

ОДЕСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ імені І.І.МЕЧНИКОВА

(повне найменування вищого навчального закладу)

Факультет математики, фізики та інформаційних технологій

(повне найменування інституту, назва факультету (відділення))

Кафедра математичного забезпечення комп'ютерних систем

(повна назва кафедри (предметної, циклової комісії))

Дипломна робота

на здобуття ступеня вищої освіти «магістр»

(освітньо-кваліфікаційний рівень)

на тему: «Аналіз часових рядів статистичними методами Data Mining»

«Time series analysis with statistical Data Mining methods»

Виконав: студент денної форми навчання

Спеціальності 126 – Інформаційні системи та технології.

(шифр і назва напрямку підготовки, спеціальності)

Гайдей Роман Вікторович

(прізвище, ім'я, по-батькові)

Керівник

кандидат фіз.мат. наук, доц.

Петрушина Т. І.

(науковий ступінь, вчене звання, прізвище та ініціали, підпис)

Рецензент

зав. каф. МЗКС, д-р тех. наук професор

Малахов Є. В.

(науковий ступінь, вчене звання, прізвище та ініціали)

Рецензент

зав. каф. КІСМ, к-т тех. наук, доцент

Шапорин Р. О.

(науковий ступінь, вчене звання, прізвище та ініціали)

Рекомендовано до захисту:

Захищено на засіданні ЕК № ____

Протокол засідання кафедри

протокол № __ від «__» ____ 2018р.

№ ____ від «__» ____ 2018 р.

Оцінка ____ / ____ / ____

(за національною шкалою, шкалою ECTS, бали)

Завідувач кафедри

Голова ЕК

Є.В. Малахов

(підпис)

(прізвище, ініціали)

О.О. Арсірій

(підпис)

(прізвище, ініціали)

Одеса – 2018

АННОТАЦИЯ

В дипломной работе рассматривается анализ временных рядов статистическими методами Data Mining. Объектом исследования стали колебания курсов валют на финансовых рынках.

Целью работы было построение прогнозирующей модели и осуществление прогноза движения курса валют на ближайшее время. Построение прогноза, основанного на техническом анализе и подкрепленного математически, имеет критическую важность для трейдеров, получающих прибыль за счет разницы курсов валют, поскольку в этой сфере нет общепринятой модели, а торговые стратегии строятся на субъективном понимании рынка и самостоятельно выведенных закономерностях.

Предметом исследования стали регрессионные методы Data Mining, в частности, семейство ARIMA моделей, техники предварительной обработки данных, а также средства анализа данных языка Python.

В ходе исследования были проанализированы данные котировок валютной пары EUR/USD, которые предоставлены базой торгового терминала MetaTrader4, проведен анализ изменений заработной платы, данные о которой предоставлены Единым архивом экономических и социологических данных, выполнена программная реализация построения ARIMA модели, осуществлен прогноз для обоих наборов данных и проведен сравнительный анализ модели и полученного прогноза для каждого из случаев.

АНОТАЦІЯ

У дипломній роботі розглядається аналіз часових рядів статистичними методами Data Mining. Об'єктом дослідження стали цінові коливання курсів валют на фінансових ринках.

Метою дослідження була побудова прогнозуючої моделі та здійснення прогнозу руху курсу валют на найближчий час. Побудова прогнозу, що базується на технічному аналізі та підкріпленій математично, є проблемою критичної важливості для трейдерів, що отримують прибуток за рахунок різниці курсів валют, оскільки у цій сфері немає загальноприйнятої моделі, а торгові стратегії базуються на суб'єктивному розумінні валютного ринку та власноруч знайдених закономірностях.

Предметом дослідження стали регресивні методи Data Mining, зокрема, сімейство ARIMA моделей, техніки попередньої обробки даних, а також засоби аналізу даних мови програмування Python.

