ISSN 1814-1196
Научный вестник НГТУ
том 77, № 4, 2019, с. 85–98

http://journals.nstu.ru/vestnik
Science Bulletin of the NSTU
Vol. 77, No. 4, 2019, pp. 85–98

ИНФОРМАТИКА,
ВЫЧИСЛИТЕЛЬНАЯ ТЕХНИКА
И УПРАВЛЕНИЕ

INFORMATICS,
COMPPUTER ENGINEERING
AND MANAGEMENT

УДК 519.254

DOI: 10.17212/1814-1196-2019-4-85-98

Автоматический подбор опережающих индикаторов для прогнозирования состояния регионального рынка труда*

А.Ю. ТИМОФЕЕВА

630073, РФ, г. Новосибирск, пр. Карла Маркса, 20, Новосибирский государственный технический университет

a.timofeeva@corp.nstu.ru

При построении прогнозных моделей на основе опережающих индикаторов возникает проблема отбора наиболее информативных переменных из множества всех потенциальных предикторов. Эта проблема может быть решена встроенными методами, такими как регрессия LASSO, или методами фильтрации, например, с помощью подхода на основе корреляций. В работе ставится задача сравнения эффективности этих методов с альтернативными подходами к анализу временных рядов (модель ARIMA, Хольта–Уинтерса, экспоненциального сглаживания). Для этого предложен алгоритм построения прогнозных моделей, включающий автоматический подбор опережающих индикаторов. Для проведения эмпирического исследования из официальных статистических данных отобраны показатели, пригодные для прогнозирования состояния регионального рынка труда. Они описывают такие индикаторы, как денежная масса, структура баланса кредитных организаций и индекс цен. Производилось псевдовневыборочное прогнозирование ряда показателей, характеризующих ситуацию на регистрируемом рынке труда Новосибирской области за период с 2015 по 2018 г. Использовались прямые многошаговые прогнозы на 6 месяцев вперед. Оказалось, что устойчивая модификация percentile-lasso не дает никаких преимуществ с точки зрения средних абсолютных прогнозных ошибок. В большинстве случаев лучшие результаты получены с помощью регрессии LASSO с выбором параметра регуляризации по правилу одной стандартной ошибки на основе 10-блочной кроссвалидации со случайным формированием блоков. За счет автоматического подбора опережающих индикаторов удалось уменьшить ошибки прогнозирования по сравнению с альтернативными методами. Тем самым предложенный алгоритм признан пригодным к использованию для прогнозирования состояния регионального рынка труда.

Ключевые слова: отбор предикторов, регрессия LASSO, percentile-lasso, отбор признаков на основе корреляций, опережающие индикаторы, рынок труда, прогнозирование, ARIMA, модель Хольта–Уинтерса, алгоритм STL

* *Статья получена 22 сентября 2019 г.*

Исследование выполнено при финансовой поддержке РФФИ/РГНФ, грант № 17-32-01087 а2.

ВВЕДЕНИЕ И ПОСТАНОВКА ЗАДАЧИ

При построении прогнозных моделей на основе опережающих индикаторов количество потенциальных предикторов, как правило, велико. С учетом возможного числа лагов переменных это порождает задачу высокой размерности [1].

Для снижения размерности задачи применяются два принципиально разных подхода. Первый предполагает выделение признаков на основе метода главных компонент [1]. Это позволяет оставить в качестве входных данных только небольшое количество ортогональных переменных. Однако кросс-корреляция регрессоров в больших наборах данных может привести к неточным прогнозам, и, следовательно, меньший набор скорее всего обеспечивает меньшую среднюю ошибку прогноза. Так, в работе [2] обнаружено, что включение большего числа предикторов для оценки главных компонент делает их менее полезными для прогнозирования. Тем самым извлечение признаков для построения прогнозных моделей на основе опережающих индикаторов нежелательно.

Второй подход заключается в отборе наиболее информативных предикторов. Чаще всего отбор предикторов осуществляется совместно с построением прогнозных моделей с привлечением встроенных методов, таких как регрессия LASSO [3]. За счет регуляризации некоторые коэффициенты регрессии обнуляются. Тем самым получается разреженное решение, включающее только существенные признаки. LASSO успешно применяется для тактического прогнозирования продаж с использованием макроэкономических опережающих индикаторов [4]. При прогнозировании макроэкономических показателей LASSO было сочтено полезным для выбора релевантных предикторов [5]. В исследовании [6] с помощью многоступенчатой регрессии LASSO решается проблема сверхвысокой размерности пространства маркетинговых переменных, влияющих на продажи розничного магазина на уровне единиц товарной номенклатуры. Модель LASSO становится все более привлекательной в свете проблем с большими данными [3].

