ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ

ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ

ВЫСШЕГО ОБРАЗОВАНИЯ

«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»

Факультет «Санкт-Петербургская школа экономики и менеджмента»

Подразделение (департамент, лаборатория, кафедра) \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Применение "Random Forest Algorithm" в задачах корпоративного управления**

КУРСОВАЯ РАБОТА

по направлению подготовки 38.03.01 «Экономика»

Образовательная Программа «Экономика»

Санкт-Петербург

2019 г.

Работа сдана в ОСУП

«\_\_\_» \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_20\_\_ г.

Защита состоялась

«\_\_\_» \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_20\_\_ г.

Состав комиссии и подписи:

1. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

2. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

3. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Оценка за работу \_\_\_\_\_\_\_\_\_ (\_\_\_)**

оценка по пяти и (десяти) бальной

системе

Студента (-ки) группы № \_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(Фамилия И.О., подпись)

Руководитель

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(должность, степень, Фамилия

И.О., подпись)

Content

[Introduction 3](#_Toc134540634)

[Chapter 1. Organizational Issues of Risk Management Risk management principles based on the ALARP concept 5](#_Toc134540635)

[Risk-management strategy in terms of the ALHARP concept 7](#_Toc134540636)

[The Role of Internal Controls in ERM Analysis 9](#_Toc134540637)

[Chapter 2. Creating a random forest. Random subspace method. 14](#_Toc134540638)

[Selection and aggregation methods 14](#_Toc134540639)

[Bootstrap 14](#_Toc134540640)

[Bagging 16](#_Toc134540641)

[Random Forest Algorithm 17](#_Toc134540642)

[Implementation 18](#_Toc134540643)

[Conclusion 25](#_Toc134540644)

[Literature 26](#_Toc134540645)

[Appendix 28](#_Toc134540646)

[Appendix 1 28](#_Toc134540647)

[Appendix 2. 30](#_Toc134540648)

# Introduction

The modern economy is characterized by high volatility in cash flows associated with the valuation of assets and future profits.

In most economic relations, it is assumed how much income these relations will bring, and how much profit will be directed to development in the future. At the same time, the proportions of profit distribution are to a greater extent associated with the adoption of competent management decisions and previously accepted obligations.

The company, as a rule, when doing business makes production and financial plans for several years ahead, but when they are implemented, the expected result is often not achieved.

The volatility of future cash flows is associated with a number of endogenous and exogenous factors that provoke the failure to achieve such results. As an agent contributing to the achievement, the company selects and uses various types of consulting services, automatic control systems, planning systems, while the costs associated with attracting agents must be at least commensurate with the added value received.

Preventive measures to prevent the receipt of the expected income (both of a speculative nature and in the field of pure risks) are incorporated in the company's modern corporate risk management systems and are based on a number of key rules, the main ones of which are: ALARP principle, AHARP principle, proper management organization, orientation on shareholder value and integration with internal controls. This chapter reveals the main content of these principles.

The purpose of this article is to explore the possibility of using random forest algorithm in risk management assessment. To do this, it was necessary to solve the following tasks:

1. Studied the theoretical and practical foundations of corporate risk management.

2. Evaluate the effectiveness of corporate risk management using the linear regression method (ANOVA) and the classification tree (CHAID).

3. Compare the results obtained by the traditional econometric method and the random forest method.

# Chapter 1. Organizational Issues of Risk Management Risk management principles based on the ALARP concept

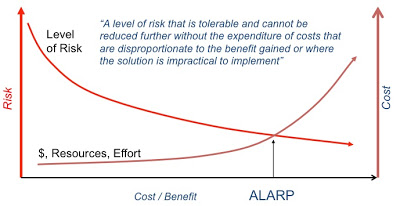
A key concept in risk management is the ALARP (Reduce As Much As Possible) principle. Essentially, this principle is that risks should be reduced to a level where the costs of risk mitigation measures balance the benefits (Figure 1).[](http://3.bp.blogspot.com/-EAfVfZJSWfY/TsNb_bU0yhI/AAAAAAAAAVo/BGjZ8z30HF8/s1600/ALARP+1.jpg)

Fig. 1. The principle of risk management

In practice, this principle is much more difficult to follow:

1. A database is required for a period of at least two years preceding the risk management event, which will be the basis for conducting a comparative analysis of the effectiveness of risk management processes.

