

УДК 519.257

DOI: 10.35211/1990-5297-2023-1-272-62-65

*Д. С. Князев, А. П. Порхун, Е. Г. Шведов, В. Г. Барабанов***ПРИМЕНЕНИЕ ГРАДИЕНТНОГО УСИЛЕНИЯ ДЛЯ ОПТИМИЗАЦИИ МОДЕЛИ
МЕТЕОРОЛОГИЧЕСКОГО ПРОГНОЗИРОВАНИЯ НА ДОЛГОСРОЧНЫЙ ПЕРИОД****Волгоградский государственный технический университет**

E-mail: dimagold.dg53@gmail.com

Рассматривается регрессионная модель, прогнозирующая погоду на долгосрочный период. Предлагается способ оптимизации модели. Описывается методика обучения модели на основе алгоритма градиентного усиления. Оценивается ее точность прогнозирования и сравнивается с предыдущей версией модели.

Ключевые слова: модель, градиентное усиление, базовый алгоритм, ДПП, выборка, функция потерь, аппроксимация.

*D. S. Knyazev, A. P. Porkhun, E. G. Shvedov, V. G. Barabanov***APPLICATION OF GRADIENT BOOSTING TO OPTIMIZE THE METEOROLOGICAL
FORECASTING MODEL FOR THE LONG-TERM PERIOD****Volgograd State Technical University**

A regression model predicting the weather for a long-term period is considered. A method for optimizing the model is proposed. The method of training the model based on the gradient boosting algorithm is described. Its prediction accuracy is evaluated and compared with the previous version of the model.

Keywords: model, gradient boosting, basic algorithm, LTWF, sampling, loss function, approximation.

Способность прогнозирования предстоящих изменений макропогоды на период от месяца и более, а также ее аналитика, чрезвычайно важна для регулирования процессов в жилищно-коммунальном и сельском хозяйствах, а также жизнеобеспечения населения, в таких аспектах как среднемесячные показатели температуры и осадков. Было разработано множество моделей [1]. Сегодня лидером в разработки моделей прогнозирования погоды является European Centre for medium-range weather forecasts (ECMWF) [2]. Недостатками моделей ECMWF являются: необ-

ходимость высокотехнологичного оборудования и большое количество физических фактов, которые необходимы для прогнозирования. В связи с данной проблемой нами предложена математическая модель долгосрочного прогнозирования погоды (ДПП) на основе регрессионного анализа и алгоритме градиентного усиления как наиболее выгодная. Объектом исследования данной статьи являются метеорологические данные города Москва с 01.01.2016 по 01.01.2022. На графике изображена зависимость температуры воздуха от времени (рис. 1) [3].

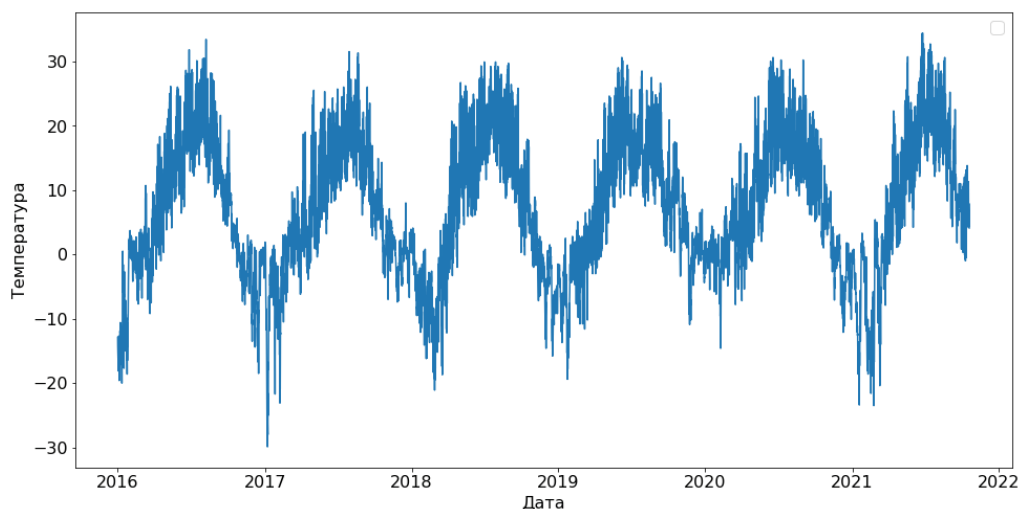


Рис. 1. Метеорологические данные города Москвы

Для получения модели ДПП г. Москва использовались пять признаков, описывающих погоду:

