DOI: 10.26794/2220-6469-2018-12-3-82-89
УДК 330.4(045)
JEL C43, C53, C55, C81

Прогнозирование волатильности финансовых временных рядов ансамблями деревьев

О.С. Видмант,
Финансовый университет,
Москва, Россия
https://orcid.org /000-0001-7331-4068

АННОТАЦИЯ

Применение нового инструментария для анализа экономических данных в последнее десятилетие привело к значительному улучшению прогнозирования. Это обусловлено как актуальностью поставленного вопроса, так и развитием технологий, которые позволяют применять более сложные модели, не прибегая к промышленным вычислительным мощностям. Постоянная волатильность мировых индексов вынуждает всех игроков финансового рынка совершенствовать модели риск-менеджмента, в то же время пересматривая политику инвестирования капитала. Ужесточающиеся нормативы ликвидности и прозрачности по отношению к финансовой сфере также подталкивают участников экспериментировать с защитными механизмами и создавать прогностические алгоритмы, способные не только снизить потери от волатильного изменения финансовых инструментов, но и получить выгоду от краткосрочных инвестиционных манипуляций.

В статье рассматривается возможность повышения эффективности вычислений при прогнозировании волатильности моделями ансамблей деревьев с использованием различных методов анализа данных. В качестве ключевых точек роста эффективности изучается возможность агрегирования данных финансовых временных рядов с использованием нескольких методов расчета и прогнозирования дисперсии: Standard, EWMA, ARCH, GARCH, а также анализируется возможность упрощения вычислений при снижении корреляционной зависимости между рядами. Применение расчетных методик демонстрируется на основе массива данных исторических цен (Open, High, Low, Close) и показателей объема (Volumes) торгов фьючерса на индекс RTS с пятиминутным временным интервалом и годовым набором исторических данных. Предлагаемая методика позволяет сократить мощностные/временные затраты на обработку данных при анализе краткосрочных позиций на финансовых рынках и выявить риски с определенным уровнем доверительной вероятности.

Ключевые слова: прогнозирование волатильности; эффективность; ансамбли деревьев; риски; финансовые инструменты; анализ данных; агрегирование

Для цитирования: Видмант О.С. Прогнозирование волатильности финансовых временных рядов ансамблями деревьев. *Мир новой экономики*. 2018;12(3):82-89. DOI: 10.26794/2220-6469-2018-12-3-82-89

DOI: 10.26794/2220-6469-2018-12-3-82-89
UDC 330.4(045)
JEL C43, C53, C55, C81

Forecasting the Volatility of Financial Time Series by Tree Ensembles

O.S. Vidmant,
Financial University,
Moscow, Russia
https://orcid.org /000-0001-7331-4

ABSTRACT

The use of new tools for economic data analysis in the last decade has led to significant improvements in forecasting. This is due to the relevance of the question, and the development of technologies that allow implementation of more complex models without resorting to the use of significant computing power. The constant volatility of the world indices forces all financial market players to improve risk management models and, at the same time, to revise the policy of capital investment. More stringent liquidity and transparency standards in relation to the financial sector also encourage participants to experiment with protective mechanisms and to create predictive algorithms that can not only reduce the losses from the volatility of financial instruments but also benefit from short-term investment manipulations. The article discusses the possibility of improving the efficiency of calculations in predicting the volatility by the models of tree ensembles using various methods of data analysis. As the key points of efficiency growth, the author studied the possibility of aggregation of financial time series data using several methods of calculation and prediction of variance: Standard, EWMA, ARCH, GARCH, and also analyzed the possibility of simplifying the calculations while reducing the correlation between the series. The author demonstrated the application of calculation methods on the basis of an array of historical price data (Open, High, Low, Close) and volume indicators (Volumes) of futures trading on the RTS index with a five-minute time interval and an annual set of historical data. The proposed method allows to reduce the cost of computing power and time for data processing in the analysis of short-term positions in the financial markets and to identify risks with a certain level of confidence probability.

