Fairness in Credit Scoring: A Case Study

Introduction:

Credit scoring systems critically influence access to financial services, and the growing use of machine learning (ML) in credit decisions has raised concerns about discriminatory outcomes. By learning from historical loan data, ML models may **"consolidate existing bias and prejudice against groups defined by race, sex, [etc.]"**. Prior studies and legal precedents underscore this risk. For instance, Wells Fargo's lending algorithm was shown to give higher risk scores to Black and Latino applicants than to similarly qualified white applicants, causing significantly higher denial rates for minority borrowers. In the U.S., laws like the Equal Credit Opportunity Act (ECOA) and the Fair Housing Act explicitly prohibit using attributes such as race, gender, or ethnicity in lending decisions. These regulations embody the ethical imperatives (reflected in codes like the ACM Code of Ethics) to ensure **algorithmic fairness**. Thus, auditing and mitigating bias in credit scoring models is both a legal and societal necessity. This case study was undertaken to systematically **audit a credit scoring model for fairness**, apply ethical frameworks (e.g. ACM ethics, the concept of "fairness through awareness"), and implement bias mitigation, with the goal of demonstrating that credit AI can be both accurate and equitable.

Dataset Description

A publicly available **Kaggle credit scoring dataset** was used (from adityarajsharma). This dataset contains a range of applicant

information (demographic and financial features) on the order of a few thousand records. Key columns include personal demographics (e.g. gender, marital status, education level, and employment status) as well as financial attributes (e.g. income, debt ratios, and a geographic ZIP code). The raw data did not include a ready-made credit *score* or a binary default label. To facilitate modeling, a **synthetic credit score** was constructed by combining relevant financial indicators into a single numeric index, representing each individual's creditworthiness. A **binary target variable** was then defined by thresholding this score to simulate a loan *approval* (good credit) vs *denial* (poor credit) outcome. The rationale was to create a realistic credit-approval scenario: individuals above a score threshold are labeled "approved," others "denied." Standard preprocessing (e.g. handling missing values, encoding categories) was applied. This transformation yielded a clean dataset with both continuous features (e.g. income) and categorical features (one-hot encoded), and a binary class label for loan outcome. These choices allowed us to apply classification models and fairness metrics meaningfully, while reflecting how an actual scoring system might operate.

Initial Bias Audit

The initial audit established a **baseline model** and measured its fairness. We trained a Random Forest classifier on the raw features to predict the binary target. On this model, standard classification accuracy was high, but fairness analysis revealed serious disparities. In particular, we computed the **Disparate Impact** (DI) ratio (the approval rate for a protected group over that of a reference group). We found DI ≈ 0.55 for Black applicants, well below the 0.8 fairness threshold, indicating that Black borrowers were far less likely to be approved than White borrowers. Further, **SHAP (Shapley Additive Explanation)**

analysis of the Random Forest showed that geographic ZIP code and income were among the top predictive features. This is concerning because ZIP code can act as a proxy for race or socioeconomic status. The combination of a low DI and proxy indicators signaled clear **bias** against minority applicants. Ethical concerns (noted in regulatory guidance) include such proxy bias and unequal denial rates. To address this, we applied two standard fairness mitigation strategies: a pre-processing **reweighing** scheme (assigning sample weights to balance protected groups) and a post-processing label-adjustment (tweaking predictions to equalize opportunities). These yielded trade-offs: accuracy dropped modestly (≈3%), but key fairness metrics (especially equal-opportunity improvements) increased (on the order of +25%). Throughout, we consulted ethical frameworks (e.g. the ACM Code of Ethics and the principle of "fairness through awareness") to guide decisions, ensuring the model's actions aligned with norms against disparate impact. Overall, the work demonstrated how to detect bias (DI and SHAP analyses) and how basic mitigation methods can significantly improve fairness (albeit at a slight cost to accuracy).

