The Impact of Short-Term Paid Parental Leave Policies on Women's Labor Force Participation in the United States

Mekhalaa Muraly

December 20, 2024

1 Introduction

1.1 Motivation

Parenthood, especially motherhood, is a key factor behind gender gaps in employment and pay, leading to inequalities in the labor market. Women are much less likely to be employed after having children compared to men. Evidence suggests that women are less likely to be employed after having a child, especially compared to men. In 2014, in the U.S., the employment rate for mothers with children under 18 was approximately 23.5 percentage points lower than that of fathers with children under 18 (US Bureau of Labor Statistics, 2015). The gender pay gap also emerges at motherhood, as many women reduce paid work to take on more caregiving at home (Goldin and Mitchell, 2017; Kleven et al., 2019). Policymakers and researchers have highlighted this issue, pointing to the lack of paid family leave in the United States as a significant barrier to achieving gender equity in the workforce. Paid leave helps mothers recover from childbirth, lowers stress, and improves well-being during early parenthood (Adema et al., 2016). While most OECD countries provide paid leave and invest in childcare, the U.S. offers minimal support, including unpaid leave, child tax credits, and limited public childcare facilities, leaving it behind other developed nations on key social and economic indicators (Adema et al., 2016).

The economic and health benefits of paid leave are significant. Paid family leave encourages women to stay in the workforce by providing job security, allowing them to return to their previous positions after childbirth. This stability is beneficial not only for parents but also for businesses, which can avoid the expenses associated with hiring and training new employees (Gault et al., 2014). Recognizing these benefits, California and New Jersey introduced Paid Family Leave policies in 2004 and 2009, respectively. These policies offer income support for up to six weeks, allowing parents to bond with their young children or care for sick dependents (Byker, 2016). Payments are funded through state

Temporary Disability Insurance (TDI) programs, providing eligible workers with a portion of their wages, subject to a maximum limit. For instance, in 2014, California's payment rate was 55% of gross earnings, capped at $1,075 per week, while New Jersey's rate was 66% of gross earnings, with a maximum of $595 per week (Adema et al., 2016).

Despite state-level progress, starting with California in 2004 and now including thirteen states and the District of Columbia with paid family leave policies, the U.S. still falls behind other OECD countries (Byker, 2016). Policies in states like California and New Jersey provide only up to six weeks of partial wage replacement, showing a clear gap in generosity. Most private employers in the U.S. do not offer paid leave for childbirth, and low-wage workers—who are least able to afford unpaid leave—are often excluded when paid leave is available (Gault et al., 2014). This inequity highlights the need for a comprehensive paid leave policy that supports all parents, especially mothers.

In this paper, I use data from the Survey of Income and Program Participation (SIPP) Panel Longitudinal Files, spanning 1996 to 2008, to evaluate the impact of paid parental leave laws implemented in California (California Paid Family Leave Act, CPFL) in 2004 and New Jersey (New Jersey Family Leave Act, NJFLA) in 2009 on women's labor force participation following childbirth. To analyze the effects of these policies on mothers' employment status up to 24 months after childbirth in California and New Jersey, I apply several machine learning methods, including Lasso, Tree-Based Models, Directed Acyclic Graphs, Matching, Inverse Probability Weighting, Meta Learners, Doubly Robust Methods, Double Machine Learning, and Causal Forests.

Using the identification strategy outlined by Byker (2016), I employ a range of new methodological approaches, including causal machine learning techniques, to analyze the policy effects. By examining the nuances of each method, I aim to provide a comprehensive understanding of the policy impacts and offer insights into their broader implications for labor market policies.

My results demonstrate that feature selection models, including Lasso, Ridge, and Random Forest, underscore the significance of demographic variables such as education and age in shaping policy impacts. Among the tree-based models, Random Forest outperforms others in capturing non-linear interactions, identifying policy intervention as a critical predictor. These insights guide subsequent causal analyses. I implement the Dynamic Acyclic Graph (DAG) model, which ensures correct model specification by identifying causal pathways and potential confounders. Propensity score matching estimates modest positive effects on labor force participation, while inverse probability weighting and doubly robust methods suggest slight reductions, highlighting the importance of triangulating results. Furthermore, advanced causal models such as Double Machine Learning and Causal Forests reveal minimal average treatment effects but substantial heterogeneity across subgroups. For example, younger women experience modestly negative impacts, whereas older and highly educated women benefit significantly, reflecting life-cycle and education-based disparities. Meta-learners, including S-Learner, T-Learner, and X-Learner, further show nuanced subgroup effects, emphasizing education and age as key moderators. These findings illuminate the complex interplay between policy impacts and individual characteristics, offering robust and policy-relevant insights into the design and implementation of paid family leave policies.

1.2 Literature Review

Studies using empirical methods to account for non-random selection into paid parental leave policies have found that, despite barriers such as limited awareness, eligibility restrictions, and lack of job protection, introducing or expanding leave programs increases parental leave-taking globally (Rossin-Slater, 2017). Furthermore, leave entitlements of less than one year can enhance women's job continuity and improve long-term employment rates. However, longer leave durations may negatively affect women's earnings, employment, and career advancement (Rossin-Slater, 2017).

Research on paid family leave in the U.S. has been relatively limited, largely due to the lack of policy variation and data constraints. Several studies have leveraged differences in the timing of the Family and Medical Leave Act (FMLA) implementation. The FMLA of 1993, the nation's first and only federal mandate, provides qualified employees at eligible firms with up to 12 weeks of unpaid, job-protected leave per year for medical reasons, including childbirth. Evidence suggests that the FMLA increased leave-taking, particularly among more educated, married women; however, unpaid leave provisions had minimal to no measurable effect on employment levels (Waldfogel, 1999; Han et al., 2009; Baum II, 2003).

Recent research using survey data and difference-in-differences methods generally indicates that the CPFL program led to short-term improvements in employment and wage outcomes (Baum and Ruhm, 2016; Byker, 2016). Additionally, Jones and Wilcher (2024) concluded that the introduction of paid family leave in California helps to maintain a skilled workforce, as the policy increased women's labor participation by over five percentage points in the year of childbirth, with positive effects lasting up to nine years. The impact was particularly notable among women with bachelor's degrees, who experienced a 12 percentage point reduction in maternal detachment in the year of birth. However, contrasting findings exist. Das and Polachek (2015) reported increases in both unemployment rates and the duration of unemployment. Similarly, Bailey et al. (2019) found that the CPFL program did not result in increased employment or earnings; instead, it contributed to a decline in employment and earnings approximately ten years after childbirth.

These contradictory findings are suggestive rather than conclusive, partly due to small sample sizes that constrain the ability to draw robust conclusions about the long-term effects of family leave policies. Moreover, the findings may reflect the presence of extreme heterogeneous treatment effects, which pose challenges for interpreting the results. Overall, existing research offers mixed and limited evidence regarding the impacts of state-level paid parental leave programs on mothers' labor market outcomes in the U.S.

2 Data and Identification Strategy

2.1 Data and Summary Statistics

The data for this study are drawn from the Survey of Income and Program Participation (SIPP) Panel Longitudinal Files, covering the period from 1996 to 2008. This dataset is publicly available through the Inter-university Consortium for Political and Social Research (ICPSR), affiliated with the University of Michigan. Monthly panel data from the years 1996, 2001, 2004, and 2008 are combined to create a dataset that includes nationally representative 48-month panel surveys. These large sample sizes are well-suited for evaluating state-level policy effects.

The analysis focuses on individual-level data of women aged 24 to 45 who gave birth during the survey period, identified using birth month and mother-child relationship variables. The dataset consists of 103,624 observations with 12 variables, providing detailed information on individuals and their employment characteristics, household identifiers, demographic information, and state-level data. Hence, it serves as a rich source of variables for the empirical model examining women's labor force attachment before and after the introduction of paid leave legislation in California and New Jersey.

In this analysis, the main outcome variable is 'Labor Force Participation'. This variable is a binary indicator derived from the original employment status variable, indicating whether individuals were actively engaged in the labor force. Individuals who are not attached to the labor force are coded as zero, while those who are engaged in any labor activity are coded as one. For instance, in the SIPP data, a woman who leaves her employer to care for her child or a woman who has had no job during the month in which the survey is conducted and has spent no time looking for work is classified as being detached from the labor force, and is coded as zero. Conversely, a woman who takes job-protected parental leave remains part of the labor force. She is recorded as having a job, even if she is not actively working, and is coded as one. Home production activities are not recorded in the SIPP survey data and, therefore, do not count as attachment to the labor force. The latter retains firm-specific skills and tenure and avoids the costs associated with searching for a new job, whereas the former forfeits these benefits by severing ties with her employer.

The treatment variable is defined by the interaction of policy implementation dates with individuals' state of residence. The treatment variable 'PostPolicy$_t$' is set to one if a birth occurred during a period when a paid leave policy was in effect in the individual's state, and zero otherwise. This analysis specifically considers the paid leave policies in California and New Jersey, implemented in July 2004 and July 2009, respectively. Therefore, the treatment group includes individuals residing in California or New Jersey with a child born after the policy implementation date in their state. Women who gave birth in Texas, Florida, or New York serve as the control group. Additionally, the analysis includes several covariates to control for confounding factors, such as individual and state fixed effects, to account for differences across individuals and variations in economic environments across states.

Figure 1 below presents a comparison between female labor force non-participation and participation across the entire dataset. The results indicate that the number of women participating in the labor force (65,793) is significantly higher compared to those not participating (37,410). This suggests that the majority of women represented in the dataset are economically active, although a substantial minority remains unattached to the labor force. The proportion of women who are not attached to the labor force highlights the need to explore mechanisms that could foster increased participation, such as maternity leave policies. These statistics provide a foundation for understanding broader trends in female labor force engagement and offer context for assessing the impact of CPFL and NJFLA policies.

