

Valuing User Data by Predictive Contribution: A Framework for Privacy Compensation under Heterogeneous Risk Preferences in Platform Markets

Yi Gao

Ningbo University

Tao Chen

chentao@nbu.edu.cn

Ningbo University

Ying Jin

Ningbo University

Article

Keywords: data valuation, machine learning, privacy compensation, predictive utility, SHAP values

Posted Date: September 30th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10623666/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Valuing User Data by Predictive Contribution: A Framework for Privacy Compensation under Heterogeneous Risk Preferences in Platform Markets

Abstract

User data are a key production asset in platform markets, yet existing valuation approaches rarely connect attribute-level predictive contribution with the privacy costs borne by users, nor do they reflect heterogeneous risk preferences in disclosure. This study develops an integrated framework that values user data by predictive contribution rather than sensitivity labels, and links valuation to privacy compensation. We evaluate predictive utility using machine-learning models, quantify attribute-level contributions with SHapley Additive exPlanations (SHAP), and combine these estimates with privacy-sensitivity levels and heterogeneous risk-preference types to derive differentiated compensation. We further conduct data-volume sensitivity analyses to examine diminishing marginal gains in predictive performance and incorporate predictive utility, aggregate compensation, and platform operating costs into a unified valuation function. Empirically, transactional attributes account for 81% of the calibrated aggregate privacy value, while sensitive demographic attributes contribute little to prediction; predictive utility stabilizes once 40–60% of the training sample is used. These results separate privacy sensitivity from economic contribution and show how compensation rules can be adjusted across risk-preference types. The framework offers a structured basis for differentiated compensation, selective data collection, and platform data governance.

Keywords

data valuation; machine learning; privacy compensation; predictive utility; SHAP values

1 Introduction

In the digital economy, a fundamental disconnect has emerged: while user-generated data serves as the lifeblood of platform enterprises, the mechanisms for valuing this asset remain significantly detached from both its economic contribution and the privacy costs borne by its originators (Arrieta-Ibarra et al., 2018; Zuboff, 2015). Digital platforms control the aggregation, analysis, valuation, and commercialization of user data, whereas users bear the privacy costs associated with the authorized use of their personal and behavioral information (Fainmesser et al., 2023). Meanwhile, the new goods and services are increasingly bartered for data (Veldkamp, 2023). Data and data-driven services have created huge value in various fields, resulting in a large number of data and service markets (Peng et al., 2025). In the data market, data exchange occurs freely, yet this process often neglects the intrinsic value of the data (Baghcheband et al., 2025). The routine conversion of everyday activities into commercially exploitable data can produce unequal data relations between platforms and individuals (Couldry & Mejias, 2019). Developing a transparent valuation mechanism is therefore essential not only for improving data market efficiency but also for determining how the benefits and costs of data use should be distributed among platforms, users, and data consumers.

The distinctive economic properties of data make this distributive problem difficult to resolve.

Data is an intangible asset, and how to value it in a reasonable manner is crucial for the data market. This is especially true for platform enterprises: they rely heavily on users' personal data to fuel their business models, using it to enable targeted information delivery and precision marketing, while also allowing partner enterprises linked to the platform to reach specific target audiences. However, as users' digital footprints grow, so do concerns over data security and online privacy.

Recently, several studies have examined data valuation from multiple perspectives, giving rise to asset-based, utility-based, contribution-based, and privacy-aware valuation approaches.

At the enterprise level, Gao and Zhang (2024) developed a big-data cooperative asset valuation model that integrates multidimensional data sources, data assets, and inter-organizational cooperative relationships to assess the value of Internet platform enterprises. At the level of individual data units, Ghorbani and Zou (2019) introduced Data Shapley to measure the marginal contribution of each observation to machine-learning performance. Research has also increasingly considered data value from the perspective of individual privacy. Mazzarello et al. (2025) found that the perceived benefits and privacy concerns associated with data are not fixed properties of a particular dataset but vary according to data combinations, transaction stages, and application contexts. Voloch and Hirschprung (2026) further developed a behavioral experimental framework to estimate the monetary value users assign to personal information from observable leakage decisions, showing that privacy valuation varies with the type of personal data, the platform of exposure, the incentive offered, and the timing of the offer.

However, despite these advances, two important disconnects remain. The first is the disconnect between utility and cost. Utility-based and contribution-based approaches quantify the predictive value captured by platforms or data buyers but rarely translate this value into compensation for the users who bear the associated privacy costs (Gao & Zhang, 2024; Ghorbani & Zou, 2019; Tian et al., 2019). Privacy-aware approaches, in turn, often treat compensation as an acquisition cost without establishing how the acquired data contribute to prediction and decision-making. The second is a contribution–preference disconnect. Model-based attribution methods provide an observable measure of the contribution made by particular data attributes, whereas privacy-compensation studies emphasize users’ subjective and heterogeneous attitudes toward disclosure (Mazzarello et al., 2025; Voloch & Hirschprung, 2026). Treating only contribution as relevant can neglect differences in perceived privacy loss, while relying only on sensitivity classifications or stated preferences can disconnect compensation from the economic value that the data actually generate. Consequently, existing

approaches provide limited guidance on how platforms can reconcile predictive value creation, privacy-risk compensation, and operational sustainability within a common valuation rule.

Furthermore, this study addresses the following questions: How can platform-generated user data be valued in a way that reconciles their predictive contribution with the heterogeneous privacy costs borne by users?

To answer this, we break it down into two specific questions. First, how does the predictive contribution of a data attribute differ from its nominal privacy-sensitivity classification? Second, how can predictive utility, attribute-level privacy risk, heterogeneous risk preferences, and platform operational costs be incorporated into a single data valuation framework?

Therefore, this study proposes an integrated valuation framework for platform-generated user data in a data ecosystem. This framework further examines the joint role of predictive utility and user privacy in data valuation. A set of machine learning algorithms were employed to evaluate data utility across classification and regression tasks. The results were integrated into a weighted composite index, enabling the valuation framework to accommodate different business applications and decision objectives. Additionally, SHAP analysis was introduced to identify the contribution of individual data attributes to model predictions. These attribute-level contributions were combined with privacy-sensitivity levels and users' heterogeneous privacy-risk preferences to estimate differentiated privacy compensation. Finally, data volume sensitivity experiments were designed to examine diminishing marginal gains in predictive utility, and predictive utility, aggregate privacy compensation, and platform operational costs were incorporated into a unified valuation function. In this function, aggregate privacy compensation and platform operational costs jointly form the cost baseline, while predictive utility determines a proportional utility premium above that baseline.

The principal contributions of this study are threefold. Firstly, an integrated data valuation

framework is proposed, in which the predictive utility, privacy risk, heterogeneous user preferences, and platform costs are considered. The framework connects the value generated from user data with the costs borne in producing and authorizing that value. Secondly, the privacy valuation mechanism is developed using SHAP-based feature contributions, thereby linking privacy compensation to the economic value generated by individual data attributes rather than treating predefined sensitivity classifications as the only valuation basis. The empirical results show that transactional attributes generate a substantially larger share of predictive value than the selected demographic attributes, accounting for approximately 81% of the calibrated aggregate privacy value. This indicates that privacy sensitivity and economic contribution are related but conceptually distinct dimensions. Thirdly, heterogeneous privacy-risk preferences are incorporated into the compensation mechanism to enable differentiated compensation for risk-seeking, risk-neutral, and risk-averse users. At the same time, the data-volume sensitivity analysis further demonstrates that indiscriminate data accumulation can increase privacy-compensation and operational costs without generating commensurate predictive benefits. This finding provides insights for platform enterprises seeking to build data collection practices that are both sustainable and equitable.

2 Literature review

Our research focuses on the valuation of data assets held by platform enterprises, specifically examining three interrelated areas: data valuation, data utility assessment in machine learning, privacy valuation, and compensation. Accordingly, each of these aspects is reviewed in detail below.

2.1 Data valuation

With the advent of the information age, data has experienced a big boom, and research on

data valuation is gradually emerging. Research, particularly in the areas of business and information systems, has been exploring the topic of data valuation (Hafner et al., 2025). Despite the increasing demand for data valuation, the field remains in its infancy, with several gaps between academic approaches and practical requirements (Li et al., 2019; Meierhofer et al., 2022). Data valuation research initially developed by extending conventional asset-valuation principles to data and information resources. Early studies commonly applied the cost, income, and market approaches used for tangible and intangible assets, treating data value as the cost of acquisition and processing, the discounted income expected from data use, or the price of comparable data products (Laney, 2017; Moody & Walsh, 1999). As data economics and business intelligence have advanced, increasing attention has been directed toward the distinctive value assessment to data (Xu et al., 2023).

Currently, data valuation approaches are identified in three types: market-based valuation, economic models, and dimension-based models (Fleckenstein et al., 2023). In the market value-based methods, this approach estimates the value of data in terms of cost and revenue when buying and selling data or data-intensive businesses, thereby reflecting the economic value of data. This method is often associated with data pricing methods. The query-based pricing framework is a query-oriented approach designed for data markets, where pricing is determined based on buyers' query demands for data in order to prevent arbitrage (Koutris et al., 2015). Heckman et al. (2015) proposed a valuation model for a dataset and explored the pricing mechanism of the data market. Economic models refer to an approach that estimates economic returns by providing data and measuring their impact. The economic dimension is one of the most popular dimensions used to determine the value of a data asset. Kim et al. (2023) proposed a data valuation model based on the discounted cash flow (DCF) method from the income approach. Farboodi et al. (2025) developed a methodology for valuing financial data using a sufficient statistic approach that exploits equilibrium asset return moments.

However, dimensional model-based data valuation is more complex. It can be conducted based on utility functions, emphasizing the alignment between user perception and actual needs. Tian et al. (2019) addressed the conflicts of interest between buyers and sellers in online data markets by constructing utility functions for data revenue and privacy cost respectively, and designed a trading mechanism aimed at maximizing the seller's utility. Chen et al. (2023) approached the issue from the perspective of data quality and proposed a comprehensive utility evaluation method that covers three dimensions: data volume, usability, and completeness.