1. P_a – барическая тенденция;
2. P_o – атмосферное давление на уровне станции;
3. P – атмосферное давление, приведенное к среднему уровню моря;
4. U – относительная влажность (%) на высоте 2 м над поверхностью земли.
5. $dayofyear$ – день в году.

Фрагмент данных для обучения модели приведены на рис. 2.

	$dayofyear$	P_o	P	U	P_a
0	291	743.6	757.4	86.0	0.4
1	291	743.2	756.9	63.0	0.2
2	291	743.0	756.7	85.0	1.3
3	291	741.7	755.5	93.0	-0.9
4	291	742.6	756.3	84.0	-1.1
...	...	...	...	...	...
16933	1	763.2	779.3	73.0	0.5
16934	1	762.7	778.8	77.0	0.3
16935	1	762.4	778.5	78.0	0.2
16936	1	762.2	778.3	78.0	0.0
16937	1	762.2	778.2	78.0	0.4

Рис. 2. Данные для обучения модели

На основе анализа графика ДПП г. Москвы (рис. 1) примем допущение, что функция аппроксимации описывается законом косинуса, следовательно, необходимо сделать тригоно-

метрическое преобразование признака $dayofyear$ (рис. 3).

	$\cos_dayofyear$	P_o	P	U	P_a
0	0.262962	743.6	757.4	86.0	0.4
1	0.262962	743.2	756.9	63.0	0.2
2	0.262962	743.0	756.7	85.0	1.3
3	0.262962	741.7	755.5	93.0	-0.9
4	0.262962	742.6	756.3	84.0	-1.1
...	...	...	...	...	...
16933	1.000000	763.2	779.3	73.0	0.5
16934	1.000000	762.7	778.8	77.0	0.3
16935	1.000000	762.4	778.5	78.0	0.2
16936	1.000000	762.2	778.3	78.0	0.0
16937	1.000000	762.2	778.2	78.0	0.4

Рис. 3. Данные для обучения модели после преобразования признака

Для получения регрессионной модели ДПП г. Москва использовались данные на двух выборках: обучающая и тестовая. На обучающей выборке модель проводит анализ и аппроксимацию. Выборки: $data_train$ – обучающая выборка, $data_test$ – тестовая выборка приведены на рис. 4. Также для обучения модели необходимо еще раз разделить выборку: на признаки x и на переменную, которую мы хотим прогнозировать y , в нашем случае y – температура воздуха. Таким образом задаем четыре переменные: две для обучающей выборки (X_train , y_train) и для тестовой (X_test , y_test) (рис. 4).