Однако результаты отбора опережающих индикаторов на основе встроенных методов сильно зависят от типа прогнозной модели. В отличие от этого методы фильтрации дают более общий результат, свободный от модельных предположений. Одним из таких многомерных методов является подход на основе корреляций (CFS) [7]. Он обеспечивает компромисс между релевантностью признаков с точки зрения их влияния на отклик и их избыточностью в смысле взаимной корреляции. Этот подход предложен для решения задач классификации. Его применимость к отбору признаков в условиях их сильной корреляции, в частности при подборе опережающих индикаторов, показана в работе [8].

Здесь ставится задача сравнения эффективности встроенных методов (регрессии LASSO) и методов фильтрации (отбор на основе корреляций CFS) по сравнению с альтернативными подходами к прогнозированию состояния регионального рынка труда. Для этого предложен алгоритм построения прогнозных моделей на основе автоматического отбора опережающих индикаторов.

1. РЕГРЕССИЯ LASSO

В условиях переизбытка входных данных оценки регрессии, полученные методом наименьших квадратов (МНК), часто оказываются нестабильными. Во избежание этого используется метод регуляризации, который заключается в наложении дополнительных ограничений на переменные для предотвращения излишней сложности модели. Смыслом процедуры является сжатие вектора коэффициентов регрессии β при ее подгонке так, чтобы в среднем эти коэффициенты оказались по абсолютной величине несколько меньше, чем это было бы при оптимизации по МНК.

LASSO минимизирует сумму квадратов ошибок, накладывая ограничение на сумму абсолютных значений параметров модели. Оценка определяется решением задачи оптимизации:

$$\hat{\beta} = \arg\min \left[\sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{m} \beta_j x_{ij} \right)^2 + \lambda \|\beta\| \right], \qquad (1)$$

где y_i, x_{ij} – i-е наблюдаемое значение отклика j-го фактора; $\beta = (\beta_1, ..., \beta_m)$ – вектор параметров; n – объем выборки; m – число факторов; λ – параметр регуляризации, который контролирует величину штрафа за сложность.

При этом достигается некоторый компромисс между ошибкой регрессии и размерностью используемого признакового пространства, выраженного суммой абсолютных значений коэффициентов $\|\beta\|$. В ходе минимизации (1) некоторые коэффициенты сокращаются до нуля. Таким образом, интерпретируемость модели увеличивается путем удаления нерелевантных переменных, которые не связаны с откликом.

Результаты оценивания чувствительны к выбору параметра регуляризации. Для его выбора широко используются два подхода. Значение, которое дает минимальную среднюю ошибку при перекрестной проверке, обозначается через $\lambda = \min$. Правило одного стандартного отклонения ($\lambda = 1\text{se}$) является альтернативным способом выбора значения параметра регуляризации. Идея состоит в том, чтобы взять простейшую (наиболее разреженную) модель, ошибка которой находится в пределах одного стандартного отклонения от минимальной ошибки [3].

Процедура перекрестной проверки (кросс-валидации) предполагает обучение и тестирование модели на разных наборах данных. Это позволяет ослабить эффект переобучения. Поскольку в случае с временными рядами макроэкономических показателей, которые довольно короткие, получение дополнительных данных затруднительно или невозможно, то имеющийся объем данных разбивается на части для оценивания и контроля ошибки. В зависимости от того, каким образом происходит это разбиение, можно выделить несколько наиболее распространенных процедур: кросс-валидация по K блокам, валидация последовательным случайным сэмплированием, поэлементная кросс-валидация (LOOCV).

Комбинация LASSO и перекрестной проверки со случайным разбиением оказывается нестабильной в том смысле, что повторное применение процедуры часто дает разные результаты [9]. Более того, устойчивость регуляриза-

ции зависит от свойств набора данных. Поэтому необходимо быть осторожным при использовании LASSO как метода построения прогнозной модели.

В работе [10] показано, что модель, выбранная с помощью LASSO, может быть чрезвычайно чувствительной к выбору блоков, используемых для перекрестной проверки. Следствием этой чувствительности является то, что результаты анализа могут лишиться интерпретируемости. Чтобы преодолеть такую нестабильность отбора модели с помощью LASSO, в [10] предлагается метод, называемый percentile-lasso. Модель, получаемая с помощью percentile-lasso, соответствует LASSO регрессии, оцененной с использованием соответствующего перцентиля возможных «оптимальных» значений параметра регуляризации. Показано, что percentile-lasso может значительно улучшить как стабильность выбора модели, так и ошибку выбора модели по сравнению с LASSO. Важно отметить, что при применении к реальным данным percentile-lasso, в отличие от LASSO, дает интерпретируемые результаты, то есть результаты, которые являются робастными к выбору блоков наблюдений для перекрестной проверки.