In summary, the three methods discussed above represent the main approaches to valuing data as an economic asset. They are all based on a common premise: data is primarily regarded as a tradable or organizational asset. In platform settings, however, the individuals whose activities generate the fundamental data are rarely treated as participants in value creation and distribution. Their privacy costs may appear as compliance or acquisition expenses but are not connected to the decision value that platforms and partner enterprises derive from their data.

This limitation produces the utility–cost disconnect. These approaches can estimate the cost, revenue, or organizational value associated with data, but they provide limited guidance on how the predictive value captured by platforms should be related to the privacy costs borne by users.

2.2 Data utility assessment in machine learning

In machine learning scenarios, other specialized methods for data valuation have also been proposed. Understanding the value of specific data for model training has become essential for ensuring fair compensation to data owners (Baghcheband et al., 2025). As machine learning becomes the standard tool for extracting value from data, quantifying the contribution of individual data points or attributes has become a focal point of research. The Shapley value, a concept from cooperative game theory, has emerged as a principal method for this assessment

due to its axiomatic fairness in attributing payouts to players. Data Shapley values each training data point based on its marginal contribution to model performance, providing a robust framework for data valuation that is both model-agnostic and theoretically grounded (Ghorbani & Zou, 2019).

Building on this foundation, recent studies have sought to refine the application of Shapley values to improve computational efficiency and extend applicability. Han et al. (2026) proposed improvements to the valuation framework for vertical federated learning, addressing the challenge of valuing data contributions across distributed silos while preserving privacy. Their work highlights the necessity of aligning data valuation with actual predictive utility rather than volume or storage costs. Furthermore, the interpretability provided by Shapley values extends beyond valuation; it serves as a critical tool for explaining model predictions (Lundberg & Lee, 2017). This interpretability allows researchers to disentangle the contribution of specific features, which is essential for understanding the economic value embedded in different data attributes.

While these approaches successfully quantify predictive utility, they often treat data as a technical input, disconnected from the economic context of privacy loss and user compensation, thereby limiting their direct application in platform user data valuation. Therefore, a valuation framework for platform user data must preserve the explanatory advantages of predictive contribution while incorporating the costs and preferences associated with privacy.

2.3 Privacy valuation and compensation

The economic valuation of data is inextricably linked to the privacy costs borne by individuals who disclose their personal information. Unlike tangible assets, data involves a trade-off between the benefits of disclosure and the potential harm of privacy invasion. The foundational work by (Acquisti et al., 2016) delineated the economics of privacy, highlighting

the persistent difficulty of satisfying privacy expectations in digital environments. They noted that privacy concerns are not static; rather, they fluctuate based on the sensitivity of the information and the perceived benefits of the transaction.

More recent literature has attempted to quantify this privacy compensation mechanism, moving beyond simple sensitivity classifications. Lei et al. (2024) designed a privacy-preserving personalized pricing algorithm and demonstrated that privacy protection can coexist with near-optimal revenue performance. Furthermore, behavioral economics suggests that individual decisions regarding privacy are often influenced by risk attitudes. Drawing on prospect theory (Kahneman & Tversky, 1979), recent research has characterized privacy decision-making as a function of individual risk attitudes, with risk-seeking users demanding lower compensation for disclosure and risk-averse users requiring substantially higher premiums (Dai et al., 2024; Lu, 2022). These differences are not merely attitudinal; they have measurable economic consequences for platforms that seek to acquire user data at minimum cost. A compensation mechanism that treats all users identically will either overpay risk-seeking users, raising acquisition costs, or inadequately reward risk-averse users, reducing participation rates. Differentiated compensation schemes that account for heterogeneous risk preferences can therefore improve both the efficiency and equity of data acquisition.

Despite these advances, the integration of privacy valuation with utility-based data valuation remains neglected. Most of the privacy valuation literature treats compensation as a function of data sensitivity and individual preferences, without linking compensation amounts to the predictive contribution that data attributes make to platform business models (Rumbold & Pierscioneck, 2018; Skatova et al., 2023). Conversely, Shapley-based and utility-centered approaches to data valuation rarely translate their contribution measures into concrete compensation mechanisms that account for user heterogeneity. This produces the contribution–preference disconnect. Compensation based only on predictive contribution may fail to

adequately account for privacy risk, while compensation based only on sensitivity or stated preference may fail to reflect the value created for platforms and data consumers.

Therefore, this study proposes an integrated data valuation framework that combines predictive utility, privacy valuation, heterogeneous user preferences, and platform operational costs, thereby bridging the separation between data value creation and privacy compensation in existing literature.

3 Methods

3.1 Research design and framework

This study considers a data trading ecosystem comprising three types of actors: a single digital platform, multiple platform users, and partner enterprises.

- **Platform.** The platform acts as the central intermediary. It collects authorized user data, processes and integrates the records, evaluates their predictive utility, estimates privacy compensation, and incorporates storage and processing costs into the final valuation. It then provides the processed data products or analytical outputs to partner enterprises.
- **Platform user.** Platform users are the original data providers. They contribute personal and behavioral information generated during platform use and authorize the platform to use selected attributes. In return, they receive privacy compensation determined by the contribution of their data and their privacy-risk preferences.
- **Partner enterprise.** Partner enterprises are the consumers of the data product. They use processed data or analytical results to improve customer identification, demand forecasting, marketing decisions, and other business activities. Their willingness to pay depends on the expected decision value generated by the data.

The specific transaction process of the circulation and trading of user data is as follows, as illustrated in **Fig. 1**.

[Insert **Fig. 1** about here]

The primary objective of this study is to develop an integrated valuation method for platform-generated user data by jointly considering two dimensions of value: *data utility*, defined as the capacity of data to improve partner enterprises' predictions and decisions, and *user utility*, reflected in the compensation required for users' disclosure and authorized use of personal information.

The research design is structured around three analytical steps. First, data utility is quantified using the predictive performance of machine-learning models. Second, SHAP values are used to measure the importance of individual data attributes to the prediction outcome. The resulting importance scores are assigned as weights to the corresponding privacy levels, thereby estimating the privacy value factor embodied in the data provided by users at each privacy level. These factors are then combined with users' heterogeneous privacy-risk preferences to determine privacy compensation. Third, predictive utility, privacy compensation, and platform costs are incorporated into a unified data valuation model.

Through these steps, the study establishes a systematic procedure for converting user-generated platform data into an estimable product. This procedure sequentially quantifies data utility, derives privacy factors from attribute-level contributions, determines risk-adjusted compensation, and produces an integrated valuation that reconciles utility creation with the platform's cost burden.

3.2 Data utility quantification

To better reflect the practical context in which most platform enterprises operate (e.g., e-commerce, finance, and social networking), this study adopts predictive performance of

machine learning models as evaluation metrics. Data utility is the decision-supportive capacity of platform-generated user data. Its economic value is realized when the information embedded in the data improves predictions and reduces uncertainty surrounding business decisions of the platform and its partner enterprises.

Because platform data may serve different types of business decisions, this study evaluates predictive utility through both classification and regression tasks. Classification models assess whether the data can distinguish between users who exhibit a target behavior and those who do not, whereas regression models examine whether the data can predict the magnitude of a continuous outcome. To reduce dependence on the assumptions and limitations of any single model, multiple algorithms are compared within each task.

The F1 score is used as the principal measure of classification-based utility, while the coefficient of determination (R^2) represents regression-based utility. These two measures are subsequently integrated into a weighted composite index. In addition, the effect of data volume is examined by estimating model performance across incrementally larger training samples, thereby identifying how data utility changes with sample size and whether additional data yield diminishing marginal returns in terms of predictive gain.

3.2.1 Classification Modeling

The classification task was designed as a binary prediction problem: identifying whether a given user would make a purchase (positive class) or not (negative class). Four classification algorithms representing distinct modeling paradigms were selected for comparative evaluation:

- **Logistic Regression.** A linear classifier serving as the baseline model, configured with a maximum of 1,000 iterations and L2 regularization.

- Random Forest. An ensemble learning method based on bagged decision trees, configured with 100 trees and a maximum tree depth of 10 to balance model complexity and generalization performance.
- Gradient Boosting Decision Tree (GBDT). A sequential ensemble method, configured with 100 trees, a maximum depth of 5, and a learning rate of 0.1.
- XGBoost (eXtreme Gradient Boosting). A highly optimized gradient boosting implementation, configured with 100 trees, a maximum depth of 6, and a learning rate of 0.1.

Among them, XGBoost and Random Forest were identified as the primary algorithms of interest: XGBoost for its predictive precision and Random Forest for its robustness.

Each model was trained on the training set and evaluated on the held-out test set. Model performance was assessed using the following metrics derived from the confusion matrix:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

The Area Under the Receiver Operating Characteristic Curve (AUC) was also computed.

The F1-score was adopted as the primary evaluation metric for the classification task, as it provides a balanced measure that harmonizes precision and recall, thereby mitigating the risk of misleading assessments arising from class imbalance.

To evaluate generalization performance and model stability, five-fold stratified cross-validation was conducted. For each fold, the F1-score was computed on the validation subset, and the mean and standard deviation across the five folds were reported. Feature importance

was extracted directly from the tree-based models using their built-in impurity-based importance measures; for Logistic Regression, the absolute values of the learned coefficients were employed as a proxy for feature importance. All importance scores were normalized to the $[0,1]$ range to facilitate cross-model comparison. Pearson correlation coefficients were computed between each feature and the binary target variable, with corresponding p-values, to assess linear associations. Statistical significance was determined at $\alpha = 0.05$.

3.2.2 Regression Modeling

The regression task aimed to predict user purchase frequency, a continuous outcome variable, extending the analytical scope beyond binary classification to capture the intensity of user purchasing behavior. Prior to modeling, descriptive statistics—including the mean, standard deviation, median, skewness, and kurtosis—were computed, and the Shapiro–Wilk test was applied to assess the normality assumption of the target variable. The target variable exhibited significant deviation from normality ($p < 0.001$), informing the selection of non-parametric and tree-based modeling approaches.

