Применение методов машинного обучения для прогнозирования индекса потребительских цен

Т.В. Азарнова, Н.Г. Аснина, А.И. Колосов, А.В. Лепендин

***Аннотация*— В статье проведен анализ эффективности применения методов машинного обучения для прогнозирования индекса потребительских цен, являющегося важнейшим макроэкономическим индикатором инфляции и ключевым статистическим показателем для правительств и центральных банков в процессе оценки стабильности цен. В качестве информационной базы исследования используются ежедневные данные по большому количеству товаров и услуг (более 12 млн ед.) с июня 2020 года по май 2022 года. Сложность использования данных для прогнозирования заключается в присутствии разрывов в наблюдениях. С помощью специальных методов обработки пропусков, ресемплирования и устранения выбросов, данные по отдельным товарам приводятся к виду непрерывных временных рядов с равноотстоящими наблюдениями. В качестве методов машинного обучения для прогнозирования используются методы PyAF, StatsForecastAutoARIMA, Prophet, рекуррентные нейронные сети LSTM. В статье также методами машинного обучения анализируется влияние на инфляцию таких факторов как: цена на нефть марки Brent, ставка центрального банка РФ, курс доллара к рублю, уровень инфляция за месяц по данным Росстата, ВВП, сальдо торгового баланса и исследуется возможность использования данных факторов в моделях прогнозирования.**

***Ключевые слова*— Индекс потребительских цен, прогнозирование индекса потребительских цен, анализ влияния экзогенных факторов на темп инфляции, методы машинного обучения.**

I. Введение

Индекс потребительских цен (ИПЦ) (индекс инфляции) – это индекс, измеряющий динамику цен на потребительские товары и услуги. ИПЦ отражает изменение среднего значения цен на выбранную группу товаров в текущем периоде по сравнению с предыдущим [1]. В ОСРД МВФ указаны рекомендации по ежемесячной публикации ИПЦ, не позднее одного-двух месяцев после сбора данных. ИПЦ является одним из самых значимых статистических показателей для принятия экономических решений, в первую очередь в сфере денежно-кредитной политики. ИПЦ часто фигурирует в законодательстве как показатель инфляции, в соответствие с которым проводятся корректировки различных платежей, от заработной платы до пособий по социальному страхованию, эта процедура носит название индексация. МОТ отмечает, что с течением времени сфера использования ИПЦ расширяется, сейчас ИПЦ является макроэкономическим индикатором инфляции и ключевым статистическим показателем для правительств и центральных банков в процессе оценки стабильности цен. В условиях глобализации торговли национальные правительства, центральные банки и международные организации придают большое значение качеству и точности национальных ИПЦ, а также их международной сопоставимости [2-6].

Публикация инфляционных данных, принятие последующих решений на их основе, выступления представителей финансовых структур, могут приводить не только к временным скачками цен на различные активы, но и вызывать более глобальные движения капитала, отдельные сферы экономики могут в относительно короткий срок потерять инвестиции заинтересованных лиц, а другие сферы получить крупные вложения. [6] В связи с этим большую актуальность приобретают эффективные методы прогнозирования ИПЦ. Прогнозированию ИПЦ посвящен целый ряд Российских и зарубежных исследований. В исследовании, представленном авторами Faiga Kharimah, Mustofa Usman, Widiarti and Faiz AM. Elfaki в работе [7], осуществляется поиск лучшей авторегрессионной модели для прогнозирования индекса потребительских цен (ИПЦ). Проводится оценка стационарности данных на основе визуализации ряда, построения функции автокорреляции ACF и проведения теста на единичный корень. Модель временного ряда определяется с помощью автокорреляционной ACF и частной автокорреляционной PACF функций. Выбор модели осуществляется с использованием критериев: среднеквадратическая ошибка (MSE), информационный критерий Акаике (AIC) и байесовский информационный

Статья получена 15 ноября 2024.

Т.В. Азарнова, д.т.н., проф., зав. каф. математических методов исследования операций, ФГБОУ ВО "Воронежский государственный университет" (e-mail: ivdas92@mail.ru)

Н.Г. Аснина, к.т.н, доцент, заф. каф. систем управления и информационных технологий в строительстве, ФГБОУ ВО "Воронежский государственный технический университет" (e-mail: andrey050569@yandex.ru)

А.И. Колосов, к.т.н, доцент, проректор по учебной работе ФГБОУ ВО "Воронежский государственный технический университет" (e-mail: andrey050569@yandex.ru)

А.В. Лепендин, аспирант каф. математических методов исследования операций, ФГБОУ ВО "Воронежский государственный университет" (e-mail: lepend1n8691@gmail.com)

критерий (BIC). Для построения модели использовались данные ИПЦ по городу Бандар-Лампунг с 2009 по 2013 год. На основе оценки по совокупности критериев лучшей моделью, была признана модель ARIMA (1,1,0). В работе Tien-Thinh Nguyen, Hong-Giang Nguyen, Jen-Yao Lee, Yu-Lin Wang, Chien-Shu Tsai [8] используются многомерная линейная регрессия (MLP), регрессия опорных векторов (SVR), модель распределенных лагов (ARDL) и многомерные адаптивные регрессионные сплайны (MARS) для прогнозирования ИПЦ США. В основу проведенного исследования легли данные по ИПЦ США с января 2017 года по февраль 2022 года, в качестве факторов, оказывающих влияние на ИПЦ рассматривались: цена на сырую нефть, мировая цена на золото и эффективная ставку федерального фонда. Среди используемых моделей MLR, SVR, ARDL и MARS наиболее высокую точность продемонстрировал алгоритм MARS. В статье Sibai Fadi, El-Moursy, Ali Sibai Ahmad [9] авторы прогнозируют индекс потребительских цен Саудовской Аравии с помощью шести методов машинного обучения (ML), используя инструмент анализа и добычи данных Orange 3 и основываясь на опубликованных исторических данных ИПЦ с января 2013 года по ноябрь 2020 года. Сравнивается производительность дерева решений (Tree), k-ближайших соседей (kNN), линейной регрессии (LR), нейронных сетей (NN), случайного леса (RF) и опорных векторов (SVM). Сопоставление результатов расчетов с прогнозом ИПЦ Международного валютного фонда (МВФ) на 2021–2024 годы и фактическими ИПЦ на 2021–2024 годы показывает, что модель нейронной сети многослойного персептрона превосходит по основным оценочным критериям другие модели машинного обучения, дает результаты близкие к фактическому ИПЦ и может использоваться для прогнозирования ИПЦ на срок до 3 лет. В работе Oren Barkan, Jonathan Benchimol, Itamar Caspi, Eliya Cohen, Allon Hammer, Noam Koenigstein [10] представлена иерархическая архитектура на основе рекуррентных нейронных сетей для прогнозирования дезагрегированных компонентов инфляции индекса потребительских цен. Авторы разработали новую модель иерархической рекуррентной нейронной сети (HRNN), которая использует информацию с более высоких уровней иерархии ИПЦ для улучшения прогнозов на более волатильных нижних уровнях. Оценки, основанные на большом наборе данных индекса CPI-U США, показывают, что модель HRNN значительно превосходит широкий спектр известных базовых уровней прогнозирования инфляции.