2. ОТБОР НА ОСНОВЕ КОРРЕЛЯЦИЙ

Метод CFS (Correlation-based Feature Selection) предложен в работе [7]. Признаки выбираются так, чтобы обеспечить наибольшую корреляцию с откликом и наименьшую взаимосвязь между самими признаками. Тем самым решается следующая оптимизационная задача:

$$\frac{\sum\limits_{i\in S_k} R_i}{\sqrt{k+2\sum\limits_{i,j\in S_k, i\neq j} r_{ij}}} \to \max\limits_{S_k}, \tag{2}$$

где R_i – абсолютное значение коэффициента корреляции между i-м признаком и откликом; r_{ij} – абсолютное значение коэффициента корреляции между i-м и j-м признаком; S_k – подмножество из k признаков. При анализе временных рядов экономических показателей чаще всего как отклик, так и предикторы измеряются в количественных шкалах, поэтому допустимо использование обычного коэффициента корреляции Пирсона.

Представим задачу (2) как задачу нелинейного целочисленного программирования:

$$\frac{\sum\limits_{i-1}^{m} R_i^2 x_i + 2\sum\limits_{i\neq j} R_i R_j x_i x_j}{\sum\limits_{i-1}^{m} x_i + 2\sum\limits_{i\neq j} r_{ij} x_i x_j} \to \max\limits_{x_1,\ldots,x_m}, \tag{3}$$

где $x_i \in \{0,1\}$, $i = 1,..., m$. Если $x_i = 1$, то i-й признак входит в оптимальный набор, иначе $x_i = 0$. Задача (3) является дробно-полиномиальной задачей целочисленного программирования. Поиск глобального оптимума путем перебора всех комбинаций приводит к NP-полной задаче. На практике получили

распространение процедуры пошагового отбора переменных, которые позволяют снизить количество вычислений, но не обеспечивают достижения оптимального набора входных переменных ввиду «жадных» стратегий. Такие алгоритмы считаются эвристическими, поскольку не являются гарантированно точными или оптимальными, но достаточны для решения поставленной задачи. Далее используется эвристический алгоритм прямого поиска.

3. АЛГОРИТМ ПОСТРОЕНИЯ ПРОГНОЗНЫХ МОДЕЛЕЙ

Особенностью анализа временных рядов является присутствие неслучайных составляющих (тренда, сезонности, цикличности). Предлагается вычислять значения показателей в процентах к соответствующему месяцу предыдущего года. Такое преобразование позволяет сразу исключить нелинейный долгосрочный тренд (при его наличии) и сезонность, а также обнаружить во временном ряду изменения, вызванные циклическими колебаниями, связанными с кризисом. Однако при этом из исходных данных теряется целый год, что может быть существенно при ограниченном размере временного ряда.

Такая проблема возникла с показателями регионального рынка труда. Из-за ограниченного объема исходных данных их предварительная обработка включала только исключение сезонной компоненты. Для этого предлагается использовать один из способов разделения временного ряда на компоненты тренда, сезонности и остатки – алгоритм STL с использованием Loess [11].

При осуществлении прогнозирования на основе регрессионных моделей предполагается использовать не итеративные, а прямые многошаговые прогнозы. Для построения таких прогнозов в модели с распределенными лагами все переменные лагируются на h периодов для получения прогноза на h шагов вперед. Горизонт прогнозирования h фиксирован. Это позволяет использовать коэффициенты регрессии непосредственно для вычисления прогноза без итераций.

Тогда алгоритм построения прогнозной модели и вычисления прогнозов на основе опережающих индикаторов можно представить в следующем виде.

Шаг 1. Декомпозиция временного ряда показателя регионального рынка труда y_t с помощью алгоритма STL:

$$y_t = \tilde{y}_t + S_t,$$

где S_t – сезонная компонента; $\tilde{y}_t$ – временной ряд за исключением сезонной компоненты (тренд, сглаженный с помощью Loess, и остатки).

Шаг 2. Построение прогнозной модели на основе опережающих индикаторов:

$$\tilde{y}_t = \alpha + \sum_{k=1}^{K} \beta_k x_{t-h}^{(k)} + \varepsilon, \qquad (4)$$

где α, β_k – параметры, подлежащие оцениванию; h – глубина запаздывания; $x_{t-h}^{(k)}$ – k-й экономический показатель, рассматриваемый как опережающий

индикатор, представленный в процентах к соответствующему месяцу предыдущего года, взятый с лагом h; ε – случайная ошибка.

Шаг 3. Построение прогноза на h периодов вперед:

$$\hat{\tilde{y}}_{T+h} = \hat{\alpha} + \sum_{k=1}^{K} \hat{\beta}_k x_T^{(k)},$$

где $\hat{\alpha}$, $\hat{\beta}_k$ – оценки параметров; T – длина временного ряда.

Шаг 4. Добавление оценки сезонности $\hat{S}_t$ и получение итогового прогноза:

$$\hat{y}_{T+h} = \hat{\tilde{y}}_{T+h} + \hat{S}_t.$$

На шаге 2 используются описанные выше методы автоматического отбора предикторов. При использовании регрессии LASSO в модель (4) включаются все возможные индикаторы. При использовании метода CFS модель (4) строится только по отобранному подмножеству признаков.

