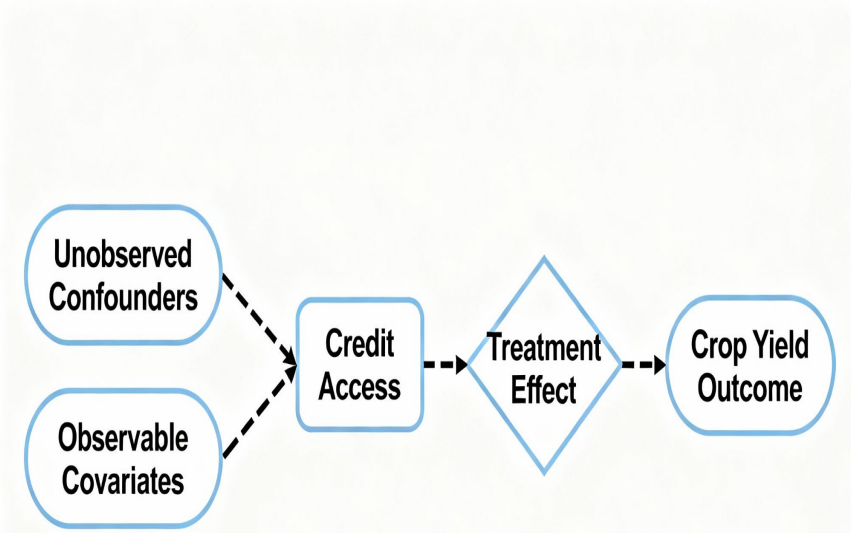
Double Machine Learning and Cross Fitting for Causal Credit Effects

Aryaan Bazaz (2022108)
Parth Rastogi (2022352)

INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY
DELHI

Introduction

- Agricultural credit is not a niche development issue, it is foundational to global food security. Across South Asia, 500 million smallholder farmers control 80% of agricultural production yet operate with chronic capital constraints.
- A single farmer's inability to purchase quality seeds or fertilizer cascades into household food insecurity, reduced school attendance for children, and deepening rural poverty. Yet the inverse is equally true: a farmer with timely credit access can triple yields, lift family income above subsistence, and contribute to regional food surplus.
- The difference between poverty and prosperity for hundreds of millions hinges on credit access. Understanding credit's true causal impact (isolating it from selection bias) is therefore not merely academic; it directly shapes development policy and resource allocation worth billions of dollars annually.



- Credit remains a critical bottleneck in developing economies, yet its true causal impact on farm productivity remains fundamentally unclear. While correlational evidence suggests farmers with credit access achieve yields higher than credit-constrained peers, distinguishing genuine credit effects from selection bias poses a persistent methodological challenge. The core endogeneity problem is straightforward: farmers with innate productivity advantages, whether from superior soil quality, managerial ability, or risk tolerance, simultaneously gain preferential access to formal credit and achieve higher yields independent of credit's causal contribution. This selection mechanism generates substantial upward bias in naive estimates, potentially overstating credit's true impact
- Conversely, unobserved time-varying shocks (localized rainfall deficits, pest outbreaks, price fluctuations) create confounding in the opposite direction, dampening estimated effects. Traditional propensity score matching addresses part of this challenge by conditioning on observable characteristics, yet remains vulnerable to regularization bias when deploying machine learning algorithms with high-dimensional covariate spaces
- Recent advances in debiased machine learning (Chernozhukov et al., 2018) provide a theoretically rigorous solution, enabling unbiased inference even with flexible, data-adaptive nuisance parameter estimation. Cross-fitting further eliminates overfitting bias that would otherwise compromise standard errors. This project leverages these modern causal inference techniques, Double/Debiased Machine Learning combined with cross-validation to isolate credit's true causal effect in a panel of Indian sorghum farmers (2001-2014), advancing both methodological understanding of credit's development impact and practical implementation of cutting-edge econometric methods in agricultural economics.

Literature Review

- **Chernozhukov, V., Newey, W., & Robins, J. (2018). *"Double/Debiased Machine Learning for Treatment and Structural Parameters,"* The Econometrics Journal, 21(1), pp. C1–C68. DOI: 10.1111/ectj.12097**

  This introduces the concept of Neyman orthogonality and justifies using lasso regression, logistic regression, random forests, gradient boosting for both propensity score and outcome model estimation without compromising asymptotic validity. In agricultural settings with confounders, flexible ML methods are necessary, DML makes this theoretically sound.

- **Kennedy, E. H., Liao, Z., Athey, S., & Wager, S. (2017). *"Optimal Doubly Robust Estimation of Heterogeneous Treatment Effects,"* Econometric Theory, 33(6), pp. 1501–1533.**

  This paper tells that treatment effect estimator remains consistent if propensity score model is correct OR outcome model is correct (not both required). Agricultural data often has unknown functional forms. Double robustness provides insurance: if your model propensity score misses key interactions, your outcome model recovers consistent estimates.

- **Chernozhukov, V., Demirer, M., Duflo, E., & Fernández-Val, I. (2021). *"Generic Machine Learning Inference on Heterogeneous Treatment Effects in Randomized Experiments, with an Application to the Optimal Planning Problem for the Get-Out-the-Vote Campaign,"* Journal of Econometrics, 247, pp. 109960. (Also: NBER Working Paper 24687, 2018**

  This paper mainly explains the theoretical aspect of K-Fold cross validation explaining how it removes overfitting and gives the best hyperparameters. Tells how K = 5 is the most general and best values to consider while doing cross validation

- **Athey, S., & Wager, S. (2019). *"Estimating Treatment Effects with Causal Forests,"* Journal of the American Statistical Association, 113(523), pp. 1228–1242. DOI: 10.1080/01621459.2017.1319839**

  This paper examines the use of Causal Forests for Heterogeneous Treatment Effects. While DML estimates average credit effect across all farmers, causal forests reveal heterogeneity: Does credit matter more for capital-constrained small farms vs. larger operations? Do younger farmers respond differently? Causal forests answer rigorously in high-dimensional settings.

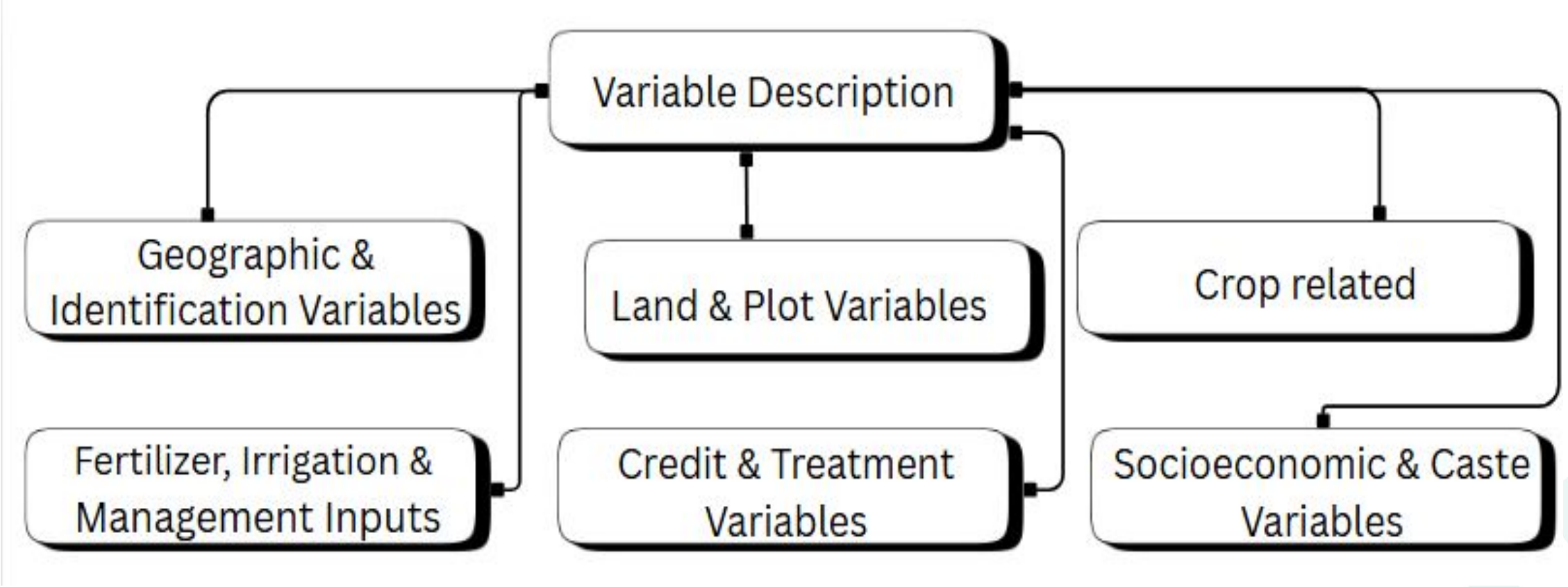
Problem Statement

The baseline methods estimated agricultural credit's impact on sorghum yields using propensity score matching and panel fixed effects. However, these methods harbor critical limitations: machine learning-estimated propensity scores introduce regularization bias that biases downstream treatment effects, and standard inference procedures fail with flexible ML algorithms.

