Міністерство освіти і науки України
Харківський національний університет радіоелектроніки

Факультет Комп'ютерних наук
(повна назва)

Кафедра Програмної інженерії
(повна назва)

АТЕСТАЦІЙНА РОБОТА
Пояснювальна записка

другий (магістерський)
(рівень вищої освіти)

Дослідження статистичних методів для аналізу замовлень користувачів в інтернет-магазинах
(тема)

Виконав: студент 2 курсу, групи ІПЗм-17-1.
спеціальності 121- Інженерія програмного забезпечення
(код і повна назва спеціальності)
Освітньо-наукової програми
Інженерія програмного забезпечення
(повна назва освітньої програми)

Кушнаренко О.О.
(прізвище, ініціали)
Керівник проф. Власенко Л.А.
(посада, прізвище, ініціали)

Допускається до захисту

Зав. кафедри, проф. _____ З.В.Дудар

2019 р.

Харківський національний університет радіоелектроніки

Факультет Комп'ютерних наук

Кафедра Програмної інженерії

Рівень вищої освіти другий (магістерський)

Спеціальність 121-Інженерія програмного забезпечення
(код і повна назва)

освітньо-професійна програма Інженерія програмного забезпечення
(повна назва)

ЗАТВЕРДЖУЮ:
Зав. кафедри _____
(підпис)
«___»_____ 20 ___ р.

ЗАВДАННЯ
НА АТЕСТАЦІЙНУ РОБОТУ

студентові Кушнаренко Олександр Олександрович
(прізвище, ім'я, по батькові)

1. Тема роботи Дослідження статистичних методів для аналізу замовлень користувачів в інтернет-магазинах

затверджена наказом по університету від "___" _____20 ___ р № _____
заповнюється вручну після отримання наказу

2. Термін подання студентом роботи до екзаменаційної комісії

3. Вихідні дані до роботи *Статистичні методи, пояснювальна записка. Використовувати середовище розробки PyCharm та операційну систему Windows.*

4. Перелік питань, що потрібно опрацювати в роботі *мета роботи, аналіз проблемної галузі, постановка задачі, аналіз продажів у інтернет магазинах, аналіз факторів, що впливають на безпеку руху, огляд існуючих методів для аналізу даних, огляд методів прогнозування, дослідження засобів покращення існуючих методів.*

5. Перелік графічного матеріалу із зазначенням креслеників, схем, плакатів, комп'ютерних ілюстрацій (слайдів) *титул, мета роботи, наукові та практичні завдання роботи, вирішення проблем аналізу даних про продажі інтернет магазинів, запропонований метод аналізу даних, платформа реалізації, програмна реалізація алгоритму та результати дослідження, висновки*

6 Консультанти розділів роботи

Найменування розділу	Консультант (посада, прізвище, ім’я, по батькові)	Позначка консультанта про виконання розділу	
		підпис	дата
Спецчастина	проф. Власенко Л. А.		

КАЛЕНДАРНИЙ ПЛАН

№	Назва етапів роботи	Терміни виконання етапів роботи	Примітка *
1.	Аналіз предметної галузі		
2.	Огляд існуючих методів		
3.	Алгоритми прогнозування		
4.	Підготовка пояснювальної записки		
5.	Спецчастина		
6.	Підготовка презентації та доповіді		
7.	Попередній захист		
8.	Нормоконтроль, рецензування		
9.	Занесення диплома в електронний архів		
10.	Допуск до захисту у зав. кафедри		
* заповнюється вручну після виконання чергового пункту			

Дата видачі завдання _____ _____2019 р.

Студент _____
(підпис)

Керівник роботи _____ __проф. Власенко Л. А._____
(підпис) (посада, прізвище, ініціали)

РЕФЕРАТ/ABSTRACT

Пояснювальна записка до атестаційної роботи: 71 с., 3 табл., 10 джерел, 2 додатки.

ПРОГНОЗУВАННЯ, РЕГРЕСІЙНІ МОДЕЛІ, АВТОРЕГРЕСІЙНІ МОДЕЛІ, ЕЛЕКТРОННА КОМЕРЦІЯ.

Об’єктом дослідження є моделі прогнозування у контексті електронної комерції.

Метою роботи є дослідження та розробка алгоритмів, що можуть бути використани при аналізі замовлень в інтернет магазинах.

Методи розробки базуються на алгоритмах для обробки даних минулого досвіду та на побудові моделей на основі проблемної області.

У результаті роботи були досліджені та проаналізовані алгоритми, що будуть використані у майбутньому для реалізації комплексної системи.

FORECASTING, REGISTRATION MODELS, AUTHORIZATION MODELS, ELECTRONIC COMMERCE.

The object of the research is the forecasting model in the context of e-commerce.

The purpose of the work is to research and develop algorithms that can be used in the analysis of orders in online stores.

Development methods are based on algorithms for processing past experience data and building models based on a problem area.

As a result of the work, the algorithms that will be used in the future for the implementation of the complex system were investigated and analyzed.

ЗМІСТ

ВСТУП

Електронна комерція або e-commerce, безповоротно увійшла в наше повсякденне життя. Сьогодні в Інтернеті можна купити все: від посуду і косметики до автомобіля. І якщо ще пару років тому електронна торгівля викликала підозру і недовіру, то сьогодні їй віддає перевагу все більший і більший відсоток населення всієї Землі.

Так як зараз відбувається справжній бум електронної торгівлі, особливо на ринках країн, що розвиваються, роздрібні торговці все частіше розглядають електронну комерцію в якості ключового елемента в своїх стратегіях глобального розширення. Можливість брати участь в міжнародній електронній комерції допомагає стимулювати зростання електронних торгових майданчиків і альтернативних Інтернет-каналів. Середньорічні темпи зростання ринку електронної комерції в світі за даними eMarketer, основного порталу з досліджень маркетингу, MI та комерції, складають близько 18-20% в рік. Це приблизно 3-4% від загального рітейлу в Росії і до 10-12% в США і інших розвинених країнах; таким чином, середній світовий рівень становить приблизно 6-8%. За деякими прогнозами частка електронної комерції в загальному ритейлі досягне 20% в найближчі кілька років [1].

Для постійного збільшення продажів інтернет магазини повинні постійно покращувати сервіс для своїх покупців. Допомагає рухатись у правильному напрямку аналітика. Веб аналітика інтернет магазину має на увазі постійний збір, аналіз і інтерпретацію даних, роботу з основними метриками.

Цілі моніторингу:

- поліпшення якості роботи ресурсу;
- оптимізація ефективності онлайн-бізнесу;
- отримання інформації для прийняття рішень;
- оптимізація бізнес-процесів.

Для збору інформації необхідно вибрати одну або кілька систем статистики. Це спростить моніторинг і дозволить отримати коректну інформацію. Сьогодні в Рунеті найбільшою популярністю користуються: Google Analytics і Яндекс.Метрика. Які плюси цих сервісів і чому так багато інтернет магазинів їх використовує? Великі функціональні можливості для відстеження конверсій, наявність спеціального модуля для інтернет-магазинів «Електронна торгівля», звіти персоналізовані. Але конверсія – одна з величезної кількості метрик які можуть допомогти вести бізнес.

Як оптимізувати бізнес процеси знаючи лише конверсію? Бізнес потребує більше метрик для своїх цілей – оптимізувати роботу складів, оптимізувати роботу с постачальниками товарів, прогнозувати – які регіони потребують наявність тих чи інших товарів та коли вони цього потребують. Сучасні сервіси повинні агрегувати величезні об’єми даних, вміти з ними працювати не втрачаючи перфомансу, а також представляти великий об’єм функціоналу, який зможе масштабуватися під особливості бізнесу, може бути постійно модифікуємим та актуальним, мати легкий процес підтримки. Наприклад, однією з найважливіших задач таких сервісів є вміння спрогнозувати продажі бізнесу. На важливість такого функціоналу дуже впливає проблема залежування товарів на складах, шо є одним з шагів де бізнес втрачає найбільше грошей. Звичайно, прогнозування продажів може залежати від великої кількості параметрів, особливостей бізнесу, особливостей країн у яких він працює. Саме тому дослідження методів та створення такого сервісу може мати великий економічній інтерес і потенціал.

1 АНАЛІЗ ПРЕДМЕТНОЇ ГАЛУЗІ

1.1 Цілі прогнозування

Електронна комерція стала невід'ємною частиною сучасної економіки. Все більше споживачів купують товари за допомогою мережі Інтернет, а комерційні організації так чи інакше використовують можливості даної мережі при здійсненні підприємницької діяльності. Загальний світовий обсяг продажів в одному тільки споживчому сегменті електронної комерції перевищив позначку в 1 трлн дол.

У багатьох сферах діяльності вивчення різного роду процесів відіграє найважливішу роль. Лікар за поточними симптомами намагається визначити подальший розвиток хвороби, інженер - наскільки міцним повинно бути крило літака, щоб витримати відповідне навантаження, власник магазину - в яких обсягах проводити закупівлю на наступний тиждень. Прикладів можна навести безліч, однак всі вони мають одну загальну деталь - передбачити точний результат неможливо. Наприклад, по поточному стану колонії бактерій можна сказати, яка в точності буде популяція цієї колонії через деякий проміжок часу. Займатися прогнозуванням таких процесів можна по-різному. У загальному випадку тут може бути два кардинально різні підходи. Перший, назвемо його ґрунтовним, спирається на досвід людини, його пізнання в деякій області, наукові методи і т.д. Другий - безпідставний метод - фактично спирається лише на віру людини. До методів другого типу можна віднести ворожіння на кавовій гущі, різні пророцтва та ін. Звернемося до методів першого типу. Що є досвід людини? По суті, це результат дій в минулому. Тобто, можна говорити про деяке наборі статистичних даних, аналізуючи який ми робимо висновки про те, як процес буде розвиватися в майбутньому. При цьому, як було сказано вище, результат прогнозу не є точною величиною - він може випадково коливатися близько деякого, найбільш ймовірного значення.

Продаж у сфері електронної комерції означає управління витратами, пропозицією та прибутком, щоб забезпечити клієнтам якісний досвід, зберігаючи

конкурентні переваги. Щоб досягти цього, бізнес повинен знайти баланс між зменшенням запасів запасів і витратами на зберігання. Продавці електронної комерції можуть використовувати прогнозування рекламних ресурсів, щоб допомогти знайти цей баланс. Продавці можуть забезпечити відповідний запас, прогнозуючи попит на продукцію та ймовірні продажі. Це змінить складність залежно від ніші, періоду, віку бізнесу та існуючих даних, але витрачання час на прогнозування продажів допоможе оптимізувати управління запасами та зменшити витрати.

Прогнозування продажів може допомогти такими способами для керування бізнесом:

– встановлення цілей: встановлення обсягу продажів, який ви хочете досягти в порівнянні з обсягом попереднього року, може мотивувати і направляти вас щодня. Прогнозування продажів може допомогти встановити ці цілі;

– якщо ви шукаєте інвесторів: аналогічно попередньому випадку з іграшками, якщо ви знаєте, що відбудеться неминуче зростання продажів, вам можуть знадобитися кошти, щоб підготувати себе з точки зору запасів і ресурсів. Хороший прогноз продажів є ключем до отримання доступу до інвесторівж;

– людські ресурси: існують певні періоди, які потребують більшої кількості співробітників, і вам буде потрібно, щоб це було заплановано. Знаючи, коли можуть настати ці періоди, ви заощадите багато грошей і уникнете того, що ви завалені роботою;

– бюджетування: ефективне розподіл ресурсів на основі потреб є ключовим завданням, яке неможливо здійснити без планування майбутніх продажів. Пам'ятайте приклади, які ми вже бачили.

У більшості випадків можна використовувати одну з двох основних моделей прогнозування запасів; кількісне прогнозування та якісне прогнозування. Кожен має свої власні плюси і мінуси.

Кількісне прогнозування. Тут створюється математичну модель для прогнозування майбутнього обсягу продажів на основі наявних даних. Чим більше даних є, тим точнішими будуть прогнози. Для більш точних прогнозів, принаймні рік даних про продаж слід порівнювати з високими та низькими показниками, піками, сезонними тенденціями та продажами, а також іншими зростаннями та падінням та періодом продажу.

Хороший процес кількісного прогнозування буде включати:

– встановити базовий попит;

– створити графік продажів за попередні роки;

– визначити вершини і долини і спробувати знайти події які їм відповідають;

– визначити періоди або повороти, які мають значення для бізнесу (щомісяця, щоквартально, раз на два місяці тощо);

– знайти відсоток збільшення продажів;

– використовуючи відсоток збільшення можно побудувати прогнози продажів для кожного наступного періоду. повторюючи це на основі продукту за продуктом можна отримати найточніші результати.

Якісне прогнозування є економічною моделлю, яка фокусується на прогнозах для економіки, ринкових сил і потенційного попиту. Можна досягти якісного прогнозування навіть без річних даних з власного магазину, оскільки більше на нього впливає економічний клімат, тобто зовнішній фактори. Ось декілька факторів, на яких будується якісне прогнозування:

– нові технології, які можуть змінити спосіб виготовлення або використання продукту;

– нові моделі продуктів;

– життєвий цикл продукту;

– нові версії продукту (нові кольори, особливості тощо);

– зміна цін;

– зміни доступності;

– публічне сприйняття.

Для деяких продуктів ці розрахунки відносно легкі. Наприклад, якщо магазин продає комп'ютери або технології, нові моделі виходять кожні 8-12 місяців, і бізнесу доведеться оновлювати інвентар і позбуватися старих запасів до цієї дати. З іншого боку, деякі продукти мають набагато довший життєвий цикл. Наприклад, кухонне начиння майже ніколи не виходить із стилю, що дозволить запасати їх протягом багатьох років або навіть десятиліть. Проте, суспільне сприйняття змінюється. Там, де яскраво-зелений може бути хітом один рік яскраво-синій може бути популярним наступний, повністю змінюється тенденції продажів.

