ORIGINAL PAPER

DOI: 10.26794/2587-5671-2022-26-5-132-148
UDC 336.051/336.64(045)
JEL G32, C14, C63

Business Valuation with Machine Learning

P.S. Koklev
Saint Petersburg State University, Saint Petersburg, Russia

ABSTRACT

The aim of the article is to test the hypothesis about the applicability of machine learning **methods** to train models that allow to accurately predict the market capitalization of an enterprise based on data contained in three main forms of financial statements: *Income statement, Balance sheet, and Cash flow statement*. **The scientific novelty** of the study lies in the proposal of an alternative approach to the actual finance problem — business valuation. The conducted empirical study allows us to test the hypothesis under consideration. We train various models using the most popular machine learning **methods** (*LASSO*, *Elastic Net*, *KNN*, *Random Forest*, SVM, and others). To determine the best approach for assessing the value of a company, the effectiveness of different methods is compared based on the R^2 performance metric (86,7% for the *GBDT*). Financial statements data of *NYSE* and *NASDAQ* companies are used. The study also addresses the problem of the interpretability of the trained models. The most important features are identified — the forms of financial statements and their specific items that have the greatest impact on market capitalization. Three independent ways to determine feature importance indicate the significance of the information contained in the *Income statement*. In particular, *Comprehensive income* was the most important item for accurate predictions. Robust methods of variable normalization and missing data imputation are also highlighted. Finally, various ways of improving the developed models are recommended to achieve even higher accuracy of forecasts. The study **concludes** that machine learning can be applied as a more accurate, unbiased, and less costly approach to value a company. Feature importance analysis can also be used to understand and further explore the value creation process.

Keywords: business valuation; relative valuation; DCF; machine learning; artificial intelligence; big data; regression analysis; gradient boosting; decision trees

For citation: **Koklev P.S. Business valuation with machine learning. *Finance: Theory and Practice*. 2022;26(5):132-148. DOI: 10.26794/2587-5671-2022-26-5-132-148**

INTRODUCTION

Artificial intelligence, which has a strong impact and completely changes the economic sectors and the field of scientific knowledge, is rightfully called the electricity of the 21st century.[1] Artificial intelligence has also influenced the financial sector and the theory of finance. Machine learning, which embodies the most productive form of implementing the idea of artificial intelligence in practice, has long been used by financial institutions to solve various problems. The first studies on the use of neural networks to predict the dynamics of stock prices were published back in the 90s last century [1]. It is the financial sector that has become the main "testing ground" for testing new methods. The most favorable factor for this has been the abundance and variety of data aggregated by financial institutions in their usual activities. In addition, it is financial markets that are considered the most attractive object of application of the intellectual efforts of leading specialists in the field of information and mathematical sciences, since the reward for the successful application of statistical methods in this area will be maximum.

Hundreds of articles are published annually, which record cases of successful application of artificial intelligence methods in the financial sector [2]. However, some topics are undeservedly ignored. Thus, aspects of using machine learning for business valuation are not sufficiently considered by the scientific community. This may be due to a number of reasons, one of which is the difficulty of modeling the subjective process of business

[1] Stanford Business. URL: https://www.gsb.stanford.edu/insights/andrew-ng-why-ai-new-electricity (accessed on 29.11.2021).

valuation by an analyst. Some researchers believe that the valuation process cannot be formalized [3]. However, it is precisely the ability to include non-linear relationships between variables that are a property of complex processes (for example, the process of forming the value of an enterprise) that is one of the main strengths of machine learning. It has been proven that some methods of statistical learning allow one to approximate continuous functions of any type and complexity [4]. That is why the complexity and subjectivity of the business valuation process are not an obstacle, but a reason for using Machine Learning methods (ML) to predict the company's market capitalization.[2]

This paper is devoted to testing the hypothesis about the applicability of machine learning methods to predict the market capitalization of an enterprise. To do this, we conducted an empirical study and subsequent analysis that compared the effectiveness of different supervised learning approaches to predict/assign a company's market capitalization based on its financial statements for the last eight quarters. Thus, the data contained in the Balance sheet, Income statements, and Cash flow statements.

The first part of the main section is devoted to a brief review of the literature and the background of the application of machine learning in the context of the research problem. Next, the methodological aspects of empirical research are considered. Finally, the study itself is based on data from 3,945 *NASDAQ* and *NYSE* companies, and its results are discussed in the third part. As a result of comparative analysis, the most effective training methods were identified, which make it possible to predict the company's value with high accuracy. The effectiveness of the methods is compared based on out of sample the R^2 metric, i.e. data not used in model training.

The high accuracy of predictions can be used to solve various problems. Estimating the value of private enterprises, determining the placement price of shares before an *IPO*, creating investment strategies,[3] and developing new approaches to reflecting investments in the capital of other enterprises in accounting are far from a complete list of possible points for applying research results. In addition to the models themselves, the analysis of the significance of features is also valuable, which makes it possible to single out independent variables — the forms of financial statements and their specific items that most strongly affect the final cost forecast. This allows you to identify the most valuable information contained in the financial statements, which, in turn, can be used both by companies to better disclose the most relevant information to investors, and by the regulator when developing new standards. In addition, the importance of attributes gives insight into the value-creation process of a company. Finally, in the course of the work, the author was able to determine the best methods for preprocessing financial statements data: independent variables normalization and missing data imputation.

1. MACHINE LEARNING FOR BUSINESS VALUATION

1.1. Prerequisites for ML Methods for Business Valuation

Let us consider the main factors that indicate that statistical learning can be successfully used to determine the market capitalization of a company. The term "machine learning" usually has the following meanings:

1. Application of a diverse set of non-parametric statistical forecasting methods

[2] In the framework of the study, a company's value is understood as the sum of the company's market capitalization and the book value of liabilities. Thus, in the presence of data on the book value of liabilities, the problem of forecasting the company's value is identical to the problem of forecasting market capitalization.

[3] In the context of developing investment strategies, a positive difference between the forecast and actual cost will be interpreted as an undervaluation, and a negative difference as an overvaluation of the company.

that can take into account non-linear relationships between independent variables.

2. Using regularization is a way to punish complex models to prevent overfitting problems.[4]

3. The choice of the optimal set of hyperparameters among the many possible variants of the model specification. *Grid Search* and *Random Search* are the most popular algorithms for hyperparameter optimization. Hyperparameter fitting is typically performed using a validation set obtained from *k-fold cross-validation*.

In addition to applying advanced algorithms, machine learning focuses on validating a model through k-fold cross-validation and using test sets to provide an unbiased assessment of the quality of a model. Thus, the emphasis shifts from the estimation of the model parameters $\hat{\beta}$ to the result of the forecast $\hat{y}$. To assess the value of a company, the primary task is to determine the function $h(x)$, which allows for a highly accurate prediction. It is the prediction, and not the estimate of the coefficients $\hat{\beta}$ that is of great practical importance. Specializing in prediction problems, machine learning is ideally suited to solve this problem. The multivariate nature of *ML* methods allows them to be much more flexible than traditional econometric approaches. This flexibility helps to better approximate the unknown function that represents the value creation process of a company.

Several properties of machine learning methods make them ideal for modeling processes with an unknown or indeterminate shape. Variety comes first. Even within the framework of one of the many dissimilar methods, the researcher has the opportunity to choose an infinite number of different model specifications. For example, the learning rate of a hyperparameter can take on an infinite number of values from the region of positive real numbers. In addition, the selection of the best models through k-fold cross-validation allows you to control the problem of overfitting and avoid false discoveries resulting from unsupervised testing of many different model specification options (Gu, Kelly, & Xiu, 2020).