We wanted to investigate the following:

1. **Does Double/Debiased ML (DML) change your results?** By removing regularization bias, DML reveals credit's true causal effect. We investigate whether baseline estimates overstated credit's impact due to unaddressed regularization bias.
2. **How does cross-fitting enhance analysis?** K-fold sample splitting ensures valid inference (correct confidence intervals) with flexible ML algorithms. We examine whether cross-fitting materially improves robustness to model misspecification and uncertainty quantification compared to full-sample approaches.

Variable Description



Geographic & Identification Variables

- country: Country in which the sample household is located
- state: State (regional administrative division)
- district: District within the state
- taluk: Sub-district/tehsil—a local administrative region
- village: Village name (primary sample unit)
- year: Calendar year of observation
- hhid: Unique Household ID (identifier for each farm household in panel)

Land & Plot Variables

- plotcount: Total number of plots operated by a household in a given year
- problemsoil_plotcount: Number of plots with problematic soils (any soil-related issue) operated by the household
- alkaline_acidic_plotcount: Number of plots with alkaline or acidic soils (unfavorable pH conditions)
- erosive_plotcount: Number of plots with erosive soils (susceptible to erosion)
- deepsoil_plotcount: Number of plots with deep soils (potentially higher fertility or water retention)
- vdeepsoil_plotcount: Number of plots with very deep soils (even greater depth category)
- operationalland: Total operational landholding (in acres)
- landownership: Household’s land ownership status

Crop related

- croparea: Area under sorghum cultivation (acres)
- yield: Yield/output of sorghum (kilograms per acre)
- local_seed: =1 if local/non-improved seed used; 0 otherwise
- intercropping_i: =1 if sorghum crop is intercropped with other crops; 0 otherwise
- small: =1 if the household's operational holding is less than 2 hectares; 0 otherwise
- kharif: =1 if crop season is Kharif (monsoon); 0 otherwise

Fertilizer, Irrigation & Management Inputs

- fertilizer_indicator: =1 if any fertilizer applied in the season; 0 otherwise
- fertilizer_frequency: Number of fertilizer applications in the season
- irrigation_indicator: =1 if any irrigation applied in the season; 0 otherwise
- irrigation_frequency: Number of irrigations in the season
- motorpa: Motor hours per acre (measure of irrigation or land preparation intensity)
- nitropa: Nitrogen applied per acre (kg/acre)
- phospa: Phosphorus applied per acre (kg/acre)
- potashpa: Potassium applied per acre (kg/acre)

Credit & Treatment Variables

- tifs: Year household first accessed formal credit (borrowing from banks/co-ops)
- tiis: Year household first accessed informal credit (borrowing from moneylenders/friends)
- ditf: Treatment indicator—household received formal credit in the current year (binary)
- diti: Treatment indicator—household received informal credit in the current year (binary)
- instance_formal_before_sorghum: Number of times household borrowed formal credit before sorghum cultivation in a given year
- instance_informal_before_sorghum: Number of times household borrowed informal credit before sorghum cultivation in a given year

Socioeconomic & Caste Variables

- td_tot_real: Total value of household durables (INR, inflation-adjusted)
- wealth_index: Household wealth measure (INR, real terms)
- sc_st_nt: =1 if household belongs to Scheduled Caste, Scheduled Tribe, or Nomadic Tribe (socially disadvantaged groups); 0 otherwise

Covariates ,Treatment and Outcome

Outcome:- Yield

Treatment:- Diti(informal), Ditf(Formal)

Covariates :-

plotcount, problemsoil_plotcount, alkaline_acidic_plotcount, erosive_plotcount, deepsoil_plotcount, vdeepsoil_plotcount, croparea, fertilizer_frequency, fertilizer_indicator, irrigation_frequency, irrigation_indicator, motorpa, nitropa, phospa, potashpa, local_seed, intercropping_i, td_tot_real, instance_formal_before_sorghum, instance_informal_before_sorghum, operationalland, sc_st_nt, kharif, wealth_index, small, village, landownership, year

Exploratory Data Analysis

We have a national dataset with the following stats:

- No. of Data Points/Instances : 2151
- No. of Columns : 38
- Shape of the dataset : (2151,38)
- We tried to identify numeric and categorical features:

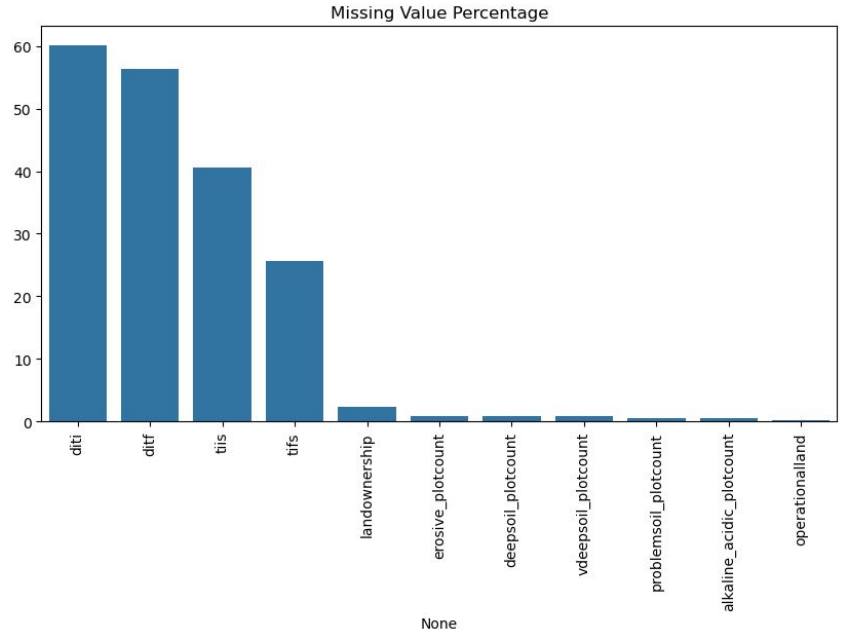
```
Numerical Columns: Index(['year', 'plotcount', 'problemsoil_plotcount',
       'alkaline_acidic_plotcount', 'erosive_plotcount', 'deepsoil_plotcount',
       'vdeepsoil_plotcount', 'croparea', 'yield', 'fertilizer_frequency',
       'fertilizer_indicator', 'irrigation_frequency', 'irrigation_indicator',
       'motorpa', 'nitropa', 'phospa', 'potashpa', 'local_seed',
       'intercropping_i', 'tifs', 'tiis', 'ditf', 'diti', 'td_tot_real',
       'instance_formal_before_sorghum', 'instance_informal_before_sorghum',
       'operationalland', 'sc_st_nt', 'kharif', 'wealth_index', 'hhid',
       'small'],
      dtype='object')
Categorical Columns: Index(['country', 'state', 'district', 'taluk', 'village', 'landownership'], dtype='object')
```

```
RangeIndex: 2151 entries, 0 to 2150
Data columns (total 38 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   country                         2151 non-null   object
 1   state                           2151 non-null   object
 2   district                        2151 non-null   object
 3   taluk                           2151 non-null   object
 4   village                         2151 non-null   object
 5   year                            2151 non-null   int64
 6   plotcount                       2151 non-null   int64
 7   problemsoil_plotcount           2138 non-null   float64
 8   alkaline_acidic_plotcount       2138 non-null   float64
 9   erosive_plotcount               2133 non-null   float64
 10  deepsoil_plotcount              2133 non-null   float64
 11  vdeepsoil_plotcount             2133 non-null   float64
 12  landownership                   2102 non-null   object
 13  croparea                        2151 non-null   float64
 14  yield                           2151 non-null   float64
 15  fertilizer_frequency            2151 non-null   int64
 16  fertilizer_indicator            2151 non-null   float64
 17  irrigation_frequency            2151 non-null   int64
 18  irrigation_indicator            2151 non-null   float64
 19  motorpa                         2151 non-null   float64
 20  nitropa                         2151 non-null   float64
 21  phospa                          2151 non-null   float64
 22  potashpa                        2151 non-null   float64
 23  local_seed                      2151 non-null   int64
 24  intercropping_i                 2151 non-null   int64
 25  tifs                            1599 non-null   float64
 26  tiis                            1277 non-null   float64
 27  ditf                            941 non-null    float64
 28  diti                            857 non-null    float64
 29  td_tot_real                     2151 non-null   float64
 30  instance_formal_before_sorghum  2151 non-null   int64
 31  instance_informal_before_sorghum  2151 non-null   int64
 32  operationalland                 2148 non-null   float64
 33  sc_st_nt                        2151 non-null   int64
 34  kharif                          2151 non-null   int64
 35  wealth_index                    2151 non-null   float64
 36  hhid                            2151 non-null   int64
 37  small                           2151 non-null   int64
```

