Банк России

Влияние негативных новостей на восприятие инфляции населением

Серия докладов об экономических исследованиях

№ 111 / февраль 2023

А. Евстигнеева
Д. Карпов

Евстигнеева Алина
Банк России, Департамент денежно-кредитной политики
E-mail: EvstigneevaAG@cbr.ru

Карпов Даниил
Банк России, Волго-Вятское главное управление
E-mail: dankarpov90@gmail.com

Авторы выражают признательность Ксении Юдаевой, Александру Морозову, Вадиму Грищенко, Генриху Пеникасу, а также анонимным рецензентам и участникам внутренних исследовательских семинаров Банка России за полезные комментарии и предложения.

Фото на обложке: Мария Багаева

107016, Москва, ул. Неглинная, 12
Тел.: +7 495 771-91-00, +7 495 621-64-65 (факс)
Официальный сайт Банка России: www.cbr.ru

Оглавление

Резюме

В этом исследовании представлен новый подход для определения новостей, которые могут оказывать наибольшее влияние на формирование инфляционных ожиданий населения. Он заключается в оценке роли негативных сообщений, касающихся экономической ситуации в широком смысле. Такой подход учитывает гипотезу о рациональной невнимательности домохозяйств: люди лишь частично обращают внимание на статистику по инфляции, так как потребление такой информации требует значительных усилий. В своей работе для анализа мы учли только те новости, которые «слишком дорого игнорировать», определив их с помощью опросов ВЦИОМ об основных страхах населения.

Мы разбили новости на 10 крупных тем: бедность, инфляция, продовольственные проблемы, закредитованность, экономический кризис, геополитика, курс рубля, пандемия COVID-19, внутренняя нестабильность, безработица. Кроме того, дополнительно мы включили тему, соответствующую интенсивности информационного присутствия Банка России в новостном потоке. Ее мы определили как долю сообщений, в которых отдельные члены Совета директоров или Банк России в целом комментируют инфляцию, от общего числа новостей.

Для определения значимости вклада каждой из этих тем в формирование инфляционных ожиданий населения мы построили четыре модели, три из которых позволяют оценить важность отдельных регрессоров (Lasso, Random Forest, XGBoost) и одна – построить взаимосвязи между переменными, представив массив данных как ориентированный граф (Bayesian Structure Learning). В качестве зависимых переменных мы использовали данные фонда «Общественное мнение» (ФОМ). Мы также проверили робастность результатов через оценку моделей с зашумленными данными, а также с альтернативными зависимыми переменными.

Все модели возвращают близкие результаты с высоким вкладом трех главных новостных потоков в формирование восприятия инфляции: инфляция, экономический кризис, курс рубля. Важным выводом нашей работы является получение подтверждений о том, что население может воспринимать вопрос о будущих ценах как более широкий вопрос об экономическом прогнозе в целом. При этом, отвечая на вопрос о прошлой инфляции, респонденты с высокой степенью вероятности могут держать в уме новости о конкретных ценах. Кроме того, мы обнаружили различия в восприятии инфляции и формировании инфляционных ожиданий между подгруппами респондентов со сбережениями и без них.

Ключевые слова: денежно-кредитная политика, текстовый анализ, инфляционные ожидания.

JEL-классификация: C83, E52, D83.

1. Введение

Инфляционные ожидания домохозяйств в последние годы привлекали много внимания исследователей. Высокий интерес академического сообщества во многом объясняется: 1) неокейнсианским подходом, согласно которому процесс инфляции является «вперед смотрящим», то есть инфляция в значительной мере зависит от инфляционных ожиданий; 2) растущей ролью коммуникации центральных банков, которая влияет на достижение целей денежно-кредитной политики через канал ожиданий. Вопрос о том, как домохозяйства формируют свои инфляционные ожидания, остается актуальным для исследователей по всему миру. Ответ на него позволит выстраивать более эффективный диалог с обществом, что будет способствовать достижению целей ценовой стабильности.

В этой работе мы опирались на три основных научных направления. Во-первых, мы начали с «эпидемиологической»[1] модели инфляции, предложенной в работе *Carroll, 2003*. В свое время этот подход изменил представления экономистов о том, что же управляет инфляционными ожиданиями. Автор выразил серьезное сомнение, что ожидания всех агентов рациональны. Он обосновал, что домохозяйства могут формировать свои ожидания, читая мнения профессионалов рынка в новостных СМИ. При этом ожидания аналитиков являются рациональными. Дополнительно К. Кэрролл сделал важный вывод: население не склонно постоянно следить за сообщениями СМИ об инфляции. Люди делают это лишь в критических ситуациях, когда динамика цен выходит из-под контроля и начинает их беспокоить. В более поздней работе *Pfajfar and Santoro, 2013* частично оспариваются выводы К. Кэрролла и доказывается, что новости об инфляции играют весьма скромную роль в точности прогнозов населения о будущей инфляции. Как было показано в *Larsen et al., 2019*, в целом сообщения СМИ являются качественными индикаторами для предсказания инфляционных ожиданий. Мы также приняли во внимание большой корпус исследований, посвященных вопросу о том, информация о каких конкретно ценах может в наибольшей степени влиять на ожидания (*Sussman et al., 2022* о ценах на энергоносители, а также *Arora et al., 2013* о ценах на продовольствие). В этой работе мы приводим дополнительные доказательства того, что СМИ играют важную роль в формировании инфляционных ожиданий населения и одновременно новости о конкретных ценах могут не иметь ключевое значение.

Во-вторых, важной отправной точкой для нашей работы стали исследования, посвященные гипотезе рациональной невнимательности домохозяйств (*Sims, 2003*). Она заключается в том, что экономические агенты не способны проанализировать всю имеющуюся информацию, но могут выбирать отдельные блоки для обработки. Наиболее полное обобщение подобных публикаций представлено в недавней работе *Maćkowiak et al., 2021*. Согласно гипотезе, если издержки от невладения информацией невысоки, ее можно проигнорировать. Такое

[1] Об «эпидемиологических» моделях говорят в тех случаях, когда рассматриваются модели, первоначально разработанные в эпидемиологии и описывающие распространение некоторого явления схожим с вирусом образом.

поведение домохозяйств особенно характерно для развитых стран с продолжительным успешным опытом обеспечения ценовой стабильности, где люди могут себе позволить лишь частично обращать внимание на статистику цен. Обработка такой информации требует существенных усилий, а потенциальные финансовые потери от ее игнорирования в странах с устойчиво низкой инфляцией стремятся к нулю (*Mankiw et al., 2004*). Для развивающихся стран ситуация выглядит иначе. В частности, *Arora et al., 2013* показали, что высокие и незаякоренные инфляционные ожидания в Перу связаны с недостатком знаний населения о том, как работает денежно-кредитная политика и в чем ее цели, а также с гиперреакцией на новостные сообщения.

Эта ситуация может дополнительно усугубляться за счет того, что центральные банки говорят с обществом на языке, который требует специальных экономических знаний. Как выяснил *Haldane, 2017b*, коммуникация центральных банков понятна лишь 5–10% населения.

В-третьих, наше исследование стало возможным благодаря растущему числу алгоритмов работы с большими данными, развитию машинной обработки естественных языков. Сегодня текстовый анализ делает возможным более глубокое понимание коммуникации центральных банков. Говоря об инфляционных ожиданиях как точке приложения усилий в этом направлении, мы можем отметить недавние работы *Sahu and Chattopadhyay, 2020*, *Tilly and Livan, 2021*, а также *Angelico et al., 2022*. Выводы, к которым приходят эти авторы, мы учли непосредственно при формулировании гипотезы. В частности, о том, что разные виды новостей по-разному влияют на инфляционные ожидания, о том, что инфляционные ожидания могут формироваться неинфляционными сообщениями, и о том, что негативные новости сильнее влияют на формирование инфляционных ожиданий.

Согласно нашему подходу, основанные на новостях модели инфляционных ожиданий могут быть улучшены с помощью снижения шума (а новости представляют собой высоко зашумленные данные). Для этого мы взяли лишь те сообщения, которые способны привлечь максимальное внимание аудитории. Наше решение обработать новостные потоки – это сужение перечня тем до негативных, то есть тех, которые «слишком дорого игнорировать». При этом, следуя логике обозначенных выше работ, мы вносим в массив данных новости широкого экономического контекста.

Наша гипотеза заключается в том, что резко негативные новости широкого экономического контекста определенным образом влияют на восприятие населением инфляции. Чтобы проверить это предположение, мы использовали большие данные сообщений российских СМИ. Для этого мы собрали базу новостей за 2014–2022 гг. из 28 СМИ. В базе представлено 7,8 млн сообщений.

Наша работа вносит вклад в литературу по нескольким направлениям. Во-первых, мы предлагаем новый метод фильтрации значимых для экономических агентов новостей. Во-вторых, мы предоставляем новые данные в изучение развития «эпидемиологической» теории инфляционных ожиданий, а также о том, каким образом домохозяйства могут формировать инфляционные ожидания через

сообщения СМИ. Кроме того, мы создали открытую базу российских новостей, которая может быть использована для различных NLP-исследований для России.

2. Данные

2.1. База новостей российских СМИ

Нам не удалось найти уже имеющуюся в открытом доступе базу новостей, которая отвечала бы целям исследования и включала региональные медиа. Поэтому мы создали базу новостей российских СМИ с широким покрытием экономической тематики. Процесс сбора и фильтрации данных доступен в репозитории проекта на GitHub[2]. Процесс составления базы мы условно разделили на две части – сбор данных и тематическое моделирование.