2. Most of the costs associated with risk management are implicit costs. And in this case, the calculation of the economic efficiency of the process is possible with a high degree of conventionality.

3. Not all risk management measures give a positive effect immediately after implementation: some processes do not pay off during the first two or three years, which calls into question the possibility of operational control over the feasibility of the measures being implemented.

Determining the ALARP scope is the starting point for building any scorecard.

The main problems in developing such a system that meets the requirements of risk benchmarking are:

Scope - What is the scope of the developed system of indicators? The system can be developed for the entire organization or for a division, whether all risk management programs or only part of them are considered.

Indicators - what specific indicators should be included: based on historical or planned data, is there enough information and what other information needs to be collected?

Measuring system - how performance is measured.

Duration - what period will be subject to research? Should historical data be used for analysis?

Performance – what specific performance indicators of the organization are being studied in relation to benchmarking. How have they been affected by risk management?

Quality - What will signal that risk management is good?

The next principle, no less important in ERM (Enterprise Risk Management) - the AHARP principle (As High As Reasonably Practicable - to overestimate as much as appropriate) refers to positive risks, in contrast to the ALARP principle.

|  |
| --- |
| [Описание: http://4.bp.blogspot.com/-vMzvtecZw3c/TpgxuSEKbxI/AAAAAAAAAQM/MQgjCXWlvYw/s400/Slide01.jpg](http://4.bp.blogspot.com/-vMzvtecZw3c/TpgxuSEKbxI/AAAAAAAAAQM/MQgjCXWlvYw/s1600/Slide01.jpg) |
| Figure 2. AHLARP model |

The combined effect of the two concepts can be seen in Fig. 2. RTP (Risk Critical Point) marks a turning point in risk management when the positive effects of risk begin to outweigh the negative effects. This item indicates the cost-effectiveness of the risk management process. It is at this point that we need to start bearing some of the cost of risk management if we are to turn corporate risk management into a value-creating tool rather than a loss recovery mechanism. In other words, if the risk is negative, then further injection of funds into its management will not bring benefits; if the risk is positive, then the economic efficiency of the risk will be exclusively positive until the growth rate of management costs exceeds the growth rate of the risk itself.

The only difference is that with a negative risk, at least some positive effect is impossible if it is realized, and when managing such risks, the strategy of minimizing losses will be the best solution to manage them, positive risk is hidden opportunities that are not provided for in the current planning system companies that, in the event of an appropriate situation, are able to create added value for the company.[20]

# **Risk-management strategy in terms of the ALHARP concept**

The risk management strategy using the ALHARP model is to reduce the impact of uncertainty about the event.

Figures 3 and 4 clearly demonstrate, using the ALHARP concept, how to channel the impact of the likely outcomes of an event into a positive outcome.

|  |
| --- |
| [Описание: http://3.bp.blogspot.com/-pXrdLAA5vYQ/Tpgx16sBknI/AAAAAAAAAQ0/gRUZNocHM64/s400/Slide06.jpg](http://3.bp.blogspot.com/-pXrdLAA5vYQ/Tpgx16sBknI/AAAAAAAAAQ0/gRUZNocHM64/s1600/Slide06.jpg) |
| Figure 3. Lack of resources stimulates an increase in the likelihood of negative consequences |

|  |
| --- |
| [Описание: http://4.bp.blogspot.com/-X1SAU_wo4FY/Tpgx3aDZS1I/AAAAAAAAAQ8/orvFH5dcs0o/s400/Slide07.jpg](http://4.bp.blogspot.com/-X1SAU_wo4FY/Tpgx3aDZS1I/AAAAAAAAAQ8/orvFH5dcs0o/s1600/Slide07.jpg) |
| Figure 4. Timely and sufficient use of risk management resources |
|  |

It should be noted that the amount of costs (investments) for risk management is not determined by the general formula of economic efficiency and must be determined individually in each specific case, however, the general principle of risk management is as follows: volatility requires more resources, and simple typical risks are subject to standard management methods.

# The Role of Internal Controls in ERM Analysis

Internal control and its organization are an integral attribute of the ERM system - almost all recommendations in the field of risk management do not ignore the organization of internal control. In practice, in companies that meet the requirements of corporate business standards, the most common and recommended is a three-level management organization model that seamlessly combines both risk management and internal control functions: the 3LOD model (3 levels of protection).