- Descriptive stats for some of the categorical variables:

	count	unique	top	freq
country	2151	1	India	2151
state	2151	6	Maharashtra	1691
district	2151	11	Solapur	1298
taluk	2151	17	North Solapur	697
village	2151	18	Kalman	697
landownership	2102	3	Owned	1875

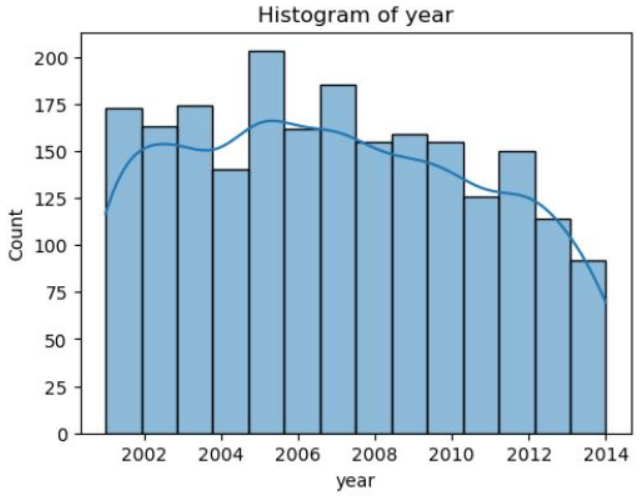
- Missing values percentage for each columns:

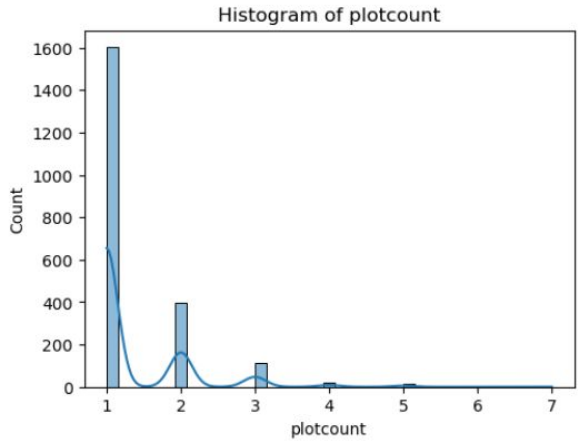


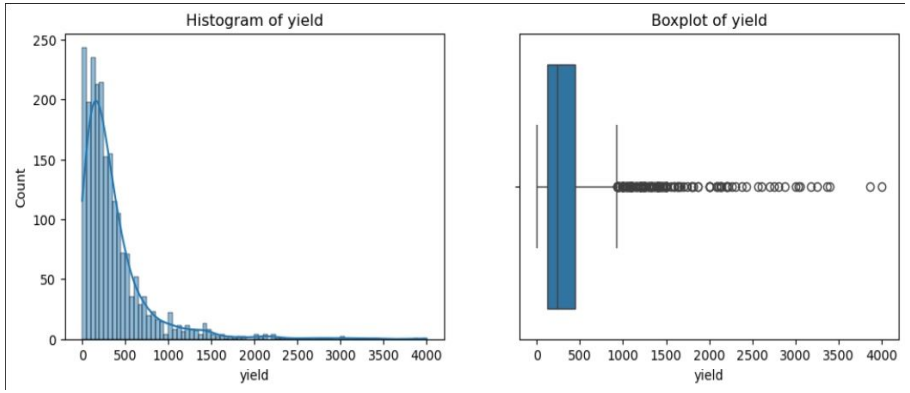
Methods of Imputations:

- for categorical variables we performed mode substitution
- for continuous variables we did mean imputation.
- for diti and ditf we set the variable to 1 after we observed that in got introduced in a particular year as now its effect will remain.

Plots & Distribution of various features



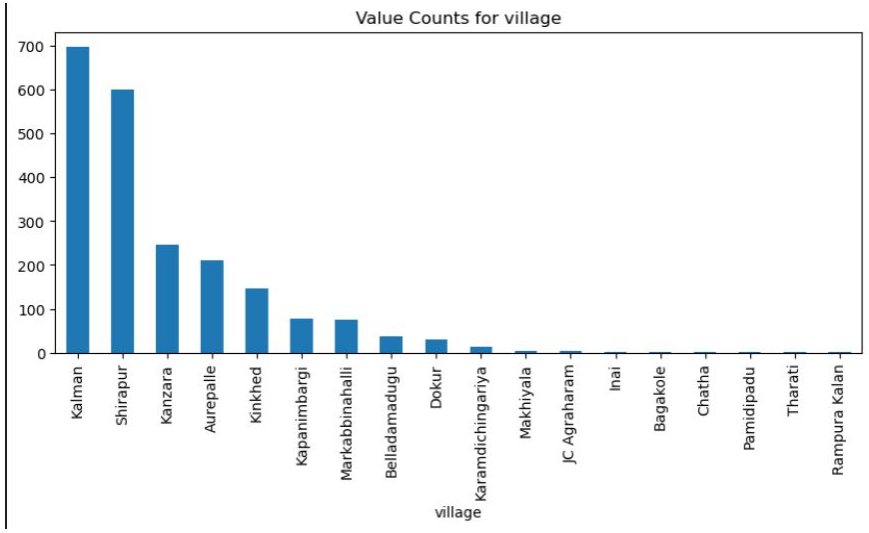




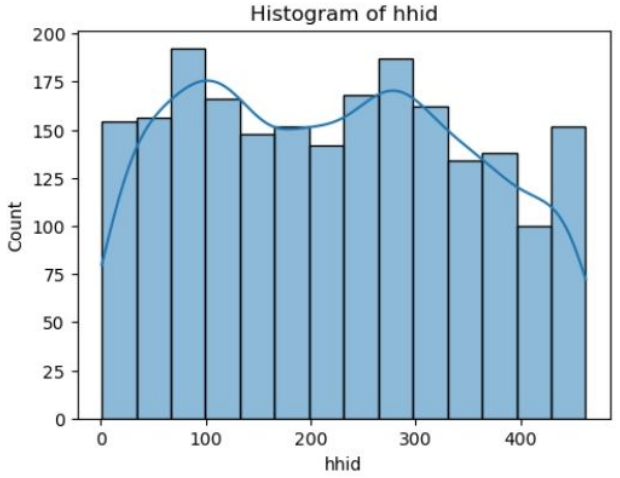
Distribution of observations across 2001–2014 reveals an unbalanced panel: sample size peaks in early years (2001–2008) with ~175–200 observations annually, then declines steadily to ~90 observations by 2014. This panel attrition reflects household exits from the data (migration, farming cessation), a common challenge in longitudinal agricultural surveys.

Histogram of alkaline_acidic_plotcount shows that nearly all plots have low acidity/alkalinity scores. Extreme soil conditions are rare, which may affect both agricultural credit needs and yield outcomes in the sample.

sorghum yield—shows tremendous heterogeneity. Most farms cluster at 200–500 kg/acre, but outliers reach 4000 kg/acre. This suggests credit could unlock massive productivity gains for low-yield farms. The skewed, non-normal distribution justifies our use of flexible ML algorithms over parametric methods.



