International Journal of Social Trends
Volume 4, Issue 6 (Nov-Dec; 2026)

Review Article · Economics and Public Policy

Prediction, Causation and Policy: A Review of Machine Learning in Development Economics and Public Policy

Jovana Diabaté [1], **Luka Chanda** [2]

[1] University of Zambia, ZAMBIA
[2] University of Zimbabwe, ZIMBABWE

ABSTRACT

Background. Machine learning (ML) has moved from the margins of economics to the centre of applied work. It is used to measure poverty from satellites and mobile phones, to target social transfers, to estimate treatment effects with many covariates and to study who benefits most from public programmes. Objective. This review asks four questions: which policy problems are prediction problems; how far new data and ML can substitute for traditional measurement; how ML can be combined with causal inference; and what limits apply when predictions are used to allocate public resources. Methods. We synthesise methodological work in econometrics and empirical applications in development and public economics, giving priority to peer-reviewed studies that report out-of-sample performance or policy-relevant counterfactuals. Figures report published results; diagrams are the authors' design. Findings. ML performs well when the task is prediction and when the target is aggregate, such as village wealth, where models explain most of the variation. It adds less when the target is an individual life outcome. Phone-based targeting improved on feasible alternatives in an emergency programme but fell short of a full social registry. Double/debiased ML and causal forests make high-dimensional causal inference more credible, but do not replace identification. Implications. The value of ML for policy depends on matching the tool to the question, validating predictions against the outcome policy actually cares about, and anticipating strategic responses, distribution shift and distributional consequences.

Keywords: Machine Learning; Econometrics; Development Economics; Poverty Measurement; Targeting; Causal Inference; Heterogeneous Treatment Effects

INTRODUCTION

Economics came late to machine learning. For most of the twentieth century, the discipline's empirical tradition was organised around estimating parameters with a causal or structural interpretation: the elasticity of labour supply, the return to schooling, the effect of a price change on demand. Machine learning, by contrast, grew out of a culture that judged models by their ability to predict new data, with little interest in the interpretation of individual coefficients [1]. A widely read essay in the Journal of Economic Perspectives introduced economists to tools such as regression trees, cross-validation and the lasso, arguing that the growth of large datasets would make these methods indispensable [2]. Others argued that new data sources would change the questions economists could ask, not only the methods they used [3].

Within a decade the picture had changed. Leading economists published accessible surveys of ML for applied researchers [4, 5], econometricians developed methods that use ML to estimate causal parameters with valid inference [6-8], and development economists showed that satellite images and mobile-phone

records could measure poverty where surveys were absent [9, 10]. Public agencies began to use the same tools to allocate scarce resources, including cash transfers during the COVID-19 pandemic [11].

This review organises that literature around four research questions, set out below. Our aim is not to catalogue every application but to clarify what kinds of policy problems ML can help with, what evidence exists on its performance, and where its limits lie. We write for economists and policy analysts who know the language of causal inference and want a critical guide to where ML fits within it.

Research Questions

RQ1. Which policy problems are prediction problems, and how should they be distinguished from causal ones?

RQ2. How far can new data sources and ML substitute for, or complement, traditional economic measurement?

RQ3. How can ML be combined with causal inference to estimate average and heterogeneous effects credibly?

RQ4. What practical and ethical limits apply when ML predictions are used to allocate public resources?

PREDICTION POLICY PROBLEMS (RQ1)

The $\hat{y}$–$\hat{\beta}$ Distinction

A useful starting point is the distinction between problems that require a good prediction, $\hat{y}$, and problems that require a good estimate of a parameter, $\hat{\beta}$ [4]. Standard econometrics focuses on $\hat{\beta}$: given a model $y = \beta x + \varepsilon$, the goal is an unbiased and consistent estimate of β, usually interpreted causally. ML focuses on $\hat{y}$: the goal is to predict y for new observations as accurately as possible, and the model's internal parameters are of secondary interest. The two goals are not the same. A model can predict well while its coefficients are unstable and uninterpretable; in an illustration using housing data, lasso models fitted to different subsamples selected quite different variables while achieving similar predictive accuracy [4]. Treating such coefficients as causal estimates would be a mistake.

Kleinberg and co-authors argued that many policy decisions depend primarily on $\hat{y}$, and called these "prediction policy problems" [12]. Their leading example concerns joint-replacement surgery for elderly patients. The benefit of surgery accrues over years, so for a patient who will die soon after the operation the surgery is of little value. Deciding whom to operate on therefore depends on predicting mortality, a pure prediction task, rather than on estimating the causal effect of surgery, which is already well understood. Similar logic applies to many allocation problems: deciding which defendants to release before trial depends on predicting their behaviour if released [13], and targeting anti-poverty transfers depends on predicting who is poor [11].

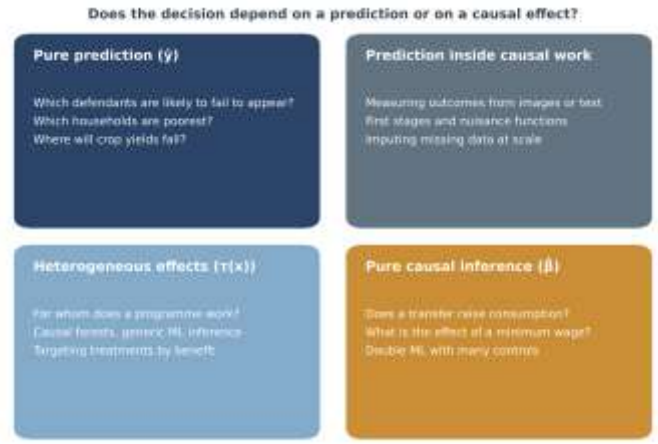


Figure 1: A typology of policy problems by whether the decision depends on a prediction, a causal effect, or both. Source: authors' design, drawing on [12], [4] and [14].

