Differentially Private Analysis of U.S. Household Income Statistics using Laplace and Gaussian Mechanisms

Muttaki I Bismoy
Shomitro K Ghosh
Farhan Tanvir &
Ahmad Jadallah

01 Motivation & Problem

Why this project?

- Public socioeconomic datasets risk privacy leakage
- Household income is highly sensitive
- Need for privacy-preserving statistical releases
- Differential Privacy (DP) solves reconstruction & inference attacks

02 Dataset Overview

Dataset: U.S. Household Income Database

(70,000+ geographic regions)

Includes:

Mean, median, stdev income, population, geolocation, land/water area.

========== HEAD ==========

	id	State_Code	State_Name	State_ab	County	City	Place	Type	Primary	Zip_Code	Area_Code	ALand	AWater	Lat	Lon	Mean	Median	Stdev	sum_w
0	1011000	1	Alabama	AL	Mobile County	Chickasaw	Chickasaw city	City	place	36611	251	10894952	909156	30.771450	-88.079697	38773	30506	33101	1638
1	1011001	1	Alabama	AL	Talladega County	Childersburg	Childersburg city	City	place	35044	256	31919335	652240	33.291877	-86.340599	39421	25400	43141	1642
2	1011002	1	Alabama	AL	Calhoun County	Anniston	Choccolocco	CDP	place	36207	256	30159923	239225	33.674346	-85.710918	73511	54847	62988	554
3	1011003	1	Alabama	AL	Mobile County	Wilmer	Chunchula	CDP	place	36587	251	4671130	21008	30.927194	-88.208200	34753	300000	28467	55
4	1011004	1	Alabama	AL	Mobile County	Citronelle	Citronelle city	City	place	36522	251	66930189	713078	31.097269	-88.249843	56102	48865	44810	892

From dataset documentation:

- hi_mean = mean household income
- hi_median
- hi_stdev
- pop, ALand, AWater, lat, lng, state, city
- 325,260 records across U.S.

03 Key Data Fields

Main Fields Used

- hi_mean, hi_median, hi_stdev – Income statistics used for DP analysis
- pop (sum_w) – Population / sampling weight for regional scaling
- lat, lng – Spatial coordinates for geographic income mapping
- ALand, AWater – Land and water area for regional characteristics
- State, County, City identifiers – Administrative grouping for EDA
- Zip_Code, id – Unique geographic identifiers

Household Income Feilds:

- **hi_mean:** The mean household income of record
- **hi_stdev:** The standard deviation of household income
- **hi_samples**: The number of income records
- **hi_sample_weight**: Sum of the samples weights

Location Information:

- **ALand:** Square area of land.
- **AWater:** Square area of water.
- **pop:** Population of location
- **male_pop:** Male population
- **female_pop:** Female population
- **lat:** Location latitude
- **lng:** Location longitude

Location Key Fields:

- **UID:** Unique golden oak Id for every location record
- **STATEID:** State Census Bureau ID.
- **area_code:** Defined via heuristic.
- **zip_code:** Defined via heuristic.

Other Location Feilds Include:

- **place:** Closest place as reported by the U.S. Census Bureau.
- **city:** Closest city to record defined via heuristic.
- **county:** Closest County as reported by the U.S. Census Bureau.
- **state_ab:** State abbreviated name
- **state:** Full state name
- **type:** LocationClassification {City, Village, Town, CPD, ..., etc.}
- **Primary:** Specifies if the location is a tract or a block
- Income Database Information Summary

