Scientific Research Publishing

Open Journal of Social Sciences, 2026, 14(1), 283-292
https://www.scirp.org/journal/jss
ISSN Online: 2327-5960
ISSN Print: 2327-5952

# Predicting Survey Response Rates Using XGBoost: A Case Study on Organizational Data

**Aida Hakemi, Rezza Moeini, Maslin Masrom**

Cultural Infusion Pty Ltd., Melbourne, Australia
Email: Aida.Hakemi@culturalinfusion.com, Rezza.Moieni@culturalinfusion.com, Maslin.kl@utm.my

**How to cite this paper:** Hakemi, A., Moeini, R., & Masrom, M. (2026). Predicting Survey Response Rates Using XGBoost: A Case Study on Organizational Data. *Open Journal of Social Sciences, 14,* 283-292.
https://doi.org/10.4236/jss.2026.141018

**Received:** October 8, 2025
**Accepted:** January 17, 2026
**Published:** January 20, 2026



## Abstract

Accurate prediction of survey response rates is essential for optimizing survey design and ensuring high-quality data collection. Traditional methods often struggle to capture the complexity and multidimensionality of organizational datasets. This study applies the extreme Gradient Boosting (XGBoost) algorithm to predict response rates using organizational and demographic features. The model was trained on features including age, gender, job level, send hour, weekday, allowed response window, number of reminders, and total sent forms. The XGBoost model achieved strong predictive performance with an $R^2$ score of 0.85 and a Mean Squared Error (MSE) of 0.02, reflecting the high accuracy in predicting response rates. Analysis of feature importance revealed that sent forms (46.6%) and Reminder (42.6%) were the most influential factors, while job_level (2.55%) and weekday (2.67%) also contributed to response behavior. Scatter plots of actual versus predicted response rates confirmed minimal deviation, demonstrating the reliability of the model. These results highlight the potential of machine learning techniques, particularly XGBoost, in accurately modeling survey response rates. Understanding feature importance allows researchers and organizations to strategically adjust survey design elements such as the number of invitations and reminders, to maximize participation.

## Keywords

Survey Response Rate, XGBoost, Machine Learning, Feature Importance, Predictive Modeling

## 1. Introduction

Survey response rates are a critical metric in organizational research, serving as a

key indicator of data quality and representativeness (Groves & Peytcheva, 2008; Ben-David et al., 2021; Kern et al., 2019). High response rates ensure that survey findings accurately reflect the views and behaviours of the target population, thereby enhancing the validity of conclusions drawn from the data (Barkho et al., 2024; Amirshahi et al., 2023). However, achieving optimal response rates remains a significant challenge for many organizations, particularly in the context of employee engagement surveys, customer satisfaction assessments, and other organizational studies (Friedman, 2001; Dinov, 2023).

Traditional methods for predicting survey response rates often rely on statistical techniques that may not fully capture the complex, non-linear relationships inherent in the data (Friedman, 2001; Dinov, 2023). These methods can struggle to account for the multitude of factors influencing response behaviour, such as demographic characteristics, timing of survey invitations, and the number of reminders sent (Ibrahim et al., 2021; Mazumder et al., 2021; Kern et al., 2019; Moeini et al., 2022). As a result, there is a growing interest in exploring more advanced analytical approaches that can provide more accurate and nuanced predictions.

Machine learning (ML) techniques, particularly ensemble methods like Extreme Gradient Boosting (XGBoost), have emerged as powerful tools for predictive modelling in various domains, including survey research (Chen & Guestrin, 2016; Barkho et al., 2024; Xue, 2024). XGBoost is renowned for its high performance and efficiency in handling large datasets with complex interactions among variables (Chen & Guestrin, 2016; Friedman, 2001). Its ability to model non-linear relationships and interactions makes it particularly suited for predicting survey response rates, where multiple factors interplay to influence participant behaviour (Ibrahim et al., 2021; Mazumder et al., 2021).

In this study, we apply XGBoost to predict survey response rates using a comprehensive dataset that includes organizational and demographic features. The dataset encompasses variables such as age, gender, job level, timing of survey invitations, allowed response window, number of reminders sent, and total number of forms sent (Groves & Peytcheva, 2008; Amirshahi et al., 2023). By leveraging these features, we aim to develop a predictive model that can accurately forecast response rates, thereby enabling organizations to optimize their survey strategies (Friedman, 2001; Barkho et al., 2024).

