Machine Learning in Casual Inference: A Survey and Application in Economics

Pengxiang Zhou Yanyu Chen

Abstract—Based on the counterfactual framework, this paper compares machine learning in causal inference and explore the method dealing with issues when assumptions are not satisfied, such as confounding latent variables. Intersection of econometrics and machine learning provides insight into heterogeneous treatment effect of economic policy.

Index Terms—Casual Inference, Counterfactual Framework, Uplift Tree, Meta Learning

I. Introduction

Following the potential outcomes and counterfactual framework, we then posit the existence of potential outcomes $Y_i\left(1\right)$ and $Y_i\left(0\right)$ corresponding respectively to the response the subject would have experienced with and without the treatment, W_i. Counterfactual framework assumes unconfoundedness $\left\{Y_i\left(1\right), Y_i\left(0\right)\right\} \perp W_i | X_i$ and overlap $p\left(x\right) = pr\left(W_i = 1 | X_i = x\right) \in \left(0, 1\right)$, where $p\left(x\right)$ is the propensity score of being treated. We define the treatment effect at x as $\tau\left(x\right) = E\left[Y_i\left(1\right) - Y_i\left(0\right) | x_i = x\right]$.

Uplift Modeling is developed as an incremental modeling to improve the problems of response modeling. Meta-algorithms[1] use base learners, such as random forest and BART to estimate the outcomes separately for units under control and those under treatments and then take difference between them, and we have T-learner[1] in the binary treatment case and separate model approach in the multi-treatment[2] case. X-learner[1] use two base learners but impute the treatment effects for the individuals in the treated group, based on the control outcome estimator, and the treatment effects for the individuals in the control group, based on the treatment outcome estimator. Results show that it has great performance when the size of treatment group is small. Several researches estimate heterogeneous treatment effect using machine learning method, such as SVM[3], KNN[4], BART[5], tree[6], random forests[7], neural network[8, 9], etc. Moreover, the method mentioned in[3] estimate the outcome using single predictive model with all features and treatment indicator. S-learner methods is also purposed in multi-treatment experiment, such as contextual treatment selection[2], which performs the split that brings the greatest increase in expected response. Tree-based algorithms are widely used in uplift models but always performs with different splitting criterions[7, 10] and also double machine learning(subsampled)[7, 11]. There are methods developed to meet with the challenges from confounding features using propensity score[12] and non-randomized experiment using deep instrumental variables[13] and IV forest[14].

Uplift modeling is a predicting way about the incremental impact of a treatment on behavior in economic policies. Davis J, Heller S. B.(2017) [15] used the causal forest method to analyze the effect of youth employment plan, but found that there was no significant heterogeneous treatment effect on the impact of youth employment plan on the number of violent crimes arrested within two years, but there were systematic differences on the impact of youth employment plan on the employment of adolescents. Datong P.Hou * et al. (2017) [16] analyzed the treatment effect of Demand Response, Smart Home Automation and other power-saving incentives by using causal decision tree, K nearest neighbor matching and lasso/ridge regression, etc., and found the heterogeneous treatment effect of power-saving scale on climate temperature, and the higher the temperature, the larger the power-saving scale.

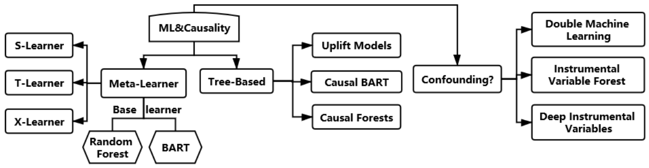

Fig. 1: Literature Review

This paper mainly makes a systematic and in-depth study on the application of machine learning in causal inference, mainly including the elaboration of the theory of uplift models and the policy analysis of land right confirmation using uplift models.

II. Method

A. Counterfactual Framework

Denote the processing variable as W_i, If received processing allocation, then $W_i = 1$, if not, $W_i = 0$. Denote the covariate as X_i. Denote the response variables as Y_i^{obs}.

$$Y_i^{obs} = \left\{ \begin{array}{l} Y_i\left(0\right) \ if \ W_i = 0 \\ Y_i\left(1\right) \ if \ W_i = 1 \end{array} \right.$$

Causal forest is based on a counterfactual framework and random control trial, so the non-obfuscation allocation mechanism is as follows.

First is overlap assumptions. is the probability that the individual is treated.

$$p\left(x\right)\in\left(0,1\right),p\left(x\right)=pr\left(W_i=1|X_i=x\right)$$

Second, it needs to be non-confounding.

$$\left\{Y_i^{(1)},Y_i^{(0)}\right\}\perp W_i|X_i$$

Finally, we get the processing effect.

$$\tau\left(x\right)=E\left[Y_i^{(1)}-Y_i^{(0)}|x_i=x\right]$$

B. Meta-learner

There are three main methods of meta-learning, as follows.

