Application of machine learning methods in big data analytics at management of contracts in the construction industry

Marina Valpeters^{1,*}, Ivan Kireev² and Nikolay Ivanov³

¹SAS Institute; Stanislavsky Street, 21, bld.1, Moscow, 109004, Russia

²AO Raiffeisenbank, Troutskaya Street, 17, bld. 1, Moscow, 129090, Russia

³Moscow State University of Civil Engineering; Yaroslavskoe Shosse, 26, Moscow, 129337, Russia

Abstract. The number of experts who realize the importance of big data continues to increase in various fields of the economy. Experts begin to use big data more frequently for the solution of their specific objectives. One of the probable big data tasks in the construction industry is the determination of the probability of contract execution at a stage of its establishment. The contract holder cannot guarantee execution of the contract. Therefore it leads to a lot of risks for the customer. This article is devoted to the applicability of machine learning methods to the task of determination of the probability of a successful contract execution. Authors try to reveal the factors influencing the possibility of contract default and then try to define the following corrective actions for a customer. In the problem analysis, authors used the linear and non-linear algorithms, feature extraction, feature transformation and feature selection. The results of investigation include the prognostic models with a predictive force based on the machine learning algorithms such as logistic regression, decision tree, randomize forest. Authors have validated models on available historical data. The developed models have the potential for practical use in the construction organizations while making new contracts.

1 Introduction

Nowadays the methods of big data are actively used in dealing problems in different fields of the economy [1]. In the financial sector, customer credit scoring is calculated while loan approving and fraud actions are monitored. There is a potential of the usage of big data in the construction industry as well. This potential begins to be gradually unlocked in such tasks as choosing of a new area for construction, monitoring of "smart" houses, etc. [2]. One more task in the construction industry, where methods of big data could be applied, is the assessment of the probability of successful execution of the contract at a stage of a choice of contract holder organization. In case of contract signing a customer always face a risk that schedulable period of commissioning of the object will be missed, the cost of separate operations or of all construction will increase, the contractor will refuse

* Corresponding author: mvalpeters@gmail.com

execution of the undertaken obligations. The ability to estimate similar risks in time will allow a customer to take necessary measures to prevent extra time and material losses.

2 Research purpose

The purpose of a research is dedicated to testing of a hypothesis that on the basis of some parameters of the contract it is possible to predict the probability of its default. For achievement of this goal, different methods of machine learning will be used to determine the probability of successful completion of the contract. Then the received results will be analyzed, an estimate of the efficiency of the used methods will be made and the most suitable method for the task will be chosen. The factors indicating possible default of the contract will be revealed and recommendations about corrective actions in order to decrease the risk level of contract non-execution will be provided. And finally, the estimation of the possibility of prediction of the negative finals of the performance of the contract will be made, in particular, failure to meet a deadline and cost overrun will be predicted.

3 Basic propositions of a research

For the possibility of applying of machine learning methods, the volume of information on contracts has to be considerable. Data from the official site of a unified information system in the sphere of purchases in the information and telecommunication Internet (further – the official site of EIS) [3] meet these requirements. The chosen Internet resource provides free access to full and reliable information about contract system in the sphere of purchases and purchases of goods, works, services, by separate types of legal entities. The machine learning model of assessment of successful execution of the contract will be developed, based on the data on contracts in the sphere of construction This machine learning model will use the range of available financial and non-financial parameters.

4 Data loading and parsing

The attractiveness of the official site of EIS as a data source, besides the reliability and large volumes of data, is also that there is an access to open data on the website. Open data is the idea that some data should be freely available to everyone to use and to republish as they wish, without restrictions from copyright, patents or other mechanisms of control [4]. In accordance with the key features of open data, information from the official site EIS can be used, reused and extended, including associations with other data sets [5].

Access to the open data from the official site of EIS is provided by means of the public FTP server [6]. The data on the FTP server which is stored in the HTML format is presented in the form of the directories containing the normative help information, information on bank guarantees, complete regional unloading of the information published on the official site of EIS, etc. For the task solving, data of only one region, the city of Moscow, is selected because its volume of contracts is already enough for machine learning model (date of the first dataset publication is 07.01.2014, the dataset is updated daily). Besides contracts, the register of unfair suppliers is of some interest. The register includes a list of companies which have avoided contract signing in case of a tender victory or terminated the contract.