У ході дослідження були проаналізовані дані котирувань валютної пари EUR/USD, які були надані відкритою базою даних торгового терміналу MetaTrader4, здійснений аналіз змін заробітної плати в Росії, дані про яку були надані Єдиним архівом економічних та соціологічних даних, виконана програмна реалізація побудови ARIMA моделі для обох наборів даних і здійснений порівняльний аналіз отриманої моделі та прогнозу для кожного з випадків.

ABSTRACT

The research deals with the time series analysis by using statistical Data Mining methods. Changes of quotes of currency pairs in financial markets are the object of study.

The goal of this research is to build a prediction model and make a forecast of the price movement for the nearest future. Forecast, based on the technical analysis and that is mathematically supported, has a critical importance for all traders whose profit is formed by the exchange difference as there is no common trading approach and each trader has to form his strategy using his personal understanding of financial markets and the dependencies found.

Regression methods of Data Mining such as ARIMA models, techniques of data preprocessing and Python data analysis tools became the subject of this research.

During the research EUR/USD quotes and salary growth in Russia were analyzed. Currency quoted were provided by Metatrader4 database, and salary growth data were provided by the Joint Economic and Social Data Archive. For each of the data sets a software that builds a prediction model and makes a forecast was implemented. After that the results for each data set were compared for accuracy and forecast stability.

ЗМІСТ

	стор.
ПЕРЕЛІК СКОРОЧЕНЬ, УМОВНИХ ПОЗНАЧЕНЬ ТА ТЕРМІНІВ.....	7
ВСТУП.....	8
1 ОГЛЯД ТА АНАЛІЗ АНАЛОГІВ.....	11
2 DATA MINING ТА МАШИННЕ НАВЧАННЯ.....	15
2.1 Цілі та задачі Data Mining.....	15
2.2 Машинне навчання.....	18
2.3 Типи ознак.....	20
2.4 Висновки.....	21
3 ЧАСОВІ РЯДИ ТА ЇХ ПРОГНОЗУВАННЯ.....	23
3.1 Задача прогнозування часового ряду.....	23
3.2 Автокореляція.....	24
3.3 Стаціонарність.....	27
3.4 Моделі ARMA та ARIMA. Теорема Волда.....	30
3.5 Моделі SARMA и SARIMA.....	33
3.6 Вибір ARIMA параметрів. Критерий Акаїке.....	34
3.7 Здійснення прогнозу.....	37
3.8 Аналіз залишків.....	38
3.9 Висновки.....	40
4 ПРАКТИЧНА РЕАЛІЗАЦІЯ.....	42
4.1 Особливості початкових даних та їхня попередня оброб.....	42
4.2 Прогнозування руху курсу валют.....	46
4.3 Прогнозування заробітної плати.....	57
4.4 Висновки.....	66
ВИСНОВКИ.....	67
СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ.....	69
ДОДАТОК А КОД МЕТОДУ КОНВЕРТАЦІЇ ДАНИХ КОТИРУВАНЬ...	71
ДОДАТОК В ЧАСТИНА ФАЙЛУ З ДАНИМИ КОТИРУВАНЬ ПІСЛЯ ЕРЕТВОРЕНЬ.....	75

ДОДАТОК С КОД ПОБУДОВИ МОДЕЛІ ARIMA ДЛЯ РЯДУ ВАЛЮТНОЇ ПАРИ З ДАНИМИ ЗА 3 ТИЖНІ.....	76
ДОДАТОК D КОД ПОБУДОВИ МОДЕЛІ ARIMA ДЛЯ РЯДУ ВАЛЮТНОЇ ПАРИ З ДАНИМИ ЗА 1 ТИЖДЕНЬ.....	80
ДОДАТОК E КОД ПОБУДОВИ ПРОГНОЗУ ЗАРОБІТНЬОЇ ПЛАТИ.....	84

ПЕРЕЛІК СКОРОЧЕНЬ, УМОВНИХ ПОЗНАЧЕНЬ ТА ТЕРМІНІВ

Скорочення

ЕЕТ – Східноєвропейський час. Влітку та взимку співпадає з Київським часом.

