# ПРЕДСКАЗАНИЕ ИНДЕКСА МОСБИРЖИ ПРИ ПОМОЩИ МЕТОДА ОПОРНЫХ ВЕКТОРОВ

Специальность: *Экономика и управление народным хозяйством*
Направление: *Математические методы в экономике*

Автор: *Е. В. АНУФРИЕВА, лаборант-исследователь Центра макроэкономических исследований НИФИ Минфина России (г. Москва)*

В данной работе проводится анализ предсказательной силы цены индекса S&P500, немецкого индекса DAX, китайского индекса SSEC, а также обменных курсов рубль – доллар США, рубль – евро, и рубль – китайский юань для предсказания значения индекса Московской Биржи. Поскольку применение методов машинного обучения в финансах становится все более распространенным, в данной работе для анализа используется метод опорных векторов. Горизонт исследования составляет чуть более 13 лет (2006–2019 гг.), согласно принятой в машинном обучении практике, построение модели происходит на обучающей выборке, а ее проверка – на тестовой. Оценка модели производилась на тестовой выборке с использованием традиционных коэффициентов математической статистики.

The goal of this paper is to predict the price of MOEX stock exchange index with the help of machine learning techniques, namely, using the Support Vector Machine. The length of the study period is almost 13 years (from 2006 to 2019). Following the common techniques of machine learning, study period is divided into training (where we construct the model) and test samples (where the model is tested). Obtained predictions are tested with common in econometrics metrics, such as MSE, MAE, etc.

**Ключевые слова:** индекс МосБиржы, метод опорных векторов, российский фондовый рынок.

JEL: C22, E44, G10

**Keywords:** MOEX index, return, SVM, Russian financial market.

## Введение

Предсказание значений индексов на финансовых рынках – одна из важнейших и трудноисполнимых задач как для участников рынка, так и для исследователей. Важность практического применения возможности оценить

актив на финансовом рынке описана, например, в работе Шабалина [8], а также в работе Боргоякова [7]. Основная трудность заключается в том, что временной ряд значений индексов по своей сути не является стабильным. Кроме того, рыночные индексы представляют собой сложную, динамически изменяющуюся систему, влияние на которую оказывают множество факторов (помимо непосредственно связанных с финансовым рынком, влияние оказывают также и политические решения, и настроения инвесторов, и решения центральных банков).

Многочисленные исследования показали, что обычные нелинейные методы и статистические регрессии не могут как-то сильно улучшить прогноз значения ценового индекса. Возможность предсказания значений индекса Мосбиржи, например, при помощи метода главных компонент исследована в работе Ануфриевой [6]. В тоже самое время, многие исследования в последнее время используют нейронные сети и методы машинного обучения в целом для построения моделей предсказания движения рынка. Метод опорных векторов, однако, имеет некоторые преимущества перед простыми нейронным сетями: вычисления не зависят от изначально заданного пространства, и в целом, при применении метода опорных векторов риск того, что модель будет подвержена излишней спецификации (overspecification), ниже. Применение метода опорных векторов особенно оправдано, в случае если набор данных не является линейным. К основным преимуществам перед другими подходами можно отнести использование функции ядра, которая позволяет исследователю выбирать наиболее оптимальные пороговые значения для функции разделяющей гиперплоскости, в методе опорных векторов есть также параметр C и регуляризирующий параметр ε (используется при построении регрессии на основе метода опорных векторов). Кроме того, решение является единственным, потому что задача оптимизации является выпуклой. В то же время, в моделях, которые используют нейронные сети наблюдается несколько решений, которые связаны с локальными минимумы, поэтому результаты таких моделей не являются робастными по отношению к изменению выборки.

**Обзор литературы**

Существует большое количество различных исследований и разных видов моделей, призванных описать и предсказать поведение рыночных индексов. Тем не менее, многие существующие модели, которые пытаются предсказывать ситуацию на финансовом рынке не всегда показывают адекватные результаты. В последнее время, с развитием методов машинного обучения и их практического применения в финансовой сфере, появляется все больше исследований, которые включают в себя методы машинного обучения,

для того, чтобы попытаться описать и предсказать поведение различных инструментов на финансовом рынке.

