Exploring the Complementarity Between Traditional Econometric Methods and Machine Learning – An Application to Adoption and Disadoption of Conservation Practices

Zhushan Du, Hongli Feng, and J. Arbuckle

Working Paper 25-WP 668
February 2025

Center for Agricultural and Rural Development
Iowa State University
Ames, Iowa 50011-1070
www.card.iastate.edu

An updated version of this working paper is forthcoming in *Applied Economics*

Zhushan Du is PhD Student, Department of Economics, Iowa State University, Ames, Iowa, 50011. E-mail: zhushand@iastate.edu.

Hongli Feng is Assistant Professor, Department of Economics, Iowa State University, Ames, Iowa, 50011. E-mail: hfeng@iastate.edu.

J. Arbuckle is Professor, Department of Sociology and Criminal Justice, Iowa State University, Ames, Iowa, 50011. E-mail: arbuckle@iastate.edu.

Exploring the Complementarity Between Traditional Econometric Methods and Machine Learning – An Application to Adoption and Disadoption of Conservation Practices

Zhushan Du, Hongli Feng, J. Arbuckle

An updated version of this working paper is forthcoming in Applied Economics

Author name and affiliations

Zhushan Du, zhushand@iastate.edu, Department of Economics, Iowa State University, United States

Hongli Feng, hfeng@iastate.edu, Department of Economics and Center for Agricultural and Rural Development, Iowa State University, United States

J. Gordon Arbuckle, arbuckle@iastate.edu, Department of Sociology and Criminal Justice, Iowa State University, United States

Abstract:

This study explores the potential complementarity between traditional econometric methods and machine learning in analyzing the adoption and disadoption of a key conservation practice in agriculture, cover crops. While the adoption of conservation practices has been widely examined, the literature on their disadoption is limited. Using a unique longitudinal panel survey of Iowa farmers, we compare logistic regression models with a Random Forest algorithm to examine factors driving conservation adoption and disadoption. Our findings show that while traditional logistic regression models offer interpretability grounded in economic theory, Random Forest provides superior predictive power and reveals complex, non-linear relationships among key factors such as past adoption behavior, cost-share participation, and farmer perceptions. SHAP (SHapley Additive exPlanations) analysis identifies adoption scale, past adoption behavior, and environmental factors as primary drivers of disadoption. Farmers with larger previous cover crop acreage and consistent adoption history are significantly less likely to disadopt, while the long-term impact of cost-share programs on continued use appears limited. By combining machine learning's predictive power with the interpretability of traditional econometrics, the study provides a deeper understanding of the drivers behind conservation decisions, which are crucial for informing policy design that promotes more sustainable adoption of conservation practices.

Key words:

Adoption; Conservation practices; Disadoption; Logistic regression; Machine learning

JEL codes: Q18, Q20, Q28

Introduction

Conservation practice adoption in agriculture is a widely discussed topic, with most literature focusing on initial adoption rather than continued use (Claassen et al., 2019). However, it is crucial to recognize that adoption decisions are not always permanent, and users may not always

maintain the technology or practice due to various factors, such as liquidity and labor constraints, or unsatisfactory yield response (Razafimahatratra et al., 2021). Understanding the circumstances related to discontinuing or alternating adoption behaviors could help with policy-making to promote more widespread long-term adoption, particularly for practices like cover crops that provide both private benefits to farmers and public benefits (Mitchell et al., 2017; Wulanningtyas et al., 2021).

Despite the importance of post-initial adoption behavior, research related to conservation practice disadoption remains limited. For the few studies that examine the adoption and disadoption of conservation practices, the primary data analysis methods are traditional econometric methods, such as logistic regression and bivariate probit models (Dunn et al., 2016; Neill & Lee, 2001). However, these methods often struggle to fully capture non-linear relationships and the complexity of factors influencing farmers' decisions due to their reliance on strict assumptions.

Machine learning (ML) models have become extensively used in many areas, from physics to genetics to social sciences. These models identify patterns and relationships through iterative processes and adjusting internal parameters to minimize prediction error, making them well-suited for capturing non-linear relationships and modeling interactions that traditional econometric methods might overlook. There can be potential complementarity between traditional econometric methods and machine learning models. While econometric methods provide a solid foundation based on economic theory, machine learning models can identify patterns and relationships that may not be easy to detect from a theoretical perspective or may even be seemingly counterintuitive.

This study compares a traditional econometric method - logistic regression - with a commonly used machine learning model – Random Forest - to examine factors influencing adoption and disadoption of conservation practice, specifically cover crops, while evaluating the strengths and limitations of each approach. Our research seeks to answer two main questions: what are the key factors driving the adoption and disadoption of conservation practice, and how can machine learning methods complement traditional econometric approaches and improve our understanding of farmers' decision-making?

We focus on four key factors identified as significant in the literature: past adoption behavior, complementarity with other conservation practices such as no-till, government support, and crop insurance. These factors have significant policy implications for promoting sustained adoption. We integrate data from the Iowa Farm and Rural Life Poll (IFRLP), a longitudinal survey of Iowa farmers conducted annually since 1982, with weather data from the National Oceanic and Atmospheric Administration (NOAA) and land data from the Natural Resource Conservation Service (NRCS). The IFRLP panel data documented information on individual-level conservation adoption and farmers' experiences, providing advantages over previous studies that relied on aggregate-level data (Plastina et al., 2024) or cross-sectional surveys with recall data (Dunn et al., 2016) and novel insights of adoption dynamics.

Our Random Forest model demonstrates superior predictive power (92.1% accuracy versus 85.1% for logistic regression) and greater robustness to outliers. It reveals complex, non-linear relationships in factors such as farmers' concern on yield impact that traditional logistic regression fails to capture. Our SHAP (SHapley Additive exPlanations) analysis, quantifying each variable's contribution to model's predictions, shows that farmers with larger cover crop acreage and recent adoption experience are less likely to disadopt. While traditional econometric

methods offer clear interpretation through marginal effects, ML approaches complement these insights through tools like accumulated local effect (ALE) plots and SHAP values, which help uncover complex patterns in adoption and disadoption.

Our study makes two key contributions to the literature. First, we advance beyond previous conservation practice adoption and disadoption studies that relied on cross-sectional surveys, recall data, or aggregate-level analysis by using longitudinal panel data that tracks actual adoption behaviors over time. We identify the key drivers of disadoption of cover crop by use SHAP values. Second, while machine learning methods have been shown to be useful in various economic applications, they have not been applied to conservation practice adoption decisions. This study demonstrates how integrating ML techniques can provide additional insights that traditional econometric methods may overlook. The integration of machine learning with econometric methods offers a promising direction for future research related to conservation policy design and evaluation and can be critical in informing effective policy interventions and promoting the sustainable adoption of conservation practices in agriculture.

Literature Review

Key Driving Factors of Adoption and Disadoption of Conservation Practices

Existing research has extensively examined factors underlying the adoption of conservation practices, focusing on economic considerations with perceived benefits and high costs (Arbuckle & Roesch-McNally, 2015; Pannell et al., 2006), however, post-adoption behavior related to conservation practices has received much less attention. While studies in general agricultural technology disadoption have shown that factors such as financial stability, training exposure, and profitability uncertainty can influence technology disadoption (Barrett et al. 2022; Moser and Barrett, 2006; Olivia et al. 2020), the literature focusing on conservation practices disadoption is limited. Using aggregate-level data, Plastina et al. (2024) found significant disadoption rates:

45.6% and 44.7% of U.S. counties experienced disadoption of cover crops and no-till, respectively. Previous studies have identified various factors influencing disadoption, including landowner support, willingness to learn (Dunn et al., 2016), land characteristics, farm size, and climate (Wade & Claassen, 2017), but face several limitations. Dunn et al. (2016) relied on a single cross-sectional survey with recall data of cover crop usage, while Wade and Claassen (2017) constructed panels from seasonal recall data, both potentially subject to memory bias. Wallander et al. (2017) used field-level data and focused on transition probabilities between tillage states rather than analyzing the underlying factors driving these changes. Our study advances beyond these approaches by using longitudinal panel data that tracks actual adoption behaviors over time, providing more reliable and more in-depth analysis of adoption dynamics.

Our analysis focuses on four factors. First, past adoption behavior shapes future adoption decisions through behavioral inertia, as farmers tend to maintain current practices due to awareness, perceived risks, and the costs associated with change (Prokopy et al., 2008). While habits and positive experiences can promote continued use (Dayer et al., 2018), farmers may cease using a practice if results are unsatisfactory (Pannell & Claassen, 2020). Second, complementarity with other conservation practices may influence the adoption or disadoption of another. For example, no till and cover crops can be synergistic, enhancing soil health and productivity (Mitchell et al., 2017; Wulanningtyas et al., 2021), leading farmers to adopt and maintain multiple complementary practices (Gong et al., 2021). Third, government support through cost-share programs support significantly increases adoption rates and contributes to the long-term maintenance of conservation practices (Claassen et al., 2019; Park et al., 2023). However, their effectiveness in promoting sustained use of conservation practices beyond the expiration of cost-share programs is less clear (Dayer et al., 2018). Finally, crop insurance may

discourage cover crop adoption, as farmers might perceive cover crops as a redundant risk management strategy. Empirical evidence indicates that crop insurance participation can disincentivize cover crop adoption (Connor et al., 2022; Delay, 2019).

Beyond these primary factors, studies have identified weather variability, particularly growing season precipitation, as another important factor in adoption decisions (Lee & McCann, 2019; Park et al., 2023). Cover crops can help manage both extreme rainfall through reduced erosion and drought conditions through improved soil moisture retention (Kaye & Quemada, 2017; Won et al., 2024).

Traditional Models and Alternative Approaches

Previous research has used a variety of econometric methods to investigate the adoption and disadoption of conservation practices, including bivariate probit models for sequential adoption (Neill & Lee, 2001; Khanna, 2001), Heckman probit model and double selectivity model to address sample selection bias (Khanna, 2001; Razafimahatratra et al., 2021), and multinomial logistic models to examine joint adoption and complementarity between practices (Gong et al., 2021). Traditional econometric models have several limitations. They often assume specific functional forms, which can lead to biased estimates if these assumptions are violated (Wooldridge, 2010). While non-parametric methods, such as kernel regression and generalized additive models can offer more flexibility, they may be less efficient with high-dimensional data and highly sensitive to the addition of covariates (Storm et al., 2020). It is important to note that while endogeneity is a common concern in econometric analysis, it is less emphasized in our study since our primary focus is on predictive modeling rather than causal inference.

Machine learning (ML) approaches have gained significant attention in economics due to their potential for expanding applied econometric methods, offering greater flexibility, requiring fewer assumptions, and capturing complex, non-linear relationships that are difficult to model

with theoretical approaches (Harding & Lamarche, 2021; Mullainathan & Spiess, 2017; Storm et al., 2020). They are well-suited for prediction tasks, handle high-dimensional data, and help with variable selection while avoiding overfitting (Hastie et al., 2009). However, these advantages can come at the cost of interpretability (Storm et al., 2020). To address this issue, researchers have developed various techniques such as partial dependence plots (PDP), accumulated local effect (ALE) plots, and SHapley Additive exPlanations (SHAP) values (Molnar, 2019; Lundberg & Lee, 2017).