Данная статья посвящена разработке алгоритма прогнозирования индекса потребительских цен в формате «месяц к месяцу» на основе российских данных за период с июня 2020 года по май 2022 года по основным группам товаров, учитываемым в компонентах ИПЦ Росстата. В рамках исследования проанализированы факторы, оказывающие влияние на изменение цен товаров и услуг, разработан алгоритм обработки большого объема данных, включающий восстановление пропусков в данных, преобразование их в временные ряды и построение прогнозов ИПЦ.

II. Постановка задачи

В основе проведенного исследования лежат ежедневные данные по большому количеству товаров и услуг (более 12 млн ед.) с июня 2020 года по май 2022 года. В роли целевой переменной рассматривается синтетический индекс потребительских цен в формате месяц к месяцу, рассчитанный на основе отдельных позиций ИПЦ Росстата. Для расчета индекса потребительских цен на отдельные товары используется формула (1) [2,15,16]:

$$\text{ИПЦ т.} = \frac{P_t}{P_{t-1}} \quad (1)$$

где P_t – цена на товар/услугу на конец месяца t, P_{t-1} - цена на товар/услугу на конец месяца t-1, для расчета индекса цен на группу товаров - формула (2) (индекс Карли):

$$\text{ИПЦ гр. т.} = \frac{\sum_{i=1}^{N} \text{ИПЦ товара}_i}{N}, \quad (2)$$

где N – количество товаров в группе.

В исходном наборе данных не содержатся непосредственно временные ряды для реализации прогнозных моделей, ряды необходимо сформировать. Особенностью используемого набора данных является то, что период наблюдения, частота фиксации изменения цен, начало и окончание наблюдений отличаются для различных товаров, наблюдения могут периодически прерываться (наблюдаемые товары могут появляться и исчезать в продаже).

Гистограмма для длины непрерывных (без прерываний наблюдений) рядов данных по различным товарам (по оси x указывается количество наблюдений (изменений цены) в одном временном ряду (ед.), по оси y - количество товаров с соответствующим количеством наблюдений (млн. ед.)), показывает, что около 10 млн. товаров из 12-ти имеющихся в базе имеют длину не более 10 наблюдений, причем, 6,7 млн товаров – не более 2 наблюдений (рис. 1).

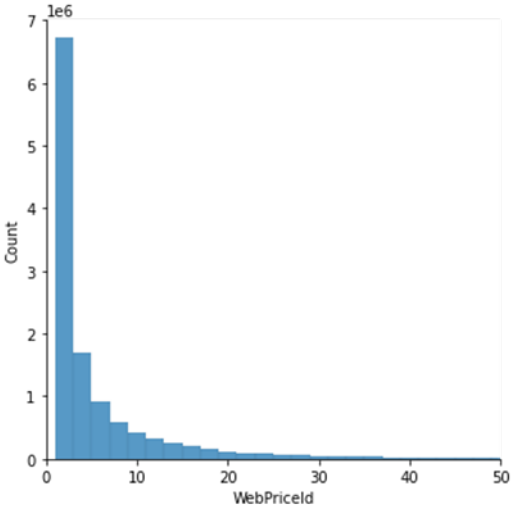


Рис. 1: Гистограмма длины непрерывных рядов наблюдений в исходном наборе данных

Анализ исходного набора данных показал, что достаточным для прогноза количеством наблюдений, обладают товары, информация о которых сосредоточена в первых 2-3 миллионах записей. На основании оценки

актуальности наблюдений для прогноза из данных записей был проведен отбор записей для построения прогнозных моделей.

Разрабатываемый в исследовании алгоритм должен формировать прогноз в режиме ежедневного обновления данных. Например, при поступлении данных за 1 сентября, модель должна прогнозировать ИПЦ на конец сентября. При поступлении данных за 2 сентября этот прогноз должен быть уточнен и т.д., при поступлении данных за 30 сентября цикл прогнозирования ИПЦ на конец сентября завершается. При поступлении данных за 1 октября начнется цикл прогнозирования ИПЦ на октябрь.

В представленном наборе данных содержится информация о фактах изменения цены на товары и услуги. Он состоит из ежедневных наблюдений за ценами с июня 2020 года по май 2022 года включительно. Факт изменения цены (либо появления или выхода из продажи) однозначно идентифицируется уникальным номером товара WebPriceId и датой наблюдения DateObserve, таблица описания полей набора данных приведена ниже.

Таблица 1: Описание полей набора данных

WebPriceId	Уникальный номер товара/услуги
DateObserve	Дата наблюдения
StockStatus	Статус товара/услуги на дату наблюдения (InStock – в продаже, OutOfStock – отсутствует в продаже)
CurrentPrice	Цена товара/услуги на дату наблюдения (Если StockStatus = OutOfStock – значение отсутствует)

На рисунке 2 приведен фрагмент набора данных.