When Prediction Is Not Enough

Figure 1 shows that the boundary between prediction and causation is porous. Many applications use prediction as an input into causal work, for example by constructing outcome measures from images or text [15, 16], or by estimating the nuisance functions required by modern causal estimators [7]. Others aim at a hybrid quantity, the conditional average treatment effect $\tau(x)$, which is a causal estimand but whose estimation resembles a prediction problem [8]. Athey emphasised that most important policy questions involve some causal component, because policy changes the environment in which predictions were learned [14]. A prediction that a household is poor does not tell us whether a transfer will reduce its poverty, and a prediction that a defendant will fail to appear does not tell us whether a reminder message would prevent it.

Even pure prediction problems contain hidden causal assumptions. Evaluating a prediction rule against observed outcomes is straightforward only when outcomes are observed for everyone. In the pretrial context, rearrest is observed only for released defendants, so evaluating an algorithm that would release different people requires assumptions about what detained defendants would have done. Kleinberg et al. addressed this by exploiting the quasi-random assignment of cases to judges of differing leniency [13]. In health care, a model trained on past treatment decisions may learn the biases of past doctors rather than patients' true needs, a problem Mullainathan and Obermeyer described as automating "moral hazard and error" [17]. A related study showed that an algorithm trained to predict health-care costs as a proxy for need systematically underestimated the needs of Black patients [18]. The lesson for RQ1 is that identifying a prediction policy problem requires checking that the predicted outcome is the quantity on which the decision should depend.

THE MACHINE LEARNING TOOLKIT FOR ECONOMISTS

Table 1 summarises the methods most frequently used in applied economics, with their typical uses and the caveats that matter for policy work. All share a common logic: flexible functional forms are allowed, and

overfitting is controlled by regularisation and by evaluating performance on data not used for estimation [5, 19].

Table 1: Machine-learning methods commonly used in applied economics.

Method	Core idea	Typical economic use	Policy-relevant caveat
Lasso and ridge	Penalise coefficient size to reduce overfitting; lasso sets some coefficients to zero	Selecting controls; prediction with many covariates	Selected variables unstable; not causal
Regression trees	Recursively split the sample to form groups with similar outcomes	Interpretable rules; heterogeneity	High variance; sensitive to small data changes
Random forests	Average many trees grown on bootstrap samples and random feature subsets	Prediction; basis for causal forests	Less interpretable; extrapolate poorly
Gradient boosting	Add trees sequentially to correct previous errors	High-accuracy prediction on tabular data	Requires careful tuning; opaque
Neural networks	Layers of nonlinear transformations learned from data	Images, text and other unstructured data	Data-hungry; hard to explain
Double/debiased ML	Use ML for nuisance functions within an orthogonal moment with cross-fitting	Treatment effects with many controls	Needs unconfoundedness or another identification strategy
Causal forests	Forests adapted to estimate conditional treatment effects	Heterogeneous effects in experiments	Noisy in small samples; multiple testing

Note. Sources: [20], [21], [19], [7], [8], [5].

Three features of these methods deserve emphasis. First, tuning parameters, such as the penalty in the lasso or the depth of trees, are chosen empirically by cross-validation, which replaces the analyst's judgement about functional form with a data-driven procedure [2]. Second, the methods handle settings with more covariates than observations, which are common when data come from administrative records, text or images. Third, their outputs are not directly interpretable as structural or causal parameters [4]. These features make ML powerful for prediction and measurement but require care when its outputs feed into economic interpretation.

MEASUREMENT: POVERTY FROM SPACE AND PHONES (RQ2)

The Measurement Gap

Measuring living standards in low-income countries has traditionally relied on household surveys, which are expensive, infrequent and often unrepresentative at fine geographic scales. Many countries have gone years between consumption surveys, leaving governments and donors with little information about where the poor live. ML combined with new data sources offers a way to fill this gap.

Night Lights and Daytime Imagery

An early contribution showed that satellite measurements of night-time light intensity could be used to improve estimates of economic growth, particularly in countries whose national accounts are weak [22]. Night lights, however, are a coarse measure: they distinguish electrified from unelectrified areas but say little about variation in living standards among the poor, where lights are often absent altogether.

Jean et al. addressed this limitation with a transfer-learning approach. They trained a convolutional neural network to predict night-time lights from high-resolution daytime images, so that the network learned to recognise features such as roads, roofs and farmland that correlate with economic activity, and then used these features to predict survey-based consumption and asset wealth. Across five African countries, the model explained up to 75 percent of the variation in local economic outcomes [9]. Subsequent work using publicly available imagery extended the approach across many African countries and over time [23], and a review in Science documented a rapidly growing literature using satellite data to measure agricultural yields, infrastructure and living standards [24].

Mobile-Phone Metadata

Mobile-phone records offer a complementary source. Blumenstock, Cadamuro and On linked survey data on wealth to the call-detail records of the same individuals in Rwanda and showed that features of phone use, such as the volume, timing and network structure of calls, predicted individual wealth; aggregated, the predictions reproduced the geographic distribution of wealth estimated from national surveys [10]. Phone data have the advantage of describing individuals rather than places, which matters for targeting; they have the disadvantage of excluding people without phones and raising serious privacy concerns.

Global Microestimates

Chi et al. combined satellite imagery, mobile-phone network data, topographic maps and aggregated social-media connectivity data to produce estimates of relative wealth for 135 low- and middle-income countries at a resolution of 2.4 kilometres [25]. Trained against household surveys, the models explained between 56 and 70 percent of the variation in wealth across 56 countries with survey data, and the estimates were validated against independent census data in 15 countries. The authors released the estimates publicly, and they have since been used by governments and humanitarian organisations.