Keywords: volatility forecasting; efficiency; tree ensembles; risks; financial instruments; data analysis; aggregation

For citation: Vidmant O.S. Forecasting the volatility of financial time series by tree ensembles. *Mir novoi ekonomiki = Word of the new economy*. 2018;12(3):82-89. (In Russ.). DOI: 10.26794/2220-6469-2018-12-3-82-89

С развитием технологического прогресса исследователи получили возможность анализировать потоки данных огромного объема и находить благодаря этому новые закономерности в различных областях исследования. Делегируя математические операции вычислительным системам, исследователи смогли получить много интересных данных об особенностях связей между множествами объектов. Использование симбиоза математико-информационных моделей позволило успешно решать многие вопросы в области экономики (https://web.stanford.edu/~leinav/pubs/Science2014.pdf), метеорологии (https://www.geosci-model-dev-discuss.net/gmd-2015–273/gmd-2015–273.pdf), маркетинга (http://web.media.mit.edu/~yva/papers/sundsoy2014big.pdf) и потребления (https://www.elen.ucl.ac.be/Proceedings/esann/esannpdf/es2013–45.pdf), что, в свою очередь, привело к необходимости анализа/агрегирования больших потоков информации. На передний план вышла проблема очистки данных от случайных возмущений, так называемая проблема зашумления данных, а также проблема снижения неизбежно возникающего роста мощностных/временных затрат на их обработку.

Эффективный анализ и управление исходными данными позволяют значительно сократить финансовые издержки на обработку данных, а также кардинально пересмотреть мощностные затраты. В качестве факторов, влияющих на оценку эффективности обработки временных рядов, в статье рассматривается объем данных и точность прогнозирования. В дальнейшем на основе такого анализа возникает

возможность пересмотра стратегии управления рисками в зависимости от предпочтений инвестора с помощью снижения/увеличения временных или мощностных ограничений, направленных на обработку информации.

Для анализа финансового рынка в статье используются модели ансамблей деревьев, разработанные с учетом необходимости устранения зашумления данных и уменьшения корреляционных связей между финансовыми временными рядами, с разделением массива данных на обучающую и контрольную выборки.

ДЕРЕВЬЯ РЕШЕНИЙ

Деревья решений — это модели, используемые для получения результата качественного характера в задачах регрессии и классификации [1]. В основе методики заложен алгоритм разделения (расщепления) данных на несколько сегментов (категорий) при использовании различных моделей расчета в каждом из сегментов (так называемых ответвлений). Общая схема алгоритма решения задачи методом деревьев решений представлена на *рис. 1*.

На *рис. 1* показано, как скоринговая задача классификации решается с помощью выделения так называемых терминальных узлов (терминальными узлами на *рис. 1* являются данные, способные в дальнейшем расщепиться на отдельные категории согласно методу разделения), в которых для сегментирования данных используется единый алгоритм. В основе алгоритма расщепления данных лежат признаки количественного характера, такие как коэффициент Джини или функция информационного прироста (http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.57.9764&rep=rep1&type=pdf).

Одними из сильных сторон модели деревьев решений являются визуализация и возможность интуитивного анализа, однако у этой модели есть и слабые стороны. Основной ее недостаток — малая обобщающая способность, а также быстрая переобучаемость модели при отсутствии ручной калибровки параметров. Именно поэтому для улучшения точности прогнозирования прибегают к более сложным ансамблевым моделям деревьев.

АНСАМБЛЕВЫЕ МЕТОДЫ

Ансамблевые методы деревьев являются логическим продолжением модели деревьев решений [2]. Эти методы могут включать построение и анализ множества классификаторов, значения которых при обработке дают более взвешенные оценки. Эти модели значительно превосходят по качеству прогнозирования линейные модели, или модели отдельных деревьев классификации [3], а также могут успешно противостоять как зашумлению данных, так и переобучению.

