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 квартальных валовых инвестиций 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. 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 в целом согласуются с экономической теорией. В частности, важными детерминантами оказываются ВВП и рыночный индекс (положительное влияние), процентная ставка (отрицательное влияние). Прогнозы моделей автора превосходят по качеству годовые прогнозы темпов роста инвестиций, 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 get the best score. 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. Estimates of the coefficients of common accepted key investment factors generally correspond with economic theory. In particular, the important determinants are GDP and the market index (positive impact), interest rate (negative impact). 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.

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

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 различных переменных и бороться с проблемой переобучения и «проклятием размерности» (Stock и Watson, 2011). Однако результаты многих популярных методы машинного обучения зачастую сложно или невозможно интерпретировать. При макроэкономическом прогнозировании необходимо понимать, насколько прогнозные модели адекватны и соответствуют экономической теории. Поэтому может быть оправдано использование регуляризации, т.е. наложения штрафов за неадекватно высокие параметры модели. Методы регуляризации, такие как LASSO или Ridge, а также их модификации, с одной стороны позволяют уменьшить возможности 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 Кудрин и Гурвич (2014) отмечалось снижение инвестиционной активности 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 ведет к отставанию от остального мира. Недостаточная инвестиционная привлекательность российской экономики, главным образом вызванная, по мнению авторов статьи, слабостью рыночных механизмов, является главным ограничением для устойчивого роста. В работе Орешкин (2018) отмечается, что при существующих демографических ограничениях единственной возможностью повысить темпы роста экономики является увеличение объема 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 Идрисов и Синельников-Мурылев (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).

Однако в сфере краткосрочного прогнозирования интерес к ним, кажется, не очень высок.ЧЕМ НИБУДЬ ПОДТВЕРДИТЬ НАДО "КАЖЕТСЯ".. В ЛИТЕРАТУРЕ БОЛЬШЕ РАБОТ ПО ВВП ИНФЛЯЦИИ ЧЕМ ПО ИНВЕСТИЦИЯМ? 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, модель акселератора, 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 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, работе Li и Chen (2014) авторы исследова-

ли возможности LASSO, эластичной сети и группированного LASSO в этой области макроэкономического прогнозирования в сравнении с динамическими факторными моделями. Данные включали 107 различных месячных макроэкономических показателей американской экономики с 1959 г. по 2008 г., и авторы подробно рассматривали результаты для 20 показателей. В результате при прогнозах на один шаг вперёд методы регуляризации показывали в среднем лучшее качество (в смысле RMSFE), чем динамические факторные модели для 18, 15 и 19 переменных из 20 для LASSO, эластичной сети и группированного LASSO соотвественно. Кроме этого, комбинация каждого из 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 с применением методов регуляризации (в терминах статьи «мягких ограничений») может быть полезно для определения факторов, влияющих на инфляцию в США. Модели, допускающие изменение набора объясняющих переменных с течением времени (в том числе переменные выбирались с помощью LASSO), показывали в среднем качество лучше, чем модели с фиксированным набором факторов.

В работе Байбуза (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 12 кварталов вперед. Помимо методов регуляризации, также рассматривались ансамблевые методы машинного обучения (случайный лес и бустинг). Модели с регуляризацией показали достаточно плохие прогнозы относительно бенчмарков (моделей 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 Фокин и Полбин (2019) посвящена совмещению векторной авторегрессии и LASSO-регуляризации (VAR-LASSO модель) для моделирования основных показателей российской экономики: ВВП, потребления, инвестиций в основной капитал, а также экзогенного шока цены на нефть. По результатам тестирования модель авторов показала хорошие прогнозные свойства по сравнению с обычной VAR моделью, а также прогнозами министерства экономического развития. Кроме того, были построены функции импульсного отклика на шок изменения цены нефти.

Далее работа построена следующим образом. В разделе 1 содержится методология исследования: приведено описание используемых методов машинного обучения (Ridge, LASSO, Post-LASSO, Adaptive LASSO, метод эластичной сети, метод пик-плато, случайный лес, бустинг), альтернативных методов прогнозирования, используемых для сравнения качества, а также описание описание использованных данных, способы их трансформации и описание построения прогнозов. В разделе 2 содержатся эмпирические результаты исследования и их обсуждение. Кроме того, 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 Методология

Можно выделить несколько групп методов работы с большими объемами данных, используемых для прогнозирования макроэкономических переменных, которые отличаются разным подходом к решению т.н. проблемы «проклятия размерности», которое и

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

является основным препятствием использования большого количества переменных при прогнозировании макроэкономических показателей.

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 (англ. Least Absolute Shrinkage and Selecting Operator), 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 т.н. 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 переменной, $,\varepsilon_i \sim N(0, \sigma^2)$ — это независимые и одинаково распределенные случайные ошибки, при этом все переменные стандартизированы, т.е. имеют нулевое математическое ожидание и единичную дисперсию, или, в матричном виде:

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

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 переменной, $\mathcal{E} \in \mathbb{R}^n$ — это вектор независимых и одинаково распределенных случайных ошибок.

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

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

1.1.1 Ridge

Как было уже сказано выше, МНК-оценки для моделей с высокой размерностью часто имеют низкое смещение, но высокую дисперсию. Регрессия Ridge позволяет уменьшать дисперсию оценок коэффициентов, однако, делая их смещенными. 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, т.к. $X'X$ ОШИБКА В ПДФ матрица необратима. Однако решение существует для модели Ridge в случае, если $\lambda \neq 0$.

Существенным 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 (англ. Least Absolute Shrinkage and Selecting Operator) 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 регрессии гребня, 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 переменных 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, Adaptive LASSO.

1.1.3 Post-LASSO

В общем случае оценки, полученные при помощи регуляризации, будут смещены. В работе Belloni и Chernozhukov (2011a) показано, что потенциально менее смещённые оценки LASSO могут быть получены при использовании метода Post-LASSO. Использование оператора 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 не оказались равны нулю. В этом случае мы получаем т.н. оценку Post-LASSO.

Формально это можно записать следующим образом:

$$\hat{\beta}^{\text{LASSO}} \in \arg\min_{\beta \in \mathbb{R}^p} \left\| Y - X\beta \right\|_2^2 + \lambda \left\| \beta \right\|_1, \tag{6}$$

$$\hat{\beta}^{\text{Post-LASSO}} \in \arg\min_{\beta \in \mathbb{R}^p} \left\| Y - X\beta \right\|_2^2, \text{ где } \beta_j = 0, \text{ если } \hat{\beta}_j^{\text{LASSO}} = 0. \tag{7}$$

Полученные таким образом оценки потенциально приводят к меньшему смещению, однако, как эмпирически показали авторы работы Belloni и Chernozhukov (2011a), при высокой зашумленности данных модель становится нестабильной и оценки коэффициентов могут оказаться даже дальше от истинных, чем у LASSO.

1.1.4 Адаптивный LASSO

В исследовании Zou (2006) показано, что в некоторых ситуациях LASSO может неверно исключить (т.е. оценивать коэффициенты как ноль) переменные. В некоторых случаях выбор параметра штрафа λ, показывающего наилучшее качество оценки, приводит к выбору «мусорных» переменных, вместе с этим также возможны случаи, когда при правильном отборе переменных выбранные коэффициенты были неоправданно высоки и приводили к относительно плохим прогнозам.

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

$$\hat{\beta}^{\text{Adaptive LASSO}} \in \arg\min_{\beta \in \mathbb{R}^p} \left\| Y - X\beta \right\|_2^2 + \lambda \left\| w\beta \right\|_1, \tag{8}$$

где вектор $w \in \mathbb{R}^p$ — 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:

$$w_j = \frac{1}{(|\beta_j^{\text{init}}|)^\gamma}, \tag{9}$$

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 выбирать параметр γ с помощью кросс-валидации, однако это было не очень разумным, учитывая небольшой объем 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 Эластичная сеть

В работе Zou и Hastie (2005) была показана неустойчивость LASSO в выборе переменных, которая обусловлена неопределенностью параметров при оценке ковариационной матрицы. Для решения этой проблемы авторы предложили метод, известный как эластичная сеть. Это общий случай LASSO и регрессии гребня. Главные отличия этих двух методов состоят в следующем: LASSO позволяет занулять коэффициенты, зато Ridge дает более стабильные оценки для высококоррелированных переменных, в то время как оценки LASSO могут сильно поменяться при добавлении новых наблюдений.