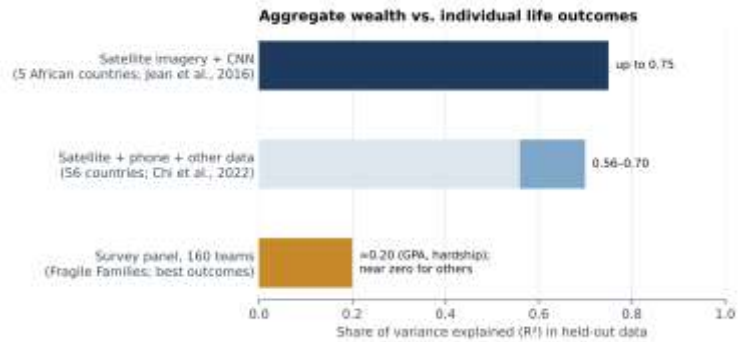


Figure 2: Variance explained by ML predictions in three settings. Bars show reported out-of-sample R^2; for Chi et al. the coloured segment shows the range across countries. Data sources: [9], [25], [26].

Table 2: Selected studies using new data and ML to measure economic well-being.

Study	Data	Setting	Target	Reported performance
Henderson et al. (2012)	Night-time lights	Global, country and sub-national	GDP growth	Improves growth estimates where national accounts are weak
Blumenstock et al. (2015)	Call-detail records linked to survey	Rwanda	Individual and regional wealth	Reproduces regional wealth distribution
Jean et al. (2016)	Daytime satellite images (transfer learning)	Five African countries	Consumption and assets	Up to 75% of variation explained
Yeh et al. (2020)	Public satellite imagery	Many African countries	Village asset wealth over time	Most cross-sectional variation explained; changes harder

Study	Data	Setting	Target	Reported performance
Chi et al. (2022)	Satellite, phone, topographic and connectivity data	135 countries, 2.4 km grid	Relative wealth	R^2 0.56–0.70 in 56 countries; census validation in 15
Naik et al. (2017)	Street-level images	US cities	Physical urban change	Identifies predictors of neighbourhood improvement

Note. Sources in row order: [22]; [10]; [9]; [23]; [25]; [16].

Text, Images and Other Unstructured Data

Beyond poverty mapping, ML has opened unstructured data to economic measurement more generally. Gentzkow, Kelly and Taddy reviewed the use of text as data in economics, describing how documents are represented numerically, how dimensionality is reduced, and how text-based measures have been used to study topics such as media slant, economic policy uncertainty and central-bank communication [15]. In urban economics, computer-vision models trained on human ratings of street-level images have been used to measure the perceived quality of neighbourhoods and to study which neighbourhoods improve over time; one such study found that density of highly educated residents and proximity to the city centre and to other attractive neighbourhoods predicted physical improvement [16].

These measures share a structure with the poverty maps described above. A model is trained to reproduce a costly measurement, such as a survey response or a human rating, from a cheap and abundant input, and then applied at scale. Their validity therefore rests on the training measurement and on the stability of the relationship between inputs and the measured quantity. When the relationship shifts, as when the meaning of words changes over time or the appearance of neighbourhoods changes for reasons unrelated to their quality, the measure can drift without any visible signal. Periodic re-validation against the underlying measurement is the main safeguard.

Complement or Substitute?

The evidence suggests that new data and ML complement rather than replace surveys. Every model in Table 2 was trained on survey data, so the quality of predictions depends on the quality and coverage of the surveys used for training. Predictions also perform better for levels than for changes: models that explain most cross-sectional variation in wealth often explain much less of its change over time [23], which limits their use for evaluating programmes. Finally, predictions are typically better at distinguishing very poor from very rich places than at ranking households within poor areas, where policy decisions are often hardest. Urban economists have reached similar conclusions about new measures of city life: they offer unprecedented coverage and timeliness but need validation against traditional data [27]. Our answer to RQ2

is therefore qualified: new data and ML can extend the reach of measurement substantially, but only when anchored to high-quality ground truth.

TARGETING PUBLIC RESOURCES

Targeting is the clearest case in which ML predictions translate directly into the allocation of public money. Many developing countries target anti-poverty programmes using proxy-means tests, which predict consumption from a small number of easily observed household characteristics. Hanna and Olken reviewed the trade-off between such targeted programmes and universal transfers, showing that targeting can raise the welfare impact of a fixed budget but that proxy-means tests make substantial errors of both inclusion and exclusion [28]. ML promises more accurate predictions, potentially reducing these errors.

The most detailed evaluation of ML targeting comes from Togo. When the COVID-19 pandemic struck, the government launched an emergency cash-transfer programme but lacked a recent social registry. Aiken et al. trained ML models on mobile-phone metadata, using a phone survey to measure consumption for a sample of subscribers, and used the predictions to prioritise transfers to the poorest. Compared with geographic targeting options the government could feasibly have used, the ML approach reduced errors of exclusion, poor people wrongly left out, by 4 to 21 percent. Compared with a hypothetical comprehensive social registry, however, it increased exclusion errors by 9 to 35 percent (Figure 3) [11].



Figure 3: Relative change in exclusion errors from phone-based ML targeting in Togo's emergency cash-transfer programme, compared with two benchmarks. Data source: [11].

The result captures both the promise and the limits of ML targeting. Where good administrative data are lacking and speed matters, ML can do better than the realistic alternatives. It is not, however, a substitute for investment in social registries and household surveys, which remain more accurate where they exist. Moreover, phone-based targeting excludes people without phones, may misclassify people whose phone use differs from the training sample, and raises questions of consent and privacy that are absent from survey-based targeting.

Strategic Responses

Once a prediction determines eligibility, people have an incentive to change the behaviour on which predictions are based. This is a version of the Lucas critique: relationships estimated under one policy regime may not hold once policy changes [29]. If phone use predicts wealth and wealth determines eligibility, poorer-looking phone use may become more common among those who are not poor. Björkegren, Blumenstock and Knight studied this problem in a field experiment and proposed estimators

that remain accurate when individuals adapt their behaviour to the rule [30]. The broader point is that targeting rules should be evaluated not only for their accuracy on historical data but also for their robustness once deployed.