Four representative regression models were selected for comparative evaluation: **Ridge Regression** as a regularized linear baseline; **Random Forest** as a bagging-based ensemble model; **GBDT** and **XGBoost** as boosting-based ensemble models. This selection enables a comparison between linear and nonlinear learning approaches while remaining consistent with the model settings adopted for the classification task. **Ridge Regression** was implemented with L_2 regularization; **Random Forest** was configured with 100 trees; **GBDT** was configured with 100 trees, a maximum depth of 5, and a learning rate of 0.1; **XGBoost** was configured with 100 trees, a maximum depth of 6, and a learning rate of 0.1.

All models were fitted on the training set and evaluated on the testing set using a comprehensive suite of regression metrics: Mean Absolute Error (MAE), Mean Squared Error

(MSE), Root Mean Square Error (RMSE), R^2 and Adjusted R^2 . The respective formulas are provided below:

$$MAE = \frac{1}{n} \sum_{t=1}^n |y_t - \hat{y}_t| \quad (5)$$

$$MSE = \frac{1}{n} \sum_{t=1}^n (y_t - \hat{y}_t)^2 \quad (6)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{t=1}^n (y_t - \hat{y}_t)^2} \quad (7)$$

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (8)$$

The coefficient of determination (R^2) measures the proportion of variation in the dependent variable explained by the regression model relative to a mean-based benchmark. An R^2 value of 1 indicates perfect predictive performance, whereas $R^2 = 0$ indicates that the model performs no better than predicting the mean of the dependent variable. A value between 0 and 1 represents partial predictive power, with a larger value indicating better model performance. A negative R^2 indicates that the model performs worse than the mean-based benchmark.

The prediction error distribution was further characterized by its mean, standard deviation, and 95% confidence interval. Error source decomposition—partitioning total prediction error into bias and variance components—was performed to diagnose model behavior. Five-fold cross-validation was conducted for each model, with the mean and standard deviation of R^2 across folds reported as indicators of generalization capability. Then, SHAP analysis was performed for Random Forest, GBDT, and XGBoost using the TreeExplainer algorithm on a randomly drawn subset of 200 test samples.

3.2.3 Weighted Composite Index of Data Utility

To integrate classification- and regression-based predictive performance into a unified measure, this study constructs a volume-dependent Weighted Composite Index (WCI) of data utility. Let $F1(n)$ denote the F1-score obtained from a classification model trained on n observations and let $R^2(n)$ denote the corresponding regression performance. Because a negative R^2 indicates that the model performs worse than a mean-only predictor, it is truncated at zero before aggregation. The resulting data utility measure is defined as:

$$W_{con}(n) = \alpha \cdot F1(n) + \beta \cdot R_{trunc}^2(n), \quad R_{trunc}^2 = \max(0, R^2) \quad (9)$$

subject to $\alpha + \beta = 1$, $\alpha, \beta \geq 0$.

Here, $W_{con}(n) \in [0,1]$ represents the overall predictive utility of a data product at sample size n . A value closer to 1 indicates that the data product provides strong predictive performance across the relevant decision tasks, whereas a value closer to 0 indicates limited predictive utility. Writing the index as a function of n also captures the fact that predictive utility changes with data volume.

The weight coefficients α and β reflect the relative importance of classification and regression tasks in the intended application. A larger α is appropriate when the data product is primarily used for categorical identification, such as purchase or fraud prediction, whereas a larger β is appropriate when continuous-outcome forecasting, such as purchase frequency or customer expenditure, is more important. Because the classification and regression tasks are treated as equally important in the present empirical setting, this study adopts $\alpha = \beta = 0.5$ as a neutral benchmark. If the data product supports only classification, the weights reduce to $(\alpha, \beta) = (1,0)$; if it supports only regression, they reduce to $(\alpha, \beta) = (0,1)$.

As n increases, both $F(n)$ and $R^2(n)$ may improve, but their marginal gains are expected to decline as model performance approaches a task-specific saturation level. The data-volume experiment therefore evaluates $W_{con}(n)$ across progressively larger training samples to determine whether additional data continues to generate meaningful predictive value.

3.3 Privacy value estimation and compensation

In the data trading scenario of platform enterprises, the authorization of user data fundamentally involves the concession of personal privacy. When their personal data are commercially exploited, they incur privacy costs. In this study, users authorize the platform to collect selected data attributes; the platform, in turn, converts these data into economic gains through predictive analytics. Privacy compensation thus functions as a monetary transfer that offsets the utility loss and perceived risk users bear for this authorization.

The proposed framework separates the privacy sensitivity of features (the type of personal information disclosed) from their predictive contribution, measured by model-based feature attribution. Then, these two dimensions are combined to derive the privacy value factors that underpin the compensation mechanism.

3.3.1 Privacy sensitivity classification

Data sensitivity is not uniform but heavily dependent on the context and the potential harm of disclosure. The privacy value of individual data attributes was anchored in a systematic privacy sensitivity classification framework. This privacy sensitivity classification framework draws upon existing research on privacy classification and is consistent with the spirit of global privacy paradigms, such as the General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA). This study categorizes sensitive data into 10 distinct levels, from L_0 to L_9 . For instance, L_0 represents public information (e.g., nicknames) with zero privacy content, while L_9 represents highly sensitive information such as personal diseases or sexual orientation. This framework assigns each data attribute a privacy sensitivity level reflecting the degree of personal intrusion associated with its disclosure. In contrast to conventional approaches that treat privacy sensitivity levels as directly determining

compensation weights, this study employs the sensitivity classification primarily as a categorical identifier and instead derives the quantitative privacy value of each attribute from its empirically measured contribution to predictive performance, as quantified by SHAP values. The specific classification of privacy levels is shown in **Table 1**.

Table 1 Statistical Classification Table of Sensitive Data Categories.

Level	Category	Exemplary Data Elements
L ₀	Public information	Internet-public data (nicknames, auto-assigned account IDs)
L ₁	Demographic information	Name, ethnicity, place of origin, gender
L ₂	Physical characteristics	Age, height, weight
L ₃	Preference information	Interests, shopping preferences, dietary preferences
L ₄	Financial information	ID documents, credit records, income, bank account numbers
L ₅	Contact information	Social media profiles, home/work addresses, email, phone numbers
L ₆	Static profile information	Family structure, occupation, education, social relationships
L ₇	Dynamic behavioral data	Cookies, location traces, online behavior, consumption records
L ₈	Precision data	Timestamps, precise geographic coordinates
L ₉	Sensitive personal data	Health conditions, sexual orientation

3.3.2 SHAP-Based feature contribution analysis

To objectively quantify the privacy value factor of each attribute, this study utilizes the SHAP (SHapley Additive exPlanations) value. Grounded in cooperative game theory, the SHAP value fairly allocates the contribution of each feature to the model's prediction. For the m^{th} data attribute of user i , its SHAP value is denoted as $\phi_{i,m}$. To obtain the overall utility at

the attribute level, the average absolute SHAP value across all N users is calculated:

$$\bar{\Phi}_m = \frac{1}{N} \sum_{i=1}^N |\Phi_{i,m}| \quad (10)$$

where N is the total sample size. The absolute value is taken to reflect the magnitude of contribution.

A larger $\bar{\Phi}_m$ indicates that the corresponding attribute exerts a higher average influence on model predictions and, consequently, generates greater incremental value for the platform's predictive applications. Consequently, the baseline privacy value V_m of the m^{th} attribute is formulated as:

$$V_m = \delta \cdot \bar{\Phi}_m \quad (11)$$

where $\delta > 0$ is a calibration coefficient that maps SHAP magnitudes to monetary units, ensuring that compensation amounts fall within feasible ranges. The expression formulation is as follows: $\delta = \frac{\text{Total Compensation Budget}}{\sum_m \Phi_m}$. This formulation satisfies two essential properties:

- (1) Attributes with larger predictive contributions receive proportionally higher privacy valuations.
- (2) The privacy values of all attributes are additive, facilitating aggregation at the individual user level.

3.3.3 Privacy compensation functions by risk attitude

Recognizing that users exhibit heterogeneous attitudes toward privacy risk, a heterogeneity that directly affects the cost to the platform of obtaining user data, this study classifies users into three risk-preference types, drawing on prospect theory (Kahneman & Tversky, 1979) and prior privacy compensation research (Dai et al., 2024; Lu, 2022). Three distinct compensation functions were designed, each mapping the degree of privacy disclosure θ to a compensation coefficient $C(\theta) \in [0,1]$:

- Risk-seeking users: These individuals are willing to bear risks and possess a higher risk tolerance. To incentivize this group, a sub-linear privacy compensation mechanism is designed:

$$C_1(\theta) = 1 - \frac{1}{1 + k_1\theta} \quad (12)$$

where $k_1 > 0$ controls the growth rate. This reflects the psychological trait that demands lower compensation for an initial privacy disclosure, and their compensation requirements do not rise sharply as the level of disclosure increases.

- Risk-averse users: These users are relatively conservative and unwilling to take risks. A logarithmic compensation mechanism is designed:

$$C_3(\theta) = \frac{\ln(k_3 * \theta + 1)}{L_{max}} \quad (13)$$

where $k_3 > 0$ determines the growth rate, and L_{max} is a normalization factor ensuring $C_3(\theta) \in [0,1]$. This function exhibits a steeper slope initially, reflecting their high sensitivity to minor privacy concessions.

- Risk-neutral users: Risk-neutral users are assumed to require an intermediate level of compensation between risk-seeking and risk-averse users. Their compensation coefficient is defined as the arithmetic mean of the two corresponding compensation functions:

$$C_2(\theta) = \frac{C_1(\theta) + C_3(\theta)}{2} \quad (14)$$

This function represents an intermediate compensation level by taking the arithmetic mean of the compensation required by risk-seeking and risk-averse users at each level of privacy disclosure.

Finally, for an individual user i with risk type C_i and authorized attribute set $\mathcal{M}_i$, the total privacy compensation is computed as:

$$F_i = C_i \cdot \sum_{m \in \mathcal{M}_i} V_m \quad (15)$$

3.4 Integrated data valuation model

A sustainable data trading ecosystem requires a rational pricing strategy that guarantees profitability for the platform while remaining acceptable to data consumers. If the final selling price of a data product falls below the total expenditure incurred during the data acquisition and processing phases, the platform will lack the economic incentive to facilitate the transaction. Therefore, the formulation of an integrated valuation model must systematically aggregate both the tangible operational costs and the intangible utility values.