In this model, risk management and compliance are included in the second layer of protection, implemented against the backdrop of impeccable operational management, in which business units are held accountable for the risks that arise in the course of their activities, and risk management provides units with the necessary risk management tools that they adopt. (limit structure, key risk indicators, etc.). [21]

Table 1. Organizational model "Three lines of defense"

|  |  |  |
| --- | --- | --- |
| **Management levels** | **Which departments are represented** | **Functions** |
| 1 | business units | formation of a correct understanding among business units of all the risks inherent in their direct activities |
| 2 | integrated active risk management system. | formation of appropriate risk management tools of the first line of defense, as well as internal regulatory documents on risk management.  Regulations for the interaction of structural divisions in the implementation of various business processes |
| 3 | Inside audit | Analysis of the performance of structural subdivisions of their duties, identification of violations recorded in the course of activities |

The main purpose of applying this model is the need to avoid duplication of risk management and internal control functions, as well as the need to avoid conflicts of interest between management and stakeholders. An analysis of the theory and practice of organizing integrated management systems at the corporate level made it possible to highlight the following controversial points:

1. The implementation of full-fledged internal control and risk management is quite expensive.[14]

2. Potential stakeholders evaluate the company based on the results reflected in the company's financial statements [15], which are often subject to various manipulations [16].

The motives that may guide company representatives who decide to misrepresent their financial statements may be different. Among the main reasons that encourage the use of illegal methods of influencing the financial result, we can distinguish:

* The desire to influence the decisions of stakeholders, for example, can lead to an overestimation or underestimation of the results of earnings, depending on what the company's goals are at that particular moment.
* Conflict of interests of interested parties and the company as a whole. Shareholders, especially minority shareholders, are most often interested in receiving the maximum possible amount of dividends, while the company strives for a competitive advantage, financial independence, increased creditworthiness and solvency, increase in market share, etc. To achieve the company's goals, it is necessary to reinvest profits, which may cause dissatisfaction among shareholders, so the company's management may underestimate financial results, hiding part of the profit in reserves and other reporting items.
* The desire to “embellish” financial results to attract investors, for example, on the eve of an additional issue or on the eve of a major project requiring investment.
* An attempt to underestimate profits in order to reduce tax payments (the desire to present the minimum possible value of the financial result in the name of reducing the tax burden). [13]
* Obtaining government subsidies [8]
* Fulfillment of the requirements of contractual obligations, for example, the terms of loan agreements, which often contain a minimum amount of the company's assets.[12]
* The desire to present the company in the most favorable light in anticipation of mergers and acquisitions (can be explained by both overestimation and underestimation of some indicators).
* Justification of the cost of services/goods offered by monopoly companies. In a monopoly, high profit figures clearly indicate that the company is making super profits, which cannot but arouse the interest of the antimonopoly service. In order to avoid attracting the interest of the state and public organizations, the company may seek to artificially lower profits.[7]
* the desire of managers to receive planned remuneration, which is often based on the achievement of certain indicators by the company, etc.

Regardless of the motives behind such actions, unattainable returns are not possible, since reading financial statements, calculating economic indicators based on them, evaluating company performance, as well as making forecasts for future value if the data presented in profitability are not reliable.

All of the above activities associated with the desire of managers to maximize their profits have nothing to do with obtaining financial income, but in the absence of proper incentives focused on the managerial basis of management, the company begins to achieve its value.

First of all, this is realized through an orientation towards short-term goals to the detriment of long-term ones. Behavioral theory of profit maximization requires that in the case of agency conflict, management is equivalent to whether to maximize profits or management costs, so it can be assumed that poor management will reflect the growth of short-term debt, having a significant impact on management. income and expenses.

Thus, as an explanation of usage, we have chosen:

1. Sales volume, as evidence of the scale of production.

2. Operating profit (EBIT) - profit from core activities.

3. Binary variable characterizing investments in intangible assets.

4. Administrative expenses.

5. Long term liabilities

6. Profitability of fixed assets.

Current liabilities are taken as a variable that needs to be explained.

The use of statistical methods for constructing trees did not give the desired result - rare signs of long-term management interest are not significant enough to build model (Fig. 5). (Annex 1).