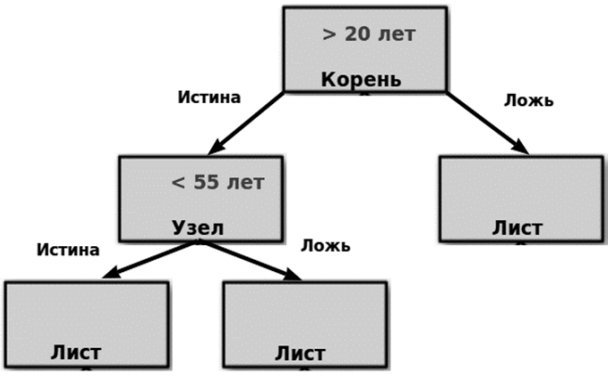

Рис. 1 / Fig. 1. **Пример расщепления данных с использованием модели деревьев решений / Example of data splitting using decision tree model**

В данной статье рассматривается использование четырех ансамблевых моделей:

Boosting Methods:

- AdaBoost (AB) [4].
- Gradient Boosting (GBM) [5].

Bagging Method:

- Random Forests (RF) [6].
- Extra trees (ET) [7].

Основанные на алгоритмах манипулирования выборками *Bagging Methods* применяются к каждому дереву ансамбля. Они заключаются в формировании на основе оригинальных данных случайной обучающей выборки, которая сохраняет размерность оригинала и в дальнейшем корректирует итоговый результат на основе взвешивания всех полученных ответов.

Boosting Methods основаны на постоянном итеративном взвешивании и коррекции наблюдений в выборке при постоянном отслеживании результативности проведенной операции. Операции коррекции должны устранять ошибки, которые были сделаны на предыдущих шагах. Классификаторы в дальнейшем взвешиваются в зависимости от допущенного количества ошибок.

МЕТОДЫ РАСЧЕТА ВОЛАТИЛЬНОСТИ

В качестве агрегирующих информацию из финансовых временных рядов моделей мы будем использовать четыре хорошо известных и ставших классическими метода расчета дисперсии:

$$\text{Standard } \sigma_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(r_i - \bar{r}\right)^2.$$

Таблица 1 / Table 1

Формат использованных данных фьючерса РТС (SPFB.RTS) за 2016–2017 гг. / The format of the futures RTS (SPFB.RTS) for 2016–2017

TICKER	DATE	TIME	OPEN	HIGH	LOW	CLOSE	VOL
SPFB.RTS	2016:08:10	20:25:00	94 270	94 310	94 250	94 260	883
SPFB.RTS	2016:08:10	20:30:00	94 260	94 340	94 240	94 310	1118
––––––-	––––––-	–––––	–––––	–––––	–––––	–––––	–––––
SPFB.RTS	2017:08:09	10:10:00	104 190	104 290	104 160	104 220	5694
SPFB.RTS	2017:08:09	10:15:00	104 220	104 230	103 940	103 940	9484

ARCH: $\sigma_n^2 = \alpha\theta^2 + (1-\alpha)\frac{1}{n}\sum_{i=1}^{n} r_i^2$. [8]

EWMA: $\sigma_n^2 = \lambda\sigma_{n-1}^2 + (1-\lambda)r_n^2$.

GARCH: $\sigma_n^2 = \alpha\theta^2 + (1-\alpha)(\lambda\sigma_{n-1}^2 + (1-\lambda)r_n^2$, [9]

где r_i — доходность i-го дня; $\bar{r}$ — выборочное среднее значение доходности; α — коэффициент веса; θ — дисперсия временного интервала $n-1$; λ — весовой коэффициент.

ИСПОЛЬЗОВАНИЕ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ПРОГНОЗИРОВАНИЯ ВОЛАТИЛЬНОСТИ

Для прогнозирования волатильности на 5-минутных временных отрезках воспользуемся набором данных по фьючерсу РТС (SPFB.RTS) (https://www.finam.ru/profile/mosbirzha-fyuchersy/rts/export/) с 5-минутным интервалом и годовым массивом исторических показателей (09.08.2016–09.08.2017). Иллюстрация структуры исходных данных приведена в *табл. 1*.