Терміни

Брокер – компанія, що надає своїм клієнтам технічний доступ до торгівлі на фінансових ринках.

Торгівельний термінал – програма, надана брокером, що надає можливість отримати дані о котируваннях валютних пар, акціях, індексах, а також здійснювати торгівельні операції за ними.

Тренд – стійкий рух ціни у певному напрямку тривалий час.

Котирування валютної пари А/В (А по відношенню до В) – ціна у валюті В, за якою здійснюється продаж 1 од. валюти А.

Тайм фрейм – часовий інтервал, що використовується торгівельним терміналом для групування значень котирувань при побудові графіку коливань ціни чи запису значень котирувань до бази.

Треjder – учасник ринку, що здійснює обмінні валютні операції.

Точка входу – шаблон коливань ціни, знайдений трейдером самостійно чи за допомогою засобів технічного аналізу, при виникненні якого трейдер впевнений у подальшому руху ціни у ступені достатньої для здійснення будь-яких обмінних операцій.

ВСТУП

Комп'ютеризація нашого суспільства значно збільшила наші можливості у збереженні даних з різних джерел. На сьогоднішній день ми живемо у світі, де майже кожен у кожному аспекту нашого життя щоденно накопичується величезна кількість даних. Відтак, сайти онлайн-кінотеатрів можуть зберігати дані о переглянутих користувачем фільмів та виставлених їм оцінках, банки зберігають інформацію о своїх клієнтах та дані о виконаних транзакціях за їхніми рахунками. Однак, чи є у цих даних якась користь? А якщо є, чи можливо її якось вилучити?

У пошуках відповідей на ці питання виникла така наука як аналіз даних і такий напрямок як Data Mining. Аналіз даних розглядає методи попередньої обробки даних (підготовку даних, зібраних із різних джерел, обробку відсутніх значень, пониження розмірності, видалення надлишкових даних) і опис задачі у термінах машинного навчання. В свою чергу Data Mining об'єднує під собою різні алгоритми, що дозволяють знаходити залежності (шаблони), що представляють собою нові знання, які неявно присутні у великих обсягах даних.

Алгоритми Data Mining знайшли своє застосування у рішенні різноманітних задач у таких сферах діяльності як медицина, торгівля, банківська сфера, пошукові системи, економіка та ін.

Одним із видів задач, у яких застосовуються методи Data Mining, є задача прогнозування, яка полягає у передбаченні майбутніх подій. Даний вид задач спрямований на визначення тенденцій динаміки об'єкту чи події на основі ретроспективних даних, тобто його аналізу у минулому та теперішньому. Прогнозувати можливо різні події: від землетрусів до епідемій та рейтингів телесеріалів.

Задача прогнозування часто розв'язується при аналізі часових рядів для передбачення майбутніх значень ряду. Дана задача є актуальною на

фінансових ринках, де вміння прогнозувати рух ціни грає ключову роль в отриманні прибутку на різниці курсів.

Даний спосіб заробітку активно розвивався після подій «Чорної середи» 16 вересня 1992 року, коли відбулося значне різке здешевлення фунту стерлінгів. Події цього дня пов'язують з діяльністю Джорджа Сороса, фонд якого одноразово продав близько 5 мільярдів фунтів стерлінгів в обмін на 15 мільярдів німецьких марок, що і викликало різке падіння курсу. На фоні цього падіння учасники ринку також почали продавати британський фунт, що дозволило Джорджу Соросу дочекавшись, доки курс знизиться ще сильніше, здійснити зворотній обмін, заробивши при цьому по різних оцінках від 1 до 1,5 мільярдів доларів США.