Despite their original specialization, analytics developed for big data are especially effective when working with small datasets (3,945 observations are considered a small dataset). They are noticeably superior to traditional social science methods: the *least squares method (OLS)* for regression problems and *logistic regression* for classification problems, which in their original form are not able to take into account the non-linear influence of independent variables on the dependent one. The ability to take into account the multidimensional nature of data is the main reason for the superiority of non-parametric AI methods.

Another disadvantage of traditional methods is the tendency to overfit. The resulting *OLS model* tends to make highly inaccurate out-of-sample predictions. In addition, the interpretability of the results is no longer considered an advantage of the *OLS* method. The use of the Tikhonov regularization method —*Ridge* regression [6], *Lasso* regression [7], as well as their combination — *Elastic Net*, allows increasing the quality of forecasting while maintaining the possibility of a straightforward interpretation of the obtained coefficients $\hat{\beta}$. С учетом всех недостатков некоторые исследователи полагают, что использование простой линейной регрессии в социальных науках должно быть сведено к минимуму (Hindman, 2015).

Given all the shortcomings, some researchers believe that using simple linear regression in the social sciences should be minimized (Hindman, 2015).

The explanatory variables used in the study, the financial statement items, are often similar and highly correlated.[5] The

[4] Overfitting is a phenomenon when the built model explains well the value of a company from the training sample, but is not able to make high-quality forecasts for companies that are not involved in model training.

[5] An example of correlated variables would be revenue and gross profit. A more extreme example is the balance sheet at the date of the last quarterly report and the balance sheet at

consequence of non-strict multicollinearity is a high standard error of the estimated coefficients, which deprives the *OLS* method of its main advantage — interpretability. Also, the *OLS* method stops working when the number of variables approaches the number of observations. Using *ML* allows you to successfully work with data that is pathological for traditional methods and select viable models.

Although machine learning has great potential for predicting the value of a company, it still has some drawbacks and limitations. The predictions made by applying the models are measurements. By themselves, the measurements do not indicate the fundamental mechanism that forms the market capitalization of a company. The process of assigning a prediction is often opaque. It is rather difficult even for specialists to describe the logic of the formation of a particular prediction in an accessible language. The complexity of communication is one of the barriers to applying machine learning to some financial problems. Fortunately, the problem of the interpretability of the results is given special attention in the literature. Modern methods for assessing the feature importance, considered in an empirical study, make it possible to mitigate the "black box" problem (Carvalho, Pereira, & Cardoso, 2019).

1.2. Literature review

An important issue is the definition of independent variables with the help of which cost forecasting should be carried out. The theory of asset valuation gives quite definite indications. Scientists and practitioners use a variety of approaches to assessing the value of a company. The leading researcher in this field, A. Damodaran, identifies four methods (Damodaran, Valuation approaches and metrics: a survey of the theory and evidence, 2007).

the date of the penultimate quarterly report. Due to the use of data from the last eight quarters, virtually every variable is highly correlated with the other seven variables in the X – design matrix .

The first approach is *Discounted Cash Flow* valuation, *DCF*. This approach determines value by discounting the expected cash flows generated by the assets of companies.

Another method, the accounting valuation method, uses data on the book value of assets.

The third and most commonly used method is relative valuation, which involves the use of market data on the value of similar companies [11]. Stock multiples are used to determine the value of companies.

Finally, the contingency method uses models to value financial options.

Due to the strong theoretical foundations, it is the discounted cash flow valuation that receives special attention in the scientific community. The basics of the approach were proposed by A. Marshall and E. Böhm-Bawerk, who considered the concept of present value in the first half of the 20th century. E. Böhm-Bawerk was the first to demonstrate the explicit calculation of the cost of an annuity [12]. The method is based on the assumption that assets with large and predictable cash flows should have a higher value than assets with low and volatile cash flows. In other words, the value of an asset is an increasing function of expected cash flows and a decreasing function of a discount rate that reflects the risk and uncertainty of the cash flows.

It should be noted that in practice, when applying the *DCF* method to determine expected cash flows, forecast their growth rates, return on equity, etc., financial reporting data is used. Therefore, the training of statistical models that predict the company's value must be carried out on these data.

This paper is linked to a number of studies from the 2000s on the application of machine learning to financial markets. In 2009, G. Atsalakis and K. Valavanisb reviewed the existing literature on the use of neural networks to predict stock market dynamics [13]. Later, in 2018, F. Xing, E. Cambria, and R. Welsch published a review of the application of NLP natural language

processing methods for financial forecasting (Xing, Cambria, & Welsch, 2018). Among the works directly related to asset valuation, the publication of 2015 by B. Park and J. Bae, devoted to the use of machine learning, in particular the *AdaBoost* algorithm, to predict the value of the residential real estate in Virginia, stands out (Park & Bae, 2015).

The closest in meaning to forecasting the capitalization of an enterprise by statistical methods are works devoted to forecasting stock multipliers. In one of the first publications on this topic, M. Kisor and V. Whitbeck used the data of the profit growth rate, the dividend payout ratio, and the standard deviation of the change in earnings per share (EPS) to determine the ratio of capitalization to net profit, the P/E (Whitbeck & Kisor, 1963). multiplier (Whitbeck & Kisor, 1963). The sample consisted of 135 companies. The regression equation, obtained in 1963, was as follows:

$$P/E = 8.2 + 1.5g + 6.7(Payout\ ratio) - 0.2\sigma_{eps}.$$

Forecasting of multipliers for Russian stock market companies is considered in (Коклев, 2020). The author prefers to predict the $EV/Sales$, multiple, arguing that a positive revenue value allows using as large a sample of companies as possible.

H. Joshi and R. Chauha used multiple determinants to predict the value of coefficients using *Ridge* and *Lasso* regression. The adjusted R-squared was 70%. A separate test sample was not used to check the results (Joshi & Chauha, 2020).

M. Le Claire, A. Alford, C. Penman, D. Nissim, A. Damadoran and many other authors have developed multiplier forecasting models (Liu, Nissim, & Thomas, 2002). Within the framework of these studies, independent variables, as a rule, are various coefficients, integral indicators characterizing the activity, financial stability and efficiency of the enterprise.

2. RESEARCH METHODOLOGY

To test the hypothesis under consideration, it is necessary to solve the following problem: to train a model for assessing the market value of the company's equity capital. Without a priori information about the effectiveness of one or another approach, it seems appropriate to consider the maximum possible number of currently existing machine learning methods. Determining the best method can be done by comparing the accuracy of the predictions for the test set data. In the most general terms, we describe the forecast value of the company as:

$$\hat{y} = h(x),$$

where h — company value function obtained using machine learning; $\hat{y}$ — the company's market capitalization forecast; x — the company's feature vector consisting of financial statement data.

2.1. Used machine learning methods

There is a wide range of high-quality literature that describes in detail each approach both in terms of describing the method and the computational algorithms themselves, which make it possible to obtain a model for a given training set (Hastie, Tibshirani, & Friedman, 2016). Therefore, we will not dwell on each of the methods considered in the work, but only briefly list them. *Table 1* shows twelve different approaches to represent the idea of using *ML* to value a company as representatively as possible.