Цей метод прогнозування також може допомогти прийняти правильні рішення при розміщенні замовлень. Наприклад, якщо бізнес знає, що вартість продукту зменшується, можна зменшити поточне замовлення до мінімуму. Якщо він знає, що витрати збільшені, можна замовити кілька витків, якщо витрати на зберігання та управління не перевищують зростання ціни.

Якісне прогнозування може допомогти контролювати ринковий попит і ціноутворення. Знаючи, що ринок насичений і є надлишок, можна знизити ціни, щоб збільшити продажі. А знаючи, що магазин єдиний постачальник, і не може берегти свій товар на складі, він можете підвищити ціни для задоволення ринкового попиту.

Основною метою прогнозування є перегляд і оцінка майбутніх наслідків, які можуть вплинути на прогнозовану змінну, або знайти відповіді на різні сценарії "що-якщо". Таким чином, зауваження відомих менеджерів «Якби я тільки знав, я б вибрав іншу стратегію», можна хоча б частково уникнути. Але в цілому визнається, що керівники та практикуючі фахівці мають недостатню формальну освіту в бізнес-прогнозуванні, і вони схильні відмовлятися від застосування цих моделей на основі математичних методів. Відповідно до різних питань, що обговорюються в літературі, прогнозування бізнесу може враховувати 3 категорії факторів: організаційні та екологічні змінні, які, як відомо, впливають на прогнозування;

вплив додаткових змінних, специфічних для конкретної фірми та середовища; і нехтують взаємозв'язку між різними аспектами організаційного прогнозування. Зазвичай, крім відомих факторів, які можуть впливати на залежну змінну, необхідно враховувати різні типи викидів.

У світі, що постійно розвивається, прогнозування нових і надзвичайно нових продуктів є ключовим для економічного добробуту. Прогнозування продажів нових продуктів має вирішувати основні проблеми, викликані відсутністю даних і невизначеністю того, як проривні технології та продукти будуть прийняті споживачами.

У сучасних ринкових умовах, коли конкуренція зростає з кожним днем і скорочується термін служби виробу, компанії повинні бути швидкими, гнучкими і гнучкими, щоб конкурувати. Це надконкурентне середовище означає, що влада зараз знаходиться в руках кінцевого споживача, тому підприємства, які здатні передбачити побажання клієнта, завжди будуть на крок попереду. Компанії, які прагнуть зберегти свою присутність на ринку, повинні бути здатні заздалегідь аналізувати та інтерпретувати ринкові зміни та зміни потреб клієнтів, що дозволяє їм належним чином здійснювати свою діяльність. Тепер це необхідно для підприємств передбачити товари або послуги, які будуть затребувані клієнтами, і зробити необхідну підготовку для постачання їх заздалегідь. Відповідно, дослідники звернули свою увагу на методи прогнозування попиту і вже розробили низку методів для цієї мети. По-перше, у застосуваннях часових рядів майже неможливо заздалегідь знати, яка модель прогнозування буде найкращою для даного набору даних. Це відповідає загальноприйнятому факту, що жодна модель не є найкращою для всіх ситуацій за будь-яких обставин. З цієї причини значення «кращого» можна тлумачити різними способами, наприклад, що найкращою моделлю є модель, прогнози якої добре працюють і демонструють незначну варіабельність продуктивності, незалежно від того, який часовий ряд аналізується.

Незважаючи на різноманітні та прогресивні методи, реалізовані в комерційних програмних продуктах, вони рідко використовуються у цілях малого бізнесу. Їхня вартість може бути однією з причин, але не тільки. Дійсно, для

отримання оптимізованої автоматичної обробки, впровадження таких систем на величезних і налаштованих базах даних може бути дуже вибагливим. Більше того, і, можливо, головна причина, бізнес хоче і повинен тримати контроль над своїми прогнозами. Жодна компанія не погоджується на те, щоб дозволити прогнозування програмного забезпечення, хоча це дуже точно. Фактично, автоматичні прогнози з програмного забезпечення, якщо вони існують, зазвичай використовуються як вихідні для остаточних прогнозів практиків. Кожна компанія використовує свої власні поради для виконання найкращого прогнозу. Основний прогноз, як правило, складається з базового прогнозу, видобутого з конкретного програмного забезпечення або, в основному, продажів минулого року. Результат може бути дуже точним, оскільки враховуються сезонність та вплив основних пояснювальних змінних.

1.2 Виявлення проблем та актуалізація рішень

Часові ряди є дуже зручним механізмом опису процесів, що працюють за принципом «чорний ящик». У таких системах ми не маємо можливості «зазирнути» всередину процесу, а бачимо лише кінцевий результат. Розділяють два види часових рядів: безперервні і дискретні. У безперервних рядах передбачається, що безліч значень параметра t - континуум, в дискретних – деяка дискретна множина. На практиці це означає, що в безперервних тимчасових рядах досліджуваний параметр можна виразити як деяку функцію від часу, а в дискретних ми маємо справу з набором вимірювань в деякі моменти часу. Таким чином, можна виділити два основних способу побудови дискретних часових рядів [3]. Зауважимо, що в разі безперервного часового ряду вимірювання теоретично також можуть проводитися безперервно, однак, на практиці, фіксування результатів вимірювань все одно

відбувається дискретним чином. Тому в подальшому ми будемо мати справу тільки з дискретними тимчасовими рядами.

Визначимо компоненти, які прийнято виділяти у тимчасових рядах:

– тренд - динаміка, що характеризує загальний розвиток часового ряду;

– циклічна компонента - динаміка, що має фазу зростання і зменшення, період якої займає досить великий проміжок часу;

– сезонна компонента - це регулярні коливання рівнів ряду в певний час доби, тижні, сезону і т. д. пов'язана з сезонними явищами (наприклад, погодними умовами) і людськими ритмами (наприклад, з фазами неспання і сну);

– календарні ефекти - скачки часового ряду, пов'язані з деякими передбачуваними календарними подіями (наприклад, святами або вихідними);

– аномальні явища (викиди) - непередбачувані скачки, що призводять до різких, але короткочасних відхилень ряду від загальної тенденції розвитку;

– структурні зрушення - непередбачувані скачки, що призводять до відхилень ряду від загальної тенденції розвитку, які позначаються на всьому його подальшу поведінку;

– випадкова компонента - безладні рухи досить великої частоти, пов'язані з впливом великої кількості невідомих факторів.

Деякі тимчасові ряди являють собою ту чи іншу компоненту в чистому вигляді, проте на практиці такі ряди зустрічаються вкрай рідко. Загалом тимчасової ряд може представляти собою адитивну, мультиплікативну або змішану комбінацію деяких з цих компонент. Крім того, далеко не всі тимчасові ряди мають досить просту структуру для розкладання на зазначені компоненти. Аналіз часового ряду проводиться на основі побудованої моделі, параметри якої підлягають оцінці.

Оскільки часовий ряд складається з компонент, можна спробувати виявити регулярну складову, зробити з її допомогою прогноз і видалити з часового ряду. Тоді залишиться тільки випадкова компонента. Розглянемо її як реалізацію деякої випадкової величини. Використовуючи статистичний критерій Колмогорова-

Смірнова, можна перевірити гіпотезу про те, що дана вибірка має якийсь конкретний розподіл. Припустимо, що нам вдалося «підібрати» деякий закон розподілення для нашої випадково величини. Тоді можна побудувати довірчий інтервал, який з наперед заданої ймовірністю α буде містити в собі значення такої випадкової величини. скрививши отриманий коридор відповідно до виділених компонентами, можна отримати кордон, усередині якої з ймовірністю α будуть лежати значення - в тому числі і прогнозоване - тимчасового ряду. На цьому, здавалося б простому, шляхи виникає одна проблема: в точності виявити компоненти часового ряду не завжди можливо. Тому на практиці користуються методами, що дозволяють наближено обчислювати значення цих компонент.

Тренд. При аналізі тренду виділяють чотири його основні види:

– поліноміальний тренд:

$$\tau_{\text{e}} = \alpha_0 + \alpha_1 t + \cdots + \alpha_0 t^p$$

Цей тип тренда застосовують для опису часових рядів з плавною, повільно змінюється з часом динамікою;

– експоненціальне тренд:

$$\tau_t = e^{a0} + \alpha_1 t + \cdots + \alpha_0 t^p$$

Такий тип тренда доцільно застосовувати для швидко зростаючих часових рядів;

– гармонійний тренд:

$$\tau_t = A\cos(2\pi f t + \varphi)\ ,$$

де A - амплітуда коливань,

f – частота,

φ - зсув по фазі.

Застосування гармонійного тренда виправдано, якщо в поведінці часового ряду чітко проглядається періодичність;

– тренд, який виражається логістичної функцією:

$$\tau_t = \frac{k}{1 + be^{-at}}$$

Цей тренд вигідно відрізняється від інших тим, що при t → ∞ має асимптоту k. часто зустрічається при аналізі часових рядів демографічних показників.

Оцінка параметрів тенденцій першого і другого виду не складає труднощів: досить застосувати метод найменших квадратів, який по суті зводиться до вирішення системи лінійних рівнянь.

Сезонна компонента. Для моделювання сезонної компоненти можна скористатися наступною формулою

$$v_t = \delta_{1t}\gamma_1 + \cdots + \delta_{ht}\gamma_h$$

де δ та γ - фіктивні змінні, кожна з яких відповідає своєму сезону.

Трактувати цю модель можна в такий спосіб: відхилення від тренда в j-му сезоні дорівнює λj. Нескладно помітити, що, тому що будь-яке спостереження може належати єдиному сезону, то $1 = \delta_{1t} + \cdots + \delta_{ht}$. Таким чином, при використанні МНК не вдасться однозначно оцінити параметри. Вирішувати цю проблему можна двома способами:

– покласти одну з змінних рівною нулю: в цьому випадку дана модель буде описувати сезонні відхилення щодо тренда одного з сезонів;

– більш вдалий спосіб полягає в покладенні на коефіцієнти λ1,. . . , Λh обмеження: $0 = \lambda_1 + \cdots + \lambda_h$ При цьому вплив сезонної компоненти як би центруеться - вплив на тренд дорівнює нулю.

Викиди, структурні зрушення, календарні ефекти. При моделюванні викидів можна дотримуватися двох кардинально різних методів:

– оскільки викиди непередбачувані, а їх вплив короткостроково, їх можна виключити з спостережень. У статистиці такі дані пов'язані з поняттям outlier-ів - даних, які різко виділяються із загальної маси. Цей прийом пов'язаний з одним тонким моментом: кількість викидів не повинно бути сильно великим, інакше, це може вказувати на неоднорідність даних в цілому;

– з іншого боку, можна спробувати врахувати аномальні викиди. Для цього можна скористатися фіктивними змінними: для кожного викиду заведемо власну фіктивну змінну δ_t Яка буде рівна одиниці в момент часу t.

Календарні ефекти можна моделювати аналогічно сезонної компоненті, проте слід пам'ятати, що в даному випадку буде потрібно завести за фіктивною змінною на кожен день, в якому даний ефект проявляється. З огляду на, що свята можуть зрушуватися або переноситися, облік цієї компоненти може привести до більш складних обчислень на стадії оцінки коефіцієнтів моделі. На щастя, моделювання цієї компоненти не завжди необхідно.

Загальний вигляд моделі. Припустимо, що нам вдалося виділити і промоделювати всі компоненти, присутні в ряді. Виникає важливе питання: як їх скомпонувати? Порівняно простою є модель наступного вигляду

$$x_t = \mu_t + \varepsilon_t, t = 1, \dots, N,$$

де μ_t - систематична складова,

ε_t - випадкова, що вдає із себе білий шум.

Як правило, μ_t моделюється як що складається з перерахованих вище компонент. Розрізняють дві моделі:

– адитивна. Нехай, для визначеності, модель тимчасового ряду містить три компоненти: тренд τ_t,сезонну компоненту v_t і випадкову складову ε_t. Тоді адитивна модель матиме такий вигляд:

$$x_t = \tau_t + \varepsilon_t + v_t$$

– мультиплікативна. В аналогічних позначеннях модель такого ряду можна записати як:

$$x_t = \tau_t * v_t * \exp(\varepsilon_t)$$

Тобто, досить про логарифмувати вихідний ряд. Це перетворення дозволяє залишитися в рамках лінійної регресії і, до того ж, обмежитися розглядом тільки адитивної моделі.

Прогнозування . У статистичних моделях залежність майбутнього значення від минулого задається в вигляді деякого рівняння. До них відносяться:

– регресивні моделі (лінійна регресія, нелінійна регресія);

– авторегресійні моделі (AR, ARMA, ARIMA, ARIMAX, GARCH, ARDLM);

– модель експоненціального згладжування;

– модель за вибіркою максимального подібності.

У структурних моделях залежність майбутнього значення від минулого задається в вигляді певної структури і правил переходу по ній. До них відносяться:

– нейромережеві моделі;

– моделі на базі ланцюгів Маркова;

– моделі на базі класифікаційно-регресійних дерев.

