sustainability

MDPI

Article

Data-Driven Public R&D Project Performance Evaluation: Results from China

Hongbo Li, Bowen Yao * and Xin Yan *

School of Management, Shanghai University, Shanghai 200044, China; ishongboli@gmail.com
* Correspondence: y19720606@shu.edu.cn (B.Y.); yjjjx@shu.edu.cn (X.Y.)

Abstract: In public R&D projects, to improve the decision-making process and ensure the sustainability of public investment, it is indispensable to effectively evaluate the project performance. Currently, public R&D project management departments and various academic databases have accumulated a large number of project-related data. In view of this, we propose a data-driven performance evaluation framework for public R&D projects. In our framework, we collect structured and unstructured data related to completed projects from multiple websites. Then, these data are cleaned and fused to form a unified dataset. We train a project performance evaluation model by extracting the project performance information implicit in the dataset based on multi-classification supervised learning algorithms. When facing a new project that needs to be evaluated, its performance can be automatically predicted by inputting the characteristic information of the project into our performance evaluation model. Our framework is validated based on the project data of the National Natural Science Foundation of China (NSFC) in terms of four performance measures (i.e., Accuracy, Recall, Precision, *F*1 score). In addition, we provide a case study that applies our framework to evaluate the project performance in the logistics and supply chain area of NSFC. In conclusion, this paper contributes to the body of knowledge in sustainability by developing a data-driven method that equips the decision-maker with an automated project performance evaluation tool to make sustainable project decisions.

Keywords: public R&D project; performance evaluation; machine learning; logistics and supply chain

Citation: Li, H.; Yao, B.; Yan, X. Data-Driven Public R&D Project Performance Evaluation: Results from China. *Sustainability* **2021**, *13*, 7147. https://doi.org/10.3390/su13137147

Academic Editors: Juliang Zhang, Wenchao Wei, Yefei Yang and António Abreu

Received: 9 May 2021
Accepted: 22 June 2021
Published: 25 June 2021

1. Introduction

To promote technology innovation, government departments usually fund a large number of R&D projects each year. For instance, for the three types of funding (General Program, Young Scientists Program and Regional Program) in the National Natural Science Foundation of China (NSFC), the number of funded projects and the funded amount increased by 11.85% and 13.13%, from 2015 to 2019, respectively (as shown in Figure 1). On the one hand, scientific research benefits from the growth of funding. On the other hand, the increasing number of projects also aggravates the complexity and workload for project assessment [1]. Therefore, to improve the decision-making process and ensure the sustainability of public investment, it is indispensable to effectively evaluate the project performance [2].

Many effective project performance assessment methods have been proposed [3–10]. Among these methods, expert estimation is the most widely used method. This method is based on subjective evaluation and experts score the projects according to the achievement of the projects. Data envelopment analysis (DEA) is another popular approach for project performance assessment [11,12]. Hsu and Hsueh use DEA to evaluate the R&D efficiency of IT projects [13]. Johns and Yu test the efficiency of public R&D projects in 109 universities in China using DEA [14]. However, these methods suffer from some deficiencies. Some of these methods tend to be time-consuming, costly, or subjective, which is not conductive to the sustainability of public investment.

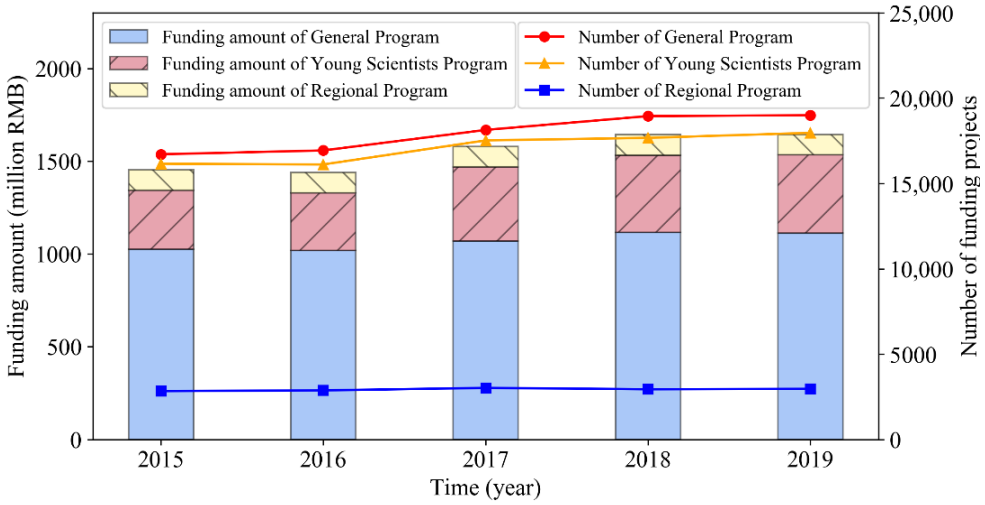

Figure 1. The amount and number of three types of NSFC funds during 2015–2019.

Generally speaking, project evaluation aims to measure the outcomes and impacts of a completed project. As a type of project evaluation, government-funded project evaluation has the similar characteristics. On the other hand, government-funded project evaluation also has its own characteristics. In government-funded project evaluation, government management departments usually rely on standardized procedures to conduct performance evaluation. The objective is to assess the project outcome and provide a basis for project decision-making. In addition to the project topic and the evaluation objective, the main difference between China's public project evaluation and project evaluation in previous research lies in the evaluation methods. In China, the main method of public project evaluation is expert judgement. While in the project evaluation literatures, DEA and other quantitative methods are being gradually adopted [15].

Recent years have witnessed rapid progress in artificial intelligence (AI), and more and more attention has been paid to using machine learning and other AI related techniques to evaluate the performance of R&D projects. Cho et al. propose a framework for R&D performance evaluation. Their framework combines an Analytical Hierarchy Process (AHP) and Bayesian Network [16]. Liu and Hu use a tree-structured growing self-organizing maps (TGSOM) network and spatial data mining to measure the R&D performance [17]. Costantino et al. use an artificial neural network (ANN) to extract expert opinions from historical data, and this process requires no expert participation [18]. Liu et al. present a data-driven evidential reasoning rule model that concentrates on criterion-comprehensive evaluation and funding recommendations [19]. Jang proposes a machine learning model to estimate the level of outputs of the public R&D projects by using the data of national funded research projects in South Korea [20]. Machine learning makes the project evaluation process automatic by extracting knowledge hidden in the historical data. Thereby, machine learning reduces the consumption of manpower, material resources, and capital.

