Использование методов машинного обучения для прогнозирования инвестиций в России*

Михаил Гареев

Российская Академия Народного Хозяйства и Государственной Службы, mkhlgrv@gmail.com

Аннотация

В работе построены прогнозы темпов роста квартального валового накопления основного капитала в России с помощью методов машинного обучения (методы регуляризации, ансамблевые методы) на горизонте до 8 кварталов. Тестируемые методы показывают качество в терминах RMSFE выше, чем у простых альтернативных моделей (модель авторегрессии, модель случайного блуждания), причем лидерами оказываются ансамблевые методы (бустинг и случайный лес). Последнее согласуется с результатами других работ по применению больших данных в макроэкономике. Получено, что удаление из выборки наблюдений, которые относятся ко времени до кризиса 1998 г., нетипичных для последующего периода времени, не ухудшает краткосрочные прогнозы методов машинного обучения. Оценки коэффициентов общепринятых ключевых факторов инвестиций, полученные с помощью методов регуляризации, в целом согласуются с экономической теорией. Прогнозы моделей автора превосходят по качеству годовые прогнозы темпов роста валового накопления основного капитала, публикуемые Министерством экономического развития.

Ключевые слова: прогноз инвестиций, машинное обучение, лассо, бустинг, случайный лес. **JEL Codes:** C53, E22.

Abstract

The work forecasts the growth rate of quarterly gross investment in Russia using machine learning methods (regularization methods, ensemble methods) over a horizon of up to 8 quarters. The tested methods show quality in terms of RMSFE higher than that of simple alternative models (autoregressive model, random walk model), ensemble methods (boosting and random forest) get the best score. The last statement is consistent with the results of other works on the application of big data in macroeconomics. The data obtained indicate that the time before the 1998 crisis did not worsen, and forecasts for machine learning methods are not reduced. Estimates of the

*Автор выражает благодарность А.В. Полбину и анонимным рецензентам за полезные правки и замечания

coefficients of generally accepted key investment factors, results using regularization methods, in general, is consistent with economic theory.

It was found that the removal of observations from the sample that belong to the time before the 1998 crisis and are atypical for a subsequent period of time does not worsen short-term forecasts of machine learning methods. Regulariastion methods estimates of the coefficients of common accepted key investment factors generally correspond with investment theory. The forecasts of the author's models are superior in quality to the annual forecasts of investment growth rates published by the Ministry of Economic Development.

Введение

В настоящее время доступны огромные массивы макроэкономических и финансовых данных, и многие прикладные экономические исследования сводятся к извлечению из них полезной информации. Существует обширная литература по применению различных методов работы с большими объёмами данных в макроэкономике. Разработаны модели, позволяющие извлекать информацию из сотен переменных и бороться с проблемой переобучения и так называемым «проклятием размерности» (см. работу Stock и Watson (2011)). Однако результаты многих популярных методы машинного обучения зачастую сложно или невозможно интерпретировать. Исследователям необходимо понимать, насколько прогнозные модели адекватны и соответствуют экономической теории, поэтому может быть оправдано использование регуляризации, то есть наложения штрафов за неадекватно высокие коэффициенты при предикторах. Методы регуляризации, такие как LASSO (англ. Least Absolute Shrinkage and Selecting Operator) или Ridge, а также их модификации, с одной стороны, несколько уменьшают риск переобучения моделей, а с другой стороны, сохраняют интерпретируемость. Однако зачастую в прикладных исследованиях именно слабо интерпретируемые методы (например, ансамблевые — бустинг, случайный лес) показывают наивысшее качество прогнозов.

Инвестиции являются одним из важнейших факторов долгосрочного роста и часто обсуждаются экономистами. В работе Кудрина и Гурвича (2014) отмечалось снижение инвестиционной активности в России в 2007–2013 гг., ограничение притока иностранного капитала и импорта технологий в связи с событиями в Крыму и на Юго-Востоке Украины. Замедление динамики инвестиций ведёт к отставанию от остального мира. Недостаточная инвестиционная привлекательность российской экономики, главным образом вызванная, по мнению авторов статьи, слабостью рыночных механизмов, является основным ограничением для устойчивого роста. В работе Орешкина (2018) отмечается, что при существующих демографических ограничениях единственной возможностью повысить темпы роста экономики является увеличение объёма и качества инвестиций. По оценкам Министерства экономического развития целевой темп роста ВВП в $3,0$ — $3,7\%$ может быть достигнут при увеличении доли инвестиций в ВВП до 25 — 30%, причем основным фактором этой трансформации должно быть повышение сбережений домохозяйств. В статье Идрисова и Синельникова-Мурылева (2014) уделяется внимание институциональным факторам улучшения инвестиционного климата в России. В числе возможных областей ускоренного роста инвестиций называются сфера высоких технологий (Аганбегян, 2016) и инфраструктура (Орешкин, 2018).

Ясно, что, как и на любой экономический показатель, на инвестиции влияют многие факторы, и, поскольку принципиально невозможно учесть все из них, зачастую для прогнозирования могут использоваться относительно простые модели (например, акселераторная модель, подразумевающая, что уровень инвестиций определяется текущим уровнем выпуска и его лагами). Однако, возможно, из обширных данных макроэкономической статистики получится извлечь информацию, которая поможет построить относительно стабильные прогнозы темпов роста инвестиций. Автор надеется, что в какой-то мере данная работа, исследующая применение методов машинного обучения для прогнозирования темпов роста валового накопления основного капитала на период до 8 кварталов, восполнит пробел, существующий в этой области.

Существует множество примеров успешного использования машинного обучения при макроэкономическом прогнозировании. Так, в работе Li и Chen (2014) авторы исследовали возможности LASSO, Elastic Net и Group LASSOв области макроэкономического прогнозирования в сравнении с динамическими факторными моделями (ДФМ). Данные включали 107 различных месячных показателей американской экономики с 1959 г. по 2008 г., и авторы подробно рассматривали результаты для 20 самых важных показателей. Было получено, что при прогнозах на один шаг вперёд методы регуляризации показывали в среднем лучшее качество (в смысле MSFE), чем динамические факторные модели для 18, 15 и 19 переменных из 20 для LASSO, Elastic Net и Group LASSO. Кроме этого, комбинация методов регуляризации и ДФМ для каждого из 20 показателей давала более качественные прогнозы, чем ДФМ.

В статье Bai и Ng (2008) авторы среди прочего показывают, что применение методов регуляризации может быть полезно для определения факторов, влияющих на инфляцию в США. Модели, допускающие изменение набора объясняющих переменных с течением времени (переменные отбирались в том числе с помощью LASSO), показывали в среднем качество лучше, чем модели с фиксированным набором факторов.

В работе Байбузы (2018) исследовались возможности методов работы с большими данными для прогнозирования инфляции в России. Автор использовал в качестве предикторов 92 макроэкономических показателя с 2012 г. по 2016 г. и строил вневыборочные прогнозы инфляции на горизонте от 1 до 24 месяцев вперёд. Помимо методов регуляризации, также рассматривались ансамблевые методы машинного обучения (случайный лес и бустинг). Модели с регуляризацией показали достаточно плохие прогнозы относительно бенчмарков (модель случайного блуждания, AR(1) и AR(p)) и ансамблевых методов, однако комбинанция AR(1) и LASSO показывала лучшее качество при прогнозе на 1 месяц вперёд.

Статья Фокина и Полбина (2019) посвящена совмещению векторной авторегрессии и LASSO-регуляризации (модель VAR-LASSO) для моделирования основных показателей российской экономики: ВВП, потребления, инвестиций в основной капитал с учётом экзогенного шока цены на нефть. По результатам тестирования модель авторов показала хорошие прогнозные свойства по сравнению с обычной моделью VAR, а также прогнозами Министерства экономического развития. Кроме того, были построены функции импульсного отклика на шок изменения цены на нефть.

Далее работа построена следующим образом. В разделе 1 содержится методология исследования. Во-первых, приведено описание используемых методов машинного обучения (Ridge, LASSO, Post-LASSO, Adaptive LASSO, Elastic Net, Spike and Slab, случайный лес, бустинг) и альтернативных методов прогнозирования, используемых для сравнения качества. Во-вторых, описаны использованные данные, способы их трансформации и методы построения прогнозов. В разделе 2 содержатся эмпирические результаты исследования и их обсуждение. Автором разработано веб-приложение[1], которое позволяет как воспроизвести результаты работы, так и задать собственную спецификацию моделей относительно границ тренировочной выборки и горизонтов прогнозирования, а также построить прогнозы темпов роста инвестиций.

[1]Приложение доступно по ссылке https://mkhlgrv.shinyapps.io/investment_forecasting/

1 Методология

Можно выделить несколько групп методов работы с большими объёмами данных, используемых для прогнозирования макроэкономических переменных, которые отличаются разным подходом к решению проблемы «проклятия размерности», являющейся основным препятствием для использования большого количества переменных.

Один из подходов к работе с большим количеством предикторов состоит в использовании методов регуляризации. Как было отмечено выше, их идея состоит в том, чтобы при оценке параметров использовать функцию штрафа за увеличение коэффициентов. Наиболее популярны и исследованы в литературе два виды регуляризаторов (штрафных функций) - LASSO, или l_1 регуляризатор (Tibshirani, 1996), и Ridge (регрессия гребня, регуляризатор Тихонова, или l_2 регуляризатор) (Hoerl и Kennard, 1970).

Существует огромное множество и слабо интерпретируемых методов машинного обучения. В данной работе подробно рассмотрены ансамблевые методы — случайный лес и бустинг.