	WebPriceId	DateObserve	StockStatus	CurrentPrice
0	1	2020-06-25 19:23:21.010	InStock	49.0
1	1	2020-09-08 07:39:23.593	OutOfStock	NaN
2	1	2020-09-08 15:39:01.663	InStock	49.0
3	1	2020-09-08 23:42:25.007	OutOfStock	NaN
4	1	2020-09-09 07:38:41.163	InStock	49.0

Рис. 2: Структура набора данных

III. Модели и методы прогнозирования индекса потребительских цен

Для преобразования исходных данных в ежедневные временные ряды реализуются следующие действия: в качестве начала временного ряда выбирается наблюдение с наиболее ранней датой DateObserve; в качестве окончания временного ряда выбирается наблюдение с наиболее поздней датой DateObserve и статусом товара StockStatus = OutOfStock (если наиболее последнее наблюдение имеет статус товара StockStatus = InStock, это значит, что товар на настоящий момент в продаже по последней известной цене); для удаления выбросов используется подход, при котором исключаются строки, в которых значение в столбце 'CurrentPrice' больше, чем в три раза выше или ниже среднего значения 'CurrentPrice' товара с тем же уникальным номером товара/услуги; проводится ресемплирование по заданной частоте времени, результат ресемплирования для одного товара представлен на рисунке 3:

Рис. 3: Ресемплирование данных

Для прогнозирования цен использовалось несколько моделей [11, 12, 13].

Прогноз цен 10 товаров с помощью PyAF показал среднюю ошибку около 4% на тестовой выборке (рис. 4):

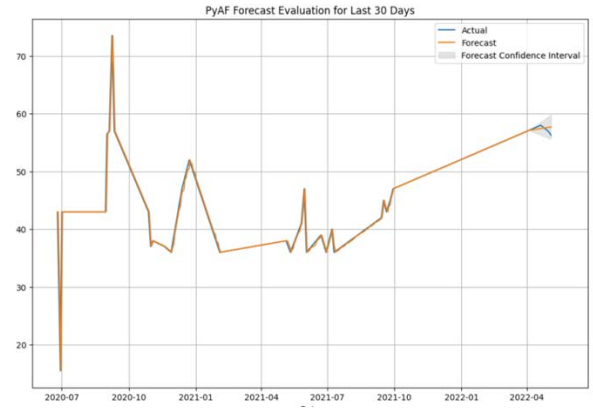


Рис. 4: Результаты прогнозирования с помощью PyAF

Прогноз цен 10 товаров с помощью StatsForecastAutoARIMA показал среднюю ошибку около 8% на тестовой выборке (рис.5).

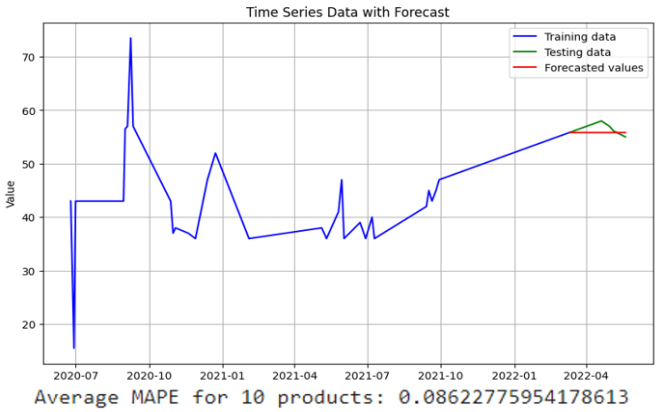
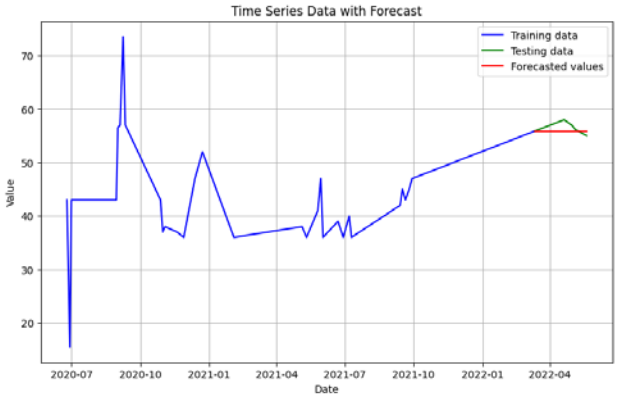


Рис. 5: Результаты прогнозирования с помощью StatsForecastAutoARIMA

Оптимальный прогноз цен 10 товаров с помощью StatsForecast был получен при использовании модели

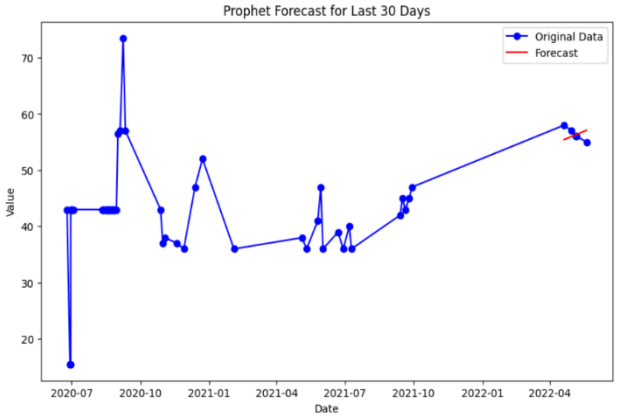
AutoCES, прогноз показал среднюю ошибку около 2% на тестовой выборке (рис. 6):




```
Average MAPE for 10 products: 0.020457098093304635
```