Название метода связано с тем, что эластичная сеть позволяет «ловить всю большую рыбу» по сравнению с двумя частными случаями. С одной стороны, в регрессии эластичной сети возможно зануление коэффициентов, но, с другой стороны, модель должна получаться более стабильной. Задача имеет следующий вид:

$$\hat{\beta}^{\text{Elastic Net}} \in \arg\min_{\beta \in \mathbb{R}^p} \left\| Y - X\beta \right\|_2^2 + \lambda(\alpha \left\| \beta \right\|_1 + (1-\alpha) \left\| \beta \right\|_2^2), \tag{10}$$

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 сокращается до LASSO, а при $\alpha = 0$ — до Ridge. Параметр α мог быть выбран с помощью кросс-валидации, однако это не очень оправдано при небольшой выборке, поэтому параметр равен $0,5$, как равноудаленный от двух крайних случаев. Такой же параметр был выбран, например, в работе Байбуза (2018).

1.1.6 Регрессия пик-плато

Байесовский подход в эконометрике предполагает наличие априорного распределения для каждого параметра модели, которое в соответствии с имеющимися данными корректируется и в результате получается апостериороное распределение. Такой подход, например, получать оценки параметров, даже если переменных больше, чем наблюдений, или позволяет напрямую ставить вопрос о вероятности того, что какой-то коэффициент в модели равен нулю. В этой работе используется байесовский метод регрессии пик-плато.

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 - регрессии LASSO (см. 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:

$$\begin{cases} (Y_i | x_i, \beta, \sigma^2) \sim N(x_i^t \beta, \sigma^2), \\ (\beta | \gamma) \sim N(0, \Gamma), \\ \gamma \sim \pi(\cdot), \\ \sigma^2 \sim \mu(\cdot), \end{cases} \tag{11}$$

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 регрессия пик-плато (Spike and Slab) в версии, изложенной в работе Ishwaran и Rao (2005), в соответствии с которым предполагается следующее:

$$\begin{cases} (\beta_j|\tau_j, \rho_j^2) \sim N(0, \tau_j \cdot \rho_j^2), \\ \tau_j \sim (1-w)\delta_{v_0}(\cdot) + w\delta_1(\cdot), \\ (\rho^{-2}|\alpha_1, \alpha_2) \sim \Gamma(\alpha_1, \alpha_2), \\ w \sim U[0;1] \end{cases} \tag{12}$$

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 задает 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 усредненных предсказаний этих методов. В основе метода случайного леса (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, они обладают и существенным недостатком — 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 деревом — это бинарная классификация потенциальных заемщиков банком — вернёт заемщик кредит или нет), их можно использовать и для задач регрессии. Однако сами по себе деревья в настоящее время используются редко, зато часто их объединяют в композицию ансамблевых методов.

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 случайные 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 $\text{RSS} = \sum_{i=1}^{n}(y_i - \hat{y}_i)$) (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 признается 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 некоторая 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). Метод случайного леса, рассмотренный ниже, не так сильно подвержен этой проблеме. «Высаживание» случайного леса состоит из двух этапов:

1. данные случайным образом разбиваются на N подвыборок (с повторениями) по $n_0 < n$ наблюдений, на каждой строится решающее дерево (при этом при построении дерева на одной подвыборке для выбора доступны только p_0 переменных, случайным образом выбираемых из p регрессоров отдельно для каждой подвыборки (как правило, в задачах регрессии $p_0 = p/3$));

2. в качестве прогнозного значения $\hat{y}_i$ выбираются усреднённое значение $\hat{y}_i$ по всем деревьям.

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 особенно существенным при прогнозировании макроэкономических или финансовых данных.

1.2.2 Бустинг

Метод градиентного бустинга, предложенная в работе Friedman (2001), как и случайный лес, принадлежит к числу ансамблевых методов (т.е представляет собой композицию нескольких простых моделей одного типа). Однако, в отличии от случайного леса, процедура бустинга позволяет производить последовательное улучшение моделей 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. Так происходит и в этой 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 $\eta \in (0,1)$:

$$M_2 = M_1 + \eta m_2 \tag{13}$$

Параметр η отвечает за скорость обучения (при высокой скорости модель склонна переобучаться). В данной работе он выбран равным $0,2$. Такое значение позволяет получать достаточно консервативную модель.

Итоговым результатом для N итераций является модель $M_N = M_1 + \sum_{i=2}^{N} \eta^{i-1} m_i$. При большом количестве итераций модель склонна переобучаться на тренировочной выборке.

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 вперед записывается следующим образом:

$$\hat{y}_{t+h} = y_t. \tag{14}$$

1.3.2 Авторегрессия

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

$$y_t = a_0 + a_1 y_{t-1} + a_2 y_{t-2} + \dots a_p y_{t-p} + \epsilon_t. \tag{15}$$

На практике выбор количества лагов обычно производится при помощи информационных критериев (AIC, BIC).

1.4 Данные

Прогнозируемой переменной является квартальный прирост валового накопления основного капитала в России в постоянных ценах. Он построен из трех рядов валового накопления, 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). Квартальный ряд инвестиций в основой капитал, очевидно, имеет сезонность и приводится к стационарному виду через взятие 4-ой разности логарифма. Такой подход учитывает мультипликативность и возможные структурные сдвиги в сезонности. Проведенный расширенный тест Дики-Фулера не обнаруживает в трансформированном ряде нестационарности. Таким образом, после преобразований временной ряд 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. Некоторые из выбранных показателей отсылают к классическим детерминантам инвестиций, поэтому ниже, при изложении результатов, им будет уделено пристальное внимание. 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 либо путем взятия первой разности логарифма (для рядов без сезонности), либо четвертой разности логарифма (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 приведен в таблице 2.2. На рисунке 1 приведен график стандартизированных 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 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 начальной до конечной дат тренировочной выборки, где конечная дата принимает значения от 1-го квартала 2012 г. до 4-го квартала 2018 г. отдельно для прогнозирования на горизонте h от 0 до 8 кварталов (при $h = 0$ изучается задача не прогнозирования инвестиций, а предсказания текущего значения в рамках построенных моделей, выявления ключевых детерминант текущей динамики инвестиций).

. При этом, соответственно, для моделей с $h = 0$ максимальная правая граница тренировочной выборки — 4-ый квартал 2018 г., для моделей прогноза на 1 квартал — 3-ий квартал 2018 г., ..., для моделей прогноза на 8 кварталов — 4-ый квартал 2016 г. Таким образом, для первой начальной даты (1996.I) окно тренировочной выборки расширялось от 65 до 84 ($h = 8$) и 92 ($h = 0$) наблюдений, а для второй начальной даты (2000.I) — от 49 до 68 ($h = 8$) и 76 ($h = 0$) наблюдений.

После обучения для каждой тренировочной выборки строились вневыборочные прогнозы от 0 до 8 кварталов для первого наблюдения, следующего за тренировочной выборкой, и для новой итерации граница тренировочной выборки сдвигалась на одно наблюдение вправо.

Параметры штрафа в соответствующих моделях выбирались при помощи кросс-валидации на тренировочном окне для каждой спецификации отдельно. Кросс-валидация внутри тренировочной выборки происходила с фиксированным окном и шагом, равным 1. Стоит отметить, что для всех использованных в работе моделей при построении прогнозов на кварталы $t + 1, t + 2, \dots, t + 8$ использовались истинные значения ряда валового накопления основного капитала в момент времени t, а для $h = 0$ соответственно, эти значения не были доступны.

Отдельно стоит отметить построение моделей-бенчмарков. Модель случайного блуждания согласно определению строила прогнозы на h шагов вперед в виде текущих значений стационарного ряда инвестиций, а при $h = 0$ по аналогии с другими моделями в

качестве объясняющей переменной использовались значения ряда в момент $t-1$. Обучение модели авторегрессии на тренировочном окне заключалось в подборе количества лагов p при помощи AIC и, после этого, оценивании коэффициентов с помощью МНК. После этого для каждого элемента тестовой выборки строились отдельно прогноз для $h = 0$ и рекурсивные прогнозы на 8 шагов вперед, т.е. последовательная оценка всех 8 прогнозных значений через друг друга вместо отдельного оценивания 8 уравнений. Согласно работе Faust и Wright (2013), прогнозы в этом случае получаются более точными. Как и для остальных моделей, при построении прогнозов на момент времени $t+h(h>0)$ использовались истинные, а не оцененные с помощью модели для $h = 0$ значения ряда инвестиций в момент времени t.