However, for government-funded public R&D projects, there are very few studies applying machine learning to project evaluation, not to mention the research that focuses on China's public R&D projects. In fact, after decades of development, the project management departments of the government in China have accumulated a large amount of public R&D project data, and many of them have been published on the internet, such as the National Natural Science Fund Big Data Knowledge Management Service Portal (http://kd.nsfc.gov.cn, accessed on 20 December 2019). Therefore, we propose a data-driven performance evaluation framework for public R&D projects. In our framework, we collect structured and unstructured data related to completed projects from multiple websites. Then, these data are cleaned and fused to form a unified dataset. We train a project performance evaluation model based on multi-classification supervised learning algorithms by extracting the project performance information implicit in the dataset. When we face

a new project that needs to be evaluated, its performance can be automatically predicted by inputting the characteristic information of the project into our performance evaluation model. Furthermore, although our framework is validated based on the management science related project data from China, our framework can be easily extended to evaluate other types of projects (Section 2).

The contributions of our paper are two-fold. First, we develop a framework that is able to collect and process public R&D project data by integrating web crawlers, regular matching, and image recognition techniques. Second, to construct an effective data-driven R&D project performance evaluation model, we compare eleven machine learning algorithms and adopt four model performance measures (i.e., Accuracy, Recall, Precision and $F1$ score) on the management science-related project data of the National Natural Science Foundation of China. The proposed data-driven public R&D project performance evaluation framework helps to reduce human cost and improve work efficiency in the process of project performance evaluation.

The rest of the paper is structured as follows. In Section 2, we present our data-driven framework for public R&D project performance evaluation. Section 3 describes how the data are collected and processed. In Section 4, we train our model based on different machine learning algorithms. Section 5 presents our experimental results. In Section 6, we perform a case study to show the process of applying our framework to evaluate the logistics and supply chain related projects. Section 7 concludes the paper.

2. A Data-Driven Framework for Public R&D Project Performance Evaluation

For human experts, when evaluating a completed public R&D project, they usually consider many aspects, such as the output, benefit, and quality of the project. Inspired by the expert evaluation process, given a project i that needs to be evaluated, suppose that we can extract n features reflecting the performance of the project from n dimensions. We put the features in an n-dimensional vector $\boldsymbol{x}_i = \left(x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(m)}, \ldots, x_i^{(n)}\right)^T$, where $x_i^{(m)}$ represents the m-th dimension of project i. For project i, its evaluation result given by an expert is denoted as $y_i \in Y$ (in this paper y_i is also called a label), where Y is the set of all the possible evaluation results. According to the requirements of the project funding agencies, y_i may be a numerical value (such as 80, 90, etc.) or a discrete value (such as excellent, good, etc.). This paper uses the latter. Therefore, the evaluation of the performance of public R&D projects can be regarded as a function fitting process. Various factors that affect the project performance can be taken as features $\boldsymbol{x}_i$, and the project performance level evaluated by experts is taken as a label y_i. Then, a function f that maps features $\boldsymbol{x}_i$ to label y_i can be fitted. From the perspective of machine learning, the above project performance evaluation process is a multi-classification problem, i.e., given the features of a project, we need to predict its performance.

Therefore, multi-classification algorithms can be used to estimate the final performance of a public R&D project. To design an effective machine learning-based performance evaluation method for public R&D projects, the following three core problems need to be addressed: (1) how to automatically collect public R&D project data, (2) how to extract effective features from the obtained data, and (3) how to train an effective multi-classification prediction model. In this situation, we propose a data-driven framework for performance evaluation of public R&D projects. This framework consists of the following three stages (Figure 2).

Stage 1: data collection (Sections 3.1 and 3.2). This stage aims at using web crawlers to collect project-related data from the website of project management departments and academic databases. Assuming that the data containing a total of P projects are obtained, then we construct a dataset $D = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_P, y_P)\}$, where the n-dimensional vector $\boldsymbol{x}_i = \left(x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(m)}, \ldots, x_i^{(n)}\right)^T$ is the set of project i's features ($\boldsymbol{x}_i$ is also called a sample or instance). y_i is the label of project i, i.e., y_i denotes the evaluation result of project i.

Stage 2: data processing (Sections 3.3 and 3.4). In this stage, we clean and merge the dataset D. We remove the duplicate values, blank values and outliers in D. We also standardize and normalize some features to eliminate the impact of the feature scale. For the unstructured data such as images and text data, image recognition and natural language processing methods are utilized to extract information and convert them into structured data. Then, we perform descriptive statistical analysis on the resulting dataset.

Stage 3: model training (Section 4). We first select a part of the data from dataset D as the training set D^{tr}. Since the dataset D has an imbalanced distribution, synthetic minority oversampling technique (SMOTE algorithm) is used to deal with it [21]. Prediction algorithms are used to construct n machine learning models $\{f_1, f_2, \ldots, f_m, \ldots, f_n\}$ based on the training set. Then, we use the remaining data in D as test set D^{te} to evaluate the performance of the models.