Для оценки качества прогнозов, построенных с помощью опережающих индикаторов, имеет смысл сравнить эти результаты с альтернативными вариантами прогнозирования, основанными только на динамике самих показателей рынка труда. В качестве альтернативных подходов для полученных временных рядов показателей регионального рынка труда с исключенной сезонностью строились модели ARIMA, экспоненциального сглаживания (ES) и Хольта–Уинтерса [12].

Предложенный алгоритм реализован в статистической среде RStudio [13]. Алгоритм STL реализован через функцию stl в стандартном пакете stats. Для автоматического подбора структуры ARIMA-модели использовалась функция auto.arima{forecast}. Оптимальные значения параметров модели Хольта–Уинтерса найдены путем минимизации квадрата ошибки предсказания на один шаг вперед с помощью функции HoltWinters{stats}. С помощью функции forecast из одноименного пакета осуществлялось прогнозирование на h шагов вперед как по модели, оцененной по методологии ARIMA, так и по модели Хольта–Уинтерса. Далее эта возможность использовалась для сопоставления результатов прогнозирования с помощью опережающих индикаторов с альтернативными методами.

4. РЕЗУЛЬТАТЫ ЭМПИРИЧЕСКОГО ИССЛЕДОВАНИЯ

Статистическая информация о ситуации на регистрируемом рынке труда взята с сайта Федеральной службы по труду и занятости [14]. Выбраны следующие прогнозируемые переменные:

y_1 – численность безработных граждан (в тыс. чел.);

y_2 – заявленная работодателями потребность в работниках;

y_3 – уровень регистрируемой безработицы;

y_4 – коэффициент напряженности на рынке труда.

Из базы данных ЕМИСС [15] выбраны показатели, которые можно отнести к опережающим индикаторам. Одним из критериев выбора было наличие данных с ежемесячной динамикой. В результате выделены три основных индикатора: денежная масса, структура баланса кредитных организаций и индекс цен.

Показатели общенационального масштаба для индикатора «Денежная масса»:

– денежный агрегат М2;

– денежный агрегат М1;

– денежный агрегат М0;

– переводные депозиты населения;

– переводные депозиты нефинансовых и финансовых (кроме кредитных) организаций.

Показатели общенационального масштаба для индикатора «Структура баланса кредитных организаций»:

– чистые иностранные активы;

– требования к нерезидентам;

– обязательства перед нерезидентами;

– требования к органам государственного управления субъектов РФ и органам местного самоуправления;

– требования к другим финансовым организациям;

– обязательства перед Центральным банком;

– депозиты, включаемые в широкую денежную массу;

– депозиты, не включаемые в широкую денежную массу.

Показатели регионального масштаба для индикатора «Индекс цен»:

– базовый индекс потребительских цен на товары и услуги;

– индексы цен приобретения машин и оборудования инвестиционного назначения (по отраслям).

При использовании региональных показателей в качестве региона взята Новосибирская область (НСО), поскольку основной целью выступало прогнозирование показателей рынка труда НСО.

Всем перечисленным показателям соответствует по одной переменной, кроме индексов цен приобретения машин и оборудования инвестиционного назначения. Доступны данные по 92 видам экономической деятельности, в том числе по 16 разделам ОКВЭД, и суммарный индекс по всем видам деятельности.

Для оценки качества моделей, построенных разными методами, осуществлялось псевдовневыборочное прогнозирование. Для этого исходный временной ряд разделялся на обучающий набор, по которому строились модели, и тестовый набор, для которого осуществлялось прогнозирование. Исходный временной ряд насчитывал 78 месяцев (с января 2012 г. по июнь 2018 г.).

Для того чтобы снизить случайность полученных результатов, псевдовневыборочное прогнозирование производилось многократно. С этой целью обучающий набор постоянно расширялся на одну точку данных. Первоначальный обучающий набор включал данные за 37 месяцев: с января 2012 г. по январь 2015 г. включительно. Следующий набор включал уже 38 точек: с января 2012 г. по февраль 2015 г. включительно. И так далее вплоть до декабря 2017 г. Таким образом, получилось всего 36 наборов.

Горизонт прогнозирования выбран равным шести месяцам, $h = 6$. Следовательно, из тестового набора использовалось одно наблюдение, соответствующее шести месяцам вперед. Для первого набора это был июль 2015 г., а для последнего набора – июнь 2018 г. (последняя точка данных).