Machine learning has gained interest in various economic fields, including macroeconomics, finance, environmental, and agricultural economics (Jarmulska, 2022; Prest et al., 2023; Shao et al., 2021; Yoon, 2021; Wang et al., 2021), with comparisons often favoring ML methods. Random Forest has outperformed logit models in predicting fiscal stress and systemic banking crises with higher accuracy (Jarmulska, 2022; Wang et al., 2021), while support vector machines and Random Forest have provided more accurate predictions of Chinese hog inventory compared to ordinary least squares (Shao et al., 2021). Ramsey and Bergtold (2021) showed that artificial neural networks can provide superior model fit and prediction accuracy compared to logit and probit models maintaining the ability to estimate economically meaningful measures such as marginal effects and willingness-to-pay.

Given these trade-offs between ML and traditional econometric methods, our study combines these two methods to leverage the strengths of both approaches - the theoretical rigor and interpretability of econometrics with the predictive power and flexibility of ML, which provides an opportunity to enhance our understanding of conservation practice adoption.

Description of Models

Logistic Regression and Its Foundation in Economics Theory

Logistic regression is a widely used statistical method across many fields. It is widely applied to model binary choice decisions, with its application grounded in economic theory through McFadden's (1972) connection between random utility theory and the logistic functional form. The theoretical grounding represents one of the key differences between traditional econometrics and machine learning approaches. While machine learning takes a more data-driven approach that focuses on optimizing prediction accuracy, econometric applications of logistic regression are guided by economic theory, which not only guides model specification but also provides economic interpretation of the estimated parameters.

We begin our analysis with the random utility model (RUM) framework. Consider a decision-maker i faced with a binary choice between two alternatives, $j = \{0, 1\}$, where 1 represents adoption and 0 represents non-adoption. According to the random utility model, the utility that decision-maker i derives from alternative j can be expressed as:

$$U_{ij} = V_{ij} + \epsilon_{ij},$$

where U_{ij} is the total utility, V_{ij} is the systematic component of utility, which is a function of observable covariates and unknown parameters to be estimated, and ϵ_{ij} is the random component of utility which contains all the unobserved determinants of the utility.

Several key assumptions and restrictions are necessary to derive the logistic functional form the random utility framework. First, the random components ϵ_{ij} are assumed to be independently and identically distributed (IID) across alternatives and decision makers, following an extreme value (Gumbel Type I) distribution. Second, the systematic component is generally V_{ij} is specified as a linear function of observed attributes, although can accommodate

some non-linearities (e.g., through variable transformations): $V_{ij} = \beta' X_i$, where β is a vector of coefficients to be estimated and X_i is a vector of observed variables affecting the utility of alternative j for decision-maker i.

The decision-maker chooses the alternative that maximizes their utility. Thus, alternative 1 (adoption) is chosen if and only if: $U_{i1} > U_{i0}$. Substituting the utility function and after re-arrangement, the probability of decision maker i choosing adoption P_{i1} becomes

$$P_{i1} = Prob(\epsilon_{i0} - \epsilon_{i1} < V_{i1} - V_{i0})$$

Following Train (2009), the probability of decision maker i choosing adoption P_{i1} and non-adoption P_{i0} is $P_{i1} = \frac{e^{\beta^T X_i}}{1+e^{\beta^T X_i}}$ and $P_{i0} = \frac{1}{1+e^{\beta^T X_i}}$, respectively, which corresponds to the logit choice probability in the binary response case.

The parameter β is estimated using the maximum likelihood estimation (MLE). The log-likelihood function for a total sample of N observations is given by:

$$L(\beta) = \sum_{i=1}^{N} \{y_i \log P(y_i = 1|X_i; \beta) + (1 - y_i) \log(1 - P(y_i = 1|X_i; \beta)\}$$

where y_i is the observed choices ($y_i = 1$ for adoption and $y_i = 0$ for non-adoption), and X_{ij} is the observed characteristic of the individual i.

Machine Learning Models

Logistic Regression with Regularization, LASSO and Ridge. In machine learning (ML) models, the primary goal is to achieve accurate prediction of outcomes based on available explanatory variables ("features" in the ML terminology). Model performance is measured by a loss function quantifying the difference between predictions and actual outcomes. To ensure the model generalizes well to unseen data, the dataset is randomly divided into a training subset (80% of

the data) and a testing subset (20% of the data). The model is trained on the training subset by minimizing the loss function and then evaluated on the testing subset to assess performance.

To address potential issue of overfitting and multicollinearity, we use two regularization techniques: Least Absolute Shrinkage and Selection Operator (LASSO) and Ridge regression. These methods improve the performance of logistic regression by adding penalty terms to the loss function, which discourage overly complex models and enhance generalization to new data.

LASSO adds a penalty term, which is the sum of the absolute values of the coefficients multiplied by a regularization parameter λ to the loss function. LASSO can help shrinking some coefficients to zero, effectively removing less important features from the model. In mathematical terms, the loss function of LASSO (L1-norm regulation) is:

$$L(\beta) = -\sum_{i=1}^{N}\{y_i \log P(y_i = 1|X_i;\beta) + (1 - y_i)\log(1 - P(y_i = 1|X_i;\beta)\} + \lambda\sum_{i=1}^{K}|\beta_k|$$

Ridge regression, on the other hand, adds a penalty term which is the sum of the squared values of the coefficients multiplied by a regularization parameter λ, to the loss function. Ridge regression does not perform feature selection, as it shrinks all coefficients towards zero but rarely sets them exactly to zero. The loss function of Ridge (L2-norm regulation) is:

$$L(\beta) = -\sum_{i=1}^{N}\{y_i \log P(y_i = 1|X_i;\beta) + (1 - y_i)\log(1 - P(y_i = 1|X_i;\beta)\} + \lambda\sum_{i=1}^{K}\beta_k^2$$

The strength of the penalty is controlled by the regularization parameter λ, which requires tuning to optimize the model performance. For more information, please refer to Appendix B.

Random Forest. Random Forest is an ensemble learning method used for classification and regression task, building on decision trees. A decision tree predicts outcomes through a tree-like

structure where each internal node represents a decision based on an input feature, each branch represents the outcome, and each leaf node represents a final prediction.

The construction of a decision tree begins with the selection of the most informative feature as the root node. The data is then split into subsets based on this feature, recursively repeating the process for each subset until reaching leaf nodes. For example, as shown in Figure 1, a decision tree might start with past adoption behavior as the root node, followed by cost-share program participation and no-till adoption at subsequent levels, generating predictions of conservation practice adoption likelihood.

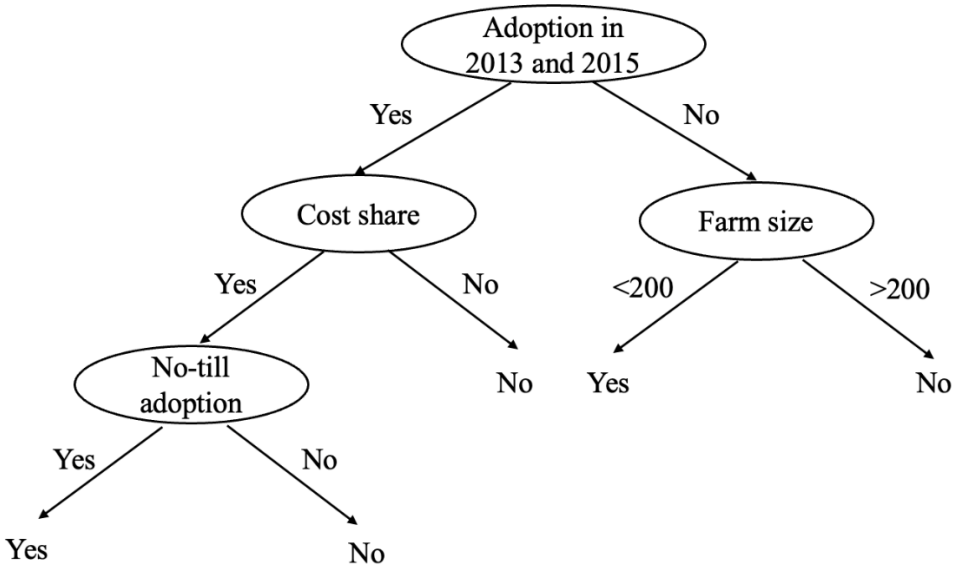

Figure 1. Decision tree structure, illustrating the hierarchical prediction process for cover crop adoption/disadoption, based on past adoption behavior, cost-share participation, no-till adoption, and farm size.

Individual decision trees can be prone to overfitting, especially when trees become too complex and are unstable as small changes in the data can significantly change tree structures. Random Forest, an ensemble learning method, addresses these limitations by leveraging multiple

decision trees to improve prediction accuracy and reduce overfitting. Each decision tree in the ensemble is trained on a different bootstrap sample of the data, and predictions are aggregated across all trees to make the final decision. This approach mitigates the instability and overfitting often associated with individual decision trees. For the detailed explanation of the Random Forest algorithm, please refer to Appendix A.

Metrics Used to Assess the Performance of the Model

We implement models using Python, applying hyperparameter tuning and validation to ensure robustness. Logistic Regression models with LASSO and Ridge regularization are optimized using 5-fold cross-validation, while the Random Forest model is enhanced with bootstrap resampling and out-of-bag (OOB) validation. Implementation and validation details are in Appendix B, with further methodological explanations in Breiman (2001) and Hastie et al. (2009).

Model performance is assessed on a separate testing dataset using key metrics. The first one is accuracy, which measures the proportion of correct predictions. Given the imbalanced dataset (115 adopters vs. 386 non-adopters), we also use the Area Under the Receiver Operating Characteristic Curve (AUC-ROC), which assesses model performance across all classification thresholds (Fawcett, 2006).

Data

This study primarily uses survey data from the Iowa Farm and Rural Life Poll (IFRLP), supplemented with county-level data on weather patterns and agricultural land characteristics.

Survey Data

This research used data from the IFRLP, an annual longitudinal panel survey of Iowa farmers. We analyze data from the 2014, 2016, and 2018 waves that tracked farmers' adoption of cover crops. The IFRLP is unique due to its long-running nature and its focus on Iowa farmers'

perspectives and behaviors related to various agricultural issues, including conservation practices and farm management. The IFRLP's consistent tracking of conservation practices provides measures of individual-level conservation practice adoption over time that are usually unavailable in other datasets, making it particularly valuable for our research.

The IFRLP surveys use a modified Dillman (2014) approach, distributing surveys in late January or early February. The method involves an initial mailing with a cover letter, followed by a reminder postcard after two weeks, and a second survey copy to non-respondents two weeks later. The survey was distributed to 2,218, 2,089, and 2,115 farmers in 2014, 2016, and 2018, with consistent response rates of about 50% (1,128, 1,039, and 1,061 responses, respectively) (Arbuckle 2015; Arbuckle 2016; Arbuckle 2018).