1) T-Learner: T, which stands two, is the way that the traditional machine learning model is used for causal inference. T-learner obtains two models by modeling the control group and the experimental group respectively, and calculates the difference between the predicted values of the two models as the estimation of HTE for each sample. It can be divided into two steps. The first step is to use a base learner, such as regression model or any supervised learning model to estimate the control the response function. And the estimated function is denoted as $\hat{\mu}_0$

$$\mu_0\left(x\right)=E\left[Y\left(0\right)|X=x\right]$$

Second, treatment response function is estimated with a potentially different base learner. And the estimated function is denoted as $\hat{\mu}_1$

$$\mu_1\left(x\right)=E\left[Y\left(1\right)|X=x\right]$$

Then we can get T-learner as follows.

$$\hat{\mathcal{T}}_T\left(x\right)=\hat{\mu}_1\left(x\right)-\hat{\mu}_0\left(x\right)$$

However, there are some problems with it. The model of the control group could not learn the pattern of the experimental group, and the model of the experimental group could not use the data of the control group. The two models are completely isolated, which leads to the possibility that the two models may have their own deviation, resulting in a large error in the prediction. In addition, T-learner requires treatment to be a discrete value. And in most cases, treatment effect is small compared with response, so the estimation deviation in response will have a great impact on treatment.

2) S-Learner: S, which stands single, is to model the control group and the experimental group together, and add the experimental groups as features to the training features. Then, the imputation method is used to calculate the difference predicted by the model if the sample entered the experimental group versus the control group as an estimate of the impact of experiment. By using any base learner on the entire dataset, the combined response function is estimated as follows.

$$\mu\left(x,w\right)=E\left[Y^{obs}|X=x,W=w\right]$$

We denote the estimator as$\hat{\mu}$. The CATE estimator is then given as follows.

$$\hat{\mathcal{T}}_s\left(x\right)=\hat{\mu}\left(x,1\right)-\hat{\mu}\left(x,0\right)$$

However, there are some problems with it. The essence of S-learner is to fit response. If the tree is used as the base learner, the final HTE can be simply understood as the sample falls on different leaf nodes and the sample difference of leaf nodes. However, since the tree itself is to model the outcome rather than the treatment effect, it is likely that an effective way to classify the population cannot be learned in this case.

3) X-Learner: X-Learner integrates T-Learner and S-Learner. There are three steps of X-learner. First, use some algorithms, such as supervised learning algorithm or regression algorithm, to estimate the response functions, which is the same with T-learner.

$$\mu_0\left(x\right)=E\left[Y\left(0\right)|X=x\right]$$

$$\mu_1\left(x\right)=E\left[Y\left(1\right)|X=x\right]$$

And denote the estimated functions $\hat{\mu}_0$ and $\hat{\mu}_1$.

Second, impute the individual treatment effects in the treated group according to the control outcome estimator and the. And impute the individual treatment effects in the control group according to the treatment-outcome estimator.

$$\tilde{D}_i^1=Y_i^1-\hat{\mu}_0\left(X_i^1\right)$$

$$\tilde{D}_i^0=\hat{\mu}_1\left(X_i^1\right)-Y_i^0$$

Then estimate $\mathcal{T}\left(x\right)$ in two ways: using the imputed treatment effects as the response variable in the control group to get $\hat{\mathcal{T}}_0\left(x\right)$ and similarly in the treatment group to get $\hat{\mathcal{T}}_1\left(x\right)$.

Third, get the CATE estimate through a weighted average of the two estimates in stage 2 as follows.

$$\hat{\mathcal{T}}\left(x\right)=g\left(x\right)\hat{\mathcal{T}}_0\left(x\right)+\left(1-g\left(x\right)\right)\hat{\mathcal{T}}_1\left(x\right)$$

Here $g\left(x\right)\in\left[0,1\right]$ is a weight function.

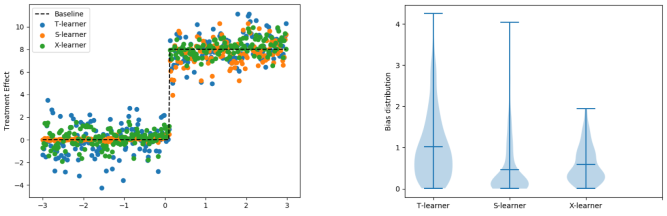

Fig. 2: Treatment effect

C. Tree-based Model

1) Causal Tree: The advantage of trees over linear models is that they are compatible with feature types, especially considering the fact that there are a lot of discrete features such as gender, region and so on. The general decision/regression tree is a fitting of Y such

as RMSE or cross-entropy, etc. We chose to maximize $Y(1) - Y(0)$ as the cost Function, and divided the local population through the tree to maximize the local experimental effect (positive or negative).