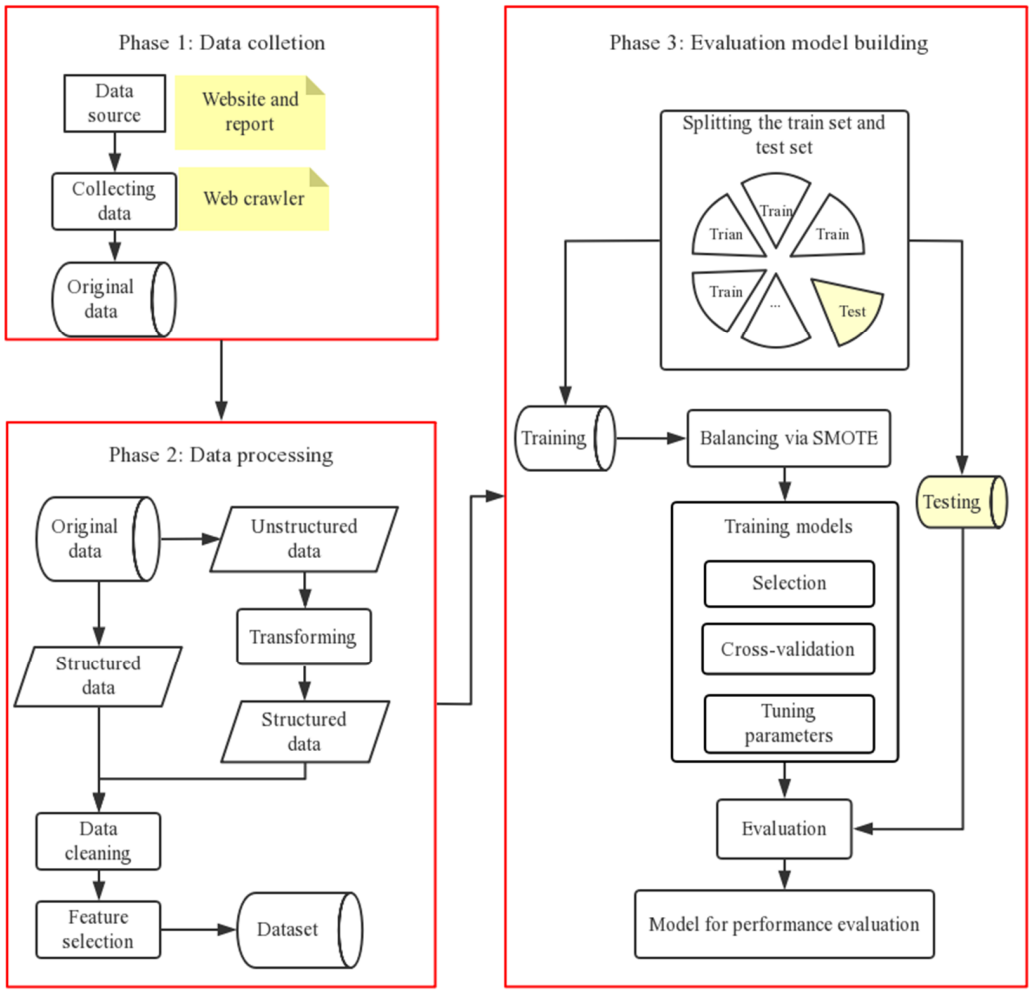

Figure 2. Data-driven framework for public R&D project evaluation.

Our data-driven framework has good applicability and scalability, and its sub-processes can be easily adjusted according to actual situations. For example, this framework is not only applicable to the projects of NSFC, but also can be applied to other types of public R&D projects by slightly modifying the related data format. In the following, we will apply our framework to the project data of NSFC.

3. Data Collection and Processing

3.1. Automatic Data Collection

We implemented a web crawler that consists of three phases (Figure 3) in Python [22] to automatically collect general program project data of the NSFC Management Science Department from the NSFC website (http://www.nsfc.gov.cn, accessed on 20 December 2019). These projects were completed between 2016 and 2017. In the first phase, the crawler generates the URL of each project. In the second phase, the crawler calls the Selenium library to get access to each URL automatically. The crawler is set to sleep for a while after accessing a URL to reduce the load on the website sever as much as possible. Then, the obtained HTML source codes are parsed with the Pyquery library and the regular expression matching method. In the third phase, the matched data are transformed as a key-value format and stored in the Mongo database.

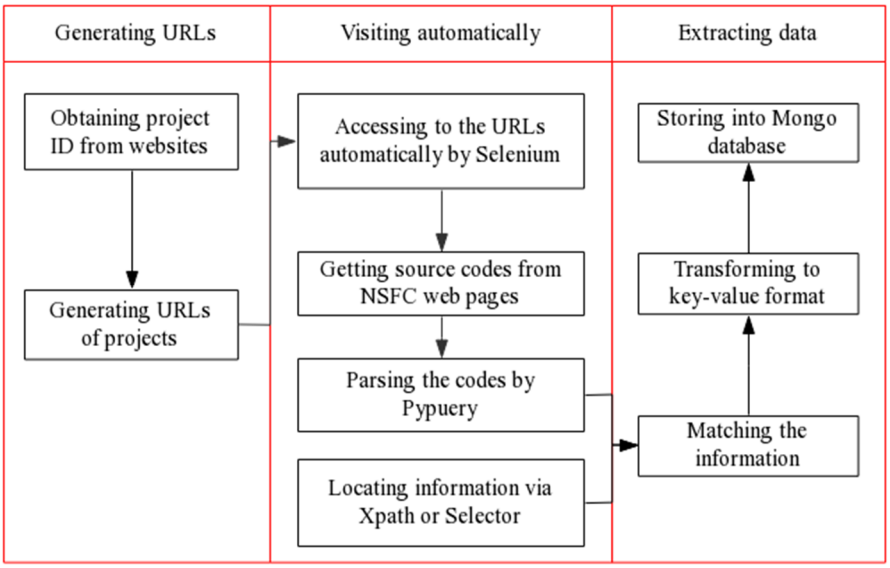

Figure 3. Automatic data collection process using a web crawler.

For each project, the following data are collected using our crawler: (1) structured data: project ID, project category, project funding, evaluation result and project outputs, and (2) unstructured data: project abstract in text format and concluding report in the form of pictures. These features are mainly selected according to the NSFC post-evaluation index system [23].

In addition, we also developed crawlers to collect journal impact factor data from the website of the China National Knowledge Infrastructure (CNKI, www.cnki.net accessed on 20 December 2019), Chaoxing Journal Network (http://qikan.chaoxing.com/, accessed on 21 December 2019) and Core journal query system (http://corejournal.lib.sjtu.edu.cn/, accessed on 25 December 2019). Then, we calculated the average impact factor for each project based on the journals that the project publishes its results in. The average impact factors are combined with the above-mentioned structured data to form the first part D_1 of the original dataset.

3.2. Extracting Information from Unstructured Data

In this section we process the collected unstructured data (the abstract texts and concluding report images) to extract structured data.

