НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«Національний університет біоресурсів і природокористування України»

Факультет інформаційних технологій

«На правах рукопису»
УДК №1939 "С" від 30.12.2022

«До захисту допущено»
ВО завідувача кафедри
_____ ?????
«___»_____20__ р.

Магістерська дисертація

на здобуття ступеня магістра

зі спеціальності 121 Інженерія програмного забезпечення

на тему: «Аналіз та прогнозування відтоку клієнтів інтернет провайдера з використанням технологій машинного навчання»

Виконав:
студент VI курсу, групи ІПЗ-22004м
ПЯВЧИК МАКСИМ ОЛЕКСАНДРОВИЧ _____

Керівник:
Доцент кафедри інформаційно-комунікаційних технологій та систем, к.т.н.,
Ніколаєнко Дмитро Володимирович
_____

Рецензент:
доцент кафедри ТК, к.т.н., доцент
???

Засвідчую, що у цій дипломній роботі немає запозичень з праць інших авторів без відповідних посилань.
Студент _____

Київ – 2023 року

Національний технічний університет України
«Національний університет біоресурсів і природокористування України»
Факультет інформаційних технологій

Рівень вищої освіти – другий (магістерський) за освітньо-професійною програмою

Спеціальність (освітня програма) – 121 «Інженерія програмного забезпечення»

ЗАТВЕРДЖУЮ
ВО завідувача кафедри
_____ ??????
«___»_____20___ р.

ЗАВДАННЯ
на магістерську дисертацію студенту
Пявчику Максиму Олександровичу

1. Тема дисертації «Аналіз та прогнозування відтоку клієнтів інтернет провайдера з використанням технологій машинного навчання», керівник роботи Ніколаєнко Дмитро Володимирович, кандидат економічних наук, к.т.н., затверджені наказом по університету від «?» листопада 2023 р. №???
2. Термін подання студентом роботи ?? грудня 2023
3. Вихідні дані до роботи:
4. Зміст роботи
   1) Концепція та особливості побудови FOG мереж;
   2) FOG-мережі з інтелектуалізованою системою управління;
   3) Аналіз конфігурацій та особливостей використання програмних засобів для моделювання FOG мережі;
   4) Імітаційна модель для дослідження характеристик FOG-мережі;
   5) Висновки

5. Перелік ілюстративного матеріалу (із зазначенням плакатів, презентацій тощо) Презентація-захист на тему: «Розробка моделі FOG-мережі з інтелектуалізованою системою управління»

- Плакат №1 (слайд) Вступний слайд для розкриття назви теми та привітання;
- Плакат №2 (слайд) Тема роботи, мета, об’єкт та предмет дослідження, завдання дослідження;
- Плакат №4 (слайд) Постановка завдання;
- Плакат №5 (слайд) Об’єкти в;
- Плакат №6 (слайд) Архітектура FOG-мережі використана в роботі;
- Плакат №8 (слайд) Висновки;

6. Дата видачі завдання 10 жовтня 2020 року

Календарний план

№ з/п	Назва етапів виконання дипломної роботи	Термін виконання етапів роботи	Примітка
1	Аналіз отриманого завдання	01.09.2020-07.09.2020	Виконав
2	Визначення мети дипломної роботи та розробка змісту	28.09.2020-15.10.2020	Виконав
3	Написання вступної частини дипломної роботи	16.10.2020-27.12.2020	Виконав
4	Написання першого розділу.	15.01.2021-15.04.2021	Виконав
5	Написання другого розділу.	16.04.2021-20.06.2021	Виконав
6	Написання третього розділу.	21.06.2021-02.07.2021	Виконав
7	Написання четвертого розділу.	05.07.2021-18.09.2021	Виконав
8	Написання загального висновку	05.10.2021-11.11.2021	Виконав
9	Оформлення дипломної роботи	12.11.2021-24.11.2021	Виконав
10	Підготовка презентації до захисту	27.11.2021-10.12.2021	Виконав

Студент Максим ПЯВЧИК

Керівник роботи Дмитро НІКОЛАЄНКО

РЕФЕРАТ

Текстова частина дипломної роботи: ?? с., рис 29., джерел 20.

Метою роботи є пошук та виявлення факторів, що змушують користувачів перейти до конкуруючого інтернет провайдера.

В даній роботі розглянуто технології машинного навчання та порівняння їхніх характеристик.

У практичній частині зімітовано машинне навчання на основі нейроної мережі та логічнорї регресії у інструменті Google Colab.

МАШИННЕ НАВЧАННЯ, МОВА ПРОГРАМУВАННЯ PYTHON, ML, BIG DATA

ABSTRACT

The purpose of the work is to find and identify factors that force users to switch to a competing Internet provider.

This work examines machine learning technologies and compares their characteristics.

In the practical part, machine learning based on a neural network and logical regression was simulated in the Google Colab tool.

MACHINE LEARNING, PYTHON PROGRAMMING LANGUAGE, ML, BIG DATA

Зміст

ПЕРЕЛІК СКОРОЧЕНЬ

JDK — Java Developer Kit;

ЦОД — центр обробки даних;

ЦП — Центральний процесор;

LAN — Local Area Network;

WAN — Wide Area Network;

IoT — Internet of Things;

QoS — Quality of Service;

NFC — Near field communication;

SIM — Subscriber Identification Module;

LTE — Long-Term Evolution;

AIDC — Automatic Identification and Data Capture;

MIPS — Million Instructions Per Second;

ІСУ — Інтелектуальна система управління;

ЦПР — Центр прийняття рішення;

ЕКГ — Електрокардіографія;

IP — Internet Protocol;

SDN — Software define network;

NFV — Network Functions Virtualization;

API — Application programming interface;

CP/CMS — Control Program/Cambridge Monitor System;

AWS — Amazon Web Services;

IDE — Integrated Drive Electronics;

GUI — Graphical User Interface;

ML – Machine Learning;

№ з/п	Назва етапів виконання магістерської роботи		
1	Аналіз задачі		
2	Розгляд моделей та методів, що використовуються для знаходження маси відтоку клієнтів		
3	Вибір та порівняння декількох методів		
4	Процес передбачення «маси» відтоку клієнтів		
5	Бізнес аналіз та побудова прогнозування на основі отриманих даних		
6	Висновок по роботі		

ВСТУП

Актуальність теми. Україна живе в умовах воєнного стану уже більше 600 днів і це означає, що усі головні індустрії, а особливо сфера електронних комунікацій, мають бути готовими до різних негативних сценаріїв та тривалої відсутності енергопостачання в осінньо-зимовий період.

Задля того щоб українці залишалися на зв'язку, вже в перші дні війни було напрацьовано кілька дієвих рішень, серед яких – впровадження національного роумінгу - для забезпечення тимчасової можливості громадянам України у разі відсутності покриття мобільного оператора, який надає їм послуги, користуватися мережею іншого оператора безкоштовно, лише здійснивши простий алгоритм вибору доступної мережі. 7 березня 2022 року мобільними операторами було запущено послугу національного роумінгу, яка діє і сьогодні та стала дуже потрібною під час аварійних відключень електроенергії та в інших випадках. У складні місяці перепадів з енергопостачанням щодня послугою національного роумінгу користувалося близько 2,2 млн українців.

Але це не є панацеєю для мереж кожного оператора і після перших блекаутів 2022 року стало очевидним, що готуватися треба системно і потрібно накопичувати необхідні резерви автономності мереж операторів та бути в готовності забезпечити зв'язком українців як мінімум на час 3-добової відсутності електроживлення мережі.[1]

Зважаючи на військовий час та проблеми для бізнесу, що він несе дуже велика частина постійних клієнтів почала перетікати до конкурентів, що пропонували більш вигідніші умови, або мали можливість інвестувати більше коштів в енергонезалежність через розмір обороту прибутку в компанії, тому малі інтернет провайдери «домомережі», чий бізнес-процес був зосереджений на певній групі клієнтів, що проживала на обмежені географічній ділянці почали терміново впроваджувати зміни в архітектуру побудови бізнесу та процесів, задля збереження основного ядра клієнтської бази.

Використання машинного навчання в покращенні аналізу бізнес-процесів має велику актуальність у сучасному світі та постійних проблемах сьогодення. Ось декілька ключових аргументів, що пояснюють цю актуальність:

1. Збільшення обсягу даних: Сучасні компанії збирають величезні обсяги даних з різних джерел, включаючи трансакційні дані, дані про клієнтів, дані з соціальних мереж, даних сенсорів тощо. Машинне навчання може допомогти вилучити цінну інформацію із цього потоку даних та аналізувати її для прийняття бізнес-рішень.
2. Покращення передбачення: Машинне навчання може допомогти в удосконаленні передбачення різних аспектів бізнес-процесів, таких як попит на продукти, ціни, виробництво та логістика. Аналітичні моделі можуть прогнозувати майбутні події на основі історичних даних і змінювати стратегію бізнесу відповідно до цих передбачень.
3. Виявлення аномалій: Машинне навчання може допомогти виявляти аномалії та невідповідності в бізнес-процесах, які можуть вказувати на проблеми або можливості для оптимізації. Наприклад, виявлення шахрайства, дефектів у виробництві чи несправностей в постачанні.
4. Автоматизація рутинних завдань: Машинне навчання може бути використане для автоматизації рутинних завдань і процесів в компанії, що дозволяє працівникам більше уваги приділяти стратегічним аспектам бізнесу. Наприклад, автоматизація обробки документів, обслуговування клієнтів або реагування на клієнтські запити.
5. Покращення персоналізації: Машинне навчання може допомогти аналізувати дані про клієнтів та розробляти індивідуальні підходи до обслуговування клієнтів. Це дозволяє компаніям створювати більш ефективні маркетингові кампанії та поліпшувати зв'язок із клієнтами.

Загалом, машинне навчання відкриває безліч можливостей для оптимізації бізнес-процесів, підвищення ефективності і покращення

конкурентоспроможності компаній. Розуміння та використання цих технологій може стати ключовим чинником успіху у сучасному бізнес-середовищі.

Як вже було зазначено вище проблема відтоку клієнтів стала дуже гостро з приходом віялових відключень та аварій в електричних мережах, до яких більшість з інтернет провайдерів не були підготовлені.
Інтернет-провайдери вживають різноманітні стратегії і практики для того, щоб перешкодити відтоку клієнтів. Декілька засобів, якими вони можуть здійснювати це, включають:

Покращення обслуговування клієнтів: Надання якісних послуг та підтримки може суттєво вплинути на задоволеність клієнтів. Інтернет-провайдери намагаються забезпечити швидке вирішення проблем, які виникають у клієнтів, і надають проактивну підтримку для виявлення і вирішення можливих проблем.

Персоналізовані пропозиції: Використовуючи дані про клієнтів, провайдери можуть створювати індивідуалізовані пропозиції і пакети послуг, які відповідають потребам кожного клієнта. Це може зробити послуги більш привабливими для клієнтів і зменшити ймовірність їхнього відтоку.

Програми лояльності: Впровадження програм лояльності і нагород для клієнтів може стимулювати їх залишатися з даним провайдером і не переходити до конкурентів.

Маркетинг та комунікація: Інтернет-провайдери можуть використовувати маркетингові кампанії та комунікаційні стратегії для того, щоб показати клієнтам переваги своїх послуг, а також надавати інформацію про нові послуги та акції.

Аналітика та машинне навчання: Використання аналітики та машинного навчання дозволяє інтернет-провайдерам аналізувати дані клієнтів та передбачати, які клієнти можуть виявити намір змінити провайдера. Це дозволяє провайдерам вживати заходів для утримання цих клієнтів, наприклад, запропонувати їм індивідуальні умови або знижки.

Постійне вдосконалення послуг: Відслідковуючи відгуки та реакції клієнтів, інтернет-провайдери можуть постійно вдосконалювати свої послуги та інфраструктуру, щоб задовольнити потреби клієнтів і підвищити їхнє задоволення.

Ці практики і стратегії спільно допомагають інтернет-провайдерам зменшити відтік клієнтів та зберегти існуючу клієнтську базу, що є важливим для підтримання стійкості і конкурентоспроможності на ринку послуг доступу до Інтернету.

Мета та задачі дослідження. Метою роботи є пошук та виявлення факторів, що змушують користувачів перейти до конкуруючого інтернет провайдера.

Для досягнення мети було поставлено та вирішено певні задачі, а саме:

1. Аналіз існуючих методів машинного навчання, що використовуються для вирішення проблеми з переходом клієнтів до конкурентів.
2. Проаналізувати отримані дані, що були отримані з відкритих джерел.
3. Використовуючи відкриті дані провести їх очищення та приводження до таких, які можна використовувати для машинного навчання.
4. Промоделювати навчання декількома методами, шо можуть бути використані для аналізу відтоку клієнтів.
5. На основі вихідних аналітичних даних визначити показники, що найбільше впливають на вибір відтоку абонентів.
6. Створити бізнес-сценарії, що можуть запобігти негативній тенденції в майбутньому.
7. Використовуючи випадкову вибірку перевірити життєздатність концепції запропонованого бізнес-процесу.

Об’єкт дослідження – процес відтоку клієнтів у оператора послуг Інтернет зв’язку.

Предмет дослідження – методи статистичного аналізу великих даних, а саме: метод дерева рішень, асоціативних правил та bagging.

Наукова новизна одержаних результатів. Новизною в роботі є комплексний підхід до вирішення завдань та групування декількорї методівв та їх переваг задля отримання найкращої вибірки по запобіганню відтоку абонентів Інтернет провайдера.

