НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ
імені ІГОРЯ СІКОРСЬКОГО»

Навчально-науковий інститут телекомунікаційних систем

Кафедра телекомунікацій

«На правах рукопису»

УДК __ 004.021 ____

«До захисту допущено»
Завідувач кафедри
_____ Сергій КРАВЧУК
«___»_____2021 р.

Магістерська дисертація

на здобуття ступеня магістра
за освітньо-професійною програмою «Інженерія та програмування інфокомунікацій»
зі спеціальності 172 «Телекомунікації та радіотехніка»

на тему: «Використання алгоритмів машинного навчання для передбачення відтоку абонентів оператора мобільного зв'язку»

Виконав:
студент VI курсу, групи ТЗ-01мп
Раченчук Іван Геннадійович _____

Керівник:
директор НН ІТС,
д.т.н., професор кафедри ТК, Ільченко М. Ю. _____

Консультант з розділу:
Доцент кафедри ТК, к.т.н.
Міночкін Д.А. _____

Рецензент:
Професор кафедри ІКТС НН ІТС, д.т.н, с.н.с.,
Скулиш М.А. _____

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

Київ – 2021 року

Національний технічний університет України

«Київський політехнічний інститут імені Ігоря Сікорського»

Інститут телекомунікаційних систем
Кафедра телекомунікацій

Рівень вищої освіти – другий (магістерський)

Спеціальність – 172 «Телекомунікації та радіотехніка»

Освітньо-професійна програма «Інженерія та програмування інфокомунікацій»

ЗАТВЕРДЖУЮ

Завідувач кафедри

_____ Сергій КРАВЧУК

«___»_____2021 р.

ЗАВДАННЯ
на магістерську дисертацію студенту

Раченчуку Івану Геннадійовичу

1. Тема дисертації «Використання алгоритмів машинного навчання для передбачення відтоку абонентів оператора мобільного зв'язку», науковий керівник директор ІТС, академік НАНУ, доктор технічних наук, професор, Ільченко Михайло Юхимович, затверджені наказом по університету від «04» листопада 2021 р. № 3672-с.

2. Термін подання студентом дисертації 10.12.2021 р.

3. Об’єкт дослідження: надання вірних рішень щодо класифікування абонентів

4. Предмет дослідження: методи, алгоритми та моделі машинного навчання, XGBoost

5. Перелік завдань, які потрібно розробити:

1. Аналіз методів та алгоритмів машинного навчання. Бустинг XGBOOST
2. Відтік абонентів у сфері телекомунікацій
3. Використання машинного навчання для передбачення відтоку абонентів оператора мобільного зв’язку
4. Розробка стартап-проекту.

6. Орієнтовний перелік ілюстративного матеріалу
Слайд №1 Назва;
Слайд №2 Вступ. Актуальність та мета;
Слайд №3-5 Загальні відомості про аналіз даних та машинне навчання. Огляд задач;
Слайд №6-7 Аналіз відтоку абонентів. Основні причини та методи запобігання;
Слайд №8-9 Проведення аналізу практичних моделей;
Слайд №10 Результати аналізу;
Слайд №11 Висновки;

7. Орієнтовний перелік публікацій

8. Консультанти розділів дисертації

Розділ	Прізвище, ініціали та посада консультанта	Підпис, дата	
		завдання видав	завдання прийняв
1	Доцент Міночкін Д.А.	**05.10.2021**	**15.10.2021**
2	Доцент Міночкін Д.А.	**15.10.2021**	**21.10.2021**
3	Доцент Міночкін Д.А.	**22.10.2021**	**25.11.2021**

9. Дата видачі завдання “ 29 ” 09 2020 р.

Календарний план

№ з/п	Назва етапів виконання магістерської дисертації	Термін виконання етапів магістерської дисертації	Примітка
1	Опрацювання літературних джерел з теми роботи	27.09.2021	виконано
2	Аналіз вимог завдання, вибір методів та засобів рішення поставлених завдань	05.10.2021	виконано
3	Дослідження відомостей про машинне навчання, його методи та моделі, градієнтний бустинг XGBoost	15.10.2021	виконано
4	Систематизація інформації про відтік абонентів, його причини та способи запобігання	21.10.2021	виконано
5	Опрацювання вихідних даних для розробки моделі	10.11.2021	виконано
6	Розробка моделей та аналіз результатів	25.11.2021	виконано
7	Оформлення пояснювальної записки	06.12.2021	виконано

Студент Іван РАЧЕНЧУК

Науковий керівник дисертації Михайло ІЛЬЧЕНКО

РЕФЕРАТ

Робота виконана на 100 сторінках, містить 45 ілюстрацій, 13 таблиць. При підготовці використовувалася література з 17 різних джерел.

Актуальність: відтік є метрикою, що показує абонентів, котрі припиняють співпрацю з компанією або послугою, також відома як відтік абонентів. Слідуючи цій метриці, більшість компаній могли б намагатися зрозуміти причину відтоку абонентів та усунути ці фактори. Це обумовлено великими об’ємами даних та неможливістю опрацьовувати їх вчасно. Об'єм даних про абонентів в сегменті телекомунікацій, які збирають оператори зв'язку, можуть багато в чому сприяти переходу від реактивної до проактивної позиції. Поява складних методів штучного інтелекту і аналітики даних допомагає використати ці дані для ефективнішого вирішення проблеми відтоку абонентів.

Мета роботи: є використання алгоритмів та моделей машинного навчання для надання вірних рішень щодо класифікації абонентів оператора мобільного зв’язку, що дозволить зменшити втрати прибутку за рахунок передбачення сприятливих до відтоку абонентів та надання їм відповідних послуг.

Задачі дослідження:

- Проаналізувати існуючі моделі, методи та алгоритми машинного навчання, градієнтний бустинг XGBoost.
- Проаналізувати відтік абонентів, основні причини та методи запобігання. Проаналізувати існуючі рішення щодо використання машинного навчання для аналізу абонентів та передбачення відтоку.
- Вдосконалити процедуру аналізу абонентів оператора мобільного зв’язку, яка використовує дані системи білінгу та використовує для аналізу весь спектр даних.
- Розробити удосконалену модель машинного навчання, що дозволить надавати вірні рішення щодо класифікації сприятливих до відтоку абонентів з високою надійністю.

- Розробка імітаційної моделі, проведення навчання моделі з метою покращення результатів та перевірки ефективності.
- Розробка стартап-проекту по темі дисертації.

Об’єкт дослідження: надання вірних рішень щодо класифікування абонентів оператора мобільного зв’язку.

Предмет дослідження: алгоритми, методи та моделі машинного навчання, бібліотеки мови програмування Python, градієнтний бустинг XGBoost.

Методи дослідження: дослідження проведені в ході написання дисертації базуються на математичному аналізі, теорії ймовірностей. Моделювання моделей проведено на основі відповідних ним бібліотек машинного навчання мови програмування Python.

Ключові слова: машинне навчання, Python, білінг, класифікація, SKLearn, XGBoost, churn, відтік.

ABSTRACT

Work carried out on 100 pages containing 45 figures, 13 tables. The paper was written with references to 17 different sources.

Relevance: Churn is a metric that shows customers who stop doing business with a company or a particular service, also known as customer attrition. By following this metric, what most businesses could do was try to understand the reason behind churn numbers and tackle those factors. The wealth and the amount of customer data that carriers collect can contribute a lot to shift from a reactive to a proactive position. The emergence of sophisticated machine learning and data analytics techniques further help leverage this rich data to address churn in a much more effective manner.

Purpose of the bachelor's thesis: is to describe using of machine learning algorithms and models to predict customer churn based on the customer's data available that will reduce revenue losses due to churned customers.

