π-Economy. 2024, Том 17, № 1. С. 27–40.
π-Economy. 2024, Vol. 17, No. 1. Pp. 27–40.

Научная статья
УДК 338.27
DOI: https://doi.org/10.18721/JE.17102

МОДЕЛИ ПРОГНОЗИРОВАНИЯ С ПРИМЕНЕНИЕМ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА В ОТРАСЛЕВОЙ ЭКОНОМИКЕ

В.И. Мян ✉

Казахский национальный технический университет имени К.И. Сатпаева, Алма-Ата, Казахстан

✉ beromyan@gmail.com

Аннотация. Инструменты экономического регулирования оказывают значительное влияние в управлении социальным уровнем жизни в стране и в мировом сообществе. Одним из приоритетных направлений экономического менеджмента является отраслевая экономика. На сегодняшний день существует достаточно большое количество методов прогнозирования в экономике и, в частности, в отраслевой экономике. Методы варьируются согласно целям прогнозирования и входным данным. Наиболее популярными считаются статистически-эконометрические методы и методы искусственного интеллекта. Данная статья посвящена прогнозированию в сельскохозяйственной отрасли экономики с применением методов искусственного интеллекта. Рассматриваются несколько подходов, состоящих из прогнозного блока и комбинации эмпирического и прогнозного блоков. В качестве эмпирического блока используется система имитационного моделирования продукционного процесса сельскохозяйственного посева AGROTOOL. В прогнозном блоке используются три метода машинного обучения: Алгоритм случайного леса (Random Forest algorithm), Гребневая регрессия (Ridge Regression method) и Лассо-регрессия (Lasso Regression method). Данные методы машинного обучения были выбраны в связи с мультиколлинеарностью данных. Соответственно, алгоритм случайного леса рассматривался в двух вариациях: с использованием метода главных компонент (Principal Component Analysis) и без метода главных компонент. В данной работе помимо анализа прогнозных моделей также был проведен анализ по подбору ключевых параметров методов Random Forest PCA и Ridge Regression – n_components и α соответственно. По итогам численных экспериментов можно сделать вывод о том, что в зависимости от подаваемых на вход прогнозного блока данных, которыми могут быть значения природных факторов (влагозапас почвы, температура, солнечная радиация, среднесуточная удельная влажность воздуха, удельная влажность насыщения листа и др.), либо результаты формульных вычислений, наиболее эффективными методами машинного обучения являются Random Forest и Ridge Regression. Таким образом, применение метода Ridge Regression наиболее эффективно с данными, пред обработанными AGROTOOL, тогда как с помощью метода Random Forest маленькая доля погрешности была получена при прогнозировании без применения AGROTOOL и с объединенными данными.

Ключевые слова: прогнозирование в экономике, искусственный интеллект, Агротул, случайный лес, метод главных компонент, ридж-регрессия, лассо-регрессия

Для цитирования: Мян В.И. (2024) Модели прогнозирования с применением искусственного интеллекта в отраслевой экономике. π-Economy, 17 (1), 27–40. DOI: https://doi.org/10.18721/JE.17102

Research article
DOI: https://doi.org/10.18721/JE.17102

FORECASTING MODELS USING ARTIFICIAL INTELLIGENCE IN SECTORAL ECONOMY

V.I. Myan ✉

Kazakh National Technical University named after K.I. Satpayev, Alma-Ata, Kazakhstan

✉ beromyan@gmail.com

Abstract. The instruments of economic regulation have a significant impact in managing the social standard of living in the country and in the global community. One of the priority areas of economic management is the sectoral economy. To date, there are quite a large number of forecasting methods in the economy and, in particular, in the sectoral economy. The methods vary according to the forecasting objectives and input data. Statistical econometric methods and artificial intelligence methods are considered the most popular. This article is devoted to forecasting in the agricultural sector of the economy using artificial intelligence methods. Several approaches are considered, consisting of a predictive block and a combination of empirical and predictive blocks. The system of crop simulation AGROTOOL is used as an empirical block. The predictive block uses three machine learning methods: the Random Forest algorithm, Ridge Regression method, and Lasso Regression method. These machine learning methods were chosen due to the multicollinearity of the data. Thus, the random forest algorithm was considered in two variations: with and without the Principal Component Analysis (PCA) method. In this paper, in addition to analyzing predictive models, the author also analyzed the selection of key parameters of the random forest PCA and ridge regression: n_components and α methods, respectively. Based on the results of numerical experiments, it can be concluded that depending on the input of the forecast block of data, which may be the values of natural factors (soil moisture content, temperature, solar radiation, average daily specific humidity, specific humidity of leaf saturation, etc.), or the results of formula calculations, the most effective methods of machine learning are random forest and ridge regression. The use of the ridge regression method is most effective with the data preprocessed by AGROTOOL. At the same time, the random forest method produced a small fractional error when forecasting without AGROTOOL and with the combined data.

Keywords: forecasting in economics, artificial intelligence, AGROTOOL, random forest, principal component method, ridge regression, lasso regression

Citation: Myan V.I. (2024) Forecasting models using artificial intelligence in sectoral economy. π-Economy, 17 (1), 27–40. DOI: https://doi.org/10.18721/JE.17102

Введение

Экономическое регулирование является необходимым аспектом в области управления социальным уровнем жизни в стране и в мире. Одним из ключевых направлений экономического менеджмента является отраслевая экономика. Важно подчеркнуть значимость влияния каждой отрасли на социально-экономическое развитие в целом.

Актуальность исследования

Поэтому применение современных методов искусственного интеллекта для прогнозирования экономического развития отрасли является чрезвычайно актуальной задачей. Учитывая особую значимость сельскохозяйственной отрасли, как одну из первостепенных, напрямую влияющей на формирование цен на продукты потребления, в данной работе анализируется эффективность прогнозных моделей машинного обучения в отраслевой экономике на примере прогнозирования урожайности пшеницы. Согласно [1] прогнозирование цены на зерно сильно коррелирует с величиной урожайности.

Литературный обзор

Методы прогнозирования урожайности весьма разнообразны. Часто урожайность прогнозируется на основе исторических и временных данных с использованием статистических методов [2].

