МИНОБРНАУКИ РОССИИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г. ЧЕРНЫШЕВСКОГО»

Кафедра математической экономики

Использование данных новостной аналитики для анализа инвестиционных рисков

АВТОРЕФЕРАТ МАГИСТЕРСКОЙ РАБОТЫ

студентки 2 курса 247 группы

направления 09.04.03-Прикладная информатика

механико-математического факультета

Сарсеновой Айжан Зинетуллаевны

Научный руководитель
профессор, д.э.н., профессор _____ В. А. Балаш

Зав. кафедрой
профессор, д.ф.-м.н _____ С.И. Дудов

Саратов 2019

СОДЕРЖАНИЕ

Стр.

ВВЕДЕНИЕ

В последнее время, чтобы принять верное и взвешенное решение и разработать грамотную стратегию поведения, связанную непосредственно с различными механизмами, функционирующих на финансовых и биржевых рынках, необходимо учитывать все большее количество факторов. Наибольший интерес для финансовых аналитиков представляет изучение неопределенности рыночного процесса, ключевым параметром которого является волатильность, характеризующая рыночный процесс количественно.

Волатильность для многих специалистов стала синонимом слова «риск». Понимание и прогнозирование изменения цены – это залог успешного планирования как для инвестора, так и для остальных игроков рынка, а, как известно, прогнозирование есть предшествующий этап планирования. Именно прогнозирование является наиболее сложным видом деятельности в экономической сфере, оно проявляется в качестве основного и завершающего этапа исследований, главные результаты которых закладывают в дальнейшую программу деятельности. Существует стратегическое планирование, которое осуществляется на основе среднесрочных и долгосрочных прогнозов, и текущее планирование, осуществляемое на базе краткосрочных прогнозов.

Актуальность выбранной темы заключается в том, что изучение волатильности и факторов, влияющих на неё, является важным элементом в организации управления как отдельным хозяйствующим субъектом, так и экономикой в целом. **Цель данной работы** является исследование взаимосвязи между интенсивностью новостного потока и объемом биржевых торгов, то есть определение влияния новостей на волатильность.

Задачами, которые будут рассмотрены и решены в ходе исследования, станут:

1. Графическое представление и описание поведения временных рядов новостей и биржевых торгов.
2. Выделение и удаление неслучайных (детерминированных) составляющих временных рядов.
3. Сглаживание и фильтрация полученных временных рядов.
4. Исследование случайной составляющей временного ряда, построение и проверка адекватности математической модели для её описания.

5. Прогнозирование развития изучаемого процесса на основе имеющегося временного ряда.
6. Исследование взаимосвязи между различными временными рядами.

В представленной работе исследование затрагивает сравнительно новую, но набирающую огромную популярность в сфере информационных технологий – Data Science (наука о данных). Data Science занимается задачами анализа, обработки и представления данных, также принимает немаловажную роль в развитии и формировании искусственного интеллекта. В данной работе используются различные методологии науки о данных и на основе теоретической базы Data Science проводится анализ новостных потоков и объемов биржевых торгов. Стоит заметить, что в прогнозировании временного ряда в представленной работе используется также искусственная нейронная сеть. Сравнительно недавнее «зарождение» Data Science как науки и популяризация данного раздела информатики характеризуют **научную новизну** данной работы.

Характеристика материалов исследования. Для реализации различных методов прогнозирования был использован язык программирования R, предназначенный для интеллектуального анализа и визуализации данных. В 2012 году язык R занял 1 место (30,7%) в опросе от KDNuggets[1][2] как программное обеспечение, которое широко применяется для обработки данных, то есть данный язык программирования зарекомендовал себя как надежный инструмент для статистической обработки данных.

Структура выпускной квалификационной работы. Магистерская работа содержит сокращения и обозначения, введение, 2 раздела: «Аналитический обзор моделей и методов прогнозирования», «Анализ биржевых торгов на основе новостных потоков», заключение, список использованных источников из 44 наименований и одно приложение – «Программная реализация анализа биржевых торгов на основе новостных потоков». Общий объем работы – 75 страниц.

[1] https://www.kdnuggets.com/

[2] Лидирующий сайт в области бизнес-аналитики, Big Data, Data Mining, науки о данных и машинного обучения.

1 Основное содержание работы

В введении содержится краткое описание волатильности, описываются актуальность и цель работы, а также задачи исследования.