The IFRLP experiences panel attrition due to various factors common to panel survey research including retirement, death, lack of interest, and privacy concerns (Lynn 2018). To maintain similar numbers of responses over time, new random samples are added annually. Nevertheless, our sample differs from the overall farmer population in two aspects: due to the panel survey design, participants tend to be older than average, and due to the survey's focus on commercial, production agriculture, non-commercial farmers (e.g., hobby farmers) are less likely to continue participation in the Poll. For example, the average 2017 IFRLP farm size in acres was 432 acres, compared to the Iowa mean of 355 acres in the 2017 Census of Agriculture (USDA NASS, 2017).

While this might initially appear as a limitation, these characteristics may actually strengthen our analysis. USDA's Census of Agriculture definition of a farm (any place from which $1,000 or more of agricultural products were produced and sold, or normally would have been sold) creates a downward bias by including many small-scale farmers who have no or

limited production. More importantly, while farms over 500 acres represent only 22% of Iowa farms, they manage approximately 72% of the state's farmland (USDA NASS, 2017). Thus, our sample's bias toward larger operations allows us to offer perspectives more relevant to the majority of agricultural land where conservation practices could be implemented.

Our analysis focuses on 519 farmers who participated in all three survey waves. Robustness checks indicate that key variables in our merged dataset show consistency between the sub-sample and overall IFRLP sample (see Appendix C for details). We also integrate county-level weather and land quality data. Weather data from the National Oceanic and Atmospheric Administration (NOAA) is used to calculate county-level precipitation during the growing season (May to August). The land quality data is sourced from the United States Department of Agriculture Natural Resource Conservation Service (NRCS) SSURGO database. In this study, land quality is represented by the percentage of land classified as Land Capability Classes (LCC) I or II, which are the most suitable for crop production.

Selection of Explanatory Variables Based on Traditional Econometrics Approach and the ML Approach

This study uses a hybrid approach to variable selection, integrating traditional econometric methods with machine learning (ML) techniques. Traditional econometric approaches typically rely on theory-driven variable selection, prioritizing interpretability and causal inference, with variables chosen based on domain knowledge and prior research. In contrast, ML approaches are primarily data-driven, capable of handling a large number of potential predictors and identifying complex patterns and interactions that may not be apparent in theory alone. Techniques such as Recursive Feature Elimination (RFE) iteratively remove less important variables, allowing the model to focus on the most predictive variables. By utilizing the strengths of both

methodologies, we ensure that our variable selection is theoretically relevant and empirically robust when used to compare different models.

Our hybrid method offers two primary advantages. First, it allows us to compare models using a consistent set of variables, ensuring that differences in predictive performance are not due to input variation. This approach aligns with several recent studies (Jarmulska, 2022; Yoon, 2021). Second, it leverages the strengths of both approaches - using ML to refine variable selection while maintaining the theoretical relevance of the variables, which is crucial for making informed policy recommendations.

Our initial variable selection is guided by domain knowledge and theoretical relevance. We then conduct RFE using a Random Forest classifier to further refine the variable selection process. Figure A3 illustrates the feature importance as determined by this method using all available variables in our dataset, which helped identify potentially important predictors that were then evaluated against theoretical relevance to develop our final set of variables. The detailed RFE analysis in Appendix D reveals both overlaps and differences from our theory-driven choices. For example, RFE prioritized some variables not included in the final model, such as some specific crop related variables and farm sizes at different time points, which could introduce multicollinearity despite their predictive power. Conversely, we retain variables like crop insurance purchase, cost-share participation, and no-till adoption due to their theoretical relevance and our purpose of empirically investigating their potential significance in adoption decisions, even though they are ranked relatively low by the RFE.

Table 1 presents summary statistics for the explanatory variables selected through our hybrid approach. The dependent variable is the binary adoption decision in 2017, with a total sample size of 501 after filtering out missing values (386 non-adopters and 115 adopters). Past

adoption behavior, an important predictor constructed based on the adoption status in 2013 and 2015 is categorized as follows: non-adoption in both years (0,0), adoption in 2015 but not in 2013 (0,1), adoption in 2013 but not in 2015 (1,0), and adoption in both years (1,1), with non-adoption in both years as the reference group.

We include several policy-relevant variables: crop insurance purchase decision in 2015, cost-share program participation in 2013, and no-till adoption in 2017. The dummy variable for cost-share received four years ago allows us to assess the long-term effectiveness of the cost-share program and determine whether respondents discontinue cover crop use after the cost-share payment contract ends. Additionally, a binary indicator for no-till adoption in 2017 is included to evaluate the complementarity of conservation practices. Additional variables include farm operation characteristics (e.g., farm size, rent land share), land quality measures (e.g., Land Capability Classification share), weather effects (growing season precipitation), and farmers' perceptions of cover crops, which are discussed in detail in Appendix E.

Table 1. Descriptive statistics of the explanatory variables (N=501)

Variable	Mean	SD	Min	Max
Past adoption (0, 1)	0.11	0.31	0	1
Past adoption (1, 0)	0.09	0.28	0	1
Past adoption (1, 1)	0.09	0.29	0	1
Cost-share participation in 2013	0.07	0.26	0	1
Farm experience	38.93	12.45	3	87
Cover crop acres in 2015	23.98	92.23	0	1000
Total number of livestock	148.75	731.53	0	11360
LCC	55.87	22.1	11.4	93.4
Precipitation in 2017	392.34	83.03	167.5	632.69
Highly erodible land indicator	0.49	0.5	0	1
Purchase crop insurance in 2015	0.72	0.45	0	1
Lack of knowledge regarding cover crop	2.84	0.96	1	5
Cover crop learning interest	3.52	0.79	1	5
Cover crop reducing yield	2.72	0.72	1	5
USDA service center visit	2.27	3.99	0	40
Adoption of no-till in 2017	0.43	0.5	0	1

Percentage of farm income of net household income	3.46	1.4	1	5
Percentage of rented land	0.2	0.31	0	1
Farm Size	525.1	605.69	2	4000

Pairwise correlation analysis suggests that most correlations among the explanatory variables are relatively low, indicating that they capture distinct dimensions of the factors influencing adoption decisions. Higher correlations for traditional econometric models may require adjustments to avoid multicollinearity, though these correlations pose less of an issue in machine learning models like Random Forest. A detailed correlation matrix and analysis, along with a correlation chart (Figure A4), are provided in Appendix E.

Results

Comparison of Predictive Performance

Table 2 presents the performance metrics of four models: Logistic Regression, LASSO, Ridge, and Random Forest. In terms of prediction power, Random Forest outperforms all logistic regression models with an accuracy of 92.1% and an AUC-ROC of 0.952. This suggests that Random Forest is the most effective at making accurate predictions of farmers who adopt conservation practices and those who disadopt. The base Logistic Regression model shows a high accuracy of 85.1% but a moderate AUC-ROC score of 0.754. The LASSO maintains this accuracy while improving the AUC-ROC to 0.759, and Ridge achieves a slightly lower accuracy (82.2%) but a higher AUC-ROC of 0.774.

Table 2. Comparison of model prediction accuracy and AUC-ROC scores

Model	Accuracy	AUC-ROC
Logistic Regression	0.851	0.754
Logistic Regression (L1 - LASSO)	0.851	0.759
Logistic Regression (L2 - Ridge)	0.822	0.774
Random Forest	0.921	0.952

Marginal Effects of Variables Based on the Logistic Regression

Table 3 presents the results of marginal effects from the logistic regression with a 400-observation training subset. Model diagnostics indicate heteroskedasticity (Breusch-Pagan test, $p < 0.001$) and proper model specification (link test: linear term $p < 0.001$; squared term $p = 0.114$), and we estimate the model using heteroskedasticity-robust standard errors. Results indicate that past adoption behavior is a strong predictor of adoption in 2017. Farmers who adopted cover crops in both previous periods (1, 1) show a 26.5% higher likelihood of adoption in 2017 compared to those who never adopted. Moreover, farmers who adopted in the later period (0, 1) demonstrate a 32.6% increase in adoption probability, relative to non-adopters in both periods. While our model uses adoption in 2017 as the dependent variable (with adoption = 1 and non-adoption = 0), it also helps us understand the potential for past adopters to discontinue the practice, which we refer to as disadoption. Based on the full set of respondents, farmers who were non-adopters in 2013 but adopted in 2015 have a 30.9% possibility of disadopting by 2017, suggesting that recent adopters have a relatively high likelihood of continuing the practice. For farmers who adopted in 2013 but not in 2015, the chance of being a non-adopter (disadopting) in 2017 reaches 79.1%. Consistent adopters (adopted in 2013 and 2015) show the lowest disadoption rate, with only a 28.9% chance of becoming non-adopters (disadopting) in 2017.

The adoption of no-till in 2017 is positively associated with cover crop adoption, increasing the likelihood by 10.7%, indicating the potential complementarities between different conservation practices. A lack of knowledge regarding cover crops significantly decreases the adoption probability by 5.3%, and cover crops learning interest significantly increases the adoption probability by 4.4%. Cost-share participation in 2013 decreases the adoption probability by 3.6%, but it is not statistically significant.

Table 3. Marginal effect of variables – logistic regression

Variable	Marginal Effect	Standard Deviations
Past adoption (0, 1)	0.326***	(0.042)
Past adoption (1, 0)	0.052	(0.058)
Past adoption (1, 1)	0.265***	(0.049)
Cost-share participation in 2013	-0.036	(0.061)
Farm experience	0.001	(0.002)
Cover crop acres in 2015	0.0002	(0.000)
Number of livestock	0.000006	(0.000)
LCC	-0.0004	(0.001)
Precipitation in 2017	0.00007	(0.000)
Highly erodible land indicator	0.021	(0.041)
Purchase crop insurance in 2015	-0.024	(0.049)
Lack of knowledge regarding cover crop	-0.053***	(0.019)
Cover crop learning interest	0.044*	(0.026)
Cover crop reducing yield	0.014	(0.028)
USDA service center visit	-0.002	(0.006)
Adoption of no-till in 2017	0.107***	(0.037)
Farm income share	-0.009	(0.017)
Rent land share	0.007	(0.069)
Farm size	0.00001	(0.000)
Number of observations:	400	
Pseudo R-square:	0.333	

Marginal Effects of Variables Based on ML Models

Given the non-linear and interactive nature of Random Forests, calculating marginal effects using traditional methods is not feasible. To address this interpretation challenge and enable comparison with logistic regression models, we use Accumulated Local Effects (ALE) plots to visualize the marginal effects of variables (Apley & Zhu, 2020). ALE plots illustrate how changes in a specific explanatory variable affect model predictions by focusing on small intervals of the independent variable and accumulating effects across these intervals.