In most cases, it is convenient to develop machine learning models, execute ETL-process (Extract, Transform, Load - ETL), parse XML files with help of an interpreted high-level programming language Python [7]. It is a popular language for the purposes of data science with a free license, as well simple in a study and contains such software

packages as pandas, scikit-learn and Tensorflow that make Python a reliable option for the modern applications in the field of machine learning [8]. During solving the task, the data was automatically loaded using Python from two directories on the FTP-server ("fcs_regions/Moskva/contracts/" and "/fcs_fas/unfairSupplier/").

The size of the directory with the downloaded contracts is 22.3 GB. The directory contains ZIP folders with XML files, the quantity of such ZIP folders is 3180. Each ZIP folder contains information on contracts within one period of time (a day) as a number of XML files, the quantity of XML files in each ZIP folder is various.

There is no division into regions for unfair suppliers, the volume of the downloaded directory is 727,7 Mb. The directory includes 737 ZIP folders with XML files. In each folder there are several XML files.

4.1 Parsing unfair organization data

After data load the imported XML files were analyzed, important attributes were selected and the algorithm of file parsing was developed. The purpose of parsing is the extraction of the necessary information and its transformation to attractive representation, which is suitable for the further work. Pandas module was used for data processing. Pandas is a high-level Python library for data analysis [8] which allows working with data in a tabular style, providing the high speed of work.

Set of imported unfair supplier attributes is presented in Table 1. It should be pointed out that not all attributes in Table 1 were used in the developed machine learning models. A part of attributes was taken for the best understanding of information, another part – for its possible use in the further investigation.

Two additional attributes were calculated:

- "end_date" – date of exclusion of the organization from the list of unfair organizations. It is calculated as a date of inclusion in the list (attribute "publish_date") plus 2 years. If the organization was included in the list of the unfair organizations several times, the maximum date is taken;
- "inn_count" – count of times when an organization gets into the list of unfair organizations.

Table 1. Unfair organization attributes.

Attribute Name	Attribute Description
inn	TIN taxpayer reg.No.
fullName	Organization Full name
okpd_code	Russian Classification of Products by Economic Activities Code
okpd_name	Russian Classification of Products by Economic Activities Name
publishDate	Date of inclusion into the list of the unfair organizations
price	Contract price on which organization was included in the list of the unfair organizations
createReason	The reason for inclusion in the list of the unfair organizations
state	Record's status

The total data set includes 13147 lines. The data set is stored in the `dishonest_customers.pickle` file.

4.2. Extraction (parsing) of contracts data

The downloaded XML files with contracts consists of 4 types: "contract", "contractCancel", "contractProcedure", "contractProcedureCancel". One contract can have files of all these types. The "contract" files contain basic information available at the time of the signing of the contract: contract number, date, customer, contract product, sum, etc. The "contractProcedure" files contain information about actual contract payments, documents on acceptance of work, information about the reasons and dates of prescheduled contract termination. The "contractCancel" files contain information about the cancellation of the contract. The "contractProcedure" files contain information about contract procedure cancellation and the reasons for cancellation. For instance, the maturity date of contract stage or sum is incorrect. Files "contractCancel" and "contractProcedureCancel" were excluded from data processing, as their volume from the total amount of files is about 1.5%. Their processing is labor-consuming and, moreover, it is supposed that these files have no significant effect on the result of the first approach. However these files will be considered at further machine learning model development.

Files with identical names and identical contents were found while analysing downloaded data. In data processing, only one file is taken from a group of identical files in a random way.

In the "contract" and "contractProcedure" files the version control of changes (attributes "schemeVersion" and "VersionNumber") is realized. For the "contract" files the information was taken from the last available version. In order to get the actually paid amount on the contract, it is necessary to summarize amounts within one "contractProcedure" file and then summarize amounts from all "contractProcedure" files belonging to the contract. The "contractProcedure" file contains payments for execution of one contract stage. In order to calculate actually contract paid sum correctly, it is necessary to take into account version control within one stage of the contract. For this purpose in the "contractProcedure" files the maximum value of the attribute "schemeVersion" is selected corresponding to the specified date of a contract stage (attributes "endDate" or attributes "month", "year"). Then for the defined value of the attribute "schemeVersion" the maximum value of the attribute "versionNumber" is selected. So files, where a group of the attributes "EndDate"- "schemeVersion"- "versionNumber" fit with selected values, are considered in the calculation of contract stage paid amount.

Table 2 introduces the instance of data from the "contractProcedure" files. According to these data, the sum of a contract stage is equal 250.

Table 2. Example of data from "contractProcedure".

