ageism have affected processes across all domains, including financial services. Datasets based on previous decisions made in credit analysis and loan processes are affected by this bias. This bias continues into the machine learning models trained on these datasets.

To address bias in creditworthiness analysis and develop a machine learning model that can make explainable predictions, there are several questions that need to be answered.

- How can the existence of bias in the dataset be observed?
- How can a machine learning model be trained to be unbiased?
- What fairness metrics are suitable for assessing bias?

To answer these questions in this research, fairness and types of bias in machine learning need to be mapped onto the existing framework of loan decisions, to statistically prove the bias exists. The bias needs to be removed from the model, without losing the crux of the creditworthiness calculation and loan granting decision making factors. The trained models need to perform well, be interpretable, and transparent. The performance of the model and the fairness of the model should not be a

trade-off. The trained model should perform well and also be fair.

3 Literature Review

One of the key services that contribute as a driver of the economy is lending. Credit access is pivotal to economic growth. It allows people to invest, innovate, and expand their personal lives or businesses. Loans allow for growth in the money supply. Lending itself is also a huge business. They contribute to the circulation of money supply in the economy. According to CEIC's Economic Database, the European Union Household Debt reached 6,956.2 billion USD in January 2025. [1] In just the US alone, credit card debt was $1.211 trillion in 2024 [2]. Household loans exert the largest contribution to economic activity [3].

Financial institutions provide this service under heavy regulations with the aim of limiting risks, as they have a strong contribution to the economy. There are several regulatory bodies that help manage risk through oversight and regulations. This regulation is first and foremost rooted in the credit analysis and loan approval processes performed by financial institutions. Credit management systems include restrictions on credit limits for new entries, limits based on longevity of employment, adjustments based on repayment behavior, and credit history among other things. The systems are dependent on an individual's gathered data. One new increasing addition to the systems in the finance sector has been the integration of machine learning solutions. The finance industry has reported a high rate of integration rates. In 2023, investment in Artificial Intelligence and Machine Learning by the finance industry reached $35 billion, with projections of $126.4 billion by 2028 [4].

Bank of England reported that 72% of financial institutes are using or developing machine learning integrations [5].

Traditional approaches to credit analysis and lending processes are based on limited data, and they are not adequate for credit analysis of many "thin-file" consumers [6]. Additionally, they lack personalization and require human involvement.

A 2013 study by the Federal Trade Commission reported in their survey, that 26% of consumers reported errors in credit reports and for 13% it resulted in denials or higher interest rates [6]. Additionally, for credit analysis and lending processes, the decision needs to be interpretable. To overcome these, the adoption of machine learning in financial sectors is consistently increasing. Financial institutes are relying on it to reach more data-driven conclusions, which are explainable, in a shorter amount of time.

Studies emphasize the importance of fairness of decisions in credit analysis and lending processes, as they create an unfair risk and systematic risk. Fairness is important for capital flow and to reduce the chances of the creation of bubbles or concentrated risks. Transparency and fair analysis processes in the financial sector are also important for trust-building in consumers [7]. From a legal point of view as well, fair credit practices align with anti-discrimination laws. That is why explainable processes are of importance to the finance sector.

Literature review is important for this research, to establish the existence of bias in the credit analysis and lending processes, the need of machine learning in financial sector, the pitfalls of traditional processes, the machine learning based algorithmic approach best suited for financial sector, the techniques for ensuring mitigating of bias, and the existing research gap. The following sections aim at providing research and literature-based evidence to support the importance of this research and verify the approach that this research takes.

3.1 Fairness in Credit Analysis and Lending Processes

Traditionally, financial applications are handled through human involvement. The process consists of collecting information regarding an applicant's credit history through data such as current debt, employment, assets, payment history, and savings. While most of the process has a structure to it, due to human involvement as discussed earlier there can be errors. Moreover, the literature supports that error is not the only concern in credit analysis and lending processes. Human involvement has led to bias in credit analysis and lending processes. This creates a risk in the processes due to the lack of fairness. As stated earlier, fairness is important in credit analysis and lending processes due to multi-fold reasons that affect the economy and stability of financial institutes.

Strong evidence exists regarding the causal impact of culture on economic outcomes [8]. The beliefs and preferences that exist in culture affect several economic agents. The culture factor casually impacts financial development [9] and regulation [7] among other phenomena.

Research shows that there is a large disparity in credit analysis and lending processes across different demographic groups. Talavera [10] found that women are more likely to be denied bank credit. In European countries, interest rates and credit usage across different genders and ethnicities exist that are not explainable by credit scores [11], [12]. In the United States, African Americans pay higher interest rates and face more mortgage loan rejections even with no differences in earning and credit scores [13], [14]. In Japan, sexism has been part of credit worthiness analysis, with a woman applicant being scored lower than a man while other application details were kept the same except gender [15].

Additionally Belluci, Borisov, and Zazzaro's [16] research found that in Italy,

women face tighter credit constraints. Alesina, Lotti, and Mistrulli [11] also found that in Italy women borrowers pay a higher interest rate.

This bias also raises regulatory and legal concerns. Discriminatory lending practices are a core issue in financial ethics and regulation. Fair lending laws by the Federal Reserve Board and EU Non-Discrimination Directives clearly prohibit unfair treatment of borrowers based on gender, race, and location. European Union GDRP, and the Fair Credit Reporting Act (FCRA) in the United States require asymmetry of information between lenders and borrowers i.e. transparency in credit scoring. Community Reinvestment Act (CRA) in the United States encourages banks to meet the credit needs of all communities, especially low and moderate income areas. European Union Charter of Fundamental Rights prohibits discrimination in all sectors, including financial services. In the European Union, a suite of anti-discrimination directives (e.g., Directive 2000/43/EC on Racial Equality, Directive 2004/113/EC on gender equality in access to goods and services) applies to financial institutions. Consumer Credit Directive 2 (CCD2) strengthens consumer protections, including fairness in credit scoring, and all European Union members have to transpose it into their laws by 20th November 2025.

Inherited cultural, ethnic, and gender bias' existence in credit analysis and lending processes is strongly supported by literature. The inherited bias exists due to human involvement and thus exists the need to delegate the process to an algorithmic approach. Machine Learning, with its advances and its growth in use and trust in the financial sector, paves the path for this.

3.2 Machine Learning for Credit Analysis and Loan Processes

Algorithmic Credit Analysis and Loan Processes have existed for quite some time. These rule-based models; which make use of IF-THEN scenarios, while automate the analysis and the process and have the potential to eliminate human error, fall short on account of their limitation to the predefined scenarios and are still prone to human bias given that it is translation of the same inherited bias that plagues the traditional process. In Japan, J. Score; an algorithmic tool, made different decisions regarding credit on applications with all input except gender being the same [15]. Additionally, these algorithms due to their fixed set of rules are only able to take into account specific predefined data points and fall short on thin file customers.

Statistical models used to determine credit scores and automate loan processes have been used for a long time now as well. They assign numerical values based on a specific set of categories pertaining to credit history. This is an approach with which bias is eradicated in mathematical models. However, they introduce the problem of information asymmetry as the numerical representation of creditworthiness does not provide the explainability [17]. Therefore, these credit scoring models do not meet all the requirements of law and financial sector needs.

With data availability and machine learning leaps, credit analysis and loan processes can achieve the four crucial things required to be addressed in this research; lower decision times, lower error rates, explainability of decision, and elimination of bias. Machine Learning solutions in this scenario can reduce the workload of loan officers, increase the reliability of credit analysis and loan eligibility, and mitigate risks.

Due to reliance on data-driven learning, to deliver optimal performance across different datasets and scenarios, comprehensive analysis of suitable machine learn-

ing techniques is important. Numerous studies have researched this area and have taken into account the effectiveness of different machine learning algorithms and modifications to improve credit analysis.

The major machine learning algorithms that have made advancements and provide a promising solution include Support Vector Machine (SVM), Random Forest (RF), and Multilayer Perceptron (MLP) [18], [19], [17], [20], [21], [22],

Baesens et al. [19] studied accuracy (ACC) and Area under the Curve (AUC) across 8 different credit scoring datasets and found SVM and MLP, LR, and LDA the best performing machine learning approaches. Lessman et al. [19] found that Random Forest should be the benchmark classifier and found that Random Forest and Multilayer Perceptron to be the best performing and versatile classifiers. Both studies focus on basic models.

Li and Chen [21] found ensemble models to be better at accuracy than baseline models. They used five different metrics; ACC, AUC, Kolmogorov-Smirnob statistics, Brier score, and operating time on Random Forest, LightGBM, Ada, and XG Boost. Their study showed Random Forest to be the best performing on the five metrics.

Trivedi [22] studied five different machine learning approaches; Random Forest, Decision Trees, Bayesian, Naive Bayesian, and SVM, on the German Credit Dataset and found Random Forest to be the best-performing model across five different evaluation metrics as well.

Ampountolas [18] found that tree-based and ensemble models in microfinance perform better than others and that Random Forests have the best accuracy compared to MLP, Boosting algorithms, Decision Tree, KNN, and Extra Tree.

Ong and Lee [17] did a comparison of five different machine learning models; Logistic Regression, SVM, Random Forest, MLP, and LDA using five different metrics; ACC, F1, Precision, Recall, and AUC on four different datasets. They found RF

and MLP to be best performing on larger datasets, and LR to be best performing on smaller datasets.

Studies across different datasets and metrics support the use of Random Forest for credit analysis and loan granting processes.

3.2.1 Research Gap

While some studies have carried out modifications in different machine learning techniques to tackle credit analysis and loan processes, and some have carried out algorithmic comparisons to determine which algorithm is best suited for these finance sector use cases, and other studies exist in dealing with bias in machine learning algorithms stemming from data-driven learning, there is a gap in research for recognizing bias in datasets for credit analysis and loan process and eradicating the bias in the training of the machine learning algorithms for this finance sector use case.

By addressing this research gap, the study aims to contribute to the advancement and incorporation of machine learning solutions in the finance sector while addressing the inherited bias that exists in the domain. Such research would provide insight into how bias can be detected in a dataset and a model, and how it can eradicate to ensure fairness and lowered risk in credit analysis and loan processes.

3.3 Bias in Machine Learning

There are several examples of bias in machine learning solutions based on protected attributes such as sex, race, age, marital status. A facial recognition software was found to be biased against dark skinned women. Amazon had to scrape their recruiting tool as it had bias against women.

A lot of research work has been done to define bias and recognize how it gets

into machine learning, how to identify and quantify it, and how to eradicate it [23]. As machine learning is data-driven learning, the bias originates from the quality of the data. If the data are biased, then the machine learning solution trained on the data is biased. As the data are human-generated, the bias in the data comes from human decisions. As discussed earlier this can be inherited bias, or bias in data from errors.

The term bias in machine learning not only refers to unfair decisions, but also the distribution of the data. In Machine Learning theory, the bias is distinguished into three types [24]:

1. Covariate Shift: Bias that occurs in the trained machine learning algorithm if one of the features is not uniformly covered in the data.

2. Sample Selection Bias: Bias that occurs based on the correlation between a subset of features and the label. This is the type of bias that is being addressed in this research. In the datasets, if the applicants of a protected group were systematically discriminated against, then this feature would correlate with the label that we want to predict.

3. Imbalance Bias: Bias that occurs if one label has more examples than the other.

As the bias this research is addressing falls in the sample selection bias category, we need to define what we mean by protected attribute. A protected or sensitive attribute is an attribute that divides the whole population into two groups (privileged and unprivileged) that have differences in terms of decision. For example, in the case of credit analysis and loan processes datasets, based on protected attribute sex, male is privileged and female is unprivileged.