Research objectives:

- Analyze existing models, methods, and algorithms for machine learning, XGBoost gradient boosting.
- Analyze subscriber churn, its main causes and methods of prevention. Analyze existing solutions regarding the use of machine learning for subscriber analysis and churn prediction.
- Improve the subscriber analysis procedure for a mobile operator, which uses data from the billing system and uses the full range of data for analysis.
- Develop an improved machine learning model that will allow you to provide correct decisions regarding the classification of favorable to churn subscribers with high reliability.
- Development of a simulation model, conducting a study of the model in order to improve the results and test the effectiveness.
- Development of a start-up project on the topic of the thesis.

Object of research: Providing correct decisions regarding the classification of mobile operator subscribers

Subject of research: algorithms, methods and models of machine learning, Python programming language libraries, XGBoost gradient boosting.

Research methods: the research conducted in the course of writing this dissertation is based on mathematical analysis, probability theory. Modeling of models was carried out on the basis of machine learning libraries of programming language Python corresponding to them.

Keywords: machine learning, Python, billing, classification, SKLearn, XGBoost, churn.

ЗМІСТ

ПЕРЕЛІК СКОРОЧЕНЬ

API	Application programming interfaces, прикладний програмний інтерфейс.
IoT	Internet of Things, Інтернет речей.
ML	Machine Learning, машинне навчання.
GPRS	General Packet Radio Service, загальний сервіс пакетної радіопередачі.
EDGE	Enhanced Data Rates for GSM Evolution, технологія передачі даних, що забезпечує передачу інформації в мережі мобільного зв’язку.
HSPA	High Speed Packet Access, протокол високошвидкісного пакетного доступу.
VoLTE	Voice over LTE, передача голосу по мережі LTE.
EE	британський оператор мобільного зв’язку та інтернет-провайдер
BSS	Business support system, система підтримки бізнесу, компоненти для проведення операцій з абонентами.
AI	Artificial Intelligence, штучний інтелект.
DL	Deep Learning, глибинне навчання.
CLIQUE	Clustering In QUEst, модель кластеризації на основі сітки.

ВСТУП

Зазвичай компанії приділяють більше уваги залученню абонентів, але значно меншої – їх утриманню. Згідно статистиці, залучення нового абонента потребує в 5 разів більше витрат, ніж утримання існуючого, а також, згідно дослідженню Bain & Company, збільшення рівня утримання абонентів на 5% може збільшити прибуток від 25% до 95%.

Відтік – це метрика, що показує абонентів, котрі припиняють співпрацю з компанією або послугою, також відома як відтік абонентів. Слідуючи цій метриці, більшість компаній могли б намагатися зрозуміти причину відтоку абонентів та усунути ці фактори за допомогою планів дій.

Але що, якщо було б можливо попередньо знати, що конкретний абонент, скоріш за все, залишить бізнес, та мати можливість вчасно вчинити необхідні дії, щоб цьому запобігти? Не існує якоїсь однієї причини, є їх сукупність, які так чи інакше призвели до незадоволення абонента. Вчитись на минулому та мати «під рукою» стратегічну інформацію для покращення майбутнього досвіду – все це про машинне навчання.

Коли йдеться про сегмент телекомунікацій, тут відкриваються широкі можливості. Багатство і об'єм даних про абонентів, які збирають оператори зв'язку, можуть багато в чому сприяти переходу від реактивної до проактивної позиції. Поява складних методів штучного інтелекту і аналітики даних допомагає використати ці багаті дані для ефективнішого вирішення проблеми відтоку абонентів.

Метою дисертації є проведення аналізу використання алгоритмів машинного навчання для запобігання відтоку абонентів оператора мобільного зв’зку. Для досягнення поставленої мети необхідно вирішити наступні задачі:

1. Провести аналіз моделей машинного навчання для бінарної класифікації.
2. На основі отриманих даних сформувати рекомендації щодо вибору моделі класифікації.

Об’єктом дослідження є надання вірних прогнозів щодо класифікування, а предметом – методи, алгоритми машинного навчання та моделі класифікації, а також бустинг XGBoost.

РОЗДІЛ 1
АНАЛІЗ МЕТОДІВ ТА АЛГОРИТМІВ МАШИННОГО НАВЧАННЯ. XGBOOST

1.1 Загальні відомості

Машинне навчання (англ. machine learning) - це метод аналізу даних, що автоматизує будування аналітичних моделей. Це відгілля від штучного інтелекту, побудоване на ідеї, що системи можуть вчитись на даних, ідентифікувати та приймати рішення з мінімальним втручанням людини.

Машинне навчання - це також наука змусити комп’ютери працювати, не будучи запрограмованими. Машинне навчання настільки розповсюджене, що ми навіть не знаємо, що використовуємо його в тих чи інших задачах. Багато дослідників вважають, що це кращий спосіб досягти прогресу в дорозі до штучного інтелекту на рівні з людиною.

Оскільки об'єм даних зростає кожного дня, то і проаналізувати їх з високою швидкістю та точністю неможливо. Більше 80% даних неструктуровані - це аудіо, відео, фотографії, документи, графіки та ін. Знайти закономірності в даних таких об'ємів дуже складно для мозку людини. Масивні дані, час, витрачений на обчислення нескінченно збільшується - саме тут вступає в дію машинне навчання, допомагаючи людям з обробкою значних даних за мінімальний період часу. Використовуючи штучний інтелект, люди хотіли побудувати більш якісні та інтелектуальні машини. Якщо поглянути на це зі сторони - дуже схоже наче дитина вчиться сама в себе. Так в машинному навчанні була розроблена нова можливість для комп'ютерів. І вже зараз машинне навчання існує в багатьох сегментах технологій, що ми навіть не розуміємо цього при його використанні.

Машинне навчання всюди - від медицинської діагностики на основі розпізнавання зображень до навігації безпілотних автомобілів. ML еволюціонує як дисципліна до такого ступеня, що в даний час дозволяє бездротовим мережам вчитись та отримувати знання, взаємодіючи з даними. Попередній інтерес та

обговорення доцільності розвитку 5G стандартів за допомогою протоколів ML привернули увагу інженерів та дослідників зі всього світу.

Ми стали свідками того, як мобільні та бездротові системи стали важливою частиною соціальної інфраструктури, мобілізуючи наше повсякденне життя та полегшуючи цифрову економіку різними способами. Однак, ML та бездротова мережа 5G сприймаються як різні області дослідження, незважаючи на потенціал, котрий вони можуть мати, коли використовуються разом. Фактично, вплив мобільного та бездротового мережевого зв'язку з підтримкою ML вже проявився в послугах на основі місцезнаходження, мобільного периферійного кешування, аналітиці Big Data та управлінні мережевим трафіком.

ML чудово підходить для складних проблем, коли значні рішення потребують ручного налаштування, і для проблем, вирішення яких взагалі відсутнє при використанні традиційних методів. Ці проблеми можливо вирішити вивчивши дані, змінивши звичайне програмне забезпечення, що містить довгі списки правил, підпрограмами ML, котрі автоматично навчаються на попередніх даних.

Важливим відмінністю ML від традиційних когнітивних алгоритмів є автоматичний витяг функцій. Завдяки цьому можна відмовитись від дорогих розробок функцій вручну. ML може знаходити аномалії, прогнозувати майбутні сценарії, адаптуватися до змін середовища, надавати представлення про складні проблеми з великими об'ємами даних та виявляти закономірності, котрі людина може пропустити.

Задачі, що стосуються машинного навчання, пропонують принципово оперативного визначення, а не тільки визначення поля в когнітивних термінах. Це слідує зі слів Алана Тьюринга в його статті “Обчислювальна техніка та інтелект”, в якій питання “Чи можуть машини думати?” замінено на “Чи можуть машини робити те, що і люди?”. В області аналізу даних машинне навчання використовується для розробки складних моделей та алгоритмів, котрі піддаються

прогнозуванню. Що стосується комерційної сфери - це називають прогнозуючою аналітикою.