2.1.1. Сбор данных

Мы отобрали 28 ключевых российских СМИ (26 федеральных и 2 региональных, чей охват аудитории сопоставим с федеральными) на основе данных Медиалогии. Эта компания публикует ежемесячные рейтинги самых популярных российских СМИ, включая данные в региональном разрезе. Такие рейтинги дают довольно полное представление о предпочтениях россиян в вопросах потребления информации. После составления списка мы извлекли базовую информацию обо всех опубликованных в этих медиа сообщениях за период с января 2014 по август 2022 года. Информация состояла из следующих данных: название СМИ, заголовок новости, время публикации и ссылка на новость (URL). Дополнительно извлекалась категория новости. Более 91% информации было получено с помощью Feedly API.

Важно отметить, что для трех информагентств (Интерфакс, Регнум и РБК) изначально были выгружены только экономические новости – благодаря настройке RSS этих сайтов, что способствовало формированию словаря слов-маркеров (тегов) экономической темы. Это позволило на втором этапе – этапе тематического моделирования – применить алгоритмы обучения с учителем.

Следующий шаг – получение текста новости по ссылке. Были проанализированы HTML-коды сайтов СМИ для определения текста новости в общем коде. Далее полученный текст новости очищался от «мусора» (символы, отличные от знаков препинания, букв и цифр) и сохранялся в базу данных.

Было получено 7 млн 779 тыс. новостей из 28 крупных российских СМИ за период почти в девять лет. Названия СМИ, их принадлежность к федеральному/региональному уровню и количество включенных в базу новостей приведены в Приложении 1.

[2] https://github.com/evstalina/russian_news_database.

2.1.2. Тематическое моделирование

Следующий этап – выделение новостей, связанных с экономикой. Эту задачу можно решить двумя методами работы с большими данными: обучение с учителем (используя размеченные данные) и без учителя.

Как правило, для выделения различных тем из новостного массива путем так называемой мягкой кластеризации используют модели без учителя, в частности, Latent Dirichlet allocation (LDA, латентное размещение Дирихле, описан в *Blei et al., 2003*). Однако благодаря тому, что некоторые новости уже были размечены, стало возможным провести разметку и для всех остальных новостей. Это потенциально позволит повысить качество тематического моделирования. Мы применили оба подхода и выбрали лучший.

Перед собственно тематическим моделированием мы провели стандартную предварительную обработку текста, которая включала токенизацию (то есть разбиение текста на отдельные слова и знаки препинания) и лемматизацию (или приведение слова к базовой форме). Далее в зависимости от задачи были получены численные метрики: для обучения с учителем – TF-IDF, для обучения без учителя – Count Vectorizer.

TF-IDF (от англ. TF – term frequency, IDF – inverse document frequency) – статистическая мера, которая используется для оценки важности слова в контексте документа, являющегося частью коллекции документов или корпуса (*Salton, 1988*):

$$TF - IDF = \frac{n_t}{\sum_k n_k} \log\left(\frac{1+n}{1+df_t}\right), \text{ где} \qquad (1)$$

n_t – количество вхождений слова t в документ,

$\sum_k n_k$ – количество слов в документе,

n – количество документов в датасете,

df_t – количество документов, в которых встречается слово t.

TF-IDF дает информацию о важности конкретного слова для определения различий между группами текстов. Эта мера несет в себе больше информации, чем простой «мешок слов» (bag of words – множество слов с количественными характеристиками без учета грамматики, порядка слов (*Harris, 1954*)), но при этом работает намного быстрее, чем word2vec[3], и иные методы, основанные на нейронных сетях, что было важно на большом наборе данных. TF-IDF использовалась для моделей с учителем. Дополнительно была осуществлена фильтрация TF-IDF: слово должно было употребляться в корпусе как минимум 100 раз – так мы отсекли узкоспециализированные термины и опечатки. К тому же слово должно как максимум входить в 80% документов – так мы отсекли наиболее частотную лексику (например, предлоги и союзы, которые не играют никакой роли

[3] Семейство алгоритмов, использующих нейронные сети для получения векторного представления слов.

для определения темы текста). Фильтрация одновременно позволяет повысить скорость обучения модели за счет снижения размерности итоговой матрицы.

Для обучения без учителя (в исследовании использовалось LDA) при предварительной обработке мы применили Count Vectorizer – метод, конвертирующий пул документов в матрицу, где столбец соответствует конкретному слову, а строчка – тексту. Далее мы использовали модификацию модели LDA – Guided LDA (или Seeded LDA, который является представлением классической LDA и описан в *Jagarlamudi et al., 2012*). Суть модели заключается в том, что словарь для темы используется для того, чтобы вызвать смещение к нужному, а также для смещения выбора тем в документе к темам, относящимся к предварительно отобранным словам. Таким образом, Guided LDA подходит для выделения экономической темы в корпусе новостей.

Словарь экономических слов был составлен авторами на основе словаря наиболее частых слов в изначально экономических новостях трех выше упомянутых СМИ. Дополнительно словарный список был очищен экспертами от слов, не относящихся к экономике. Полученный словарь был использован для формирования темы 0. Остальные темы не имели начального смещения в словаре, поэтому формировались стандартными алгоритмами LDA. Обучение выполнено на основе выборки в 100 тыс. новостей по 25 источникам, представленным в равных долях в выборке (исключая три ранее упомянутых СМИ, чтобы избежать смещения распределения тем в сторону экономических новостей).

Размеченная выборка для моделей с учителем была сформирована на основе списка самых популярных категорий новостей (которые покрыли почти половину от всех новостей базы), сгруппированных по шести основным темам (см. Приложение 1), одна из которых посвящена экономике. Обучение выполнено на подвыборке в 120 тыс. новостей, каждая из шести тем представлена равными долями. Отложенная выборка составляла одну треть.

Для обучения с учителем использовались следующие методы: метод опорных векторов (support vector machine, SVM, *Boser et al., 1992*), softmax-регрессия, метод случайного леса (random forest, *Breiman, 2001*) и градиентный бустинг (XGBoost, *Chen & Guestrin, 2016*). Выбор моделей объясняется множеством факторов: результаты softmax-регрессии можно интерпретировать как базовое решение, метод опорных векторов – классический метод для задач текстового анализа; метод случайного леса является основным ансамблевым методом над решающими деревьями, а градиентный бустинг – в данный момент самый популярный метод в задачах классификации, который часто дает лучший результат.

Всего было оценено пять моделей: четыре 6-классовые модели и одна 2-классовая для сравнения влияния количества тем на значение метрики. Для 2-классовой модели использовалась спецификация, при которой 0 – экономическая новость, 1 – остальные категории, соответственно, тестовая часть для этой модели также представлена была двумя классами. Все модели с учителем были обучены на кросс-валидации (оценка модели, при которой данные делятся на k равных частей, обучение происходит на k – 1 части, а оценка качества на k-части, при этом

каждая из k-частей служит как для оценки качества, так и для обучения, процедура повторяется k раз) с пятью разделениями и предварительным перемешиванием.

Качество Guided LDA-модели проверялось на основе тестовой размеченной выборки для двухклассовой модели с учителем. Для возможности оценки метрик по экономической теме темы Guided LDA также были сгруппированы: 0 – экономика, 1 – остальные темы. В этом случае использование классических метрик для задач тематического моделирования (например, когерентность) не имело смысла, так как есть размеченный датасет для проверки качества. Основными метриками были выбраны precision, recall и F-score для экономической темы.

Guided LDA-модели с 7 и 10 темами оказались лучшими по F-score, для последующего сравнения была взята модель с 7 темами. Обученные модели с учителем показали лучшие результаты, чем Guided LDA, что подтверждает то, что модели с подкреплением (reinforcement learning) дают лучший результат. Сравнение результатов моделей представлено в Приложении 1.

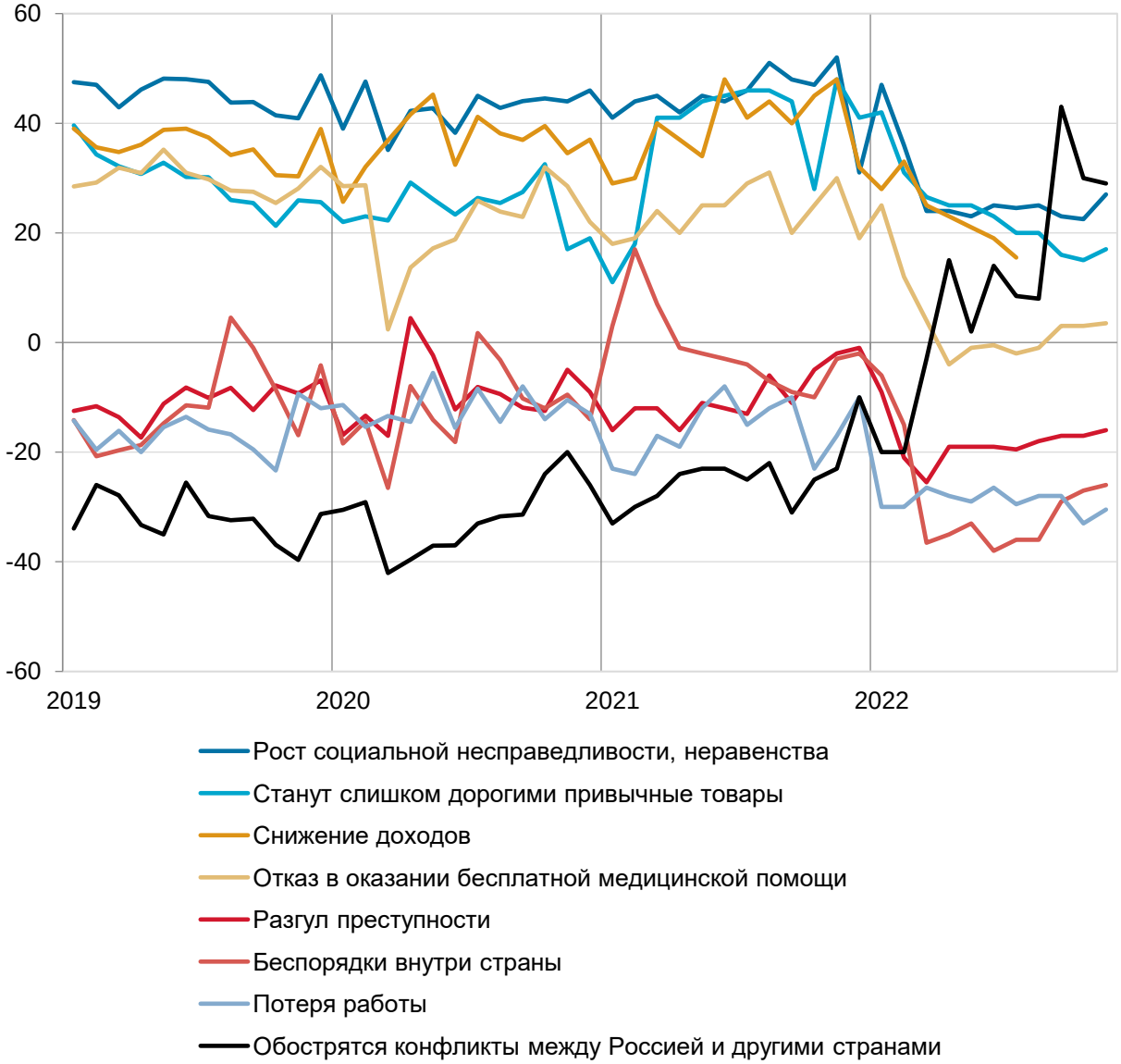
Наибольший F-score по экономическим темам показали модели SVM и XGBoost для шести классов, показав равный результат – 0,94 на тестовой выборке. Однако XGBoost дает лучший результат по precision, а также, что немаловажно, намного быстрее формирует прогноз, в отличие от SVM, поэтому для дальнейшей работы была выбрана именно эта модель. Всего в корпусе представлено 1,567 млн экономических новостей из 28 СМИ за период с начала 2014 по август 2022 года.