The software libraries used to build models within each method are also listed: *scikit-learn* for traditional algorithms and *CatBoost* for gradient boosting. In addition to *CatBoost,* there are several other popular gradient-boosting libraries on decision trees: *XGBoost, H2O* and *LightGBM* from *Microsoft*. The choice in favor of the *Yandex* project is explained by the higher accuracy of the resulting models both with the default value of the hyperparameters and with their optimized value. Also, the learning process with *CatBoost* is usually faster (Prokhorenkova, Gusev, Vorobev, Dorogush, & Gulin, 2017). The last column shows the number of models tested under each method. To determine the best set of hyperparameters, a large number of possible

Table 1

Employed machine learning methods

Method	Algorithm implementation	Loss function	Number of specifications investigated
Least Squares Method, OLS	scikit-learn*	MSE	10 000
Ridge	scikit-learn	$MSE + \alpha \|w\|_2^2$	10 000
Lasso	scikit-learn	$MSE + \alpha \|w\|_1$	2000
Elastic Net	scikit-learn	$MSE + \alpha\rho \|w\|_1 + \frac{\alpha(1-\rho)}{2} \|w\|_2^2$	2000
Stochastic Gradient Descent, SGD	scikit-learn	MSE, Huber loss, epsilon insensitive, squared epsilon insensitive	2000
Huber	scikit-learn	Huber loss	2000
Support Vector Machine, SVM	scikit-learn	epsilon-insensitive	3000
K-Nearest Neighbors, KNN	scikit-learn	–	2000
Decision Tree, DT	scikit-learn	MSE, MAE, Poisson loss	10000
Random Forest, RF	scikit-learn	MSE, MAE, Poisson loss	500
Extremely Randomized Trees, ERT	scikit-learn	MSE, MAE	500
Gradient Boosted Decision Trees, GBDT	CatBoost**	MSE	100

Source: compiled by the author.

Note: * Scikit-learn is a library for machine learning in Python. URL: https://scikit-learn.org/ (accessed on 15.11.2021); ** CatBoost is a gradient boosting on decision trees library, developed by Yandex. URL: https://catboost.ai/ (accessed on 17.11.2021).

specifications must be tested. It should be noted that *Table 1* lacks *neural networks*. Due to the heterogeneity of financial reporting data,[6] this method turned out to be ineffective ($R^2 = -8.0\%$).

[6] The heterogeneity of the data lies in the fact that different items of financial statements carry different economic meanings. Neural networks, on the contrary, are most effective when working with homogeneous data. For example, images, where each feature represents a pixel.

2.2. Evaluation of the quality of models

There are good reasons to believe that regression results can often be the result of unsupervised testing of many model specification options [22]. To obtain the desired result, the researcher retroactively chooses the "best" model. The formation of a separate test sample deprives the

unscrupulous scientist of the opportunity to fit the model to the data to obtain sensational results. Within the framework of this study, the general population is divided into two incompatible populations: training and testing. The test sample consists of a random sample of 20% of the data — 788 companies. It is the R-squared calculated for companies i, included in the test sample τ, that will be a reliable indicator characterizing the quality of forecasts. A comparison of the effectiveness of various *ML* methods will be carried out using this indicator.

$$R^2_{oos} = 1 - \frac{\sum_{i\in\tau}\left(y_i - \hat{y}_i\right)^2}{\sum_{i\in\tau}\left(y_i - \bar{y}_i\right)^2}.$$

Data from the remaining 3,155 companies will be used for model training and cross-validation in the hyperparameter optimization process using a random search algorithm. The essence of random search is to compare many different specifications within the method by randomly selecting hyperparameters from a given distribution (Bergstra & Bengio, 2012). The choice of the best model from many different specification options is based on the R-squared R^2_{cv}, calculated for the test set, which is a dynamically changing part of the training set. For the test sample, a tenth of the training sample will be sequentially removed.

2.3. Feature Importance

Various ways of assessing the feature importance are designed to level the problem of interpretability, allowing you to evaluate the contribution of each independent variable — financial statement items both to the overall quality of the model and to a specific prediction. We will consider three different approaches:

1. Prediction Values Change.
2. Permutation Importance.
3. SHAP (SHapley Additive exPlanations).

In the framework of the first method, the importance of a variable is determined by calculating the average change in the prediction when the feature changes. The larger the average change in the prediction, the more importance is assigned to the feature. This process is performed for each variable.

The *Permutation Importance* method works as follows: the observations for the variable being evaluated are randomly shuffled. That is, one column of the X, plan matrix is randomly shuffled, the rows of which consist of company feature vectors. The rest of the columns remain unchanged. After the rearrangement, the predicted values of capitalization are calculated on the basis of the changed matrix. A variable is considered important if the accuracy of the predictions is significantly reduced compared to the original, unchanged matrix. On the other hand, a feature is considered unimportant if its permutation did not lead to a significant decrease in R^2.Thus, the process of random permutation is repeated three hundred times for each variable [24]. A high number of permutation iterations for each feature is necessary to obtain narrow confidence intervals for estimating the average quality degradation of the model.

The *SHAP* method is an approach motivated by the idea of Shapley value, a principle from game theory that allows you to calculate the optimal distribution of winnings between players, taking into account their contribution to the final result [25]. In the context of interpreting statistical learning models, *SHAP* allows you to get a local (for a particular observation x) addition of the contribution of each financial statement item to the final capitalization forecast compared to the baseline (the average prediction of the model). *Fig. 1* shows an illustration of the method for the biopharmaceutical company *Biogen Inc.*

The vertical axis lists financial statement items and their values, sorted in descending order of absolute *SHAP*. That is, the variables that had the greatest impact on *Biogen's* capitalization prediction are presented. Thus, according to the method, the value of *Net operating income* of $ 2.36 billion three

Fig. 1. **SHAP values based on feature importance for *Biogen Inc.* Numbers are given in US dollars. GBDT model**

Source: compiled by the author.

quarters ago (*netIncomeFromContinuingOperations_t-3*) increased the predicted capitalization of the company by $ 7.95 billion compared to the base level.

The most negative impact was made by the value of *Comprehensive income* in the last quarter (*comprehensiveIncomeNetOfTax_t-1*), which lowered the forecasted value of the company by $ 2.23 billion. The total contribution of local values for the remaining 315 features that are not shown individually on the graph is $ 10.76 billion with a positive sign (the last line in *Fig. 1*). The final capitalization forecast in this example was $ 42.4 billion.

In addition to analyzing the forecast for a single company, it is also possible to assess the overall importance of each of the features by calculating the mean absolute *SHAP* value based on all observations. To assess the feature importance within each of the three methods, a choice was made in favor of using the entire general population, consisting of the union of the training and test sets.