Изображение выглядит как диаграмма

Автоматически созданное описание

Figure 5. Classification tree built by the CHAID method in SPSS

Therefore, we resorted to the random forest algorithm and implemented it in the Python programming language.

# Chapter 2. Creating a random forest. Random subspace method.

In the random subspace method, the main algorithms are trained on various feature description subsets, which are also randomly selected. This method is effective in problems with a large number of features and a relatively small number of objects, as well as in the presence of redundant non-informative features. In these cases, feature-based algorithms may have better generalization ability than all-feature based algorithms. RSM allows you to reduce the correlation between trees and avoid overfitting.[17]

A well-known algorithm that uses both RSM and bagging is called Random Forest. The feature section at the top of the random forest is searched among a random subset of features, and each tree in the composition is trained on the sample obtained using the bootstrap operation.[9]

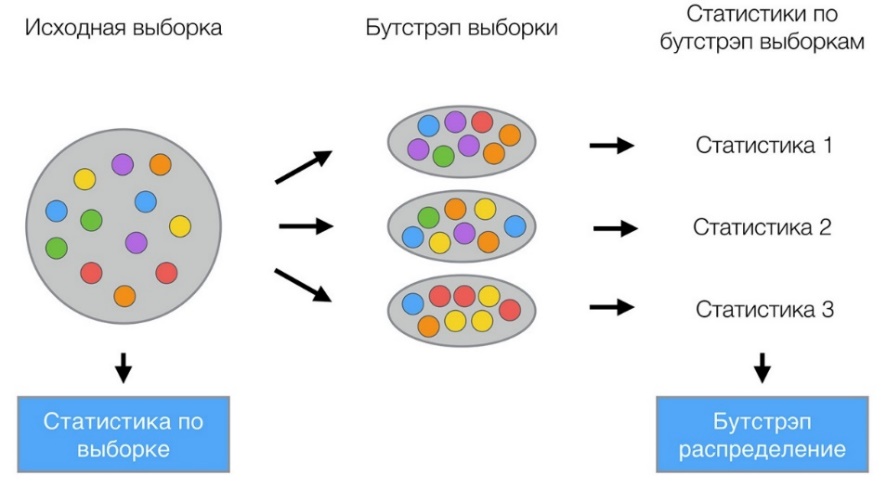
# Selection and aggregation methods

# Bootstrap

This method is described in detail at <https://habr.com/en/company/ods/blog/324402/#butstrep>. The method is schematically shown in Figure 6.

Statistics (for a sample) - a measurable numerical function of a sample that does not depend on the unknown parameters of the distribution of sample elements.

Figure 6 Bootstrap scheme



# Bagging

The bagging method was proposed by L. Breiman in 1996. From the initial training sample of length N, various training subsamples of the same length N are formed using the previously described bootstrap method. At the same time, some objects fall into the subsample several times, some - not even once. Basic algorithms trained on subsamples are combined into a composition using simple voting.[17] The effectiveness of bagging can be explained by two circumstances:

1. Due to the difference in the underlying algorithms, their errors cancel each other out when voting. [17]

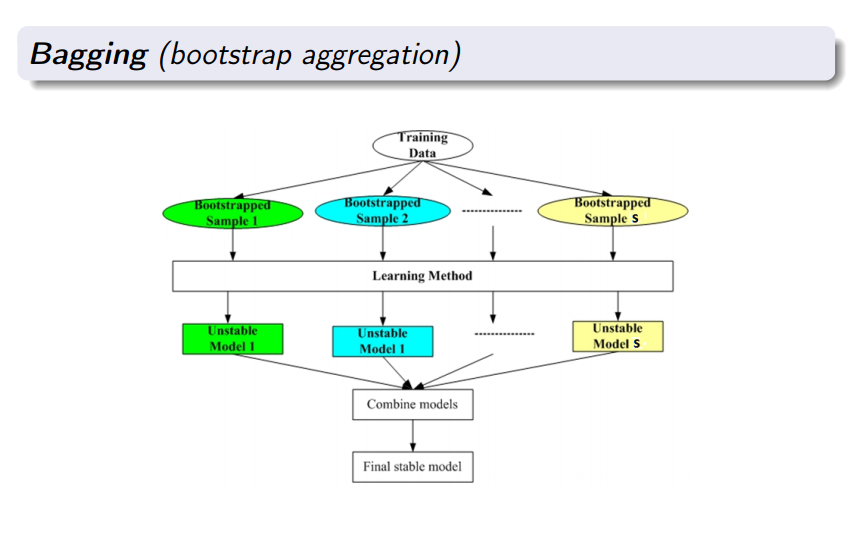
2. Outlier objects may not fall into some training subsamples. [17]

Figure 7 Scheme description of Bagging.

