Default of Credit Card Interpretability Using SHAP

Shixin Zhang
Szhang655@wisc.edu

Problem Formulation

As I will be working in the banking industry upon graduation, I chose the Default of Credit Card Clients Dataset from Kaggle for this analysis. Understanding and predicting credit card defaults is crucial for banking and financial institutions to mitigate risks and make more informed lending decisions. This dataset contains information on several payment information and demographic factors of credit card holders in Taiwan from April 2005 to September 2005.

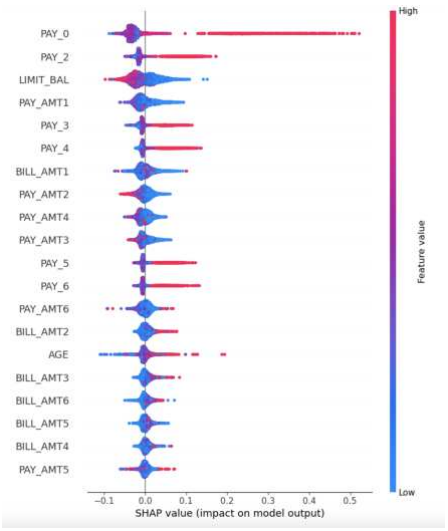
Interpretability in credit risk area is important because banks must ensure fairness and comply with regulatory requirements while building trust with customers. This exercise aims to enhance model interpretability using SHAP to understand which factors influence default predictions and how factors are interacted with each other. Financial institutions, customers, and regulators are likely to be benefiting from an effective interpretation in this area for them to assess risks, design credit scoring model, and ensure transparency and fairness in lending.

Method Application

The dataset has 30,000 entries and 25 variables. The target variable is ‘default’ which indicates whether the customer defaults on their next payment. The ‘ID’ column is dropped from the features, resulting in 23 input features: 'LIMIT_BAL', 'SEX', 'EDUCATION', 'MARRIAGE', 'AGE', 'PAY_0', 'PAY_2', 'PAY_3', 'PAY_4', 'PAY_5', 'PAY_6', 'BILL_AMT1', 'BILL_AMT2', 'BILL_AMT3', 'BILL_AMT4', 'BILL_AMT5', 'BILL_AMT6', 'PAY_A'. Before fitting the model, the data is preprocessed by standard scaling the numerical features. Twenty percent of the data is then taken out from the dataset for testing. Random Forest Classifier is then chosen because of its robustness in handling structured financial dataset. The model achieved 81.6% accuracy.

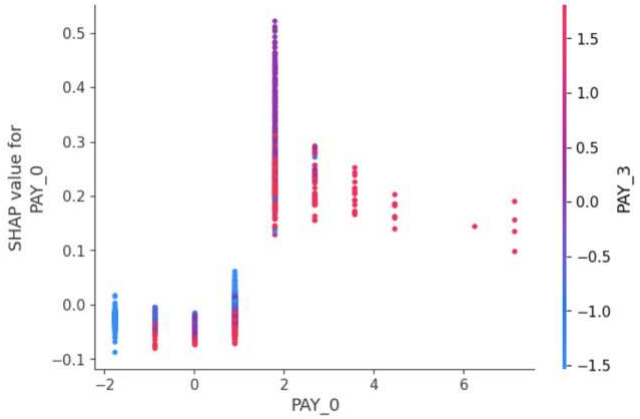
Discussion

SHAP is then applied using TreeExplainer to explain individual predictions and global feature importance.

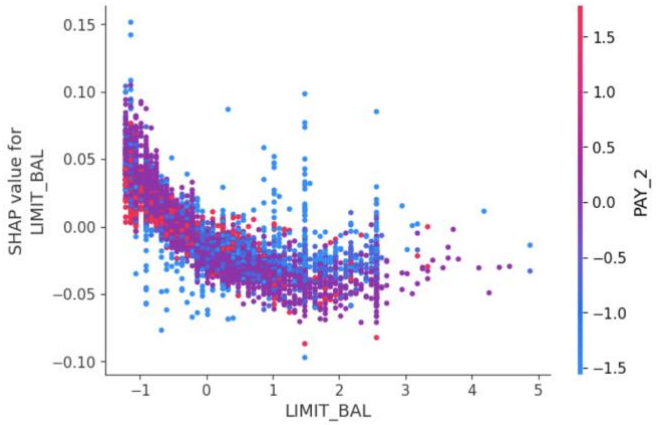


The SHAP summary plot indicates that PAY_0 which is the payment status in the most recent month, PAY_2 which is the payment status two months ago, and LIMIT_BAL which is the amount of given credit were the most influential features in determining default. This aligns with financial intuition that past repayment is a strong indicator of future default as higher values of PAY_0 and PAY_2 significantly increases the likelihood of default from the summary plot. In addition, higher credit limits is associated with lower risk.

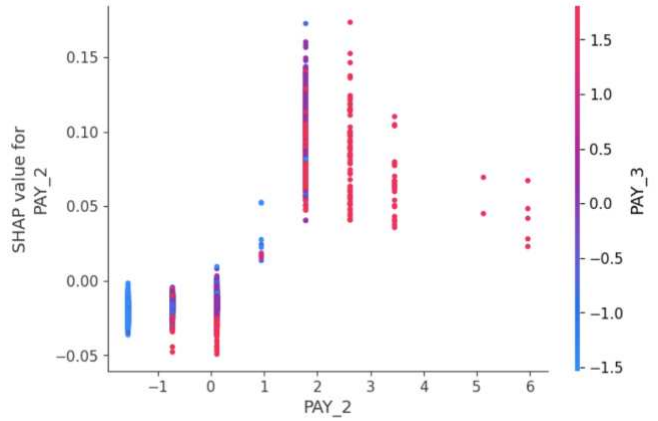
SHAP dependence plots for the top three features illustrates how key features interact.



PAY_0 dependence plot above shows that as payment delays increase, risk of default also increases. The interaction with PAY_3 shows that customers who were previously delinquent for three months has greater default probability when their most recent payments were late.



