OXFORD JOURNAL
OF STUDENT
SCHOLARSHIP

A Machine Learning Approach to Understanding the Determinants of Female Labour Force Participation Rate in India

Meit Dabas
meitdaabs@gmail.com

ABSTRACT

Female labour force participation (FLFP) in India remains low despite improvements in education and economic growth, forming what is often referred to as the "FLFP puzzle." This study examines which factors most strongly predict women's participation in the labour market, and whether these patterns reflect individual characteristics or deeper constraints.

Using microdata from the Periodic Labour Force Survey (PLFS), the analysis applies logistic regression, Random Forest, and XGBoost models to evaluate the relative importance of demographic, socio-economic, and household-level variables. Across models, marital status, sector of residence, and life-cycle effects (captured through age) consistently emerge as the strongest predictors of participation, while education, religion, and social group play a comparatively limited role.

Comparative evaluation using multiple performance metrics shows that non-linear models outperform logistic regression, indicating the presence of complex relationships not fully captured by linear specifications. However, recall remains below 0.5 across all models, suggesting that fewer than half of actual participants are correctly identified. This highlights the limits of observable characteristics in fully capturing participation behaviour.

These findings indicate that FLFP is more strongly associated with household context, labour market conditions, and lifecycle dynamics than with individual attributes alone. At the same time, the results demonstrate that predictive models, while useful in identifying consistent patterns, cannot fully explain participation outcomes.

Overall, the study shows how machine learning methods can complement traditional approaches by uncovering stable predictive relationships, while also highlighting the need for richer data and integrated predictive–causal frameworks to better understand the drivers of female labour force participation in India.

May 2026
Vol 7. No 1.

INTRODUCTION

Female labour force participation (FLFP) is widely recognized as an important driver of economic growth, household welfare, and gender equality. Greater participation of women in the workforce increases household incomes, expands the effective labour supply, and contributes to broader economic development. Consequently, understanding the factors that influence women's participation in the labour market has become a central question in labour economics and development research.

Despite significant improvements in female education and health outcomes over the past few decades, India continues to exhibit relatively low levels of female labour force participation compared with many other developing economies, with an estimated participation rate of approximately 41.7% (2023-2024). A substantial proportion of working-age women therefore remain outside the labour market. This phenomenon has often been described as the "Indian FLFP puzzle," referring to the persistence of low participation rates despite sustained economic growth and rising educational attainment among women.

A wide range of factors have been proposed to explain this pattern, including social norms, marital status, educational attainment, household responsibilities, and labour market opportunities. However, an important question remains: Which factors most strongly predict female labour force participation in India? And how accurately can we predict them?

Previous research has largely relied on traditional econometric approaches to study the determinants of FLFP. While these models provide valuable insights, they often assume linear relationships between variables and may not fully capture the complex interactions between socio-economic characteristics and labour market outcomes.

Recent advances in machine learning offer new tools to examine such relationships in large-scale datasets. By applying predictive modelling techniques to labour force survey data, it becomes possible to identify the most influential determinants of FLFP and explore patterns that may not be easily captured through conventional statistical methods.

LITERATURE REVIEW

The determinants of female labour force participation in India have been studied previously using traditional econometric methods on NSSO and PLFS data. A central finding in this literature is the importance of household income and structural change. Mehrotra and Parida (2017) highlight a strong negative income effect arising from rising rural wages and household incomes, combined with agricultural mechanisation that displaced women from traditional roles. Similarly, Deshpande and Singh (2021) argue that women are not voluntarily "dropping out" but are being "pushed out" by structural shifts, particularly the decline in agricultural employment without corresponding expansion of suitable jobs in manufacturing or services.

A second strand of the literature emphasises the role of social norms and intra-household constraints.

Klasen and Pieters (2015) show that, unlike in many other countries, higher education and income in India are often negatively associated with FLFP, reflecting strong cultural norms and restrictions on women's work, especially in urban areas. These findings suggest that labour supply decisions are not determined solely by individual characteristics, but are shaped by household bargaining dynamics and social expectations.