Иногда используются данные, полученные со спутников [3]. На сегодняшний день перспективным считается использование искусственного интеллекта в качестве инструмента прогнозирования урожайности [4]. Однако, не все методы искусственного интеллекта можно считать эффективными для решения большинства проблем, связанных с урожайностью пшеницы. Это объясняется тем фактом, что на качество прогноза сильно влияет правильный подбор данных и методов.

В своем исследовании авторы [5] продемонстрировали, что процесс прогнозирования пшеницы более эффективен при использовании данных из нескольких источников. Однако уровень точности зависит от сорта пшеницы, местоположения и временных параметров процесса обучения модели.

Наиболее популярным типом данных, используемым для прогнозирования урожайности, являются климатические данные. Например, авторы [6] в качестве входных данных для экспериментальной части загружают метеорологические данные. Популярность метеорологических данных обусловлена частотой их появления, объемом данных и доступностью.

Рассматривая методы искусственного интеллекта, такие как нейронные сети, следует отметить, что этот тип методов, как правило требует достаточно большой объем данных подаваемых на вход [7]. Однако, в данном исследовании, в силу того, что процесс созревания пшеничного зерна достаточно продолжителен по времени [8], нет большого объема данных.

Поэтому для прогнозирования урожайности было решено использовать методы машинного обучения, которые позволяют работать с небольшим объемом данных и делать достаточно точные прогнозы [9]. Согласно [10], основными методами машинного обучения являются регрессия, классификация и кластеризация, которые могут быть выбраны в зависимости от области исследования и поставленной задачи.

Поскольку желаемым результатом исследования является конкретное значение спрогнозированной урожайности, выбор пал на регрессию (а не на кластеризацию или классификацию). По мнению авторов [11], наиболее популярными и эффективными методами машинного обучения с использованием регрессии являются Neural Network Regression [12], ElasticNet [13], Random Forest Regression [14], LASSO [15] и Ridge Regression [16].

Прогнозирование урожайности пшеницы за счет использования методов машинного обучения является не только актуальным направлением, но, прежде всего, помогает улучшить качество прогноза будущего урожая. Помимо прочего, сравнение нескольких методов машинного обучения показало, что правильно подобранный метод и параметры метода, относительно ряда данных, положительно влияют на эффективность прогнозирования, как в комбинации с другими методами, так и по отдельности.

Цель исследования

Цель данного исследования заключается в разработке моделей прогнозирования с применением искусственного интеллекта для оценки урожайности пшеницы.

Объектом исследования являются методы машинного обучения, используемые для прогнозирования на примере оценки урожайности пшеницы.

Ключевые вопросы данного исследования следующие:

– Какой из рассмотренных методов машинного обучения наиболее эффективен?

– Является ли гибридная модель эффективнее отдельно взятых методов?

– Какие значения параметров методов машинного обучения являются оптимальными для обеспечения хорошего прогноза урожайности пшеницы?

По мнению автора, новизна данного исследования заключается в реализации методов машинного обучения с данными, которые предварительно обрабатываются прикладными математическими формулами, представленными в виде AGROTOOL. Результаты исследования автор в дальнейшем планирует использовать для прогнозирования цен на пшеничное зерно.

Методы и материалы

Данная работа предполагает два подхода к решению поставленной задачи.

Первый подход основан на применении следующих методов машинного обучения:

– Алгоритм случайного леса (Random Forest algorithm),
– Гребневая регрессия (Ridge Regression method),
– Лассо-регрессия (Lasso-Regression method).

Каждый алгоритм применяется отдельно и называется прогнозным блоком.

Второй подход подразумевает комбинацию прогнозного блока и эмпирического блока, где инструментом выступает программное обеспечение AGROTOOL.

Прогнозный блок

В прогнозном блоке рассматривается три метода машинного обучения: Алгоритм случайного леса (Random Forest algorithm), Гребневая регрессия (Ridge Regression method) и Лассо-регрессия (Lasso-Regression method). Наряду с алгоритмом случайного леса также применена его модификация с предобработкой метода главных компонент (англ. principal component analysis, PCA).

Алгоритм машинного обучения Случайный лес (англ. Random Forest) впервые был представлен в статье Лео Бреймана в 2001 году [17]. Позже, исследования были продолжены Адель Катлер в работе [18], где Random Forest определен, как ансамбль из деревьев решений, в котором каждое такое дерево решений зависит от набора случайных величин. Следовательно, Случайный лес обладает низкой вероятностью переобучения модели, даже в случае превышения количества признаков над количеством наблюдений, что положительно сказывается на точности прогноза [19].

Используемые природные данные (влагозапас почвы, температура, солнечная радиация и др.) сильно коррелированы между собой. Учитывая эту особенность, важно выделить основные компоненты без больших потерь данных. Одним из способов такой предобработки данных является Метод главных компонент (англ. principal component analysis, PCA), который был введен английским математиком Карлом Пирсоном в 1901 году. Идея метода состоит в ортогональном преобразовании коррелированных переменных в набор линейно некоррелированных переменных, то есть главных компонентов. Главный компонент представляется в виде линейной комбинации взвешенных исходных переменных. Таким образом, каждый последующий компонент имеет максимально возможную дисперсию, не учитываемую предыдущими главными компонентами, при условии ортогональности, что и приводит к нулевой корреляции главных компонент между собой [20].

Гребневая регрессия (англ. Ridge Regression) впервые была представлена Hoerl и Kennard в 1970 году [21]. Метод гребневой регрессии является одним из методов линейной регрессии, применяемый в случае мультиколлинеарности регрессионной модели. Данный метод позволяет регуляризировать параметры, тем самым, способствуя уменьшению дисперсии прогноза, и, как следствие, увеличению точности прогнозной модели [22].

Лассо- регрессия (англ. LASSO, Least Absolute Shrinkage and Selection Operator – Regression) действует по аналогичному принципу с гребневой регрессией, за исключением того, что значения коэффициентов Лассо-регрессии могут быть максимально уменьшены или приравнены нулю, тем самым признаки с нулевыми значениями могут просто исключаться из модели [23].

Эмпирический блок

Эмпирический блок реализован за счет программного обеспечения AGROTOOL.