3.4 Finding Bias: Counterfactual Analysis, Statistical Tests

Machine Learning solutions used in decision-making solutions for the financial sector impact human lives and have legal and ethical consequences. Unfairly biased algorithms against protected groups have existed in the past. Kuser et al. [25] developed a framework to model fairness in machine learning algorithms.

Counterfactual Analysis is a method used to explore what-if scenarios. It asks the question, "Would a model's prediction remain the same if a sensitive attribute had been different, all else being equal?"

Kusner et al., [25] formally defines it as: predictor $\hat{Y}$ is counterfactually fair if under any context $X = x$ and $A = a$,

$$P(\hat{Y}_{A\leftarrow a}(U) = y|X = x, A = a) = P(\hat{Y}_{A\leftarrow a_0}(U) = y|X = x, A = a)$$

for all y and for any value a_0 attainable by A.

Kasirzadeh and Smart's [26] study in use cases of counterfactual fairness shows that counterfactual analysis can be applied in machine learning to become aware of the impacts of the dataset and helps in causal explanations of the sensitive attributes in a dataset.

Verma et al. [27] further explore different algorithms and modifications of counterfactual analysis and determine that counterfactual fairness analysis helps attain explainability in machine learning algorithms by providing causal explanations of sensitive features and allows for understanding the social implications of the dataset.

A lot of research exists to help quantify the existence of bias in a dataset being utilized for training a machine learning algorithm. Fairness measures can be mathematically derived from a dataset.

Hardt et al. [28] propose a simple, interpretable, and actionable framework for

measuring and removing discrimination based on protected attributes for binary predictors called Equal Opportunity. It measures unfairness rather than proving fairness.

Predictive parity means that a predictive model should have equal positive predictive value across different sensitive groups. In the framework proposed by Chouldechova [29], it focuses on calibration within decision outcome i.e. for everyone who got a specific prediction, the actual outcome should happen at the same rate across groups. Similar to their work, Corbett-Davies et al. [30] propose *Predictive Equality*, where the proposed framework requires that the false positive rate (FPR) to be the same across groups. Additionally, Berk et. al [31] introduces Treatment Equality. *Treatment equality* holds when the ratio of false positives to false negatives is equal across protected groups. Together, these three statistical tests answer three questions:

- Are decisions equally reliable? (PP)
- Are false positives balanced? (PE)
- Are the types of errors balanced? (TE)

Another framework, *Absolute Between ROC Area* is a fairness metric that quantifies the difference in model performance between subgroups over the entire threshold space. It measures the average absolute difference between ROC curves for two (or more) demographic groups. This framework is presented in a study by Gardener et al. [32] that compared how ROC curves differ across thresholds, not just their total AUC. It also captures the performance disparity.

In this research, these statistical tests alongside counterfactual analysis can be utilized to analyze datasets and models for recognition of unfairness and quantifying biases.

Fairness Measure	Citation	Treatment equality (TE)
Treatment equality (TE)	Berk et al. [31]	$(-\infty, \infty)$
Absolute Between-ROC3 Area (ABROCA)	Gardner et al. [32]	$[0, 1]$
Equal opportunity	Hardt et al. [28]	$[0, 1]$
Predictive parity (PP)	Chouldechova [29]	$[0, 1]$
Predictive equality (PE)	Corbett-Davies et al. [30]	$[0, 1]$

Table 3.1: Statistical tests for measuring fairness

3.5 Mitigating Unfairness and Bias

Across the available literature, mitigating bias has been a topic that has been largely addressed. The approaches to mitigating unfairness and bias can be categorized into three groups [33]:

1. Individual Fairness which requires the model to give similar predictions to similar individuals

2. Group Fairness which aims to treat the groups with different protected sensitive attributes equally

3. Max-Min Fairness which tries to maximize the minimum expected utility across groups

This research mainly focuses on group fairness as it tries to tackle the presence of inherited bias. Group fairness can further be achieved through pre-processing, post-processing, and in-processing i.e. model training. Zhao et al. [33] encourages isolating the influence of correlated features. Adebayo and Kagal [34] also support this approach. In their study, they focus on how model predictions change with the exclusion of individual features. Their framework introduces a feature-wise analysis method to diagnose bias. Hooker [35] endorses experiments that involve removing sensitive attributes to assess influence. Lipton et al. [36] in their study experimented with controlled interventions and feature suppression.

A popular framework for mitigation of bias is SHAP proposed by Lundberg et. al [37]. SHAP implements marginal feature removal to assess impact. SHAP is model-agnostic and this research focuses on preprocessing methodologies to cater to bias.

Kamiran and Calders [38] introduce various preprocessing methods to reduce discrimination, including the removal or relabeling of sensitive features. Bertsimas et al. [39] in their study investigate fairness through sensitive feature impact analysis, explicitly supporting feature-wise comparisons like *Leave One Feature Out (LOFO)* for identifying disparate influences.

Hardt et al.'s [40] and [28] finance targeted study uses principles of altered inputs and subgroup slices for model fairness.

These studies serve as the groundwork for formulating an approach to mitigate bias through controlled interventions. While feature isolation is important, it is also important to not completely take away the feature correlation from the original dataset. This is where Ensemble Learning, which is widely accepted as a benchmark for credit analysis and loan processing, combined with a slicing approach comes into this research.

3.6 Datasets

There are several datasets, that have been in the past research and studies used for training machine learning algorithms for credit analysis and loan processes. These datasets have also been used as benchmarks.

1. Australian Credit Approval (AC) [41]
2. German Credit Data (GC) [42]
3. Japanese Credit Screening (JC) [43]

Dataset	Samples	Features	Class Distribution
AC	690	14	383 non-defaulters, 307 defaulters
GC	1000	20	700 non-defaulters, 300 defaulters
JC	690	15	307 non-defaulters, 383 defaulters
LC	1,770,000	72	1,420,000 non-defaulters, 350,000 defaulters
GMSC	150,000	11	140,000 non-defaulters, 10,000 defaulters

Table 3.2: Overview of Datasets

4. Lending Club Loan Data 2007 to 2020Q3 (LC) [44]

5. Give Me Some Credit Competition Data (GMSC) [45]

German Credit Data includes 1,000 instances with 20 features covering personal and financial information such as checking account status, credit amount, employment, and housing. It has feature attributes of sex, personal status, and foreign work which is crucial to this research. The dataset has over 63 citations. As feature documentation is available and not anonymized, it has easy interpretability for the use case of this research.

The dataset has been explored both in terms of using as a benchmark to evaluate different machine learning algorithms, notably by Ong and Lee [17] and has been used in studies for recognition of bias and unfairness in datasets.

Other popular datasets are, Australian Credit Approval which has a small size and limited feature documentation. Japanese Credit Screening which has similar drawbacks. Lending Club Loan Data which has an under-represented rejection class. Give Me Some Credit Competition Data which has a similar issue.

There are other datasets that have sensitive features such as gender and race, but they do not fit the use of the research.

4 Approach

4.1 Counterfactual Fairness Analysis

Counterfactual Fairness Analysis utilizes counterfactual samples to determine if the model makes a different decision based on sensitive attributes. If one of the sensitive attributes is counterfactual while the rest of the attributes remain the same, and the model makes a different decision, the model is biased. A model is considered counterfactually fair if its prediction remains the same when we change a sensitive attribute while keeping all other characteristics constant.

For a given individual X with attributes $X = (X_s, X_o)$ where:

- X_s = Sensitive attributes (e.g., gender, personal status, race)
- X_o = Other non-sensitive attributes (e.g., income, credit score, employment history)

A model $f(X)$ is counterfactually fair if:

$$f(X_s, X_o) = f(X'_s, X_o)$$

for all possible counterfactual versions X'_s (e.g., changing "Male" to "Female" while keeping income, credit score, etc., unchanged).

The research uses counterfactual fairness analysis to determine which of the factors in the dataset create a bias in the decision. The steps to achieve this are:

- Train model on the original datasets.
- Identify sensitive attributes in the datasets.
- Generate counterfactual samples of the datasets. For a given counterfactual dataset, only one of the sensitive attributes is changed at a time.
- Predict on the counterfactual samples using the model trained on original datasets, and calculate the bias.

4.2 Statistical Tests

On the datasets, several of the most popular group fairness measures can be used to determine how fair the model's results are. The fairness measures are chosen from different citations.

- Treatment Equality (TE):

  Treatment equality is computed based on False Negative (FN) and False Positive (FP) of the protected group (prot.) and non-protected (non-prot.) groups.

- Absolute Between-ROC3 Area (ABROCA):

  It measures the divergence between the protected (ROCs) and non-protected group (ROCs) curves across all possible thresholds $t \in [0, 1]$ of false positive rates (FPR) and true positive rates (TPR). The absolute difference between the two curves is calculated to account for cases where the curves intersect.

- Equal opportunity (EO):

  With respect to the protected attribute S and class attribute Y, a predicted outcome $\hat{Y}$ satisfies EO if:

$$EO = |P(\hat{Y} = -|Y = +, S = \overline{s}) - P(\hat{Y} = -|Y = +, S = s)|$$

- Predictive parity (PP):

  Predictive parity is satisfied if both groups; protected and non-protected, have equal precision

$$PP = |P(Y = +|\hat{Y} = +, S = s) - P(Y = +|\hat{Y} = +, S = \overline{s})|$$

- Predictive equality (PE):

  Predictive equality is satisfied if for both groups, the false positive rate is equal

$$PE = |P(\hat{Y} = +|Y = -, S = s) - P(\hat{Y} = +|Y = -, S = \overline{s})|$$

In all fairness measures, a higher value indicates a larger difference in predictions between the two groups, so the model is less fair, i.e., 0 stands for no discrimination.

4.3 Models

The results from different literature show that in creditworthiness analysis and other finance use cases for machine learning simple models such as decision trees, rather than deep machine learning models have better predictive ability.

This research focuses on three models:

- Logistic Regression
- Multi Layer Perceptron
- Support Vector Machine with Random Forest

The goal is to identify which model works best with the given datasets. These models are focused on in this research because existing literature supports them.

However, based on existing literature verifying that generally, SVM with Random Forest gives transparent and auditable results, which is crucial for this paper, the main focus is on SVM with Random Forest.

4.4 Logistic Regression

Logistic Regression is a supervised machine learning algorithm. It is used often for binary classification problems. It predicts the probability of a dependent variable by modeling the log of the odds of the event as a linear combination of the input features. Logistic Regression is modelled by:

$$log(\frac{p}{p-1}) = w^T x + b$$

where:

- $(p = P(y = 1|x))$ is the probability of class 1 given features x
- w is the vector of weights
- b is the bias

the predicted probability p is given by applying the sigmoid function:

$$p = \frac{1}{1 + 2^{-(w^T x + b)}}$$

The model is trained by maximizing the likelihood, or minimizing the log loss over training data.

4.5 Multi Layer Perceptron

Multi Layer Perceptron or MLP, is a feed forward artificial neural network. It consists of an input layer, one or more hidden layers, and an output layer. The

nodes in a layer are fully connected to the nodes in the next layer. It applies a non linear activation function to its weighted sum of inputs. The weights are adjusted through backpropagation using gradient descent optimization. It is also a supervised machine learning algorithm. For an input vector x, an MLP with one hidden layer computes:

$$(h = f(W_1 x + b_1), y = g(W_2 h + b_2))$$

where:

- W_1, W_2 are weight matrices
- b_1, b_2 are bias vectors
- f is the activation function in the hidden layer (e.g., ReLU)
- g is the output activation (e.g., sigmoid for binary classification)

4.6 Ensemble Learning: LOFO with SVM and DT

SVM is used to identify the most important features by analyzing support vectors or weight coefficients, and Random Forest is used as a classifier. This hybrid approach is backed by different citations.