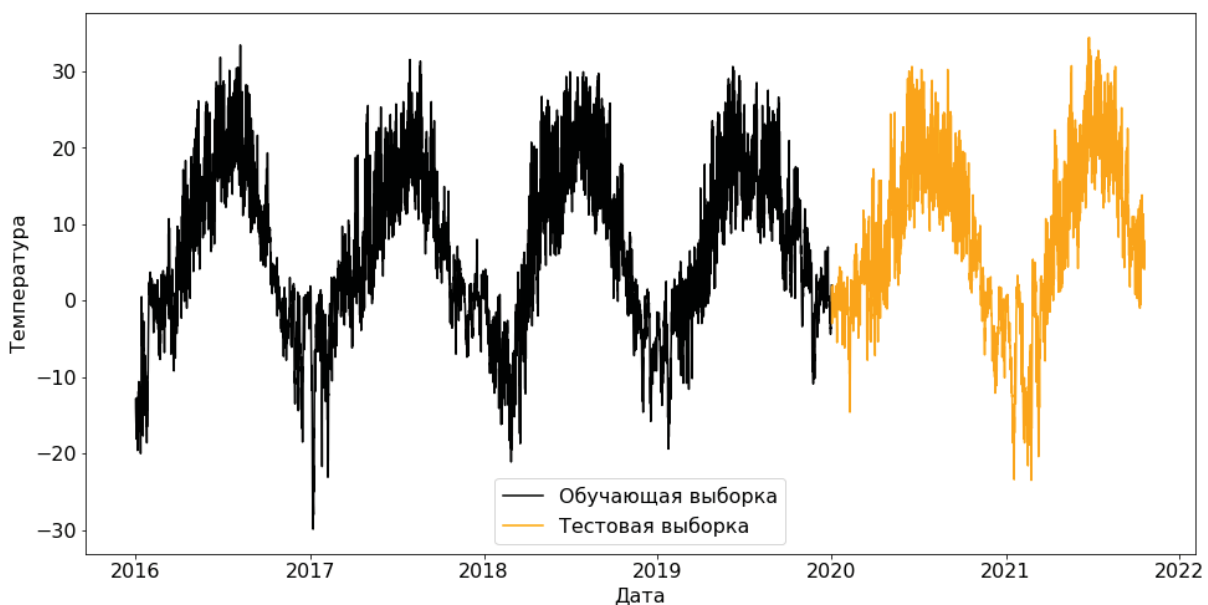


Рис. 4. График ДПП после разбиения выборки

Для улучшения показателей точности модели ДПП был выбран алгоритм обучения – градиентное усиление. Суть метода заключается в построении ансамбля «слабых» моделей – базовых алгоритмов (максимальное количество разбиений – 3). Модели строятся не параллельно, а последовательно. Данный метод ищет приближение $\hat{F}(x)$ в виде взвешенной суммы

функций $h_m(x)$ из некоторого класса «слабых» моделей:

$$\hat{F}(x) = \sum_{m=1}^M \gamma_m h_m + \text{const}, \quad (1)$$

где M – количество моделей, γ_m – коэффициент.

С помощью перекрестной проверки было найдено оптимальное количество базовых алгоритмов $M=3000$ (рис. 5).

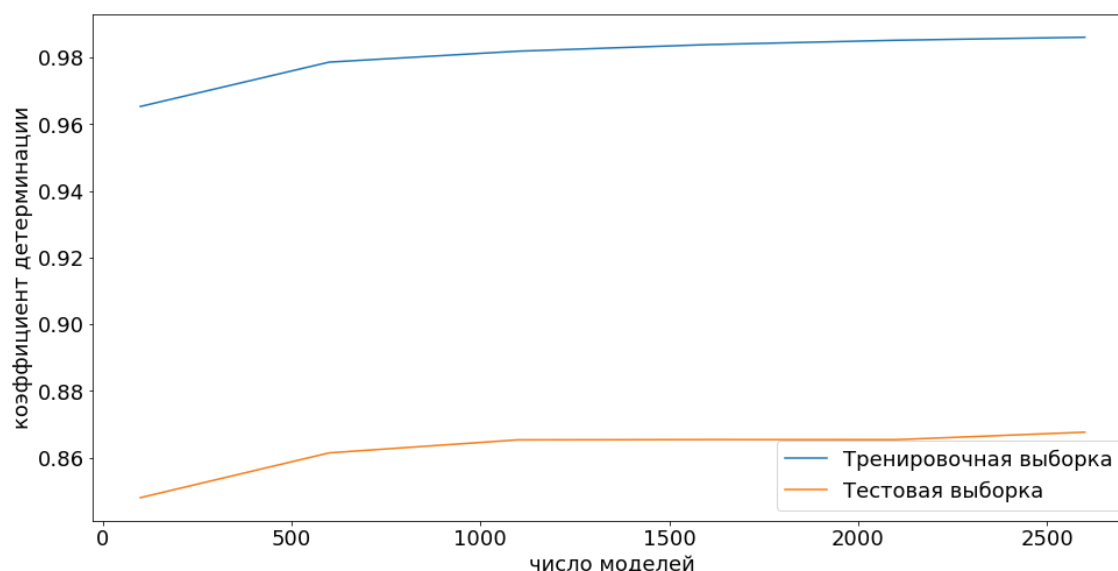


Рис. 5. Подбор оптимального количества моделей

Коэффициент γ_m аппроксимируется с помощью метода градиентного спуска – алгоритм численной оптимизации, позволяющий найти глобальный минимум функции потерь (2):

$$L_{MSE} = \frac{1}{n} \sum_{i=q}^n (y_i - F(x_i))^2 \quad (2)$$

На рис. 6 в колонке learn приведены минимальные значения функции потерь (2) для соответствующего базового алгоритма.