Методология построения прогнозов модели случайного леса в целом не отличалась от методологии для моделей регуляризации. Гиперпараметры для обучения модели случайного леса в данной работе выбраны следующим образом:

- Количество деревьев $N = 100$,
- Количество случайно выбранных переменных, доступных для построения дерева в каждой из N подвыборок $p_0 = p/3$, где p — общее количество объясняющих переменных (стандартная величина для задачи регрессии).

Гиперпараметры для обучения модели бустинга следующие:

- Количество деревьев $N = 100$,
- Скорость обучения $eta = 0,2$,
- Количество случайно выбранных переменных, доступных для построения дерева в каждой из N итераций $p_0 = p/3$.

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

$$\text{RMSFE} = \sqrt{\frac{\sum_{t=1}^{T}(\hat{y}_t - y_t)^2}{T}}. \tag{16}$$

Такая метрика вполне типична для задач регрессии.
Все вычисления 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 авторегрессии;
- xgboost для модели градиентного бустинга.

Таблица 1: RMSFE (стартовая дата — 1996.I)

Модель	0	1	2	3	4	5	6	7	8
Random Walk	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
AR	0.93	0.92	0.85	0.79	0.73	0.68	0.67	0.68	0.61
Adaptive LASSO	0.91	0.99	0.91	0.79	0.84	0.75	0.77	0.61	**0.50**
Boosting	0.87	0.93	0.94	0.76	0.70	**0.62**	0.65	0.59	0.62
Elastic Net	0.92	1.00	0.85	0.84	0.80	0.76	0.74	0.69	0.56
LASSO	0.93	1.01	0.89	0.83	0.81	0.82	0.72	0.71	0.55
Post-LASSO	1.02	1.04	0.95	0.80	0.93	0.79	0.78	0.72	0.54
Random Forest	**0.84**	**0.90**	**0.70**	**0.74**	**0.69**	0.65	**0.61**	**0.54**	0.52
Ridge	0.95	0.98	0.88	0.88	0.87	0.81	0.79	0.70	0.61
Spike and Slab	0.87	1.00	0.84	0.82	0.79	0.77	0.74	0.63	0.58

Таблица 2: RMSFE (стартовая дата — 2000.I)

Модель	0	1	2	3	4	5	6	7	8
Random Walk	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
AR	0.99	1.00	1.09	1.20	1.17	1.16	1.13	1.02	0.88
Adaptive LASSO	1.08	1.00	0.64	0.97	0.84	0.89	0.89	0.63	0.63
Boosting	0.93	1.04	**0.58**	0.70	0.68	**0.58**	**0.58**	0.69	0.61
Elastic Net	0.93	0.78	0.62	0.96	0.76	0.81	0.82	0.67	0.59
LASSO	0.95	0.91	**0.58**	0.99	0.74	0.88	0.83	0.66	**0.57**
Post-LASSO	1.05	0.94	0.72	1.03	0.91	1.04	0.98	0.75	0.63
Random Forest	**0.87**	**0.78**	0.59	**0.73**	**0.62**	0.67	0.60	**0.57**	0.58
Ridge	1.11	0.88	0.64	0.79	0.78	0.76	0.79	0.68	0.66
Spike and Slab	0.91	0.85	0.64	0.80	0.80	0.91	0.83	0.66	0.60

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 процесса случайного блуждания для горизонта прогнозирования от нуля до восьми кварталов представлены в таблицах 1 для стартовой даты в 1-ом квартале 1996 г. и 2 для стартовой даты в 1-ом квартале 2000 г.

Для двух разных левых границ тренировочной выборки наилучшее качество почти всегда показывают модель случайного леса и бустинга. Этого можно было ожидать — часто слабо интерпретируемые методы машинного обучения выигрывают при различных прогнозов. Например, так произошло и в работе Байбуза (2018), где при прогнозировании 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. 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 серьезно улучшить качество прогнозов. Интересно, что метод эластичной сети, 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 показывают качество не ниже, чем у случайного блуждания, исключение составляет лишь $h = 0, 1$, когда некоторые модели оказываются хуже, чем случайное блуждание. 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. при $h > 0$.

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 RMSE подтверждается и на графике: визуально кажется, что бустинг и случайный лес лучше повторяют движение прогнозируемого ряда. На графике не отображены прогнозы случайного блуждания, так как они тривиальны и получаются путем 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 (Diebold и Mariano, 1995; Diebold и Mariano, 2002)) для сравнения качества моделей. Так как размеры тестовых выборок очень скромные — от 19 до 27 наблюдений, — в работе использовалась скорректированная для маленьких выборок статистика теста Диболда—Мариано (Harvey, Leybourne и Newbold, 1997). Нулевая гипотеза теста состоит в том, что две модели обладают одинаковым качеством прогнозов. Альтернативная гипотеза для каждой из спецификаций была односторонней, т.е. состояла в том, что модель с наименьшим на выборке RMSFE дает более качественные прогнозы.

На рисунке 3 изображены значения статистики Диболда-Мариано для различных моделей и горизонтов прогнозирования. Зеленый цвет означает превосходство урезанной модели (использовались наблюдения с 1-го квартала 2000 г.), красный — превосходство полной модели (использовались наблюдения с 1-го квартала 1996 г.), синий — статистическое равенство качества двух моделей на уровне значимости в 1% при односторонней альтернативной гипотезе. Видно, что модели машинного обучения, тренированные на урезанных данных, дают прогнозы не хуже, чем аналогичные модели на расширенных данных при прогнозе на один год (h⊠4), только модель Ridge ухудшает качество при $h = 0$.

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, т.е. 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 вперед.

При этом модель авторегрессии демонстрирует значительное ухудшение 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 уже отмечалось выше, методы регуляризации ценны тем, что, с одной стороны, позволяют получать прогнозы на основе большого количества объясняющих переменных, а, с другой стороны, сохраняют интерпретируемость. Было бы интересно узнать, сколько всего предикторов и какие именно из них отбираются при разных спецификациях моделей, в особенности — при тех спецификациях, которые дают хорошие прогнозы.

Сколько предикторов выбиралось моделями и как это число менялось с течением времени? Было бы излишне приводить это число для всех моделей, учитывая, что обычно по качеству лидируют слабо интерпретируемые методы, автором для примера отобрана модель LASSO. Стоит отметить, что число предикторов в ней точно совпадает с числом переменных в модели Post-LASSO.

На рисунке 4 отображено количество отобранных LASSO переменных для двух разных левых границ тренировочной выборки.

В целом количество довольно нестабильно, что в целом типично для LASSO-моделей (Zou и Hastie, 2005; De Mol, Giannone и Reichlin, 2008). При этом можно заметить, что для $h = 0,1$ количество выбранных переменных сильно стабильнее для модели, которые тренировались на данных с 1-го квартала 1996 г. (на расширенной выборке).

Однако, какие же именно предикторы выбираются? В таблицах 3 и 4 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 объяснении текущих инвестиций главным и почти полностью определяющим фактором оказывается ВВП в постоянных ценах, что в целом согласуется с классической акселераторной теорией (введена больше века назад в работе Clark (1917), расширена в работе Guitton (1955)), согласно которой существует оптимальное отношение капитала к выпуску μ, к этому соотношению стремятся максимизирующие прибыль фирмы, и в этом случае единственным показателем, определяющим уровень инвестиций, является выпуск. Другие, менее существенные объясняющие переменные — индекс реального оборота розничной торговли, ИПЦ, индекс цен производителей промышленных товаров и заявленная потребность в работниках.

При прогнозировании на один квартал вперед ВВП по-прежнему играет самую важную роль, но при $h = 1$ еще одним важным фактором оказывается лаговое значение инвестиций. Этот вывод в целом также согласуется с акселераторной моделью с предложением об отсутствии мгновенной подстройки уровня капитала к оптимальному (подстройка происходит с лагами, поэтому уровень текущих инвестиций определяется уже текущим ВВП и предыдущими значениями ВВП и инвестиций). Кроме этого, для прогноза на один квартал важными переменными для предсказания являются индекс реальных денежных доходов населения, агрегат М2 и индекс цен на строительно-монтажные работы. Ключевые предикторы для остальных h можно также найти в таблицах 3 и 4, но, по-видимому, с ростом горизонта интерпретируемость прогнозов падает. Самыми важными переменными при среднесрочном прогнозировании оказываются ИПЦ и кредиторская задолженность.