Figure 2 shows the ALE plots derived from our Random Forest model. A positive slope indicates increasing prediction probability with higher variable values, while the magnitude of the slope shows the strength of impact. For binary variables (past adoption behavior, cost-share

participation, crop insurance, highly erodible land, and no-till adoption), the x-axis represents the binary state (0 = absence, 1 = presence), with the y-axis showing the accumulated impact on predicted adoption probability. For past adoption behavior, we implemented customized ALE calculations where the y-axis shows the difference in predicted probability compared to the reference group of never-adopters (0,0), which has a baseline probability of 12.0%. The steeper slopes for Past adoption (0,1) and (1,1), align with the logistic regression results, showing stronger effects on continued adoption.

When comparing logistic regression results to ALE plots, several discrepancies can be observed in both the signs and magnitudes of variable effects, mainly due to ML models' ability to capture non-linear relationships and their robustness to multicollinearity. For example, precipitation shows a U-shaped effect in the ALE plot, indicating that both extreme rainfall conditions can increase the likelihood of adopting cover crops, which reflect cover crops' ability to manage both moisture extremes (Kaye & Quemada, 2017). Our results align with Park et al. (2023), who found a similar pattern using growing season precipitation and its squared term, though our Random Forest model reveals this pattern through a more flexible functional form rather than an explicit quadratic specification. Farmers' concerns about yield reduction show a non-linear relationship, where moderate levels reduce adoption likelihood while higher concerns might actually increase adoption likelihood, possibly reflecting increasing awareness and engagement with the practice. The percentage of household income from farming also shows non-linear effects, with higher percentage generally decreases adoption likelihood.

Potential multicollinearity can play a crucial role. While logistic regression models can produce unstable and unreliable coefficient estimates when variables are highly correlated, Random Forests mitigate this issue through random feature selection at each split, leading to

more robust results that better reflect real-world relationships. This explains why some variables like cost-share participation and crop insurance purchase show opposite effects - negative in logistic regression but positive in ALE plots. Similarly, USDA service center visits reveal more nuanced relationships in the ALE plots that aren't captured by logistic regression results.

These discrepancies demonstrate the value of combining traditional econometric methods with ML approaches to better understand the dynamics of adoption and disadoption decisions and inform policy implications. ML models can capture detailed patterns that might be obscured in traditional logistic regression analyses.

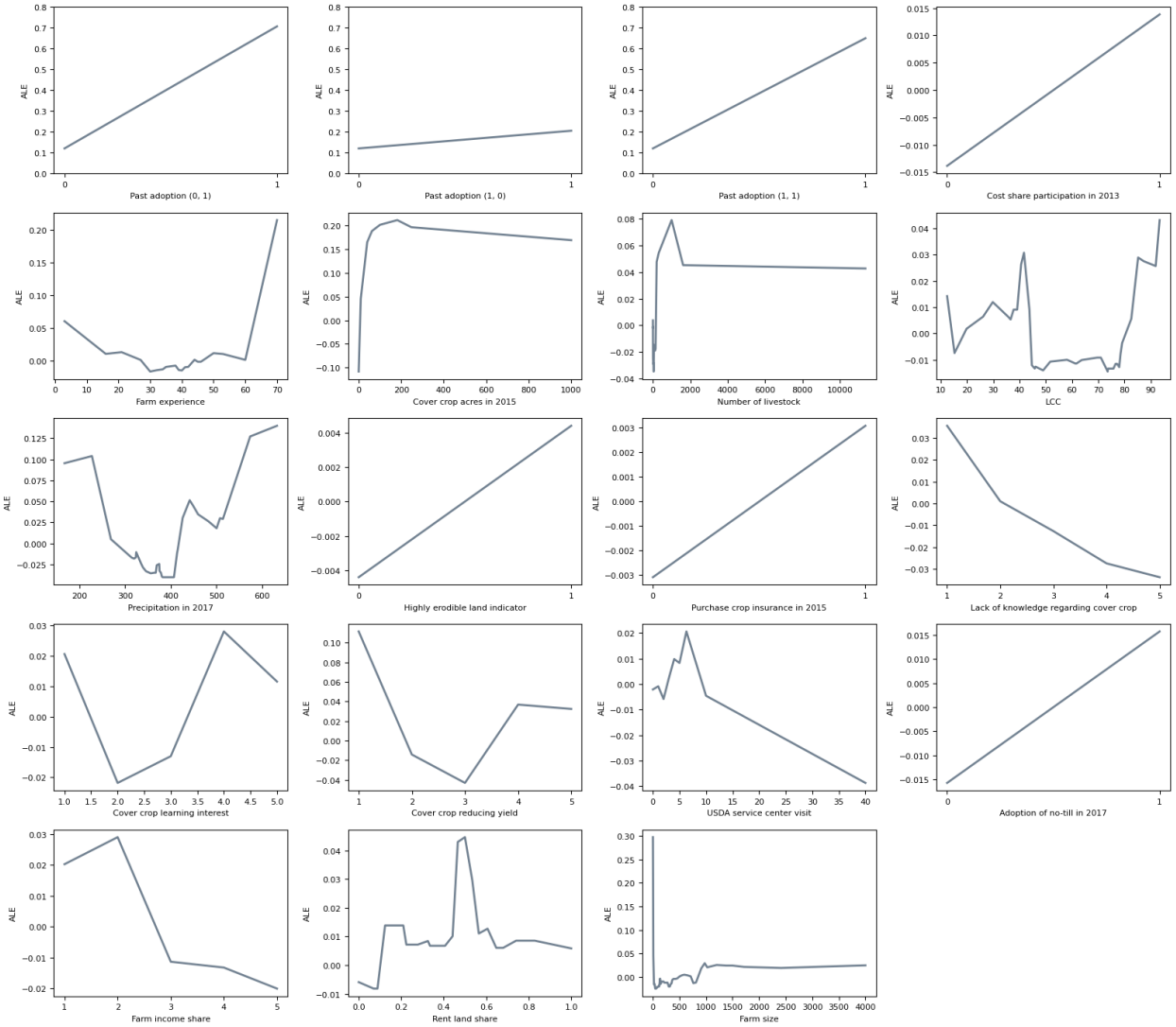

Figure 2. Accumulated local effects (ALE) plot from Random Forest model, showing variable impacts on the likelihood of conservation practices adoption.

Drivers of Disadoption

To investigate the factors driving disadoption, we examine SHAP (SHapley Additive exPlanations) values from our Random Forest model (Lundberg & Lee, 2017), focusing on features that influence farmers' decisions to discontinue cover crop use (the non-adoption class). Figure 3 presents SHAP values from our final model using selected theoretical and empirically relevant variables, which differs from Figure A3 in the Appendix that shows feature importance rankings from an initial Random Forest model using all available variables. Features in Figure 3 are ordered by overall importance, with those at the top having the greatest impact on predictions. Each individual dot represents an observation. The x-axis shows how much each feature shifts the predicted probability of disadoption from the average prediction, with positive values indicating increased likelihood. The visualization uses a color gradient from blue (low values) through purple to red (high values) to represent feature values. Horizontal alignment of dots indicates consistent impacts and vertical spreading showing varying impacts of similar feature values.

Cover crop acres in 2015 shows the strongest impact, with larger acres (red dots) generally associated with lower likelihood of disadoption, suggesting that farmers with more extensive adoption are less likely to disadopt. Consistent with our earlier findings, past adoption behavior is another crucial predictor, with farmers who adopted in 2015 or both 2013 and 2015 showing a strong negative association with disadoption. Environmental factors, particularly precipitation in 2017, show similar patterns to our previous results, with both high and low precipitation encourage continued adoption.

Results also reveals that larger farm sizes tend toward slight disadoption, and higher levels of cover crop learning desire show a slight negative association with disadoption. Policy-related variables including cost-share participation and crop insurance purchase show relatively modest impacts on disadoption decisions, appearing at the bottom of the feature ranking. It suggests that while these policy instruments may influence initial adoption, their long-term effect on continued use may be limited.

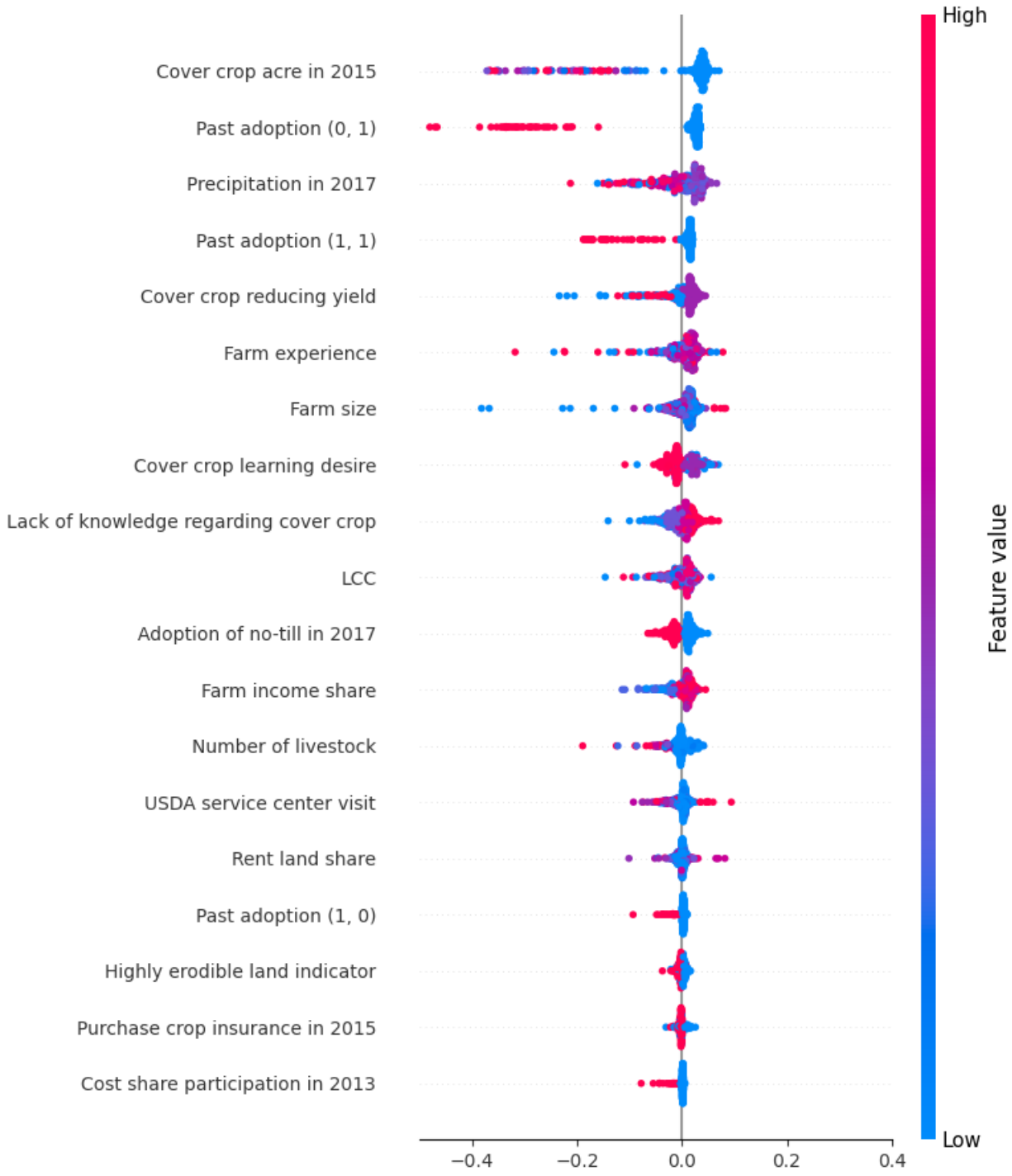

Figure 3. SHAP values from the Random Forest model, showing feature importance and impact on cover crop disadoption. Red and blue dots represent high and low feature values.