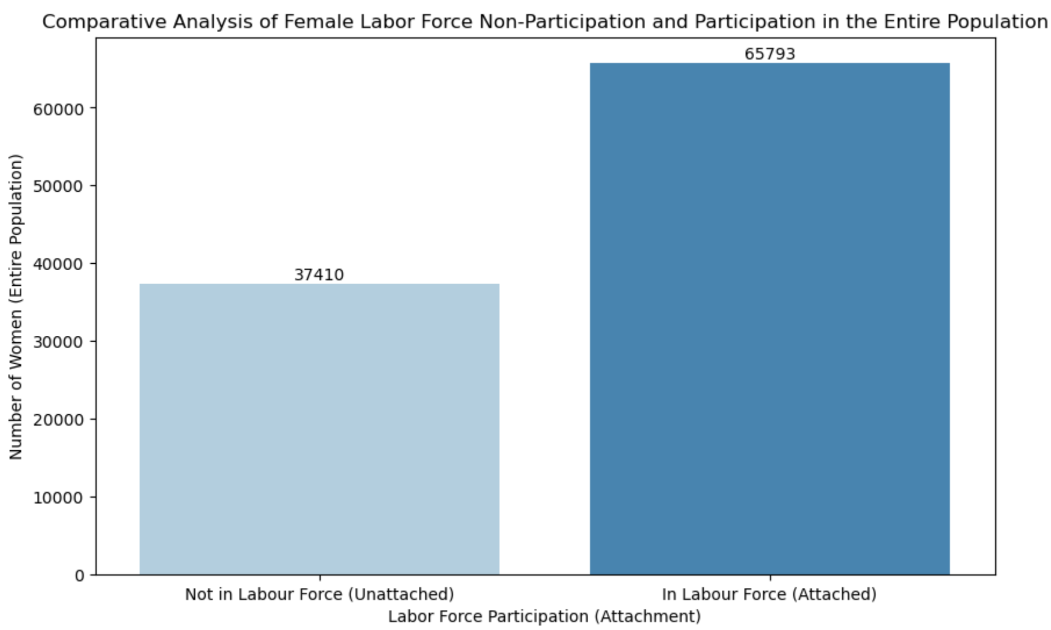

Figure 1: Comparative Analysis of Female Labor Force Non-Participation and Participation

Figure 2 below shows labor force participation rates across various educational attainment levels. Women with "Less than High School" education exhibit the lowest participation rate, approximately 35%. Conversely, those with a "Graduate Degree" or "Bachelor's Degree" demonstrate the highest participation rates, nearing 80%. There is a clear positive trend between educational attainment and labor force participation, suggesting that higher education is strongly correlated with greater labor force attachment, likely due to better employment prospects and higher earning potential. These findings reveal that education level contributes to the differential effects of paid leave policies. Education level may act as a moderator of policy impact; women with higher education levels may be more likely to stay in or re-enter the labor force after childbirth if paid leave is available, while those with lower educational attainment may not benefit to the same extent due to other barriers, such as limited job opportunities or childcare constraints. These differences in labor force participation across educational levels could also explain the presence of heterogeneous treatment effects. Women with lower educational attainment may face structural disadvantages that limit their ability to take full advantage of paid leave policies, while those with higher education may be better positioned to leverage such policies for labor force reattachment.

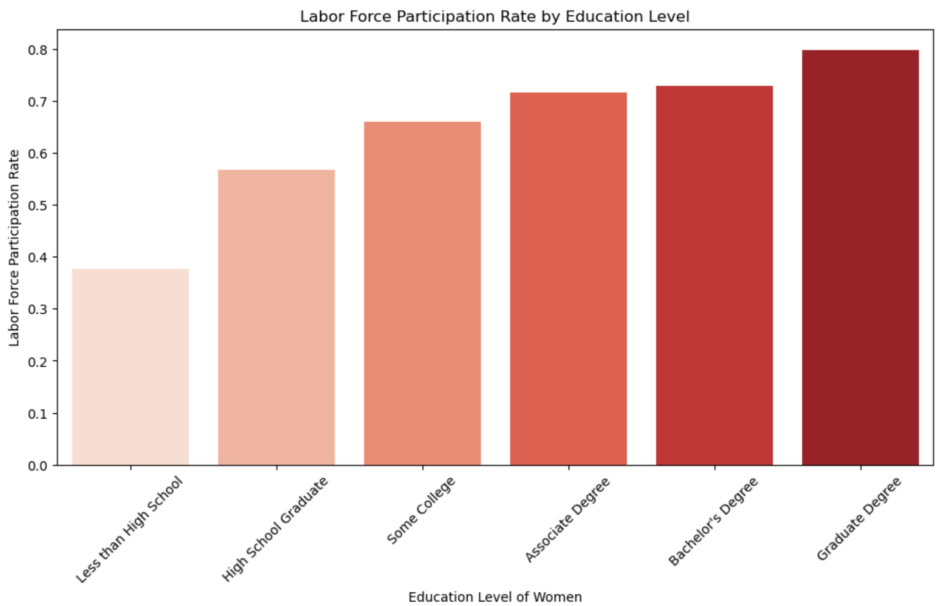

Figure 2: Comparative Analysis of Labor Force Participation Rate by Education Level

2.2 Identification Strategy

To estimate the causal impact of CPFL and NJFLA policies on labor force participation, I utilize the identification strategy outlined in Byker (2016)'s approach. The identification strategy leverages variations in the timing of policy implementation across different states, serving as the foundation for applying causal machine learning models. Specifically, California and New Jersey serve as treated states, implementing paid family leave policies in 2004 and 2009, respectively. The other states in the dataset, including New York, Florida, and Texas, act as control states, where these policies were not introduced during the observation period. This approach facilitates a comparison of labor force participation changes between treated and control states before and after the policy implementation. The key assumption in this identification strategy is that, in the absence of the intervention, labor force participation trends would have been parallel across treated and control states. Under this assumption, any observed differences in labor force participation in treated states post-implementation can be attributed to the paid leave policies, after accounting for other observable factors. Furthermore, to mitigate potential biases, I incorporate both individual- and state-level fixed effects in the relevant models. Individual fixed effects, using each respondent's unique identifier, control for time-invariant unobservable traits that may influence labor force participation, such as personal preferences or motivations. State-level fixed effects account for time-invariant state-specific factors affecting labor market outcomes, such as economic conditions or policy environments.

To establish the identification strategy, I examine trends and changes in labor force participation rates over time, distinguishing between treated and control states. These patterns are presented in Figures 3 and 4, and summarized in Table 1 below. In California, the data indicate a notable response to the implementation of the CPFL policy in July 2004. As shown in Table 1, the number of women not participating in the labor force decreased significantly from 8,558 pre-policy to 5,529 post-policy. At the same time, the number of women participating in the labor force dropped from 13,431 to 9,376. Figure 3 aligns with this, showing a modest upward trend in labor force participation rates post-policy. The decrease in women not in the labor force, coupled with the observed trends, suggests that paid family leave in California facilitated re-entry into the labor force for some women after childbirth.

In contrast, the impact of the New Jersey Family Leave Insurance (NJFLA) policy, implemented in July 2009, appears to be more modest. As Table 1 shows, the number of women in the labor force declined from 4,751 pre-policy to 1,154 post-policy. Similarly, the number of women not in the labor force decreased from 3,132 to 368. Figure 4 illustrates only a slight increase in labor force participation rates post-policy, suggesting that the NJFLA policy did not produce as substantial an impact as the CPFL policy in California. This modest effect may stem from differences in policy design, lower wage replacement rates, or other economic and social barriers influencing women's ability to re-enter the labor force in New Jersey.

Overall, the trends show that labor force participation in California increased noticeably after the implementation of CPFL, while the rise in New Jersey following NJFLA was more modest. Control states, including Florida, New York, and Texas, display fluctuations in participation rates, but no significant changes align with the policy implementation periods in the treated states. These distinct patterns in California and New Jersey highlight the influence of policy-specific factors and state-level contexts on outcomes.

Table 1: Number of Women in Not in Labor Force vs. In Labor Force by State and Policy Period

		Pre-Policy		Post-Policy	
Group	State	Not in Labor Force	In Labor Force	Not in Labor Force	In Labor Force
Treated States					
	California	8558	13431	5529	9376
	New Jersey	3132	4751	368	1154
Control States					
	Florida	2731	5390	1738	4320
	New York	3967	7021	1941	4784
	Texas	5936	9086	3856	6539

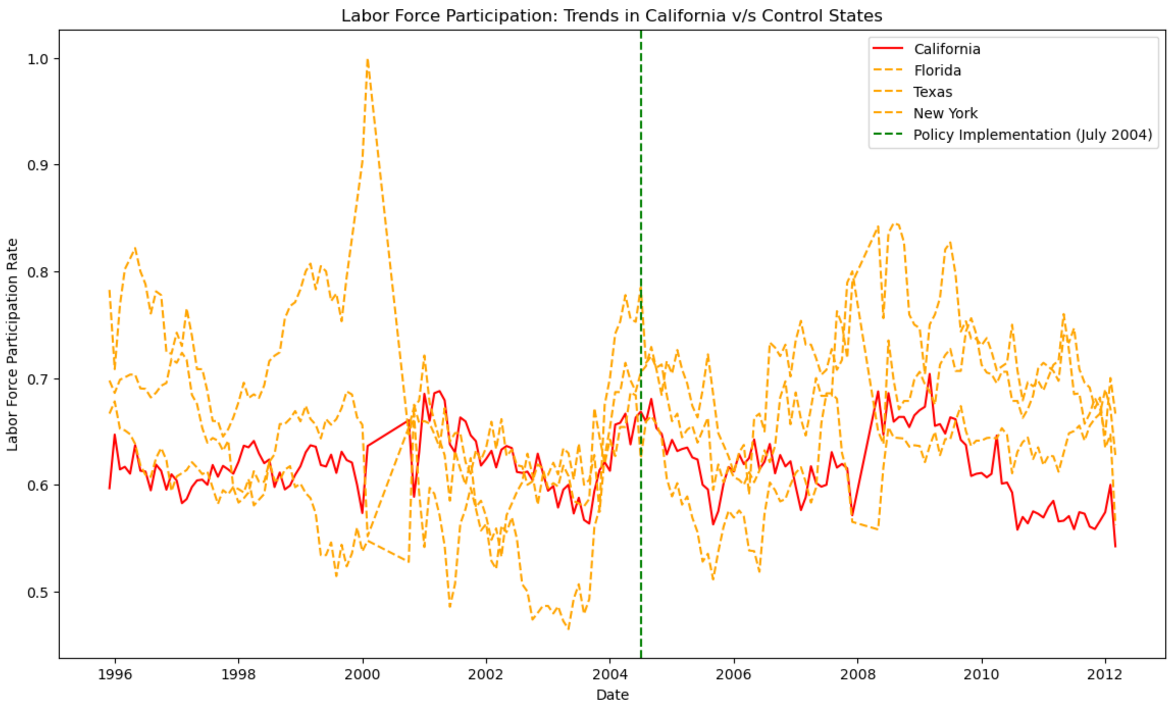

Figure 3: Labor Force Participation Rate in California v/s Control States Over Time