Denote the leaf node - local sample as S_l.

$$S_l = (X_i, Y_i, T_i) : X_i \in X_l$$

Denote the average of Y in AB group as $\hat{\mu}_t(S_l)$.

$$\hat{\mu}_t(S_l) = \frac{1}{N_{l,t}} \sum_{T_i = t, i \in S_l} Y_i$$

The loss function is as follows.

$$\hat{\mathcal{T}}(S_l) = \hat{\mu}_1(S_l) - \hat{\mu}_0(S_l)$$

$$F(S_l) = N_l * \hat{\mathcal{T}}^2(S_l)$$

$$\max \sum_{i=1}^{l} F(S_l)$$

The biggest problem with trees is overfitting, because every split must bring Information Gain. This is about the most classic bias-Variance Trade off in ML. The smaller the node of the tree partition, the smaller the estimation bias of the sample but the greater the variance. The traditional decision tree generally solves the problem of overfitting through several methods, such as cross-validation to determine the depth of the tree and stopping growth with parameters such as the minimum sample size of leaf nodes. Recently, researchers proposed two approaches to overfitting, Honest Approach and Variance Penalty. Honest Approach divides the training sample into train and est, and trains the model with train and gives the estimation of each leaf node with est. Variance penalty adds the variance of the leaf node directly into the cost function. In this paper, we use variance penalty to deal with overfitting. The final loss function is as follows.

$$F(S_l) = N_l * \hat{\mathcal{T}}^2(S_l) - N_l \left(\frac{Var(S_{l,1})}{p} + \frac{Var(S_{l,0})}{1-p} \right)$$

2) Uplift Tree: The goal of most classifications is to achieve a higher accuracy rate based on a given data set. However, in more practical cases, such as whether to mail or treat a patient, the purpose is to predict the category. Instead of modeling them in terms of their categories, we should model them in terms of the probability of change due to our actions, performing actions on the most profitable objects. Three different methods are implemented in the package to quantify the differences, namely KL, ED, and CHi.

- The Kullback-Leibler (KL) divergence is given by:

  $$KL(P:Q) = \sum\nolimits_{k=left,right} p_k \log \frac{p_k}{q_k}$$

  Where, p is the sample mean of the treatment group, and q is the sample mean of the control group. k is the leaf used to calculate p and q.
- The Euclidean Distance is given by:

  $$ED(P:Q) = \sum\nolimits_{k=left,right} (p_k - q_k)^2$$
- Finally, the χ^2-divergence is given by:

  $$\chi^2(P:Q) = \sum\nolimits_{k=left,right} \frac{(p_k - q_k)^2}{q_k}$$

3) Double Machine Learning: The aim of the Heterogeneous Treatment Effect is to quantify the differences between different groups of people, and then to conduct differentiation experiments by means of crowd orientation or numerical strategy, or to modify the experiments. Double Machine Learning (DML) takes Treatment as a feature and calculates the difference effect of the experiment by estimating the influence of features on the target.

Although Machine Learning (ML) is good at giving accurate predictions, economics pays more attention to unbiased estimation of characteristics' influence on targets. DML combines the method of economics with machine learning, and gives the unbiased estimation of the influence of characteristics on the target with arbitrary ML model under the framework of economics.

DML model can be divided into three steps.

First, the residual $\tilde{Y}$, $\tilde{T}$ are obtained by fitting Y and T with any ML model.

$$\tilde{Y} = Y - l(x) \;\; where \; l(x) = E(Y|x)$$

$$\tilde{T} = T - m(x) \;\; where \; m(x) = E(T|x)$$

Second, for $\tilde{Y}$, $\tilde{T}$, use any ML model to fit $\hat{\theta}$. The $\theta(x)$ fit can be a parametric model or a nonparametric model which can be fitted directly. As a non-parametric model only accepts input and output, the following transformation is required. The model target becomes $\tilde{Y}/\tilde{T}$ and the sample weight is $\tilde{T}^2$.

$$\tilde{Y} = \theta(x)\tilde{T} + \varepsilon$$

$$arg \min E\left[\left(\tilde{Y} - \theta(x) \cdot \tilde{T} \right)^2 \right]$$

$$E\left[\left(\tilde{Y} - \theta(x) \cdot \tilde{T} \right)^2 \right] = E\left[\tilde{T}^2 \left(\frac{\tilde{Y}}{\tilde{T}} - \theta(x) \right)^2 \right]$$

Third step is cross-fitting. An important step for DML to ensure unbiased estimation is cross-fitting to reduce the estimation deviation caused by overfitting. First, divide the total sample into two parts: sample 1 and sample 2. Use sample 1 to estimate the residuals, sample 2 to estimate $\hat{\theta}^1$. Next, use sample 2 to estimate the residuals, sample 1 to estimate $\hat{\theta}^2$. Then take the averages to get the final estimate. The K-fold can be further used to increase the robustness of the estimate.

$$sample_1, sample_2 = sample\_split$$

$$\theta(x) = \hat{\theta}^1 + \hat{\theta}^2$$

III. Simulation Experiments

We use X-learner, T-learner, S-learner, Forest double machine learning and also casual forest to predict the treatment effect for 20 times with different sample size ranging from 1000 to 5000. To measure the accuracy of the individual treatment effect, we use average square error of individual treatment effect, average absolute error and average absolute percentage error to compare the machine learning method in casual inference.