The decision trees in random forest are additionally treated to Leave One Feature Out (LOFO) on the dataset, where sensitive attributes are left out one by one in the training process, and in different combinations to train models without bias in the dataset.

The resulting machine learning model is expected to have a low bias with a high accuracy.

Notation:

- Let $D = \{X, y\}$ be the dataset, where:
  - $X = (X_s, X_o)$ consists of:
    - X_S: Sensitive attributes (e.g., gender)
    - X_O: Other attributes (e.g., income, credit score, employment history)
  - y is the target variable (e.g., loan approval)
- Let $f_\theta(X)$ be trained Random Forest model parameterized by θ
- Let $T = T_1, T_2, ..., T_m$ be the set of decision trees in the Random Forest,
- Let $S = S_1, S_2, ..., S_k$ be all subsets of the sensitive attributes X_s,
  - including single-feature omissions and multiple-feature omissions

$$\hat{y} = \frac{1}{m} \sum_{j=1}^{m} T_j(X_O \cup (X_s \setminus S_j)), S_j \in S$$

where each tree is trained on a different subset of features with at least one sensitive attribute removed.

4.7 Evaluation

This research evaluates the model on two major criteria; fairness measures and machine learning evaluation metrics. The fairness measures are used on the final models to evaluate if they have bias.

For evaluations of the machine learning models in terms of their estimation accuracy, the following are used:

- Accuracy (ACC): This measures the overall correctness of the model's predictions, indicating the proportion of total correct predictions (both defaulters and non-defaulters). However, in imbalanced datasets, accuracy can be misleading because the model might be biased towards the majority class.

- F1 Score (F1): This metric is the harmonic mean of precision and recall, providing a single measure that balances the two. It is especially useful when the costs of false positives and false negatives are different, which is often the case in credit scoring.

- Precision (P): This indicates the proportion of true positive predictions among all positive predictions. High precision means that when the model predicts a borrower as risky, it is often correct, reducing the number of good borrowers incorrectly classified as risky.

- Recall (R): This measures the model's ability to identify all actual risky borrowers. High recall ensures that most risky borrowers are correctly identified, minimizing the number of risky borrowers that are missed.

- Area Under the ROC Curve (AUC): This metric provides an overall performance measure by illustrating the trade-off between true positive rate and false positive rate across different thresholds. A higher AUC indicates better model performance in distinguishing between defaulters and non-defaulters.

5 Datasets in Credit Scoring

The dataset utilized in this study is Germany Credit Data. It has 1000 entries of applications. The dataset was published in UCI repository. It was formed by Professor Dr. Hans Hofmann at the University of Hamburg under the Institute for Statistics and Econometrics. This dataset classifies people described by a set of attributes as good or bad credit risks. There are 20 feature attributes and 1 target attribute. There are both categorical and numerical attributes present. From the feature attributes, 7 are numerical and 13 are categorical.

5.1 Understanding the Dataset

The attributes in German Credit Data cover the factors that are taken into account for credit analysis at a bank for a loan. The features in the dataset are:

Table 5.1: Features in German Credit Data

	Feature Name	Description and Values	Type
1.	Status of existing checking account	The categories are based on ranges of amounts in the account kept for at least a year: • A11: Amount < 0 • A12: $0 <$ Amount < 200 • A13: Amount > 200 • A14: No checking account	Categorical

2.	Duration in months	The terms of financing, i.e., how many months the loan is provided for	Numerical
3.	Credit history	Representing whether previous loans exist, and what pattern was present in repayments • A30: No credit taken or all credits paid by duly	Categorical
4.	Purpose	The reason for applying for the loan: • A40: Car (new) • A41: Car (used) • A42: Furniture/equipment • A43: Radio/television • A44: Domestic appliances • A45: Repairs • A46: Education • A47: Vacation • A48: Retraining • A49: Business • A410: Others	Categorical
5.	Credit amount	The amount in the savings account, kept for at least a year, based on ranges: • A61: Amount $<$ 100 • A62: 100 $<=$ Amount $<$ 500 • A63: 500 $<=$ Amount $<$ 1000 • A64: Amount $>=$ 1000 • A65: Unknown/No Savings Account	Categorical

6.	Savings account	The amount in the savings account, kept for at least a year, based on ranges: • A61: Amount < 100 • A62: 100 <= Amount < 500 • A63: 500 <= Amount < 1000 • A64: Amount >= 1000 • A65: Unknown/No Savings Account	Categorical
7.	Present employment since	The duration in years, based on the ranges, the applicant has present employment since: • A71: Unemployed • A72: Less than 1 year • A73: Between 1 and 4 years • A74: Between 4 and 7 years	Categorical
8.	Instalment rate	The installment rate in the percentage of disposable income	Numerical
9.	Personal status and sex	The relationship status of the application and their sex: • A91: Male, divorced/separated • A92: Female, divorced/separated/married • A93: Male, single • A94: Male, married/widowed • A95: Female, single	Categorical
10.	Other debtors/guarantors	If the application is co-applied or if someone is acting as a guarantor: • A101: None • A102: Co-applicant • A103: Guarantor	Categorical
11.	Present residence since	The time the applicant has been living at the present residence for years	Numerical

12.	Property	Assets owned by the applicant other than a savings account: • A121: Real estate • A122: Building society savings agreement/life insurance • A123: Car or other • A124: Unknown/No Property	Categorical
13.	Age in years	The age of the applicant in years.	Numerical
14.	Other installment plans	Any other installments the applicant is paying: • A141: To the bank • A142: To stores	Categorical
15.	Housing	The type of housing the applicant lives in: • A151: Rent • A152: Own • A153: Free	Categorical
16.	Number of existing credits	The number of existing credits the applicant has with this bank	Numerical
17.	Job	The employment status and job type of the applicant: • A171: Unemployed/unskilled, non-resident • A172: Unskilled, resident • A173: Skilled employee/official • A174: Management, self-employed, highly qualified employee or officer	Categorical
18.	Number of people liable	The number of people being liable to provide maintenance for Numerical	
19.	Telephone	• A191: None • A192: Yes, registered under the applicant's name	Categorical

20.	Foreign worker	Whether the applicant is a foreign worker or not • A201: Yes • A202: No	Categorical

The target attribute has values 1 or 2, where 1 is good credit and 2 is bad credit. In the dataset, it is easy to see the relevance of some features to the target. However, it is harder to see how some features like gender, and foreign worker are correlated to our target variable.

To understand the role of each feature better in relation to the target, the long-standing statistical method in credit analysis, Weight of Evidence [46] is commonly used. Weight of Evidence (WOE) score, which quantizes the predictive power of an independent feature to a dependent one, can be used to derive Information Value (IV). WOE focuses on the odds ratio of good to bad. Information value reveals the overall strength of a variable. According to Siddiqi (2006), the IV in credit scoring can be interpreted as in [46]:

- Less than 0.02, the predictor is not useful
- 0.02 to 0.1, the predictor has a weak relationship
- 0.1 to 0.3, the predictor has a medium strength relationship
- 0.3 to 0.5, the predictor has a strong relationship
- 0.5, suspicious relationship

Table 5.2: Feature Information Value

Feature	Information Value
Status of checking account	0.666
Instalment Rate	0.293

Savings account	0.196
Purpose of Loan	0.166
Duration of Loan Repayment	0.165
Credit amount	0.119
Property	0.113
Age	0.093
Foreign worker	0.087
Employment Length	0.086
Housing	0.085
Credit history	0.058
Personal Status and Sex	0.045
Other debtors	0.032
Other instalment plans	0.026
Number of existing credits	0.013
Telephone	0.01
Job	0.009
Present residence since	0.004
Number of liable people	0.00004

From Information Value, it is interpretable that most features have a medium to weak relationship with the target. This implies that foreign worker, personal status, and gender have roughly the same contribution as the other features. Thus, our sensitive attributes are Personal status and sex, along with Foreign worker.

With the sensitive attributes identified on the basis of whether protected attributes account for as much Information Value towards the target attribute as other domain-relevant attributes, it is deducible that the dataset contains bias. Henceforth, any model that is trained on this dataset without accounting for bias in

pre-processing, training, or post-processing is expected to be prone to bias.

6 Data Exploration and Analysis

Prior to performing training of models and for a better understanding of the relationship of the dataset, certain checks and analyses are necessary and are a core need in machine learning. Exploration of the dataset verified that the dataset is clean on the basis of no missing values, no wrong or duplicate values, and no duplicate features. Out of the 1000 records in the dataset, all 1000 are unique. Given there are 1000 records, and 21 attributes in total, i.e., 20 feature attributes and 1 target attribute, all 21000 values are present.

To understand the dataset and the relationship of the dataset with respect to distribution and correlation, further data analysis is required.

6.1 Feature Distribution

As discussed above, of the total samples in the dataset, 700 are of good credit, whereas 300 are of bad credit. That provides an insight into the class imbalance. As the study focuses on sensitive attributes and bias, it is important to understand data bias further by also seeing a representation of the different values of the features, especially sensitive features.

Understanding the distribution of the dataset on the whole as well as on the feature level is important. Models are affected by feature distribution directly and indirectly. Highly skewed feature distribution can dominate model behavior. Many machine learning algorithms, in the case of the study, Logistic Regression (LR) and

Linear Discriminant Analysis (LDA), assume the input features follow a normal distribution. In the case of a violation, the models can expect a degradation of performance. Other machine learning algorithms, in the case of this study, Support Vector Machine (SVM), are sensitive to feature scales. Understanding the feature distribution helps understand how scaling needs to be performed on the dataset before proceeding with these models. Additionally, understanding feature distribution allows for identifying and handling outliers.

Feature Distribution helps recognize class imbalance or helps to identify underrepresented groups in the data. Identified imbalances can be resolved in data preprocessing steps. Additionally, one of the goals in the credit analysis and loan granting use case is the interpretability of the model. Knowing how different features behave in the dataset can make the model itself more explainable as well. Additionally, understanding the feature distribution can also help analyze if the dataset represents the real-world scenario.

6.2 Numerical Features

As the dataset has both numerical and categorical attributes, the manner in which their distributions are analyzed is slightly different. Numerical attributes' feature distribution can be analyzed using histograms, box plots, and Q-Q plots. Of the numerical attributes in the dataset, some have continuous values; duration in months, credit amount, and age. Others have discrete values; installment rate percentage, present residence since, number of existing credits at the bank, and number of people liable for.

Taking a look at Fig. 6.1, it can be seen which attributes have continuous values and which have discrete values (Fig. 6.2). Duration in months, Credit amount, and Age have continuous values. Their histograms also show that the distribution is right skewed. Most of the loans are in the range of one to three years. Only a

low number of samples in the dataset exist for longer ones. The credit amounts applied for mostly range between 1000 to 5000 in amount. Larger credit amounts exist, but are far fewer in number. Lastly, age, it can be seen that most of the credit analysis data samples exist for people aged 20 and 40 years old. The samples go up to seventy years of age, but they are underrepresented. One thing to note here is that this right skewness itself is representative of the real world, as the majority of the people applying for loans and undergoing credit analysis fall between this age, apply for those amounts, and the number of months for the loan.

