УДК 336

ПРИМЕНЕНИЕ СТАТИСТИЧЕСКОЙ ТЕОРИИ МАШИННОГО ОБУЧЕНИЯ ПРИ ПРОГНОЗИРОВАНИИ ФОНДОВОГО ИНДЕКСА

Я.В. Федорова, РГЭУ (РИНХ), e-mail: fyv21@mail.ru

Т.Н. Шарыпова, РГЭУ (РИНХ), e-mail: tnt@mail.ru

Л.К. Попова, РГЭУ (РИНХ), e-mail: popova_plk@mail.ru

Е.Н. Лозина, РГЭУ (РИНХ), e-mail: lozinaen@mail.ru

Аннотация. В современном мире существует много вопросов, касающихся прогнозирования показателей, связанных с финансовым рынком. Модель, которая может точно предсказать направление движения рынка, в настоящее время очень ценна и актуальна. В данной работе представлена линейная регрессионная модель прогнозирования финансового индекса Доу-Джонса на фондовом рынке. Ключевым вопросом в работе фондового рынка является прогнозирование периодов увеличения (роста) или уменьшения индексной функции на определенный период времени. Модели, которые могут предсказывать направление движения рынка с высокой точностью, могут быть построены с использованием статистической теории машинного обучения.

Ключевые слова: фондовый рынок, финансовый индекс Доу Джонса

APPLICATION OF THE STATISTICAL THEORY OF MACHINE LEARNING IN STOCK INDEX FORECASTING

Ya.V. Fedorova, RSEU (RINH), e-mail: fyv21@mail.ru

T.N. Sharypova, RSEU (RINH), e-mail: tnt@mail.ru

L.K. Popova, RSEU (RINH), e-mail: popova_plk@mail.ru

E.N. Lozina, RSEU (RINH), e-mail: lozinaen@mail.ru

Abstract. In the modern world, there are many questions concerning the prediction of indicators related to the financial market. A model that can accurately predict the direction of market movement is currently very valuable and relevant. This paper presents a linear regression

model for predicting the Dow Jones financial index in the stock market. The key issue in the work of the stock market is the prediction of periods of increase (growth) or decrease of the index function for a certain period of time. Models that can predict the direction of market movement with high accuracy can be built using the statistical theory of machine learning.

Keywords: stock market, Dow Jones financial index

Индекса Доу-Джонса это один из самых известных и старейших индексов среди американских рыночных индексов. Его основное назначение заключается в отслеживании развития промышленной составляющей американских фондовых рынков. Индекс охватывает 30 крупнейших компаний США, которые и создают весь экономический климат не только США, но и значительную часть всего мирового экономического потенциала.

Целью исследования являлась реализации предсказательной модели для прогнозирования цены закрытия индекса Доу-Джонса на фондовом рынке, используя модели статистического анализа.

Для реализации предсказательной модели были взяты исходные данные из открытых платформ Quandl и AlphaVantage, обеспечивающих доступ к более чем 9 миллионам бесплатных наборов данных [5], [6].

В течение каждого торгового дня в период с понедельника по пятницу цена акции изменяется и регистрируется в режиме реального времени. Пять характеристических значений (дата, цена открытия, цена закрытия, наибольшее и наименьшее значений цен, суммарное количество проданных за день акций) показывают изменение цены за один день и являются ключевыми показателями предсказательной модели.

При создании модели и анализе данных рассчитываем промежуточные данные, такие как: цена открытия за предыдущий день; цена закрытия за предыдущий день; самая высокая цена за предыдущий день; самая низкая цена за предыдущий день и т.п.

Для предсказания цен закрытия индекса Доу-Джонса на преобразованных данных воспользуемся стандартной функцией SGDRegressor из библиотеки sklearn. Для подбора параметров воспользуемся функцией GridSearchCV.

Полученный результат представлен на рис. 1:

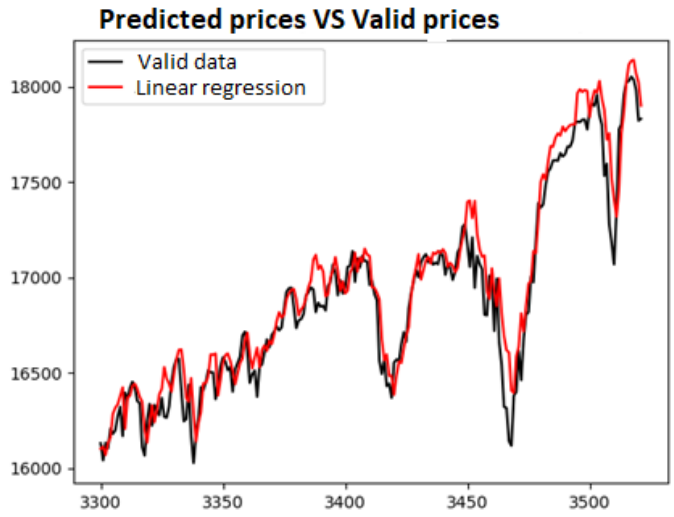

Рис. 1. Сравнение цен

MSE: 50415.891 Точность на обучающем наборе: 0.96

Точность на тестовом наборе: 0.92

Для увеличения точности предсказания нашей модели проверим являются ли наши данные стационарными, так как нестационарные данные могут привести к ухудшению работы регрессии.