3. EMPIRICAL STUDY BASED ON NYSE AND NASDAQ DATA

3.1. Data

The independent variables used to calculate capitalization are quarterly financial statements (Form 10-Q) for the last eight quarters received in November 2021. The general population consists of 3,945 companies listed on the *NYSE* and *NASDAQ*. Financial statements provided by *Alpha Vantage*.[7] *In total, the plan X matrix contains 329 columns, 328 of which are items in the balance sheet, income statement, and cash flow. The set of predictors also includes information*

[7] Alpha Vantage — an API service financial data. URL: https://www.alphavantage.co// (accessed on 28.11.2021).

about the company's industry affiliation. Thus, the feature vector for a single company has the following form:

$$x = \begin{pmatrix} x_1 \\ x_2 \\ \dots \\ \dots \\ x_{329} \end{pmatrix} = \begin{pmatrix} \\ \text{Total Assets}_{t=0} \\ \text{Total Current Assets}_{t=0} \\ \dots \\ \dots \\ \text{Cash at the end of the year}_{t=-7} \end{pmatrix}.$$

3.1.1. Missing data

For most companies, there are missing values for some characteristics. There can be several reasons for missing data. Thus, the level of disclosure of information may vary from company to company. On the other hand, the specifics of the company's economic activity may not imply the presence of a specific article. For example, a company that does not use finance leases will have missing data for the feature capital Lease Obligations in each of the eight quarters. Almost each of the considered *ML* methods cannot be used in the presence of missing data: model training and subsequent prediction are impossible. The obvious solution is to exclude corporations with missing data from the sample. An alternative option is to remove the variable with missing data. However, the disadvantages of such radical solutions significantly outweigh their main advantage — simplicity.

The presence of missing data for a given feature may be a unifying characteristic for a certain stratum of companies. For instance, for companies of the same industry. Therefore, removing companies with missing data makes the sample unrepresentative of the stock market as a whole. The results and conclusions obtained would extend only to a specific subset of companies — those with a low number of missing data. Obviously, the value of the results of such a study would be much lower. In addition, a decrease in the number of observations leads to a noticeable decrease in the quality of the resulting models.

In the publication of 2001, M. Banko and E. Brill, using the example of the problem of natural language processing, showed that it is the amount of data for training, and not the choice of method, that is the main determinant of the effectiveness of machine learning [26]. Working with a small data set, there is every reason to believe that the exclusion of companies from the sample will significantly reduce the predictive power of the trained models. Taking into account the above factors, the removal of a company from the sample will be applied only in the most pathological cases — in the absence of data for ninety or more reporting items.

Replacing missing data with imputed values avoids deleting more companies. The most commonly used methods are measures of central tendency imputation: mean or median. This approach is also seen as insufficient for the task of company valuation, where even a small increase in the predictive power of the model is given with great difficulty.

Let us use a more complex imputation method — the *Iterative Imputation* algorithm. The essence of the method is to train a set of auxiliary models, where each feature is predicted using the rest. Thus, for each feature and for each company, the missing data are replaced by predictive values [27]. An important component is the definition of a training method for auxiliary models. With no reason to favor one or the other, *GBDT* seems to make the most sense. To date, *GBDT* is considered the most advanced method and the default choice for working with heterogeneous data (Munkhdalai, Munkhdalai, Namsrai, Lee, & Ryu, 2019). Iterative imputation was applied before training the final capitalization forecasting model within each of the twelve methods considered, with the exception of *GBDT* itself. The fact is that the implementation of this algorithm by the *CatBoost* library has native support for missing data. Thus, the imputation for this method

is optional. In addition, *GBDT* also does not require feature scaling/standardization.

3.1.2. Feature scaling

Feature scaling or standardizing is a necessary step in data preprocessing. The order of magnitude of financial statement items is quite high and varies greatly. Typical values for many articles can be in the billions. With rare exceptions, the effectiveness of *ML* methods is noticeably reduced if the dimensions of the features differ greatly. Many optimization algorithms and cost functions require the normalization of independent variables. There are a fairly large number of options for data standardization, the best of which is very difficult to determine in advance. In total, seven transformation options were considered and presented in the *scikit-learn* library. Experiments with the training set have shown that the most stable and preferred method of data transformation is the *Quantile Transformation*. This approach to data standardization uses the transformation of non-parametric features into a uniform distribution with values from zero to one. For example, after transformation, the maximum value of the *income* variable will be equal to one, and the minimum value will be equal to zero. The median company's income will be 0.5. For alternative standardization methods (*Standard Scaler*, *Max Absolute Scaler*, etc.), the R-squared R^2_{is} for the final models, calculated on the basis of training data, was often in the negative zone and was noticeably inferior as a quantile transformation.

3.2. Results and discussions

The effectiveness of each of the considered methods is compared in *Table 2* and in *Fig. 2*. As expected, due to the inability to take into account non-linear relationships, traditional methods turned out to be relatively inefficient (R^2_{oos} = 20.8% for *OLS*). For the same reason, the application of various regularization methods to the linear model did not lead to a significant improvement in the results. Moreover, the effectiveness of *Lasso* and *OLS* turned out to be almost identical. Only with the help of *Ridge* regression was it possible to slightly improve the result — up to 22.5%. Changing the *MSE* loss function to *Huber* only worsens the quality of forecasts (17.0%).

The Huber loss function [29] is used for a large number of outliers, which is relevant for financial reporting in the raw format. After the variables were standardized, the dataset no longer included extreme observations. Therefore, the use of this loss function is not necessary. The use of stochastic gradient descent *SGD* with different loss functions did not lead to the desired result (17.6%). The *SVM,* used mostly for classification problems, has predictably proved to be inefficient for the regression problem. R-squared was only 21.7%. A significant improvement in the quality of models was achieved using non-parametric methods: *k-nearest neighbors* and *decision tree*. The share of explained variance for *KNN* is 44.3%, and for the decision tree — 41.4%. The use of ensembles of decision trees made it possible to raise the accuracy of predictions to a new level of quality. For the classic *Random Forest* algorithm, the coefficient of determination is already 73.2%. For its even more randomized modification, the method of *Extremely Randomized Trees*, the R-squared turned out to be less: 64.9%. Finally, the *Gradient Boosted Decision Trees* method (*GBDT*) made it possible to obtain highly accurate predictions of the capitalization of companies from the test sample ($R^2_{oos} = 86.7\%$).

Due to the fact that the latter model is able to give the most qualitative assessment of the company's value, it is of particular interest from the point of view of the analysis of the feature importance. The bar chart (*Fig. 3*) shows the most important reporting items, with changes in the values of which the prediction value of the company's value changes the most (the *Prediction Values Change* method). The information is presented in the context of each financial statement: *Income statement, Cash flows,* and *Balance*

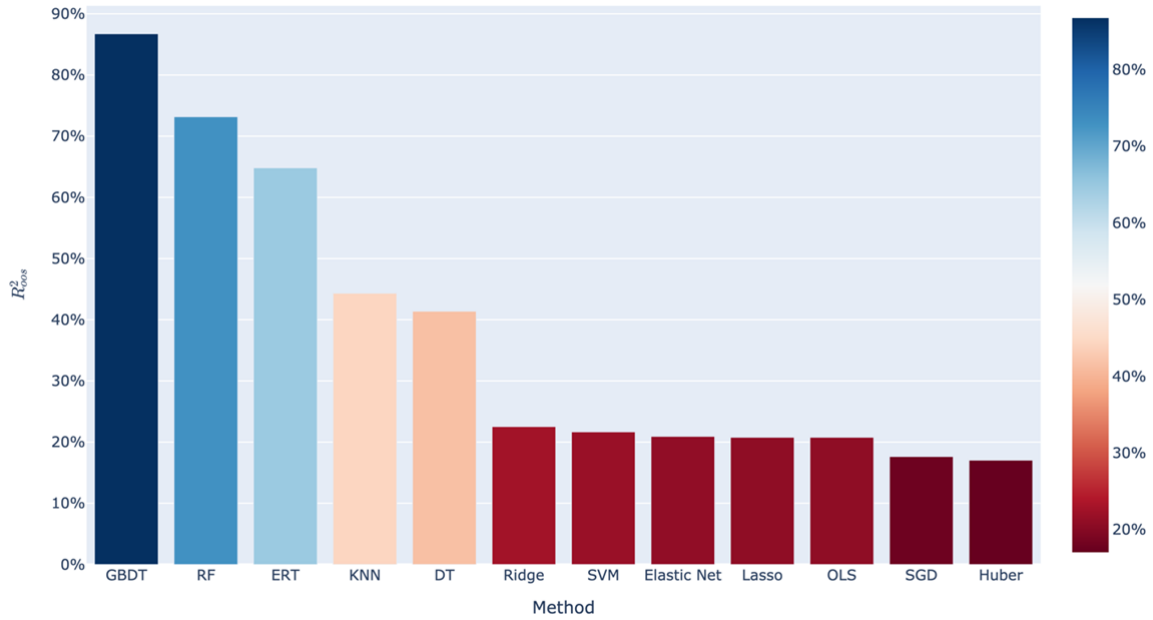

Fig. 2. **Machine learning methods performance comparison based on the out of sample R-squared**

