МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение
высшего образования
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО»

Кафедра теории функций и стохастического анализа

ПРОГНОЗИРОВАНИЕ НА ОСНОВЕ ФИНАНСОВОЙ СТАТИСТИКИ

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

Студентки 4 курса 412 группы
направления 01.03.02 — Прикладная математика и информатика

механико-математического факультета
Моисеевой Вероники Александровны

Научный руководитель
доцент, к. ф.-м. н., доцент _____ М. Г. Плешаков

Заведующий кафедрой
д. ф.-м. н., доцент _____ С. П. Сидоров

Саратов 2023

Введение. Задача прогнозирования поведения некоторой системы стоит перед различными сферами науки и бизнеса уже долгое время и является актуальной в современном мире. Это может быть прогноз погоды, прогноз активности пользователей на сайте или прогноз изменения цен акций компании на бирже.

Основная проблема в задаче анализа и прогнозирования заключается в построении модели, адекватно отражающей динамику финансовых временных рядов. Рыночный механизм, характеризующийся огромным количеством постоянно меняющихся связей, зависит от множества внешних факторов, способных существенно повлиять на всю структуру его зависимостей, причем воздействие может быть самым разнообразным. Появление тех или иных внешних факторов, как правило, никак не отражается в предыстории финансового временного ряда, но вызывает значительное нарушение его динамики. Именно в этом состоит особенность практически всех финансовых временных рядов.

Выбор модели, как правило, осуществляется эмпирическим методом на основе некоторого заданного универсального семейства предикторов, возможности которого позволяют описать любой временной ряд. Выбор того или иного семейства моделей отражает специфику решения задачи прогнозирования. Наиболее ярким представителем практического подхода является технический анализ. Кроме того, в последнее время получили широкое распространение регрессионные методы, а также методы и на нейронных сетях. Общая черта, объединяющая оба подхода, заключается в попытке построения единой модели для финансового временного ряда. Однако построение такой модели наталкивается на ряд принципиальных трудностей, обусловленных природой финансового временного ряда.

Существует множество моделей и соответствующих им методов прогнозирования, на данный момент насчитывается свыше 100 классов моделей прогнозирования. Существующие модели временных рядов широко используются в процессе изучения динамики реальных явлений различной природы. Они зачастую применяются в исследованиях динамики грузо - и пассажиропотоков, товарных и складских запасов, миграционных процессов, анализе химических процессов, моделировании разнообразных природных событий.

Наиболее активно модели временных рядов применяются в анализе финансовых рынков, при оценке изменений финансовых показателей, прогнозировании цен на различные товары, курсов акций, соотношений курсов валют и т. п.

Актуальность работы. Задача прогнозирования будущих значений временного ряда на основе его исторических значений является основой для финансового планирования в экономике и торговле, планирования, управления и оптимизации объемов производства, складского контроля. Хороший прогноз временного ряда на много шагов вперед может дать бизнесу серьезные конкурентные преимущества, участникам фондовой биржи делать наилучшие решения по покупке или продаже акций, или позволить ученым смоделировать какое-либо природное явление и провести эксперимент.

Целью бакалаврской работы является исследование основных методов прогнозирования финансовых временных рядов и построение модели, с помощью которой можно будет сделать наиболее точный прогноз.

Для достижения данной цели можно выделить ряд **задач**:

— рассмотреть понятия временного ряда;
— изучить основные методы анализа и прогнозирования временных рядов;
— оценить применимость каждого метода для прогнозирования выбранного временного ряда;
— выбрать наиболее подходящие методы;
— применить теоретические знания на практике и продемонстрировать работу выбранных моделей на практике;
— проанализировать полученные результаты.

В качестве **объекта исследования** выступает задача прогнозирования временных рядов.

Предметом исследования являются методы прогнозирования временных рядов.

В ходе исследования применялись следующие методы: анализ, моделирование, изучение научной литературы и ее обобщение.

Структура и содержание бакалаврской работы. Работа состоит из введения, трех разделов, заключения, списка использованых источников, содержащего 20 наименований и одного приложения. *В первом разделе* при-

водится определение временного ряда, его компонент, классификация и требования к исходной информации. *Во втором разделе* подробное описываются модели прогнозирования временных рядов. Рассмотрены следующие модели: регрессионные, авторегрессии, скользящего среднего, ARCH, GARCH, ARIMA, SARIMA, рекурентные нейронные сети. *В третьем разделе* приведены результаты работы моделей и их сравнительный анализ. Общий объем работы 52 страницы. в том числе две таблицы и 27 рисунков.