Методи прогнозування попиту на статистику / математику, що застосовуються на практиці, можна розділити на три потоки: часові ряди (також звані методи екстраполяції), причинний і зважений комбінований прогноз. Методи з часовими рядами - це ті, які визначають закономірності (тенденції, сезонність, циклічність і випадковість) і передбачають їх у майбутньому. Ці методики мають більш високу прогностичну точність у стабільних ринках. Приклади включають: ковзну середню; просте експоненціальне згладжування; Два параметри Хольта; і три параметра Winters. Часові ряди легко розробляються і вимагають мінімального обсягу даних, але не можуть передбачити раптових змін у попиті. Вона також може передбачати лише один-три періоди вперед з будь-яким ступенем точності. ARIMA - це більш просунута технологія часових рядів, яка поєднує в собі часові ряди і елементи регресії. Цей метод є більш точним для прогнозування попиту в довгостроковій перспективі. Вона може моделювати тенденцію / цикл, сезонність, а також інші фактори, що впливають на попит (пояснювальні змінні), але вимагає більшої кількості даних і є більш складним для розвитку. Причинно-наслідкові прийоми припускають, що майбутні продажі пов'язані зі змінами в інших змінних (ціна, акції, серед інших). Прикладами є регресія і ARIMAX (розширення ARIMA). Ці методи вимагають більшої кількості даних і є більш складними для їх розробки. З іншого боку, вони можуть включати інтервенційні змінні (з використанням фіктивних змінних).

Зважене комбіноване прогнозування об'єднує методи (тобто часові ряди, причинний і / або судження) і створює єдиний прогноз. Це можна зробити, задавши

кожному рівному або різному вазі. Ця комбінація перевершує більшість окремих прогнозів, оскільки упередження серед методів компенсують один одного. Навіть вважаючи, що точність прогнозування є важливою, практикуючі практики все ще практикують простіші методи прогнозування. Однією з причин є те, що моделі, які є результатом маркетингових досліджень, важко реалізувати на практиці. Це є суттєвою проблемою структурних моделей, оскільки «велика кількість спостережень на практиці, великий масив стану та контрольних змінних, а також частота прийняття рішень може зробити застосування структурних моделей неможливим». У наступному розділі ми обговорюємо великомасштабні / комп'ютерно-інтенсивні методи прогнозування попиту, які застосовуються в літературі з можливістю стати методами, що використовуються в маркетинговій практиці.

Основна увага цього розділу приділяється методам прогнозування попиту, типам даних та майбутнім дослідженням у сфері маркетингу. Для цього ми описали класифікацію моделей прогнозування попиту в маркетингу. Вони були розділені на два підходи: статистика / математика і великі дані / комп'ютерно-інтенсивні методи. Було зроблено два розділи для обговорення статистики / методів на основі математики, один про літературу та інший про практику. Розділ про прогнозування попиту в маркетинговій літературі пояснює різні класифікації знайдених методів: по системах попиту; за рівнем даних (сукупним або індивідуальним) та періодом запуску продукту (до або після запуску); їх цілі (прогнозування, вимірювання та тестування). Описано також типи моделі (описовий, структурний та редукований). Основна мета цього розділу полягає в тому, щоб протиставити прості методи, що використовуються практиками в прогнозуванні попиту, до більш складних моделей, наявних у маркетинговій літературі. Причина в тому, що моделі, які є результатом маркетингових досліджень, важко реалізувати на практиці.Зазначається, що застосування таких методів у дослідженнях прогнозування попиту на маркетинг все ще має розвиватися.

Регресивні моделі прогнозування.

Дані моделі є одними з найстаріших, однак зараз знаходять не саме велике застосування. Існує багато задач, які потребують вивчення відносини між двома і більше змінними. Для вирішення таких завдань використовується регресійний аналіз. В даний час регресія отримала широке застосування, включаючи завдання прогнозування і управління. Метою регресійного аналізу є визначення залежності між вихідної змінної і безліччю зовнішніх факторів (регресорів). При цьому коефіцієнти регресії можуть визначатися за методом найменших квадратів або методу максимальної правдоподібності.

Основні види регресійній моделі:

- проста лінійна регресія (linear regression);
- множинна регресія;
- нелінійна регресія.

Лінійна регресійна модель.

Найпростішим варіантом регресійній моделі є лінійна регресія. В основу моделі покладено припущення, що існує дискретний зовнішній фактор X (t), який впливає на досліджуваний процес Z (t), при цьому зв'язок між процесом і зовнішнім фактором лінійна. Модель прогнозування на підставі лінійної регресії описується рівнянням:

$$Z(t) = \alpha 0 + \alpha 1 X1(t) + \varepsilon t,$$

де α0 і α1 – коефіцієнти регресії,

εt - помилка моделі.

Для отримання прогнозних значень Z (t) в момент часу t необхідно мати значення X (t) в той же момент часу t, що рідко здійснимо на практиці.

Множинна регресійна модель.

На практиці на процес Z (t) впливають цілий ряд дискретних зовнішніх факторів X1(t)…Xs(t) . Тоді модель прогнозування має вигляд:

$$Z(t) = \alpha 0 + \alpha 1 X1(t) + \alpha s Xs(t) + \cdots + \varepsilon t,$$

де α0 і α1 - коефіцієнти регресії,

εt - помилка моделі.

Недоліком даної моделі є те, що для обчислення майбутнього значення процесу Z (t) необхідно знати майбутні значення всіх X(t), що майже нездійсненно на практиці.

Нелінійна регресійна модель.

В основу нелінійної регресійної моделі покладено припущення про те, що існує відома функція: Z (t) = F (X (t), A), де Z (t) - вихідний процес, X (t) - зовнішній фактор, від якого залежить процес Z (t), A - функція, параметри якої необхідно визначити в рамках побудови моделі прогнозування. Наприклад, можна припустити, що Z (t) = α1cos (X (t)) + α0. тоді для побудови моделі досить визначити параметри A = | α0, α0 | , Де α0 і α1 - коефіцієнти регресії. Однак на практиці рідко зустрічаються процеси, для яких вид функціональної залежності між процесом Z (t) і зовнішнім фактором X (t) заздалегідь відомий. У зв'язку з цим нелінійні регресійні моделі застосовуються рідко.

Авторегресійні моделі.

В основу авторегресійних моделей закладено припущення про те, що значення процесу Z (t) лінійно залежить від деякої кількості попередніх значень того ж процесу Z (t-1), ..., Z (t-p). Розглянемо основні види авторегресійних моделей:

- модель змінного середнього;
- модель з умовною гетероскедастичних;
- модель з розподіленим лагом.

Авторегресійна модель змінного середнього.

У статистиці і обробці сигналів модель авторегресійного змінного середнього (autoregressive moving average, ARMA), звана іноді моделлю Боксу-Дженкінса, застосовується для дослідження часових рядів.

Маючи тимчасової ряд X_t, модель авторегресійного змінного середнього дозволяє пояснити і, можливо, передбачити майбутні значення ряду. Модель складається з двох частин: авторегрессионной (AR) частини і змінного середнього (MA). Для згадки моделі зазвичай використовується позначення ARMA (p, q), де p

- порядок регресійної частини, а q - порядок змінного середнього. По суті своїй авторегресійна модель є полюсним фільтром з нескінченною імпульсною характеристикою, витлумаченим в контексті аналізу часових рядів. Для того, щоб модель була стаціонарної потрібно накласти деякі обмеження на параметри моделі.

Авторегресійна модель змінного середнього.

Особливістю даної моделі є припущення, що умовне значення дисперсії залежить від попередніх значень ряду і від попередніх значень дисперсії. В такому випадку дисперсія моделі описується наступним рівнянням:

$$\sigma_t^2 = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \cdots + \alpha_q \varepsilon_{t-q}^2 + \beta_1 \sigma_{t-1} + \cdots + \beta_p \sigma_{t-p},$$

де α_q, β_p - невід'ємні коефіцієнти моделі,

εt - помилка моделі.

Наведене рівняння називається моделлю GARCH (p, q) і має два параметра: p характеризує порядок авторегресії квадратів залишків (внесок попередніх значень ряду); q - кількість попередніх оцінок залишків (внесок попередніх значень дисперсії). Найбільш часте застосування дана модель отримала в фінансовому секторі, де з допомогою неї моделюється волатильність. На сьогоднішній день існує ряд модифікацій моделі під назвами NGARCH, IGARCH, EGARCH, GARCH-M і інші.

Авторегресійна модель з розподіленим лагом.

Авторегресійна модель з розподіленим лагом (autoregressive distributed lag models, ARDLM) переважно описується в книгах з економетрики. Часто при моделюванні процесів на досліджувану змінну впливають не тільки поточні значення процесу, а й його лаги, тобто значення часового ряду, попередні досліджуваному моменту часу. Модель авторегресії розподіленого лага описується рівнянням:

$$Z(t) = \varphi_0 + \varphi_1 Z(t-l-1) + \varphi_2 Z(t-q-2) + \cdots + \varphi_p Z(t-l-p) + \varepsilon^t,$$

де $\varphi_0 \ldots \varphi_p$ - коефіцієнти моделі,

l - величина лага.

Дана модель позначається як ARDLM (p, l) і найчастіше застосовується для моделювання економічних процесів.

Модель експоненціального згладжування (exponential smoothing, ES) найчастіше застосовується для моделювання фінансових і економічних процесів. В основу експоненціального згладжування закладена ідея постійного перегляду прогнозних значень в міру надходження фактичних. Модель ES привласнює експоненціально убутні ваги спостереженнями в міру їх старіння. Таким чином, останні доступні спостереження мають більший вплив на прогнозне значення, ніж старші спостереження. Функція моделі ES має вигляд:

$$Z(t) = S(t) + \varepsilon_t,$$

$$S(t) = a * Z(t-1) + (1-a) * S(t-1) + \varepsilon_t,$$

де α - коефіцієнт згладжування $0 < \alpha < 1$,

початкові умови визначаються як $Z(0) = S(1)$.

У даній моделі кожне наступне згладжені значення S (t) є зваженим середнім між попереднім значенням тимчасового ряду Z (t) і попереднього згладженого значення S (t-1).

Головною метою моделювання часових рядів є ретельне зіставлення та ретельне вивчення минулих спостережень часових рядів з метою розробки відповідної моделі, що описує невід'ємну структуру серії. Потім за допомогою цієї моделі створюється прогноз. Іншими словами, прогнозування часових рядів можна виразити як здатність прогнозувати майбутні значення шляхом розуміння минулого. Передбачення еволюції деяких змінних у різних галузях, таких як фінанси, страхування і океанографія, є дуже цінним і безцінним; цей метод робить науковців і дослідників більш суворими в побудові та встановленні найбільш адекватної моделі для своїх часових рядів.

Інші моделі прогнозування.

Можна використовувати ряд інших моделей, таких як моделі на основі ланцюгів Маркова, модель на класифікаційно-регресійних деревах, модель на основі генетичного алгоритму для задач прогнозування часових рядів, але з певних

причин було прийнято рішення не розглядати їх детально. Наприклад, моделі на ланцюгах Маркова грунтуються на тому, що прогноз майбутнього значення залежить тільки від поточного значення, тобто модель буде набагато менш гнучка і з великою ймовірністю менш точна. Так само ряд подібних моделей не володіє достатньою методологічною базою, тобто має недостатньо докладний опис як моделей, так і їх можливостей.

Таблиця 1.1 - Порівняння моделей прогнозування

Модель і метод	Переваги	Недоліки
Регресивні моделі і методи	Простота, гнучкість, прозорість моделювання; однаковість аналізу і проектування.	складність визначення функціональної залежності; трудомісткість знаходження коефіцієнтів залежності; відсутність можливості моделювання нелінійних процесів (для нелінійної регресії)
Моделі і методи експоненціального згладжування	Простота моделювання; однаковість аналізу і проектування	Недостатня гнучкість; вузька застосовність моделей

Кінець таблиці 1.1

Модель і метод	Переваги	Недоліки
Нейромережеві моделі і методи	Нелінійність моделей; масштабованість, висока адаптивність; однаковість аналізу і проектування; безліч прикладів застосування	Нелінійність моделей; масштабованість, висока адаптивність; однаковість аналізу і проектування; безліч прикладів застосування

2.1 Постановка задачі

Для формування вимог, був проведений ретельній аналіз задачі, були визначенні потреби та умови , враховуючи можливо конфліктні вимоги. Був проведений збір даних про алгоритми прогнозування та вимоги для реалізації алгоритму. Була оцінена сфера електронної коммерції, виявлені її основні вимоги та проблемі, які необхідно вирішити.

Бібліотека має бути розроблена на мові програмування Python з застосуванням бібліотек. Мова розробки Python має велику кількість бібліотек для роботи з даними. Користувач — людина, яка має, мав, або, можливо, буде мати доступ в систему для здійснення операцій по прогнозуванню продажів. У системі повинна бути присутня лише одна бізнес роль – стандартний користувач системи.

Бібліотека, що проектується, має містити зручні інтерфейси для виклику алгоритмів з клієнтського коду. Бібліотека має задовольняти наступні нефункціональні умови:

– бути стійкою до помилок;

– мати зручні інтерфейси для застосування їх іншими розробниками програмних застосувань;

– бібліотека повинна бути гнучкою, для зручного підключення до проекту;

– система повинна бути відкритою для взаємодії з іншим програмним забезпеченням;

– зовнішні помилки не повинні впливати на роботу програмної системи;

– програмна система не повинна включати будь-яку інфраструктуру, використання якої не несе прямої вигоди.

2 РЕАЛІЗАЦІЯ ПРОДУКТУ

2.1 Прогнозування продажів

Прогнозування продажів - це процес оцінки майбутнього рівня продажів за певний період. Точне прогнозування продажів має вирішальне значення для розумного управління бізнесом, оскільки дозволяє планувати попит і ефективно управляти грошовими потоками та запасами. Прогнозування рекламних ресурсів - це процес прогнозування того, коли і скільки потрібно замовити. Вона використовує комбінацію таких факторів, як попередня історія продажів, тенденції продажів і попиту, а також середній час для отримання нових запасів для визначення оптимального часу для переупорядкування запасів.