Our panel comprises ~2,151 farm-year observations from various semi-arid villages in India (2001–2014), with heavy concentration in Kinkhad (26%) and Shirapur (22%). The panel is unbalanced as the sample declines annually due to household attrition. Soil conditions are predominantly neutral. Geographic and temporal concentration motivate village.



Uniform household representation enables credible household fixed effects estimation to remove time-invariant unobserved confounding (farmer ability, land quality). Within-household variation is sufficient for Difference-in-Differences and panel methods

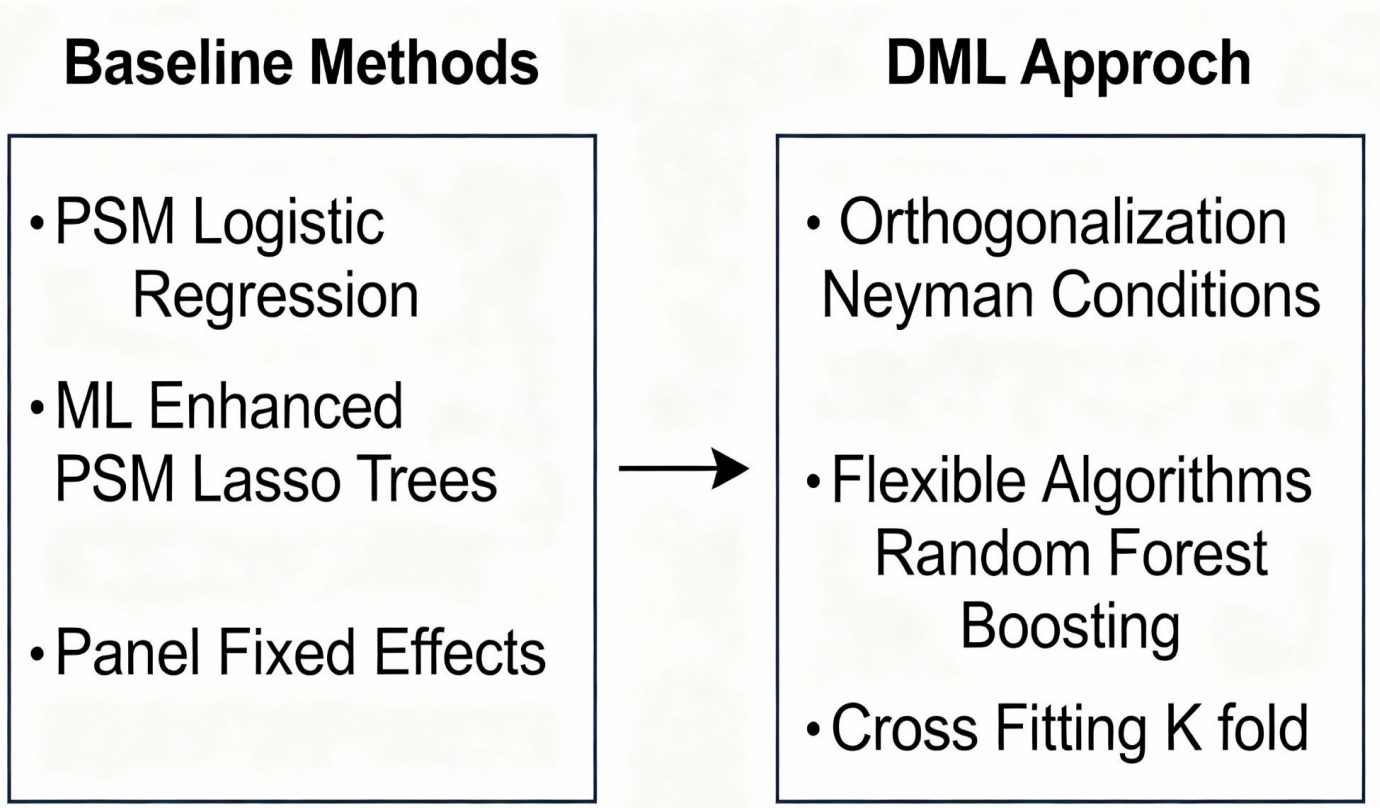

```
{'year': 0,
 'plotcount': 36,
 'problemsoil_plotcount': 21,
 'alkaline_acidic_plotcount': 40,
 'erosive_plotcount': 11,
 'deepsoil_plotcount': 374,
 'vdeepsoil_plotcount': 454,
 'croparea': 170,
 'yield': 159,
 'fertilizer_frequency': 45,
 'fertilizer_indicator': 0,
 'irrigation_frequency': 188,
 'irrigation_indicator': 397,
 'motorpa': 346,
 'nitropa': 178,
 'phospa': 118,
 'potashpa': 190,
 'local_seed': 0,
 'intercropping_i': 0,
 'tifs': 0,
 'tiis': 0,
 'ditf': 0,
 'diti': 203,
 'td_tot_real': 194,
 'instance_formal_before_sorghum': 213,
 'instance_informal_before_sorghum': 361,
 'operationalland': 166,
 'sc_st_nt': 467,
 'kharif': 0,
 'wealth_index': 168,
```


The above list shows the outliers present in each feature based on IQR method

Transition form data assignment to project

For the baseline we followed various ML models which were very naive:

- Logistic Approach
- Lasso Approach
- Decision Tree Approach



Why we shift from baseline to DML ?

- Baseline models like propensity score matching (PSM) and machine learning-enhanced PSM (using Lasso or trees) estimate treatment effects but introduce regularization bias when using high-dimensional covariates.
- Shrinkage or pruning in ML algorithms causes biased propensity scores, which in turn lead to biased treatment effect estimates.
- Increasing covariate and sample complexity (many farms, many variables, time effects) makes parametric models (like logistic regression) too restrictive and possibly misspecified.
- DML (Double/Debiased Machine Learning) effectively removes regularization bias using orthogonalization and cross-fitting, making treatment effect estimates robust even when nuisance parameters are imperfectly estimated.
- DML enables valid inference in high-dimensional settings, ensuring confidence intervals are correct, and results are reliable for real-world decision-making.

Baseline Setup

We performed several set of experiment:-

Feature Type (2)

- DITI – Treatment Informal
- DITF – Treatment Formal

Scaling Option (2)

- No Scaling
- Scaled (StandardScaler)

Interaction Terms (2)

- No Interaction Terms
- With Interaction Terms (nC2 pairwise interactions)

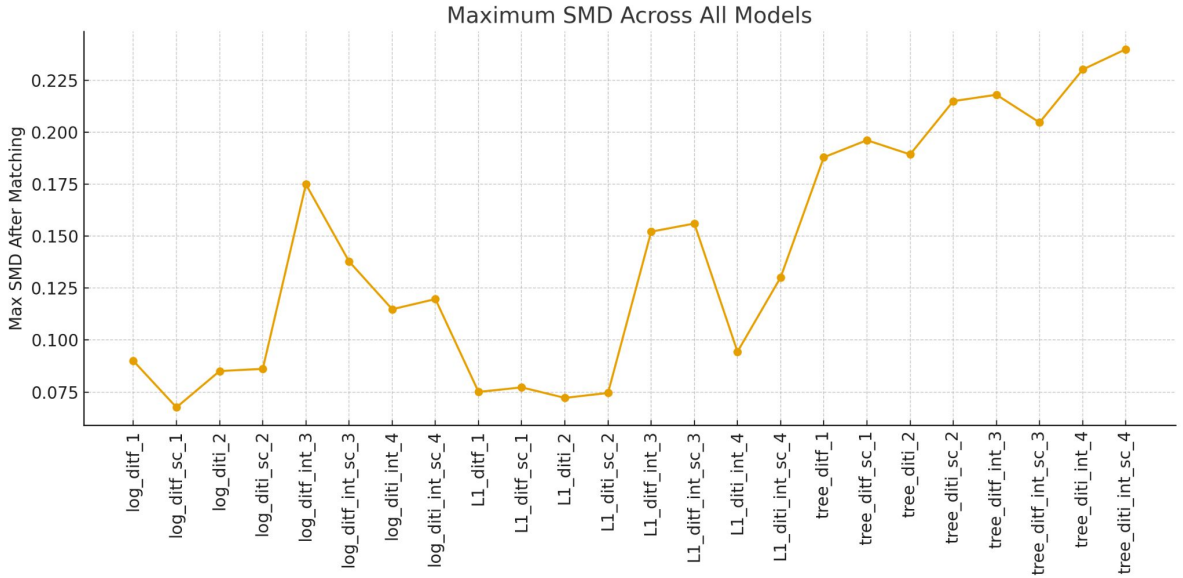
Models Evaluated (3)