04 Methodology Overview

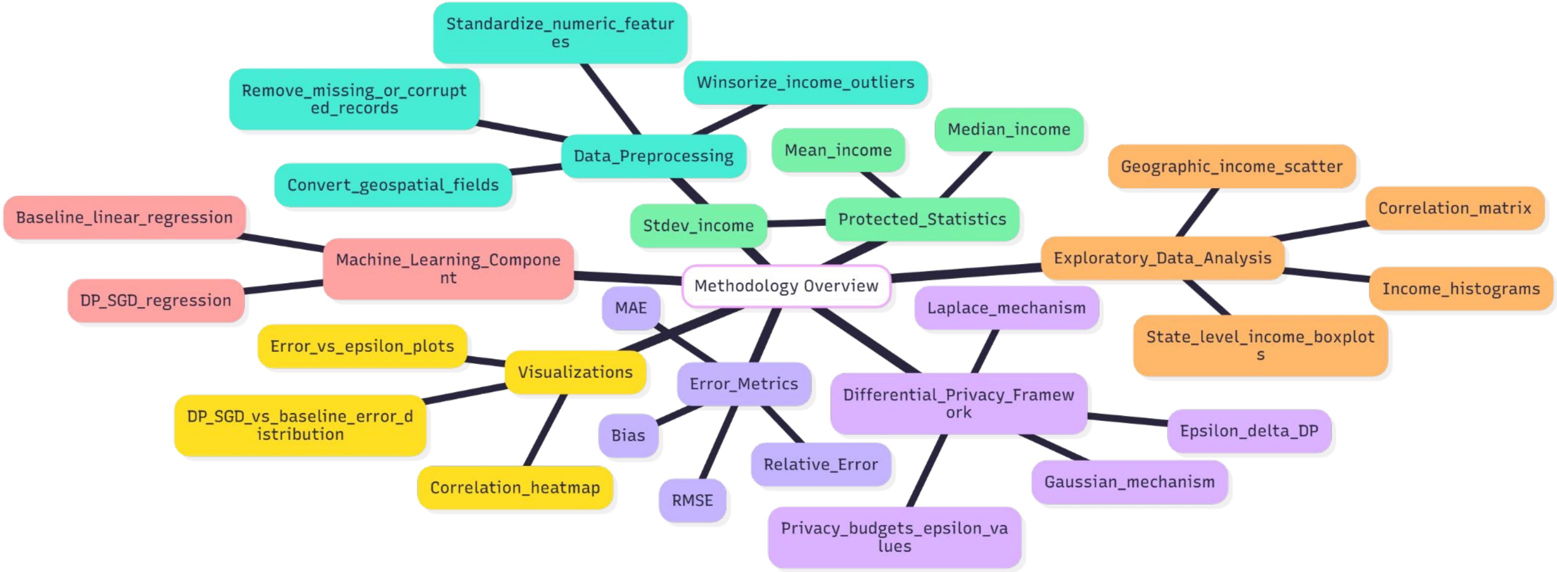


Fig 2. Methodology Overview

1. Data cleaning & preprocessing
2. Exploratory Data Analysis
3. Apply Laplace & Gaussian mechanisms
4. Compute errors: Bias, MAE, RMSE, Relative Error
5. Train Baseline Regression vs DP-SGD
6. Evaluate privacy–utility trade-off

05 Exploratory Data Analysis

Highlight findings:

- **Income is right-skewed**
- **Strong correlation between mean, median & stdev**
- **Income is right-skewed**

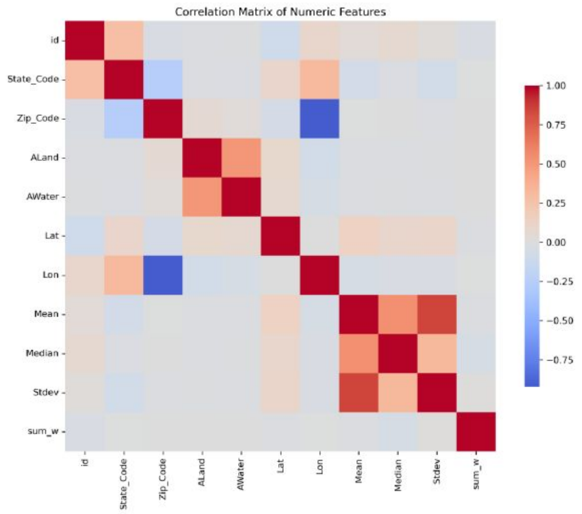


Fig 2. Correlation Matrix of Numeric Features

06 Income Distributions

Income Distributions histograms

- **All income distributions are heavy-tailed**
- **High skew → DP noise impact increases**