На сьогоднішній день у цій сфері не існує загальноприйнятої стратегії торгівлі, яка б дозволила гарантовано заробляти на ціновому русі. Стратегії торгівлі, що використовуються трейдерами зараз, загалом базуються на їхньому власному суб'єктивному розумінні ринку, самостійно знайдених закономірностях у цінових коливаннях чи даних технічних індикаторів та, як правило, ніяк не обґрунтовані математично.

Цим активно користуються різні фінансові брокери, рекламуючи стратегії власної розробки за допомогою реклами чи різних каналів «успішних» підставних трейдерів, створюючи ілюзію легкого заробітку. Дані стратегії обіцяють гарантований і стабільний прибуток, однак, на практиці, призводять до повної втрати грошових коштів через низьку точність прогнозу. З цього випливає, що закономірності коливання цін актуально шукати самостійно, спираючись на технічний аналіз і математично обґрунтовані висновки, що надає даній темі велике практичне і наукове значення.

Враховуючи вищесказане, об'єктом дослідження стали цінові коливання валютних пар на фінансових ринках.

Метою даного дослідження стало формулювання задачі прогнозування у термінах машинного навчання та побудова прогнозуючої моделі для передбачення цінового руху.

Предметом дослідження стали регресивні методи аналізу даних, зокрема, сімейство моделей класу ARIMA, а також бібліотеки аналізу даних мови Python.

У рамках дослідження необхідно виконати наступні задачі:

- проаналізувати початкові дані та виконати їхню попередню обробку;
- побудувати графік часового ряду, що аналізується, та виконати його візуальну оцінку;
- сформулювати задачу в термінах машинного навчання та побудувати прогнозуючу модель;
- провести аналіз отриманої моделі на предмет наявності будь-яких істотних недоліків;
- візуалізувати отриману модель. Оцінити якість апроксимації. Здійснити прогноз та проаналізувати його.

ВИСНОВКИ

У ході роботи були проаналізовані дані про курс валют EUR/USD та заробітну плату, створено програмну реалізацію на мові Python побудови прогнозуючих моделей та здійснено прогноз курсу валют до кінця дня та заробітної плати на наступні два роки.

Для цього необхідно було виконати попередній аналіз даних та обробку даних. Дані про заробітну плату не вимагали будь-якої попередньої обробки, тоді як у даних про значення курсу EUR/USD був відсутній ряд записів, а формат дати містив роздільники, що не дозволяють перетворити значення в тип даних `DateTime` автоматично при зчитуванні. Для вирішення цих проблем додатково була створена програмна реалізація перетворення роздільників та заповнення відсутніх записів на мові C#. Після, вже на мові Python, вихідні дані були візуалізовані. В обох випадках при візуальному аналізі було видно, що вихідні ряди виявилися нестационарними. Формально це підтверджувалося критерієм Дікі-Фуллера, який відхилив гіпотезу про нестационарність, внаслідок чого був потрібен ряд додаткових перетворень для приведення кожного ряду до стаціонарного виду.

Після цього були побудовані корелограми автокореляційної та часткової автокореляційної функцій, за якими були визначені початкові наближення параметрів p, P, q, Q . Після чого для кожної моделі з отриманої множини був розрахований критерій Акаїке, після чого оптимальною вважалася модель з найменшим значенням цього критерія. Отримані моделі апроксимували початкові дані досить добре, однак, результати апроксимації ряду заробітної плати були дещо краще внаслідок малої кількості зовнішніх факторів і яскраво вираженої сезонності. При цьому якість апроксимації помітно погіршилась при зменшенні кількості даних в навчальній вибірці.

Отриманий прогноз курсу валют виявився вірним лише на короткостроковому інтервалі, що не дивно, так як на коливання курсу впливають економічні новини, реакція учасників ринку на які неможливо спрогнозувати заздалегідь. У той же час прогноз для заробітної плати виявився стабільним на довгостроковому інтервалі та співпав з реальними даними.