РОЗДІЛ 1
ВИКОРИСТАННЯ ТЕХНОЛОГІЇ BIG DATA В СФЕРІ ТЕЛЕКОМУНІКАЦІЙНИХ ПОСЛУГ

1.1 Введення в технології BIG DATA

«Великі дані (Big Data) – позначення структурованих и неструктурованих даних величезних обсягів і значного розмаїття, що піддаються ефективній обробці програмних інструментів, які горизонтально масштабуються та з’явились у кінці 2000-х років, і альтернативних традиційних систем управління базами даних і рішенням класу рішень Business Intelligence».[2]

Використання технології Big Data в телекомунікаціях має глибокий вплив на галузь та створює низку переваг та можливостей. Вона дозволяє операторам зв'язку збирати, аналізувати та використовувати великі обсяги даних для оптимізації діяльності та покращення обслуговування клієнтів. Основні аспекти використання Big Data в телекомунікаціях можна розділити на наступні пункти:

- Аналіз поведінки клієнтів: Великі обсяги даних дозволяють операторам ретельно аналізувати поведінку своїх клієнтів. Вони можуть вивчати, які послуги користувачі використовують найбільше, коли і де вони це роблять. Це допомагає створювати персоналізовані пропозиції та забезпечувати кращий досвід обслуговування.
- Оптимізація мережі: Аналіз даних дозволяє операторам вдосконалити управління мережею, прогнозувати навантаження і розподіляти ресурси ефективніше. Це зменшує перевантаження мережі та покращує якість обслуговування.
- Боротьба з шахрайством і кібербезпека: Аналіз Big Data використовується для виявлення шахраїв та зловмисників, а також для виявлення можливих кіберзагроз. Це сприяє забезпеченню безпеки мережі та даних клієнтів.

Розробка нових продуктів та послуг: Аналіз Big Data даних дозволяє операторам розробляти нові, інноваційні послуги та продукти, що відповідають ринковим потребам і підвищують конкурентоспроможність.

1.2 Аналіз поведінки клієнтів

Аналіз поведінки клієнтів - це процес збору, обробки і інтерпретації даних, які вказують на те, як клієнти взаємодіють і використовують продукти або послуги компанії. Цей аналіз дозволяє розуміти потреби, уподобання та звички клієнтів, а також прогнозувати їхню майбутню поведінку. Це важливий інструмент для бізнесу, оскільки допомагає підвищити задоволеність клієнтів, оптимізувати маркетингові зусилля та приймати рішення на основі даних.

Поведінка людини досить невизначений фактор, що може вносити дужу значні відхилення в моделювання відтоку, але все ж використовуючи статистику можна прийти до спільного множника в шаблонах поведінки клієнтів та лояльності до використання певних послуг зі зростанням проміжку часу. [3]

1.3 Оптимізація мережі

Оптимізація мережі в телекомунікаціях може значно покращити користувацький досвід, і це має велике значення для задоволення клієнтів і підвищення їхньої лояльність, що в свою чергу забезпечить формування збільшення кількості клієнтів. Ось яким чином оптимізація мережі сприяє покращенню користувацького досвіду:

Підвищення якості зв'язку: Оптимізація мережі допомагає зменшити перевантаження і знижує відставання в мережі. Це призводить до зменшення втрат дзвінків, переривань під час розмов і зниження шуму в мережі, що поліпшує якість зв'язку для клієнтів.

Забезпечення високої швидкості передачі даних: Оптимізована мережа може підвищити швидкість передачі даних, особливо в мережах мобільного Інтернету. Це означає, що користувачі можуть швидше завантажувати сторінки веб-сайтів, стрімінгові відео, застосунки та інші контенти.

Забезпечення доступності послуг: Оптимізація мережі дозволяє забезпечити більшу доступність послуг, особливо в регіонах з низьким покриттям сигналу. Це допомагає користувачам зберігати зв'язок навіть в умовах поганої мережі.

Менша витрата енергії та ресурсів: Оптимізація мережі дозволяє зменшити споживання енергії і оптимально використовувати ресурси мережі. Це може призвести до зниження вартості послуг для клієнтів.

Можливість розширення послуг: Завдяки оптимізації мережі оператори можуть впроваджувати нові послуги та функції, які покращують користувацький досвід. Це може включати в себе швидкісний доступ до нових технологій, таких як 5G, або покращені послуги для Інтернету речей.

Впровадження мереж з великою пропускною спроможністю: Оптимізація мережі дозволяє створювати мережі з великою пропускною спроможністю, що забезпечує можливість підключення більшої кількості пристроїв і підтримує попит на підключені пристрої та сервіси.[4]

1.4 Боротьба з шахрайством і кібербезпека

Інтернет став неодмінною частиною сучасного життя, принісши безліч переваг, включаючи спрощений доступ до інформації, комунікацію та онлайн-покупки. Проте, ця епоха також супроводжується зростаючими загрозами шахрайства в Інтернеті. Кіберзлочинці намагаються використовувати нові та вдосконалені методи, щоб обдурити користувачів та отримати незаконний доступ до їхніх особистих даних та фінансових ресурсів.

Шахрайство в Інтернеті має багато форм і проявів. Однією з найпоширеніших атак є фішинг, який полягає у відправці підроблених листів або повідомлень, що намагаються переконати користувачів надати конфіденційну інформацію, таку як паролі або номери кредитних карток. Кіберзлочинці докладають зусиль, щоб ці підроблені повідомлення були максимально схожі на легітимні, і користувачі попадають в пастку обману.

Додатковим видом кібершахрайства є розповсюдження шкідливого програмного забезпечення, такого як віруси, троянці та шпигунські програми. Це програмне забезпечення може нанести шкоду комп'ютерам та іншим пристроям користувачів, вкрасти їхні дані або навіть вимагати викуп за розблокування.

Зростає популярність атаки, відомої як рансомвар. Рансомвар блокує доступ до файлів чи пристрою користувача, а потім вимагає викуп у вигляді криптовалюти в обмін на розблокування. Ця атака може бути особливо небезпечною для користувачів, оскільки їм доводиться вибирати між викупом та втратою важливих даних.

Серйозним викликом для корпорацій і великих організацій є атаки на їхні системи. Кіберзлочинці можуть намагатися викрасти конфіденційну інформацію, вимагати викуп або завдавати шкоди великим компаніям, щоб збільшити свій прибуток або вплив.

Шахраї також використовують соціальні мережі для обману користувачів. Вони створюють фейкові профілі та поширюють дезінформацію, що може вплинути на громадську думку та викликати паніку серед користувачів.

Все частіше накладаються обмеження та регулювання доступу до мережі Інтернет державами. Мережа Інтернет стала необхідною частиною сучасного суспільства, забезпечуючи доступ до інформації, комунікації та можливостей для користувачів по всьому світу. Проте, разом із цим зростає необхідність обмеження та регулювання доступу до Інтернету державою. Це важливо для забезпечення безпеки, захисту прав та інтересів громадян та для контролю над різноманітними аспектами використання мережі.

Попередження зловживання: Держава може обмежувати доступ до Інтернету для запобігання зловживанням, таким як діяльність кіберзлочинців, розповсюдження шкідливого контенту або терористичні загрози. Це допомагає забезпечити безпеку і захист користувачів.

Захист особистих даних: Регулювання доступу до Інтернету дозволяє державі встановити стандарти щодо збереження та обробки особистих даних користувачів. Це важливо для захисту приватності та попередження незаконного використання особистої інформації.

Фільтрація контенту: Деякі країни встановлюють фільтри контенту для блокування доступу до сайтів іноземних ЗМІ або інформації, яка може бути розцінена як шкідлива або небажана. Це може бути зроблено з метою контролю над інформаційним простором і впливу на суспільну думку.

Захист від кібератак: Держава може вживати заходів для захисту критично важливих інфраструктур від кібератак та інших кіберзагроз. Це включає в себе регулярні аудити безпеки та створення нормативних актів для захисту від цих загроз.

Обмеження діяльності на Інтернеті: Деякі держави можуть обмежувати доступ до соціальних мереж, новинних ресурсів чи інших платформ в рамках цензури або політичних мотивів. Це викликає дискусії щодо свободи слова та прав людини.

Захист від дитячої порнографії та образи: Регулювання доступу до Інтернету допомагає боротися з поширенням дитячої порнографії та образливого контенту, забезпечуючи захист прав неповнолітніх. [5]

1.5 Збір та обробка даних для роботи з Big DATA

Сучасний світ характеризується вибуховим зростанням обсягу даних, які зберігаються, передаються та аналізуються. Цей обсяг і різноманітність даних, включаючи текст, зображення, відео та структуровані дані, спричинили народження поняття Big Data. Використання цих даних може призвести до відкриття нових можливостей для бізнесу, науки та суспільства. Однак для ефективного використання Big Data необхідні процеси збору та обробки даних.

Збір даних - перший та важливий етап у роботі з Big Data. Дані можуть бути зібрані з різних джерел, включаючи веб-сайти, соціальні мережі, датчики,

мобільні пристрої та багато інших. Важливо забезпечити, щоб зібрані дані були репрезентативними та якісними, оскільки якість вихідних даних визначає точність та надійність аналізу.

Обробка даних - наступний крок у роботі з Big Data. Основною метою обробки даних є виділення корисної інформації зі сирових даних та їх перетворення в зрозумілу форму. Цей процес може включати в себе фільтрацію, сортування, агрегацію, структурування та інші операції. Для обробки великих обсягів даних часто використовуються спеціалізовані платформи та інструменти, такі як Hadoop, Apache Spark, а також мови програмування, такі як Python та R.

Одним із ключових аспектів обробки даних є здатність виявляти патерни, тенденції та взаємозв'язки між даними. Це може допомогти в розумінні поведінки користувачів, прогнозуванні ринкових тенденцій та прийнятті стратегічних рішень. Машинне навчання та аналітика даних грають важливу роль у виявленні цих паттернів.

Після обробки даних, отримані результати можуть використовуватися для прийняття рішень, розробки нових продуктів та послуг, а також для оптимізації процесів в різних галузях, включаючи медицину, фінанси, телекомунікації та інші. Big Data також грають важливу роль у наукових дослідженнях, дозволяючи аналізувати великі обсяги даних та виявляти нові знання та залежності.

1.6 Аналіз даних в Big Data

З плином часу сучасний світ перетворився на справжнє "суцільне озеро даних," де обсяг та різноманітність інформації зростають з кожним днем. Ця практично нескінченна кількість даних, відома як Big Data, відкриває нові можливості для аналізу та використання інформації для розв'язання різних завдань. Аналіз даних в Big Data відіграє важливу роль у багатьох аспектах життя, від бізнесу та науки до медицини та громадської політики.

Однією з ключових особливостей Big Data є розмаїття джерел і форм даних. Це включає в себе структуровані дані, такі як таблиці та бази даних, а також

невлаштовані дані, такі як текст, зображення, відео, аудіо та дані з сенсорів. Аналіз такої різноманітної інформації вимагає спеціалізованих методів та інструментів.

Одним з підходів до аналізу Big Data є машинне навчання. Це поле штучного інтелекту дозволяє комп'ютерам вчитися та робити прогнози на основі великих обсягів даних. Машинне навчання використовується для виявлення паттернів, класифікації даних, прогнозування трендів та прийняття рішень на основі статистики та аналізу.

Інший підхід - це аналітика даних, яка використовується для виявлення залежностей та взаємозв'язків між даними. Аналітика даних допомагає виявляти складні структури в інформації, розуміти, як дані впливають на рішення та прогнози, та допомагає в прийнятті стратегічних рішень.

Аналіз даних в Big Data також має велике значення в бізнесі. Великі компанії використовують дані для вдосконалення стратегії маркетингу, підвищення ефективності логістичних процесів, прогнозування попиту та багатьох інших завдань. Аналітика даних може допомогти виявити нові ринкові можливості та підвищити конкурентоспроможність.

У науці аналіз даних використовується для відкриття нових знань та залежностей. Великі набори даних дозволяють дослідникам проводити більш детальні експерименти, виявляти тенденції та структури в інформації, які можуть залишитися непоміченими в менших обсягах даних.

Загалом, аналіз даних в Big Data відкриває нові горизонти для розв'язання складних завдань.

1.7 Майбутнє у Big Data

Big Data, або великі дані, залишаються однією з найбільш значущих технологічних тенденцій сучасності. Передбачається, що ця галузь буде розвиватися надзвичайно швидко і матиме значний вплив на різні сфери життя, від бізнесу і науки до медицини та громадської політики.[6]

Однією з ключових тенденцій майбутнього Big Data є зростання обсягу даних. Це стосується як кількості вже існуючих даних, так і генерації нових. З розвитком Інтернету речей (IoT), мобільних пристроїв та сенсорів з'явиться ще більше джерел для збору даних. Велика кількість даних створюватиме нові виклики у сферах збереження, обробки та аналізу.

Другою важливою тенденцією є зростання важливості кібербезпеки в Big Data. Зі збільшенням обсягу даних зростає й ризик їхнього втрати або крадіжки. Для захисту великих наборів даних, буде потрібно вдосконалення систем захисту, шифрування та моніторингу.

Ще однією тенденцією є розвиток технологій обробки Big Data в реальному часі. Здатність аналізувати дані миттєво набуває все більшого значення для прийняття рішень в реальному часі. Це може вплинути на такі галузі, як фінанси, медицина та телекомунікації.

Також спостерігається зростання використання штучного інтелекту (AI) і машинного навчання (ML) в аналізі Big Data. Ці технології допомагають виявляти патерни та залежності в даних, а також робити прогнози. Вони можуть бути використані для автоматизації процесів та вдосконалення прийняття рішень.

Майбутнє Big Data також пов'язане з розвитком інфраструктури для зберігання та обробки даних. Обчислювальні хмари та технології розподіленого зберігання даних будуть грати важливу роль у забезпеченні масштабованості та доступності для великих наборів даних.

1.8 Висновки

В розділі розглянуте поняття Big Data та його складові. Введення в технології Big Data сприяє створенню нових можливостей та вирішенню складних завдань у сучасному світі. Основні висновки стосуються важливості цих технологій та їх впливу на різні сфери життя:

- Зростання обсягу даних: Big Data відкриває нам новий світ, де обсяг і різноманітність даних швидко зростають. Це створює безліч нових

можливостей, але вимагає від нас адекватних засобів збереження, обробки та аналізу.