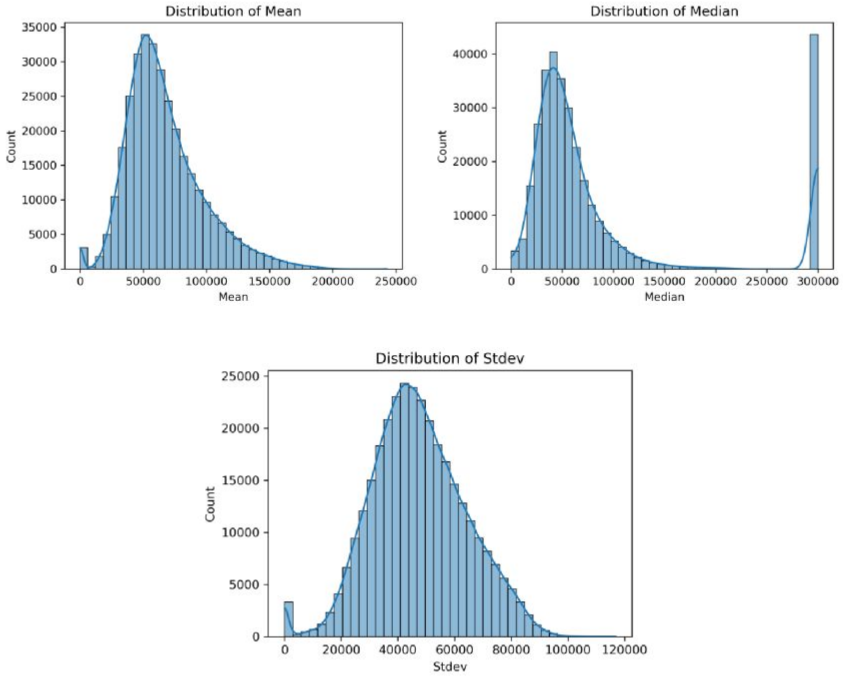


Fig 3: Histograms for Mean, Median and Standard Deviation of household income. All three exhibit right-skewed heavy-tail distributions typical in socioeconomic variables

07 Geographic Distribution

Visual showing strong clustering

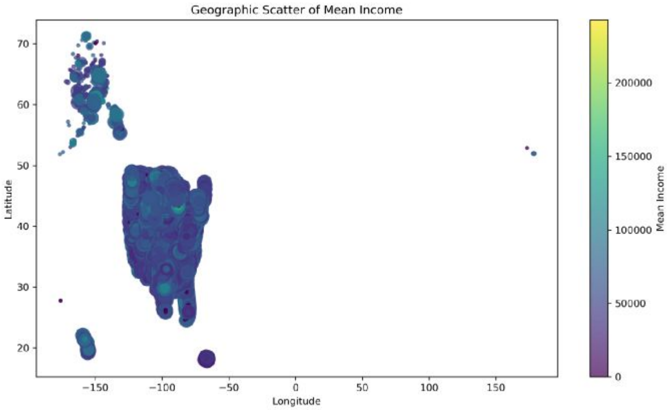


Fig 4. Geographic scatter plot of mean household income across U.S. regions. Color intensity corresponds to income, revealing strong spatial socioeconomic clustering

Key insights:

Coastal states show **higher income**

Rural areas show **lower spread**

08 Differential Privacy Framework

- **(ε, δ)-Differential Privacy**
- **Laplace Mechanism for pure DP**
- **Gaussian Mechanism for approximate DP**

Laplace noise: Laplace(0, Δf / ε)