Рис. 6: Результаты прогнозирования с помощью StatsForecastAutoCES

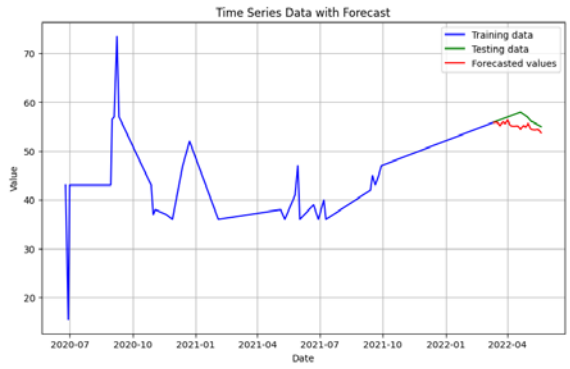
Самая низкая ошибка была получена при прогнозировании цен товаров с помощью рекуррентной нейронной сети LSTM (между 1 и 2%). Структура нейронной сети: LSTM слой с 64 нейронами, полносвязный слой с 64 нейронами, слой dropout с коэффициентом 0.05 и полносвязный слой с 1 нейроном (при изменении структуры, ее усложнении, зачастую увеличивалось время обучения и ошибка) (рис. 7):




```
Average MAPE for 10 products: 0.009003434829265036
```

Рис. 7: Результаты прогнозирования с помощью LSTM

Самый быстрый прогноз был получен с помощью Prophet (10 товаров за 7 секунд), обучение модели также не потребовало предварительного ресемплирования данных.




```
Average MAPE for 10 products: 0.04779732372632432
```

Рис. 8: Результаты прогнозирования с помощью Prophet

Таким образом, были проанализированы результаты прогнозов различных моделей и было принято решение использовать Prophet для прогнозирования инфляции по данным, содержащим наблюдения цен множества товаров. Prophet имеет приемлемую величину ошибки (менее 5%) и предоставляет высокую скорость прогноза без необходимости проведения ручного ресемплирования и сложной дополнительной настройки данных (рис. 9):

	PyAF	StatsForecast AutoARIMA	StatsForecast AutoETS	StatsForecast AutoCES
Требует ресемплирования	Да	Да	Да	Да
MAPE (10 товаров)	0.0418	0.0862	0.0204	0.0594
Время (10 товаров) с.	1320	433	66	36
MAPE (100 товаров)	0.0393	0.0781	0.0225	0.0573
Время (100 товаров) с.	8960	2820	713	421

	AutoARIMA + SARIMAX	LSTM простой	LSTM с доп. слоями	Prophet
Требует ресемплирования	Да	Да	Да	Нет
MAPE (10 товаров)	0.0197	0.02238	0.009	0.0478
Время (10 товаров) с.	106	70	123	7
MAPE (100 товаров)	0.0208	0.0148	0.0119	0.0369
Время (100 товаров) с.	1132	681	890	44

Рис. 9: Сравнение ошибки и времени формирования прогнозов различных моделей

Результаты расчета инфляции за апрель 2021 года и январь 2022 года на основании 10000 наблюдений представлены на рисунках 10 и 11:

```
print("В среднем инфляция за месяц составила:", mean(InflationPerMonthArr1),"%")
print('Время, потребовавшееся для прогнозирования цен на товары в цикле =' ,testtime1,"секунд" )
print('Инфляция рассчитывается за месяц, следующий за выбранная датой: ' ,prevmonth )

В среднем инфляция за месяц составила: 1.600487246715331 %
Время, потребовавшееся для прогнозирования цен на товары в цикле = 186.2921826839447 секунд
Инфляция рассчитывается за месяц, следующий за выбранная датой:  2021-03-31
```

Рис. 10: Результаты расчета инфляции за апрель 2021 года на основании 10000 наблюдений

```
print("В среднем инфляция за месяц составила:", mean(InflationPerMonthArr1),"%")
print('Время, потребовавшееся для прогнозирования цен на товары в цикле =' ,testtime1,"секунд" )
print('Инфляция рассчитывается за месяц, следующий за выбранная датой: ' ,prevmonth )

В среднем инфляция за месяц составила: 0.8304291368266918 %
Время, потребовавшееся для прогнозирования цен на товары в цикле = 166.10291695594788 секунд
Инфляция рассчитывается за месяц, следующий за выбранная датой:  2021-12-31
```

Рис. 11: Результаты расчета инфляции за январь 2022 года на основании 10000 наблюдений

При сравнении рассчитанных выше значений с официальными данными Росстата, можно заметить, что полученный в работе результат за апрель 2021 года, равный 101.6%, превышает 100.58% из таблицы Росстата. Если же сравнивать показатели ИПЦ за январь 2022 года, то результат 100.83% близок к данным Росстата — 100.67%.

При увеличении количества наблюдений и товаров для расчета инфляции до 50000, результаты за январь 2022 года приведены на рисунке 12.

```
print("В среднем инфляция за месяц составила:", mean(InflationPerMonthArr1),"%")
print('Время, потребовавшееся для прогнозирования цен на товары в цикле =' ,testtime1,"секунд" )
print('Инфляция рассчитывается за месяц, следующий за выбранная датой: ' ,prevmonth )
print('Количество товаров и услуг, участвующих в расчете', len(groups_by_ticker))

В среднем инфляция за месяц составила: 1.9960487530883722 %
Время, потребовавшееся для прогнозирования цен на товары в цикле = 606.4397473335266 секунд
Инфляция рассчитывается за месяц, следующий за выбранная датой:  2021-12-31
Количество товаров и услуг, участвующих в расчете 1087
```

Рис. 12: Результаты расчета ИПЦ за январь 2022 года на основании 50000 наблюдений

Рассмотрим результаты расчетов ИПЦ для корзины товаров. Путем ресемплирования данных 303 товаров, можно получить динамику общей стоимости (суммы) этих товаров (рис.13):

Рис. 13: Динамика суммы цен на товары с октября 2020 г. по июнь 2022 г.