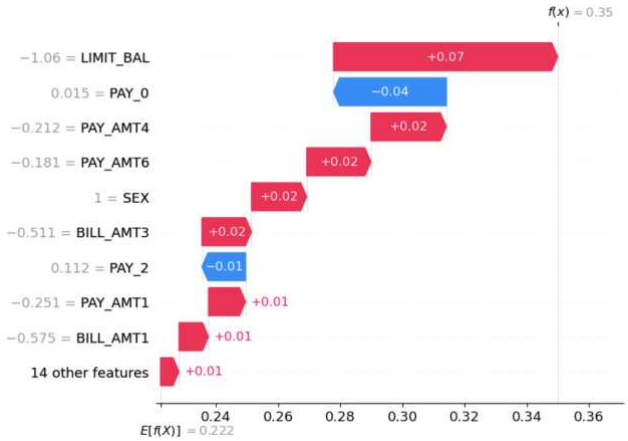
LIMIT_BAL dependence plot above shows that customers with lower credit limits have a higher risk of default.



PAY_2 dependence plot shows a similar trend as PAY_0, indicating the significance of short-term repayment history.

Force plots and waterfall plots are used to examine individual predictions to gain a more granular understanding.

The SHAP force plot was generated for an individual sample showing how different features contribute the prediction towards or away from the default. Recent payment status push the prediction towards higher default risk. While credit limit and further amount payments inversely contribute to default risk.



The waterfall plot breaks down feature contributions quantitatively. LIMIT_BAL(-1.06), PAY_0(0.015), PAY_AMT4(-0.212), and PAY_AMT6(-0.181) were the dominant factors, explaining how the model arrived at its decision.

Overall, short-term repayment history is the most important factor in determining default risk. Credit limit given inversely influences default probability. Demographic features have minimal impact. This makes sense as financial behavior is more impactful than demographic factors.

SHAP provides both global and local explanations. Globally, SHAP provides financial institutions ability to understand which factors contribute most to risk assessment through feature importance analysis. Locally, SHAP gives explanations to individual predictions so that specific lending decisions are transparent. Insights from SHAP aligns with existing financial intuition which increases trust in model's decisions. Interpretability is crucial in banking where understanding factors to default risk will lead to better intervention techniques.

SHAP also comes with limitations. Calculating SHAP values can be computationally extensive. For this relatively small dataset covering only few months of transaction data, the computation took about an hour. However, datasets are usually much larger covering years of transaction history in real-world setting. Additionally, further validation through external datasets or real-world application is necessary to confirm robustness and generalizability.

Additional testing on alternative models such as XGBoost or Logistic Regression could help to compare the interpretability and predictive results. Incorporating with other interpretability models like LIME could provide complementary or comparative perspectives. Evaluating fairness of machine models is also important to help identify any potential biases in the model so that model decisions do not discriminate certain groups.

The findings provide insights from model decision-making for financial institutions to refine lending policies and improve transparency. This approach enhances trust in automated decision-making which drives digital transformation in banking sector. As I start my career in banking, these techniques will be invaluable in developing digital solutions that balance profitability, risk and fairness.

Interpretable ML Homework 1

Elaine Chiu

February 2025

1 Motivation

Research shows that how well kids do in math and reading early on can really affect their future academic success. Because of this, there's been a lot of attention on the factors that influence a child's learning. For instance, Darling-Hammond (2000) points out that things like teacher qualifications, course design, and class size have a clear impact on learning. But we also think that a child's socioemotional skills and ability to focus might be just as important. This analysis aims to look into how specific cognitive and behavioral skills relate to academic achievement.

2 Data Explanation

This analysis uses data from the Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), which was collected by the National Center for Education Statistics (NCES). It followed 21,260 children who started kindergarten in the 1998–99 school year all the way through to the spring of 2007, by which time most were in eighth grade. The key variables used in this study are as follows:

- *math.irt*: The math item response theory (IRT) score when most children were in kindergarten. Scores were computed based on the pattern of correct and incorrect

answers to estimate a child's ability parameter (θ), ranging from 10 to 117.

- *motor*: The general motor skill score, ranging from 0 to 17.
- *attention*: The child's attention level, measured on a scale from 1 (unable to attend) to 5 (complete and full attention).
- *income*: The weekly household income, ranging from \$0 to \$1,000,000.
- *sibling*: The number of siblings within the household.

3 Methods

We hypothesize that higher motor and attention skills contribute positively to academic achievement. Additionally, family income serves as a proxy for family resources and is expected to be positively correlated with academic performance. Prior literature suggests that a greater number of siblings may promote literacy development through increased family interactions, potentially leading to a positive correlation with math scores.

To examine these relationships, we fit a linear regression model and compute SHAP values. The regression equation is formulated as follows:

$$math.irt = \beta_0 + \beta_1 motor + \beta_2 attention + \beta_3 income + \beta_4 sibling.$$

Table 1 presents the estimated regression coefficients along with their corresponding t-values.

All independent variables, except for *sibling*, are statistically significant and positively correlated with math scores. We now analyze the SHAP values to further interpret the model.

Variable	Estimate	t-value
Intercept	5.022	11.296
Motor	0.879	42.380
Income	0.392	25.954
Attention	0.244	25.226
Sibling	-0.0547	-7.466

Table 1: Regression coefficients and their corresponding t-values.

4 Results and SHAP Analysis

The running of SHAP does not cost more than a minute as I only consider a few features.

4.1 Interpreting the Global SHAP Plot

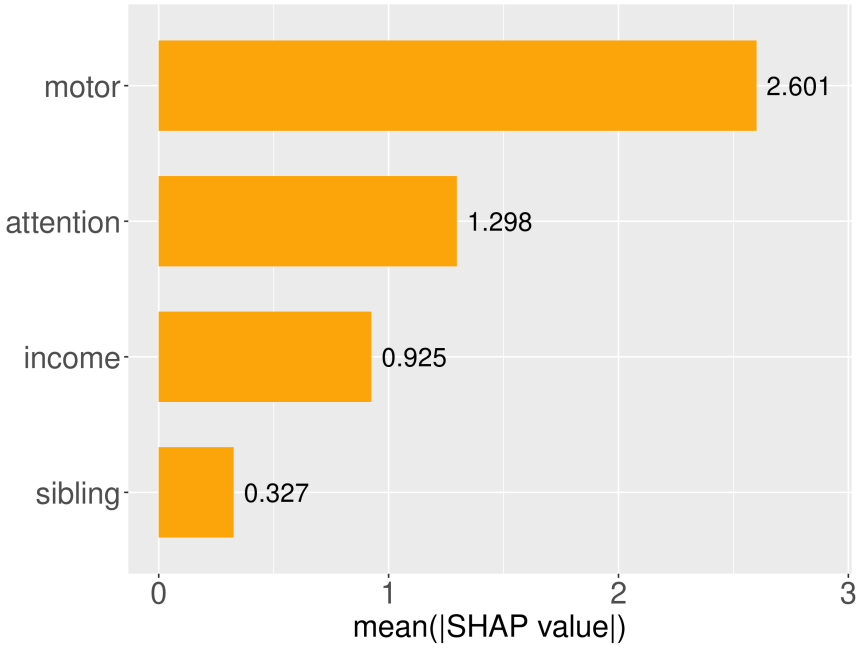


Figure 1: Global SHAP Summary Plot.