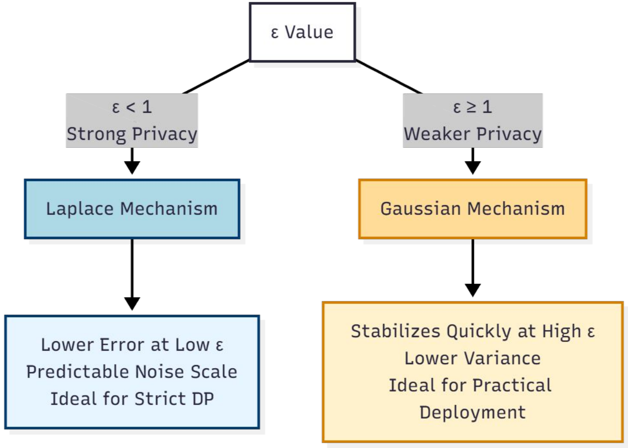
Gaussian noise: σ = sqrt(2 ln(1.25/δ)) * Δf / ε

ε values tested: {0.05, 0.1, 0.5, 1, 2, 5}

We apply DP to:

- **Mean household income**
- **Median household income**
- **Standard deviation**

09 Interpreting Laplace vs Gaussian Mechanisms



Why Laplace Performs Better at Low ε (Strong Privacy):

- Laplace noise is scaled by $\Delta f / \varepsilon$, producing a *tighter* noise distribution.
- Works well when ε is tiny because the added noise grows linearly and remains predictable.
- This method yields fewer extreme values, which in turn leads to lower MAE, RMSE, and Relative Error at low values.
- This method is more appropriate for pure DP scenarios that demand strict guarantees.

Why Gaussian Stabilizes Faster at High ε (Weaker Privacy):

- Gaussian noise adds variance proportional to $\sigma = \sqrt{(2 \ln(1.25/\delta))} \cdot (\Delta f / \varepsilon)$.
- At small ε, the variance is large → more outliers → higher errors.
- At moderate and high ε, variance rapidly shrinks, causing the Gaussian mechanism to stabilize and outperform Laplace.
- Gaussian handles high-dimensional or aggregated statistics more smoothly, improving performance at $\varepsilon \geq 1$.

10 Error Metrics Used

Define:

- **Bias**
- **Mean Absolute Error (MAE)**
- **Root Mean Squared Error (RMSE)**
- **Relative Error**

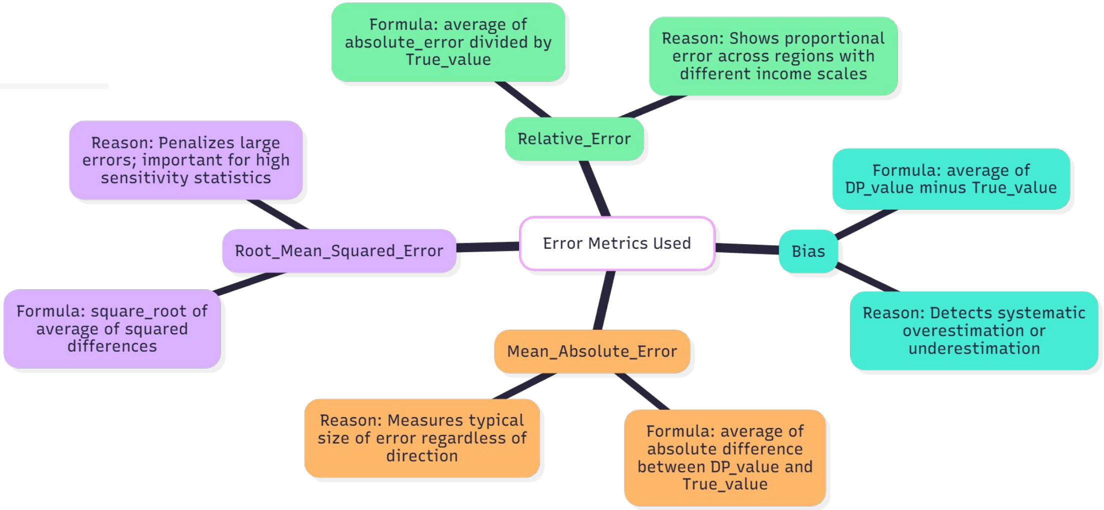


Fig 5. Bias, MAE, RMSE and RE are the error matrices used for this project. Their formula and reason behind using them are stated in this figure.

