УДК 004.89

Сравнение методов машинного обучения для задачи прогнозирования в среде Wolfram Mathematica

Хлебородова Лидия Дмитриевна
Сахалинский государственный университет
Студент

Осипов Геннадий Сергеевич
Сахалинский государственный университет
д.т.н., заведующий кафедрой Информатики

Аннотация
Дана формальная постановка задачи прогнозирования (предсказания) по результатам наблюдений – обучающей выборки, представленной как большой массив исходных данных. Сформулированы критерии оценки качества прогнозирования. Разработан программный комплекс, который позволяет определить наилучший метод машинного обучения, позволяющий решить задачу, обеспечивая оптимальное значение показателя качества прогнозирования.
Ключевые слова: прогнозирование, методы машинного обучения.

Comparison of machine learning methods for forecasting tasks in the Wolfram Mathematica environment

Khleborodova Lidiya Dmitrievna
Sakhalin State University
Student

Osipov Gennady Sergeevich
Sakhalin State University
Doctor of Technical Sciences, Head of the Department of Computer Science

Abstract
A formal formulation of the forecasting problem (prediction) based on the results of observations is given – a training sample presented as a large array of initial data. Criteria for assessing the quality of forecasting are formulated. A software package has been developed that allows you to determine the best machine learning method that allows you to solve the problem by providing the optimal value of the prediction quality indicator.
Keywords forecasting, machine learning methods.

Введение

В настоящее время для решения задач прогнозирования на больших данных (Big Data) широко применяется концепция использования методов машинного обучения. Поэтому актуальной является проблема определения наилучшего (оптимального) метода по выбранному критерию качества, обучения, например, минимальному значению ошибки предсказания.

Постановка задачи

Имеется выборка котировок ценной бумаги [1, 2] за определенный период времени. Многомерный временной ряд оформлен в виде массива данных, представленных на рисунке 1.

	Date	Open	High	Low	Close	Amount	Volume	Close+
2	20.01.2019	243,99	244,74	241,06	243,25	2541960	617617305,1	**243,73**
3	21.01.2019	244,8	246,9	243,5	**243,73**	2966460	727341922,9	242,65
4	22.01.2019	243,52	244,72	242,11	242,65	2712340	659072971,4	240,82
5	23.01.2019	242	242,64	240,03	240,82	2103330	507203050,3	240,87
6	24.01.2019	241,75	243,5	239,15	240,87	2594210	626355852,6	239,1
7	25.01.2019	241,18	242,28	238,38	239,1	2936390	703924933,3	236,1
8	26.01.2019	239,87	241,37	236	236,1	3474100	826500817,5	234,5
					…			
1000	14.10.2021	321	321,55	316,75	318	2734760	871617550	318
1001	15.10.2021	318	319,55	316,6	318	3446200	1097277610	315,9
1002	16.10.2021	319	319,85	310,75	315,9	6157270	1939588227	315,1
1003	17.10.2021	315,15	317,7	312,75	315,1	4402740	1388166557	318,5

Рисунок 1. Обучающая выборка

Здесь x = {Open, High, Low, Close, Amount, Volume} – входные данные, где:

Open – цена открытия торговой сессии (дня);

High – наивысшая цена за торговый день;

Low – наименьшая цена;

Close – цена закрытия сессии;

Amount – количество сделок (покупок/продаж) с ценной бумагой;

Volume – суммарная стоимость сделок.

y = {Close+} – цена закрытия на следующий день.

Требуется на основании данных x об итогах текущей сессии сделать прогноз цены закрытия y ценной бумаги на следующий день, т.е. установить соответствие $x \rightarrow y$.

Таким образом целью настоящего исследования является разработка программной системы, позволяющей сравнить методы машинного обучения с целью обоснования выбора оптимального метода прогнозирования.

Метод решения

Для решения задачи исследовались следующие методы машинного обучения:

1. Линейная регрессия Linear Regression (LR)

2. Дерево решений Decision Tree (DT)
3. Градиентный бустинг Gradient Boosted Trees (GBT)
4. Метод ближайшего соседа Nearest Neighbors (NN)
5. Метод случайного леса Random Forest (RF)
6. Гауссовский процесс Gaussian Process (GP)

В качестве инструментария выбрана система символьной математики (компьютерной алгебры) Wolfram Mathematica [3].

Основные критерии оценки качества прогноза:

1. Среднеквадратичная ошибка (Root Mean Square Error)

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}\left(y_i - \tilde{y}_i\right)^2}{n}},$$

где y_i – цена закрытия на следующий день, $\tilde{y}_i$.– расчетное (предсказанное) значение цены закрытия.

2. Средняя процентная ошибка (Mean Absolute Percentage Error)

$$\mathrm{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\frac{\left|y_i - \tilde{y}_i\right|}{y_i}\cdot 100\%$$

3. Ошибка экстраполирования (Relative Extrapolation Error)

$$\mathrm{REE} = \frac{\left|y_{n+1} - y^*\right|}{y^*}\cdot 100\%$$

где y^* – эталонное (проверочное) значение цены, y_{n+1} – значение цены закрытия на текущий (предшествующий прогнозному) день.