Программный комплекс AGROTOOL создан Агрофизическим научно-исследовательским институтом (АФИ). По определению, взятому из документации программы, AGROTOOL является системой имитационного моделирования продукционного процесса сельскохозяйственного посева [24]. Данная модель разработана на основе эмпирических формул и нелинейных дифференциальных уравнений в частных производных, и, как следует из Информационное обеспечение модели [25], формально может быть записана в виде системы конечно-разностных уравнений:

$$x(k+1) = f\left(x(k),\ a,\ w(k),\ u(k)\right),$$
$$x(0) = x_0,\ k = 0, 1, \ldots, T,$$

где k – номер шага (номер суток), $x(k)$, $x(k + 1)$ – векторы состояния модели на двух соседних шагах, a – вектор параметров модели, $w(k)$ – вектор неконтролируемых внешних воздействий (погода), $u(k)$ – вектор управляющих воздействий (агротехника), x_0 – начальное условие, а T – время окончания процесса моделирования, совпадающее с днём уборки урожая.

На вход программы AGROTOOL подаются значения начальных параметров, которые описывают состояние природных факторов, таких как среднесуточная удельная влажность воздуха, максимальная и минимальная температуры воздуха, коэффициент ослабления радиации, среднесуточная скорость ветра и другие.

AGROTOOL выдает результат урожайности, в виде переменной Yield, что происходит на этапе окончания процесса моделирования. Однако, помимо результата урожайности, AGROTOOL сохраняет ежедневные результаты процесса моделирования таких факторов, например, как биомасса корня, листовой индекс, суммарные температура воздуха и радиация.

Следует отметить, что параметры системы AGROTOOL изменяются с шагом в 1 день [24], что позволяет отслеживать развитие растения в динамике, улучшая качество прогноза даже в случае различных природных катаклизмов.

Численные эксперименты

В данной работе было проведено 12 численных экспериментов, которые были поделены на 3 группы согласно данным, подаваемым на вход:

– Прогнозирование с помощью методов машинного обучения;

– Прогнозирование с помощью комбинации методов машинного обучения и эмпирического блока;

– Прогнозирование на основе объединенных данных.

Прогнозирование с помощью методов машинного обучения. В данной группе экспериментов на вход подавались значения природных факторов, в количестве 47 переменных (табл. 1), которые генерировались внутри интерфейса программы AGROTOOL [25].

Таблица 1. Природные факторы для прогнозирования с помощью первого подхода: применение методов машинного обучения

Table 1. Natural factors for forecasting using the first approach: application of machine learning methods

Description (English) – Описание		Имя поля
Biomass of the generative organs	Биомасса генеративного органа	Ebiom#1
Daily average air humidity	Среднесуточная удельная влажность воздуха	Qavg
Specific saturation humidity of the leaf	Удельная влажность насыщения листа	Qleaf
Linearization coefficient in the Magnus formula	Коэффициент линеаризации в формуле Магнуса	BTair
Vapor pressure deficit	Дефицит давления пара	Def
Air heat conductivity (above soil)	Проводимость приземного слоя воздуха по температуре	DsoilT
Air moisture conductivity (above soil)	Проводимость приземного слоя воздуха по влажности	DsoilQ
Heat conductivity of the air layer at soil surface	Проводимость припочвенного слоя воздуха по теплу	ResTemp
Long-wave radiation balance	Баланс длинноволновой радиации	Rnl
Day length	Длина дня	DayLength
The ratio of total precipitation to total transpiration (moisture ratio)	Отношение суммарных осадков к суммарной транспирации (коэффициент увлажнения)	SPR_STR
Air temperature, min	Минимальная температура воздуха	Tmin

Окончание таблицы 1

Description (English) – Описание		Имя поля
Air temperature, max	Максимальная температура воздуха	Tmax
Relative humidity	Относительная влажность воздуха	Q
Attenuation coefficient (of solar radiation)	Коэффициент ослабления радиации	Kex
Daily average wind speed	Среднесуточная скорость ветра	WindSpeed
Crop absorbed Shortwave Radiation	Поглощенная посевом КВР	RshPlant
Crop projectivity	Коэффициент проективного покрытия	crop
Nitrogen uptake by roots	Поглощение азота корнями	NRUpt
Development Stage Number	Текущая фаза	Ifase
Leaf water potential	Водный потенциал листа	PLeaf_list
Cumulative water storage (of meter thick layer) of soil	Суммарный влагозапас метрового слоя почвы	WS100
Infiltration of water beyond a meter thick layer of soil	Инфильтрация воды за пределы метрового слоя	Infiltration
Ammonium concentration in soil (total)	Концентрация аммония в почве (всего)	N_am_Total
Concentration of nitrogen in soil (total)	Концентрация азота в почве (всего)	N_nit_Total
Concentration of minerals in soil (total)	Концентрация минералов в почве (всего)	N_min_Total
Concentration of humidity in soil (total)	Концентрация влажности в почве (всего)	C_hum_ Total
Soil absorbed Shortwave Radiation	Поглощенная почвой КВР	RshSoil
Soil temperature	Температура почвы	TSoil_0
Volumetric soil water content (in the layer)	Объемная влажность слоя	WW_iter_1
The water flow through the lower boundary of the layer	Поток воды через нижнюю границу слоя	DownFlow_ iter_1
Soil water potential	Водный потенциал слоя	WPot_iter_1
Soil temperature (layers)	Температура почвы (слои)	TSoil_iter_1
Water Storage (1m)	Влагозапас слоя	WS_iter_1
Concentration of humidity in the layer	Концентрация влажности в слое	C_hum_ iter_1
Microorganisms' biomass in the layer	Биомасса микроорганизмов в слое	MBiom_ iter_1
Ammonium concentration in the layer	Концентрация аммония в слое	N_am_iter_1
Concentration of nitrogen in the layer	Концентрация азота в слое	N_nit_iter_1
Volumetric soil water content (in the layer)	Объемная влажность слоя	WW_iter_2
The water flow through the lower boundary of the layer	Поток воды через нижнюю границу слоя	DownFlow_ iter_2
Soil water potential	Водный потенциал слоя	WPot_iter_2
Soil temperature (layers)	Температура почвы (слои)	TSoil_iter_2
Water Storage (1m)	Влагозапас слоя	WS_iter_2
Concentration of humidity in the layer	Концентрация влажности в слое	C_hum_ iter_2
Microorganisms' biomass in the layer	Биомасса микроорганизмов в слое	MBiom_ iter_2
Ammonium concentration in the layer	Концентрация аммония в слое	N_am_iter_2
Concentration of nitrogen in the layer	Концентрация азота в слое	N_nit_iter_2

Для сбора данных, генерируемых внутри AGROTOOL была использована версия программного обеспечения AGROTOOL V.4. Система программного обеспечения осуществляла работу в DOS-режиме. Схема модели построена в среде визуального программирования Rational Rose. Динамическая модель реализована на объектно-ориентированном языке Turbo Pascal в нотации

системы Delphi, что касается базы данных: стационарная база данных реализована на СУБД Access, а оперативная – в системе Excel [25].