Далее представлены результаты псевдовневыборочного прогнозирования в виде средних абсолютных прогнозных ошибок (MAE) для каждого года с 2015 по 2018 г.:

$$MAE_j = \frac{1}{n(T_j)} \sum_{T_j} \left| \hat{y}_{T_j+h|T_j} - y_{T_j+h} \right|, \quad j = 2012, \ldots, 2018,$$

где $\hat{y}_{T_j+h|T_j}$ – прогноз для периода $T_j + h$ по модели, построенной по данным, доступным до момента T_j включительно, $T_{2015} = 37, \ldots, 42$, $T_j = 43 + 12(j - 2016), \ldots, 54 + 12(j - 2016)$, $j = 2016,\ 2017$, $T_{2018} = 67, \ldots, 72$; $n(T_j)$ – число месяцев в j-м году, для которых строился прогноз.

Исходный набор потенциальных опережающих индикаторов состоял из 55 переменных, поскольку после очистки данных индексы цен приобретения машин и оборудования инвестиционного назначения остались только по 40 видам экономической деятельности (в том числе по разделам ОКВЭД) и суммарный индекс по всем видам деятельности.

Для отбора существенных предикторов сначала использовалась регрессия LASSO. Как отмечалось выше, результаты оценивания модели (4) путем решения задачи (1) чувствительны к выбору параметра регуляризации. Произведено сравнение двух вариантов выбора параметра – на основе минимума кросс-валидации и по правилу одной стандартной ошибки. Процедура перекрестной проверки реализовывалась двумя альтернативными способами: LOOCV (поэлементная) и 10-блочная со случайным формированием блоков. Следовательно, рассматривалось четыре варианта: минимум поэлементной кросс-валидации (LOOCV $\lambda = \min$), одна стандартная ошибка поэлементной кросс-валидации (LOOCV $\lambda = 1\text{se}$), минимум 10-блочной кросс-валидации (10-foldCV $\lambda = \min$), одна стандартная ошибка 10-блочной кросс-валидации (10-foldCV $\lambda = 1\text{se}$).

В табл. 1 сравниваются средние абсолютные ошибки прогнозирования при разных вариантах оценивания регрессии LASSO для четырех анализируемых показателей рынка труда. В целом результаты не сильно отличаются. Однако для всех показателей наиболее существенные различия наблюдались в 2015 г. Для численности безработных и уровня безработицы лучше работает правило одной стандартной ошибки и LOOCV, для потребности в работниках и коэффициента напряженности на рынке труда – минимум кросс-валидации. Если обобщать результаты, то для большинства показателей почти за весь период наименьшие ошибки давала 10-блочная кросс-валидация и выбор параметра регуляризации по правилу одной стандартной ошибки.

Однако 10-блочная кросс-валидация со случайным формированием блоков дает случайные результаты, следовательно, нестабильна. Для получения более устойчивых прогнозов использован новый подход percentile-lasso.

В работе [10] предлагается задавать уровень квантиля для оптимального параметра регуляризации в диапазоне от 0,75 до 1. Оказалось, что выбор квантиля существенно не сказывается на результатах прогнозирования состояния рынка труда. Явные расхождения наблюдаются только в 2015 г. Причем для численности безработных лучший результат обеспечивает выбор квантиля уровня 0,95. Для других показателей, наоборот, наименьшее значение квантиля предпочтительнее.

Таблица 1

Table 1

MAE при разных процедурах кросс-валидации для LASSO

MAE under various cross-validation procedures for LASSO

Отклик	Процедура CV	2015 г.	2016 г.	2017 г.	2018 г.
y_1	10-foldCV $\lambda = \min$	2,134	0,763	0,967	1,794
	10-foldCV $\lambda = 1se$	1,328	**0,660**	**0,926**	**1,563**
	LOOCV $\lambda = \min$	2,059	0,771	0,983	1,889
	LOOCV $\lambda = 1se$	**1,245**	0,666	0,937	1,668
y_2	10-foldCV $\lambda = \min$	**3887**	2641	2089	2436
	10-foldCV $\lambda = 1se$	6546	**2168**	**1960**	**2119**
	LOOCV $\lambda = \min$	4060	2653	402	2414
	LOOCV $\lambda = 1se$	7384	2214	2096	2173
y_3	10-foldCV $\lambda = \min$	0,199	0,051	0,070	**0,074**
	10-foldCV $\lambda = 1se$	0,132	0,054	0,069	0,087
	LOOCV $\lambda = \min$	0,190	**0,050**	**0,068**	0,079
	LOOCV $\lambda = 1se$	**0,129**	0,055	**0,068**	0,083
y_4	10-foldCV $\lambda = \min$	0,216	0,113	0,064	0,053
	10-foldCV $\lambda = 1se$	0,242	**0,065**	**0,054**	**0,049**
	LOOCV $\lambda = \min$	**0,147**	0,114	0,085	0,053
	LOOCV $\lambda = 1se$	0,267	0,071	0,066	0,051

В табл. 2 объединены результаты прогнозирования по percentile-lasso при заданных уровнях квантилей оптимального параметра регуляризации. Если сравнивать эти результаты с обычной реализацией LASSO (табл. 1), то они в большинстве случаев оказываются хуже или примерно такие же.