Первый раздел «Аналитический обзор моделей и методов прогнозирования» содержит информацию о методах и математических моделях прогнозирования. Прогнозирование является одним из важных аспектов в организации управления любым хозяйствующим субъектом и отдельной отраслью в целом. Развитие методов и инструментов прогнозирования непосредственно связано с развитием информационных технологий, в частности, с ростом объемов хранимых данных и усложнением методов и алгоритмов прогнозирования. Одним из таких методов, применяемых в прогнозировании, является регрессионный анализ, который также используется для анализа временных рядов, тестирования гипотез и выявления скрытых взаимосвязей в данных.

Планирование и принятие управленческого решения практически всегда начинаются с прогнозирования, поэтому задача прогнозирования является одной из распространенных проблем в любой области экономики. Примеров прикладных задач прогнозирования в экономике огромное количество: прогнозирование потребительского спроса, объемов грузоперевозок, финансовых потоков компаний, курсовой стоимости акций, цен на недвижимость. Безусловно, область применения постановки задачи прогнозирования не ограничивается только экономической и финансовой сферами, данная проблема рассматривается в медицине, фармакологии, также сейчас набирает популярность политическое прогнозирование.

Прогнозирование направлено на определение тенденций динамики конкретного объекта или события на основе ретроспективных данных, то есть анализа его состояния в прошлом и настоящем. Обычно задачу прогнозирования делят на две подзадачи:

1. выбор модели прогнозирования;
2. анализ точности построенного прогноза.

По оценкам зарубежных и отечественных систематиков насчитывается более 100 моделей и методов прогнозирования, хотя самих базовых методов не так много. Большая часть моделей – это вариация уже существующего метода с некоторым дополнением или улучшением, то есть некоторые мето-

ды относят скорее к отдельным приемам или процедурам прогнозирования, а другие представляют собой набор отдельных приемов, отличающихся от базовых или друг от друга количеством частных приемов и последовательностью их применения.

В данном разделе рассматриваются такие методы и модели прогнозирования как прогнозная экстраполяция, метод наименьших квадратов, линейная регрессия, метод экспоненциального сглаживания, модели стационарных временных рядов, авторегрессионные модели со скользящими средними в остатках, модели рядов, содержащие сезонную компоненту, адаптивные методы прогнозирования, методы Хольта и Уинтерса, а также прогнозирование с использованием нейронных сетей, искусственного интеллекта и генетических алгоритмов.

Во втором разделе «Анализ биржевых торгов на основе новостных потоков» была реализована практическая часть магистерской диссертации. Данный раздел посвящен исследованию взаимосвязи между интенсивностью новостного потока и объемом биржевых торгов. Биржевой рынок и константное состояние цен – довольно редкое явление, точнее практически невозможное. На биржевом рынке постоянно происходят колебания цен, которые называют волатильностью. Волатильность – это статистический показатель, характеризующий изменение цены за некоторый промежуток времени. Это очень важный показатель, так как представляет собой меру риска использования финансового инструмента за определенный интервал времени.

Для расчета волатильности обычно применяется выборочное стандартное отклонение, что позволяет инвесторам с некоторой точностью определить риск приобретения финансового инструмента.Тем не менее использование выборочного стандартного отклонения не всегда дает верный результат, данная выборочная статистика обладает рядом недостатков. Полученная таким образом оценка волатильности будет состоятельной, то есть будет сходится к истинному значению, только в случае асимптотики, что требует бесконечно большой объем данных. Также этот показатель не позволяет идентифицировать изменение волатильности во времени. Одним из популярных примеров отображения волатильности в финансово-математических моделях является уравнение Блэка-Шоулза, однако в этом случае волатильность подразумева-

ют как постоянную величину, что на практике бывает не всегда. Сама оценка волатильности складывается из исследований исторических данных, которые позволяют проследить за динамикой цен в прошлом, и, исходя из этих данных, строится гипотеза о потенциальной изменчивости.

Известно, что различные виды активов имеют периоды низкой и высокой волатильности, то есть в некоторые моменты времени цена может меняться быстро, тогда как в другое время она практически не изменяется. Безусловно, такие колебания скорее всего не возникают на пустом месте, должно быть существуют некоторые компоненты, оказывающие влияние на волатильность. Такими факторами фундаментального характера могут быть релизы важных экономических отчетов, решения в отношении монетарной политики, принимаемые Центральными банками стран, какие-либо события, происходящие в политической сфере государства – одним словом новости.

Новостные данные были предоставлены одним из ведущих провайдеров Big Data-аналитики для финансовых учреждений – RavenPack [1]. RavenPack преобразовывает массивы неструктурированных больших данных, состоящих из новостей масс-медиа и информации из социальных сетей, в структурированные данные, в состав которых входят даже индикаторы настроения и внимания общественности к СМИ. Платформа, предоставляющая новостную аналитику по более чем 28 тыс. компаниям со всего мира, называется Raven Pack News Analitics (RPNA). RPNA анализирует каждую новость, публикуемую профессиональными поставщиками новостей (такими как Dow Jones или Reuters), а также сотни финансовых сайтов, онлайновых газет и даже блогов.