По данным таблицы видно, что с каждой

последующей обученной моделью средняя квадратическая ошибка уменьшается.

Для обучения моделей h_m используется алгоритм дерева принятия решения – непараметрический контролируемый метод обучения, предсказывающий значение целевой переменной, изучая простые правила принятия решений, выведенные из характеристик данных [5]. В качестве оценки разбиения набора данных используется функция (2) (рис. 7).

0:	learn: 10.0152527	total: 14.3ms	remaining: 14.3s
1:	learn: 9.5966521	total: 27ms	remaining: 13.5s
2:	learn: 9.1936823	total: 35.3ms	remaining: 11.7s
3:	learn: 8.8239176	total: 43.1ms	remaining: 10.7s
4:	learn: 8.4865855	total: 50.3ms	remaining: 10s
5:	learn: 8.1749861	total: 57.4ms	remaining: 9.51s
6:	learn: 7.8770109	total: 64ms	remaining: 9.07s
7:	learn: 7.6027946	total: 70.5ms	remaining: 8.74s
8:	learn: 7.3515481	total: 77.2ms	remaining: 8.51s
9:	learn: 7.1183556	total: 83.9ms	remaining: 8.3s
10:	learn: 6.9057941	total: 90.5ms	remaining: 8.13s
11:	learn: 6.7130499	total: 96.8ms	remaining: 7.97s
12:	learn: 6.5289489	total: 103ms	remaining: 7.85s
13:	learn: 6.3675443	total: 110ms	remaining: 7.71s
14:	learn: 6.2179513	total: 116ms	remaining: 7.63s
15:	learn: 6.0800039	total: 123ms	remaining: 7.57s

Рис. 6. Фрагмент таблицы значений функции потерь базовых алгоритмов

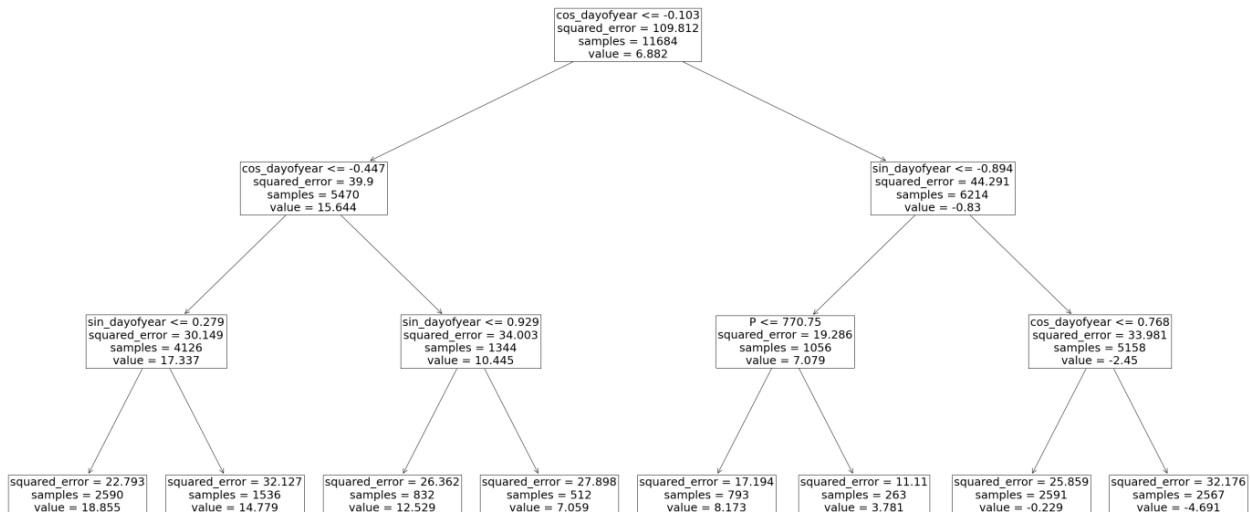


Рис. 7. Дерево принятия решений базового алгоритма

Результат обучения ансамбля моделей представлен на графике (рис. 8). На нем приведены тренировочная и тестовая выборки, а также спрогнозированные на них данные.