Так, например, в статье Карассо и др. [1] авторы используют один из самых популярных классификаторов в машинном обучении – метод опорных векторов – для предсказания цены на медь. Полагая, что методы машинного обучения позволят получить больше информации из имеющихся данных, авторы оценивают поведение цен на медь на Лондонской бирже металлов. На горизонте равном почти 12 годам (с января 2005 года по июнь 2017 г.) авторы, таким образом, анализируют 2980 наблюдений цены металла. Как это принято в машинном обучении, авторы разделяют получившуюся выборку на обучающую и тестовую. Для оценки модели и полученных прогнозов авторы использовали MSE. Результаты применения данного метода показали, что предсказания модели на коротком горизонте являются достаточно точными, а с ростом горизонта предсказания увеличивается и кумулятивная ошибка построенных прогнозов. В целом, авторы также пришли к выводу, что значение гиперпараметров, которые используются при спецификации метода опорных векторов также оказывают существенное влияние на получаемые результаты. В целом, на горизонте до 20 дней, значения прогнозов модели демонстрируют маленькую ошибку.

В статье Самсудина и др. [4] авторы сравнивают способность двух разных методов машинного обучения оценить и предсказать поведение временных рядов. Авторы отмечают, что общей характеристикой временных рядов является то, что они часто оказываются нелинейными и в них есть выбросы. Исследователи для оценки таких данных выбрали для сравнения модель, основанную на искусственной нейронной сети и метод опорных векторов. Используя данные по ценам закрытия IBM, ежедневные данные по которой в этой базе данных собраны с 1961 по 1962 гг. На выбранном горизонте авторы не наблюдают ни четко выраженного тренда, ни сезонности (в отличие, например, от данных по солнечным пятнам или пассажирским перевозкам). Исследователи нормируют выбранные наборы данных на отрезке [0, 1]. Авторы также отмечают, что выбор гиперпараметров при построении модели оказывает большое значение на результаты. Результаты построения моделей авторы сравнивают при помощи MAE и RMSE. Исследователи отмечают, что метод опорных векторов показывает лучшие результаты, чем искусственная нейронная сеть для большинства из выбранных наборов данных.

В работе Найака и др. [3] авторы применяют методы машинного обучения к временному ряду обменного курса индийской рупии. Определение гиперпараметров для модели производилось при использовании различных техник, принятых в машинном обучении: поиск по сетке, случайный поиск,

генетический алгоритм для оптимизации, метод роя частиц, алгоритм оптимизации подражанием муравьиной колонии и др. Авторы отмечают, что предсказание обменного курса является одной из самых популярных сфер для исследования. В целом, нейронные сети показывают лучшие результаты при предсказании значений обменных курсов, чем обычные статистические модели. Тем не менее, авторы отмечают, что метод опорных векторов имеет ряд преимуществ перед простыми моделями искусственных нейронных сетей: для метода опорных векторов не имеет значение строение входного пространства, кроме того, такие модели также менее склонны к переобучению. Помимо метода опорных векторов авторы также используют метод k ближайших соседей (KNN), чтобы определить направление изменения. В работе рассматривается три пары обменных курсов: курс индийской рупии к доллару США, британскому фунту и к евро. В целом авторы приходят к выводу, что методы машинного обучения имеют существенное преимущества и более сильную предсказательную силу, чем обычные статистические регрессии.

Чен и Хао [5] в своем исследовании исследуют возможность предсказания значений индексов акций. В своем анализе авторы также рассматривают два подхода к оценке индексов: метод опорных векторов и метод k ближайших соседей. В отличие от предыдущих рассмотренных работ, здесь исследователи взвешивают рассматриваемые переменные в зависимости от их важности в алгоритме классификации (доли, которую они помогают объяснить). В работе оцениваются и предсказываются значения двух китайских индексов – Шанхайский индекс и Шеньженьский индекс. В целом, предложенные авторами модели показывают хорошую предсказательную способность не только на коротких (1–5 дней) и средних (10–15 дней), но также и на длинных горизонтах предсказания (20–30 дней). В анализ авторы также включают несколько технических индикаторов: скользящее среднее, экспоненциальное скользящее среднее, индикатор MACD, балансовый объем и несколько других индикаторов. Также в анализ включаются принятые в литературе факторы: цена открытия и закрытия, наибольшая и наименьшая цены, а также объем торгов. Все данные нормализованы на отрезке [-1, 1]. Полученные результаты оказались робастными, и в целом, хорошо предсказывают поведение китайских индексов акций.

