Selection of Big Data Monitoring Indicators for Investment Risk in Photovoltaic Projects: A Case Study of the Solarbao Platform (Postprint)

Authors: Yang Yang, Lin Hui, Hu Guangwei

Date: 2017-11-08T00:00:00+00:00

Abstract

Objective: In the process of constructing a monitoring model for photovoltaic project investment risks, this study proposes a systematic selection scheme for identifying big data application monitoring indicators for internet finance platforms, and validates it using actual cases.

Method: Employing a big data monitoring model, multi-source heterogeneous data from the Solarbao platform was integrated. Expert judgment served as the basis for project investment risk analysis. CHAID decision trees were utilized to induce multi-dimensional monitoring indicator combinations, while R-Q type factor analysis was applied to extract key indicators for investment risk identification.

Results: Eight indicator combinations for monitoring photovoltaic project investment risks and ten key indicators for identifying investment risks were obtained.

Limitations: The specialized indicators in the R-Q type factor analysis require further subdivision and the establishment of a dynamic update mechanism.

Conclusion: This selection scheme satisfies the indicator collection requirements of big data monitoring models and provides valuable reference for investors evaluating photovoltaic project risks, platforms screening suitable projects, and regulatory authorities detecting systemic risks in this sector.

Full Text

Preamble

ChinaXiv Partner Journal, Issue 276, 2016, No. 11

Research on Selection of Big Data Monitoring Indicators for Photovoltaic Project Investment Risk: A Case Study of Solarbao Platform

Yang Yang[1], Lin Hui[1], Hu Guangwei[2]
[1](School of Business, Nanjing University, Nanjing 210093, China)
[2](School of Information Management, Nanjing University, Nanjing 210093, China)

Abstract

[**Objective**] This study proposes a systematic selection scheme for big data application monitoring indicators for internet financial platforms in the context of constructing a photovoltaic project investment risk monitoring model, and verifies it through a real-world case. [**Methods**] We applied a big data monitoring model to integrate multi-source heterogeneous data from the Solarbao platform, used expert judgment as the basis for project investment risk analysis, employed CHAID decision trees to induce multi-dimensional monitoring indicator combinations, and utilized R-Q factor analysis to extract key indicators for identifying investment risks. [**Results**] The analysis yielded eight monitoring indicator combinations for photovoltaic project investment risk and ten key indicators for risk identification. [**Limitations**] The professional indicators in the R-Q factor analysis require further refinement and a dynamic update mechanism. [**Conclusions**] The proposed selection scheme meets the indicator collection requirements of big data monitoring models and offers valuable insights for investors assessing photovoltaic project risks, platforms screening suitable projects, and regulators identifying systemic risks in this domain.

Keywords: Big data monitoring index; Photovoltaic project; Investment risk; CHAID decision tree; R-Q mode factor analysis

Classification Codes: F276.3; F830; TM62

2. Construction of Big Data Monitoring Model for Photovoltaic Project Investment Risk

The application environment of a big data monitoring model constitutes an important foundation for selecting monitoring indicators. This study adopts the Solarbao platform as the application environment for the monitoring model. Solarbao is an internet-based financial service platform operating on a physical asset leasing model, specializing in photovoltaic project financing leases with numerous investment projects, making it an ideal research subject. The platform combines features of crowdfunding and financial leasing: investors purchase solar panels through the platform and entrust them to a leasing company that rents the panels to power generation enterprises. These enterprises then use electricity sales revenue and government subsidies to pay the lease fees. Throughout this process, customers never physically handle the solar panels but receive returns through periodic rental payments—an upfront investment followed by later

returns—thereby imbuing the platform with internet finance characteristics. The application environment for the photovoltaic project investment risk big data monitoring model is illustrated in [Figure 1: see original paper].

As shown in [Figure 1: see original paper], the Solarbao platform serves as an information intermediary between photovoltaic projects and investors. Information regarding financing targets, power station operators, operating conditions, and government subsidies is displayed through the platform, revealing the fundamental conditions for project financing and subsequent operations. This provides the basis for selecting photovoltaic project investment risk monitoring indicators and forms the data foundation for the big data monitoring model. The photovoltaic project investment risk big data monitoring model is presented in [Figure 2: see original paper].