Метрикой качества был выбран коэффициент детерминации:

$$R^2 = 1 - \frac{\sum_{i=0}^{n-1} (y_n - z_n)^2}{\sum_{i=0}^{n-1} (y_n - \bar{y})^2}, \quad (3)$$

где z_n – спрогнозированные значения, y_n – наблюдаемые значения.

Значение коэффициента детерминации на тестовой выборке:

$$R^2 = 0.867.$$

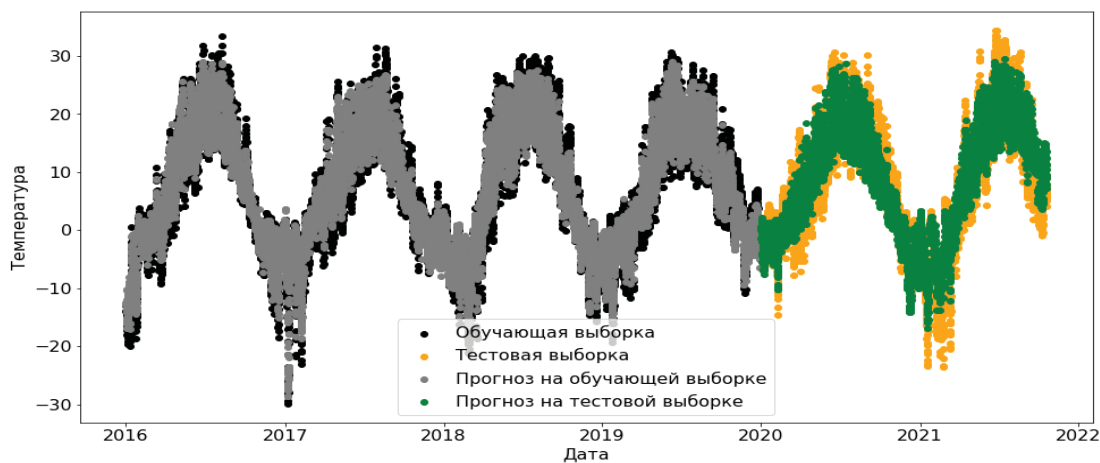


Рис. 8. Результат обучения модели (1)

В результате исследования, подбора алгоритма и его гиперпараметров была оптимизирована модель ДПП. Точность прогнозирования была улучшена по сравнению с прошлой версией на 4 %, что является хорошим результатом. По сравнению с другими моделями ее отличает небольшая стоимость и простота. Модель можно продолжать улучшать за счет добавления новых признаков, влияющих на погоду и увеличения объема данных.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Угрюмов, А. И. Долгосрочные метеорологические прогнозы / А. И. Угрюмов. – СПб. : ЗАО «НПП Система», 2006. – 84 с.
2. Технологии динамико-статистических долгосрочных метеорологических прогнозов: современное состояние и перспективы. URL: <http://seakc.meteoinfo.ru/training> (дата обращения: 13.02.22)
3. Погода в Москве (ВДНХ) – ПП5. URL: [https://tp5.ru/Погода_в_Москве_\(ВДНХ\)](https://tp5.ru/Погода_в_Москве_(ВДНХ)) (дата обращения: 15.02.22)
4. Коновалов, В. И. Идентификация и диагностика систем : учеб. пособие / В. И. Коновалов ; Томский политехнический университет. – Томск : Изд-во Томского политехнического университета, 2010. – 163 с.
5. Decision Trees - scikit-learn 1.1.3 documentation. URL: <https://scikit-learn.org/stable/modules/tree.html> (дата обращения 10.09.2022).
6. Gradient Boosting - mlcourse.ai. URL: https://mlcourse.ai/book/topic10/topic10_gradient_boosting.html (дата обращения 15.10.2022)