Robustness Checks

We conduct robustness checks by examining the models' robustness to outliers, using Z-score method. The Z-score method identifies outliers by calculating the standard deviation from the mean for each data point, with points beyond 3 standard deviations considered outliers. After removing the detected outliers, we re-evaluate the models using accuracy and AUC-ROC.

Table 4 shows the differences in model performance after outlier removal (calculated as post-removal performance minus pre-removal performance). The Random Forest model shows greater stability with the smallest change in accuracy and is the only model to show an improvement in AUC-ROC. In contrast, the logistic regression models show larger changes in accuracy and substantial decreases in AUC-ROC, indicating less stability and lower robustness to outliers. Random Forest's predictions are less affected by extreme values in the dataset, potentially due to its ensemble nature and use of multiple decision trees.

Table 4. Model performance change after outlier removal

	Accuracy Difference between before and after outlier removal	**AUC-ROC Difference between before and after outlier removal**
Logistic Regression	0.046	-0.067
Logistic Regression (L1 - LASSO)	0.046	-0.102
Logistic Regression (L2 - Ridge)	0.075	-0.058
Random Forest	0.035	0.016

Conclusion

This study explores the complementarity between traditional econometric methods and machine learning in analyzing farmers' adoption and disadoption of conservation practices, specifically cover crops. Using a unique longitudinal panel of Iowa farmers, we find that Random Forest achieves higher predictive accuracy (92.1%) compared to logistic regression (85.1%), shows greater robustness to outliers and captures complex non-linear relationships.

Prediction power is crucial for identifying outcomes of interest, with significant implications for policy design. Prediction-based analyses have improved decision-making and program effectiveness across various fields, such as teacher selection in education (Chalfin et al., 2016), bail decisions in criminal justice (Kleinberg et al., 2018), and poverty targeting (Engstrom et al., 2022). In the context of conservation programs, improved predictive accuracy may facilitate more effective program implementation. First, better prediction of potential disadopters could allow for early intervention strategies before disadoption occurs, reducing both direct financial costs and indirect environmental costs. Second, conservation agencies could potentially identify farmers most likely to maintain long-term adoption, optimizing program funding allocation by targeting resources (e.g., cost-share programs). Third, improved predictions might also help conservation professionals better focus their outreach and technical assistance efforts on farmers requiring additional support for continued adoption.

The economic impact of this improved accuracy could be substantial. In our sample alone, the 7% improvement in accuracy represents better prediction for approximately 35 farmers. At a cost-share rate of $61.5/acre and given an average farm size of 432 acres (USDA NRCS, 2024), the total amount of conservation funding at stake would be sizeable. For example, if each farm enrolled a quarter of its acreage in a cost-share program, the funding for 35 farmers would amount to $230,000. When scaled to the broader population and considering the current substantial county-level disadoption rate (Plastina et al., 2024), improved prediction accuracy could enhance the potential efficiency of millions in conservation funding while reducing administrative burden and improving environmental outcomes through continued adoption.

Our findings reveal several important factors influencing conservation practice persistence. Past adoption behavior is a strong predictor, particularly for recent adopters. The

adoption of no-till is positively associated with cover crop adoption, suggesting complementarity between these conservation practices. The SHAP analysis further demonstrated that farmers with larger previous cover crop acreage are less likely to disadopt. The Random Forest reveals complex non-linear relationships, such as the U-shaped effect of precipitation and the detailed impact of farmers' yield concerns. While cost-share participation and crop insurance show some impacts on adoption decisions, the long-term influence on continued use appears limited.

It is important to note that these findings should be interpreted with caution. Our sample primarily represents larger operations and farmers who consistently participated in surveys, those who may be more engaged in agricultural conservation practices or care more about environmental issues. Our sample may underestimate true disadoption rates as farmers who discontinue conservation practices may also be more likely to drop out of the survey.

While machine learning models offer advantages in identifying complex relationships and achieving superior predictive accuracy, their "black box" nature limits understanding the underlying mechanisms. This poses challenges in agricultural economics, where causality is crucial for policy design. Traditional econometric methods, grounded in economic theory, provide clear insights into how specific factors impact decisions. Our analysis emphasizes the potential of complementing these approaches, leveraging machine learning's predictive power while maintaining the interpretability of econometric models. Such integration represents a promising direction for achieving more comprehensive insights into farmer decision-making and improving policy effectiveness.

Declaration of Funding

This work was supported by the Iowa Agriculture and Home Economics Experiment Station at Iowa State University; and the Iowa Nutrient Research Center under Grant [AWD-024728].

Data Availability Statement

The survey data used in this study was collected with approval of the Institutional Review Board at Iowa State University for the protection of human subjects in research. Summary reports and investigator contact information can be found at https://ext.soc.iastate.edu/programs/iowa-farm-and-rural-life-poll/. To protect respondents' privacy, the survey data is not publicly available. However, selected data may be made available upon request to the authors, subject to human subjects protections.

The land classification data was obtained from the Natural Resources Conservation Service (NRCS). This dataset contains soil information and was used to construct the Land Capability Classification (LCC) variable. This dataset is not publicly available. Researchers interested in replicating the study can contact nri@wdc.usda.gov to request access to this dataset.

The weather data was obtained from the National Oceanic and Atmospheric Administration (NOAA). This dataset contains daily weather information and was used to construct variables for precipitation, growing degree days (GDD), and stressful degree days (HDD).

Disclosure of Interest

No potential competing interest was reported by the author(s).

References

Apley, D. W., & Zhu, J. (2020). Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(4), 1059-1086.

Arbuckle, J. G. (2015). Iowa Farm and Rural Life Poll: 2014 Summary Report. Ames, IA: Iowa State University Extension and Outreach.

Arbuckle, J. G., & Roesch-McNally, G. (2015). Cover crop adoption in Iowa: The role of perceived practice characteristics. *Journal of Soil and Water Conservation*, 70(6), 418-429.

Arbuckle, J. G. (2016). Iowa Farm and Rural Life Poll: 2016 Summary Report. Ames, IA: Iowa State University Extension and Outreach.

Arbuckle, J. G. (2018). Iowa Farm and Rural Life Poll: 2018 Summary Report. Ames, IA: Iowa State University Extension and Outreach.

Barrett, C. B., Islam, A., Mohammad Malek, A., Pakrashi, D., & Ruthbah, U. (2022). Experimental evidence on adoption and impact of the system of rice intensification. *American Journal of Agricultural Economics*, 104(1), 4-32.

Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5-32.

Chalfin, A., Danieli, O., Hillis, A., Jelveh, Z., Luca, M., Ludwig, J., & Mullainathan, S. (2016). Productivity and selection of human capital with machine learning. *American Economic Review*, 106(5), 124-127.

Claassen, R., Bowman, M., McFadden, J., Smith, D., & Wallander, S. (2019). Conservation-practice adoption rates vary widely by crop and region. EIB-197. U.S. Department of Agriculture, Economic Research Service. Retrieved January 23, 2025, https://www.ers.usda.gov/publications/pub-details/?pubid=44030

Connor, L., Rejesus, R. M., & Yasar, M. (2022). Crop insurance participation and cover crop use: Evidence from Indiana county-level data. *Applied Economic Perspectives and Policy*, 44(4), 2181-2208.

Dayer, A. A., Lutter, S. H., Sesser, K. A., Hickey, C. M., & Gardali, T. (2018). Private landowner conservation behavior following participation in voluntary incentive programs: Recommendations to facilitate behavioral persistence. *Conservation Letters*, 11(2), e12394.

DeLay, N. (2019). The impact of federal crop insurance on the conservation reserve program. *Agricultural and Resource Economics Review*, 48(2), 297-327.

Dillman, D. A., Smyth, J. D., & Christian, L. M. (2014). *Internet, phone, mail, and mixed-mode surveys: The tailored design method*. John Wiley & Sons.

Dunn, M., Ulrich-Schad, J. D., Prokopy, L. S., Myers, R. L., Watts, C. R., & Scanlon, K. (2016). Perceptions and use of cover crops among early adopters: Findings from a national survey. *Journal of Soil and Water Conservation*, 71(1), 29-40.

Engstrom, R., Hersh, J., & Newhouse, D. (2022). Poverty from space: Using high resolution satellite imagery for estimating economic well-being. *The World Bank Economic Review*, 36(2), 382-412.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861-874.

Gong, S., Bergtold, J. S., & Yeager, E. (2021). Assessing the joint adoption and complementarity between in-field conservation practices of Kansas farmers. *Agricultural and Food Economics*, 9(1), 30.

Harding, M. C., & Lamarche, C. (2021). Small steps with big data: using machine learning in energy and environmental economics. *Annual Review of Resource Economics*, 13, 469-488.

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The Elements of Statistical Learning: Data mining, Inference, and Prediction* (Vol. 2, pp. 1-758). New York: springer.

Jarmulska, B. (2022). Random forest versus logit models: Which offers better early warning of fiscal stress? *Journal of Forecasting*, 41(3), 455-490.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2018). Human decisions and machine predictions. *The Quarterly Journal of Economics*, 133(1), 237-293.

Kaye, J. P., & Quemada, M. (2017). Using cover crops to mitigate and adapt to climate change. A review. *Agronomy for Sustainable Development*, 37, 1-17.

Khanna, M. (2001). Sequential adoption of site-specific technologies and its implications for nitrogen productivity: A double selectivity model. *American Journal of Agricultural Economics*, *83*(1), 35-51.

Lee, S., & McCann, L. (2019). Adoption of cover crops by US soybean producers. *Journal of Agricultural and Applied Economics*, 51(4), 527-544.

Lundberg, S., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *arXiv.* Retrieved January 23, 2025, https://doi.org/10.48550/arXiv.1705.07874

Lynn, P. (2018). Tackling panel attrition. The Palgrave handbook of survey research, 143-153.

McFadden, D. (1972). Conditional logit analysis of qualitative choice behavior.

Mitchell, J. P., Shrestha, A., Mathesius, K., Scow, K. M., Southard, R. J., Haney, R. L., ... & Horwath, W. R. (2017). Cover cropping and no-tillage improve soil health in an arid irrigated cropping system in California's San Joaquin Valley, USA. *Soil and Tillage Research*, *165*, 325-335.