The model features a hierarchical information aggregation architecture. Risk monitoring based on this architecture involves four steps: (1) Data extraction and preprocessing. We employ a GoldenGate-based method for extracting multi-source heterogeneous data, capturing data in real-time from platform online logs and storing it in Trail format files. This step also addresses issues such as missing data and outliers. (2) Data-level aggregation. We utilize Hadoop' s MapReduce parallel computing framework to accelerate data loading. The framework automatically aggregates and sorts data, outputting final results to a SQL Server database and information assistance system. The MapReduce framework enhances data loading speed, providing technical support for real-time, dynamic data acquisition in the big data monitoring model. (3) Information-level aggregation. Data mining techniques are applied to select photovoltaic project investment risk monitoring indicators in real-time and dynamically, presenting multi-dimensional correlations and key characteristics of risk monitoring indicators and proposing similarity matching schemes for specific risk types. Previous studies on project risk identification have employed various data mining techniques, including association rules, Kano models, Kansei engineering, CHAID decision trees, and R-Q factor analysis. A comparison of these techniques appears in .

As shows, association rules are unsuitable for risk judgment extraction and exhibit high model complexity; Kano models focus more on extracting customer preferences for project risks; Kansei engineering lacks the capacity to handle multi-dimensional data; CHAID decision trees and R-Q factor analysis are relatively applicable. These methods can use expert judgment as the basis for project investment risk analysis, employ CHAID decision trees to dynamically induce multi-dimensional monitoring indicator combinations, and utilize R-Q factor analysis to dynamically extract key risk identification indicators. HBase offers efficient distributed concurrent processing, easy scalability, and dynamic flexibility. Consequently, we store the foundational data for data mining in one-dimensional or multi-dimensional forms in HBase, supporting real-time, dynamic analysis by the data mining engine while responding to multi-condition rapid combination queries in data monitoring. Hive backs up the foundational

data to facilitate offline computation and indicator optimization. (4) Decision-level aggregation. We establish an index mechanism for monitoring indicators in HBase, enabling low-latency, dynamic extraction of monitoring data during data import. A Join engine then establishes mapping relationships between risk types and cross-table data to monitor photovoltaic project investment risks in real-time. Through these steps, we construct a big data application monitoring model with dynamic monitoring processes and real-time results.

3.1. Using CHAID Decision Trees to Induce Multi-Dimensional Monitoring Indicator Combinations

CHAID decision trees (Chi-square Automatic Interaction Detection Decision Trees) feature supervised feature extraction and description capabilities. The fundamental concept involves spontaneously constructing decision rules from training datasets to classify other datasets. Each non-leaf node represents a feature attribute, each branch represents the output value of this attribute, and each leaf node stores a class. Rule formation begins at the root node, testing the feature attributes of items to be classified and selecting branches based on chi-square test results from the CHAID algorithm until reaching a leaf node, which stores the final classification result. Since CHAID algorithms use chi-square test results as branching criteria, decision tree pruning is unnecessary.

Based on basic information from Solarbao platform photovoltaic projects, this study uses CHAID decision trees to dynamically induce multi-dimensional monitoring indicator combinations. As shown in [Figure 3: see original paper], basic information refers to project details displayed in the Solarbao platform' s investable project list interface. We define these as basic indicators, comprising annualized return, product unit price, investment lock-up period, and interest payment method.

Each basic indicator contains two attribute types representing decision tree branches, while project investment risk type represents leaf nodes, including low-risk, medium-risk, and high-risk categories, as detailed in .

presents a sample of Expert Judgment Questionnaire A for photovoltaic project investment risk. If an expert considers a photovoltaic project with annualized return $ $7%, product unit price $ $2,000 RMB, lock-up period $ $90 days, and lump-sum principal and interest payment at maturity as low-risk, they would mark "LR" in the questionnaire.