1.1 Методы регуляризации

Пусть задана стандартная линейная модель регрессии:

$$y_i = x_i^{'}\beta + \varepsilon_i, \tag{1}$$

где для наблюдения $i = 1, \dots, n$ y_i — это значения объясняемой переменной, $x_i \in \mathbb{R}^p$ — это значения p объясняющих переменных, $\beta \in \mathbb{R}^p$ — это вектор из p коэффициентов, $\varepsilon_i \sim N(0, \sigma^2)$ — это независимые и одинаково распределённые случайные ошибки. Все переменные стандартизированы, то есть имеют нулевое математическое ожидание и единичную дисперсию. В матричном виде можно записать модель следующим образом:

$$Y = X\beta + \mathcal{E}, \tag{2}$$

где $Y \in \mathbb{R}^n$ — это значения объясняемой переменной, $X \in \mathbb{R}^{n \times p}$ — это матрица значений объясняющих переменных, $\beta \in \mathbb{R}^p$ — это вектор коэффициентов, $\mathcal{E} \in \mathbb{R}^n$ — это вектор независимых и одинаково распределенных случайных ошибок.

Модель называется разреженной линейной моделью с высокой размерностью (Belloni и Chernozhukov, 2011a; Belloni и Chernozhukov, 2011b) в случае, если возможно, что $p \geq n$, но при этом только $s < n$ элементов вектора β не равны нулю. Высокая размерность означает, что оценивание модели стандартным МНК может либо приводить к нестабильным оценкам коэффициентов, либо же вовсе невозможно, если $p \geq n$.

Если линейная модель имеет большую размерность, было бы полезно иметь способ оценки коэффициентов с меньшей, чем при МНК, дисперсией (пусть даже и смещённых). При этом, если модель является разреженной, то, помимо оценки коэффициентов, также полезно было бы каким-то образом идентифицировать и отбрасывать нулевые элементы вектора β. Одним из возможных способов решений этих задач является использование методов регуляризации. Общая идея таких методов — это введение некоторого штрафа (оператора регуляризации), который бы препятствовал переобучению, вызванному неоправданно высокими оценками коэффициентов, смещая их к нулю.

1.1.1 Ridge

МНК-оценки для моделей с высокой размерностью обычно имеют низкое смещение, но высокую дисперсию. Регрессия Ridge позволяет уменьшать дисперсию оценок коэффициентов, однако, делая их смещёнными. Формально задача выглядит следующим образом:

$$\hat{\beta}^{\text{Ridge}} \in \arg\min_{\beta \in \mathbb{R}^p} \lambda \left\| Y - X\beta \right\|_2^2 + \lambda \left\| \beta \right\|_2^2. \tag{3}$$

К стандартному минимизируемому функционалу наименьших квадратов добавляется функция штрафа, которая состоит из l_2-нормы вектора β, умноженной на штрафной параметр λ. С ростом λ оценки коэффициентов становятся все ближе к нулю, а при $\lambda = 0$ задача сводится к обычному МНК. Важным свойством такого подхода является наличие аналитического решения:

$$\hat{\beta}^{\text{Ridge}} = \left(X'X + \lambda I_p \right)^{-1} X'Y. \tag{4}$$

В случае, если $rank(X) < p$, МНК не имеет решения, так как матрица $X'X$ необратима. Однако решение существует для модели Ridge в случае, если $\lambda \neq 0$.

Недостатком этого метода является то, что он не может отбирать ненулевые коэффициенты в разреженной модели.

1.1.2 LASSO

Метод LASSO был популяризован после работы Tibshirani (1996), однако и до этого встречался в литературе (Santosa и Symes, 1986). Оценка на основе LASSO выглядит следующим образом:

$$\hat{\beta}^{\text{LASSO}} \in \arg\min_{\beta \in \mathbb{R}^p} \left\| Y - X\beta \right\|_2^2 + \lambda \left\| \beta \right\|_1. \tag{5}$$

В этом случае в качестве оператора регуляризации выступает сумма абсолютных значений коэффициентов модели, умноженная на параметр λ. Так же, как и в Ridge, при достаточно малых значениях λ размер штрафа незначителен и результаты оценки похожи на результаты стандартной линейной регрессии, однако при достаточно больших значениях λ в модели вовсе не оказывается объясняющих переменных. Использование LASSO, в отличии от Ridge, позволяет выбирать из общего набора переменных лишь несколько наиболее важных и отбрасывать остальные. При этом аналитического решения модели уже не существует.

Методы регуляризации позволяют в явном виде получать оценки коэффициентов при предикторах. Однако можно ли их корректно интерпретировать? Хорошо известно, что коэффициенты, которые оцениваются обычным LASSO, часто нестабильны во времени и при добавлении новых наблюдений могут резко меняться. Это подтверждается, например, в работах Zou и Hastie (2005) и De Mol, Giannone и Reichlin (2008). Существуют несколько модификаций классического LASSO, которые призваны улучшить его статистические свойства, например, Post-LASSO и Adaptive LASSO.

1.1.3 Post-LASSO

В общем случае оценки методов регуляризации, будут смещены. В работе Belloni и Chernozhukov (2011a) показано, что потенциально менее смещённые, чем у LASSO, оценки могут быть получены при использовании метода Post-LASSO. Оператор LASSO позволяет убрать из рассмотрения лишние переменные. В этом случае естественным кажется после отбора переменных рассмотреть ещё дополнительно и обычную линейную регрессию, используя только те предикторы, коэффициенты при которых оказались не равны нулю. Таким образом мы получаем оценку Post-LASSO. Формально это можно записать следующим образом:

$$\hat{\beta}^{\text{Post-LASSO}} \in \arg\min_{\beta \in \mathbb{R}^p} \left\| Y - X\beta \right\|_2^2, \text{ где } \beta_j = 0, \text{ если } \hat{\beta}_j^{\text{LASSO}} = 0. \tag{6}$$

Такие оценки могут привести к меньшему смещению, однако, как с помощью симуляции показали Belloni и Chernozhukov (2011a), при высокой зашумленности данных модель становится нестабильной и оценки коэффициентов могут оказаться даже дальше от истинных, чем оценки LASSO.

1.1.4 Adaptive LASSO

В исследовании Zou (2006) показано, что в некоторых ситуациях LASSO может неверно исключать (то есть оценивать коэффициенты как ноль) переменные. В некоторых случаях значение параметра штрафа λ, показывающее наилучшее качество оценки, приводит к выбору «мусорных» переменных, вместе с этим также возможна ситуация, когда при правильном отборе переменных ненулевые коэффициенты оказываются неоправданно высоки, что приводит к относительно плохим прогнозам.

Из-за этого предлагается другая версия LASSO — Adaptive LASSO, в котором используется уже взвешенная функция штрафа. В этом случае при определённых предпосылках метод отбирает верные переменные, и, кроме того, можно говорить о состоятельности полученных таким образом оценок коэффициентов. Предложенное Zou (2006) усовершенствование состоит в предварительном взвешивании коэффициентов вектора β — составных частей оператора регуляризации $\frac{\lambda}{n} \left\| \beta \right\|_1$. Задача нахождения оценок вектора коэффициентов при помощи адаптивного LASSO формально описывается следующим образом:

$$\hat{\beta}^{\text{Adaptive LASSO}} \in \arg\min_{\beta \in \mathbb{R}^p} \left\| Y - X\beta \right\|_2^2 + \lambda \left| w' \beta \right|_1, \tag{7}$$

где $w \in \mathbb{R}^p$ — это вектор весов коэффициентов. В соответствии с результатами авторов, при правильном выборе весов оценки, полученные таким образом, оказываются состоятельными. Но каким образом выбирать веса? Можно воспользоваться следующей формулой:

$$w_j = \frac{1}{(|\beta_j^{\text{init}}|)^\gamma}, \tag{8}$$

где β_j^{init} — это первоначальная оценка коэффициента β_j, полученная, например, при помощи Ridge, γ — дополнительный параметр. С ростом параметра γ важность взвешивания штрафов повышается (при $\gamma = 0$ задача

сводится к обычному LASSO). Добавление весов позволяет предварительно указать оператору LASSO на те переменные, добавление которых нежелательно (чем меньше абсолютное первоначальной оценки коэффициента j, тем больше штраф за появление коэффициента в модели).

Можно было бы выбирать параметр γ с помощью кросс-валидации, однако это было бы не очень разумным при небольшом объёме доступной выборки. В соответствии с рекомендациями автора метода в данной работе используется значение $\gamma = 0,5$, а первоначальные веса рассчитываются из оценок Ridge.

1.1.5 Elastic Net

В работе Zou и Hastie (2005) была показана неустойчивость LASSO в выборе переменных, которая обусловлена неопределённостью параметров при оценке ковариационной матрицы. Для решения этой проблемы авторы предложили метод Elastic Net. Он обобщает LASSO и Ridge. Главные отличия этих двух методов состоят в следующем: LASSO позволяет занулять коэффициенты, зато Ridge даёт более стабильные оценки для высококоррелированных переменных, в то время как оценки LASSO могут сильно поменяться при добавлении новых наблюдений. При применении метода Elastic Net с одной стороны, возможно зануление коэффициентов (преимущество LASSO), но, с другой стороны, оценки коэффициентов могут получаться относительно стабильными (преимущество Ridge). Задача Elastic Net имеет следующий вид:

$$\hat{\beta}^{\text{Elastic Net}} \in \arg\min_{\beta \in \mathbb{R}^p} \left\| Y - X\beta \right\|_2^2 + \lambda(\alpha \left\| \beta \right\|_1 + (1-\alpha) \left\| \beta \right\|_2^2), \tag{9}$$

Оператор регуляризации состоит из взвешенной суммы операторов LASSO и Ridge. В такой постановке задачи появляется уже два параметра — λ, как и выше, отвечающих за важность штрафов, и α. При $\alpha = 1$ модель сводится к LASSO, а при $\alpha = 0$ — к Ridge. Параметр α может быть выбран с помощью кросс-валидации, однако это не очень оправдано при небольшой выборке, поэтому в данной работе его значение установлено на уровне $0,5$, равноудалённом от двух крайних случаев. Такое же значение было выбрано, например, в работе Байбузы (2018).