Однако стоит обратить внимание на переменную доли валового накопления основного капитала в ВВП, которая оказывается важна для нескольких горизонтов с отри-

Таблица 3: Основные предикторы в модели LASSO для $h = 0, \dots, 4$ (левая граница тренировочной выборки — 1-ый квартал 2000 г.)

	0	1	2	3	4
1	ВВП в постоянных ценах 0.237	ВВП в постоянных ценах 0.046	М2 (на конец квартала) 0.023	Индекс потребительских цен -0.111	Индекс потребительских цен -0.292
2	Индекс реального оборота розничной торговли 0.057	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 потребительских цен 0.024	Индекс реальных денежных доходов населения 0.024	Индекс реальной зарплаты -0.015	М2 (на конец квартала) 0.052	Индекс реальной зарплаты -0.068
4	Индекс цен производителей промышленных товаров 0.006	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	Заявленная потребность в работниках (в среднем за квартал) 0.005	Индекс цен на строительно-монтажные работы 0.007	Кредиторская задолженность (в среднем за квартал) -0.012	Доля валового накопления основного капитала в ВВП (номинал) -0.036	Индекс цен на строительно-монтажные работы -0.051

Таблица 4: Основные предикторы в модели LASSO X для $h = 5, \dots, 8$ (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

цательным знаком. Наличие этой переменной фактически свидетельствует о том, что существует некоторое долгосрочное отношение инвестиций к выпуску, при нарушении которого происходит корректировка

Ниже рассмотрено, как оценки коэффициентов в модели LASSO некоторых переменных, встречающихся в литературе по объяснению инвестиций, менялись с изменением границ тренировочной выборки.

Во-первых, как было отмечено выше, основным фактором при $h = 0,1,2$ является ВВП, и это вполне согласуется с акселераторной моделью. На рисунке 5 показаны значения коэффициента при ВВП для $h = 0,1$ Явно видна тенденция к увеличению значения коэффициента с течением времени. Вместе с тем кризисному периоду 2014–2015 г. соответствует резкое снижение значимости ВВП при прогнозировании на один квартал вперед.

Как сказано выше, модель гибкого акселератора предполагает, что текущий уровень инвестиций определяется не только ВВП, но и лагами самих инвестиций. На рисунке 6 отображены значения коэффициента при инвестициях для $h = 1$ (на остальных горизонтах прогнозирования переменная почти никогда не выбирается LASSO). В целом значение коэффициента стабильно, но в 2016 г. происходит падение значимости (можно заметить, что в 2015–2016 гг. значимость инвестиций и ВВП двигалась разнонаправленно при прогнозировании на один квартал).

Интересным показателем в среднесрочном прогнозировании является доля инвестиций в ВВП. Значение коэффициента при этой переменной отображено для $h = 7,8$ на рисунке 7. Стабильно отрицательное значение говорит о том, что существует некоторое стабильное соотношение ВВП и инвестиций: при относительно чрезмерных инвестициях, растущих несоразмерно ВВП, через некоторое время происходит корректировка , и темпы роста инвестиций снижаются, и наоборот. Значимость в абсолютном смысле уменьшается в 2013–2014 гг. (так как прогнозирование осуществляется на 7–8 кварталов вперед, это соответствует 2015–2016 гг.) и восстанавливается одновременно с восстановлением стабильности экономики.

Несколько известных моделей объяснения инвестиций: модель q Тобина (Tobin, 1969), модель денежного потока (Grunfeld, 1960) в качестве одного из факторов инвестиций рассматривают рыночную стоимость фирмы как выражение будущих прибылей. При макрорасширении этих моделей в качестве объясняющей переменной можно использовать рыночный индекс, конкретно в работе одной из переменных является индекс РТС (с 2011 - индекс Московской биржи) На рисунке 8 показаны значения соответствующего коэффициента для $h = 1,2,3$ (на остальных горизонтах не выбирался). Несмотря на нестабильное поведение, знак остается положительным и на таких горизонтах LASSO всегда отбирает эту переменную, что в целом не противоречит перечисленным моделям.

Что касается процентной ставки (в качестве одной из ставок в экономике использовалась ставка межбанковского рынка), то на рисунке 9 показано изменение коэффициента при $h = 1,2,3$ (на остальных горизонтах переменная почти никогда не выбирается моделью LASSO). Знак остается отрицательным, что совпадает с ожиданиями, но в целом поведение коэффициента относительно нестабильно, причем не только в периоды кризиса.

2.2 Сравнение с прогнозом министерства экономического развития

Министерство экономического развития ежегодно осенью публикует обширные прогнозы социально-экономических показателей[2]. Автор сравнил результаты прогнозов соб-

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

ственных 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 года t, то есть прогноз годового изменения получался из прогнозов для $h = 2$ (соответствует 1-му кварталу года t), ..., 5 (соответствует 4-му кварталу года t).

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 10 представлены прогнозы МЭР и моделей автора. Можно видеть, что в целом никто 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 большого набора макроэкономического 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. Вместе с тем, некоторые из ключевых факторов инвестиций (ВВП, лаги инвестиций, фондовый индекс, процентная ставка, доля инвестиций в ВВП) вполне значимы при прогнозировании с помощью методов регуляризации и имеют ожидаемые знаки. 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 учетом неоднородности выхода статистической информации (т.н. проблема «рваного края»).

Список литературы

Bai, J. and S. Ng (2008). "Forecasting economic time series using targeted predictors". *Journal of Econometrics* 146.2, pp. 304–317.

Belloni, A. and V. Chernozhukov (2011a). "High dıimensional sparse econometric models: An introduction". *Inverse Problems and High-Dimensional Estimation*. Springer, pp. 121–156.

Belloni, A. and V. Chernozhukov (2011b). "☒1-penalized quantile regression in high-dimensional sparse models". *The Annals of Statistics* 39.1, pp. 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, pp. 217–235.

De Mol, C., D. Giannone, and L. Reichlin (2008). "Forecasting using a large number of predictors: Is Bayesian shrinkage a valid alternative to principal components?" *Journal of Econometrics* 146.2, pp. 318–328.

Diebold, F. X. and R. S. Mariano (2002). "Comparing predictive accuracy". *Journal of Business & economic statistics* 20.1, pp. 134–144.

Diebold, F. X. and R. S. Mariano (1995). "Comparing Predictive Accuracy". *Journal of Business and Economic Statistics* 13.3, pp. 253–263.

Faust, J. and J. H. Wright (2013). "Forecasting inflation". *Handbook of economic forecasting*. Vol. 2. Elsevier, pp. 2–56.

Friedman, J. H. (2001). "Greedy function approximation: a gradient boosting machine". *Annals of statistics*, pp. 1189–1232.

Grunfeld, Y. (1960). *The determinants of corporate investment*. University of Chicago Press.

Guitton, H. (1955). "Koyck (LM)-Distributed Lags and Investment Analysis." *Revue économique* 6.6, pp. 127–128.

Harvey, D., S. Leybourne, and P. Newbold (1997). "Testing the equality of prediction mean squared errors". *International Journal of forecasting* 13.2, pp. 281–291.

Hoerl, A. E. and R. W. Kennard (1970). "Ridge regression: Biased estimation for nonorthogonal problems". *Technometrics* 12.1, pp. 55–67.

Ishwaran, H. and J. S. Rao (2005). "Spike and slab variable selection: frequentist and Bayesian strategies". *The Annals of Statistics* 33.2, pp. 730–773.

Kvisgaard, V. H. (2018). "Predicting the future past. How useful is machine learning in economic short-term forecasting?" MA thesis.

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

Liaw, A. and M. Wiener (2002). "Classification and regression by randomForest". *R news* 2.3, pp. 18–22.

Santosa, F. and W. W. Symes (1986). "Linear inversion of band-limited reflection seismograms". *SIAM Journal on Scientific and Statistical Computing* 7.4, pp. 1307–1330.