Прогнозирование с помощью комбинации методов машинного обучения и эмпирического блока. В данной группе экспериментов на вход прогнозного блока, состоящего из методов машинного обучения, подавались данные, в количестве 23 переменных (табл. 2), содержащие результаты процесса имитационного моделирования программой AGROTOOL, то есть, данные, прошедшие предобработку программой AGROTOOL [25].

Таблица 2. Природные факторы для прогнозирования с помощью второго подхода: комбинация методов машинного обучения и эмпирического блока
Table 2. Natural factors for forecasting using the second approach: a combination of machine learning methods and an empirical block

Description (English) – Описание		Имя поля
Biomass of the generative organs	Биомасса генеративного органа	Ebiom#1
Leaf biomass	Биомасса листьев	Lbiom
Biomass of stems	Биомасса стеблей	Sbiom
Total aboveground biomass	Суммарная надземная биомасса	SumBiom
Root biomass	Биомасса корней	Rbiom
Physiological time	Физиологическое время	Ph_Time
Leaf index	Листовой индекс	LAI
The assimilates that accumulated during a single model step	Ассимилянты, накопленные за шаг модели	PrimAss
Average daily temperature	Среднесуточная температура воздуха	Tavg
Accumulated radiation	Суммарная радиация	SumRad
Accumulated precipitation	Суммарные осадки	SumPrec
Root Depth	Глубина проникновения корней	HRoot
Water reserves in the meter layer	Влагозапас в метровом слое	WCSoil
Transpiration	Транспирация растением	Eplant
Evaporation	Испарение из почвы	Esoil
The ratio of total precipitation to total transpiration	Отношение суммарных осадков к суммарной транспирации	SumPr/SumE
Biological Time	Биологическое время	BioTime
Storage Pool of Nitrogen	Пул запасного азота	SNPool
The proportion of primary assimilates directed to the root	Доля первичных ассимилянтов, идущих в корень	CRS
Total nitrogen level	Уровень суммарной концентрации азота	SumNLow
The total amount of nitrogen in the plant	Суммарное количество азота в посеве	SumNPlant
Water potential of the soil at the depth of seeding	Водный потенциал почвы на глубине заделки семян	WPotSeed
Soil temperature at the depth of seeding	Температура почвы на глубине заделки семян	TPotSeed

Прогнозирование на основе объединенных данных. Под объединенными данными мы подразумеваем, слияние двух баз данных:

1) данные, генерируемые внутри программы AGROTOOL;

2) данные, выраженные в виде результатов имитационного моделирования программой AGROTOOL, таким образом, общее количество переменных, полученных объединением данных из табл. 1 и табл. 2, составило 70, но, с учетом повторяющейся переменной Y – Ebiom вышло 69 переменных.

Следовательно, основной целью на выходе каждого эксперимента были предсказанные результаты биомассы колоса (Ebiom).

Описание выполненных численных экспериментов

Все эксперименты были реализованы по следующему алгоритму:

1. Загрузка основных библиотек: Pandas, Numpy, MatPlotlib и Scikit Learn;
2. Загрузка данных;
3. Распределение данных в качестве Y и X;
4. Разбиение данных на тестовые и тренинговые, где test_size = 0.5 и random_state = 0;
5. Стандартизация данных;
6. Реализация методов машинного обучения: Random Forest, Random Forest PCA, Ridge Regression, LASSO;
7. Оценка результатов обучения посредством MAE, MSE, RMSE и R^2.

Технические сведения

Для реализации методов прогнозного блока были использованы язык Python и блокнот Jupyter из Anaconda Navigator версии Anaconda 3 (64-bit) от 20 октября 2021 года. В свою очередь, блокнот Jupyter использовался посредством веб-браузера Google Chrome версии 100.0.4896.88.

Касательно технической базы, эксперименты проводились на ноутбуке изготовителя Lenovo с оперативной памятью 8,00 ГБ, с 64-разрядным процессором и операционной системой Windows 11.

Подбор параметров для методов Random Forest PCA и Ridge Regression

Надо заметить, что для методов Random Forest PCA и Ridge Regression некоторые ключевые параметры, а именно, n_components и α соответственно, подбирались опытным путем. При этом параметр α в LASSO был одинаков для всех экспериментов и равен 0.1.

Так, параметр n_components определяет количество ключевых компонентов, которые необходимо учитывать при реализации Random Forest PCA, в случае, если не задавать параметр вручную, будут учитываться все компоненты, подаваемые на вход. В свою очередь параметр α в методах Ridge Regression и LASSO отвечает за регуляризацию параметров, накладывая верхний порог на значения входных данных, α варьируется от 0 до бесконечности.

В связи с тем, что данные параметры являются опциональными, одна из важных задач данного исследования заключалась в определении значений параметров n_components и α, при которых будут получены наиболее эффективные результаты прогнозирования.

Таким образом, были проанализированы варианты прогнозирования урожайности пшеницы методом Random Forest PCA для 1-ой группы экспериментов с данными, исключающими комбинацию с эмпирическим блоком; для 2-ой группы экспериментов, состоящей из комбинации методов машинного обучения и эмпирического блока; и для 3-ей группы экспериментов, реализуемых на основе объединенных данных.

В экспериментах с параметрами были рассмотрены n_components равные 6, 9, 12 и 15.