The global SHAP summary plot in Figure 1 illustrates the overall importance of features in predicting math scores. Features are ranked by their average absolute SHAP value, which represents their contribution to model predictions. The motor skill and income variables have the highest SHAP values, indicating they are the strongest predictors. Attention also

plays a meaningful role, while the number of siblings shows a small but negative effect.

We may connect the results with previous studies. First, for the negative association between the number of siblings and the academic achievement, Downey (2001) proposed the "resource dilution hypothesis", which argues that as the number of siblings increases, parental resources—such as time, attention, and financial support—are divided among children, leading to lower intellectual development. Secondly, for the positive association between attention level and math achievement, this finding is aligned with previous literature. For example, Anobile et al. (2013) discovered that children with higher sustained visual attention tend to perform better in mathematics. Thirdly, for the positive association between motor skill and math achievement. Cameron et al. (2016) mentioned that the fine motor skills, which involve coordinating small muscles, directly linked to how well a kid can manage self-care, writing, and manipulating small objects. A kid with a lower fine motor skill will find the school more challenging.

4.2 Interpreting the Individual SHAP Plot

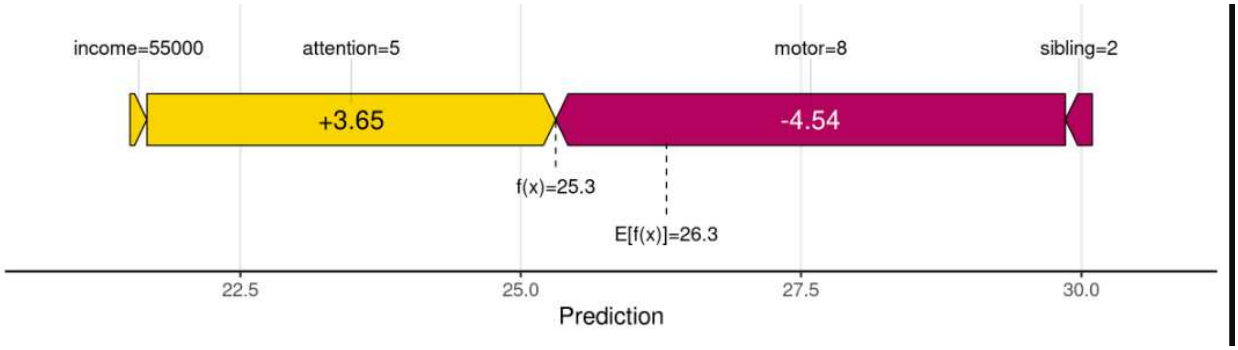


Figure 2: A sample individual SHAP plot.

The individual SHAP plot in Figure 2 explains how each feature contributes to a specific student's predicted math score. Positive SHAP values indicate features that increase the prediction, while negative values indicate features that lower the prediction. For instance, a student with high motor skills and attention might receive a higher predicted math score due to strong positive contributions from these features. In contrast, if a student comes from

a low-income household with many siblings, these factors may contribute negatively to their predicted score.

5 Discussion: Who Benefits from Interpretability

SHAP values enhance interpretability by providing both global and local insights into a model's predictions. While regression coefficients indicate the average effect of each feature across all data points (thus in this sense, always global), SHAP values explain individual predictions, illustrating how each feature contributes to specific outcomes. And I think the "local" interpretability is particularly beneficial for education studies, which discuss the heterogeneous effect pretty often because a particular intervention or even test assessment usually has different effect on kids with varying ability.

- **Educators:** Teachers can identify students who may need additional support, using SHAP insights to tailor learning strategies. See, for example, Liu et al. (2017).
- **Policymakers:** Decision-makers can design data-driven policies that prioritize interventions in areas with the most impact.
- **Parents:** Families can better understand factors influencing their child's academic performance and advocate for needed resources.

Finally, although I believe that the use of SHAP value assists the interpretation of models, it is worth exploring if a causal interpretation can come out of it. For example, Heskes et al. (2020) considers the combination of the SHAP value with the do-calculus and aims to detect the direct and indirect effects among variables.

Intepretable AI and Education Data Code Appendix

Elaine Chiu

2025-02-10

Load the libraries and data cleaning

```
# 1) Install and load shapr
# install.packages("shapr")
library(shapr)

# install.packages("shapviz")
library(shapviz)

library(ggplot2)



# 2) load the data
data.t <- read.csv("education.data.csv")

data.t$math.irt = data.t$C1R4MSCL
data.t$motor = data.t$C1CMOTOR
data.t$income = data.t$WKINCOME
data.t$attention = data.t$C6ATTLVL
data.t$sibling = data.t$P1NUMSIB


## only keep the data with variables we want to discuss, also filter out kids without data
## in the last round, we don't talk about the missing mechanism of this longitudinal data set
analysis.data = data.t[,c("math.irt","motor","income","attention","sibling")]
analysis.data$sample = complete.cases(analysis.data)
analysis.data.complete = subset(analysis.data, sample==TRUE)
```

compute some summary statistic on the independent variables

```
summary(analysis.data.complete$math.irt)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   -9.00   19.86   25.19   26.30   31.32   93.23
```

```
summary(analysis.data.complete$motor)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    -9.0    10.0    13.0    11.9    15.0    17.0
```

```r
summary(analysis.data.complete$attention)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   3.000   4.000   3.865   5.000   5.000
```

```r
summary(analysis.data.complete$sibling)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   1.000   1.000   1.472   2.000  11.000
```

```r
summary(analysis.data.complete$income)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0   24000   45000   56074   70000 1000000
```