2.2. Определение негативной семантики

Выявление типа новостей, которые максимально влияют на принятие домохозяйствами решений, может быть одной из самых сложных задач для исследователей (*Larsen et al., 2019*). Как отмечают *Coibion et al., 2019*, использование для этих задач новостей по денежно-кредитной политике не выглядит многообещающим из-за низкого интереса общества к этой теме. Мы подтвердим выводы этого исследования в следующих главах.

Наша гипотеза (см. подробнее Введение) строится на предпосылке, что люди могут формировать свои инфляционные ожидания на основе более общих новостей и с уклоном в негативные, поэтому наша методика отбора исходит из правила о включении новостей, которые «слишком дорого игнорировать». Чтобы максимально точно определить такие сообщения, мы используем опросные данные Всероссийского центра изучения общественного мнения (ВЦИОМ) об «индексе страхов» россиян (рис. 1).

Рис. 1. Индекс страхов россиян, данные опросов[4]



Для выделения негативных новостей мы вновь использовали Guided LDA. Как уже было описано выше, этот алгоритм требует на вход (или *посев*) список экспертно выбранных якорных токенов, описывающих тему. В качестве якорных токенов на этот раз мы использовали соответствующие варианты ответов из опроса ВЦИОМ (рис. 1) и их синонимы.

[4] Индекс страхов показывает, насколько высокой кажется россиянам вероятность наступления той или иной проблемы. Индекс строится на основе вопроса «Как вы оцениваете вероятность появления следующих проблем в вашей жизни?», измеряется в пунктах и может колебаться в пределах от -100 до 100. Ответу «Полностью уверен, что случится» присвоен коэффициент 1, ответу «Скорее случится» – коэффициент 0,5, ответу «Скорее не случится» – (-0,5), ответу «Полностью уверен, что не случится» (-1). Чем выше значение индекса, тем более вероятным кажется россиянам наступление проблемы. Данные представлены на основании всероссийских телефонных опросов «Спутник». Статистическая погрешность не превышает 2,5%.

2.3. Посев негативных якорных токенов

В качестве следующего шага мы произвели посев полученных якорных токенов с помощью модели Guided LDA. При этом мы разбили восемь изначальных тем ВЦИОМ на 10 блоков для посева. Это было сделано для улучшения объясняющей силы итоговых моделей и из-за высокой концентрации тем внутри отдельных вопросов ВЦИОМ. Например, ответу об инфляции потенциально соответствует несколько подтем: о собственно ценах, а также о долговой нагрузке и продовольственных проблемах. Кроме того, мы выделили курс рубля в отдельную тему из-за его высокой важности для россиян[5]. Кроме того, мы объединили темы роста преступности и внутренней нестабильности. В результате на выходе из Guided LDA мы получили список из 10 тем с вероятностным распределением слов внутри каждой из них. Следуя рекомендациям *Martin and Johnson, 2015*, мы отдали предпочтение существительным в наборе финальных токенов для уменьшения зашумленности моделей. Также для оценки возможной роли коммуникации Банка России по теме денежно-кредитной политики мы добавили тему, соответствующую коммуникации. Ее мы определили как новости, в которых отдельные члены Совета директоров или Банк России в целом комментируют инфляцию.

2.4. Новостные временные ряды

После определения критериев отнесения новостей к темам мы рассчитали частотность каждой темы в корпусе по следующей формуле:

$$Freq_{i,t} = \frac{N_{i,t}}{\sum_k N_{k,t}}, \text{ где} \qquad (2)$$

$\mathrm{Freq_{i,t}} -$ частотность темы i в месяц t,
$\mathrm{N_{i,t}} -$ количество новостей, содержащих тему i в месяц t,
$\sum_{\mathrm{k}} N_{\mathrm{k,t}} -$ общее количество новостей в месяце t.

Полученные новостные временные ряды по 10 отобранным темам представлены на рис. 2–12.

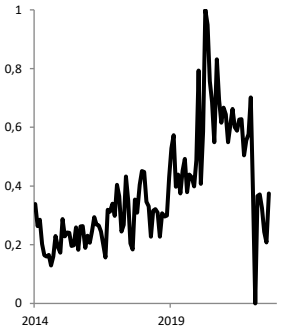
Наш подход отличается от предложенного в работе *Larsen et al., 2019* и других исследователей, в которых внимание СМИ к теме измеряется по весу LDA-юнита в корпусе как едином документе. По нашему мнению, такой подход может приводить к завышению значения темы. Этот эффект может быть связан со структурой новости. Когда журналист описывает событие в одном материале, он упоминает его несколько раз: в заголовке, лиде, после чего приводит различные цитаты, мнения и прочие расширения информации, и каждый из этих блоков может содержать интересующую тему. Если объединить все заметки СМИ в единый

[5] По опросам фонда «Общественное мнение» по заказу Банка России, до 18% россиян учитывают курс рубля при формировании своих инфляционных ожиданий. См. подробнее: https://www.cbr.ru/analytics/dkp/inflationary_expectations/#a_102610.

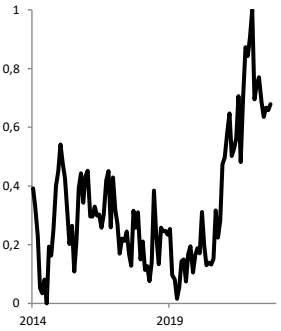
массив, алгоритм LDA будет воспринимать одну новостную заметку как несколько упоминаний одной и той же темы, что может отразиться на финальном качестве модели.

Рис. 2–12. Новостные временные ряды[6]

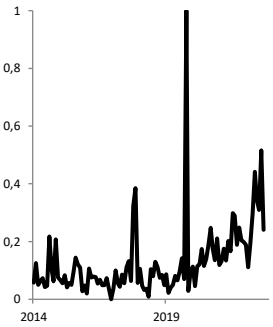
а) Т1: Бедность



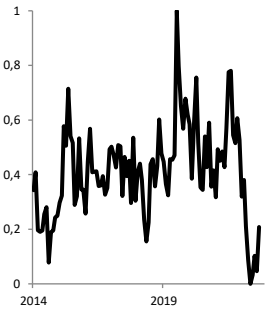
б) Т2: Инфляция



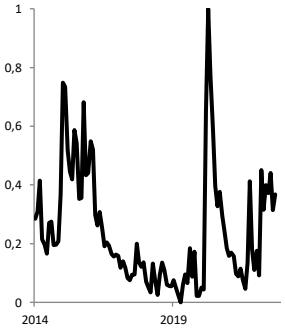
в) Т3: Продовольственные проблемы



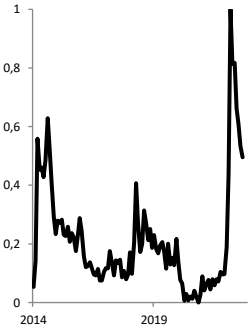
г) Т4: Закредитованность



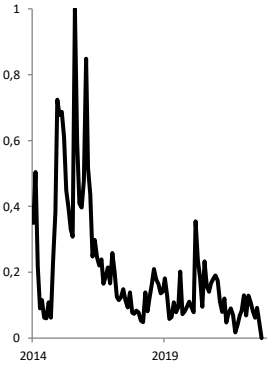
д) Т5: Экономический кризис



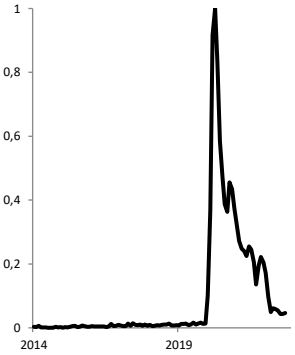
е) Т6: Геополитика



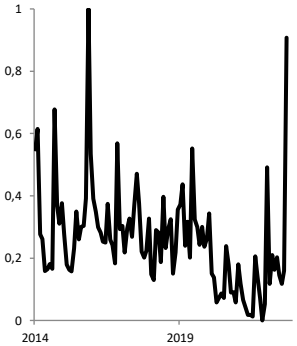
ж) Т7: Курс рубля



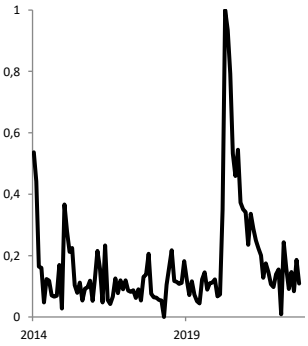
з) Т8: COVID-19



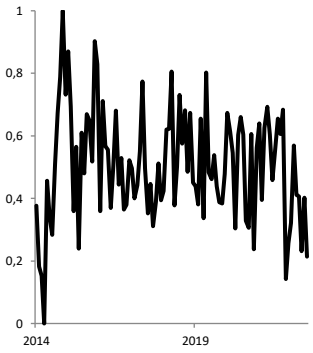
и) Т9: Нестабильность



к) Т10: Безработица



л) Т11: Коммуникация Банка России



[6] Данные на графиках нормализованы.