EndDate	SchemeVersion	versionNumber	paid
2017-10-25	7.2	0	100
2017-10-25	7.3	0	50
2017-10-25	7.3	1	50
2017-10-25	7.3	1	200

In some files contract expiration date and contract stage expiration date are defined in a format dd.mm.yyyy. Then in the files "contract", "contractProcedure" the fields "endDate" are populated. For other contracts, the expiration date has another format as mm.yyyy. Then for them, fields "month" and "year" are populated. This feature was taken into account while data processing. The extracted data from the "contract" and "contractProcedure" files were aggregated according to the contract number (Table 3).

Table 3. Contract attributes.

Attribute Name	Attribute Description	XML source file
reg_num	Contract ID	contract
currentContract Stage	Status of the contract. The accepted values: ET – execution is stopped, EC — execution is completed, E - execution	contract, contractProcedure
signDate	Date of signing of the contract	contract
startDate	Start date of performance of the contract	contract
price	Contract price	contract
paid	Actually paid sum	contractProcedure
code	Russian Classification of Products by Economic Activities Code	contract
name	Russian Classification of Products by Economic Activities Name	contract
endDate	Scheduled date of contract execution, format dd.mm.yyyy	contract
month	Scheduled month of contract execution	contract
year	Scheduled year of contract execution	contract
terminationDate	Date of contract termination	contractProcedure
inn	TIN taxpayer reg.No. of Organization contract holder	contract
organizationName	Organization contract holder full name	contract

The data set on contracts is stored in the contracts.pickle file.

5 Data preparation for simulation

On the previous steps, financial and non-financial measures on which the machine learning will be made were prepared. These measures will be treated as predictors. The following predictors were selected: contract duration, contract price, actually paid amount, product, flag that organization is included in the list of the unfair organizations. At the current investigation phase only three predictors are used in the model: contract duration, contract price, paid amount. The status of the contract was selected as the target variable.

As the most of machine learning models use data in a numeric format, it is necessary to execute preliminary transformations of all loaded data that don't have numeric a format.

First of all contracts dates were investigated.

At first, all the fields containing different contract dates were transformed. Contract execution date (attribute "enddate") was calculated as the attribute "enddate" if this field is populated or as a concatenation of attributes 01-"month"-"year" other like. Then format of attributes "signDate", "startDate", "enddate", "terminationDate" was changed from string to date. Also errors (misprints) in dates were corrected. And finally, dates were transferred to duration in days according to the calculation algorithms described in Table 4.

The fields "paid" and "price" were transformed into a numeric format.

Table 4. Algorithms of data transformation.

Attribute Name	Attribute Description	Calculation algorithm
contract_duration	contract duration in days	calculated as the difference between "enddate" and "startDate"
fact_contract_duration	the actual duration of the contract in days	calculated as the difference of "terminationDate" and "startDate"
start_delay	time between signing and start of the contract in days	calculated as the difference of signDate and startDate

Contracts with the status "EC" (successful execution) and "ET" (termination) were selected for machine learning modeling. Contracts under execution were not included into a data set as it is unknown how they will proceed to completion. The statuses of the selected contracts have been decoded by the rule: EC = 0, ET = 1.

The population of the fields "paid" (see table 2), "price" (Table 2), "contract_duration" (Table 3) was estimated. It is respectively 100%, 85%, and 67%. Such attributes population is considered to be acceptable. Let's look through the pattern of data distribution for the fields "paid" and "price" (Fig. 1, 2).

Fig. 1. Initial distribution of values of the field "Paid".

Fig. 2. Initial distribution of values of the field "Price".

Distribution of the amounts is similar to logarithmically normal distribution [9]. Many values are displaced to the left and have small size. Also there are a few extremely great values in the right area. For work of some machine learning models, it is recommended to use instead of values their logarithm [10]. The result of transformation of these fields is presented in Figures 3 and 4.

As can be seen from figures, the diagram became more symmetric and is more similar to a normal distribution. There is some emission around 0 for the field "paid" that could be considered noncritical.

Fig. 3. Distribution of values of the field "Paid" in logarithmic form

Fig. 4. Distribution of values of the field "Price" in logarithmic form.

Now let's look through the histogram of distributions of values of the field "contract_duration" (Fig. 5). Values are not varying, there are no strong distortions therefore logarithmation isn't required. However, if we look at a dispersion of values (minimum, maximum, the 1 and 99 percentiles), it is obvious, that there are large emissions which are better to be removed.

Fig. 5. Initial distribution of values of the field "contract_duration".