# Random Forest Algorithm

**The random forest algorithm** is a machine learning algorithm that combines the bagging method (a combination of learning models increases the overall result) and the random subspace method. [11]

The algorithm is used for classification, regression and clustering problems.

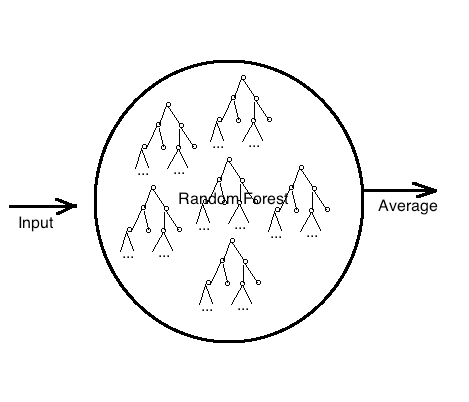


Figure 10. Scheme of random forest.

# Implementation

The implementation of the random forest algorithm will be carried out in the Python programming language, in the Jupyter development environment, since this language already contains all the necessary modules to simplify our task (pandas, skitlearn, etc.). The main one for us is skitlearn, since it is responsible for machine learning and, in our case, Random forest (from sklearn.ensemble import RandomForestRegressor).

As data, we take a sample of companies in the metallurgical sector in the period from 2013 to 2017, the number of observations is 1280.

import pandas as pd - we connect the pandas module (a simple tool for processing and analyzing data in Python).

features = pd.read\_csv(fa.csv') - read our data file

features.head(5) – output the first 5 lines (Fig. 11)

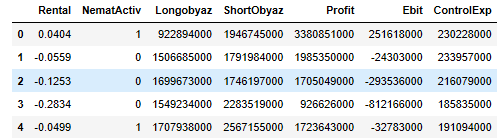


Figure 11 The first 5 rows of our data.

print('The shape of our functions is:', features.shape) - the number of rows and columns

features = pd.get\_dummies(features) - get rid of dummy variables (if any), since Python does not work well with them.

It is necessary to separate the data into functions and purposes. Goals are the predicted values and features are all the columns used by the model to make predictions.

import numpy as np - connect the NumPy module (this module adds support for large multidimensional arrays and matrices, as well as a large library of high-level and very fast math functions for working with these arrays)

labels = np.array(features['ShortObyaz']) - write the data we want to predict

features= features.drop('ShortObyaz', axis = 1) - remove targets from our data, leaving only features

feature\_list = list(features.columns) - store feature names

features = np.array(features) - convert to NumPy array

We split the data into training and test sets. During training, we let the model match the responses, in our case, short-term debt, so that it learns to predict them based on the features. We assume that there is a relationship between all features and target values, and the goal of our model is to study these relationships in the learning process. Then, when it comes time to evaluate the model, we ask it to make a prediction on the test dataset, in which case only feature access is provided (no response access). Since there are answers for the test set, you can compare the resulting predictions with the true values to evaluate the accuracy of the model. Typically, when training a model, we randomly split the data into train and test sets to get an idea of all the data points. To debug the code, random\_state will be set to 42 (the split will always be the same), if we do not need it, then we simply remove this parameter. We will test 75% of the sample against 25% of the sample. To do this, we divide the data into test and training.

from sklearn.model\_selection import train\_test\_split enable traffic to separate test and train data

train\_features, test\_features, train\_labels, test\_labels = train\_test\_split(features, labels, test\_size=0.25, random\_state=42)

It is advisable to review the form of all data to make sure everything is done correctly. The number of features is expected to match the number of columns and the number of rows for the respective training, testing, and labeling features:

print('Training features shape:', train\_features.shape)

print('Training labels shape:', train\_labels.shape)

print('Shape of test features:', test\_features.shape)

print('Test label shape:', test\_labels.shape)

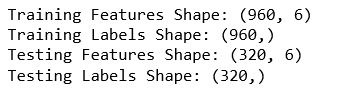


Figure 12. Data form.