Source: compiled by the author.

Table 2

Machine learning methods performance comparison based on the out of sample R-squared

Method	*OLS*	*Ridge*	*Lasso*	*Elastic Net*	*Huber*	*SGD*	*SVM*	*KNN*	*DT*	*RF*	*ERT*	*GBDT*
R^2_{oos}, %	20.8	22.5	20.8	20.9	17.0	17.6	21.7	44.3	41.4	73.2	64.9	86.7

Source: compiled by the author.

sheet. The horizontal axis measures the relative feature importance.

We note that the *Balance sheet* data are of the least value for the model. Interestingly, the book value of intangible assets is by far the most important variable of this reporting form. This contradicts the opinion of some researchers who consider the book value of intangible assets, in particular, the *Goodwill* item, to be the least informative when building *DCF* models. Within the framework of the cash flow statement, the most informative signs are changes in operating assets for different quarters.

Fig. 3 also shows that the total importance of the *Income statement* significantly exceeds the total importance of the other two reports. Although subject to manipulation, depending on the accounting policy, it is the data at the bottom of the report that is the most important in assessing the value of the company. In accordance with the goals and objectives of financial accounting, *Comprehensive income, Net profit, EBITDA* can indeed be considered key indicators of an enterprise's performance, which are necessary when making investment decisions. *Sales revenue*, on the contrary, is not among the important features. These observations may also indirectly point to the benefits of using earnings-based stock multiples.[8]

[8] Contrary to popular belief about the preference for multiples calculated based on revenue.

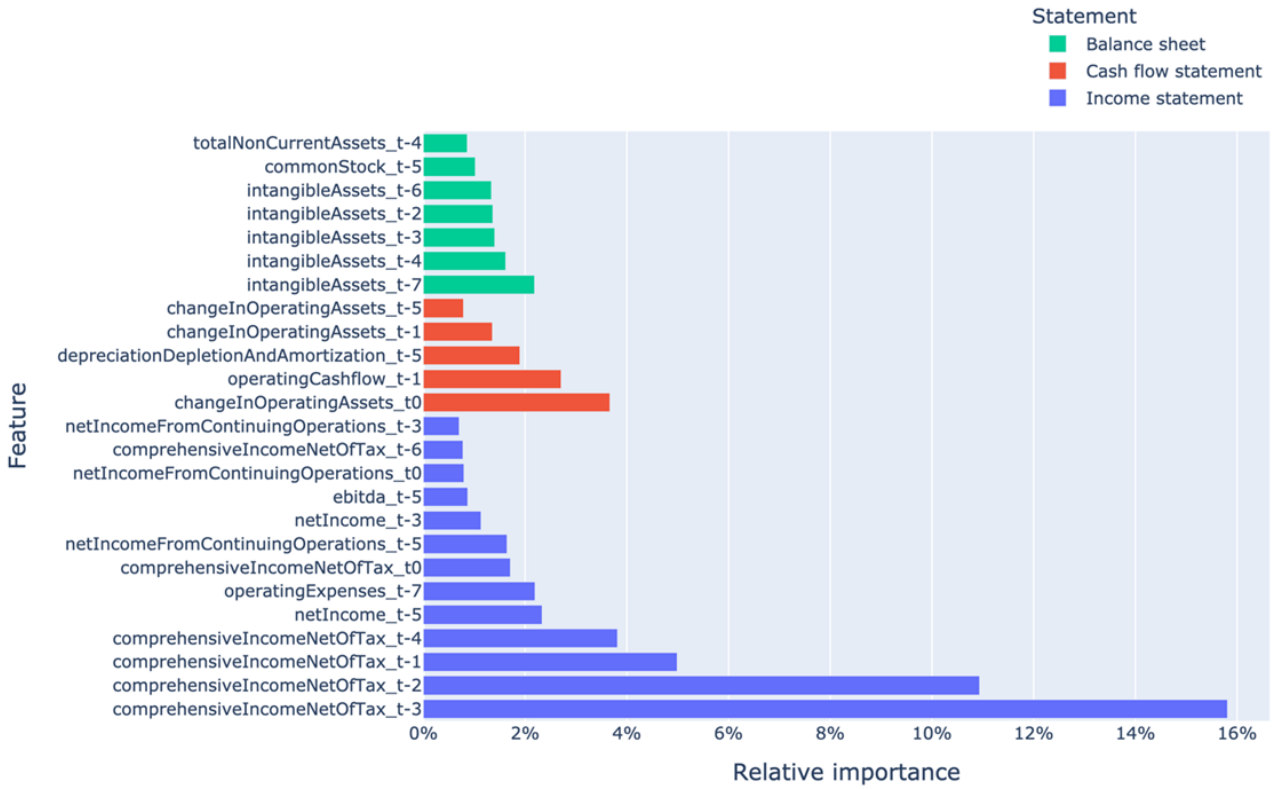

Fig. 3. **Relative feature importance. Prediction values change method**

Source: compiled by the author.

Let us evaluate the influence of features by the *Permutation Importance* method. Important features are those whose random permutation led to a significant decrease in the quality of the model. The variables whose permutation resulted in the largest average drop in R-squared are shown in *Fig. 4*. The average decrease in the quality of the model is calculated based on three hundred random permutations for each feature. Confidence intervals are based on two standard deviations.

The main conclusions obtained by different methods for assessing the feature importance overlap in many respects, which increases their reliability. The low significance of the *Balance sheet* data is obvious: the cumulative significance of all seven report items does not exceed the value of the fifth most important *Income statement* item. Quarterly *Comprehensive income* data remains the key variable in each of the feature importance methods. For the *Cash flow statement*, the main fields are *Changes in operating assets* and *Operating cash flow*.

Finally, we calculate the importance of variables using the *SHAP* method. *SHAP* values of each individual observation for the most important features are shown in (*Fig. 5*), which shows the influence of local feature values on the prediction. The features are listed along the vertical axis. Observations with a high value of a given feature are highlighted in red, and those with a low value are highlighted in blue. The position of the points along the horizontal axis shows the magnitude of the additive contribution of the value of the variable to the final prediction compared to the baseline.

