Application of machine learning technology to analyze the probability of winning a tender for a project

Nikita Kultin, docente, kultin_nb@spbstu.ru
Danila Kultin, Master of Engineering and Technology, kultin.dn@edu.spb.stu.ru
Roman Bauer, Bachelor of Engineering and Technology, bauer.rv@edu.spb.stu.ru

Peter the Great Saint-Petersburg Polytechnic University, Russia, 195251, Saint-Petersburg, Polytechnicheskaya, 29

Abstract. The possibility of using machine learning technology to solve the problem of project analysis in order to support the decision to participate in the tender for the implementation of the project is substantiated and shown using a specific example. The approaches are described and the process of solving the problem of binary classification of projects using libraries of the Python language is shown. Attention is paid to the problem of choosing an algorithm for constructing a membership function, the problem of generating and analyzing input data, and evaluating the accuracy of a solution. It is shown that for the considered problem the best solution is provided by the logistic regression algorithm.

Key words:
machine learning; artificial intelligence; project risk analysis; project management; tender for the implementation of the project.

Introduction. In modern conditions, many enterprises in various industries are guided by the implementation of projects whose customer is the “state”. In accordance with the legislation of the Russian Federation, the executors of such projects are selected based on the results of open tenders. Participation in the tender implies a significant amount of preparatory work related to the preparation of tender documentation, including the development of a technical and economic proposal. Since the result of the tender may be a loss (failure to receive an order for the project), the costs associated with preparing for the tender can be considered as risky. It is possible to reduce the likelihood of losses associated with losing a tender by “sifting” projects at the first stage of preparation for participation in the tender by analyzing them in detail before making the final decision to participate in the tender and starting work on the development of a feasibility study. It is possible to realize the “sifting” of projects on the basis of machine learning technology, solving the classification problem. As a result of solving the classification problem, the analyzed project can be classified as “prospective”, in this case, it is advisable to continue the preparation for bidding, or “unpromising”.

The purpose of this study is to justify the possibility of using machine learning technology to solve the problem of analyzing projects with the aim of deciding whether to participate in a tender for its implementation, to develop an algorithm for classifying projects to assess the possibility of obtaining a contract for a project.

Materials and research methods. To assess the possibility of obtaining a contract for an engineering project, most companies use qualitative risk analysis methods. In practice, the most widely used method of expert assessments and the method of analogy [1]. The advantages and disadvantages of these methods are shown in Table 1.

Table 1. Qualitative risk analysis methods of projects

Method	Advantages	Disadvantages
Expert assessment method	- Use of reliable knowledge - Simple calculations	- Limited expert pool - Subjective ratings
Method of analogy	- Use of reliable knowledge - Using knowledge inside the company	- Low forecast accuracy - Lack of consideration of individual project features

Conducting a quantitative risk assessment of the pre-project stage of work requires significant costs. In order to reduce the cost of assessing the risks of the pre-project stage of work, it is proposed to assess the likelihood of obtaining a contract for the implementation of the project by solving the binary classification problem.

To solve the binary classification problem, machine learning methods will be used. Based on the existing characteristics of the projects, the algorithm will construct a separating hyperplane [2] between projects of the classes “prospective” (the probability of obtaining a contract is high) and “unpromising” (the probability of obtaining a contract is low).

To achieve this goal it is necessary to solve the following tasks:

- Perform statistical analysis of data on the characteristics of projects
- Based on the results of the analysis, select an adequate family of algorithms
- From the family of algorithms based on the quality criterion, select the best algorithm
- Using training and test data samples, determine the quality indicator of the algorithm

Project attributes. A large number of factors influence the fact of obtaining a contract for the implementation of the project. Information about the degree of influence of each factor is recorded in the vector of project attributes. Each component of the vector is a sign that describes the project, as well as the relationship of the customer and the contractor in the framework of this project. Thus, the vector of project attributes is a description of the pre-project stage of work.