MACHINE LEARNING FOR CAUSAL INFERENCE (RQ3)

Selecting Controls: Double Selection and Double ML

A common problem in applied work is choosing which control variables to include when estimating a treatment effect under an unconfoundedness assumption. When the number of potential controls is large, informal selection invites specification searching, while including all of them may be infeasible. Belloni, Chernozhukov and Hansen showed that naive use of the lasso to select controls can produce biased estimates, because variables weakly related to the outcome but strongly related to the treatment may be dropped. Their "post-double-selection" method selects controls that predict either the outcome or the treatment and then estimates the effect by ordinary least squares on the union of selected variables, delivering valid inference [6].

Double/debiased machine learning generalises this idea [7]. The method estimates the conditional expectations of the outcome and the treatment given covariates using any sufficiently accurate ML method, forms residuals, and estimates the treatment effect from a regression of residualised outcome on residualised treatment (Figure 4). Two ingredients make the procedure work. Neyman-orthogonal moment conditions make the estimate of the treatment effect insensitive to small errors in the ML estimates of the nuisance functions, and cross-fitting, in which nuisance functions are estimated on one part of the sample and residuals computed on another, removes the bias that arises from overfitting. Under regularity conditions, the resulting estimator is approximately normal with a standard error that can be computed in the usual way.

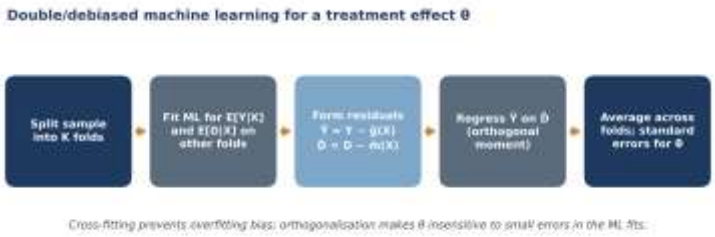


Figure 4: The double/debiased machine-learning procedure with cross-fitting. Source: authors' design, based on [7].

Double ML is an improvement in estimation, not in identification. It allows flexible and data-driven adjustment for observed confounders, but if important confounders are unobserved, the estimate remains biased. Its main contribution is to remove a source of researcher discretion and functional-form misspecification, not to solve the fundamental problem of selection on unobservables.

Heterogeneous Treatment Effects

Policy often requires knowing not only whether a programme works on average but for whom it works best. Traditional subgroup analysis is vulnerable to specification searching, because researchers can test many subgroups and report those with significant results. Athey and Imbens proposed "causal trees", which partition the sample to find subgroups with different treatment effects using one part of the data and estimate effects within those subgroups on another, a procedure they called honest estimation [31]. Wager and Athey extended the approach to causal forests, which average many causal trees and yield estimates of conditional treatment effects with valid confidence intervals [8], and generalised random forests provide a broader framework for estimating heterogeneous parameters [32].

Applications illustrate both the value and the difficulty of this work. Davis and Heller used causal forests to study a summer youth employment programme in Chicago, finding heterogeneity in effects on outcomes such as employment and violent-crime involvement, and showing that the young people who benefited most were not those that simple observable characteristics would have suggested [33]. Because estimates of individual-level effects are noisy, Chernozhukov et al. proposed methods that focus on robust features of heterogeneity, such as the average effect for the most and least affected groups, with valid inference in randomised experiments [34].

From Heterogeneity to Policy Rules

Knowing that effects differ does not immediately tell a policymaker whom to treat. Policy learning methods estimate treatment-assignment rules that maximise expected welfare, subject to constraints such as budget limits or a requirement that the rule be simple. Kitagawa and Tetenov developed "empirical welfare maximisation" for experimental data [35], and Athey and Wager extended policy learning to observational data with guarantees on the regret of the estimated rule [36]. These methods make explicit the link between causal estimation and decision-making that the prediction-policy framing leaves implicit.

Panel Data and Comparative Case Studies

ML ideas have also influenced methods for panel data. Synthetic control constructs a weighted combination of untreated units that matches a treated unit's pre-treatment trajectory [37]; it can be viewed as a regularised prediction of the treated unit's counterfactual outcomes. Matrix-completion methods treat missing potential outcomes in panel data as missing entries in a matrix and use low-rank regularisation to impute them, nesting synthetic control and fixed-effects approaches as special cases [38]. These methods illustrate how the prediction toolkit can be repurposed for causal estimands when identification rests on stable relationships across units and time.

Table 3: Machine-learning approaches to causal questions.

Approach	Estimand	Identification assumption	Key reference	Example of use
Post-double-selection lasso	Average treatment effect	Unconfoundedness given selected controls	Belloni et al. (2014)	Effects with many potential controls

Approach	Estimand	Identification assumption	Key reference	Example of use
Double/debiased ML	Average or local average effect	Unconfoundedness or instrument validity	Chernozhukov et al. (2018)	Programme evaluation with rich administrative data
Causal trees and forests	Conditional average treatment effect	Randomisation or unconfoundedness	Athey & Imbens (2016); Wager & Athey (2018)	Heterogeneity in youth employment programmes
Generic ML inference	Features of effect heterogeneity	Randomisation	Chernozhukov et al. (2018, NBER)	Most- and least-affected groups in experiments
Policy learning	Welfare-maximising assignment rule	Randomisation or unconfoundedness	Kitagawa & Tetenov (2018); Athey & Wager (2021)	Choosing whom to treat under a budget
Matrix completion	Effects in panel data	Low-rank structure of untreated outcomes	Athey et al. (2021)	Policy changes across regions over time

Note. Sources: [6], [7], [31], [8], [34], [35], [36], [38].

LIMITS AND RISKS (RQ4)

The Predictability of Life Outcomes

A sobering test of what prediction can achieve came from the Fragile Families Challenge. Salganik and colleagues invited researchers to predict six life outcomes for children and families, including children's school grades, household material hardship and eviction, using a rich longitudinal dataset collected over fifteen years. The 160 participating teams used a wide range of methods, including many ML approaches. Even the best predictions were not very accurate: for grade point average and material hardship the best models explained only about a fifth of the variance in held-out data, and for the other outcomes they explained almost none of it. Moreover, the best submissions were only slightly better than a simple benchmark model using four variables chosen by a domain expert [26].