11 Results: Bias vs ε

Key points:

- **Laplace bias high at low ε**
- **Gaussian stabilizes faster**
- **Median most sensitive**

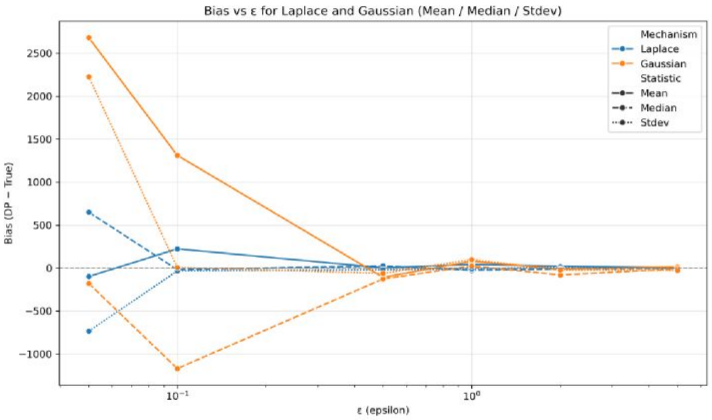


Fig 6. Bias vs ε for Laplace and Gaussian

12 Results: MAE & RMSE Trends

Insights:

- **Error decreases exponentially as ε increases**
- **Gaussian performs better at mid-high ε**
- **Laplace competitive at ε < 1**

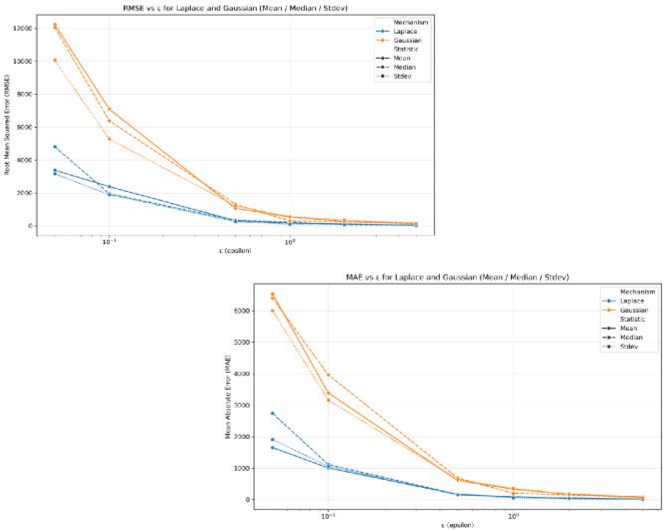


Fig 7. MAE and RMSE for Laplace and Gaussian mechanisms across varying privacy budgets (ϵ). Each panel compares mean, median and standard deviation perturbations

13 Relative Error Comparison

Observations:

- **Relative error extremely high at ε=0.1**
- **Drops near zero at ε≥10**
- **Mean least sensitive to noise**