3.2. Using R-Q Factor Analysis to Extract Key Monitoring Indicators

R-Q factor analysis (R-Q Mode Factor Analysis) is a multivariate statistical method combining R-type and Q-type factor analysis. Due to the dual relationship between these two approaches, variable points and sample points can be projected onto the same factor space, with sample point types conve-

niently explained by neighboring variable points. This dual relationship has been mathematically proven, allowing Q-type results to be derived from R-type results through orthogonal transformation—a mathematical process known as correspondence analysis. In a uniformly scaled factor space, the proximity between sample points and variable points on the correspondence analysis graph indicates the degree of explanation provided by that variable factor.

Based on basic and professional information from Solarbao platform photovoltaic projects, this study employs R-Q factor analysis to dynamically extract key indicators for identifying project investment risks. In the example shown in [Figure 4: see original paper], professional information refers to financing details that appear after clicking on any project in the Solarbao platform' s investable project list. We define these as professional indicators, including photovoltaic project operator information, construction progress schedules, relevant qualifications, and project security safeguards. Variable points consist of both basic and professional indicators, as listed in .

Sample points represent project investment risk types, including low-risk, medium-risk, and high-risk categories. provides a sample of Expert Judgment Questionnaire B for photovoltaic project investment risk. If an expert believes that indicator A1, "higher annualized investment return," characterizes medium-risk and high-risk photovoltaic projects, they would check the corresponding options.

During questionnaire administration, if power sector experts encounter difficulties understanding specialized financial terminology in Questionnaires A and B, survey administrators provide detailed explanations to ensure accurate comprehension. Based on the survey results, we obtain expert judgments and statistical data on project investment risks, then apply CHAID decision trees and R-Q factor analysis to extract multi-dimensional monitoring indicator combinations and key indicators from expert judgments, thereby supporting the selection of big data monitoring indicators for project investment risk.

4. Empirical Analysis and Results Discussion

We distributed 85 copies of Questionnaire A and 32 copies of Questionnaire B to participating experts. We recovered 68 valid Questionnaire A responses (80% valid response rate). Since each Questionnaire A contains 16 judgment results, the CHAID decision tree modeling sample size was 1,088. We recovered 30 valid Questionnaire B responses (approximately 93.8% valid response rate). With 12 judgment results per Questionnaire B, the R-Q factor analysis sample size was 360.

4.1. Summarizing Multi-Dimensional Monitoring Indicator Combinations Based on CHAID Decision Tree

The CHAID decision tree achieved an overall classification accuracy of 66.9%, with low-risk project classification accuracy at 84.2%, medium-risk at 45%, and

high-risk at 70.2%. The decision tree generated eight decision rules, presented in .

In , the eight decision rules follow an "if-then" format. For example, Rule 1 states: If a photovoltaic project' s annualized return $ $7%, product unit price $ $2,000 RMB (or >2,000 RMB), lock-up period $ $90 days, and interest payment method is monthly interest payment, the project is classified as low-risk. Rule 2 states: If annualized return >7%, product unit price >2,000 RMB, lock-up period >90 days, and interest payment method is lump-sum principal and interest payment at maturity (or monthly interest payment), the project is classified as high-risk. In these rules, each risk type is characterized by four monitoring indicators, yielding eight multi-dimensional monitoring indicator combinations based on the CHAID decision tree.

4.2. Extracting Key Monitoring Indicators Based on R-Q Factor Analysis

presents the aggregated expert judgment results from Questionnaire B and the R-Q factor analysis outcomes. The aggregated expert judgment results show the percentage of experts who checked each indicator for each risk type relative to the total number of returned questionnaires. The R-Q factor analysis results indicate the correlation between monitoring indicators and project investment risk types, measured by chi-square values.

Based on the R-Q factor analysis results, the correspondence analysis graph of project investment risk types and monitoring indicators is shown in [Figure 5: see original paper].