(1) Extracting information from texts. Because it is difficult to utilize the texts directly, we need to transform the abstract texts to numeric features that describe the text's information [24,25]. We adopted the TF-IDF model to construct word vectors. For each project's abstract in dataset D_0, the corresponding text features are represented by vector C with many dimensions. We add C to our original dataset and form a new dataset D_{text}. Text

mining produces a large number of features. This brings a large operating load when building prediction models. Therefore, text mining is only used as an auxiliary method to improve the predictive ability after our prediction model is built.

(2) Extracting information from images. To extract information from project conclusion reports that are in the format of images, we use the Tesseract library to recognize the text in pictures [26]. Tesseract library is able to extract single-line texts from images. However, in the concluding reports of NSFC, many texts are placed in fixed-form tables that cannot be recognized by the Tesseract library. To solve this problem, we designed a method that clips and cuts tables based on pixel positioning. The results of random sampling and inspection show that this method is valid. The features that can be extracted from the concluding reports include: the number of published papers, which databases the papers are indexed, talent training and international exchanges, etc. The above collected data are stored in dataset D_2.

3.3. Data Cleaning and Fusion

We cleaned datasets D_1 and D_2 by removing duplicate values, blank values and outliers. Then, we used "Project ID" as the primary key to merge D_1 and D_2 into the final dataset D. In D, the label $y_i \in Y = \{\text{Premium, Excellent, Good, Average}\}$. Note that there is no project evaluated as "Poor" in years 2016 and 2017, so we do not consider "Poor" in this paper. For the sake of simplicity, Premium, Excellent, Good and Average are coded as 4, 3, 2 and 1, respectively. During data fusion, for features that are highly correlated, we only preserved one feature. For features that have zero values in more than 90% instances, if the corresponding label values are similar, then we removed such features. Table 1 shows the features and the label in dataset D. Note that the timing of using our performance evaluation method is after the project is finished, so the data listed in Table 1 are collected after the projects are finished.

Table 1. Features and label in dataset D.

Features	**Description**
Project ID	Unique identifier and the primary key
Project funding	Expenditure of a project (ten thousand RMB); continuous variable
Number of papers	Number of papers published by a project; continuous variable
Average impact factors (English)	Average of impact factors of English journals where a project has published papers; continuous variable
Average impact factors (Chinese)	Average impact factors (Chinese)
Weighed impact factors	Average weighted impact factors of all journals where a project has published papers, the number of SCI papers is the weight; continuous variable
Papers published in English	Number of papers published in English in a project; continuous variable
Papers published in Chinese	Number of papers published in Chinese in a project; continuous variable
Academic reports	Number of academic reports; continuous variable
Journals	Number of papers published in journals; continuous variable
Conference papers	Number of papers published in conference proceedings; continuous variable
SCI papers	Number of papers published in SCI-indexed journals; continuous variable
EI papers	Number of papers published in EI-indexed journals; continuous variable
PKU papers	Number of papers published in Core journals of Peking University; continuous variable
CSSCI papers	Number of papers published in CSSCI-indexed journals; continuous variable
Number of doctors	Number of doctors cultivated in a project; continuous variable
Number of masters	Number of masters cultivated in a project; continuous variable
International meeting	Number of international conferences attended; continuous variable
Awards	Number of awards obtained by a project; continuous variable
Last evaluation results	Applicant's last project evaluation results; discrete variable
Text	Features related to text
Label	**Description**
Evaluation results	Evaluation results of a project given by experts, involving 5 levels: Premium, Excellent, Good, Average and Poor

3.4. Descriptive Statistics for the Collected Data

There are 1199 samples in dataset D. Table 2 shows the descriptive statistics for dataset D. The second and third columns of Table 2 show the quantity and proportion of projects according to the project evaluation results. It can be seen that most of the projects are evaluated as "Excellent" or "Good" level, and the proportions of both levels are similar. Table 2 also shows the mean values of some representative features. We can see that these values vary among different evaluation result levels. For the projects with Premium level, they have higher mean values in these features, especially in features of the number of SCI papers and the number of papers published in English.

Table 2. Comparison of the mean values of some features in different classes.

Evaluation Results	Count	Frequency	Mean Value of Representative Features				
			Number of Papers	Number of Papers Published in English	SCI Papers	Number of Doctors	Number of Masters
Premium	51	4.25%	27.92	16.39	13.35	2.15	3.67
Excellent	492	41.03%	23.43	7.61	5.12	1.53	4.71
Good	632	52.71%	18.70	3.76	1.67	1.05	4.40
Average	24	2.00%	8.87	1.50	0.13	0.58	4.33

Figure 4 shows the word cloud for project abstracts in English. In Figure 4, the larger the font size, the more frequently the word appears. In Figure 4, "Supply chain", "Enterprise" and "Model" are popular keywords in the Management Science Department of NSFC.

Figure 4. Word cloud for project abstracts.

4. Model Training

4.1. Dealing with Imbalanced Data

We can see from Table 2 that the distribution of label values is imbalanced, which tends to cause the prediction results of many classification algorithms to be very poor. To deal with this problem, two strategies exist: down-sampling and over-sampling. In the down-sampling method, samples belonging to the major classes will be decreased to a level similar to the minor classes [27]. However, there are some flaws in the down-sampling method. For example, the deleted samples of major classes may contain critical information in classification, which may lead to information losses; the trained model may lack universality, enlarging the characteristics of the minor classes. In the over-sampling methods, SMOTE (Synthetic Minority Oversampling Technique) [28] is a well-accepted method to deal with imbalanced data. Therefore, we used the SMOTE algorithm to process

the dataset [21] before training a prediction model. By using the nest-neighbor algorithm, SMOTE generates new small class samples that are added to the dataset. In doing so, the obtained model may have a higher prediction ability [28].

4.2. Measures for Model Performance Evaluation

Based on the true and predicted values of the labels, we constructed a confusion matrix $C = (c_{yy\prime})_{4\times4}$, where the element $c_{yy\prime}$ denotes the number of projects whose true label value is y are predicted to be y' ($y, y\prime \in Y$). Based on the confusion matrix, we used the following four measures to evaluate our performance prediction models [29].