Each attribute of the project must be associated with a numerical value and put it in the corresponding component of the vector. However, not all project characteristics can be expressed numerically. There are categorical (one value from a finite set), binary (value 1 or 0) and other complex types (date, time, geographical location). Complex types can be represented by a set of simple features with a numerical representation [3]. To bring categorical features to numerical representation, the binarization method was used [4]. Each value from the set of values of a categorical attribute is associated with its own attribute. For example, the set of “Upcoming project work” is associated with the signs “Start-up and adjustment work”, “Design and survey work”, etc. If the type of work corresponding to the characteristic needs to be performed within the framework of the project, then the characteristic takes the value “1”, otherwise “0”. Methods for converting features are presented in Table 2.

Table 2. Methods for converting of attributes

Characteristic category	Characteristic name	Influence factor	Representation type	Numeric presentation method
Attributes of the project	Preliminary budget (budget)	Preliminary assessment of income from the implementation of project	Numerical	Value of the budget specified by the customer
	Type of work (kw	Assessment of the	Categorical	Binarization

	_...)	complexity of work on the project		
	Preliminary project implementation period (rz_date)	Duration estimation	Date and time	The number of days between pre-specified project start and end dates
	Place of implementation (rz_place)	Assessment of travel expenses, climate, infrastructure	Geographic	Latitude and longitude (GPS coordinates). It is represented by two real values
Characteristics of the executing company	Project manager (app_manager)	Evaluation of the impact of the compiled project description on further work	Numeric	Number of contracts for applications out of accepted by the manager / Number of all applications accepted by him
	Sales Manager (sale_manager)	Evaluation of work experience	Numeric	Number of applications that the manager in the company conducted
	Responsible Manager (res_manager)	Assessment of experience and specialization of the manager	Categorical	Binarization
Characteristics of the customer's company	Client's office (cl_office)	Estimation of expenses for business trips	Numerical	Distance to the client's office in kilometers
	The legal form of the client's company (kc _...)	Assessment of the complexity of interaction between companies	Categorical	Binarization
Other attributes	Collaboration efficiency (coop_eff)	Assessment of the impact of past experience in collaboration	Numeric	Percentage of implemented joint projects from all attempts to work with a specified customer
	Application submission date (app_date)	Estimation of the urgency of the project	Date/time	The number of days between the date of submission of the application and the preliminary start date of the project
Target attribute	Contract conclusion (is_proj)	Whether the contract for the implementation of the project was concluded	Binary	Takes the value "1" if the contract was concluded, "0" if it was not

The set of vectors corresponding to the projects is a matrix, the rows of which are the sets of values of the characteristics of a particular project, and the columns are the sets of values of a specific attribute in all projects of an engineering company [3, 4].

Statistical analysis of attributes. In order to assess the quality of the source data and make

sure that they can be used to assess the possibility of obtaining a contract for the implementation of the project, it is necessary to perform a statistical analysis of the characteristics. For convenient presentation of the matrix of traits and subsequent statistical studies of individual traits, the Pandas, Numpy, Seaborn libraries of the Python language were used [3].

To obtain a reliable picture, it is necessary to exclude duplication of information in the source data. To solve this problem, it is necessary to construct a matrix of pair correlations of attributes. If the value of the correlation coefficient of attributes modulo more than 0.4, it is concluded that the information in these signs is duplicated and one of the columns of the matrix should be deleted [5]. After this, it is necessary to evaluate the type of data distribution in individual features and the presence of anomalous values. In fig. Figure 1 shows histograms of the distribution of features. The given assessment is taken into account when choosing the parameters of the algorithm for training.