Three key findings emerge: (1) Low-risk projects cluster with indicators A4, A7, A8, A10, A11, and A12, indicating that projects with higher interest payment frequency, longer lessee company establishment duration, more complete lessee company qualifications, larger expected annual power generation during the project term, greater proportion of per-kWh subsidies in electricity sales price, and richer on-site photos are classified as low-risk. (2) High-risk projects cluster with indicators A1, A2, A3, and A5, suggesting that projects with higher annualized investment returns, higher financing product unit prices, longer lock-up periods, and longer total project investment periods are classified as high-risk. (3) For medium-risk project identification, based on chi-square statistical significance, indicators A3 and A10 show chi-square values of 0.8 and 0.9 respectively for medium-risk type, while indicators A6 and A9 also exhibit positive chi-square values. We therefore classify these four indicators as auxiliary monitoring indicators for medium-risk projects.

Further analysis reveals: (1) Higher annualized investment returns, higher financing product unit prices, and longer lock-up periods indicate high-risk projects, consistent with Rule 2 from the CHAID decision tree. (2) Higher interest payment frequency indicates low-risk projects, consistent with Rules 1 and 5 from the CHAID decision tree. This demonstrates that three rules

from the CHAID decision tree model are validated by the R-Q factor analysis results, confirming the reliability of the selection scheme.

Conclusion

Selecting appropriate monitoring indicators represents a core challenge in constructing photovoltaic project investment risk monitoring models for internet financial platforms. This study built a real-time, dynamic big data monitoring model for photovoltaic project investment risk and proposed a corresponding indicator selection scheme. By integrating multi-source heterogeneous data from the Solarbao platform and using expert judgment as the basis for risk analysis, we employed CHAID decision trees to induce multi-dimensional monitoring indicator combinations and R-Q factor analysis to extract key risk identification indicators. The research yielded eight "if-then" style monitoring indicator combinations and ten key indicators for photovoltaic project investment risk, demonstrating the scheme's feasibility and compliance with model requirements. The validation of three CHAID decision tree rules through R-Q factor analysis confirms the scheme' s reliability. The limitation lies in the need for further refinement and dynamic updating of professional indicators in the R-Q factor analysis.

Currently, internet finance in the power new energy sector plays an increasingly important role in photovoltaic project financing, though accumulating credit risks among financing entities cannot be ignored. Strengthening photovoltaic project investment risk monitoring has become an objective necessity and urgent priority for promoting sustainable, healthy development of internet finance in this domain. The proposed big data monitoring indicator selection scheme addresses practical risk monitoring needs, providing a valuable tool for investors to assess project risks, platforms to screen suitable projects, and regulators to identify systemic risks, thereby possessing both theoretical significance and practical value.

References

[1] Scholtens B. Finance as a Driver of Corporate Social Responsibility [J]. Journal of Business Ethics, 2006, 68(1): 19-33.

[2] Climent F, Soriano P. Green and Good? The Investment Performance of US Environmental Mutual Funds [J]. Journal of Business Ethics, 2011, 103(2): 275-287.

[3] Graham A, Maher J J, Northcut W D. Environmental Liability Information and Bond Ratings [J]. Journal of Accounting Auditing & Finance, 2001, 16(2): 93-116.

[4] Thomas S, Repetto R, Dias D. Integrated Environmental and Financial Performance Metrics for Investment Analysis and Portfolio Management [J]. Corporate Governance: An International Review, 2007, 15(3): 421-426.

[5] Pope D G, Sydnor J R. What' s in a Picture? Evidence of Discrimination from Prosper.com [J]. Journal of Human Resources, 2008, 46(1): 53-92.

[6] Duarte J, Siegel S, Young L. Trust and Credit: The Role of Appearance in Peer-to-Peer Lending [J]. Review of Financial Studies, 2012, 25(8): 2455-2484.

[7] Luan Chunyu, Dai Rongjia. Study on the Ecological Management Mode and Control Strategy of Internet Financial Information [J]. Information Science, 2015, 33(5): 48-52.

[8] Zhang Lichao, Fang Junming, Tang Qinneng. Research on Risk Identification in the Early Warning of Industry Competitive Intelligence: A Case Study of Photovoltaic Power Generation Industry in China [J]. Information Studies: Theory & Application, 2011, 34(10): 52-55.