Следовательно, percentile-lasso не дает существенных преимуществ в качестве прогнозирования, однако требует гораздо большего вычислительного времени. Поэтому далее результаты LASSO с 10-блочной кросс-валидацией со случайным формированием блоков и выбором параметра регуляризации по правилу одной стандартной ошибки признаются как лучшие.

Таблица 2

Table 2

MAE для percentile-lasso

MAE for percentile-lasso

Отклик	Уровень квантиля	2015 г.	2016 г.	2017 г.	2018 г.
y_1	0,95	1,613	0,730	0,950	1,772
y_2	0,8	4494	2578	2084	2407
y_3	0,8	0,150	0,050	0,065	0,070
y_4	0,75	0,209	0,120	0,064	0,051

Сопоставим результаты использования метода CFS с альтернативными методами (табл. 3). Преимущества отбора на основе корреляций проявляются при прогнозировании уровня безработицы и коэффициента напряженности на рынке труда. При сравнении этих результатов со средними ошибками регрессии LASSO10-foldCV $\lambda = 1\text{se}$ (см. табл. 1) оказывается, что если не принимать во внимание несколько провальных прогнозов, LASSO обеспечивает лучшие прогнозы. Для показателя уровня регистрируемой безработицы LASSO имеет абсолютное преимущество. Для показателя заявленной работодателями потребности в работниках LASSO обеспечивает лучшие результаты на всем горизонте прогнозирования, кроме 2015 г. Коэффициент напряженности на рынке труда лучше всего прогнозируется с помощью LASSO в 2016–2017 гг., а численность безработных – в 2015–2016 гг.

Таблица 3

Table 3

MAE для CFS и альтернативных методов

MAE for CFS and alternative methods

Отклик	Метод прогнозирования	2015 г.	2016 г.	2017 г.	2018 г.
y_1	Модель ARIMA	2,254	4,314	2,147	1,328
	Модель Хольта–Уинтерса	2,300	2,159	**0,893**	**0,660**
	ES	**1,341**	2,895	1,583	0,926
	CFS	2,069	**1,019**	0,905	1,694

Окончание табл. 3

End of Tab. 3

Отклик	Метод прогнозирования	2015 г.	2016 г.	2017 г.	2018 г.
y_2	Модель ARIMA	6361	3943	3060	2455
	Модель Хольта–Уинтерса	2770	2945	2684	4621
	ES	**1760**	**2482**	**2596**	3404
	CFS	17620	3215	2853	**2251**
y_3	Модель ARIMA	0,235	0,116	0,123	0,103
	Модель Хольта–Уинтерса	0,203	0,174	0,160	**0,090**
	ES	**0,183**	0,066	0,136	0,119
	CFS	0,216	**0,063**	**0,072**	**0,090**
y_4	Модель ARIMA	0,162	0,117	0,102	0,068
	Модель Хольта–Уинтерса	0,132	0,135	0,080	0,058
	ES	**0,088**	0,114	0,096	**0,031**
	CFS	0,440	**0,097**	**0,072**	0,036

Таким образом, показана возможность прогнозирования состояния рынка труда НСО на основе опережающих индикаторов путем их автоматического подбора на основе регрессии LASSO. Выделенные экономические показатели на полгода раньше по сравнению с уровнем безработицы реагируют на циклические изменения в экономике. Это позволяет предсказать изменения на рынке труда на основе динамики показателей денежной массы, структуры баланса кредитных организаций и индексов цен.

ЗАКЛЮЧЕНИЕ

Таким образом, в работе предложен алгоритм построения прогнозных моделей, описывающих состояние регионального рынка труда, новизна которого состоит в автоматическом подборе опережающих индикаторов. Для этого использованы методы отбора признаков на базе фильтров (отбор на основе корреляций) и встроенные методы (регрессия LASSO). В ходе эмпирического исследования сравнивалось качество прогнозирования с использованием предложенного алгоритма и альтернативными методами (с помощью моделей ARIMA, Хольта–Уинтерса, экспоненциального сглаживания). Исследованы различные процедуры кросс-валидации, применяемые для выбора параметра регуляризации в регрессии LASSO, а также устойчивая модификация percentile-lasso. Оказалось, что лучший результат обеспечивает регрессия LASSO с 10-блочной кросс-валидацией со случайным формированием блоков и выбором параметра регуляризации по правилу одной стандартной ошибки. Сопоставление результатов прогнозирования с помощью опережающих индикато-

ров с альтернативными методами показало преимущества использования индикаторов для долгосрочного прогноза (на полгода вперед), особенно при прогнозировании уровня безработицы в НСО.

СПИСОК ЛИТЕРАТУРЫ

1. *Stock J., Watson M.* Forecasting using principal components from a large number of predictors // Journal of the American Statistical Association. – 2002. – Vol. 297. – P. 1167–1179. – DOI: 10.1198/016214502388618960.