В нашем подходе одна тема может присутствовать в тексте одной новости только один раз. При этом заметка СМИ может одновременно принадлежать к разным темам. В качестве дополнительного обоснования своего подхода приведем данные в пользу распространенного «чтения по заголовкам» среди пользователей. Об этом свидетельствуют как академические исследования (*Gabielkov et al., 2016*), так и многочисленные опросы и маркетинговые исследования (например, проведенные American Press Institute (API) и Associated Press-NORC Center, а также Nielsen Norman Group). По нашему мнению, читатель воспринимает статью как одну тему или несколько, но не как несколько вхождений одной и той же темы, особенно если он читает только заголовки.

Некоторые временные ряды сильно коррелированы (рис. 2–12, Табл. 1). В частности, T8: пандемия COVID-19 и T10: безработица (коэффициент корреляции Пирсона 0,81), T1: бедность и T8: пандемия COVID-19 (0,76), а также T1: бедность и T10: безработица (0,64). Корреляции возникли, несмотря на то, что в этих темах нет общих слов или их незначительно мало. Иначе говоря, существует достаточное количество сообщений СМИ, в которых темы бедности, пандемии и безработицы присутствуют одновременно в тесной семантической связи. Мы интерпретируем это следующим образом: во время пандемии (2020–2021 гг.) и особенно в острую ее фазу начала 2020 г., Россия, как и многие другие страны, вводила ограничения на работу мест массового скопления людей (локдауны). Соответственно, много людей, занятых в сфере услуг, опасались потерять работу и ухудшить свое материальное положение. Поэтому эти темы возникали в новостях одновременно. Хотя стандартные подходы к работе с временными рядами требуют от нас объединения этих тем в одну общую, мы все же сознательно сохраняем их как три разные темы, пожертвовав частично качеством моделей ради их большей интерпретируемости.

Табл. 1**.** Корреляционная матрица для новостных тем

	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11
T1	1,00	0,22	0,21	0,43	0,04	-0,46	-0,30	**0,76**	-0,36	**0,64**	0,05
T2	0,22	1,00	0,36	-0,13	0,18	0,19	0,04	0,10	-0,14	0,01	0,19
T3	0,21	0,36	1,00	-0,06	0,05	0,18	-0,22	0,13	-0,18	0,11	-0,11
T4	0,43	-0,13	-0,06	1,00	-0,08	-0,63	-0,02	0,25	-0,16	0,17	0,13
T5	0,04	0,18	0,05	-0,08	1,00	0,16	0,57	0,39	-0,02	0,57	0,20
T6	-0,46	0,19	0,18	-0,63	0,16	1,00	-0,01	-0,37	0,15	-0,27	-0,23
T7	-0,30	0,04	-0,22	-0,02	0,57	-0,01	1,00	-0,10	0,20	0,06	0,36
T8	0,76	0,10	0,13	0,25	0,39	-0,37	-0,10	1,00	-0,43	**0,81**	0,03
T9	-0,36	-0,14	-0,18	-0,16	-0,02	0,15	0,20	-0,43	1,00	-0,34	0,03
T10	0,64	0,01	0,11	0,17	0,57	-0,27	0,06	0,81	-0,34	1,00	0,09
T11	0,05	0,19	-0,11	0,13	0,20	-0,23	0,36	0,03	0,03	0,09	1,00

2.5. Зависимые переменные

В России регулярные опросы об инфляционных ожиданиях населения с 2009 г. проводит фонд «Общественное мнение» (ФОМ) по заказу Банка России. Опрос ФОМ включает длинный перечень вопросов, направленных на выявление отношения респондентов к ценовой динамике, ожиданий по ценам и потребительских стратегий. В том числе ФОМ публикует два важных для нас ряда данных: инфляционные ожидания домохозяйств на 12 месяцев вперед и наблюдаемую населением инфляцию в течение последних 12 месяцев. ФОМ задает вопросы «Как, по вашему мнению, в целом будут меняться цены в следующие 12 месяцев (год)?» и «Как, по вашему мнению, в целом менялись цены за прошедшие 12 месяцев (год)?». Чтобы квантифицировать по сырым данным опросов переменные об ожидании, Банк России использует вероятностный метод квантификации, описанный в статье *Berk, 1999*, а также *Carlson and Parkin, 1975*. Подробная методика доступна на сайте Банка России.

2.6. Предварительная обработка временных рядов

Прежде всего мы исключили календарный фактор из полученных данных, разделив каждое значение временного ряда на количество дней в месяце. Далее мы применили трансформацию Бокса – Кокса (*Box and Cox, 1964*) для сглаживания амплитуды колебаний и очистили данные от сезонного фактора. После этого мы применили Min-Max Scaling, чтобы все независимые переменные поступали в модель с одинаковым весом. Все временные ряды новостных индексов были лагированы на 1 месяц. То есть мы оцениваем ожидаемую и наблюдаемую населением инфляцию за месяц t по новостям t – 1, предполагая, что новости в прошлом месяце повлияли на их отношение к теме инфляции в текущем месяце.

Для дальнейшего анализа мы применили тест Дики – Фуллера о стационарности временных рядов для 11 независимых переменных и двух зависимых. Стационарность была отвергнута на 5%-ном уровне значимости для всех проверенных временных рядов, кроме ряда для коммуникации Банка России. Чтобы избежать проблем со спорными результатами регрессий, мы взяли первые разности по всем нестационарным рядам. После этого все ряды прошли тест Дики – Фуллера на стационарность. Кроме того, для Bayesian Structure Learning мы трансформировали первые разности в дискретные данные, присвоив категориальные факторы, соответствующие квартилям.

Для методов машинного обучения мы не приводили данные к стационарным, как рекомендовано в последних работах по этой теме. В частности, *Ahmed et al., 2010* объясняют, что важен абсолютный уровень переменных, который при переводе рядов в стационарный формат теряется, и в итоге алгоритмы машинного обучения теряют доступ к важной информации. Преимущества использования сырых данных для support vector machines продемонстрированы в *Kim, 2003*, для нейронных сетей – в *Kondratyev, 2018*. В исчерпывающем обзоре *Petelin at al., 2022* приведены самые подробные рекомендации об обработке временных рядов для

применения алгоритмов машинного обучения, и решение задачи отбора признаков (feature selection) предложено как показание к выбору в пользу сырых рядов, не прошедших дополнительных шагов по приведению рядов к стационарному виду. Именно эту задачу (отбора признаков) мы решаем в настоящем исследовании.

Наконец, мы применили бутстреп (*Efron, 1979*), разбив данные на небольшие подвыборки. На наш взгляд, это позволило частично сгладить недостатки небольшого количества данных в нашем датасете, а также рассчитать критически важные метрики качества моделей.

3. Методы

В этой работе мы решаем задачу определения вклада тематических новостных индексов в формирование наблюдаемой и ожидаемой инфляции. Формально это задача отбора признаков (feature selection). Она может быть решена с помощью регрессионных моделей, позволяющих оценивать важность независимых переменных и сравнивать ее между собой, а также с помощью более сложных моделей определения причинно-следственных связей между переменными. Мы выбрали четыре модели: три классические, которые используются наиболее часто для решения схожих задач (Lasso, Random Forest, XGBoost), а также одну нестандартную – Bayesian Structure Learning, которая получила большое распространение в последние годы в таких областях, как биоинформатика и медицинские науки. В этом разделе мы описываем спецификации отобранных моделей.

3.1. Lasso

Least Absolute Selection and Shrinkage Operator (Lasso) был предложен *Tibshirani, 1996* и на сегодня является одним из самых популярных методов регуляризации (то есть наложения дополнительных ограничений на параметры модели для предотвращения ее избыточной сложности и определения наиболее значимых предикторов). Функция потерь может быть описана следующим образом:

$$\sum_{i=1}^{n}(y_i - \sum_j x_{ij}\ \beta_j)^2 + \lambda \sum_{j=1}^{p}|\beta_j| \quad \text{, где} \qquad (3)$$

n – количество наблюдений,
p – количество признаков (независимых переменных),
λ – параметр регуляризации, штраф за сложность.

Минимизируя функцию потерь, Lasso обнуляет не вносящие вклада в работу модели переменные и отбирает полезные признаки. Обнуление признака означает приравнивание коэффициента перед ним к нулю, тогда как полезные признаки получают положительные коэффициенты. Lasso был отобран в качестве базовой модели ввиду своей простоты, распространенности и высокой интерпретируемости.

3.2. Random Forest

Random Forest – ансамблевый метод, который позволяет решать задачи регрессионного анализа с помощью множества решающих деревьев. Алгоритм предложен в работе *Breiman, 2001*. Мы используем спецификацию модели с 50 решающими деревьями, а также минимальным размером наблюдений для принятия решения в два наблюдения. Мы также используем K-Fold Cross Validation для предотвращения переобучения модели со значением k = 10. Для оценки вклада переменных мы выбрали метод выключения (permutation feature importance). Он основан на измерении увеличения ошибки прогноза после поочередного выключения независимых переменных. Если ошибка увеличивается, модель штрафует переменную.