Molnar, C. (2019). Interpretable machine learning: *A Guide for Making Black Box Models Explainable*. Retrieved January 23, 2025, https://christophm.github.io/interpretable-ml-book/

Moser, C. M., & Barrett, C. B. (2006). The complex dynamics of smallholder technology adoption: the case of SRI in Madagascar. *Agricultural Economics*, 35(3), 373-388.

Mullainathan, S., & Spiess, J. (2017). Machine learning: an applied econometric approach. *Journal of Economic Perspectives*, 31(2), 87-106.

Neill, S. P., & Lee, D. R. (2001). Explaining the adoption and disadoption of sustainable agriculture: The case of cover crops in northern Honduras. *Economic Development and Cultural Change*, 49(4), 793-820.

Oliva, P., Jack, B. K., Bell, S., Mettetal, E., & Severen, C. (2020). Technology adoption under uncertainty: Take-up and subsequent investment in Zambia. *Review of Economics and Statistics*, *102*(3), 617-632.

Pannell, D. J., & Claassen, R. (2020). The roles of adoption and behavior change in agricultural policy. *Applied Economic Perspectives and Policy*, 42(1), 31-41.

Pannell, D. J., Marshall, G. R., Barr, N., Curtis, A., Vanclay, F., & Wilkinson, R. (2006). Understanding and promoting adoption of conservation practices by rural landholders. *Australian Journal of Experimental Agriculture*, 46(11), 1407-1424.

Park, B., Rejesus, R. M., Aglasan, S., Che, Y., Hagen, S. C., & Salas, W. (2023). Payments from agricultural conservation programs and cover crop adoption. *Applied Economic Perspectives and Policy*, 45(2), 984-1007.

Plastina, A., Sawadgo, W., & Okonkwo, E. (2024). Pervasive disadoption substantially offsets new adoption of cover crops and no-till. *Choices: The Magazine of Food, Farm, and Resource Issues*, *39*(02).

Prest, B. C., Wichman, C. J., & Palmer, K. (2023). RCTs against the machine: Can machine learning prediction methods recover experimental treatment effects?. *Journal of the Association of Environmental and Resource Economists*, 10(5), 1231-1264.

Prokopy, L. S., Floress, K., Klotthor-Weinkauf, D., & Baumgart-Getz, A. (2008). Determinants of agricultural best management practice adoption: Evidence from the literature. *Journal of Soil and Water Conservation*, 63(5), 300-311.

Ramsey, S. M., & Bergtold, J. S. (2021). Examining inferences from neural network estimators of binary choice processes: Marginal effects, and willingness-to-pay. *Computational Economics*, 58(4), 1137-1165.

Razafimahatratra, H. M., Bignebat, C., David-Benz, H., Bélières, J. F., & Penot, E. (2021). Tryout and (Dis) adoption of conservation agriculture. Evidence from Western Madagascar. *Land Use Policy*, 100, 104929.

Shao, Y., Xiong, T., Li, M., Hayes, D., Zhang, W., & Xie, W. (2021). China's missing pigs: Correcting china's hog inventory data using a machine learning approach. *American Journal of Agricultural Economics*, 103(3), 1082-1098.

Storm, H., Baylis, K., & Heckelei, T. (2020). Machine learning in agricultural and applied economics. *European Review of Agricultural Economics*, 47(3), 849-892.

Train, K. E. (2009). *Discrete choice methods with simulation*. Cambridge university press.

United States Department of Agriculture, National Agricultural Statistics Service. (2017). *2017 Census of Agriculture: Iowa, Volume 1, Chapter 1, State Level Data* (p. 71). Retrieved

January 23, 2025, https://www.nass.usda.gov/Publications/AgCensus/2017/Full_Report/Volume_1,_Chapter_1_State_Level/Iowa/st19_1_0071_0071.pdf

United States Department of Agriculture, Natural Resources Conservation Service. (2024). Iowa EQIP payment rates: Fiscal year 2024. Retrieved January 23, 2025, https://www.nrcs.usda.gov/sites/default/files/2023-12/fy24-iowa-eqip.pdf

Wade, T., & Claassen, R. (2017). Modeling no-till adoption by corn and soybean producers: Insights into sustained adoption. *Journal of Agricultural and Applied Economics*, *49*(2), 186-210.

Wallander, S., Bowman, M., Beeson, P., & Claassen, R. (2017). Farmers and habits: the challenge of identifying the sources of persistence in tillage decisions.

Wang, T., Zhao, S., Zhu, G., & Zheng, H. (2021). A machine learning-based early warning system for systemic banking crises. *Applied Economics*, *53*(26), 2974-2992.

Won, S., Rejesus, R. M., Goodwin, B. K., & Aglasan, S. (2024). Understanding the effect of cover crop use on prevented planting losses. *American Journal of Agricultural Economics,* 106(2), 659-683.

Wooldridge, J. M. (2010). Econometric analysis of cross section and panel data. MIT press.

Wulanningtyas, H. S., Gong, Y., Li, P., Sakagami, N., Nishiwaki, J., & Komatsuzaki, M. (2021). A cover crop and no-tillage system for enhancing soil health by increasing soil organic matter in soybean cultivation. *Soil and Tillage Research*, 205, 104749.

Yoon, J. (2021). Forecasting of real GDP growth using machine learning models: Gradient boosting and random forest approach. *Computational Economics*, *57*(1), 247-265.

Appendix

Appendix A. Random Forest Algorithm

Random Forests train each decision tree on a different subset of the data using bootstrap sampling (randomly selecting samples with replacement) and combines the predictions from different trees, as shown in Figure A1. This process, known as bagging (bootstrap aggregating), can reduce overfitting and improve accuracy. Overfitting is reduced because each tree is trained on a slightly different dataset, along with the random selection of features at each node, thereby smoothing out the noise and capturing more robust patterns in the data. The basic steps for Random Forest are as follows (Hastie et al., 2009):

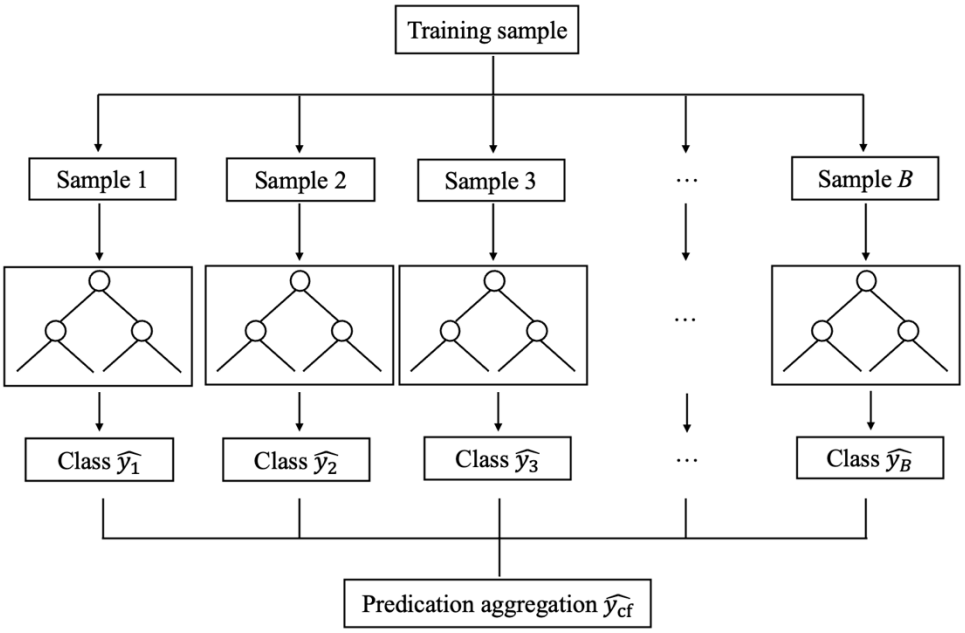

Figure A1. Random Forest algorithm, illustration showing how multiple decision trees are trained on bootstrap samples and combined through majority voting to make final predictions.

1. Bootstrap sampling: From original training set $\{X_1, X_2, \dots, X_N\}^T$ with labels $\{y_1, y_2, \dots, y_N\}$, generate a bootstrapped sample set, B of size N, where N is the total number of observations (farmers). In our setting, X_i is a vector of observed variables for

individual $i; = \{0, 1\}$, y_i is the binary choice where $y_i = 1$ represents adoption and $y_i = 0$ represents non-adoption. Each bootstrap sample B_b (Sample 1, Sample 2, … Sample B in Figure A1) is created by random selection with replacement.

2. Tree building: For each bootstrap sample B_b, construct a decision tree T_B using a subset of features selected randomly at each node. For each node, randomly select a subset of features d form the total features D. Choose the best feature and threshold to split the node based on a criterion called entropy which is calculated as follows:

$$Entropy(D) = -\sum_{i=1}^{c} q_i \log q_i$$

where the term q_i represents the proportion of instances in the dataset that belong to class i. In our binary case of adoption prediction, there are two classes: non-adoption (0) and adoption (1) in 2017. The feature and threshold that minimize the entropy are chosen for splitting the node, effectively creating two new subsets of data. This process continues recursively for each new node until a stopping criterion is met (such as a maximum tree depth or minimum number of samples in a node).

3. Prediction aggregation: To predict whether a farmer will adopt the practice (1) or not (0) in 2017, we aggregate the predictions from all trees in the forest (B trees). Let $\widehat{y_b}$ be the class prediction of the b-th tree. The final class prediction of the Random Forest $\widehat{y_{RF}}$ is determined by taking the majority vote of the predictions from all B trees. This means the outcome predicted by more than half of the trees becomes the final prediction. This can be mathematically represented as:

$$\widehat{y_{RF}} = majority\ vote\{\widehat{y_b}(x)\}$$

Appendix B. Implementation of Machine Learning Models

We begin our analysis by implementing Logistic Regression models with LASSO (L1 regularization) and Ridge (L2 regularization). To find the best settings (hyperparameters) for these models, we use "GridSearchCV" function in Python, which systematically evaluates different combinations of hyperparameters. As shown in Figure A2, we use a 5-fold cross-validation. This process involves first dividing our entire dataset into training and test sets. The training data is then further split into five folds. During each iteration, four folds are used for training the model, while the remaining fold serves as a validation set. This process is repeated five times, with each fold taking a turn as the validation set. After determining the optimal hyperparameters through cross-validation, we evaluate the model's final performance on the test set. This evaluation provides an estimate of the model's effectiveness on unseen data.