Основное содержание работы. В **первом разделе** приводится определение временного ряда, его компонент, классификация и требования к исходной информации.

Временной ряд — это последовательность упорядоченных во времени числовых показателей, характеризующих уровень состояния и изменения изучаемого явления. Анализ таких рядов является непростой и важной задачей, решение которой дает полезные результаты для той или иной области. Одной из основных целей анализа временных рядов является прогнозирование его поведения. Прогноз будущих значений на основе прошлых наблюдений позволяет наиболее эффективно принимать решения в настоящем.

Работа с временными рядами предполагает два аспекта: анализ временного ряда, т.е. понимание его структуры и закономерностей и моделирование и построение прогноза на будущее.

При анализе временных рядов принято выделять следующие *компоненты*: тренд, сезонная компонента, случайная составляющая.

Выделение этих компонент – один из первых этапов анализа. Таким образом модель временного ряда можно описать, как $Y = T + S + \varepsilon$ – *аддитивная* модель и $Y = T * S * \varepsilon$ – *мультипликативная* модель. Наиболее распространенной считается вторая модель, которая, в свою очередь, сводится к первой логарифмированием.

Требования к исходной информации: сопоставимость, однородность, устойчивость.

Во **втором разделе** приводятся методы прогнозирования временных рядов.

Регрессионные модели. Регрессионные модели прогнозирования одни из старейших. Регрессионными моделями являются:

— простая линейная регрессия;

— множественная регрессия;

— нелинейная регрессия.

Частным случаем является модель широко используемой линейной регрессии:

$$Y = \beta_0 + \beta_1 x_1 + ... + \beta_n x_n + \varepsilon,$$

где $\beta_1, ..., \beta_n$ - подбираемые коэффициенты регрессии.

Модели авторегрессии – это класс моделей временных рядов, в которых текущее значение моделируемой переменной задается функцией от прошлых значений самой этой переменной.

Простейшая модель автокоррелированного стационарного ряда, которая часто используется на практике, имеет вид:

$$y_t = \alpha y_{t-1} + \varepsilon_t.$$

Такая модель называется *моделью авторегрессии первого порядка (AR(1))*. *Модели скользящего среднего* – это класс моделей временных рядов, в которых моделируемая величина задается функцией от прошлых ошибок.

Модель скользящего среднего первого порядка (MA(1)) записывается в виде:

$$y_t = \varepsilon_t - \beta \varepsilon_{t-1},$$

Модель авторегрессии$(ARMA(p, q))$ - скользящего среднего порядков p и q имеет вид:

$$y_t = \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + ... + \alpha_p y_{t-p} + \varepsilon_t - \beta_1 y_{t-1} - \beta_2 y_{t-2} - ... - \beta_q y_{t-q}.$$

Традиционные модели временных рядов, такие как модель ARMA, не могут адекватно учесть все характеристики, которыми обладают финансовые временные ряды, и требуют расширения.

Модель ARCH. Авторегрессионная условная гетероскедастичность или $ARCH$ - это метод, который явно моделирует изменение дисперсии во времени во временном ряду. В частности, метод $ARCH$ моделирует дисперсию на временном шаге как функцию от остаточных ошибок от среднего процесса

(например, нулевого среднего).

Модель GARCH. Модель $GARCH$ (обобщенная модель $ARCH$) является альтернативной модификацией модели $ARCH$, позволяющей получить более длинные кластеры при малом числе параметров. Модель $ARMA$ зачастую позволяет получить более сжатое описание временных зависимостей для условного математического ожидания, чем модель AR. Подобным же образом модель $GARCH$ дает возможность обойтись меньшим количеством параметров по сравнению с моделью $ARCH$, если речь идет об условной дисперсии.