Добавим в данную таблицу рассчитанные значения доходности и модуля доходности, необходимые для расчета среднего уровня волатильности, и определим бинарный класс для значения каждой из строк временного ряда по формуле

$$class_n = r_n > Me(|dif|),$$

где коэффициент размытия *dif* варьируется в пределах:

$$dif = \{r_{n-1}, \ldots, r_{n-n+1}\}$$

в зависимости от предпочтений исследователя.

Классификация строится следующим образом: в случае если доходность временного промежутка

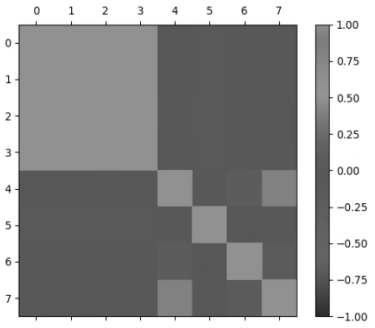

Рис. 2 / Fig. 2. **Корреляционная матрица / Correlation matrix**

n (r_n) окажется больше медианы (*Me*) модулей доходности временных промежутков, классу присваивается значение 1, в обратном случае — 0. В работе используется $dif = \{r_{n-1}, r_{n-2}\}$, что позволяет строить классификацию по самым свежим данным, повышая, таким образом, чувствительность модели.

Отобразим получившуюся структуру набора данных в *табл. 2*.

Со статистической точки зрения данные представляют массив с размерностью (42135,12). Классы распределены неравномерно в следующем соотношении:

- **Class 0**: 28 179 (67%).
- **Class 1**: 13 754 (32%).

Оценим корреляционную зависимость между исходными данными для выявления информации, которой в дальнейшем можно будет пренебречь без ухудшения прогностической способности модели, и визуализируем ее с помощью корреляционной матрицы (*рис. 2*).

На *рис. 2* продемонстрирована сильная линейная зависимость между финансовыми рядами Open, High, Low, Close, а значит, данные соответствующих этим показателям рядов являются избыточными и не вносят дополнительной полезной информации, а возможно, и способствуют снижению продуктивности (*табл. 3*).

Изучим асимметрию и гистограммы эмпирического распределения для выявления наилучшего

Таблица 2 / Table 2

Набор данных для решения задачи классификации / Dataset to solve the classification problem

TICKER	DATE	TIME	OPEN	HIGH	LOW	CLOSE	VOL	RET	Abs(RET)	Class
SPFB.RTS	2017:08:09	10:10:00	104190	104290	104160	104220	5694	0,002694	0,002694	0
SPFB.RTS	2017:08:09	10:15:00	104220	103230	103940	103940	9484	–0,00048	0,00048	1
SPFB.RTS	2017:08:09	10:20:00	103940	104020	103820	103990	10955	–0,00096	0,00096	0
SPFB.RTS	2017:08:09	10:25:00	103990	104140	103980	104090	6054	–0,00239	0,00239	1
SPFB.RTS	2017:08:09	10:30:00	104100	104370	104080	104340	10702	–0,00019	0,00019	1

Таблица 3 / Table 3

Корреляционная взаимосвязь рядов / Correlation relationship of time series

	OPEN	HIGH	LOW	CLOSE	VOL	RET	Abs(RET)	Class
OPEN	1	0,999	0,999	0,999	–0,309	0,009	–0,001	–0,012
HIGH	0,999	1	0,999	0,999	–0,025	0,001	–0,007	–0,009
LOW	0,999	0,999	1	0,999	–0,035	0,009	–0,002	–0,001
CLOSE	0,999	0,999	0,999	1	–0,030	0,001	–0,001	–0,001
VOL	–0,030	–0,025	–0,035	–0,030	1	0,004	0,092	0,400
RET	0,009	0,010	0,009	0,010	0,004	1	–0,028	–0,005
Abs(RET)	–0,001	–0,007	–0,002	–0,001	0,092	–0,028	1	0,062
Class	–0,012	–0,009	–0,001	–0,001	0,400	–0,005	0,062	1

временного ряда, который в дальнейшем необходимо оставить для исследований.