Полученные данные – это новости, выходившие в период с 1 января 2015 года по 31 сентября 2015 года. Сами данные представлены в обычном текстовом формате с расширением *.txt*.

Каждую новость RPNA представляет как массив, состоящий из следующих атрибутов: время выхода, id компании, релевантность новости, взвешенное настроение новости, тип новости, категория события и название компании; в соответствии с рисунком 1.1 представлен вид этих новостных данных.

[1] https://www.ravenpack.com/

При этом для анализа необходимы только название компании и дата выхода новости, связанной с этой фирмой.

	A	B	C	D	E	F	G
1	IDN_TIME	ITEM_ID	RELEVANCE	SENTIMENT	ITEM_TYPE	ITEM_GENRE	BCAST_REF
2	01 JAN 2015 00:00:04.307	nfa.service1.2015010	0.288675	-1	ARTICLE	NOT DEFINED	0700.HK
3	01 JAN 2015 00:00:04.307	nfa.service1.2015010	0.288675	-1	ARTICLE	NOT DEFINED	EXPE.O
4	01 JAN 2015 00:00:04.307	nfa.service1.2015010	1	-1	ARTICLE	NOT DEFINED	LONG.O
5	01 JAN 2015 00:03:47.095	nfa.service1.2015010	1	0	ARTICLE	NOT DEFINED	MGROS.IS
6	01 JAN 2015 00:23:05.716	nfa.service1.2015010	1	1	ALERT	NOT DEFINED	SYX.N
7	01 JAN 2015 00:23:07.328	nfa.service1.2015010	1	1	ARTICLE	NOT DEFINED	SYX.N
8	01 JAN 2015 00:30:00.498	nfa.service1.2015010	1	-1	ARTICLE	NOT DEFINED	BRO.N
9	01 JAN 2015 00:30:01.026	nfa.service1.2015010	1	0	ALERT	NOT DEFINED	BRO.N
10	01 JAN 2015 00:36:12.753	nfa.service1.2015010	1	-1	ARTICLE	NOT DEFINED	AIR.PA
11	01 JAN 2015 00:36:12.753	nfa.service1.2015010	1	-1	ARTICLE	NOT DEFINED	AIRA.KL
12	01 JAN 2015 00:36:12.753	nfa.service1.2015010	0.57735	-1	ARTICLE	NOT DEFINED	GE.N
13	01 JAN 2015 00:36:12.753	nfa.service1.2015010	0.57735	-1	ARTICLE	NOT DEFINED	SAF.PA
14	01 JAN 2015 01:00:00.823	nfa.service1.2015010	0.282843	0	ARTICLE	NOT DEFINED	AVT.V
15	01 JAN 2015 01:05:39.450	nfa.service1.2015010	1	0	ARTICLE	NOT DEFINED	600000.SS
16	01 JAN 2015 01:05:41.044	nfa.service1.2015010	1	0	ARTICLE	NOT DEFINED	600005.SS
17	01 JAN 2015 01:05:42.576	nfa.service1.2015010	1	0	ARTICLE	NOT DEFINED	600006.SS
18	01 JAN 2015 01:05:44.043	nfa.service1.2015010	1	0	ARTICLE	NOT DEFINED	600007.SS
19	01 JAN 2015 01:05:45.450	nfa.service1.2015010	1	0	ARTICLE	NOT DEFINED	600008.SS
20	01 JAN 2015 01:05:47.059	nfa.service1.2015010	1	0	ARTICLE	NOT DEFINED	600009.SS
21	01 JAN 2015 01:05:48.544	nfa.service1.2015010	1	0	ARTICLE	NOT DEFINED	600010.SS
22	01 JAN 2015 01:05:49.950	nfa.service1.2015010	1	0	ARTICLE	NOT DEFINED	600011.SS

Рисунок 1.1 — Новостные данные

Информация об объеме биржевых торгов компаний была получена из базы данных сервиса Yahoo!Finance[2] в соответствующий новостям период, то есть с 1 января 2015 года по 31 сентября 2015 года, по каждой компании. В соответствии с рисунком 1.2 показан вид биржевой информации об Apple Inc., представленный следующими полями: дата торгов, цена открытия, максимум цены, минимум цены, цена закрытия, скорректированная цена закрытия[3], объем торгов. Для анализа нас интересуют только дата и объем торгов.