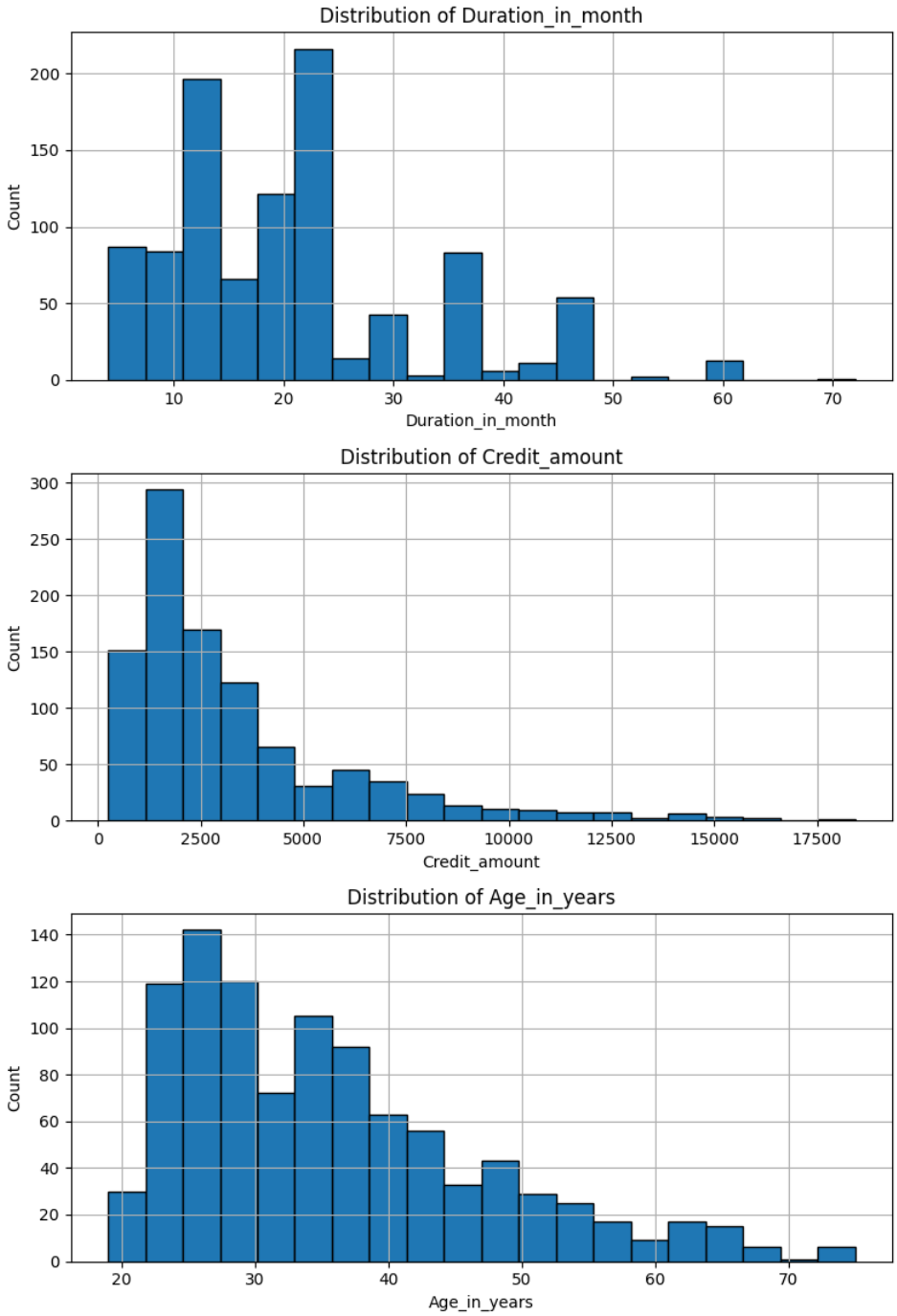

Figure 6.1: Histogram for Distribution of Continuous Numerical Features

For the discrete numerical attributes, in installment rate and present residence since, the distribution is less deviating and more so between normal and multimodal. Taking a look at Fig. 6.2, With respect to the number of existing credits at the bank

and the number of people liable to the applicant, it is again right skewed. Installment rates are between 1% and 4%, with 4% being the most represented in the samples. The sum of the other values' samples is equivalent to the number of samples, where 4% is the rate. In the dataset, the number of liable people the applicant has is 1 or 2, with a high majority having one. Most of the samples have at least one existing credit at the bank. The other values in the sample are 2 to 4, but are far less common.

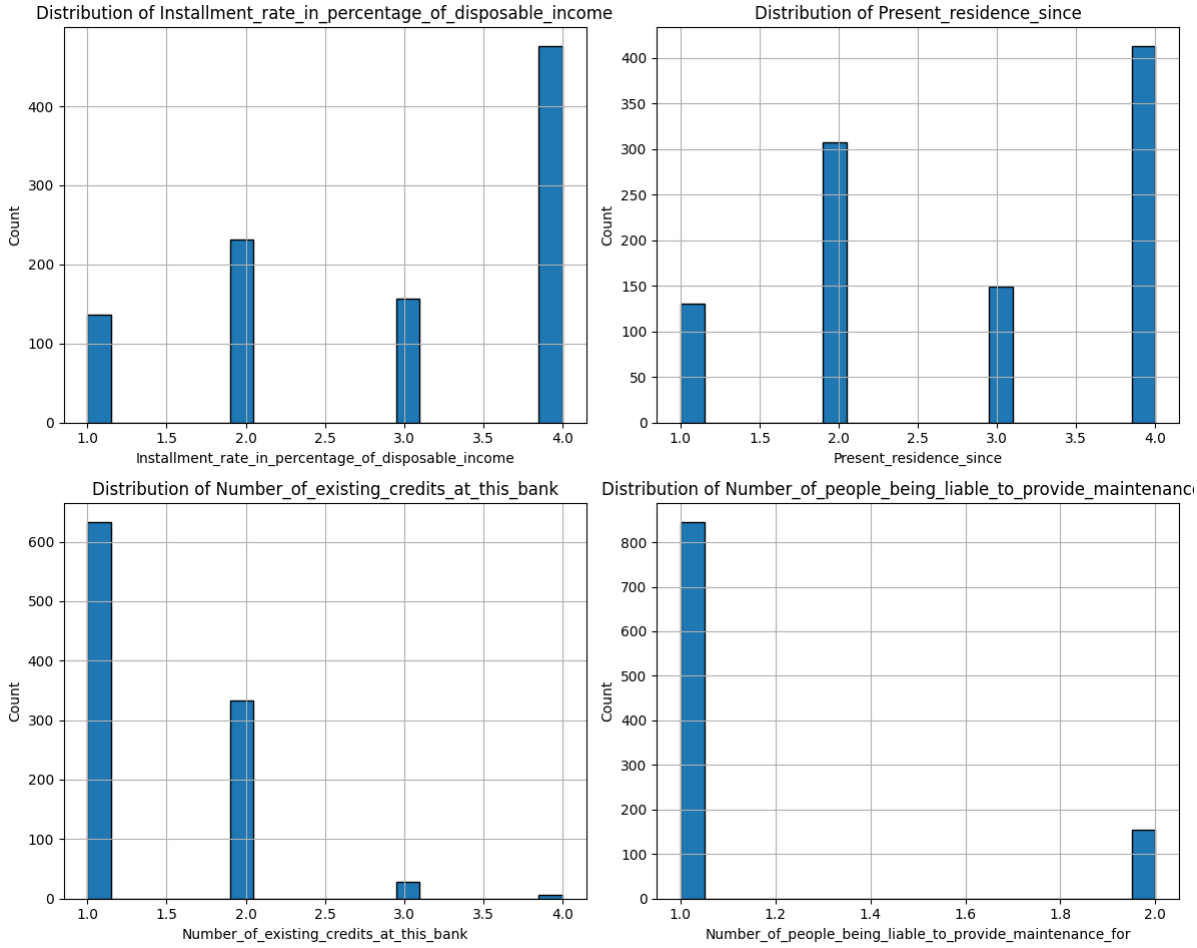

Figure 6.2: Histogram for Distribution of Continuous Numerical Features

Another important aspect of feature distribution is checking for outliers in each feature. To visualize the numerical features and analyze whether the features have

outliers and how spread out their values are, the box plot is observed (Fig. 6.3). As seen in histograms, most features have some degree of right-skewness. That implies that higher values are less common in the samples present in the German Credit Dataset. The installment rate feature has the most compact distribution. This aligns with the real world as it is a tightly controlled attribute. Age and duration show the widest ranges, i.e., there is variability in these features, which again aligns with the real world.

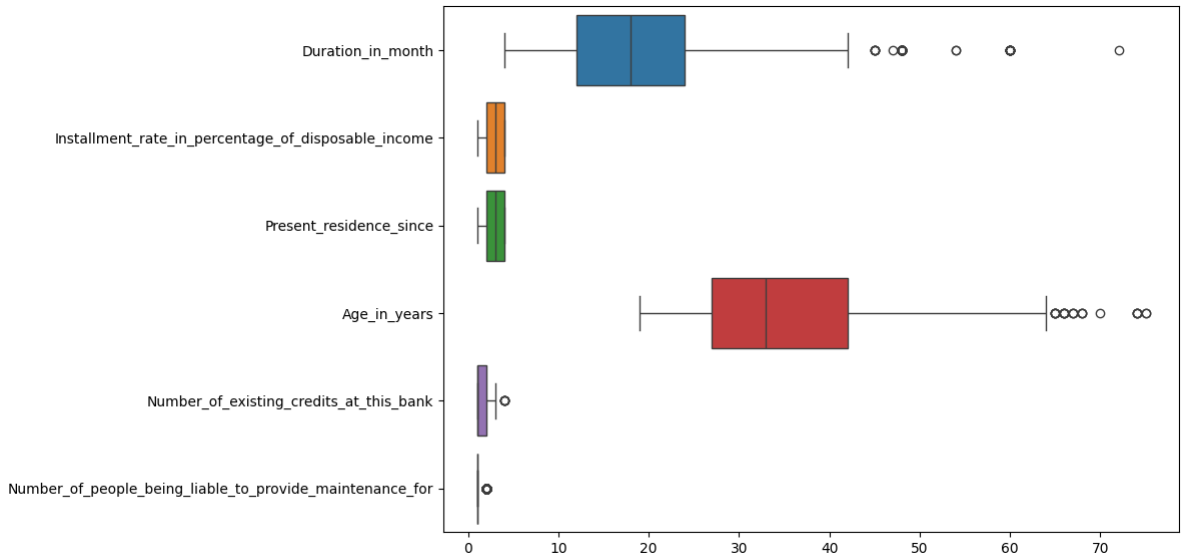

Figure 6.3: Boxplot of numerical features

Looking at the box plot for credit amount (Fig. 6.4), the range of credit amounts is diverse. It ranges from about 500 to 15,000. The interquartile range shows that 50% of the credit amounts fall between approximately 1,500 to 4,500. The majority of loans are concentrated in the lower to middle range. The distribution is highly right skewed, which is typical for financial data. It is important to note that there is a significant number of high value outliers. These deviate significantly from the typical loan amount, but this suggests that the dataset covers diverse customer types.