Stock, J. H. and 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, pp. 267–288.

Tobin, J. (1969). "A general equilibrium approach to monetary theory". *Journal of money, credit and banking* 1.1, pp. 15–29.

Zou, H. (2006). "The adaptive lasso and its oracle properties". *Journal of the American statistical association* 101.476, pp. 1418–1429.

Zou, H. and T. Hastie (2005). "Regularization and variable selection via the elastic net". *Journal of the royal statistical society: series B (statistical methodology)* 67.2, pp. 301–320.

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

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

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

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

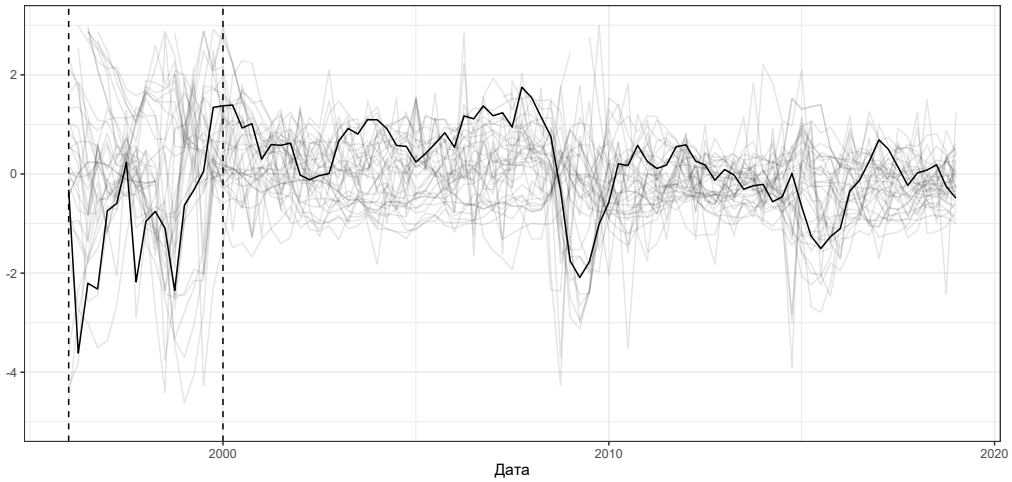
Орешкин, М. С. (2018). “Перспективы экономической политики”. *Экономическая политика* 13.3.

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

Таблица 5: 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 объема 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

Рис. 1: Используемые в работе временные ряды



Приведены трансформированные и стандартизированные значения. Выделены значения прогнозируемой переменной — валового накопления основного капитала. Пунктиром отмечены две левые границы тренировочных выборок.

Рис. 2: Прогнозы изменения инвестиций

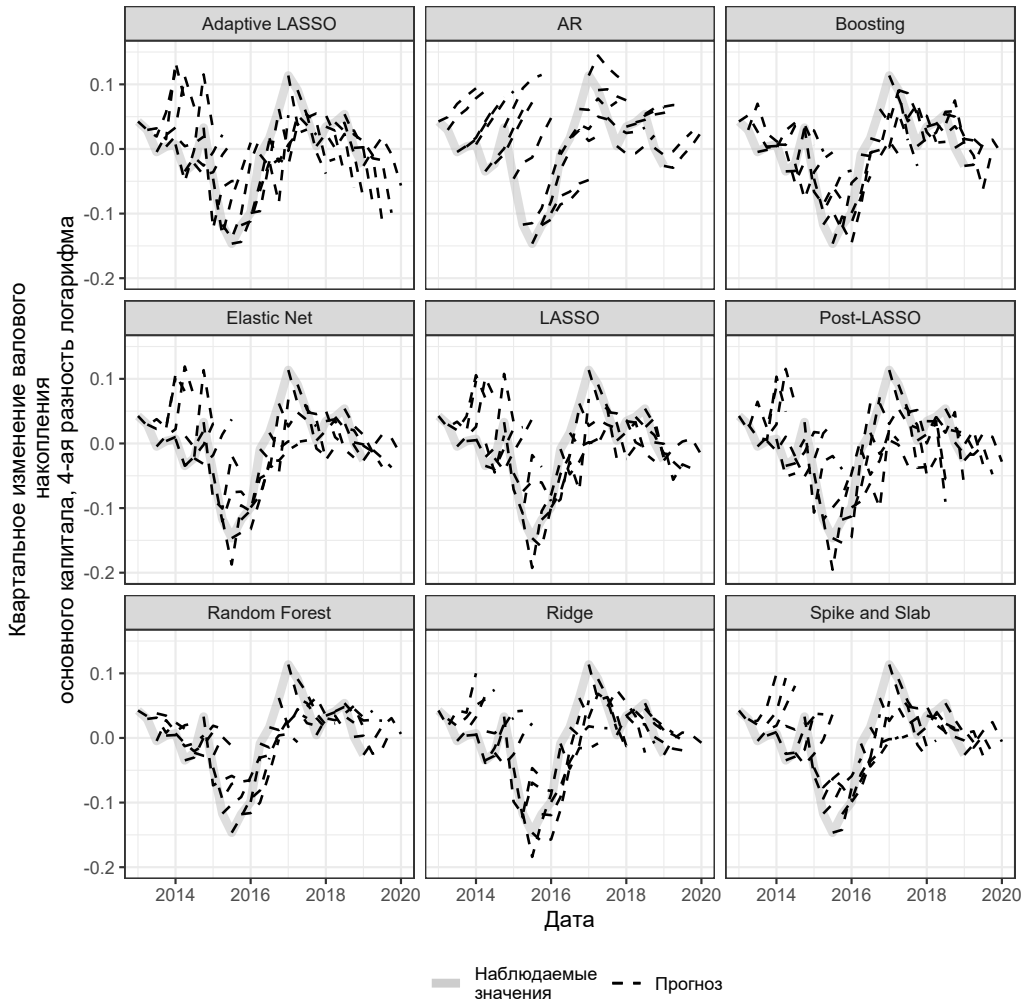
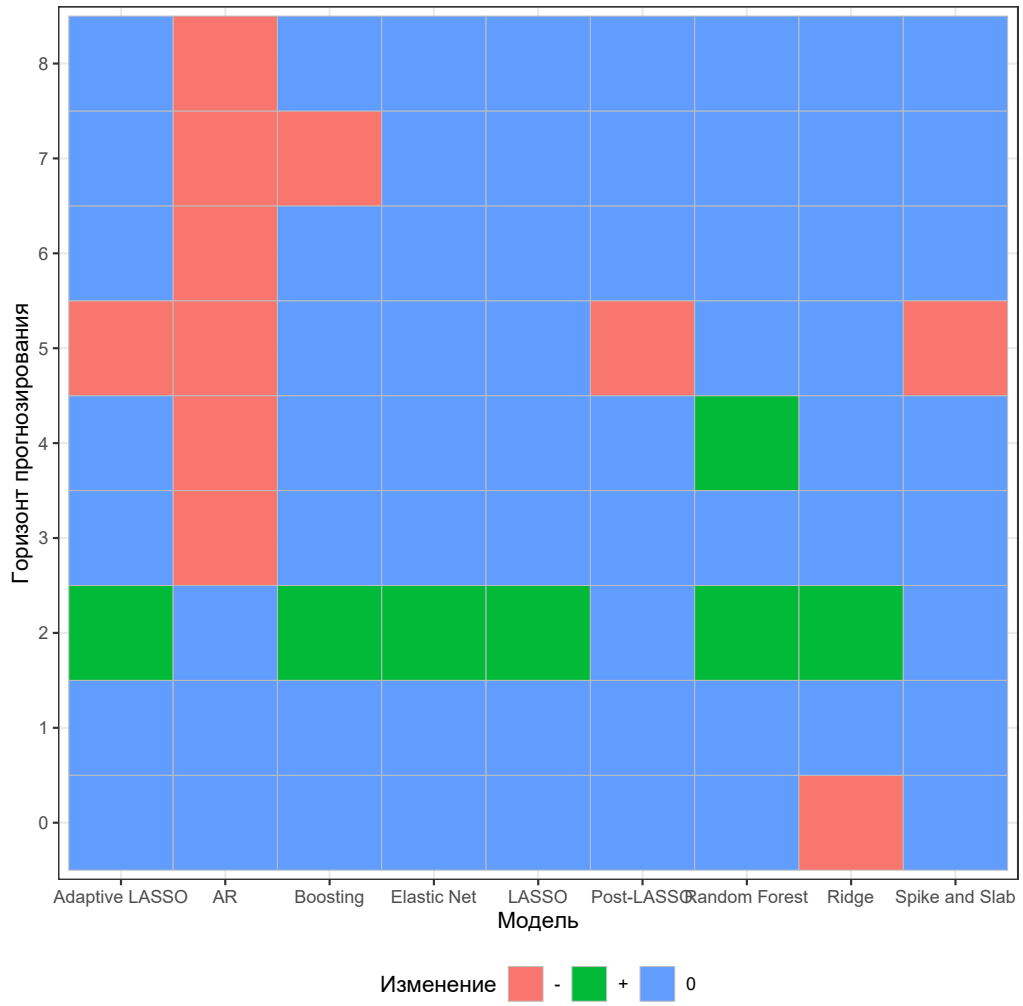