While these studies provide important insights, they rely almost exclusively on linear models (probit/logit) that report average effects and struggle to capture complex interactions or rank the relative importance of predictors. Very few studies have applied machine learning techniques to Indian FLFP microdata.

This study fills that gap by using logistic regression together with Random Forest and XGBoost on recent PLFS data. By comparing feature importance, permutation importance, and SHAP values across models, it identifies which factors *most strongly predict* participation and how accurately the models perform.

DATA

Dataset Overview

This research uses data from the Periodic Labour Force Survey (PLFS) conducted by the National Sample Survey Office (NSSO) under the Ministry of Statistics and Programme Implementation, Government of India. The dataset used in this study contains information on 415,549 individuals across 101,957 households surveyed across India. The survey collects detailed demographic and socio-economic information for each household member, including age, gender, marital status, educational attainment, religion, and social group, along with information on employment status and labour market participation.

Given its large sample size and extensive coverage of socio-economic characteristics, the PLFS provides a suitable dataset for examining the determinants of female labour force participation in India.
The variables utilised are:

- Age (continuous variable): Age of the individual in years.
- General_Education_Level (13 integer options): Educational attainment level of the individual, ranging from no formal education to postgraduate or professional degree.
- Marital_Status (5 integer options): The marital status of the individual, either never married, currently married, widowed, divorced, or separated.
- Sector (2 integer options): Sector of residence of the individual, either rural or urban.
- Religion (6 integer options): Religious affiliation of the individual, either Hindu, Muslim, Christian, Sikh, Buddhist/Jain, or other religions.
- Social_Group (4 integer options): Social category of the individual, either General, Other Backward Classes (OBC), Scheduled Castes (SC), or Scheduled Tribes (ST).
- Household_Expenditure (continuous variable): Monthly household consumption expenditure used as a proxy for household economic status.
- log_expenditure (continuous variable): The natural logarithm of household expenditure.

- Age_squared (continuous variable): Square of the age variable.

Sample Selection

The analysis is restricted to female individuals of working age, i.e. 15-59 years. Observations with missing values in key variables used in the analysis were excluded. After applying these filters, the final analytical sample consists of 135,500 observations.

The dependent variable is a binary indicator of female labour force participation constructed using the usual principal activity status recorded in the survey, where individuals classified as employed or actively seeking employment are coded as participating in the labour force.

Limitations

Although the PLFS provides detailed information on demographic and labour market characteristics, some limitations should be noted. First, the survey is cross-sectional, limiting the ability to draw causal conclusions about the determinants of female labour force participation. Second, certain relevant factors, such as specific social norms, individual household decision-making, and local labour market conditions, are not directly observable in the dataset. Finally, women's economic activity may be underreported, particularly in cases of informal or unpaid work.

METHODOLOGY

Data Preprocessing

Prior to model estimation, the dataset was prepared for machine learning analysis through several preprocessing steps. Variables describing individual characteristics, including marital status, religion, social group, education level, and sector of employment, are categorical in nature. To incorporate these variables into the models, they were transformed into binary indicators using one-hot encoding, ensuring that no artificial ordinal relationships are imposed between categories.

To capture potential non-linear relationships between age and labour force participation, an additional variable representing age squared (Age_squared) was constructed. This transformation allows the models to account for possible curvature in the relationship between age and participation outcomes.

After preprocessing, the modelling dataset consisted of 135,500 observations and 28 features, including both continuous variables and encoded categorical indicators. The dataset was then divided into training and testing subsets, with the training data (70%) of 94,850 observations used to estimate model parameters and the testing data (30%) of 40,650 observations used to evaluate predictive performance on unseen observations.

All preprocessing and model estimation procedures were implemented using the Python programming language, primarily through the scikit-learn libraries.