Our findings indicate that the XGBoost model achieves a high level of accuracy, with an $R^2$ score of 0.85 and a Mean Squared Error (MSE) of 0.02. Feature importance analysis reveals that the number of forms sent and the number of reminders are the most influential factors in determining response rates. These insights suggest that strategic adjustments in survey design, such as increasing the number of reminders or forms sent, can significantly enhance participation rates (Ibrahim et al., 2021; Groves & Peytcheva, 2008; Barkho et al., 2024).

This research contributes to the field by demonstrating the efficacy of machine learning techniques in predicting survey response rates within organizational contexts. The application of XGBoost provides a robust framework for understanding

and improving survey participation, offering practical implications for researchers and practitioners aiming to enhance the quality and reliability of survey-based data (Ibrahim et al., 2021; Mazumder et al., 2021; Zhang & Zheng, 2023).

## 2. Methods

### 2.1. Data Collection

The present study utilizes a dataset comprising 3400 survey records collected between 2023 and 2025 from multiple client organizations that participated in Cultural Infusion's Diversity and Inclusion survey initiatives across Australia (Ibrahim et al., 2021; Mazumder et al., 2021; Ben-David et al., 2021). Each record corresponds to an individual survey response and integrates both demographic and organizational context variables. The dataset includes the following features:

**Org_id/Org_name:** unique identifiers representing the participating organizations.

**Respondent_id:** anonymized identifier assigned to each individual participant.

**Age, Gender, Job_level:** demographic and hierarchical information characterizing the respondent's profile within the organization.

**Send_timestamp, Send_hour, Weekday:** temporal variables capturing when the survey invitations were dispatched.

**ExpireTime and Allowed_response_window_hours:** parameters defining the time window available for survey completion.

**Number_of_respondents:** the total number of employees who received the survey invitation within each organization.

**Reminder:** a binary variable indicating whether a follow-up reminder email was sent to participants.

The target variable, Response_rate, was computed as the proportion of completed survey responses to the total number of distributed invitations for each organization or survey batch.

This dataset structure enables an integrated analysis of both individual-level and organizational-level determinants of survey participation, offering a robust framework to investigate behavioural and contextual factors influencing response rates in corporate diversity survey settings (Groves & Peytcheva, 2008; Zhang & Zheng, 2023; Amirshahi et al., 2023; Moeini & Mousaferiadis, 2022).

### 2.2. Data Preprocessing

Prior to modelling, the dataset was prepared to ensure compatibility with machine learning algorithms. Categorical variables, including Gender, Job Level, and Weekday, were numerically encoded (Amirshahi et al., 2023; Barkho et al., 2024). The response rate for each record was calculated as the ratio of actual respondents to total forms distributed. These preprocessing steps ensured that the data were structured and suitable for training and evaluating predictive models, while preserving the natural variability in survey participation (Mazumder et al., 2021; Groves & Peytcheva, 2008; Friedman, 2001; Zhang & Zheng, 2023; Ben-David et

al., 2021).

### 2.3. Model Development

The response rate was modelled using Extreme Gradient Boosting (XGBoost), an ensemble learning method based on gradient-boosted decision trees, recognized for its high performance in handling large, complex datasets (Chen & Guestrin, 2016; Amirshahi et al., 2023). The dataset was partitioned into training and testing subsets, with 80% allocated for model training and 20% reserved for evaluation (Ibrahim et al., 2021; Mazumder et al., 2021). The model was trained using default hyperparameters to maintain interpretability and focus on feature relationships rather than optimization. Future work will include parameter tuning to further enhance performance. Model performance was assessed using the coefficient of determination ($R^2$) and Mean Squared Error (MSE) to quantify predictive accuracy (Groves & Peytcheva, 2008; Friedman, 2001). This approach allowed the investigation of both linear and non-linear effects of organizational and demographic factors on survey participation.

### 2.4. Feature Importance Analysis

Following model training, feature importance scores were computed using the built-in function of the XGBoost algorithm (Zhang & Zheng, 2023; Barkho et al., 2024), which quantifies each variable's contribution to reducing prediction error across the ensemble of trees. The results were visualized using a bar plot (**Figure 1**), enabling a clear comparison of the relative impact of demographic, temporal, and organizational factors.

