The objective of this project is to model real income probabilistically rather than through point estimates. Conventional regression approaches, including gradient boosting methods such as XGBoost, estimate only the conditional mean of the response variable, providing no insight into uncertainty or distributional shape. However, income data are strictly positive, right-skewed, and heteroskedastic, characteristics that mean-based models cannot adequately capture.

To address these limitations, we employ XGBoostLSS [2], an extension of XGBoost that models the entire conditional distribution by jointly estimating its parameters as functions of the predictors. This framework retains the predictive strength and computational efficiency of XGBoost while enabling richer probabilistic interpretation of the outcomes.

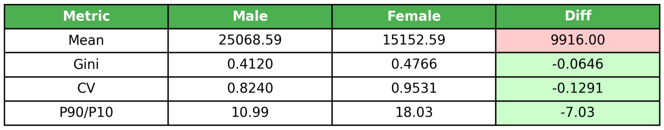
After an extensive exploratory data analysis, we conducted feature engineering to improve model expressiveness. Continuous variables such as `age`, `childs`, and `prestg10` were binned (following [1]) or log-transformed to reduce skewness, and selected categorical variables were combined to capture potential interactions [1]. Mean target encoding with out-of-fold cross-validation and smoothing was applied to categorical variables to ensure robust and unbiased encoding.

Among the likelihood options implemented in the `xgoostlss` library, the Gamma distribution provided the best fit [1] to the training income data. The model employed L2 stabilization to ensure the numerical stability of gradient and Hessian calculations, an exponential response function to guarantee the positivity of distributional parameters, and was optimized using the CRPS loss.

Model performance was optimized using nested cross-validation with an outer 5-fold and inner 5-fold structure. Hyperparameter tuning in each inner loop was conducted via Bayesian optimization with Optuna, with a budget of 3 hours per inner cross-validation. With the optimized hyperparameters [2], the final model achieved a mean CRPS of 7949.84 across outer folds, and an overall OOF CRPS of 7948, indicating consistent generalization performance across folds.

Finally, to isolate the distributional effects of demographic factors, we trained a simplified conditional model using only `year` and `gender` as predictors, under the same Gamma likelihood and optimization framework.

1. Females display higher real income inequality in 1980. Their Gini coefficient (0.4766 vs. 0.4120), coefficient of variation (0.9531 vs. 0.8240), and P90/P10 ratio (18.03 vs. 10.99) are all higher than for males, indicating greater dispersion. The density plot [2] confirms this, showing a sharper low-income peak and a longer right tail for females. Overall, inequality is higher among females despite lower average income levels.



Metric	Male	Female	Diff
Mean	25068.59	15152.59	9916.00
Gini	0.4120	0.4766	-0.0646
CV	0.8240	0.9531	-0.1291
P90/P10	10.99	18.03	-7.03

2. Based on the Monte Carlo simulation with 1,000 samples per gender, the probability that a male earned more than a female was 69.1% in 1980 and 65.2% in 2010, showing a slight reduction in the gender income gap over 30 years. The figures display overlapping income distributions with males consistently shifted rightward, and income difference histograms centered around $9,357 (1980) and $8,898 (2010). The probability is calculated by comparing paired random samples from each gender's fitted gamma distribution—the proportion where male income exceeds female income directly estimates $P(\text{Male} > \text{Female})$. The income difference histograms [3, 4] confirm this: approximately 69% and 65% of samples fall above zero (light blue bars) in 1980 and 2010 respectively, indicating persistent, though slightly narrowing, gender wage gap.

Appendix

A. Feature Engineering

Table 1: Features Used in XGBoostLSS Model

Feature Type	Features
Original Features	
Temporal	`year`
Demographic	`age`, `gender`, `maritalcat`
Occupational	`occrecode`, `prestg10`, `wrkstat`
Education	`educcat`
Family	`childs`
Binned Features	
Categorical bins	`age_bin`, `childs_bin`, `prestg10_bin`
Interaction Features	
Two-way interactions	`occrecode_gender`, `occrecode_educcat`, `wrkstat_gender`, `occrecode_prestg10_bin`, `educcat_prestg10_bin`, `maritalcat_age_bin`
Log-Transformed Features	
Logarithmic	`age_log`, `prestg10_log`, `childs_log`
Total Features	**21**

Note: Binning specifications: `age_bin` bins=[18-30, 31-50, 51+]; `childs_bin` bins=[0-2, 3-5, 6+]; `prestg10_bin` bins=[16-30, 31-50, 51+]. Log transformations applied as $\log(x+1)$ to handle zero values.