Logistic Regression

Logistic regression serves as the baseline model for analysing the relationship between socioeconomic characteristics and FLFP. Because the dependent variable is binary, the model estimates the probability that an individual participates in the labour force as a function of the explanatory variables described in Section 3 as $P(FLFP_i = 1) = \frac{1}{1+e^{-(\beta_0+\beta_1 Age_i+\beta_2 Age_i^2+\ldots+\beta_k X_{ik})}}$.

Model parameters are estimated using maximum likelihood. To ensure convergence during estimation, the maximum number of iterations was increased. The estimated coefficients are transformed into odds ratios as $Odds\ Ratio = e^{\beta j}$, and marginal effects as $\frac{\partial P}{\partial X_j} = \beta_j \cdot P(1-P)$, and are computed to examine how changes in explanatory variables are associated with the probability of participation.

Model hyperparameters:

- 'max_iter': 5000
- 'penalty': 'l2'
- 'solver': 'lbfgs'

Random Forest

To complement the baseline logistic regression model, a Random Forest classifier is employed to capture nonlinear relationships and interactions between explanatory variables. Random Forest models construct an ensemble of decision trees and aggregate their predictions, allowing the model to flexibly capture complex patterns in the data while reducing overfitting. The model is trained using the same explanatory variables described in Section 3 with the following hyperparameters:

- 'n_estimators': 200
- 'max_depth': 10
- 'random_state': 42
- 'n_jobs': −1

To identify the most influential predictors of female labour force participation, feature importance scores derived from the trained forest are examined. Feature importance measures reflect the contribution of each variable to predictive performance within the model and do not imply causal relationships. In addition, permutation importance is computed as $PI_j = E[L(y, f(x)) - L(y, f(X^{\pi_j}))]$, to evaluate the relative contribution of each predictor by measuring the change in model performance when individual features are randomly shuffled, considering that standard feature importance measures can sometimes be biased toward variables with greater variability.

eXtreme Gradient Boosting

The analysis further employs eXtreme Gradient Boosting (XGBoost). XGBoost is a gradient boosting algorithm that builds decision trees sequentially, with each new tree correcting errors made by previous trees. This approach often provides strong predictive performance and is well suited for capturing complex nonlinear relationships and interactions among predictors. The model is trained using the same

set of explanatory variables described in Section 3. To interpret the contribution of individual predictors, feature importance scores are examined, and SHAP (SHapley Additive exPlanations) values are used to analyse how individual variables influence predicted participation outcomes. The XGBoost model was implemented with the following model hyperparameters:

- 'n_estimators': 300
- 'max_depth': 5
- 'learning_rate': 0.05
- 'subsample': 0.8
- 'colsample_bytree': 0.8

Model hyperparameters were selected through a combination of domain-informed choices and empirical validation to balance predictive performance and generalizability. For instance, tree-based models were constrained using parameters such as maximum depth and minimum samples per split to prevent overfitting, while ensemble size was increased to stabilize predictions. Similarly, logistic regression was allowed sufficient iterations to ensure convergence given the dimensionality introduced by one-hot encoding. Across models, parameters were chosen to capture non-linear relationships without excessively fitting noise in the data, ensuring that the results reflect robust predictive patterns rather than artifacts of model complexity.

Model performance is evaluated using five complementary metrics, including accuracy, precision, recall, and F1-score, alongside ROC-AUC. ROC-AUC measures the model's ability to distinguish between participants and non-participants across thresholds. Precision and recall provide additional insight into classification performance by capturing the accuracy of predicted participation and the proportion of actual participants correctly identified. The F1-score summarizes this trade-off.

RESULTS

Descriptive Statistics

The final dataset with 135,500 working-age women reveals that the overall female labour force participation rate is 35.24% in the unweighted sample and 32.92% using survey weights.