1.2 Задачі машинного навчання

Корінь проблеми полягає у виявленні шаблонів, котрі також повинні бути виявлені на основі мережевої діяльності 4G в якості джерела даних для виявлення поведінки мережі як системи та користувачів як компонентів, з можливістю власного вибору, котрий міг би змінити потреби в послугах, заснованих на їх діяльності. При обміні інформацією між мережею та пристроями користувачів можна отримати надзвичайно великий об'єм даних. Такий сценарій показує джерело різного роду інформації, яку можна об'єднати для отримання результатів.

В аналізованій системі необхідно брати до уваги безліч змінних, котрі можна визначити та реєструвати з журналів, виконуючих процеси реєстрації користувачів, викликів, передачі, призначення IP, потоку даних та ін. Кореляція між безліччю змінних, котрі знаходяться під управлінням в межах одного аналізу, не може бути виконана без використання обчислювальної допомоги. Процес вивчення нового сховища даних є новим завданням, оскільки на початку процесу немає чіткого розуміння критеріїв пошуку. Аналіз за часовою шкалою виявляє зміни в системі та надає представлення про діяльність в мережі на основі змін значень кожної змінної. При даному сценарії є необхідною підтримка алгоритмів машинного навчання для виявлення та ідентифікації шаблонів в мережі, підлеглих аналізу.

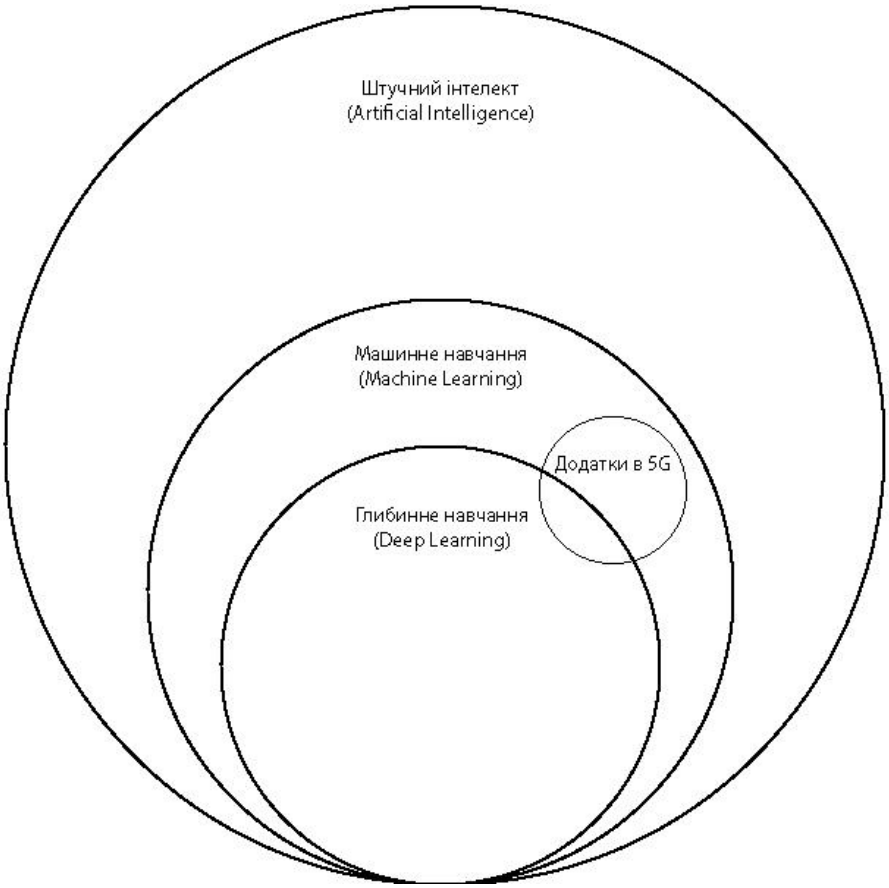


Рис. 1.1 Співвідношення між штучним інтелектом (AI), машинним навчанням (ML) та глибоким навчанням (DL).

В мобільних та бездротових мережах існує безліч параметрів, деякі з них встановлюються з використанням евристичних алгоритмів, оскільки це дозволяє пришвидшити рішення задачі в тих сценаріях, коли точне рішення не може бути знайдене. Для такого роду проблем алгоритм МН (наприклад, нейронна мережа) може вносити свій вклад, передбачаючи параметри та оцінюючи функції на основі наявних даних. Наступні покоління технологій мобільної та бездротової мереж також потребують використання оптимізації для мінімізації (або максимізації) деяких об'єктивних функцій.

З рисунку 1.1, машинне навчання є частиною штучного інтелекту. Тобто машинне навчання вважається штучним інтелектом, але не весь штучний інтелект вважається МН. Наприклад, символічну логіку - механізми правил, експерті системи та графіки знань - можна назвати ШІ, але жодне з них не є МН. Одним із

аспектів, що відділяє машинне навчання від графіків знань та експертних систем, полягає в його особливостях змінювати себе під впливом великих об'ємів даних. Тобто машинне навчання є динамічним та не потребує втручання людини для внесення змін. Це робить його менш залежним від людини.

Машинне навчання пов'язане з аналізом даних. Аналіз даних - область математики та інформатики, що включає в себе розробку методів обробки даних. Гарним прикладом аналізу даних є підрахунок мінімального та максимального значень або відношення декількох величин, але при умові, що висновки нададуть важливі знання про дані та дозволять вирішити поставлені задачі. Але слід зазначити, що аналіз даних не є машинним навчанням в чистому виді, а важливою відмінністю є здатність вирішувати широкий спектр завдань без конкретного плану та опису алгоритму рішення.

Основними узагальненими задачами машинного навчання є:

– Регресії, що виконує функцію прогнозування значень мітки за набором пов’язаних компонентів. Таким чином, нехай X – безліч даних, що є описами деяких об’єктів, а Y – безліч можливих рішень для X. Мітка може приймати будь яке значення, а не лише обиратися з кінцевого набору значень – значення відомі тільки на об’єктах вибірки $XY = \{(x_1, y_1) \dots (x_n, y_n)\}, x \in X, y \in Y$. Алгоритми регресії моделюють залежність міток від пов’язаних компонентів, щоб визначити закономірність змін міток при різних значеннях компонентів. В такому випадку, на вхід алгоритму буде надходити набір прикладів з мітками відомих значень , а результатом буде функція $a: X \to Y$, котра вміє прогнозувати значення мітки залежно від набору вхідних компонентів. Прикладом для сценаріїв регресії можна вважати прогнозування продажу тарифу мобільного оператора в залежності від рекламного бюджету.
– Двійкова / багатокласова класифікація, що прогнозує розподіл елементів даних по двом чи більше категоріям (класам). В основі алгоритму

класифікації є подання набору прикладів з мітками, кожна з яких представляє собою ціле число “0” або “1” для двійкової класифікації, де при багатокласовій класифікації відбувається перетворення початкових текстових міток в числовий ключ. Але для класифікації також характерне розділення об'єктів на пересічні, непересічні та нечіткі класи, де для пересічних один об'єкт може належати кільком класам, непересічні - тільки одному, а для випадку нечіткого класифікування - об'єкт належить всім класам з певним ступенем належності. В результаті роботи алгоритму, буде отримано класифікатор, що вмітиме прогнозувати клас нових екземплярів без мітки. Прикладом двійкової класифікації є розподіл коментарів соціальної мережі Twitter за тональністю - позитивні чи негативні.

- Виявлення аномалій за допомогою аналізу головних компонентів. Виявлення аномалій на основі АГК дозволяє створювати моделі, коли отримати дані для навчання з одного класу не є складним завданням, але отримати достатню вибірку аномальних значень навпаки складно. Оскільки аномалії за своєю сутністю є рідкісною подією, то можуть виникати труднощі при сборі репрезентативної вибірки даних, які використовуються для моделювання.
- Ранжування створює засоби ранжування, беручи за основу набір прикладів з мітками. Ці набори містять в собі групи екземплярів, що можуть бути оцінені за заданими критеріями, а мітки ранжування для кожного екземпляру є числовим рядом, наприклад, {0, 1, 2, 3, 4}. Засіб ранжування навчається ранжувати нові групи екземплярів з невідомими оцінками для кожного екземпляру.
- Прогнозування використовує попередні дані часових рядів, щоб робити прогнози про поведінку в майбутньому. Прикладом прогнозування є звичайний прогноз погоди.