Прогнозування продажів допомагає при плануванні бізнесу, бюджетуванні та налаштуванні цілей. Після того, як бізнес добре зрозуміє, як можуть виглядати ваші майбутні продажі, він може розпочати розробку стратегії поінформованої закупівлі, щоб переконатися, що пропозиція відповідає попиту клієнтів. Завдяки прогнозуванню продажів, можна визначити та виправити будь-які перегини в конвеєрі продажів заздалегідь, щоб забезпечити надійність роботи бізнесу протягом усього періоду. Коли справа доходить до управління запасами, більшість власників електронної комерції дуже добре знають, що занадто мало або занадто багато запасів може бути шкідливим для операцій. Купівля занадто великої кількості запасів може мати значний негативний вплив на рух грошових коштів та потенціал заробітку, тоді як надто малий запас може означати, що клієнти не можуть робити покупки, що призводить до зниження продажів.

Роздрібні підприємства змушені ефективно використовувати свої ресурси та приймати обґрунтовані стратегічні рішення на майбутнє для того, щоб вижити і збільшити свої доходи, особливо коли конкуренція потійно зростає. Оскільки всі прогнози передбачають принаймні деякі ступінь невизначеності, підприємствам необхідно робити оцінки з метою мінімізацією невизначень. Підприємства повинні робити прогнози, що охоплюють багато змінних, наприклад вимоги до сировинних

матеріалів, оптимальні рівні запасів, вимоги до запозичень. Однак для того, щоб будь-який з них оцінювався, необхідно спочатку прогнозувати рівень попиту, який буде на ринку, і, відповідно, перспективні продажі компанії. Таким чином, прогнози ринку попиту є необхідним попередником всі інші оцінки, необхідні для даної операції. Точні прогнози дозволяють доцільно цілеспрямованість діяльності компанії і полегшення їх досягненню цілей. Крім того, прогнозування продажів має велике значення для компаній, що приймають стратегічні рішення щодо майбутніх інвестицій. Наприклад, обсяги продажів використовуються в комбінації з прогнозами маржі для оцінки майбутніх доходів компанії та використання разом з ними прогноз обороту для оцінки майбутніх активів компанії. Прогнози відіграють вирішальну роль в управлінні бізнесом та стратегічному плануванні. Рішення щодо управління, прийняті на кожному рівні бізнесу, прямо або опосередковано пов'язані з прогнозами. Без корисних передбачень діяльність з планування та контролю не може бути виконана ефективно. Поганий прогноз негативно впливає на здатність організацій та компаній досягати своїх цілей, оскільки це призводить до таких проблем, як збільшення витрат на запас і неможливість задовольнити попит, що, в свою чергу, може призвести до втрати частки ринку.

У цьому розділі ми зосереджуємося на описі окремих моделей і комбінації прогнозів методи, що використовуються в цьому аналізі. По-перше, взаємозв'язок між моделями пространства станів та методів експоненціального згладжування. Далі, ARIMA, ARFIMA, ANN, і процеси моделювання ANFIS коротко пояснюються. Нарешті, узагальнено наявну літературу, що стосується комбінованих прогнозів і найбільш часто використовуваних методів. Для кожної розглянутої моделі існує дуже багато варіацій, запропонованих у літературі, тому немає сенсу розглдати їх всі. Тому буде розглянуто базову версію кожної моделі. Методи експоненціального згладжування застосовувалися більше п'яти десятиліть, завдяки їх простоті і тому, що вони не вимагають складних обчислень. Особливо вигідно у випадках, що часто зустрічаються, які вимагають техніки прогнозування, яка є надійною і здатною виробляти прогнози для багатьох змінних у найкоротші терміни. Експоненціальні методи згладжування полягають у визначенні окремо

форм, прийнятих двома компонентами даних, а саме - тенденціями і сезонності. Беручи до уваги різні структури тенденцій і сезонності, можна визначити п'ять компонентів тренду і три сезонні компоненти, що дають п'ятнадцять комбінацій, кожен з яких представляє інший метод експоненціального згладжування. Деякі з цих методів є дуже популярними і загальновідомі своїми спеціальними назвами, такими як просте експоненціальне згладжування; Лінійний метод Хольта; адитивний метод Холт-Вінтерса; і мультиплікативний метод Холт-Вінтерса. Незважаючи на те, що вони використовувалися протягом багатьох років, експоненціальні методи згладжування не були поміщені в статистичні рамки до недавнього появи нових методологічних розробок. Кожен метод експоненціального згладжування відповідає двом моделям просторового стану: один з адитивними і один з мультиплікативними. Хоча точкові прогнози, отримані моделями, однакові, вони призводять до різних інтервалів прогнозування. З метою диференціації моделей з адитивними та мультиплікативними помилками третій компонент - на додаток до тенденції та сезонності - включений до аналізу. Додавання цієї складової помилки призводить до того, що кількість досліджуваних моделей простору станів дорівнює 30. Модель складається з адитивних помилок, адитивних тенденцій і відсутності сезонності. Моделі простору станів формуються двома групами рівнянь. До першої групи належать рівняння вимірювання (або спостереження), які визначають спостережувані дані. До другої групи належать рівняння стану (або переходу), які визначають, як змінюються неспостережувані компоненти або стани (рівень, тенденція, сезонні). Існують певні обмеження на значення, які можуть мати параметри згладжування: деякі є традиційними обмеженнями, призначеними для отримання зважених середніх величин з рівнянь, інші - для усунення чисельних нестабільностей при оцінці моделі.

Методи прогнозування попиту на статистику / математику, що застосовуються на практиці, можна розділити на три частини: часові ряди (також звані методи екстраполяції), причинний і зважений комбінований прогноз. Методи з часовими рядами - це ті, які визначають закономірності (тенденції, сезонність, циклічність і випадковість) і передбачають їх у майбутньому. Ці методики мають

більш високу прогностичну точність у стабільних ринках. Приклади включають: ковзну середню; просте експоненціальне згладжування; Два параметри Хольта; і три параметра Winters. Часові ряди легко розробляються і вимагають мінімального обсягу даних, але не можуть передбачити раптових змін у попиті. Вона також може передбачати лише один-три періоди вперед з будь-яким ступенем точності. ARIMA - це більш просунута технологія часових рядів, яка поєднує в собі часові ряди і елементи регресії. Цей метод є більш точним для прогнозування попиту в довгостроковій перспективі. Вона може моделювати тенденцію / цикл, сезонність, а також інші фактори, що впливають на попит (пояснювальні змінні), але вимагає більшої кількості даних і є більш складним для розвитку. Причинно-наслідкові прийоми припускають, що майбутні продажі пов'язані зі змінами в інших змінних (ціна, акції, серед інших). Прикладами є регресія і ARIMAX (розширення ARIMA). Ці методи вимагають більшої кількості даних і є більш складними для їх розробки. З іншого боку, вони можуть включати інтервенційні змінні (з використанням фіктивних змінних). Зважене комбіноване прогнозування об'єднує методи (тобто часові ряди, причинний і / або судження) і створює єдиний прогноз. Це можна зробити, задавши кожному рівному або різному вазі. Ця комбінація перевершує більшість окремих прогнозів, оскільки упередження серед методів компенсують один одного. Навіть вважаючи, що точність прогнозування є важливою, найчастіше все ще практикують простіші методи прогнозування. Однією з причин є те, що моделі, які є результатом маркетингових досліджень, важко реалізувати на практиці. Це є суттєвою проблемою структурних моделей, оскільки «велика кількість спостережень на практиці, великий масив стану та контрольних змінних, а також частота прийняття рішень може зробити застосування структурних моделей неможливим».

2.2 Опис досліджуємих даних

Набір даних для аналізу складається з 500000 замовлень інтернет магазину з Великобританії за період року. Кожен запис складається з номеру замовлення, номери товарів, опис товарів, кількість товарів, дата замовлення, ціна товарів, номер замовника, країна замовника.

Набори даних часових рядів зазвичай дуже великі. Висока розмірність, висока кореляція ознак і велика кількість шумів, які можуть бути присутніми в часових рядах, представляють виклик завданням інтелектуального аналізу даних. Висока розмірність таких часових рядів збільшує час доступу до даних та час обчислення, необхідний алгоритмам інтелектуального аналізу даних. Крім того, методи візуалізації повинні застосовувати методи скорочення та агрегації даних, щоб впоратися з великим обсягом даних, які не можуть бути деталізовані одночасно. Вищезгадані причини роблять застосування методів обробки та прогнозування безпосередньо на необроблених даних часових рядів громіздкими. Щоб подолати цю проблему, оригінальні «необроблені» дані повинні бути замінені більш високим рівнем представлення, що дозволяє ефективно обчислювати дані, і витягує функції вищого порядку.

Для обробки такого об’єму даних необхідно провести попередню обробку даних, виділити дані які нас існують, повністю її перебати, группувати за атрібутами та зберегти у зручному для подальшої оборки виду. Для цього будуть використяні вбудовані модулі та алгоритми Python. Дані збережені у форматі xls. Для того щоб почати обробку даних – були проведені наступні операції: вибрати усі необхідні атрибути, створити агреговану змінну Сума, помноживши кількість на ціну, яка дає загальну суму грошей, витрачених на продукт / елемент у кожній операції, розділити змінну InvoiceDate на дві змінні Date і Time - це дозволяє різні операції, створені тим самим споживачем в один день, але в різний час, розглядатись окремо, фільтрувати будь-які транзакції, з якими не пов'язана країна.

У таблиці 2.1 наведено опис цих функцій у наборі даних.

Таблиця 2.1 – Опис даних

Назва поля	Опис
InvoiceNo	Номер транзакції. Номінальний, 6-значний інтегральний номер, однозначно призначений для кожної транзакції.
StockCode	Код продукту (елемента). Номінальний, 5-значний інтегральний номер, однозначно призначений кожному окремому продукту.
Description	Назва продукту (елемента). Номінальний. Кількість: Кількість кожного продукту (елемента) за кожну операцію. Числові.
InvoiceDate	Дата і час замовлення. Числові, день і час, коли було створено кожну транзакцію.
UnitPrice	Ціна за одиницю. Числова, Ціна продукту за одиницю в фунтах стерлінгів.
CustomerID	Номер клієнта. Номінальний, 5-значний інтегральний номер, унікально призначений кожному клієнту.
Country	Назва країни. Номінальна, назва країни, в якій проживає кожен клієнт.

У таблиці 2.2 наведено приклад набора даних.

Таблиця 2.2 – Приклад даних

InvoiceNo	StockCode	Description	Quantity	UnitPrice	Country
536375	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	3,39	United Kingdom
536375	22752	SET 7 BABUSHKA NESTING BOXES	2	7,65	United Kingdom

Продовження таблиці 2.2

InvoiceNo	StockCode	Description	Quantity	UnitPrice	Country
536375	21730	GLASS STAR FROSTED T-LIGHT HOLDER	6	4,25	United Kingdom
536376	22114	HOT WATER BOTTLE TEA AND SYMPATHY	48	3,45	United Kingdom
536376	21733	RED HANGING HEART T-LIGHT HOLDER	64	2,55	United Kingdom
536377	22632	HAND WARMER RED POLKA DOT	6	1,85	United Kingdom
536377	22633	HAND WARMER UNION JACK	6	1,85	United Kingdom
536378	22386	JUMBO BAG PINK POLKADOT	10	1,95	United Kingdom
536378	85099C	JUMBO BAG BAROQUE BLACK WHITE	10	1,95	United Kingdom

У контексті штучного інтелекту та науки про дані, бази даних відіграють центральну роль як у навчанні, так і в оцінці. Це відбувається незалежно від того, чи використовується набір даних для побудови предиктора, який буде розгортатися як частина системи, або використовується для задавання наукових питань і досягнення наукових висновків. В обох випадках специфічні властивості набору даних можуть мати глибокий вплив на якість вивченого предиктора або якість наукових висновків.

Для зручності обробки та для необхідно визначити до яких груп товарів належать товари. У даних які використовуються є лише один спосіб – за назвою товару. Так як не існує баз для навчання для конкретного магазину, необхідно створити універсальний класифікатор та спробувати класифікувати дані. Бажано, щоб конкретний продукт був категоризований точно так само в різних магазинах. Існують універсальні таксономії, такі як таксономія Google, які можна використовувати для відображення продуктів на універсальні категорії в роздрібній мережі. Існує два підходи:

– Ручне відображення - це тягне за собою перегляд назв продуктів і призначення категорій з універсальної таксономії вручну. Типовий агрегатор електронної комерції має близько 50–75 мільйонів списків продуктів. Це виключає можливість робити це вручну, оскільки це займе дуже величезну кількість часу і схильні до людських помилок.

– Категоризація на основі правил - Інший варіант полягає в написанні правил категоризації на основі категорій роздрібної торгівлі та назв продуктів. Але скільки таких правил потрібно зробити для вичерпного набору категорі. Занадто багато. І навіть при дуже великій кількості правил, важко сказати, чи є вона вичерпною чи ні. Таким чином, цей параметр не є масштабованим і приносить обмежений успіх.

Підхід для машинного навчання використовує методи класифікації тексту для класифікації кожного продукту до категорії на основі назви / опису продукту. Перш ніж перейти до деталей методу, треба подивитись на проблеми, що виникають у впровадженні підходу ML. Наявність навчальних даних - алгоритми ML потребують навчальних даних для вивчення правильних відповідей, перш ніж вони почнуть прогнозувати. У більшості програм ML ці дані надходять з історичних бізнес-процесів. Але в цьому випадку вона недоступна. У таких випадках дуже часто створюються навчальні дані вручну або за допомогою певного евристичного методу. Для проблеми класифікації текстів повинна існувати суттєва репрезентація кожної категорії, яку намагаються передбачити. Ручний підхід є трудомістким і схильним до помилок, як обговорювалося раніше. Ось як виглядають дані навчання в нашому випадку.