- Роль кібербезпеки: Зі зростанням обсягу даних, збільшується й ризик їхнього втрати або крадіжки. Захист даних стає надзвичайно важливим, і важливо вдосконалювати системи кібербезпеки.
- Швидкість та реальний час: Можливість аналізу даних в реальному часі набуває важливості, особливо для прийняття рішень в реальному часі. Це стосується багатьох галузей, включаючи фінанси, телекомунікації та медицину.
- Використання штучного інтелекту та машинного навчання: Технології штучного інтелекту та машинного навчання допомагають виявляти патерни та залежності в даних. Вони стають важливою частиною аналізу Big Data та допомагають приймати більш інформовані рішення.
- Інфраструктура для зберігання та обробки даних: Розвиток інфраструктури, такої як обчислювальні хмари та технології розподіленого зберігання даних, дозволяє забезпечити масштабованість та доступність для великих наборів даних.
- Вплив на бізнес та суспільство: Big Data впливає на багато сфер життя, від бізнесу і науки до медицини та громадської політики. Він допомагає вирішувати складні завдання та приймати стратегічні рішення.

Усі ці висновки підкреслюють важливість вивчення та впровадження технологій Big Data в сучасному світі. Збільшення обсягу даних, зростання важливості кібербезпеки та потреба у реальному аналізі даних роблять Big Data важливою складовою сучасної інформаційної та технологічної епохи.

Розділ 2
АЛГОРИТМИ, ЩО ВИКОРИСТОВУЮТЬСЯ ДЛЯ ЗАПОБІГАННЯ ВІДТОКУ КЛІЄНТІВ

Для запобігання відтоку клієнтів в інтернет-провайдерах і компаніях інших сфер використовують різні алгоритми машинного навчання. Деякі з них включають:

- Логістична регресія: Цей алгоритм використовується для прогнозування ймовірності відтоку клієнтів на основі різних факторів, таких як тривалість користування послугою, частота використання послуги, цінова політика та інші.
- Дерево рішень: Цей метод дозволяє створювати моделі, які аналізують набір правил для визначення, чому клієнти можуть покинути компанію. Вони допомагають визначити найбільш впливові фактори на відтік і вжити заходи для їх зменшення.
- Випадковий ліс: Випадковий ліс - це ансамбль дерев рішень, який може допомогти виявити більш складні взаємозв'язки між факторами та відтоком клієнтів. Він дозволяє створити більш точні моделі для прогнозування відтоку.
- Метод опорних векторів (SVM): SVM може бути використаний для виділення клієнтів, які мають найбільшу ймовірність відтоку, і для визначення, які фактори їх характеризують. Він допомагає створити моделі, які розрізняють клієнтів, які залишаються, і тих, які відходять.
- Нейронні мережі: Глибоке навчання і нейронні мережі можуть бути використані для аналізу великих обсягів даних і виявлення складних залежностей між факторами та відтоком клієнтів.

2.1 Логістична регресія

Логістична регресія - це один з найпоширеніших алгоритмів машинного навчання, який використовується для розв'язання задач класифікації та прогнозування ймовірності подій. Цей алгоритм здобув популярність завдяки своїй ефективності та великій кількості застосувань у різних галузях, включаючи медицину, фінанси, маркетинг та багато інших.

Однією з основних особливостей логістичної регресії є її використання у задачах бінарної та багатокласової класифікації. У бінарній класифікації алгоритм визначає, до якого з двох класів належить об'єкт, використовуючи логістичну функцію для оцінки ймовірності належності до одного з класів. У багатокласовій класифікації логістична регресія може бути розширена для визначення належності до більше ніж двох класів.

Основним інструментом, що використовується у логістичній регресії, є логістична функція. Вона перетворює значення лінійної комбінації вхідних факторів у значення між 0 та 1, які інтерпретуються як ймовірності. Ця функція допомагає моделі вирішувати задачі класифікації, де потрібно визначити, чи належить об'єкт до певного класу.

Ще однією особливістю логістичної регресії є можливість враховувати взаємозв'язки між факторами та виразити їх в аналітичних моделях. Завдяки цьому, логістична регресія може враховувати не тільки лінійні, а й нелінійні залежності між факторами та цільовими змінними.

Логістична регресія знаходить широке застосування в аналізі даних та прийнятті рішень. У медицині вона використовується для прогнозування ризику захворювань, у фінансах - для оцінки кредитного ризику, а в маркетингу - для аналізу клієнтської поведінки та прогнозування продажів.

2.2 Дерево рішень

Дерево рішень є одним із популярних алгоритмів машинного навчання, який використовується для прийняття рішень в різних галузях, від бізнесу до науки та медицини. Цей алгоритм має кілька особливостей, що роблять його важливим інструментом в аналізі даних та прийнятті рішень.

Дерево рішень - це графічна модель, яка представляє собою деревоподібну структуру з вузлами та гілками. Кожен вузол відповідає певному тесту на одному з факторів даних, а гілки вказують на різні варіанти відповідей. Під час подачі нового об'єкта на вхід алгоритму, він проходить від кореня до листя дерева, виконуючи тести та визначаючи кінцевий результат.

Однією з основних переваг дерева рішень є його інтерпретованість. Модель може бути легко візуалізована та зрозуміла як фахівцям, так і неспеціалістам. Це робить його корисним інструментом для виявлення закономірностей в даних та пояснення прийнятих рішень.

Дерева рішень можуть бути використані в задачах класифікації та регресії. У задачах класифікації, дерево визначає до якого класу належить об'єкт, поділяючи дані на підгрупи згідно з факторами. У задачах регресії, дерево рішень передбачає числове значення на основі вхідних даних.

Однак дерева рішень мають певні недоліки. Вони можуть схильні до перенавчання, коли модель дуже точно підлаштовується до навчальних даних і не може узагальнювати на нові дані. Для подолання цього недоліку використовуються методи, які зменшують глибину дерева, обмежуючи кількість розділів та розгалужень.

Дерева рішень також використовуються для важливості факторів, допомагаючи визначити, які з них найбільше впливають на рішення моделі.

Дерева рішень - це потужний інструмент машинного навчання, який знаходить широке застосування у багатьох галузях. Їх інтерпретованість,

здатність до класифікації та регресії роблять їх важливим інструментом для аналізу даних та прийняття рішень.

2.3 Випадковий ліс

Випадковий ліс - це потужний алгоритм машинного навчання, який використовується для класифікації та регресії. Він є формою ансамблевого навчання, що об'єднує багато дерев рішень, які працюють як колектив, для отримання більш точних результатів. Випадковий ліс вважається одним з найефективніших алгоритмів машинного навчання та знаходить застосування в багатьох галузях.

Ансамбель випадкового лісу складається з великої кількості дерев рішень, які навчаються на різних підмножинах даних. Під час класифікації, кожне дерево дає свій голос щодо класу, до якого може належати об'єкт, і результат обчислюється шляхом голосування. У випадку регресії, кожне дерево робить свій внесок у прогноз числової величини, і результат обчислюється як середнє значення всіх дерев.

Однією з основних переваг випадкового лісу є його висока точність. Завдяки ансамблю дерев рішень, цей алгоритм здатний виявляти складні взаємозв'язки між факторами та цільовими змінними, що дозволяє йому досягати високої точності у класифікації та регресії. Крім того, випадковий ліс має властивість робити важливість факторів, що допомагає визначити, які з них найбільше впливають на модель.

Іншою важливою особливістю випадкового лісу є його стійкість до перенавчання. Із великою кількістю дерев в ансамблі, алгоритм має меншу схильність до перенавчання, коли модель підлаштовується під навчальні дані і не може узагальнювати на нові дані.

Випадковий ліс має безліч застосувань у різних галузях. Він використовується для передбачення клієнтської поведінки, виявлення шахраїв, аналізу медичних даних та багатьох інших завдань. Важливо враховувати, що правильна настройка гіперпараметрів, таких як кількість дерев та їх глибина, грає важливу роль у досягненні найкращих результатів.

Загалом, випадковий ліс - це потужний і універсальний алгоритм машинного навчання, який відзначається високою точністю, стійкістю до перенавчання та великою кількістю застосувань. Він є важливим інструментом для вирішення складних завдань аналізу даних та прийняття рішень.

2.4 Метод опорних векторів (SVM)

Метод опорних векторів (SVM) є одним із найпоширеніших алгоритмів машинного навчання, використовуваних для задач класифікації та регресії. Вперше запропонований Вапником у 1963 році, SVM став важливим інструментом для багатьох дисциплін, включаючи комп'ютерний наук, статистику, фінанси та біологію. Його основні принципи та застосування вражають різноманітністю та ефективністю.

Основний принцип SVM полягає в тому, щоб знайти оптимальну гіперплощину, яка розділяє об'єкти різних класів в просторі ознак. Ця гіперплощина вибирається таким чином, щоб максимізувати відстань між нею та найближчими об'єктами кожного класу, які називаються опорними векторами. Цей підхід називається максимальним інтервалом та допомагає забезпечити найкращу узагальненість моделі на нових даних.

SVM також може використовуватися для регресії, де завдання полягає в тому, щоб знайти гіперплощину, яка найкраще підходить для розподілу числових значень. В цьому випадку SVM називається SVM для регресії або SVR. Він також

використовує принцип максимізації інтервалу, але застосовується до регресійних завдань.

SVM має безліч застосувань. Він використовується для класифікації текстових документів, виявлення облич у зображеннях, прогнозування фінансових ринків, аналізу генетичних даних та багатьох інших завдань. Його ефективність полягає в тому, що він здатний працювати з нелінійними залежностями між факторами та цільовими змінними завдяки ядровій функції, яка перетворює дані в вищорозмірний простір.

Однією з важливих особливостей SVM є його стійкість до перенавчання. Це означає, що модель здатна досягати високої точності на нових даних, навіть якщо вона була навчена на обмеженому об'ємі даних. Це робить SVM важливим інструментом для задач, де обмежені ресурси для навчання.

2.5 Нейронні мережі

Нейронні мережі - це один із найбільш обговорюваних та впливових алгоритмів машинного навчання в сучасному світі. Ці штучні нейронні системи інспіровані біологічною нейронною мережею людського мозку та відкривають безмежні можливості для аналізу даних, розпізнавання образів, передбачення та багатьох інших сфер.

Основна ідея нейронних мереж полягає в тому, щоб імітувати роботу нейронів у мозку. Нейрони приймають вхідні сигнали, обробляють їх і генерують вихідний сигнал. Нейронні мережі складаються зі штучних нейронів, які з'єднані в мережу та працюють разом для вирішення складних завдань. Кожен нейрон обчислює лінійну комбінацію вхідних даних та передає результат через функцію активації для нелінійного перетворення.

Нейронні мережі вражають своєю здатністю вивчати складні залежності в даних. За допомогою процесу навчання, вони адаптуються до вхідних даних та

навчаються визначати закономірності та шаблони. Однією з найпоширеніших форм навчання є навчання з вчителем, де модель навчається на основі пар вхідних та вихідних даних.

Однією з ключових переваг нейронних мереж є їх здатність до глибокого навчання, що призвело до розвитку глибокого навчання або нейронних мереж з багатьма шарами, відомими як глибокі нейронні мережі. Це дозволяє їм виявляти ще більше складних залежностей та вирішувати завдання, які раніше вважалися надзвичайно складними.

Нейронні мережі знаходять застосування в різних галузях, включаючи комп'ютерне зорове розпізнавання, обробку природної мови, голосове розпізнавання, біоінформатику, автономне водіння, фінанси та багато інших. Вони стали ключовим інструментом у вирішенні складних завдань, де інші алгоритми машинного навчання досягають обмежень.

Зростаюча популярність нейронних мереж та їхній успіх в різних галузях свідчать про їх великий потенціал у майбутньому машинного навчання.

2.6 Висновки до розділу 2

В цьому розділі було наведено найвідоміші алгоритми машинного навчання, що використовуються для вирахуванню відтоку клієнтів будь то Інтернет провайдера чи магазину одягу. Методи та алгоритми, а саме: Нейронні мережі, Метод опорних векторів (SVM), Випадковий ліс, Дерево рішень, Логістична регресія можна застосовувати для опису взаємодії між вхідними визнаннями та цільовою змінною.

Розділ 3

ПРОГНОЗУВАННЯ ВІДТОКУ КЛІЄНТІВ ІНТЕРНЕТ ПРОВАЙДЕРА З ВИКОРИСТАННЯМ ЬЕХНОЛОГІЙ МАШИННОГО НАВЧАННЯ

3.1 Прогнозування з використанням нейроної мережі

Прогнозування відтоку клієнтів полягає в тому, щоб визначити, чому клієнти залишають бізнес. Буде розглянуто відтік клієнтів у телекомунікаційному бізнесі, а саме в сфері послуг Інтернет провайдера. Буде створена модель глибокого навчання, щоб передбачити відтік і використовувати точність, запам'ятовування, показник f1 для вимірювання ефективності даної моделі. Для простоти та зручності буде використаний існуючий інструментарій для тестування технологій машинного навчання, а саме Google Colab. [7]

Перш за все потрібно імпортувати бібліотеки, щоб потім використовувати їх в подальшому прогнозуванні. В даній роботі використовуються стандартні бібліотеки, такі як: pandas, matplotlib, numpy. [8] [9] [10]

3.2 Пошук даних

Дані про клієнтів телекомунікаційними послугами було використано з відкритих джерел тільки для навчання та не мало на меті використання інформації задля збагачення або компрометування організації, всі можливі збіги це вигадка та не є існуючою інформацією. [11]

3.3 Завантаження даних

В відкритому віконці інструменту Google Colab [7] потрібно вписати наступні команди:

```
df1 = pd.read_csv("customer_churn_study_mat.csv")
df1.sample
```

Після внесення строк коду у віконце потрібно натиснути на емблему трикутника для запуску скрипта.