Основные результаты

В качестве примера на рисунке 2 представлены типовые операторы обучения на обучающей выборке и прогнозирования (экстраполирования) на следующий торговый период.

…

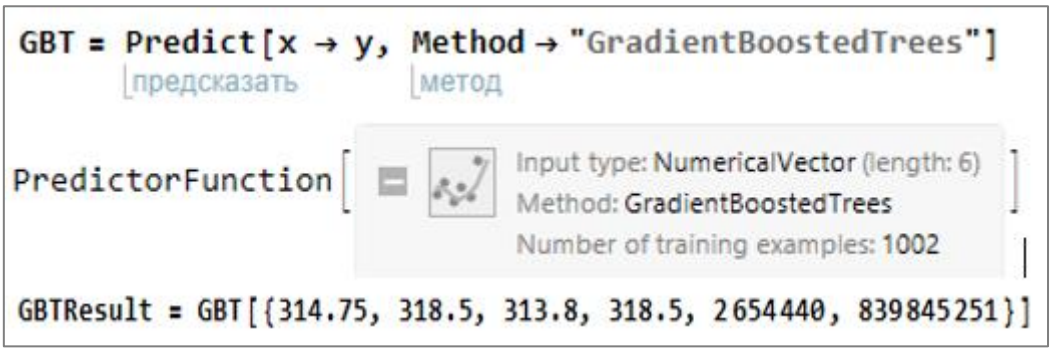

Рисунок 2. Операторы обучения и прогнозирования

Сравнение результатов прогноза в виде столбиковой диаграммы приведено на рисунке 3.

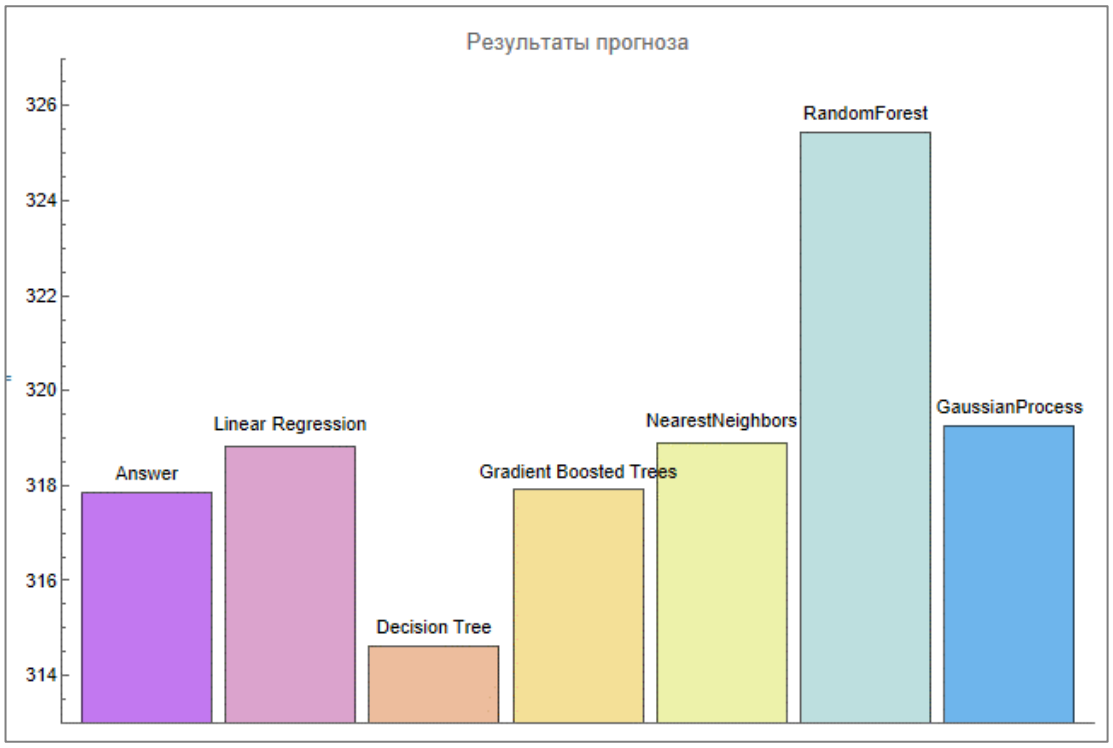

Рисунок 3. Результаты прогноза на следующий день

Среднеквадратичная ошибка прогноза для обучающей выборки показана на рисунке 4.