(1) Accuracy (A). Accuracy represents the proportion of correctly classified samples to the total samples. The accuracy of model f on dataset D is calculated as follows:

$$A(f;D) = \frac{1}{m}\sum_{y=1}^{4}\sum_{y\prime=1}^{4} I(y = y\prime) \times c_{yy\prime} \tag{1}$$

where m is the number of projects in dataset D. The indicator function $I(\cdot)$ equals 1 if the calculation result in the brackets holds, otherwise it equals 0. The larger the value of accuracy, the higher the accuracy of the prediction.

(2) Recall (R). Recall is the ratio of the number of samples correctly predicted as a certain class to the total number of samples with that class in the dataset. Higher value of the recall means a better model. In our multi-classification problem, we average the recall of each class as follows:

$$R(f;D) = \frac{1}{4}\sum_{y=1}^{4}\frac{\sum_{y\prime=1}^{4} I(y = y\prime) \times c_{yy\prime}}{\sum_{y\prime=1}^{4} c_{yy\prime}} \tag{2}$$

(3) Precision (P). Precision is the ratio between the projects that are correctly classified and the projects that are classified to the corresponding class. The higher the precision is, the higher the proportion of samples whose predicted results are consistent with the experts' judgments. The precision of model f is calculated as follows:

$$P(f;D) = \frac{1}{4}\sum_{y'=1}^{4}\frac{\sum_{y=1}^{4} I(y = y') \times c_{yy'}}{\sum_{y=1}^{4} c_{yy'}} \tag{3}$$

(4) $F1$ score. Recall and precision are contradictory. When a model's precision is high, the recall is usually low, and vice versa. Therefore, we use $F1$ score that considers both precision and recall. The higher the $F1$ score is, the better performance the model has. The $F1$ score of model f obtained on dataset D is calculated as follows:

$$F1(f;D) = \frac{2 \times P \times R}{P + R} \tag{4}$$

4.3. Model Selection

To obtain an appropriate multi-classification algorithm, we tested 11 algorithms by calling the Scikit-learn library in Python [30]. Dataset D is divided into training and test set with a proportion of 7:3. This results in 839 samples in the training set and 360 samples in the test set. Based on the training set, the 11 classification algorithms are used to train 11 candidate prediction models. The confusion matrixes and performance measures of these models that are calculated on the test set are shown in Figure 5 and Table 3, respectively. Each subgraph in Figure 5 corresponds to the confusion matrix of the prediction result obtained by an algorithm. In each confusion matrix, the X-axis represents the result of the model prediction, the Y-axis represents the true labels, and the number in each cell represents the number of projects with the true label y predicted as y' (so the elements

in the main diagonal of the matrix represent the number of correct results predicted by the model).

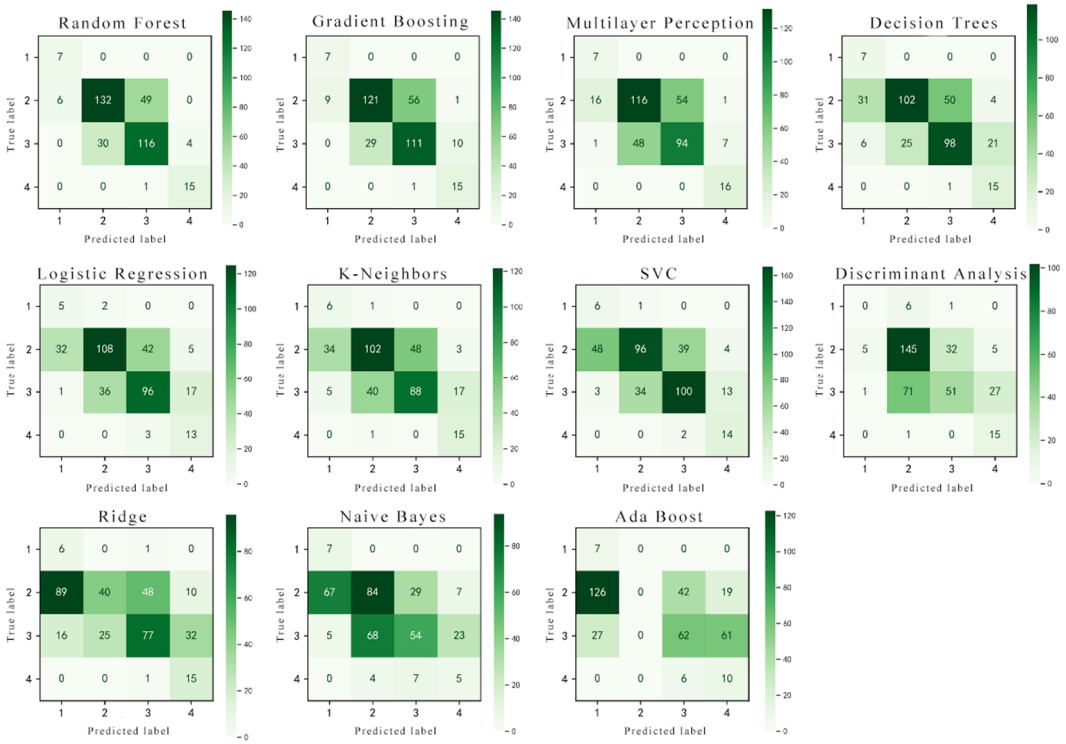

Figure 5. Confusion matrixes for 11 classification algorithms.

Table 3. Performance measures for 11 classification algorithms.

No.	Algorithms	Accuracy	Recall	Precision	*F*1 Score
1	Random Forest	***0.75***	***0.85***	***0.71***	***0.76***
2	Gradient Boosting	0.71	0.83	0.62	0.68
3	Multilayer Perception	0.65	0.81	0.58	0.64
4	Decision Trees	0.62	0.78	0.50	0.53
5	Logistic Regression	0.60	0.73	0.50	0.52
6	K-Neighbors	0.59	0.73	0.48	0.51
7	SVC	0.58	0.75	0.43	0.44
8	Discriminant Analysis	0.59	0.51	0.39	0.40
9	Ridge	0.38	0.63	0.38	0.35
10	Naive Bayes	0.42	0.53	0.34	0.32
11	Ada Boost	0.22	0.51	0.18	0.19

It can be seen from Figure 5 and Table 3 that the Random Forest model (Row 1, Column 1) performs best. Therefore, the following analysis will be based on the Random Forest model. To further improve the performance of the Random Forest, we used a grid search to tune its parameters and a 10-fold cross validation to ensure the effectiveness of the model. In the grid search, the number of sub-classifiers is set from 300 to 1000 with 100 as the interval, and the maximum depth is set between 40 and 110 with 10 as the interval. The minimum number of samples for leaf nodes is 2, 3, 4, 5, and 10. The minimum number of samples for internal nodes is 1, 2, 3, 4, and the remaining parameters adopt the default values. The final obtained best model f_{best} has the following parameter settings: the number of sub-classifiers is 900; the maximum depth is 110; the minimum number of samples for leaf nodes is 2; the minimum number of samples for internal nodes is 2.