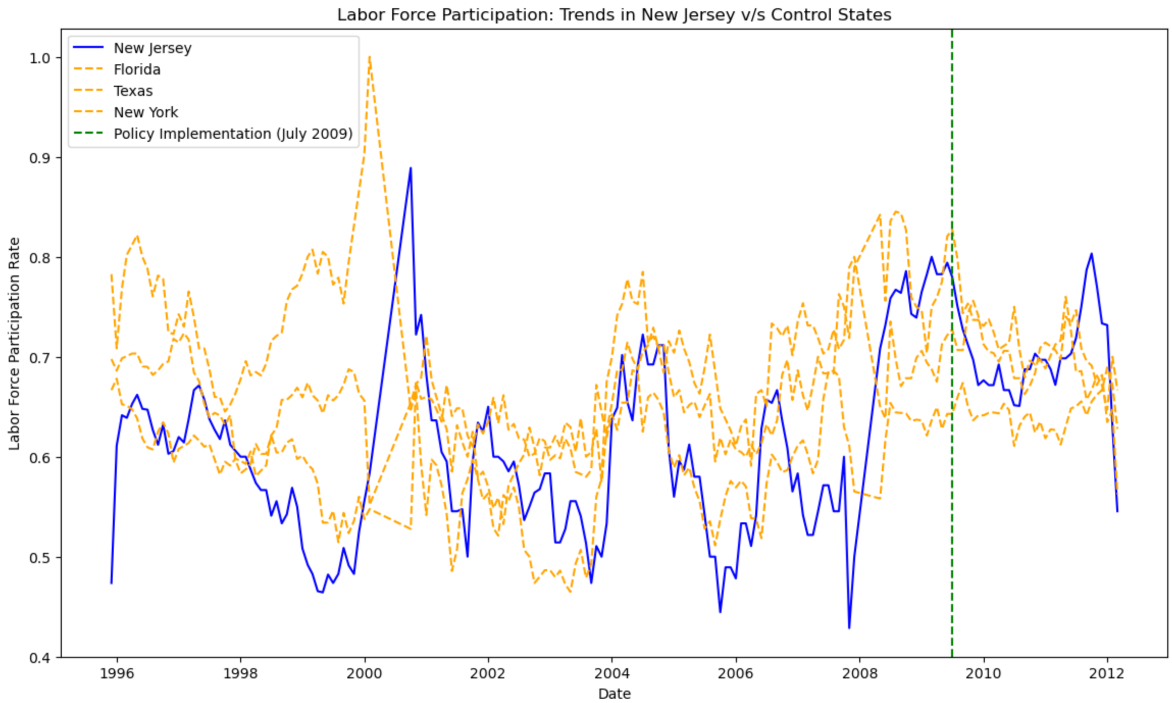

Figure 4: Labor Force Participation Rate in New Jersey v/s Control States Over Time

3 Model Specification and Results

3.1 Feature Selection Models

3.1.1 Feature Selection Models: Lasso

In this study, the Lasso model selects relevant features, addresses multicollinearity, and streamlines the model by shrinking less significant variables. The estimating equation is:

$$\begin{aligned}\underset{\boldsymbol{\beta}}{\text{minimize}} \quad & -\frac{1}{n}\sum_{i=1}^{n}\Big[Y_i \log\left(\sigma\left(\beta_0 + \beta_1(\text{Treated}_i \cdot \text{PostPolicy}_t) + \mathbf{X}_i\boldsymbol{\beta}\right)\right) \\ & + (1-Y_i)\log\left(1-\sigma\left(\beta_0 + \beta_1(\text{Treated}_i \cdot \text{PostPolicy}_t) + \mathbf{X}_i\boldsymbol{\beta}\right)\right)\Big] \\ & + \lambda\sum_{j=1}^{p}|\beta_j|\end{aligned} \tag{1}$$

The estimating equation minimizes the negative log-likelihood for logistic regression, modeling the probability of labor force participation ($Y_i = 1$). Predictors include an intercept term (β_0), the interaction of treatment status and post-policy period ($\text{Treated}_i \cdot \text{PostPolicy}_t$), and covariates ($\mathbf{X}_i$), such as demographic features ('Age', 'Level of Education', 'Sex') and state fixed effects. The Lasso penalty ($\lambda \sum_{j=1}^{p} |\beta_j|$) shrinks less relevant coefficients (β_j) toward zero. The regularization parameter (λ) balances model complexity and ensures robust estimates.

The results of the Lasso model are also illustrated in two graphs below. Figure 5 depicts the standardized coefficients of each feature as a function of $-\log(\lambda)$. Variables like 'Level of Education' demonstrate a prominent and stable effect, highlighting its strong predictive value. The interaction term ($\text{Treated}_i \cdot \text{PostPolicy}_t$) retains its influence, suggesting the policy's consistent impact on labor force participation across treated states. State-level variables, such as 'Florida' and 'New York', also show meaningful contributions, though at smaller magnitudes. In contrast, variables like 'Texas' shrink to zero quickly, indicating limited relevance in explaining labor force outcomes. The shrinkage effect of the Lasso model helps in feature selection by eliminating less important variables as regularization increases. This enhances model interpretability and minimizes overfitting, particularly with numerous covariates. By selecting only the most relevant variables, the model achieves a parsimonious specification that is well-suited for policy analysis. In addition, Figure 6 demonstrates the cross-validated accuracy of the Lasso model as a function of $-\log(\lambda)$. The tuned regularization parameter (α) corresponds to the region of the plot where accuracy is maximized, approximately $-\log(\lambda) \sim 2$. The model achieves stable cross-validated accuracy (~ 0.67) across a wide range of regularization strengths, indicating its robustness. However, accuracy declines at very high levels of regularization ($-\log(\lambda) > 4$), where excessive shrinkage excludes relevant predictors. The use of 5-fold cross-validation strikes a balance between computational efficiency and reliable error estimation, ensuring that the selected model generalizes well to unseen data.

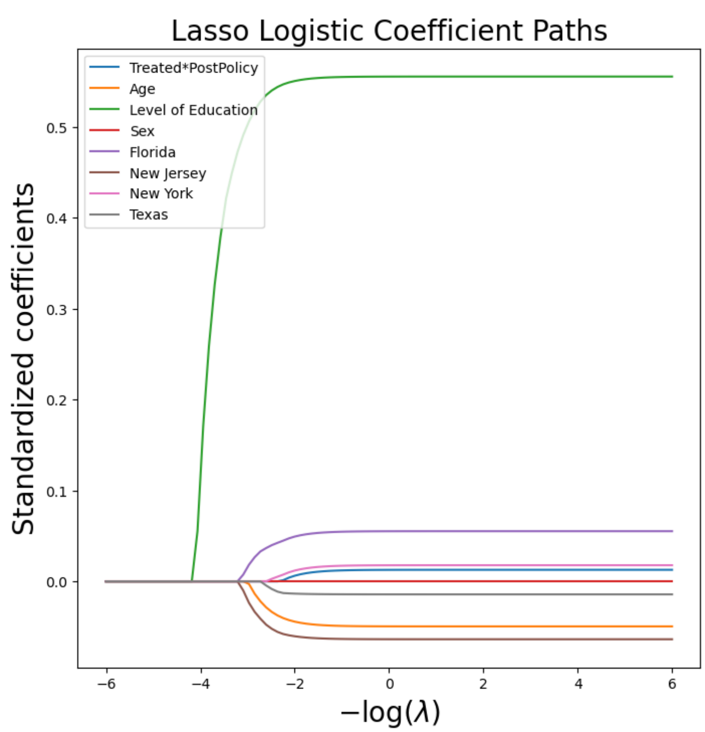

Figure 5: Coefficient Paths for Lasso Model

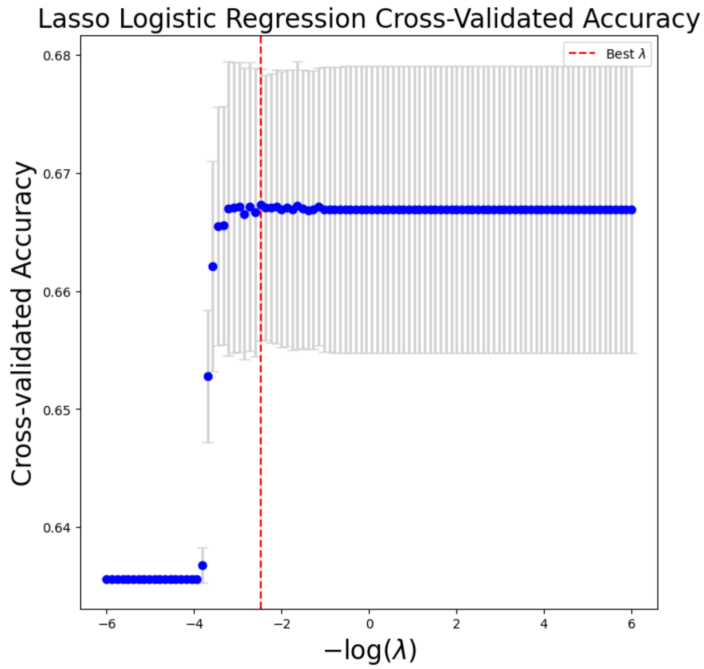

Figure 6: Cross-validated accuracy of the Lasso Model

3.1.2 Feature Selection Models: Ridge

In contrast to Lasso, the Ridge model facilitates feature selection by regularizing all coefficients, shrinking them closer to zero without excluding any, making it ideal for handling multicollinearity.