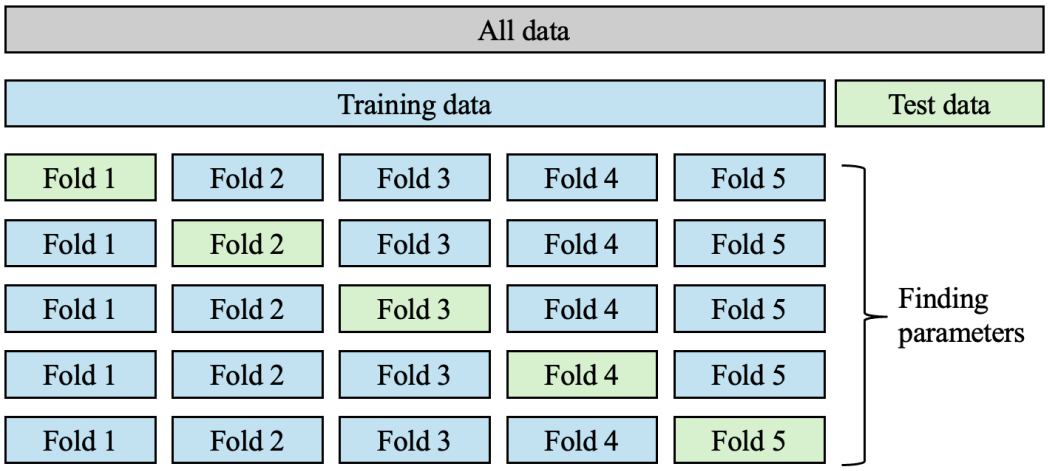

Figure A2. Five-fold cross-validation process, where training data is split into five parts for model validation, with four folds used for training and one for validation in each iteration.

For the Logistic Regression models, we explore different optimization algorithms and different penalty strengths (λ values) to identify the best model parameters. We conduct a

comprehensive search by testing different optimization algorithms available in our software. The full range of settings we test are detailed in Table A1.

Table A1. Hyperparameter test set for machine learning models

Machine learning model	Hyperparameter	Hyperparameter test set
Logistic regression – L1 regulation	Optimization algorithms	["liblinear", "saga"]
	Penalty strengths (λ values)	[0.01, 0.1, 1, 10, 100]
Logistic regression – L2 regulation	Optimization algorithms	["newton-cg", "lbfgs", "sag", "saga"]
	Penalty strengths (λ values)	[0.01, 0.1, 1, 10, 100]
Random Forest	Number of trees	[50, 100, 200]
	Maximum depth of the tree	[0, 5, 10]
	Minimum samples required to split a node	[2, 5, 10]
	Minimum samples required at a leaf node	[1, 2, 4]

For the Random Forest model, we explore various settings of hyperparameters, including number of trees (how many decision trees are created; For example, if set to 100, the model will create 100 different trees); maximum depth of the tree (the maximum number of levels in each decision tree which represent how complex each tree could become; For instance, a max depth of 5 means that from the root node to end node, there are at most 5 levels); minimum samples required to split a node (the minimum number of samples required to split an internal node; For example, if set to 10, a node must have at least 10 samples before it can be split further); minimum samples required at a leaf node (the minimum number of samples required to be at an end node; For instance, if set to 5, every leaf node must contain at least 5 samples).

We implement a robust approach combining bootstrap resampling, grid search for hyperparameter tuning, and out-of-bag (OOB) validation for Random Forest model. This process involved 50 iterations of bootstrap resampling, where each iteration created a subset of the training data by sampling with replacement. In each iteration, a bootstrap sample is created,

containing 80% of the original training data. Grid search with 5-fold cross-validation is performed on this bootstrap sample to find the optimal hyperparameters for the Random Forest model. The best-performing model from the grid search (based on mean cross-validation score) is then evaluated on the OOB samples - the 20% of data not included in the bootstrap sample. After completing all 50 iterations, we select as our final model the one that performed best on its respective OOB samples.

This approach offers several advantages: It thoroughly explores different subsets of the data through bootstrap sampling. It performs extensive hyperparameter tuning for each bootstrap sample. By repeating this process multiple times, it increases the likelihood of finding a globally optimal model. The combined approach with bootstrap sampling and OOB validation, creates a robust process that is particularly effective for ensemble models like Random Forest.

For logistic regression models, we only use 5-fold cross-validation. Each fold of the data is used for both training and validation, which is particularly suitable for linear models like logistic regression. We believe cross-validation is sufficient for Logistic Regression models for several reasons. Logistic Regression is a simpler, parametric model compared to Random Forest, with fewer hyperparameters to tune. Its performance is often more stable across different sub dataset. The main hyperparameters for Logistic Regression (optimization algorithms and penalty strengths C values) can be effectively tuned using GridSearchCV with cross-validation.

Appendix C. Sample Robustness Checks – Data Representativeness

This section presents robustness checks performed to assess the representativeness of the merged dataset compiled from the Iowa Farm and Rural Life Poll (IFRLP) surveys conducted in 2014, 2016, and 2018. The merged dataset is compared with the unselected respondents from each respective survey year to identify potential differences in key farm characteristics.

For categorical variables, such as gender and gross farm sales categories, Chi-squared tests are used to determine if there are significant differences in the distribution of these variables between the merged dataset and the omitted respondents. For continuous variables, such as farm size, owned land acreage, and rented land acreage, t-tests are used to compare the means between the merged dataset and the omitted respondents. In addition to the t-tests, standardized mean differences (SMD) are calculated for the continuous variables. The SMD is a measure of effect size that quantifies the magnitude of the difference between two means in terms of standard deviation units.

Table A2. Omitted sample check – merged data v.s. omitted Poll 2014

Variable	Merged data	Omitted Poll 2014	Chi-Squared Test P-value	T-test P-value	SMD P-value
Categorical					
Gender	1.035	1.104	0.000***		
Gross farm sales for 2013	6.556	6.004	0.002***		
Continuous					
Age in 2013	63.872	65.723		0.004**	0.865
Number of days worked off-farm in 2013	73.667	68.975		0.444	0.964
Acre of cover crop planted in 2013	18.707	11.926		0.149	0.929
Acres received cost-share in 2013	6.724	2.627		0.008**	0.868
Largest acreage ever planted	20.484	15.221		0.267	0.946
Cover crop acre percentage (Acre of cover crop planted in 2013/Farm size)	4.611	3.618		0.349	0.955
Cost share percentage (Acres received cost-share in 2013/ Acre of cover crop planted)	6.211	3.535		0.034	0.897
Acre of owned land in 2013	278.721	218.462		0.005**	0.864
Acre of rented land in 2013	192.127	165.116		0.185	0.936
Farm size in 2013	470.848	383.578		0.006**	0.869

Tables A2, A3, and A4 compare the merged dataset with the omitted respondents from the 2014, 2016, and 2018 IFRLP surveys, respectively. Chi-squared tests reveal significant differences in gender distribution and gross farm sales categories across all three years, with the merged dataset having a higher proportion of male respondents and higher average gross farm sales. T-tests indicate that farms in the merged dataset have on average, larger acreages of owned land, total farm size, and acres received cost-share for cover crops in 2013 (Table A2), as well as larger average farm sizes and owned land acreages in 2015 (Table A3) and 2017 (Table A4). However, the SMD p-values for these variables are not statistically significant, suggesting that the magnitude of the differences is not substantial. These findings imply that the merged dataset remains largely representative of the overall IFRLP sample. The general representativeness of the merged dataset, as demonstrated by the robustness checks, lends support for the validity of the findings presented in the main text of the paper.

Table A3. Omitted sample check – merged data v.s. omitted Poll 2016

Variable	**Merged data**	**Omitted Poll 2016**	**Chi-Squared Test P-value**	**T-Test P-value**	**SMD P-value**
Categorical					
Gender	1.041	1.105	0.000***		
Gross farm sales for 2015	5.691	5.085	0.001***		
Continuous					
Age in 2015	65.887	64.409		0.027*	0.88
Total number of livestock	374.212	300.967		0.484	0.965
Acre of owned land in 2015	326.212	207.139		0.000***	0.750
Acre of rented land in 2015	185.532	162.520		0.317	0.950
Farm size in 2015	511.744	369.659		0.000***	0.804

Table A4. Omitted sample check – merged data v.s. omitted Poll 2018

Variable	Merged data	Omitted Poll 2018	Chi-Squared Test P-value	T-Test P-value	SMD P-value
Categorical					
Gender	1.039	1.102	0.000***		
Gross farm sales for 2017	5.617	5.176	0.098		
Age in 2017	5.813	5.475	0.000***		
Raise livestock in 2017 or not	0.277	0.269	0.821		
Percentage of net household income from the farm	3.419	3.101	0.010*		
Rented from others or not in 2017	0.493	0.429	0.052		
Acres rented from others in 2017	1.335	1.201	0.093		
Number of landlords in 2017	1.335	1.437	0.134		
Number of years rented from the landlords	2.398	2.167	0.069		
Continuous					
Acre of owned land in 2017	178.451	140.994		0.000***	0.822
Acre of rented land in 2017	68.365	63.678		0.560	0.971
Farm size/Total acre in 2017	246.816	204.673		0.003**	0.857

Appendix D. Recursive Feature Elimination (RFE) Analysis

The Recursive Feature Elimination (RFE) is conducted using a Random Forest classifier to identify the most predictive variables in our dataset. The process involved iteratively removing less important features to focus the model on the variables that contributed most to predictive accuracy.

Figure A3 illustrates the feature importance as determined by the RFE method using a Random Forest classifier with all available variables in our dataset. This initial variable selection process helped identify potentially important predictors, which were then evaluated against theoretical relevance to develop our final set of variables for analysis. Comparing these ML-selected features with our theory-driven choices (shown in Table 1) revealed both overlaps and

differences. Significant overlap is observed in variables related to past adoption behavior, land characteristics, farm characteristics, and environmental factors. The RFE method tended to prioritize some variables not included in our final model, such as detailed crop-specific variables (corn and soybean acreage share) and multiple time points for the same variable (e.g., farm size in different years and Growing Degree Days or Stress Degree Days in different years). While these variables may enhance predictive accuracy, they could potentially introduce multicollinearity issues. Conversely, our theory-driven approach included some variables ranked lower by RFE, such as crop insurance purchase, cost-share participation, no-till adoption, and farmers' perceptions of cover crops. We retain these variables due to their theoretical importance and potential policy relevance, despite their lower predictive power as identified by the ML approach.