3.3. XGBoost и Kernel SHAP

Extreme Gradient Boosting (XGBoost, экстремальный градиентный бустинг) предложен в работе *Chen & Guestrin, 2016*. Это техника машинного обучения для задач классификации и регрессии, которая строит модель предсказания в форме ансамбля слабых предсказывающих моделей – обычно деревьев решений. Метод доказал свою эффективность для небольших и средних размеров данных. В нашей спецификации мы использовали максимальную глубину дерева в размере 2, а среднюю квадратическую функцию как функцию потерь. Для выявления важности переменных мы применили Kernel SHAP (SHapley Additive exPlanation), описанный в *Lundberg and Lee, 2017*. Метод достраивает специальные линейные регрессии для взвешивания признаков. Важность признаков квантифицируется как значения Шепли из теории игр, то есть это средний ожидаемый предельный вклад одного игрока (в нашем случае – новостного индекса) после рассмотрения всех возможных комбинаций. Значения Шепли рассчитываются следующим образом:

$$\varphi_i(v) = \sum_{C \subseteq N-i} \frac{|C|!(n-|C|-1)!}{n!}[v(N \backslash C) - v(C)] \text{, где} \qquad (4)$$

φ – значение Шепли,
n – количество признаков (независимых переменных),
C – подмножество признаков без i – того признака,
$|C|$ – размер датасета,
$\frac{|C|!\,(n-|C|-1)!}{n!}$ может быть интепретировано как вероятность того, что в любой перестановке переменные из группы C опережают успешного игрока i.

3.4. Bayesian Structure Learning

Bayesian Structure Learning стоит особняком среди описанных выше методов. Это вероятностная графовая модель, в которой в виде ориентированного ациклического графа (directed acyclic graph, DAG) описаны взаимосвязи между

переменными в качестве вершин и их зависимостей в качестве ребер. Если ребер между переменными нет, это интерпретируется как независимость переменных. Байесовские сети используются для установления причинно-следственных отношений (см. *Pearl, 1988* как одну из первых работ, посвященных изучению возможностей этого класса моделей). Мы применили несколько протоколов обучения и с помощью байесовского информационного критерия (Bayesian information criterion, BIC) выбрали лучший: PC algorithm (*Spirtes and Glymour, 1991*), Grow-Shrink для марковских сетей (*Margaritis, 2003*) и Incremental Association Markov Blanket (*Tsamardinos et al., 2003*). Оценка моделей была произведена с помощью хорошо зарекомендовавшей себя в предыдущих исследованиях библиотеки bnlearn в R (*Scutari, 2009*).

4. Результаты

Для алгоритма BSL в качестве отбора переменных приведены значения True/False, где первое соответствует наличию ребра между соответствующим новостным индексом и зависимой переменной, а второе – его отсутствию (Табл. 2). В качестве метода оценок вклада переменных для Lasso использованы стандартные веса – коэффициенты модели, для Random Forest – метод выключения переменных (permutation feature importance), для XGBoost – значения Шепли. Все полученные значения были пропорционально преобразованы к сумме 1 для удобства сравнения. Жирным шрифтом в каждом столбце выделены топ-3 переменных, отобранных моделями. Несмотря на разные модели, различия в предобработке данных и разные алгоритмы выявления важности признаков, модели показали схожие результаты. Наиболее часто отбирались такие новостные темы, как инфляция, экономический кризис и курс рубля. Наиболее значимыми новостями в модели для инфляционных ожиданий при этом скорее были курс рубля и инфляция, а в модели наблюдаемой инфляции – инфляция и экономический кризис. Переменная, отражающая интенсивность присутствия Банка России в новостном потоке, не оказывала значимого влияния на восприятие инфляции населением ни в одной из моделей.

Табл. 2. Важность новостных индексов для всех моделей

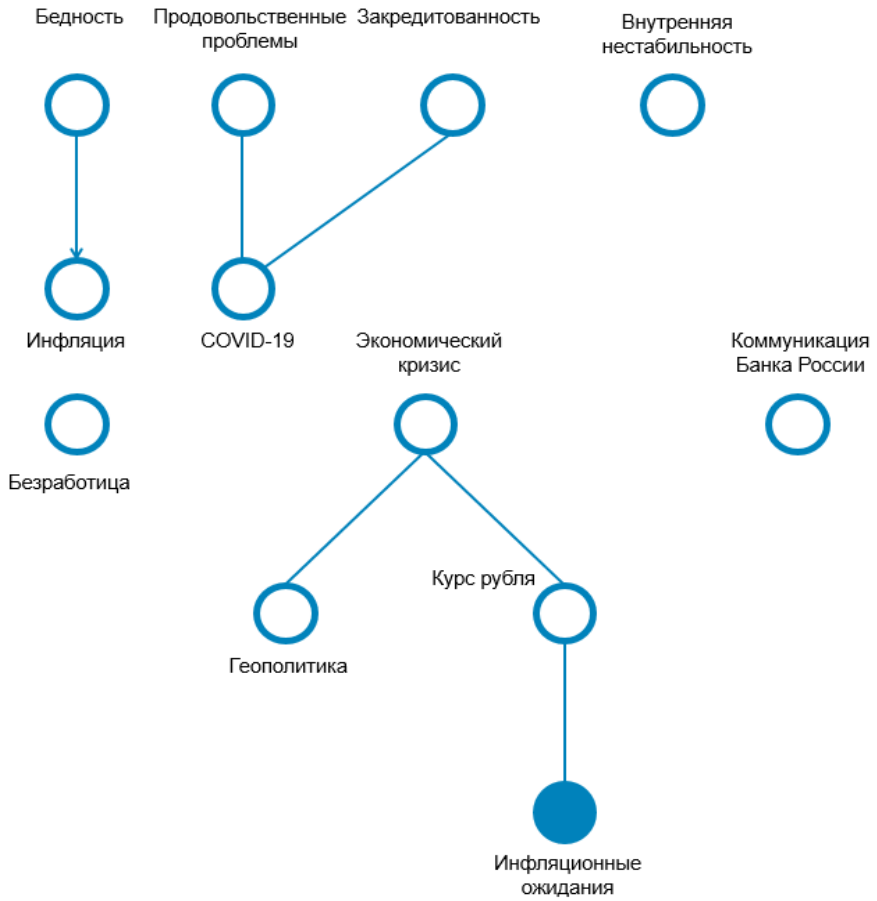
Новостной индекс	**Инфляционные ожидания**				**Наблюдаемая инфляция**			
	Lasso	**RF**	**XGBoost**	**BSL**	**Lasso**	**RF**	**XGBoost**	**BSL**
Бедность	0,00	0,02	0,05	False	**0,24**	0,06	0,11	False
Инфляция	**0,32**	**0,25**	**0,34**	False	**0,14**	**0,20**	**0,25**	**True**
Продовольственные проблемы	0,00	0,02	0,02	False	0,03	0,02	0,01	False
Закредитованность	0,00	0,02	0,01	False	0,00	0,04	0,05	False
Экономический кризис	**0,29**	0,02	0,05	False	**0,18**	**0,18**	**0,16**	False
Геополитика	0,00	0,02	0,00	False	0,13	0,08	0,06	False
Курс рубля	**0,39**	**0,42**	**0,26**	**True**	0,00	**0,27**	**0,22**	False

Пандемия COVID-19	0,00	0,15	**0,25**	False	0,04	0,09	0,10	False
Внутренняя нестабильность	0,00	0,02	0,01	False	0,00	0,03	0,02	**True**
Безработица	0,00	0,05	0,01	False	0,13	0,03	0,01	False
Коммуникация	0,00	0,01	0,00	False	0,12	0,01	0,01	False
Adjusted R^2	0,23	0,74	0,83		0,43	0,81	0,67	
MAE	0,03	0,06	0,06	0,02	0,01	0,01	0,01	0,03

В качестве метрик качества моделей мы выбрали скорректированный коэффициент детерминации (Adjusted R^2) и среднюю абсолютную ошибку (для BSL ошибка определялась для зависимых переменных в качестве тестируемой вершины графа). Random Forest продемонстрировал более высокий R^2 как для наблюдаемой, так и для ожидаемой инфляции. Наименьшую ошибку при этом для ожидаемой инфляции показал BSL, а для наблюдаемой – Random Forest. Более подробные данные о качестве полученных моделей приведены в Приложении 2.

В качестве результатов работы BSL мы представляем полученные вероятностные графы связей между переменными (рис. 13, рис. 14). Модель затруднилась определить направленность для некоторых ребер, что мы связываем с одновременностью происходящих процессов (по крайней мере внутри одномесячного временного интервала, который был предложен модели).