В 1-ой группе экспериментов:

– при n_components равному 6 Mean Absolute Error составил 0.228, Root Mean Squared Error – 0.483, Mean Squared Error – 0.233 и R^2 – 0.767;

– при n_components равному 9 Mean Absolute Error составил 0.236, Root Mean Squared Error – 0.486, Mean Squared Error – 0.237 и R^2 – 0.764;

– при n_components равному 12 Mean Absolute Error составил 0.229, Root Mean Squared Error – 0.511, Mean Squared Error – 0.261 и R^2 – 0.739;

– при 15 n_components Mean Absolute Error составил 0.233, Root Mean Squared Error – 0.512, Mean Squared Error – 0.262 и R^2 – 0.738.

В результате можно однозначно сказать, что для метода Random Forest PCA в 1-ой группе экспериментов наиболее оптимальным количеством главных компонент является n_component равный 6.

Во 2-ой группе экспериментов результаты были следующими:

– при n_components равному 6: Mean Absolute Error составил 0.304, Root Mean Squared Error – 0.563, Mean Squared Error – 0.317 и R^2 – 0,683;

– при n_components равному 9: Mean Absolute Error составил 0.307, Root Mean Squared Error – 0.580, Mean Squared Error – 0.336 и R^2 – 0,664;

– при n_components равному 12: Mean Absolute Error составил 0.306, Root Mean Squared Error – 0.578, Mean Squared Error – 0.334 и R^2 – 0,666;

– при n_components равному 15: Mean Absolute Error составил 0.317, Root Mean Squared Error – 0.574, Mean Squared Error – 0.329 и R^2 – 0,671.

Исходя из полученных результатов можно сделать вывод, что оптимальным параметром для метода Random Forest PCA в экспериментах с комбинацией методов машинного обучения и эмпирического блока является n_components равный 6.

В 3-ей группе экспериментов был проведен аналогичный анализ по подбору наилучшего параметра n_components. Результаты анализа отображены ниже:

– при n_components равному 6: Mean Absolute Error равен 0.421, Root Mean Squared Error – 0.777, Mean Squared Error – 0,604 и R^2 – 0.396;

– при n_components равному 9: Mean Absolute Error равен 0.402, Root Mean Squared Error – 0.737, Mean Squared Error – 0,544 и R^2 – 0.456;

– при n_components равному 12: Mean Absolute Error равен 0.392, Root Mean Squared Error – 0.722, Mean Squared Error – 0,522 и R^2 – 0.479;

– при n_components равному 15: Mean Absolute Error равен 0.395, Root Mean Squared Error – 0.734, Mean Squared Error – 0,538 и R^2 – 0.462.

Согласно вышеописанным результатам, применение параметра n_components равного 12 является оптимальным решением для эффективной реализации метода Random Forest PCA с объединенными данными.

Подобное сравнение параметров было проведено и для метода Ridge Regression с параметром α, из которого можно сделать вывод, что наиболее эффективным значением параметра α с данными из 1-ой группы экспериментов будет α равный 1.0:

– при α равному 0.5: Mean Absolute Error равен 0.190, Root Mean Squared Error – 0.256, Mean Squared Error – 0.066 и R^2 – 0.934;

– при α равному 1.0: Mean Absolute Error равен 0.181, Root Mean Squared Error – 0.246, Mean Squared Error – 0.060 и R^2 – 0.940;

– при α равному 6.0: Mean Absolute Error – 0.197, Root Mean Squared Error – 0.272, Mean Squared Error – 0.074 и R^2 – 0.926;

– при α равному 9.0: Mean Absolute Error – 0.213, Root Mean Squared Error – 0.289, Mean Squared Error – 0.084 и R^2 – 0.916.

Подбор α для 2-ой группы экспериментов, для прогнозирования посредством комбинации методов машинного обучения и эмпирического блока осуществлялся на основе следующих результатов:

– при α равному 0.5: Mean Absolute Error – 0.139, Root Mean Squared Error – 0.201, Mean Squared Error – 0.041 и R^2 – 0.959;

– при α равному 1.0: Mean Absolute Error – 0.150, Root Mean Squared Error – 0.219, Mean Squared Error – 0.048 и R^2 – 0.952;

– при α равному 6.0: Mean Absolute Error – 0.164, Root Mean Squared Error – 0.251, Mean Squared Error – 0.063 и R^2 – 0.937;

– при α равному 9.0: Mean Absolute Error – 0.173, Root Mean Squared Error – 0.267, Mean Squared Error – 0.071 и R^2 – 0.929.

Итог анализа показал, что наиболее эффективным является α со значением 0.5.

С 3-ей группой экспериментов был проведен анализ аналогичным способом на основе результатов, приведенных ниже:

– при α равному 0.5: Mean Absolute Error – 0.167, Root Mean Squared Error – 0.215, Mean Squared Error – 0.046 и R^2 – 0.954;

– при α равному 1.0: Mean Absolute Error – 0.166, Root Mean Squared Error – 0.220, Mean Squared Error – 0.049 и R^2 – 0.951;

– при α равному 6.0: Mean Absolute Error – 0.174, Root Mean Squared Error – 0.241, Mean Squared Error – 0.058 и R^2 – 0.942;

– при α равному 9.0: Mean Absolute Error – 0.179, Root Mean Squared Error – 0.249, Mean Squared Error – 0.062 и R^2 – 0.938.

Так, α равный 0.5 является наилучшим параметром для применения метода Ridge Regression для объединенных данных.

Результаты численных экспериментов

В данном разделе представлены результаты численных экспериментов по прогнозированию урожайности пшеницы.

Для численного сравнения полученных результатов использовался открытый исходный код документа Scikit-learn Regression metrics [26], с помощью которого были вычислены средняя абсолютная ошибка (MAE), среднеквадратическая ошибка (MSE), среднеквадратическое отклонение (RMSE) и коэффициент детерминации R^2.

В связи с проведением нескольких групп экспериментов результаты по оценке качества численных экспериментов также были ранжированы соответственно группам.

Таким образом при вычислении средней абсолютной ошибки (MAE) выяснилось, что при прогнозировании с помощью:

– 1-го подхода применение методов машинного обучения с использованием метода Random Forest показала долю погрешности равную 0.095, Random Forest PCA – 0.228, погрешность при прогнозировании Ridge Regression равна 0.181 и LASSO – 0.208. 2-го подхода комбинация методов машинного обучения и эмпирического блока при реализации метода Random Forest была равна 0.104, Random Forest PCA – 0.317, Ridge Regression – 0.139 и LASSO – 0.178.