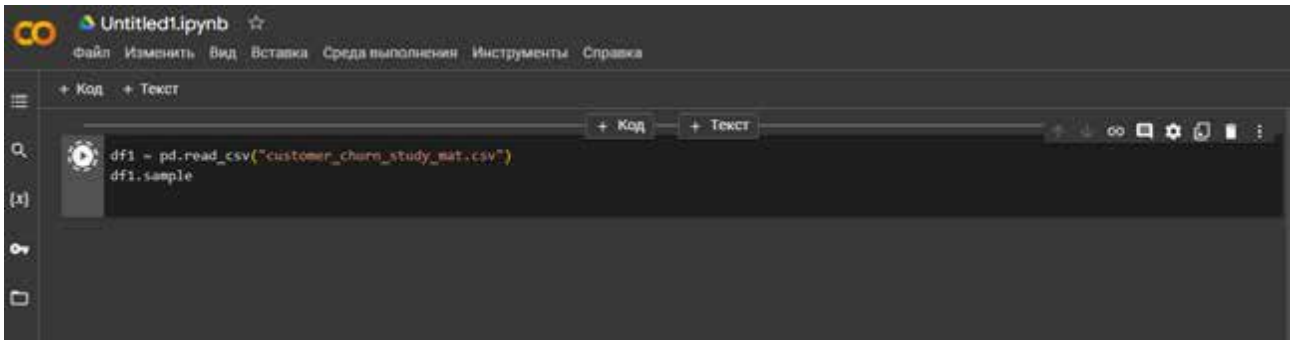

Рисунок 3.1 Завантаження даних

3.4 Очищення даних

Зазвичай дані це купа сміття та корисної інформацію, щоб отримати тількти важливе, потрібно провести процес очищення даних. В даному випадку потрібно видалити всі стобці, що будуть створювати лише лишній клопіт для нашої моделі.

```
df1.drop('customerID',axis='columns',inplace=True)
df1.dtypes
```

```
Out[254…  gender             object
          SeniorCitizen       int64
          Partner            object
          Dependents         object
          tenure              int64
          PhoneService       object
          MultipleLines      object
          InternetService    object
          OnlineSecurity     object
          OnlineBackup       object
          DeviceProtection   object
          TechSupport        object
          StreamingTV        object
          StreamingMovies    object
          Contract           object
          PaperlessBilling   object
          PaymentMethod      object
          MonthlyCharges    float64
          TotalCharges       object
          Churn              object
          dtype: object
```

Рисунок 3.2 Отриманий результат після видалення стовбця «customerID»

3.5 Перетворення строк в номери

Важливо пам’ятати, що машини розуміють тільки мову чисел, тому всі значення, що є строками потрібно перетворювати на числові значення.

```
df1[pd.to_numeric(df1.TotalCharges,errors='coerce').isnull()]
```

Out[258...

	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurit
488	Female	0	Yes	Yes	0	No	No phone service	DSL	Ye
753	Male	0	No	Yes	0	Yes	No	No	No interne servic
936	Female	0	Yes	Yes	0	Yes	No	DSL	Ye
1082	Male	0	Yes	Yes	0	Yes	Yes	No	No interne servic
1340	Female	0	Yes	Yes	0	No	No phone service	DSL	Ye
3331	Male	0	Yes	Yes	0	Yes	No	No	No interne servic
3826	Male	0	Yes	Yes	0	Yes	Yes	No	No interne servic
4380	Female	0	Yes	Yes	0	Yes	No	No	No interne servic
5218	Male	0	Yes	Yes	0	Yes	No	No	No interne servic
6670	Female	0	Yes	Yes	0	Yes	Yes	DSL	N
6754	Male	0	No	Yes	0	Yes	Yes	DSL	Ye

Рисунок 3.3 Отриманий результат від перетворень строк на цифри

3.6 Візуалізація даних

Машинне навчання в даний час відіграє важливу роль у багатьох галузях, включаючи бізнес, медицину, фінанси та науку. Процес розробки та налагодження моделей машинного навчання може бути складним завданням, і важливим аспектом є візуалізація даних. Візуалізація надає можливість відображати інформацію у вигляді графіків, діаграм, зображень та інших візуальних елементів. Це допомагає аналізувати та розуміти дані, виявляти

закономірності та патерни, а також спрощує взаємодію з моделями машинного навчання.

Перш за все, візуалізація допомагає зрозуміти дані. Наочна подача інформації у вигляді графіків або графічних представлень може значно спростити сприйняття даних та допомогти виявити важливі особливості. Наприклад, графік розсіювання може показати взаємозв'язок між двома змінними, що допомагає визначити, чи існує кореляція між ними.

Друга важлива функція візуалізації - це виявлення аномалій та викидів в даних. Графіки та діаграми можуть виділити значення, що виходять за межі звичайних варіантів, що важливо при виявленні помилок або аномалій в даних.

Окрім цього, візуалізація даних допомагає в підготовці та очищенні даних перед навчанням моделі. Вибірка некоректних або відсутніх даних може бути визначена завдяки аналізу графічних подань.

Важливою перевагою візуалізації в машинному навчанні є її роль у взаємодії з моделями та результатами. Візуальне представлення даних та результатів прогнозу допомагає спрощувати комунікацію між дослідниками, бізнес-аналітиками та іншими зацікавленими сторонами.

Загалом, візуалізація даних в машинному навчанні є необхідною складовою для розуміння, аналізу та взаємодії з даними та моделями. Вона допомагає виявляти важливі патерни, викиди та аномалії, спрощує процес підготовки даних та комунікації результатів. Тому важливо вдосконалювати навички візуалізації даних для досягнення кращих результатів в машинному навчанні.

Машинне навчання включає в себе багато завдань, включаючи класифікацію, регресію, кластеризацію та виявлення аномалій. Виявлення аномалій важливо в багатьох галузях, таких як фінанси, медицина, кібербезпека та виробництво. Однією з потужних та ефективних стратегій виявлення аномалій є використання візуалізації даних.

Візуалізація даних надає можливість представити дані у вигляді графічних зображень, що дозволяє аналізувати їх на основі візуальних подань. Основними типами візуалізації для виявлення аномалій є графіки розсіювання, графіки залежності, гістограми та коробкові діаграми. Ці графічні інструменти дозволяють візуалізувати взаємозв'язки між ознаками, розподіл даних та виявлення викидів.

Графіки розсіювання особливо корисні для виявлення аномалій, оскільки вони показують розташування точок даних у виміряному просторі. Аномальні точки зазвичай виділяються та можуть бути помічені як викиди на графіку. Графіки залежності можуть розкривати залежності між ознаками та дозволяють виявити аномалії, де дані не ведуть себе так, як очікувалося.

Гістограми та коробкові діаграми надають інформацію про розподіл даних та дозволяють виявити викиди на основі статистичних метрик. Аномальні дані зазвичай відображаються поза основним розподілом або як викиди в коробкових діаграмах.

Важливою перевагою візуалізації для виявлення аномалій є можливість спростити процес виявлення. Замість складних обчислень та алгоритмів, аналітик може використовувати свій зоровий аналіз для виявлення аномалій у даних.

Багато інструментів для візуалізації даних, таких як бібліотеки Python, R та інші, надають можливість створювати складні візуалізації з легкістю. Вони також дозволяють інтерактивність та можливість подробного дослідження даних.

Виявлення аномалій є важливим завданням в аналізі даних та машинному навчанні. Аномалії можуть бути ознакою ненормальних або потенційно проблемних ситуацій, які потребують уваги та виправлення. Одним з підходів до виявлення аномалій є використання бібліотеки Pandas в поєднанні з візуалізацією даних.

Pandas - це потужна бібліотека для роботи з даними в мові програмування Python. Вона надає структури даних, такі як DataFrame, що дозволяють легко завантажувати, обробляти та аналізувати дані. Для виявлення аномалій з Pandas, спочатку важливо завантажити та підготувати дані для аналізу.

Після завантаження даних в DataFrame, ви можете використовувати різні методи та функції для візуалізації даних. Один із найпростіших способів - це побудова графіків розсіювання, які дозволяють вам відобразити взаємозв'язок між двома ознаками даних. Якщо ви спостерігаєте велику концентрацію точок даних в одній області графіку та деякі точки розташовані подалі від цієї області, це може свідчити про наявність аномалій.

Гістограми - ще один корисний інструмент для аналізу розподілу даних. Гістограми дозволяють вам побачити, як дані розподілені по значеннях та чи є викиди або незвичайні патерни у розподілі.

Крім того, коробкові діаграми - це інший спосіб виявлення аномалій. Вони показують медіану, квартилі та викиди у розподілі даних. Якщо є точки даних, які виходять за межі "вус", це може свідчити про аномалії.

Після візуалізації даних і виявлення потенційних аномалій, ви можете подальше дослідження та аналізувати ці аномалії, щоб зрозуміти їхні причини та вплив на ваші дані. Далі можна використовувати різні аналітичні методи та моделі для подальшого виявлення та обробки аномалій.

У величезному обсязі даних, які часто зустрічаються в аналітиці, важливо вміти використовувати бібліотеку Pandas та візуалізацію даних для ефективного виявлення аномалій. Це допомагає виявляти потенційні проблеми та забезпечує якість та достовірність аналізу даних.

3.7 Візуалізація відтоку вибірці

Використовувючи вибірку можна створити графік в якому буде показаний наглядний приклад відтоку у порівнянні х плином часу, що абоненти користуються послугами.

```
a1_churn_no = df2[df2.Churn=='No'].MonthlyCharges
a1_churn_yes = df2[df2.Churn=='Yes'].MonthlyCharges

plt1.xlabel("Monthly Charges")
plt1.ylabel("Number Of Customers")
plt1.title("Customer Churn Prediction Visualiztion")

bd_sug_men = [113, 85, 90, 150, 149, 88, 93, 115, 135, 80, 77, 82, 129]
bd_sug_women = [67, 98, 89, 120, 133, 150, 84, 69, 89, 79, 120, 112, 100]

plt1.hist([mc1_churn_yes, mc1_churn_no], rwidth=0.95,
color=['green','red'],label=['Churn=Yes','Churn=No'])
plt.legend() [17]
```

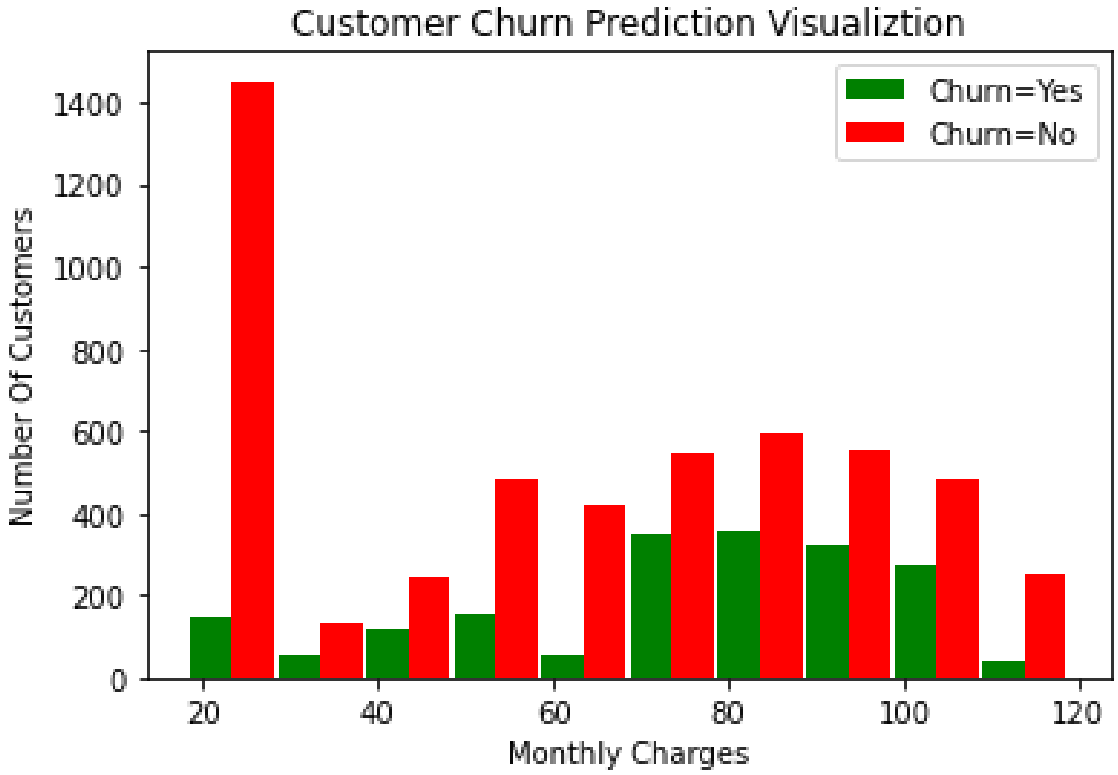

Рисунок 3.4 Візуалізація відтоку

def print_unique1_col_values(df2):

```
    for column in df2:
        if df2[column].dtypes=='object':
            print(f'{column}: {df2[column].unique()}')
```