Рис. 13. Граф для инфляционных ожиданий

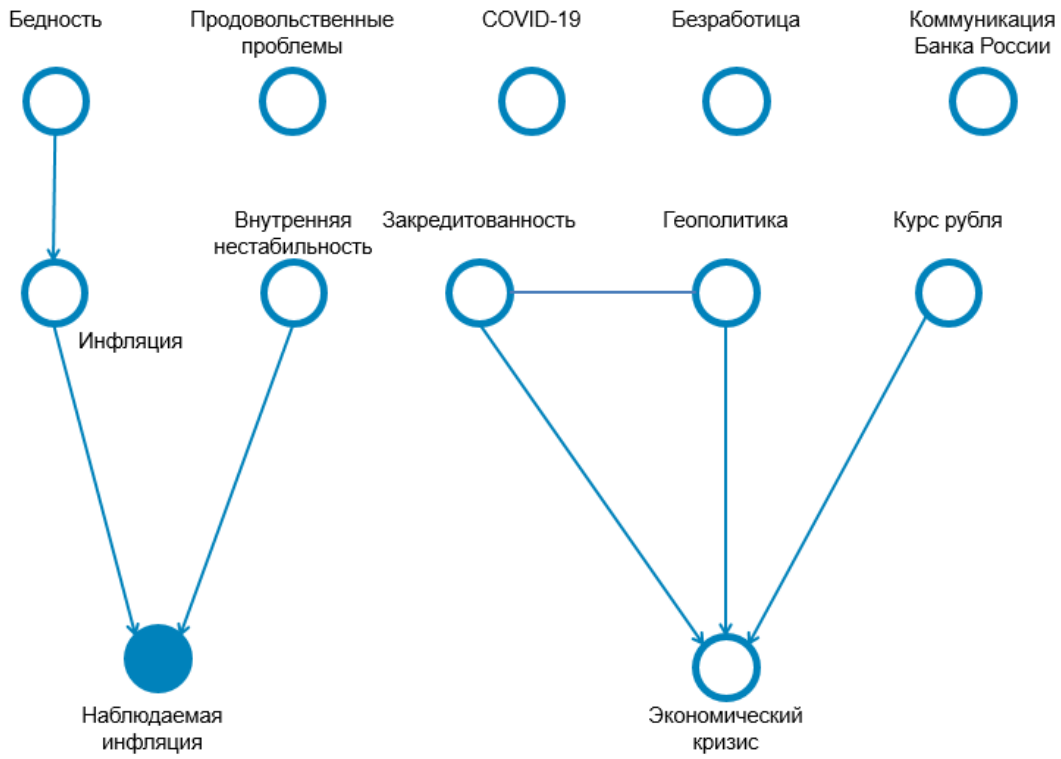


Анализируя полученную структуру (рис. 13) для инфляционных ожиданий, можно предположить, что респонденты, отвечая на вопрос о будущей динамике цен, как минимум принимают во внимание экономические условия в целом, которые для России исторически тесно связаны с геополитикой и курсом национальной валюты. Другими словами, вопрос о будущих ценах люди могут воспринимать как более широкий вопрос о перспективах экономики вообще. При этом новости о снижении уровня жизни и инфляции оказались тесно связанными.

Граф для наблюдаемой инфляции (рис. 14) может указывать на то, что респонденты при ответе на вопрос о прошлой инфляции могут вспоминать реальные колебания цен, а также учитывать ситуацию внутри страны в целом, включая информацию о бедности и продовольственной безопасности. Обнаруженные взаимосвязи могут отражать логику восприятия инфляции как оценку текущих условий жизни.

Разная кластеризация отобранных переменных на графах свидетельствует о возможных различиях в формировании представлений о будущей и прошлой инфляции.

Рис. 14. Граф для наблюдаемой инфляции



5. Проверки робастности

В этой части мы описываем проверки робастности полученных моделей и их результатов. Чтобы их провести, мы рассчитали набор альтернативных регрессий и моделей и проанализировали результаты (см. Приложение 3).

Первый набор измененных данных – описанные выше зависимые и независимые переменные плюс три случайно сгенерированные переменные белого шума вида ARIMA (0, 0, 0). Для моделей, которые использовали нестационарные ряды, в белый шум был дополнительно внесен случайный тренд. С помощью этого эксперимента мы проверяем, останется неизменным выбор моделей в пользу отобранных ранее новостных индексов или внесение шума исказит оценки. По результатам тестов прошли проверку Random Forest и Lasso: все три выбранные ими наиболее значимые переменные сохранились. XGBoost в спецификации с наблюдаемой инфляцией показал большую значимость пандемии, чем экономического кризиса (что наблюдалось в базовой модели). Альтернативная BSL была устойчива к зашумливанию данных и не создала новых «ребер».

Второй блок проверки робастности – оценка разработанных моделей с новыми переменными отклика. В качестве альтернативного измерения инфляционных ожиданий и наблюдаемой инфляции мы использовали данные ФОМ в разрезе подгрупп респондентов со сбережениями и без них. Таким образом, у нас получилось четыре новых зависимых переменных: инфляционные ожидания для респондентов со сбережениями, инфляционные ожидания для респондентов без сбережений, наблюдаемая инфляция для респондентов со сбережениями, наблюдаемая инфляция для респондентов без сбережений. Практически все результаты оказались робастными, то есть топ-3 наиболее важных признаков в основном сохранились. Изменения произошли в части: 1) Lasso-модели для наблюдаемой инфляции для домохозяйств без сбережений (геополитика стала существенно более значимым фактором); 2) Lasso-модели для наблюдаемой инфляции для домохозяйств со сбережениями (вместо темы инфляции в топ-3 вошла безработица); 3) Lasso-модели для ожидаемой инфляции для домохозяйств без сбережений (коммуникация БР стала значимым фактором, но с минимальным коэффициентом); 4) XGBoost-модель для наблюдаемой инфляции для домохозяйств без сбережений показала большую значимость темы бедности, чем пандемии. При этом оценки значимости для подгрупп со сбережениями и без них несколько изменились, что можно отнести к различиям в восприятии ценовой динамики между теми, у кого есть сбережения, и теми, у кого их нет.

6. Дискуссия

В качестве потенциального ограничения полученных результатов мы можем декларировать небольшой размер имеющегося в нашем распоряжении объема накопленных наблюдений об инфляционных ожиданиях и наблюдаемой инфляции. Опросы ФОМ регулярно проводятся в России лишь с 2014 года. Мы учли эти ограничения при моделировании, но более длинные ряды данных и получение данных по отдельным респондентам могли бы улучшить качество полученных оценок.

В качестве направлений для будущих исследований мы считаем возможным провести аналогичные расчеты для развитых стран, а также сравнить вклад

новостных индексов и классических товаров-маркеров в формирование восприятия цен населением.

Кроме того, мы считаем перспективным более подробно изучить то, как влияет информационное присутствие Банка России в СМИ на восприятие инфляции населением, в том числе с точки зрения значения динамики общей финансовой грамотности.

Заключение

В настоящем исследовании мы предложили новый подход для определения новостей, которые могут оказывать наибольшее влияние на формирование восприятия цен населением. Мы подтвердили выводы прошлых исследователей о том, что информация СМИ оказывает существенное влияние на инфляционные ожидания домохозяйств. При этом мы дополнили эту литературу, построив модели оценки важности признаков новостных тем на исключительно негативных новостях.

Мы создали базу новостей из 28 ведущих российских СМИ за период 2014–2022 гг., в которую вошло 7,8 млн новостей. Мы провели два этапа тематического моделирования. На первом были отсеяны все сообщения, которые не относятся к экономической тематике. На втором мы выделили частотность отдельных новостных потоков. Релевантные негативные новостные потоки были определены с помощью регулярного опроса ВЦИОМ о страхах россиян. Так, мы получили 10 временных рядов негативных новостей: бедность, инфляция, продовольственные проблемы, закредитованность, экономический кризис, геополитика, курс рубля, пандемия COVID-19, внутренняя нестабильность, безработица (названия тем определили эксперты на основе вероятностного распределения слов внутри каждой темы). Дополнительно мы добавили ряд с новостями, отражающий интенсивность присутствия Банка России в новостном потоке.

Далее мы использовали четыре модели, позволяющие оценить важность отдельных независимых переменных (Lasso, Random Forest, XGBoost) или построить ориентированный граф причинно-следственных связей (Bayesian Structure Learning). Результаты свидетельствуют о том, что есть три основные темы в новостях СМИ, которые вносят наибольший вклад в формирование восприятия цен населением: это собственно новости об инфляции, а также об экономическом кризисе и курсе рубля. При этом есть существенные различия в формировании представлений о будущей и прошлой инфляции. Отвечая на вопрос о будущей инфляции, респонденты могут в большей мере ориентироваться на прогноз об экономике в целом (в их представлении он может быть скорее связан с внешними факторами и курсом рубля), а при ответе на вопрос о прошлой инфляции – вспоминать новости о ценах, а также уровне жизни в стране в целом (новости об уровне бедности, внутренней обстановке в стране, продовольственной безопасности).

Полученные результаты робастны к двум видам проверок: зашумления данных и использования альтернативных методов оценки зависимых переменных.

Мы сообщаем также о нескольких дополнительных результатах.

Первое. Негативные новости СМИ, связанные с экономической ситуацией в широком плане, могут выступать значимым предиктором будущей динамики инфляционных ожиданий населения и воспринимаемой им ценовой картины на коротком горизонте в 1 месяц.

Второе. Существуют различия в восприятии инфляции между подгруппами населения со сбережениями и без них. Прежде всего отличия характерны для восприятия подгруппы без сбережений. Согласно оценкам моделей, такие граждане склонны обращать большее внимание на новости о геополитике и уровне бедности. Для подгруппы со сбережениями в среднем более значимым фактором выступает курсовая динамика.

Третье. Мы приводим свидетельства в пользу того, что на восприятие населением инфляции большое влияние может оказывать их ощущение экономической ситуации в стране в целом. Прямое сравнение влияния новостных индексов и товаров-маркеров может стать важной темой для будущих исследований.

Четвертое. Интенсивность информационного присутствия Банка России в СМИ, по всей видимости, не оказывает существенного влияния на восприятие инфляции населением.

Наша работа вносит вклад в литературу в следующих областях:

1) Мы привели новые свидетельства о высокой роли новостей в формировании восприятия инфляции населением. Этот опыт может быть учтен будущими исследователями при разработке новых моделей, базирующихся на больших данных новостей при прогнозировании макроэкономических переменных.

2) Мы создали открытую базу российских новостей, которая может быть использована для различных NLP-исследований для России.

3) Мы выделили отдельные темы негативных новостей, которые могут иметь доминирующее влияние на восприятие инфляции. Среди них оказались новости об инфляции и отдельных ценах, курсовой динамике и экономической ситуации в целом. Именно эти темы могут обладать наибольшей «эпидемиологической силой», которая легче преодолевает невнимание людей к событиям в экономике.