A. Simulation with confounding latent variables

We experiment on a simulated dataset where the marginal distribution of X is a mixture of Gaussians, with the hidden variable Z determining the mixture component. We generate synthetic data by the following process:

$$z_i \sim Bern\left(0.5\right); x_i|z_i \sim N\left(z_i, \sigma^2_{Z_1} Z_i + \sigma^2_{Z_0}\left(1 - Z_i\right)\right)$$

$$t_i|z_i \sim Bern\left(0.75z_i + 0.25\left(1 - z_i\right)\right)$$

$$y_i|t_i, z_i \sim Bern\left(Sigmoid\left(3\left(z_i + 2\left(2t_i - 1\right)\right)\right)\right)$$

Where $\sigma_{Z_0} = 3, \sigma_{Z_1} = 5$, and Sigmoid is the logistic sigmoid function. This generation process introduces hidden confounding between t and y as t and y depend on the mixture assignment z for x.

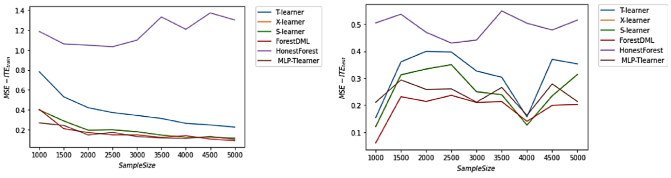

Fig. 3: Simulation With Confounding Latent Variables

IV. Analyze the Treatment Effect of Land Confirmation

This paper mainly focuses on how land ownership confirmation affects farmers' land transfer decision, whether land ownership confirmation will promote farmers' income growth, and what factors affect the treatment effect of land ownership confirmation policy. When the community/village completes the land ownership confirmation, $W_i = 1$ is taken as the treatment group. The covariates include the degree of agricultural mechanization (whether threshing machines are available or not, whether organic agricultural implements are available), household endowment (total household assets, land area owned by the household, hours of work and farming per day), the number of children and elderly people, and the educational level of the household, which are recorded as X_i. We take the income of farmers as a potential outcome. The average treatment effect of land title confirmation is denoted as $\tau\left(x\right) = E\left[Y_i^{(1)} - Y_i^{(0)}\right]$ treatment effect function is denoted as $\tau\left(x\right) = E\left[Y_i^{(1)} - Y_i^{(0)}|X_i = x\right]$.

A. Dataset and Features

In 2009 and 2010, the No. 1 document of the CPC Central Committee pointed out that we should continue to do a good job in land contract management, fully implement the "four to four households" of contracted plots, areas, contracts and certificates, and stabilize and expand the trial scope of registration of contracted management right of rural land. The survey data used in this paper are samples of farmers who had collectively allocated farmland, forest land, pasture or fish pond or rented farmland, forest land, pasture or fish pond from others in 2011 in the China Health and Retirement Longitudinal Study Database.

A total of 4,554 samples were used after the missing values were removed from the samples of farmers who had collectively allocated farmland, woodland, pasture or fish pond or rented farmland, forest land, pasture or fish pond from others. The variables used in this paper to estimate the impact of the work of confirming, registering and certifying the right of contracted land management on the land circulation decision-making and income of peasant households are shown in the table 1.

B. Test on Assumptions and Modifications

First we use overlap test and find that the estimated propensity scores is close to one or zero in figure 4.

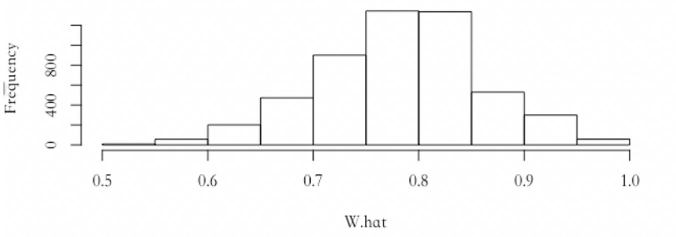

Fig. 4: Overlap Test

Derive from figure 5, we derive that the covariates are balanced across the treated and control group through plotting the inverse-propensity weighted histograms of all samples.

Suppose $\left\{Y_i^{(1)}, Y_i^{(0)}\right\} \perp W_i|X_i, Y = \tau\left(x\right) * W + c\left(x\right) + \varepsilon, W = f\left(x\right) + \eta$, but the estimated $\tau\left(x\right)$ is biased if there exists confounding variable. To be accurate, the bias resulting from confounding variable is $\left(\frac{1}{n}\sum W_i^2\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum f\left(x_i\right)\left(c\left(x_i\right) - \hat{c}\left(x_i\right)\right)\right)$. Here we apply the method of double machine learning. Thus, $Y - E\left(Y|X\right) = \tau\left(x\right) * \left(W - E\left(W|X\right)\right) + \varepsilon$, namely $\tilde{Y} = \tau\left(x\right) * \tilde{W} + \varepsilon$ where $\tilde{Y} = Y - E\left(Y|X\right)$, $\tilde{W} = W - E\left(W|X\right)$. In our model, we use leave one out cross validation to estimate regression tree for $E\left(Y|X\right)$ and $E\left(W|X\right)$, and then apply subsampled honest tree to estimate the heterogeneous treatment effect. We denote it as ForestDML for convenience.