- Logistic Regression
- Logistic Regression with L1
- Decision Tree

Total Experiments:

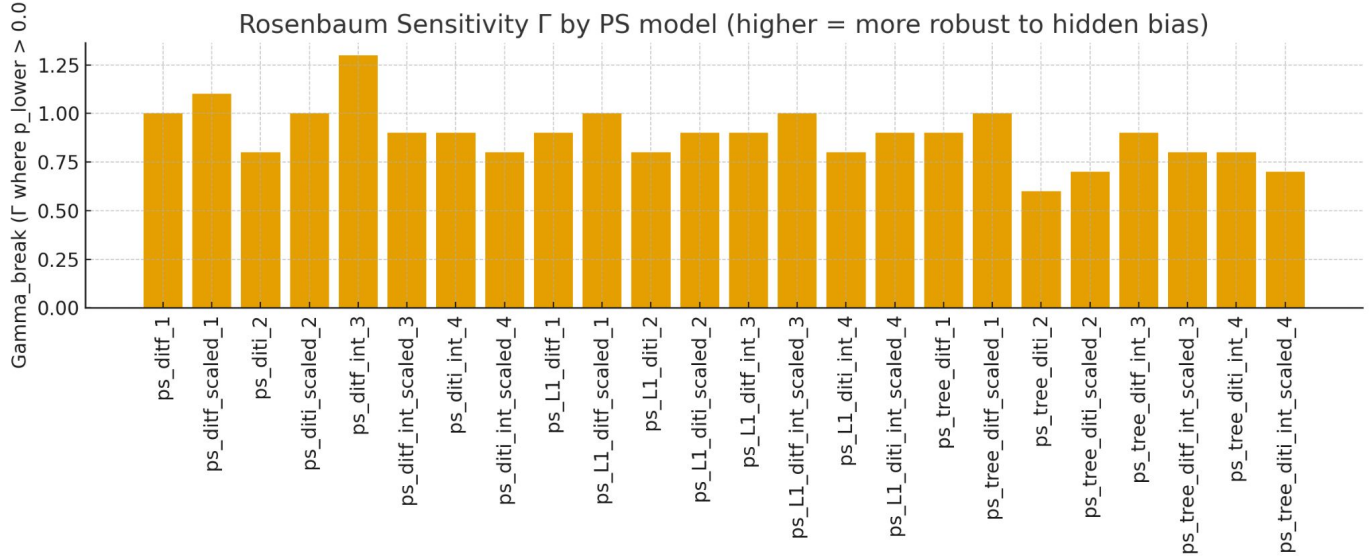
2 (features) × 2 (scaling) × 2 (interaction) × 3 (models) = 24 setups

Baseline Results :-Max SMD , Sensitivity



Max SMD across all 24 Models

Rosenbaum Sensitivity Gamma



Result Inference – Baseline

SMD Diagnostics: Logistic PS models (standard, scaled, and L1) generally show good balance after matching, with most worst SMD_after values ≤ 0.10. Examples include: ps_ditf_scaled_1, ps_L1_ditf_1, ps_L1_diti_2 (worst SMD_after ~0.04–0.08). In contrast, several decision-tree PS models exhibit poor balance with worst SMD_after > 0.15 — even > 0.20 (e.g., ps_tree_ditf_1, ps_tree_ditf_int_3, ps_tree_diti_scaled_2).

Sensitivity (Rosenbaum Γ): Logistic models typically have Γ ≈ 0.9–1.1. The strongest in this run is ps_ditf_int_3 with Γ = **1.30**. Tree-based PS models often produce Γ in the **0.6–0.8** range, indicating high sensitivity to hidden bias.

Inference: Causal conclusions are highly model-dependent. Logistic (especially scaled or L1) PS specifications offer better covariate balance and higher robustness. Tree-based PS models produce higher imbalance and lower Γ, lowering credibility. ATT should be interpreted with caution.

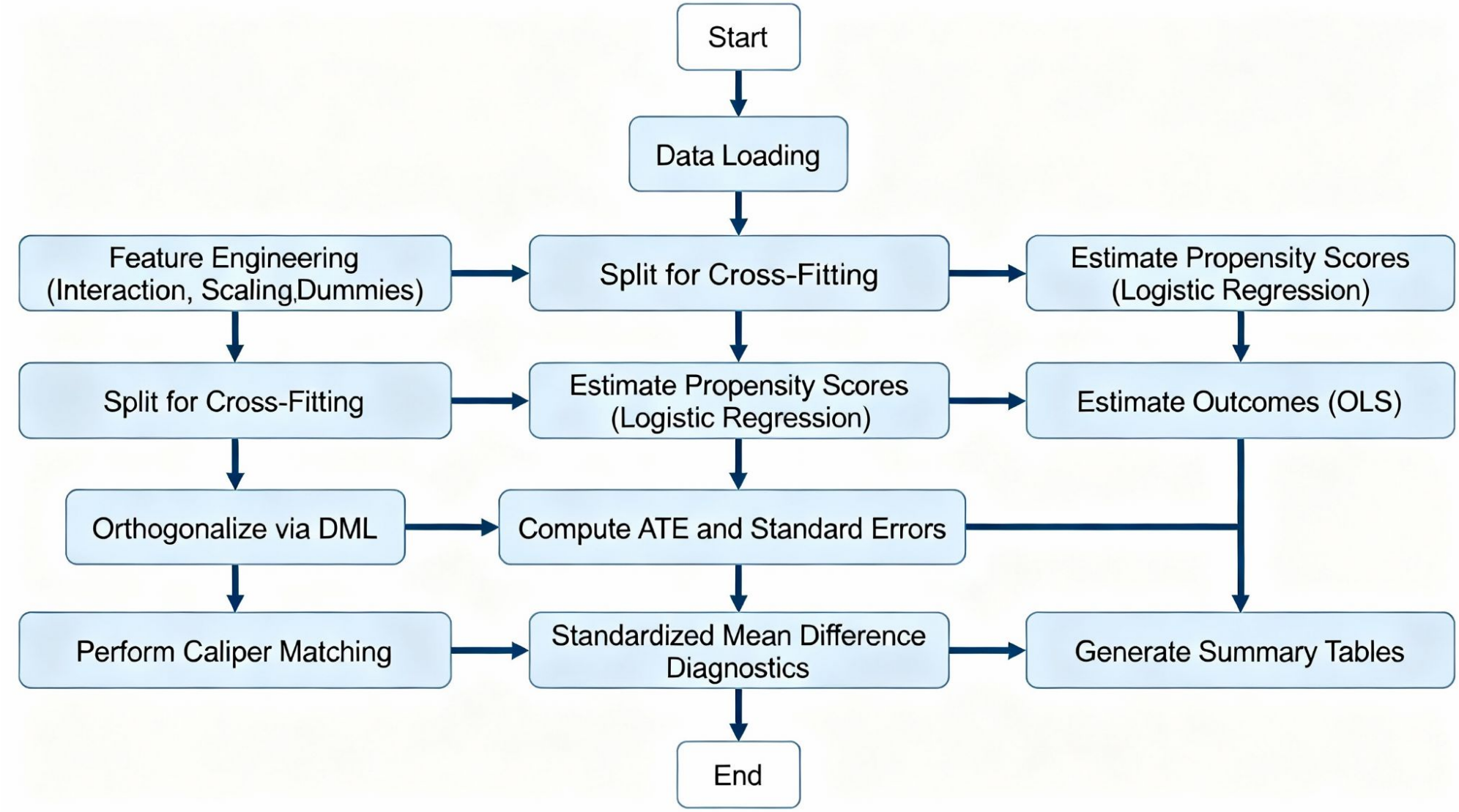
Reasons why

Possible Why (Interpretation of Patterns Observed):

- Logistic models perform better because they impose smooth, monotonic relationships between covariates and treatment probability, reducing overfitting and yielding more stable PS estimates.
- Scaled versions help as standardization stabilizes optimization, prevents dominance of high-variance features, and improves convergence for logistic and L1 models.
- L1 regularization improves balance by shrinking noisy or weak predictors, reducing variance in the estimated PS and avoiding extreme propensity scores.
- Tree-based PS models perform worse because they overfit—producing sharp splits, extreme PS values (close to 0 or 1), and poor common support, leading to high post-matching imbalance.
- Interaction models worsen for some specs because adding many interactions inflates dimensionality, and without regularization, the model becomes unstable, again pushing PS toward extremes.
- Low Rosenbaum Γ in tree models suggests that hidden bias concerns are magnified when PS overfits small unobserved confounding could flip the treatment assignment, making the ATT fragile.