The foundational baseline of data valuation consists of two primary cost components: hard costs and privacy compensation costs. First, the hard costs (denoted as P) encompass the tangible operational expenditures associated with data aggregation, storage, and algorithmic processing. These include the amortization of hardware infrastructure (e.g., server clusters and enterprise-grade storage arrays), cloud service fees, computational power consumption, and algorithmic maintenance. While the marginal replication cost of a data product is near zero, these initial fixed and variable processing expenses must be amortized and factored into the baseline valuation. Second, the platform must account for the aggregated privacy compensation paid to the data providers. Assuming a data product is generated from n distinct users, the total privacy compensation cost $F(n)$ is the sum of the individual compensation derived in the previous section:

$$F(n) = \sum_{i=1}^n F_i \quad (16)$$

The hard costs P and the aggregated privacy compensation $F(n)$ constitute the absolute cost floor of the data product.

However, the intrinsic commercial value of the data product transcends its mere production cost. It is fundamentally driven by the decision-empowerment capabilities it provides to partner enterprises. Therefore, this study introduces the concept of a *utility premium*. The volume-constrained data utility W_{con} serves as a multiplier that scales the baseline costs. The higher the predictive accuracy of the data product, the greater the utility premium it commands in the market.

By integrating these dimensions, the final integrated data valuation function G is formally defined as:

$$G = (1 + W_{con}) \times (F_n + P) \quad (17)$$

In this function, $F_n + P$ represents the cost baseline of the data product, consisting of aggregate privacy compensation and platform operational costs. The constant term 1 ensures that when predictive utility is zero ($W_{con} = 0$), the estimated value remains equal to the cost baseline rather than falling to zero. The term W_{con} represents the utility premium generated by the predictive performance of the data. Therefore, a higher level of predictive utility proportionally increases the platform-side valuation above its cost baseline. This function represents the platform's valuation. If the pricing falls below this value, selling data would incur a loss, thereby deterring the platform from proceeding with the transaction. Only when the valuation exceeds this sum is there room for profitability, which simultaneously safeguards user privacy rights and ensures the platform's operational sustainability.

4 Experimental evaluation

4.1 Dataset

This study utilizes a dataset derived from the Alibaba Tianchi platform, comprising partial user characteristic data from the Douyin e-commerce ecosystem in 2024. The dataset contains

1,000 valid samples and 15 attributes, capturing a comprehensive view of user behaviors. The features encompass basic demographic information (e.g., age, gender, location, income), platform interaction metrics (e.g., time spent on site, pages viewed, last login days), and transactional records (e.g., average order value, total spending, product category preference).

These attributes contain all privacy sensitivity levels from L_0 to L_8 in the ten-level classification framework established in Section 3.3.1, except for the L_9 level (which includes sensitive personal data), because the platform rarely has legitimate access to such deeply personal information. A comprehensive mapping of each attribute to its corresponding privacy sensitivity level is presented in **Table 2**.

Table 2 Table of Privacy Levels Corresponding to Attributes of Data Set.

Attribute	Description	Data Type	Privacy Level
User ID	User's nickname on the platform	String	L_0
Age	User's age	Numeric	L_2
Gender	User's gender (Male/Female)	String	L_1
Location	User's location (Urban/Rural)	String	L_5
Income	User's income level	Numeric	L_4
Interests	User's interests (e.g., sports, fashion, technology)	String	L_6
LAST Days Ago	Days since last login	Numeric	L_8
Average Order Value	Average order value	Numeric	L_7
Total Spending	Total amount spent by the user	Numeric	L_7
Product Category Preference	Preferred product category	String	L_3

Time Spent on Site	Time spent on the site while logged in minutes	Numeric	L ₈
Minutes			
Pages Viewed	Number of pages viewed during the session	Numeric	L ₇
Newsletter	Whether the user subscribed to marketing		
Subscription	notifications (TRUE/FALSE)	String	L ₇

4.2 Data preprocessing and benchmark settings

The primary objective is to predict user purchasing behavior, defined as a binary classification task (purchase vs. non-purchase), and user purchase frequency, defined as a regression task. To ensure data quality and enhance model expressiveness, preliminary data preprocessing was conducted. This included the removal of identifiers (e.g., User ID), label encoding for categorical variables, and standardization of numerical features to eliminate scale effects. Furthermore, based on domain knowledge, three interactive features were engineered: **Spending Per Minute, Order Value to Income Ratio, and Pages Per Minute.**

The dataset was randomly partitioned into an 80% training set (800 samples) and a 20% testing set (200 samples), with stratified sampling applied in the regression task to maintain target variable distribution consistency.

To ensure comprehensive and fair benchmark, representative models from various learning paradigms were selected for both classification and regression tasks. For the classification task, Logistic Regression, Random Forest, GBDT, and XGBoost were employed. For the regression task, Ridge Regression, Random Forest, GBDT, and XGBoost were similarly selected. The specific hyperparameter configurations are detailed in Section 3.2. All models were trained on the training set, with training time recorded, and evaluated on the testing set. To assess generalization capability, five-fold stratified cross-validation was adopted to compute the mean and standard deviation of the F1-score (for classification) and R^2 (for regression).

4.3 Data volume sensitivity experiment

To investigate the relationship between data volume and predictive utility, and to test the diminishing marginal returns of data utility, a systematic volume sensitivity experiment was conducted. Holding the test set constant ($n = 200$), training subsets of progressively increasing size were randomly sampled from the full training set ($n = 800$) at the following eight levels: 16, 40, 80, 160, 320, 480, 640, and 800. These levels span the full range from 2% to 100% of the training set, enabling observation of model behavior across the complete data availability spectrum.

At each volume level, all four models were independently trained and evaluated on the fixed test set. The coefficient of determination (R^2) was recorded as the primary outcome metric.

5 Results

5.1 Performance of classification model

This subsection systematically evaluates the predictive accuracy of four classification algorithms on the binary purchase prediction task. Experiments were conducted under a stratified 80:20 train-test split, and all models were evaluated using five-fold cross-validation. The detailed classification results for each model are presented in **Table 3**.

Table 3 Detailed performance results of each model.

Model	Accuracy	Precision	Recall	F1-score	AUC	TP	FP	TN	FN	CV_F1 Mean	CV_F1 Std
XGboost	0.98	0.977	0.992	0.984	0.998	130	3	66	1	0.983	0.007
Logistic Regression	0.975	0.963	1.000	0.981	0.997	131	5	64	0	0.967	0.013
GBDT	0.965	0.955	0.992	0.973	0.996	130	6	63	1	0.981	0.010

Model	Accuracy	Precision	Recall	F1-score	AUC	TP	FP	TN	FN	CV_F1 Mean	CV_F1 Std
Random Forest	0.945	0.941	0.977	0.958	0.990	128	8	61	3	0.955	0.017

Table 3 presents the detailed classification performance of the four evaluated models. XGBoost achieved the best overall performance ($F1 = 0.984$, $Accuracy = 0.980$, $Precision = 0.977$, $Recall = 0.992$, $AUC = 0.992$). Five-fold cross-validation confirmed the stability of this result, with a mean F1-score of 0.983 ± 0.007 . Logistic Regression followed closely with $F1 = 0.981$ and $Recall = 1.000$, though its relatively lower precision (0.963) suggests a slight susceptibility to false positives. GBDT ($F1 = 0.973$) and Random Forest ($F1 = 0.958$) also delivered strong performance, with all four models $F1 > 0.95$. These results indicate that the ensemble tree-based models, particularly XGBoost, are highly effective at extracting predictive signals from platform user behavioral data.

To assess the importance of every feature, we computed the correlation coefficients for all features. **Table 4** presents the distribution of feature importance across the 15 features and different models.

Table 4 Correlation coefficient and importance of each feature under different algorithms.

Feature	Correlation Coefficient	XGBoost Feature Importance (Normalized)	Logistic Regression Feature Importance (Normalized)	Random Forest Feature Importance (Normalized)	GBDT Feature Importance (Normalized)
Average Order Value	-0.596	0.350	0.673	0.3580	0.6700
Total Spending	-0.213	0.117	0.183	0.0728	0.1716
Interests	-0.167	0.029	0.017	0.0965	0.0057
Spending Per Minute	-0.075	0.048	0.004	0.0421	0.0339
Age	-0.067	0.011	0.012	0.0204	0.0053
Order Value to Income Ratio	-0.053	0.334	0.035	0.2543	0.1071
Income	0.044	0.012	0.0157	0.0446	0.0024
Gender	0.033	0.000	0.0082	0.0030	0.0002
Time Spent on Site Minutes	0.0325	0.015	0.0047	0.0248	0.0007
Las Login Days Ago	-0.031	0.0049	0.0044	0.0184	0.0006
Pages Per Minute	0.0188	0.0356	0.0214	0.0219	0.0020
Location	-0.0187	0.0054	0.0096	0.0106	0.0004
Newsletter Subscription	0.0186	0.0099	0.0024	0.0034	0.0009
Product Category Preference	0.0071	0.0071	0.0040	0.0109	0.0009
Pages Viewed	-0.004	0.0184	0.0031	0.0172	0.0003

Table 4 reports the Pearson correlation coefficients between each feature and the binary purchase target, along with the normalized feature importance scores derived from all four models. **Average Order Value** emerged consistently as the most influential predictor across all algorithms, with normalized importance scores of 0.350 (XGBoost), 0.673 (Logistic Regression), 0.358 (Random Forest), and 0.67 (GBDT). **Total Spending** ranked second, while **Order Value to Income Ratio** also demonstrated notable contributions. The analysis revealed a significant negative correlation between **Average Order Value** and purchase behavior ($r = -0.5968, p < 0.001$), indicating that lower average order values were associated with higher purchase likelihood in this dataset. **Gender** and **Last Login Days Ago** exhibited negligible

correlations with the target variable, suggesting limited standalone predictive power for these features.