On the other hand, we use cluster-robust subsampled honest tree for sensitivity analysis. Cluster-robust subsampled honest tree gets certain amount of sample from

TABLE I: Definition of Variables in Dataset

Notation	Description
Income	Annual household income (unit: yuan)
Land ownership confirmation	Denoted as 1 if recorded, otherwise it is 0
Land transfer	Denoted as 1 if the farmer transfers its land, otherwise it is 0
Type of land transfer	Denoted as -1 if the farmer rents in land, rent out land as 1, otherwise 0
Amount of the transferred land	Amount of the transferred land (unit: mu)
Thresher	Denoted as 1 if it has a thresher, otherwise it is 0
Organic driving farm tools	Denoted as 1 if it has organic driving farm tools, otherwise it is 0
Education	The higher the level of education, the higher the value
Number of the elderly	Number of the elderly people in the family
The number of children	The number of children
Rent	Average rent of land (unit: yuan)
Wealth	Total assets of households (unit: YUAN)
Land	Amount of the land, including cultivated land, woodland, pasture, and pond (unit: mu).
Labor supply	The average number of hours a day spent on farming and working
Dummy variable	Province dummy variable

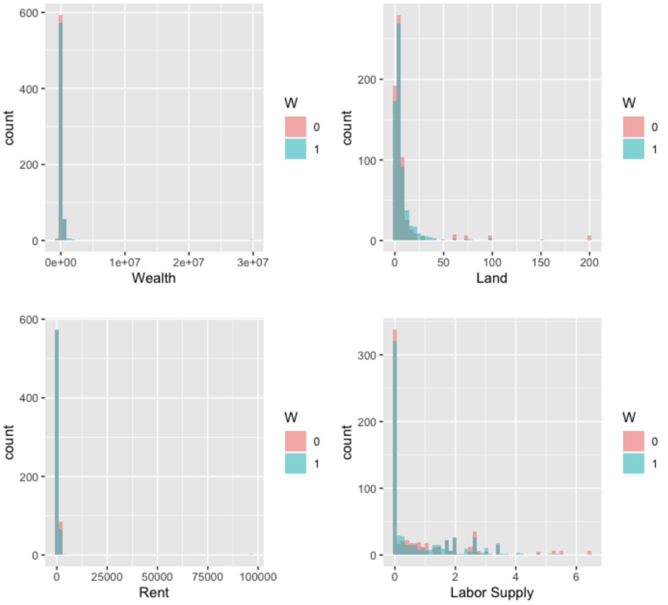

Fig. 5: Balance Test on Covariates

all the clusters(different provinces in out paper) and use the sample data to estimate. We assume that the farmers' income from province $P_i \in \{1, \cdots, j\}$ is Y_i and different province shares the same weight, then the average and variance of the farmers' income is $\hat{\mu}_j = \frac{1}{n_j} \sum_{\{i:P_i=j\}} Y_i$, $\hat{\mu} = \frac{1}{\mathcal{J}} \sum_{j=1}^{\mathcal{J}} \hat{\mu}_j, \hat{\sigma}^2 = \frac{1}{\mathcal{J}(\mathcal{J}-1)} \sum_{j=1}^{\mathcal{J}} \left(\hat{\mu}_j - \hat{\mu}\right)^2$ Using B trees, estimation of the farmers' income is

$$\hat{\mu} = \frac{1}{B} \sum_{b=1}^{B} \sum_{i=1}^{n} \frac{Y_i * 1\left(X_i \in L_b\left(x\right), i \in S_b\right)}{\left|i : X_i \in L_b\left(x\right), i \in S_b\right|}$$

where $L_b\left(x\right)$ is the leave of b^{th} tree and S_b is the subsample for b^{th} tree.

C. Policy analysis in land confirmation

1) Average treatment effect estimation: The estimated results of the model are shown in table 2 with sample size at 5208. After the confirmation of land ownership, the income growth of farmers is about 2300 yuan / year.

TABLE II: Estimation of Average Treatment Effect

Model	Average treatment effect	95% CI
ForestDML	2336.3 1855.7	[-1300.7,5973.3]
Cluster-Robust ForestDML	2319.3 315.6	[-4179.1,8817.6]

Furthermore, we obtain the distribution of individual treatment effects in figure 6, the left subplot represents the results from ForestDML and the other represents cluster-robust ForestDML. We find that the average treatment effect of the cluster robust subsampled honest forest considering the heterogeneity of farmers in different provinces is similar to that of the ordinary forest. It is worth noting that the range of individual processing effects estimated by ordinary forest is twice as large as that of cluster robust subsampled honest forest, but the variance and the length of 95% confidence interval is small.