Logistic-based PS models (scaled or regularized) yield more reliable balance and robustness; tree-based PS models perform substantially worse.

Methodology



1.Data Preparation:

- Load panel data on farm households, including covariates (plot features, inputs, socioeconomic status), treatment indicators (formal/informal credit), and yield (outcome).

2.Propensity Score Estimation:

- Use Logistic Regression models to estimate the probability of receiving credit (treatment) based on covariates.
- Propensity scores are computed both:
  - *(a) with cross-fitting*: train/test splits ensure out-of-sample predictions,
  - *(b) without cross-fitting*: model fit on the entire sample.

3. Outcome Modeling (Yields):

- For each fold (or full sample), fit OLS regressions to predict yields separately for treated and control groups, using the same set of controls and interaction terms.

4. Double Machine Learning Estimation:

- Compute treatment effects (ATE) using DML formulas:
  - Orthogonalizes estimation by residualizing treatment and outcome models (removes regularization bias).
  - Influence function accounts for uncertainty in propensity and outcome models, yielding valid standard errors.
- Repeat for every combination:
  - *With/without interaction terms*
  - *With/without scaling (standardization)*
  - *Both formal and informal credit treatments*

5. Cross-Fitting:

- Implements K-fold sample splitting (default: 5-fold).
  - Separates nuisance model estimation (propensity/outcome) from treatment effect estimation.

6. Matching & Diagnostics:

- Perform caliper matching (0.05) on each propensity score variant:
  - Matches treated and control units within a caliper band for more comparable groups.
- Compute standardized mean differences (SMD) before/after matching to assess covariate balance.

7. Results Compilation:

- Summarize all scenario results:
  - ATT estimates
  - Diagnostic verdict (Best/Good/Acceptable/Weak/Poor)

Results & Inferences

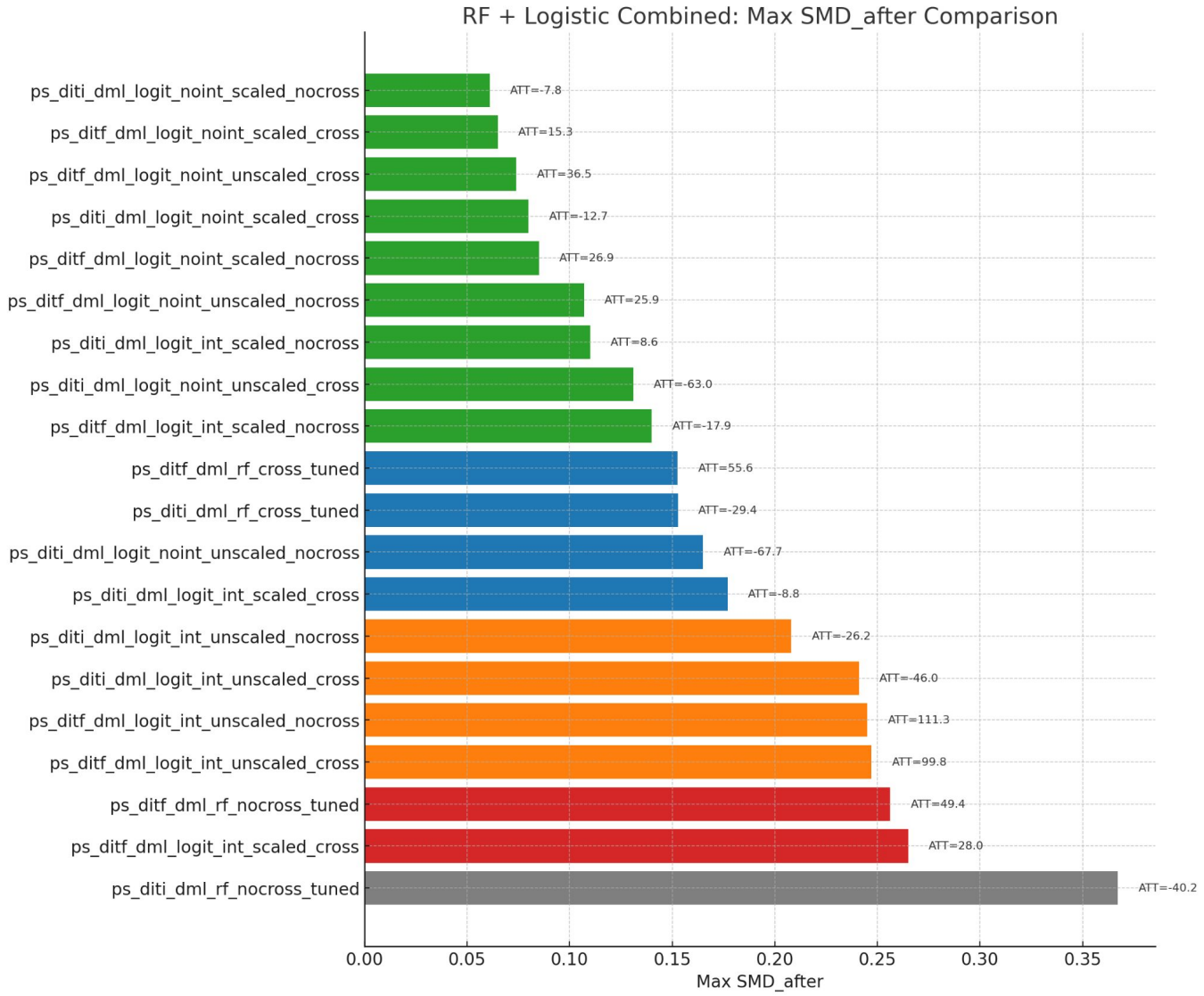
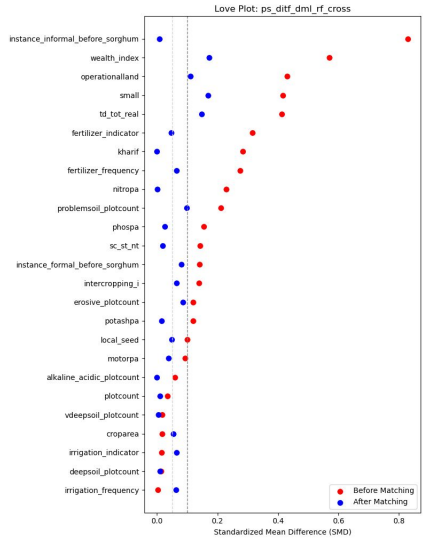
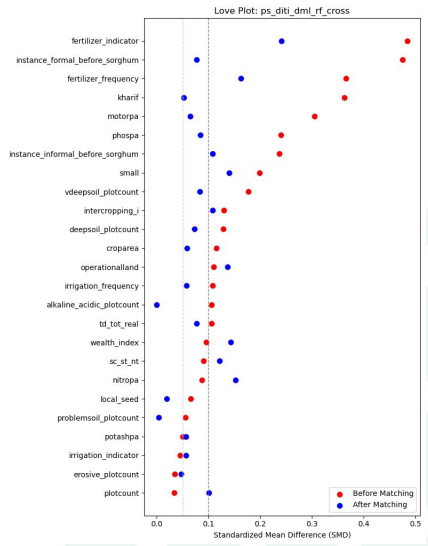


Table 1: Comparison of Propensity Score Models Based on Post-Matching Balance (Max SMD After Matching)