Как можно видеть из *табл. 3* и *рис. 2*, часть данных имеет сильную корреляцию, что может воздействовать негативно на качество модели и скорость расчета/затраченное количество мощности. Кроме того, в гистограммах наблюдается присутствие толстых хвостов. Поэтому в качестве временного ряда, отвечающего нашим потребностям, будет использоваться ряд Close, который может рассматриваться как усредненная версия других рядов по показателю асимметрии.

Далее будет проведено сравнение результатов, полученных на полном наборе данных, с результатами расчета на данных, упрощенных с учетом корреляционной зависимости.

На основе массива данных, представленных в *табл. 2*, произведем расчет с использованием ансамблевых моделей. Результаты расчета представлены в *табл. 5*.

Как видно из *табл. 5*, на основе анализа первичных (необработанных) данных можно получить максимальный уровень предсказания чуть выше 67%. Также следует отметить, что для расчета использовалось все 7 временных рядов (см. *табл. 3*).

Далее воспользуемся различными методами расчета волатильности, в частности:

- Standard.
- EWMA.
- ARCH.
- GARCH.

Проверим агрегирующую способность данных моделей, создав таблицу из двух временных рядов, где первый столбец отражает значения волатильности, рассчитанные на основе одной из четырех упомянутых моделей, а второй — бинарный признак классификации (*табл. 6*).

На основе каждой из таблиц, приведенных под общим названием *табл. 6*, рассчитаем прогностические показатели для четырех ансамблевых моделей деревьев решений (*табл. 7*).

В *табл.* 7 с данными расчетов показано, что уровни предсказания сопоставимы с уровнем, полученным на основе анализа исторических данных. Таким образом, снизив объем расчетов в 7 раз, мы получили сопоставимый результат.

Применим стандартный (Standard) метод расчета дисперсии, показавший лучший с точки зрения прогностической способности результат, к ряду исторических данных, и оценим прогностические способности ансамблевых моделей (*табл. 8*).

Можно отметить, однако, что при таком подходе оптимизация вычислительного процесса не доведена

Таблица 4 / Table 4

Оценка асимметрии / Estimation of asymmetry

Open	0,226482
Hight	0,226276
Low	0,226640
Close	0,226477
Volume	2,973177
Returns	–1,474955
Abs_ret	14,947850
Class	0,732747

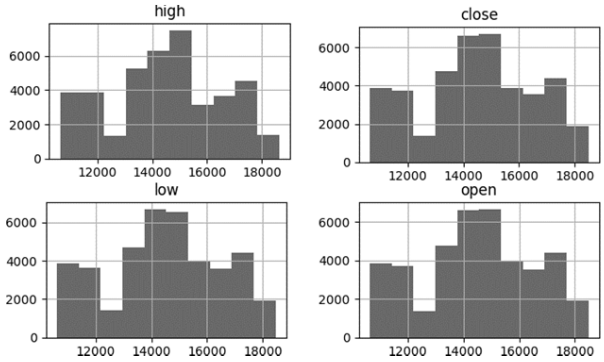

Рис. 3 / Fig. 3. **Гистограммы распределения временных рядов / A histogram of the distribution of the time series**

Таблица 5 / Table 5

Прогностические результаты моделей на основе исторических данных / Predictive results of models based on historical data

Ансамбли деревьев	AB	GBM	RF	ET
Уровень прогнозирования	0,670184	0,670333	0,650450	0,637423

Таблица 6 / Table 6

Таблицы на основе моделей волатильности и бинарного признака классификации / Tables based on models of volatility and binary trait classification

Arch	Class		Garch	Class		Ewma	Class		Stand	Class
9,29e-07	0		1,01 e-06	0		6,40 e-07	0		6,47 e-07	0
9,32 e-07	1		1,14 e-06	1		6,67 e-07	1		6,11 e-07	1
9,35 e-07	0		1,24 e-06	0		3,25 e-07	0		6,13 e-07	0
9,39 e-07	1		6,99 e-07	1		3,44 e-07	1		5,82 e-07	1