Через эти результаты мы делаем вклад в лучшее понимание природы инфляционных ожиданий населения в России.

Список литературы

1. Ahmed, N.K., Atiya, A.F., Gayar, N.E., and El-Shishiny, H., 2010. 'An empirical comparison of machine learning models for time series forecasting'. Econometric reviews, 29(5–6), pp. 594–621.

2. Angelico, C., Marcucci, J., Miccoli, M. and Quarta, F., 2022. Can we measure inflation expectations using Twitter? Journal of Econometrics, 228(2), pp. 259–277.

3. Arora, V., Gomis-Porqueras, P. and Shi, S., 2013. The divergence between core and headline inflation: Implications for consumers' inflation expectations. Journal of Macroeconomics, 38, pp. 497–504.

4. Berk, J.M., 1999. Measuring inflation expectations: a survey data approach. Applied Economics, 31(11), pp. 1467–1480.

5. Bird, S., Klein, E. & Loper, E., 2009. Natural language processing with Python: analyzing text with the natural language toolkit, "O'Reilly Media, Inc."

6. Blei, D.M., Ng, A.Y. and Jordan, M.I., 2003. Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), pp. 993–1022.

7. Boser, Bernhard E., Isabelle M. Guyon, and Vladimir N. Vapnik, 1992. "A training algorithm for optimal margin classifiers". Proceedings of the fifth annual workshop on Computational learning theory.

8. Box, G.E.P. and Cox, D.R., 1964. An analysis of transformations, Journal of the Royal Statistical Society, Series B, 26, pp. 211–252.

9. Breiman, L., 2001. Random forests. Machine learning, 45(1), pp. 5–32.

10. Carlson, J.A. and Parkin, M., 1975. Inflation expectations. Economica, 42(166), pp. 123–138.

11. Carroll, C.D., 2001. The epidemiology of macroeconomic expectations.

12. Carroll, C.D., 2003. Macroeconomic expectations of households and professional forecasters. the Quarterly Journal of economics, 118(1), pp. 269–298.

13. Cavallo, A., Cruces, G. and Perez-Truglia, R., 2014. Inflation expectations, learning and supermarket prices (No. w20576). National Bureau of Economic Research.

14. Coibion, O., Georgarakos, D., Gorodnichenko, Y. and Van Rooij, 2019. How does consumption respond to news about inflation? Field evidence from a randomized control trial (No. w26106). National Bureau of Economic Research.

15. Chen, T. & Guestrin, C., 2016. XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 785–794).

16. Maksym Gabielkov, Arthi Ramachandran, Augustin Chaintreau, Arnaud Legout. Social Clicks: Whatand Who Gets Read on Twitter? ACM SIGMETRICS / IFIP Performance 2016, Jun 2016, AntibesJuan-les-Pins, France. hal-01281190.

17. Efron, B., 1979. Computers and the theory of statistics: thinking the unthinkable. SIAM review, 21(4), pp. 460–480.

18. Del Giovane, P. and Sabbatini, R., 2004. La divergenza tra inflazione rilevata e percepita in Italia.

19. Jagarlamudi, J., Daumé III, H. and Udupa, R., 2012. Incorporating lexical priors into topic models. In Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, pp. 204–213.

20. Haldane, A., 2017. A little more conversation, a little less action. Bank of England-Speech.

21. Harris, Z. S., 1954. Distributional structure. Word, 10(2-3), pp. 146–162.

22. Kim, K.J., 2003. Financial time series forecasting using support vector machines. Neurocomputing, 55(1-2), pp. 307–319.

23. Kondratyev, A., 2018. Learning curve dynamics with artificial neural networks. Available at SSRN 3041232.

24. Larsen, V.H., Thorsrud, L.A. and Zhulanova, J., 2021. News-driven inflation expectations and information rigidities. Journal of Monetary Economics, 117, pp. 507–520.

25. Lundberg, S.M. & Lee, S.I., 2017. A unified approach to interpreting model predictions. Advances in neural information processing systems, 30.

26. Lyziak, T., 2010. Measurement of perceived and expected inflation on the basis of consumer survey data (No. 5). Bank for International Settlements.

27. Maćkowiak, B., Matějka, F. and Wiederholt, M., 2021. Rational inattention: A review.

28. Mankiw, N.G., Reis, R. and Wolfers, J., 2003. Disagreement about inflation expectations. NBER macroeconomics annual, 18, pp. 209–248.

29. Margaritis D., 2003. Learning Bayesian Network Model Structure from Data. Ph.D. thesis, School of Computer Science, Carnegie-Mellon University, Pittsburgh, PA. Available as Technical Report CMU-CS-03–153.

30. Martin, F. and Johnson, M., 2015, December. More efficient topic modelling through a noun only approach. In Proceedings of the Australasian Language Technology Association Workshop 2015, pp. 111–115.

31. Pearl, J., 1988. Probabilistic reasoning in intelligent systems: networks of plausible inference. Morgan kaufmann.

32. Petelin, G., Cenikj, G. and Eftimov, T., 2022. Towards understanding the importance of time-series features in automated algorithm performance prediction. Expert Systems with Applications, 119023.

33. Pfajfar, D. and Santoro, E., 2013. News on inflation and the epidemiology of inflation expectations. Journal of Money, Credit and Banking, 45(6), pp. 1045–1067.

34. Rehurek, R. & Sojka, P., 2011. Gensim–python framework for vector space modelling. NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic, 3(2).

35. Sahu, S. and Chattopadhyay, S., 2020. Epidemiology of inflation expectations and internet search: an analysis for India. Journal of Economic Interaction and Coordination, 15(3), pp. 649–671.

36. Salton, G. & Buckley, C., 1988. Term-weighting approaches in automatic text retrieval. Information processing & management, 24(5), pp. 513–523.

37. Sedova A. & Mitrophanova P., 2018. Topic Modelling of Russian Texts based on Lemmata and Lexical Constructions. Computational Linguistics and Computational Ontologies. 132-144. 10.17586/2541-9781-2017-1-132-144.

38. Scutari, M., 2009. Learning Bayesian networks with the bnlearn R package. arXiv preprint arXiv:0908.3817.

39. Sims, C.A., 2003. Implications of rational inattention. Journal of monetary Economics, 50(3), pp. 665–690.

40. Sussman, N. and Zohar, O., 2022. Have Inflation Expectations Become Unanchored? The Role of Oil Prices and Global Aggregate Demand. International Journal of Central Banking, 18(2), pp. 149–192.

41. Spirtes, P. and Glymour, C., 1991. An algorithm for fast recovery of sparse causal graphs. Social science computer review, 9(1), pp. 62–72.

42. Tibshirani, R., 1996. Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58(1), pp. 267–288.

43. Tilly, S. and Livan, G., 2021. Predicting market inflation expectations with news topics and sentiment. arXiv preprint arXiv:2107.07155.

44. Tsamardinos, I., Aliferis, C.F., Statnikov, A.R. and Statnikov, E., 2003, May. Algorithms for large scale Markov blanket discovery. In FLAIRS conference (Vol. 2, pp. 376–380).

Приложения

Приложение 1. База новостей российских СМИ и тематическое моделирование

Табл. 3. Состав базы новостей

Название СМИ	Покрытие	Количество включенных новостей
«Коммерсант»	федеральный	334 754
«Аргументы и факты»	федеральный	18 957
E1.ru	региональный	117 911
«Ведомости»	федеральный	58 536
«Утро.ру»	федеральный	182 579
«Московский комсомолец»	федеральный	1 784 184
Forbes	федеральный	28 090
«Банки.ру»	федеральный	183 399
BBC Russia	федеральный	80 841
Интерфакс	федеральный	86 913
«Лента.ру»	федеральный	593 999
«Финам.ру»	федеральный	74 907
Деловая газета «Взгляд»	федеральный	492 891
«Эхо Москвы»	федеральный	562 774
«Комсомольская правда»	федеральный	414 135
«Парламентская газета»	федеральный	252 769
«Медуза»	федеральный	96 575
Регнум	федеральный	44 204
«Pravda.ru»	федеральный	236 773
«Republic.ru»	федеральный	45 491
360tv.ru	федеральный	537 391
Svpressa.ru	федеральный	205 910
Finanz.ru	федеральный	272 249
«Российская газета»	федеральный	586 372
Svoboda.org	федеральный	95 596
«Финмаркет»	федеральный	11 368
«Фонтанка»	региональный	363 596
РБК	федеральный	16 106

Табл. 4. Агрегирование ключевых слов в темы

Подтемы	**Метка темы**	**Количество новостей**	**Количество новостей в теме**
Экономика	0	298 343	423 277
Экономика и финансы	0	5248	
Бизнес	0	83 082	
Финансы	0	19 297	
Строительство	0	6263	
Недвижимость	0	8138	
Финансы и инвестиции	0	2906	
Политика	1	400 108	807 252
Международная политика	1	10 015	
Политика в России	1	6059	
В мире	1	211 104	
Мир	1	161 736	
Новости мир	1	18 230	
Общество	2	901 518	922 131
Новости – общество	2	11 931	
Государство и общество	2	8682	
Культура	3	115 380	115 380
Происшествия	4	663 231	663 231
Спорт	5	220 919	220 919

Табл. 5. Результаты обучения моделей с учителем для решения задачи классификации новостей

Метрики качества моделей		**2 класса**	**6 классов**			
		XGBoost		**SVM**	**Logistic regression (softmax)**	**Random Forest**
Экономические новости	precision	0.9381	0.9363	0.9267	0.8658	0.8886
	recall	0.9064	0.9482	0.9548	0.9249	0.9240
	F-score	0.9219	0.9422	0.9405	0.8944	0.9060
Прочие	precision	0.9598	0.9388	0.9344	0.8757	0.9006
	recall	0.9472	0.9390	0.9347	0.8765	0.9013
	F-score	0.9534	0.9389	0.9345	0.8758	0.9006