5. Results and Discussion

5.1. Main Results

The prediction results obtained with f_{best} are as follows: the accuracy A is 0.78, the recall R is 0.87, the precision P is 0.73 and the $F1$ score is 0.78. Next, we further examined the performance of f_{best}. We divided the test set into four parts according to the labels, and the prediction results of f_{best} in each part are shown in Figure 6. Most of the results of f_{best} are consistent with the experts. For each class, f_{best} produces satisfactory classification results. f_{best} performs particularly well in the projects with "Average" and "Premium" labels.

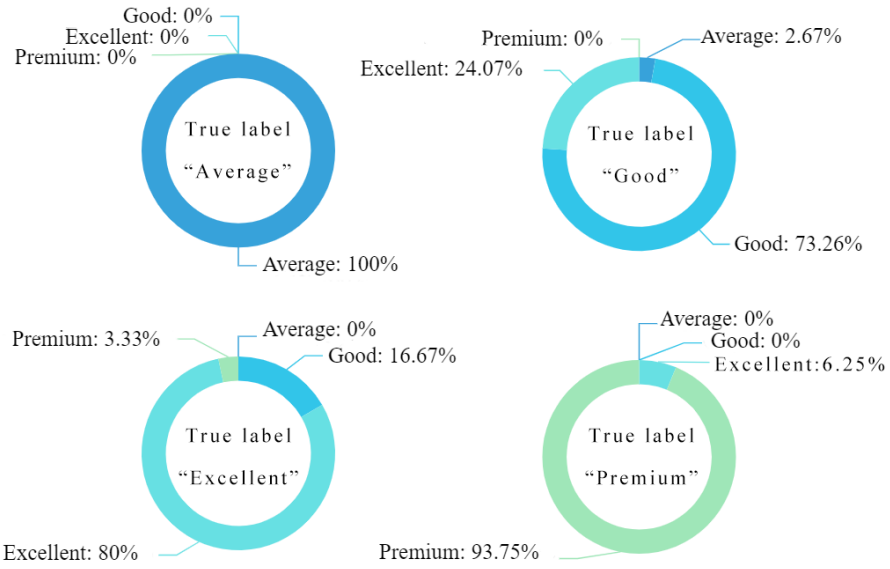

Figure 6. Prediction results of the Random Forest for different classes.

Relatively speaking, the Random Forest's ability to classify "Excellent" and "Good" is not as good as the other two classes. The reasons may be attributed to: (1) the features used by the model are not as rich as that of the experts. The data used in the model in this paper are public and all collected from the internet. However, in addition to the public data, there are many undisclosed dimensions in the project, which causes our model to be inferior to the expert review process to a certain extent. (2) Experts have other subjective factors of the projects in their judgment. It can be seen from the descriptive statistics in Section 3.4 that the two more extreme performance levels of "Premium" and "Average" have their own characteristics in each dimension. Among them, the performance of "Premium" in all dimensions is prominent. "Average" is obviously lagging behind other classes in all dimensions, while the two categories of "Excellent" and "Good" are in the middle of the four categories, and there will be little distinction between them. For two projects with similar achievement levels, there may be cases where the objective scores are similar, and experts believe that one of the projects has higher value and significance, or is more innovative, and thus subjectively prefer that project. This is what a machine learning algorithm can hardly learn.

In addition, in our data, the evaluation score (label) is ordinal, which means that misclassifying different classes has different costs. For example, misclassifying "Premium" as "Average" is worse than misclassifying it as "Excellent". Therefore, we perform an additional experiment using ordinal regression that is able to deal with ordinal variables. To compare the ordinal regression with the previous Random Forest, we reported the results based on the total cost TC of classification. Specifically, we first introduced a cost matrix $C\prime = \left(c\prime_{yy\prime}\right)_{4\times4}$, where its element $c\prime_{yy\prime}$ represented the cost of classifying y to $y\prime$ (Figure 7). Then, the total cost $TC = \sum_{y=1}^{4}\sum_{y\prime=1}^{4} c_{yy\prime} \times c'_{yy\prime}$. A lower value of TC means a better model. The results of the additional experiment are as follows: the ordinal regression

model has a total cost of 99, while the total cost of the Random Forest f_{best} is 81. This means that the previous Random Forrest model achieves better classification results than the ordinal regression model.

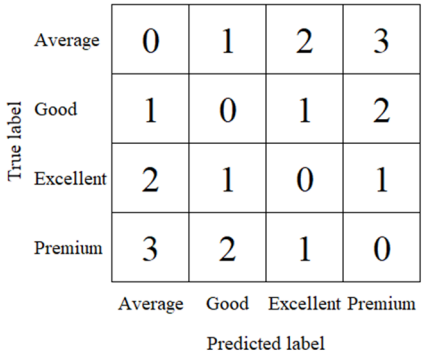

Figure 7. Classification cost matrix.

5.2. The Importance of Features

In this subsection, we discuss the impact of different features on the prediction results. Given a feature, we modify its value into a set of noise values. The importance of the feature is obtained by comparing the prediction results before and after the modification [31]. Specifically, this process is as follows:

Input: prediction model f, training set D_{train}, pre-determined parameter K.
step 1: compute the *f1* score of f.
step 2: for each feature j in D_{train}:
  let $k = 1$.
  step 2.1: while $k \leq K$, repeat:
    randomly disturb feature j to generate dataset $\widetilde{D_{k,j}}$.
    compute the *f1* score $F1_{k,j}$ of model f on dataset $\widetilde{D_{k,j}}$.
  $k = k + 1$.
  step 2.2: the importance i_j of feature j is calculated as:

$$i_j = F1 - \frac{1}{K}\sum_{k=1}^{K} F1_{k,j} \tag{5}$$

Output: the importance i_j for each feature j.