до конца: не использована возможность снижения объема вычислений за счет должного учета корреляционных зависимостей внутри временного ряда. Проанализировав таблицу корреляционных зависимостей (см. *рис. 2*), можно сделать вывод, что имеется существенное количество избыточных данных, которые могут снижать эффективность вычислений из-за высокой линейной зависимости между временными рядами. Исключим из анализа часть фрагментов ряда (корреляционная зависимость, близкая к единице, позволяет оставить лишь независимый фрагмент, удалив все фрагменты, приблизительно линейно от него зависящие) и получим скорректированный набор данных (*табл. 9, рис. 4*).

Произведем перерасчет ансамблевых моделей на скорректированном наборе данных, и заново оценим прогностическую способность моделей (*табл. 10*).

Расчетные показатели (см. *табл. 10*) позволяют нам сделать вывод, что скорректированный с точки зрения удаления линейно коррелированных фрагментов набор данных при расчете дисперсии на основе стандартного метода, а также с использованием модели градиентного бустинга продемонстрировал лучший результат как с точки зрения распределения временных/мощностных ресурсов, так и с точки зрения уровня достоверности прогноза (*рис. 5*). Для построения достоверного прогнозирования одного и того же уровня в 74% случаев было использовано на 42% меньше данных, что значимо снижает уровень временных/мощностных затрат на производимые расчеты.

Таблица 7 / Table 7

Прогностические показатели ансамблевых моделей для каждой из моделей дисперсии / Predictive indicators of ensemble models for each of the dispersion models

Тип модели дисперсии	AB	GBM	RF	ET
Standard	0,669827	0,670125	0,669976	0,669737
EWMA	0,669827	0,669469	0,668992	0,668217
ARCH	0,669946	0,669946	0,669916	0,668574
GARCH	0,669856	0,669469	0,669320	0,668396

Таблица 8 / Table 8

Прогностические показатели ансамблевых моделей на основе совокупности исторических данных и Standard модели расчета дисперсии / Predictive indicators of ensemble models based on a set of historical data and the Standard model of dispersion calculation

Вид дисперсии	AB	GBM	RF	ET
Уровень прогнозирования	0,735050	0,740893	0,718208	0,698682

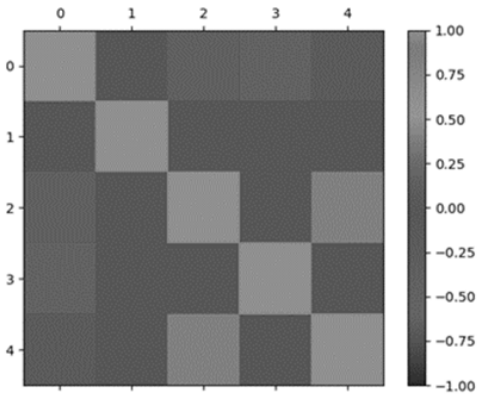

Рис. 4 / Fig. 4. **Корреляционная матрица / Correlation matrix**

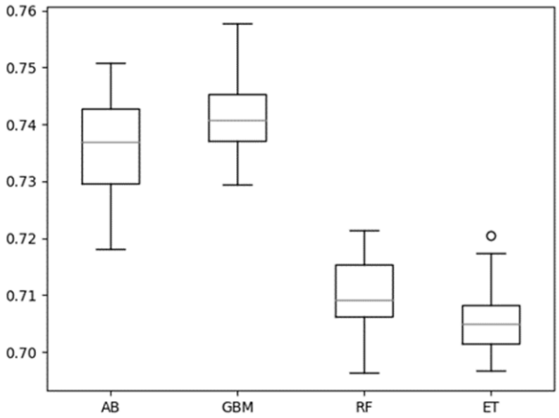

Рис. 5 / Fig 5. **Уровни достоверности ансамблей / Ensemble confidence levels**

Таблица 9 / Table 9

Скорректированный набор данных / A corrected set of data

TIME	CLOSE	VOL	Var	Class
10:10:00	104 220	5694	6,47 e-07	0
10:15:00	103 940	9484	6,11 e-07	1
10:20:00	103 990	10 955	6,13 e-07	0
10:25:00	104 090	6054	6,09 e-07	1
10:30:00	104 340	10 702	5,82 e-07	1