Run a linear regression model

```r
model <- lm(
  math.irt ~ motor+income+attention+sibling,
  data   = analysis.data.complete)
summary(model)
```

```
##
## Call:
## lm(formula = math.irt ~ motor + income + attention + sibling,
##     data = analysis.data.complete)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -37.507  -5.288  -1.138   3.916  71.141
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.022e+00  4.446e-01  11.296  < 2e-16 ***
## motor        8.792e-01  2.074e-02  42.380  < 2e-16 ***
## income       3.919e-05  1.510e-06  25.954  < 2e-16 ***
## attention    2.439e+00  9.668e-02  25.226  < 2e-16 ***
## sibling     -5.467e-01  7.323e-02  -7.466 9.01e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.354 on 9639 degrees of freedom
## Multiple R-squared:  0.3001, Adjusted R-squared:  0.2998
## F-statistic:  1033 on 4 and 9639 DF,  p-value: < 2.2e-16
```

Compute the SHAP values

```r
# 'analysis.data.complete' is your final dataset after filtering.

# Create a shorter reference
df <- analysis.data.complete

# The features
```

```r
x <- df[, c("income","motor","sibling","attention")]

# The outcome is
y <- df$math.irt



# We'll use the entire feature set as the "background" distribution
bg_x <- x

# Suppose we want to explain the model's predictions for the first 10 rows
explain_x <- x[c(1,7,9,2200,3000,4000),]


baseline_value <- mean(y)

# Empirical approach
explain1 <- explain(
  model = model,
  x_explain = explain_x,
  x_train = bg_x,
  approach = "empirical",
  phi0 = mean(y),
  n_MC_samples = 1e2
)
```

```
## Success with message:
## max_n_coalitions is NULL or larger than or 2^n_features = 16,
## and is therefore set to 2^n_features = 16.

##

## -- Starting `shapr::explain()` at 2025-02-10 22:44:00 --------------------------

## * Model class: <lm>

## * Approach: empirical

## * Iterative estimation: FALSE

## * Number of feature-wise Shapley values: 4

## * Number of observations to explain: 6

## * Computations (temporary) saved at:
## '/tmp/Rtmpx61OhQ/shapr_obj_2b840923ed9.rds'

##

## -- Main computation started --

##

## i Using 16 of 16 coalitions.
```

Visualization - the global importance of the variables

```r
shp <- shapviz(explain1)
```

```
global_graph = sv_importance(shp, show_numbers = TRUE, number_size = 6)

# Assuming `shap_viz` is already created
global_graph <- global_graph+
  theme(
    text = element_text(size = 20),        # Increases all text size
    axis.title = element_text(size = 20),  # Increases axis titles
    axis.text = element_text(size = 20),   # Increases axis labels
    legend.text = element_text(size = 20), # Increases legend text
    legend.title = element_text(size = 20) # Increases legend title
  )




# Save the plot
ggsave("global_graph.png", plot = global_graph, width = 8, height = 6, dpi = 300)
```

The visualization of each single observation

```
explain_x[1,]
```

```
##   income motor sibling attention
## 2 120000    14       2         5
```

```
id_1.graph = sv_force(shp,row_id = 1)
ggsave("id_1.graph.png", plot = id_1.graph, width = 8, height = 6, dpi = 300)


explain_x[2,]
```

```
##    income motor sibling attention
## 21  50000    10       1         4
```

```
id_2.graph = sv_force(shp,row_id = 2)
# Assuming `shap_viz` is already created
id_2.graph <- id_2.graph+
  theme(
    text = element_text(size = 20),        # Increases all text size
    axis.title = element_text(size = 20),  # Increases axis titles
    axis.text = element_text(size = 20),   # Increases axis labels
    legend.text = element_text(size = 20), # Increases legend text
    legend.title = element_text(size = 20) # Increases legend title
  )

ggsave("id_2.graph.png", plot = id_2.graph, width = 8, height = 6, dpi = 300)


explain_x[3,]
```

```
##    income motor sibling attention
## 31  55000     8       2         5
```

```
id_3.graph = sv_force(shp,row_id = 3)
ggsave("id_3.graph.png", plot = id_3.graph, width = 8, height = 6, dpi = 300)
```

References

Anobile, G., Stievano, P., and Burr, D. C. (2013). Visual sustained attention and numerosity sensitivity correlate with math achievement in children. *Journal of Experimental Child Psychology*, 116(2):234–246.

Cameron, C. E., Cottone, E. A., Murrah, W. M., and Grissmer, D. W. (2016). How are motor skills linked to children's school performance and academic achievement? *Child Development Perspectives*, 10(2):93–98.

Darling-Hammond, L. (2000). Teacher quality and student achievement. *Education Policy Analysis Archives*, 8(1):1–44.

Downey, D. B. (2001). Number of siblings and intellectual development: The resource dilution explanation. *American Psychologist*, 56(6-7):497–504.

Heskes, T., Sijben, E., Bucur, I. G., and Claassen, T. (2020). Causal shapley values: Exploiting causal knowledge to explain individual predictions of complex models. *arXiv preprint arXiv:2011.01625*.

Liu, M., McKelroy, E., Corliss, S. B., and Carrigan, J. (2017). Investigating the effect of an adaptive learning intervention on students' learning. *Educational Technology Research and Development*, 65:1605–1625.

Finding Evidence of ML biases from the SHAP Value

Eunjee Kim

February 13, 2025

1 Problem Formulation

The purpose of this project is to examine whether machine learning (ML) algorithms exhibit biases regarding sensitive features in training data. Among various types of biases, this study focuses on statistical bias using explainable ML techniques. Statistical bias, in this context, refers to the biases introduced by the characteristics of the training data, particularly the gender distribution. In other words, this study aims to detect statistical bias that arises due to gender imbalance in the training data, utilizing modern interpretation methods in ML.

2 Methodology

We use a publicly available student performance dataset from two Portuguese schools, sourced from the UC Irvine Machine Learning Repository (Dua and Graff 2019). The dataset includes various student attributes, and the goal is to build a performance predictor using multiple ML models. The best-performing algorithm is selected for further analysis.

Since the mean and variance of the target variable (Portuguese scores) differed by gender, we standardized the target variable within each gender group to remove these effects. Specifically, each score was adjusted by subtracting the mean and dividing by the standard deviation within the corresponding gender group.

