

DO VOLUNTEERS REPRESENT THE FULL POPULATION? VOLUNTEER BIAS IN CLICKSTREAM-BASED LEARNING ANALYTICS.

Author: Lennox Ruffiner Supervisors: Dominik Glandorf, Tanja Käser



INTRODUCTION:

Research on student learning frequently relies on participants who **actively choose** to take part in studies. However, these volunteers may not **fully represent** the **wider student population**. When this occurs, **predictive models** and **analytical conclusions** built on volunteer data may succeed for study participants but **fail for the rest of the class**. This study investigated whether **digital behavioral data** from volunteers in a university course could **reliably predict** academic outcomes for the **entire student cohort**.

METHODOLOGY:

- Matched volunteers to anonymized clickstream records and split the full course cohort into **volunteer and non-volunteer subgroups** using platform logs.
- **Derived 20 behavioral features** from raw clickstreams (e.g., regular peak study time, course anticipation, seek direction; see Fig. 1).
- Used **volunteer-only course grades** as the label and trained a **supervised model** on volunteer data to predict performance from behavioral features.
- Compared **behavioral feature distributions** between volunteers and non-volunteers, and compared **predicted-grade distributions** for non-volunteers against the **population-wide grade distribution**.
- Assessed **model confidence** on non-volunteers using **Gaussian Process predictive variance**.

$$\text{speeding up ratio} = \frac{\text{speeding_up_watch_time}}{\text{total_watch_time}}, \quad \text{where } \text{watch_speed} > 1 \Rightarrow \text{speeding up.}$$

Fig 1.1: Formula for the SpeedingUpRatio feature.