2. Boivin J., Ng S. Are more data always better for factor analysis? // Journal of Econometrics. – 2006. – Vol. 132, N 1. – P. 169–194. – DOI: 10.1016/j.jeconom.2005.01.027.

3. *Tibshirani R.* Regression shrinkage and selection via the lasso: a retrospective // Journal of the Royal Statistical Society: Series B (Statistical Methodology). – 2011. – Vol. 73, N 3. – P. 273–282. – DOI: 10.1111/j.1467-9868.2011.00771.x.

4. Tactical sales forecasting using a very large set of macroeconomic indicators / Y.R. Sagaert, E.H. Aghezzaf, N. Kourentzes, B. Desmet // European Journal of Operational Research. – 2018. – Vol. 264, N 2. – P. 558–569. – DOI: 10.1016/j.ejor.2017.06.054.

5. *Bulligan G., Marcellino M., Venditti F.* Forecasting economic activity with targeted predictors // International Journal of Forecasting. – 2015. – Vol. 31, N 1. – P. 188–206. – DOI: 10.1016/j.ijforecast.2014.03.004.

6. *Ma S., Fildes R., Huang T.* Demand forecasting with high dimensional data: the case of SKU retail sales forecasting with intra-and inter-category promotional information // European Journal of Operational Research. – 2016. – Vol. 249, N 1. – P. 245–257. – DOI: 10.1016/j.ejor.2015.08.029.

7. *Hall M.A.* Correlation-based feature selection for machine learning: PhD thesis. – Hamilton: University of Waikato, 1999.

8. *Timofeeva A.Y., Mezentsev Y.A.* Forecasting using predictor selection from a large set of highly correlated variables // CEUR Workshop Proceedings. – 2019. – Vol. 2416: Information Technology and Nanotechnology: Data Science. – P. 10–18.

9. *Lund K.V.* The Instability of cross-validated LASSO: Master's thesis / Faculty of Mathematics and Natural Sciences, University of Oslo. – Oslo, 2013.

10. *Roberts S., Nowak G.* Stabilizing the lasso against cross-validation variability // Computational Statistics and Data Analysis. – 2014. – Vol. 70. – P. 198–211. – DOI: 10.1016/j.csda.2013.09.008.

11. STL: a seasonal-trend decomposition procedure based on loess / R.B. Cleveland, W.S. Cleveland, J.E. McRae, I. Terpenning // Journal of Official Statistics. – 1990. – Vol. 6. – P. 3–73.

12. *Brockwell P.J., Davis R.A., Calder M.V.* Introduction to time series and forecasting. – New York: Springer, 2002. – 425 p.

13. Open source and enterprise-ready professional software for data science [Electronic resource]. – URL: https://rstudio.com/ (accessed: 12.12.2019).

14. Статистическая информация о ситуации на регистрируемом рынке труда [Электронный ресурс] // Роструд: web-сайт. – URL: https://www.rostrud.ru/rostrud/deyatelnost/?CAT_ID=6293 (accessed: 12.12.2019).

15. ЕМИСС – Единая межведомственная информационно-статистическая система [Электронный ресурс]: web-сайт. – URL: https://fedstat.ru/ (accessed: 12.12.2019).

Тимофеева Анастасия Юрьевна, кандидат экономических наук, доцент кафедры теоретической и прикладной информатики Новосибирского государственного технического университета. Основное направление научных исследований – развитие методов статистического анализа объектов стохастической природы, в том числе социально-экономических явлений. Имеет более 100 научных публикаций. E-mail: a.timofeeva@corp.nstu.ru.

Timofeeva Anastasia Yuryevna, PhD (Eng.), an associate professor at the theoretical and applied informatics department, Novosibirsk State Technical University. The main direction of scientific research is of methods for statistical analysis of objects of stochastic nature, including socio-economic phenomena. He is the author of more than 100 publications. E-mail: a.timofeeva@corp.nstu.ru.

DOI: 10.17212/1814-1196-2019-4-85-98

Automatic selection of leading indicators for regional labor market forecasting[*]

A.Yu. TIMOFEEVA

Novosibirsk State Technical University, 20 K. Marx Prospekt, Novosibirsk, 630073, Russian Federation

a.timofeeva@corp.nstu.ru

Abstract

The problem of selection of most informative variables from a set of candidate predictors arises when constructing forecast models based on leading indicators. This problem can be solved by embedded methods, such as the LASSO regression, or filter methods, for example, using the correlation-based feature selection. The task of this paper is to compare the performance of these methods with alternative approaches to the time series analysis (ARIMA, the Holt-Winters model, and exponential smoothing). For this, an algorithm for constructing forecast models is proposed, including automatic selection of leading indicators. To conduct an empirical study, indicators suitable for regional labor market predicting were selected from official statistics. They describe indicators such as money supply, a balance sheet structure of credit institutions and a price index. A pseudo-out-of-sample forecasting of a number of indicators characterizing the situation in the registered labor market of the Novosibirsk Region for the period from 2015 to 2018 was carried out. Direct multi-step forecasts were computed for horizons of 6months. It turned out that a stable modification of the LASSO, the percentile-lasso, does not give any advantages in terms of average absolute forecast errors. In most cases, the best results were obtained using the LASSO regression with the choice of the regularization parameter according to the one standard error rule based on block cross-validation with 10 blocks selected at random. Due to the automatic selection of leading indicators, it was possible to reduce forecasting errors in comparison with alternative methods. Thus, the proposed algorithm is appropriate for regional labor market predicting.