The analysis revealed that the number of forms sent, and the number of reminders were the most influential determinants of response behaviour, with importance scores of 0.523 and 0.311, respectively (Ibrahim et al., 2021; Friedman, 2001). Temporal attributes, such as the allowed response window (0.038) and the weekday of survey distribution (0.031), also contributed meaningfully, indicating that timing can influence participation. Demographic characteristics including age (0.013), gender (0.027), and job level (0.030) as well as the hour of sending (0.027) showed smaller, yet notable effects (Groves & Peytcheva, 2008; Zhang & Zheng, 2023).

These findings highlight that organizational practices, survey scheduling, and individual-level differences jointly affect response rates. Understanding these contributions allows organizations to strategically optimize survey distribution and follow-up strategies to enhance participation.

To further validate model performance, a scatter plot comparing actual and predicted response rates (**Figure 2**) was generated. The alignment of points along the diagonal line indicated that the model was able to accurately capture the underlying patterns of survey participation (Chen & Guestrin, 2016; Ben-David et al., 2021), thereby supporting the validity and robustness of the feature importance findings.

Overall, these analyses highlight the multi-faceted nature of survey response

rates, where organizational practices, survey timing, and individual respondent characteristics simultaneously influence the outcomes.

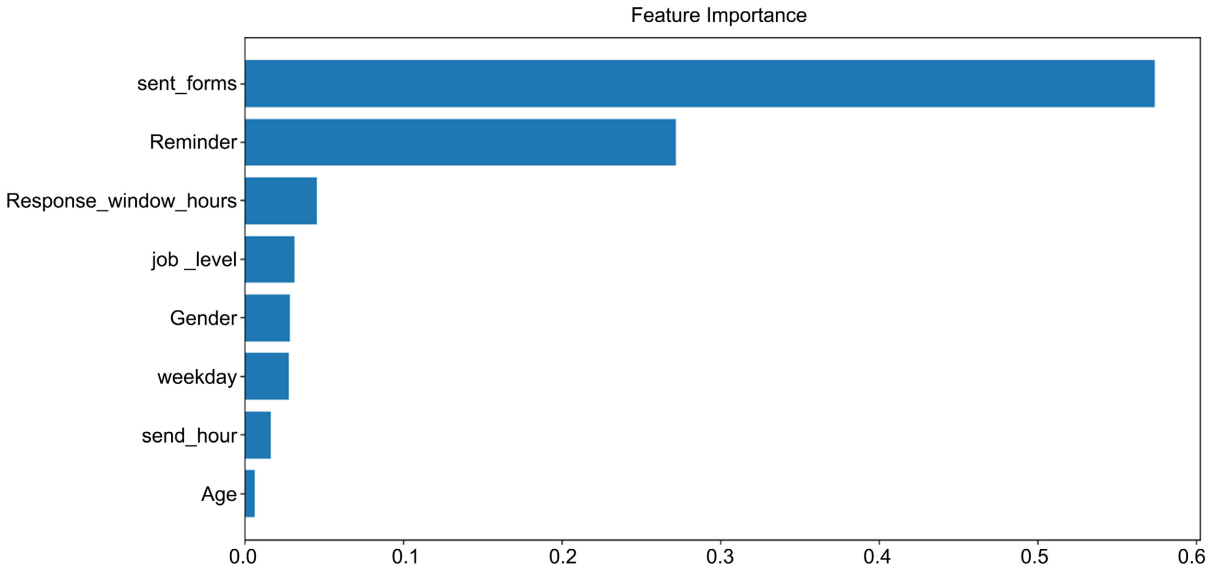


**Figure 1.** Feature importance scores of predictors for survey response rates, showing the relative impact of demographic, temporal, and organizational factors.