The magnitude and sign of the influence for most items are natural: a high positive value of the item, as a rule, increases the value prediction. Once again, the item of *comprehensive income* turned out to be the most informative in the prediction formation. Note that the mean absolute

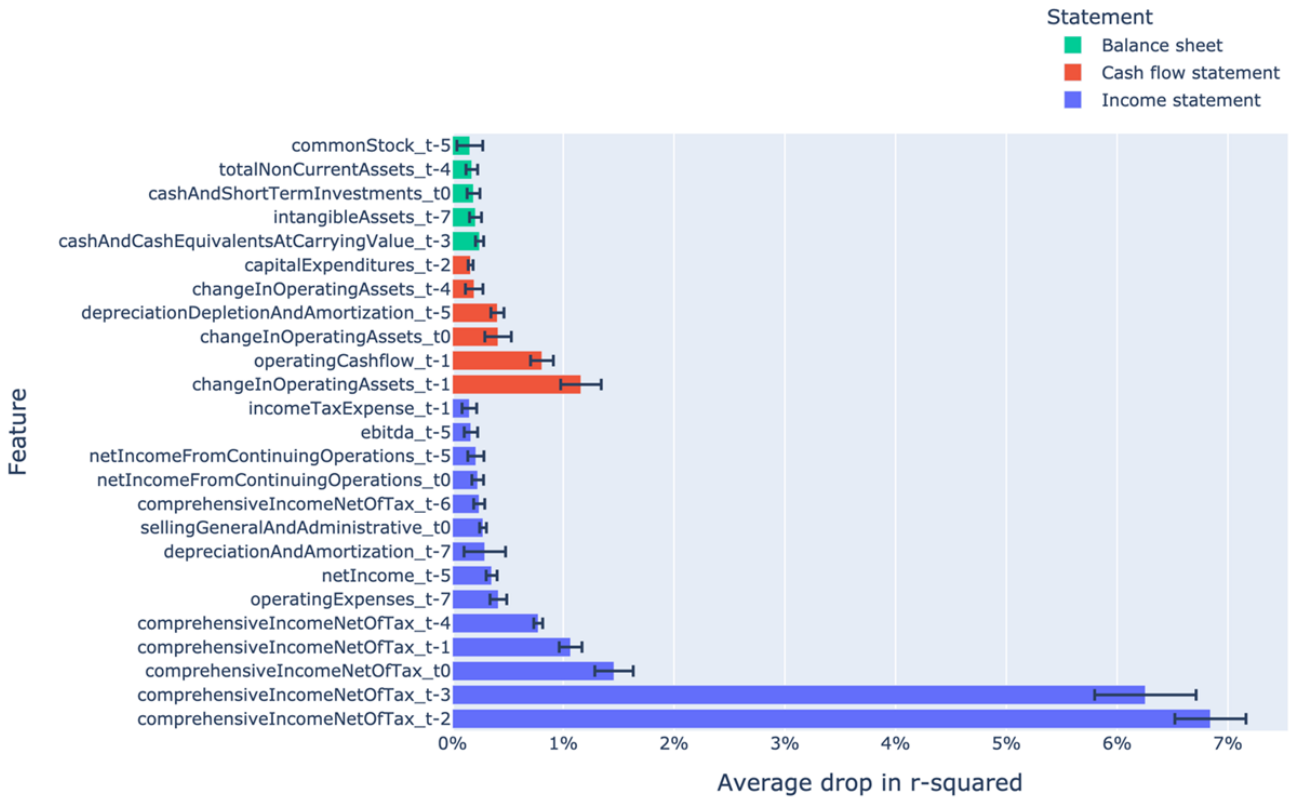

Fig. 4. **Feature importance. Permutation importance method**

Source: compiled by the author.

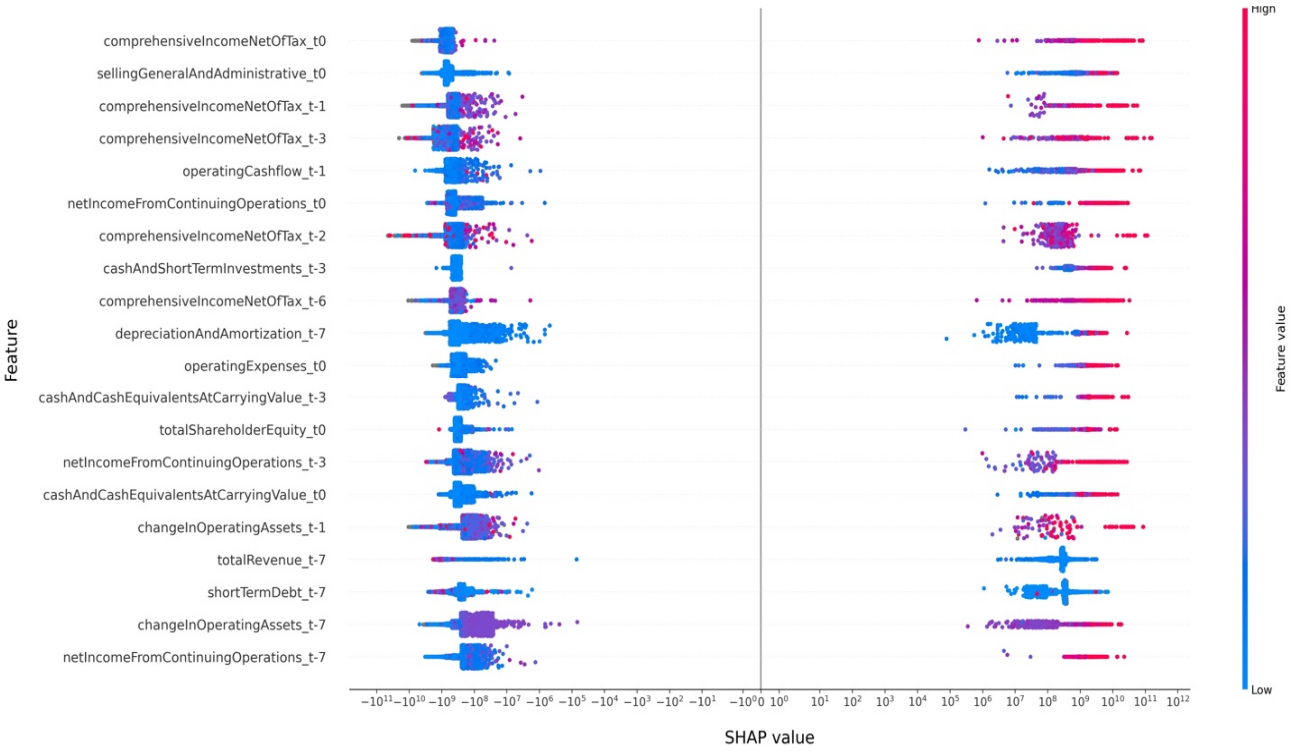

Fig. 5. **SHAP values for the most important features**

Source: compiled by the author.