Приложение 2. Качество полученных моделей

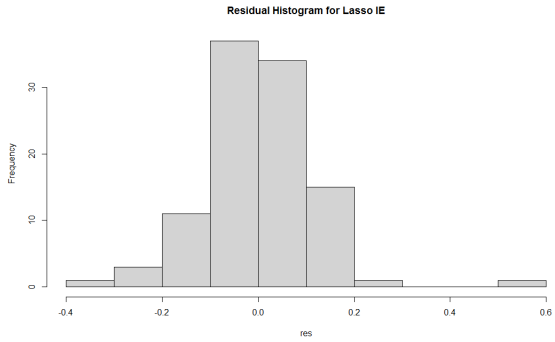
Табл. 6. Тесты Дики – Фуллера для временных рядов в моделях Lasso и BSL

Временной ряд	**Статистика Дики – Фуллера**	**p-значение**
Бедность	-6,1192	<0,01
Инфляция	-3,5895	0,03751
Продовольственные проблемы	-3,9019	0,01665
Закредитованность	-4,8159	<0,01
Экономический кризис	-4,0759	<0,01
Геополитика	-4,2891	<0,01
Курс рубля	-4,0910	<0,01
Пандемия	-3,5200	0,04372
Внутренняя нестабильность	-5,6691	<0,01
Безработица	-3,9125	0,01614
Инфляционные ожидания	-5,0091	<0,01
Наблюдаемая инфляция	-4,2700	<0,01

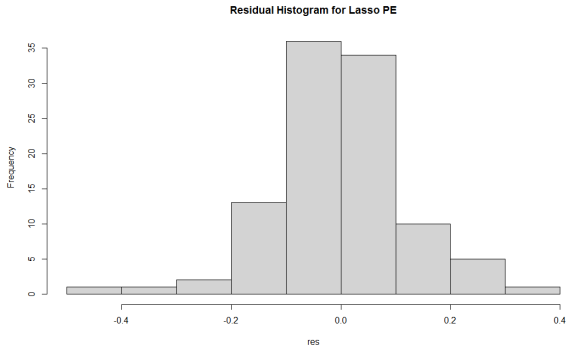
Рис. 15–20. Анализ остатков моделей

1.) Lasso

а) инфляционные ожидания



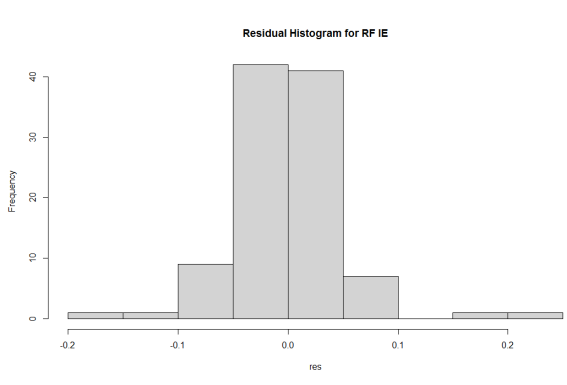
б) наблюдаемая инфляция

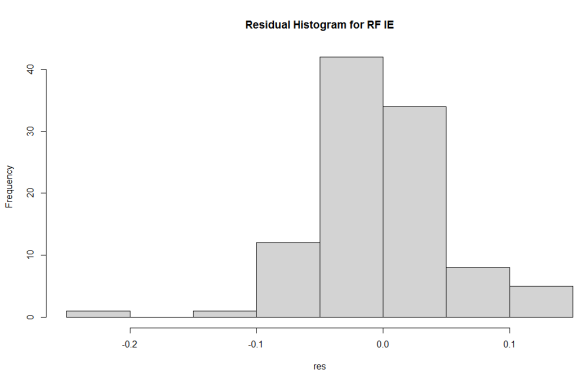


2.) Random Forest

а) инфляционные ожидания

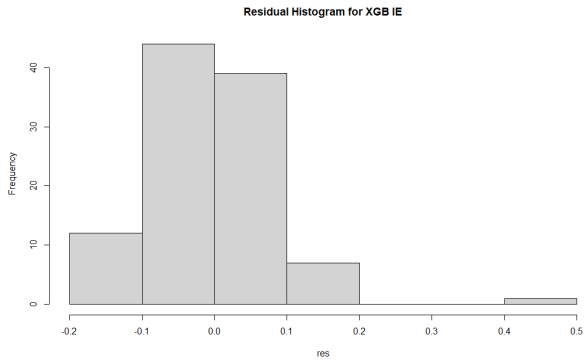
б) наблюдаемая инфляция



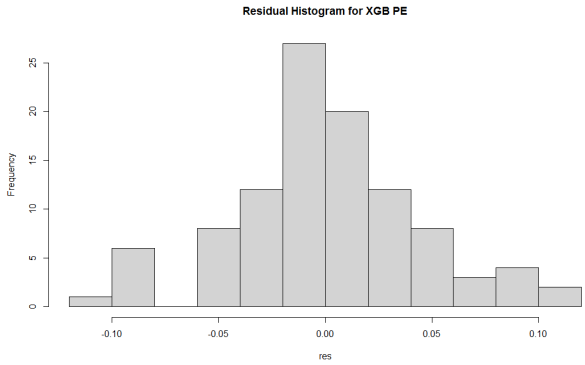


3.) XGBoost

а) инфляционные ожидания



б) наблюдаемая инфляция



Приложение 3. Проверки робастности результатов

Табл. 7. Альтернативные модели с зашумлением

Новостной индекс	Инфляционные ожидания				Наблюдаемая инфляция			
	Lasso	RF	XGBoost	BSL	Lasso	RF	XGBoost	BSL
Бедность	0,00	0,02	0,05	False	0,18	0,05	0,09	False
Инфляция	0,11	0,24	0,34	False	0,20	0,21	0,30	True
Продовольственные проблемы	0,00	0,01	0,02	False	0,01	0,01	0,00	False
Закредитованность	0,00	0,02	0,01	False	0,00	0,03	0,02	False
Экономический кризис	0,42	0,02	0,05	False	0,18	0,15	0,10	False
Геополитика	0,00	0,03	0,00	False	0,13	0,07	0,07	False
Курс рубля	0,46	0,40	0,26	True	0,00	0,30	0,29	False
Пандемия COVID-19	0,00	0,16	0,25	False	0,05	0,08	0,12	False
Внутренняя нестабильность	0,00	0,02	0,01	False	0,00	0,02	0,01	True
Безработица	0,00	0,05	0,01	False	0,14	0,02	0,00	False
Коммуникация БР	0,01	0,01	0,00	False	0,12	0,01	0,00	False
Шум 1	0,00	0,01	0,00	False	0,00	0,01	0,00	False
Шум 2	0,00	0,01	0,00	False	0,00	0,01	0,00	False
Шум 3	0,00	0,01	0,00	False	0,00	0,01	0,00	False
Adjusted R^2	0,17	0,72	0,83		0,43	0,72	0,66	

Табл. 8. Модели с альтернативными переменными для ожидаемой инфляции

Новостной индекс	Инфляционные ожидания респондентов со сбережениями				Инфляционные ожидания респондентов без сбережений			
	Lasso	RF	XGBoost	BSL	Lasso	RF	XGBoost	BSL
Бедность	0,07	0,09	0,05	False	0,00	0,04	0,06	False
Инфляция	0,13	0,15	0,25	False	0,27	0,34	0,38	True
Продовольственные проблемы	0,00	0,02	0,02	False	0,00	0,01	0,00	False
Закредитованность	0,00	0,03	0,03	False	0,00	0,01	0,01	False
Экономический кризис	0,20	0,08	0,02	False	0,69	0,02	0,05	False
Геополитика	0,00	0,02	0,00	False	0,00	0,01	0,00	False
Курс рубля	0,49	0,27	0,24	True	0,00	0,31	0,20	False
Пандемия COVID-19	0,00	0,18	0,30	False	0,00	0,16	0,26	False
Внутренняя нестабильность	0,00	0,04	0,04	False	0,00	0,01	0,02	True
Безработица	0,00	0,10	0,03	False	0,00	0,07	0,03	False
Коммуникация БР	0,11	0,01	0,03	False	0,04	0,01	0,00	False
Adjusted R^2	0,25	0,56	0,72		0,19	0,70	0,83	

Табл. 9. Модели с альтернативными переменными для наблюдаемой инфляции

Новостной индекс	**Наблюдаемая инфляция респондентов со сбережениями**				**Наблюдаемая инфляция респондентов без сбережений**			
	Lasso	**RF**	**XGBoost**	**BSL**	**Lasso**	**RF**	**XGBoost**	**BSL**
Бедность	0,30	0,08	0,08	False	0,15	0,06	0,14	False
Инфляция	0,12	0,13	0,24	True	0,20	0,21	0,33	True
Продовольственные проблемы	0,00	0,01	0,00	False	0,00	0,01	0,00	False
Закредитованность	0,00	0,08	0,02	False	0,00	0,01	0,03	False
Экономический кризис	0,20	0,15	0,10	False	0,23	0,23	0,14	False
Геополитика	0,12	0,04	0,05	False	0,15	0,15	0,04	False
Курс рубля	0,00	0,28	0,33	False	0,00	0,21	0,24	False
Пандемия COVID-19	0,00	0,10	0,15	False	0,05	0,07	0,07	False
Внутренняя нестабильность	0,00	0,03	0,02	False	0,00	0,02	0,02	False
Безработица	0,15	0,05	0,00	False	0,09	0,02	0,00	False
Коммуникация БР	0,12	0,06	0,00	False	0,13	0,01	0,00	False
Adjusted R^2	0,33	0,74	0,60		0,46	0,77	0,64	