The contrast with the poverty-mapping results in Figure 2 is instructive. Aggregate outcomes, such as the average wealth of a village, are much more predictable than individual outcomes, which are shaped by idiosyncratic events that no dataset records. Policies that target individuals on the basis of predicted outcomes should therefore expect substantial errors, even with excellent data and methods. This is an argument for humility, for combining prediction with other information, and for designing programmes that tolerate misclassification.

Distribution Shift and External Validity

ML models learn relationships present in their training data. When the population or the environment changes, those relationships may no longer hold. A model trained on survey data from one year may perform poorly after an economic shock that changes the relationship between observable features and wealth, and a model trained in one country may not transfer to another. External validity, long a concern in programme evaluation, is equally important for prediction [14]. Validation on data from the setting where the model will be used, and monitoring after deployment, are essential.

Fairness and Distributional Consequences

Economists have traditionally evaluated targeting by its efficiency in reaching the poor. ML targeting adds new distributional questions. Errors may be concentrated among groups that are under-represented in training data, such as people without phones or people in remote areas, and predictions may encode biases in the outcome used for training [18]. Because the welfare consequences of exclusion errors fall on those excluded, evaluations should report error rates by relevant subgroup as well as in aggregate.

Privacy and Governance

Many new data sources contain sensitive information about individuals. Phone-based targeting requires access to call-detail records; satellite and street-level imagery can reveal information about homes and neighbourhoods. Using such data in public programmes requires legal authority, safeguards against misuse and meaningful consent where feasible. These issues are not peripheral: they determine whether the public will accept data-driven allocation, and therefore whether its efficiency gains can be realised.

Table 4: Risks of using ML predictions for public allocation, with mitigation strategies.

Risk	Mechanism	Illustration	Mitigation
Low individual predictability	Individual outcomes depend on unobserved events	Fragile Families Challenge	Use predictions as one input; design for error
Label bias	Training outcome is a biased proxy for the target	Cost used as a proxy for health need	Validate labels; choose outcomes carefully
Strategic manipulation	People change behaviour to meet eligibility rules	Phone-use features in targeting	Robust features; manipulation-proof estimators
Distribution shift	Relationships change across time or place	Shocks altering wealth predictors	Local validation; monitoring; retraining
Coverage gaps	Some groups absent from data	People without phones	Combine sources; complementary outreach
Privacy	Sensitive data used for allocation	Call-detail records	Legal safeguards; aggregation; consent

Note. Sources: [26], [18], [30], [23], [11].

CASE STUDY: ALGORITHMS AND JUDICIAL RELEASE DECISIONS

The most complete economic analysis of a prediction policy problem concerns pretrial release in New York City [13]. Judges deciding whether to release defendants before trial are legally required to consider the risk that a defendant will fail to appear in court. Kleinberg and co-authors trained a gradient-boosted decision-tree model on hundreds of thousands of cases to predict this risk from information available to the judge at the time of the decision, and then asked how outcomes would change if release decisions followed the model's predictions.

Two econometric problems complicated the analysis. The first was the selective labels problem already mentioned: outcomes are observed only for released defendants. The authors addressed it by exploiting variation across judges in leniency. Because cases are assigned to judges in a quasi-random way, the defendants released by a lenient judge but detained by a stricter one reveal what happens when the marginal defendant is released. The second was the possibility that judges use information not recorded in the data, such as the defendant's demeanour. The authors examined this possibility directly and found that judges' departures from the algorithm's ranking did not appear to reflect valuable private information; rather, judges released many defendants whom the algorithm rated as high risk.

The simulations suggested that a release rule based on the algorithm could reduce crime among released defendants substantially at the current detention rate, or reduce the jail population substantially without increasing crime, and that these gains could be achieved while reducing the share of detained defendants who were Black or Hispanic [13]. The study also illustrates the boundaries of the prediction policy framing. Its conclusions apply to the outcome the algorithm predicts, and they assume that the relationship between features and outcomes would remain stable under a new release rule. Whether such gains materialise in practice depends on how judges use the predictions, a question the simulation cannot answer.

For economists, the case offers a template for evaluating prediction policy problems. It specifies the decision and the objective, confronts the problem of unobserved counterfactual outcomes with a credible research design, tests whether human decision-makers possess information the algorithm lacks, and reports distributional consequences alongside efficiency gains. Few applied studies meet all four standards, and adopting them would raise the quality of evidence on ML in policy considerably.

PRACTICAL GUIDANCE FOR APPLIED RESEARCHERS

Drawing on the literature reviewed above, we offer six recommendations for economists and policy analysts using ML in empirical work.

Define the estimand first. Before choosing a method, state whether the goal is a prediction, an average causal effect, a conditional effect or a decision rule. The choice of method follows from the estimand, not the reverse [4, 5].

Separate model selection from evaluation. Use held-out data, cross-validation or cross-fitting so that reported performance and inference are not contaminated by overfitting. In causal work, honest estimation and cross-fitting serve the same purpose [7, 31].

Report benchmarks. Compare ML models with simple, transparent alternatives. The Fragile Families Challenge showed that a four-variable model came close to the best ML submissions [26]; in many settings a well-specified linear model will perform similarly to a complex one, and the difference should be reported.

Validate the target. Check that the predicted outcome is the quantity the policy cares about, and examine how its measurement differs across groups [18]. Where possible, validate predictions against independent data, as Chi et al. did with census data [25].

Anticipate deployment. Ask how individuals and agencies will respond to the use of predictions, how the population may change, and how the system will be monitored and updated [29, 30].

Pre-specify where possible. In experiments, pre-specify the heterogeneity analyses to be conducted and use methods with valid inference for the features of heterogeneity reported [34]. Flexible methods expand the space of possible analyses, which increases rather than decreases the importance of transparency about what was planned.