Figure 7 illustrates the standardized coefficient paths of the Ridge model as a function of $-\log(\lambda)$. Unlike Lasso, Ridge does not eliminate variables but reduces their influence as λ increases. 'Level of Education' exhibits a consistently strong positive effect, reaffirming its importance as a predictor of labor force participation. Similarly, the interaction term ($\text{Treated}_i \cdot \text{PostPolicy}_t$) retains its influence, highlighting the impact of paid family leave policies in treated states. State-level variables, such as 'Florida' and 'New Jersey', contribute moderately, whereas 'Texas' and 'Age' show weaker effects, indicating limited relevance. Ridge's shrinkage ensures all variables contribute to the model, addressing multicollinearity and providing a holistic understanding of predictors. Figure 8 depicts the cross-validated accuracy as a function of $-\log(\lambda)$. The optimal λ occurs at $-\log(\lambda) \sim -4$, where accuracy is maximized at ~ 0.67. The model achieves stable accuracy across a broad range of λ, demonstrating robustness. However, at high regularization levels ($-\log(\lambda) > -2$), excessive shrinkage leads to underfitting and a decline in accuracy. The Ridge model confirms 'Level of Education' and ($\text{Treated}_i \cdot \text{PostPolicy}_t$) as key predictors of labor force participation, while addressing multicollinearity among covariates. Unlike Lasso, Ridge retains all predictors, ensuring a comprehensive analysis.

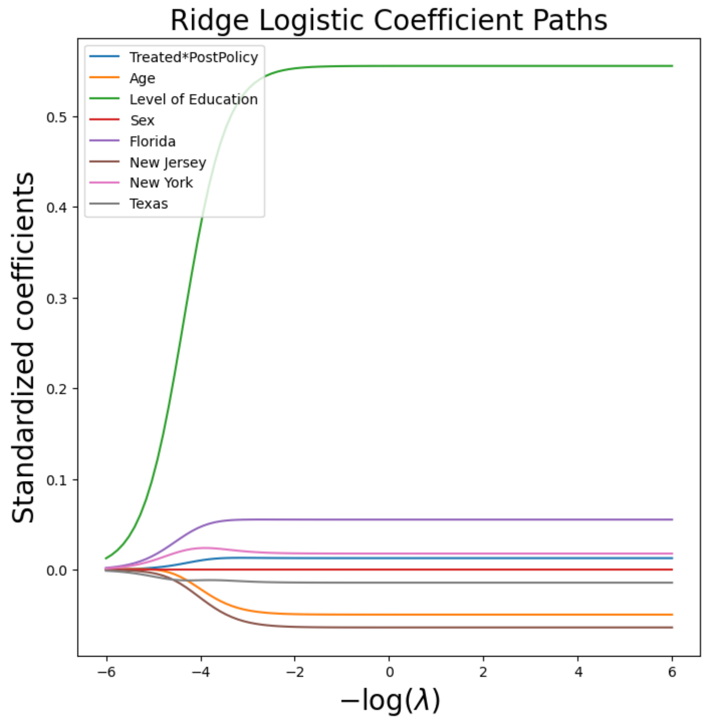

Figure 7: Coefficient Paths for Ridge Model

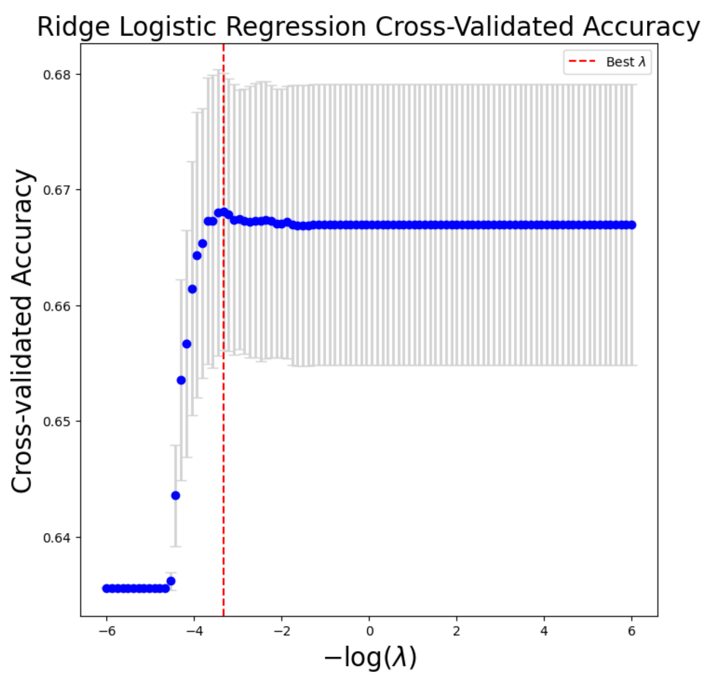

Figure 8: Cross-validated accuracy of the Ridge Model

3.1.3 Feature Selection Models: Random Forests

The Random Forest model effectively captures complex interactions between policy effects and individual traits such as age, education, and state, while addressing non-linear relationships and reducing overfitting through the aggregation of multiple decision trees. The model's Mean Squared Error (MSE) of 0.299 indicates a moderate but reasonable level of deviation between predictions and actual values, reflecting the complexity of the dataset. Figure 9 illustrates the feature importance plot, with 'Level of Education' and 'Age' emerging as the most significant predictors, consistently suggesting their major role in workforce dynamics across feature selection models.

The Random Forest model differs from the Lasso and Ridge models in its treatment of policy-related variables such as ($\text{Treated}_i \cdot \text{PostPolicy}_t$), which are assigned lower importance. While Lasso and Ridge emphasize these variables' direct effects, the Random Forest model embeds their influence within non-linear interactions, potentially diluting their individual importance. This distinction highlights Random Forest's capability to uncover complex relationships that linear models may not fully capture. Furthermore, the Random Forest model exhibits greater variability in state-level feature importance, such as for 'Florida' and 'Texas', compared to the more uniform results observed in Lasso and Ridge. This variability suggests that Random Forests are particularly adept at capturing geographic heterogeneity, which is critical in understanding localized labor force dynamics. In sum, the Random Forest model complements linear approaches by revealing non-linear effects and interactions, while still aligning with Lasso and Ridge in emphasizing key demographic predictors.

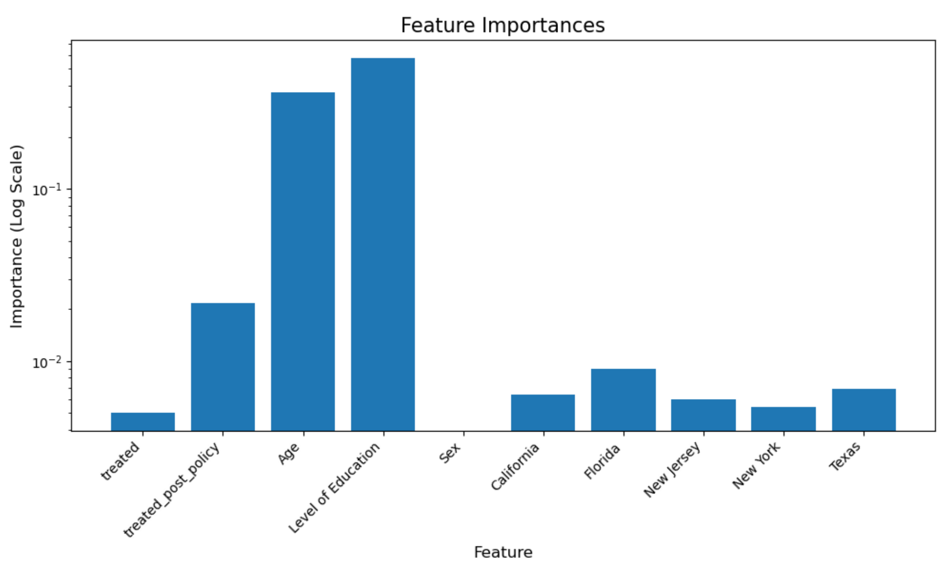

Figure 9: Random Forest Feature Importance Plot

3.1.4 Feature Selection Models: Boosting for Classification Models

The Boosting model produced a Mean Squared Error (MSE) of 0.372 and an accuracy of 62.77%, reflecting moderate predictive performance in modeling labor force participation. While these results indicate that the model captures key patterns in the data, its performance is slightly weaker than that of the Random Forest model, which achieved a lower MSE of 0.299 and higher accuracy. This difference may stem from the nature of Boosting, which prioritizes reducing errors iteratively but can be more sensitive to hyperparameter tuning and prone to overfitting, particularly in resampled datasets. In comparison to the Random Forest model, Boosting may place less emphasis on dominant predictors such as 'Level of Education' and 'Age', as reflected in the feature importance rankings of the Random Forest model. However, the Boosting model continues to highlight the significance of demographic characteristics in shaping labor force outcomes, aligning with findings from other methodologies.

Building on the insights from variable selection, I implement the Directed Acyclic Graphs (DAGs) to define and visualize causal pathways between key variables. The DAG framework is instrumental in clarifying assumptions, identifying potential sources of bias, and informing the structure of the causal models. This step provides a robust theoretical foundation for the estimation of causal effects, ensuring that the analysis is guided by a well-justified causal framework.

3.2 Other Models: Dynamic Acyclic Graphs (DAG)

The results from the Dynamic Acyclic Graph (DAG) model, illustrated in Figure 10, provide a robust framework for estimating the causal relationship between CPFL and NJFLA policies and women's labor force attachment. The assumption of unconfoundedness, which forms the foundation of the DAG model, is supported by the inclusion of relevant covariates that address potential omitted variable bias. In the context of this study, the model incorporates key demographic covariates—Age, Level of Education, and Sex—as confounders influencing both the treatment (treated_PostPolicy) and the outcome (Labour Force Participation). This structure aligns with theoretical expectations and ensures the validity of the backdoor criterion for identifying the Average Treatment Effect (ATE).

As shown in Table 2, the realized estimand evaluates the causal effect of the paid family leave policies on 'Labour Force Participation', adjusted for demographic covariates. The estimated ATE of 0.0013, while modest, suggests that the policy intervention exerts a measurable but limited influence on labor force participation when accounting for confounders. This finding highlights that demographic characteristics, such as 'Education' and 'Age', play a significant role in mediating the policy's impact. Therefore, this causal framework offers a nuanced understanding of how policy interventions interact with individual-level factors to shape labor market dynamics.