В работе Лозинской и Жемчужникова [2] производится анализ возможности применения нейронных сетей и метода опорных векторов для предсказания российского рынка. Авторы оценивают качество полученных при помощи методов машинного обучения предсказаний. Выборка значений индекса рассматривается на горизонте 2012–2016 гг., и так же, как и в других

работах, авторы разделяют выборку на две: тестовую и обучающую. Помимо фундаментальных характеристик индекса, авторы, следуя зарубежным исследованиям, включают в анализ также и технические индикаторы. Зависимая и объясняющие переменные нормируются на отрезке [-1, 1]. Качество построенных моделей оценивалось при помощи общепринятых в статистики метрик: MAE и MSE, а также несколько других, например, $R^2$. Авторы пришли к выводу, что большее число наблюдений в обучающей выборке способствует снижению ошибки прогноза. Метод опорных векторов показал лучшие результаты, чем простая нейронная сеть, однако обнаруженные различия в получаемых результатах оказались незначительными.

**Методология**

Для предсказания индекса Мосбиржи используется метод опорных векторов (SVM). Метод опорных векторов позволяет преобразовать заданный вектор параметров в пространство более высокой размерности, посчитать линейную регрессию в таком пространстве и поострить разделяющую гиперплоскость с наибольшим «зазором». Предположения модели заключаются в том, что каждый из параметров имеет одинаковый вклад в зависимую переменную.

Метод опорных векторов можно видоизменить для построения регрессий. Принцип работы алгоритма остается таким же, как и в задачи классификации. Основная идея работы данного алгоритма заключается в следующем. Задан набор точек:

$$\{(x_1, c_1), (x_2, c_2), \dots (x_n, c_n)\}$$

Где каждый из $x_i$ – это некоторый вектор размера m, обычно нормализованный на отрезке [0, 1] или на [-1, 1]. Если данные не будут нормализованы, то точки с большими значениями очень сильно повлияют на классификатор. В случае линейной неразделимости классов, алгоритму позволяется допускать ошибки на обучающей выборке. Величина ошибки задается при помощи дополнительных переменных $\xi_i \geq 0,\ 1 \leq i \leq n$. Построение задачи линейной оптимизации сводится к минимизации квадрата нормали перпендикуляра к разделяющей гиперплоскости. Таким образом, система уравнений (1) выглядит как:

$$\begin{cases} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \xi_i \to min_{w,b,\xi_i} \\ c_i(w \cdot x_i - b) \geq 1 - \xi_i, \qquad 1 \leq i \leq n' \\ \xi_i \geq 0, \qquad 1 \leq i \leq n \end{cases} \qquad (1)$$

где C – это параметр настройки метода, который позволяет регулировать соотношение между накопленной ошибкой и максимизацией ширины разделяющей полосы. Решением данной задачи является наименьшее возможное значение $\frac{1}{2}\|w\|^2$, и наименьшая возможная ошибка классификатора.

Нелинейный классификатор также использует ядра, которые представляют собой нелинейные функции. Скалярные произведения далее заменяются функцией выбранного ядра в пространстве большой размерности. В данной работе, при оценки модели исследуются несколько видов ядер: линейное, уравнение ядра которого можно представить как $k(x, x') = x'x$, полиномиальное вида $k(x, x') = (x \cdot x')^d$, сигмоид: $k(x, x') = tanh(\kappa x \cdot x' + c)$, радиальная базисная функция, где уравнение ядра имеет вид: $k(x, x') = exp(-\gamma\|x - x'\|^2)$. К основным преимуществам SVM можно отнести то, что задачи выпуклого квадратичного программирования имеют единственное решение. Кроме того, число нейронов скрытого слоя определяется автоматически – оно равно количеству опорных векторов.