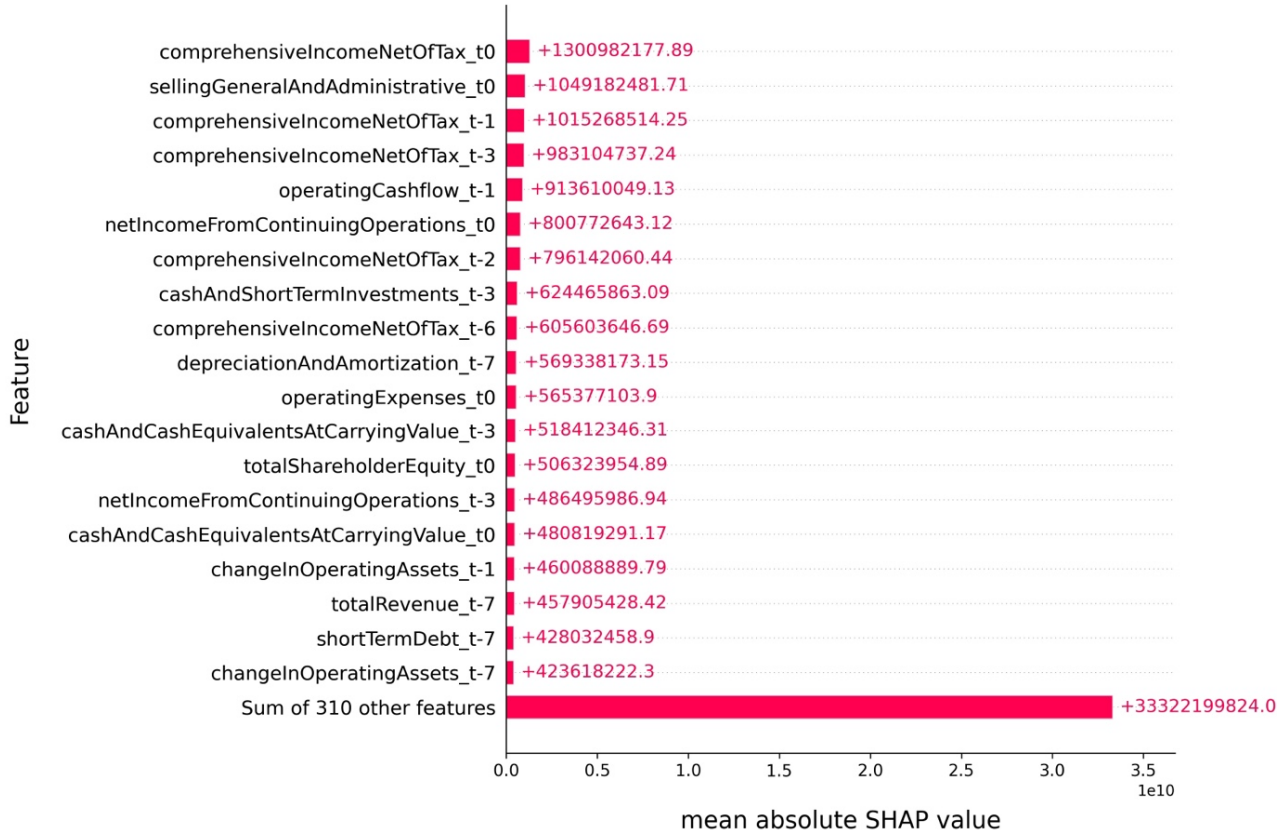

Fig. 6. **Mean absolute SHAP values for the most important features**

Source: compiled by the author.

SHAP value turned out to be the largest for the *comprehensive income* data for the last period (closest to today). Conversely, the contribution of low values for most variables is associated with smaller capitalization. The exceptions are *Revenue, Short-term debt*, and *Total liabilities*.

Thus, the model, and hence investors, encourage low-leverage companies. This may indicate that, in practice, for most companies, the current value of the tax shield is lower than the loss in value from the risk of financial distress resulting from the use of borrowed funds.

The negative impact of revenue is more difficult to explain. This may be an artifact of the selected data set. The top twenty important factors included only revenue data for the seventh quarter. Revenue figures for other quarters were relatively unimportant. The contribution of the variable characterizing the company's industry was also insignificant.

Features sorted by mean absolute SHAP value are summarized in a bar chart (*Fig. 6*). It can be seen that the absolute importance of some features is quite high, but their relative contribution to the final prediction is, as a rule, small. The total average contribution of the nineteen most important features is incomparably lower than the contribution of the remaining 310 features (the last line of *Fig. 6*). Thus, to form a highly accurate prediction, not individual items are used, but the data of the entire financial statements for each of the eight quarters.

3.3. Ideas for further consideration of the problem

Improvement of models obtained using machine learning methods can be carried out in several directions at once.

1. Engineering of new features.
2. Search for new features.

3. Creation of a dynamic version of the model.

In theory, machine learning methods are able to extract the necessary information from financial reporting data on their own, without calculating auxiliary features. However, this is only applicable to large datasets (millions of observations). Working in the mode of several thousand observations, the calculation of liquidity, turnover, and profitability of assets allows you to extract more information from the existing data set.

In addition to financial statements, other data, including unstructured data, may be useful for valuation. Extraction of textual information from annual reports [30], the use of substitute variables describing corporate governance, and the company's dividend policy, also improve the quality of predictions [31]. Including financial statement data for companies in both developed and emerging markets will provide a larger sample and a more generalized model (capable of making predictions for companies of different sizes, sectors, and regions). As a result, the developed models can be applied to the complex task of evaluating companies in the Russian stock market (Abramishvili, Lvova, & Voronova, 2019).

Fluctuations in the stock market indicate a strong influence of macroeconomic factors on the market capitalization of companies. Creating a dynamic version of the model will take into account macroeconomic variables and significantly increase the sample. The idea is that a company at two different points in time represents two different observations. For example, a dataset might include ten observations from *Apple Inc.* at different points in time. Each of the ten observations will have its own feature vector, consisting of the company's financial statements for the last eight quarters. In addition, this approach will allow expanding the feature vector with macroeconomic data corresponding to the moment of observation. As a result of applying this procedure for each company, the sample size will increase by about ten times. Together with taking into account macroeconomic factors, an increase in the sample will significantly increase the accuracy of capitalization predictions.

CONCLUSIONS

Let us list the main results and conclusions of the study.

1. Due to the ability to take into account non-linear relationships, machine learning methods significantly outperform traditional econometric approaches and are able to provide accurate estimates of the value of companies outside the training sample.

2. The economic benefits of using machine learning methods are enormous. *ML* allows you to digest a complex set of financial statement data and saves a lot of labor costs for highly qualified personnel. Instead of spending dozens of hours creating complex, multi-page *MS Excel* documents with the calculation of the value of a single company using the *DCF* method, an analyst can get an accurate and unbiased assessment of hundreds and even thousands of companies in a few seconds.

3. The resulting models can be used to solve applied problems of financial management, corporate finance, accounting, and investment analysis when creating trading strategies. A positive difference between the predicted capitalization and the actual one can be a criterion for including the company's shares in the portfolio.

4. Various ways of assessing the feature importance unanimously indicate the special value of the *Income statement* data and, in particular, the item's *Comprehensive income*. Further research on the feature importance will provide a better understanding of the company's value-creation process.

5. Improvement of the developed models can be carried out in the following areas: the use of a larger sample, and the creation, addition, and engineering of new features. The proposed dynamic version of the model will raise the already high accuracy of

company valuation to a qualitatively new level, presumably leaving the expert level far behind.

The results of the work confirm the considered hypothesis. The high value of R^2 for companies from the test sample clearly indicates the great potential of using machine learning methods to assess the value of an enterprise by predicting its market capitalization based on financial statements.