Рис. 3: Тест Диболда-Мариано



H_0: увеличение тренировочной выборки за счет данных до кризиса 1998 г. не изменяет качество моделей

Рис. 4: Количество выбранных переменных в модели LASSO

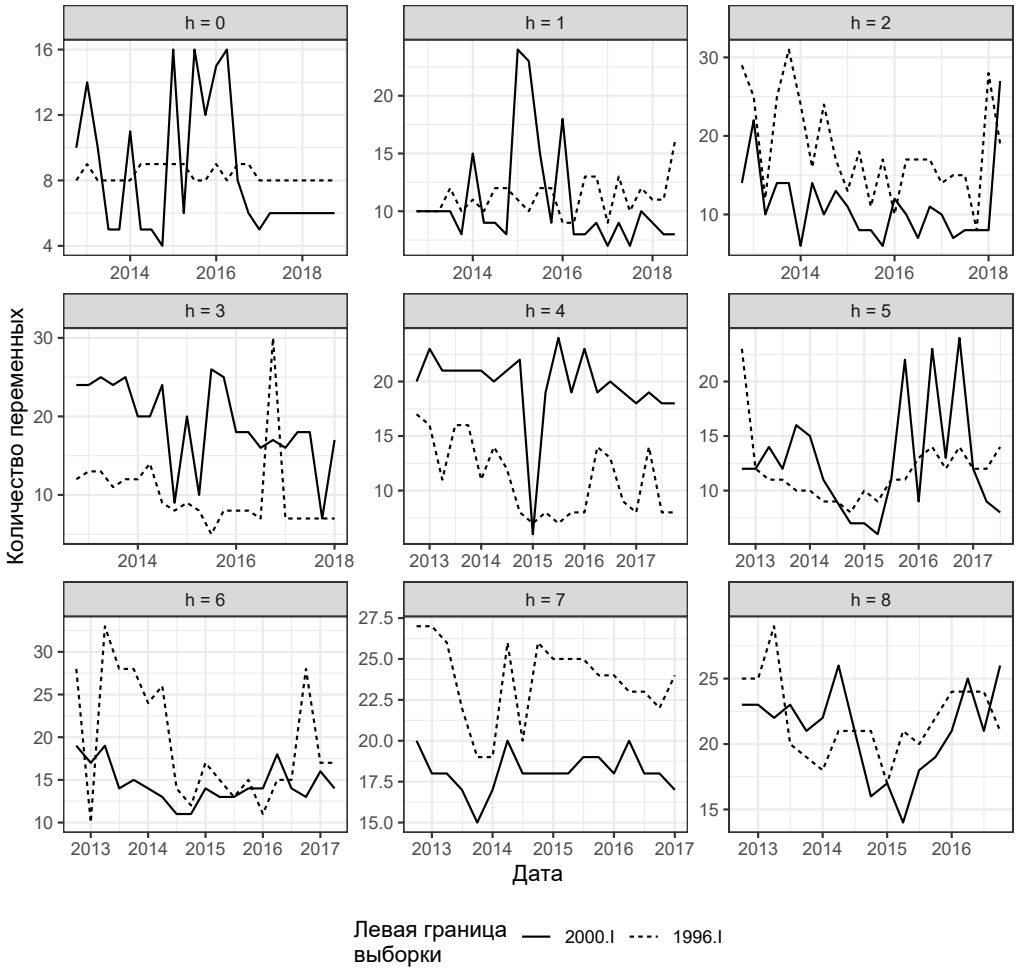


Рис. 5: Коэффициент при ВВП в модели LASSO

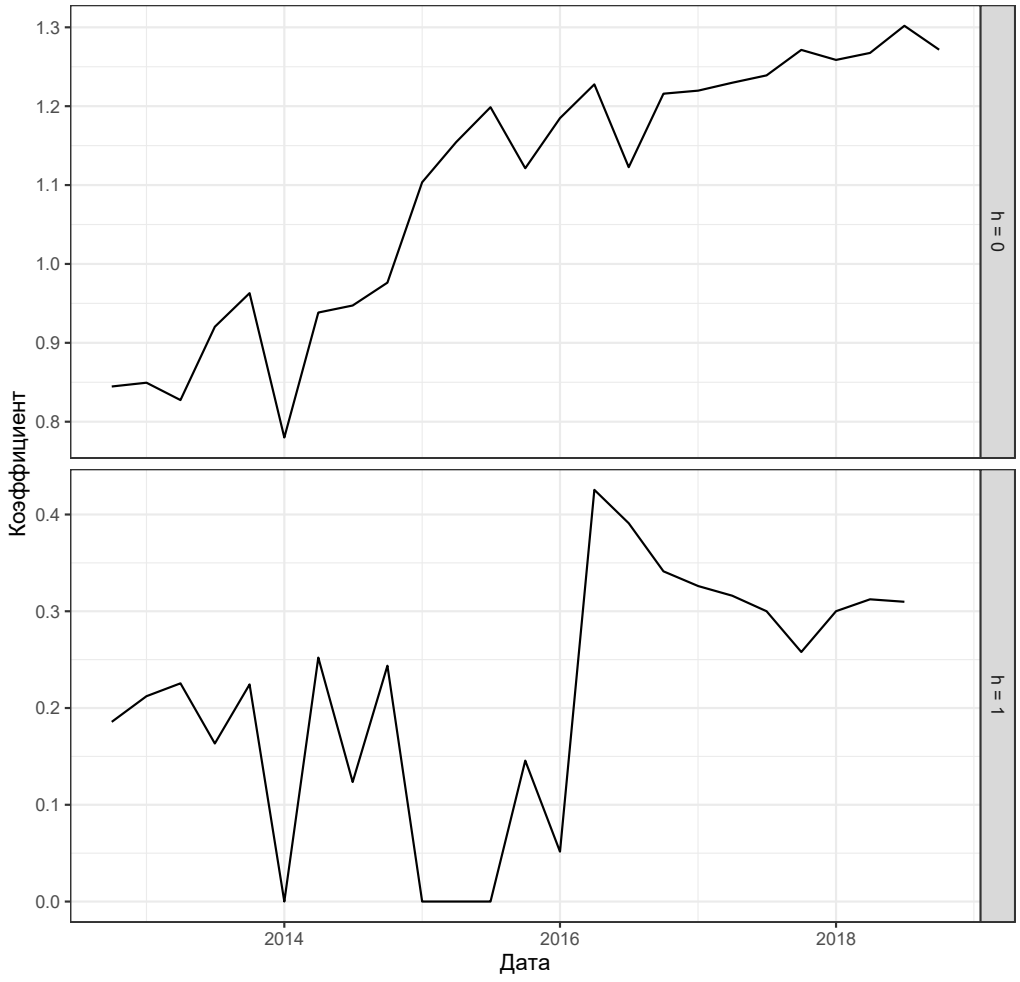