По полученным данным проводится расчет ИПЦ корзины товаров с использованием многопроцессорной обработки. Прогноз модели строится на 90 дней вперед. В анализ попадают только те товары, по которым имеется данные на выбранный месяц. Время обучения модели сокращается (117 вместо более 160 секунд для 10000 наблюдений), на рисунке 14 приведены результаты расчётов май 2022 года.

```
ticker
4       3.021562
6       2.031455
7       1.236226
8       2.069500
16      2.510959
          ...
297    -1.275008
299     0.291996
300    -2.169875
301   -11.719765
302     0.007079
Name: percentage_change, Length: 90, dtype: float64
Инфляция на корзину товаров за май 2022 составила: 0.1815196196111876 %
Время на обучение моделей для 269 товаров: 122.40044522285461 секунд
Количество учтенных в расчете инфляции товаров/услуг: 90
```

Рис. 14: Прогнозируемая инфляция за май 2022 г.

Результат расчета инфляции за июнь 2022 года отражен на рисунке 15:

```
ticker
2       0.585689
3     -60.211798
4       2.838330
6       1.926782
7       0.069647
          ...
298    -4.337187
299     0.281754
300    -2.146454
301   -12.847392
302     0.006850
Name: percentage_change, Length: 156, dtype: float64
Инфляция на корзину товаров за июнь 2022 составила: -0.5038135588633834 %
Время на обучение моделей для 269 товаров: 122.40044522285461 секунд
Количество учтенных в расчете инфляции товаров/услуг: 156
```

Рис. 15: Прогнозируемая инфляция за июнь 2022 г.

Значение индекса потребительских цен в мае и июне 2022 года, полученное по моделям, показывает хорошее приближение к динамике ИПЦ (рис.16).

	январь	февраль	март	апрель	май	июнь	июль
2024	100,86	100,68					
2023	100,84	100,46	100,37	100,38	100,31	100,37	100,63
2022	100,99	101,17	107,61	101,56	100,12	99,65	99,61
2021	100,67	100,78	100,66	100,58	100,74	100,69	100,31
2020	100,40	100,33	100,55	100,83	100,27	100,22	100,35
2019	101,01	100,44	100,32	100,29	100,34	100,04	100,20
2018	100,31	100,21	100,29	100,38	100,38	100,49	100,27
2017	100,62	100,22	100,13	100,33	100,37	100,61	100,07

Рис. 16: ИПЦ по данным Росстата

С увеличением скорости работы моделей можно попробовать загрузить большее количество данных. Рассмотрим 100000 записей, содержащих информацию по динамике цен 1885 товаров. Предсказание строится на 90 дней вперед, в итоговых результатах учитываются товары, по которым имелись данные на даты, не более чем на 90 дней отстоящие от предсказываемых. Получили предсказание инфляции за июль 2022 года, которое расходится с результатами от Росстата (рис. 17):

```
ticker
2       0.601688
4       2.851992
7       0.071919
15     -2.630441
16      2.392736
          ...
1877    2.890831
1878   -8.259846
1881    2.214011
1883   -5.419174
1884   -0.090550
Name: percentage_change, Length: 1018, dtype: float64
Инфляция на корзину товаров за июль 2022 составила: 0.7780302335970976 %
Время на обучение моделей для 1796 товаров: 848.7924137115479 секунд
Количество учтенных в расчете инфляции товаров/услуг: 1019
```

Рис. 17. Прогнозируемая инфляция за июль 2022 г.

Далее были рассмотрены 100000 записей, начиная с 200000-ой, проведено ресемплирование, сформирована сумма всех товаров (рис. 18).

Рис. 18: Временной ряд суммы цен 1634 товаров

Прогноз инфляции за апрель 2022 года приведен на рисунке 19.

```
4837     0.705460
4872     1.575023
4873     2.701128
4907     2.528264
4910     0.322614
4968     1.887232
4969     1.890060
5068    -0.853332
5071    -0.553510
5103     1.901685
5105     2.977631
5109     1.501132
5187     3.731215
Name: percentage_change, dtype: float64
Инфляция на корзину товаров за июнь 2022 составила: 1.8856986121212237 %
Время на обучение моделей для 1587 товаров: 849.9971542358398 секунд
Количество учтенных в расчете инфляции товаров/услуг: 48
```

Рис. 19: Прогнозируемая инфляция за апрель 2022 г.

Несмотря на то, что в расчетах инфляции было учтено только 48 товаров из 1634, результаты расчета оказались близки к значениям из таблицы Росстата.

Возвращаясь к подходу, когда в первую очередь выбирается дата прогноза, а затем для всех товаров прогнозируется цена на две даты с разницей в месяц, можем наблюдать схожий прогноз (рис. 20):

```
print("В среднем инфляция за месяц составила:", mean(InflationPerMonthArr1),"%")
print('Время, потребовавшееся для прогнозирования цен на товары в цикле =' ,testtime1 )
print('Выбранная для расчета ИПЦ дата: ' ,prevmonth )

В среднем инфляция за месяц составила: 1.982973117737981 %
Время, потребовавшееся для прогнозирования цен на товары в цикле = 872.9904885292053
Выбранная для расчета ИПЦ дата:  2022-03-31
```

Рис. 20: Результаты расчета ИПЦ за апрель 2022 года на основании 100000 наблюдений

IV. Оценка влияния экономических факторов на ИПЦ

Для оценки влияния на ИПЦ таких факторов как: цена на нефть марки Brent, ставка центрального банка РФ, курс доллара к рублю, дата, инфляция за месяц по данным Росстата, ВВП, сальдо торгового баланса в работе используется несколько методов машинного обучения. Фрагмент информации по факторам представлен на рисунке 21.

	Oil	rate	Close	date	inflation	gdp	tradebalance
0	50.30	12.50	33.422295	2009-04-30	0.69	-11.0	4.69
1	64.98	12.00	31.866024	2009-05-31	0.57	-11.8	8.04
2	68.11	11.50	31.003659	2009-06-30	0.60	-10.7	8.90
3	70.08	11.00	31.496465	2009-07-31	0.63	-8.7	10.95
4	69.02	10.75	31.819557	2009-08-31	0.00	-9.6	11.99
...	...	...	...	...	...	...	...
171	87.29	12.00	95.389346	2023-08-31	0.28	5.2	5.49
172	95.86	13.00	96.330428	2023-09-30	0.87	5.6	11.00
173	86.82	15.00	97.176854	2023-10-31	0.83	5.1	15.29
174	81.72	15.00	90.486181	2023-11-30	1.11	4.5	9.43
175	77.69	16.00	90.962143	2023-12-31	0.73	4.4	8.68

176 rows × 7 columns

Рис. 21: Данные по факторам с 2009 г. по 2023 г.