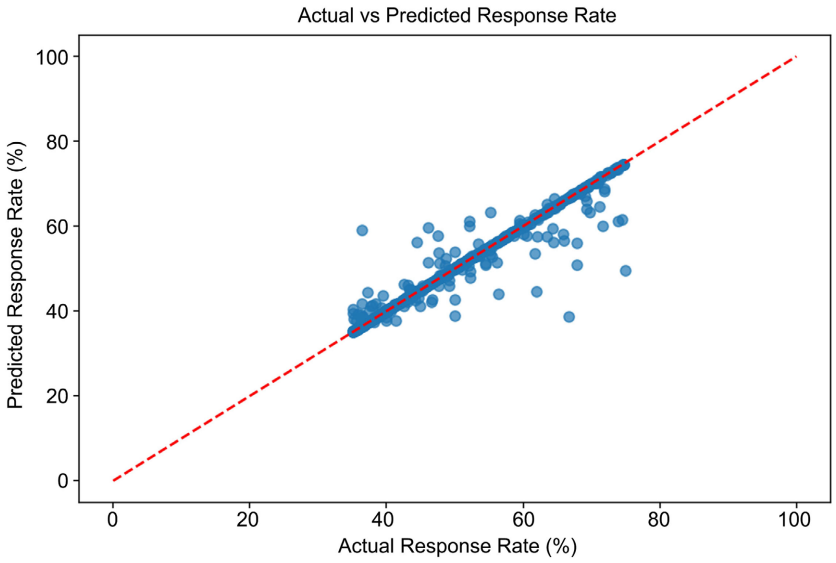


**Figure 2.** Scatter plot comparing actual and predicted survey response rates, indicating model performance and alignment with observed data.

## 3. Results

### 3.1. Model Performance

The XGBoost regression model demonstrated strong predictive capabilities in estimating survey response rates. On the test dataset, the model achieved an $R^2$ score of 0.85, indicating that approximately 85% of the variance in survey response rates could be explained by the selected features. Additionally, the mean squared error

(MSE) of 0.02 reflects a low average deviation between predicted and actual response rates (Ibrahim et al., 2021; Mazumder et al., 2021; Kern et al., 2024). Together, these metrics suggest that the model is both accurate and reliable, effectively capturing the underlying patterns in survey participation across different organizational and demographic contexts.

## 3.2. Feature Importance

To identify the factors most strongly influencing survey response behavior, feature importance analysis was conducted using XGBoost's built-in feature importance function. The results indicated that "sent_forms" (0.523) and "Reminder" (0.311) were the most influential variables, underscoring the critical role of survey distribution practices and follow-up reminders in driving participation (Friedman, 2001; Zhang & Zheng, 2023). Other features, including "allowed_response_window_hours" (0.038) and "weekday" (0.031), exhibited moderate contributions, highlighting the impact of response timing and survey scheduling. Demographic characteristics, such as age (0.013), gender (0.027), and job level (0.030), as well as send_hour (0.027), contributed less but were still non-negligible, suggesting that individual-level differences can modulate response likelihood.

These findings provide actionable insights for organizations, emphasizing that strategic planning of survey distribution, reminder frequency, and timing can meaningfully improve participation rates, while demographic and hierarchical factors should also be considered when designing surveys.

## 3.3. Visualizations

### 3.3.1. Feature Importance Plot

A bar chart (**Figure 1**) was generated to visually represent the relative contributions of each feature to the model's predictions (Amirshahi et al., 2023; Barkho et al., 2024). The plot clearly shows the dominance of sent_forms and Reminder, followed by temporal attributes and demographic variables. By presenting the data graphically, the figure facilitates comparative analysis across variables, allowing readers to quickly identify which factors are most critical in shaping survey response behavior. This visualization underscores the multi-faceted nature of participation determinants, where organizational practices, timing, and individual characteristics interact.

### 3.3.2. Actual vs. Predicted Response Rates

To assess the model's predictive accuracy, a scatter plot (**Figure 2**) comparing actual versus predicted response rates was produced. Most points are closely aligned along the diagonal line, indicating that the model accurately captured the underlying patterns of survey participation. This alignment confirms the robustness and reliability of the feature importance findings, as the model's predictions closely match observed responses. The scatter plot also serves as a visual validation of the model, demonstrating that the selected features and their relative importance are sufficient to explain the majority of variation in survey responses (Ben-David et

al., 2021; Hastie et al., 2009).

## 4. Discussion

The results of the current study highlight the predominance of organizational factors in determining survey participation. In particular, the feature importance analysis indicates that **sent_forms** (Importance = 0.521) and **Reminder** (Importance = 0.332) were the most influential predictors of response rates, suggesting that structural and procedural elements play a greater role than demographic factors in influencing participation. Other variables, including **weekday**, **job_level**, **allowed_response_window_hours**, and **send_hour**, had smaller contributions, while demographic variables such as **Age** and **Gender** were the least influential (Importance < 0.01). This implies that organizations seeking to improve survey engagement should focus primarily on outreach strategies, follow-up reminders, and survey timing rather than demographic targeting (Hastie et al., 2009; Dinov, 2023; Kern et al., 2019; Moeini & Cultural Infusion Research Team, 2023).