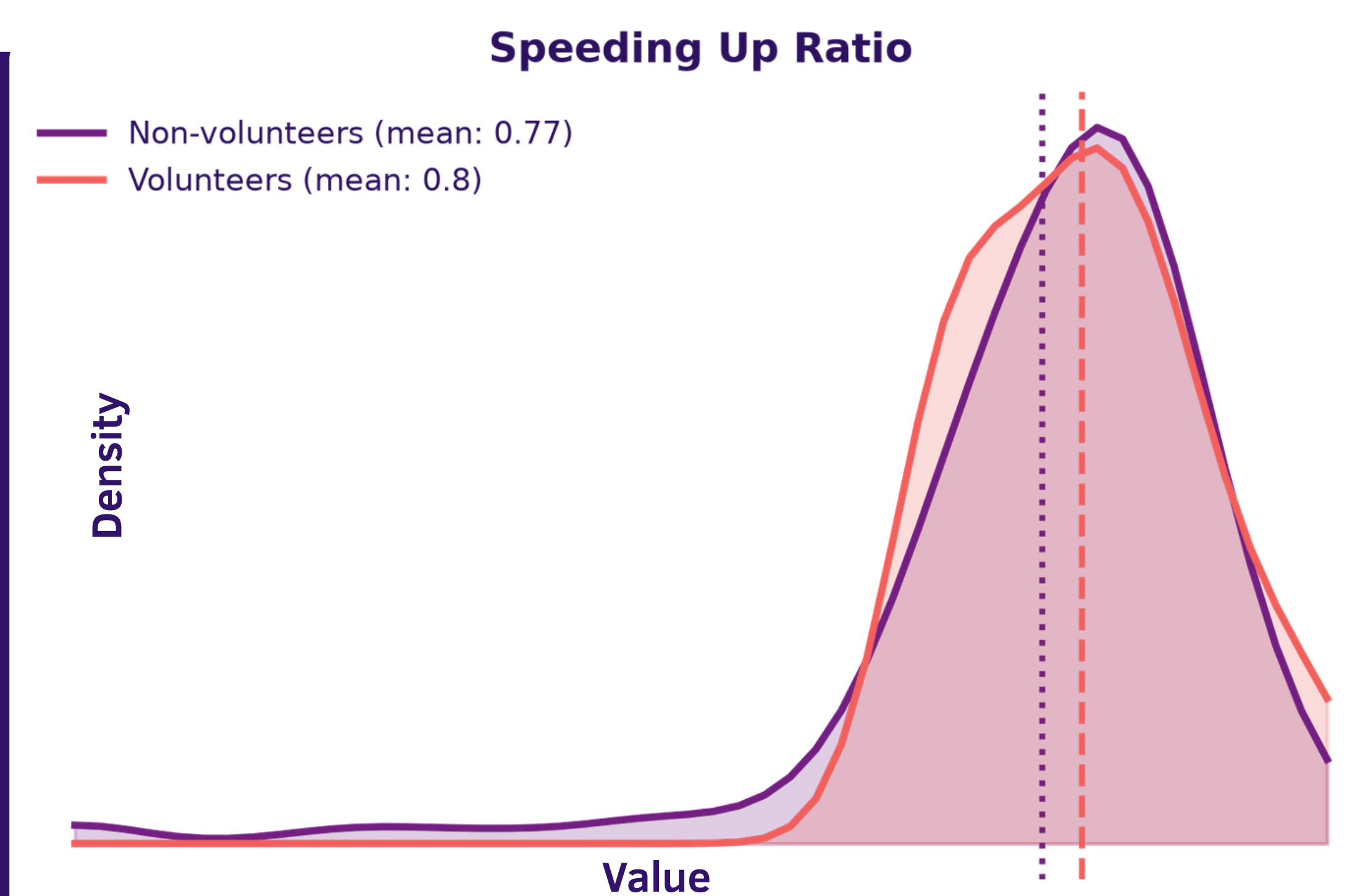


Fig 1.2: Distribution of the SpeedingUpRatio feature.

$$\begin{aligned} \sigma_{volunteers} &= 0.79 \\ \sigma_{non-volunteers} &= 0.77 \end{aligned}$$

Fig. 2: Mean predictive standard deviation from GP.