ANSWERS TO THE RESEARCH QUESTIONS

The table below summarises our answers to the four research questions.

Table 5: Summary of findings by research question.

Research question	Short answer	Main supporting evidence
RQ1: Which problems are prediction problems?	Those where the decision depends on an outcome that the action does not change, provided the predicted outcome is the right target	Prediction policy framework; pretrial and health-care applications; label-bias evidence
RQ2: Can new data and ML replace traditional measurement?	They extend coverage and timeliness, especially for aggregate wealth, but depend on surveys for training and validation	Satellite and phone poverty mapping; global microestimates
RQ3: How can ML support causal inference?	By flexibly estimating nuisance functions and heterogeneous effects within a credible identification strategy	Double selection, double ML, causal forests, policy learning
RQ4: What limits apply in allocation?	Individual outcomes are hard to predict; predictions are vulnerable to manipulation, shift, bias and coverage gaps	Fragile Families Challenge; Togo targeting; manipulation-proof ML

Note. Summary of evidence reviewed in the sections above.

Three cross-cutting lessons emerge. First, matching the tool to the question matters more than the sophistication of the tool. Many disappointments in applied ML stem from treating a causal question as a prediction problem or from predicting the wrong outcome. Second, ML is most valuable when combined with economic reasoning about identification, incentives and welfare. The methodological advances of the

past decade, such as double ML and policy learning, are successful precisely because they embed ML within economic and statistical frameworks rather than replacing them. Third, the performance of ML in policy depends on institutions: on investment in high-quality ground-truth data, on legal frameworks governing data use, and on the capacity of agencies to monitor and revise predictive systems.

DIRECTIONS FOR FUTURE RESEARCH

Several research directions follow from this review. The first is evaluation of deployed systems. Much of the evidence on ML in policy comes from retrospective analyses or simulations; there are few randomised evaluations comparing ML-based allocation with alternatives in real programmes. Such evaluations would reveal how predictions perform once people respond to them and how agencies use them in practice.

The second direction is the measurement of change. Poverty maps are increasingly accurate in cross-section, but tracking changes over time, which is needed for programme evaluation and crisis response, remains difficult [23]. Combining frequent but noisy signals, such as satellite and phone data, with infrequent but accurate surveys is a promising area for methodological work.

The third direction concerns welfare-based evaluation of predictive systems. Accuracy metrics such as R^2 or the area under a curve do not translate directly into welfare. Evaluations should report the welfare consequences of errors, weighted by the circumstances of those affected, and should compare predictive allocation with alternatives such as universal programmes [28]. Policy-learning methods provide a natural framework for such comparisons [35, 36].

The fourth direction is the integration of text and other unstructured data into economic measurement. Text from news, social media, policy documents and administrative records offers new measures of economic sentiment, uncertainty and policy [15], and images of streets and buildings offer new measures of urban change [16]. Validating these measures and understanding their biases is a substantial agenda in its own right.

Finally, there is the question of capacity. The benefits of ML for public policy accrue only to governments that can collect, manage and analyse data. Investment in statistical capacity, data infrastructure and analytical skills in low- and middle-income countries is a precondition for many of the applications reviewed here, and the distribution of that capacity will shape who benefits from the methods.

CONCLUSION

Machine learning has become part of the economist's toolkit, and its uses in development economics and public policy are increasingly consequential. The evidence reviewed here supports a balanced assessment. ML excels at prediction and measurement, particularly of aggregate outcomes, and it can improve the targeting of resources when better data are unavailable [9, 11, 25]. Combined with careful identification, it makes causal inference more flexible and transparent, and it enables the study of who benefits from policy [7, 8, 36]. At the same time, individual outcomes remain hard to predict [26], predictions can encode biases and invite strategic responses [18, 30], and no algorithm substitutes for a credible research design or for investment in high-quality data.

The central contribution of the economics literature has been to place machine learning within a framework that asks what decision a prediction informs, what causal assumptions it relies on and what the welfare consequences of its errors are [4, 5, 12]. That framework is likely to remain the discipline's most important contribution as ML continues to spread through public policy.

REFERENECS

[1] Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical Science, 16*(3), 199–231.

[2] Potharaju, B. (2026). Infrastructure-level intelligence: Embedding AI into data movement and validation layers. *International Journal of Computational and Experimental Science and Engineering*.

[3] Varian, H. R. (2014). Big data: New tricks for econometrics. *Journal of Economic Perspectives, 28*(2), 3–28.

[4] Mallempati, A., Jaladi, D. S., & Jegajothi, B. (2026). HDLF-DGC: A hybrid deep learning framework for intelligent cybersecurity and data governance in distributed web systems. In *2026 Third International Conference on Innovations in Cybersecurity and Data Science (ICICDS)* (pp. 16–21). IEEE.

[5] Einav, L., & Levin, J. (2014). Economics in the age of big data. *Science, 346*(6210), 1243089.

[6] Potharaju, B., & Ravipudi, J. (2023). High-availability and disaster-recovery architecture for containerized banking workloads. *International Journal of Data Engineering and Intelligent Computing*, 6(1), 1–32.

[7] Rahman, M. S., Kabir, A. A., Reon, M. A. H., & Rashid, S. U. (2025). From Financial Transactions to Clinical Records: A Unified Machine Learning Methodology for Real-Time Anomaly Detection and Institutional Decision Support In High-Stakes Domains. *International Journal of Information Technology and Management Information Systems (IJITMIS)*, *16*(6), 23-42.

[8] Karim, N. B., & Moreau, L. (2026). Quantifying the Prospective Impact of AI-Assisted Early Intervention in US Small-Business Lending: Default Avoidance, Credit-Access Preservation, and Small-Business Employment Effects Across Community Lending Portfolios. *The Metascience*, *4*(2), 33-44.

[9] Potharaju, A. (2026). Policy-driven authorization architectures for modern financial regulatory platforms. *Journal of International Crisis and Risk Communication Research*, 9(4), 36–44.