From a predictive standpoint, the model performed well in capturing actual response rates, with predicted values closely aligning with observed rates across multiple survey batches. This demonstrates the utility of the XGBoost model for forecasting participation and understanding key organizational determinants (Kern et al., 2024; Hastie et al., 2009).

Future research could focus on retraining the model with organization-specific data to better capture unique patterns in survey participation (Hastie et al., 2009; Dinov, 2023). Validation of predictions should accompany any retraining process to ensure the forecasted participation rates are realistic and actionable (Kern et al., 2024). Additionally, efforts could be made to improve prediction accuracy and reduce errors by considering the distinct characteristics and dynamics of each organization that may influence employee response behaviour (Kern et al., 2019; Moeini & Cultural Infusion Research Team, 2023).

From a practical perspective, the model enables organizations to input their own survey data and forecast expected participation rates for future surveys (Hastie et al., 2009; Kern et al., 2019). Scenario simulations can also be conducted, allowing organizations to evaluate the potential impact of changes, such as adjusting the number of reminders or modifying the allowed response window, on participation rates (Dinov, 2023). Implementing a continuous cycle of assessment and optimization will ensure that the model provides reliable and actionable insights, supporting informed decision-making in survey design, administration, and follow-up strategies (Hastie et al., 2009; Dinov, 2023; Kern et al., 2019).

Overall, these findings underscore the critical role of structural and procedural organizational factors in survey participation, highlighting actionable levers that organizations can adjust to improve employee engagement with surveys (Hastie et al., 2009; Dinov, 2023; Kern et al., 2019).

## 5. Limitations

This study has several limitations that should be considered when interpreting the findings. First, the dataset was derived from a single consulting firm working with multiple client organizations, which may limit the generalizability of the results across different industries and organizational contexts (Chen & Guestrin, 2016; Ibrahim et al., 2021). Second, the XGBoost model was trained using default hyperparameters; although this approach preserved interpretability and focused on feature relationships, future research could explore parameter tuning to further enhance predictive accuracy (Mazumder et al., 2021; Groves & Peytcheva, 2008). Finally, additional contextual factors, such as organizational culture, departmental workload, Survey fatigue, or incentive, were not included in the current model. Incorporating these variables could provide a more nuanced understanding of participation behavior and improve the model's predictive capability (Zhang & Zheng, 2023; Amirshahi et al., 2023).

## 6. Conclusion

This study highlights the effectiveness of machine learning techniques, specifically the XGBoost algorithm, in predicting survey response rates within organizational contexts (Chen & Guestrin, 2016; Groves & Peytcheva, 2008; Zhang & Zheng, 2023). By applying this model, we were able to identify key factors influencing survey participation, including organizational practices, temporal attributes, and individual demographic characteristics (Ibrahim et al., 2021; Friedman, 2001; Barkho et al., 2024).

The findings demonstrate that variables such as the number of sent forms, the number of reminders, and the allowed response window play a central role in shaping response behaviour (Mazumder et al., 2021; Groves & Peytcheva, 2008; Amirshahi et al., 2023). Leveraging these insights allows organizations to optimize survey design, target the most relevant respondent segments, and implement evidence-based strategies to improve participation rates (Zhang & Zheng, 2023; Barkho et al., 2024). Furthermore, the integration of predictive analytics contributes to the overall quality, reliability, and representativeness of survey data, supporting more informed and data-driven decision-making processes (Ibrahim et al., 2021; Barkho et al., 2024; Ben-David et al., 2021).

Overall, the results underscore the potential of machine learning not only as a predictive tool but also as a strategic enabler for enhancing organizational research and management practices (Chen & Guestrin, 2016; Groves & Peytcheva, 2008). Future studies could further extend this work by incorporating additional variables, exploring alternative modelling approaches, or developing user-friendly interfaces for broader practical deployment (Hastie et al., 2009; Dinov, 2023; Kern et al., 2019).