Тем не менее, в общем случае применение метода опорных векторов не сильно улучшает результаты предсказания. Однако, настройка гиперпараметров, в частности параметра C, регуляризирующего параметра ε или выбор ядра, позволяют существенно улучшить результат. В качестве алгоритма, при помощи которого отбираются оптимальные параметры, выбран поиск по сетке (Grid Search). Данный метод позволят выбрать гиперпараметры модели (то есть, те параметры, значение которых невозможно оценить исходя из данных). Оценка модели производится при помощи языка Python и пакета Scikit-learn.

**Данные**

В данном исследовании используются ежедневные данные по индексу Мосбиржи (IMOEX). Рассматриваемый период составляет чуть больше 13 лет: с 11 мая 2006 года по 17 сентября 2019 года, таким образом в выборке оказалось 3473 наблюдения. В рассматриваемый период попадает кризис 2008 года и период рецессии 2014 года. В качестве объясняющих факторов в модели выбираются обменные курсы рубля к валютам стран – торговым партнёрам, статистика объема экспорта и импорта, по которым является наибольшей: это страны еврозоны, Китай и США. На рассматриваем периоде были собраны данные по таким показателям, как цена индекса S&P500, цена немецкого индекса DAX – важнейшего фондового индекса Германии, цена индекса SSEC (Shanghai Stock Exchange Composite) – фондового индекса Китая. Также были собраны данные по обменному курсу пар рубль – доллар США, рубль – евро, и рубль – китайский юань. Поскольку на российскую экономику огромное

влияние оказывает цена на сырую нефть (например, более 99% от всех собранных экспортных пошлин в РФ приходится на природный газ, нефть и продукты, выработанные из нефти, кроме того добывающая отрасль также является одной из самых «нагруженных» налогами отраслью в РФ). Помимо этого, также используются данные по ставкам трехмесячных казначейских облигаций США. Данная переменная включается как некоторый ориентир для безрисковой ставки. В частности, многие исследователи и представители бизнеса используют модель CAPM для определения справедливой стоимости актива, в формулу которого входит безрисковая ставка (чаще всего ставка по Казначейским облигациям), а также различные премии за риск (например, страновой). Для заполнения пропусков данных используется линейная интерполяция: значения в пропусках заменяются средним значением двух ближайших значений. Помимо этих переменных в модель также включается лаг зависимой переменной.

Данные по объему торговли, ценам открытия и закрытия, а также минимальным и максимальным значениям по индексу Мосбиржи взяты с сайта Мосбиржи. Источником данных по индексам DAX, S&P500 и SSEC, а также по цене сырой нефти марки Brent стал терминал Thompson Reuters Eikon. Данные по обменному курсу всех трех пар валют взяты с сайта Центрального Банка РФ. Данные по трехмесячной ставке казначейских облигаций взяты с сайта Федерального Резерва США.

Далее, следуя литературе [2,3] данные нормализируются на отрезке [-1; 1] (необходимо для функции активации) при применении следующей формулы (2):

$$x_{\text{норм}} = -1 + \big(1 - (-1)\big) \times \frac{(x - x_{\text{мин}})}{(x_{\text{макс}} - x_{\text{мин}})}, \quad (2)$$

где $x$ – это изначальное значение переменной, $x_{\text{мин}}$ и $x_{\text{макс}}$ – минимальное и максимальное значение переменной.

Также данные разделяются на обучающую и тестовую выборки.

**Результаты применения модели**

При применении метода поиска по сетке, выбранные гипрепараметры приняли следующие значения: $C = 8.15$, $\varepsilon = 4.43 \times 10^{-4}$. Оптимальная функция для ядра оказалась линейной: $k(x, x') = x'x$. Для валидации построенной модели были построены несколько метрик качества:

1) $MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$

Этот коэффициент считает ошибку как абсолютную разницу между реальным и предсказанным значениями. Это линейный коэффициент, поэтому

вес каждой ошибки является одинаковым, однако такого, например, не происходит при подсчете RMSE. Кроме того, штраф за большую ошибку здесь не является таким же большим, как в MSE. В целом MAE является менее чувствительным к выбросам, чем MSE.