5.2 Performance of regression model

As shown in **Table 4**, the importance of each feature is fully reported, which also reflects its individual data utility. To further improve the accuracy and completeness of the findings, this study introduces the coefficient of determination (R^2) to achieve a more precise prediction and fitting of the dependent variable *purchase frequency*, and to better quantify the overall data utility of the dataset. The detailed results are provided in **Table 5**.

Table 5 Detailed performance results of each model.

Model	MAE	MSE	RMSE	R^2	Adj. R^2
GBDT	0.5666	0.9354	0.9671	0.9669	0.9642
XGboost	0.6466	1.1921	1.0918	0.9578	0.9544
Random Forest	0.8192	1.9433	1.3940	0.9313	0.9257
Ridge	3.4487	17.7402	4.2119	0.3729	0.3218
Regression					

Table 5 presents the regression performance of all evaluated models. The GBDT model achieved the best overall results ($R^2 = 0.9669$, $Adj. R^2 = 0.9642$, $RMSE = 0.9671$, $MAE = 0.5666$). Its strong generalization capacity was confirmed through five-fold cross-validation, which yielded a mean R^2 of 0.9368 ± 0.0140 . XGBoost ranked second ($R^2 = 0.9578$), followed by Random Forest at 0.9313. In contrast, Ridge Regression attained only $R^2 = 0.3729$, demonstrating that the relationship between user features and purchase frequency is predominantly nonlinear and that linear models are inadequate for this prediction task. GBDT demonstrated the best performance across all metrics, with an MAE of only 0.5666, indicating the smallest prediction error. XGBoost and Random Forest ranked next, and the linear model exhibited the worst performance.

These results suggest that, for the purchase frequency regression prediction task in the present study, the GBDT or XGBoost model should be prioritized to ensure the accuracy and reliability of the predictions.

5.3 SHAP analysis

To further explain the model's decision-making mechanism and clarify the direction and degree of each feature's influence on the model output, SHapley Additive exPlanations (SHAP) was employed. SHAP analysis was conducted on the three models (GBDT, RF, and XGboost) using a random subset of 200 test samples. The specific results are illustrated in **Fig. 2** and **Table 6**.

[Insert **Fig. 2** about here]

Fig. 2 presents the direction and magnitude of each feature's contribution to the outputs of different models. In this figure, the color gradient indicates feature values, with red representing higher values and blue representing lower values. In terms of feature importance, **Average Order Value** and **Total Spending** emerge as the two most critical variables influencing the predictive outcomes. Furthermore, a closer examination of the **Average Order Value** distribution reveals that its high values (red) do not unilaterally contribute to positive predictions. Although a subset of high-value samples is positioned on the right side with positive SHAP values, a substantial proportion of red samples clusters on the left, corresponding to negative SHAP values. This suggests that within our study population, an excessively high average order value is conversely associated with a decrease in the model's predicted target. Moreover, the feature importance rankings demonstrate a high degree of consistency across the three evaluated models. The corresponding numerical values are detailed in **Table 6**.

Table 6 Average absolute SHAP value of each feature.

Feature	XGBoost	Random Forest	GBDT
Average Order Value	3.9555	3.8878	4.0806
Total Spending	1.8011	1.3821	2.0217
Interests	0.4247	0.5865	0.3319
Order Value to Income Ratio	0.2955	0.3962	0.2877
Spending Per Minute	0.2938	0.4065	0.2493
Income	0.0497	0.0284	0.0500
Time Spent on Site Minutes	0.0403	0.0438	0.0333
Pages Viewed	0.0402	0.0301	0.0113
Age	0.0240	0.0239	0.0125
Product Category Preference	0.0240	0.0122	0.0186
Pages Per Minute	0.0193	0.0232	0.0165
Newsletter Subscription	0.0174	0.0056	0.0170
Last Login Days Ago	0.0162	0.0170	0.0125
Location	0.0123	0.0098	0.0052
Gender	0.0023	0.0070	0.0021

Table 6 reports the mean absolute SHAP values of all 15 features across the three models, providing a quantitative ranking of feature contributions. These values are used in the subsequent evaluation to calculate the privacy value.

5.4 Data volume analysis

To systematically evaluate the effect of data volume on predictive utility, a data volume sensitivity experiment was conducted. Four regression models were independently trained on progressively larger training subsets, with sample sizes ranging from 16 to 800, while the test set was kept fixed. Model performance was assessed using the coefficient of determination (R^2). The detailed numerical results across different data volume levels are summarized in

Table 7, while the visual trajectories of model performance variations are illustrated in **Fig. 3**.

Table 7 Model performance (R^2) across different training data volumes.

Sample sizes	XGBoost	GBDT	RF	Ridge
16	0.5903	0.5918	0.4214	-0.6921
40	0.4404	0.5051	0.5106	0.0961
80	0.5699	0.6334	0.632	0.1606
160	0.6649	0.7194	0.7076	0.2692
320	0.8518	0.8888	0.7993	0.3490
480	0.895	0.9169	0.8814	0.3852
640	0.9537	0.9544	0.9172	0.3722
800	0.9579	0.9669	0.9352	0.3729

The results confirm that when training sample size increased, the predictive performance of all models rose and gradually stabilized. At a sample size of 16, the R^2 value for GBDT and XGBoost reached 0.5918 and 0.5903, respectively, indicating their ability to extract valuable predictive signals from limited data. Conversely, the Ridge Regression demonstrated severe instability under small sample condition and plateaued at a low performance (0.3729) even after sufficient training.

[Insert **Fig. 3** about here]

As illustrated in **Fig. 3**, with the initial increase in data volume, the curves of all models exhibit an upward trend that subsequently levels off. This illustrates that the performance improvement of the models adheres to diminishing marginal utility. For XGBoost, increasing the sample size from 16 to 160 yielded a substantial R^2 gain (from 0.590 to 0.665), whereas the increase from 640 to 800 produced a negligible improvement (from 0.954 to 0.958). This suggests that beyond a specific data threshold (about 40% to 60% of the complete training set),

the marginal value of additional data starts to decline.

5.5 Privacy value estimation

The privacy values obtained from the SHAP analysis and the corresponding compensation estimates across the three risk-preference types are presented below.

Based on previous literature (Dai et al., 2024; Lu, 2022) , this study categorizes users into three risk preference types: risk-seeking, risk-neutral, and risk-averse, with proportions of 30%, 40%, and 30%, respectively. Given a platform user base of $n = 1000$ users, the aggregate privacy compensation budget under the specified risk-type distribution amounted to $F_n = 5240$ CNY.

The privacy value of each attribute was computed as $V_m = \delta \cdot \bar{\phi}_m$, where $\delta = \frac{\text{Total Compensation Budget}}{\sum_m \phi_m} = 0.7476$ is determined by the platform's total privacy compensation budget relative to the aggregate SHAP contributions across all attributes.

Table 8 Mean absolute SHAP values ($\bar{\phi}_m$) and corresponding privacy values (V_m) of each feature.

Feature	$\bar{\phi}_m$	V_m (CNY)
Average Order Value	3.9746	2.971
Total Spending	1.7350	1.297
Interests	0.4477	0.335
Order Value to Income Ratio	0.3265	0.244
Spending Per Minute	0.3165	0.237
Income	0.0427	0.032
Time Spent on Site Minutes	0.0391	0.029
Pages Viewed	0.0272	0.020

Feature	$\bar{\phi}_m$	V_m (CNY)
Age	0.0201	0.015
Product Category Preference	0.0183	0.014
Pages Per Minute	0.0197	0.015
Newsletter Subscription	0.0133	0.010
Last Login Days Ago	0.0152	0.011
Location	0.0091	0.007
Gender	0.0038	0.003
Sum	7.0088	5.240

Table 8 reports the mean absolute SHAP values ($\bar{\phi}_m$) for GBDT, XGBoost, and Random Forest, along with their corresponding privacy values (V_m).

The results demonstrate significant heterogeneity in attribute-level privacy values. Transactional attributes, namely **Average Order Value** (2.971 CNY) and **Total Spending** (1.297 CNY), hold the highest privacy values, jointly accounting for approximately 81% of the total user privacy value. Conversely, demographic and behavioral features like **Gender** (0.003 CNY), **Location** (0.007 CNY), and **Last Login Day Ago** (0.011 CNY) yield negligible contributions, consistent with their marginal SHAP importance detailed in **Table 8**. This distribution indicates that transactional features primarily drive model predictions and the platform’s data-derived business value. Overall, the total aggregate privacy value per user across all 15 features is $\sum V_m = 5.240$ CNY.

[Insert **Fig. 4** about here]

The **Fig. 4** illustrates the changes in compensation for the three risk attitude types across varying levels of privacy disclosure (θ). The compensation for risk-averse users is relatively high, with even slight privacy disclosure resulting in significant compensation. Risk-seeking users receive lower compensation, but the disparity between them and risk-averse users

diminishes at higher disclosure levels, with risk-neutral users situated in between. It can be observed that as privacy disclosure increases, the platform's compensation for the three user types tends to converge, consistent with theoretical expectations.

5.6 Integrated data valuation

Applying the preceding empirical results to the integrated data valuation function G , as shown in equation(17). The Weighted Composite Index was computed as $W_{con}(n) = \alpha \cdot F1(n) + \beta \cdot R_{trunc}^2(n)$ with equal weights ($\alpha = \beta = 0.5$), using the models with optimal performance: XGBoost for classification ($F = 0.984$) and GBDT for regression ($R^2 = 0.9669$), resulting in $W_{con} = 0.975$. The total privacy compensation, estimated in Section 5.5, is $F_n = 5240$ CNY, and the fixed cost is $P = 2000$ CNY $P = 2,000$ RMB. Therefore, we obtain $G = (1 + W_{con}) \times (F_n + P) = 14299$ CNY . The utility premium ($1 + W_{con}$) magnifies the cost baseline by a factor of 1.975, thereby capturing the decision-empowerment value that the data product delivers to partner enterprises.

These results confirm that the proposed framework successfully synthesizes data utility and privacy valuation into a unified, operationally tractable data valuation.