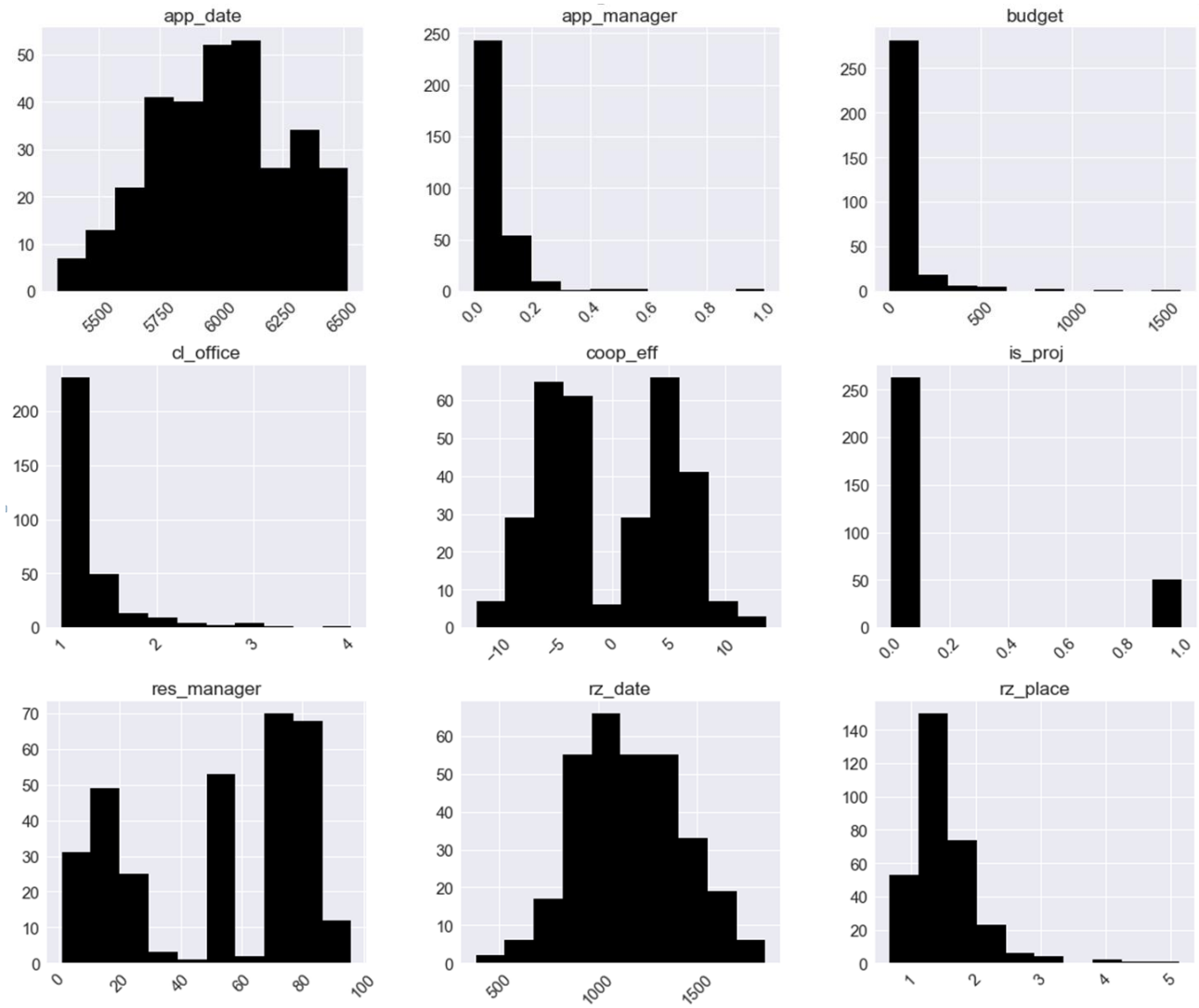

Figure 1. Attribute value distribution charts

It is also necessary to normalize all numerical signs so that when using these data in training, the relative weights of the signs are not biased. Normalization is performed according to the formula$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}}$ [4].

Building a learning algorithm. After selecting and analyzing the characteristics, we obtain a certain set X and its subset X^l. This subset X^l is the set of all known pairs $(x_i, y_i)_{i=1}^{l}$, where x_i- is a characteristic description of a specific project of the company from the set of projects X, and y_i- is

one of two values of the target characteristic from the set of possible responses $Y = \{0, 1\}$, corresponding to the response on the object x_i [6].