– Кластеризація, що проводить групування окремих екземплярів даних в кластери зі схожими характеристиками. Також дану задачу використовують для виявлення неможливих логічних зв'язків в наборі даних, які неможливо помітити переглядом чи при спостеріганні за даними. Описати кластеризацію можна наступним чином: нехай X є множиною даних, що містить опис деяких об’єктів; Y є безліччю кластерів, відмічених мітками; також визначена функція відстані між об’єктами з початкової множини $X: f(x, x)$, і є деяка навчальна вибірка об’єктів $X_0 = \{x_n, y_n\}, x \in X$. Слід зазначити, що вхідні та вихідні дані напряму залежать від методу машинного навчання, рівно як і число кластерів заздалегідь невідомо і задається суб'єктивно. Надалі відбувається розбиття навчальної вибірки на кластери, де приписується номер кластера y_i для кожного x так, щоб близькі об’єкти належали одному кластеру, а об’єкти різних кластерів істотно відрізнялися за метрикою f. Будується алгоритм $a: X \to Y$, який ставить кожному $x \in X$ ставить у відповідність номер кластера $y \in Y$, що показано на рис.1.4 та рис.1.5.

Тобто немає чіткого критерію якості кластеризації, а існує лише ряд евристичних критеріїв, що виконують кластеризацію за одними даними, але з різними результатами. При цьому, незважаючи на описані складності, кластеризація допомагає досягти покращити розуміння даних простим розбиттям вибірки на групи схожих об’єктів, що спрощує подальшу обробку даних через застосування особливих методів аналізу до кожного окремого кластеру. Також вдається виявити аномалії та нові нетипові об'єкти, які не вдається віднести до жодного з кластерів. Прикладом сценаріїв для використання кластеризації є розподіл користувачів тарифів мобільного зв'язку на сегменти, беручи до уваги об'єм послуг обраних тарифів.

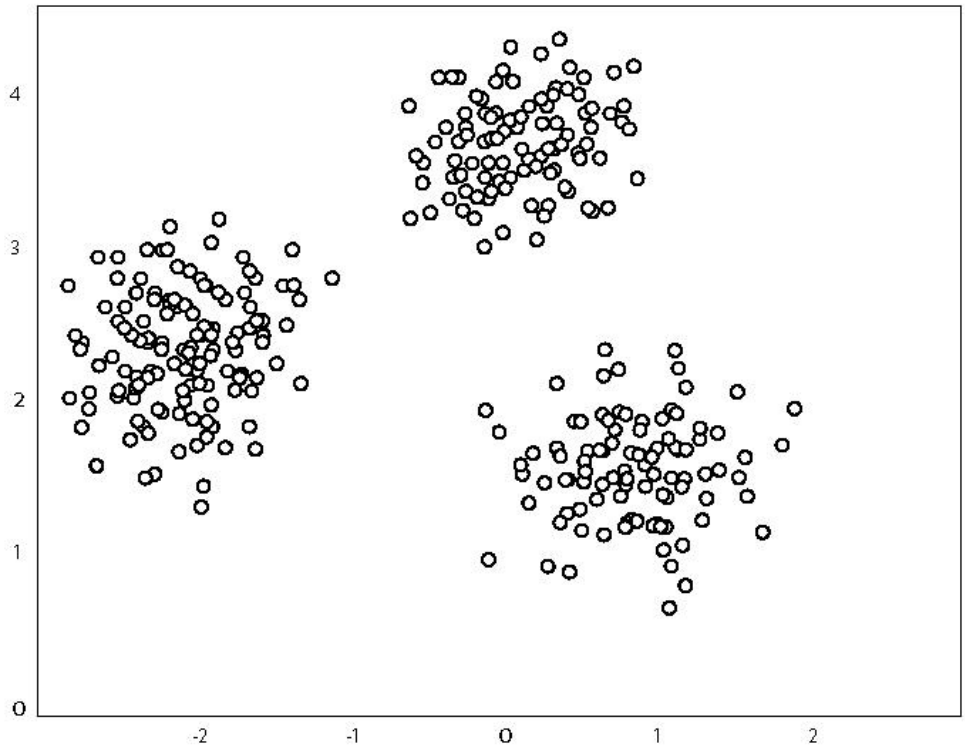


Рис. 1.2 Кластеризація. Приклад заданих початкових даних

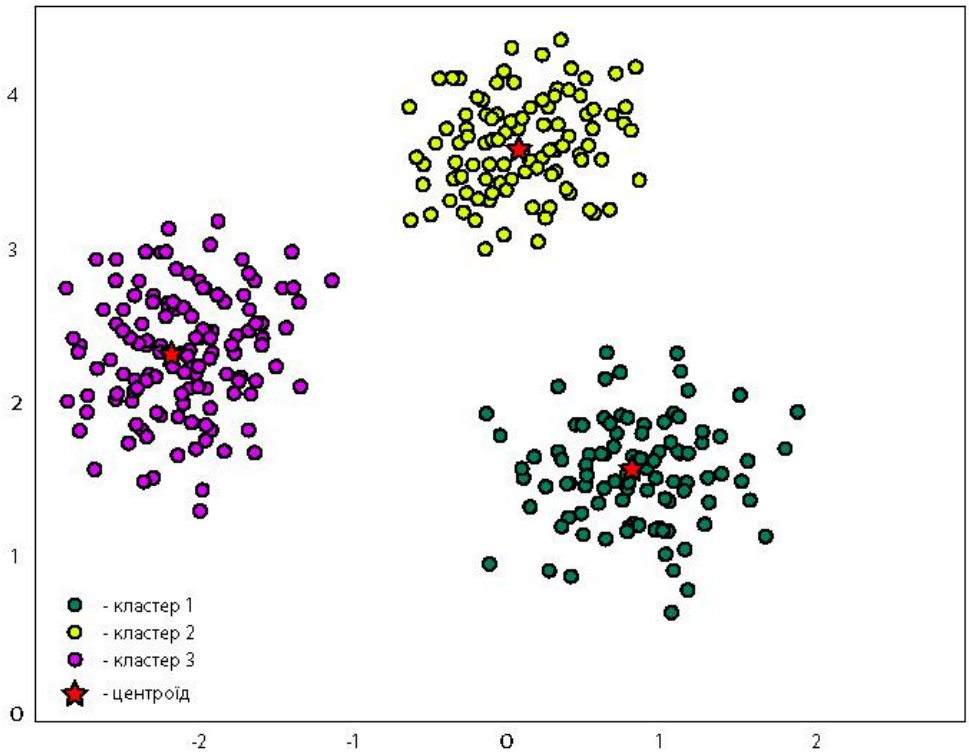


Рис. 1.3 Кластеризація. Графічне зображення результату кластеризації

З вищевказаного опису задач можна зробити висновок, що для розробки машинному навчанню необхідні оцифровані дані, оскільки вони є ключем до якості та повноти рішення.

1.3 Моделі та алгоритми машинного навчання

Моделі описують переконання про те, як функціонує та чи інша система, вони пояснюють систему та вивчають наслідки різних компонентів на прогнозування поведінки.

1.3.1 Математичні моделі

Математична модель – це опис системи, що використовує математичні поняття та мову, а математичне моделювання – процес розробки математичної моделі. В загальному випадку, математичні моделі складаються з відносин та змінних. Тобто модель описує систему набором змінних та рівнянь, які встановлюють відносини між змінними. Змінні можуть бути багатьох типів, наприклад, цілі або дійсні числа, логічні значення або строки. Незважаючи на те, що змінні представляють деякі властивості системи, фактична модель являє собою набір функцій, описуючих відношення між різними змінними. Ці відношення можуть бути описані алгебраїчними операторами, функціями, диференціальними операторами і т.д. Вивчаючи моделі важливо виявляти широкі категорії з моделей.