In [144]:

```
print_unique_col_values(df3)
```

```
gender: ['Female' 'Male']
Partner: ['Yes' 'No']
Dependents: ['No' 'Yes']
PhoneService: ['No' 'Yes']
MultipleLines: ['No phone service' 'No' 'Yes']
InternetService: ['DSL' 'Fiber optic' 'No']
OnlineSecurity: ['No' 'Yes' 'No internet service']
OnlineBackup: ['Yes' 'No' 'No internet service']
DeviceProtection: ['No' 'Yes' 'No internet service']
TechSupport: ['No' 'Yes' 'No internet service']
StreamingTV: ['No' 'Yes' 'No internet service']
StreamingMovies: ['No' 'Yes' 'No internet service']
Contract: ['Month-to-month' 'One year' 'Two year']
PaperlessBilling: ['Yes' 'No']
PaymentMethod: ['Electronic check' 'Mailed check' 'Bank transfer (automatic)'
 'Credit card (automatic)']
Churn: ['No' 'Yes']
```

```
df3.replace('No internet service','No',inplace=True)
df1.replace('No phone service','No',inplace=True)
[12]
```

```
for col in df3:
```

```
    print(f'{colomn}: {df2[colomn].unique()}')
```

```
gender: ['Female' 'Male']
SeniorCitizen: [0 1]
Partner: [1 0]
Dependents: [0 1]
tenure: [ 1 34  2 45  8 22 10 28 62 13 16 58 49 25 69 52 71 21 12 30 47 72 17 27
  5 46 11 70 63 43 15 60 18 66  9  3 31 50 64 56  7 42 35 48 29 65 38 68
 32 55 37 36 41  6  4 33 67 23 57 61 14 20 53 40 59 24 44 19 54 51 26 39]
PhoneService: [0 1]
MultipleLines: [0 1]
InternetService: ['DSL' 'Fiber optic' 'No']
OnlineSecurity: [0 1]
OnlineBackup: [1 0]
DeviceProtection: [0 1]
TechSupport: [0 1]
StreamingTV: [0 1]
StreamingMovies: [0 1]
Contract: ['Month-to-month' 'One year' 'Two year']
PaperlessBilling: [1 0]
PaymentMethod: ['Electronic check' 'Mailed check' 'Bank transfer (automatic)'
 'Credit card (automatic)']
MonthlyCharges: [29.85 56.95 53.85 ... 63.1  44.2  78.7 ]
TotalCharges: [  29.85 1889.5   108.15 ...  346.45  306.6  6844.5 ]
Churn: [0 1]
```

3.8 Навчання за допомогою тестової вибірки

Машинне навчання (ML) відіграє ключову роль у сучасному світі, дозволяючи комп'ютерам вчити і покращувати свої здібності на основі даних. Однією з найважливіших складових ML є навчання моделей на вибірках даних. Тестова вибірка відіграє важливу роль в цьому процесі, і ця стаття розгляне процес навчання моделі машинного навчання з використанням тестової вибірки, його важливість та кращі практики.

- Розуміння ролі тестової вибірки

Тестова вибірка - це набір даних, який використовується для оцінки продуктивності навченої моделі. Вона розділяється на дві основні частини: навчальну та тестову вибірки. Навчальна вибірка використовується для навчання моделі, тобто для визначення внутрішніх параметрів моделі на основі доступних даних. Після навчання моделі тестова вибірка використовується для оцінки її продуктивності та здатності робити прогнози на нових даних, які вона раніше не бачила.

- Важливість тестової вибірки

Тестова вибірка є критично важливою, оскільки вона відображає реальну продуктивність моделі на нових, реальних даних. Якщо модель буде успішно працювати на тестовій вибірці, це свідчить про її здатність генералізувати та робити прогнози на невідомих даних. Однак, якщо модель показує відмінну продуктивність на навчальній вибірці, але погану на тестовій, це може свідчити про перенавчання - ситуацію, коли модель вивчає шум у даних, замість відображення справжніх залежностей.

- Поділ даних на навчальну та тестову вибірку

Першим кроком у використанні тестової вибірки є поділ доступних даних на навчальну та тестову вибірки. Зазвичай цей поділ відбувається випадковим чином, але важливо дотримуватися деяких правил. Наприклад, доречно визначити відсоток даних, які виділяються для тестової вибірки, зазвичай це 20-30% від загального обсягу даних.

- Навчання та оцінка моделі

Після поділу даних модель може бути навчена на навчальній вибірці. Це включає в себе підбір параметрів моделі, оптимізацію та навчання на основі доступних даних. Після завершення навчання модель може бути протестована на тестовій вибірці.

Оцінка моделі на тестовій вибірці проводиться за допомогою різних метрик продуктивності, таких як точність, відновлення, F-міра тощо, залежно від типу задачі машинного навчання. Ці метрики дають змогу визначити, наскільки добре модель працює на тестових даних.

- Перекресна перевірка (крос-валідація)

Крім тестової вибірки, існує ще один метод оцінки моделі, який називається крос-валідацією. Крос-валідація використовується для більш об'єктивної оцінки продуктивності моделі, особливо коли доступна велика кількість даних.

Перекресна перевірка (крос-валідація) - важливий метод навчання моделей машинного навчання. Ця техніка дозволяє об'єктивно оцінювати продуктивність моделі, поділяючи дані на кілька підвибірок для навчання та тестування. Крос-валідація допомагає уникнути перенавчання та забезпечити кращу генералізацію моделі на нові дані. Вона важлива при роботі з

обмеженими наборами даних та дозволяє збільшити достовірність результатів. Цей метод є стандартною практикою в машинному навчанні та сприяє покращенню надійності та ефективності моделей.

3.9 Навчання моделі за допомогою нейроної мережі

```
import tensorflow as tf
from tensorflow import keras



model = keras.Sequential([
    keras.layers.Dense(26, input_shape=(26,), activation='relu'),
    keras.layers.Dense(15, activation='relu'),
    keras.layers.Dense(1, activation='sigmoid')
])

# opt = keras.optimizers.Adam(learning_rate=0.01)

model.compile(optimizer='adam',
              loss='binary_crossentropy',
              metrics=['accuracy'])

model.fit(X_train, y_train, epochs=100)
```

```
Epoch 1/100
176/176 [==============================] - 0s 1ms/step - loss: 0.4822 - ac
curacy: 0.7623
Epoch 2/100
176/176 [==============================] - 0s 1ms/step - loss: 0.4269 - ac
curacy: 0.8000
```

Epoch 3/100

176/176 [==========] - 0s 1ms/step - loss: 0.4182 - ac

curacy: 0.7984

Epoch 4/100

176/176 [==========] - 0s 1ms/step - loss: 0.4153 - ac

curacy: 0.8046

Epoch 5/100

176/176 [==========] - 0s 1ms/step - loss: 0.4127 - ac

curacy: 0.8078

Epoch 6/100

176/176 [==========] - 0s 1ms/step - loss: 0.4108 - ac

curacy: 0.8073

Epoch 7/100

176/176 [==========] - 0s 1ms/step - loss: 0.4084 - ac

curacy: 0.8057

Epoch 8/100

176/176 [==========] - 0s 1ms/step - loss: 0.4070 - ac

curacy: 0.8108

Epoch 9/100

176/176 [==========] - 0s 1ms/step - loss: 0.4059 - ac

curacy: 0.8107

Epoch 10/100

176/176 [==========] - 0s 1ms/step - loss: 0.4043 - ac

curacy: 0.8107

Epoch 11/100

176/176 [==========] - 0s 1ms/step - loss: 0.4037 - ac

curacy: 0.8110

Epoch 12/100

176/176 [==========] - 0s 1ms/step - loss: 0.4020 - ac

curacy: 0.8114

```
Epoch 13/100
176/176 [==============================] - 0s 1ms/step - loss: 0.3996 - ac
curacy: 0.8128
Epoch 14/100
176/176 [==============================] - 0s 1ms/step - loss: 0.3992 - ac
curacy: 0.8132
Epoch 15/100
176/176 [==============================] - 0s 1ms/step - loss: 0.3982 - ac
curacy: 0.8119
Epoch 16/100
176/176 [==============================] - 0s 1ms/step - loss: 0.3973 - ac
curacy: 0.8105
Epoch 17/100
176/176 [==============================] - 0s 1ms/step - loss: 0.3955 - ac
curacy: 0.8128
Epoch 18/100
176/176 [==============================] - 0s 1ms/step - loss: 0.3939 - ac
curacy: 0.8126
Epoch 19/100
176/176 [==============================] - 0s 1ms/step - loss: 0.3936 - ac
curacy: 0.8149
Epoch 20/100
176/176 [==============================] - 0s 1ms/step - loss: 0.3930 - ac
curacy: 0.8155
Epoch 21/100
176/176 [==============================] - 0s 1ms/step - loss: 0.3920 - ac
curacy: 0.8151
Epoch 22/100
176/176 [==============================] - 0s 1ms/step - loss: 0.3912 - ac
curacy: 0.8148
```