Exploratory Analysis

We explored patterns in the data to understand potential sources of bias. The distribution of the synthetic credit scores was roughly **normal with a slight right skew**. Most individuals had moderate-to-good scores, while very high or very low scores were rarer; this indicated a fairly balanced population for analysis. Next, we segmented scores by demographic groups:

- **Gender:** Boxplots revealed *no significant gender difference* in median credit score. Males and females showed nearly identical medians and interquartile ranges, suggesting that in this dataset gender itself was not a strong predictor of score.

- **Marital status:** We grouped applicants by Single/Married/Divorced. Singles had a slightly higher median score, but all groups' score ranges overlapped extensively. Marital status alone did not strongly distinguish creditworthiness.
- **Education level:** Higher education correlated with better scores. Applicants with Master's or PhD degrees tended to have higher median credit scores than those with only high school or bachelor's degrees. This likely reflects higher income potential and financial stability at advanced education levels.
- **Employment status:** Employment status showed a pronounced effect: employed or self-employed borrowers had much higher median scores, whereas unemployed individuals had significantly lower scores. This underscores that a stable income is key to creditworthiness in our data.

Finally, we applied **K-Means clustering** on financial and demographic features. The clustering uncovered distinct customer segments (e.g. "high-risk" vs "low-risk" profiles). Such segmentation suggests that individuals naturally group into different risk cohorts; for example, one cluster included mostly high-income, high-score customers, while another had lower-income, lower-score borrowers. These insights (e.g. the importance of income stability and education) helped in understanding which factors might drive bias and informed our later modeling steps.

Modeling and Fairness Metrics

For this, we developed a logistic regression model and examined its fairness. A **correlation heatmap** was used to select features: highly correlated predictors were removed to reduce multicollinearity and simplify the model. The resulting logistic

model achieved reasonable prediction accuracy. To quantitatively assess fairness, we computed group fairness metrics. In particular, we focused on **Statistical Parity Difference (SPD)**, **Disparate Impact (DI)**, and **Equal Opportunity Difference (EOD)**. SPD is defined as the difference in the rate of positive outcomes between protected and reference groups (a value of 0 indicates parity). DI is the ratio of those rates (fairness ideally at 1). EOD measures the difference in true-positive rates between groups (0 is fair). For the initial logistic model, these metrics revealed imbalances. For example, SPD and EOD were significantly different from zero and DI was well below 1 for the unprivileged group. This confirmed that even a simpler linear model inherited biases from the data. However, quantifying these disparities provided clear targets for mitigation in the next steps.

Explainability and Group-Based Fairness

Next we focused on **model explainability and intersectional fairness**. We applied SHAP to the trained Random Forest model to interpret feature contributions at a granular level. SHAP values distribute a feature's contribution to each prediction, enabling both global and local explanations. The SHAP summary plot highlighted several influential features (e.g. income, existing debts, ZIP code) and confirmed earlier findings: ZIP code emerged as a top predictor, suggesting it encoded sensitive information. We also performed a **gender-based fairness analysis** by calculating approval and true positive rates for subgroups (e.g. female vs. male, intersected with income levels). Textual examination of these group metrics showed continued disparities. For instance, applicants in the "female, low-income" subgroup had a notably lower approval rate and true positive rate than those in "male, high-income," reflecting an intersectional bias. Although plots cannot be shown here, these findings align with numerical fairness scores: for example, the SPD and EOD

between these intersectional groups were far from zero, indicating unfair gaps.

Using SHAP and subgroup analysis surfaced the complexity of bias: it is not always captured by a single protected attribute but can emerge when attributes combine. It also illustrated a challenge: improving fairness at one level (e.g. by race) may not automatically ensure fairness for all subgroups (e.g. low-income women). Moreover, while SHAP enhanced transparency, interpreting the results demanded care. We noted that simply removing protected attributes was insufficient, because other features (like ZIP code) can act as proxies. In summary, explainability work confirmed the presence of gender- and income-based disparities and underscored the difficulty of achieving intersectional fairness.