There is a target function $y^*: X \to Y$, that must be restored from known values on the set X^l. The task of learning the algorithm is to build a decisive function $a: X \to Y$, that would approximate the objective function $y^*(x)$, and not only on objects of the set X^l, but on the whole set X [6].

To apply machine learning algorithms, we divide the set X into three subsets: a training set, a test set, and objects without answers. Elements of the training set will be used to solve the learning problem and are elements of the subset X^l but do not participate in the learning of the algorithm. They help identify the problem of retraining. Retraining refers to a situation where the quality of the algorithm on data that did not participate in training drops sharply. To avoid the problem of retraining, the subset X^l needs to be divided into training and test samples so that the ratio of the values of the target attribute in the objects of both samples is the same.

When analyzing the literature on the research topic, it was revealed that the set task is well solved by logistic regression algorithms and support vector methods [7, 8]. The implementation of these algorithms was taken from the ScikitLearn library of the Python language.

To obtain the decisive function a, we will minimize the error functional Q on the training set of 295 vectors. The type of error functional for logistic regression and the support vector method are presented in Table 3 [3, 6, 9].

Table 3. Error functional of selected algorithms

	Logistic regression method	Support vector method
Error functional	$\frac{1}{2}ww^T + C\sum_{j=1}^{n}\log(\exp\left(-y_i\left(x_i^T w + b\right)\right) + 1)$, x_i- algorithm response, y_i- valid answer, w - weights vector, b - distance between dividing plane and origin, C - penalty for total error, includes L2 regularization	$\frac{1}{2}ww^T + C\sum_{i=1}^{n}\zeta_i$, on condition $y_i(w^T\phi(x_i) + b) \geq 1 - \zeta_i$. x_i, y_i, w, b и C have the same meaning as in logistic regression, ζ_i- error value at each object.

Three metrics were used to evaluate the algorithms: accuracy (T), completeness (P), f-measure (F). Accuracy shows the proportion of class objects correctly classified by the decisive function among all objects that the function has assigned to this class. Completeness shows the proportion of class objects correctly classified by the decisive function among all objects of this class in the test sample. The F-measure is the harmonic mean between accuracy and completeness. The values of the quality metrics of the algorithms in the test set are presented in Table 4.

Table 4. Comparison of algorithm quality metrics

	Logistic regression method			Support vector method (polynomial loss function)			Number of vectors
Metrics	T	P	F	T	P	F	
Value for class “1”	1.00	0.73	0.84	0.57	0.73	0.64	11
Value for class “0”	0.97	1.00	0.98	0.96	0.93	0.95	84
Average	0.97	0.97	0.97	0.92	0.91	0.91	95

Evaluation of the quality of the algorithm. To obtain a more objective assessment of the quality of the algorithms, cross-validation was conducted on 7 sets of test and training samples [10]. During the cross-validation process, the entire data set (matrix X^l) is divided into several parts - in

this case, seven. Next, one part of the data becomes a test sample, and the algorithm is trained on the remaining six. Then, accuracy is evaluated on the test sample and the result is recorded. Then the next of seven pieces of data becomes a test sample. In such an artificial way, the algorithm is checked for adaptability to data from different sets [3, 4]. The cross-validation results for the algorithms are shown in Table 5.

Table 5. Cross validation results

Quality ratings	1	2	3	4	5	6	7
Logistic regression method	0.95	0.93	0.95	0.95	0.90	0.86	0.88
Support vector method	0.91	0.93	0.95	0.95	0.93	0.86	0.86

Analysis of the data shows that the logistic regression algorithm is more suitable for a given data structure and better solves the problem.

The algorithm was tested on data on 53 projects, applications for which were received by the project department of the enterprise and were being processed at the time when the algorithm evaluated them using known parameters. After receiving information on the results of the work on these applications, quality metrics for evaluating the algorithm were measured. The measurement results are shown in Table 6.