REFERENCES

1. Kryzanowski L., Galler M., Wright D.W. Using artificial neural networks to pick stocks. *Financial Analysts Journal*. 1993;49(4):21–27. DOI: 10.2469/faj.v49.n4.21
2. Cao L. AI in finance: A review. *SSRN Electronic Journal*. 2020. DOI: 10.2139/ssrn.3647625
3. Damodaran A. Investment valuation: Tools and techniques for determining the value of any asset. Hoboken, NJ: John Wiley & Sons, Inc.; 2012. 992 p.
4. Cybenko G. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*. 1989;2(4):303–314. DOI: 10.1007/BF02551274
5. Gu S., Kelly B., Xiu D. Empirical asset pricing via machine learning. *The Review of Financial Studies*. 2020;33(5):2223–2273. DOI: 10.1093/rfs/hhaa009
6. Tikhonov A.N. On the solution of ill-posed problems and the regularization method. *Doklady Akademii nauk*. 1963;151(3):501–504. URL: http://www.mathnet.ru/links/76d17d1b225aa6609693b033d8ad3c25/dan28329.pdf (In Russ.).
7. Tibshirani R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*. 1996;58(1):267–288. DOI: 10.1111/J.2517–6161.1996.tb02080.x
8. Hindman M. Building better models: Prediction, replication, and machine learning in the social sciences. *The Annals of the American Academy of Political and Social Science*. 2015;659(1):48–62. DOI: 10.1177/0002716215570279
9. Carvalho D.V., Pereira E.M., Cardoso J.S. Machine learning interpretability: A survey on methods and metrics. *Electronics*. 2019;8(8):832. DOI: 10.3390/electronics8080832
10. Damodaran A. Valuation approaches and metrics: A survey of the theory and evidence. Hanover, MA: Now Publishers Inc.; 2007. 104 p.
11. Pinto J.E., Robinson T.R., Stowe J.D. Equity valuation: A survey of professional practice. *Review of Financial Economics*. 2019;37(2):219–233. DOI: 10.1002/rfe.1040
12. Böhm-Bawerk E. Recent literature on interest (1884–1899): A supplement to "Capital and interest". New York: The MacMillan Co.; 1903. 151 p.
13. Atsalakis G.S., Valavanisb K.P. Surveying stock market forecasting techniques — Part II: Soft computing methods. *Expert Systems with Applications*. 2009;36(3):5932–5941. DOI: 10.1016/j.eswa.2008.07.006
14. Xing F.Z., Cambria E., Welsch R.E. Natural language based financial forecasting: A survey. *Artificial Intelligence Review*. 2018;50(1):49–73. DOI: 10.1007/s10462–017–9588–9
15. Park B., Bae J.K. Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data. *Expert Systems with Applications*. 2015;42(6):2928–2934. DOI: 10.1016/j.eswa.2014.11.040
16. Whitbeck V.S., Kisor M., Jr. A new tool in investment decision-making. *Financial Analysts Journal*. 1963;19(3):55–62. DOI: 10.2469/faj.v19.n3.55
17. Koklev P.S. Impact of the state ownership in equity on company value. *Tendentsii razvitiya nauki i obrazovaniya*. 2020;(60–8):14–18. (In Russ.). DOI: 10.18411/lj-04–2020–154
18. Joshi H., Chauha R. Determinants and prediction accuracy of price multiples for South East Asia: Conventional and machine learning analysis. *Indonesian Capital Market Review*. 2020;12(1):42–54. DOI: 10.21002/icmr.v12i1.12051
19. Liu J., Nissim D., Thomas J. Equity valuation using multiples. *Journal of Accounting Research*. 2002;40(1):135–172. DOI: 10.1111/1475–679X.00042

20. Hastie T., Tibshirani R., Friedman J. The elements of statistical learning: Data mining, inference, and prediction. 2nd ed. New York: Springer-Verlag; 2016. 767 p. (Springer Series in Statistics). DOI: 10.1007/978–0–387–84858–7
21. Prokhorenkova L., Gusev G., Vorobev A., Dorogush A.V., Gulin A. CatBoost: Unbiased boosting with categorical features. In: Proc. 32nd Int. conf. on neural information processing systems (NIPS'18). (Montréal, December 3–8, 2018). New York: Curran Associates Inc.; 2018:6639–6649. URL: https://proceedings.neurips.cc/paper/2018/file/14491b756b3a51daac41c24863285549-Paper.pdf
22. Ioannidis J., Doucouliagos C. What's to know about the credibility of empirical economics? *Journal of Economic Surveys*. 2013;27(5):997–1004. DOI: 10.1111/joes.12032
23. Bergstra J., Bengio Y. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*. 2012;13(2):281–305. URL: https://www.jmlr.org/papers/volume13/bergstra12a/bergstra12a.pdf
24. Breiman L. Random forests. *Machine Learning*. 2001;45(1):5–32. DOI: 10.1023/A:1010933404324
25. Shapley L.S. A value for n-person games. In: Kuhn H.W., Tucker A.W., eds. Contributions to the theory of games. Vol. II. Princeton, NJ: Princeton University Press; 2016:307–318. DOI: 10.1515/9781400881970–018
26. Banko M., Brill E. Scaling to very very large corpora for natural language disambiguation. In: Proc. 39th Annu. meet. of the Association for Computational Linguistics (ACL'01). (Toulouse, July 06–11, 2001). Stroudsburg, PA: Association for Computational Linguistics; 2001:26–33. DOI: 10.3115/1073012.1073017
27. Buck S.F. A method of estimation of missing values in multivariate data suitable for use with an electronic computer. *Journal of the Royal Statistical Society. Series B (Methodological)*. 1960;22(2):302–306. DOI: 10.1111/j.2517–6161.1960.tb00375.x
28. Munkhdalai L., Munkhdalai T, Namsrai O.-E., Lee J.Y., Ryu K.H. An empirical comparison of machine-learning methods on bank client credit assessments. *Sustainability*. 2019;11(3):699. DOI: 10.3390/su11030699
29. Huber P.J. Robust estimation of a location parameter. In: Kotz S., Johnson N.L., eds. Breakthroughs in statistics: Methodology and distribution. New York: Springer-Verlag; 1992:492–518. (Springer Series in Statistics). DOI: 10.1007/978–1–4612–4380–9_35
30. Sehrawat S. Learning word embeddings from 10-K filings for financial NLP tasks. *SSRN Electronic Journal*. 2019. DOI: 10.2139/ssrn.3480902
31. Kovalev V.V., Drachevsky I.S. Dividend policy as a factor for managing company value: Comparing trends in emerging markets. *Vestnik Sankt-Peterburgskogo universiteta. Ekonomika = St Petersburg University Journal of Economic Studies (SUJES)*. 2020;36(1):95–116. DOI: 10.21638/spbu05.2020.105
32. Abramishvili N.R., Lvova N.A., Voronova N.S. Is it possible to assess the corporate market value in the emerging market? In: New challenges of economic and business development — 2019: Incentives for sustainable economic growth. Proc. 11th Int. sci. conf. (Riga, May 16–18, 2019). Riga: University of Latvia; 2019:12–21. URL: https://dspace.lu.lv/dspace/handle/7/48896 (дата обращения: 18.12.2021).

ABOUT THE AUTHOR

Petr S. Koklev — postgraduate student of the Department of Credit Theory and Financial Management, Saint Petersburg State University, Saint Petersburg, Russia
https://orcid.org/0000-0003-2594-7973
koklevp@gmail.com

The author read and approved the final version of the manuscript.

Conflicts of Interest Statement: The author has no conflicts of interest to declare.
The article was submitted on 24.01.2022; revised on 11.02.2022 and accepted for publication on 27.04.2022.