Кроки для категоризації продуктів можна класифікувати як попередню обробку даних, оцінку моделі, валідацію та оптимізацію моделі тощо.

Крок 1: Випадкова вибірка. Кількість унікальних продуктів у різних категоріях продуктів різко змінюється. Як правило, розподіл продукції за категоріями відповідає принципу парето, що передбачає, що 80% продукції міститься в 20% категорій. Метод грубої сили для боротьби з цією асиметрією

полягає у виключенні категорій продуктів з дуже малим представництвом з аналізу. Але це призводить до втрати даних, що не рекомендується. Навпаки, випадкова вибірка, Stratified Random Sampling, щоб бути точним, використовується для забезпечення справедливого представлення кожної категорії.

Стратифікована випадкова вибірка також може бути використана для швидкого прототипу моделі, що бере зразок, що містить усі категорії з великого набору даних.

Крок 2: Попередня обробка даних. Попередня обробка для текстових даних сильно відрізняється від звичайного набору даних. Текстові дані спочатку повинні бути перетворені в числове представлення, перш ніж до неї будуть застосовані алгоритми ML. Методи, які досягають цієї мети, називаються методами векторизації тексту. Найпопулярнішими є «Мішок слів», «TF-IDF» і «Word2Vec». У більшості випадків вони призводять до того, що називається матрицею Term-Document (TDM). Методи векторизації тексту поєднуються з етапами попередньої обробки, як n-грами, видалення стоп-слів, витіснення тощо. Видалення стоп-слів виключає знаки пунктуації та інші несуттєві слова, такі як предмети та сполучники (про, серед, так чи інакше) з TDM. Стовбування і лемметізація підрізають слово до його кореня. Множинне число стає сингулярним, різні напружені варіанти зводяться до їх простої нинішньої форми.

Крок 3: Навчання моделі класифікації. Після перетворення текстових даних на числове представлення, він готовий застосовувати класифікаційні моделі. Найбільш популярним алгоритмом, який використовуються для класифікації тексту, є наївний баєсів класифікатор. Інші класифікаційні методи, такі як випадкові ліси з методами підсилення, також можуть бути випробувані.

Наївний баєсів класифікатор є методом класифікації на основі ймовірності. Він обчислює умовну ймовірність того, що блок слів (назва продукту) належить до певного класу (категорія продукту). Назва продукту присвоюється категорії продуктів, для яких умовна ймовірність найвища. Він має кілька параметрів моделі, які можна налаштувати. Продуктивність та точність моделей багато в чому залежить від правильних значень параметра. На цьому етапі для кожного параметра

надається діапазон значень, і створюються кілька моделей, щоб побачити, які значення параметра дають найкращий результат. Цей метод необхідний для точного налаштування та пошуку оптимальних значень параметрів моделі.

2.3 Прогнозування рядів

Моделювання часових рядів - це динамічна дослідницька область, яка за останні десятиліття привернула увагу науковців. Основною метою моделювання часових рядів є ретельне збирання та ретельне вивчення минулих спостережень часових рядів для розробки відповідної моделі, що описує невід'ємну структуру серії. Дана модель потім використовується для створення майбутніх значень для серії, тобто для прогнозування. Таким чином, прогнозування часових рядів можна назвати актом прогнозування майбутнього шляхом розуміння минулого. У зв'язку з необхідністю прогнозування часових рядів у багатьох практичних галузях, таких як бізнес, економіка, фінанси, наука та інженерія тощо, необхідно належним чином пристосувати відповідну модель до основних часових рядів. Очевидно, що успішне прогнозування часових рядів залежить від відповідної моделі моделі. Багато років дослідники робили багато зусиль для розробки ефективних моделей для підвищення точності прогнозування. Як результат, у літературі розвивалися різні важливі моделі прогнозування часових рядів. Хоча існує багато підходів у існуючій літературі, які намагалися розглядати замовлення інтернет магазинів з конкретними моделями прогнозування, такими як ковзаюча середня, експоненціальне згладжування, ланцюг Маркова та інші, результатів було недостатньо або результати мали недостатню точність. Знаючи, що ARIMA є одним з найпотужніших підходів до прогнозування, в основному використовується для фінансових часових рядів, ця стаття має на меті побачити, чи можуть дані про інтренет магазин адаптуватися до моделей ARIMA для оцінки та прогнозування.

Процедура моделювання ARIMA була однією з лінійних моделей, які найчастіше використовуються у прогнозуванні часових рядів. Моделі ARIMA базуються на визначенні структури автокореляцій у даних. Передбачається, що прогнозована змінна складається з лінійної комбінації власних залагованих змінних та помилок. Це припущення також вказує на слабкість моделей ARIMA: вони не можуть охоплювати нелінійні структури. Проте вони досягли значних успіхів у багатьох програмах прогнозування. Модель ARIMA складається з вдосконаленого теоретичного підходу, який включає ітераційний триетапний процес побудови моделі: ідентифікація моделі, оцінка параметрів і діагностична перевірка.

Було вибрано для використання моделі ARIMA замість інших машинознавчих з наступних причин: вони забезпечують хороші показники прогнозування, коли використовується невелика / велика кількість даних (тобто, часові ряди), з точки зору короткострокового прогнозування, ці моделі є відносно більш надійними та ефективними, ніж більш складні. Суто статистичний метод, який використовується ARIMA, вимагає лише попередніх даних часового ряду, щоб узагальнити прогноз, після чого він може підвищити точність прогнозування за допомогою мінімального числа параметрів.

Кілька дослідників використовували моделі ARIMA для прогнозування цін на акції. Як приклад, автори [4] описують, як ARIMA може легко обробляти такий формат даних і як він добре підходить для прогнозування часових рядів. Ще один приклад представлений [21], де автори описали модель ARIMA з валютою як функцію, що використовується для прогнозування цін на сировинні товари. ARIMA також використовувалася в галузі сільського господарства, де автори [6] обговорювали, як вони використовували його для прогнозування цін на сільськогосподарські культури. Наприклад, робота в [5] описує модель ARIMA для прогнозування ціни електроенергії протягом тижнів, що має вхідне вейвлет-перетворення, застосоване до часових рядів, пов'язаних з цінами на електроенергію.

Слід зазначити, що, незважаючи на прийняту модель, коли екзогенні змінні не використовуються, існує межа передбачуваності, тоді як відповідний вибір

таких змінних призводить до поліпшення показників прогнозування. Більш того, автори зосереджувалися на прогнозуванні ціни на житлово-комунальні послуги та електронні продукти, використовуючи новини про настрої від Baidu, відомого портального веб-сайту в Китаї, аналізуючи зміст новин і вилучаючи ключові події, пов'язані з продуктами. Результати показують, що на прогнозовані ціни на продукти можуть впливати фактори настрою.

Часовий ряд - це послідовність вимірювань однієї і тієї ж змінної, виконаних у часі. Зазвичай вимірювання проводяться в рівномірно розподілені часи - наприклад, щодня або щомісяця. Модель ARIMA (p, d, q), в аналізі часових рядів, є процесом, що відповідає даним часових рядів, і передбачає майбутні точки в серії. ARIMA добре підходить для запропонованого випадку використання, оскільки дозволяє працювати не тільки з кортежами, але і з зовнішньою інформацією, включеною в модель як екзогенні особливості. Наприклад, ми бачимо, чи додає нова екзогенна функція покращує загальну точність чи ні. Це може бути корисним для розуміння того, як модель реагує на різні зовнішні особливості і наскільки вони впливають на точність прогнозу. Оскільки в нашому дослідженні ми з'ясували, що зовнішня функція, наприклад, більш докладну інформацію про ARIMA можна знайти в літературі. Модель являє собою комбінацію з трьох частин AR, I, MA. На ці частини впливають параметри (p, d, q). Нижче пояснення кожного:

Авторегресивна модель (AR): модель авторегресії AR вказує, що вихідна змінна залежить лінійно від власних попередніх значень і від стохастичного терміна. Порядок авторегресійної моделі - це кількість безпосередньо попередніх значень у рядах, які використовуються для прогнозування значення в даний час. Позначення AR (p) вказує на авторегресивну модель порядку p. Модель AR (p) визначається як:

$$X_t = c + \sum_{i=1}^{p} \varphi_i X_{t-1} + \varepsilon_t,$$

де $\varphi_1, \ldots, \varphi_p$ - параметри моделі,

c - постійна, εt - білий шум.

Інтегрований (I): Він вказує на ступінь диференціювання порядку d серії. Різниця в статистиці - це трансформація, що застосовується до даних часових рядів, щоб зробити її стаціонарною. Оформлений в [16] процес {Yt} є стаціонарним, якщо ймовірнісні розподіли випадкових векторів (Yt1, Yt2, ..., Ytn) та (Yt1 + l, Yt2 + l, ..., Ytn + l) однакові довільні моменти часу t1, t2,. . . , tn, всі n, а всі лаги або відведення l = 0, ± 1, ± 2,. . .. Стаціонарні часові ряди не залежать від часу, коли спостерігається серія. Для диференціації даних обчислюється різниця між послідовними спостереженнями. Математично це показано як:

$$y_t^{`} = y_t - y_{t-1},$$

де y_t, y_{t-1} є значеннями серії, відповідно, в момент часу t, t-1,

$y_t^{`}$ - диференційоване значення ряду в момент часу t.

Слід зазначити, що друга різниця Підхід ARIMA вимагає, щоб дані були стаціонарними. Коли часові ряди вже є стаціонарними, оцінюється модель ARMA. Навпаки, якщо часовий ряд не є стаціонарним, то серія повинна бути перетворена, щоб стати стаціонарною (порядок інтегрування "I" означає кількість разів, що серія повинна бути диференційована для отримання стаціонарності), і ми працюємо з ARIMA моделі. Якщо часовий ряд є стаціонарним, то приймати перші відмінності не потрібно. Ряд інтегрується з нульового порядку (d = 0), і ми задаємо модель ARMA, оскільки диференціювання не усуває існуючу в даних сезонну структуру. Для прогнозування часових рядів у моделях ARIMA звичайно використовується наступне перетворення даних:

$$\Delta y_t = \log(Y_t) - \log(Y_{t-1})$$

Це перетворення дозволяє зменшити мінливість ряду, а трансформована змінна може бути інтерпретується як наближення швидкості росту.

Модель ковзного середнього (MA): В аналізі часових рядів, модель ковзного середнього MA є загальним підходом для моделювання одновимірних часових рядів. Модель змінного середнього визначає, що вихідна змінна залежить лінійно від поточних та різних минулих значень стохастичного терміна. Позначення MA

(q) належить до моделі ковзних середніх порядку q. Модель MA (q) визначається як:

$$X_t = \mu + \varepsilon_t + \theta_1\varepsilon_{t-1} + \cdots + \theta_q\varepsilon_{t-q},$$

де μ - середнє значення ряду,

$\theta_1,\ldots,\theta_q$ - параметри моделі,

εt - білий шум.

Далі буде показано, як розраховані основні параметри ARIMA (p, d, q). Для обчислення параметрів ARIMA необхідний крок попередньої обробки і оцінюється, яка комбінація призводить до найкращих результатів. Потім буде показано алгоритм, який використано для використання ARIMA з обраними даними. Як зазвичай виконується для використання ARIMA-моделей для задач прогнозування, встановлено інновації рівними нулю. Буде зроблено прогноз, використовуючи підхід, що складається з декількох кроків, з прямою (також називається незалежною) стратегією, яка складається з прогнозування кожного горизонту незалежно від інших.

Далі буде введено стадію попередньої обробки. Введення такого кроку - це правильно відформатовані часові ряди. Я використав часові ряди, що складаються з (ціна, дата) кортежів з послідовними щоденними записами. Використано ці кортежі, щоб з'ясувати, чи є певний часовий ряд стаціонарним (d), розміром кластерів корельованих записів (q) та частково корельованими записами (p). Коротше кажучи, якщо часовий ряд є стаціонарним, то приймати перші відмінності не потрібно. Ряд інтегрується з нульового порядку, отримуючи d = 0, і задаємо модель ARMA. І навпаки, якщо ряд не є стаціонарним, необхідний подальший розрахунок. Коротше кажучи, параметри p та q говорять про те, наскільки великими є кластери записів, які мають подібну поведінку у часі. Оцінку параметрів можна здійснити за методом Бокса та Дженкінса, щоб знайти найкраще відповідність часового ряду з використанням ARIMA. Він складається з етапу ідентифікації моделі, етапу оцінки параметрів і етапу оцінки моделі.