[10] Mullainathan, S., & Spiess, J. (2017). Machine learning: An applied econometric approach. *Journal of Economic Perspectives, 31*(2), 87–106.

[11] Athey, S., & Imbens, G. W. (2019). Machine learning methods that economists should know about. *Annual Review of Economics, 11*, 685–725.

[12] Movva, C. K., & Mupparapu, H. K. (2026). MVVM architectural patterns for enterprise Android kiosk applications: State management and lifecycle handling. *International Journal of Emerging Trends in Computer Science and Information Technology*, 7(2), 320–330.

[13] Belloni, A., Chernozhukov, V., & Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies, 81*(2), 608–650.

[14] Rashid, S. U., Al Muzakkiruddin, S., & Møller, C. (2026). Detecting and Correcting Temporal Drift in Credit-Risk and Fraud-Detection Models at Community Banks, Credit Unions, and CDFIs: An Automated Model-Risk-Monitoring Framework Below the Supervisory Threshold. *The Computertech*, 26-37.

[15] Potharaju, A. (2026). Microservices-based modernization of financial regulatory systems. *International Journal of Computer Science and Engineering Innovations*, 2(1), 68–85.

[16] Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal, 21*(1), C1–C68.

[17] Kumar, M. S., Yuvaraj, N., & Manoj, M. S. (2026). Intelligent orchestration of master data and privacy compliance via agentic AI for customer lifecycle optimization. In *2026 IEEE 18th International Conference on Computational Intelligence and Communication Networks (CICN)* (pp. 1715–1723). IEEE.

[18] Wager, S., & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association, 113*(523), 1228–1242.

[19] Mallempati, A., & Jaladi, D. S. (2023). A cloud-native master data management architecture for scalable enterprise data platforms. *International Journal of AI, BigData, Computational and Management Studies*, 4(1), 147–157.

[20] Yuvaraj, N., & Kumar, M. S. (2025). Agentic AI for zero-touch customer experience: A governance-constrained framework for autonomous service systems. *American International Journal of Computer Science and Technology*, 7(2), 100–111.

[21] Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B., & Ermon, S. (2016). Combining satellite imagery and machine learning to predict poverty. *Science, 353*(6301), 790–794

[22] Potharaju, A. (2026). Embedded compliance architecture for real-time financial systems modernization. *Journal of Information Systems Engineering and Management*, 11(2s), 1147–1158.

[23] Blumenstock, J., Cadamuro, G., & On, R. (2015). Predicting poverty and wealth from mobile phone metadata. *Science, 350*(6264), 1073–1076.

[24] Aiken, E., Bellue, S., Karlan, D., Udry, C., & Blumenstock, J. E. (2022). Machine learning and phone data can improve targeting of humanitarian aid. *Nature, 603*(7903), 864–870.

[25] Kleinberg, J., Ludwig, J., Mullainathan, S., & Obermeyer, Z. (2015). Prediction policy problems. *American Economic Review, 105*(5), 491–495.

[26] Shammah, R. S., Sultana, S., Faruq, O., Kowser, S., Uddin, A., & Haque, S. A. (2026, July). Data-Driven Decision Support Framework for Engineering and IT Project Management Using Integrated Business Analytics and Management Information Systems. In *2026 IEEE International Conference on Artificial Intelligence and Mechatronics Systems (AIMS)* (pp. 1-6). IEEE.

[27] Karim, N. B., & Kovac, T. (2025). Cloud-Based Big-Data Architectures for AI-Driven Financial Planning and Analysis and Early Risk Signaling in Resource-Constrained Community Financial Institutions. *The Metascience, 3*(3), 29-40.

[28] Potharaju, B. (2026). Architecting national-scale regulatory reporting platforms: A high-availability infrastructure framework for compliance-critical systems. *Journal of International Crisis & Risk Communication*.

[29] 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.

[30] Movva, C. K., & Manohar, V. (2021). OCF protocol implementation for smart home IoT applications: An Android SDK development case study. *International Journal of Artificial Intelligence, Data Science, and Machine Learning*, 2(1), 89–96.

[31] Aluri, G. N. C., & Mupparapu, H. K. (2021). Spring Batch framework patterns for large-scale data processing in cloud-native microservices: A GCP deployment study. *International Journal of AI, BigData, Computational and Management Studies*, 2(4), 130–142.

[32] Athey, S. (2017). Beyond prediction: Using big data for policy problems. *Science, 355*(6324), 483–485.

[33] Aluri, G. N. C., & Manohar, V. (2026). Model context protocol servers as orchestration layers for generative AI agents in enterprise software delivery. *International Journal of AI, BigData, Computational and Management Studies*, 7(2), 337–345.

[34] Gentzkow, M., Kelly, B., & Taddy, M. (2019). Text as data. *Journal of Economic Literature, 57*(3), 535–574.

[35] Yuvaraj, N., Kumar, M. S., & Manoj, M. S. (2026). Unified agentic AI architecture for data lineage, governance, and predictive workflow automation in enterprise systems. In *2026 IEEE 18th International Conference on Computational Intelligence and Communication Networks (CICN)* (pp. 1817–1828). IEEE.

[36] Naik, N., Kominers, S. D., Raskar, R., Glaeser, E. L., & Hidalgo, C. A. (2017). Computer vision uncovers predictors of physical urban change. *Proceedings of the National Academy of Sciences, 114*(29), 7571–7576.

[37] Jaladi, D. S., Mallempati, A., & Jegajothi, B. (2026). A hybrid federated privacy-aware deep learning framework for secure data governance in cloud systems. In *2026 4th International Conference on Sustainable Computing and Smart Systems (ICSCSS)* (pp. 1629–1636). IEEE.

[38] Zuberi, N., Rahman, M. M., Rashid, S. U., Ahsan, M. N., Rahman, M. D., & Mollah, M. H. O. R. (2026). A robust machine learning framework for detecting temporal drift in financial fraud prevention. *Scientific Reports*.