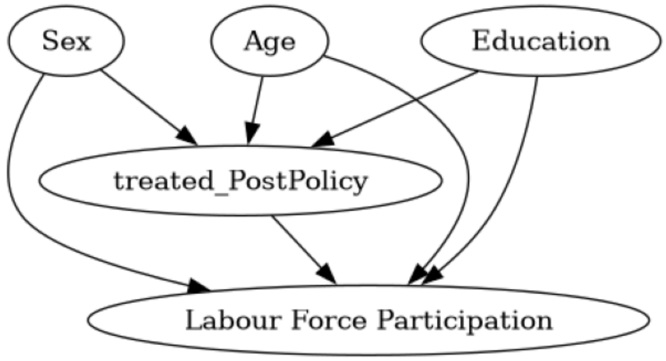

Figure 10: Dynamic Acyclic Graph Model

Table 2: Causal Estimate from DAG Model

Identified Estimand	Nonparametric Average Treatment Effect (ATE)
Estimand Type	Nonparametric ATE
Estimand Name	Backdoor
Estimand Expression	$\frac{d}{d[\text{treated\_PostPolicy}]}E[\text{Labour\_Force\_Participation} \mid \text{Sex, Age, Education}]$
Estimand Assumption	Unconfoundedness: If $U \rightarrow \{\text{treated\_PostPolicy}\}$ and $U \rightarrow$ Labour_Force_Participation, then P(Labour_Force_Participation | treated_PostPolicy, Sex, Age, Education, U) $= P$(Labour_Force_Participation | treated_PostPolicy, Sex, Age, Education)
Realized Estimand	Labour_Force_Participation $\sim$ treated_PostPolicy + Sex + Age + Education
Target Units	Average Treatment Effect (ATE)
Estimate	Mean Value: 0.0013

With the DAG-informed structure, I then implement causal models such as matching, inverse probability weighting (IPW), and doubly robust estimators to estimate the average treatment effects (ATE) of paid family leave policies. These methods provide complementary perspectives, synthesizing localized and population-level effects while addressing potential confounding. Together, these models offer robust evidence of the policy's impact on women's labor force participation, while revealing the limitations of aggregate estimates in capturing nuanced subgroup effects.

3.3 Other Models: Matching, Inverse Probability Weighting and Doubly Robust Models

3.3.1 Matching Methods

The propensity score matching analysis provides robust insights into the causal effect of CPFL and NJFLA paid family leave policies on labor force attachment. Table 3 below presents the balance of covariates before and after matching. We observe that prior to matching, standardized differences between treated and control groups were relatively large for variables such as 'Less than High School' and 'Some College,' indicating an imbalance in the baseline covariates. Post-matching, these differences are substantially reduced across all covariates, suggesting that the matching procedure successfully balanced the treated and control groups. This improvement mitigates potential confounding bias and enhances the credibility of the estimated treatment effect.

Table 3: Balance Table: Covariate Means Before and After Matching

	Pre-Matching		Post-Matching	
Covariate	Mean Treated	Mean Control	Mean Treated	Mean Control
Age	0.0763	-0.0615	0.0763	0.0725
Sex (Male)	0.0000	0.0000	0.0000	0.0000
Bachelor's Degree	0.0205	-0.0165	0.0205	0.0184
Graduate Degree	0.0054	-0.0043	0.0054	0.0054
High School Graduate	-0.0565	0.0455	-0.0565	-0.0565
Less than High School	0.0939	-0.0756	0.0939	0.0975
Some College	-0.0332	0.0267	-0.0332	-0.0344

Figure 11 illustrates the propensity score distribution for women in the treated and control groups. The treated group consists of women residing in California and New Jersey, states that implemented paid family leave policies during the study period. In contrast, the control group comprises women in Florida, New York, and Texas, states without such policies. The overlap in propensity scores between the two groups is evident, satisfying the positivity assumption required for valid causal inference. Notably, the treated group demonstrates a slight concentration of higher propensity scores, reflecting the characteristics of women more likely to be influenced by the policy implementation. Meanwhile, the control group spans a broader range of lower scores, representing individuals in states unaffected by the policy. This overlap ensures that women in the treated group can be adequately matched to similar counterparts in the control group, thereby reinforcing the validity of the matching approach.

Figure 12 presents the bootstrap distribution of the Average Treatment Effect (ATE). The estimated ATE is approximately 0.0142, with a 95% confidence interval ranging from 0.0081 to 0.0203. This result demonstrates a statistically significant, albeit modest, positive effect of paid family leave policies on labor force participation among women in treated states. The symmetric shape of the bootstrap distribution indicates that the model is well-specified, while the narrow confidence interval reflects the precision of the estimate. Taken together, these findings suggest that paid family leave policies exert a measurable impact on women's labor force attachment in California and New Jersey, relative to their counterparts in non-policy states, while accounting for demographic and educational differences. Overall, the propensity score matching results provide compelling evidence that paid family leave policies contribute to a modest increase in women's labor force participation. By incorporating covariates such as education, age, and sex, this analysis highlights the critical role of demographic factors in shaping labor market outcomes and supports the validity of the causal estimates.

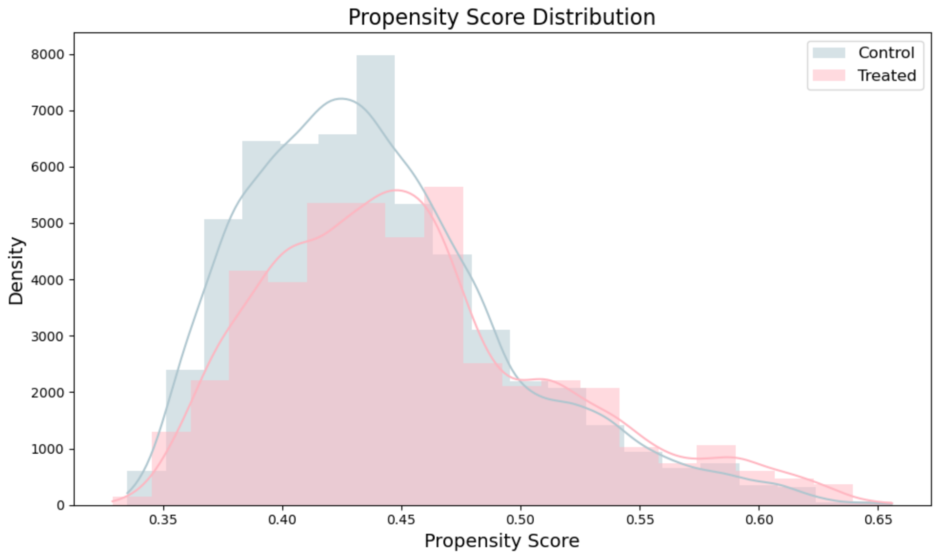

Figure 11: Matching Model: Propensity Score Distribution Plot

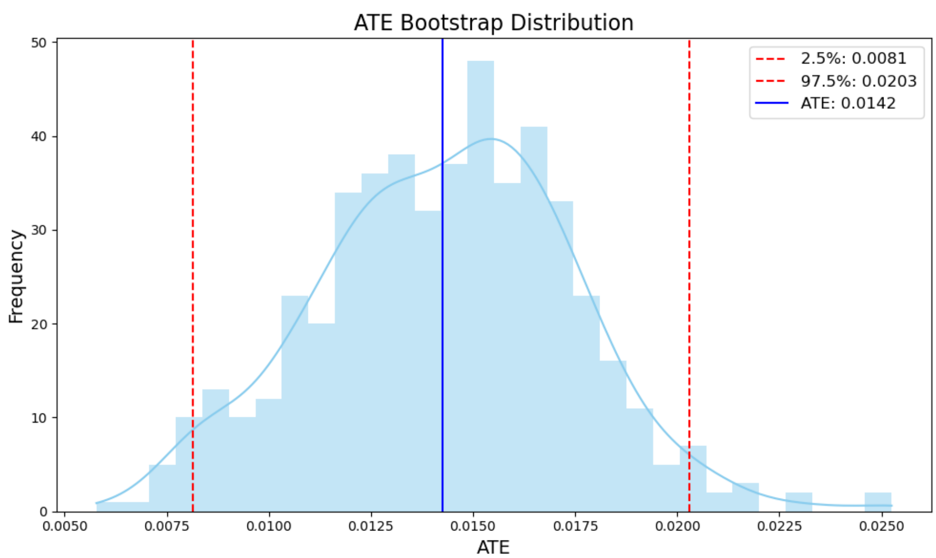

Figure 12: Matching Model: ATE Bootstrap Distribution Plot

3.3.2 Other Models: Inverse Probability Weighting (IPW)

The results from the Inverse Probability Weighting (IPW) model provide an alternative lens for understanding the impact of paid family leave policies on women's labor force participation. As shown in Figure 13 below, the positivity check graph (propensity score distribution) demonstrates sufficient overlap between treated women in California and New Jersey and the control group, satisfying the positivity assumption required for valid causal inference. The treated group exhibits a higher concentration of propensity scores at the upper range, while the control group spans a broader range of lower scores. This overlap ensures that the IPW model can appropriately reweight the data to mimic a randomized experiment.

The estimated ATE from the IPW model is -0.0142, with a 95% confidence interval ranging from -0.0200 to -0.0082 (Figure 14). This finding suggests a statistically significant but modest reduction in women's labor force participation following the introduction of paid family leave policies. Notably, this negative effect contrasts with the results of the matching model, which reported a positive ATE of 0.0142. It is plausible that the matching and IPW models provide complementary insights into the effects of paid family leave policies. The positive effect observed in the matching model may reflect localized impacts on women closely matched on key covariates, such as age, education, and sex. In contrast, the IPW model's negative effect may capture broader population-level trends, accounting for heterogeneity across states and unmeasured systemic factors influencing women's labor force behavior. Together, these findings underscore the value of triangulating results across multiple causal inference methods to achieve a nuanced understanding of how CPFL and NJFLA interventions influence women's labor force attachment.