Таблица 10 / Table 10

Прогностические показатели моделей на основе скорректированных данных / The predictive indicators of models based on adjusted data

Вид дисперсии	AB	GBM	RF	ET
Уровень прогнозирования	0,735855	0,741132	0,709295	0,706343

В век развивающихся технологий и усложняющихся финансовых процессов решение задач прогнозирования приобретает все большую актуальность [10]. Разработка эффективных методов прогнозирования волатильности и риска может не только существенно сократить финансовую нагрузку на капитал, но и позволит переосмыслить всю систему риск-менеджмента. Использование моделей машинного обучения дает возможность исследовать финансовые активы на предмет наличия определенных паттернов на отдельных временных промежутках; агрегирование и предварительная обработка «сырых» данных позволяет обнаружить не учтенные ранее зависимости.

В работе исследовалась возможность прогнозирования волатильности ликвидного финансового инструмента на основе разработанных автором методов и метрик классификации. Полученные результаты убедительно демонстрируют эффективность применения классификационных ансамблевых моделей машинного обучения для прогнозирования волатильности на 5-минутных временных интервалах. Кроме того, была подтверждена гипотеза о возможности улучшения уровня прогнозирования посредством агрегирования исходных данных согласно предложенной автором методике.

REFERENCES

1. Quinlan J.R. Induction of Decision Trees. *Machine Learning*. 1986;(1):81–106.
2. Sollich P., Krogh, A. Learning with ensembles: How overfitting can be useful. *Advances in Neural Information Processing Systems*. 1996;8:190–196.
3. Yoo W., Ference B.A., Cote M.L., & Schwartz A.A. Comparison of Logistic Regression, Logic Regression, Classification Tree, and Random Forests to Identify Effective Gene-Gene and Gene-Environmental Interactions. *International Journal of Applied Science and Technology*. 2012;2(7):268.
4. Freund Y., Shapire R. A decision-theoretic generalization of on-line learning and an application to boosting. Proceedings of the Second European Conference on Computational Learning Theory; 1995:23–37.
5. Friedman J. Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics*. 2001;29(5):1–39.
6. Breiman L. Random Forests. *Machine Learning*. 2001;45:5–32.
7. Geurts P., Erns D., Wehenkel L. Extremely randomized trees. *Machine Learning*. 2006;63:3–42.
8. Engle R.F. Estimating Time Varying Risk Premia in the Term Structure: The Arch-M Model. *Econometrica*. 1987;55(2):391–407.
9. Raileanu L.E., Stoffel K. Theoretical Comparison between the Gini Index and Information Gain Criteria. *Annals of Mathematics and Artificial Intelligence*. 2004;41(1):77–93.
10. Bollerslev T. Generalized Autoregressive Conditional Heteroskedasticity. *Journal of Econometrics*. 1986;31:307–327.
11. Sundsoy P., Bjelland J., MIqba A. Big Data-Driven Marketing: How machine learning outperforms marketers' gut-feeling. Massachusetts Institute of Technology. *Lecture Notes in Computer Science*. 2014;8393:367–374.

ИНФОРМАЦИЯ ОБ АВТОРЕ

Видмант Олег Сергеевич — аспирант Департамента анализа данных, принятия решений и финансовых технологий, Финансовый университет, Москва, Россия
Vis.oleg@mail.ru

ABOUT THE AUTHOR

Oleg S. Vidmant — Postgraduate student, Department of data analysis, decision-making and financial technologies, Financial University, Moscow, Russia
Vis.oleg@mail.ru