Сглаживание временного ряда, т.е. замена фактических уровней расчетными значениями, имеющими меньшую колеблемость, чем исходные данные, является простым методом выявления тенденции развития. В ряде случаев при графическом изображении временного ряда тренд прослеживается недостаточно отчетливо. Поэтому ряд сглаживают, на график наносят сглаженные значения и, как правило, тенденция проявляется более четко. Некоторые методы анализа и прогнозирования требуют в качестве предварительного условия сглаживание временного ряда. Сглаживание временных рядов используется при устранении аномальных наблюдений. Существуют следующие методы сглаживания: *метод простой скользящей средней*, *метод взвешенной скользящей средней*, метод экспоненциального сглаживания.

Модель ARIMA. Модель $ARIMA$, представляет собой алгоритм прогнозирования, основанный на концепции, согласно которой данные предыдущих значений временного ряда могут использоваться только для прогнозирования будущих значений. $ARIMA$ представляет собой класс моделей, которые «демонстрируют» данный временной ряд на основе его предыдущих значений. Данная модель является расширением модели $ARMA(p, q)$ для нестационарных временных рядов, которые сводятся к стационарным взятием разностей d-го порядка.

Основная идея модели $ARIMA$: обрабатывать последовательность данных, сформированную предсказанным объектом со временем, как случайную последовательность, и использовать определенную математическую модель, чтобы приблизительно описать эту последовательность. Как только эта модель идентифицирована, будущая стоимость может быть предсказана из про-

шлых и текущих значений временного ряда.

Модель SARIMA. Модель SARIMA - обобщение $ARIMA$-модели на временные ряды, в которых имеется ярко выраженная сезонная компонента. Дополнительно в такой модели вводятся сезонные параметры (P,D,Q,s), позволяющие учесть циклические колебания процесса.

Искусственные нейронные сети. В основе нейросетевых технологий лежит идея о том, что функционирование биологического нейрона можно промоделировать относительно простыми математическими моделями, а вся глубина и гибкость человеческого мышления и другие важнейшие качества нервной системы определяются не сложностью нейронов, а их большим числом и наличием сложной системы связей между ними.

Рекуррентная нейронная сеть (RNN). Последовательная модель обычно предназначена для преобразования входной последовательности одной предметной области в выходную последовательность другой. Рекуррентная нейронная сеть (RNN) подходит для этой цели и продемонстрировала прорыв в решении таких задач, как распознавание рукописного текста, распознавание речи и машинный перевод.

Однако простые персептронные сети, которые линейно объединяют текущий элемент ввода и последний элемент вывода, могут легко потерять долгосрочные зависимости. Чтобы решить эту проблему, исследователи создали специальный нейрон с гораздо более сложной внутренней структурой для запоминания долгосрочного контекста, называемого ячейкой Long-short Term Memory (LSTM). Он достаточно умен, чтобы узнать, как долго он должен запоминать старую информацию, когда использовать новые данные и как объединить старую память с новым входом.

Рекуррентная сеть LSTM-типа. Рекуррентная нейронная сеть глубокого обучения(LSTM) - это особый вид нейронных сетей, которые обычно способны различать долгосрочные зависимости. Модель LSTM, была разработана для предотвращения проблем с долгосрочными зависимостями, и с этим они обычно справляются очень хорошо. Модель имеет способность улавливать закономерности в данных временных рядов и вследствие этого может использоваться для прогнозирования будущего тренда данных.

В **третьем разделе** приведен сравнительный анализ моделей прогно-

зирования.

Для прогнозирования использовались следующие алгоритмы:

— скользящая средняя;
— линейная регрессия;
— LSTM;
— ARIMA;
— prophet.

Для построения моделей прогнозирования, описанных в теоретической части работы, был использован язык программирования Python и его основные библиотеки: *scikit-learn*, *NumPy*, *SciPy*, *pandas*, *matplotlib*.

Для работы были использованы исторические данные о курсах акций компании. Используемый датасет для прогнозирования цен закрытия акций включает в себя следующие величины

1. дата;
2. цена открытия;
3. ценовой максимум;
4. ценовой минимум;
5. цена закрытия;
6. объем торгов.

Прогнозирование финансовых рядов будет выполняться на основе характеристик цен закрытия акций компании за период 08.10.2013-08.10.2018 с шагом в один день. Информация о ценах акций компаний за выбранный промежуток времени была получена с помощью открытого источника Kaggle.

Для начала загрузим набор данных и определим целевую переменную. Расчет прибыли или убытка обычно определяется ценой закрытия акции в течение дня, поэтому мы будем рассматривать цену закрытия в качестве целевой переменной.