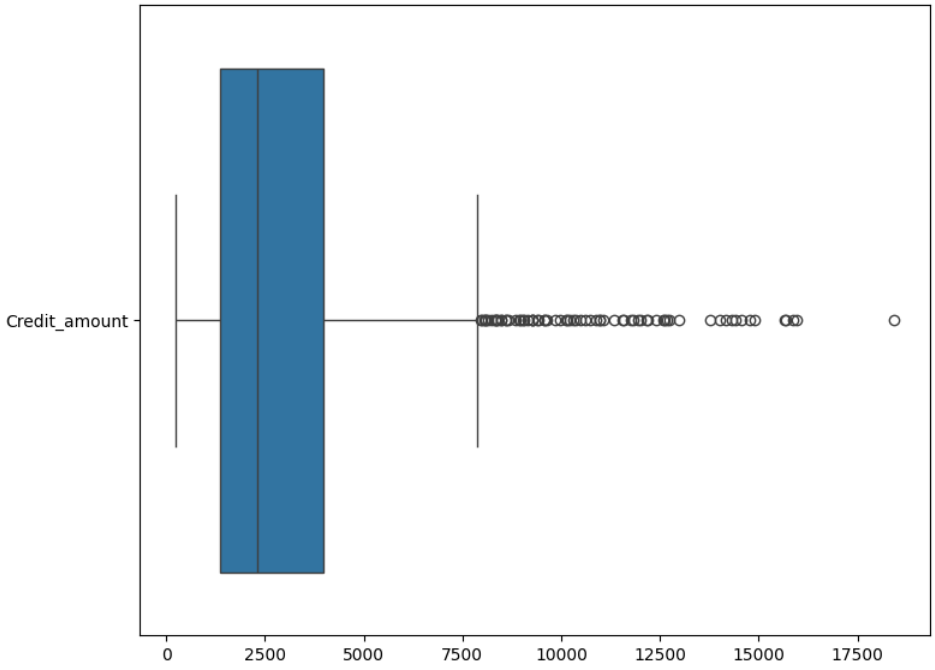

Figure 6.4: Boxplot of credit amount

6.3 Categorical Features

The dataset comprises of 13 categorical features. To understand their distribution, take a look at bar charts that describe the distribution both in terms of the number of samples and the number of samples for each value of the target attribute.

Looking into the distribution of the medium information value categorical features (Fig. 6.5), it is clear to see that bad credit is underrepresented for at least half of the categories of each feature. However, taking a look at each of the features individually provides insight into why that is the case. In the status of the existing checking account, it is clear to see that the higher bracket, i.e., A13, where the applicant has had over 200 amount in their account for at least a year, has the least "bad" credit. Where the applicant had 0 to 200 amount in their account (A11, A12), the "bad" credit is almost equal to the "good" credit outcome.

The distribution of purpose mainly shows the trend of what most people apply

loans for. It is also notable here that A46, i.e., education loans, have equal "bad" credit outcomes as "good" credit. These can be due to other factors associated with education loans, e.g., age, amount, and employment status, etc., as applicants for education loans are often students. Also, for most purposes of loans, the rate of "good" credit outcome is better than "bad".

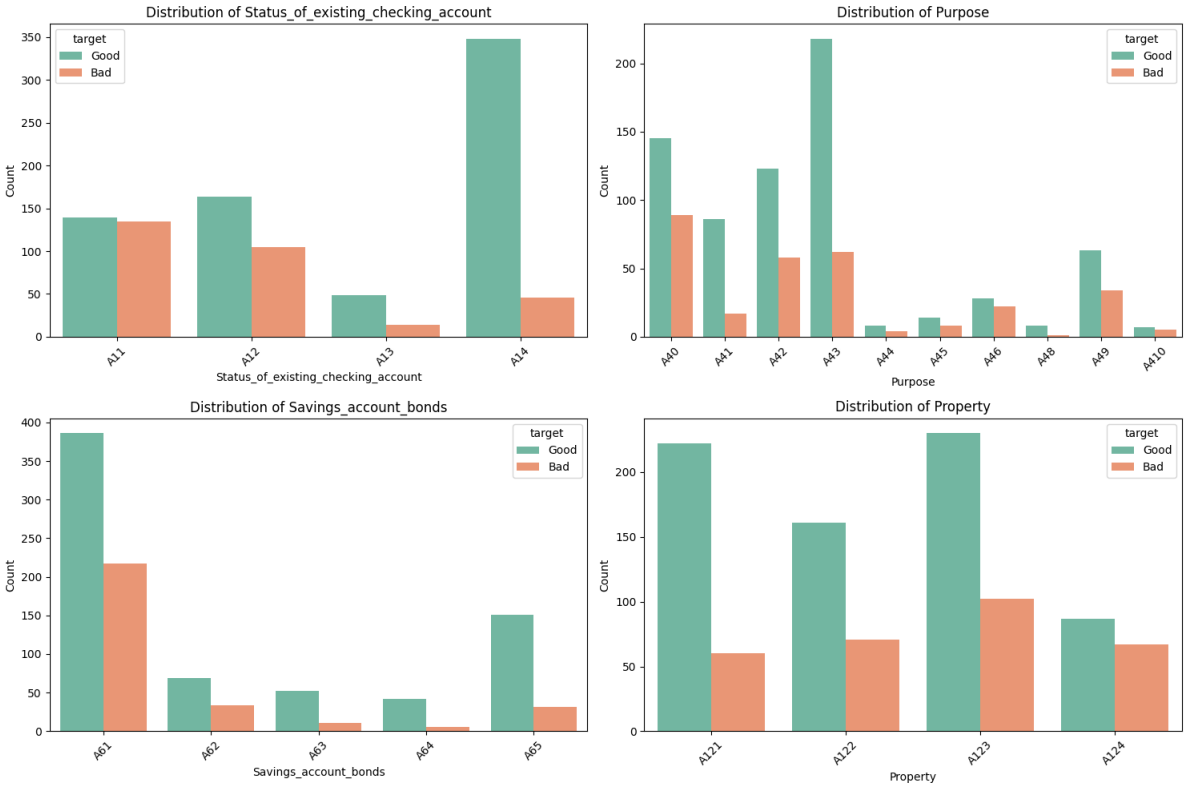

Figure 6.5: Histograms of medium information valued categorical features

The distribution of savings accounts/bonds shows that for all savings accounts, the "good" outcomes outweigh the "bad" ones.

The most notable, perhaps, is the distribution of outcomes on the attribute property. Here, it can be seen that with A124, the good and bad outcomes are the same. This implies that having no property/assets, most, provided property's medium information value, affects the outcomes. On the other hand, in the dataset,

most people own either real estate or have life insurance or a car. It can be seen that for the rest of the three property categories, not only is the ratio of good to bad similar, but also roughly the number of applicants in both good and bad outcomes is similar.

For the categorical features with medium information value, their category distributions differ substantially between Good and Bad credit. Status of checking account has good separation between categories. In Purpose, several categories show distinct distributions. Property indicates clear class separation based on property ownership type.

For the weak information value categorical features (Fig. 6.6), the charts show a significant imbalance in categories. For Job, most of the data for good credit is concentrated in A173, i.e. Skilled Employee, and bad credit is relatively spread across others. Both classes are prominent in the graph. For the distribution in the Telephone attribute, both categories have similar distributions. In the Other installment plans attribute, the majority of the samples present in the dataset have value A143 (None), so there is low variance in that attribute. In other debtors/guarantors, the majority of the samples are A101 (None), and there is low distribution across other categories. However, samples with A102 (co-applicant) there is a slightly higher proportion of Bad credit, but overall low separation. For present employment since, A73 and A75 are the most common for Good credit, i.e. applicants having present employment for more than a year.

Bad credit appears more spread out across A71, A72, and A73, i.e., between zero to four years of present employment. There is not a strong separation, however. In Housing, A152 (own housing) dominates both Good and Bad classes. Very few have A151 (rent) or A153 (free). The variation is low in Housing. In the Credit History attribute, A34 and A32 (see table for explanations) are dominant among Good credits. A32 and A34 also have a fair amount of Bad credit outcomes, but

A32 shows a relatively larger portion of Bad than Good. These features have low information value due to class imbalance within the feature values and overlapping distributions between "Good" and "Bad" targets. They are not good predictors on their own but might still be useful in interaction with other features or within ensemble models.

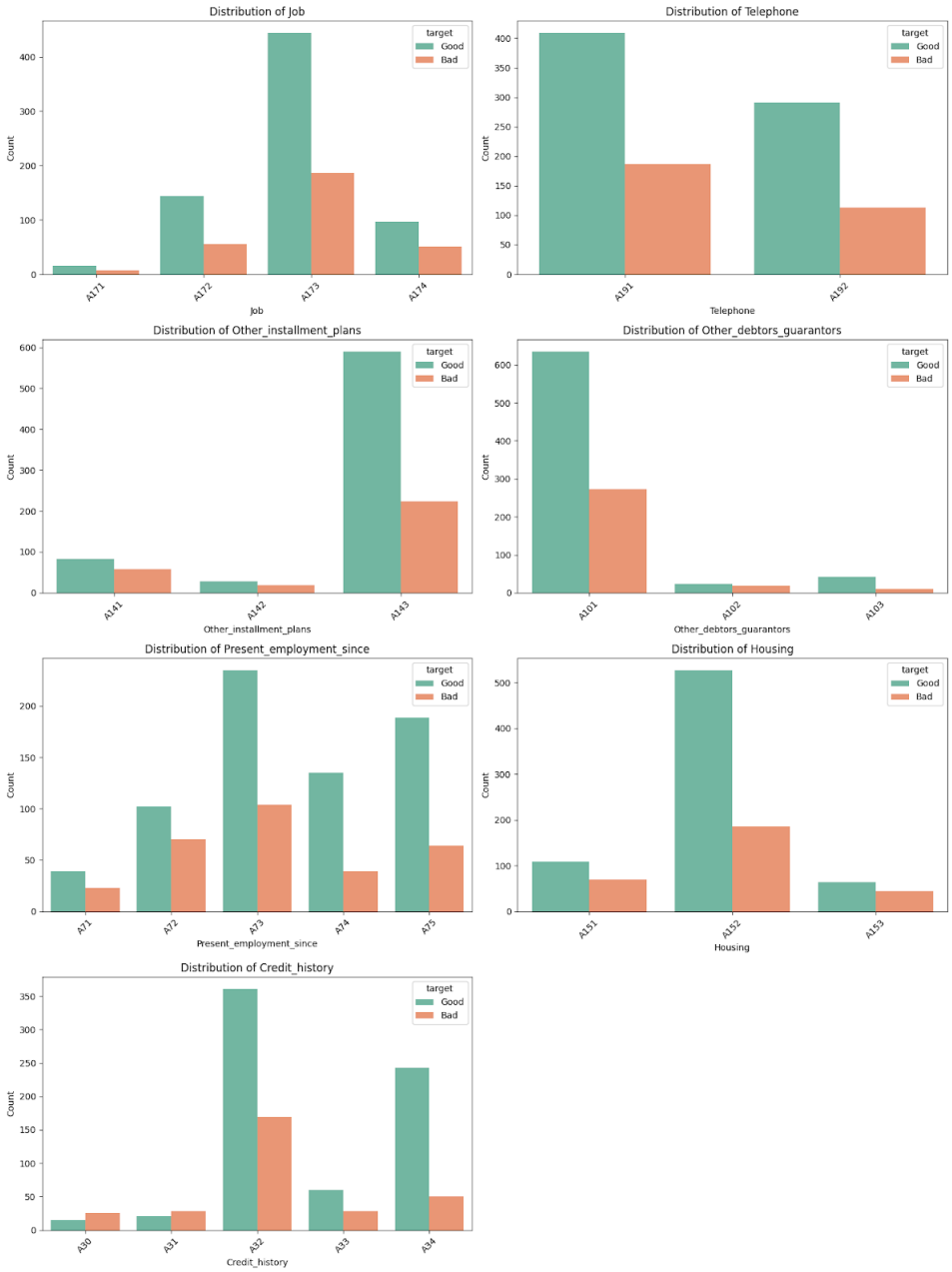

Figure 6.6: Histograms of weak information valued categorical features

Of these categorical features, two are the sensitive attributes under this study (Fig. 6.7). For Personal Status and Sex, categories represented in the samples in the dataset are:

- A91: male, divorced/separated
- A92: female, divorced/separated/married
- A93: male, single
- A94: male, married/widowed

It is important to note that in the dataset, there are no single, female applicants. A93 has the highest count overall, with a large portion labeled as Good, but also a significant number of Bad. A94 (male, married/widowed) also shows a positive skew but with fewer cases. A91 (male, divorced/separated) is the smallest group, but it still leans towards Good. One thing to note here is that A92 (female), despite having significantly fewer applicants, has equal Bad outcomes as to A93 (male, single). Based on the bar plots, there is some gender-based signal in the data.