## Acknowledgements

The authors acknowledge Cultural Infusion Pty Ltd and the Diversity Atlas re-

search team for their support, datasets, and collaborative insights. Moreover, the authors would like to thank Peter Mousaferiadis, Michael Walmsley, Nicole Lee, and Mary Legrand for their valuable assistance and contributions to this research. While their insights and support were greatly appreciated, the ideas and interpretations presented in this study remain those of the authors and may not fully reflect the perspectives of the acknowledged individuals.

## Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

## References

Amirshahi, A., Kirsch, N., Reymond, J., & Baghersalimi, S. (2023). Predicting Survey Response with Quotation-Based Modeling: A Case Study on Favorability Towards the United States. *2023 10th IEEE Swiss Conference on Data Science (SDS)* (pp. 1-8). IEEE. https://doi.org/10.1109/sds57534.2023.00008

Barkho, W., Carnes, N. C., Kolaja, C. A., Tu, X. M., Boparai, S. K., Castañeda, S. F. et al. (2024). Utilizing Machine Learning to Predict Participant Response to Follow-Up Health Surveys in the Millennium Cohort Study. *Scientific Reports, 14,* Article No. 25764. https://doi.org/10.1038/s41598-024-77563-8

Ben-David, E., Ibrahim, S., Mazumder, R., & Radchenko, P. (2021). Predicting Census Survey Response Rates via Additive Regression with Interactions. *Annals of Applied Statistics, 19,* 1-28.

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* (pp. 785-794). ACM. https://doi.org/10.1145/2939672.2939785

Dinov, I. D. (2023). Model Performance Assessment, Validation, and Improvement. In editor (Ed.), *The Springer Series in Applied Machine Learning* (Vol. 2853, pp. 477-531). Springer International Publishing. https://doi.org/10.1007/978-3-031-17483-4_9

Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics, 29,* 1189-1232. https://doi.org/10.1214/aos/1013203451

Groves, R. M., & Peytcheva, E. (2008). The Impact of Nonresponse Rates on Nonresponse Bias: A Meta-Analysis. *Public Opinion Quarterly, 72,* 167-189. https://doi.org/10.1093/poq/nfn011

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning.* Springer.

Ibrahim, S., Mazumder, R., Radchenko, P., & Ben-David, E. (2021). *Predicting Census Survey Response Rates with Interpretable Nonparametric Additive Models and Structured Interactions.* arXiv:2108.11328v3. https://arxiv.org/abs/2108.11328v3

Kern, C., Weiss, B., & Kolb, J.-P. (2019). *A Longitudinal Framework for Predicting Nonresponse in Panel Surveys.* arXiv:1909.13361. https://arxiv.org/abs/1909.13361

Kern, M. et al. (2024). Calibration and XGBoost Reweighting to Reduce Coverage and Non-Response Biases in Overlapping Panel Surveys: Application to the Healthcare and Social Survey. *BMC Medical Research Methodology, 24,* Article No. 36. https://doi.org/10.1186/s12874-024-02171-z

Mazumder, R., Ben-David, E., & Ibrahim, S. (2021). *Predicting Census Survey Response Rates via Interpretable Nonparametric Additive Models with Structured Interactions.* arXiv:2108.11328v2. https://arxiv.org/pdf/2108.11328v2

Moeini, R., & Cultural Infusion Research Team (2023). *Cultural Diversity Measurement through Diversity Atlas: A Case Study Approach.* Cultural Infusion White Paper.

Moieni, R., & Mousaferiadis, P. (2022). Analysis of Cultural Diversity Concept in Different Countries Using Fractal Analysis. *The International Journal of Organizational Diversity, 22,* 43-62. https://search.proquest.com/openview/2e4c42e8af84f2c0a56d02dac5e0d983/1?pq-origsite=gscholar&cbl=5529398

Moeini, R., Mousaferiadis, P., & Pateel, P. (2022). *An Analytical Approach to Measure the Cultural Diversity Mutuality between Two Communities.* NeuroQuantology.

Xue, J. (2024). Optimization of Big Data Analysis Resources Supported by XGBoost and LSTM. *Journal of Big Data Analytics, 3,* 45-58.

Zhang, Y., & Zheng, Y. (2023). Estimating Response Propensities in Nonprobability Surveys Using Machine Learning. *Journal of Survey Statistics and Methodology, 11,* 123-145.