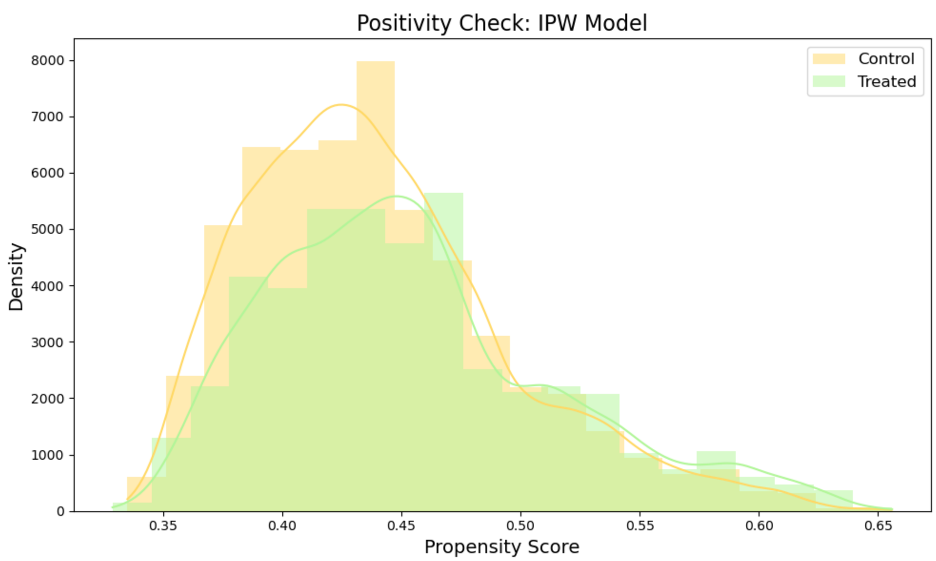

Figure 13: IPW Model: Positivity Check Plot

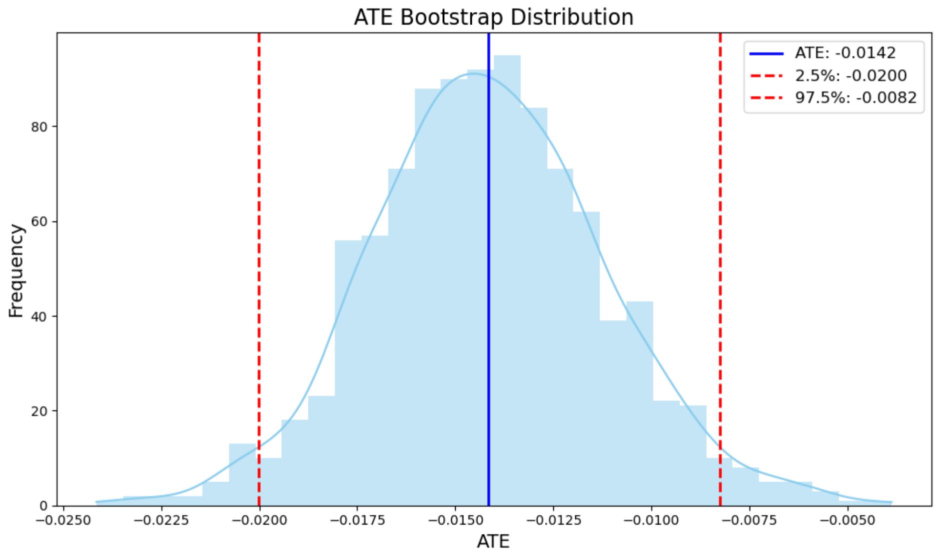

Figure 14: IPW Model: ATE Bootstrap Distribution Plot

3.3.3 Other Models: Doubly Robust (DR) Model

The analysis of the doubly robust model focuses on the covariates of education and gender to explore the heterogeneity in policy effects, informed by their prominence in feature selection from machine learning models like LASSO, Ridge, and Random Forests. These methods consistently identified education and gender as key predictors of labor force participation, highlighting their structural importance in labor market dynamics. By incorporating these covariates, the doubly robust model offers a deeper understanding of how the CPFL and NJFLA policy impacts vary across demographic groups. The DR model is more nuanced as it combines the propensity score modeling and outcome regression to estimate the ATE. The estimated ATE of -0.0140, with a 95% confidence interval ranging from -0.0181 to -0.0089 (Figure 15 below), shows a statistically significant reduction in women's labor force participation associated with the paid family leave policies. The precision of the confidence interval highlights the robustness of the DR approach, leveraging dual-model specifications to account for potential biases in the treatment assignment and outcome processes.

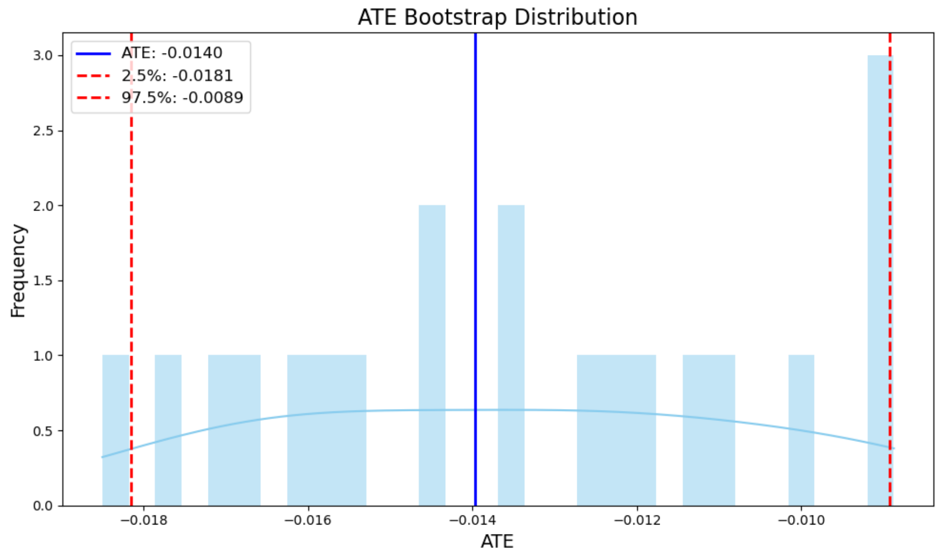

Figure 15: Doubly Robust Model: ATE Bootstrap Distribution

The positivity check for education levels (Figure 16) demonstrates a reasonable overlap in propensity scores across all education categories, including 'Bachelor's Degree', 'Graduate Degree', and 'Less than High School'. The alignment of median propensity scores between treated and control groups across education strata suggests adequate balance, ensuring comparability. However, the variability in propensity scores for higher education categories, such as 'Graduate Degree', reflects underlying heterogeneity in policy impact across educational attainment levels – women with higher educational attainment experience stronger positive effects of the policy. This suggests that the policy disproportionately benefits more educated women, potentially due to better access to job-protected leave or higher baseline labor force participation rates. Similarly, the distribution of age by treatment status (Figure 17) reveals overlapping patterns between treated and untreated groups, satisfying the positivity assumption. While younger age groups show slightly higher density among untreated women, older age groups exhibit comparable distributions. This balance supports the validity of the DR estimates by mitigating confounding effects attributable to age differences. The plot also reveals that younger women in the treated group exhibit a higher propensity for labor force attachment under the policy compared to their untreated peers, likely reflecting their stage in life where workforce participation is critical to household income and career-building.

The DR model findings align with the IPW results, both indicating a reduction in women's labor force participation following the policy, while contrasting with the positive ATE from the matching model. By integrating population-level trends from IPW with localized effects from matching, the DR model highlights the context-dependent nature of policy impacts, revealing variations across subgroups and methodological approaches.

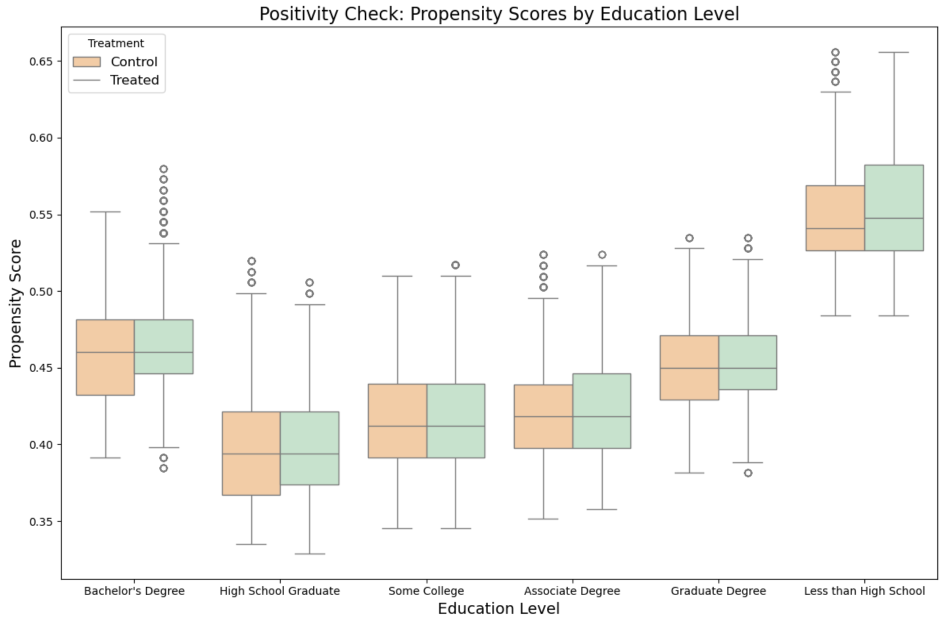

Figure 16: Doubly Robust Model: Propensity Scores by 'Level of Education'

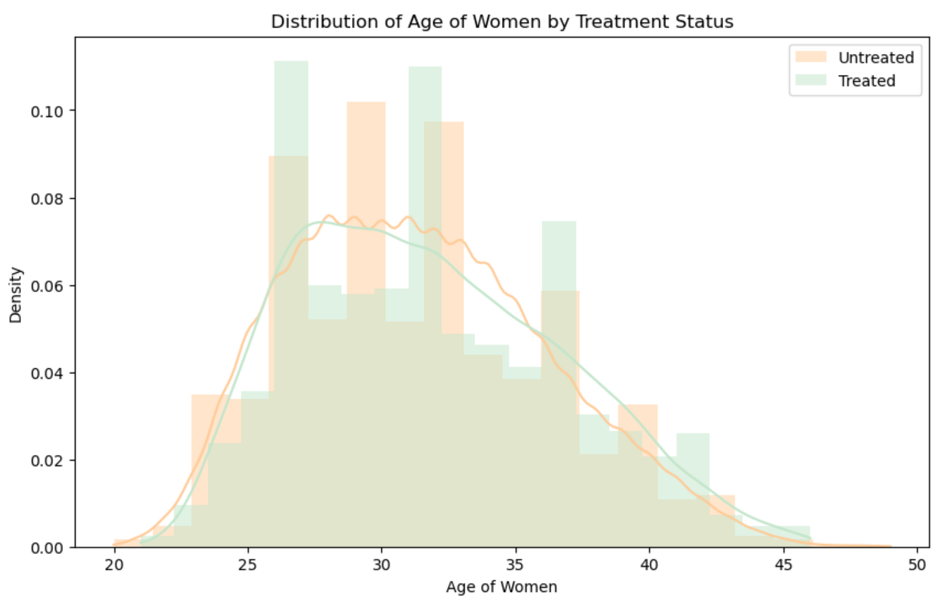

Figure 17: Doubly Robust Model: Positivity Check Plot for Covariate 'Age'

3.4 Advanced Causal Models

To explore heterogeneous treatment effects, advanced causal models like Double Machine Learning (DML) and Causal Forests are applied. These methods extend the causal analysis by incorporating machine learning techniques to model complex relationships and identify subgroup-specific effects. While DML focuses on balancing robustness and interpretability, Causal Forests highlight variability in treatment effects across demographic groups, uncovering disparities that aggregate models might overlook.