In Foreign Worker attribute, the categories are:

- A201: Yes
- A202: No

Almost all applicants are A201 (foreign workers). It is important to note here that the Good outcomes to Bad outcomes in foreign workers are almost the same in number. And although very few samples have value A202, they have almost no Bad outcomes.

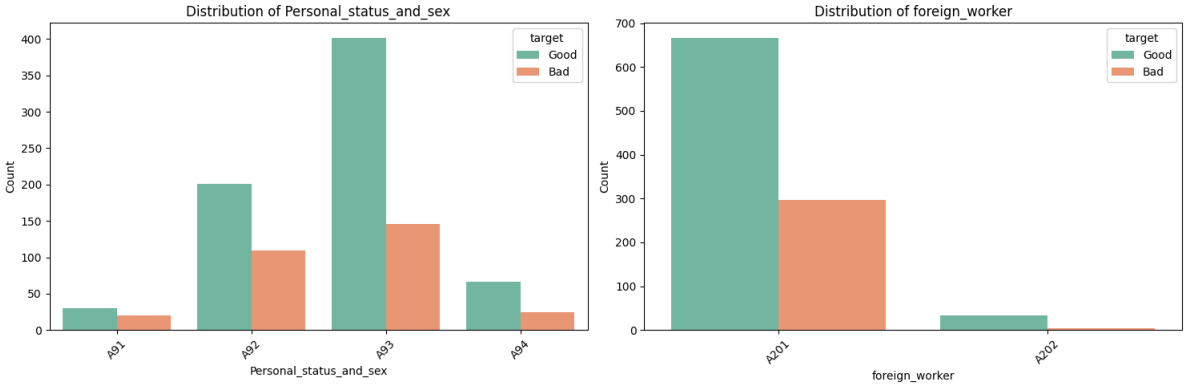

Figure 6.7: Histograms of sensitive attributes

6.4 Feature Correlation

Another important aspect of data analysis is finding multicollinearity in different features of the dataset. For numerical features, observing the correlation heatmap (Fig. 6.8), it can be seen that generally there is low multicollinearity. The heatmap reveals that most numerical features are only weakly correlated, except for a few moderate relationships. The data supports the use of these features together in predictive models.

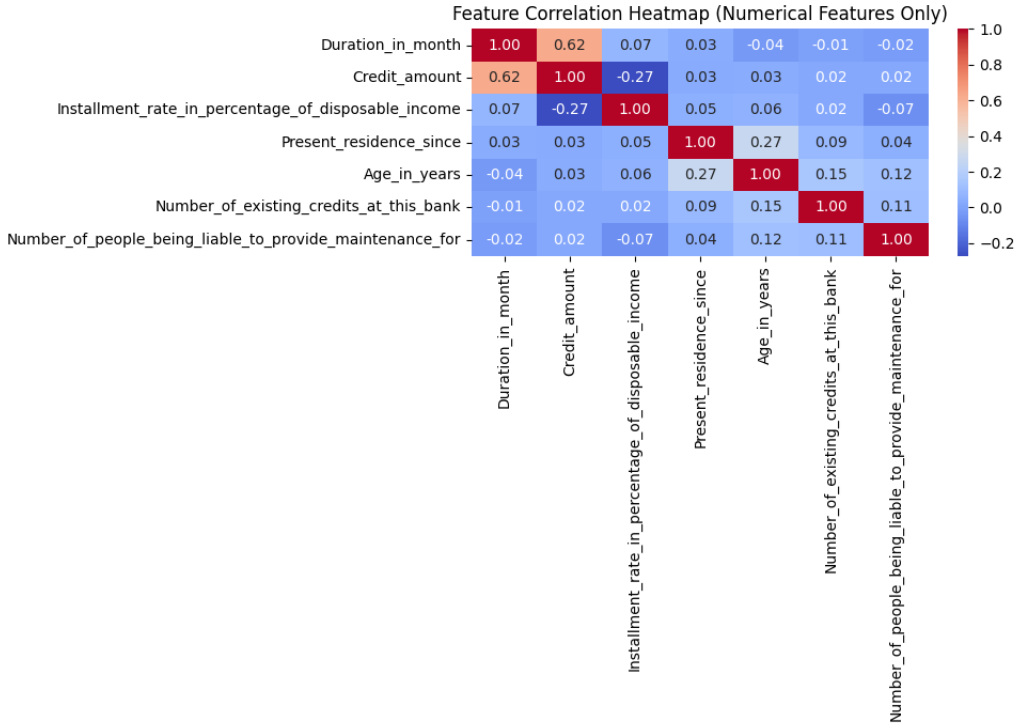

Figure 6.8: Feature correlation matrix of numerical attributes

The strongest correlations are Duration in months and Credit amount, at 0.62. Applicants taking loans over a longer duration tend to request higher credit amounts. This is a moderately strong positive correlation and quite intuitive. Credit amount and Installment rate have a negative correlation of -0.27. Higher credit amounts are slightly associated with lower installment rates as a percentage of income. This could imply that high loan applicants have more disposable income. Most other variables show very weak or no linear correlation. Since most features are not strongly correlated, each can contribute unique information, aiding interpretability.

7 Experimental Models

Having built the knowledge and understanding of the dataset, the next steps are to move towards training and evaluation of the models determined best for the use case of credit analysis for loan applications, alongside the proposed solution, keeping in mind the goal of reducing bias. This section discusses the models, how they are trained, their evaluation based on the metrics decided earlier, results of counterfactual analysis on the trained models, alongside the statistical tests to determine how biased the resulting models are.

7.1 Data Preparation

Before heading into the model details, first, this section discusses how the dataset is prepared to be fed to the models for training.

7.1.1 One Hot Encoding

The first thing that needs to be catered to is that the dataset is both numerical and categorical. For the categorical columns to be understood by the models for training, they need to be in some numerical form.

One hot encoding creates binary columns for each category of the categorical attributes. Suppose a categorical variable x can take values from a finite set of k unique categories:

$$x \in \{c_1, c_2, ..., c_k\}$$

Then the one-hot encoded vector $e(x) \in R_k$ is defined as:

$$e(x)_i = \{1 \text{ if } x = c_1 \text{ , } 0 \text{ otherwise, for } i = 1, 2, ..., k\}$$

7.1.2 Z-score Normalization

Features with a large scale of values can dominate models and make other features less influential. Many machine learning algorithms perform better when features are on the same scale. Normalization or Standardization ensures faster convergence and better model performance. The data is normalized using z-score normalization. It standardizes the features by transforming them to have zero mean and unit variance.

Given a feature , the standardization value z is calculated

$$z = \frac{x - \mu}{\sigma}$$

where μ is the mean and σ is the standard deviation of the feature.

7.1.3 Train/Test Split

To prevent overfitting of the model, and to estimate how the model performs on unseen data, the dataset is split into two: the training set and the test set. The ratio of the sets is 80 and 20 respectively. This also helps in model evaluation being done realistically. Additionally, while splitting the ratio of 7 : 3 that is present in the current dataset is retained in each set to ensure each set sees both classes.

7.1.4 Class Balancing

As seen in the sections before, there is a class imbalance in the dataset. To ensure balanced learning of both classes, the dataset needs to be balanced. This can be achieved by introducing class weights. The weight for a class i is computed as:

$$W_i = \frac{n_{samples}}{n_{classes} \times n_i}$$

where,

- $n_{samples}$ = total number of samples.
- $n_{classes}$ = total number of classes.
- n_i = number of samples in class i.

These weights multiply the penalty C for misclassification of samples in each class.

7.2 Model Evaluation

With the data evaluated, prepared, and models trained, the focus falls on how well the models are performing, alongside how biased they are. While tackling bias and ensuring transparency, it is as important for the model to also perform well, as otherwise, the intended use case is rendered useless.

The three models under this research are Logistic Regression, Multi-Layer Perceptron, and a hybrid SVM and Random Forest. The metrics used for evaluating the models are Accuracy, Precision, Recall, F1-Score, and AUC.

The hybrid model has the highest accuracy at 80%, and without much difference, logistic regression is in second place with 79.5% accuracy. The MLP model is the least accurate model at 74.5%. However, the MLP model has the highest precision

at 0.88, indicating fewer false positives. Next in precision is the hybrid model, and then logistic regression. Logistic Regression has the highest recall, indicating it is the best at capturing true positives, and is closely followed by the hybrid model. The MLP model's recall is much lower compared to these two. Both the Hybrid Model and the Logistic Regression have similar F1-Scores, indicating balanced performance between precision and recall. The biggest difference can be seen in the evaluation metrics in the AUC. The hybrid model has an AUC of 0.97, showcasing its ability to distinguish best between the good and bad credit outcomes. The MLP is second with 0.86 AUC, and logistic regression has 0.79 AUC.

Logistic Regression is strong in recall and F1-Score, with decent accuracy. However, AUC is the weakest, suggesting it's not the best at ranking predictions. MLP has the highest precision, but suffers from low recall and F1-score. It's likely missing many true positives. The hybrid model shows the best accuracy and AUC, and ties for the best F1-score. It effectively balances all aspects of performance, making it the most well rounded and reliable choice among the three.

Table 7.1: Evaluation Metrics Values

Model	Accuracy	Precision	Recall	F1-Score	AUC
LR	0.7950	0.83	0.89	0.86	0.7976
MLP	0.7450	0.88	0.74	0.80	0.8659
Hybrid	0.8000	0.84	0.88	0.86	0.9714

7.2.1 Counterfactual Analysis (CFA)

The second question the research is aimed at is "What fairness metrics are suitable for assessing bias?". One way to analyse the trained models is counterfactual analysis. A non biased model would not change its outcome from negative to positive when only the value of the sensitive attribute is changed.

The percentage of predictions that changed from bad to good when "female" was

changed to a "male" category is represented by Female CF Change in the table. The MLP model had the highest percentage at 11.61%, and logistic regression closely followed it at 10.32%. This means that about 1 in 9 applicants in this sample, simply changing the gender attribute, resulted in a more favorable credit analysis outcome for MLP, and 1 in 10 for Logistic Regression. The Hybrid model, on the other hand, has only a 0.97% change. This makes the Hybrid model the least biased towards gender, whereas MLP and LR have some bias.

For Foreign Worker, the percentage of predictions that changed from bad to good, when Foreign Worker attribute was set to no, while keeping the rest the same, the logistic regression model showed the most bias. It had a percentage of 17.24, meaning about 1 in 6 applicants, when treated as nonforeign workers, had a "good" outcome. For MLP, the same percentage was 13.81%. This implies that about 1 in 7 applicants had the outcome change from bad to good. Both logistic regression and MLP show that they are sensitive to foreign worker status, and thus hold a bias resulting from the data. The Hybrid model, on the other hand, has a much lower percentage of 1.77% for foreign workers.