Класифікація окремих моделей за цими категоріями одразу вказують на деякі з основних елементів їх структури. За їх структурою можна використовувати декілька критеріїв класифікації для математичних моделей. Наприклад, статична та динамічна моделі або детерміністична або стохастична моделі. Статичні моделі не враховують варіації з часом, тоді як динамічні моделі в явному вигляді пов’язані на взаємодії з часом. В детерміністичних моделях всі математичні і логічні відношення між елементами фіксовані, в результаті чого, всі відношення

повністю визначають рішення. В стохастичній моделі хоча б одна змінна є випадковою.

В інженерії математичні моделі використовуються для максимізації певного виходу. Наприклад, спостережувана система може потребувати певних вхідних даних, а відношення між входами та виходами може залежати від других змінних, таких як змінні прийняття рішення, змінні стану та випадкові змінні. Змінні рішень також називають незалежними змінними. Змінні не залежить одна від одної, оскільки змінні стану залежать від рішення, вхідних та вихідних змінних. Крім того, вихідні змінні залежать від стану системи.

На відміну від спостережуваних змінних, приховані змінні є змінними, котрі не спостерігаються безпосередньо, а визначаються з інших змінних. Математичні моделі, що направлені на пояснення спостережуваних змінних з точки зору латентних змінних, називаються латентними змінними моделями. Такі моделі використовуються в машинному навчанні / штучному інтелекті, фізиці, менеджменті та біоінформатиці.

Одною з переваг використання прихованих змінних є те, що вони можуть використовуватись для зменшення розмірності даних. Велика кількість спостережуваних змінних може бути агреговано в моделі для представлення базової концепції, що облегшує розуміння даних. Вони виступають функцією аналогічною до функції наукових теорій. Однак, приховані змінні пов’язують спостережувані реальні дані з символічними модульованими даними.

Проблеми математичного моделювання часто класифікуються на моделях “чорного ящика” та “білого ящика”, в залежності від того, наскільки доступна інформація про систему. Модель “чорного ящика” – система, про яку зовсім немає доступної інформації, в той час як в моделі “білого ящика” вся необхідна інформація доступна. Практично всі системи знаходяться між цими двома моделями, тому ця концепція існує тільки в якості інтуїтивного опори для прийняття рішення про підхід. Важливо використовувати якнайбільш апріорну інформацію для створення більш точної моделі. Правильне використання цієї

інформації дозволить моделі вести себе коректно. Зазвичай така інформація надходить в формі знань типу функцій, пов'язаних з різними змінними. В моделях "чорного ящику" відбувається оцінювання функціональної форми відношень між змінними та числові параметри в цих функціях.

Будь яка модель, що не є чистим "білим ящиком", містить деякі параметри, котрі можуть бути використані для підгону моделі до системи, яку вона повинна описати. Якщо моделювання відбувається за допомогою машинного навчання, то оптимізація параметрів буде називатись навчанням. В звичайному моделюванні через явно задані математичні функції параметри зазвичай визначаються апроксимацією кривої.

1.3.2 Статистична модель

Під статистичною моделлю для машинного навчання розуміють математичну модель, яка втілює набір статистичних припущень, що стосуються генерації вибіркових даних. Статистична модель являє собою процес генерації даних (в ідеальній формі). Статистична модель задається як математична залежність між одним або кількома випадковими змінними та іншими невипадковими змінними.

Статистична модель зазвичай розглядається як пара (S, P), де S – набір можливих спостережень (вибірка), а P – набір розподілу ймовірності на S. Нехай, існує істинний розподіл ймовірності, який виконується завдяки генеруванню даних для спостережень. P обирається для представлення набору, який містить розподіл, наближений до істинного. Слід зазначити, що немає потреби містити істинне значення розподілу в множині значень P. Набір $P = \{ P_\theta : \theta \epsilon \Theta \}$. Набір Θ описує параметри моделі. Параметризація, як правило, потрібна для відокремлених значень параметрів, які приведуть до різних розподілів, $P_{\theta_1} = P_{\theta_2} \Rightarrow \theta_1 = \theta_2$ (повинна бути ін'єкційною). Параметризація, задовольняюча вимоги, вважається ідентифікованою.

Якщо задано статистичну модель (S, P) з $P = \{ P_\theta : \theta \in \Theta \}$, то вважається, що вона параметризована Θ має кінцевий розмір. Береться до уваги, що $\Theta \subseteq R^k$, де k – додатне ціле число, що називають розмірністю моделі. Статистична модель є напівпараметричною, якщо має як кінцеві, так і нескінченні параметри, та є непараметричною, якщо набір параметрів Θ не має кінцевих розмірів.

1.3.3 Алгоритми машинного навчання

Алгоритмом в машинному навчанні називають процедуру, що виконується над даними для створення моделі машинного навчання. Алгоритми машинного навчання виконують функцію “розпізнавання шаблонів”. Алгоритми вчаться по даних або вкладаються в набір даних. Існує безліч алгоритмів машинного навчання. Так, наприклад, є алгоритми класифікації як k-найближчі сусіди.

Узагальнений алгоритм процесу машинного навчання зображено на рис.1.4.

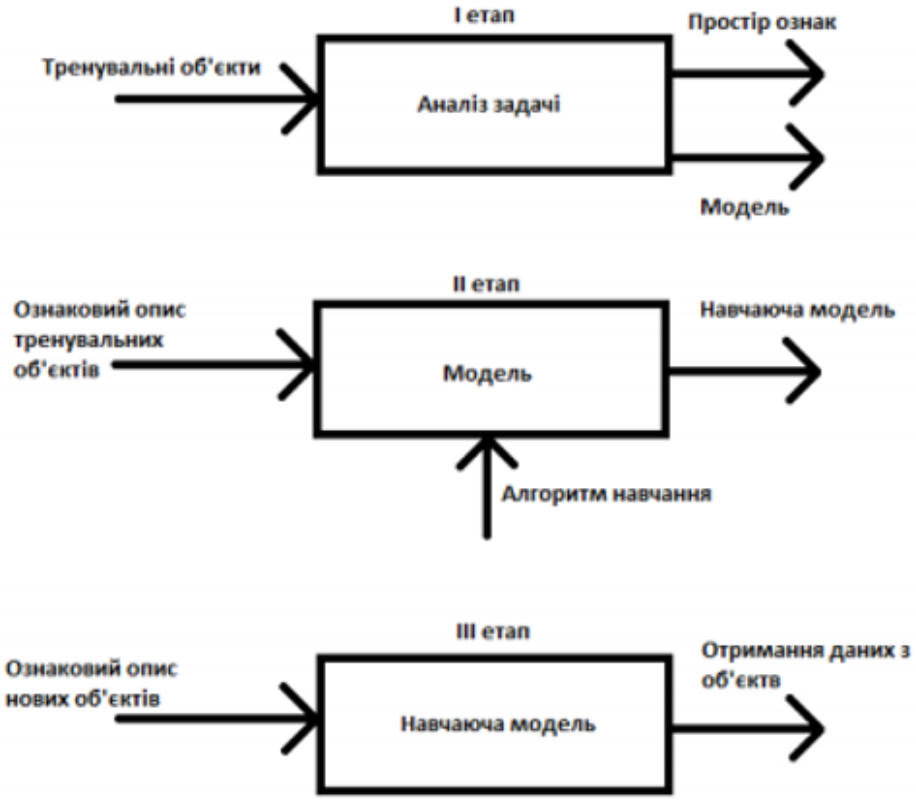


Рис.1.4 Узагальнений алгоритм проекту машинного навчання

Слід наголосити, що представлена ілюстрація відображає узагальнений випадок, в результаті якого отримано одразу працюючу модель. Але між другим та третім етапами є також чимало важливих дій – наприклад, оцінка якості моделі. Саме за результатами оцінки якості моделі буде прийнято рішення щодо переходу на наступний етап або повернення до попередніх.