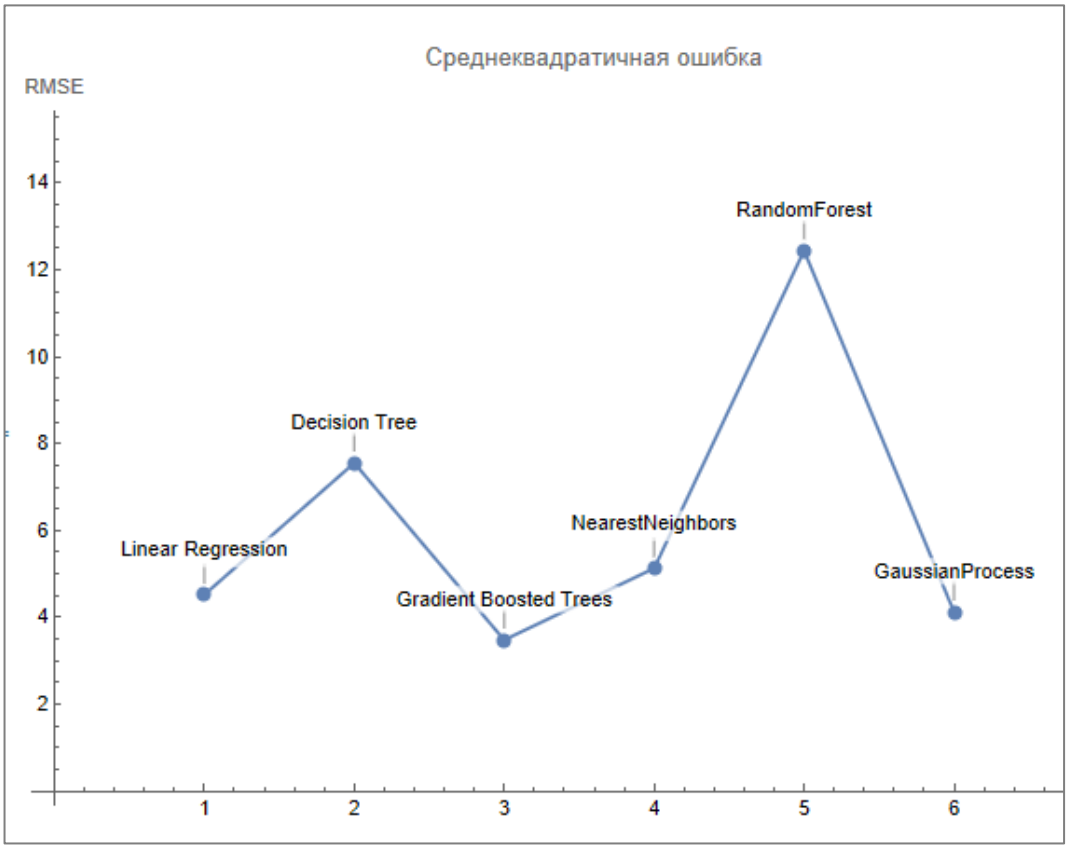

Рисунок 4. Величины среднеквадратичной ошибки

Итоговый показатели сравнения исследуемых методов машинного обучения применительно к задаче прогнозирования сведены в таблицу.

Таблица – Сравнение результатов прогноза

Метод	Затраты времени (с)		Ошибки		
	Построение модели	Прогноз	RMSE	MAPE (%)	REE (%)
Линейная регрессия	1.56	0.003	4.52	1.23	0.27
Дерево решений	0.37	0.003	7.55	2.21	1.02
Градиентный бустинг	2.21	0.007	**3.48**	**0.97**	**0.02**
Метод ближайшего соседа	0.39	0.003	5.46	1.42	0.48
Метод случайного леса	0.52	0.005	12.44	3.96	2.38
Гауссовский процесс	9.37	0.008	4.04	1.12	0.48

Выводы

Проведенное исследование и компьютерное моделирование позволяет сделать следующие выводы:

1. При использовании метода градиентного бустинга величины среднеквадратической ошибки, средней процентной ошибки и относительной ошибки экстраполирования являются минимальными. Таким образом, при решении задач такого класса оптимальным решением является применение такого метода машинного обучения, как градиентный бустинг.

2. Программная реализация задачи аналитического сравнения методов машинного обучения применительно к задаче прогнозирования на больших массивах данных является унифицированной и может быть использована в различных предметных областях.

Библиографический список

1. Ким Н. Г. Прогнозирование котировок ценных бумаг методами линейной регрессии, дерева решений и с помощью многослойной нейронной сети / Н. Г. Ким, Л. Д. Хлебородова // Студент года 2021: Сборник статей Международного учебно-исследовательского конкурса в 6-ти частях,

Петрозаводск, 19 мая 2021 года. – Петрозаводск, 2021. – С. 288-292. – DOI 10.46916/02062021-4-978-5-00174-249-4.
2. Ким Н. Г., Осипов Г. С. Исследование влияния соотношения тренировочных и тестовых данных на точность прогнозирования с помощью многослойной нейронной сети // Постулат. 2021. №10. С. 5.
3. Stephen Wolfram. An Elementary Introduction to the Wolfram Language. URL: https://www.wolfram.com/language/elementary-introduction/2nd-ed/ (Дата обращения 18.10.2021).