Далее создаем DataFrame. DataFrame - многомерных массивов с метками столбцов, который содержит только столбцы даты и цены закрытия, а затем разделить его на обучающий и тестовый набор (80% данных будем брать для обучения и 20% для тестирования). Для оценки точности прогнозируемых значений будем использовать среднеквадратичную ошибку (RMSE)

и среднюю относительную ошибку (MRPE):

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

$$MRPE = \frac{1}{N}\sum_{i=1}^{N}\frac{|y_i - \hat{y}_i|}{y_i} * 100,$$

где N — число наблюдений набора данных, y_i — фактическое значение независимой переменной для i-го наблюдения, $\hat{y}_i$ — значение, предсказанное моделью для i-го наблюдения.

Для некоторых моделей потребуются нормализованные данные, поэтому воспользуемся функцией MinMaxScaler из библиотеки scikit-learn.

Первой будем использовать модель *скользящей средней*.В результате, прогнозируемые значения находятся в том же диапазоне, что и наблюдаемые значения в обучающем наборе: сначала наблюдается тенденция к увеличению, а затем - к медленному снижению. Данный метод прост, легко применим, дает близкую к действительности картину долговременных изменений, но помощью него сложно получить хорошие прогнозы.

Рассмотрим модель *линейной регрессии*. Модель не соответствует столбцу даты и месяца. Модель рассматривает значение с той же даты месяц назад или с той же даты год назад, а не предыдущие значения.

Рассмотрим модель *ARIMA*. ARIMA представляет собой класс моделей, которые «демонстрируют» данный временной ряд на основе его предыдущих значений: его лаги и ошибки в прогнозировании, их выравнивание можно использовать для прогнозирования будущих значений. Было сделано предположение, что это связано с тем, что модель лучше работает для краткосрочных прогнозов. Попробуем увеличить обучающую выборку, тем самым уменьшить тестовую выборку, и посмотрим на результат. Показатели улучшили свои значения. Значит наше предположение о том, что модель лучше работает для краткосрочных прогнозов, имеет место быть.

Рассмотрим модель *Prophet*. Большинство методов прогнозирования требуют предварительной обработки данных перед построением модели. Данная модель представляет собой библиотеку прогнозирования временных рядов,

которая не требует предварительной обработки данных. Данный инструмент хорошо работает с рядами, которые имеют ярко выраженные сезонные эффекты, а также имеют несколько таких периодов. Prophet пытается уловить тренд и сезонность на основе прошлых данных. А так как цены на акции сильно зависят от того, что в данное время происходит на рынке, а не от сезонности, поэтому данная модель справилась с прогнозированием не очень хорошо.

Последней рассмотрим модель *Long Short Term Memory (LSTM)*. Рекуррентные нейронные сети страдают от кратковременной памяти. Если последовательность достаточно длинная, им будет сложно переносить информацию с более ранних этапов времени на более поздние, то есть слои перестают учиться. Это обычно более ранние слои. Так как эти уровни не изучаются, RNN могут забыть то, что видели в более длинных последовательностях. Преимущество LSTM перед RNN состоит в том, что у нее есть внутренние механизмы, называемые воротами, которые могут регулировать поток информации. Эти ворота могут узнать, какие данные в последовательности важно сохранить или выбросить. Благодаря этому модель может передавать важную информацию по длинной цепочке последовательностей для прогнозирования. Эта модель показала наилучшие результаты.

Теперь возьмем для исследования данные за 1 год и за 3 месяца. И реализуем все вышеперечисленные модели на новых данных и сравним результаты.

Посмотрим как изменились относительные и абсолютные ошибки. В таблице 1 представлена сравнительная характеристика абсолютных ошибок. В первой строке таблицы приведены ошибки, которые показали модели, работая с данными за 5 лет, во второй за 1 год, в третьей за 3 месяца.

Таблица 1 – Сравнительная характеристика абсолютных ошибок

Период	Скользящая средняя	Лин.регрессия	ARIMA	Prophet	LSTM
5 лет	104.5	121.1	10.4	57.4	12.3
1 год	66.4	53.6	4.9	40.3	10.3
3 месяца	53.7	7.5	21.5	177.2	13.7

В таблице 2 представлена сравнительная характеристика относитель-

ных ошибок.