The counterfactual analysis shows some influence of gender in MLP and logistic regression, but nearly nonexistent in the hybrid model. As for a foreign worker, being a non-foreign worker decreases the predicted risk of a bad outcome. Both logistic regression and MLP are sensitive to foreign worker status. Both models show the presence of bias, especially compared to the Hybrid model.

Table 7.2: Counterfactual Analysis of sensitive attributes

	LR	MLP	Hybrid
Female CF Change	10.32%	11.61%	0.97%
Foreign Worker CF Change	17.24%	13.81%	1.77%

Table 7.3: Feature Importance Coefficient of sensitive attributes

	LR	MLP	Hybrid
A91 CF Feature Importance Coefficient	0.380367	0.0041	0.114496
A92 CF Feature Importance Coefficient	0.004545	0.0053	0.013591
A93 CF Feature Importance Coefficient	-0.499435	0.0123	-0.176289
A94 CF Feature Importance Coefficient	0.009928	0.0015	-0.013247
A201 Feature Importance Coefficient	0.478152	0.0004	0.234288
A202 Feature Importance Coefficient	-0.582747	0.0024	-0.295737

7.2.2 Statistical Tests

Other than counterfactual analysis, statistical tests provide insight into the bias present in the models. For gender, the treatment equality test tells in logistic regression there is a difference of 85 in the sum of false positives and false negatives between the protected and non-protected groups. For MLP, it is 79. The Hybrid Model has the lowest at 21. A large value indicates a notable disparity in error rates between groups. The Hybrid Model is highly fair in error rates for gender. The ROC curves for the protected and non-protected groups in gender are quite similar for MLP and LR at a rate of about 2.5% average absolute difference. While these are good already, and indicate the models' ability to distinguish between good and bad credit risk is fairly consistent across gender groups, the Hybrid model has an even lower rate at 0.9%.

For logistic regression, the difference in false negative rates (FNR) between groups is about 3.7%, and the probability that a positive prediction is correct is almost the same for both groups. These indicate little disparity in the model's precision. For MLP, the difference in false negative rates (FNR) between groups is 6.5%. This means the likelihood of incorrectly denying credit to a "good" applicant

is somewhat higher. The probability that a positive prediction is correct differs by only 1% between groups, indicating very little disparity in precision. For both, the predictive equality is high at 16% and 9%, indicating the models are more likely to incorrectly approve a "bad" applicant in non protected group than protected group. For the Hybrid model, EO is 0.4%, PP is is 1.3%, and PE is just 0.2% indicating the hybrid model is much less biased compared to the other two. The likelihood of incorrectly denying credit to a "good" applicant is nearly the same for both groups in the hybrid model, it indicates very little disparity in precision, and it shows almost perfect equality in false positive rates. The Hybrid model is the least bias towards protected group under gender.

For Foreign Worker attribute, the statistical tests just like the CFA show the model to be more biased than gender. For logistic regression, there is a much larger disparity (219) in the sum of false positives and false negatives between foreign workers and non-foreign workers, indicating a significant error rate imbalance. It is nearly equal for MLP at 213. Compared to these, the hybrid model shows a much less imbalance at 57, meaning the imbalance is less by nearly 1/4th. The ROC curves for the protected and non-protected groups in gender are quite similar for all three models. The Equal Opportunity statistical test shows MLP has the highest difference rate in false negatives, at 18.5%, and LR at 11.8%, meaning foreign workers and non foreign workers have noticeably different chances of being incorrectly denied credit when they are actually "good" applicants. The Hybrid Model, on the other hand, has only a 1.8% rate. Predictive Parity indicates very little disparity in precision between the protected and non protected groups for all three models, but the values in descending order as MLP, LR, and then Hybrid. The predictive equality however, is the only statistical test that shows the Hybrid model showcasing a higher rate than both MLP and LR, indicating the Hybrid model holds some disparity in incorrectly approving "bad" applicants between groups, whereas

in both MLP and LR, it is low.

The majority of the statistical test's results align with the counterfactual analysis. They showcase that of the three trained models, the Hybrid model shows the least bias, compared to the other two models. However, what is important to note here is not only the comparison with other models, but also the Hybrid model's capability to reduce the bias that was present.

Table 7.4: Statistical Tests on sensitive attributes

	Test	LR	MLP	Hybrid
Personal Sex Status	Treatment Equality	85	79	21
	ABROCA	0.02492	0.02608	0.00890
	Equal Opportunity	0.03709	0.06514	0.00365
	Predictive Parity	0.00465	0.01022	0.01331
	Predictive Equality	0.16312	0.09015	0.00249
Foreign Worker	Treatment Equality	219	21	57
	ABROCA	0.09449	0.08303	0.08377
	Equal Opportunity	0.11844	0.18527	0.01767
	Predictive Parity	0.13737	0.062443	0.01048
	Predictive Equality	0.02027	0.074324	0.15878

8 Model and Bias Discussion

The hybrid model based on SVM and ensemble techniques (Random Forest) tackles bias in the dataset, provides good accuracy, and is interpretable and transparent. This method represents a thoughtful evolution of machine learning practices, with fairness, interpretability, and modularity at its core.

While many ensemble techniques aim to improve raw predictive performance through techniques like bagging, boosting, or stacking, this approach distinguishes itself by using intentional data selection strategies for each model in it. These data selections are not arbitrary. They are central to the design, letting the system to explicitly model and analyze the effects of sensitive attributes in data. This makes the ensemble more than just predictive. It becomes a fairness aware decision making framework.

The model performance metrics are important to this research, as the bias measurements and improvements themselves are rendered useless if the model is unable to perform well. The goal of training the hybrid model, exceeds the expectations as it not only performs as good as the other models, but also on majority of the metrics performs better than the other models. This solidifies that an ensemble method approach, specifically a Random Forest, remains a solid approach for financial services and especially for the credit analysis and loan processes.

As observed under section 7.2, amongst the models trained and evaluated, the hybrid model is most well-rounded. The models are tested under evaluation metrics

described in section 4.7 i.e. Accuracy, Precision, Recall, F1 score, and AUC. Under every evaluation metric, the hybrid model performs just as good or better than the other models. The evaluation metrics radar in Fig. 8.1 summarizes the comparison. The Hybrid model is giving the highest accuracy on the test data. It also has high recall, and F1 score. Most notably, it has a higher AUC score compared to the other models. The Hybrid model is good at predicting both the classes. It strikes the right balance between precision and recall, and the trained model has a good ability to distinguish between the classes itself. The high AUC score implies that the model has a good class separation, and has learned well from the dataset's features itself.

These evaluation results satisfy that the first requirement for this approach i.e. good performance is satisfied, and the Hybrid model does not have a trade off between fairness and performance.

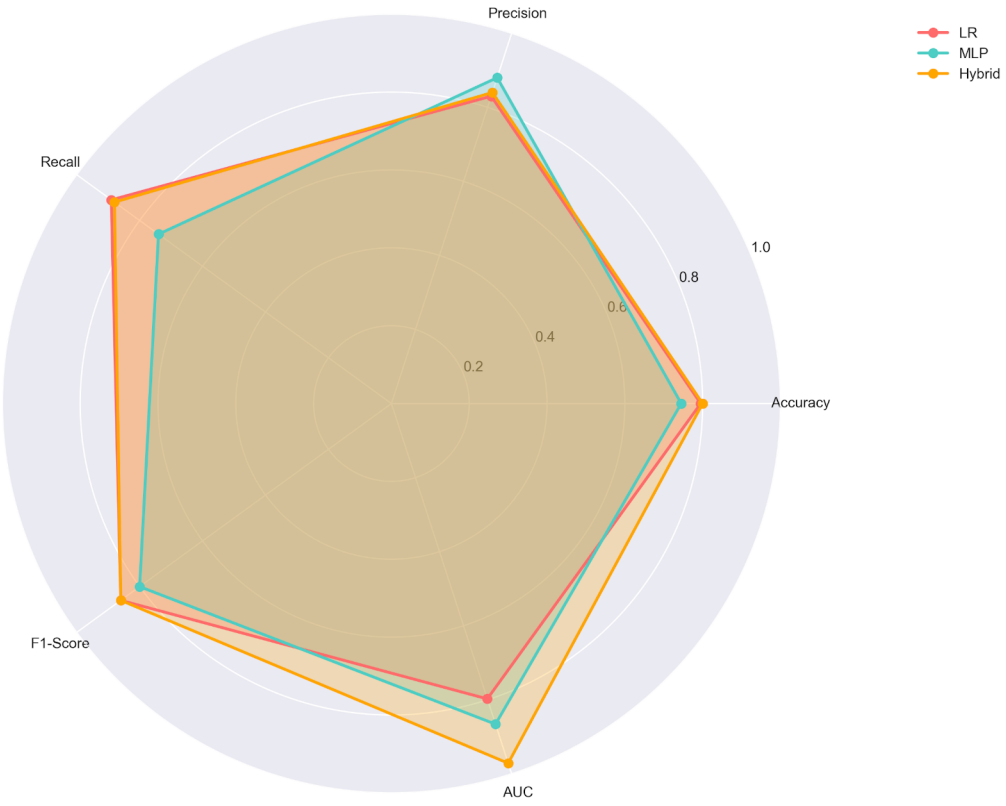

Figure 8.1: Model evaluation radar for the three trained models (LR, MLP, Hybrid)

At the heart of this method is the recognition that fairness is not a monolithic concept. Rather than imposing a single definition of fairness across a dataset with overlapping demographic attributes, this system allows for the coexistence of multiple fairness assumptions. Each configuration (tree definition) views the data differently. One configuration might say, "Let's pretend we don't have any data about foreign workers," while another might say, "Let's train the model on all data except male applicants." These are not just technical manipulations, they represent different normative choices about what information a model should or should not be allowed to consider when making predictions. This is a stark contrast to conventional Machine Learning pipelines, which often focus on maximizing accuracy without critically examining which groups are underrepresented, overrepresented, or unfairly weighted. In traditional workflows, sensitive features like race, gender, or nationality are either included and risk introducing bias, or dropped entirely and risk losing signal. The binary nature of that decision leads to ethically ambiguous models. The selective data approach offers a middle path. It allows multiple perspectives on fairness to be included, allowing the model to make better decisions in aggregate.

Additionally, this isn't just a fairness technique. It is a robust machine learning technique grounded in the theory of diversity in ensembles. Diverse models' results aggregated, e.g., via majority voting, reduce the variance of the final predictions and are less likely to overfit noise or bias in any one view of the data. What is different in this approach is that the diversity is not a result of random sampling or boosting, but intentional and interpretable. Practitioners can point to each tree and say precisely how and why it was trained differently from the others, which lends itself to transparency and reproducibility.

Consider a concrete example: a decision tree trained without any foreign worker feature may underperform slightly when compared to one that includes them, but it

might better align with certain fairness criteria or legal requirements. On the other hand, a tree trained with the full feature set might capture more information and contribute stronger outcomes on average. By having both of these models in the ensemble, the system does not have to commit to a single ethical stance or modeling strategy. Instead, it allows those models to "vote" on each prediction, leading to outcomes that are more balanced and justifiable.

One of the core strengths of this approach is the SVM aspect. The data selection directly shapes the feature selection process through SVM. Since each tree is trained on a different subset of data or features, the SVM will identify a different ranking of what is predictive or discriminative under that tree. This means each decision tree does not only vary in its input data, but also in which features it considers most relevant. Some may emphasize income related features, others might lean on housing, employment time, or credit history. This variation leads to a diverse ensemble in terms of decision logic and feature dependency, which enhances both performance and fairness.