– доля MAE в 3-ей группе экспериментов прогнозирование на основе объединенных данных с применением метода Random Forest составила 0.092, Random Forest PCA – 0.392, Ridge Regression - 0.167 и LASSO – 0.179.

Анализируя полученные результаты средней абсолютной ошибки (MAE), можно заметить, что в 1-ой, 2-ой и 3-ей группах экспериментов наименьшую погрешность показал метод Random Forest.

Однако, для данных, прошедших обработку программой AGROTOOL, можно утверждать, что более эффективным является метод Ridge Regression по результатам MSE, RMSE и R^2. Такая разница MAE с остальными показателями, вероятно, возникла по той причине, что метод оценки прогнозирования среднеквадратической ошибкой (MSE) позволяет исключать результаты измерений с большими погрешностями [27], помогая наиболее точно оценить качество прогноза при более детальном сравнении.

Так, среднеквадратическая ошибка (MSE) для групп экспериментов имеет следующие значения:

– прогнозирование с помощью 1-го подхода применение методов машинного обучения с использованием метода Random Forest показало долю MSE равную 0.028, Random Forest PCA – 0.233, Ridge Regression – 0.060 и LASSO – 0.115.

– прогнозирование с помощью 2-го подхода комбинация методов машинного обучения и эмпирического блока при реализации метода Random Forest выдало долю ошибки равную 0.047, Random Forest PCA – 0.317, Ridge Regression – 0.041 и LASSO – 0.085.

– доля среднеквадратической ошибки в 3-ей группе экспериментов прогнозирование на основе объединенных данных с применением метода Random Forest составила 0.027, Random Forest PCA – 0.522, Ridge Regression – 0.046 и LASSO – 0.088.

Данные результаты указывают на то, что для 1-ой и 3-ей группы экспериментов наилучшим методом прогнозирования является Random Forest, тогда как для 2-ой – Ridge Regression.

Следовательно, как и MSE, RMSE также остро реагирует на большие ошибки, исключая их из результатов измерений. При применении метода машинного обучения Random Forest среднеквадратическое отклонение равно 0.169, при Random Forest PCA – 0.483, Ridge Regression – 0.246 и при реализации LASSO – 0,339.

Прогнозирование вторым подходом комбинация методов машинного обучения и эмпирического блока выдало следующие значения RMSE: Random Forest – 0.218, Random Forest PCA – 0.563, Ridge Regression – 0.201 и LASSO – 0.292.

В случае прогнозирования на основе объединенных данных среднеквадратическое отклонение в Random Forest составило долю, равную 0.164, Random Forest PCA – 0.722, Ridge – Regression – 0.215 и LASSO – 0.296.

Таким образом видно, что для 1-ой группы экспериментов применение методов машинного обучения и для 3-ей группы прогнозирование на основе объединенных данных наименьшая погрешность у Random Forest, для 2-ой группы комбинация методов машинного обучения и эмпирического блока – Ridge Regression.

В свою очередь, рассматривая коэффициент детерминации, необходимо было учитывать, что значение R^2 увеличивается с добавлением новых переменных [28], а значит нижеприведенные результаты можно интерпретировать таким образом, что наилучшим методом прогнозирования для 1-ой группы применение методов машинного обучения и для 3-ей группы прогнозирование на основе объединенных данных будет Random Forest, когда для 2-ой группы комбинация методов машинного обучения и эмпирического блока, наилучшим оказался метод Ridge Regression.

Результаты коэффициента детерминации для 1-ой группы экспериментов составили следующие значения: Random Forest – 0.972, Random Forest PCA – 0.767, Ridge Regression – 0.940 и LASSO – 0.885. Во 2-ой группе экспериментов R^2 вышел со следующими результатами: Random Forest – 0.953, Random Forest PCA – 0.683, Ridge Regression – 0.959 и LASSO – 0.915. При прогнозировании, основанном на объединенных данных, результаты коэффициента детерминации были следующими: Random Forest – 0.973, Random Forest PCA – 0.479, Ridge Regression – 0.954 и LASSO – 0.912.

Заключение

Результаты численных экспериментов данного исследования оценивались посредством показателей погрешности прогноза – MAE, MSE и RMSE, и коэффициента детерминации – R^2:

1. Прогнозирование с помощью методов машинного обучения
   - Random Forest (MAE: 0.095, MSE: 0.028, RMSE: 0.169, R2: 0.972)
   - Random Forest PCA (MAE: 0.228, MSE: 0.233, RMSE: 0.483, R2: 0.767)
   - Ridge Regression (MAE: 0.181, MSE: 0.060, RMSE: 0.246, R2: 0.940)
   - LASSO (MAE: 0.208, MSE: 0.115, RMSE: 0.339, R2: 0.885)
2. Прогнозирование с помощью комбинации методов машинного обучения и эмпирического блока
   - Random Forest (MAE: 0.104, MSE: 0.047, RMSE: 0.218, R2: 0.953)
   - Random Forest PCA (MAE: 0.317, MSE: 0.317, RMSE: 0.563, R2: 0.683)
   - Ridge Regression (MAE: 0.139, MSE: 0.041, RMSE: 0.201, R2: 0.959)
   - LASSO (MAE: 0.178, MSE: 0.085, RMSE: 0.292, R2: 0.915)
3. Прогнозирование на основе объединенных данных

– Random Forest (MAE: 0.092, MSE: 0.027, RMSE: 0.164, R2: 0.973)
– Random Forest PCA (MAE: 0.392, MSE: 0.522, RMSE: 0.722, R2: 0.479)
– Ridge Regression (MAE: 0.167, MSE: 0.046, RMSE: 0.215, R2: 0.954)
– LASSO (MAE: 0.179, MSE: 0.088, RMSE: 0.296, R2: 0.912).

Согласно вышеуказанным результатам, выяснилось, что прогнозирование урожайности пшеницы методом Random Forest наиболее эффективно с данными без эмпирической основы (без Agrotool) и с объединенными данными. В то время, как прогнозирование, основанное на комбинации искусственного интеллекта и эмпирического блока (с обработкой Agrotool) имеет наименьшую погрешность с методом Ridge – Regression.