Basis - Area	% age of participation	Basis - Marital Status	% age of participation
Rural	35.47%	Never Married	18.4%
Urban	27.09%	Currently married	35.1%
		Widowed	56.6%
		Divorced/Separated	63.7%

Figure 1: Participation based on sector and marital status

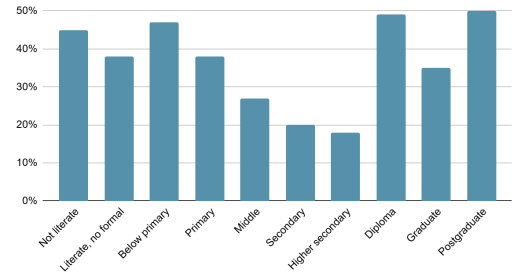
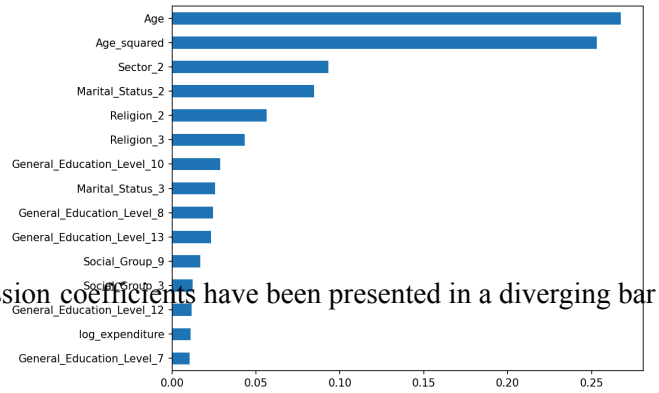


Figure 2: Participation based on education

There is a U-shaped curve in labor participation based on education level with least involvement with only higher secondary education and most with PG and even illiterate individuals.



Logistic Regression

The top 15 predictors based on the logistic regression coefficients have been presented in a diverging bar graph in Figure 3.

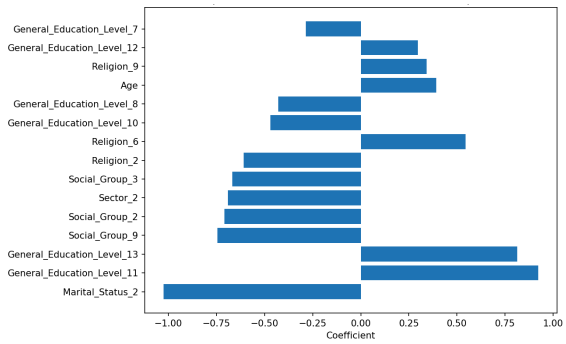


Figure 3: Logistic Regression Coefficients

Random Forest

The top 15 predictors based on permutation importance have been presented in Figure 4, and those of feature importance in Figure 5.

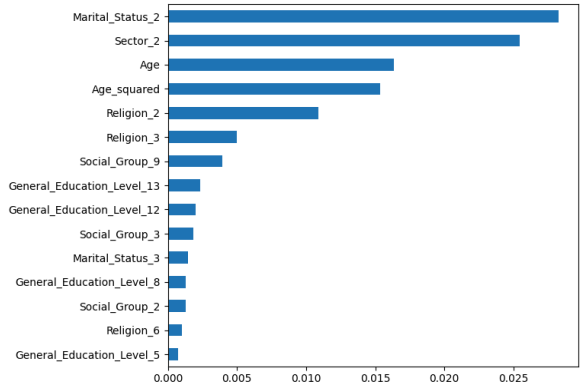


Figure 4: Random Forest Permutation Importance

Figure 5: Random Forest Feature Importance

eXtreme Gradient Boosting

The top 15 predictors based on SHAP feature importance have been presented in Figure 6 as a beeswarm plot.