Поскольку инфляция характеризует изменение цен за некоторый период, то было решено для цены на топливо и курса доллара к рублю рассчитать процентное изменение за период, в нашем случае — месяц (рис. 22).

	rate	date	inflation	gdp	tradebalance	Oil_pc	Usd_pc
171	12.0	2023-08-31	0.28	5.2	5.49	0.024290	0.054317
172	13.0	2023-09-30	0.87	5.6	11.00	0.098178	0.009866
173	15.0	2023-10-31	0.83	5.1	15.29	-0.094304	0.008787
174	15.0	2023-11-30	1.11	4.5	9.43	-0.058742	-0.068850
175	16.0	2023-12-31	0.73	4.4	8.68	-0.049315	0.005260

Рис. 22: Фрагмент преобразования данных

На рисунке 23 приведены результаты применения метода деревья решений.

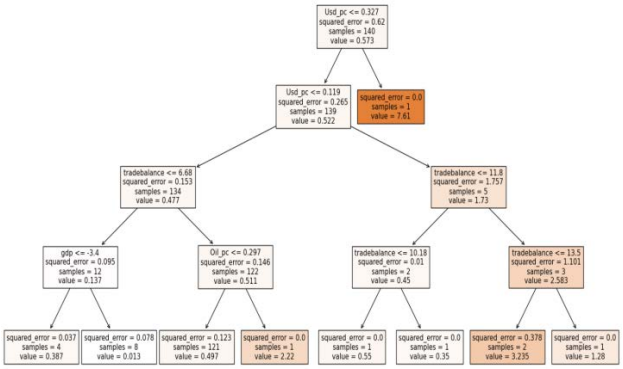


Рис. 23: Результаты применения метода деревья решений

Наблюдалась следующая закономерность: высокий коэффициент детерминации на обучающей выборке и низкий, близкий к нулю на тестовой выборке, метрики качества дерева решений представлены на рисунке 24:

```
# Calculate evaluation metrics
mae = mean_absolute_error(y_test, predictions2)
mse = mean_squared_error(y_test, predictions2)
r2 = r2_score(y_test, predictions2)

# Print evaluation metrics
print(f'Средняя абсолютная ошибка на тестовой выборке Деревом решений: {mae}')
print(f'Среднеквадратическая ошибка на тестовой выборке Деревом решений: {mse}')
print(f'Коэффициент детерминации на тестовой выборке Деревом решений: {r2}')

Средняя абсолютная ошибка на обучающей выборке Деревом решений: 0.25759031877213695
Среднеквадратическая ошибка на обучающей выборке Деревом решений: 0.11739297963400236
Коэффициент детерминации на обучающей выборке Деревом решений: 0.810543201706013
Средняя абсолютная ошибка на тестовой выборке Деревом решений: 0.27950413223140497
Среднеквадратическая ошибка на тестовой выборке Деревом решений: 0.1932710330090646
Коэффициент детерминации на тестовой выборке Деревом решений: -0.004964677381320781
```

Рис. 24: Метрики качества для дерева решений

Пример построение регрессионного дерева решений с помощью библиотеки xgboost показан на рисунке 25:

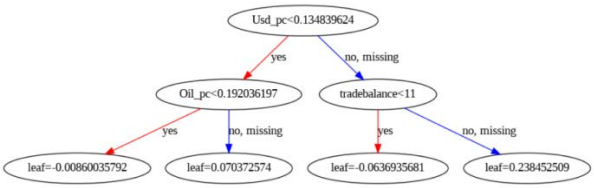


Рис. 25: Дерево решений xgboost

Метрики качества модели xgboost представлены на рисунке 26:

```
Средняя абсолютная ошибка на обучающей выборке Градиентный бустинг xgboost: 0.09220893103097166
Среднеквадратическая ошибка на обучающей выборке Градиентный бустинг xgboost: 0.013792691675005003
Коэффициент детерминации на обучающей выборке Градиентный бустинг xgboost: 0.9777404133300855
Средняя абсолютная ошибка на тестовой выборке Градиентный бустинг xgboost: 0.2466094509448324
Среднеквадратическая ошибка на тестовой выборке Градиентный бустинг xgboost: 0.16350291293470276
Коэффициент детерминации на тестовой выборке Градиентный бустинг xgboost: 0.14982266307531333
```

Рис. 26: Метрики качества для xgboost

Пример построение регрессионного дерева решений с помощью библиотеки sklearn показан на рисунке 27:

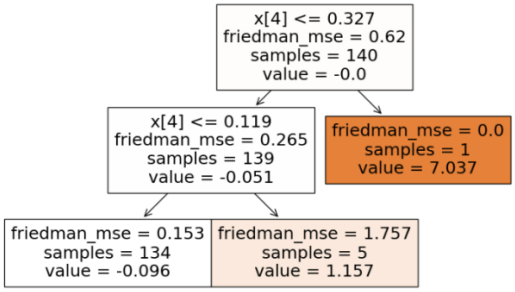


Рис. 27: Дерево решений sklearn

Метрики качества модели градиентного бустинга sklearn представлены на рисунке 28:

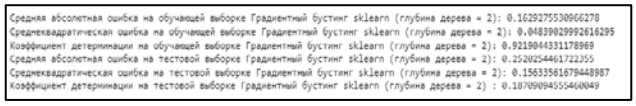


Рис. 28: Метрики качества для sklearn

Метрики качества модели случайного леса представлены на рисунке 29:

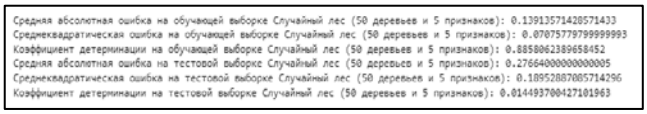


Рис. 29: Метрики качества для случайного леса

Несмотря на невысокие показатели коэффициента детерминации, было принято решение использовать модель SARIMAX для прогнозирования инфляции, которая сочетает в себе авторегрессию, интегрированность, скользящую среднюю, сезонность и учитывает экзогенные (внешние) факторы. В результате тщательного подбора параметров модели был получен результат на рисунке 30:

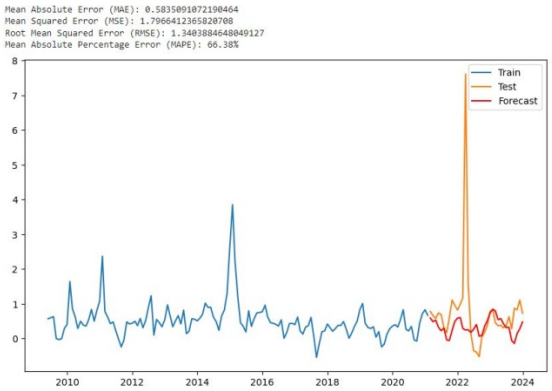


Рис. 30: Прогноз инфляции и метрики качества SARIMAX

Результаты прогнозирования SARIMAX можно учесть при прогнозировании временных рядов моделями, приведенными выше, так, AutoARIMA показала более низкую ошибку на временных рядах цен товаров и услуг, рисунок 31:

Рис. 31: Прогноз цены товара и метрики качества AutoARIMA и SARIMAX

Поведение инфляции в краткосрочном периоде может быть близко аппроксимировано с помощью простых моделей, основанных только на временном ряде инфляции. Прогнозирование инфляции с использованием других макроэкономических переменных в качестве предикторов имеет серьезные ограничения, связанные, во-первых, с потенциально большим количеством информативных предикторов, во-вторых, с продолжительностью имеющихся временных рядов. Если количество наблюдений слишком мало, особенно при большом количестве предикторов, то любые модели часто подстраиваются к случайным закономерностям в обучающей выборке. Вне обучающей выборки модели часто дают неточный прогноз. Результаты коррелируют с результатами, полученными в работе [14] при исследовании влияния других экзогенных факторов на темпы инфляции

V. Заключение

В статье проведен анализ возможности применения методов машинного обучения для прогнозирования индекса потребительских цен, измеряющего динамику цен на потребительские товары и услуги. Использование библиотеки PyAF для прогнозирования временных рядов дает высокие показатели точности, позволяет автоматически выбирать модель и не требует сложной настройки параметров, но обучение занимает продолжительное время. Библиотека StatsForecast предоставляет возможность использовать различные модели и настраивать их параметры, однако в совокупности скорость и качество прогноза можно считать только удовлетворительными. Использование рекуррентной нейронной сети с долгой краткосрочной памятью для прогнозирования временных рядов в данной задаче не дает ощутимых преимуществ. Время, затраченное на обучение нейронной сети по данным 10 товаров сопоставимо с тем, что было получено при обучении Prophet и расчете инфляции по данным около 150 товаров, однако средняя абсолютная процентная ошибка с использование LSTM оказалась самой низкой.

В работе также проанализирована возможность уточнения прогнозов индекса потребительских цен путем учета различных экзогенных факторов.

Библиография

[1] Совет управляющих федеральной резервной системы. – URL: https:// www.federalreserve.gov/

[2] Pollak, Robert A. “The Consumer Price Index: A Research Agenda and Three Proposals.” The Journal of Economic Perspectives, vol. 12, no. 1, 1998, pp. 69–78. JSTOR, http://www.jstor.org/stable/2646939. Accessed 5 Nov. 2024.

[3] Hałka A, Leszczyńska A. The Strengths and Weaknesses of the Consumer Price Index: Estimates of the Measurement Bias for Poland. Gospodarka Narodowa. The Polish Journal of Economics. 2011; 250(9):51-75. doi:10.33119/GN/101086.

[4] Schultze, Charles, L. 2003. "The Consumer Price Index: Conceptual Issues and Practical Suggestions." Journal of Economic Perspectives, 17 (1): 3–22.

[5] 2018. "Chapter 17. The Consumer Price Index." In BLS Handbook of Methods. https://www.bls.gov/opub/hom/pdf/homch17.pdf.

[6] He, L.P., Fan, G. and Hu, J.N. (2008) Consumer Price Index and Producer Price Index: Who Drives Who? Economic Research, No. 11, 44-48.

[7] Faiga Kharimah, Mustofa Usman, Widiarti and Faiz AM. Elfaki. Time series modeling and forecasting of the consumer price index bandar lampung /Sci.Int.(Lahore),27(5),4619-4624,2015

[8] Tien-Thinh Nguyen, Hong-Giang Nguyen, Jen-Yao Lee, Yu-Lin Wang, Chien-Shu Tsai. The consumer price index prediction using machine learning approaches: Evidence from the United States, Heliyon, Volume 9, Issue 10, 2023, e20730, ISSN 2405-8440, https://doi.org/10.1016/j.heliyon.2023.e20730. (https://www.sciencedirect.com/science/article/pii/S2405844023079380)

[9] Sibai Fadi, El-Moursy, Ali Sibai, Ahmad. Forecasting The Consumer Price Index: A Comparative Study of Machine Learning Methods. International Journal of Computing and Digital Systems. 15. 2210-142. 10.12785/ijcds/150137.

[10] Oren Barkan, Jonathan Benchimol, Itamar Caspi, Eliya Cohen, Allon Hammer, Noam Koenigstein. Forecasting CPI inflation components with Hierarchical Recurrent Neural Networks, International Journal of Forecasting, Volume 39, Issue 3,2023,Pages 1145-1162,ISSN 0169-2070, https://doi.org/10.1016/j.ijforecast.2022.04.009.