Fairness Mitigation

We implemented a systematic bias mitigation strategy. We used **Fairlearn's GridSearch** algorithm with a **Demographic Parity constraint**, treating the **combination of gender and income** as a composite sensitive attribute. This approach finds a model that minimizes classification error while enforcing parity in positive outcomes across groups. The **baseline model (pre-mitigation)** was re-evaluated: it had high overall accuracy, but fairness metrics remained poor (DI < 0.8, SPD < 0, EOD < 0 for the disadvantaged group). In other words, it still produced systematically lower approval and true-positive rates for protected subgroups.

After applying the Fairlearn mitigation, the **mitigated model** had only a slight drop in accuracy, but major fairness gains: DI moved close to 1.0, SPD approached 0, and the EOD gap shrank dramatically. In practical terms, approval rates and true-positive rates across the gender–income groups became more balanced.

We observed the expected trade-off: a minor reduction in predictive accuracy yielded a substantial reduction in disparity. Key takeaways from this phase were clear: bias *can* be measured and significantly reduced through fairness-aware training, and the accuracy–fairness trade-off can be acceptable (a small accuracy sacrifice for large fairness improvements). The end result was a model that produces **fairer credit decisions across demographic groups** while remaining practically useful. Importantly, our mitigated model achieved equitable approval rates and true positive rates among groups, validating the approach. As one slide summarized, *"Accuracy makes a model useful, but fairness makes it trustworthy"*.

Discussion

This case study yielded several important lessons. First, **bias propagates easily**: historical and proxy variables (like ZIP code) can embed discriminatory patterns into models unless checked. Our audits quantitatively demonstrated that disparity metrics often diverge substantially from ideal parity, reflecting underlying societal inequalities. Second, **mitigation strategies work but incur trade-offs**. The reweighing and post-processing, and the fairness-constrained training, all improved group equity at some cost to accuracy. This trade-off is intrinsic: fairness constraints can slightly reduce prediction accuracy, echoing the observation that *"a small accuracy drop for significant fairness gain"* is a typical outcome. Third, **the choice of fairness metric matters**. We found that optimizing for Demographic Parity (equal positive rates) greatly improved SPD and DI, but we also monitored Equal Opportunity (TPR gap). No single metric captures all aspects, and improving one criterion may leave others imperfect.

Finally, **real-world implications** are profound. Biased credit algorithms can perpetuate historical injustices, affecting not just

loan approvals but many downstream outcomes. Credit scores influence employment decisions, insurance rates, housing, and more. Thus, fairness in credit models has broad social impact. Our case study shows that ethical AI practices-systematic auditing, transparent explanation, and targeted mitigation can substantially reduce unwanted bias. In doing so, the model not only meets legal/ethical standards (e.g. ECOA) but also improves trust among stakeholders. We demonstrated that **performance and fairness can coexist**: with careful methodology, it is possible to achieve a model that is both accurate and equitable.

Conclusion

In conclusion, this multi-week case study highlights the critical importance and feasibility of **fair AI in credit scoring**. Starting with a standard random forest model, we detected severe bias against minority and low-income groups, driven in part by proxy variables like ZIP code. Through exploratory data analysis and explainability (SHAP) we gained insights into the root causes of bias. We then applied fairness-aware techniques (reweighing, post-processing, and optimized training with parity constraints) to correct these issues. The result was a mitigated model that delivered balanced approval rates and equal opportunity across gender–income groups, at the cost of only a slight accuracy reduction. This outcome underscores that enforcing fairness can indeed coexist with maintaining utility. Overall, the case study achieved its goals: it quantified bias with concrete metrics, applied recognized ethical frameworks, and demonstrated effective mitigation. As we reflect on the work, the key takeaway is that in algorithmic lending, **fairness is as essential as accuracy**. Building and deploying credit models demands continual ethical scrutiny; by embedding fairness analysis into the ML pipeline, practitioners can create credit systems that are not just powerful, but also just and trustworthy for all communities.