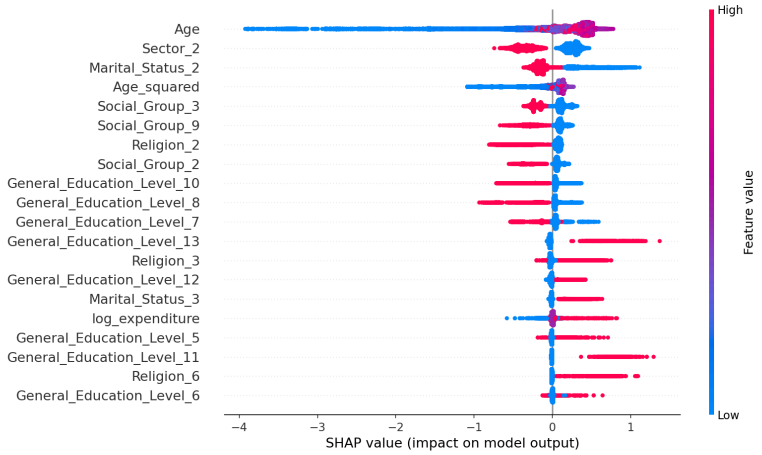


Figure 6: XGBoost SHAP feature importance plot

Model Comparison

Models are compared against ROC-AUC, Accuracy, Precision, Recall, and F1 Score in Figure 7.

Model	**Logistic Regression**	**Random Forest**	**XGBoost**
ROC-AUC	0.7475	0.7601	0.7625
Accuracy	0.7054	0.8081	0.7183
Precision	0.6063	0.6068	0.6183
Recall	0.4395	0.4661	0.4908
F1 Score	0.5096	0.5272	0.5472

Figure 7: Model Comparison

DISCUSSION

Across all three models, marital status, sector of residence, and age-related variables consistently emerge as the strongest predictors of female labour force participation. In the logistic regression, Marital_Status_2 (currently married) exhibits the largest negative coefficient (−1.025), corresponding to an odds ratio of approximately 0.36, indicating substantially lower predicted participation relative to never-married women. Sector_2 (urban residence) also shows a strong negative association (−0.69). Age enters positively, while Age_squared is negative, implying a nonlinear relationship consistent with a lifecycle pattern in which participation rises with age, stabilises, and declines at older ages.

Tree-based models capture similar relationships but differ in how importance is assigned. In the Random Forest, permutation importance identifies marital status, sector, and age as key predictors. In XGBoost, Age_squared (0.129), Marital_Status_2 (0.099), Age (0.086), and Sector_2 (0.071) are among the most influential features. Importantly, SHAP values provide directional insights, confirming that being currently married and residing in urban areas are associated with lower predicted participation, while widowed and divorced categories contribute positively. Age exhibits nonlinear effects across models, reinforcing the lifecycle interpretation.

A notable difference across models lies in the relative importance of age-related variables. While age plays a role in logistic regression, it becomes more dominant in tree-based models. This reflects differences in model structure: logistic regression captures only linear and pre-specified nonlinear effects, whereas tree-based models flexibly detect nonlinearities and interactions. The greater importance of age in these models therefore suggests that lifecycle effects are complex and interaction-driven rather than weak or inconsistent.

In contrast, marital status and sector show strong consistency across models. Both linear and nonlinear approaches identify these variables as central predictors, and their direction of association remains stable across logistic coefficients and SHAP values. These patterns are also consistent with descriptive statistics, where participation is lower in urban areas (27.09%) than in rural areas (35.47%), and varies substantially across marital categories.

Education, religion, and social groups contribute to prediction but with lower and less consistent importance. While logistic regression indicates a nonlinear relationship between education and participation, these variables are not among the top predictors in tree-based models, suggesting a more limited role in improving predictive accuracy.

These findings highlight an important distinction between predictive importance and causal interpretation. Feature importance indicates which variables the model relies on most to improve prediction, but does not by itself establish direction or causality. The use of SHAP values helps recover directional effects in nonlinear models, allowing for more consistent interpretation across approaches. At the same time, differences in variable ranking across models indicate that importance is conditional on model structure.