Коли алгоритм асоційований з обробкою інформації, дані можуть можуть бути зчитувати з джерела вводу, записуватись на прилад виводу та зберігатись для подальшої обробки. Збережені дані розглядаються як частина внутрішнього стану об’єкту, виконуючого алгоритм. На практиці, вони зберігаються в одній або декількох структурах даних. Таким чином, алгоритм повинен бути чітко визначеним, а порядок розрахунків мати вирішальне значення для функціонування алгоритму.

В мовах програмування алгоритм призначений насамперед для вираження алгоритмів в формі, яку здатен виконати комп’ютер. Проектування алгоритмів є методом для рішення задач та інженерних алгоритмів. Є частиною багатьох теорій рішення операційних дослідів, як динамічне програмування. Методики проектування та реалізації алгоритмів також називають шаблонами. Один з найважливіших аспектів проектування алгоритму є створення алгоритму, час виконання якого є ефективним.

Алгоритми машинного навчання можна описати як навчання цільової функції f, яка найкращим чином співвідносить вхідні змінні X та вихідну змінну $cY : Y = f(x)$. Оскільки невідомо, що з себе буде представляти функція f, то необхідно навчити їх за допомогою різних алгоритмів.

Одним з найбільш використовуваних алгоритмів в машинному навчанні є лінійна регресія. Як вже описувалося раніше, моделювання в першу чергу стосується мінімізації помилки моделі або якомога більш точного прогнозування.

Лінійну регресію можна представити у вигляді рівнянь, які описують пряму, найбільш точно зображаючи зв'язок між вхідними змінними *X* та вихідними змінними *Y*.

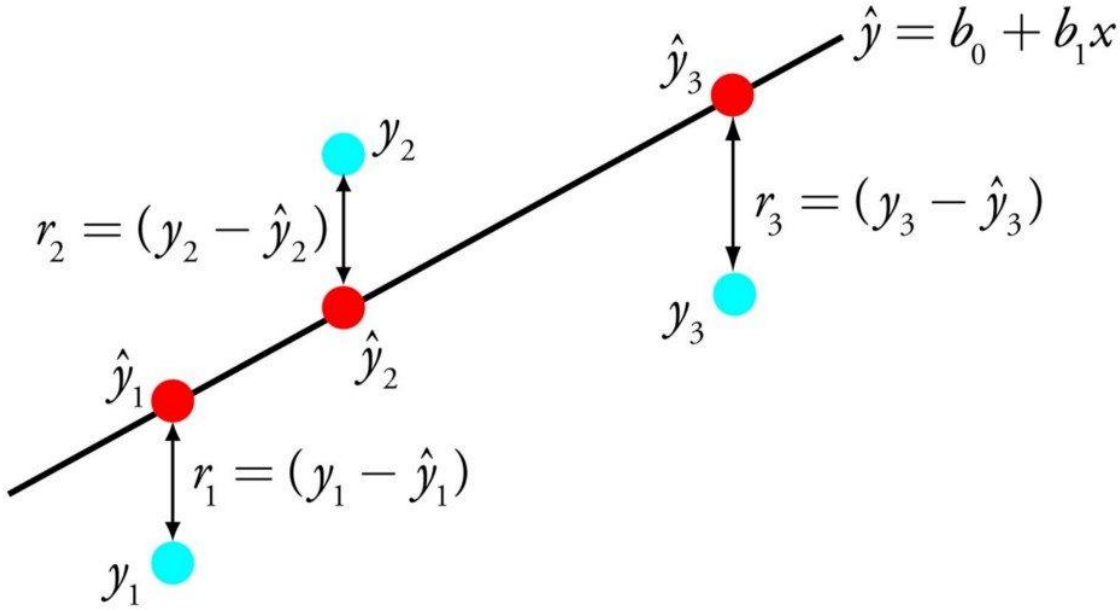


Рис.1.5 Графічне зображення використання лінійної регресії

Наприклад, $Y = B_0 + B_1 * X$. Знаючи *X* треба знайти *Y*. А мета лінійної регресії полягає в пошуку значень коефіцієнтів B_0 та B_1. Для оцінки регресивної моделі використовуються різні методи типу лінійної алгебри або методу найменших квадратів.

Також часто використовується дерево прийняття рішень. Дерево рішень можна представити у вигляді двійкового дерева, знайомого по алгоритмам та структурам даних. Кожен вузол представляє собою вхідну змінну і точку розподілу для кожної змінної, але тільки при умові, що змінна є числом. Тобто суть роботи полягає в послідовному розбитті безлічі даних на непересічні класи, які також піддаються розбиттю за будь-якими критеріями з оцінкою ефективності розбиття.

Дерево рішень складається з:

– «листя» - містять значення цільової функції;

– «гілок» - містять записи атрибутів, від яких залежить цільова функція;

– «вузлів» – містять інші атрибути, які використовуються при класифікації.

Дерева рішень найчастіше використовуються для класифікації (передбачається результат - клас, якому належать дані) та регресії (результатом є прогнозоване значення цільової функції).

Узагальнений алгоритм для побудови дерева прийняття рішень за навчальною вибіркою є наступним:

1. Береться атрибут та встановлюється в корінь дерева.
2. В «листі» даної «гілки» залишаються лише ті значення, які відповідають необхідній умові. Крок повторюється для кожного значення цього атрибута.
3. Продовжується побудова дерева із залишеного на попередньому кроці «листя».

Дерева швидко навчаються та роблять прогнози. Крім того, вони є точними для широкого спектру задач та не потребують особливої підготовки даних.

Було б неправильно не розглянути й наступний алгоритм. Дуже простий та ефективний алгоритм – K-ближніх сусідів. Модель КНС представлена набором тренувальних даних. Прогнози для нової точки робляться виходячи з результатів пошуку K-ближніх сусідів в наборі даних та сумуванні вихідної змінної для цих K екземплярів. Для того, щоб визначити схожість між екземплярами даних необхідно використати евклідові відстані (при умові, що масштаб один і той самий для всіх параметрів) – числа, які можна вирахувати на основі відмінностей з кожною вхідною змінною.

Метод K-ближніх сусідів може вимагати багато пам’яті для зберігання всіх даних, але це компенсується швидкими прогнозами. Дані для навчання також можна оновлювати, щоб прогнози залишались точними з плином часу. Даний алгоритм може погано працювати з багатовимірними даними, що негативно вплине на ефективність алгоритму при вирішення задачі.