Epoch 23/100

176/176 [==========] - 0s 1ms/step - loss: 0.3896 - ac

curacy: 0.8162

Epoch 24/100

176/176 [==========] - 0s 1ms/step - loss: 0.3897 - ac

curacy: 0.8162

Epoch 25/100

176/176 [==========] - 0s 1ms/step - loss: 0.3876 - ac

curacy: 0.8174

Epoch 26/100

176/176 [==========] - 0s 1ms/step - loss: 0.3864 - ac

curacy: 0.8187

Epoch 27/100

176/176 [==========] - 0s 1ms/step - loss: 0.3864 - ac

curacy: 0.8172

Epoch 28/100

176/176 [==========] - 0s 1ms/step - loss: 0.3846 - ac

curacy: 0.8181

Epoch 29/100

176/176 [==========] - 0s 1ms/step - loss: 0.3846 - ac

curacy: 0.8172

Epoch 30/100

176/176 [==========] - 0s 1ms/step - loss: 0.3834 - ac

curacy: 0.8187

Epoch 31/100

176/176 [==========] - 0s 1ms/step - loss: 0.3812 - ac

curacy: 0.8197

Epoch 32/100

176/176 [==========] - 0s 1ms/step - loss: 0.3815 - ac

curacy: 0.8180

Epoch 33/100

176/176 [==========] - 0s 1ms/step - loss: 0.3811 - ac

curacy: 0.8199

Epoch 34/100

176/176 [==========] - 0s 1ms/step - loss: 0.3806 - ac

curacy: 0.8178

Epoch 35/100

176/176 [==========] - 0s 1ms/step - loss: 0.3799 - ac

curacy: 0.8219

Epoch 36/100

176/176 [==========] - 0s 1ms/step - loss: 0.3787 - ac

curacy: 0.8185

Epoch 37/100

176/176 [==========] - 0s 1ms/step - loss: 0.3775 - ac

curacy: 0.8236

Epoch 38/100

176/176 [==========] - 0s 1ms/step - loss: 0.3783 - ac

curacy: 0.8212

Epoch 39/100

176/176 [==========] - 0s 1ms/step - loss: 0.3769 - ac

curacy: 0.8229

Epoch 40/100

176/176 [==========] - 0s 1ms/step - loss: 0.3760 - ac

curacy: 0.8224

Epoch 41/100

176/176 [==========] - 0s 1ms/step - loss: 0.3757 - ac

curacy: 0.8199

Epoch 42/100

176/176 [==========] - 0s 1ms/step - loss: 0.3749 - ac

curacy: 0.8260

Epoch 43/100

176/176 [==========] - 0s 1ms/step - loss: 0.3738 - ac

curacy: 0.8238

Epoch 44/100

176/176 [==========] - 0s 1ms/step - loss: 0.3727 - ac

curacy: 0.8228

Epoch 45/100

176/176 [==========] - 0s 1ms/step - loss: 0.3725 - ac

curacy: 0.8242

Epoch 46/100

176/176 [==========] - 0s 1ms/step - loss: 0.3722 - ac

curacy: 0.8245

Epoch 47/100

176/176 [==========] - 0s 1ms/step - loss: 0.3718 - ac

curacy: 0.8252

Epoch 48/100

176/176 [==========] - 0s 1ms/step - loss: 0.3716 - ac

curacy: 0.8244

Epoch 49/100

176/176 [==========] - 0s 1ms/step - loss: 0.3706 - ac

curacy: 0.8240

Epoch 50/100

176/176 [==========] - 0s 1ms/step - loss: 0.3703 - ac

curacy: 0.8224

Epoch 51/100

176/176 [==========] - 0s 1ms/step - loss: 0.3682 - ac

curacy: 0.8279

Epoch 52/100

176/176 [==========] - 0s 1ms/step - loss: 0.3695 - ac

curacy: 0.8233

Epoch 53/100

176/176 [==========] - 0s 1ms/step - loss: 0.3678 - ac

curacy: 0.8251

Epoch 54/100

176/176 [==========] - 0s 1ms/step - loss: 0.3671 - ac

curacy: 0.8261

Epoch 55/100

176/176 [==========] - 0s 1ms/step - loss: 0.3666 - ac

curacy: 0.8251

Epoch 56/100

176/176 [==========] - 0s 1ms/step - loss: 0.3656 - ac

curacy: 0.8251

Epoch 57/100

176/176 [==========] - 0s 1ms/step - loss: 0.3650 - ac

curacy: 0.8263

Epoch 58/100

176/176 [==========] - 0s 1ms/step - loss: 0.3643 - ac

curacy: 0.8268

Epoch 59/100

176/176 [==========] - 0s 1ms/step - loss: 0.3646 - ac

curacy: 0.8284

Epoch 60/100

176/176 [==========] - 0s 1ms/step - loss: 0.3641 - ac

curacy: 0.8252

Epoch 61/100

176/176 [==========] - 0s 1ms/step - loss: 0.3639 - ac

curacy: 0.8228

Epoch 62/100

176/176 [==========] - 0s 1ms/step - loss: 0.3630 - ac

curacy: 0.8299

Epoch 63/100

176/176 [==========] - 0s 1ms/step - loss: 0.3617 - ac

curacy: 0.8277

Epoch 64/100

176/176 [==========] - 0s 1ms/step - loss: 0.3622 - ac

curacy: 0.8284

Epoch 65/100

176/176 [==========] - 0s 1ms/step - loss: 0.3615 - ac

curacy: 0.8276

Epoch 66/100

176/176 [==========] - 0s 1ms/step - loss: 0.3623 - ac

curacy: 0.8263

Epoch 67/100

176/176 [==========] - 0s 1ms/step - loss: 0.3603 - ac

curacy: 0.8281

Epoch 68/100

176/176 [==========] - 0s 1ms/step - loss: 0.3600 - ac

curacy: 0.8284

Epoch 69/100

176/176 [==========] - 0s 1ms/step - loss: 0.3602 - ac

curacy: 0.8293

Epoch 70/100

176/176 [==========] - 0s 1ms/step - loss: 0.3596 - ac

curacy: 0.8288

Epoch 71/100

176/176 [==========] - 0s 1ms/step - loss: 0.3587 - ac

curacy: 0.8276

Epoch 72/100

176/176 [==========] - 0s 1ms/step - loss: 0.3585 - ac

curacy: 0.8290

Epoch 73/100

176/176 [==============================] - 0s 1ms/step - loss: 0.3581 - ac

curacy: 0.8277

Epoch 74/100

176/176 [==============================] - 0s 1ms/step - loss: 0.3582 - ac

curacy: 0.8311

Epoch 75/100

176/176 [==============================] - 0s 1ms/step - loss: 0.3573 - ac

curacy: 0.8272

Epoch 76/100

176/176 [==============================] - 0s 1ms/step - loss: 0.3575 - ac

curacy: 0.8277

Epoch 77/100

176/176 [==============================] - 0s 1ms/step - loss: 0.3573 - ac

curacy: 0.8306

Epoch 78/100

176/176 [==============================] - 0s 1ms/step - loss: 0.3564 - ac

curacy: 0.8288

Epoch 79/100

176/176 [==============================] - 0s 1ms/step - loss: 0.3550 - ac

curacy: 0.8313

Epoch 80/100

176/176 [==============================] - 0s 1ms/step - loss: 0.3550 - ac

curacy: 0.8324

Epoch 81/100

176/176 [==============================] - 0s 1ms/step - loss: 0.3548 - ac

curacy: 0.8284

Epoch 82/100

176/176 [==============================] - 0s 1ms/step - loss: 0.3552 - ac

curacy: 0.8329

Epoch 83/100

176/176 [==============================] - 0s 1ms/step - loss: 0.3556 - ac

curacy: 0.8279

Epoch 84/100

176/176 [==============================] - 0s 1ms/step - loss: 0.3534 - ac

curacy: 0.8331

Epoch 85/100

176/176 [==============================] - 0s 1ms/step - loss: 0.3533 - ac

curacy: 0.8299

Epoch 86/100

176/176 [==============================] - 0s 1ms/step - loss: 0.3536 - ac

curacy: 0.8332

Epoch 87/100

176/176 [==============================] - 0s 1ms/step - loss: 0.3536 - ac

curacy: 0.8325

Epoch 88/100

176/176 [==============================] - 0s 1ms/step - loss: 0.3505 - ac

curacy: 0.8356

Epoch 89/100

176/176 [==============================] - 0s 1ms/step - loss: 0.3517 - ac

curacy: 0.8311

Epoch 90/100

176/176 [==============================] - 0s 1ms/step - loss: 0.3513 - ac

curacy: 0.8313

Epoch 91/100

176/176 [==============================] - 0s 1ms/step - loss: 0.3525 - ac

curacy: 0.8309

Epoch 92/100

176/176 [==============================] - 0s 1ms/step - loss: 0.3495 - ac

curacy: 0.8334

```
Epoch 93/100
176/176 [==============================] - 0s 1ms/step - loss: 0.3506 - ac
curacy: 0.8270
Epoch 94/100
176/176 [==============================] - 0s 1ms/step - loss: 0.3495 - ac
curacy: 0.8350
Epoch 95/100
176/176 [==============================] - 0s 1ms/step - loss: 0.3497 - ac
curacy: 0.8327
Epoch 96/100
176/176 [==============================] - 0s 1ms/step - loss: 0.3500 - ac
curacy: 0.8338
Epoch 97/100
176/176 [==============================] - 0s 1ms/step - loss: 0.3484 - ac
curacy: 0.8343
Epoch 98/100
176/176 [==============================] - 0s 1ms/step - loss: 0.3504 - ac
curacy: 0.8325
Epoch 99/100
176/176 [==============================] - 0s 1ms/step - loss: 0.3490 - ac
curacy: 0.8325
Epoch 100/100
176/176 [==============================] - 0s 1ms/step - loss: 0.3486 - ac
curacy: 0.8368
```

Out[208]:

```
<tensorflow.python.keras.callbacks.History at 0x21818af0f10>
```

In [209]:

```
model.evaluate(X_test, y_test)
```

```
44/44 [==============================] - 0s 1ms/step - loss: 0.4932 - accur
acy: 0.7754
```

Out[209]:

```
[0.4931727349758148, 0.7754086852073669]
```

In [210]:

```
yp = model.predict(X_test)
yp[:5]
```

Out[210]:

```
array([[0.25819573],
       [0.4437274 ],
       [0.00808946],
       [0.7649808 ],
       [0.35091308]], dtype=float32)
```

In [213]:

```
y_pred = []
for element in yp:
    if element > 0.5:
        y_pred.append(1)
    else:
        y_pred.append(0)
```

In [218]:

```
y_pred[:10]
```

Out[218]:

```
[0, 0, 0, 1, 0, 1, 0, 0, 0, 0]
```

In [219]:

```
y_test[:10]
```

Out[219]:

```
2660    0
744     0
```

```
5579    1
64      1
3287    1
816     1
2670    0
5920    0
1023    0
6087    0
Name: Churn, dtype: int64
```

In [217]:

```
from sklearn.metrics import confusion_matrix , classification_report

print(classification_report(y_test,y_pred))
```

```
              precision    recall  f1-score   support

           0       0.83      0.86      0.85       999
           1       0.63      0.56      0.59       408

    accuracy                           0.78      1407
   macro avg       0.73      0.71      0.72      1407
weighted avg       0.77      0.78      0.77      1407
```

In [222]:

```
import seaborn as sn
cm = tf.math.confusion_matrix(labels=y_test,predictions=y_pred)

plt.figure(figsize = (10,7))
sn.heatmap(cm, annot=True, fmt='d')
plt.xlabel('Predicted')
```

```
plt.ylabel('Truth')
```

Out[222]:

```
Text(69.0, 0.5, 'Truth')
```

```
y_test.shape
```

Out[224]:

```
(1407,)
```

Точність

In [235]:

```
round((862+229)/(862+229+137+179),2)
```

Out[235]:

0.78

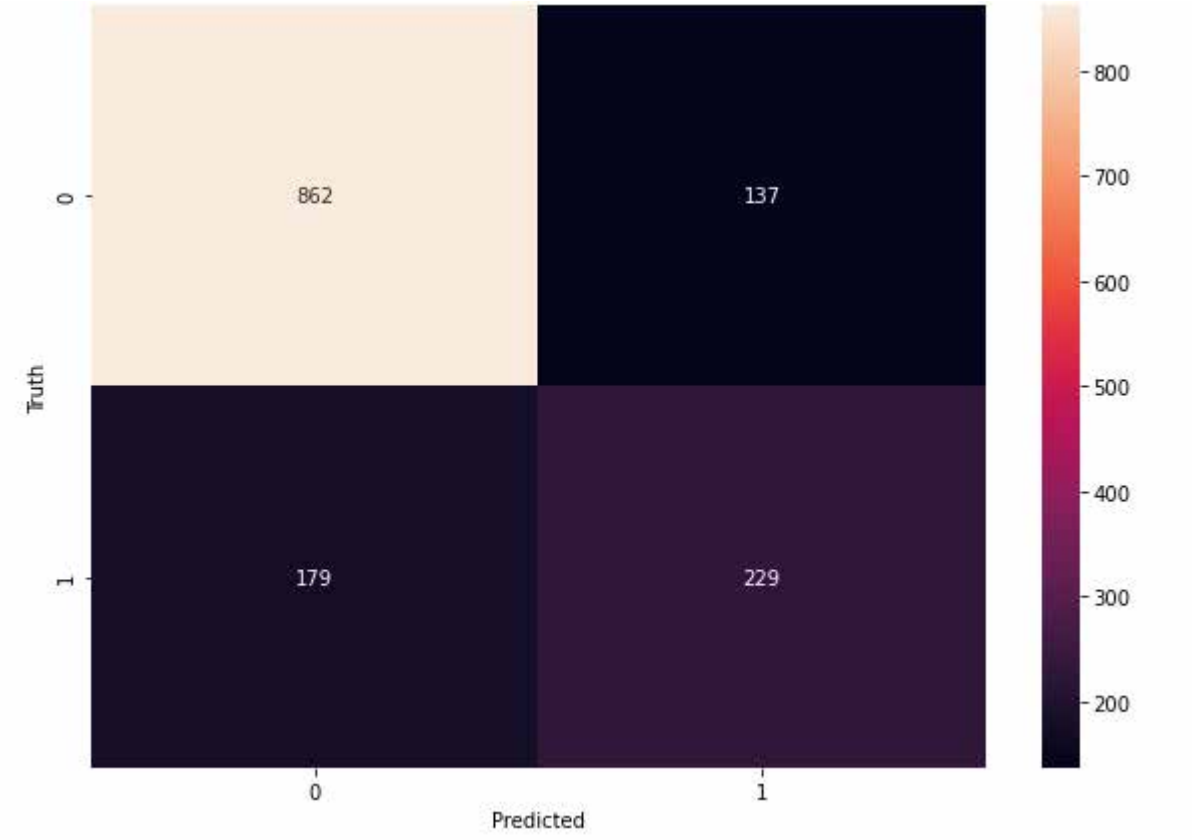

Рисунок 3.5 Візуалізація відтоку на основі моделі нейроної мережі

3.10 Висновки методу манинного навчання для відтоку клієнтів на основі моделі нейроної мережі

Вивчення відтоку клієнтів з використанням моделей нейронних мереж може призвести до різних висновків, які допоможуть покращити стратегію утримання клієнтів та зменшити втрати. Ось деякі можливі висновки:

Важливі фактори відтоку: Модель нейронної мережі допоможе визначити, які конкретні чинники або зразки поведінки клієнтів найбільше сприяють відтоку. Це може включати такі фактори, як частота використання послуг, зв'язок із службою підтримки, знижки тощо.

Сегментація клієнтів: Модель може допомогти виокремити різні сегменти клієнтів, які схильні до відтоку. Наприклад, ви можете визначити, що нові клієнти, які приєднуються на короткий термін, схильні до відтоку, або що клієнти певних географічних областей схильні до відтоку.

Рекомендації для утримання клієнтів: На основі вивчених зразків, модель може надати рекомендації щодо того, які заходи можна приймати для утримання клієнтів. Наприклад, ви можете надавати знижки або персоналізовані послуги тим клієнтам, які входять до групи ризику відтоку.

Моніторинг ефективності: Після впровадження рекомендацій моделі, важливо відстежувати їх ефективність. Модель може допомогти вам оцінити, чи зменшилися рівні відтоку серед тих клієнтів, які були включені до програми утримання.

Оптимізація ресурсів: Модель також може допомогти оптимізувати витрати на утримання клієнтів, спрямовуючи ресурси на тих клієнтів, які мають найбільший потенціал для збереження.

Постійне навчання: Важливо розуміти, що відток клієнтів - це динамічний процес, і модель потребує постійного навчання і поновлення для того, щоб враховувати зміни в поведінці клієнтів і нові фактори впливу.

Використання моделей нейронних мереж для аналізу відтоку клієнтів може бути потужним інструментом для покращення стратегії утримання клієнтів і збільшення прибутку компанії. Тим не менше, важливо дотримуватися етичних стандартів та захищати приватність клієнтів під час обробки їхніх даних.

3.11 Прогнозування відтоку клієнтів за допомогою логістичної регресії

Вибираючи постачальника телекомунікаційних послуг, клієнти зазвичай мають багато варіантів. Вони можуть вибрати будь-якого постачальника послуг і можуть відійти від поточного постачальника. Коли клієнт вирішує перейти від поточного постачальника до нового, це призводить до втрати бізнесу та доходу поточного постачальника. Відсоток клієнтів, які вийшли з мережі та відключили послугу, відомий як «відтік». Стабільна клієнтська база – запорука успіху будь-якого бізнесу. Підприємства намагаються залишати клієнтів задоволеними, утримувати їх якомога довше. Однак у реальному світі відтік клієнтів у телекомунікаційній галузі може досягати 25% на рік. Крім того, вартість залучення нового клієнта в 10 разів більша, ніж вартість утримання існуючого клієнта. Це створює серйозну проблему для власників бізнесу.

Аналіз даних про відтік клієнтів може допомогти компанії зрозуміти основні причини, чому клієнти можуть вирішити залишити компанію. Впроваджуючи методи прогнозної аналітики та застосовуючи їх до існуючих даних про відтік клієнтів із записів, можна зрозуміти ймовірність того, що клієнти змінять або припинять послугу. Потім можна працювати з клієнтами з

високою ймовірністю переходу, щоб переконатися, що вони залишаються з поточним постачальником.

Дані, використані з цього аналізу, доступні на Kaggle. Очищена копія доступна за посиланням GitHub. Він містить кілька різних частин інформації про клієнтів. Використовуючи ці дані, можна створити прогнозну модель відтоку клієнтів. І потім це можна використовувати, щоб зрозуміти, над якими клієнтами слід працювати, щоб утримати.

Загальну проблему можна підсумувати так:

1. Створення прогнозної моделі з використанням доступних даних про відтік клієнтів, щоб передбачити та знайти клієнтів, які, ймовірно, припинять послугу.