The obtained set of contracts was divided into the training and test selection in the ratio 60% to 40%. As contracts with a tag of 1 (terminated) make only 10% of the total amount of contracts, it is possible to use the option "stratify" of a scikit-learn package [7]. This option organizes splitting in such a way that in the test and training selections the ratio of the classes of successfully executed contracts (code 0) and terminated contracts (code 1)

remains the same. Thus, it is possible to avoid a situation when all terminated contracts get only in training or only in test selections.

As the considered machine learning models work badly with missing values in the fields, therefore such values were replaced by the average values of the field in the training selection. Then calculated average values in the training selection were also used for replacing empty fields in the test selection.

Fields "price", "paid" and "contract_duration" were scaled. In scaling one percent from above and from below were excluded in order to reduce the influence of emissions. Scales were chosen on the training selection and have been applied to the testing selection as well.

The performed data transformations resulted in data sets suitable for the majority of machine learning algorithms.

6 Comparison of machine learning models

In case of an expert determination of the success of the contract it is convenient to use the calculation of the probability of default [11] using the following formula:

$$\frac{1}{1 + \exp(-1 * (B_0 + B_1 * s_1 + \dots + B_i * s_i + \dots + B_n * s_n))}, \quad (1)$$

Where: B_i – the normalized model parameters get on a model input which calculation based on values of variables (financial and non-financial parameters);

s_i – the weights of parameters sum up to 0;

$\exp$ – exponential function;

n – number of parameters.

However as we work with big data and we would like to apply machine learning methods, therefore we chose a model of training based on function, which has similar principles of operation — regression.

As a comparison, the "Decision Tree" model and its improved version "Random Forest" model were probed. They are high-performed analytical models of machine learning, thus the output of their work in general is comparable.

Assessment of quality of models was executed with "roc_auc" metric, which shows, how well the model divides the selection into 2 classes with positive and negative examples. The "roc_auc" metric was chosen because it successfully works in cases when the quantity of examples of one class considerably exceeds the quantity of examples of other class. In our task the ratio of the terminated contracts to successfully executed makes 1 to 9, thus, the metrics "roc_auc" provides correct result. The alternative metric "accuracy" applied to this task does not give an exact assessment. "Roc" curves of 3 models used in the investigation is pictured in Figure 6.

7 Conclusions

The basic solution of a problem of determination of the probability of a contract default with the help of a linear model of logistic regression was developed. The basic model has shown a possibility of the task solving and has provided generalized assessment of the main results.

Fig. 6. Roc-curves of three models. Linear roc_auc_score model = 0.8764, Decision Tree roc_auc_score = 0.8554, Random Forest roc_auc_score = 0.8884.

8 Future Work

1. Probation of other models of machine learning, their comparison and assessment of efficiency. Experiments with a set of input parameters and ways of their normalizing have planned. Usage of filters on subject of the contract has planned too.
2. Assessment of a possibility of prediction of other negative finals of contract execution: extension of terms and cost increase.
3. Analysis of the influence of predictors on a probability of a contract default and thus making recommendations about the reduction of the risk level of contract failures.
4. Experiment planning of the efficiency assessment of the offered recommendations.
5. Taking into account temporary delay of data that is the basis of prediction.

References

1. M. Hilbert, DPR **34**, 1 (2016) DOI: 10.1111/dpr.12142
2. Y. Kang, J. Yu, J. Chang, SSRG-IJCE **4**, 10 (2017) DOI: 10.14445/23488352/IJCE-V4I10P101
3. The official site of a unified information system in the sphere of purchases on the information and telecommunication (2018) <http://zakupki.gov.ru>
4. *What is Open Data? The Open Data Handbook.* (2018) <http://opendatahandbook.org/guide/en/what-is-open-data/>
5. V.V. Staroverov, INJOIT **4**, 10 (2016)

6. [Ftp://free:free@zakupki.gov.ru](ftp://free:free@zakupki.gov.ru)
7. S. Raschka, V. Mirjalili, *Python Machine Learning, 2nd ed.* (Packt Publishing, Birmingham, 2017)
8. W. McKinney, *Python for Data Analysis* (O'Reilly Media, Inc., Sebastopol, 2018)
9. StatSoft Russia Knowledge Portal (2018) <http://statistica.ru/theory/lognormalnoe-raspredelenie/>
10. Open Machine Learning Course. Topic 6. Feature Engineering and Feature Selection (2018) <https://medium.com/open-machine-learning-course/open-machine-learning-course-topic-6-feature-engineering-and-feature-selection-8b94f870706a>
11. V.V. Sofronova, Journal Financial Analytics: Science and Experience, **3** (2016)