Table 6. The result of testing the algorithm on real data

Metrics	T	P	F	Number of vectors
Value for class “1”	0.094	1.000	0.172	4
Value for class “0”	1.000	0.408	0.508	49

The test results show that the algorithm coped quite well with the solution of the problem, since all projects classified by the algorithm as unpromising are indeed such. At the same time, the algorithm correctly classified only 40% of unpromising projects. The algorithm has not missed a single truly promising project.

The probabilities of belonging to classes were also analyzed, which the algorithm evaluated when working on test data. It was found that these estimates are very high, which indicates excessive "confidence" of the algorithm.

Conclusions

1. Using machine learning technology to build a binary classification algorithm for projects is possible, but requires a significant amount of information about the characteristics of projects and the results of previous tenders.
2. The constructed classification algorithm can be adapted for use in any companies engaged in project activities. Using the algorithm will allow you to rank current projects, reduce the cost of the pre-project stage.
3. In the future, it is possible to refine the algorithm in the direction of expanding the number of features, increasing the accuracy of tuning the hyperparameters of the algorithm, which will make it possible to obtain more accurate and reasonable estimates, reduce the degree of confidence of the algorithm in assessing the probability of belonging to the class, and increase the accuracy metric for the class of promising projects while maintaining the

current completeness value.

References

1. Konstantinov G. N. Strategic management: concepts. - Moscou: Business-alignment, 2009. – 239 p. (in Russ.)
2. Principles of Machine Learning. Microsoft, 2018. Available at: https://www.edx.org/course/principles-machine-learning-microsoft-dat203-2x-6
3. Training on tagged data. National Research University Higher School of Economics, Yandex School of Data Analysis // Coursera. 2018. Available at: https://www.coursera.org/learn/supervised-learning
4. Machine Learning. Yandex School of Data Analysis //Moscow Institute of Physics and Technology, 2018. Available at: http://lectoriy.mipt.ru/course/MachineLearning-L
5. Mathematics and Python for data analysis. National Research University Higher School of Economics, Yandex School of Data Analysis // Coursera. 2018. Available at: https://www.coursera.org/learn/mathematics-and-python
6. Vorontsov K.V. Mathematical teaching methods on precedents (machine learning theory), 2018. (in Russ.) Available at: http://www.machinelearning.rU/wiki/images/6/6d/Voron-ML-1.pdf
7. Min-Yuan Cheng, Nhat-Duc Hoang. Interval estimation of construction cost at completion using least squares support vector machine, Journal of Civil Engineering and Management, 20:2, 223-236, DOI: 10.3846/13923730.2013.801891
8. Min-Yuan Cheng, Nhat-Duc Hoang. Dynamic prediction of project success using evolutionary support vector machine inference model. ISARC 2008 - Proceedings from the 25th International Symposium on Automation and Robotics in Construction pp 452-458; doi:10.22260/isarc2008/0066
9. ScikitLearn // Scikit-learn developers (BSD License). 2007 - 2017. Available at: http://scikit-learn.org/stable/modules/linear_model.html#logistic-regression
10. Kaftannikov I.L., Parasich A.V. Problems of Training Set's Formation in Machine Learning Tasks. Bulletin of the South Ural State University. Ser. Computer Technologies, Automatic Control, Radio Electronics, 2016, vol. 16, no. 3, pp. 15–24. (in Russ.) DOI: 10.14529/ctcr160302

Application of machine learning technology to analyze the probability of winning a tender for a project

Annotation

The possibility of using machine learning technology to solve the problem of project analysis in order to support the decision to participate in the tender for the implementation of the project is substantiated and shown using a specific example. The approaches are described and the process of solving the problem of binary classification of projects using libraries of the Python language is shown. Attention is paid to the problem of choosing an algorithm for constructing a membership function, the problem of generating and analyzing input data, and evaluating the accuracy of a solution. It is shown that for the considered problem the best solution is provided by the logistic regression algorithm.

Key words:

machine learning; artificial intelligence; project risk analysis; project management; tender for the implementation of the project.