From a practical standpoint, selective data training is also highly flexible. Organizations can define trees based on new fairness metrics, legal mandates, or social expectations without changing the entire system. Suppose a new policy states that credit decisions must not consider marital status. The organizations can remove configurations that include marital status or add new configurations that drop those features.

Another benefit is interpretability. By observing how different trees perform across different groups, analysts can better understand where model performance disparities lie and how sensitive predictions are to certain features. If the ensemble's accuracy drops sharply when gender is removed from training, it raises a red flag that the model is overly dependent on gender for predictions. Such analyses can help audits. Decision trees are among the most interpretable machine learning models.

By design, this method limits the number of features per model, further enhancing transparency. Each tree can be inspected, visualized, and explained. Combined with documented tree rules this makes the model's behavior traceable and auditable.

9 Conclusion and Prospect

9.1 Conclusion

In this thesis, a Hybrid SVM and Random Forest model is introduced. It aims to produce a model for credit analysis that is transparent, interpretable, and fair, while also being accurate. The motivation behind this is the regulations and ethics revolving around the finance sector, and the presence of documented bias in the history of creditworthiness analysis. The work accomplished in this thesis can be divided into parts.

The first part analyses the existing literature to define bias in finance sector data as well as machine learning, access what machine learning models work best for finance usecases, how bias can be observed in machine learning models, and how it can be tackled. The introduced model is designed on the basis of existing literature. The existing literature supports that Random Forest, along with feature selection on the basis of feature importance as well as intentional leave out concept can lead to better performing and fair models. For comparison, other well performing models; Logistic Regression and Multi Layer Perceptron are decided. Methodologies for quantified analysis of bias are derived. Counterfactual analysis alongside statistical tests such as treatment equality, equal opportunity, predictive parity, predictive equality, and ABROCA are set to be used for bias analysis. Different credit based datasets are explored, and based on relevance to the thesis alongside quality of

dataset, German Credit Data available on UCI Repository is selected.

The second analyzes the dataset. It contains twenty features, both categorical and numerical, and one target attribute which defines if the credit is good or bad. The sensitive attributes under this dataset are Personal Status and Sex, and Foreign Worker status. All the features are analyzed to determine their Information Value. The sensitive attributes are found to have similar information value as many others features, which indicates presence of bias in the dataset. The dataset is further analyzed to understand distribution, balance, and correlation as these affect machine learning algorithms' performance. The dataset is found to be clean, and the distribution is found to map real-world scenarios.

The third part focuses on training and evaluation of the models. The dataset is prepared for the models using one hot encoding, normalization, splitting, and class balancing techniques. The proposed model along with logistic regression and multi layer perceptron are trained on it. The model performance evaluation is done using accuracy, precision, recall, f1 score, and AUC metrics. The comparison shows that the proposed model balances all aspects of performance, making it the most well rounded and reliable choice among the three. The counterfactual and statistical tests on the three models showcase that the proposed model is the most fair, and compared to other models shows significant reduction in bias.

9.2 Prospect

This thesis can serve as a stepping stone for eradicating bias in machines at in-processing level. It serves to provide transparency, interpretability, and fairness, and a way to avoid binary feature selection. There is still a lot more that can be explored as an extension to this research. One major future research can be how this methodology can be extended to other machine learning algorithms. Additionally, this thesis focuses on German Credit Data. It would be interesting to explore how

other datasets can be handled using this methodology, and how they react to the approach both in terms of performance and fairness. The reliability of this approach in theoretical and research scenarios is good, but how reliable it is in commercial use is also left for testing. Additionally, it would be interesting to see how well the method scales in comparison to others when the dataset is treated to bias eradicating methods prior to training itself.

References

[1] *Eu household debt, 1999 – 2022*, `https://www.ceicdata.com/en/indicator/european-union/household-debt`, 2023.

[2] J. Stavins *et al.*, "Credit card borrowing, delinquency, and personal bankruptcy", *New England Economic Review*, pp. 15–30, 2000.

[3] G. Cafiso, "The contribution of loans to economic activity.", CESifo Working Paper, Tech. Rep., 2020.

[4] "Financial sector ai spending 2024", *Statista*, 2024.

[5] C. Jung, H. Mueller, S. Pedemonte, S. Plances, and O. Thew, "Machine learning in uk financial services", *Bank of England and Financial Conduct Authority*, 2019.

[6] M. Hurley and J. Adebayo, "Credit scoring in the era of big data", *Yale JL & Tech.*, vol. 18, p. 148, 2016.

[7] P. Aghion, Y. Algan, P. Cahuc, and A. Shleifer, "Regulation and distrust", *The Quarterly journal of economics*, vol. 125, no. 3, pp. 1015–1049, 2010.

[8] S. Ongena and A. Popov, "Gender bias and credit access", *Journal of Money, Credit and Banking*, vol. 48, no. 8, pp. 1691–1724, 2016.

[9] L. Guiso, P. Sapienza, and L. Zingales, "The role of social capital in financial development", *American economic review*, vol. 94, no. 3, pp. 526–556, 2004.

[10] A. Muravyev, O. Talavera, and D. Schäfer, "Entrepreneurs' gender and financial constraints: Evidence from international data", *Journal of comparative economics*, vol. 37, no. 2, pp. 270–286, 2009.

[11] A. F. Alesina, F. Lotti, and P. E. Mistrulli, "Do women pay more for credit? evidence from italy", *Journal of the European Economic Association*, vol. 11, no. suppl_1, pp. 45–66, 2013.

[12] S. Y. Deku, A. Kara, and P. Molyneux, "Access to consumer credit in the uk", *The European Journal of Finance*, vol. 22, no. 10, pp. 941–964, 2016.

[13] K. K. Charles and E. Hurst, "The correlation of wealth across generations", *Journal of political Economy*, vol. 111, no. 6, pp. 1155–1182, 2003.

[14] P. Bayer, F. Ferreira, and S. L. Ross, "The vulnerability of minority homeowners in the housing boom and bust: Corrigendum", *American Economic Journal. Economic Policy*, vol. 9, no. 1, p. 344, 2017.

[15] R. Kashiwagi, "How to fix bias concealed in credit scores", Nomura Research Institute (NRI), Tech. Rep., 2020.

[16] A. Bellucci, A. Borisov, and A. Zazzaro, "Does gender matter in bank–firm relationships? evidence from small business lending", *Journal of Banking & Finance*, vol. 34, no. 12, pp. 2968–2984, 2010.

[17] J. C. Ong and L. S. Lee, "Credit scoring: A comparison of machine learning models and their modifications", *Applied Mathematics and Computational Intelligence (AMCI)*, vol. 14, no. 1, pp. 57–78, 2025.

[18] A. Ampountolas, T. Nyarko Nde, P. Date, and C. Constantinescu, "A machine learning approach for micro-credit scoring", *Risks*, vol. 9, no. 3, p. 50, 2021.

[19] B. Baesens, T. Van Gestel, S. Viaene, M. Stepanova, J. Suykens, and J. Vanthienen, "Benchmarking state-of-the-art classification algorithms for credit

scoring", *Journal of the operational research society*, vol. 54, no. 6, pp. 627–635, 2003.

[20] S. Lessmann, B. Baesens, H.-V. Seow, and L. C. Thomas, "Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research", *European Journal of Operational Research*, vol. 247, no. 1, pp. 124–136, 2015.

[21] Y. Li and W. Chen, "A comparative performance assessment of ensemble learning for credit scoring", *Mathematics*, vol. 8, no. 10, p. 1756, 2020.

[22] S. K. Trivedi, "A study on credit scoring modeling with different feature selection and machine learning approaches", *Technology in Society*, vol. 63, p. 101 413, 2020.

[23] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness", in *Proceedings of the 3rd innovations in theoretical computer science conference*, 2012, pp. 214–226.

[24] J. Gu and D. Oelke, "Understanding bias in machine learning", *arXiv preprint arXiv:1909.01866*, 2019.

[25] M. J. Kusner, J. Loftus, C. Russell, and R. Silva, "Counterfactual fairness", *Advances in neural information processing systems*, vol. 30, 2017.

[26] A. Kasirzadeh and A. Smart, "The use and misuse of counterfactuals in ethical machine learning", in *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 2021, pp. 228–236.

[27] S. Verma, V. Boonsanong, M. Hoang, K. Hines, J. Dickerson, and C. Shah, "Counterfactual explanations and algorithmic recourses for machine learning: A review", *ACM Computing Surveys*, vol. 56, no. 12, pp. 1–42, 2024.

[28] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning", *Advances in neural information processing systems*, vol. 29, 2016.

[29] A. Chouldechova, "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments", *Big data*, vol. 5, no. 2, pp. 153–163, 2017.

[30] S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, and A. Huq, "Algorithmic decision making and the cost of fairness", in *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*, 2017, pp. 797–806.

[31] R. Berk, H. Heidari, S. Jabbari, M. Kearns, and A. Roth, "Fairness in criminal justice risk assessments: The state of the art", *Sociological Methods & Research*, vol. 50, no. 1, pp. 3–44, 2021.

[32] J. Gardner, C. Brooks, and R. Baker, "Evaluating the fairness of predictive student models through slicing analysis", in *Proceedings of the 9th international conference on learning analytics & knowledge*, 2019, pp. 225–234.

[33] H. Zhao and G. J. Gordon, "Inherent tradeoffs in learning fair representations", *Journal of Machine Learning Research*, vol. 23, no. 57, pp. 1–26, 2022.

[34] J. Adebayo and L. Kagal, "Iterative orthogonal feature projection for diagnosing bias in black-box models", *arXiv preprint arXiv:1611.04967*, 2016.

[35] V. Xinying Chen and J. N. Hooker, "A guide to formulating fairness in an optimization model", *Annals of Operations Research*, vol. 326, no. 1, pp. 581–619, 2023.

[36] Z. C. Lipton, "The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery.", *Queue*, vol. 16, no. 3, pp. 31–57, 2018.

[37] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions", *Advances in neural information processing systems*, vol. 30, 2017.

[38] F. Kamiran and T. Calders, "Data preprocessing techniques for classification without discrimination", *Knowledge and information systems*, vol. 33, no. 1, pp. 1–33, 2012.

[39] D. Bertsimas, J. Pauphilet, and B. Van Parys, "Rejoinder: Sparse regression: Scalable algorithms and empirical performance", 2020.

[40] M. Hardt, N. Megiddo, C. Papadimitriou, and M. Wootters, "Strategic classification", in *Proceedings of the 2016 ACM conference on innovations in theoretical computer science*, 2016, pp. 111–122.

[41] *Australian credit approval*, `https://archive.ics.uci.edu/dataset/143/statlog+australian+credit+approval`, accessed Jun. 17, 2025, 2018.

[42] H. Hofmann, *Statlog (german credit data)*, UCI Machine Learning Repository, DOI: https://doi.org/10.24432/C5NC77, 1994.

[43] *Japanese credit screening*, `https://archive.ics.uci.edu/dataset/28/japanese+credit+screening`, accessed Jun. 17, 2025, 2024.

[44] *Lending club 2007-2020 q3*, `https://www.kaggle.com/datasets/ethon0426/lending-club-20072020q1`, accessed Jun. 17, 2025, 2020.

[45] "Give me some credit", Tech. Rep., 2025.

[46] H. A. Abdou and J. Pointon, "Credit scoring, statistical techniques and evaluation criteria: A review of the literature", *Intelligent systems in accounting, finance and management*, vol. 18, no. 2-3, pp. 59–88, 2011.