Using the selected ML model, we computed SHapley Additive exPlanations (SHAP) values to quantify the contribution of the sensitive feature "sex" to the prediction outcome. We observed differences in SHAP values across genders. To further investigate this discrepancy, we manipulated the female-to-male ratio through Monte Carlo resampling and trained the ML model on these resampled datasets. The female-to-male ratio was varied systematically from 0.1 to 10 while keeping the total sample size constant. This approach allows us to attribute any changes in SHAP values to the gender ratio in training data rather than other confounding factors. For robustness, we repeated the SHAP value estimation 20 times for each female-to-male ratio.

3 Discussion

Examining the summary of SHAP values across all features, "sex" does not appear to be a key predictor in the model (Figure 1). However, when plotting the SHAP value of "sex"

across gender groups (Figure 2), we observed systematic differences. Female SHAP values were generally negative, whereas male SHAP values were more centered around zero.

To determine whether this effect is due to the initial gender imbalance in the dataset (266 males vs. 383 females), we conducted Monte Carlo simulations (Figure 3). Two interesting trends emerged:

Effect of Gender Representation on SHAP Values:
The absolute SHAP values increased as the gender ratio became more imbalanced (i.e., approaching 0.1 or 10). This suggests that the underrepresentation of a group leads to a higher SHAP magnitude, meaning the model assigns greater importance to that feature when it is underrepresented. This observation is somewhat counterintuitive, as one might expect the opposite effect. The mechanism behind this remains an interesting topic for further study.

Systematic Differences in SHAP Values:
Although the target variable was standardized by gender, the SHAP values suggest that being male gives a slight advantage in performance prediction. This raises the question of whether SHAP values are fully reliable in such cases or whether alternative explainability methods should be explored.

To further investigate, we examined the built-in feature importance measure provided by the Gradient Boosting method (Figure 4). However, this measure did not provide much additional insight. One limitation of the feature importance function in Gradient Boosting is that it does not specify the sign of the feature importance, which may explain why it offers even less information than SHAP values.

4 Conclusion

This study provides empirical evidence that the representation of gender in training data influences how ML models assign importance to sensitive features. The results indicate that underrepresented groups receive greater absolute SHAP values, implying a form of statistical bias. Moreover, the systematic advantage of one gender despite target standardization suggests potential limitations of SHAP values as an interpretation tool. Future research could explore alternative explainability methods or test different ML models to generalize these findings.

References

Dua D, Graff C (2019) Uci machine learning repository. URL `http://archive.ics.uci.edu/ml`.