2) $RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$

В целом, эта метрика сильно похожа на MSE. Этот коэффициент может показать, насколько данные распределены вблизи разделяющей плоскости. В случае, если наличие больших ошибок нежелательно, использование RMSE является более оправданным, чем, например, использование MAE. Тем не менее, недостатком этого критерия является то, что с ростом выборки, на котором он применяется, RMSE также существенно возрастает, и становится больше, чем MAE.

Метрики качества для полученной модели представлены в таблице 1.

Таблица 1 – Метрики качества для оценённой модели

| Метрика качества | Тестовая выборка |
|---|---|
| MAE | 0.014 |
| RMSE | 0.017 |

Значение RMSE достаточно близкое к значению MAE (табл.1), что говорит о том, что есть несколько наблюдений, чье отличие предсказанных и реальных значений довольно заметно отличается. MAE является очень низким, и получилось гораздо меньше, чем в работе Лозинской и Жемчужникова [2], которые также исследовали индекс Мосбиржы. Их значение MAE достигало 0,024. В целом метод опорных векторов обладает высокой объясняющей способностью. Модели на основе метода опорных векторов являются устойчивыми к проблеме излишней спецификации, а решение всегда одно, потому что модель всегда сходится к глобальному минимуму.

**Заключение**

В данной работе изучена возможность объяснять и предсказывать значение цены индекса Мосбиржи при помощи методов машинного обучения. Временной горизонт анализа составляет чуть больше 13 лет (с мая 2006 года по сентябрь 2019 года), а количество ежедневных наблюдений составило 3473.

В работе применяется метод опорных векторов к временным рядам, в процессе построения модели методом поиска по сетке подбираются оптимальные значения гиперпараметров. Результаты исследования

демонстрируют достаточно маленькие значения ошибок прогноза (MAE, RMSE). Применение метода опорных векторов к временным рядам демонстрирует высокий потенциал возможностей для более точного предсказания движения цены индекса. В будущем, в качестве объясняющих факторов при построении модели можно использовать факторы из технического анализа (например, скользящее среднее), а также существенно увеличить интервалы для поиска значений гиперпараметров по сетке, что, однако, потребует наличия существенных вычислительных мощностей.

**Список литературы**


1. Carrasco, Raúl et al. Forecast of copper price series using vector support machines. 2018 7th // International Conference on Industrial Technology and Management (ICITM). – 2018. – p. 380–384.
2. Lozinskaia Agata M., Zhemchuzhnikov Victor A. MICEX index forecasting: the predictive power of neural network modeling and support vector machine // Вестник ПГУ. – Серия: Экономика. – 2017. – № 1 – p. 49–60.
3. Nayak R., Mishra D., Rath A. An optimized SVM-k-NN currency exchange forecasting model for Indian currency market // Neural Computing and Applications. – 2017. – p. 2995-3021. – 10.1007/s00521-017-3248-5.
4. Samsudin R, Shabri, A., Saad, P. A Comparison of Time Series Forecasting using Support Vector Machine and Artificial Neural Network Model // Journal of Applied Sciences. – 2010. – №10. – p. 950-958. – 10.3923/jas.2010.950.958.
5. Yingjun Chen, Yongtao Hao. A feature weighted support vector machine and K-nearest neighbor algorithm for stock market indices prediction // Expert Systems with Applications. – Volume 80. – 2017. – p. 340–355 – ISSN 0957-4174 – 10.1016/j.eswa.2017.02.044.
6. Ануфриева Е. В. Влияние макроэкономических показателей на доходность индексов российской фондовой биржи // Финансовый журнал. – 2019. – № 4. – С. 75–87. – DOI: 10.31107/ 2075-1990-2019-4-75-87.
7. Боргояков А. С. Модель оценки доходности облигаций с позиции платежеспособности фирмы-эмитента // Научно-исследовательский финансовый институт: Финансовый журнал. – 2017. – № 3. – С. 96–106.
8. Шабалин М. К. Правильная оценка стоимости инвестиционного объекта в кризисное время // Научно-исследовательский финансовый институт: Финансовый журнал. – 2016. – № 4. – С. 66–75.