Prepare the data, now we need to build our random forest.

from sklearn.ensemble import RandomForestRegressor - we include the Random Forest model in our program.

rf = RandomForestRegressor(n\_estimators=10, random\_state=42, max\_depth=5) — create a forest of 10 trees (the more topics, the more accurate the forecast, but you need to calculate manually), depth 5.

rf.fit(train\_features, train\_labels); - train the model.

Now we need to prove that our model is accurate (not for nothing that we started all this?). Let's find out how our forecast differs from the actual value. (absolute translation error)

predictions = rf.predict(test\_features) - predicted values for sample tests

errors = abs(predictions - test\_labels) - throw an exception

print('Mean Absolute Error:', round(np.mean(errors), 2)) - print our error (Fig. 13)



Figure 13 error output, for our limited small forest.

Далее построим все наши деревья графически, чтобы спрогнозировать значения.

from sklearn.tree import export\_graphviz

We then build all of our graphic trees to predict the values.

from sklearn.tree import export\_graphviz

import pydot - quick start modules.

tree1 = rf.estimators\_[0] - get first tree

tree2 = rf.estimators\_[1] - get second tree

tree3 = rf.estimators\_[2] - get third tree

tree4 = rf.estimators\_[3] - get the fourth tree

tree5 = rf.estimators\_[4] - get fifth tree

tree6 = rf.estimators\_[5] - get the sixth tree

tree7 = rf.estimators\_[6] - get the seventh tree

tree8 = rf.estimators\_[7] - get the eighth tree

tree9 = rf.estimators\_[8] - get the ninth tree

tree10 = rf.estimators\_[9] - get the tenth tree

export\_graphviz(tree1, out\_file = 'tree1.dot', feature\_names = feature\_list, rounded = True, precision = 1) - Write the first tree in .dot format

export\_graphviz(tree2, out\_file = 'tree2.dot', feature\_names = feature\_list, rounded = True, precision = 1) - Write the second tree in .dot format

export\_graphviz(tree1, out\_file = 'tree1.dot', feature\_names = feature\_list, rounded = True, precision = 1) - Write the first tree in .dot format

export\_graphviz(tree2, out\_file = 'tree2.dot', feature\_names = feature\_list, rounded = True, precision = 1) - Write the second tree in .dot format

export\_graphviz(tree3, out\_file = 'tree3.dot', feature\_names = feature\_list, rounded = True, precision = 1) - Write third tree in .dot format

export\_graphviz(tree4, out\_file = 'tree4.dot', feature\_names = feature\_list, rounded = True, precision = 1) - write the fourth tree in .dot format

export\_graphviz(tree5, out\_file = 'tree5.dot', feature\_names = feature\_list, rounded = True, precision = 1) - Write fifth tree in .dot format

export\_graphviz(tree6, out\_file = 'tree6.dot', feature\_names = feature\_list, rounded = True, precision = 1) - Write the sixth tree in .dot format

export\_graphviz(tree7, out\_file = 'tree7.dot', feature\_names = feature\_list, rounded = True, precision = 1) - Write the seventh tree in .dot format

export\_graphviz(tree8, out\_file = 'tree8.dot', feature\_names = feature\_list, rounded = True, precision = 1) - write the eighth tree in .dot format

export\_graphviz(tree9, out\_file = 'tree9.dot', feature\_names = feature\_list, rounded = True, precision = 1) - write the ninth tree in .dot format

export\_graphviz(tree10, out\_file = 'tree10.dot', feature\_names = feature\_list, rounded = True, precision = 1) - Write tenth tree in .dot format

Further, through the online converter "https://dreampuf.github.io/GraphvizOnline/" we convert the received files in .png format in the form of trees and translate. (Because for some reason "Jupyter" did not see my files in .dot format).

Examples of unloaded trees are found in the Sample Random Forest Trees application.

In theory, you can use the usual commands: (graph, ) = pydot.graph\_from\_dot\_file('tree1.dot'); graph.write\_png('tree1.png') but it says .dot file not found.

Having made a graphical display, now we can move on to the calculation part. To do this, we will remake our forest into a completely random one and increase the number of trees to 100,000 (more is possible, but it’s already hard for my processor to build 100,000).

train\_features, test\_features, train\_labels, test\_labels = train\_test\_split(features, labels, test\_size = 0.25)

rf = RandomForestRegressor(n\_estimators=100000)

rf.fit(train\_features, train\_labels);

predictions = rf.predict(test\_features)

errors = abs (predictions - test\_labels)

print('Mean Absolute Error:', round(np.mean(errors), 2)) - Figure 14.