[39] Potharaju, B., & Ravipudi, J. (2025). Automated data validation and resilient infrastructure design for enterprise regulatory reporting platforms. *International Journal of Computer Science and Engineering Innovations*, 1(1), 46–64.

[40] Karim, N. B., Tolosana, D., & Simko, R. (2025). Responsible and Legally Compliant Enterprise Artificial Intelligence in Small-Business Lending: Governance, Explainability, and Fair-Lending Safeguards for Early-Intervention Recommendation Engines. *International Journal of Modern Computing, 8*(1), 55-66.

[41] Mullainathan, S., & Obermeyer, Z. (2017). Does machine learning automate moral hazard and error? *American Economic Review, 107*(5), 476–480.

[42] Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science, 366*(6464), 447–453.

[43] Rahman, M. S., Bhowmik, P. K., Hossain, B., Tannier, N. R., Amjad, M. H. H., Chouksey, A., & Hossain, M. (2023). Enhancing fraud detection systems in the USA: A machine learning approach to identifying anomalous transactions. *Journal of Economics, Finance and Accounting Studies*, *5*(5), 145-160.

[44] Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer.

[45] Karim, N. B., Kovac, T., & Song-Chun, Z. (2025). Model Risk Management for Adaptive Credit Early-Warning Systems in Supervised SmallBusiness Lending: Aligning Machine-Learning Recalibration with Federal Model-Risk Guidance and the NIST AI Risk Management Framework. *International Journal of Acta Informatica*, *4*(1), 80-94.

[46] Mallempati, A. (2024). Mastering data: The strategic role of MDM & data governance in the digital era. *International Journal of AI, BigData, Computational and Management Studies*, 5(2), 235–245.

[47] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological), 58*(1), 267–288.

[48] Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.

[49] Manohar, V., & Mupparapu, H. K. (2025). Scalable data pipeline architecture for real-time supply chain analytics using PySpark and Snowflake. *International Journal of Artificial Intelligence, Data Science, and Machine Learning*, 6(3), 203–212.

[50] Rahman, M. M., Computer, M. M. R., Das, N., Uddin, M. S., Shammah, R. S., & Rahman, M. M. (2025, May). An Investigation of Face Detection and Feature Extraction Techniques in Deepfake Detection Using a Multi-Model Machine Learning Approach. In *2025 3rd International Conference on Business Analytics for Technology and Security (ICBATS)* (pp. 1-8). IEEE.

[51] Mupparapu, H. K., & Movva, C. K. (2020). Serverless migration patterns for compliance-critical financial applications on Microsoft Azure. *International Journal of Emerging Research in Engineering and Technology*, 1(2), 85–96.

[52] Henderson, J. V., Storeygard, A., & Weil, D. N. (2012). Measuring economic growth from outer space. *American Economic Review*, *102*(2), 994–1028.

[53] Siddiqui, M. I. H., Rahman, M. S., Kabir, A. A., Mahmud, F. U., Rashid, S. U., & Shammah, R. S. (2023). Comparative analysis of explainable machine learning models for cancer classification using cytological features. *Journal of Medical and Health Studies*, *4*(5), 110-150.

[54] Potharaju, A. (2026). Human-guided artificial intelligence for adaptive compliance automation in financial institutions. *Journal of Computational Analysis and Applications*, 35(4), 57–68.

[55] Yeh, C., Perez, A., Driscoll, A., Azzari, G., Tang, Z., Lobell, D., Ermon, S., & Burke, M. (2020). Using publicly available satellite imagery and deep learning to understand economic well-being in Africa. *Nature Communications, 11*, 2583.

[56] Movva, C. K., & Aluri, G. N. C. (2025). Enterprise mobile payment gateway architecture on IoT handheld devices: Security and reliability design. *International Journal of AI, BigData, Computational and Management Studies*, 6(1), 232–241.

[57] Burke, M., Driscoll, A., Lobell, D. B., & Ermon, S. (2021). Using satellite imagery to understand and promote sustainable development. *Science, 371*(6535), eabe8628.

[58] Aluri, G. N. C., & Manohar, V. (2025). Generative AI integration patterns for enterprise workflow automation: A practitioner framework. *International Journal of Emerging Research in Engineering and Technology,* 6(3), 165–172.

[59] Shahiduzzaman, M., Sourav, M. S. A., Kabtia, M., Shammah, R. S., Islam, M. A., Chakraborty, P., ... & Adnan, M. (2026). Attention-guided quantum-inspired extreme learning machine with grey wolf optimization for rare-attack detection in IIoT networks. *Discover Artificial Intelligence*.

[60] Potharaju, A. (2021). Designing dynamic questionnaire engines for risk-based account review workflows. *International Journal of Applied Data Science & Modern Computing*, 4(1), 1–37.

[61] Chi, G., Fang, H., Chatterjee, S., & Blumenstock, J. E. (2022). Microestimates of wealth for all low- and middle-income countries. *Proceedings of the National Academy of Sciences, 119*(3), e2113658119.

[62] Kumar, M. S., & Yuvaraj, N. (2024). Predictive customer experience orchestration using governed data pipelines and intelligent service signals. *International Journal of Emerging Trends in Computer Science and Information Technology*, 5(1), 206–215.

[63] Shammah, R. S., Rahman, I., Guria, Z. M., Kowser, S., Faruq, O., & Morshed, N. (2025, December). Scalable Machine Learning Architecture for Securing IoT Networks with Digital Governance in the AI Era of Cyber-Resilient Smart Cities. In *2025 International Conference on AI-Driven STEM Education and Learning Technologies (AISTEMEDU)* (pp. 1-7). IEEE.

[64] Potharaju, B., & Sajja, R. R. (2024). Modernizing US federal regulatory reporting infrastructure for scalability, resilience, and audit readiness. *International Journal of Data Engineering and Intelligent Computing*, 7(1), 1–36.