6 Discussion

This study develops an integrated valuation framework that transforms platform-generated user data into an economically interpretable asset by combining machine learning-based predictive utility, attribute-level privacy contributions, and heterogeneous user risk preferences. Our findings demonstrate, first, that the economic value of user data concentrates in transactional behavior rather than in nominally sensitive demographic attributes. This finding separates privacy sensitivity from economic value rather than treating the two as equivalent. Second, predictive utility follows a clear diminishing-returns trajectory with respect

to sample size, with direct consequences for how platforms should time and scale data acquisition. Third, risk-differentiated compensation functions ground the pricing mechanism in individual heterogeneity. These results support the integrated valuation function G defined in equation(17), which links predictive contribution, privacy compensation, and operational cost within one pricing rule.

6.1 Predictive contribution over privacy sensitivity classification

The observed dominance of ensemble learners, particularly XGBoost ($F1 = 0.984$) and GBDT ($R^2 = 0.9669$) as reported in **Table 3** and **Table 5**, indicates that platform user behavior is governed by threshold effects and feature interactions rather than by additive, linear relationships. Tree-based boosting methods partition the feature space according to exactly this kind of interaction structure, which linear estimators cannot represent (T. Chen & Guestrin, 2016), and this explains why Logistic Regression and Ridge Regression trail the ensemble models throughout both tasks. The SHAP analysis, which is visualized in **Fig. 2** and quantified in **Table 6**, revealed that transactional features (Average Order Value and Total Spending) exhibit the highest mean absolute SHAP value. As a direct economic consequence, these two behavioral attributes account for approximately 81% of the total aggregate privacy value, as shown in **Table 8**. Conversely, highly sensitive demographic features such as Gender (L_1) and Location (L_5) make minor contributions, with privacy values of merely 0.003 CNY and 0.007 CNY, respectively. This result overcomes limitations of traditional privacy valuation research (Acquisti et al., 2016), which typically assumes that higher data sensitivity directly correlates with higher monetary value. Instead, value tracks predictive contribution, which is a function of behavioral variance rather than of data-type classification. This distinction has a direct implication for compensation design: attribute-level, contribution-based compensation avoids both the underpayment of highly predictive but nominally low-sensitivity attributes and the

overpayment of sensitive but low-contribution ones, a distinction that fixed sensitivity-tier compensation schemes cannot draw. The result also demonstrates that stated privacy concern does not reliably predict actual disclosure behavior (Norberg et al., 2007) : because the proposed framework prices attributes according to their demonstrated predictive contribution rather than by elicited or assumed sensitivity, it avoids the measurement problem that self-reported privacy preferences introduce.

6.2 Balancing Diminishing Predictive Returns with Loss Aversion

To validate the economic principle of diminishing marginal returns in predictive modeling, this study conducted sensitivity experiments on data volume. The results show that predictive performance rises steeply at small sample sizes and flattens beyond 40% to 60% of the training set. When the model approaches its effective capacity on a specific task, additional training data will result in diminishing marginal improvements in accuracy (Sun et al., 2017). This implies that the utility premium a platform can command from a data product is bounded: beyond the saturation threshold, acquiring additional users increases privacy-compensation liability without a matching increase in predictive utility, so marginal profitability declines. Therefore, for platform enterprises, this dictates that aggressive, boundless data acquisition is economically inefficient. Platforms must balance the marginal cost of acquiring and storing additional user data against the diminishing predictive returns, thereby optimizing data procurement strategies to operate near the identified saturation threshold. Furthermore, incorporating heterogeneous user risk preferences into the compensation mechanism introduces a behavioral dimension to data valuation. The divergence in compensation curves between risk-seeking and risk-averse users confirms the principles of prospect theory (Kahneman & Tversky, 1979), according to which users exhibit loss aversion toward initial privacy concessions. The observed convergence of compensation requirements at higher

disclosure levels suggests a psychological threshold where the marginal cost of additional privacy loss becomes uniform across user segments. Platforms can leverage this convergence by designing bundled data authorization schemes, encouraging comprehensive data disclosure in a single transaction to minimize aggregate compensation payouts.

7 Conclusion

This study proposes an integrated valuation framework for platform-generated user data that jointly considers predictive utility, attribute-level privacy value, heterogeneous privacy-risk preferences, and platform operational costs. In the resulting valuation function, privacy compensation and platform operational costs establish the cost baseline, while predictive utility determines the proportional premium above that baseline. By combining classification- and regression-based utility assessment with SHAP-based feature attribution, the framework translates the decision-support capacity of data into an economically interpretable measure and incorporates the compensation required for the authorized use of personal information. The empirical analysis demonstrates that this integrated approach can distinguish the heterogeneous value embedded in different data attributes and provide a more comprehensive valuation basis than methods relying solely on data volume, production cost, or predefined privacy sensitivity levels. First, the economic value of user data depends more on its contribution to predictive performance than on its nominal privacy category. Transactional and behavioral attributes often contribute more to model predictions than demographic attributes, even when the latter are highly sensitive. In our experiments, transactional attributes account for approximately 81% of the calibrated aggregate privacy value. This shows that privacy sensitivity and economic contribution are related but distinct dimensions: the former reflects disclosure risk, while the latter captures decision value. Second, data volume affects value through diminishing returns. Increasing sample size initially improves predictive performance, but gains decline as data

approaches task requirements. Beyond this point, additional data raises costs without comparable benefits, suggesting that data acquisition should be guided by marginal utility rather than scale alone. Third, incorporating heterogeneous privacy-risk preferences enables a more realistic compensation mechanism. SHAP values measure attribute contributions, while risk preferences determine compensation levels. Their integration allows differentiated compensation for risk-seeking, risk-neutral, and risk-averse users, reflecting that users perceive privacy loss differently. Our framework not only provides practical guidance for platform enterprises, partner enterprises, and data governance authorities but also enhances the alignment between data value, privacy protection, and compensation mechanisms.

Moreover, extending the analysis to a broader range of industries, application scenarios, and data environments would further deepen our understanding of how attribute-level contributions and valuation structures vary across contexts. The integration of complementary explainability methods may also provide additional insights into attribute-level value assessment. Future research could further extend the proposed framework to different transaction settings and market mechanisms, thereby helping to establish a clearer connection between estimated data value and realized exchange outcomes in more complex data market environments.

References

- Acquisti, A., Taylor, C., & Wagman, L. (2016). The Economics of Privacy. *Journal of Economic Literature*, 54(2), 442–492. <https://doi.org/10.1257/jel.54.2.442>
- Arrieta-Ibarra, I., Goff, L., Jiménez-Hernández, D., Lanier, J., & Weyl, E. G. (2018). Should We Treat Data as Labor? Moving Beyond “Free.” *AEA Papers and Proceedings*, 108, 38–42. <https://doi.org/10.1257/pandp.20181003>
- Baghcheband, H., Soares, C., & Reis, L. P. (2025). Shapley value-based data valuation for machine learning data markets. *Discover Applied Sciences*, 7(12), 1431. <https://doi.org/10.1007/s42452-025-07328-z>

- Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. <https://doi.org/10.1145/2939672.2939785>
- Chen, Y., Bai, R., Wu, Y., Li, T., & Zhou, H. (2023). A Multidimensional Data Utility Evaluation and Pricing Scheme in the Big Data Market. *Wireless Communications and Mobile Computing*, 2023, 1–12. <https://doi.org/10.1155/2023/6217495>
- Cheong, H., Kim, B., & Vaquero, I. U. (2023). A Data Valuation Model to Estimate the Investment Value of Platform Companies: Based on Discounted Cash Flow. *Journal of Risk and Financial Management*, 16(6), 293. <https://doi.org/10.3390/jrfm16060293>
- Couldry, N., & Mejias, U. A. (2019). Data Colonialism: Rethinking Big Data’s Relation to the Contemporary Subject. *Television & New Media*, 20(4), 336–349. <https://doi.org/10.1177/1527476418796632>
- Dai, J., Li, S., Liao, Z., & Zhu, Y. (2024). Research on Data Pricing Model Based on Privacy Compensation. *Management And Administration*, 2, 10–18. <https://doi.org/10.16517/j.cnki.cn12-1034/f.20240919.001>
- Fainmesser, I. P., Galeotti, A., & Momot, R. (2023). Digital Privacy. *Management Science*, 69(6), 3157–3173. <https://doi.org/10.1287/mnsc.2022.4513>
- Farboodi, M., Singal, D., Veldkamp, L., & Venkateswaran, V. (2025). Valuing Financial Data. *The Review of Financial Studies*, 38(3), 938–980. <https://doi.org/10.1093/rfs/hhae034>
- Fleckenstein, M., Obaidi, A., & Tryfona, N. (2023). A Review of Data Valuation Approaches and Building and Scoring a Data Valuation Model. *Harvard Data Science Review*, 5(1). <https://doi.org/10.1162/99608f92.c18db966>
- Gao, H., & Zhang, J. (2024). Evaluation of the value of Internet platform Enterprises in the digital economy by the big data cooperation asset valuation model. *Heliyon*, 10(14), e34248. <https://doi.org/10.1016/j.heliyon.2024.e34248>

- Ghorbani, A., & Zou, J. (2019). Data Shapley: Equitable Valuation of Data for Machine Learning. *Proceedings of the 36th International Conference on Machine Learning*, 2242–2251. <https://proceedings.mlr.press/v97/ghorbani19c.html>
- Hafner, M., Mira Da Silva, M., & Proper, H. A. (2025). Data valuation as a business capability: From research to practice. *Information Systems and E-Business Management*, 23(3), 745–784. <https://doi.org/10.1007/s10257-025-00701-w>
- Han, X., Wang, L., Wu, J., & Fang, X. (2026). Data Valuation for Vertical Federated Learning: A Model-Free and Privacy-Preserving Method. *MIS Quarterly*, 50(1), 177–210. <https://doi.org/10.25300/MISQ/2025/19161>
- Heckman, J., Peters, E., Kurup, N. G., Boehmer, E., & Davaloo, M. (2015). *A Pricing Model for Data Markets*.
- Kahneman, D., & Tversky, A. (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, 47(2), 263. <https://doi.org/10.2307/1914185>
- Koutris, P., Upadhyaya, P., Balazinska, M., Howe, B., & Suciu, D. (2015). Query-Based Data Pricing. *Journal of the ACM*, 62(5), 1–44. <https://doi.org/10.1145/2770870>
- Laney, D. B. (2017). *Infonomics: How to Monetize, Manage, and Measure Information as an Asset for Competitive Advantage* (1st ed.). Routledge. <https://doi.org/10.4324/9781315108650>
- Lei, Y. (Murray), Miao, S., & Momot, R. (2024). Privacy-Preserving Personalized Revenue Management. *Management Science*, 70(7), 4875–4892. <https://doi.org/10.1287/mnsc.2023.4925>
- Li, Z., Ni, Y., Gao, X., & Cai, G. (2019). Value Evaluation of Data Assets: Progress and Enlightenment. *2019 IEEE 4th International Conference on Big Data Analytics (ICBDA)*, 88–93. <https://doi.org/10.1109/ICBDA.2019.8713240>
- Lu, K. (2022). *Research on Dynamic Pricing of Personal Data Assets Based on Privacy Elements* [Master dissertation, Zhongnan University of Economics and Law].