1.1.6 Spike and Slab

Байесовский подход в эконометрике предполагает наличие априорного распределения для каждого параметра модели, которое корректируется в соответствии с имеющимися данными и в результате приводится к апостериорному распределению. Такой подход, например, позволяет получать оценки параметров, даже если переменных больше, чем наблюдений, или напрямую ставить вопрос о вероятности того, что какой-то коэффициент в модели равен нулю. В данной работе используется байесовский метод Spike and Slab.

Важно отметить, что методы регуляризации могут быть описаны и через байесовский подход. При нормальном априорном распределении байесовская модель сводится к регрессии Ridge, а при двойном экспоненциальном распределении — к регрессии LASSO (см. работу De Mol, Giannone и Reichlin (2008)).

В общем виде представление модели имеет следующий вид. Пусть выполняется (2), при этом существуют следующие априорные представления о модели:

$$\begin{cases} (y_i|x_i, \beta, \sigma^2) \sim N(x_i^{'}\beta, \sigma^2), \\ (\beta|\gamma) \sim N(0, \Gamma), \\ \gamma \sim \pi(\cdot), \\ \sigma^2 \sim \mu(\cdot), \end{cases} \tag{10}$$

где $\sigma^2 > 0$ — дисперсия случайной ошибки, $\Gamma = \gamma \cdot I_k$, I_k — единичная матрица, $\pi(\cdot)$ и $\mu(\cdot)$ — априорные представления о распределениях соответствующих величин. Существуют различные версии априорных распределений.

Автором используется метод Spike and Slab в версии, изложенной в работе Ishwaran и Rao (2005), в соответствии с которой предполагается следующее:

$$\begin{cases} (\beta_j|\tau_j, \rho_j^2) \sim N(0, \tau_j \cdot \rho_j^2), \\ \tau_j \sim (1-w)\delta_{v_0}(\cdot) + w\delta_1(\cdot), \\ (\rho^{-2}|\alpha_1, \alpha_2) \sim \Gamma(\alpha_1, \alpha_2), \\ w \sim U[0;1], \end{cases} \tag{11}$$

где δ_x — дискретная мера, сконцентрированная в окрестности точки x, v_0 — число, близкое к нулю. Число v_0 и параметры Гамма-распределения α_1 и α_2 выбираются так, чтобы дисперсия j-го коэффициента γ_j имела пик в нуле и правосторонний хвост. В отличии от других, более ранних изложений модели, такая форма предполагает все используемые распределения непрерывными, а не задаёт их в виде кусочных функций, что повышает гибкость модели. В соответствии с этими предположениями производится стандартная для байесовского подхода минимизация апостериорного выборочного среднего ошибок.

1.2 Ансамблевые методы

1.2.1 Случайный лес

Случайный лес (англ. Random Forest) относится к ансамблевым методам. Общая идея ансамблевых методов — это использование на одной выборке большого количества «простых» методов регрессии или классификации и построение предсказанных значений на основе усреднённых предсказаний этих методов. В основе метода случайного леса, изложенного в работе Liaw и Wiener (2002), лежит построение решающих деревьев. Какова мотивация их использования? Линейные методы оценивания моделей обладают рядом достоинств: они могут быть быстро обучены, в них возможна работа с большим количеством признаков, их можно подвергнуть регуляризации. Вместе с тем, у них есть существенный недостаток — они могут оценивать только линейные зависимости между переменными (можно конечно, менять спецификацию модели и добавлять нелинейные компоненты, но такие преобразования должны иметь какое-либо обоснование, и, конечно, возможности этого ограничены). Решающие деревья позволяют в некоторой степени решить эту проблему. И, хотя первоначально они применялись для задач классификации (например, классическая задача, решаемая с использованием деревьев — это бинарная классификация потенциальных заёмщиков банком — вернёт заёмщик кредит или нет), их можно использовать и для задач регрессии. Однако сами по себе

деревья в настоящее время используются редко, зато их часто объединяют в композицию.

Прежде, чем перейти к изложению метода случайного леса, дадим формальное определение бинарного дерева решений. Пусть задан вектор объясняемой переменной Y и матрица объясняющих переменных X. Дерево состоит из внутренних и терминальных (листовых) вершин. Каждой внутренней вершине v приписана функция (предикат) $\beta_v : X \to \{0, 1\}$, а каждой терминальной вершине u приписан прогноз $y_u \in Y$. При этом задан алгоритм (X), который стартует из начальной вершины v_0 и переходит в левую вершину, если $\beta_{v_0}(X) = 0$ и в правую вершину, если $\beta_{v_0}(X) = 1$. Так происходит до тех пор, пока алгоритм не достигнет терминальной вершины, после чего делается прогноз объясняемой переменной. Как правило, один предикат использует только одну переменную из набора объясняющих переменных.

Каким образом строятся деревья? При построении каждой вершины переменная и её разделяющее значение выбирается таким образом, чтобы улучшение заранее заданного функционала качества было максимальным. В качестве такого функционала, например, может выступать сумма квадратов ошибок $\mathrm{RSS} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$, а в качестве прогнозного значения $\hat{y}_i$ можно использовать, среднее значение или медиану объясняемой переменной в подвыборке. В каждой вершине проверяется критерий остановки. Если он не выполнен, то вершина объявляется внутренней и разбиение продолжается, если выполнен — то вершина признаётся терминальной и ей приписывается прогноз $\hat{y}_i$. В качестве критерия остановки может использоваться максимальная глубина дерева, минимальное количество объектов в вершине, предельный прирост функционала качества или их комбинация.

Легко заметить, что деревья склонны к переобучению (например, в тривиальном случае можно построить дерево без ошибок на обучающей выборке, если в каждой терминальной вершине будет находится только по одному объекту). Случайный лес не так сильно подвержен этой проблеме. «Высаживание» решающих деревьев в случайный лес состоит из двух этапов:

1. Данные случайным образом разбиваются на N подвыборок с повторениями. Размер каждой подвыборки равен размеру всей тренировочной выборки, но часть наблюдений в ней не встречается, а часть — встречается несколько раз. На каждой подвыборке строится решающее дерево. При этом при построении дерева на одной подвыборке доступны только p_0 переменных, случайным образом выбираемых из p регрессоров отдельно для каждой подвыборки.

2. В качестве прогнозного значения $\hat{y}_i$ выбираются усреднённое значение $\hat{y}_{ij}$ по всем деревьям j $(j = 1, \dots, N)$.

Такой подход позволяет потенциально получить высокую предсказательную силу, при этом, как говорилось выше, использование множества деревьев в некотором смысле страхует модель от переобучения, но, вместе с тем, содержательная экономическая интерпретация результатов построения случайного леса тяжела или вовсе невозможна.

Кроме того, использование решающих деревьев имеет важное ограничение: предсказанные значения не могут выйти за пределы тех значений, которые наблюдались на тренировочной выборке, в то время как для линейных моделей такой проблемы не существует. Это ограничение может быть особен-

но существенным при прогнозировании макроэкономических или финансовых данных.

1.2.2 Бустинг

Градиентный бустинг (англ. boosting), предложенный в работе Friedman (2001), как и случайный лес, принадлежит к числу ансамблевых методов (то есть представляет собой композицию нескольких простых моделей одного типа). Однако, в отличии от случайного леса, процедура бустинга позволяет производить последовательное улучшение модели при использовании информации из предыдущей итерации обучения.

Бустинг фактически является некоторой общей методологией построения методов, поэтому в качестве базовой модели может использоваться почти любой вид модели, но часто, как и в случае со случайным лесом, базовая модель — это решающее дерево. Такой подход используется и в данной работе. В ходе бустинга сначала на всей тренировочной выборке обучается базовая модель M_1, после чего считаются ошибки модели и на них происходит тренировка новой модели m_2. Она прибавляется к предыдущей с коэффициентом $\eta \in (0, 1)$:

$$M_2 = M_1 + \eta m_2. \tag{12}$$

Параметр η отвечает за скорость обучения (при высокой скорости модель склонна переобучаться, а при низкой скорости для получения хороших предсказаний требуется большее число итераций).

Итоговым результатом для N итераций является модель $M_N = M_1 + \sum_{i=2}^{N} \eta^{i-1} m_i$.

1.3 Альтернативные модели

В этом подразделе описаны используемые в работе альтернативные модели, не относящиеся к методам машинного обучения.

1.3.1 Случайное блуждание

В качестве базовой модели используется простая модель случайного блуждания. Случайное блуждание — простой и понятный индикатор качества, часто используемый в прикладных экономических исследования. Эта модель предполагает, что изменения процесса во времени являются белым шумом. Формально прогноз на h шагов вперёд записывается следующим образом:

$$\hat{y}_{t+h} = y_t. \tag{13}$$

1.3.2 Авторегрессия

Ещё одна часто используемая модель для сравнения качества при прогнозировании временных рядов — модель авторегрессии порядка p AR(p). Она предполагает, что изменения процесса в момент времени t полностью зависят от p его предыдущих значений и случайной ошибки ε_t:

$$y_t = a_0 + a_1 y_{t-1} + a_2 y_{t-2} + \dots a_p y_{t-p} + \varepsilon_t. \tag{14}$$

На практике выбор количества лагов обычно производится при помощи информационных критериев — AIC или BIC.

1.4 Данные

В качестве прогнозируемой переменной используются темпы роста валового накопления основного капитала. Эта переменная построена из трёх рядов валового накопления основного капитала, предоставляемых Росстатом для разных периодов (с 1-го квартала 1995 г. по 4-ый квартал 2002 г., с 1-го квартала 2003 г. по 4-ый квартал 2010 г. и с 1-го квартала 2011 г. по 1-ый квартал 2019) путём взятия 1-ой разности логарифма по отношению к соответствующему кварталу предыдущего года. Такой подход позволяет избавиться от сезонности в ряде валового накопления основного капитала. Кроме того, он учитывает возможную мультипликативность и структурные сдвиги в сезонности. Проведённый расширенный тест Дики-Фулера не обнаруживает в трансформированном ряде нестационарности. После преобразования временной ряд уменьшается на 4 первых наблюдения и начинается с 1-го квартала 1996 г.