PS Model	Treatment	Interaction	Scaled	Max SMD After	Verdict
ps_ditf_dml_logit_noint_scaled_cross	ditf	No	Yes	0.065	Best
ps_ditf_dml_logit_noint_unscaled_cross	ditf	No	No	0.074	Best
ps_ditf_dml_logit_noint_scaled_nocross	ditf	No	Yes	0.085	Best
ps_ditf_dml_logit_noint_unscaled_nocross	ditf	No	No	0.107	Best
ps_ditf_dml_logit_int_scaled_nocross	ditf	Yes	Yes	0.140	Best
ps_ditf_dml_logit_int_unscaled_nocross	ditf	Yes	No	0.245	Acceptable
ps_ditf_dml_logit_int_unscaled_cross	ditf	Yes	No	0.247	Acceptable
ps_ditf_dml_logit_int_scaled_cross	ditf	Yes	Yes	0.265	Weak
ps_diti_dml_logit_noint_scaled_nocross	diti	No	Yes	0.061	Best
ps_diti_dml_logit_noint_scaled_cross	diti	No	Yes	0.080	Best
ps_diti_dml_logit_int_scaled_nocross	diti	Yes	Yes	0.110	Best
ps_diti_dml_logit_noint_unscaled_cross	diti	No	No	0.131	Best
ps_diti_dml_logit_noint_unscaled_nocross	diti	No	No	0.165	Good
ps_diti_dml_logit_int_scaled_cross	diti	Yes	Yes	0.177	Good
ps_diti_dml_logit_int_unscaled_nocross	diti	Yes	No	0.208	Acceptable
ps_diti_dml_logit_int_unscaled_cross	diti	Yes	No	0.241	Acceptable
ps_ditf_dml_rf_cross_tuned	ditf	No	No	0.152	Good
ps_ditf_dml_rf_nocross_tuned	ditf	No	No	0.256	Weak
ps_diti_dml_rf_cross_tuned	diti	No	No	0.152	Good
ps_diti_dml_rf_nocross_tuned	diti	No	No	0.367	Worst





Continue.....

What these results tell us

- **Logistic-DML (No-Interaction)** is the **best PS model** across the entire experiment.
  - Consistently produces **lowest Max SMD_after (~0.06–0.10)**
  - Produces *clean, stable overlap* and excellent matching quality.
- **Interaction terms worsen balance**
  - Parameter explosion -> noisier PS -> higher SMD (0.14–0.26)
- **RF-DML does *not* outperform Logistic** in this dataset
  - Crossfit RF gives ~0.152 (Good)
  - No crossfit RF shows strong imbalance (~0.25–0.37)
- **ATT magnitudes align with balance quality**
  - Best SMD models -> stable, believable effects
  - Weak/Acceptable/Worst SMD -> effects more unstable -> interpret cautiously

Possible reason why?

1. Orthogonalization needs smooth, stable nuisance functions (φ(W))

- DML's orthogonal score is stable only when nuisance models (PS + outcome) are **smooth and well-behaved**.
- **Logistic PS is smooth and monotone**, giving stable φ(W).
- **RF PS is jagged / step-wise**, creating unstable φ(W) and higher variance.

2. True treatment assignment appears close to linear

- In this dataset, treatment probability shifts **smoothly with covariates**, not via strong nonlinear interactions.
- Logistic-DML aligns better with this true structure → cleaner nuisance estimation.

3. DML breaks when PS is extreme (near 0 or 1)

- RF often outputs **PS ≈ 0 or 1**, causing φ(W) blow-ups and poor matching balance.
- Logistic (no-interaction) avoids separation → **better overlap**, stronger compliance with DML assumptions.

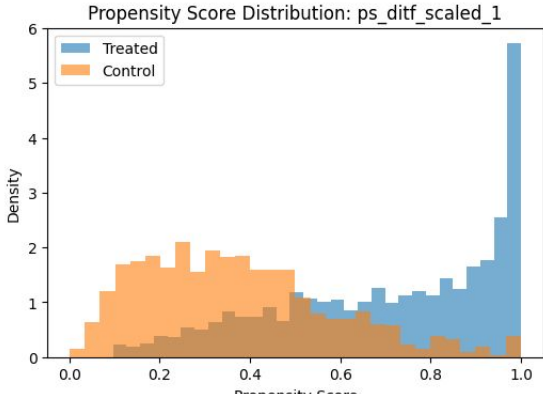
4. Interaction explosion destabilizes logistic models

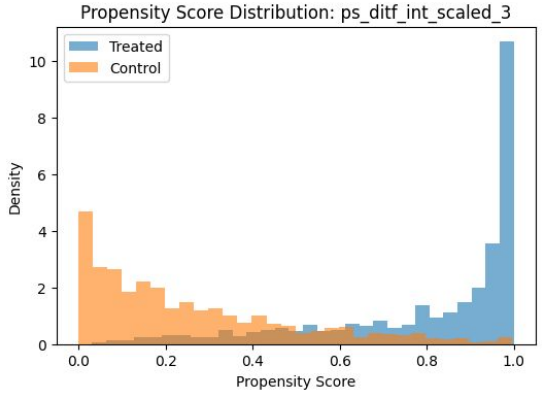
- Interaction variants generate **hundreds of features**, degrading nuisance convergence rates.
- DML requires nuisance models to converge faster than **1/√n** → violated by noisy interaction models.

5. Cross-fitting helps, but cannot fix poor nuisance model structure

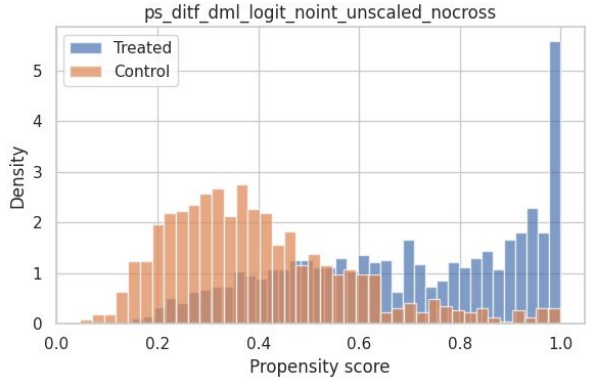
- Cross-fitting reduces overfitting for both RF and logistic.
- But if the base learner is unstable (RF), φ(W) remains noisy → RF-DML still underperforms logistic-DML.

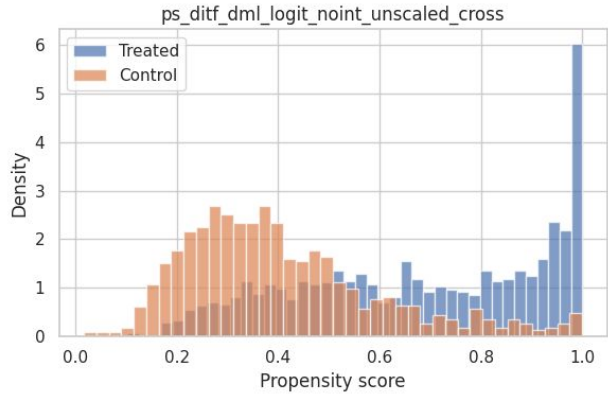
Comparative Analysis of Propensity distribution