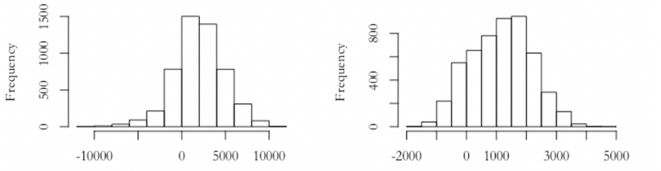

Fig. 6: Distribution of Individual Treatment Effect

Using the frequency with which features are selected to segment training data as the scoring criteria, we analyze the importance of features selected for ForestDML and Cluster-Robust ForestDML. The results are shown in Table 3. We let $V_i\left(X_k\right)$ be the importance of the covariable X_k, and let j be the depth of the tree. In our paper, we only use the segmentation with a depth of no more than 4 as the measure of the importance of the feature. Let n_j be the number of cutting with a depth of j, let n_{jk} be the number of segmentation with a depth of j as the covariable X_k as the number of segmentation features, and $W_j = \frac{j^{-2}}{\sum_{d=1}^{4} d^{-2}}$ is the preset weight. Results show that wealth, land, rent and Labor supply show greater importance and the characteristics of farm households (education level, the number of children and the elderly) show smaller importance.

2) Test on Heterogeneity: We use linear regression to analyze the heterogeneity of the ForestDML and Cluster-Robust ForestDML. Let $Y_i - \hat{y}^{-i}\left(X_i\right)$ be explained variable and let $C_i = \tau\left(W_i - \hat{w}^{(-i)}\left(X_i\right)\right)$ be explaining variable, where $\hat{y}^{-i}\left(x_i\right)$ is the estimated outcome using regression forest $\hat{w}^{(-i)}\left(X_i\right)$ is the estimated propensity score $\overline{\tau}$ is the average treatment effect $\hat{\tau}^{(-i)}\left(X_i\right)$ is the i^{th} individual treatment effect. We believe that C_i measures the impact of average treatment effect and D_i measures the impact of heterogeneous treatment effect[17]. If the coefficient of D_i is 1 shows the model can greatly capture the heterogeneous treatment effect. Results show that the ForestDML greatly estimates the average treatment effect and heterogeneous treatment effect while not cluster-robust.

TABLE III: Heterogeneity Test

	ForestDML	CRF DML
C_i	0.923* (0.715)	2.842 (0.973)
D_i	1.450** (0.822)	-0.858 (0.686)

Note: *, **, *** indicate 10%, 5% and 1% significance levels, respectively.

3) Heterogeneous Treatment Effect of Land Confirmation: We sample from the whole dataset with the top 1% of individual treatment effect (about 53 samples) and samples from the whole dataset the bottom 1% (about 53 samples), and find that the income of the top 1% of farmers increases by about 8469.34 yuan per year, while the income of the bottom 1% of farmers decreases by about 4,574.47 yuan per year. Further, we calculate the mean of all the covariates and the conditional average treatment effect. On the one hand, we find that for the top 1% of household groups in terms of individual treatment effect, the proportion of farmers who rent out their land was higher than the proportion of farmers who rent in their land.

As for the top 1% of the treatment effect farmers, wealth, education level and labor supply are above the average level. We believe that for this group, they are more inclined to rent out land. Good education and high working ability can help them find matching jobs in the industrial sector. The elderly in the family (1.45 on average) are able to take care of the children, thus eliminating the need to care for the family. In addition, we find that the average rent of land (about 46.61 yuan/mu) and the average land area (about 1.66 acres) are far below the average level, we think that for this group of the farmers, in order to meet the land property rights protection, care left-behind children, production support, farmers often transfer low-value land to relatives and friends for free or at a low price in exchange for favors[18]. Land confirmation guarantees farmers' land use right and land management right, relieve farmers' concern of losing their land, thus promoting the transaction of land means of production and labor force, and promote income growth. As for the bottom 1% of the treatment effect farmers, they owns about 2.03 mu land endowment, which is lower than the average land area of 8.42 mu, and has lower mechanization of agricultural production. However, the rent available to them is relatively high, about 574.33 yuan/year, and the proportion of farmers who rent out the land is much higher than the average level. We think land confirmation promotes farmland transfer. However, for this part of the farmers, after they rent out their land, they have greater probability of being unemployed due to weak working ability and low degree of education. Therefore, we suggest the government to ensure sufficient non-farm jobs, while not just changing the allocation of land use rights and management rights in the farmer, and prevent farmers the risk of chronic poverty.

We further analyze the distribution of individual treatment effect of the characteristics of the high importance score (including wealth, land, rent, and labor supply), see in fig 5.

We find that the lower the rent, the larger the treatment effect of land confirmation. In our opinion, the lower the land rent is, the lower the expectation of farmers on land use due to the instability of land property rights, which will further reduce the reserve price of land for farmers who rent the land. Moreover, land confirmation can effectively eliminate the risks including land disputes caused by the imperfect land system. In addition, through the full implementation of "Four to the household" for contracting the plot, area, contract and certificate, we promote land transaction and improve the exchange value of land, which further strengthen the treatment effect of land confirmation. In addition, there are systematic differences in the treatment effect of land right confirmation on farmers with different labor supply. Moreover, the greater the labor supply is, the stronger the treatment effect of land confirmation is.