В качестве предикторов был использован набор из 36 (вместе с самими значениями прогнозируемого ряда — 37) макроэкономических показателей российской экономики.

К сожалению, многие из часто используемых при макроэкономических прогнозировании индикаторов (например, широко известные PMI — опросные индексы, отражающие настроения менеджмента в различных отраслях экономики), появились в России относительно поздно, в середине 2000-ых гг. Если бы они были включены в данные, то доступных для обучения и проверки качества моделей квартальных наблюдений было бы совсем мало, и автору пришлось пожертвовать количеством предикторов ради расширения выборки: исходные значения всех используемых в работе рядов доступны как минимум с 1-го квартала 1995 г. Все они, если требовалось, были приведены к стационарному виду. Это происходило путём взятия первой разности логарифма либо по отношению к предыдущему кварталу (для рядов без сезонности), либо по отношению к соответствующему кварталу предыдущего года (для рядов с сезонностью). После всех трансформаций левой границей использованных в работе данных стал 1-ый квартал 1996 г. Полный список переменных, их источники и способы трансформации приведены в таблице 2.2. На рисунке 1 приведён график стандартизированных[2] остационаренных переменных, используемых в работе, отдельно выделен ряд прогнозируемой переменной.

1.5 Построение прогнозов

В данной работе было исследовано несколько спецификаций для каждой модели.

Были отдельно рассмотрены две стартовые даты (левые границы тренировочной выборки): 1-ый квартал 1996 г. и 1-ый квартал 2000 г. Первая дата выбрана лишь из тех соображений, что, фактически, многие макроэкономические показатели до 1995 г. вовсе недоступны, а после преобразований, проводимых с данными, минимальная доступная дата сдвигается до 1-го квартала 1996 г. Вторая дата выбрана по следующим соображениям: можно предполагать, что добавление информации периода до кризиса 1998 г. (на момент

[2]Стандартизация на рисунке нужна лишь для удобства графического представления. При оценивании моделей стандартизация данных проводилась только для методов регуляризации — для ансамблевых методов она не требуется.

1-го квартала 2000 г. самая ранняя используемая при обучении информация относится к 1-му кварталу 1999 г.) не улучшает качество моделей. Если такая гипотеза косвенно подтвердится, это может быть полезно и для других прикладных исследований.

Для каждой из спецификаций модели регуляризации обучались на окне от левой до правой границ тренировочной выборки, где правая граница принимает значения от 1-го квартала 2012 г. до 4-го квартала 2018 г., отдельно для прогнозирования на горизонте h от 1 до 8 кварталов.

При этом, соответственно, для моделей с $h = 1$ максимальная конечная дата тренировочной выборки — 3-ий квартал 2018 г., для моделей прогноза на 2 квартала — 2-ой квартал 2018 г., и так далее, для моделей прогноза на 8 кварталов — 4-ый квартал 2016 г. Таким образом, для первой начальной даты (1-ый квартал 1996 г.) окно тренировочной выборки расширялось от 65 до 84 ($h = 8$) и 91 ($h = 1$) наблюдений, а для второй начальной даты (1-ый квартал 2000 г.) — от 49 до 68 ($h = 8$) и 75 ($h = 1$) наблюдений.

После обучения для каждой тренировочной выборки строились вневыборочные прогнозы на горизонте $h = 1, 2 \dots, 8$ кварталов для первого наблюдения, следующего за тренировочной выборкой, и для новой итерации граница тренировочной выборки сдвигалась на одно наблюдение вперёд. Таким образом, для получения прогноза на квартал $t+h$ использовались только значения переменных в квартале t.

Параметры штрафа в соответствующих моделях регуляризации выбирались при помощи кросс-валидации на тренировочном окне для каждой спецификации отдельно. Кросс-валидация внутри тренировочной выборки происходила на фиксированном скользящем окном величиной в 40 кварталов и шагом, равным 1.

Методология построения прогнозов модели случайного леса и бустинга в целом не отличалась от методологии построения прогнозов моделей регуляризации, за исключением того, что для выбора параметров не использовалась кросс-валидация.

Параметры случайного леса были выбраны следующим образом. Количество доступных для построения одного решающего дерева переменных p_0 установлено на уровне, рекомендуемом авторами этого метода Liaw и Wiener (2002) для задач регресии $p_0 = p/3$, где p — количество переменных. Таким образом, $p_0 = 12$. Выбирая количество деревьев N, необходимо руководствоваться следующими соображениями. При небольшом количестве деревьев нельзя быть уверенным в том, что все наблюдения использовались в модели, однако с ростом количества деревьев растет вычислительная сложность. В данной работе используется четыре разных значения N: $100, 500, 1000, 2000$. Как будет показано ниже, результаты для этих трёх моделей достаточно близки по качеству, и ни одна из них не лидирует при всех спецификациях, при этом разница между $N = 1000$ и $N = 2000$ почти незаметна, так что дальнейшее увеличение N бессмысленно. В работе используется стандартный рекомендуемый автором метода критерий остановки при построении дерева: вершина признаётся терминальной, если дальнейшее разбиение приводит к тому, что в одной из дочерних вершин окажется меньше 5 наблюдений.

Перейдём к параметрам бустинга. Стандартное значение скорости обучения $\eta = 0,3$. Хотя такое значение обычно позволяет получить довольно консервативную модель, в работе дополнительно тестируются спецификации $\eta = 0,4$, $\eta = 0,2$ и $\eta = 0,1$. В разных спецификациях нельзя однозначно

отобрать лучшее значение параметра. Параметр N отвечает за число итераций в ходе бустинга. Обычно при выборе N останавливаются на том уровне, после которого предсказания модели перестают меняться. В данной работе для этого оказалось достаточным установить $N = 100$.

Отдельно стоит отметить построение моделей-бенчмарков. Прогнозы модели случайного блуждания на h шагов вперёд представляли собой просто текущие значения прогнозируемого ряда.

Для построения модели авторегрессии сначала для каждой спецификации модели с помощью AIC выбиралось количество используемых лагов (как правило, оно не превышало 2). После этого происходила оценка коэффициентов и строился рекурсивный прогноз на 8 шагов вперёд, то есть последовательная оценка всех 8 прогнозных значений вместо отдельного оценивания 8 уравнений. Как показано в работе Faust и Wright (2013), прогнозы при таком подходе получаются более точными.

Метрика качества, используемая в работе — это RMSFE (англ. Root Mean Squared Forecast Error):

$$\text{RMSFE} = \sqrt{\frac{\sum_{t=1}^{T}(\hat{y}_t - y_t)^2}{T}}. \tag{15}$$

Такая метрика часто встречается при рассмотрении задач регрессии (она используется, например, в работах Фокина и Полбина (2019), Байбузы (2018)).

Все вычисления[3] проводились на языке R при помощи следующих пакетов:

- glmnet для моделей Ridge, LASSO, Post-LASSO, Adaptive LASSO, Elastic Net,
- spikeslab для модели Spike and Slab,
- randomForest для модели случайного леса,
- forecast для модели AR,
- xgboost для модели бустинга.

2 Результаты и обсуждение

2.1 Качество моделей

После построения прогнозов для каждой модели на тренировочной выборке были рассчитаны значения RMSFE. Эти значения относительно показателя качества наивного прогноза (процесса случайного блуждания) для горизонта прогнозирования от нуля до восьми кварталов представлены в таблице 1 для левой границы тренировочной выборки в 1-ом квартале 1996 г. и в таблице 2 для левой границы тренировочной выборки в 1-ом квартале 2000 г.

Для двух разных начальных дат тренировочной выборки наилучшее качество почти всегда показывают модели случайного леса и бустинга. Этого можно было ожидать — часто слабо интерпретируемые методы машинного обучения выигрывают в качестве прогнозов у методов регуляризации. Например, так произошло и в работе Байбузы (2018), где при прогнозировании

[3]Код с расчётами доступен по ссылке https://github.com/mkhlgrv/investment_forecasting/

Таблица 1: RMSFE (левая граница тренировочной выборки — 1-ый квартал 1996 г.)

Модель	1	2	3	4	5	6	7	8
Случайное блуждание	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
AR	0.92	0.85	0.79	0.73	0.68	0.67	0.68	0.61
Adaptive LASSO	0.99	0.91	0.79	0.84	0.75	0.77	0.61	0.50
Elastic Net	1.00	0.85	0.84	0.80	0.76	0.74	0.69	0.56
LASSO	1.01	0.89	0.83	0.81	0.82	0.72	0.71	0.55
Post-LASSO	1.04	0.95	0.80	0.93	0.79	0.78	0.72	**0.54**
Ridge	0.98	0.88	0.88	0.87	0.81	0.79	0.70	0.61
Spike and Slab	1.00	0.84	0.82	0.79	0.77	0.74	0.63	0.58
Бустинг ($\eta = 0,1$)	0.91	0.74	0.81	0.66	0.67	0.53	**0.45**	0.62
Бустинг ($\eta = 0,2$)	**0.87**	0.76	0.79	0.71	0.68	**0.46**	0.49	0.59
Бустинг ($\eta = 0,3$)	0.96	0.79	**0.69**	0.71	0.73	0.52	0.53	0.60
Бустинг ($\eta = 0,4$)	1.01	**0.70**	0.82	0.69	**0.61**	0.54	0.48	0.59
Случайный лес ($N = 100$)	0.90	0.72	0.75	**0.65**	0.70	0.56	0.54	0.57
Случайный лес ($N = 500$)	0.88	**0.70**	0.75	0.66	0.66	0.58	0.53	0.56
Случайный лес ($N = 1000$)	0.90	**0.70**	0.76	0.66	0.66	0.58	0.55	0.55
Случайный лес ($N = 2000$)	0.90	0.70	0.75	0.66	0.67	0.58	0.54	0.56

инфляции метод случайного леса и бустинг лидировали почти во всех спецификациях, или в работе Kvisgaard (2018), в которой методы регуляризации проигрывали другим методам машинного обучения при прогнозировании ВВП и инфляции. Разница в качестве прогнозов модели случайного леса при разных значениях параметра N небольшая, что позволяет говорить о том, что дальнейшее увеличение количества деревьев не приведёт к улучшению предсказаний.