B. Hyperparameter Optimization

Table 2: Results of Bayesian hyperparameter optimization using 5×5-fold nested cross-validation.

Parameter	Search Range	Optimized Value
`eta`	$[10^{-5}, 0.3]$	0.0878
`max_depth`	$[1, 8]$	1
`gamma`	$[10^{-8}, 40]$	0.00043
`subsample`	$[0.2, 1.0]$	0.986
`min_child_weight`	$[10^{-8}, 500]$	0.0640
`reg_alpha`	$[0, 5]$	2.60
`reg_lambda`	$[0.5, 5]$	2.05
`colsample_bytree`	$[0.2, 1.0]$	0.590
`booster`	gbtree	gbtree
`tree_method`	{auto, approx, hist}	auto
`opt_rounds`	$[1, 500]$	237

C. Visualizations

Figure 1: Best-fitting distribution for the training data (negative log-likelihood = 347,828.41).

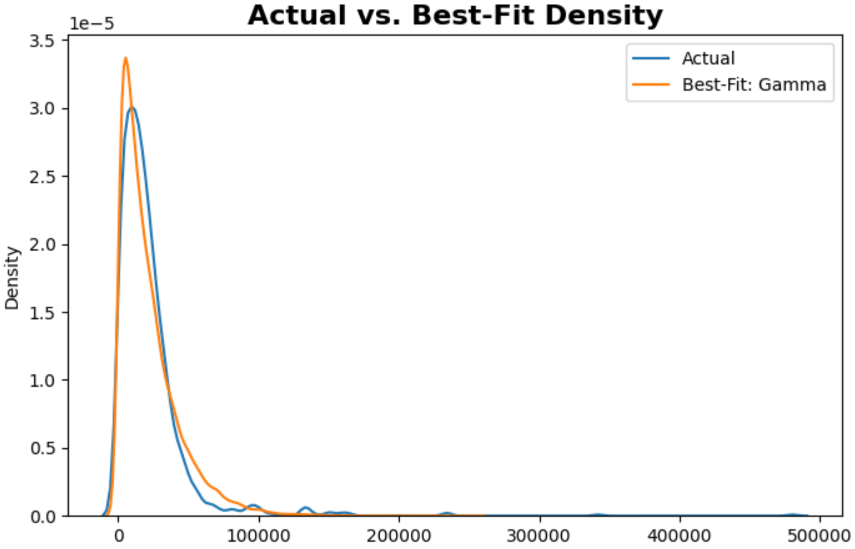