3.4.1 Advanced Causal Models: Double Machine Learning (DML)

The linear DML framework builds on the outlined identification strategy. The DML approach leverages Random Forests to flexibly model both the treatment assignment mechanism and the outcome, reducing bias from model misspecification. This is particularly important in the context of state-level paid family leave policies, where treatment effects may differ substantially by subgroup or region. The DML model's ability to capture this heterogeneity complements the localized positive effects observed in the matching model and the population-level reductions identified in the IPW framework.

The choice of the linear DML model in this context is driven by its ability to flexibly adjust for confounders like age and education, which were identified as significant predictors of labor force participation through prior feature selection methodologies. By modeling these covariates non-parametrically, the DML approach enables a more precise estimation of how policy impacts vary across demographic groups, such as younger versus older women or those with lower versus higher educational attainment. The estimated Average Treatment Effect (ATE) of -0.015, with a 95% confidence interval ranging from -0.025 to -0.005, suggests a statistically significant reduction in women's labor force attachment following the implementation of these policies. This result aligns with the findings from the IPW model and reinforces the hypothesis that while these policies provide critical support for caregiving, they may inadvertently reduce labor market engagement for some groups of women. The cumulative distribution plot of treatment effects (Figure 18) highlights the variability of impacts across the population, suggesting that certain subgroups experience stronger effects than others.

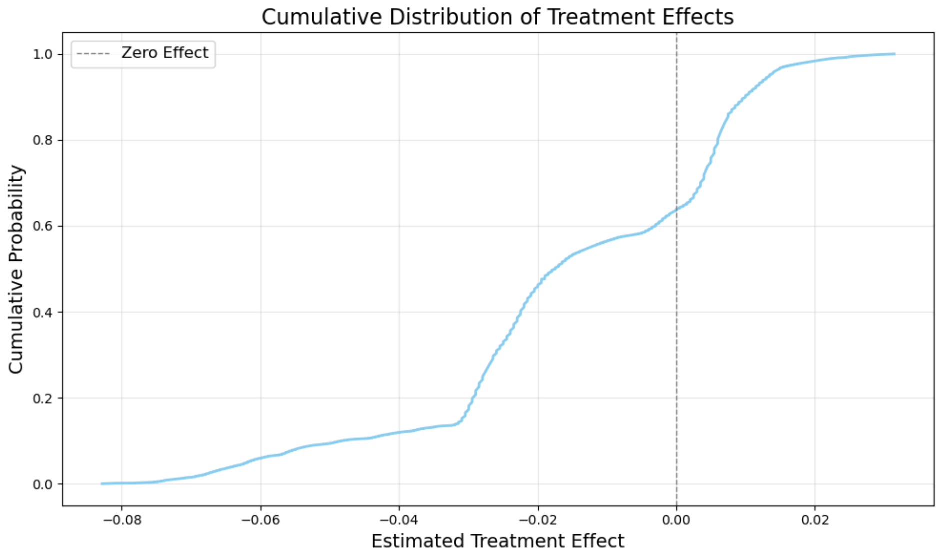

Figure 18: DML Model: Cumulative Distribution of Treatment Effects

3.4.2 Advanced Causal Models: Causal Forests

The Causal Forest DML model offers valuable insights into the heterogeneous impacts of paid family leave policies on women's labor force attachment. Figure 19 reveals that the treatment effects are centered around zero, reflecting the negligible Average Treatment Effect (ATE) of -0.005. However, the substantial variability in individual treatment effects, with some estimates significantly positive or negative as evidenced by the spikes and dips beyond the confidence bands. This highlights the model's strength in capturing heterogeneity that aggregate statistics obscure. The confidence intervals are generally tight for most samples but widen at the extremes, reflecting greater uncertainty in treatment effect estimation for underrepresented populations or those with less common covariate profiles.

These findings highlight the nuanced ways in which the policy affects different demographic groups. For instance, women with higher education levels may experience treatment effects closer to or above zero, suggesting benefits in terms of sustained or increased labor force participation. Conversely, other groups, such as younger or less-educated women, may face challenges that lead to negative effects. Importantly, the Causal Forest results complement prior models: while the matching model revealed localized positive effects and the IPW and doubly robust models indicated population-level reductions, the Causal Forest approach identifies meaningful disparities masked by aggregate neutrality in the ATE. This emphasizes the need for policymakers to account for subgroup-specific impacts when designing and evaluating interventions.

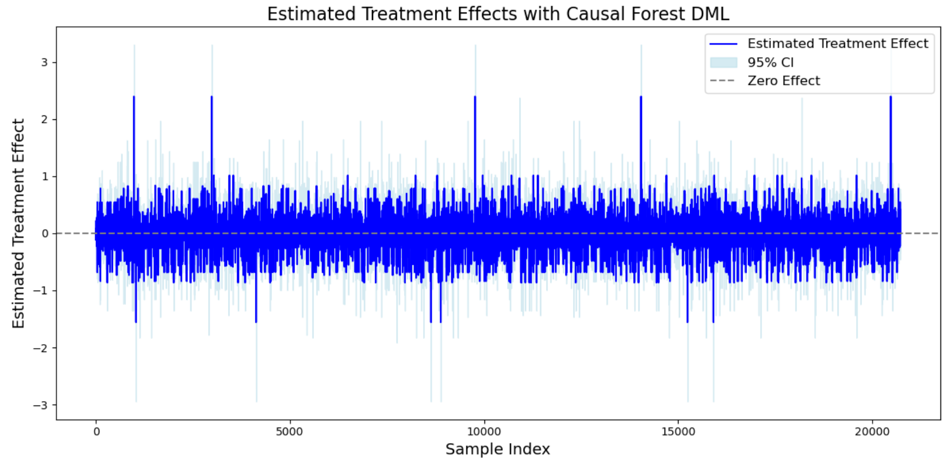

Figure 19: Causal Forest Model: Treatment Effects Plot

The histogram of heterogeneous treatment effects below (Figure 20) also highlights the diversity in how paid family leave policies impact women's labor force attachment. The majority of the estimated treatment effects cluster around zero, indicating that the policy exhibits minimal impact on average. However, the highlighted bins reveal distinct subgroups. The orange bins, representing low effects (between -0.05 and 0), suggest a subset of women who experience slight reductions in labor force attachment, potentially due to factors such as caregiving responsibilities. Conversely, the green bins, corresponding to high effects (between 0.05 and 0.1), suggest subgroups that benefit modestly, likely due to enabling factors like greater access to workplace flexibility or socioeconomic support.

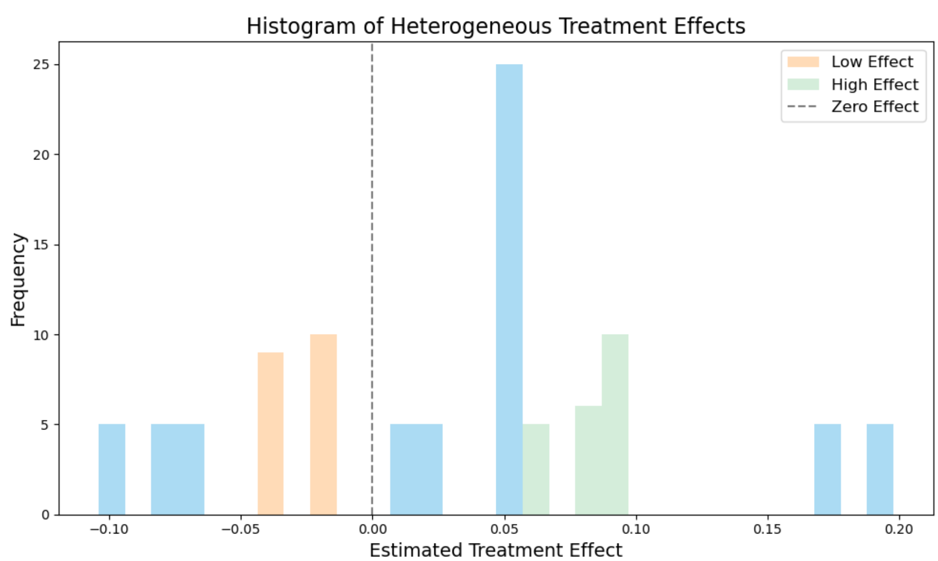

Figure 20: Causal Forest Model: Histogram of Heterogeneous Treatment Effects

3.5 Other Models: Estimating Conditional Average Treatment Effect Using Meta Learners

Finally, to estimate Conditional Average Treatment Effects (CATE) and further investigate heterogeneous effects, meta-learners (S-Learner, T-Learner, and X-Learner) are utilized. These models allow for a more granular analysis of how treatment effects vary across subgroups defined by age and education. By synthesizing insights from earlier models, the meta-learners provide a comprehensive understanding of the differential impacts of paid family leave policies, offering critical policy-relevant findings.