Figure 14. Our full forest error.

features1 = pd.read\_csv('fina.csv') - load the regressors with which we will predict

feature1 = np.array(features1) - convert them to NumPy format

pred = rf.predict(features1) - make predictions

i = 0

while i < 11:

print(pred[i]) # print them. Figure 15.

i=i+1

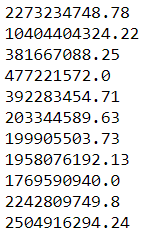
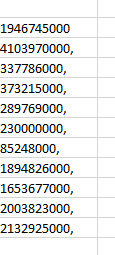


Figure 15. Our predictions for our test dataset and true values (left).

importances = list(rf.feature\_importances\_)

feature\_importances = [(feature, round(importance, 5)) for feature, importance in zip(feature\_list, Importants)]

feature\_importances = sorted(feature\_importances, key = lambda x: x[1], reverse = True)

[print('Variable: {:20} Importance: {}'.format(\*pair)) for pair in feature\_importances];

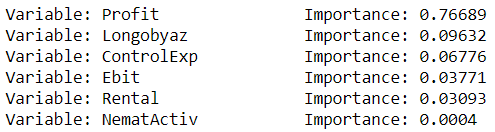


Figure 16. Importance of variables for regression.

Conclusion: the method of random trees made it possible to include in the analysis factors that are rarely found in the sample. The fact of having long-term loans, the fact of having intangible assets.

# Conclusion

Thus, in our work, we have confirmed the high efficiency of the random tree algorithm on data with a low tracking frequency. In this work, all tasks were solved:

1. When studying the basics of risk management, it is shown that the 3LOD model is the most effective form of organizing risk management, which allows reducing the degree of agency conflict and maximizing added value. However, when working with real financial data, managers, observing compliance standards, are forced to manipulate reporting. Manipulations occur between reporting periods. With this approach, the efficiency of building risk management analysis models on real data is quite low.

2. The article evaluates the effectiveness of corporate risk management using the linear regression method (ANOVA) and the classification tree (CHAID). It is shown that in regression analysis, intangible assets and administrative expenses are the most significant, while when building organizational algorithms (CHAID tree), the frequency of these parameters is insufficient.

3. As a result of the random forest algorithm, the resulting standard error turned out to be smaller (2 billion < 11 billion) and the significance of the variables increased (Figure 16).

Thus, all the tasks set are fulfilled, the goals are achieved. It is also worth noting that when constructing a random forest, there is no need to standardize the data in any way to obtain a normal distribution.