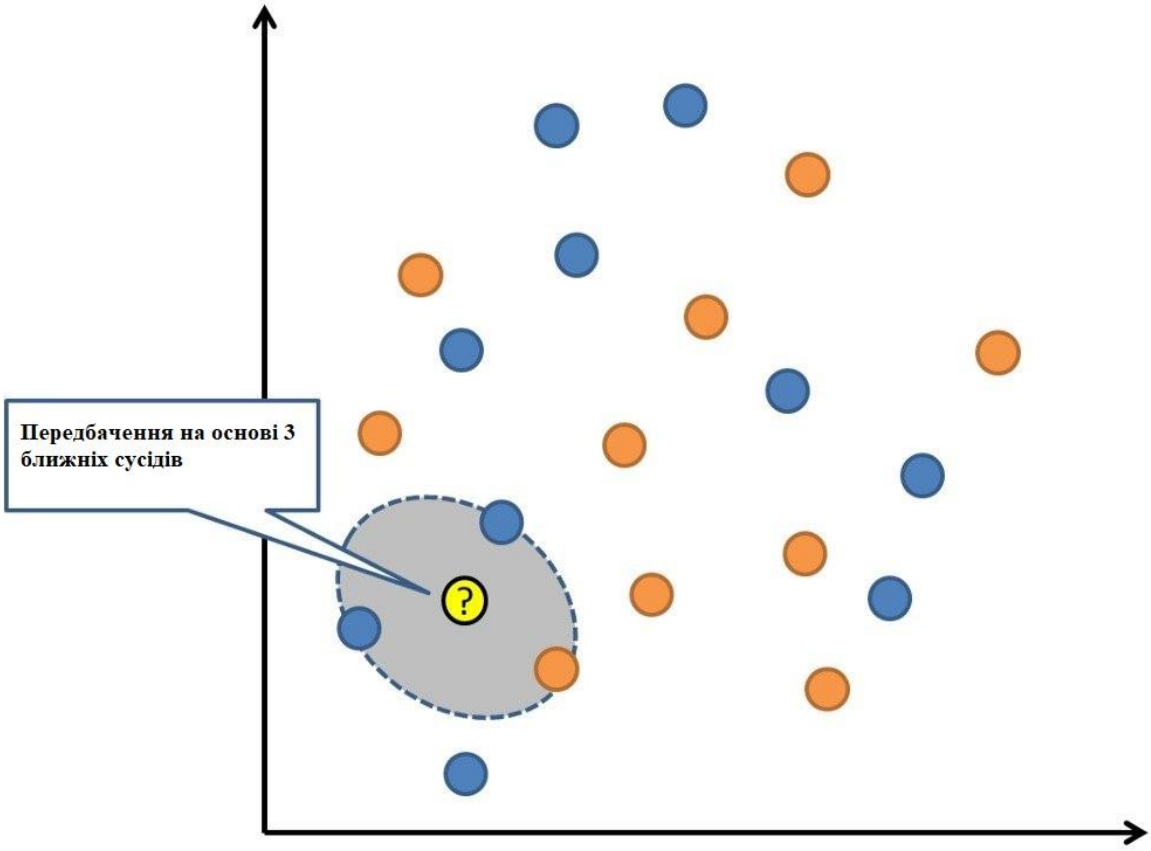


Рис. 1.6 Визначення подібності між екземплярами даних

1.4 Моделі кластеризації

Моделі кластеризацію можна класифікувати за наступними типами:

1. На основі щільності: ці моделі розглядають кластери як щільно заповнену область, що має деякі спільні та відмінні якості в порівнянні з нижчою за щільністю областю. Цим моделям властиві точність та здатність до злиття двох кластерів.
2. Ієрархічні моделі: кластери створюють деревовидну структуру на основі ієрархії. Нові кластери формуються виходячи з використання попередньо сформованого.
3. Моделі секціонування: застосовується розділення об’єктів на k кластерів, де кожен розділ формує один кластер. Цей спосіб використовується для оптимізації функції подоби цільового критерію, коли відстань є основним параметром. Прикладом є розглянутий метод k-найближчих сусідів.

4. Моделі на основі сіток: в таких моделях простір даних перетворюється в число осередків, які формують структуру сітки. Всі операції кластеризації виконуються швидко та незалежно від кількості об’єктів даних.

Розглянувши модель дерева рішень та K-найближчих сусідів, які були прикладами ієрархічної моделі та моделі секціонування, тепер звернемося до моделі кластеризації в квестах (англ. CLustering In QUEst).

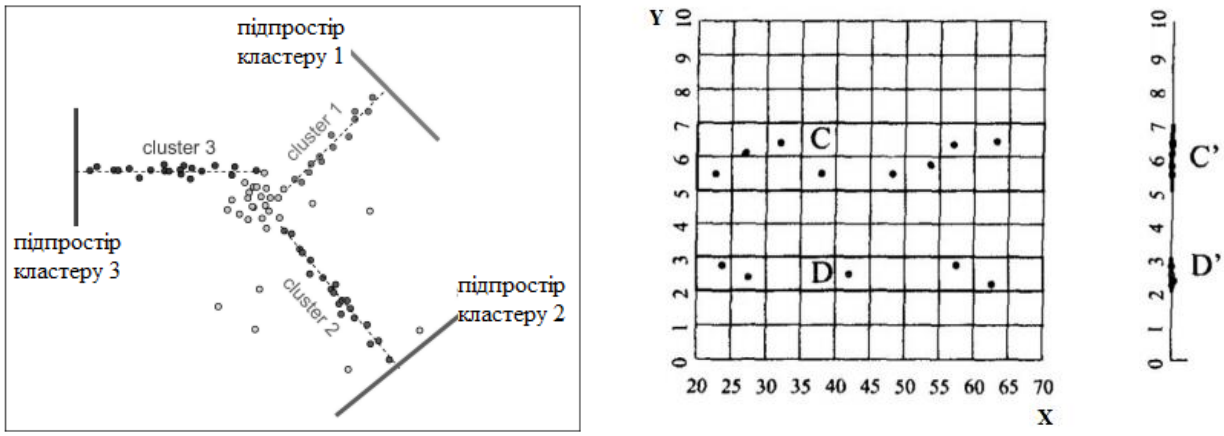


Рис.1.7 Представлення даних на основі щільності

CLIQUE – алгоритм кластеризації підпростору на основі щільності та сітки. По сітці проводиться дискретизація простору даних через сітку та оцінюється щільність, підраховуючи кількість точок в комірці сітки. Кластер на основі щільності - це максимальний набір пов’язаних за щільністю елементів в підпросторі. Елемент вважається щільним, якщо частка спільних точок даних, що містяться в елементі, перевищує вхідний параметр моделі. Кластером підпростору є набір сусідніх щільних комірок в довільному підпросторі. Він також виявляє деякі мінімальні описи кластерів. CLIQUE також автоматично визначає підпростори даних з великими розмірами, котрі дозволяють краще проводити кластеризацію, ніж початковий простір, з використанням апріорного принципу (див.рис.1.8).

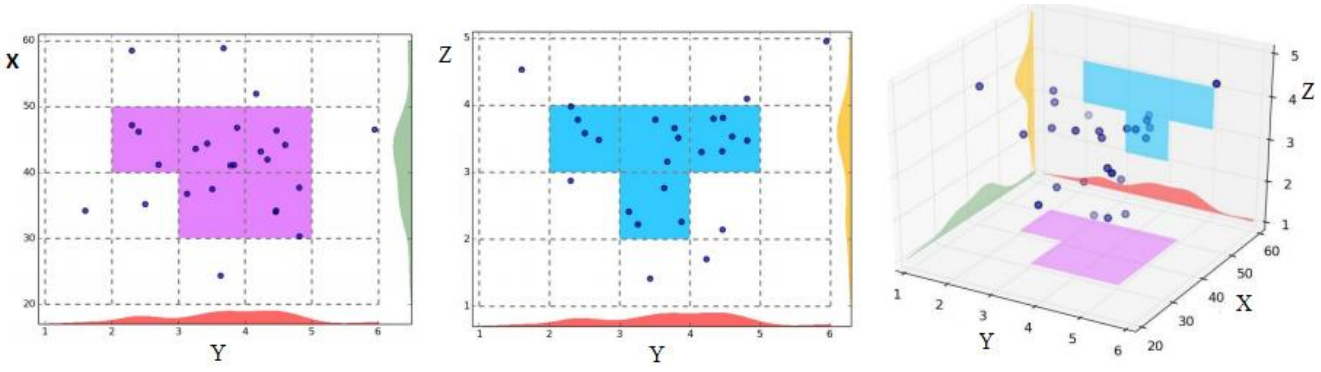


Рис.1.8 Апріорний принцип представлення даних

Апріорний принцип полягає в тому, якщо сукупність точок є кластером в k-мірному просторі, то ця сукупність також є частиною кластера в будь-яких (k-1)-мірних проекціях цього простору.