Рис. 6: Коэффициент при инвестициях в модели LASSO

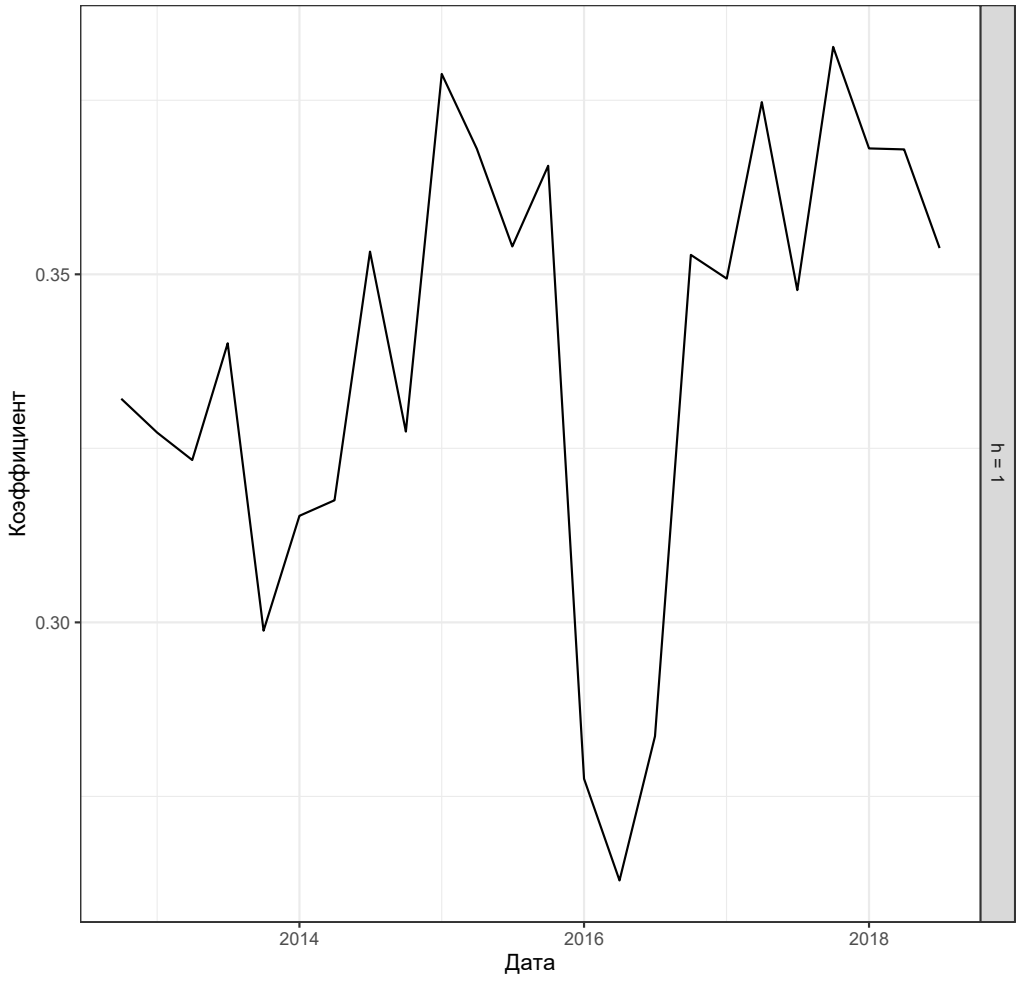


Рис. 7: Коэффициент при доле инвестиций в ВВП в модели LASSO

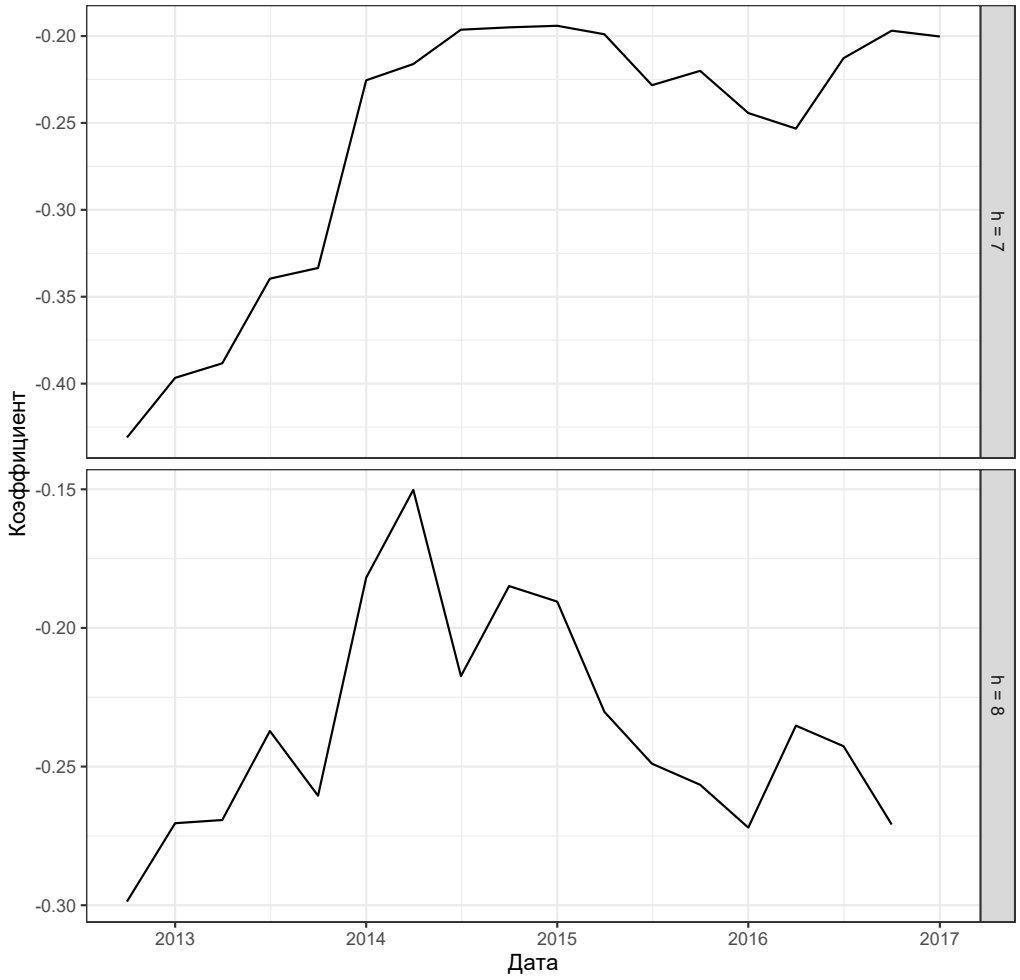


Рис. 8: Коэффициент при индексе РТС в модели LASSO

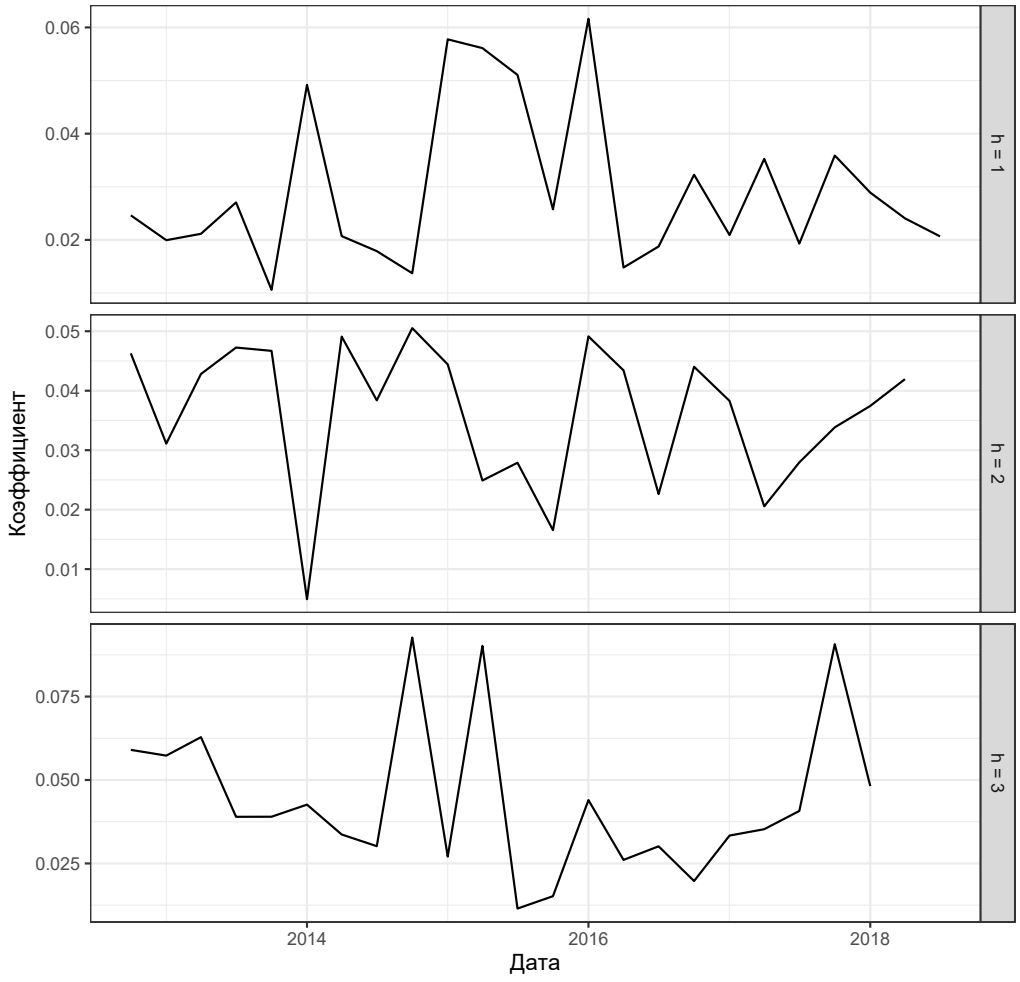


Рис. 9: Коэффициент при 7-дневной ставке МБР в модели LASSO