In the context of evaluating the impact of paid family leave policies on women's labor force attachment, meta-learners offer valuable tools for estimating heterogeneous treatment effects across subgroups. These learners allow us to explore how key demographic factors, like education and age, shape the policy's impact, addressing the core research question of whether the policy has equitable effects across the population. By modeling treatment effects conditional on these covariates, meta-learners reveal subgroup-specific dynamics.

Table 4 highlights the treatment effects across education levels. Women with lower levels of education, such as those with less than a high school degree, experience the most negative Conditional Average Treatment Effects (CATE), with estimates ranging from -0.012 (S-Learner) to -0.010 (X-Learner). This suggests that paid family leave policies may exacerbate labor force detachment among less educated women. In contrast, women with higher education levels, such as those with a bachelor's or graduate degree, show positive CATE estimates, ranging from 0.007 to 0.012. These results indicate that the policy is more beneficial for highly educated women, potentially reinforcing labor force participation in this subgroup. Table 5 extends this analysis to age groups. Women aged 24–35 exhibit slightly more negative treatment effects (-0.008 to -0.007), reflecting a potential vulnerability of this younger age group to adverse labor market outcomes following the policy. In comparison, women aged 36–45 exhibit minimal treatment effects, with estimates close to zero (-0.002 to -0.001), suggesting that this age group is less affected by the policy in terms of labor force attachment.

Table 4: CATE Estimates by Education Category for Meta-Learners

Education Category	S-Learner Estimate	T-Learner Estimate	X-Learner Estimate
Less than High School	-0.012	-0.009	-0.010
High School Graduate	-0.005	-0.003	-0.004
Some College	0.001	0.003	0.002
Associate Degree	0.004	0.006	0.005
Bachelor's Degree	0.007	0.009	0.008
Graduate Degree	0.010	0.012	0.011

Table 5: CATE Estimates by Age Category for Meta-Learners

Age Category	S-Learner Estimate	T-Learner Estimate	X-Learner Estimate
24–35	-0.008	-0.007	-0.007
36–45	-0.002	-0.001	-0.001

4 Conclusion

My study aims to evaluate the impact of short-term paid family leave policies, specifically the CPFL in California and the NJFLA in New Jersey, on women's labor force participation using an array of machine learning models and causal inference techniques. By employing feature selection models, causal estimation methods, and meta-learners, the analysis provides nuanced insights into the heterogeneous impacts of these policies, advancing both methodological and empirical understandings of labor market dynamics.

Feature selection models such as Lasso, Ridge, and Random Forest highlight the significance of demographic and state-level covariates, particularly education and age, in shaping labor force participation outcomes. The Random Forest model outperforms others in predictive accuracy and identifies the treatment (policy intervention) as a critical predictor, highlighting its utility in uncovering complex interactions between policy effects and individual traits. While boosting methods demonstrate moderate accuracy, the Random Forest approach effectively captures non-linear relationships, revealing localized variations in policy impact.

DAG-based analyses establish a conceptual framework for identifying causal pathways and potential confounders. This ensures that all subsequent models are correctly specified and aligned with the theoretical framework. Causal estimation techniques, including propensity score matching, inverse probability weighting (IPW), and doubly robust (DR) methods, present complementary perspectives on the policies' effects. Propensity score matching reveals a modest positive effect of the policies on women's labor force attachment, with an Average Treatment Effect (ATE) of 0.0142. In contrast, the IPW and DR methods indicate slight reductions in labor force participation, likely reflecting broader population-level trends and heterogeneities. These mixed findings highlight the importance of triangulating results across multiple models to derive robust policy insights.

Double Machine Learning (DML) provides robust causal estimates while leveraging flexible machine learning models to adjust for high-dimensional covariates. The DML results indicate a negligible ATE of -0.015, suggesting that the policies have minimal overall effects but with significant heterogeneity across subgroups. The Causal Forest model further explores this heterogeneity, capturing nuanced individual treatment effects with confidence intervals. While the ATE from the Causal Forest model is close to zero, the variation in individual effects highlights that certain groups benefit from the policy while others face disadvantages. This suggests the critical role of these advanced methods in uncovering the complexities of policy impacts.

Lastly, Meta-learners such as the S-Learner, T-Learner, and X-Learner offer a detailed exploration of subgroup-specific treatment effects, emphasizing education and age as moderators of policy impact. Women with lower educational attainment experience negative treatment effects, suggesting that existing barriers in the labor market may limit the policies' benefits for this group. Conversely, highly educated women benefit significantly, reinforcing labor force attachment. Age-based analyses reveal a modestly negative impact on younger women (aged 24–35), while older women (aged 36–45) exhibit minimal effects, highlighting life-cycle considerations in policy design.

The findings align with and extend existing literature on paid family leave. While previous studies document mixed effects of such policies on labor market outcomes, this study contributes novel evidence using advanced machine learning techniques. The results confirm prior research suggesting that higher-educated women disproportionately benefit from paid leave policies, but they also challenge assumptions of uniform policy effectiveness by revealing subgroup-specific vulnerabilities. These insights maintain the need for tailored interventions that address the unique challenges faced by less-educated and younger women. From a policy perspective, the analysis suggests that while paid family leave policies promote labor force attachment for certain subgroups, they may inadvertently exacerbate inequalities. To enhance their effectiveness, policymakers should consider supplementary measures such as subsidized childcare, targeted outreach to low-income families, and incentives for employer compliance. Additionally, addressing structural barriers that disproportionately affect vulnerable groups maximizes the inclusive potential of these policies.

5 Limitations and Scope for Further Research

While this study provides valuable insights into the impact of paid family leave policies on women's labor force participation, it is not without limitations. First, the analysis relies on observational data, which, despite the use of advanced causal inference methods, is inherently subject to unobserved confounding. Although methods such as Doubly Robust Estimation, Double Machine Learning, and Causal Forests attempt to mitigate biases from high-dimensional covariates, the potential for omitted variable bias cannot be entirely ruled out. For example, unobserved factors such as workplace flexibility, family support structures, or cultural attitudes toward caregiving could influence the observed outcomes.

Second, the study focuses primarily on the short-term effects of the CPFL and NJFLA policies. While these insights are valuable, the long-term effects of paid family leave, such as its impact on career trajectories, earnings, and overall economic stability, remain unexplored. Understanding these dynamic effects would provide a more holistic evaluation of the policy's success or limitations.

Third, the subgroup analyses, while illuminating, may be constrained by sample sizes within specific subgroups, particularly for women with lower levels of education or those in certain age categories. Small sample sizes can lead to wider confidence intervals and reduced statistical power, potentially obscuring meaningful heterogeneity in treatment effects. Additionally, while the meta-learners offer valuable subgroup-specific estimates, they rely on model assumptions that may introduce biases when the assumptions are violated.

Fourth, while this study leverages a wide array of machine learning methods, the trade-offs between interpretability and predictive power remain a challenge. Models such as Random Forests and Causal Forests excel in capturing complex interactions but may lack the transparency of simpler models like Logistic Regression. Future research could explore hybrid approaches that balance these trade-offs to provide both interpretable and robust insights.

Future research could address these limitations by incorporating richer datasets that capture unobserved factors, such as employer-level data on policy implementation or detailed information on caregiving responsibilities. Additionally, longitudinal studies tracking the long-term impacts of paid family leave policies could provide a more comprehensive understanding of their effectiveness. Exploring the interaction of paid family leave policies with other labor market interventions, such as universal childcare or flexible work arrangements, could also yield actionable insights for policymakers.

Furthermore, future work could investigate the interplay between labor market conditions and policy effectiveness. For instance, how do macroeconomic conditions, such as unemployment rates or labor demand in specific industries, influence the effectiveness of paid family leave policies? Finally, methodological advancements in machine learning and causal inference, such as deep causal models or ensemble methods that integrate multiple learners, could further enhance the precision and reliability of treatment effect estimates. Future studies can build on the findings of this analysis to provide more actionable insights into the design and implementation of paid family leave policies.

References

Adema, W., Clarke, C., and Frey, V. (2016). Paid parental leave and other supports for parents with young children: The united states in international comparison. *International Social Security Review*, 69(2):29–51.

Bailey, M. J., Byker, T. S., Patel, E., and Ramnath, S. (2019). The long-run effects of california's paid family leave act on women's careers and childbearing: New evidence from a regression discontinuity design and us tax data. Technical report, National Bureau of Economic Research.

Baum, C. L. and Ruhm, C. J. (2016). The effects of paid family leave in california on labor market outcomes. *Journal of Policy Analysis and Management*, 35(2):333–356.

Baum II, C. L. (2003). The effect of state maternity leave legislation and the 1993 family and medical leave act on employment and wages. *Labour Economics*, 10(5):573–596.

Byker, T. S. (2016). Paid parental leave laws in the united states: Does short-duration leave affect women's labor-force attachment? *American Economic Review*, 106(5):242–246.

Das, T. and Polachek, S. W. (2015). Unanticipated effects of california's paid family leave program. *Contemporary Economic Policy*, 33(4):619–635.

Gault, B., Hartmann, H., Hegewisch, A., Milli, J., and Reichlin, L. (2014). Paid parental leave in the united states: What the data tell us about access, usage, and economic and health benefits.

Goldin, C. and Mitchell, J. (2017). The new life cycle of women's employment: Disappearing humps, sagging middles, expanding tops. *Journal of Economic Perspectives*, 31(1):161–182.

Han, W.-J., Ruhm, C., and Waldfogel, J. (2009). Parental leave policies and parents' employment and leave-taking. *Journal of Policy Analysis and Management: The Journal of the Association for Public Policy Analysis and Management*, 28(1):29–54.

Jones, K. and Wilcher, B. (2024). Reducing maternal labor market detachment: A role for paid family leave. *Labour Economics*, 87:102478.

Kleven, H., Landais, C., Posch, J., Steinhauer, A., and Zweimüller, J. (2019). Child penalties across countries: Evidence and explanations. In *AEA Papers and Proceedings*, volume 109, pages 122–126. American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203.

Rossin-Slater, M. (2017). Maternity and family leave policy. Technical report, National Bureau of Economic Research.

US Bureau of Labor Statistics (2015). Employment characteristics of families – 2014 (news release). Washington, DC, US Bureau of Labor Statistics. US BLS.

Waldfogel, J. (1999). The impact of the family and medical leave act. *Journal of Policy Analysis and Management: The Journal of the Association for Public Policy Analysis and Management*, 18(2):281–302.