После создания модели линейной регрессии и доказательства существования связи между новостями и торгами, полученные данные были преобразованы во временные ряды, а далее производился анализ временных рядов по следующим шагам:

1. Графическое представление и описание поведения временного ряда.

[2]https://finance.yahoo.com/

[3]Цена акций в любой конкретный день торговли, в которую была внесена поправка о включении любого дистрибутива и корпоративного действия, произошедшие в любое время вплоть до следующего дня открытия. Скорректированная цена закрытия часто используется при изучении исторических данных или для подробного анализа исторических данных.

Time Period: Dec 31, 2014 - Sep 30, 2015 Show: Historical Prices Frequency: Daily Apply

Currency in USD Download Data

Date	Open	High	Low	Close*	Adj Close**	Volume
Sep 30, 2015	110.17	111.54	108.73	110.30	105.31	66,473,000
Sep 29, 2015	112.83	113.51	107.86	109.06	104.13	73,365,400
Sep 28, 2015	113.85	114.57	112.44	112.44	107.35	52,109,000
Sep 25, 2015	116.44	116.69	114.02	114.71	109.52	56,151,900
Sep 24, 2015	113.25	115.50	112.37	115.00	109.80	50,219,500
Sep 23, 2015	113.63	114.72	113.30	114.32	109.15	35,756,700
Sep 22, 2015	113.38	114.18	112.52	113.40	108.27	50,346,200
Sep 21, 2015	113.67	115.37	113.66	115.21	110.00	50,222,000
Sep 18, 2015	112.21	114.30	111.87	113.45	108.32	74,285,300
Sep 17, 2015	115.66	116.49	113.72	113.92	108.77	64,112,600
Sep 16, 2015	116.25	116.54	115.44	116.41	111.14	37,173,500

Рисунок 1.2 — Биржевая информация по Apple Inc

2. Выделение и удаление неслучайных (детерминированных) составляющих временного ряда.
3. Сглаживание и фильтрация временного ряда.
4. Исследование случайной составляющей временного ряда, построение и проверка адекватности математической модели для её описания.
5. Прогнозирование развития изучаемого процесса на основе имеющегося временного ряда.
6. Исследование взаимосвязи между различными временными рядами.

В Приложении А представлен программный код, реализующий анализ биржевых торгов на основе новостных потоков, на языке программирования R.

ЗАКЛЮЧЕНИЕ

Подводя итоги выпускной квалификационной работы, хочу отметить, что мною были представлены и решены все установленные во введении задачи и за их счет была достигнута цель работы.

В ходе исследования был изучен синтаксис языка программирования R, а после на основе полученных результатов был проведен анализ, исходя из которого был сделан вывод, что связь между новостными потоками и волатильностью биржевых торгов существует. После приведения данных к временному ряду был выполнен комплексный анализ временных рядов, что соответствует теме магистерской диссертации, то есть для проведения анализа инвестиционных рисков была использована новостная аналитика. В ходе анализа было представлены визуализация и описание поведения временного ряда. После были выделены и удалены детерминированные компоненты, чтобы в дальнейшем для анализа использовать только случайные составляющие временного ряда. И для этого была произведена сезонная декомпозиция. Так как для построения различных моделей тяжело применять случайные компоненты в «чистом» виде, то к полученным результатам было применено сглаживание и фильтрация временного ряда с помощью метода экспоненциального сглаживания и методов Хольта и Уинтерса.

Для проверки адекватности математической модели использовались различные показатели точности, в том числе информационный критерий Акаике (AIC) и информационный критерий Байеса (BIC). Также были построены ARIMA-модели. И, наконец-то, с помощью нейронных сетей была воспроизведена попытка спрогнозировать торги на основе новостей. Однако, как показало исследование, спроектированная модель оказалась недостаточной для прогнозирования объема биржевых торгов. Возможно, это связано с тем, что объемы биржевых торгов – это комплексная переменная, на которую влияют множество других факторов, поэтому рассматривать только одни новости в качестве внешнего регрессора торгов, причем включая только количественную характеристику и исключая качественные показатели вроде релевантности, взвешенное настроение статьи и т.д., не совсем верно. Однако невозможно отрицать взаимосвязь объема биржевых торгов и новостей, потому что исходя из вышеизложенных результатов, получается, что новости и торги

действительно коррелирует между собой, но для прогнозирования необходимо использовать куда сложную модель, которая будет учитывать различные факторы и явления.

Таким образом, было применено знание теоретической базы по методам прогнозирования на основе реальных данных и получен опыт решения задачи прогнозирования, затрагивающую несколько сфер наук: экономику, статистику, Data Science и Data Mining.