In our opinion, the household contract responsibility system (1978) was limited by the household registration system, while the heterogeneity of farmers in agricultural production was ignored, resulting in the small operation land area of farmers, and also fragmented and scattered plots. Land confirmation ensures the stable and permanent land management right of peasant households. Through land transaction, peasant households with stronger farming ability and longer working hours can rent in land and expand the scale of agricultural production. On the other hand, farmers with different land endowment and wealth endowment are also affected by the land confirmation system differently. Land ownership confirmation encourages farmers to choose the optimal allocation of land resources, and selects the scale of agricultural production by means of renting and renting out land. For example, farmers with strong agricultural capacity but less land endowment can choose to rent in land to achieve the adaptation of agricultural production capacity and farm area.

V. Conclusion

Based on the counterfactual framework, this paper analyzes the application of machine/deep learning in causal inference and focuses on the study and exploration of meta learner and uplift models. We also explore the

TABLE IV: Compare the top 1% and the bottom 1%

Covariate	Wealth	Land	Labor supply	Rent	Thresher	Children	Education	Land transfer	Elderly
top 1%	312090.90	1.66	1.44	46.61	0.11	2.62	4.088	0.04	1.45
bottom1%	-2541.61	2.03	0.03	574.33	0.08	2.40	2.36	0.15	0.23
Full sample	119860.30	8.42	0.97	318.33	0.11	2.89	3.07	-0.01	0.73

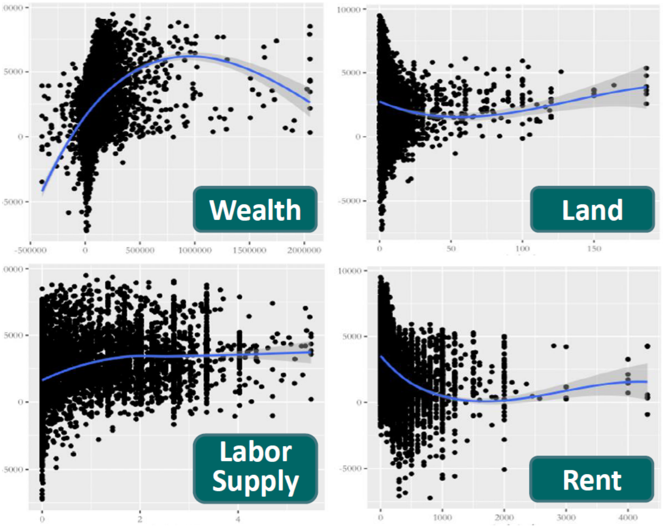

Fig. 7: Distribution of Individual Treatment Effect

method dealing with confounding latent variables, such as double machine learning method, instrument variable forest and deep instrumental variable method.

We test on the overlapping, balance assumptions and also heterogeneity test, the natural experiment, land confirmation certainly satisfies the assumptions of the counterfactual framework and has heterogeneous treatment effect. Moreover, we consider the different pattern in different province and we use cluster-robust forest double machine learning for sensitivity analysis. In conclusion, land confirmation can promote the income growth of farmers, increase the scale of land usage right transaction. In addition, we find that the confirmation of land rights eliminates the risk of land disputes after the outflow of land, but the gap between rich and poor farmers after the lease of land widens. Future direction of the research may focus on the method propose for high-dimension data and network-effect.

References

[1] S. R. Künzel, J. S. Sekhon, P. J. Bickel, and B. Yu, ”Metalearners for estimating heterogeneous treatment effects using machine learning,” Proceedings of the national academy of sciences, vol. 116, no. 10, pp. 4156-4165, 2019.

[2] Y. Zhao, X. Fang, and D. Simchi-Levi, ”Uplift modeling with multiple treatments and general response types,” in Proceedings of the 2017 SIAM International Conference on Data Mining, 2017: SIAM, pp. 588-596.

[3] K. Imai and M. Ratkovic, ”Estimating treatment effect heterogeneity in randomized program evaluation,” The Annals of Applied Statistics, vol. 7, no. 1, pp. 443-470, 2013.

[4] G. J. Hitsch and S. Misra, ”Heterogeneous treatment effects and optimal targeting policy evaluation,” Available at SSRN 3111957, 2018.

[5] D. P. Green and H. L. Kern, ”Modeling heterogeneous treatment effects in survey experiments with Bayesian additive regression trees,” Public opinion quarterly, vol. 76, no. 3, pp. 491-511, 2012.

[6] S. Athey and G. Imbens, ”Recursive partitioning for heterogeneous causal effects,” Proceedings of the National Academy of Sciences, vol. 113, no. 27, pp. 7353-7360, 2016.

[7] S. Wager and S. Athey, ”Estimation and inference of heterogeneous treatment effects using random forests,” Journal of the American Statistical Association, vol. 113, no. 523, pp. 1228-1242, 2018.

[8] U. Shalit, F. D. Johansson, and D. Sontag, ”Estimating individual treatment effect: generalization bounds and algorithms,” in International Conference on Machine Learning, 2017: PMLR, pp. 3076-3085.

[9] F. Johansson, U. Shalit, and D. Sontag, ”Learning representations for counterfactual inference,” in International conference on machine learning, 2016, pp. 3020-3029.