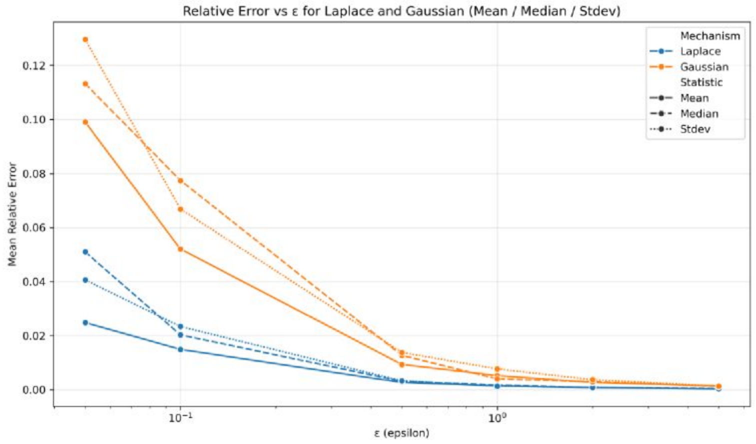


Fig 8. Relative Error for Laplace and Gaussian mechanisms across varying privacy budgets (ϵ)

14 Additional Data Insights

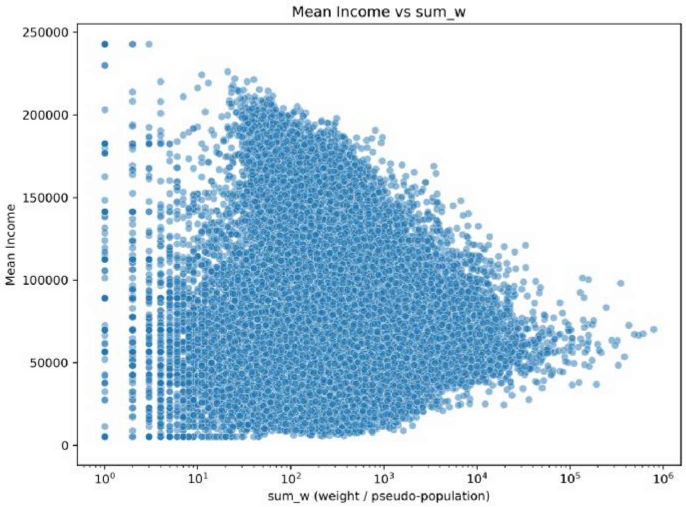


Fig 9. Scatter plot of mean income vs. pseudo-population (sum_w). Higher populations show narrower income vari ability, while low-population regions show wider spread

- **Low-population regions → higher variance**
- **DP noise impacts them more**

15 Machine Learning Component

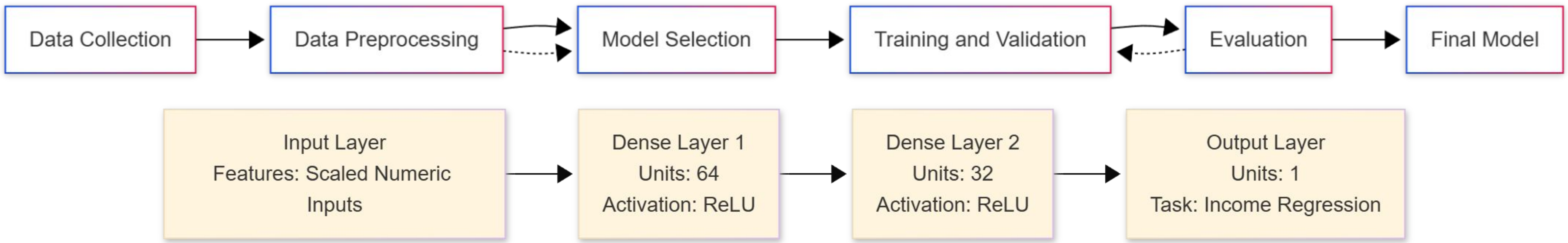


Fig 9. Diagram of ML Workflow and ML Model Architecture

Models used:

- **Baseline Linear Regression**
- **DP-SGD Regression (TensorFlow Privacy)**
- **DP-SGD settings: ε≈1, δ=1e-5, noise multiplier=1.1**

16 ML Results: Error Distribution

Findings:

- **Baseline error tightly centered**
- **DP-SGD shows heavy negative skew, wide spread**
- **DP noise strongly reduces ML utility**