2. Остаточним прогнозованим результатом для будь-якого конкретного клієнта має бути «Так» або «Ні» (двійковий вихід)

На основі результатів прогнозів компанія може вибрати відповідні дії у вигляді різних стратегій утримання клієнтів і зменшення відтоку клієнтів. У наступному розділі ми дізнаємося, який метод можна використовувати для моделювання сценарію для прогнозування бінарного результату. Ми також побачимо, як працює прогностична модель і за яких припущень її слід застосовувати.[16]

Виходячи з постановки проблеми, нам потрібна прогностична модель, яка може виконати двійкову класифікацію або передбачити тип вихідної змінної «Так/Ні» або 1/0. Однією прогностичною моделлю, яка зазвичай реалізується для бінарної класифікації та передбачення бінарного результату, є логістична регресія. Логістична регресія – це алгоритм двійкової класифікації,

що належить до моделі узагальненої лінійної регресії. Його також можна використовувати для розв’язування задач з більш ніж 2 класами. Можна використати логістичну регресію, щоб створити модель, використовуючи дані про відтік клієнтів, і використовувати її для прогнозування, чи припинить окремий клієнт із групи клієнтів послугу.[20]

Наприклад, однією зі змінних у даних може бути «річний дохід». Іншою змінною є «стать» клієнта. Результат функції логістичної регресії покаже нам, як дохід і/або стать визначають ймовірність припинення обслуговування клієнтом.

Рівняння логістичної регресії та сигмоїдна функція

Нижче наведено функцію логістичної регресії:

$$P(Y=1|X) = 1 / (1 + e^{(-1x(\beta0+\beta1X1+\beta2X2+\beta3X3+\beta4X4+\cdots+\beta nXn))})$$

де:
від β0 до βn різні коефіцієнти
від X0 до Xn є незалежними змінними, що впливають на залежну змінну,
а P (Y =1 | X) є ймовірністю позитивного результату.

Зверніть увагу на показник степеня у функції. Ось тут грає лінійна регресія $(\beta0+\beta1X1+\beta2X2+\beta3X3+\beta4X4+\cdots+\beta nXn)$.

Функція логістичної регресії є сигмоподібною функцією (графік вище). Як видно на графіку, функція має вихідні значення від 0 до 1 із переходом між рівнями. Ця характеристика функції допомагає передбачати двійкові результати. На основі значення змінних результат може бути на рівні 1 або 0,

що відповідає ймовірності того, що клієнт покине компанію або продовжує працювати в ній.

Для логістичної регресії зроблено наступні припущення:

- Двійкова логістична регресія вимагає, щоб залежна змінна була двійковою та відповідала біноміальному розподілу (наприклад, припинить клієнт послугу чи ні, Так чи Ні). Для більш ніж 2 результатів (порядкових) логістична регресія вимагає, щоб категорії залежних змінних були взаємовиключними та вичерпними.

- Спостереження мають бути незалежними одне від одного (наприклад, дані одного клієнта не повинні залежати від даних іншого клієнта, або той самий клієнт не повинен використовуватися повторно в даних)

- Мультиколінеарність між незалежними змінними не повинна існувати (наприклад, уникайте використання змінних із даних клієнта, які залежать одна від одної, скажімо, місто, штат, округ та поштовий індекс не всі незалежні)

- Лінійність незалежних змінних щодо логарифму шансів залежної змінної (наприклад, логарифм шансів ймовірності припинення клієнтом послуги має бути лінійно пов’язаний з різними змінними, такими як стать, дохід тощо)

- Великий розмір вибірки (наприклад, дані про відтік клієнтів містять 10 000 записів)

Підготовка даних починається з розуміння наявних даних для аналізу. Дані про відтік клієнтів містять 50 полів. Одним із важливих завдань є визначення полів, які можна використовувати для регресійного аналізу. Існують поля категорійних даних, як-от військовий, стать тощо, і безперервні поля числових даних, як-от перебування, вік тощо. Деякі з цих полів можуть бути неважливими для аналізу, наприклад ідентифікатор клієнта, взаємодія та UID (які пов'язані з клієнтом). сервісні взаємодії). Деякі інші поля, які не є важливими для аналізу, це широта та довгота клієнта, порядок випадків (використовується як порядковий номер). Ми також перевіримо дані на наявність нульових значень, і якщо вони будуть знайдені, їх потрібно буде обробити належним чином.

Категоричні поля даних, як-от стать, Інтернет-послуги, телефонні послуги тощо, потрібно буде перетворити на відповідні стовпці, що містять 0 або 1. Категоричні дані, як-от Так/Ні, добре/погано/потворно, не можна використовувати для математичних операцій . Отже, ці стовпці, що містять категоричні дані, зрештою стануть стовпцями з 0 або 1 записами.

Для безперервних числових даних масштаб даних для кожного стовпця різний. Наприклад, стовпець віку містить значення в діапазоні від 10 до 89, а стовпець зарплати містить цифри в 10 або 100 тисяч. Перш ніж ми зможемо використовувати ці дані для аналізу, їх потрібно нормалізувати, щоб дані були зосереджені навколо середнього значення та вимірювалися з точки зору їх відхилення від середнього. Це важливо для чисельної стабільності моделі.

3.12 Початкова модель

Як перший крок, щоб перевірити вплив, важливість і значимість різних стовпців даних відносно аналізу відтоку, буде створено початкову модель, яка міститиме всі змінні в наборі даних. Ми підготували дані на попередньому кроці таким чином, щоб включити всі змінні на даний момент.

Початкова модель надасть інформацію про те, які змінні важливі для прогнозного аналізу. Виходячи зі значущості змінних, ми повинні мати можливість виключити деякі змінні з набору даних, щоб отримати скорочену модель. Потім ми перевіримо оцінку точності, матрицю плутанини та AUC для 2 моделей, щоб порівняти їх продуктивність.

Почнемо з імпорту необхідних бібліотек. Ми використовуємо такі бібліотеки, як sklearn і statsmodels, щоб створювати моделі, щоб перевірити, наскільки добре працює прогностична модель і наскільки точно ми можемо передбачити результати або відтік клієнтів. Оскільки ми маємо велику кількість рядків у наборі даних, я вибираю, що 70% даних будуть у навчальному наборі, а 30% даних – у тестовому наборі.[19]

3.13 Аналіз вихідної моделі

У початковому підсумку моделі, створеному за допомогою моделі logit statsmodels.api, ми можемо побачити p-значення різних незалежних змінних. Використовуючи ці p-значення, ми можемо визначити, які змінні є значущими для прогнозного моделювання. Це допоможе визначити список змінних, які можна безпечно видалити без будь-якого впливу на загальну точність і матрицю помилок.

Початкова модель, що містить (містить повний набір змінних), має загальний рівень точності прогнозування 89 %. Матриця плутанини показує, що можна точно передбачити 92,7 % (специфічність) продовження обслуговування та 79,1 % (чутливість) припинення обслуговування.

Щоб отримати скорочену модель, ми можемо видалити змінні, які мають вищі p-значення, і використовувати лише ті змінні, які мають нижчі p-значення. Як правило, поріг 0,25 можна використовувати для порівняння p-значення. Це дає нам список незалежних змінних для використання у зменшеній моделі.

3.14 Зменшена модель

Скорочена модель матиме наступні незалежні змінні. Інші змінні були вилучені на основі їх низької значущості та p-значень.

«Churn», «Tenure», «Contacts», «MonthlyCharge», «Bandwidth_GB_Year», «State», «Marital», «Gender», «Techie», «Contract», «Age», «Children», «Email». ', 'Port_modem','InternetService','Phone','Timely_response','Timely_replacements', 'StreamingMovies','PaperlessBilling','PaymentMethod'

Маючи знання про значущі змінні, ми вилучили менш значущі змінні. Тепер ми створимо набір даних лише зі значущими змінними та запустимо аналіз.

3.15 Остаточний скорочений набір даних

На основі тесту на відношення правдоподібності, виконаного в коді R вище, було визначено, що стовпець State не має значення для моделі. Значення P при порівнянні зменшеної моделі та моделі з вилученим із неї стовпцем стану становило 0,4081. Ці значення можна побачити вище, у вихідних даних коду R. Оскільки p-значення високе, що означає неспроможність відхилити нульову гіпотезу, стовпець не має значення. Нульова гіпотеза для перевірки відношення правдоподібності стверджує, що коефіцієнт для стовпця State дорівнює 0. Подібний аналіз проводився для інших стовпців Timely_replacements, Timley_response, Multiple & Gender. Вони були визнані незначними. Це підводить нас до нашої остаточної зменшеної моделі. Остаточний список змінних у скороченому наборі даних наведено нижче:

Термін перебування, контакти
MonthlyCharge, Bandwidth_GB_Year
Marital, Techie
Contract, Age
Children, Port_modem
InternetService, Phone
StreamingMovies, PaperlessBilling, PaymentMethod

3.16 Порівняння моделей

Початкова та остаточна зменшені моделі не сильно відрізняються, коли йдеться про загальну продуктивність і те, наскільки точно вони передбачають результат. Однак існує велика різниця в кількості змінних, і, отже, зменшена модель є економічною, швидшою та менш ресурсомісткою. Нижче наведено порівняння продуктивності початкової та кінцевої моделей.[18]

Performance comparison of the Initial and Final model

Comparison	Initial Model	Final Reduced Model
Confusion Matrix	2040 161 167 632	2049 152 160 639
Accuracy Score	0.89067	0.896
Specificity	92.7 %	93.01 %
Sensitivity	79.1 %	79.97 %
AUC	0.955	0.954
Total independent variables	38	15
Total variables (with dummy)	97	22

Рисунок 3.6 Порівняння моделей

Змінні, які були визначені як незначущі та, отже, вилучені з прогнозного аналізу:

'State', 'Population', 'Area', 'Income', 'Gender', 'Outage_sec_perweek', 'Email', 'Yearly_equip_failure', 'Tablet', 'Multiple', 'OnlineSecurity', 'OnlineBackup', 'DeviceProtection ', 'TechSupport', 'StreamingTV', 'Timely_response', 'Timely_fixes', 'Timely_replacements', 'Reliability', 'Options', 'Respectful_response', 'Courteous_exchange', 'Evidence_of_active_listening'

3.17 **Резюме моделі**

Ми створили досить точну прогностичну модель із приблизно 89% точністю прогнозування поведінки клієнтів. Модель може визначити, чи планує клієнт відключитися, приблизно з 80% точністю. Оцінка AUC остаточної моделі становить 0,954, що дуже близько до ідеальної 1. Завдяки цій моделі компанія тепер може мати чітке уявлення про те, які клієнти можуть припинити надання послуг. Модель повідомляє нам про змінні та їхній відповідний вплив на відтік.

3.18 AUC або площа під кривою

Площа під кривою показує нам, наскільки добре наша модель прогнозує результат. Якщо площа під кривою дорівнює 0,5, наша модель нічим не відрізняється від випадкового припущення з 50% шансом правильно передбачити результат (двійковий). Коли ми наближаємося до значень AUC, близьких до 1, ми знаємо, що наш алгоритм стає кращим у прогнозах. При ідеальному 1 наша модель може передбачити результат зі 100% упевненістю.[14]

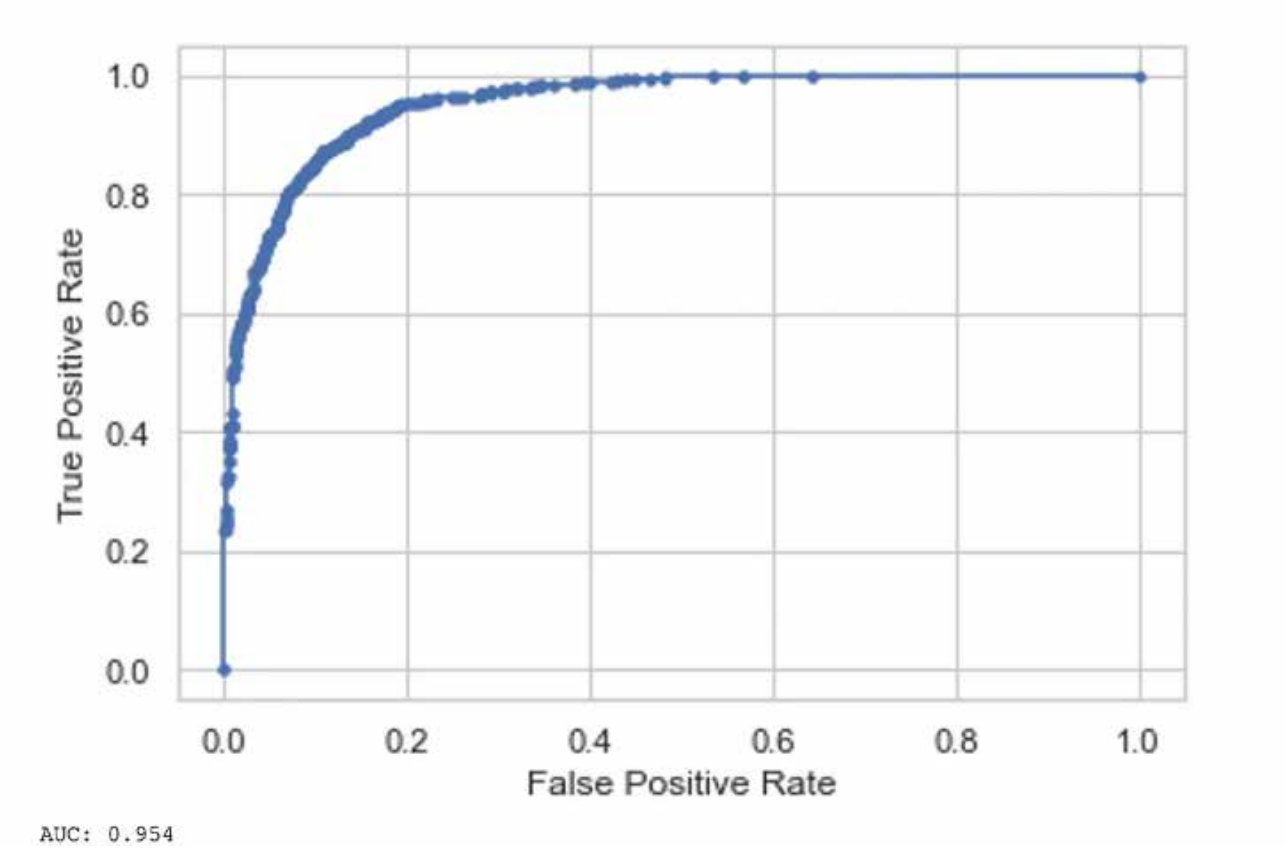

Рисунок 3.7 Діаграма AUC

3.19 Рівняння логістичної регресії

Рівняння логістичної регресії має вигляд:

$$P(Y=1|X) = \exp(y) / 1 + \exp(-y)$$

де $y = \beta 0 + \beta 1X1 + \beta 2X2 + \beta 3X3 + \ldots \beta nXn$

де n – кількість змінних у моделі, а βn – значення n-го коефіцієнта.