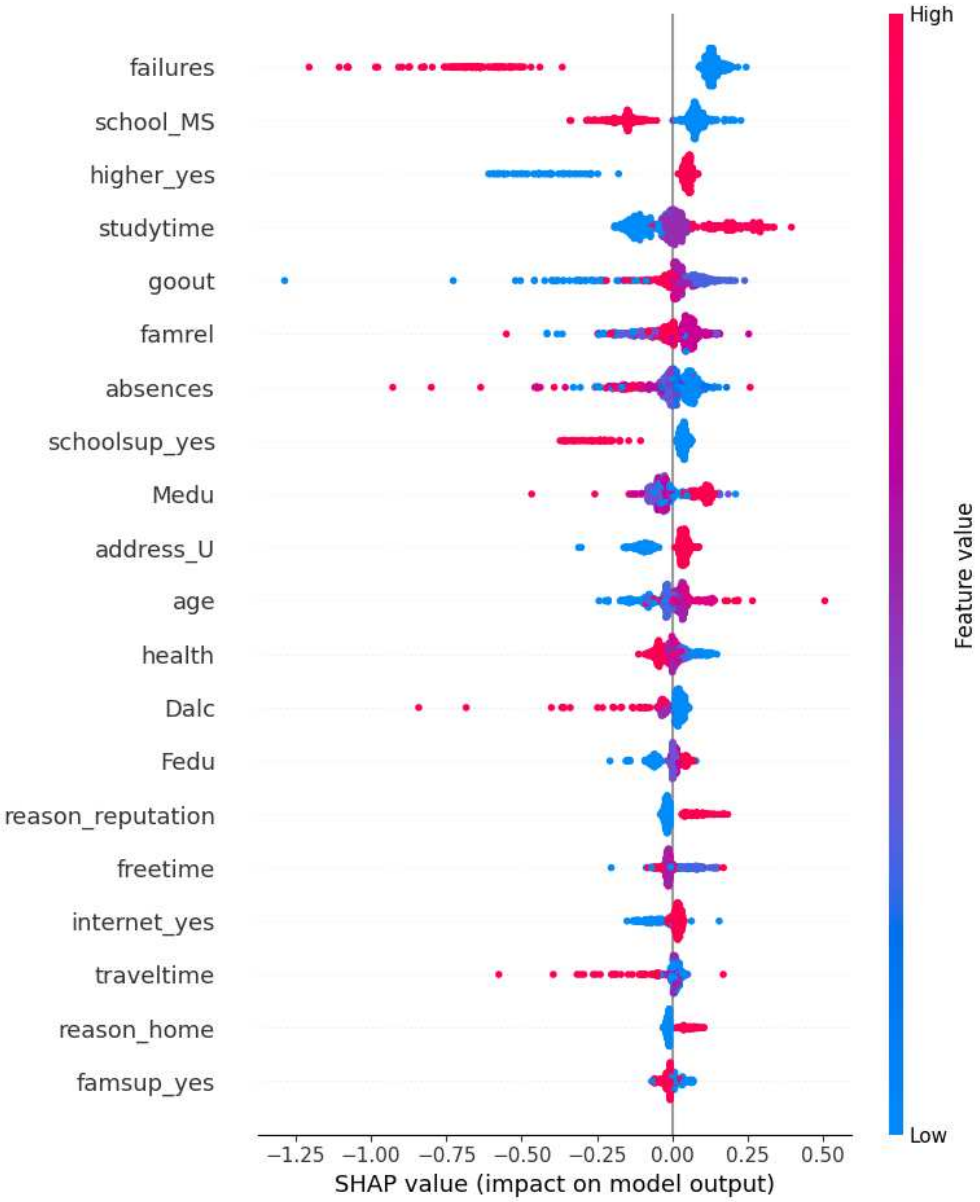


Figure 1: Summary Plot (Beeswarm Plots) for SHAP feature importance

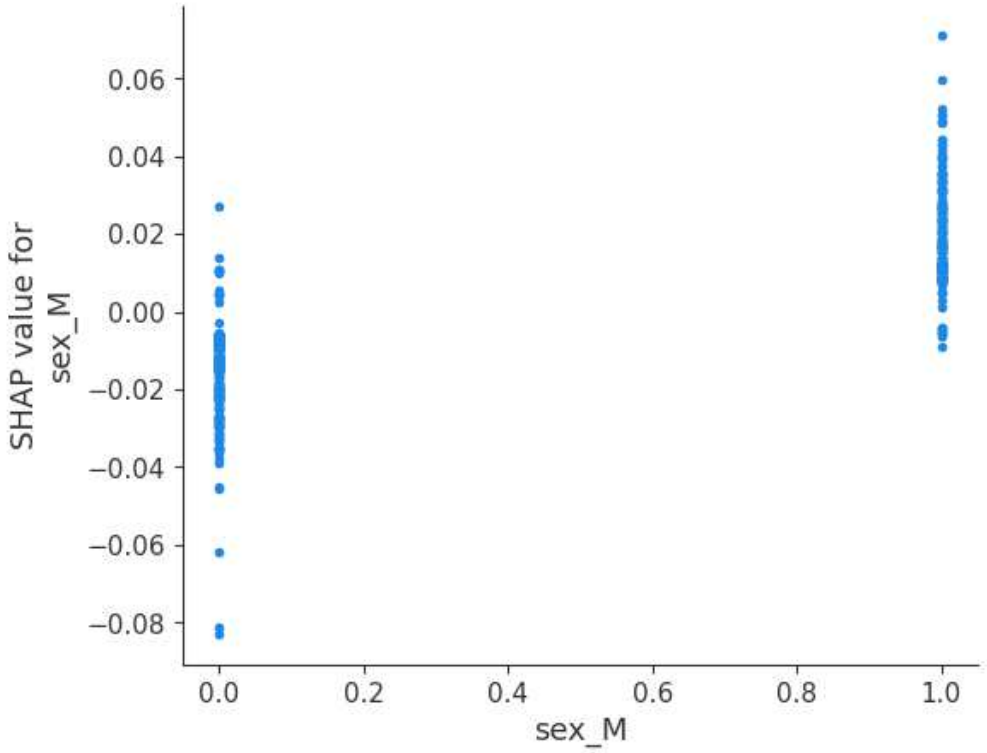


Figure 2: SHAP feature importance of sex by sex groups

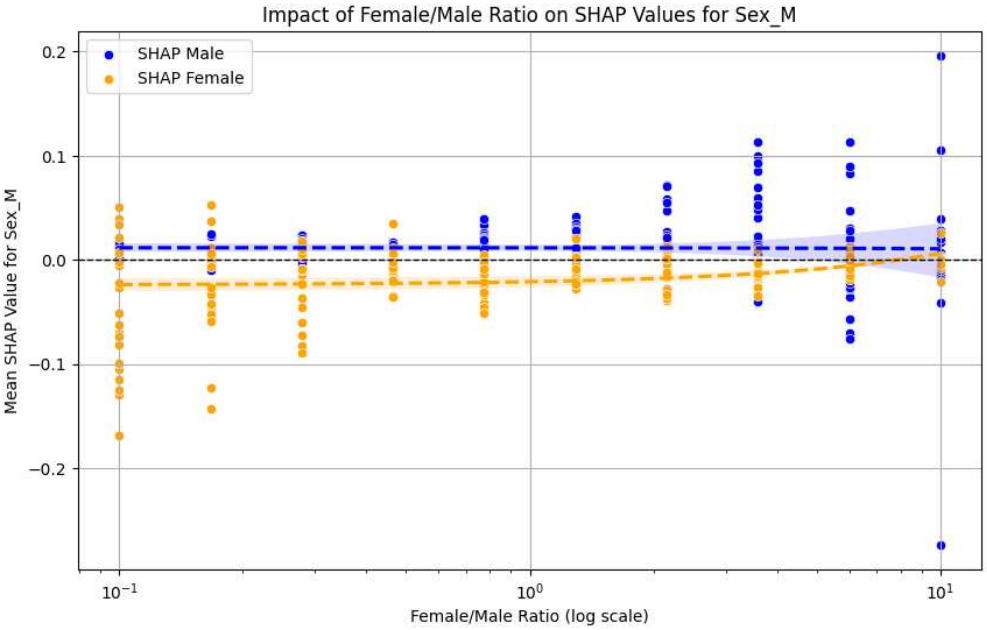


Figure 3: SHAP feature importance vs. female to male ratio

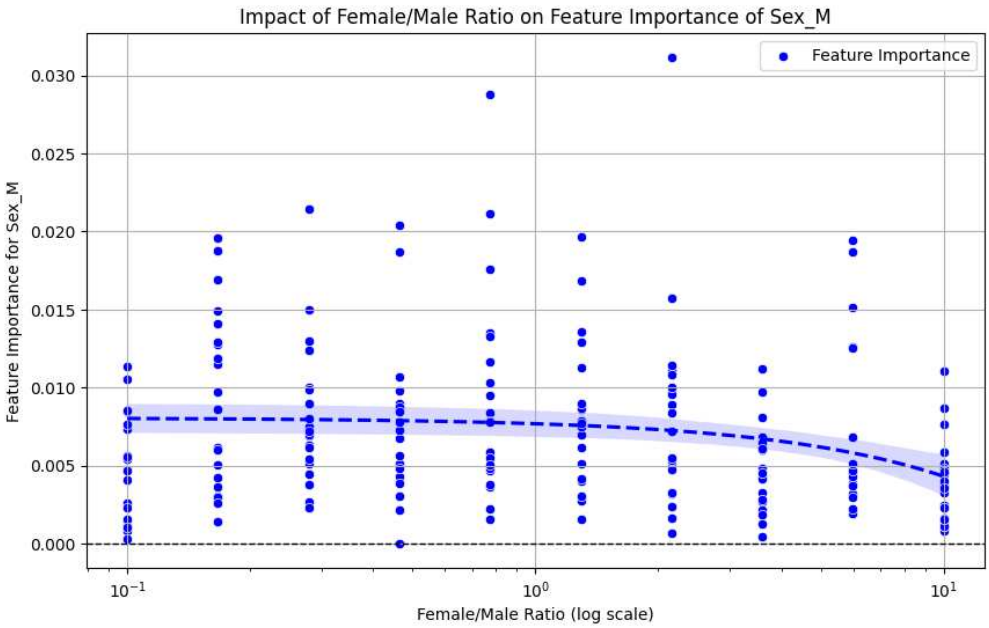


Figure 4: Built-in feature importance of Gradient Boosting

Intepretability of the LightGCN Recommender System on MovieLens Dataset

by Langtian Ma

February 13, 2025

1 Introduction

Recommender systems have become indispensable in digital platforms. However, a growing number of modern recommender systems relies on complex models like deep learning and graph neural networks. Their decision-making processes are often opaque. This lack of transparency undermines user trust, debugging [2], and raises ethical concerns, such as bias and fairness [5]. The ability to explain why recommendatons are generated has thus become a critical research area. This report implements a slightly modified version of LightGCN [4], trains it on the MovieLens 1M dataset [3] and attributes feature importance for specific predictions using the Integrated Gradients method [6]. The approach successfully identifies key features of both users and movies that inluence the predictions.