Таким образом, результаты исследования подтверждают эффективность и значимость применения искусственного интеллекта в области отраслевой экономики и, в частности, в сельскохозяйственной отрасли.

Направление дальнейших исследований

Учитывая результаты исследования по прогнозированию урожайности пшеницы, изложенного в данной статье, автор планирует продолжить работу в области прогнозирования цен на пшеничное зерно.

СПИСОК ИСТОЧНИКОВ

1. Амирова Э.Ф., Сафиуллин И.Н., Губанова Е.В., Ханнанов М.М. (2023) Особенности ценообразования на рынке зерна. *Аграрная наука*, 7, 163–167.

2. Воротников И.Л., Розанов А.В., Богатырев С.А., Ключиков А.В. (2022) Методологические особенности долгосрочного прогнозирования урожайности зерновых культур. *Аграрный научный журнал*, 11, 34–37.

3. Андреев К.П., Аникин Н.В., Бышов Н.В., Терентьев В.В., Шемякин А.В. (2019) Внедрение системы точного земледелия. *Вестник Рязанского государственного агротехнологического университета им. П.А. Костычева*, 2 (42), 74–80.

4. Medar R., Rajpurohit V.S., Shweta S. (2019) Crop yield prediction using machine learning techniques. In: *2019 IEEE 5th international conference for convergence in technology* (*I2CT*), 1–5.

5. Han J., Zhang Z., Cao J., Luo Y., Zhang L., Li Z., Zhang J. (2020) Prediction of winter wheat yield based on multi-source data and machine learning in China. *Remote Sensing*, 12 (2), 236.

6. Gümüşçü A., Tenekeci M.E., Bilgili A.V. (2020) Estimation of wheat planting date using machine learning algorithms based on available climate data. *Sustainable Computing: Informatics and Systems*, 28, 100308.

7. Алиева З. (2023). Искусственный интеллект и нейронные сети. *Scientific Collection InterConf*, 176, 193–198.

8. Жананов Б.Х., Эшонкулов Н.С., Вафоева М.Б. (2020) Вегетационный период формирования элементов урожая и показателей урожайности сортов яровой пшеницы. *Academy*, 12 (63), 28–32.

9. Васильченко А.М. (2023) Решение задач анализа данных на основе машинного обучения. *Universum: технические науки*, 9–1 (114), 50–54.

10. Palanivel K., Surianarayanan C. (2019). An approach for prediction of crop yield using machine learning and big data techniques. *International Journal of Computer Engineering and Technology*, 10 (3), 110–118.

11. Mohd T., Jamil N.S., Johari N., Abdullah L., Masrom S. (2020) An overview of real estate modelling techniques for house price prediction. In: *Charting a Sustainable Future of ASEAN in Business and Social Sciences: Proceedings of the 3rd International Conference on the Future of ASEAN* (*ICoFA*) *2019*, 1, 321–338.

12. Chen C.H., Lai J.P., Chang Y.M., Lai C.J., Pai P.F. (2023) A Study of Optimization in Deep Neural Networks for Regression. *Electronics*, 12 (14), 3071.

13. Han H., Dawson K.J. (2021) Applying elastic-net regression to identify the best models predicting changes in civic purpose during the emerging adulthood. *Journal of Adolescence*, 93, 20–27.

14. Tzenios N. (2020) Examining the Impact of EdTech Integration on Academic Performance Using Random Forest Regression. *Researchberg Review of Science and Technology*, 3 (1), 94–106.

15. Ranstam J., Cook J.A. (2018) LASSO regression. *Journal of British Surgery*, 105 (10), 1348–1348.

16. Carneiro T.C., Rocha P.A., Carvalho, P.C., Fernández-Ramírez L.M. (2022) Ridge regression ensemble of machine learning models applied to solar and wind forecasting in Brazil and Spain. *Applied Energy*, 314, 118936.

17. Breiman L. (2001) Random forests. *Machine learning*, 45, 5–32.

18. Cutler A., Cutler D.R., Stevens J.R. (2012) Random forests. *Ensemble machine learning: Methods and applications*, 157–175. DOI: http://dx.doi.org/10.1007/978-1-4419-9326-7_5

19. Чистяков С.П. (2013) Случайные леса: обзор. *Труды Карельского научного центра Российской академии наук*, (1), 117–136.

20. Shlens J. (2014) *A tutorial on principal component analysis*. DOI: https://doi.org/10.48550/arXiv.1404.1100

21. Kidwell J.S., Brown L.H. (1982) Ridge regression as a technique for analyzing models with multicollinearity. *Journal of Marriage and the Family*, 287–299.

22. Melkumova L.E., Shatskikh S.Y. (2017) Comparing Ridge and LASSO estimators for data analysis. *Procedia engineering*, 201, 746–755.

23. Hastie T., Tibshirani R., Wainwright M. (2015) *Statistical learning with sparsity: the lasso and generalizations*. NY: CRC press, 367. DOI: https://doi.org/10.1201/b18401

24. Poluektov R.A., Fintushal S.M., Oparina I.V., Shatskikh D.V., Terleev V.V., Zakharova E.T. (2002) Agrotool—a system for crop simulation. *Archives of Agronomy and Soil Science*, 48 (6), 609–635.

25. Баденко В.Л., Топаж А.Г., Якушев В.В., Миршель В., Нендель К. (2017). Имитационная модель агроэкосистемы как инструмент теоретических исследований. *Сельскохозяйственная биология*, 52 (3), 437–445.

26. Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O. Duchesnay É. (2011) Scikit-learn: Machine learning in Python. *The Journal of machine learning research*, 12, 2825–2830.

27. Avila J., Hauck T. (2017) *Scikit-learn cookbook: over 80 recipes for machine learning in Python with scikit-learn*. Packt Publishing Ltd.

28. Джеймс Г., Уиттон Д., Хасти Т., Тибширани Р. (2016) *Введение в статистическое обучение с примерами на языке R*. М.: ДМК Пресс, 456.

REFERENCES

1. Amirova E.F., Safiullin I.N., Gubanova E.V., Khannanov M.M. (2023) Osobennosti tsenoobrazovaniya na rynke zerna. *Agrarnaya nauka*, 7, 163–167.