<https://link.cnki.net/doi/10.27660/d.cnki.gzczu.2022.002591>

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17*, 4768–4777. <https://dl.acm.org/doi/10.5555/3295222.3295230>

Mazzarello, M., Duarte, F., Mora, S., Heine, C., & Ratti, C. (2025). Data Slots: Trade-offs between privacy concerns and benefits of data-driven solutions. *Humanities and Social Sciences Communications*, 12(1), 643. <https://doi.org/10.1057/s41599-025-04776-1>

Meierhofer, J., Benedech, R., Schweiger, L., Barbieri, C., & Rapaccini, M. (2022). Quantitative Modelling of the Value of Data for Manufacturing SMEs in Smart Service Provision. *ITM Web of Conferences*, 41, 04001. <https://doi.org/10.1051/itmconf/20224104001>

Moody, D., & Walsh, P. (1999). *Measuring The Value Of Information: An Asset Valuation Approach*.

Norberg, P. A., Horne, D. R., & Horne, D. A. (2007). The Privacy Paradox: Personal Information Disclosure Intentions versus Behaviors. *Journal of Consumer Affairs*, 41(1), 100–126. <https://doi.org/10.1111/j.1745-6606.2006.00070.x>

Peng, H., Miao, X., Zhang, J., Gao, Y., Deng, S., & Yin, J. (2025). Cost-aware prediction service pricing with incomplete information. *The VLDB Journal*, 34(3), 28. <https://doi.org/10.1007/s00778-025-00909-9>

Rumbold, J. M. M., & Pierscioneck, B. K. (2018). What Are Data? A Categorization of the Data Sensitivity Spectrum. *Big Data Research*, 12, 49–59. <https://doi.org/10.1016/j.bdr.2017.11.001>

Skatova, A., McDonald, R., Ma, S., & Maple, C. (2023). Unpacking privacy: Valuation of personal data protection. *PLOS ONE*, 18(5), e0284581. <https://doi.org/10.1371/journal.pone.0284581>

Sun, C., Shrivastava, A., Singh, S., & Gupta, A. (2017). Revisiting Unreasonable Effectiveness of Data in Deep Learning Era. *2017 IEEE International Conference on Computer Vision*

(ICCV), 843–852. <https://doi.org/10.1109/ICCV.2017.97>

Tang, Y., Wan, Y., & Li, J. (2026). Improving the accuracy and stability of privacy-aware Shapley data valuation. *Discover Applied Sciences*, 8(7), 750. <https://doi.org/10.1007/s42452-026-08711-0>

Tian, L., Li, J., Li, W., Ramesh, B., & Cai, Z. (2019). Optimal Contract-Based Mechanisms for Online Data Trading Markets. *IEEE Internet of Things Journal*, 6(5), 7800–7810. <https://doi.org/10.1109/IIOT.2019.2902528>

Veldkamp, L. (2023). Valuing Data as an Asset. *Review of Finance*, 27(5), 1545–1562. <https://doi.org/10.1093/rof/rfac073>

Voloch, N., & Hirschprung, R. S. (2026). The value of leaked data in online social networks. *Computers & Security*, 168, 104937. <https://doi.org/10.1016/j.cose.2026.104937>

Xu, J., Hong, N., Xu, Z., Zhao, Z., Wu, C., Kuang, K., Wang, J., Zhu, M., Zhou, J., Ren, K., Yang, X., Lu, C., Pei, J., & Shum, H. (2023). Data-Driven Learning for Data Rights, Data Pricing, and Privacy Computing. *Engineering*, 25, 66–76. <https://doi.org/10.1016/j.eng.2022.12.008>

Zuboff, S. (2015). Big other: Surveillance Capitalism and the Prospects of an Information Civilization. *Journal of Information Technology*, 30(1), 75–89. <https://doi.org/10.1057/jit.2015.5>

Figure legends

Fig. 1 Basic Logic of Data Circulation.

Fig. 2 SHAP summary plots for three models.

Fig. 3 Effect of training data volume on model performance (R^2).

Fig. 4 Privacy compensation curves for the three risk-attitude types.

Author contributions

All three authors contributed equally to the work. All the authors unanimously recognize each other's efforts and contributions, and there is no conflict of interest or contradiction among them.

Competing interest

The authors declare no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Data availability

The data used in this study were collected from publicly accessible websites. During the data collection process, we strictly adhered to ethical guidelines and refrained from collecting or storing any personally identifiable information, and the data were solely utilized for this academic research, ensuring the protection of the privacy and rights of all relevant users. The standardized analysis dataset and the complete custom code are available from the corresponding author upon reasonable request.

Informed consent

This article does not contain any studies with human participants performed by any of the authors.

Figures

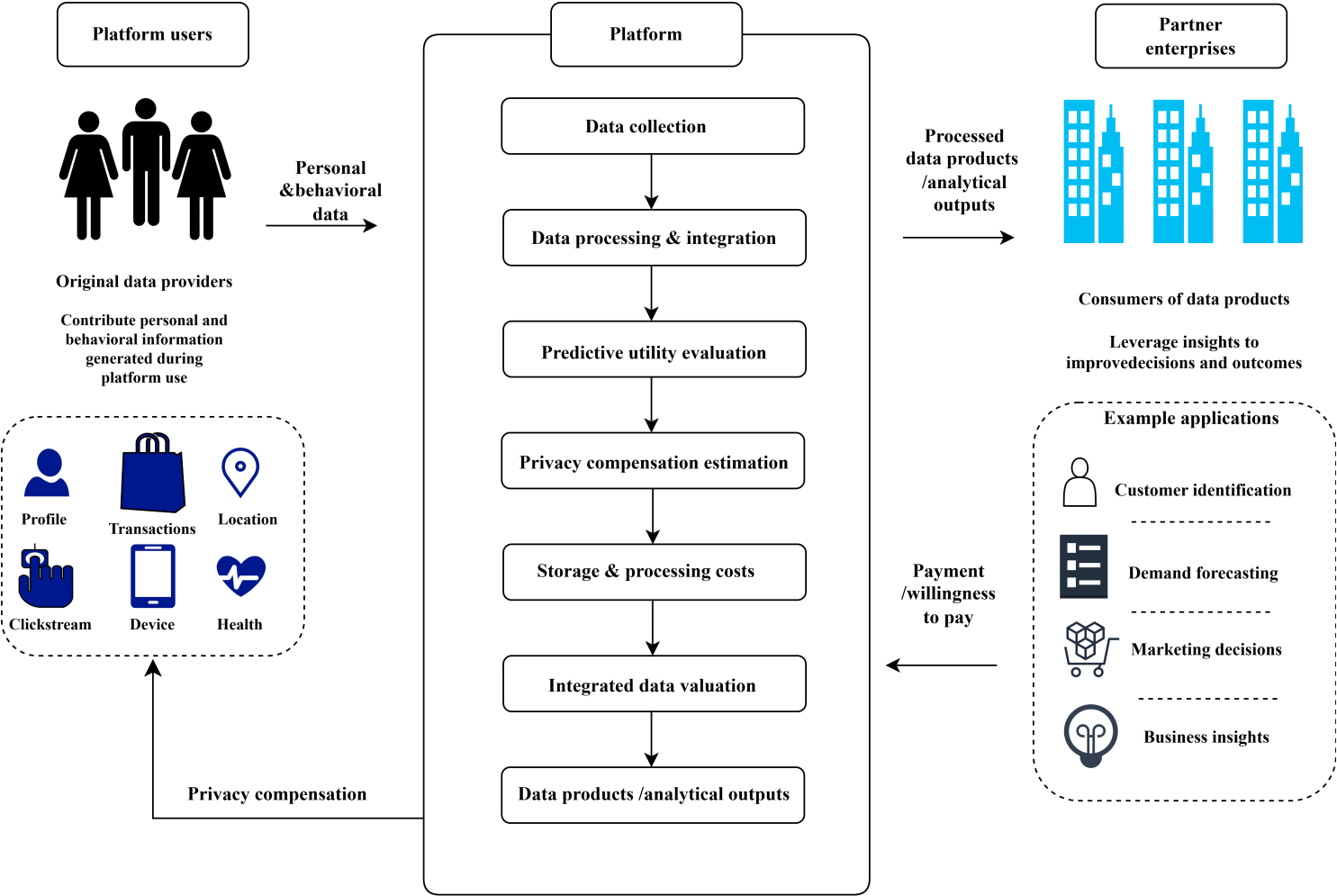


Figure 1

Basic Logic of Data Circulation.