Стоит отметить, что использование модификаций LASSO в целом почти не оправдано — в большинстве спецификаций Adaptive LASSO и Post-LASSO не превосходят «материнскую» модель по качеству. Байесовский подход к регуляризации в случае данной работы не позволил серьёзно улучшить качество прогнозов. Интересно, что метод Elastic Net, который теоретически должен использовать преимущества и Ridge, и LASSO, нередко оказывается хуже хотя бы одной из этих моделей. Впрочем, все они довольно близки в смысле ошибок прогнозов. В целом все тестируемые модели почти всегда показывают качество не ниже, чем у случайного блуждания, но при прогнозе на один квартал вперёд многие модели оказываются хуже, чем наивный прогноз.

Второй бенчмарк — модель авторегрессии — оказывается лучше некоторых методов регуляризации при полном наборе данных, однако неизменно проигрывает всем остальным моделям, если левая граница тренировочной выборки установлена на 1-ом квартале 2000 г.

Таблицы 1 и 2, показывающие качество моделей, можно дополнить с помощью теста Диболда — Мариано, изложенного в работах Diebold и Mariano (1995; 2002) для сравнения качества моделей. Так как размеры тестовых выборок очень скромные — от 17 до 25 наблюдений, — в работе использовалась скорректированная для маленьких выборок статистика теста Диболда—Мариано, предложенная Harvey, Leybourne и Newbold (1997). Нулевая гипотеза теста состоит в том, что две модели обладают одинаковым качеством прогнозов. Альтернативная гипотеза для каждой из спецификаций является

Таблица 2: RMSFE (левая граница тренировочной выборки — 1-ый квартал 2000 г.)

Модель	1	2	3	4	5	6	7	8
Случайное блуждание	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
AR	1.00	1.09	1.20	1.17	1.16	1.13	1.02	0.88
Adaptive LASSO	1.00	0.64	0.97	0.84	0.89	0.89	0.63	0.63
Elastic Net	0.78	0.62	0.96	0.76	0.81	0.82	0.67	0.59
LASSO	0.91	**0.58**	0.99	0.74	0.88	0.83	0.66	0.57
Post-LASSO	0.94	0.72	1.03	0.91	1.04	0.98	0.75	0.63
Ridge	0.88	0.64	0.79	0.78	0.76	0.79	0.68	0.66
Spike and Slab	0.85	0.64	0.80	0.80	0.91	0.83	0.66	0.60
Бустинг ($\eta = 0,1$)	0.97	0.64	0.70	0.60	0.68	**0.59**	0.60	**0.44**
Бустинг ($\eta = 0,2$)	0.90	0.63	**0.65**	0.58	0.69	0.63	0.63	0.57
Бустинг ($\eta = 0,3$)	1.10	0.69	0.66	0.66	**0.63**	**0.59**	0.62	0.52
Бустинг ($\eta = 0,4$)	1.08	0.66	0.72	**0.56**	0.67	0.66	**0.56**	0.50
Случайный лес ($N = 100$)	**0.75**	0.62	0.73	0.62	0.67	0.63	0.62	0.56
Случайный лес ($N = 500$)	0.81	0.61	0.73	0.64	0.67	0.63	0.61	0.58
Случайный лес ($N = 1000$)	0.78	0.62	0.72	0.63	0.66	0.62	0.59	0.57
Случайный лес ($N = 2000$)	0.78	0.62	0.73	0.64	0.66	0.62	0.59	0.58

односторонней, то есть состоит в том, что модель с наименьшим по выборке RMSFE даёт более качественные прогнозы.

На рисунках 3 и 4 показаны результаты тестов Диболда — Мариано для разных моделей, обученных на одной выборке и с одинаковым горизонтом прогнозирования. Зелёный цвет на рисунках означает, что модель, соответствующая строке, оказалась лучше, чем модель, соответствующая столбцу. Красный цвет означает превосходство модели, соответствующей столбцу. Чем бледнее цвет, тем больше p-значение статистики теста Диболда — Мариано, и соответственно, меньше разницы между двумя прогнозами. Если p-значение больше, чем $0,1$, разница между двумя моделями статистически незначима, и на рисунках это отражено белым цветом. Очевидно, что расположение на квадрате ячеек с красным и зелёным цветом симметрично относительно главной диагонали.

Можно заметить, что с ростом горизонта прогнозирования (при $h > 4$) прогнозы моделей начинают всё сильнее отличаться друг от друга и чаще проявляется превосходство ансамблевых методов над методами регуляризации. Однако уже при $h = 8$ прогнозы почти всех моделей статистически слабо отличаются друг от друга. Исключение составляют только модель AR, качество которой сильно ухудшается при переходе к урезанной выборке, и модель случайного блуждания, которая проигрывает большинству моделей на большинстве горизонтов прогнозирования.

Несмотря на то, что с ростом горизонта прогнозирования качество методов машинного обучения относительно случайного блуждания растёт, по-видимому, возможность извлечь из данных полезную информацию падает. При средних горизонтах ($h = 4, \dots, 6$) все ансамблевые методы сильно превосходят методы регуляризации, но при двухлетнем прогнозировании это превосходство становится статистически незаметным. Фактически это означает, что в данных недостаточно информации, чтобы модели могли строить двухлетние прогнозы, сильно отличающееся от просто среднего уровня.

На рисунках 3 и 4 можно увидеть, как переход к урезанной выборке влияет на разные методы: хорошие по качеству модели (бустинг и случайный лес) становятся ещё лучше, в то время как качество в среднем проигрывающих методов регуляризации увеличивается не так сильно или даже падает.

На рисунке 2 приведены вневыборочные прогнозы на горизонте прогнозирования до одного года ($h \leq 4$) с 1-го квартала 2013 г. по 1-ый квартал 2019 г. для урезанного набора данных. На графике отдельная линия прогноза соответствует одной тренировке модели. Разница в качестве моделей в терминах RMSFE подтверждается и на графике: визуально кажется, что бустинг и случайный лес лучше повторяют движение прогнозируемого ряда. На графике отображены прогнозы бустинга и случайного леса только для одной спецификации, потому что визуально прогнозы одинаковых моделей при изменении параметров оказываются похожими. На графике не отображены прогнозы случайного блуждания, так как они тривиальны и получаются путём сдвига наблюдений вправо.

2.1.1 Два набора данных

Из таблиц 1 и 2 видно, что данные докризисного периода почти никогда не помогают улучшить прогнозы. Чем это может быть объяснено? Докризисные наблюдения сильно нестабильны, значения переменных нетипичны для последующих переменных, и в результате обучения на них модели начинают хуже предсказывать данные последних периодов. На рисунке 1 изображены значения всех переменных, используемых в работе. Можно заметить уменьшение разброса их значений с началом века (ожидаемо, что в кризисные 2008–2009 гг. и 2014–2015 гг. ряды вновь становятся нестабильными).

Более строгая проверка гипотезы об улучшении прогнозов возможна с использованием теста Диболда—Мариано.

В таблице 3 отображены изменения RMSFE относительно модели случайного блуждания при переходе к урезанной выборке, а в скобках — значения статистики Диболда — Мариано для проверки гипотезы о равенстве двух моделей. Альтернативная гипотеза является односторонней, то есть состоит в том, что модель с меньшим показателем RMSFE показывает лучшее качество. Модели машинного обучения, которые были обучены на урезанных данных, дают прогнозы не хуже, чем аналогичные модели на расширенных данных при прогнозе на один год ($h \leq 4$). Исключение составляет только модель бустинга, прогнозы которой заметно ухудшаются при $h = 1$.

Что касается двухлетнего прогнозирования, то качество нескольких моделей также ухудшается при $h = 5, 6$, но в целом ни одна из двух спецификаций не превосходит другую постоянно. Однако, учитывая, что с ростом h, то есть увеличением разрыва между датами, в которую и на которую производится прогноз, интерпретируемость предсказаний должна падать по естественным причинам, можно утверждать, что удаление наблюдений до кризиса 1998 г., нетипичных для последующих периодов, позволяет давать более стабильные и адекватные результаты по крайней мере при прогнозе на один год вперёд.

Из методов регуляризации наиболее стабильной по отношению к уменьшению выборки оказалась модель Ridge: либо ухудшение качества было совсем незначительным ($h = 6, 8$), либо качество улучшалось. Этого можно было ожидать, учитывая, что Ridge позволяет получать устойчивые оценки.