2. Vorotnikov I.L., Rozanov A.V., Bogatyrev S.A., Klyuchikov A.V. (2022) Metodologicheskie osobennosti dolgosrochnogo prognozirovaniya urozhainosti zernovykh kul'tur. *Agrarnyi nauchnyi zhurnal*, 11, 34–37.

3. Andreev K.P., Anikin N.V., Byshov N.V., Terent'ev V.V., Shemyakin A.V. (2019) Vnedrenie sistemy tochnogo zemledeliya. *Vestnik Ryazanskogo gosudarstvennogo agrotekhnologicheskogo universiteta im. P.A. Kostycheva*, 2 (42), 74–80.

4. Medar R., Rajpurohit V.S., Shweta S. (2019) Crop yield prediction using machine learning techniques. In: *2019 IEEE 5th international conference for convergence in technology* (*I2CT*), 1–5.

5. Han J., Zhang Z., Cao J., Luo Y., Zhang L., Li Z., Zhang J. (2020) Prediction of winter wheat yield based on multi-source data and machine learning in China. *Remote Sensing*, 12 (2), 236.

6. Gümüşçü A., Tenekeci M.E., Bilgili A.V. (2020) Estimation of wheat planting date using machine learning algorithms based on available climate data. *Sustainable Computing: Informatics and Systems*, 28, 100308.

7. Alieva Z. (2023). Iskusstvennyi intellekt i neironnye seti. *Scientific Collection InterConf*, 176, 193–198.

8. Zhananov B.Kh., Eshonkulov N.S., Vafoeva M.B. (2020) Vegetatsionnyi period formirovaniya elementov urozhaya i pokazatelei urozhainosti sortov yarovoi pshenitsy. *Academy,* 12 (63), 28–32.

9. Vasil'chenko A.M. (2023) Reshenie zadach analiza dannykh na osnove mashinnogo obucheniya. *Universum: tekhnicheskie nauki*, 9–1 (114), 50–54.

10. Palanivel K., Surianarayanan C. (2019). An approach for prediction of crop yield using machine learning and big data techniques. *International Journal of Computer Engineering and Technology*, 10 (3), 110–118.

11. Mohd T., Jamil N.S., Johari N., Abdullah L., Masrom S. (2020) An overview of real estate modelling techniques for house price prediction. In: *Charting a Sustainable Future of ASEAN in Business and Social Sciences: Proceedings of the 3rd International Conference on the Future of ASEAN* (*ICoFA*) *2019*, 1, 321–338.

12. Chen C.H., Lai J.P., Chang Y.M., Lai C.J., Pai P.F. (2023) A Study of Optimization in Deep Neural Networks for Regression. *Electronics*, 12 (14), 3071.

13. Han H., Dawson K.J. (2021) Applying elastic-net regression to identify the best models predicting changes in civic purpose during the emerging adulthood. *Journal of Adolescence*, 93, 20–27.

14. Tzenios N. (2020) Examining the Impact of EdTech Integration on Academic Performance Using Random Forest Regression. *Researchberg Review of Science and Technology*, 3 (1), 94–106.

15. Ranstam J., Cook J.A. (2018) LASSO regression. *Journal of British Surgery*, 105 (10), 1348–1348.

16. Carneiro T.C., Rocha P.A., Carvalho P.C., Fernández-Ramírez L.M. (2022) Ridge regression ensemble of machine learning models applied to solar and wind forecasting in Brazil and Spain. *Applied Energy*, 314, 118936.

17. Breiman L. (2001) Random forests. *Machine learning*, 45, 5–32.

18. Cutler A., Cutler D.R., Stevens J.R. (2012) Random forests. *Ensemble machine learning: Methods and applications*, 157–175. DOI: http://dx.doi.org/10.1007/978-1-4419-9326-7_5

19. Chistyakov S.P. (2013) Sluchainye lesa: obzor. *Trudy Karel'skogo nauchnogo tsentra Rossiiskoi akademii nauk*, (1), 117–136.

20. Shlens J. (2014) *A tutorial on principal component analysis*. DOI: https://doi.org/10.48550/arXiv.1404.1100

21. Kidwell J.S., Brown L.H. (1982) Ridge regression as a technique for analyzing models with multicollinearity. *Journal of Marriage and the Family*, 287–299.

22. Melkumova L.E., Shatskikh S.Y. (2017) Comparing Ridge and LASSO estimators for data analysis. *Procedia engineering*, 201, 746–755.

23. Hastie T., Tibshirani R., Wainwright M. (2015) *Statistical learning with sparsity: the lasso and generalizations*. NY: CRC press, 367. DOI: https://doi.org/10.1201/b18401

24. Poluektov R.A., Fintushal S.M., Oparina I.V., Shatskikh D.V., Terleev V.V., Zakharova E.T. (2002) Agrotool—a system for crop simulation. *Archives of Agronomy and Soil Science*, 48 (6), 609–635.

25. Badenko V.L., Topazh A.G., Yakushev V.V., Mirshel' V., Nendel' K. (2017). Imitatsionnaya model' agroekosistemy kak instrument teoreticheskikh issledovanii. *Sel'skokhozyaistvennaya biologiya*, 52 (3), 437–445.

26. Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O. Duchesnay É. (2011) Scikit-learn: Machine learning in Python. *The Journal of machine learning research*, 12, 2825–2830.

27. Avila J., Hauck T. (2017) *Scikit-learn cookbook: over 80 recipes for machine learning in Python with scikit-learn*. Packt Publishing Ltd.

28. Dzheims G., Uitton D., Khasti T., Tibshirani R. (2016) *Vvedenie v statisticheskoe obuchenie s primerami na yazyke R*. M.: DMK Press, 456.

СВЕДЕНИЯ ОБ АВТОРЕ / INFORMATION ABOUT AUTHOR

МЯН Вероника Игоревна
E-mail: beromyan@gmail.com
Veronika I. MYAN
E-mail: beromyan@gmail.com

Поступила: 17.12.2023; Одобрена: 12.02.2024; Принята: 12.02.2024.
Submitted: 17.12.2023; Approved: 12.02.2024; Accepted: 12.02.2024.