2 Problem Formulation

Understanding which features most influnce recommendation decisions is crucial. In this report, we trained a modified LightGCN model with additional user features and item features informations on the **MovieLens 1M dataset** [3]. The MovieLens 1M dataset is a widely used benchmark dataset for evaluating recommender systems. It contains **1 millioin ratings** from **6,000 users** on **4000 movies**, collected by the GroupLens research team. Each rating is an integer between 1 (worst) and 5 (best). In addition to the user-movie interactions, the dataset provides user demographic informtion (age, gender, occupation, and zip code) and moview metadata (title and genres). We try to answer key questions such as:

- What user features most affect their recomendations?
- Which movie genres paly the most significant role in recoomendation scores?

The primary beneficiaries of this analysis include:

- End-users, who can better understand why a moview was recommended.
- Developers and researchers, who can debug and improve the model more eaily.

To investigate interpretability, we apply the Integrated Gradients method [6], which attributes importance scores to input features. This allows us to quantify how much different user and movie characteristics contribute to a specific recommendation.

3 Methods Application

3.1 Training LightGCN

LightGCN (Lightweight Graph Convolutional Network) is a simplified yet effective GCN-based model for collaborate filtering on user-item graphs [4]. We use an enhanced LightGCN which incorperate user and item side features. Let U be the number of users, M be the numebr of items, d be the dimension of embeddings, d_u be the dimension of user featrues, and d_m be the dimension of item features. The model is formulated by:

1. Trainbale Embeddings:
    - $\boldsymbol{P_u} \in \mathbb{R}^{U \times d}$ for users.

- $\boldsymbol{P_m} \in \mathbb{R}^{M \times d}$ for items.

2. Feature Projections:

   - User features $X_u \in \mathbb{R}^{U \times d_u}$ are passed through a linear larayer $\boldsymbol{W_u} \in \mathbb{R}^{d_u \times d}$.

   - Item features $X_m \in \mathbb{R}^{M \times d_m}$ are passed through $\boldsymbol{W_u} \in \mathbb{R}^{d_m \times d}$.

The initial embeddings for user u and item m are:

$$\boldsymbol{E_u^{(0)}} = \boldsymbol{P_u^{(0)}} + \boldsymbol{X_u W_u^{(0)}}, \quad \boldsymbol{E_m^{(0)}} = \boldsymbol{P_m^{(0)}} + \boldsymbol{X_m W_m^{(0)}}.$$

These are concatenated into a single matrix $\boldsymbol{E^{(0)}} = \left(\boldsymbol{E_u^{(0)^T}}, \boldsymbol{E_m^{(0)^T}}\right)^T \in \mathbb{R}^{(U+M) \times d}$. Let $\boldsymbol{A} \in \mathbb{R}^{(U+I) \times (U+I)}$ be the normalized user-item interaction matrix where edges encode interactions. The embeddings are propagated through $\boldsymbol{A}$:

$$\boldsymbol{E}^{(k+1)} = \boldsymbol{A E}^{(k)}.$$

After n steps, we average the embeddings $(\boldsymbol{E}^{(0)}, \boldsymbol{E}^{(1)}, \dots)$ to form the final node representation:

$$\boldsymbol{E}^* = \frac{1}{n+1} \sum_{k=1}^{n} \boldsymbol{E}^{(k)}.$$

Then we split $\boldsymbol{E^*}$ to $\boldsymbol{E_u^*} \in \mathbb{R}^{U \times d}$ and $\boldsymbol{E_m^*} \in \mathbb{R}^{M \times d}$. For a user-item pair (u, m), the predicted rating is computed via dot product:

$$\hat{y}_{u,i} = \boldsymbol{E_u^*}(u) \cdot \boldsymbol{E_m^*}(m).$$

A BCEWithLogitsLoss over each observed rating $r_{u,m}$ is used to trian the model:

$$\mathcal{L} = \sum_{(u,m) \in \text{data}} \text{BCE}(\hat{y}_{u,m}, r_{u,m}).$$

3.2 Applying Integrated Gradients

The Integrated Gradients (IG) method is an attribution method used to explain the prediction of neural networks by computing the importance of each input feature. Given a model f and an input x, the importance of each feature x_i is computed by:

$$\text{IG}_i(x) = (x_i - x_i') \int_{\alpha=0}^{1} \frac{\partial f(x' + \alpha(x - x'))}{\partial x_i} d\alpha,$$

where x' is a baseline input.

In our approach, we combine user features and item features to form the input x, using the averaged feature values as the baseline input x'. We then apply Integrated Gradients (IG) to attribute the predicted scores for specific movies. The IG method for the recommender system takes approximately 40 seconds to run on a MacBook Air with an M1 chip.

We select a young student (age ≤ 18) and find that the system predicts *Toy Story 2* as his most favored movie. We then run the IG method to determine the contribution of different features to this prediction. As shown in Figure 1, IG provides a clear explanation of why the recommender system suggests a particular movie by identifying the most influential user and movie features in the prediction.