[11] Бринк Х. Машинное обучение / Х. Бринк, М. Феверолф, Дж. Ричардс. – Санкт-Петербург : Питер, 2017. – 336 с.

[12] Коэльо Л.П. Построение систем машинного обучения на языке Python / пер. с англ. А. А. Слинкина / Л. П. Коэльо, В. Ричарт. – Москва : ДМК Пресс, 2016. – 302 с.

[13] Рашка С. Python и машинное обучение. – Москва : ДМК Пресс, 2017. – 418 с.

[14] Baybuza I. Inflation Forecasting Using Machine Learning Methods: / I. Baybuza // Russian Journal of Money and Finance. – 2018. – Vol. 77, № 4. – P. 42–59.

[15] Consumer Price Index Manual, 2020: Concepts and Methods. – Washington : International Monetary Fund, 2020. – 506 p. – ISBN 978-1-4843-5484-1.

[16] Stock J. Introduction to Econometrics, Global Edition / J. Stock, M. Watson . – London : Pearson, 2019. – 800 p. – ISBN 978-1-292-26445

Application of Machine Learning Methods for Forecasting the Consumer Price Index

T.V. Azarnova, N.G. Asnina, A.I. Kolosov, A.V. Lependin

***Abstract*— The article analyzes the effectiveness of machine learning methods for forecasting the consumer price index, which is the most important macroeconomic indicator of inflation and a key statistical indicator for governments and central banks in the process of assessing price stability. The research uses daily data on a large number of goods and services (more than 12 million units) from June 2020 to May 2022 as an information base. The difficulty of using data for forecasting lies in the presence of gaps in observations. Using special methods for processing gaps, resampling and eliminating outliers, data on individual goods are converted to continuous time series with equally spaced observations. The following machine learning methods are used for forecasting: PyAF, StatsForecastAutoARIMA, Prophet, LSTM recurrent neural networks. The article also uses machine learning methods to analyze the impact of such factors on inflation as the price of Brent crude oil, the rate of the Central Bank of the Russian Federation, the dollar to ruble exchange rate, the level of inflation for the month according to Rosstat data, GDP, the balance of trade, and examines the possibility of using these factors in forecasting models.**

***Keywords*— Consumer price index, forecasting of consumer price index, analysis of the influence of exogenous factors on the inflation rate, machine learning methods.**

REFERENCES

[1] Board of Governors of the Federal Reserve System. – URL: https:// www.federalreserve.gov/

[2] Pollak, Robert A. “The Consumer Price Index: A Research Agenda and Three Proposals.” The Journal of Economic Perspectives, vol. 12, no. 1, 1998, pp. 69–78. JSTOR, http://www.jstor.org/stable/2646939. Accessed 5 Nov. 2024.

[3] Hałka A, Leszczyńska A. The Strengths and Weaknesses of the Consumer Price Index: Estimates of the Measurement Bias for Poland. Gospodarka Narodowa. The Polish Journal of Economics. 2011; 250(9):51-75. doi:10.33119/GN/101086.

[4] Schultze, Charles, L. 2003. "The Consumer Price Index: Conceptual Issues and Practical Suggestions." Journal of Economic Perspectives, 17 (1): 3–22.

[5] 2018. "Chapter 17. The Consumer Price Index." In BLS Handbook of Methods. https://www.bls.gov/opub/hom/pdf/homch17.pdf.

[6] He, L.P., Fan, G. and Hu, J.N. (2008) Consumer Price Index and Producer Price Index: Who Drives Who? Economic Research, No. 11, 44-48.

[7] Faiga Kharimah, Mustofa Usman, Widiarti and Faiz AM. Elfaki. Time series modeling and forecasting of the consumer price index bandar lampung /Sci.Int.(Lahore),27(5),4619-4624,2015

[8] Tien-Thinh Nguyen, Hong-Giang Nguyen, Jen-Yao Lee, Yu-Lin Wang, Chien-Shu Tsai. The consumer price index prediction using machine learning approaches: Evidence from the United States, Heliyon, Volume 9, Issue 10, 2023, e20730, ISSN 2405-8440, https://doi.org/10.1016/j.heliyon.2023.e20730. (https://www.sciencedirect.com/science/article/pii/S2405844023079380)

[9] Sibai Fadi, El-Moursy, Ali Sibai, Ahmad. Forecasting The Consumer Price Index: A Comparative Study of Machine Learning Methods. International Journal of Computing and Digital Systems. 15. 2210-142. 10.12785/ijcds/150137.

[10] Oren Barkan, Jonathan Benchimol, Itamar Caspi, Eliya Cohen, Allon Hammer, Noam Koenigstein. Forecasting CPI inflation components with Hierarchical Recurrent Neural Networks, International Journal of Forecasting, Volume 39, Issue 3,2023,Pages 1145-1162,ISSN 0169-2070, https://doi.org/10.1016/j.ijforecast.2022.04.009.

[11] Brink H. Machine Learning / H. Brink, M. Feverolf, J. Richards. – Saint Petersburg : Piter, 2017. – 336 p..

[12] Coelho L.P. Building Machine Learning Systems in Python / trans. from English by A. A. Slinkin / L. P. Coelho, V. Richart. – Moscow : DMK Press, 2016. – 302 p..

[13] Rashka S. Python and Machine Learning. – Moscow : DMK Press, 2017. – 418 p..

[14] Baybuza I. Inflation Forecasting Using Machine Learning Methods: / I. Baybuza // Russian Journal of Money and Finance. – 2018. – Vol. 77, № 4. – P. 42–59.

[15] Consumer Price Index Manual, 2020: Concepts and Methods. – Washington : International Monetary Fund, 2020. – 506 p. – ISBN 978-1-4843-5484-1.

[16] Stock J. Introduction to Econometrics, Global Edition / J. Stock, M. Watson . – London : Pearson, 2019. – 800 p. – ISBN 978-1-292-26445