Keywords: variable selection, LASSO regression, percentile-lasso, correlation-based feature selection, leading indicators, labor market, forecasting, ARIMA, Holt-Winters model, STL algorithm

REFERENCES

1. Stock J., Watson M. Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association*, 2002, vol. 297, pp. 1167–1179. DOI: 10.1198/016214502388618960.

2. Boivin J., Ng S. Are more data always better for factor analysis? *Journal of Econometrics*, 2006, vol. 132, no. 1, pp. 169–194. DOI: 10.1016/j.jeconom.2005.01.027.

3. Tibshirani R. Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2011, vol. 73, no. 3, pp. 273–282. DOI: 10.1111/j.1467-9868.2011.00771.x.

[*] *Received 22 September 2019.*

4. Sagaert Y.R., Aghezzaf E.H., Kourentzes N., Desmet B. Tactical sales forecasting using a very large set of macroeconomic indicators. *European Journal of Operational Research*, 2018, vol. 264, no. 2, pp. 558–569. DOI: 10.1016/j.ejor.2017.06.054.

5. Bulligan G., Marcellino M., Venditti F. Forecasting economic activity with targeted predictors. *International Journal of Forecasting*, 2015, vol. 31, no. 1, pp. 188–206. DOI: 10.1016/j.ijforecast.2014.03.004.

6. Ma S., Fildes R., Huang T. Demand forecasting with high dimensional data: the case of SKU retail sales forecasting with intra-and inter-category promotional information. *European Journal of Operational Research*, 2016, vol. 249, no. 1, pp. 245–257. DOI: 10.1016/j.ejor.2015.08.029.

7. Hall M.A. *Correlation-based feature selection for machine learning*. PhD thesis. Hamilton, University of Waikato, 1999.

8. Timofeeva A.Y., Mezentsev Y.A. Forecasting using predictor selection from a large set of highly correlated variables. *CEUR Workshop Proceedings*, 2019, vol. 2416. *Information Technology and Nanotechnology: Data Science*, pp. 10–18.

9. Lund K.V. *The instability of cross-validated LASSO*. Master's thesis. Faculty of Mathematics and Natural Sciences, University of Oslo, 2013.

10. Roberts S., Nowak G. Stabilizing the lasso against cross-validation variability. *Computational Statistics and Data Analysis*, 2014, vol. 70, pp. 198–211. DOI: 10.1016/j.csda.2013.09.008.

11. Cleveland R.B., Cleveland W.S., McRae J.E., Terpenning I. STL: a seasonal-trend decomposition procedure based on loess. *Journal of Official Statistics*, 1990, vol. 6, pp. 3–73.

12. Brockwell P.J., Davis R.A., Calder M.V. *Introduction to time series and forecasting*. New York, Springer, 2002. 425 p.

13. *Open source and enterprise-ready professional software for data science*. Available at: https://rstudio.com/ (accessed 12.12.2019).

14. Statisticheskaya informatsiya o situatsii na registriruemom rynke truda [Statistical information on the situation in the registered labor market]. *Rostrud* [Federal Service for Labour and Employment]: website. Available at: https://www.rostrud.ru/rostrud/deyatelnost/?CAT_ID=6293 (accessed 12.12.2019).

15. EMISS (Unified interdepartmental information and statistical system): website. (In Russian). Available at: https://fedstat.ru (accessed 12.12.2019).

Для цитирования:

Тимофеева А.Ю. Автоматический подбор опережающих индикаторов для прогнозирования состояния регионального рынка труда // Научный вестник НГТУ. – 2018. – № 4 (77). – С. 85–98. – DOI: 10.17212/1814-1196-2019-4-85-98.

For citation:

Timofeeva A.Yu. Avtomaticheskii podbor operezhayushchikh indikatorov dlya prognozirovaniya sostoyaniya regional'nogo rynka truda [Automatic selection of leading indicators for regional labor market forecasting]. *Nauchnyi vestnik Novosibirskogo gosudarstvennogo tekhnicheskogo universiteta – Science bulletin of the Novosibirsk state technical university*, 2018, no. 4 (77), pp. 85–98. DOI: 10.17212/1814-1196-2019-4-85-98.