Таблица 2 – Сравнительная характеристика относительных ошибок

Период	Скользящая средняя	Лин.регрессия	ARIMA	Prophet	LSTM
5 лет	37.4%	43.8%	5.6%	15.8%	0.05%
1 год	20.9%	16.5%	2.1%	10.8%	0.02%
3 месяца	17.7%	2.3%	9.8%	33.1%	4.2%

Проанализировав результаты ошибок, мы видим, что в несколько раз улучшился результат у модели линейной регрессии.

Модели ARIMA и Prophet улучшили свои результаты на данных за 1 год, но показали свои худшие результаты на данных за 3 месяца.

Наименьшую относительную ошибку (0.02%) в результате исследования показала модель LSTM на данных за 1 год

Подводя итоги, можно отметить, что большинство моделей показывают наиболее приближенные к реальным значениям на данных за 1 год. На выбранных данных хорошо показали себя модели LSTM и ARIMA, а также на данных за 3 месяца модель линейной регрессии.

Заключение. В заключении данной работы хотелось бы отметить, что прогнозирование временных рядов в ближайшее время станет наиболее популярной областью для исследований и экспериментов, поскольку оно может приносить высокую материальную прибыль. Именно поэтому будет появляться все больше и больше качественных и быстрых алгоритмов прогнозирования, что предоставит людям, желающим начать исследования в этой сфере, большой интеллектуальный, информационный и алгоритмический фундамент для последующей работы и развития. Но прогнозировать финансовые временные ряды порой очень сложно, так как цены на акции не имеют определенного тренда или сезонности. Они сильно зависят от того, что сейчас происходит на рынке. В этом и состоит сложность подбора моделей для прогнозирования. Но уже сейчас существует огромное количество моделей, которые хорошо справляются с поставленной задачей.

Основу всей совокупности названных методов традиционно составляют статистические методы, применяемые для прогнозирования развития социальных и экономических явлений и процессов, построения адекватных моде-

лей временных рядов и выбора наиболее приемлемых вариантов из всех возможных способов прогнозирования. Система статистических методов изучения динамики явлений позволяет определить, как развиваются общественные явления: растут или уменьшаются их размеры, быстро или медленно происходят эти изменения и так далее. Но существуют не только статистические методы. И когда встает вопрос о прогнозировании из всех методов нужно выбрать более подходящий. Но стоит отметить, что не существует универсального метода прогнозирования временных рядов, каждый метод находит свое применение для разных типов временного ряда. Например, на рядах, которые имеют ярко выраженную тенденцию, хорошую прогностическую способность проявляют регрессионные модели; на рядах, в которых присутствует сезонная составляющая – модели экспоненциального сглаживания.

Также сейчас особое внимание уделяют прогнозированию с помощью нейронных сетей - это очень мощный и гибкий механизм прогнозирования идовольно перспективная технология, предлагающая современный подход к исследованию задач во многих областях экономики. Использование соответствующих нейросетевых методов позволяет значительно улучшить существующие на данный момент способы анализа и прогнозирования.

Перед тем как приступить к прогнозированию, была изучена и проанализирована необходимая литература, выбран датасет и модели, произведена обработка данных. В качестве исследуемых моделей были выбраны:

— скользящая средняя;
— линейная регрессия;
— LSTM;
— ARIMA;
— prophet.

И были рассмотрены разные объемы выборок (5 лет, 1 год и 3 месяца). Наилучший результат мы получили с помощью модели LSTM на данных за 1 год. Модель линейной регрессии показала худший вариант на данных за 5 лет, но один из лучших за 3 месяца. Модель ARIMA показала не очень хороший результат за 5 лет и за 3 месяца, но за то за 1 год показала один из лучших результатов. Модель скользящей средней довольно неплохо сглаживает данные, но хороший прогноз с помощью нее сделать не получится.

Модель Prophet показала самые плохие результаты.

Подбор моделей для прогнозирования является очень трудной задачей. И не существует оптимальной модели, с помощью которой можно было бы хорошо спрогнозировать любой временной ряд. Но после анализа и реализации нескольких моделей, можно подобрать ту, которая хорошо выполнит прогноз. С этой задачей довольно неплохо справляются как стохастические модели, так и модели, построенные на основе нейронных сетей.