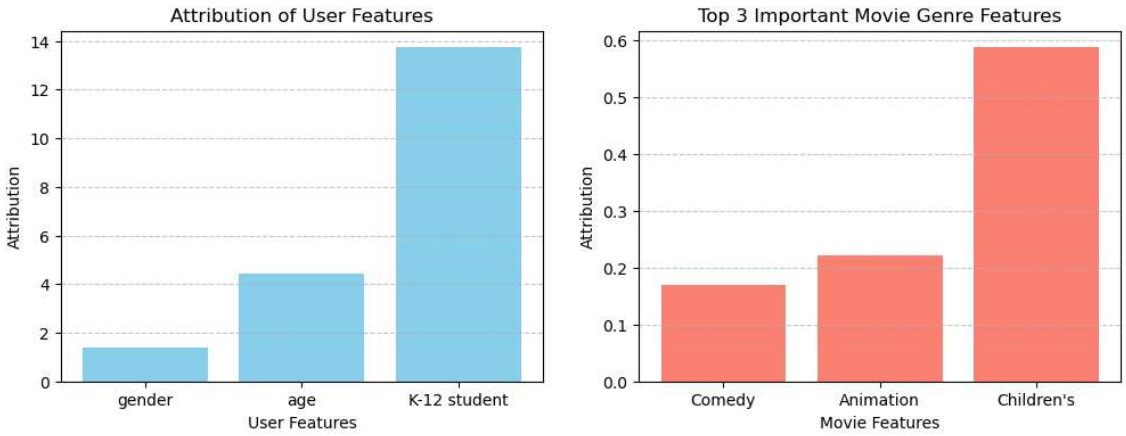


Figure 1. Feature importance of the predicted score that a teenager student would give to the movie "Toy Storey".

The attribution method can also help identify whether a user prefers certain genres. We selected a writer aged 25–30 and analyzed three movies, each including **Fantasy** as one of their genres. As shown in Figure 2, the attribution score for *Fantasy* is significantly higher than that of other genres, suggesting that the user has a strong preference for Fantasy movies.

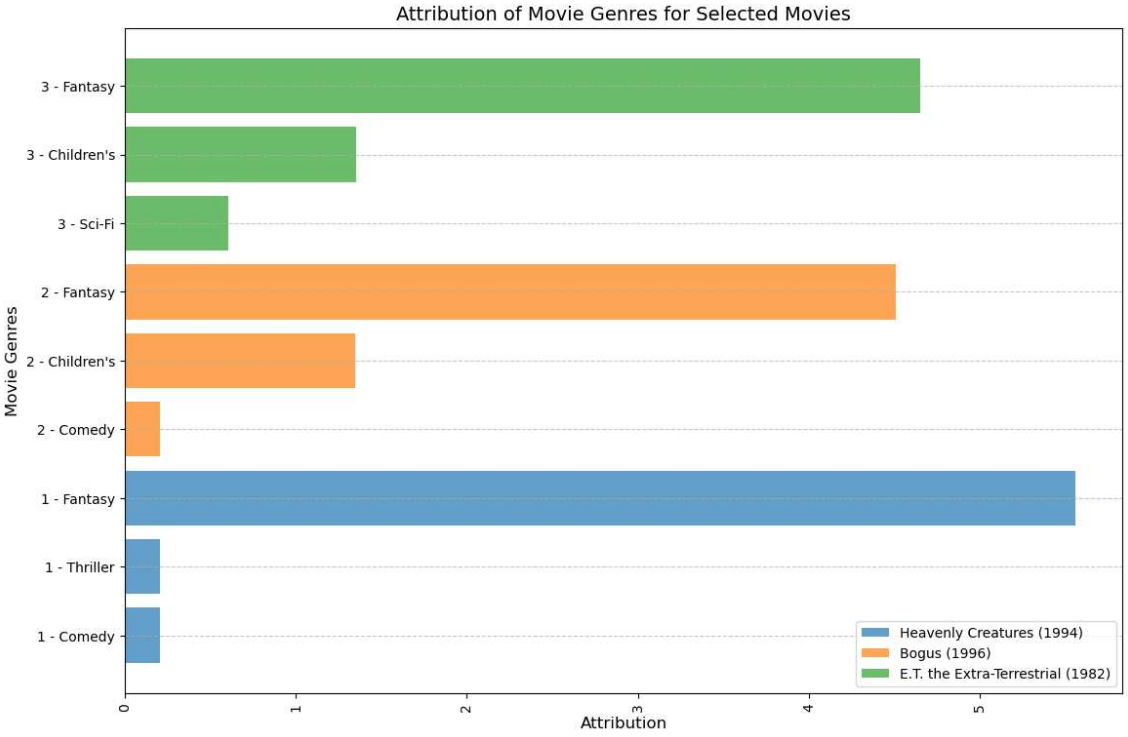


Figure 2. Top three most influential genres for selected movies. Higher values indicate greater impact on the model's recommendations.

4 Discussion

The results of the Integrated Gradients (IG) method provide insightful explanations for the recommender system's predictions. However, certain limitations remain. For example, one might assume that users who favor movies like *Toy Story 2* do so because the film appeals to young children. However, as shown in Figure 1, the user's occupation as a K-12 student receives the highest attribution, while age has a much lower attribution. This discrepancy likely arises because age and student status are highly correlated, leading to confounding in the data. This suggests that the attribution method does not support causal claims. Additionally, the results imply that the model relies on correlations rather than causal relationships, which may indicate poor out-of-distribution generalization [1].

Another observation from Figure 1 is that the scale of attributions for user features is substantially larger than that of movie features, despite feature normalization. While further experiments are needed to reveal the reason, a possible explanation is that one-hot encoding of user occupations results in a sparser feature vector than genre encoding for movies, where each movie can belong to multiple genres. This sparsity could cause the model to assign greater importance to user features during training, leading to the observed disparity in attribution scales.

To conclude, while IG offers valuable insights into the model's decision-making process, its reliance on correlation rather than causation limits its interpretability in real-world scenarios. Further experiments, such as alternative baseline choices or perturbation analysis, could enhance our understanding of the model's behavior and improve its robustness in different contexts.

Bibliography

[1] Peng Cui and Susan Athey. Stable learning establishes some common ground between causal inference and machine learning. 4(2):110-115.

[2] Alexandru L. Ginsca, Adrian Popescu, and Mihai Lupu. Credibility in Information Retrieval. 9(5):355-475.

[3] F. Maxwell Harper and Joseph A. Konstan. The MovieLens Datasets: History and Context. 5(4):19-1.

[4] Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yongdong Zhang, and Meng Wang. LightGCN: Simplifying and Powering Graph Convolution Network for Recommendation.

[5] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A Survey on Bias and Fairness in Machine Learning.

[6] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic Attribution for Deep Networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 3319-3328. PMLR.