Отримані результати показують, що в перервах між виходом новин учасники ринку ведуть торгівлю в одному ключі, внаслідок чого тенденцію руху ціни можна прогнозувати. Проте, отриманих результатів ще недостатньо для практичного застосування в реальній торгівлі. В навчальну вибірку необхідно постійно вносити нові дані про котирування в процесі торгівлі та оновлювати короткостроковий прогноз. При цьому на практиці в процесі торгівлі трейдер працює з декількома валютними пар, а не з однією, через що відслідковувати торгову ситуацію необхідно в реальному часі по 10-15 валютним парам.

Тому необхідна інтеграція з джерелом даних котирувань, відображених в торговельному терміналі певного брокера. Це дозволить програмно реалізувати торгового помічника, який в реальному часі буде збирати дані за обраними валютними парами і кожного разу оновлювати прогноз.

Крім цього, часовий інтервал за 1 годину, на якому проводився аналіз, дозволяє згладити більшу частину викидів, внаслідок чого їх безпосередній пошук в вихідних даних не здійснювався, однак, їх наявність не дозволить побудувати хорошу модель для менших тайм фреймів, через що необхідне додаткове дослідження для визначення їхнього характеру, пошуку у початкових даних, прийняття рішень щодо того, як вони впливають на коливання ціни і що саме з ними потрібно вчинити.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Чубукова И.А. Data Mining, 2-е изд. испр. — НОУ «Интуит», 2016. — 471 с. — ISBN: 9785947748192
2. Han, Jiawei, Data mining: concepts and techniques / Jiawei Han, Micheline Kamber, Jian Pei. — 3rd edition — Morgan Kaufmann, 2011 — 744 с. — ISBN 978-0-12-381479-1
3. МФТИ. Конспект лекций и учебные материалы онлайн-курса «Машинное обучение и анализ данных», [Электронный ресурс] — Режим доступа: <https://www.coursera.org/specializations/machine-learning-data-analysis>
4. База данных торгового терминала MetaTrader4 [Электронный ресурс] — Режим доступа: <https://www.metatrader4.com/ru/download>
5. Единый архив экономических и социологических данных [Электронный ресурс] — Режим доступа: <http://sophist.hse.ru/hse/nindex.shtml>
6. Stefan Lyocsa. Unit-root and stationarity testing with empirical application on industrinal production of CEE-4 countries / Stefan Lyocsa, Tomas Vyrost, Eduard Baumohl — Munich Personal RePEc Archive [Электронный ресурс], 16 March 2011 — 28 p. — Режим доступа: <https://mpira.ub.uni-muenchen.de/29648/>
7. Michael Snipes, Model selection and Akaike Information Criteria: An example from wine ratings and prices / Michael Snipes, D. Christopher Taylor // Wine Economics and Policy, Volume 3, Issue 1, June 2014 — p. 3-9
8. Iberedem A. Iwok, Justification of Wold's Theorem and the Unbiasedness of a Stable Vector Autoregressive Time Series Model Forecast — International Journal of Statistics and Probability, Vol. 6, №2 — March 2017
9. David S. Stoffer, A note on the Ljung-Box-Pierce portmanteau statistic with missing data / David S. Stoffer, Clelia M.C. Tolo — Statistics & Probability Letters 13 — North-Holland — 1992 — p. 391 — 396.
10. Гайдей Р.В. Застосування Data Mining алгоритмів для виявлення трендів на фінансових ринках // Тези доповідей XIV Міжнародної

науково-практичної конференції «Військова освіта і наука: сьогодення та майбутнє» [Електронний ресурс] – Київ – 2018 – 404 стр. – Режим доступу: <http://mil.univ.kiev.ua/page/lib/32>

11. Данил Байбак. Прогнозирование временных рядов с помощью Prophet от Facebook [Електронний Ресурс] – Режим доступу: <https://dou.ua/lenta/articles/forecasting-time-series-with-prophet/>