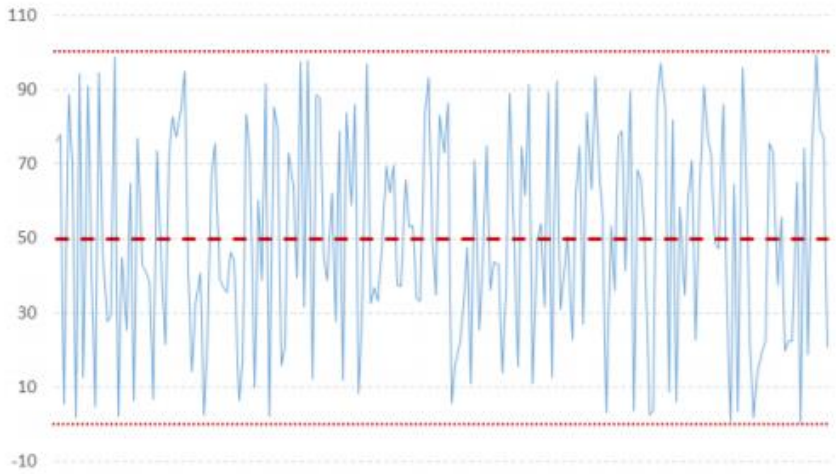

Рисунок 2.1 – Приклад стаціонарного ряду

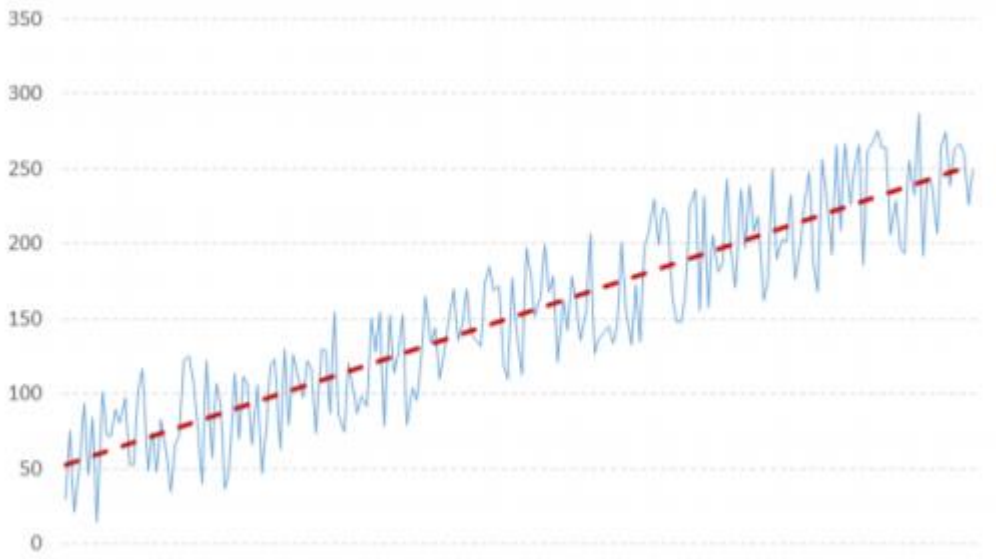

Рисунок 2.2 – Приклад не стаціонарного ряду

Підхід Боксу-Дженкінса дозволяє побудувати модель ARIMA (p,d,q) для нестаціонарного часового ряду. В даному підході виділяються такі етапи:

– Ідентифікація моделі ARIMA (p,d,q), тобто визначення величин p, d, q. Причому спочатку визначається параметр d, а потім вже p і q.

– Встановити порядок інтеграції ряду d. Тобто знайти мінімальну кількість послідовних різниць, які необхідно взяти, щоб підсумковий ряд став стаціонарним. Це мінімальна кількість і дорівнює параметру d.

– Побудова моделі ARMA(p,q) для підсумкового стаціонарного ряду. Аналізуючи автокорреляционную і приватну автокорреляционную функції, підібрати параметри p і q. Кількість коефіцієнтів на автокоррелограммі сильно відмінних від 0 дорівнює q в моделі MA. Максимальний номер коефіцієнта на приватній автокоррелограммі сильно відмінний від 0 дорівнює p в моделі AR.

У межах області часових рядів дві важливі функції стосуються лагів. По-перше, функція часткової автокореляції (PACF) - це кореляція між тимчасовими рядами та власними відсталими значеннями. Інша функція автокореляції (ACF) визначає, як точки співвідносяться один з одним, виходячи з того, скільки кроків часу вони розділені. Ці дві функції необхідні для аналізу часових рядів, оскільки вони здатні визначити ступінь відставання в авторегресійній моделі (PACF) і в моделі ковзної середньої (ACF). На рисунку 5 і 6 показаний приклад ACF і PACF. Використання цих функцій було введено в рамках підходу Box – Jenkins до моделювання часових рядів, де обчислення часткової автокореляційної функції може бути визначено відповідними лагами p в моделі ARIMA (p, d, q) і обчислення автокореляції Функція може бути визначена відповідним відставанням q в моделі ARIMA (p, d, q). Оскільки (часткова) автокореляція MA (q) (AR (p)) процесу дорівнює нулю при затримці q + 1 (p + 1) і вище, ми враховуємо вибірку (часткову) автокореляційну функцію, щоб перевірити, де вона стає нуль. Це досягається шляхом розгляду 95% довірчого інтервалу для зразка (часткового) автокореляційного ділянки. Довірчий діапазон становить приблизно $\pm 2 / \sqrt{N}$, де N відповідає розміру вибірки.

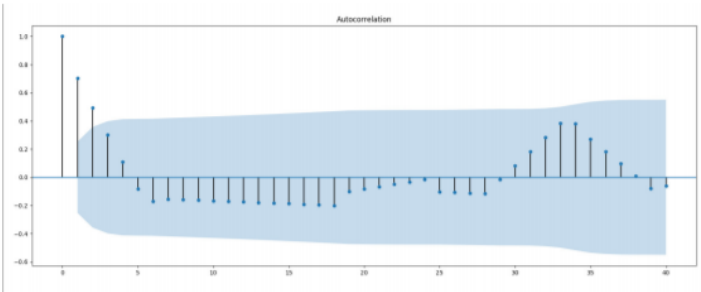

Рисунок 2.3 – Приклад ACF

Щоб визначити найкращу модель ARIMA для звмовлень, треба виконати наступні кроки:

– перевірити стаціонарність, знайти відповідне значення d.

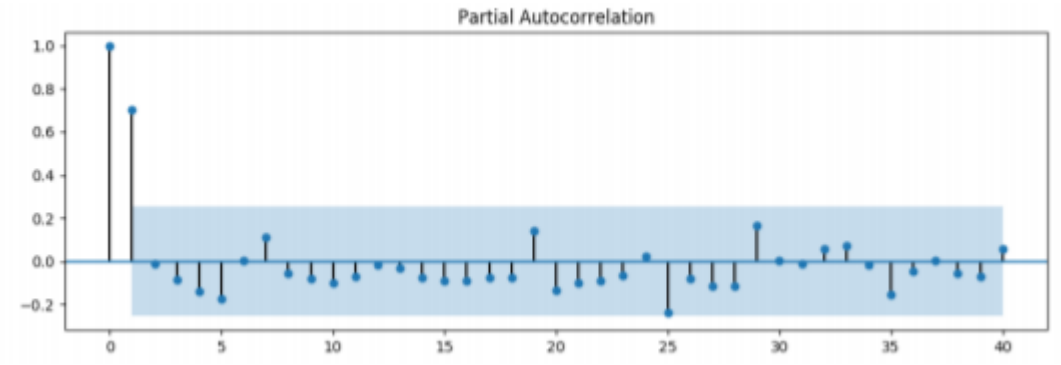

Рисунок 2.4 – Приклад PACF

– знайти p і q на основі PACF і ACF. Результат на ACF і PACF дає нам верхню межу для повторення підгонки моделі, зберігаючи кращу (p, d, q) комбінацію, засновану на найменшому значенні MSE, описуваному як:

$$MSE = \frac{1}{n} \cdot \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 ,$$

де Yˆ - вектор n прогнозів

Y - вектор спостережуваних значень змінної, що прогнозується.

2.4 Алгоритм роботи

Цей розділ описує, як застосовано ARIMA до дослідження. Як згадувалося вище, є багато різнах особливостей (як часові ряди), які можна використовувати як екзогенні змінні для нашої моделі ARIMA. Екзогенна змінна містила кортежі у вигляді (дата, значення). Таким чином, алгоритм, який ми виконуємо для кожного продукту нашої колекції:

– Розрахувати параметри ARIMA - (p, q, d), як зазначено в у попередньому розділі.

– Здійснити отримання даних

– Fit Model - побудувати модель для кожної комбінації доступних основних та зовнішніх функцій.

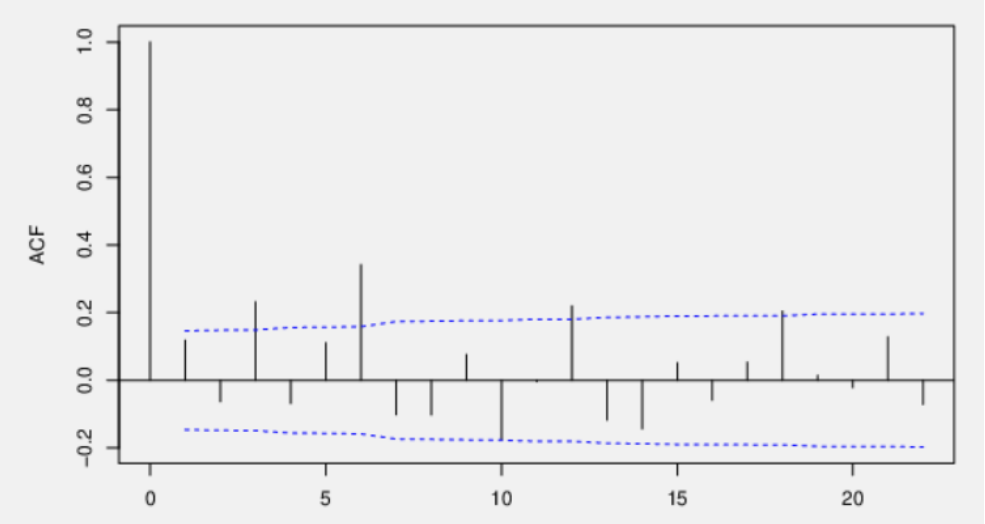

Рисунок 2.5 – Автокореляційна функція

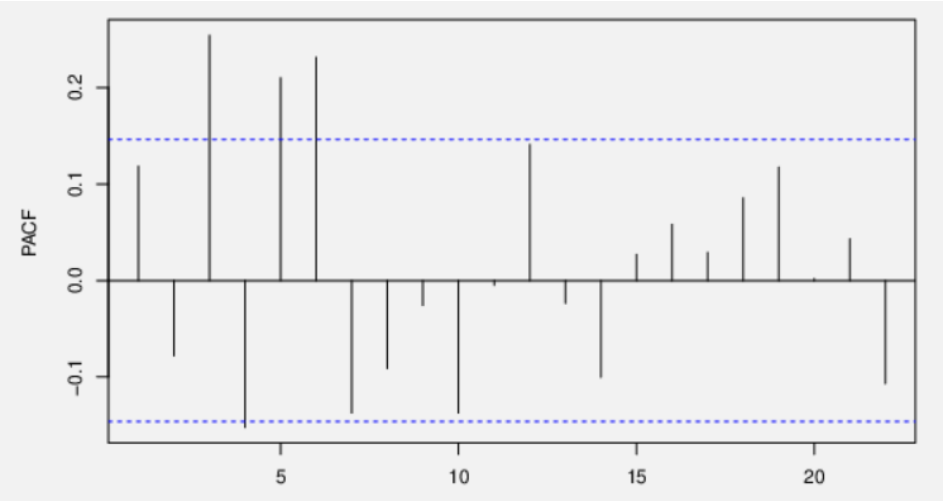

Рисунок 2.6 – Приватна автокореляційна функція

Для експериментальних цілей я обрав топ 100000 замовлень. Використав три різні розміри тестів, 10%, 20% і 30%, що стосувалися записів цін кожного замовлення. Розмір тесту відображає кількість зроблених прогнозів. Таким чином, для даного замовлення, що має часові ряди з 100 ціновими елементами, якщо у нас є тестовий набір розміром 10%, ми будемо використовувати перші 90 елементів для навчання моделі, і будемо виконувати прогноз на останніх 10 елементів. Було розраховано прогноз для двох інших розмірів тесту (20% і 30%) аналогічно. Для кожного відсотка була використана функцією перехресної перевірки часових рядів, пропонованою бібліотекою Python scikit-learn. У більшості випадків зменшення значень розміру тесту відповідає більш точним оцінкам прогнозу. Зокрема, можна спостерігати, що (ціна, дата) поєднання є найбільш перспективним параметром, що має найнижчу помилку від реального значення.

Проаналізуємо автокоррелограмму - другий графік (рис. 2.5). Значення автокореляційної функції близькі до 0, що вказує на те, що наш ряд є стаціонарним. За методом Бокса-Дженкінса побудуємо модель ARIMA(, 1,). Звернемося до другого і третього графіком, тобто до графіків автокорреляционной і приватної кореляційної функцій. Видно, при лагу 3, 6, 12 і 18 є значуще відхилення від 0. Отже, q=4. Проаналізувавши частково кореляційну функцію, можна сказати, що p=18. Виходить, що даний ряд представимо моделлю ARIMA (18,1,4). Тепер необхідно перевірити чи відповідає обрана модель нашими даними.

Реалізація, що використовується в даній роботі, полягає в ARIMA з p = 18, d = 1 і q = 4, оскільки вона була здатна виробляти хороші результати для різних періодів. Важливо відзначити, що було б краще підібрати різні моделі ARIMA для різних часових рядів. Якщо взяти реалізовану модель ARIMA - прогнози показують перспективні результати. Незважаючи на те, що прогноз дещо відхилений, рівень продажів залишається більш постійним з часом, ніж прогнозується фактичний рівень. Однак, є зауваження, що модель передбачає продаж конкретної категорії, але не враховує різні варіації продуктів, такі як кольори та розміри, та інщі атрибути. Вона також не прогнозує розподіл на країни магазинів. Але, незважаючи

на точність, ця методика прогнозування може допомогти продавцю підтримувати запаси на рівні в залежності від часу.