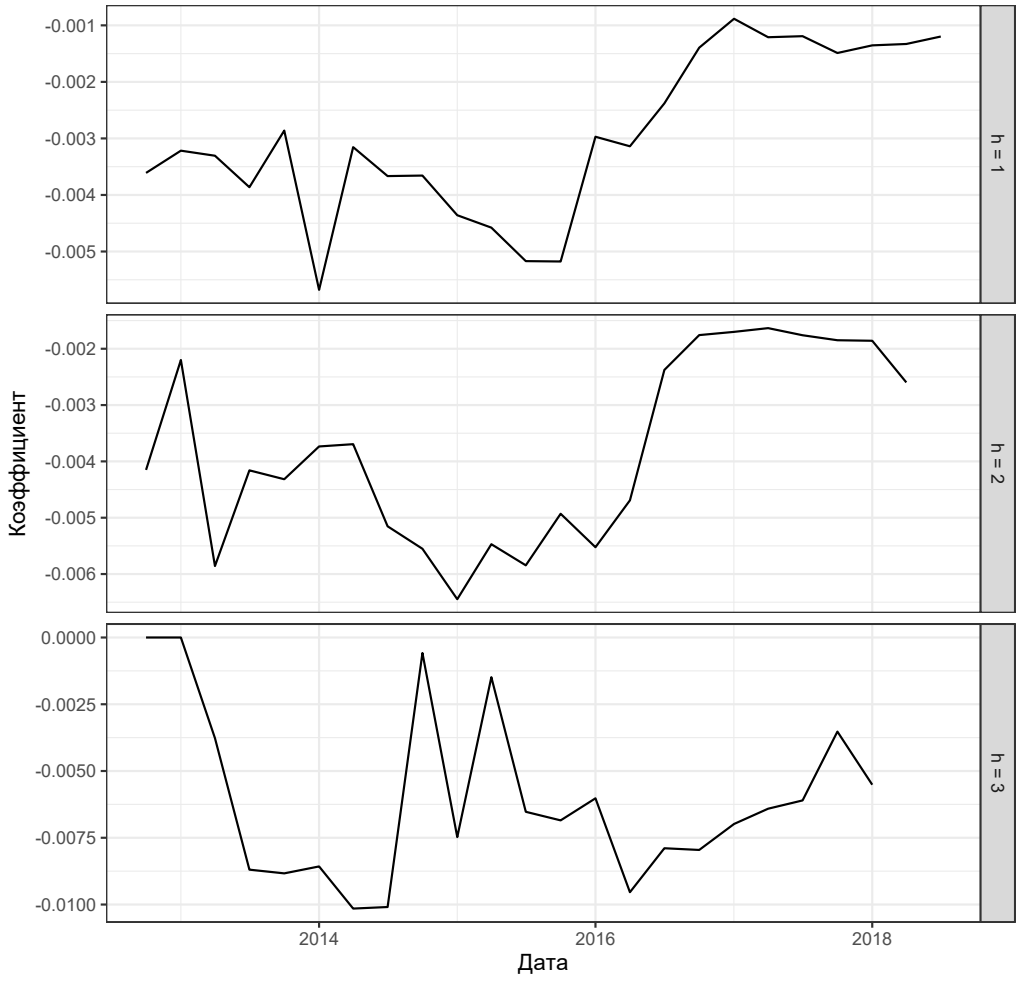
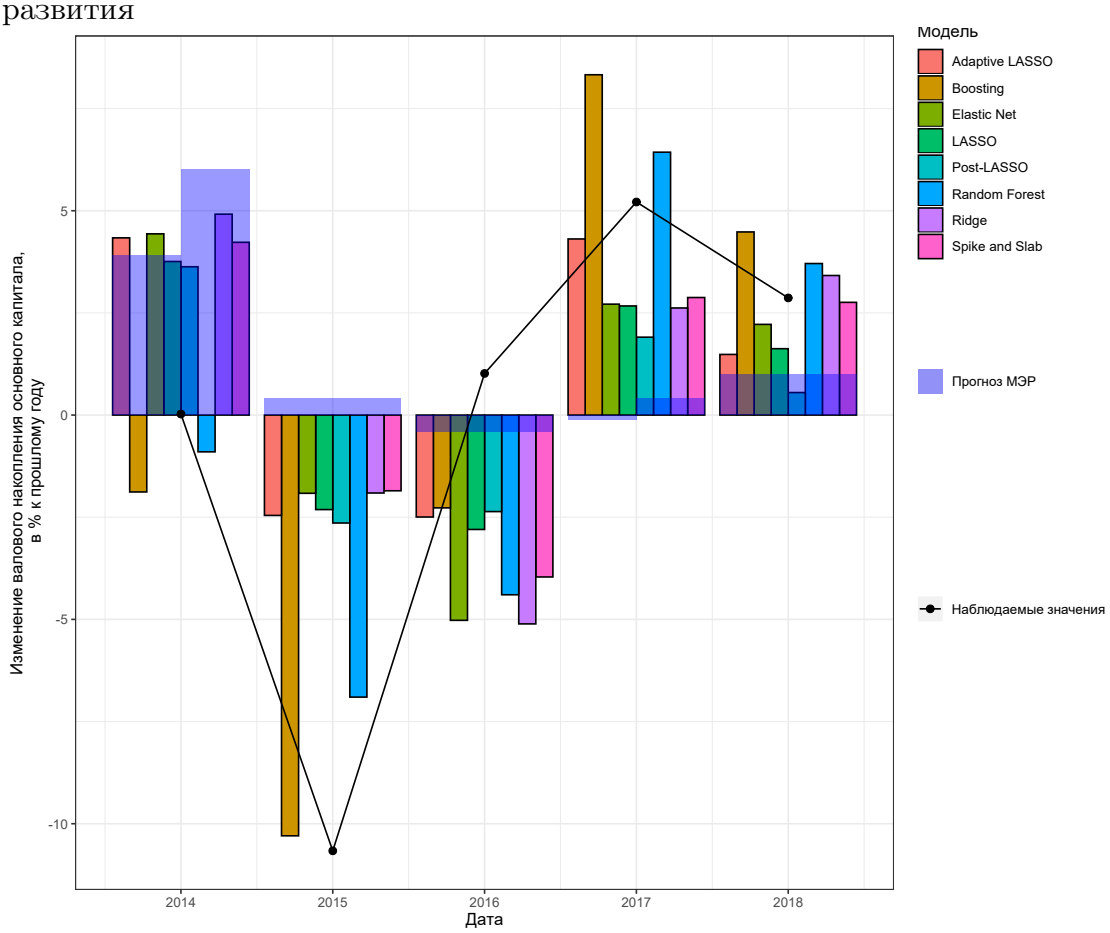


Рис. 10: Годовое изменение инвестиций: прогнозы автора и министерства экономического развития



Дата составления прогноза: 3-ий квартал предыдущего года. На 2014 и 2017 гг. показаны два вида прогнозов МЭР (базовый и консервативный).