# Literature

1. Открытый курс машинного обучения. Тема 5. Композиции: бэггинг, случайный лес. URL: <https://habr.com/ru/company/ods/blog/324402/#ansambli>
2. Skurichina M., Duin R. P. W. Limited bagging, boosting and the random subspace method for linear classifiers // Pattern Analysis & Applications. 2002. Pp. 123–126.
3. Breiman, Leo. Bagging predictors // Machine learning 24.2 (1996): 123-140.
4. Will Kehrsen. Random Forest in Python URL: <https://towardsdatascience.com/random-forest-in-python-24d0893d51c0>
5. Е.А. Соколов ФКН ВШЭ: Лекция 3. Решающие деревья URL: <https://www.hse.ru/mirror/pubs/share/215285956>
6. Артиков М. Э. Анализ эффективности применения методов классификации URL: https://moluch.ru/th/8/archive/62/2632/
7. Бычкова, С.М; Итыгилова, Е.Ю. Риски искажения учетной информации: выявление в процессе аудита // Аудиторские ведомости, 2010, №12. URL: http://www.lawmix.ru/bux/476
8. Макарова В. А. Эффективность риск-менеджмента: оценка и ее влияние на инвестиционную привлекательность бизнеса // Управление финансовыми рисками. 2015. № 04(44). с. 270- 287.
9. Breiman L. Random forests // Machine Learning. — 2001. — Vol. 45, no. 1. — Pp. 5–32
10. Попков, М.И. Автоматическая система классификации текстов для базы знаний предприятия, МГУ имени Ломоносова, 2014, Магистерская диссертация. URL: <http://master.cmc.msu.ru/files/master2014_1_popkov.pdf> 27-28
11. Шамаева Д.Р. Деревья решения для задач построения рейтинга коммерческих банков // Технические науки: проблемы и перспективы: материалы V Междунар. науч. конф. (г. Санкт-Петербург, июль 2017 г.). — СПб.: Свое издательство. С. 18-22.
12. Вахрушина, М.А. К вопросу о транспарентности финансовой отчётности. URL: http://www.finotchet.ru/article.html?id=468
13. Черкасова, Е.В; Истомин, В.С. Манипулирование финансовой отчётностью. URL: http://www.amursu.ru/attachments/article/9530/N53\_26.pdf
14. Ingley C., Van der Walt Do Board Processes Influence Director and Board Performance // Corporate Governance: An International Review. 2005. Vol. 13. No. 5. P. 632-653.
15. Ittner, C. D. Larcker D. F.*.* Coming up short on nonfinancial performance measurement. Harvard Business Review (November): 88-95.
16. Jaspal S. (2011) *Jaspal S.* Metrics to Measure Risk Management Strategy
17. Воронцов К.В. Лекции по алгоритмическим композициям, 7 октября 2012. C 15-16 URL: <http://www.machinelearning.ru/wiki/images/0/0d/Voron-ML-Compositions.pdf>
18. Scikit Learn URL: [https://scikit-learn.org](https://scikit-learn.org/)
19. Machinelearning.ru URL: Machinelearning.ru
20. Tablot J., Jakeman M. How to Perfomance Benchmark Your Risk Management, 18 November 2016, Pp. 29-31.
21. Potter P., Toburen M.. The 3 Lines of Defense for Good Risk Management, June 1 2016. URL: <http://www.rmmagazine.com/2016/06/01/the-3-lines-of-defense-for-good-risk-management/>

# Appendix

# Appendix 1

Регрессионный анализ исследовательской проблемы

|  |  |  |  |
| --- | --- | --- | --- |
| **Variables Entered/Removeda** | | | |
| Model | Variables Entered | Variables Removed | Method |
| 1 | Нематериальные активы, RUB, Рентабельность активов (ROA), %. Управленческие расходы, RUBb | . | Enter |
| a. Dependent Variable: 2013, Краткосрочные обязательства, RUB | | | |
| b. All requested variables entered. | | | |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model Summaryb** | | | | |
| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
| 1 | ,864a | ,747 | ,746 | 11060298363,6844220 |
| a. Predictors: (Constant), 2013, Нематериальные активы, RUB, 2013, Рентабельность активов (ROA), %, 2013, Управленческие расходы, RUB | | | | |
| b. Dependent Variable: 2013, Краткосрочные обязательства, RUB | | | | |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **ANOVAa** | | | | | | |
| Model | | Sum of Squares | df | Mean Square | F | Sig. |
| 1 | Regression | 271863372887974620000000,000 | 3 | 90621124295991550000000,000 | 740,791 | ,000b |
| Residual | 92114640519971390000000,000 | 753 | 122330199893720300000,000 |  |  |
| Total | 363978013407946000000000,000 | 756 |  |  |  |
| a. Dependent Variable: 2013, Краткосрочные обязательства, RUB | | | | | | |
| b. Predictors: (Constant), 2013, Нематериальные активы, RUB, 2013, Рентабельность активов (ROA), %, 2013, Управленческие расходы, RUB | | | | | | |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Coefficientsa** | | | | | | | | |
| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. | Collinearity Statistics | |
| B | Std. Error | Beta | Tolerance | VIF |
| 1 | (Constant) | 306307495,260 | 445184912,991 |  | ,688 | ,492 |  |  |
| Рентабельность активов (ROA), % | -117637167,650 | 224453939,532 | -,010 | -,524 | ,600 | ,996 | 1,004 |
| Управленческие расходы, RUB | 7,991 | ,196 | ,935 | 40,763 | ,000 | ,639 | 1,565 |
| Нематериальные активы, RUB | -6,011 | 1,071 | -,129 | -5,610 | ,000 | ,639 | 1,564 |
| a. Dependent Variable: 2013, Краткосрочные обязательства, RUB | | | | | | | | |

# Appendix 2.

Примеры деревьев случайного леса.