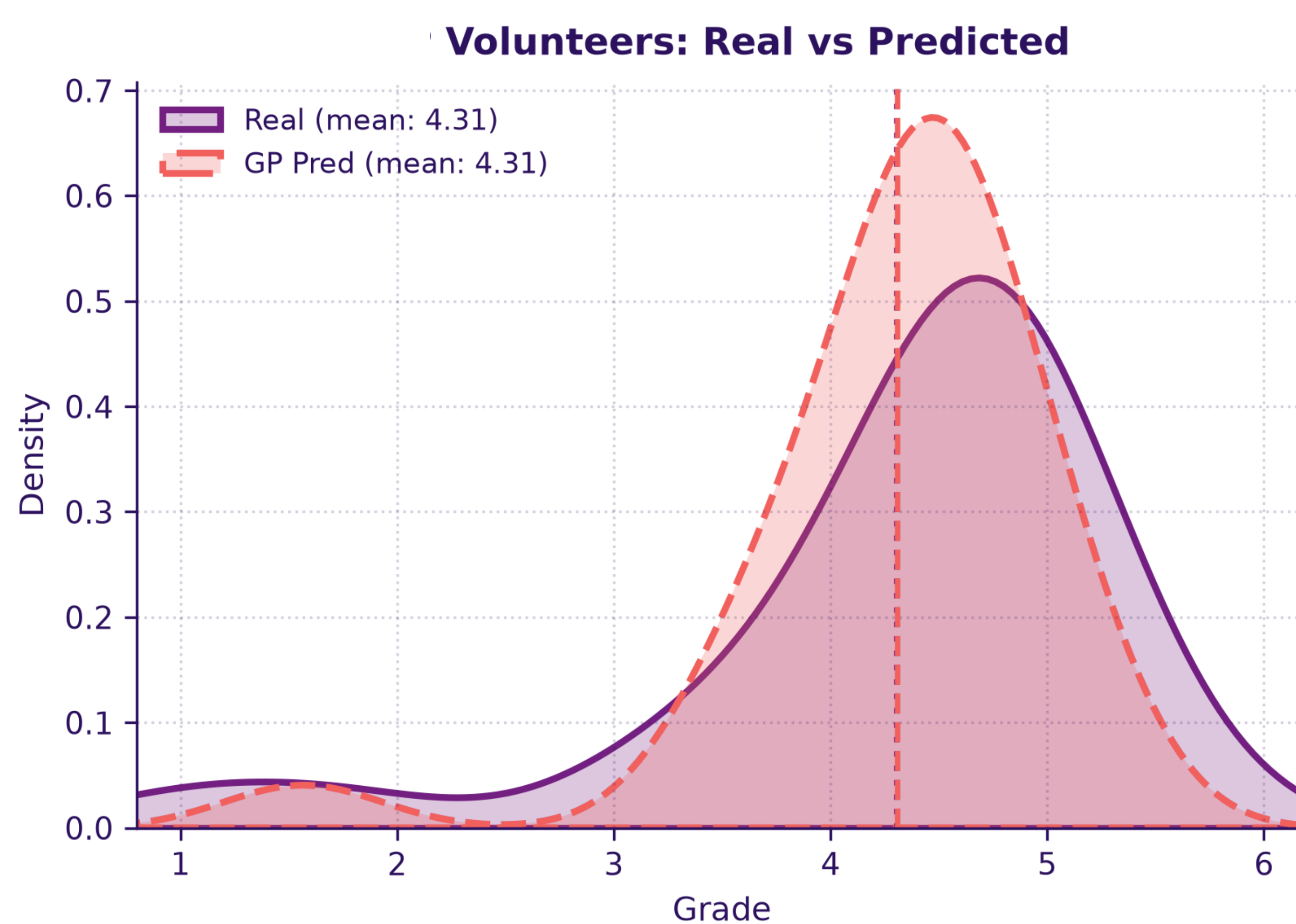
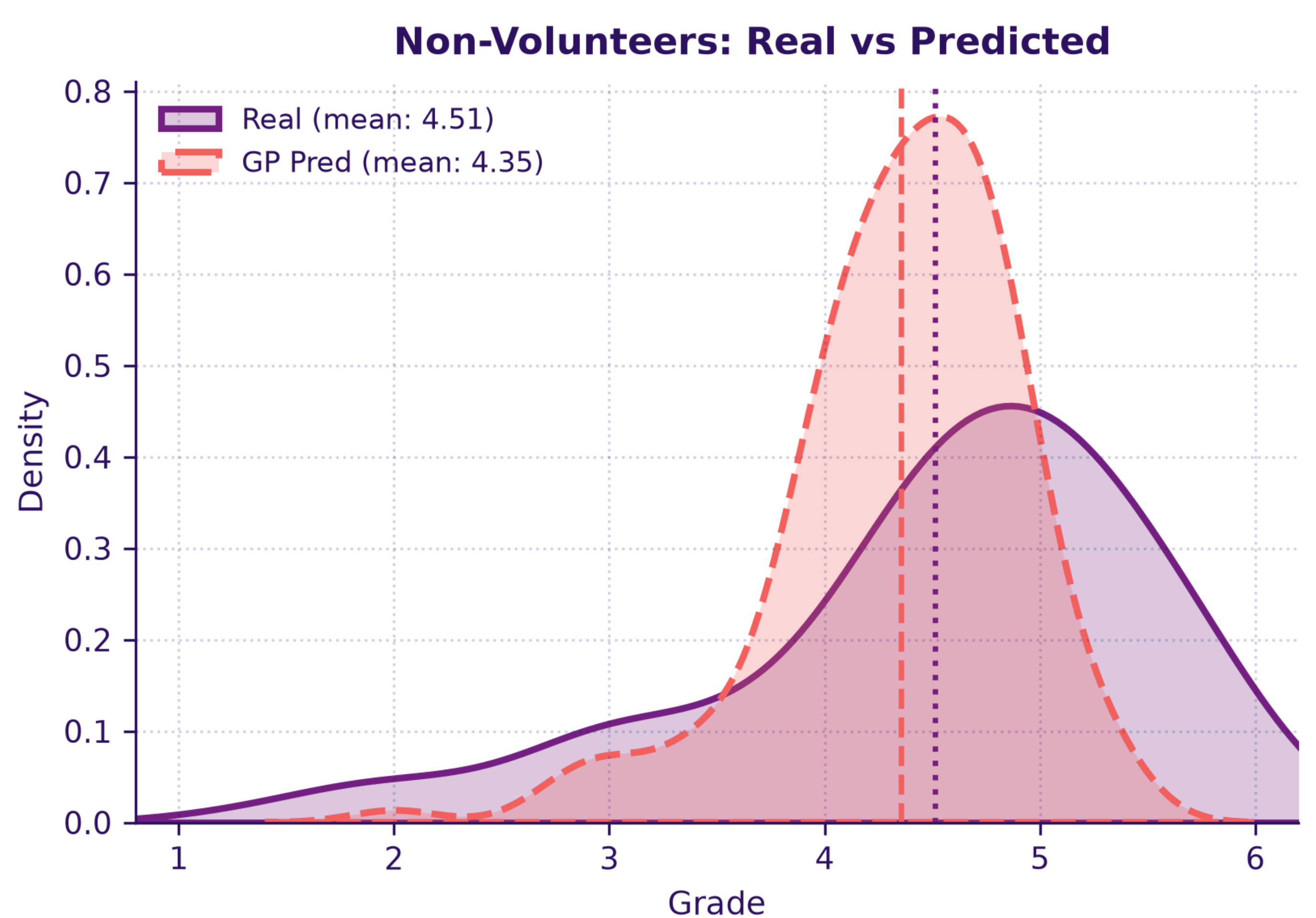


Fig. 3: Observed vs. Predicted Grade Distributions. (volunteers vs. non-volunteers)



RESULTS:

- Volunteers and non-volunteers showed **nearly identical observable online study behaviors** (Fig. 1.2).
- Models trained on volunteers performed **substantially worse** when predicting **non-volunteers' academic performance** (Fig. 3).
- Despite the **performance drop**, models reported **similar confidence** for both groups (Fig. 2).
- This implies **identical observable study behaviors** map to **different academic outcomes** depending on **volunteer status**.

CONCLUSION:

These findings show that **self-selection** can introduce a **hidden bias** into educational data, even when observable study behaviors appear similar across groups. Analyses based only on volunteers may therefore produce **inaccurate conclusions** about the broader student population while remaining **apparently reliable**.