In terms of predictive performance, non-linear models outperform logistic regression across recall and F1-score, indicating that female labour force participation is shaped by complex, non-linear relationships not fully captured by linear specifications. However, across all models, recall remains below 0.5, suggesting that a substantial proportion of actual participants are not identified. While precision remains relatively stable, this imbalance indicates that models are more effective at confirming participation than detecting it. Among the models, XGBoost performs best across metrics.

Overall, these results suggest that female labour force participation is more strongly associated with structural constraints and labour market conditions than with individual characteristics alone, while also highlighting the limitations of predictive models in fully capturing participation behaviour. This aligns with Klasen and Pieters (2015), who highlight the role of social norms and labour market conditions in

limiting women's participation, particularly in urban settings.

CONCLUSION

First, across all models, marital status, sector of residence, and lifecycle variables consistently emerge as the strongest predictors of female labour force participation. This indicates that these factors carry the most predictive information within the dataset. In contrast, individual characteristics such as education, religion, and social group contribute to prediction but with lower and less consistent importance. Taken together, these results suggest that participation is more strongly associated with household context, labour market conditions, and lifecycle dynamics than with individual attributes alone, although these relationships should be interpreted as predictive rather than casual effects.

Second, the analysis highlights that the interpretation of importance depends on the modeling approach. Logistic regression provides clear directional estimates, showing that being currently married and residing in urban areas are associated with lower predicted participation, while lifecycle effects follow a nonlinear pattern. In contrast, tree-based models assign relatively greater importance to variables such as age and age squared, reflecting their ability to capture nonlinearities and interactions with directional insights. Together, these approaches show that while key variables remain consistent, their relative importance and interpretation vary across models. At the same time, the models do not fully explain participation outcomes, indicating that additional factors such as social norms, intra-household decision-making, and local labour market conditions may also play an important role but are not directly captured in the dataset.

Third, model performance reveals an important limitation. Although non-linear models outperform logistic regression, recall remains below 0.5 across all models, indicating that fewer than half of actual participants are correctly identified. This suggests that female labour force participation is inherently difficult to predict using observable characteristics alone, and that important determinants may not be fully captured in the available data.

This opens scope for further research, including the incorporation of additional variables, the use of alternative modeling approaches, and analysis across different regions or time periods. Future work could also further examine model robustness and explore methods that better integrate predictive and causal analysis. Female labour force participation remains a complex and multifaceted field characterized by predictive patterns linked to societal and lifecycle factors, and this study demonstrates how machine learning can be used to identify consistent predictive patterns while also highlighting the limitations of such approaches.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to Ms. Rashi Sabherwal at the University of Pennsylvania for her insightful feedback and for directing me to key literature that informed this study.

REFERENCES

National Statistics Office. (2025). Periodic Labour Force Survey (PLFS), Key Employment Unemployment Indicators for (January 2024 – December 2024) (DDI-IND-NSO-PLFS-2024-24) [Data set]. Ministry of Statistics and Programme Implementation, Government of India. https://microdata.gov.in/NADA/index.php/catalog/254

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Müller, A., Nothman, J., Louppe, G., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830. https://jmlr.org/papers/v12/pedregosa11a.html

Mehrotra, S., & Parida, J. K. (2017). Why is the labour force participation of women declining in India? World Development, 98, 360–380. https://doi.org/10.1016/j.worlddev.2017.05.010

Deshpande, A., & Singh, J. (2021). Dropping out, being pushed out or can't get in? Decoding declining labour force participation of Indian women (IZA Discussion Paper No. 14639). Institute of Labor Economics (IZA). https://docs.iza.org/dp14639.pdf

Klasen, S., & Pieters, J. (2015). What explains the stagnation of female labor force participation in urban India? The World Bank Economic Review, 29(3), 449–478. https://doi.org/10.1093/wber/lhv003