Figure 2: Distribution of income per gender in 1980

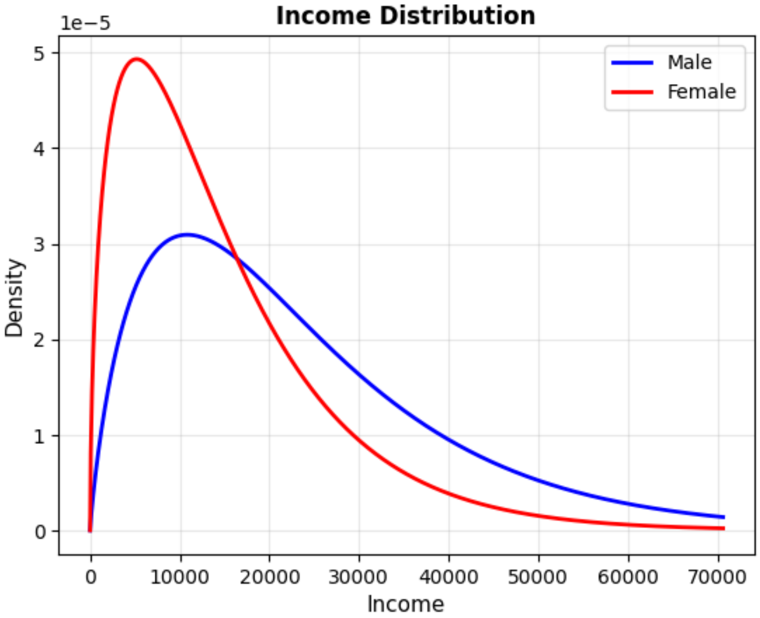
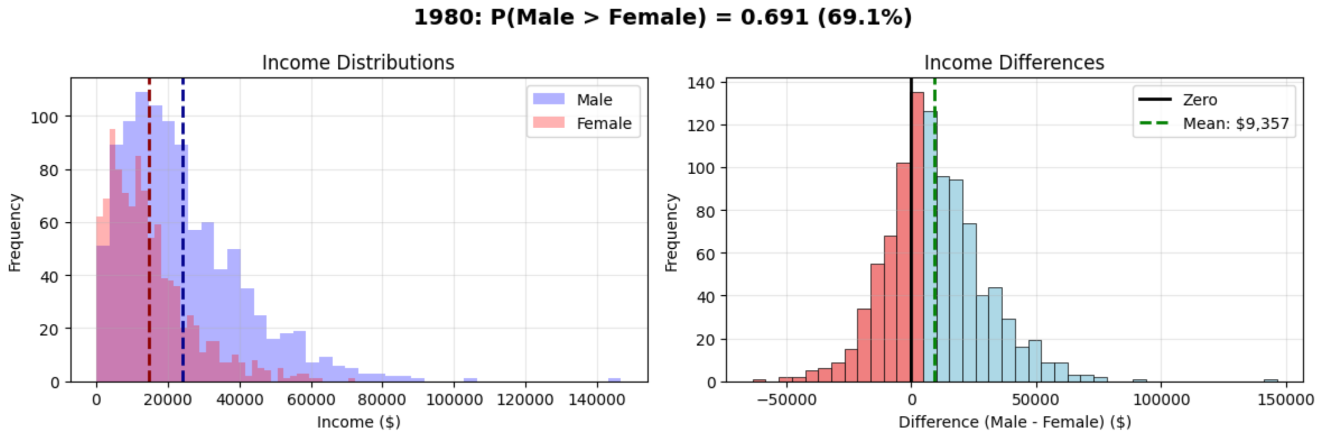


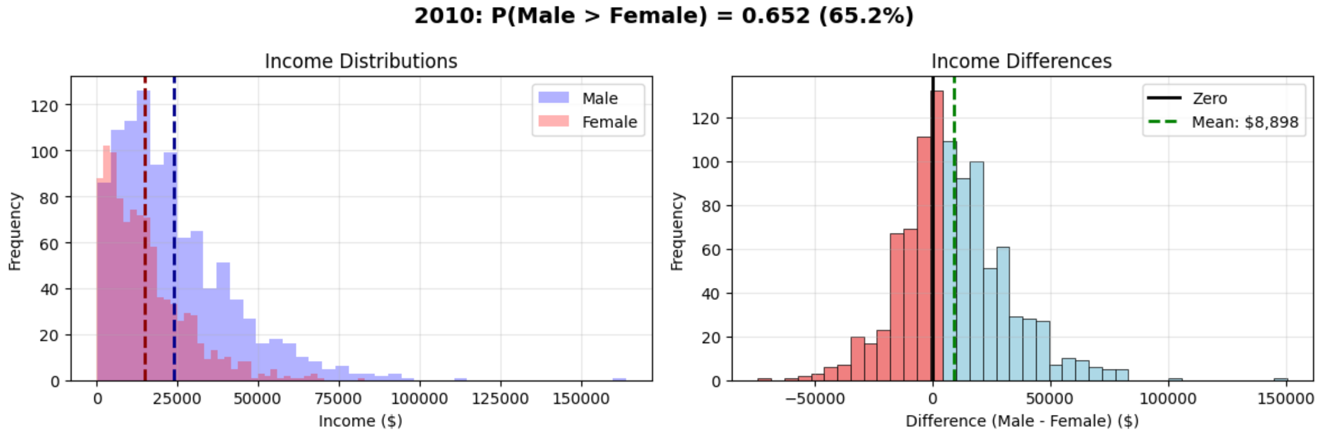
Figure 3: Income comparison by gender in 1980


1980: P(Male > Female) = 0.691 (69.1%)

Income Distributions

Income Differences

Figure 4: Income comparison by gender in 2010


2010: P(Male > Female) = 0.652 (65.2%)

Income Distributions

Income Differences

References

[1] Alex Amaguaya and Lea Bergmann. The gender pay gap in the general social survey: Statistical learning contents, October 2023.

[2] Alexander März. Xgboostlss: An extension of xgboost to probabilistic forecasting, 2019.