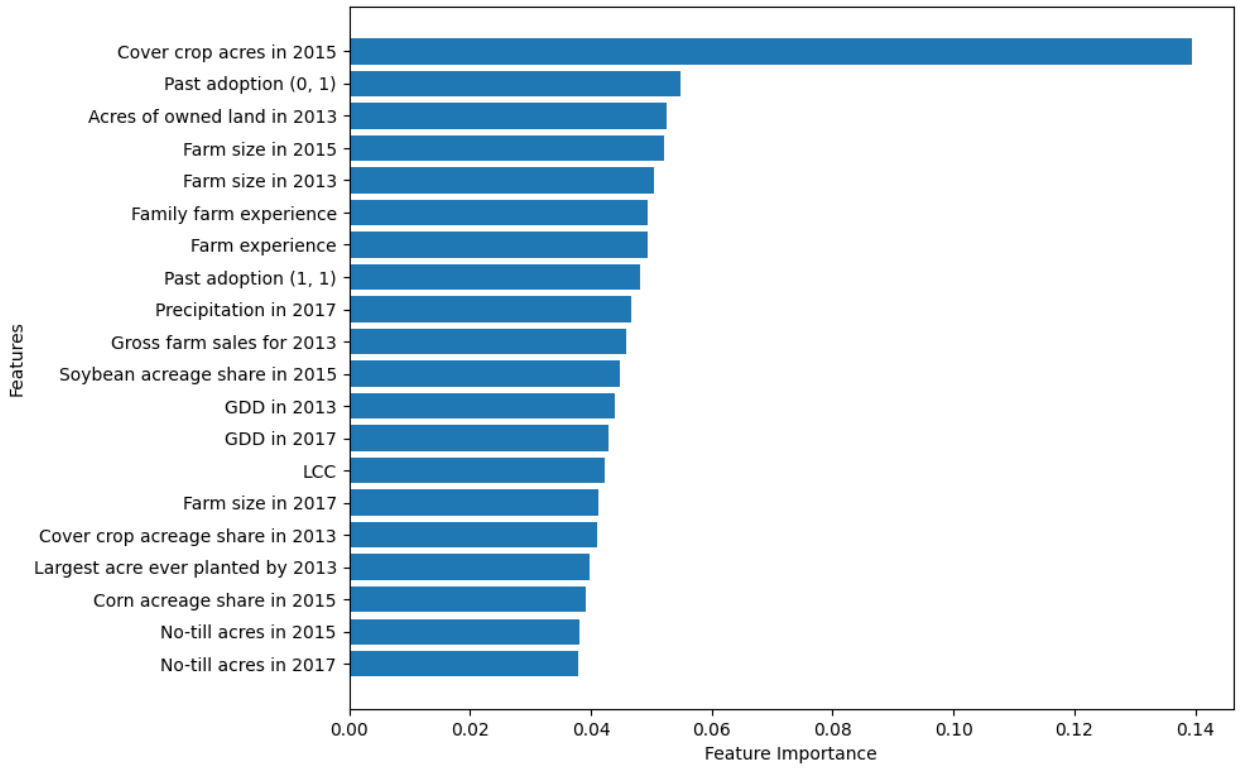

Figure A3. Feature importance rankings from Random Forest during variable selection process, showing the relative predictive power of different variables on a scale from 0 to 1. Higher values indicate stronger predictive power.

Appendix E. Description of Explanatory Variables

We focus on 4 key driving factors, as we explain earlier in the Literature section. The first is past adoption behavior, which is constructed using adoption in 2013 and 2015, categorized as follows: Past adoption (0,0) for non-adoption in both years; Past adoption (0,1) for adoption in 2015 but not in 2013; Past adoption (1,0) for adoption in 2013 but not in 2015; Past adoption (1,1) for adoption for both years. The past adoption (0,0) is the reference group.

The other three dummy variables that have been mentioned in the literature review part include crop insurance purchase in 2015 and cost-share participation in 2013. We use these specific years due to data availability. The cost-share programs for cover crops in Iowa typically last for three years (Iowa Department of Agriculture and Land Stewardship). The dummy variable for cost-share received four years ago allows us to assess the long-term effectiveness of the cost-share program and determine whether respondents discontinue cover crop use after the cost-share payment contract ends. Additionally, we include a binary indicator for no-till adoption in 2017 to evaluate the complementarity of various conservation practices.

Farm operation characteristics are represented by several variables: farm size (total acres), rent land share (percentage of rented land to total farm acres), farm income share (percentage of farm income to net household income), farming experience (years), and number of livestock. For livestock, we only include animals that can utilize cover crops in grazing systems (beef cattle, dairy cattle, goats, and sheep), converting numbers to animal units (Iowa Department of Natural Resources).

Land quality is measured by the county-level share of Land Capability Classification (LCC) I and II, which ranges from 11.4% to 93.4% in our dataset. Additionally, we include a binary indicator for highly erodible land. Farmers are more likely to adopt cover crops on highly erodible land and lower LCC land, as cover crops have been shown to reduce soil erosion and improve soil productivity (Mitchell et al., 2017; Wulanningtyas et al., 2021). Weather effects are represented by county-level precipitation during the growing season (May to August). Precipitation may have a "U-shaped" impact on soil: too little precipitation may increase the likelihood of cover crop adoption, while excessive precipitation may decrease it (Park et al., 2023).

In the questionnaire, respondents were asked to rate their agreement with several statements about cover crops. We include three statements to capture their views on cover crops while also considering multicollinearity if more are included. For each statement, respondents chose from five choices (1 = Strongly Disagree; 2 = Disagree; 3 = Uncertain; 4 = Agree; 5 = Strongly Agree). The three statements we include in our analysis are: "Landlords' lack of knowledge of or support for cover crops is a barrier to cover crop use" (lack of knowledge), "I would like to learn more about using cover crops" (cover crop learning desire), and "Cover crops reduce yields in crops that follow" (cover crop reducing yield). Finally, we include "service center visit," which records the number of times respondents visited a USDA Service Center for conservation assistance since spring 2013.

To better understand the relationships among the selected explanatory variables, we conducted a pairwise correlation analysis, as shown in Figure A4. Overall, the pairwise correlations between the variables are relatively low, indicating that the variables capture different dimensions of the factors influencing adoption decisions. This suggests that the selected

variables provide distinct and complementary information. While most correlations are relatively low, there are a few instances of stronger correlations that may require attention. For example, the correlation between farm size and rented land share is moderately high (approximately 0.5), which could be reasonable given that larger farms may have more rented land.

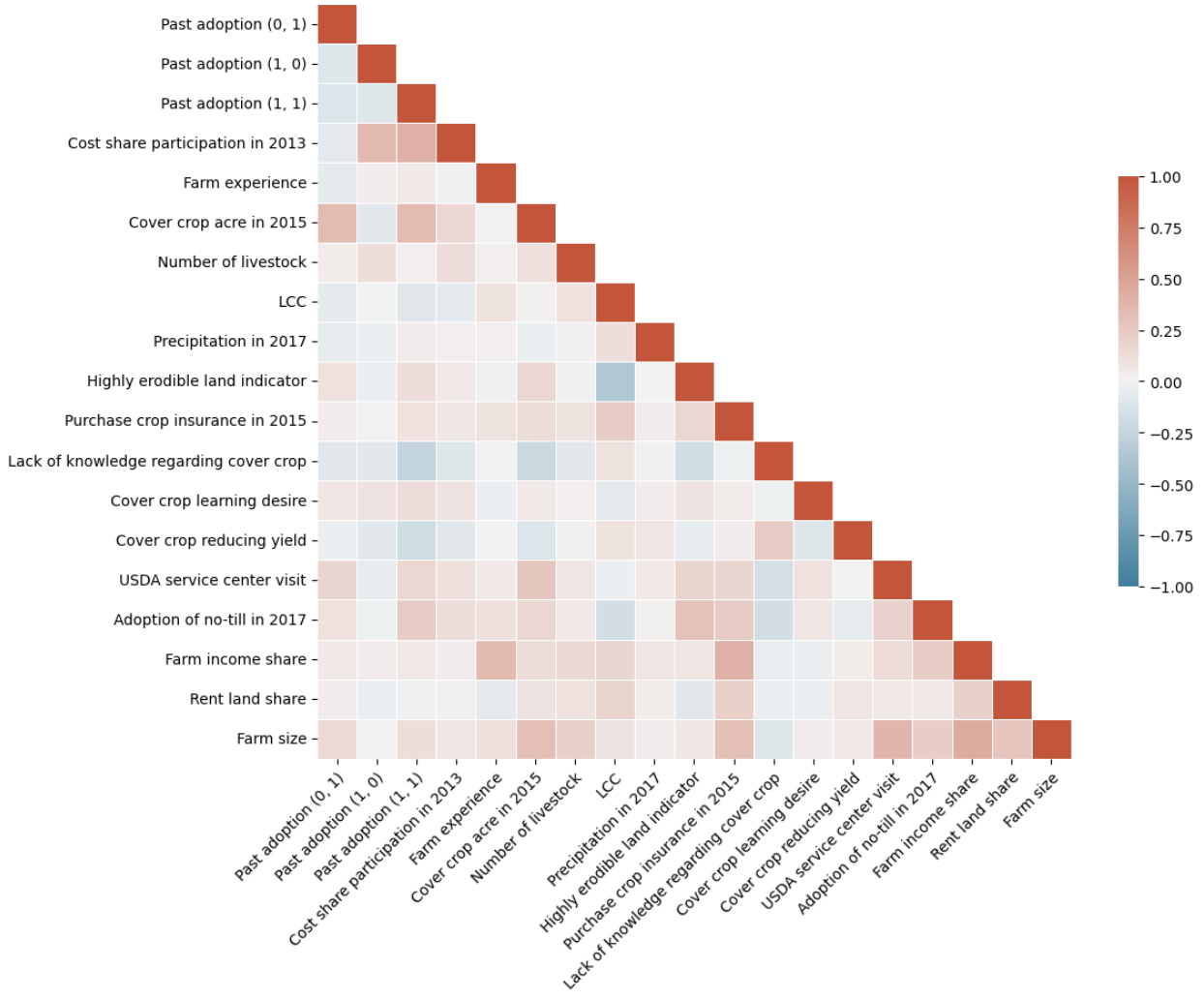

Figure A4. Pairwise correlation matrix between model variables, with darker blue showing stronger positive correlation and darker red showing stronger negative correlation on a scale from -1 to 1.

For traditional econometric models like logistic regression, high pairwise correlations can be problematic and may require excluding certain variables from the model to avoid

multicollinearity issues. However, for machine learning methods such as Random Forest, high pairwise correlations between some explanatory variables are less problematic. Our dual analysis allows us to leverage the strengths of each method while mitigating their weaknesses.

References

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction* (Vol. 2, pp. 1-758). New York: springer.

Iowa Department of Natural Resources. Animals to Animal Units Conversion., Retrieved January 23, 2025, https://www.iowadnr.gov/Portals/idnr/uploads/forms/5420020.pdf

Iowa Department of Agriculture and Land Stewardship. Guide to cover crop initiatives in Iowa.

Lundberg, S., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *arXiv.*, Retrieved January 23, 2025, https://doi.org/10.48550/arXiv.1705.07874

Mitchell, J. P., Shrestha, A., Mathesius, K., Scow, K. M., Southard, R. J., Haney, R. L., ... & Horwath, W. R. (2017). Cover cropping and no-tillage improve soil health in an arid irrigated cropping system in California's San Joaquin Valley, USA. *Soil and Tillage Research*, *165*, 325-335.

Park, B., Rejesus, R. M., Aglasan, S., Che, Y., Hagen, S. C., & Salas, W. (2023). Payments from agricultural conservation programs and cover crop adoption. *Applied Economic Perspectives and Policy*, 45(2), 984-1007.

Wulanningtyas, H. S., Gong, Y., Li, P., Sakagami, N., Nishiwaki, J., & Komatsuzaki, M. (2021). A cover crop and no-tillage system for enhancing soil health by increasing soil organic matter in soybean cultivation. *Soil and Tillage Research*, 205, 104749.