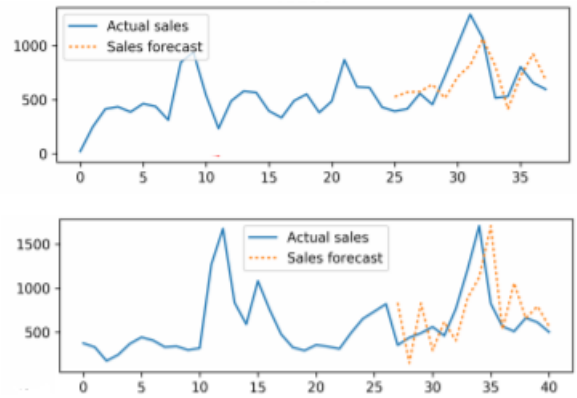

Рисунок 2.5 – Приклади часового ряду з прогнозом продажів для однієї з категорій

2.5 Використання моделі для прогнозування

Метою даного дослідження було надати компаніям по всьому світу модель або підхід до вирішення проблем часових рядів. Виходячи з кроків, зроблених у прикладі, була створена модель. Запропонована модель представлена на малюнку 6 описує найпоширеніші підходи, які використовують фахівці з аналізу даних для вирішення проблем. Метою моделі є те, щоб роздрібні торговці отримали розуміння бізнесу, яке допоможе у прийнятті важливих бізнес-рішень. Щоб ефективно обробляти дані, роздрібні торговці повинні створити сховище даних, що містить всі важливі факти.

Першим кроком для кожного аналізу є візуалізація даних. Візуалізація даних визначається, як передача інформації за допомогою графічних зображень. Перевага використання візуалізації даних полягає в тому, що графічне представлення може містити велику кількість інформації, яка може оброблятися набагато швидше, ніж сторінка слів. Які дані і відносини роздрібні торговці повинні візуалізувати, ґрунтуються на діловому домені та бізнес-питаннях, на які хоче

відповісти роздрібний продавець. Зрозуміння цього кроку вже допомагає роздрібному підприємцеві зрозуміти бізнес.

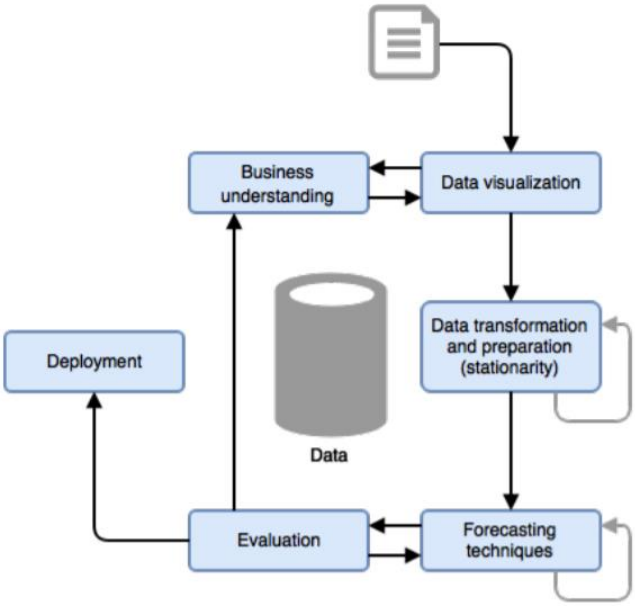

Рисунок 2.6 - Модель для отримання бізнес-знань з використанням найбільш точних методів прогнозування.

Наступним кроком є перетворення даних. Фаза перетворення є першим і відносно легким кроком в отриманні стаціонарних рядів. Наявність стаціонарних рядів має вирішальне значення для багатьох моделей прогнозування часу, і часто просте перетворення може призвести до більш високої достовірності. Якщо перетворення часового ряду не призводить до високої стаціонарної достовірності, можна використовувати інші методики. Деякі з методів, які можуть бути використані для підвищення стаціонарності: ковзне середнє значення (MA) або експоненціально зважена ковзна середня (EWMA).

Іншим кроком у моделі є фаза прогнозування. Тут може бути реалізована і протестована одна або кілька методів прогнозування, щоб побачити, який з них найкраще підходить для даної проблеми часових рядів щодо статті. Для порівняння різних методів прогнозування можна використати середньоквадратичну помилку (RMSE). Важливо відзначити, що необхідно перетворити часові ряди на початкове

представлення, щоб побачити реальні прогнози. Це залежить від вибору, зробленого на етапі перетворення та підготовки даних.

Незважаючи на те, що попередній крок автоматично вибирає оптимальну техніку прогнозування. Важливо оцінити вжиті заходи та точність прогнозів, тому що навіть незважаючи на оптимальні кроки, результати можуть бути незадовільними. Точність може бути перевірена шляхом об'єднання та порівняння прогнозу продажів з фактичними продажами, щоб побачити вплив. Перш ніж приступити до розгортання моделі, торговці повинні ретельно оцінити модель і переглянути кроки, виконані для побудови моделі. Наприкінці цієї фази продавець повинен бути впевнений, що з використанням цих методів досягаються цілі бізнесу.

Коли відомо, які методи підготовки даних і прогнозування працюють добре для даної статті, можна розгорнути програми та ПЗ для підтримки бізнес-рішень. Залежно від вимог, фаза розгортання може генерувати звіт, але й більш складний, наприклад, створення алгоритму, який роздрібний продавець може використовувати для прогнозування.

Виходячи зазначеного вище майбутня робота необхідна для:

– Перевірити модель з використанням різних наборів даних;
– Удосконалити існуючу техніку прогнозування продажів;
– Створити алгоритм, що дозволяє робити прогнози на основі моделі.

З альтернативних версій прогнозування можуть також використовуватись штучні нейронні мережі. Штучна нейронна мережа користується тим, як працює людський мозок і є, по суті, дуже спрощеним його варіантом, імітуючи можливості його паралельної обробки за допомогою математичних штучних нейронів, розподілених паралельно. Хоча примітивні версії нейронних мереж вперше з'явилися на початку 1940-х років, широко поширили своє використання майже у всіх галузях наукових досліджень і стали серйозним конкурентом класичним статистичним методам. На відміну від традиційних моделей часових рядів, нейронні мережі є підходом, що не містить моделей і керується даними, що передбачає дві важливі особливості: гнучкість і універсальність. Компроміс для

цих привабливих особливостей - це відсутність систематичної методології, за допомогою якої можна подолати невизначеності щодо побудови моделі та встановлення параметрів.

Велика кількість досліджень показала, що комбіновані прогнози показують кращу продуктивність, ніж обидві складові моделі та окремі моделі. Об'єднання різних моделей або методів якомога більше може прокласти шлях до досягнення більш точних прогнозів. Існує багато досліджень, які пропагують різні аспекти об'єднання даних часових рядів. Труднощі, з якими стикаються при комбінуванні методів, полягає у визначенні того, як, зокрема, окремі моделі повинні бути об'єднані таким чином, щоб привести до більш точних прогнозів. Було запропоновано багато методів вирішення цієї проблеми. Існує ряд складних статистичних підходів, але було зауважено, що прості методи мають тенденцію давати прогнози краще або майже настільки точні, як складні. Припустили, що прості комбінуючі методи будуть розумним підходом, коли дана проблема передбачає значну непевність. Деякі привабливі риси простих комбінованих методів полягають у тому, що їх легко зрозуміти і реалізувати, менш схильні до помилок і не залежать від будь-яких припущень: ці особливості зробили їх часто застосовуваним компонентом інструменту прогнозування. Методи комбінування, що враховуються в даному дослідженні, коротко описані нижче.

Просте середнє значення: всі прогнози на час t, що генеруються окремими моделями, мають однакову вагу.

Обрізане середнє: це середні прогнози окремих моделей, але виключає найвищі t% і найнижчі t% прогнозів.

Простий медіан: Цей метод формується як медіана прогнозів, породжених окремими моделями. Медіана комбінації менш схильна до впливу екстремальних значень, ніж проста середня комбінація, тому вона має тенденцію до отримання хороших результатів.

Найменші площі ваг: У цьому методі прогнози, сформовані за допомогою окремих моделей, регресуються проти фактичних значень. Після цього

коефіцієнти, отримані в результаті регресії, використовуються як ваги окремих моделей у комбінації.

2.6 Технології для прогнозування

Для створення моделі необхіодно обрати інструмент за допомогою якого є можливим створити легку та водночас функціональну програму. Для цього за допомогою декольких критеріїв необхідно обраті мову програмування та бібліотеки або фреймворки для використання. "Яка мова краще" – питання, яке зазвичай залежать від контексту. Краще обрати мову, яка має найбільшу підтримку з найближчого оточення. Matlab, Python і R успішно використовуються для навчання студентами основ математики та статистики. У сьогоднішньому середовищі даних, вивчення даних через аналітику великих даних є дуже потужним, особливо для цілей прийняття рішень і використання статистично даних у цій обстановці з багатими даними.

Нижче приведено коротке порівняння Matlab, Python і R. Matlab вважається не тільки комерційним чисельним обчислювальним середовищем, але й мовою програмування. Аналогічно Matlab має стандартну бібліотеку, але її використання включає матричну алгебру і велику мережу для обробки даних і побудови графіків. Вона також містить набори інструментів для завзятих учнів, але вони будуть коштувати додатково. R - програмне забезпечення, призначене для виконання статистичного аналізу та виведення графіки. R - вільне статистичне програмне забезпечення з відкритим кодом. R, безумовно, переросла своє походження, зараз хвалиться більш ніж двома мільйонами користувачів відповідно до веб-сайту R Community. MatLab має велику кількість відданих користувачів, серед яких багато університетів і кілька компаній, які мають бюджет на придбання ліцензії на програму. Незважаючи на те, що Matlab використовується у багатьох університетах, Matlab є простим для початківців, які тільки починають вивчати

мову програмування, оскільки пакет, при покупці, включає в себе все, що потрібно. Недоліком є його вартість ліцензії. Користувач повинен купувати кожен модуль і платити за нього. Недоліком є під час перехресного компіляції або перетворення Matlab в інший код мови дуже важко. Це вимагає глибоких знань Matlab для боротьби з усіма помилками. R - статистичний пакет, який намагається вирішити проблеми статистики. Існує безліч програм в R, які намагаються вирішити різні проблеми аналітики. Тим не менш, MatLab використовується для викладання різних аспектів математики, таких як числення або графічні рівняння. У полі аналітики R є кращим, ніж MatLab, коли йдеться про виконання статистичного аналізу. R є найповнішим доступним пакетом статистичного аналізу, включає всі стандартні статистичні тести, моделі та аналізи, а також надає всебічну мову для управління та обробки даних. Нові технології та ідеї часто з'являються першими в R. R має великий поріг входження - це займе деякий час, щоб звикнути до R, але не більше, ніж для інших статистичних мов. R не такий простий у використанні для новачків. Існує кілька простих у використанні графічних інтерфейсів користувача для R, які охоплюють декотрі сфери, але вони не покривають усього що потрібно.

Мова Python має різноманітне застосування в компаніях, що займаються розробкою програмного забезпечення, наприклад, в іграх, веб-фреймворках і додатках, розробці мов, прототипуванні, додатках графічного дизайну тощо. Python: легкий в навчанні, підтримується крос-платформа, велика спільнота, дуже потужний, має відкритий доступ, містить тисячі сторонніх модулів. Це надає мові більш високу чисельність над іншими мовами програмування, що використовуються в промисловості. Більшість найпоширеніших завдань програмування вже написано на скрипт, що обмежує довжину кодів, які будуть написані на Python. Завдяки потужним можливостям інтеграції процесів, система тестування модулів та покращені можливості керування сприяють збільшенню швидкості для більшості програм та продуктивності програм. Це чудовий варіант для створення масштабованих мультипотокових мережевих додатків. Але Python виконується за допомогою інтерпретатора замість компілятора, що призводить до його уповільнення, оскільки компіляція і виконання допомагають йому нормально

працювати. У порівнянні з популярними технологіями, такими як JDBC і ODBC, рівень доступу до бази даних Python виявляється трохи недорозвиненим і примітивним.