Figure 8 shows the top five most important features. It can be inferred that in our performance evaluation model, the weighed impact factors, the number of CSSCI papers and the features representing the academic achievement are important factors. The talent training is also a main factor considered by the machine learning model. Furthermore, the machine learning model also believes that the funding of the project and the applicant's previous project evaluation results are also a major influencing factor.

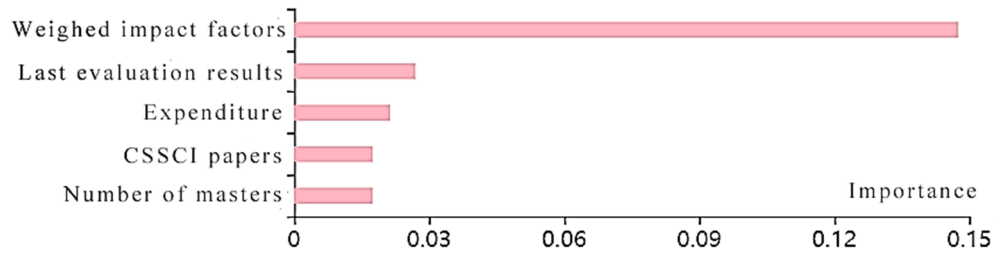

Figure 8. Top 5 important features.

5.3. Results after Considering Text Data

In Section 3.2, we constructed a dataset D_{text} that contains text data of project abstracts. In our previous experiments, we did not use D_{text}. In this subsection, we combine D_{text} with the previous dataset D, and repeat the previous data analysis process using the Random Forest algorithm to examine whether the text data can improve the performance of the prediction model. The results are shown in Table 4. It can be seen that after adding the text features, there is no obvious improvement. On the contrary, the text features negatively affect the model. Therefore, the text features of the project abstract do not play a significant role in classification. Whether the text features are added or not has little effect on the conclusions of this paper.

Table 4. Model performance comparison after considering text data using Random Forest.

Dataset	Accuracy	Recall	Precision	*F*1 Score
D	0.78	0.87	0.73	0.78
$D_{text} \cup D$	0.76	0.86	0.71	0.76

In summary, the experimental results verify the effectiveness of our data-driven R&D project performance evaluation framework, and our method has high predictive accuracy. For public R&D management departments, our model can automatically make a pre-evaluation on the performance level for newly completed projects. This helps to reduce the workload of management departments and human experts, and improve the efficiency of project evaluation.

6. Case Study: Applying Our Framework to Logistics and Supply Chain Project Evaluation

Our framework can be used to evaluate different projects with various topics. As an example, in this section, we provide a case study that applies our framework to evaluate the performance of logistics and supply chain related projects.

In 2015, a total of 29 logistics and supply chain related projects under the NSFC logistics and supply chain related code "G010303/G0212" were finished. Therefore, we extract features from these projects based on our framework. Finally, based on the features, we used our model f_{best} (see Section 4.3) to output the evaluation results for each project.

Since the 29 projects have been evaluated by experts, we compared the results obtained by our model with the experts. Among these 29 projects, 19 projects received results consistent with experts. In addition, although there are 10 projects that are wrongly evaluated by our model, our results are very close to the experts. Specifically, six (3, 1) projects that are evaluated as "Good" ("Excellent", "Premium") by the experts are classified as "Excellent" ("Good", "Excellent") by our model.

The above results reveal that facing the logistics and supply chain related projects, our framework performs well and is able to give reasonable evaluation results.

7. Conclusions

We propose a data-driven public R&D project performance evaluation framework that can effectively estimate the project performance based on multi-classification supervised learning algorithms. Web crawlers are designed to automatically collect project data from different websites. The collected structured and unstructured data are processed. Effective features are extracted from the resulting dataset. A total of 11 classification algorithms are used to construct the evaluation model. We also use different measures to estimate the performance of the model. Additionally, we also provide a case study that applies our framework to evaluate the project performance in the logistics and supply chain area of NSFC.

We validate our method based on the data of NSFC projects. The results show that among the 11 classification algorithms, the Random Forest algorithm performs best. We

also compare our model with ordinal regression that takes the ordinal categorical variables into consideration. The project performance is mainly affected by the weighted SCI impact factors, the applicant's last project evaluation results and the talent training related features. The abstract related text features have no obvious impact on the evaluation results. Case study shows that our model has a good performance when evaluating the logistics and supply chain projects of NSFC.

Our method provides a unified and scalable framework for public R&D project performance evaluation. It helps to improve the automation and intelligence level in project evaluation and support the data-driven decision-making for the management departments. Future research work will further integrate richer data and explore more effective machine learning and imbalanced data pre-processing algorithms.