[9] Bai Yang. Research on Data Aggregation of Power Equipment Condition Monitoring Based on Big Data [D]. Kunming: Kunming University of Science and Technology, 2014.

[10] Zheng Haiyan, Jin Nong, Ji Cong, et al. Data Analysis Technology and Typical Application of Electric Power User [J]. Power System Technology, 2015, 39(11): 3147-3152.

[11] Takabi H, Joshi J B D, Ahn G J. Security and Privacy Challenges in Cloud Computing Environments [J]. Security & Privacy IEEE, 2010, 8(6): 24-31.

[12] Lu Yonghe, Cao Lizhao. Book Recommendation Model Based Comprehensive Evaluation of Association Rules [J]. New Technology of Library and Information Service, 2011(2): 81-86.

[13] Wang C H, Hsuesh O Z. A Novel Approach to Incorporate Customer Preference and Perception into Product Configuration: A Case Study on Smart Pads [J]. Computer Standards & Interfaces, 2013(35): 549-556.

[14] Sun Xiaoling, Zhao Yuxiang, Zhu Qinghua. Analyzing the Demand of Online Product Review System' s Features Using Kano Model: An Empirical Study of Chinese Online Shops [J]. New Technology of Library and Information Service, 2013(6): 76-84.

[15] Wang C H. Incorporating Customer Satisfaction into the Decision-Making Process of Product Configuration: A Fuzzy Kano Perspective [J]. International Journal of Production Research, 2013(22): 6651-6662.

[16] Nagamachi M. Kansei Engineering: A New Ergonomic Consumer-Oriented Technology for Product Development [J]. International Journal of Industrial Ergonomics, 1995, 15(1): 3-11.

[17] Cheng Tiexin, Guo Tao, Qi Xin. Application of Decision Tree Classification Model in Risk Early Warning of Engineering Project Evaluation [J]. Journal of Applied Statistics and Management, 2010, 29(1): 122-128.

[18] Dey P K. Project Risk Management Using Multiple Criteria Decision Making Technique and Decision Tree Analysis: A Case Study of Indian Oil Refinery [J]. Production Planning & Control, 2012, 35(3): 1-19.

[19] Walden J, Smith J P, Dackombe R V. The Use of Simultaneous R-Q Mode Factor Analysis as a Tool for Assisting Interpretation of Mineral Magnetic Data [J]. Mathematical Geology, 1992, 24(3): 227-247.

[20] Murtagh F. The Correspondence Analysis Platform for Uncovering Deep Structure in Data and Information [J]. The Computer Journal, 2010, 53(3): 304-315.

Author Contributions

Yang Yang: Refined research propositions and ideas, drafted the manuscript, cleaned, processed, and analyzed data.
Lin Hui: Proposed research propositions, designed research framework, collected data, conducted experiments, and revised the manuscript.
Hu Guangwei: Optimized research framework, designed experimental procedures, and revised the final manuscript.

Conflict of Interest Statement

All authors declare no conflict of interest.

Supporting Data

Supporting data is self-archived by the authors. E-mail: hugw@nju.edu.cn.

[1] Yang Yang, Lin Hui, Hu Guangwei. IndexSelection.rar. Experimental data for photovoltaic project risk monitoring indicator selection.
[2] Yang Yang, Lin Hui, Hu Guangwei. SolarbaoDataExtraction.rar. Solarbao platform data extraction program package.
[3] Yang Yang, Lin Hui, Hu Guangwei. QuestionnaireA.rar. Statistical results of Questionnaire A survey.
[4] Yang Yang, Lin Hui, Hu Guangwei. QuestionnaireB.rar. Statistical results of Questionnaire B survey.
[5] Yang Yang, Lin Hui, Hu Guangwei. DecisionTree.spv. CHAID decision tree model experimental output.
[6] Yang Yang, Lin Hui, Hu Guangwei. CorrespondenceAnalysis.spv. R-Q factor analysis model experimental output.

Received: July 25, 2016
Revised: August 29, 2016

Note: Figure translations are in progress. See original paper for figures.