[10] P. Rzepakowski and S. Jaroszewicz, ”Decision trees for uplift modeling with single and multiple treatments,” Knowledge and Information Systems, vol. 32, no. 2, pp. 303-327, 2012.

[11] V. Chernozhukov et al., ”Double/debiased machine learning for treatment and causal parameters,” 2017.

[12] P. R. Hahn, J. S. Murray, and C. Carvalho, ”Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects,” arXiv preprint arXiv:1706.09523, 2017.

[13] J. Hartford, G. Lewis, K. Leyton-Brown, and M. Taddy, ”Deep IV: A flexible approach for counterfactual prediction,” in International Conference on Machine Learning, 2017, pp. 1414-1423.

[14] S. Athey, J. Tibshirani, and S. Wager, ”Generalized random forests,” The Annals of Statistics, vol. 47, no. 2, pp. 1148-1178, 2019.

[15] J. Davis and S. B. Heller, ”Using causal forests to predict treatment heterogeneity: An application to summer jobs,” American Economic Review, vol. 107, no. 5, pp. 546-50, 2017.

[16] D. P. Zhou, M. Balandat, and C. J. Tomlin, ”Estimating Heterogeneous Treatment Effects in Residential Demand Response,” arXiv preprint arXiv:1710.03190, 2017.

[17] Athey S, Wager S. Estimating treatment effects with causal forests: An application[J]. arXiv preprint6 arXiv:1902.07409, 2019.

VI. Appendix

A. Method

1) Meta-learner: Generally speaking, machine learning is to let machine be able to learn, while meta learning is to teach machine how to learn. In machine learning, the training unit is a piece of data that is used to optimize the model. The data can be divided into training set, test set and validation set. In meta-learning, the training unit is hierarchical, and the first training unit is a task. Many tasks need to be prepared for learning, and the second training unit is the data corresponding to each task. The purpose of both is to find a function, but the two functions do different things. Function in machine learning is used for features and tags to find associations between features and tags. While function in meta-learning is used to find a new f, and the new f will be applied to the specific task.

Meta-learner is a framework that use any base machine learning to estimate conditional average processing effects (CATE). The meta-algorithm uses a single base learner with a treatment indicator as a feature (e.g., S-learner), or multiple base learners for each treatment group and control group (e.g., T-learner, X-learner, and R-learner).

2) Uplift Tree: There are two main approaches to the Uplift model. The most common method is to build two two-models, one model training treatment and the other model training control data. When used, we subtracted the predicted values of the two models as the final result. However, there is a drawback here, we want to predict the difference between the experimental group and the control group, while the model training goal is only to separate the positive and negative samples within the respective data groups, and we can't learn the difference between the two groups, which leads to the probability prediction value we want may be very different from a single model. In the case of decision tree, it is not conducive to splitting the corresponding differences between the experimental group and the control group caused by action, but to separate the predicted results in the respective groups of the experimental group and the control group. The second approach is to train a model that attempts to directly model the differences between the experimental and control groups. Rzepakowski proposed three different ways to quantify the gain in divergence as the result of splitting.

$$D_{gain} = D_{after_{split}}\left(P^T, P^C\right) - D_{before_{split}}\left(P^T, P^C\right)$$

Where D represents difference, P^T and P^C represent the probability distribution of results of interest in the treatment group and the control group, respectively.

B. Simulation Measurement

$$MSE - ITE_{test} = \frac{1}{N}\sum_{i=1}^{N}\left((y_{i1} - y_{i0}) - (\hat{y}_{i1} - \hat{y}_{i0})\right)^2$$

$$MSE - ITE_{train} = \frac{1}{N}\sum_{i=1}^{N}\left((y_{i1} - y_{i0}) - (\hat{y}_{i1} - \hat{y}_{i0})\right)^2$$

$$MAE - ITE_{test} = \frac{1}{N}\sum_{i=1}^{N}\left|(y_{i1} - y_{i0}) - (\hat{y}_{i1} - \hat{y}_{i0})\right|$$

$$MAE - ITE_{train} = \frac{1}{N}\sum_{i=1}^{N}\left|(y_{i1} - y_{i0}) - (\hat{y}_{i1} - \hat{y}_{i0})\right|$$

$$MAPE - ITE_{test} = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{(y_{i1} - y_{i0}) - (\hat{y}_{i1} - \hat{y}_{i0})}{(y_{i1} - y_{i0})}\right|$$

$$MAPE - ITE_{train} = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{(y_{i1} - y_{i0}) - (\hat{y}_{i1} - \hat{y}_{i0})}{(y_{i1} - y_{i0})}\right|$$

VII. Analyze the Treatment Effect of Land Confirmation

TABLE V: Importance Score

Covariate	Forest DML	CRF DML
Wealth	0.241	0.207
Land	0.162	0.161
Rent	0.169	0.152
Labor Supply	0.119	0.179
Education	0.101	0.118
Number of Children	0.073	0.070
Number of the elderly	0.051	0.064
Thresher	0.001	0.000
Organic driving farm tools	0.000	0.000