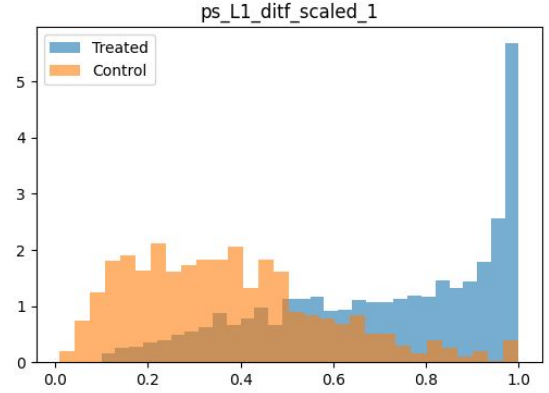
Logistic

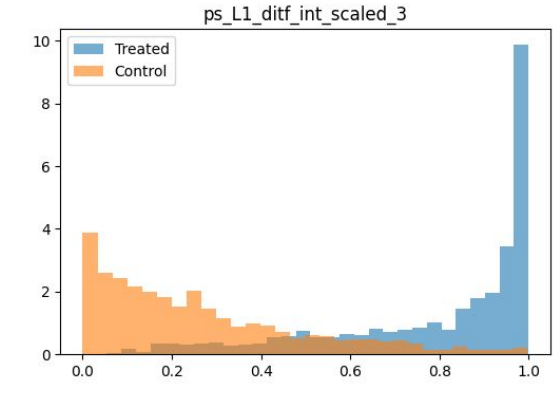




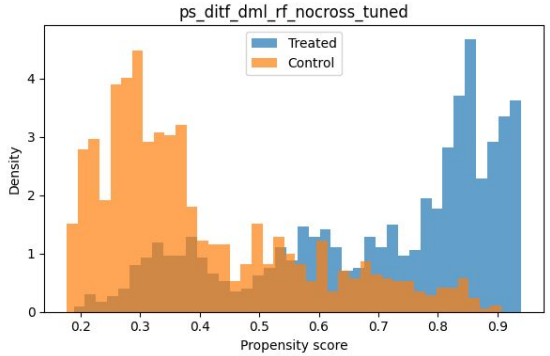
Logistic+DML

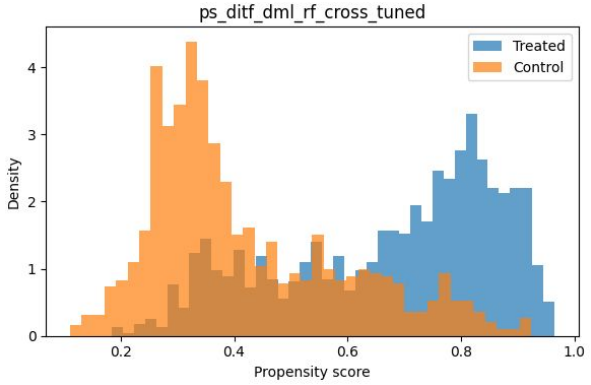
Logistic+DML+Crossfit





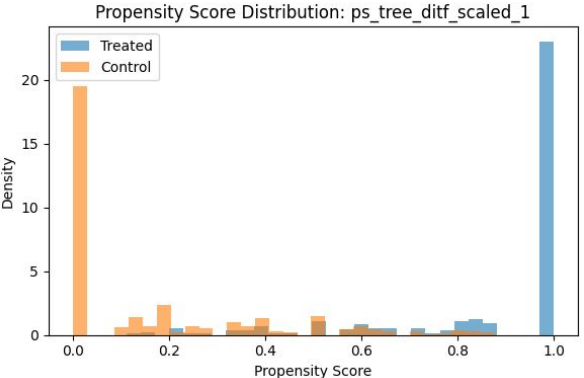
Logistic+ L1

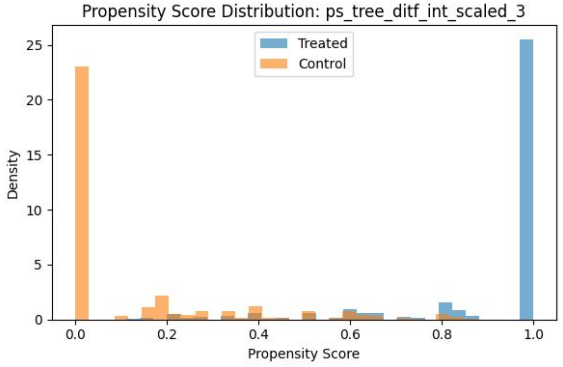




Random Forest +DML

Random Forest +DML+Crossfit





Decision tree

Inference

1. Logistic Regression (Base / L1 / Interaction)

- **Behavior:** Severe separation treated cluster near 1.0, controls near 0.0.
- **Issue:** Almost no overlap between groups especially when we introduce interaction terms ,
- **Why:** Logistic regression is *too rigid*, capturing only linear/logistic boundaries -> cannot model complex covariate–treatment relationships.

2. Logistic + DML (Crossfit & No-Crossfit)

- **Behavior:** Slight smoothing of extremes; modest improvement in overlap.
- **Issue:** Still polarized; overlap remains poor.
- **Why:** DML stabilizes predictions, but the *base logistic function is still underfit*, limiting improvements.

3. Decision Tree PS

- **Behavior:** Strong spikes at 0 and 1 (almost binary propensities).
- **Issue:** Violates positivity; unusable for matching.
- **Why:** Trees create hard splits, leading to deterministic assignment.also at times it overfit.

4. Random Forest + DML

- **Behavior:** Smooth, continuous PS; treated & control distributions overlap across 0.3–0.8.
- **Why:** RF captures *nonlinearities, interactions, and heterogeneous effects* without requiring explicit specification.

5. Random Forest + DML + Crossfit (Best Model)

- **Behavior:** Highest overlap; no extreme spikes; balanced PS distribution.
- **Why it works best:**
  - Crossfit reduces overfitting.
  - RF naturally models nonlinear covariate–treatment patterns.
  - Stabilizes both PS and outcome models -> strongest orthogonalization.

Extension Setup

Estimate causal impact of **formal (ditf)** and **informal credit (diti)** on agricultural yield using modern ML-based causal inference.

Methods Used

1. Double Machine Learning (DML)

- Propensity models: **Logistic Regression** and **Random Forest**
- Outcome models: **Linear Regression** and **Random Forest**
- **5-fold cross-fitting** to avoid overfitting
- Outputs: **ATE & ATT** via **orthogonal score construction**

2. Meta-Learners (Treatment Effect Learners)

All implemented using Random Forest regressors:

- **S-Learner** – Single joint model
- **T-Learner** – Separate treated & control models
- **X-Learner** – Imputation + re-weighting
- **DR-Learner** – Doubly-robust pseudo-outcomes
- **D-Learner** – Orthogonal pseudo-outcome regression

Each provides:

- Individual treatment effect τ_i
- **ATE = mean(τ_i)**
- **ATT = mean(τ_i | treated)**

Diagnostics & Inference

- Propensity score overlap checks
- τ (treatment effect) distribution checks
- **Z-tests, p-values, 95% CIs** for both ATE and ATT
- Final comparison table generated for each estimator

Extension Results

Method	ATE	ATE CI	ATT	ATT CI	p-value (ATT)
Formal Credit (DITF)					
DML (Logit+LR)	128.7	[-231.4, 488.8]	-22.0	[-44.19, 0.15]	0.0516
DML (RF)	6.28	[-33.1, 45.7]	8.54	[-12.9, 30.0]	0.4350
S-Learner	12.86	[11.94, 13.77]	11.32	[10.15, 12.49]	0.0000
T-Learner	18.79	[11.68, 25.90]	6.25	[-5.04, 17.55]	0.2779
X-Learner	28.54	[25.25, 31.82]	24.80	[21.13, 28.48]	0.0000
DR-Learner	21.97	[5.94, 37.99]	2.30	[-21.52, 26.13]	0.8497
D-Learner	78.96	[0.09, 157.83]	92.19	[-40.29, 224.67]	0.1726
Informal Credit (DITI)					
DML (Logit+LR)	59.55	[-90.86, 209.96]	-21.23	[-46.48, 4.02]	0.0994
DML (RF)	-21.62	[-48.07, 4.83]	-20.64	[-43.44, 2.15]	0.0759
S-Learner	1.66	[1.25, 2.07]	1.02	[0.30, 1.74]	0.0057
T-Learner	-15.96	[-22.93, -8.98]	-8.90	[-19.11, 1.32]	0.0881
X-Learner	-5.15	[-8.84, -1.47]	-13.95	[-20.21, -7.69]	0.000013
DR-Learner	-5.35	[-20.28, 9.58]	-9.70	[-33.89, 14.49]	0.4320
D-Learner	-0.57	[-59.43, 58.30]	-0.67	[-69.20, 67.86]	0.9847

Table 1: Causal Effect Estimates: ATE and ATT with 95% Confidence Intervals

Extension Inference

1. Formal Credit (DITF)

- Across the most reliable learners (X-Learner, S-Learner), formal credit has a positive and statistically significant impact on crop yield.
- Estimated ATE ranges from +13 to +29 kg, and ATT from +11 to +25 kg.
- Treatment-on-treated significance ($p \approx 0.0000$) shows that households who actually receive formal credit experience meaningful productivity gains.
- DML estimates show wider uncertainty, but the direction remains predominantly positive.

2. Informal Credit (DITI)

- X-Learner and T-Learner consistently show negative and statistically significant effects.
- Estimated ATE ≈ –5 to –16 kg, ATT ≈ –9 to –14 kg, with strong significance for X-Learner.
- This suggests that informal credit is associated with lower yields, even after adjusting for covariates and selection bias.

3. Interpretation of Meta-Learner Consensus

- The convergence of X-Learner, S-Learner, and T-Learner indicates a robust treatment signal, independent of modeling assumptions.
- Meta-learners outperform DML in stability due to their ability to model heterogeneous treatment effects directly.
- The consistency across multiple estimators strengthens the causal claim.

Overall Inference

- Formal credit improves agricultural productivity.
- Informal credit reduces agricultural productivity.
- These results are statistically robust and consistent across heterogeneous ML estimators