Author Contributions: Conceptualization, H.L.; Data curation, H.L., B.Y. and X.Y.; Formal analysis, H.L., B.Y. and X.Y.; Funding acquisition, H.L.; Investigation, H.L., B.Y. and X.Y.; Methodology, H.L., B.Y. and X.Y.; Project administration, H.L. and B.Y.; Resources, H.L. and X.Y.; Software, H.L., B.Y. and X.Y.; Supervision, H.L.; Validation, H.L.; Visualization, H.L. and B.Y.; Writing—original draft, H.L., B.Y. and X.Y.; Writing—review & editing, H.L., B.Y. and X.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (Grant Number 71602106) and the Key Soft Science Project of Shanghai Science and Technology Innovation Action Plan (Grant Number 20692192400).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available data were collected and analyzed in this study. The data can be found here: http://kd.nsfc.gov.cn/ (accessed on 20 December 2019).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Zhou, J.; Yang, L.; Xu, Z. A study on review to achieved research projects financed by the department of management sciences of NSFC. *Manag. Rev.* **2007**, *19*, 13–19. [CrossRef]
2. Liu, C. A study for allocating resources to research and development programs by integrated fuzzy DEA and fuzzy AHP. *Sci. Res. Essays* **2011**, *6*, 3973–3978. [CrossRef]
3. Eilat, H.; Golany, B.; Shtub, A. R&D project evaluation: An integrated DEA and balanced scorecard approach. *Omega* **2008**, *36*, 895–912. [CrossRef]
4. Florescu, M.; Davidescu, A.; Mosora, M.; Alpopi, C.; Nastase, M. Assessment of the research field in the European universities and analysis of the research projects impact on academic performance. *Ind. Text.* **2019**, *70*, 587–596. [CrossRef]
5. Gao, J.-P.; Su, C.; Wang, H.-Y.; Zhai, L.-H.; Pan, Y.-T. Research fund evaluation based on academic publication output analysis: The case of Chinese research fund evaluation. *Scientometrics* **2019**, *119*, 959–972. [CrossRef]
6. Sun, W.; Tang, J.; Bai, C. Evaluation of university project based on partial least squares and dynamic back propagation neural network group. *IEEE Access* **2019**, *7*, 69494–69503. [CrossRef]
7. Uzbay, T. Two new factors for the evaluation of scientific performance: U and U′. *Turk. J. Pharm. Sci.* **2019**, *16*, 115–118. [CrossRef] [PubMed]
8. Zhu, W.; Li, S.; Ku, Q.; Zhang, C. Evaluation information fusion of scientific research project based on evidential reasoning approach under two-dimensional frames of discernment. *IEEE Access* **2020**, *8*, 8087–8100. [CrossRef]
9. Park, J.; Kim, J.; Sung, S.-I. Performance evaluation of research and business development: A case study of Korean public organizations. *Sustainability* **2017**, *9*, 2297. [CrossRef]
10. Kim, W.S.; Park, K.; Lee, S.H.; Kim, H. R&D investments and firm value: Evidence from China. *Sustainability* **2018**, *10*, 4133. [CrossRef]
11. Ghapanchi, A.H.; Tavana, M.; Khakbaz, M.H.; Low, G. A methodology for selecting portfolios of projects with interactions and under uncertainty. *Int. J. Proj. Manag.* **2012**, *30*, 791–803. [CrossRef]
12. Karasakal, E.; Aker, P. A multicriteria sorting approach based on data envelopment analysis for R&D project selection problem. *Omega* **2017**, *73*, 79–92. [CrossRef]
13. Hsu, F.-M.; Hsueh, C.-C. Measuring relative efficiency of government-sponsored R&D projects: A three-stage approach. *Eval. Program. Plan.* **2009**, *32*, 178–186. [CrossRef] [PubMed]

14. Johnes, J.; Yu, L. Measuring the research performance of Chinese higher education institutions using data envelopment analysis. *China Econ. Rev.* **2008**, *19*, 679–696. [CrossRef]
15. Wang, X.; Zhang, S.; Liu, Y.; Qiao, Y.; Han, X.; Huang, H. Forty Years of Research on Science and Technology Evaluation in China: Historical and Theme evolution. *Sci. Sci. Manag. S T* **2018**, *39*, 67–80.
16. Cho, J.-H.; Lee, K.-W.; Son, H.-M.; Kim, H.-S. A study on framework for effective R&D performance analysis of Korea using the Bayesian network and pairwise comparison of AHP. *J. Supercomput.* **2013**, *65*, 593–611. [CrossRef]
17. Liu, Z.; Hu, H. SDM Techniques Based on TGSOM and its Application in R&D Performance Evaluation. In Proceedings of the 2009 Second International Workshop on Knowledge Discovery and Data Mining, Moscow, Russia, 23–25 January 2009; IEEE Computer Society: Washington, DC, USA, 2009; pp. 372–375. [CrossRef]
18. Costantino, F.; Di Gravio, G.; Nonino, F. Project selection in project portfolio management: An artificial neural network model based on critical success factors. *Int. J. Proj. Manag.* **2015**, *33*, 1744–1754. [CrossRef]
19. Liu, F.; Yang, J.-B.; Xu, D.-L.; Liu, W. Solving multiple-criteria R&D project selection problems with a data-driven evidential reasoning rule. *Int. J. Proj. Manag.* **2019**, *37*, 87–97. [CrossRef]
20. Jang, H. A decision support framework for robust R&D budget allocation using machine learning and optimization. *Decis. Support Syst.* **2019**, *121*, 1–12. [CrossRef]
21. Simsek, S.; Kursuncu, U.; Kibis, E.; AnisAbdellatif, M.; Dag, A. A hybrid data mining approach for identifying the temporal effects of variables associated with breast cancer survival. *Expert Syst. Appl.* **2020**, *139*, 112863. [CrossRef]
22. You, F.; Gong, H.; Guan, X.; Cao, Y.; Zhang, C.; Lai, S.; Zhao, Y. Design of Data Mining of WeChat Public Platform Based on Python. In Proceedings of the 3rd Annual International Conference on Information System and Artificial Intelligence, Suzhou, China, 22–24 June 2018.
23. Chen, X.; Huang, H.; Li, R. Afterwards evaluation-an effective way of strengthening the management of supported projects by NSFC. *Bull. Natl. Nat. Sci. Found. China* **2004**, *18*, 186–188. [CrossRef]
24. Singhal, A.; Kasturi, R.; Srivastava, J. Automating Document Annotation Using Open Source Knowledge. In Proceedings of the 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT), Atlanta, GA, USA, 17–20 November 2013; IEEE Computer Society: Washington, DC, USA, 2013; pp. 199–204. [CrossRef]
25. Singhal, A.; Srivastava, J. Research dataset discovery from research publications using web context. *Web Intell.* **2017**, *15*, 81–99. [CrossRef]
26. Pavaskar, A.V.; Achha, A.S.; Desai, A.R.; Darshan, K.L. Information extraction from images using Pytesseract and NLTK. *J. Emerg. Technol. Innov. Res.* **2017**, *4*, 83–84.
27. Thiruvadi, S.; Patel, S.C. Survey of data-mining techniques used in Fraud detection and prevention. *Inf. Technol. J.* **2011**, *10*, 710–716. [CrossRef]
28. Chawla, N.; Bowyer, K.; Hall, L.; Kegelmeyer, W. SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [CrossRef]
29. Agarwal, S. *Data Mining: Data Mining Concepts and Techniques*; IEEE: New York, NY, USA, 2013; pp. 203–207.
30. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
31. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]