При этом модель авторегрессии демонстрирует значительное ухудшение

Таблица 3: Тест Диболда — Мариано: две выборки

Модель	1	2	3	4	5	6	7	8
AR	-0.082	-0.243.	-0.413*	-0.441**	-0.477**	-0.458**	-0.337*	-0.271*
	(0.195)	(0.081)	(0.018)	(0.002)	(0.002)	(0.008)	(0.021)	(0.021)
Adaptive LASSO	-0.004.	0.264*	-0.181.	-0.002	-0.144*	-0.121	-0.013	-0.128
	(0.080)	(0.025)	(0.073)	(0.186)	(0.015)	(0.461)	(0.490)	(0.126)
Elastic Net	0.218	0.230*	-0.116	0.041	-0.050	-0.072	0.027	-0.037
	(0.412)	(0.046)	(0.128)	(0.453)	(0.421)	(0.379)	(0.249)	(0.250)
LASSO	0.104	0.309*	-0.153	0.074	-0.056	-0.102	0.055	-0.021
	(0.472)	(0.037)	(0.113)	(0.343)	(0.403)	(0.421)	(0.130)	(0.288)
Post-LASSO	0.105	0.229.	-0.229	0.022	-0.243*	-0.201	-0.025	-0.095
	(0.298)	(0.054)	(0.126)	(0.176)	(0.014)	(0.325)	(0.298)	(0.128)
Ridge	0.098	0.240*	0.087	0.092	0.041	-0.005	0.025	-0.051
	(0.365)	(0.050)	(0.246)	(0.189)	(0.162)	(0.249)	(0.484)	(0.217)
Spike and Slab	0.142	0.207.	0.022	-0.013	-0.139*	-0.083	-0.024	-0.021
	(0.479)	(0.063)	(0.302)	(0.418)	(0.045)	(0.381)	(0.292)	(0.442)
Бустинг ($\eta = 0,1$)	-0.058	0.100.	0.113	0.066.	-0.011	-0.054.	-0.147**	0.178*
	(0.109)	(0.083)	(0.107)	(0.069)	(0.421)	(0.083)	(0.006)	(0.034)
Бустинг ($\eta = 0,2$)	-0.033	0.131.	0.139*	0.132	-0.011	-0.162*	-0.136**	0.024
	(0.175)	(0.062)	(0.023)	(0.128)	(0.493)	(0.050)	(0.006)	(0.134)
Бустинг ($\eta = 0,3$)	-0.136*	0.097	0.022	0.047.	0.099	-0.064	-0.091	0.082
	(0.042)	(0.220)	(0.224)	(0.055)	(0.112)	(0.117)	(0.120)	(0.114)
Бустинг ($\eta = 0,4$)	-0.067	0.048	0.092	0.126.	-0.061.	-0.118*	-0.081	0.091
	(0.102)	(0.207)	(0.170)	(0.089)	(0.079)	(0.021)	(0.325)	(0.185)
Случайный лес ($N = 100$)	0.156	0.096*	0.024	0.026	0.032	-0.071	-0.076	0.009
	(0.269)	(0.041)	(0.275)	(0.124)	(0.203)	(0.269)	(0.219)	(0.147)
Случайный лес ($N = 500$)	0.074	0.096*	0.026	0.019.	-0.011	-0.052	-0.071.	-0.020
	(0.448)	(0.036)	(0.353)	(0.093)	(0.392)	(0.200)	(0.079)	(0.356)
Случайный лес ($N = 1000$)	0.114	0.075*	0.034	0.029.	-0.005	-0.044	-0.047	-0.019
	(0.308)	(0.040)	(0.279)	(0.093)	(0.486)	(0.224)	(0.256)	(0.369)
Случайный лес ($N = 2000$)	0.124	0.083*	0.027	0.027	0.008	-0.044	-0.040	-0.013
	(0.227)	(0.037)	(0.337)	(0.101)	(0.385)	(0.213)	(0.199)	(0.335)

В таблице указано изменение RMSFE относительно модели случайного блуждания при переходе к урезанной выборке.
В скобках указано p-значение статистики Диболда — Мариано (H_0: увеличение тренировочной выборки за счет данных до кризиса 1998 г. не изменяет качество моделей). p-значение: *** $\leq 0,001$ < ** $\leq 0,01$ < * $\leq 0,05$ < . $\leq 0,1$.

качества при сдвиге вправо левой границы тренировочной выборки. Видимо, эффект от недостатка информации при урезании выборки в этом случае превышает эффект от включения в выборку нестабильных наблюдений.

2.1.2 Выбор переменных в модели LASSO

Как отмечалось выше, методы, основанные на LASSO, позволяют отбирать только наиболее важные предикторы. Несмотря на то, что в данной работе LASSO проиграл в эксперименте по прогнозированию, результаты отбора релевантных регрессоров оказались интерпретируемыми и согласующимися с базовыми теоретическими представлениями о виде инвестиционной функции. В связи с этим представляется важным кратко представить результаты данного отбора.

Стоит начать с того, чтобы показать, какое количество переменных отбиралось LASSO при разных горизонтах прогнозирования и разных тренировочных выборках. На рисунке 5 изображено количество отобранных LASSO переменных для двух разных левых границ тренировочной выборки.

Количество выбранных переменных довольно нестабильно, что в целом типично для LASSO-моделей, как показано Zou и Hastie (2005) и De Mol, Giannone и Reichlin (2008). При этом можно заметить, что для $h = 3,5$ количество выбранных переменных стабильнее для моделей, которые тренировались на данных с 1-го квартала 1996 г. (на расширенной выборке), и на этих горизонтах прогнозирования LASSO показывает ухудшение качества при

Таблица 4: Основные предикторы в модели LASSO для $h = 1, \dots, 4$ (левая граница тренировочной выборки — 1-ый квартал 2000 г.)

	1	2	3	4
1	ВВП в постоянных ценах 0.046	М2 (на конец квартала) 0.023	Индекс реального оборота розничной торговли 0.228	Индекс потребительских цен - 0.292
2	Валовое накопление основного капитала 0.035	Индекс реальных денежных доходов населения 0.023	Кредиторская задолженность (в среднем за квартал) - 0.059	Кредиторская задолженность (в среднем за квартал) - 0.083
3	Индекс реальных денежных доходов населения 0.024	Индекс реальной зарплаты - 0.015	М2 (на конец квартала) 0.052	Индекс реальной зарплаты - 0.068
4	М2 (на конец квартала) 0.01	Индекс цен на строительно-монтажные работы 0.014	Индекс реальной зарплаты - 0.042	Индекс реальных денежных доходов населения 0.056
5	Индекс цен на строительно-монтажные работы 0.007	Импорт 0.016	Доля валового накопления основного капитала в ВВП (номинал) - 0.036	Индекс цен на строительно-монтажные работы - 0.051

уменьшении выборки (см. таблицу 3). Для $h = 2$ же, наоборот, количество коэффициентов в модели, обученной на урезанной выборке, колеблется не так сильно, как в модели, обученной на расширенной выборке, и это согласуется со статистически заметным улучшением качества модели при уменьшении выборки.

Однако, какие же именно предикторы выбираются? В таблицах 4 и 5 даны значения первых по модулю пяти средних стандартизированных значений коэффициентов модели LASSO для разных горизонтов прогнозирования при начальной дате в 1-ом квартале 2000 г. Так как коэффициенты стандартизированы, их значения условны, однако, чем больше вклад предиктора в изменение инвестиций, тем больше абсолютное значение коэффициента.

При прогнозировании на один квартал вперёд главными предикторами оказываются текущий ВВП и текущий лаг объясняемой переменной, что согласуется с акселераторной теорией, изложенной Clark (1917) и Guitton (1955), согласно которой существует оптимальное отношение капитала к выпуску μ, к этому соотношению стремятся максимизирующие прибыль фирмы. Подстройка происходит с лагами, поэтому текущий уровень инвестиций должен зависеть не только от текущего выпуска, но и от предыдущих инвестиций и предыдущего выпуска.

Ключевые предикторы для остальных h можно также найти в таблицах 4 и 5, но, по-видимому, с ростом горизонта интерпретируемость прогнозов падает. Самыми важными переменными при среднесрочном прогнозировании оказываются ИПЦ и кредиторская задолженность.

Стоит обратить внимание на переменную доли валового накопления основного капитала в ВВП, которая оказывается важна для нескольких горизонтов

Таблица 5: Основные предикторы в модели LASSO для $h = 5, \ldots, 8$ (левая граница тренировочной выборки — 1-ый квартал 2000 г.)

	5	6	7	8
1	Индекс потребительских цен - 0.022	Индекс потребительских цен - 0.04	Индекс потребительских цен - 0.019	Кредиторская задолженность (в среднем за квартал) - 0.013
2	Кредиторская задолженность (в среднем за квартал) - 0.007	Кредиторская задолженность (в среднем за квартал) - 0.01	Кредиторская задолженность (в среднем за квартал) - 0.011	Дебиторская задолженность (в среднем за квартал) 0.009
3	Индекс реальных денежных доходов населения 0.003	Индекс цен на строительно-монтажные работы - 0.003	Индекс цен на строительно-монтажные работы - 0.006	Индекс потребительских цен 0.007
4	Индекс цен на строительно-монтажные работы - 0.003	Дебиторская задолженность (в среднем за квартал) 0.003	Дебиторская задолженность (в среднем за квартал) 0.006	Индекс реального оборота розничной торговли 0.004
5	Просроченная дебиторская задолженность (в среднем за квартал) 0.001	Индекс цен производителей промышленных товаров - 0.002	Доля валового накопления основного капитала в ВВП (номинал) - 0.003	Индекс цен производителей промышленных товаров - 0.004

с отрицательным знаком. Наличие этой переменной фактически свидетельствует о том, что существует некоторое долгосрочное отношение инвестиций к выпуску, при нарушении которого происходит корректировка: при относительно чрезмерных инвестициях, растущих несоразмерно ВВП, через некоторое время происходит корректировка, и темпы роста инвестиций снижаются, и наоборот.

Несмотря на то, что отобранные LASSO коэффициенты и знаки при них не противоречат экономической теории, набор предикторов в целом не стабилен. Такая же ситуация характерна и для модификаций LASSO. По-видимому, последним и объясняется то, что их прогнозы уступают по качеству слабо интерпретируемым ансамблевым методам.

2.2 Сравнение с прогнозом Министерства экономического развития

Министерство экономического развития (МЭР) ежегодно осенью публикует обширные прогнозы социально-экономических показателей[4]. Автор сравнил результаты прогнозов собственных моделей с прогнозами темпов роста валового накопления основного капитала МЭР на следующий год. Поскольку МЭР прогнозирует изменения в процентах, прогнозы автора тоже были пересчитаны в процентные изменения к предыдущему году, причем в качестве даты прогноза на год $t+1$ использовался 3-ий квартал[5] года t, то есть прогноз

[4]Прогнозы МЭР доступны по ссылке http://economy.gov.ru/minec/activity/subsections/macro/prognoz/

[5]Дата прогноза — 3-ий квартал — установлена так, чтобы информация, использованная для обучения моделей, дающих прогнозы, примерно соответствовала информации, доступной

годового изменения получался из прогнозов для $h = 2$ (соответствует 1-му кварталу года $t + 1$), и так далее, $h = 5$ (соответствует 4-му кварталу года $t + 1$).