Для проверки на стационарность используем расширенный тест Дики-Фуллера (ADF). ADF — это тип статистического теста, который определяет, присутствует ли единичный корень в данных временных рядов. Рассмотрим две гипотезы:

нулевая гипотеза, H0: если не удалось отклонить, значит есть вероятность, что временной ряд является нестационарным;

альтернативная гипотеза, H1: если H0 отклонено, значит временной ряд является стационарным.

При этом используется следующее критическое значение:

$p > 0{,}05$: нельзя отвергнуть H0, значит данные имеют единичный корень и являются нестационарными; $p \leq 0{,}05$: отвергаем H0, следовательно, данные не имеют единичного корня и являются стационарными.

Вычисляем скользящее среднее и стандартное отклонение в переменные с периодом окна в один год: Выведем графики скользящего среднего относительно исходного временного ряда (рис. 2)

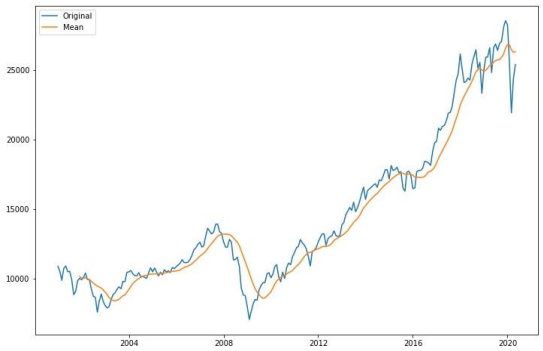

Рисунок 2. График скользящего среднего

Используя модуль statsmodels, выполняем тест ADF для нашего набора данных с помощью метода adfuller ():

Статистическое $p-$ значение теста ADF больше 0,05. Таким образом, нельзя отвергнуть H0 о существовании единичного корня, следовательно, данные являются нестационарными.

На не стационарность временного ряда, влияет тренд или сезонность. Для того чтобы сделать ряд стационарным, нужно тренд и сезонность устранить. Для этого используют методы: детрендинг, дифференцирование и разложение.

Спрогнозируем будущие значение на полученных статистических данных, используя метод авторегрессионного интегрированного скользящего

среднего (ARIMA). ARIMA - модель прогнозирования для стационарных временных рядов, основанная на линейной регрессии.

Поиск параметров для нашей модели будем производить используя «поиск по сетке», также известный как метод оптимизации гиперпараметров.

Модель с наименьшим значением AIC дает нам наиболее подходящую модель, которая определяет наши параметры.

Модель сезонного компонента ARIMA (0,1,1,12) даст нам самое низкое значение AIC - 3289.336.

Построим график цен начиная с 2008 года:

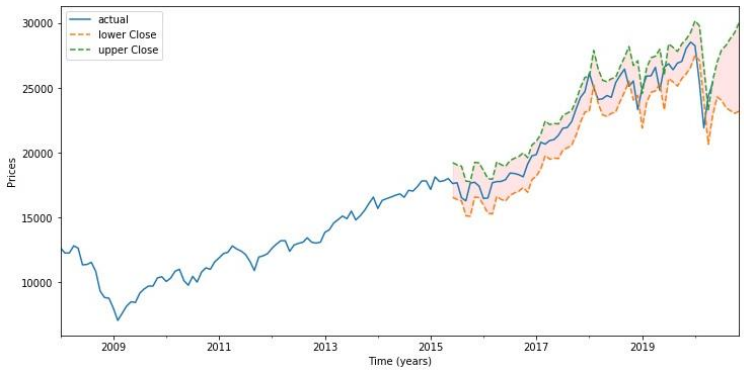

Рисунок 3. Результат прогноза цены на индекс Доу-Джонса.

Сплошная линия показывает наблюдаемые значения, в то время как пунктирные линии отображаю пятилетние скользящие прогнозы. По мере того, как прогноз на ближайшие пять месяцев уходит в будущее, доверительный интервал расширяется, чтобы отразить потерю уверенности в прогнозе.

Таким образом, реализована предсказательная модель для прогнозирования цены закрытия индекса Доу-Джонса на фондовом рынке, используя модель линейной регрессии.

Библиографический список

1. Shalev-Schwartz Sh., Ben-David Sh. Machine Learning Ideas, 2019

2. Muller, Guido. Introduction to Machine Learning with Python, 2017. (p. 59-71)
3. Documentation for the pandas library [Electronic resource]. – URL: https://pandas.pydata.org/docs/user_guide/index.html
4. Documentation on Quandl [Electronic resource] – URL: https://docs.quandl.com
5. Weiming - Mastering Python for Finance Implement Advanced State-of-the-art Financial Statistical Applications Using Python-Packt Publishing, 2019
6. Documentation on AlphaVantage [Electronic resource] – URL: https://www.alphavantage.co/documentation/