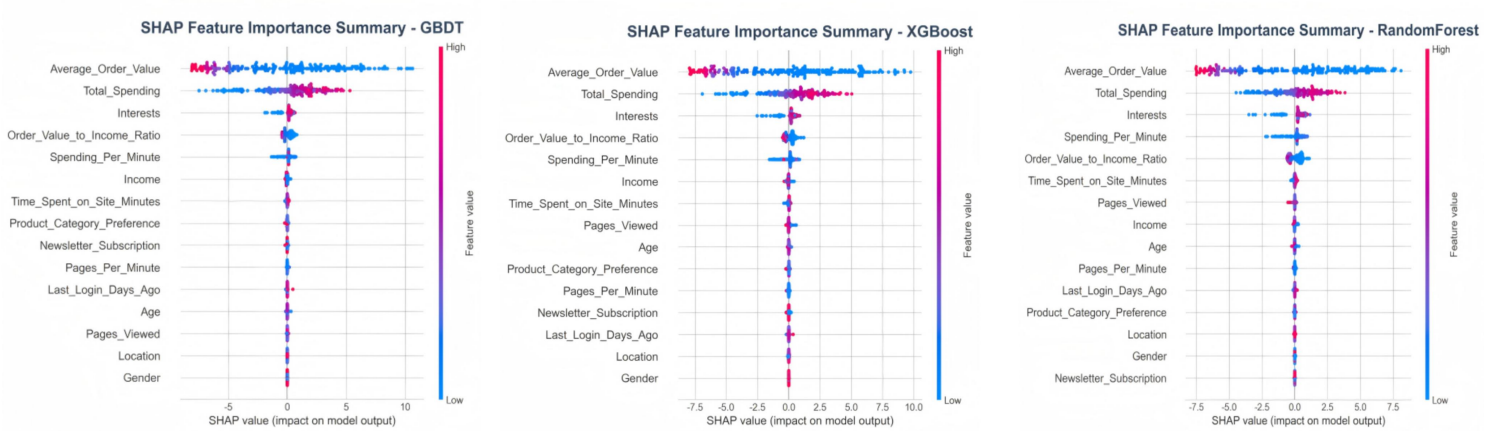


Figure 2

SHAP summary plots for three models.

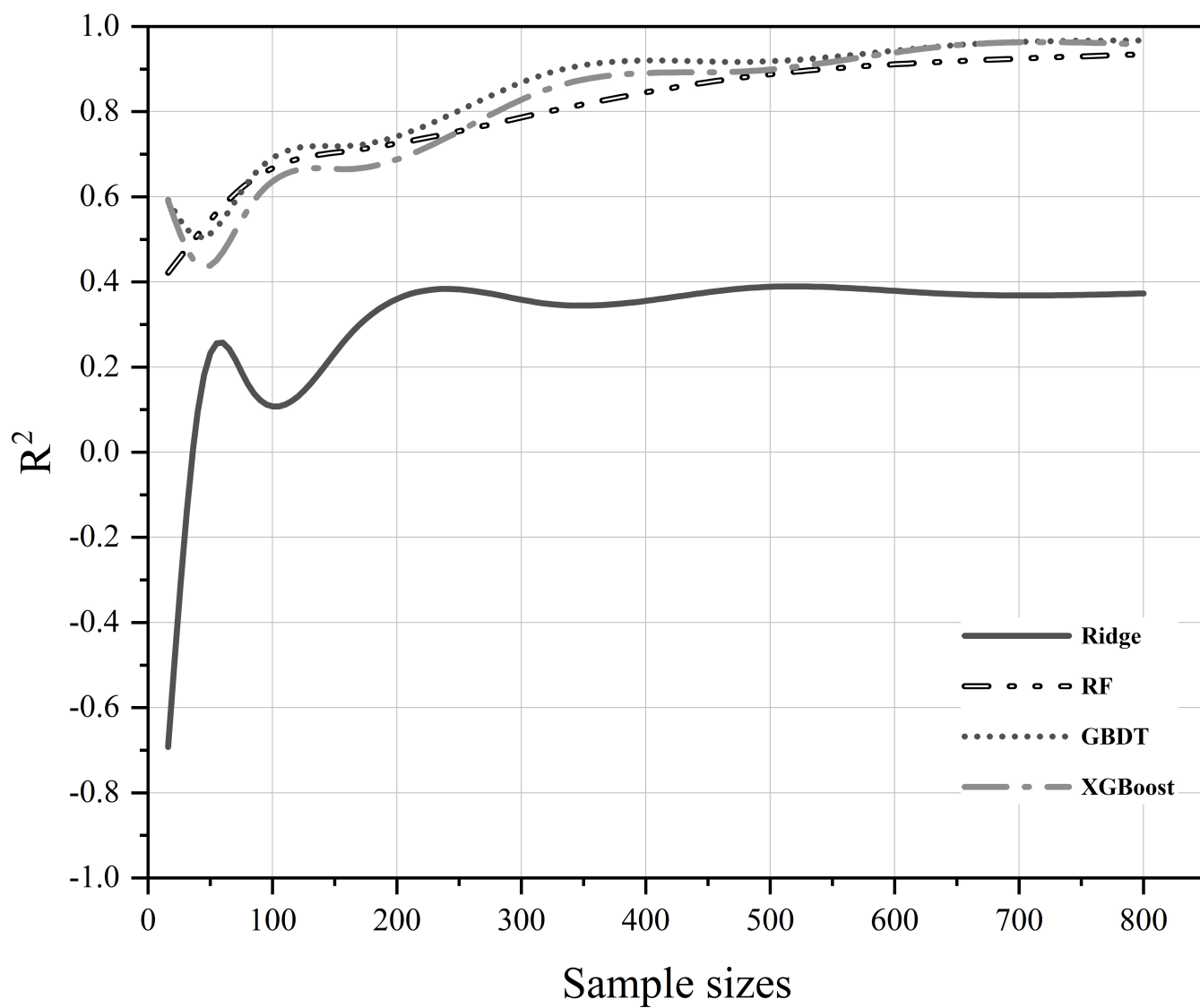


Figure 3

Effect of training data volume on model performance (R^2).

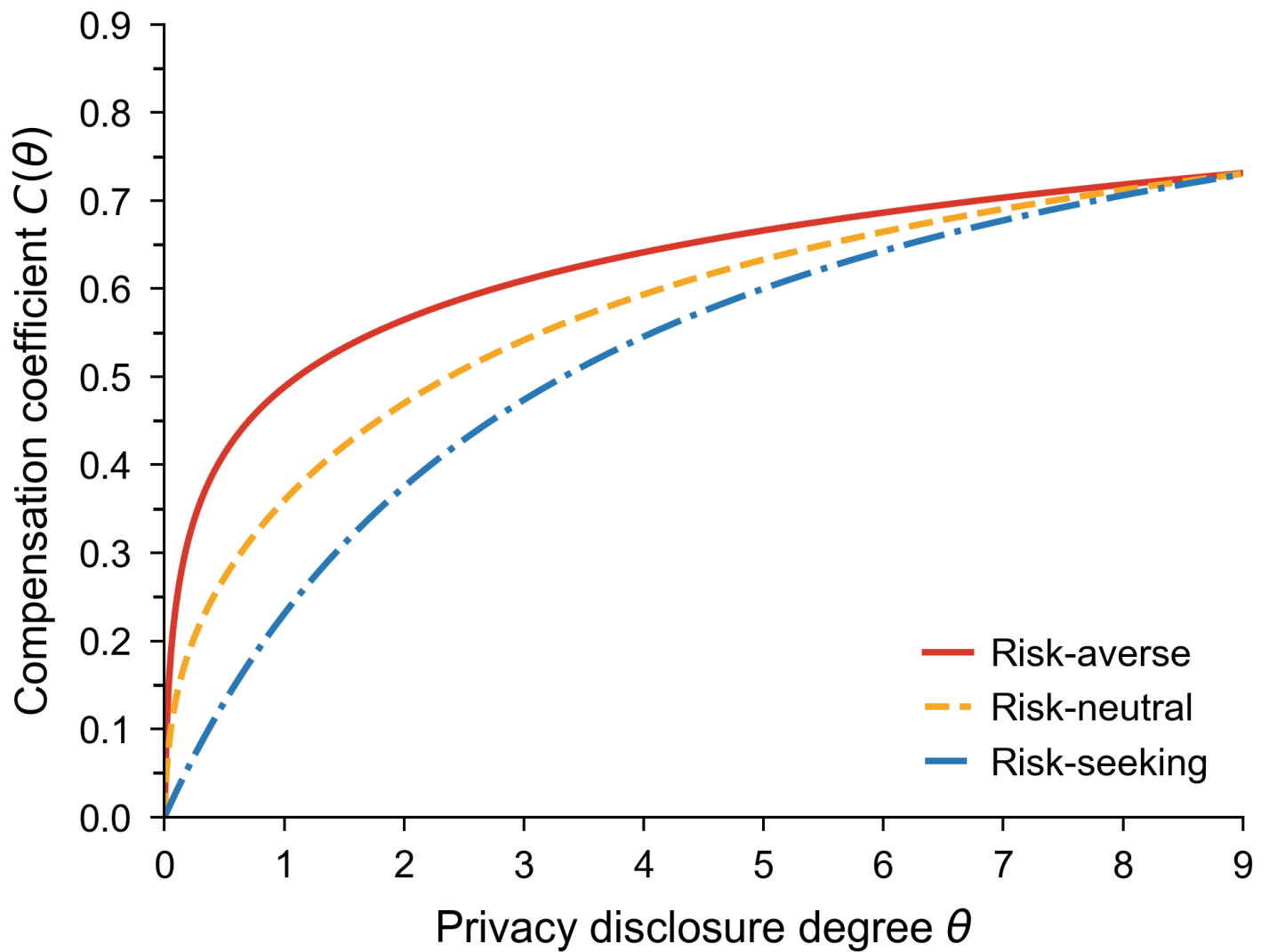


Figure 4

Privacy compensation curves for the three risk-attitude types.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Table.1StatisticalClassificationTableofSensitiveDataCategories.pdf](#)
- [Table.2TableofPrivacyLevelsCorrespondingtoAttributesofDataSet.pdf](#)
- [Table.3Detailedperformanceresultsofeachmodel.pdf](#)
- [Table.4Correlationcoefficientandimportanceofeachfeatureunderdifferentalgorithms.pdf](#)
- [Table.5Detailedperformanceresultsofeachmodel.pdf](#)
- [Table.6AverageabsoluteSHAPvalueofeachfeature.pdf](#)

- [dataset.csv](#)