Наименьшая дата, начиная с которой доступны вневыборочные годовые прогнозы — это 3-ий квартал 2013 г., и, соответственно, сравнение с прогнозом МЭР возможно с 2014 г.

На рисунке 6 представлены прогнозы МЭР и моделей автора. Из всех спецификаций ансамблевых методов отображено только по одной спецификации для одного метода — визуально прогнозы при разных значениях параметров очень схожи. Можно видеть, что в целом никакие из двух предсказателей не является безусловным лидером, однако модели бустинга и случайного леса почти всегда оказываются близки к реальным значениям изменения инвестиций, что согласуется с их высоким качеством в смысле RMSFE.

Заключение

В работе были построены прогнозы индекса валового накопления основного капитала в России с помощью некоторых методов машинного обучения и большого набора предикторов. Наилучшее качество показывают ансамблевые методы — случайный лес и бустинг, лидирующие не только над простыми эталонными моделями, но и над методами регуляризации, что согласуется с литературой по макроэкономическому прогнозированию. Использование расширенного набора данных, включающего наблюдения до кризиса 1998 г., почти никогда не позволяет улучшить прогнозы по сравнению с урезанным набором данных. Относительно невысокое качество прогнозов методов регуляризации согласуется с нестабильностью их спецификаций в разные моменты времени. По результатам сравнения прогнозов автора и Министерства экономического развития можно увидеть, что некоторые из моделей машинного обучения дают значительно более качественные предсказания краткосрочного изменения инвестиций.

Перспективным направлением дальнейшего исследования является расширение комплекса моделей для прогнозирования, а также наукастинг инвестиций с учётом неоднородности выхода статистической информации (так называемая проблема «рваного края»).

на дату публикации прогноза МЭР (обычно это происходит в ноябре). Для простоты автор предполагает, что значения всех предикторов в 3-ем квартале публикуются не позднее, чем в 4-ом квартале.

Список литературы

Bai, J. и S. Ng (2008). “Forecasting economic time series using targeted predictors”. *Journal of Econometrics* 146.2, с. 304—317.

Belloni, A. и V. Chernozhukov (2011a). “High dıimensional sparse econometric models: An introduction”. *Inverse Problems and High-Dimensional Estimation*. Springer, с. 121—156.

Belloni, A. и V. Chernozhukov (2011b). “l1-penalized quantile regression in high-dimensional sparse models”. *The Annals of Statistics* 39.1, с. 82—130.

Clark, J. M. (1917). “Business acceleration and the law of demand: A technical factor in economic cycles”. *Journal of political economy* 25.3, с. 217—235.

De Mol, C., D. Giannone и L. Reichlin (2008). “Forecasting using a large number of predictors: Is Bayesian shrinkage a valid alternative to principal components?” *Journal of Econometrics* 146.2, с. 318—328.

Diebold, F. X. и R. S. Mariano (2002). “Comparing predictive accuracy”. *Journal of Business & economic statistics* 20.1, с. 134—144.

Diebold, F. X. и R. S. Mariano (1995). “Comparing Predictive Accuracy”. *Journal of Business and Economic Statistics* 13.3, с. 253—263.

Faust, J. и J. H. Wright (2013). “Forecasting inflation”. *Handbook of economic forecasting*. Т. 2. Elsevier, с. 2—56.

Friedman, J. H. (2001). “Greedy function approximation: a gradient boosting machine”. *Annals of statistics*, с. 1189—1232.

Guitton, H. (1955). “Koyck (LM)-Distributed Lags and Investment Analysis.” *Revue économique* 6.6, с. 127—128.

Harvey, D., S. Leybourne и P. Newbold (1997). “Testing the equality of prediction mean squared errors”. *International Journal of forecasting* 13.2, с. 281—291.

Hoerl, A. E. и R. W. Kennard (1970). “Ridge regression: Biased estimation for nonorthogonal problems”. *Technometrics* 12.1, с. 55—67.

Ishwaran, H. и J. S. Rao (2005). “Spike and slab variable selection: frequentist and Bayesian strategies”. *The Annals of Statistics* 33.2, с. 730—773.

Kvisgaard, V. H. (2018). “Predicting the future past. How useful is machine learning in economic short-term forecasting?” Дис. ... маг.

Li, J. и W. Chen (2014). “Forecasting macroeconomic time series: LASSO-based approaches and their forecast combinations with dynamic factor models”. *International Journal of Forecasting* 30.4, с. 996—1015.

Liaw, A. и M. Wiener (2002). “Classification and regression by randomForest”. *R news* 2.3, с. 18—22.

Santosa, F. и W. W. Symes (1986). “Linear inversion of band-limited reflection seismograms”. *SIAM Journal on Scientific and Statistical Computing* 7.4, с. 1307—1330.

Stock, J. H. и M. Watson (2011). “Dynamic factor models”. *Oxford Handbooks Online*.

Tibshirani, R. (1996). “Regression shrinkage and selection via the lasso”. *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1, с. 267—288.

Zou, H. (2006). “The adaptive lasso and its oracle properties”. *Journal of the American statistical association* 101.476, с. 1418—1429.

Zou, H. и T. Hastie (2005). “Regularization and variable selection via the elastic net”. *Journal of the royal statistical society: series B (statistical methodology)* 67.2, с. 301—320.

Аганбегян, А. Г. (2016). “Сокращение инвестиций-гибель для экономики, подъем инвестиций-ее спасение”. *Экономические стратегии* 18.4, с. 74—83.

Байбуза, И. (2018). “Прогнозирование инфляции с помощью методов машинного обучения”. *Деньги и кредит* 77.4, с. 42—59.

Идрисов, Г. и С. Синельников-Мурылев (2014). “Формирование предпосылок долгосрочного роста: как их понимать”. *Вопросы экономики* 3, с. 4—20.

Кудрин, А. и Е. Гурвич (2014). “Новая модель роста для российской экономики”. *Вопросы экономики* 12, с. 3.

Орешкин, М. С. (2018). “Перспективы экономической политики”. *Экономическая политика* 13.3.

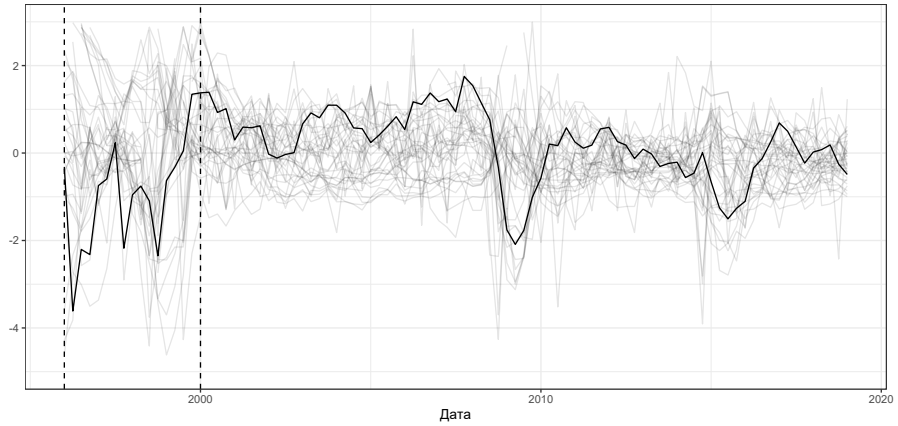
Фокин, Н. и А. Полбин (2019). “VAR-LASSO модель для прогнозирования ключевых макроэкономических показателей России”. *Деньги и кредит* 78.2, с. 67—93.

Таблица 6: Список используемых переменных

Переменная	Способ трансформации	Источник
Валовое накопление основного капитала	2	Росстат
Ввод в действие жилых домов	2	Росстат
ВВП в постоянных ценах	2	Росстат
Дебиторская задолженность (в среднем за квартал)	2	Росстат
Доля валового накопления основного капитала в ВВП (номинал)	2	Расчеты автора
Доходность 6-месячных государственных облигаций (в среднем за квартал)	0	Росстат
Доходы консолидированного бюджета	2	Росстат
Доходы федерального бюджета	2	Росстат
Задолженность в бюджет (в среднем за квартал)	2	Росстат
Задолженность поставщикам (в среднем за квартал)	2	Росстат
Заявленная потребность в работниках (в среднем за квартал)	2	Росстат
Импорт	2	Росстат
Индекс RTS/ Московской биржи (на конец квартала)	1	Bloomberg
Индекс потребительских цен	2	Росстат
Индекс реального оборота розничной торговли	2	Росстат
Индекс реального объёма сельскохозяйственного производства	2	Росстат
Индекс реальной зарплаты	2	Росстат
Индекс реальных денежных доходов населения	2	Росстат
Индекс тарифов на грузовые перевозки	2	Росстат
Индекс цен на строительно-монтажные работы	2	Росстат
Индекс цен производителей промышленных товаров	2	Росстат
Кредиторская задолженность (в среднем за квартал)	2	Росстат
Курс доллара на ММВБ/ Московской бирже (на конец квартала)	2	Росстат
M0 (на конец квартала)	2	Росстат
M2 (на конец квартала)	2	Росстат
Номинальный эффективный обменный курс (на конец квартала)	1	Bloomberg
Норма безработицы (в среднем за квартал)	2	Росстат
Официальный курс доллара (на конец квартала)	2	Росстат
Просроченная дебиторская задолженность (в среднем за квартал)	2	Росстат
Просроченная кредиторская задолженность (в среднем за квартал)	2	Росстат
Расходы консолидированного бюджета	2	Росстат
Расходы федерального бюджета	2	Росстат
Реальный эффективный обменный курс (на конец квартала)	1	Bloomberg
Ставка межбанковского рынка, 1 день (в среднем за квартал)	0	Банк России
Ставка межбанковского рынка, 7 дней (в среднем за квартал)	0	Банк России
Цена нефти Brent (на конец квартала)	1	Bloomberg
Экспорт	2	Росстат

Способы трансформации ряда x_t: 0 — без трансформации, 1 — $\ln(x_t) - \ln(x_{t-1})$, 2 — $\ln(x_t) - \ln(x_{t-4})$.