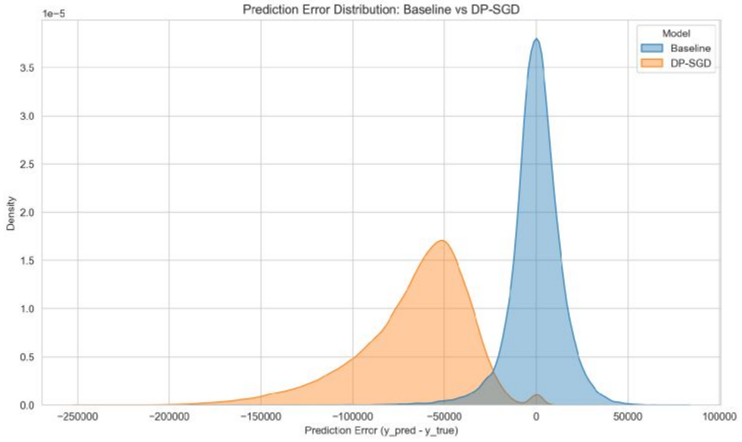


Fig 11. Prediction error density for baseline Linear Regression vs. DP-SGD model. DP-SGD exhibits significantly larger error variance and negative skew

17 ML Metrics Comparison

Summary:

- **MAE & RMSE significantly worse for DP-SGD**
- **R^2 near zero → poor predictive power**

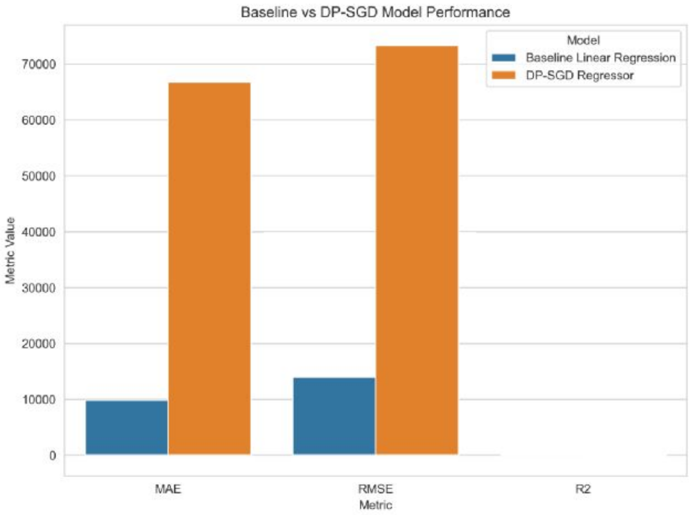


Fig 12. MAE, RMSE and $R2$ comparison between baseline and DP-SGD models. The DP-SGD model suffers significant performance degradation but ensures (ϵ,δ)-DP

18 Key Takeaways

Laplace vs Gaussian

- Laplace is better for small ε. Laplace has higher bias but lower relative error
- Gaussian performs worse at low ε. Gaussian more stable overall

Impact of DP

- Low ε → high noise, low accuracy
- Skewed income data amplifies errors
- ML under DP-SGD loses significant utility

19 Conclusion & Future Work

Conclusion

- Differential Privacy effectively protects sensitive socioeconomic statistics.
- Strong privacy (low ε) introduces high noise and reduces accuracy.
- Heavy-tailed income distributions amplify DP error.
- Gaussian mechanism more stable at higher ε; Laplace better for small ε.
- ML under DP-SGD shows significantly reduced utility.

Limitations

- Income data are highly skewed, increasing sensitivity and noise impact.
- Median and standard deviation are more sensitive to DP noise.
- DP-SGD regression shows poor predictive performance (low R^2, large error).

Future Work

- Explore Rényi Differential Privacy and advanced privacy accounting.
- Improve sensitivity bounding techniques.
- Test alternative DP machine learning architectures.

THANK YOU

Any Questions?