Я перегляну деякі з причин, чому хтось може захотіти перейти з інших мов на Python. Причини поділяються на такі категорії: фінансові, свобода, технічні, соціальні. Фінансова: вартість часто є першою причиною відключення від MATLAB. Це, звичайно, поважна причина. Плата за ліцензування швидко зростає, особливо якщо використовується на багато інструментів, і вони можуть бути значною частиною бюджету, якщо ви перебуваєте в невеликій організації. Python, безумовно, має привабливість бути безкоштовним. Це правда, що при використанні Python не доведеться платити ліцензію ні від кого, і що користувачі Python мають доступ до безлічі безкоштовних пакетів з відкритим кодом. Python дозволить вам бути більш гнучким і продуктивним у довгостроковій перспективі, що показали багато людей і компанії. Вибір Python, або будь-якої іншої мови з відкритим вихідним кодом, дозволяє запускати код «назавжди». Ви не заблоковані для даного постачальника. Не потрібно платити ліцензійний збір для того, щоб тримати сервер. Що ще важливіше, це означає, що люди, яких ви навіть не знаєте, можуть запускати ваш код Python без покупки ліцензії. Це може значно підвищити шанси на виживання проекту. Python має перевагу в першу чергу як мова програмування загального призначення. Він є чудовою мовою для наукових обчислень, і навіть має деякі особливості для цього, але це не тільки науково обчислювальна мова. Він може бути використаний для того, щоб зробити все, від побудови системи синхронізації файлів (Dropbox), служби фотозйомки (Instagram), 3-D програми для редагування відео та відео (Blender), платформи відеохостингу (YouTube), для виявлення гравітаційних хвиль . Наслідком такого різноманітного використання є те, що можна знайти інструменти для виконання майже всіх загальних завдань. Python дозволяє використовувати його для всього, від апаратного контролю до веб-API і настільних додатків. А для випадків, коли функція або бібліотека існує лише на іншій мові, Python є чудовою мовою "клею". Він може легко взаємодіяти з бібліотеками C / C ++ і Fortran, і є реалізації Python для деяких з основних інших

мов, таких як IronPython для C # і Jython для Java. Python - це мова програмування загального призначення (на відміну від R або Matlab). Вона легка в освоєнні і використанні в першу чергу тому, що мова зосереджена на читабельності. Це популярна мова в цілому, послідовно з'являючись у топ-10 мов програмування в опитуваннях StackOverflow (наприклад, результати дослідження 2015 року). Python є динамічною мовою і дуже підходить для інтерактивної розробки і швидкого прототипування з можливістю підтримки розробки великих додатків. Python також широко використовується для машинного навчання і наука про дані через відмінну бібліотечну підтримку. Вона швидко стала однією з домінуючих платформ для фахівців у галузі машинного навчання та наукових даних і користується більшим попитом, ніж навіть платформа R роботодавцями. Це означає, що ви можете виконувати свої дослідження і розробки (з'ясувати, які моделі використовувати) на тій же мові програмування, що використовується в операціях, значно спрощуючи перехід від розробки до операцій.

Є дистрибутив під назвою Anaconda Python, який можна завантажити і встановити безкоштовно. Він підтримує три основні платформи Microsoft Windows, Mac OS X і Linux. Вона включає в себе Python, SciPy і scikit-learn: все, що потрібно для вивчення, практики і використання прогнозування часових рядів з середовищем Python.

Одним з найпопулярніших інструментів для побудови моделі є statmodels. statsmodels - це модуль Python, який надає класи і функції для оцінки багатьох різних статистичних моделей, а також для проведення статистичних тестів і дослідження статистичних даних. Для кожного оцінювача доступний великий список статистичних даних результатів. Пакет випущено під відкритою ліцензією модифікованої BSD. Онлайн-документація розміщена на сайті statsmodels.org. Великий перелік описової статистики, статистичних тестів, графічних функцій і статистики результатів доступний для різних типів даних і кожного оцінювача. Він доповнює модуль статистики SciPy. Statsmodels є частиною наукового стеку Python, який орієнтований на аналіз даних, наука даних і статистику. Statsmodels побудований поверх чисельних бібліотек NumPy і SciPy, інтегрується з Pandas для

обробки даних і використовує Patsy для інтерфейсу R-подібної формули. Графічні функції засновані на бібліотеці Matplotlib. Statsmodels надає статистичний сервер для інших бібліотек Python. Окрім початкових моделей, лінійної регресії, надійних лінійних моделей, узагальнених лінійних моделей і моделей для дискретних даних, останнім випуском моделей scikits.stats є деякі основні засоби і моделі аналізу часових рядів. Це включає дескриптивну статистику, статистичні тести і кілька класів лінійних моделей: авторегресивні, AR, авторегресивні ковзні середні, ARMA. У цій статті ми хотіли б представити та дати огляд нових особливостей аналізу часових рядів statsmodels. Дані часових рядів містять спостереження, які упорядковані вздовж одного виміру, тобто часу, що накладає на дані конкретні стохастичні структури. Сучасні моделі припускають, що спостереження є безперервними, що час є дискретним і однаково рознесеним і що немає відсутніх спостережень. Цей тип даних дуже поширений у багатьох галузях, наприклад, в економіці та фінансах, національному виробництві, робочій силі, цінах, цінностях фондового ринку, обсягах продажів.

Ідентифікація процесів ARIMA (p, d, q), зокрема вибір кількості термінів, p та q. Одна з рекомендацій у методології Box-Jenkins полягає в тому, щоб подивитися на закономірності автокореляції і часткової автокореляції функцій scikits.statsmodels.tsa.arima_process містить клас, який забезпечує кілька властивостей процесів ARMA і генератора випадкових процесів. Як приклад, statsmodels / examples / tsa / arma_plots.py можуть бути використані для побудови автокореляційних і часткових автокореляційних функцій для різних моделей ARMA. Розробка statsmodels за останні кілька років була зосереджена на побудові правильних і перевірених реалізацій стандартного набору економетричних моделей, доступних в інших статистичних обчислювальних середовищах, таких як R. Проте ще є довга дорога, перш ніж Python буде на тому ж самому рівні. Рівень бібліотечної роботи з іншими обчислювальними середовищами зосереджений на статистиці та економетриці. Я вважаю, що, враховуючи багатство потужних наукових обчислень та інтерактивних дослідницьких інструментів у поєднанні з чудовою мовою Python, statsmodels можуть зробити Python передовим

середовищем для роботи зі статистикою. Подальша робота потребуватиме інтеграцію всіх цих інструментів для створення плавного та інтуїтивно зрозумілого користувальницького досвіду, порівнянного зі стандартними комерційними та відкритими статистичними продуктами.

Мова програмування Python інтерпретується, динамічно вводиться і є достатньо високорівневую. У порівнянні з іншими мовами програмування, які зазвичай використовуються для статистичних обчислень, він має як сильні, так і слабкі сторони. Не має широти доступних статистичних процедур, присутніх на мові програмування R, але замість цього містить основний стек добре розвинених наукових бібліотек. Починаючи з мов програмування загального призначення, їй не вистачає належного розуміння матричної алгебри, що робить MATLAB таким легким для початку роботи (ці функції доступні, але надаються NumPy) та науковим Python (SciPy)) бібліотеки), але вона має більше вбудованих функцій для роботи з текстом, файлами, веб-сайтами тощо. Всі Python, R і MATLAB мають відмінні графічні можливості, а також можливість інтегрувати скомпільований код для більш швидкої роботи. Звичайно, все, що може бути зроблено однією мовою, в принципі може бути зроблено в багатьох інших, тому знайомство, стиль і традиція відіграють істотну роль у визначенні того, якою мовою використовується дисципліна.

Щоб з'ясувати яке росподілення має наш ряд треба провести тести. Проведемо тест Харки - Бера для визначення номарльності розподілу, щоб підтвердити припущення про однорідність. Для цього в існує функція jarque_bera (), яка повертає значення даної статистики. Значення цієї статистики свідчить про те, нульова гіпотеза про нормальність розподілу відкидається з малою вірогідністю (probably> 0.05), і, отже, наш ряд має нормальний розподіл. Як вже було зазначено вище – ряд є стаціонарним. Згідно з алгоритмом описаним в розділі 2.4, залишилося створити код для знаодження p та q. ACF допоможе нам визначити q, так як по її коррелограмм можна визначити кількість автокореляційних коефіцієнтів сильно відмінних від 0 до моделі MA. PACF допоможе нам визначити так як по її коррелограмм можна визначити максимальний номер коефіцієнта

сильно відмінний від 0 до моделі AR. Щоб побудувати відповідні коррелограмми, в пакеті statsmodels є наступні функції: plot_acf () і plot_pacf (). Вони виводять графіки ACF і PACF, у яких по осі X відкладаються номери лагов, а по осі Y значення відповідних функцій. Потрібно відзначити, що кількість лагів у функціях і визначає число значущих коефіцієнтів. Отже, коли відомі всі параметри можна побудувати модель, але для її побудови ми візьмемо не всі дані, а тільки частину. Дані з частини не потрапили до модель ми залишимо для перевірки точності прогнозу нашої моделі. Весь код програми створений для прогнозування наведено у додатку А.

Для створення графіків та відображення (візуалізації даних) може використовуватись Matplotlib. Matplotlib - це бібліотека для побудови Python 2D, яка виробляє показники якості публікацій у різноманітних друкованих форматах та інтерактивних середовищах на різних платформах.

ВИСНОВКИ

Встановлення маркетингових стратегій є важливим процесом для майбутнього планування підприємницької діяльності компанії, будучи однією з найважливіших завдань тих, хто приймає рішення. Але стратегії повинні підтримуватися твердою інформацією про минулу діяльність компаній та перспективи розвитку, починаючи з поточної позиції на ринку. У цьому контексті прогнозування продажів може допомогти особам, які приймають рішення, впроваджувати на практиці відповідні стратегії. Система прогнозування може бути використана для поліпшення прогнозів попиту, дозволяючи перегляд закупівель або виробничих рішень, коли це необхідно, щоб уникнути високих витрат на старіння в кінці життєвого циклу продукції або уникнути відмов, що впливають на рівень обслуговування клієнтів. Було наведено і описано найбільш популярні прогнозні моделі, із решти для вирішення завдань предиктивної аналітики. На основі проведеного огляду сформульовані основні переваги та недоліки різних типів моделей, на основі яких можна приймати рішення про використанні того чи іншого виду моделей для вирішення конкретних завдань. Було виявлено, що найбільш перспективним підходом на даний момент можна вважати використання нейронних мереж, так як вони дозволяють побудувати найбільш гнучку і повну модель. У домені електронної комерції існує великий потенціал для підвищення точності прогнозування продажів.

Прогнозування продажів в індустрії є складним питанням протягом багатьох років. Було зроблено багато зусиль для підвищення точності систем прогнозування з конкретними обмеженнями в цій області. Передові технології, такі як екстремальна машина навчання, дозволили шукачам збільшити здатність систем витягувати інформацію з історичних даних, навіть якщо ці дані сильно порушені. Методи інтелектуального аналізу даних та методи екстраполяції, що базуються на “попередньому продажу”, можуть бути дуже потужними, якщо немає історичних даних. Всі ці методи будуть покращуватися знову і знову в найближчому

майбутньому. Проте інші теми також можуть бути дуже цікавими для вивчення. Дійсно, індустрія ритейлу є дуже динамічним сектором. З'являються нові ринки і, як наслідок, нові обмеження і нові вимоги до систем прогнозування. Ці еволюції є привабливими можливостями для дослідників у наступні десятиліття. Серед цих нових тенденцій, масова стратегія налаштування в даний час представляє невелику вибірку продукції, але може збільшуватися і змінювати потреби в термінах прогнозу та пропозиції. Нарешті, ще одним способом поліпшення прогнозів продажів може бути більш глибоке дослідження управління знижкою цін, просування, непроданих. Дуже корисною системою може бути система підтримки прийняття рішень, що базується на прогнозуванні продажів, щоб допомогти компаніям управляти своїми продажами, а також їх прибутки за ціною продукту. У такій системі механізм прогнозування повинен вміти точно моделювати зв'язок між продажем і ціною продукту.

Більшість моделей прогнозування часових рядів вимагають, щоб дані були стаціонарними. Після трансформації також використовувались інші методики, такі як ковзаюча середня (MA). Модель авторегресійної ковзного середнього (ARMA) була використана для прогнозування продажів. Модель показала багатообіцяючі результати, оскільки вона добре продемонструвала прогноз продажів. Проте, більш просунуті методи, запропоновані в літературі, мають потенціал, щоб перевершити її і зробити прогноз більш надійним. Використана методика прогнозування далека від досконалості, однак запропонована модель забезпечує продавців методом обробки передових проблем часових рядів у стислій і зрозумілій формі.

ПЕРЕЛІК ПОСИЛАНЬ

1. Global B2C Ecommerce Sales to Hit $1.5 Trillion This Year Driven by Growth in Emerging Markets. Asia-Pacific leapfrogs North America to become world's largest regional ecommerce market [Electronic resource]. — Mode of access: http://www.emarketer.com/Article/GlobalB2C-Ecommerce-Sales-Hit-3615-TrillionThis-Year-Driven-by-Growth-EmergingMarkets/1010575.

2. Бокс Дж., Дженкинс Г.М. Анализ временных рядов, прогноз и управление. М.: Мир, 1974. 406 с.

3. Мелас В.Б., Шпилев П.В. — Планирование и анализ для регрессионных моделей: Учеб. пособие. — СПб., 2012. — 75 с.

4. Авторегрессионная модель // The free encyclopedia «Wikipedia». Режим доступа: https://ru.wikipedia.org/wiki/Авторегресионная_модель (дата обращения 20.05.2017) .

5. Модель скользящего среднего // The free encyclopedia «Wikipedia». Режим доступа: https://ru.wikipedia.org/wiki/ Модель_скользящего_среднего (дата обращения 20.05.2017) .

6. Эконометрия: Учебное пособие / В.И. Суслов [и др.] Новосибирск: Издательство СО РАН, 2005. 744 с.

7. Prajakta S.K. Time series Forecasting using Holt-Winters Exponential Smoothing // Kanwal Rekhi School of Information Technology Journal 2004. 13 p. Режим доступа :http://www.it.iitb.ac.in/~praj/acads/seminar/04329008_ExponentialSmoothing.pdf (дата обращения 12.05.2017).

8. Чучуева А.И. Модель прогнозирования временных рядов по выборке максимального правдоподобия: дис. … канд. техн. наук. М., 2012. 155 с. [6]. Афанасьев В. Н., Юзбашев М. М. Анализ временных рядов и прогнозирование //М.: Финансы и статистика. – 2001. – Т. 228. – С. 2.

9. Вентцель А.Д. Курс теории случайных процессов — 2-е изд., доп. — М.: Наука. Физматлит, 1996.

10. Суслов В.И., Ибрагимов Н.М., Талышева Л.П., Цыплаков А.А. Эконометрия. Часть III Эконометрия - I: Анализ временных рядов — Учебное пособие — Новосибирск: Изд-во СО РАН, 2005. — 744 с.