Рис. 1: Используемые в работе временные ряды



Приведены трансформированные и стандартизированные значения. Выделены значения прогнозируемой переменной. Пунктиром отмечены две левые границы тренировочных выборок.

Рис. 2: Прогнозы изменения инвестиций

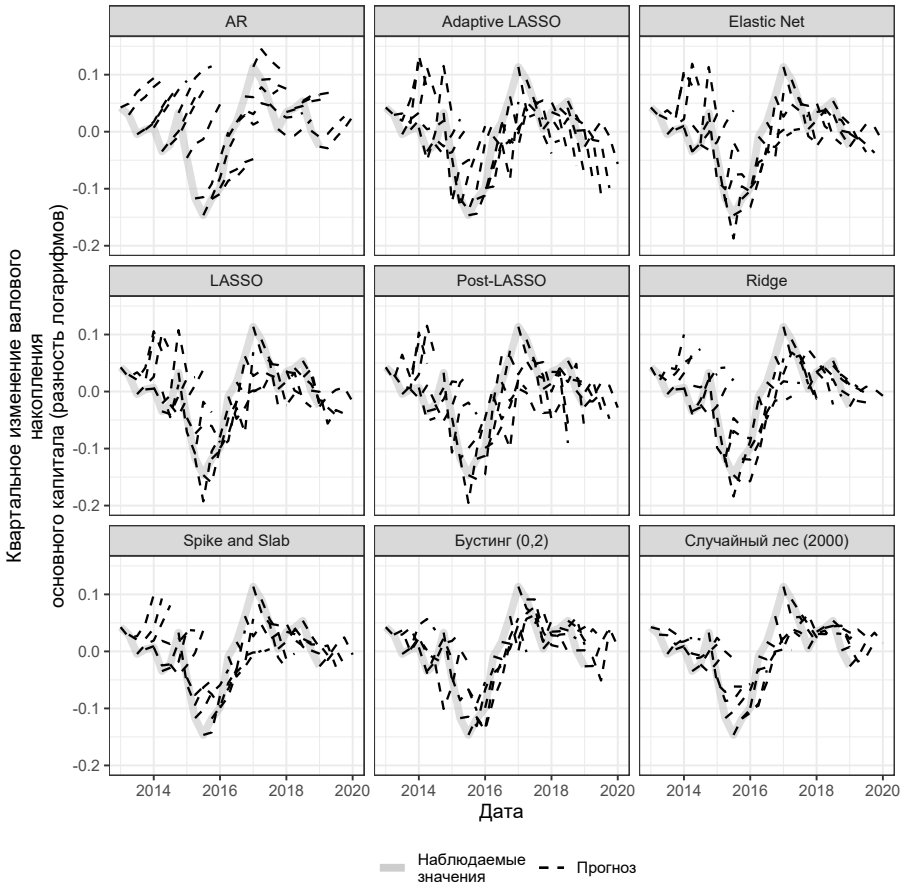


Рис. 3: Тест Диболда — Мариано: качество моделей (левая граница тренировочной выборки — 1-ый квартал 1996 г.)

H_0: две модели дают прогнозы одинакового качества. p-значение: *** $\leq 0,001$ < ** $\leq 0,01$ < * $\leq 0,05$ < . $\leq 0,1$. В скобках даны значения параметров: количество деревьев для модели случайного леса N и скорость обучения для модели бустинга η.

Рис. 4: Тест Диболда — Мариано: качество моделей (левая граница тренировочной выборки — 1-ый квартал 2000 г.)

H_0: две модели дают прогнозы одинакового качества. p-значение: *** $\leq 0,001$ < ** $\leq 0,01$ < * $\leq 0,05$ < . $\leq 0,1$. В скобках даны значения параметров: количество деревьев для модели случайного леса N и скорость обучения для модели бустинга η.

Рис. 5: Количество выбранных переменных в модели LASSO

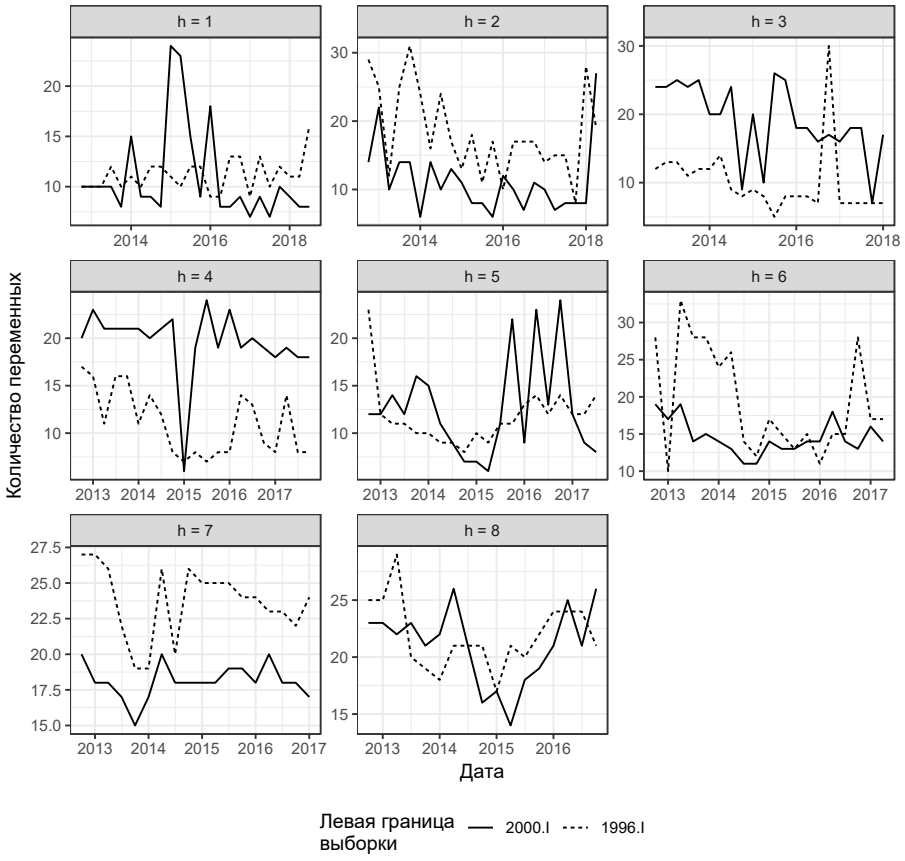
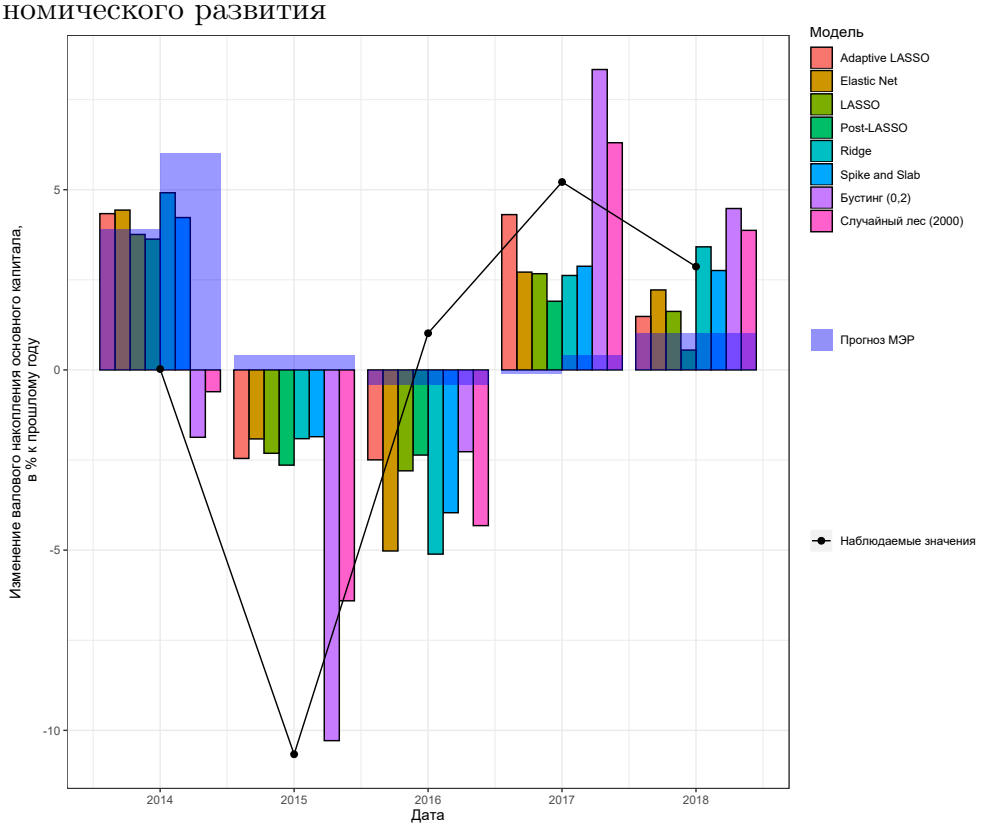


Рис. 6: Годовое изменение инвестиций: прогнозы автора и Министерства экономического развития



Дата составления прогноза на год t — 3-ий квартал года $t-1$. На 2014 и 2017 гг. показаны два вида прогнозов МЭР (базовый и консервативный). В скобках даны значения параметров: количество деревьев для модели случайного леса N и скорость обучения для модели бустинга η.