У моделі логістичної регресії, яку ми створили вище, значення n дорівнює 26, включаючи всі фіктивні змінні. Для візуальної розумності коефіцієнти буде округлено до 2 цифр після коми, щоб зробити загальне рівняння читабельним.

Рівняння з необхідними змінами, щоб зробити його читабельним:

P (Y=1|X) = 1 / (1 + exp(-1x(-3,15 x Термін перебування +0,09 x Контакти +2,28 x MonthlyCharge +0,19 x Bandwidth_GB_Year -0,02 x Age +0,0 x Children +1,04 x Techie -3,24 x Contract_One_year -3,33 x Contract_Two_Year +0,05 x Marital_Married +0,08 x Marital_Never_Married +0,13 x Marital_Separed +0,22 x Marital_widowed +0,19 x Port_modem -2,2 x InternetService_Fiber_Optic -0,62 x InternetService_None -0,37 x Phone +0 0,44 x StreamingMovies +0,15 x PaperlessBilling +0,25 x PaymentMethod_Credit_Card_automatic + 0,62 x PaymentMethod_Electronic_Check +0,22 x PaymentMethod_Mailed_Check -1,09)))

3.20 Інтерпретація коефіцієнтів

Індивідуальні коефіцієнти можна інтерпретувати як зміну співвідношення шансів для кожної одиничної зміни змінної. Завдяки нормалізації, виконаній для безперервної числової змінної, зміна одиниці таких змінних дорівнює їх стандартному відхиленню. Для категоріальних змінних зміною є перехід від 0 до 1.

Аналізуючи коефіцієнти, ми можемо побачити, як певна змінна може вплинути на відтік. Давайте візьмемо кілька прикладів і розберемося детально:

Для змінної володіння ми бачимо негативний коефіцієнт (-3,15 x володіння). Якщо користувач провів у компанії тривалий час (більше середнього

або середнього), він, швидше за все, припинить роботу, але якщо він провів у компанії менше часу, він, швидше за все, припинить роботу. Середній термін перебування на посаді становить 34,53 місяця. Стаж має стандартне відхилення 26,44 місяців. Для клієнта з кожним додатковим місяцем, проведеним у компанії, коефіцієнт шансів залишити компанію знижується на 11,2% (1-exp(-3,15/26,44). Для змінної MonthlyCharge коефіцієнт становить 2,28. З кожним доларом збільшення щомісячної плати підвищується ймовірність того, що клієнт залишить послугу. І навпаки, із кожним зменшенням місячної плати клієнт має більше шансів залишитися. Зі зменшенням щомісячної плати в доларах коефіцієнт шансів зменшується на 8,26%. Зі знижкою в 5 доларів на місяць коефіцієнт шансів знижується на 35%. Зі знижкою 10 доларів на місяць коефіцієнт шансів зменшується на 57,78%.[13]

Клієнти, які переходять із послуги DSL (за замовчуванням) на послугу оптоволоконного Інтернету, зменшать коефіцієнт шансів на 88,92%.

Клієнти, які переходять з місячного контракту (за замовчуванням) на однорічний контракт, зменшать коефіцієнт шансів на 96,08%. А ті, хто переходить на дворічний контракт, знижують коефіцієнт шансів на 96,42%.

Розділ 4
ДІЇ ЯКІ МОЖНА ВЖИТИБ ЩОБ ЗМЕНШИТИ ВІДТІК КЛІЄНГТІВ ІНТЕРНЕТ ПРОВАЙДЕРА

Завдяки наявності цієї моделі можна передбачити, чи ймовірно клієнт припинить надання послуги. Однак компанії також потрібно будувати плани та реалізовувати їх, щоб зменшити відтік працівників. Давайте детально перевіримо кілька фактів, показаних моделлю, і обговоримо, які можливі дії компанія може вжити, щоб збільшити утримання клієнтів і зменшити відтік.

a. Клієнти, які платять високі щомісячні платежі, швидше за все, припинять послугу. Компанія може запропонувати кращі пропозиції та знижки клієнтам, які, ймовірно, припинять послугу, використовуючи прогнозну модель. Виходячи з інтерпретації коефіцієнтів, надання знижок є відмінним способом знизити ймовірність втрати клієнта. Щомісячна знижка в розмірі п’яти доларів може знизити коефіцієнт шансів на 35%. Середній місячний платіж становить $172. Багато клієнтів платять більше ніж на одне стандартне відхилення вище середнього (214+ доларів, 1800 таких клієнтів у наборі даних). А деякі клієнти сплачують понад 2 стандартні відхилення від середнього ($256+ на місяць, 280 таких клієнтів у наборі даних). Ці клієнти мають високий ризик залишити постачальника послуг.[17]

b. Відповідно до аналізу, проведеного в розділі інтерпретації коефіцієнтів, якщо клієнт підписує 2-річний контракт, коефіцієнт шансів знижується на 96,42%. Якщо клієнт підписує 1-річний контракт, то коефіцієнт шансів знижується на 96,08%. Якщо компанія зможе працювати над переведенням щомісячних клієнтів на одно- або дворічні контракти, ймовірність припинення клієнтом

послуги буде значно зменшена. Це може стати величезним позитивним впливом на зниження рівня відтоку, оскільки в наборі даних є 5456 клієнтів із щомісячними контрактами (що становить понад 50% клієнтів у наборі даних).

c. Споживачі Інтернет-послуг, які мають підключення DSL, швидше за все, припинять. Можуть бути конкуренти, які надають кращі послуги, ніж DSL-інтернет, і клієнти відмовляться від послуг кращої якості з іншим постачальником. Компанія може перевірити, чи може вона запропонувати альтернативні послуги DSL і утримати цих клієнтів. Цілком ймовірно, що в певних регіонах компанія пропонує лише DSL-послуги Інтернету без наявності альтернатив. Компанії слід планувати перехід існуючих клієнтів DSL на оптоволокно або іншу кращу інтернет-технологію. Як видно з інтерпретації розділу коефіцієнтів, оновлення до опції волокна зменшує співвідношення шансів на 88,92%. Дивно, але клієнти, які не мають Інтернет-послуги, мають меншу ймовірність припинення послуги, ніж клієнти з DSL-підключенням до Інтернету. Якщо хтось переходить з DSL на відсутність Інтернету, коефіцієнт шансів знижується на 46,21%.

d. Клієнти, які є новими та не проводили багато часу з компанією, швидше за все, припинять послугу. Можливо, компанія пропонує хороші пропозиції протягом першого року обслуговування, а потім підвищує ціни. Дані показують, що клієнти, які припиняють послугу, мають середній термін близько 13,5 місяців. Можливе закінчення терміну дії пропозиції або знижки наприкінці першого року обслуговування та підвищення вартості обслуговування після першого року може бути причиною виходу клієнтів. Компанія може активно працювати над утриманням тих клієнтів, які закінчують перший рік, надсилаючи їм нові пропозиції або розширюючи для них

знижки. Як видно з інтерпретації розділу коефіцієнтів, з кожним додатковим місяцем служби коефіцієнт шансів на припинення знижується на 11,23%. Утримання клієнтів у компанії довше може допомогти зменшити відтік.

e. Додавання телефонної лінії може зменшити коефіцієнт шансів на 30,93%. Компанії слід спробувати заохотити більше клієнтів підписатися на телефонні послуги, якщо це можливо, оскільки це збільшує утримання клієнтів.

f. Клієнти потокового передавання мають високий рівень відтоку. Важливо зрозуміти, що є причиною відтоку цих клієнтів. Можливо, вони не задоволені якістю послуг. Загалом для потокового передавання потрібне високошвидкісне з'єднання без перепадів з'єднання. Необхідно провести подальший аналіз, щоб зрозуміти проблеми з клієнтами потокового передавання. Знаючи, що клієнти потокового передавання, швидше за все, припинять, компанії слід активно працювати з клієнтами, щоб усунути причину їхнього незадоволення та, як наслідок, відтоку.

g. Рівень відтоку клієнтів із «технарями» високий. Можна провести подальші розслідування, щоб зрозуміти, що цим технічно підкованим клієнтам не подобається в загальному обслуговуванні. Це може бути пов'язано з важкою у використанні технологією або погано розробленою технологією або, можливо, з меншою свободою керування, налаштування чи зміни параметрів служби, що зменшує технічну привабливість, або з низькою продуктивністю веб-сайту, або з будь-якої іншої можливої причини, яку технічно підковані клієнти не бачать. люблю. Завдяки проактивним крокам і змінам компанія може краще утримувати цих технічно підкованих клієнтів.

Отже, був проведений всебічний аналіз даних про відтік клієнтів за допомогою логістичної регресії, щоб отримати уявлення про поведінку клієнтів.

ВИСНОВКИ ПО РОБОТІ

Машинне навчання - це сфера, яка стрімко розвивається і знайшла широкий застосування в багатьох галузях. Завдяки постійному зростанню обчислювальних можливостей та доступності даних, машинне навчання стало невід'ємною частиною сучасного світу. Два із найпопулярніших методів машинного навчання - це нейронні мережі і логістична регресія. У цьому есе ми проведемо порівняння цих двох методів з точки зору їхніх особливостей, застосувань і відмінностей.

Нейронні мережі - це клас моделей машинного навчання, який був надзвичайно популярним у останні роки завдяки своїй здатності розрізняти складні залежності в даних. Вони інспіруються нейронною системою людини і складаються з шарів нейронів, які обробляють інформацію та передають її далі. Нейронні мережі здатні автоматично визначати ваги для кожного з'єднання між нейронами під час процесу навчання. Ця здатність робить їх ефективними в розв'язанні різноманітних завдань, таких як розпізнавання образів, машинний переклад, обробка природних мов, а також велика кількість завдань у сферах медицини, фінансів та технологій.

Логістична регресія, з іншого боку, є статистичним методом, який використовується для моделювання ймовірностей категорійних результатів. Вона базується на функції логітіс, яка приймає значення від 0 до 1 і використовується для передбачення ймовірностей. Логістична регресія є однією з найпростіших форм машинного навчання, але в той же час вона може бути дуже ефективною для багатьох завдань класифікації, де важлива точність передбачень.

Однією з ключових відмінностей між нейронними мережами та логістичною регресією є їх архітектура та вміст даних, з якими вони працюють. Нейронні мережі складаються зі складних структур шарів нейронів, а їхнє навчання вимагає великої кількості даних та обчислювальних ресурсів. Логістична регресія, навпаки, використовується для бінарної класифікації та її модель може бути побудована на основі невеликої кількості признаків.

Ще однією важливою відмінністю є процес навчання. У нейронних мережах навчання відбувається шляхом встановлення ваг для кожного з'єднання між нейронами. Цей процес може бути великою перевагою у вирішенні складних завдань, але водночас вимагає багато даних і часу на навчання. Логістична регресія, навпаки, навчається швидше, оскільки ваги встановлюються шляхом оптимізації функції втрат за допомогою методів, таких як градієнтний спуск.

З іншого боку, нейронні мережі можуть бути більш ефективними у вирішенні завдань, які вимагають аналізу великих обсягів даних або розпізнавання складних залежностей.

Найкращим вибором буде суміщення методу логістичної регресії та нейроної мережі.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. https://nkrzi.gov.ua/index.php?r=site/index&pg=99&id=2780&language=uk;
2. https://www.it.ua/knowledge-base/technology-innovation/big-data-bolshie-dannye;
3. https://ela.kpi.ua/bitstream/123456789/22361/1/EV2017_307-314.pdf;
4. https://hub.kyivstar.ua/articles/kyyivstar-u-2-kvartali-2023-roku-zbilshyv-investycziyi-u-telekom-merezhu;
5. УДК 340.12:004.738.5:342.7;
6. https://hub.kyivstar.ua/news/majbutne-big-data-industrii-yak-stati-zatrebuvanim-fahivczem/;
7. https://colab.research.google.com/?hl=ru;
8. https://pandas.pydata.org/;
9. https://matplotlib.org/;
10. https://numpy.org/;
11. https://www.kaggle.com/datasets/blastchar/telco-customer-chu;
12. https://github.com/codebasics/deep-learning-keras-tf-tutorial/blob/master/11_chrun_prediction/churn.ipynb;
13. https://medium.com/data-science-on-customer-churn-data/customer-churn-data-analysis-using-logistic-regression-3861e2d4d1f3;
14. Practical Statistics for Data Scientists by Peter Bruce, Andrew Bruce, and Peter Gedeck Copyright © 2020 Peter Bruce, Andrew Bruce, and Peter Gedeck;
15. Что такое отток клиентов и как с ним бороться. URL: https://ngmsys.com/blog/churn-management;
16. Логистическая регрессия. URL: https://loginom.ru/blog/logistic-regression-roc-auc;
17. 40 алгоритмов, которые должен знать каждый программист на Python. — СПб.: Питер, 2023. — 368 с.: ил. — (Серия «Библиотека программиста»).
18. Big data, data mining, and machine learning : value creation for business leaders and practitioners / Jared Dean.;
19. Машинное обучение. Патгерны проектирования: Пер. с англ./ В. Лакшманан, С. Робинсон, М. Мунн. - СПб.: БХВ-Петербург, 2022. - 448 с.: ил.
20. Eremenko K. Data Science A-Z: Real-Life Data Science Exercises Included. URL: https://www.udemy.com/course/datascience.;
21.