Основні етапи алгоритму CLIQUE:

1. Визначення підпросторів, що містять кластери.
   1.1. Розбиття простору даних та пошук кількості точок, що лежать всередині кожної комірки розділу.
   1.2. Визначення підпросторів, що містять кластери, з використанням апріорного принципу.
2. Ідентифікація кластерів.
   2.1. Визначення щільних елементів в усіх підпросторах
   2.2. Визначення пов’язаних щільних елементів в усіх підпросторах
3. Створення мінімального опису для кластерів
4. Визначення максимальних областей, котрі може покрити кожен кластер
5. Визначення мінімального покриття кожного кластеру

Переваги алгоритму:

– Автоматичний пошук підпросторів найвищої розмірності при наявності кластерів високої щільності
– Ігнорування порядку запису у вхідні дані та не припускає канонічного розподілу даних
– Масштабується лінійно до розміру вхідних даних
– Простота методу та інтерпретування результатів

Недоліки:

– Як і у всіх підходах кластеризації на основі сітки, якість результатів значно залежить від вибору кількості, ширини секцій та комірок сітки.

1.5 Моделі класифікації

Класифікаційна модель робить висновок на основі спостережуваних даних. За наявністю одного або декількох вхідних даних, модель класифікації прогнозує значення одного або декількох результатів. Результатами, як і було сказано в пункті 1.2, є мітки, які застосовуються до набору даних. Наприклад, при фільтрації повідомлень електронної пошти «спам» або «не спам».

Існує два основних підходи (способи) машинного навчання.

Першим тип - навчання з учителем або контрольований. Цей спосіб можна описати наступним чином:

– є деяка безліч об'єктів і безліч можливих реакцій системи на ці об'єкти, відповіді і об'єкти пов'язані між собою деякою невідомої залежністю;
– є кінцева сукупність прецедентів (так звана навчальна вибірка) на її основі необхідно виявити алгоритм, який згодом для будь-якого об'єкта з початкової множини дасть досить точну відповідь;
– для вимірювання точності відповідей використовується один з функціоналів якості, як правило, зав'язаний на обчисленні відхилення отриманої відповіді від очікуваного, тобто обчисленні помилки.

В контрольованій моделі навчальна вибірка вводиться в алгоритм, що дає змогу дізнатись, що є «спам», а що ні. Потім порівнюються тестова та навчальна вибірки з метою виявлення спам-повідомлень. Саме під такий спосіб навчання підпадає класифікація.

Але існує і другий спосіб навчання – навчання без учителя (або неконтрольоване). В них моделі отримують набори даних, які не є поміченими та шукають дані для кластеризації. Ці дані використовуються для пошуку даних за подібністю, виявлення шаблонів та ідентифікація відхилень в наборі даних.

Прикладом такого використання є пошук аналогічних зображень. Під такий тип навчання підпадає кластеризація.

Розглядаючи надалі на прикладі виявлення спам-повідомлень, відомо, що такі системи використовують класифікацію Байєса. При встановленні відмітки «спам» на повідомлені, слова з цього повідомлення потрапляють в загальну базу, що представляє собою набір даних і її можна вважати навчальною вибіркою. З часом складається список спам-слів та фраз. Тоді алгоритм здатен вирахувати вірогідність того, що електронний лист є «спамом» або «не спамом», і зробити свої висновки.

Моделі класифікації включають логістичну регресію (англ. Logistic regression), дерево рішень, наївний баєсів класифікатор (Naive Bayes), дерево з градієнтним бустингом (англ. gradient-boosted tree), один проти інших (англ. one-vs-rest), «випадковий ліс» (англ. random forest), k-ближніх сусідів.

Проведемо короткий опис деяких з них.

Логістична регресія приймає вхідні дані та визначає ймовірність деякого результату. Але в логістичні регресії діє «правило великого пальця», де при перевищенні ймовірності в 50%, робиться висновок, що рішення вірне. Логістична регресія є варіацією лінійної регресії, де робиться модель для розрахунку деякої залежної змінної на основі деякої незалежної змінної.

Дерево рішень також використовується і для класифікації. З назви зрозуміло, що остаточні рішення виносяться за допомогою встановлення проміжних рішень, які за своєю важливістю є некритичними самі по собі, але разом грають велику роль. Так рішеннями є «листя», а факторами на основі яких виносяться рішення – «гілки». Використовується концепція ентропії, при цьому ж розглядається частотний розподіл рішень, а потім розраховується логарифм. Коли ентропія рівна «0», всі відповіді є однаковими. Процес повторюється, але відбувається розділення кожного рішення на додаткові умови для кожного рішення, до тих пір, поки ентропія буде рівна 0.

Наступною моделлю є «випадковий ліс». Цей підхід класифікації є аналогічним до дерева рішень, за винятком поставлених умов (факторів), що містять деяку випадковість. Метою є витіснити упередженість та групові результати на основі найбільш ймовірних позитивних результатів. Такі набори позитивних відгуків називають сумками. Таку модель використовує стримінговий сервіс Netflix для рекомендацій фільмів, роблячи їх на основі людей зі схожими смаками. Рекомендації надаються з перебільшенням, використовуючи випадковість для уникання відхилення в неправильному напрямі.

Дерева з градієнтним бустингом відрізняються мають особливість, що використовується модифікація градієнтного бустингу для дерев рішень з фіксованими розмірами навчальної бази. Сам градієнтний бустинг є методом машинного навчання для задач регресії та класифікації, який створює модель прогнозування у вигляді сукупності дерев прийняття рішень. Модель будується поетапно, як і інші методи бустингу, та дозволяє оптимізувати функцію втрат. Такі моделі машинного навчання використовуються пошуковими гігантами Yahoo та Yandex в рейтингових системах.

1.6 XGBoost

Деревовидний бустинг - це високоефективний і широко використовуваний метод машинного навчання. У цій статті ми описуємо масштабовану наскрізну систему деревовидного бустинга під назвою XGBoost, яка широко використовується фахівцями з обробки даних для досягнення передових результатів при рішенні багатьох завдань машинного навчання. Ми пропонуємо новий алгоритм з урахуванням розрідженості даних і зважений квантильный ескіз для наближеного навчання дерев. Більше того, ми надаємо інформацію про моделі доступу до кеша, стискуванні даних і чергуванні для створення масштабованої системи деревовидного бустинга. Об'єднавши ці ідеї, XGBoost масштабується до мільярдів прикладів, використовуючи значно менше ресурсів, чим існуючі системи.

Рис. 1.9 Основні функції XGBoost

Машинне навчання і підходи, грунтовані на даних, стають дуже важливими у багатьох областях. Розумні класифікатори спаму захищають нашу електронну пошту, навчаючись на величезних масивах даних про спам і відгуки користувачів; рекламні системи вчаться підбирати правильні оголошення до правильного контексту; системи виявлення шахрайства захищають банки від зловмисників; системи виявлення аномальних подій допомагають фізикам-експериментаторам знаходити події, які призводять до нової фізики. Є два важливі чинники, які визначають ці успішні застосування: використання ефективних(статистичних) моделей, які відбивають складні залежності даних, і масштабовані системи навчання, які вивчають модель, що цікавить, на великих наборах даних.

Серед методів машинного навчання, використовуваних на практиці, градієнтний бустинг дерев є однією з техніки, яка яскраво проявляє себе у багатьох застосуваннях. у багатьох застосуваннях. Було показано, що деревовидний бустинг дає найсучасніші результати на багатьох стандартних еталонах класифікації. LambdaMART, варіант деревовидного бустинга для ранжирування, досягає найсучасніших результатів для завдань ранжирування. Окрім використання як самостійного предиктора, він також включений в реальні виробничі процеси для прогнозування коефіцієнта переходу по рекламі. Нарешті,

він є дефакто вибором методу ансамблю і використовується в таких завданнях, як премія Netflix.

XGBoost - масштабована система машинного навчання для бустинга дерев. Система доступна у вигляді пакету з відкритим початковим кодом . Вплив цієї системи був широко визнаний у ряді завдань машинного навчання і здобичі даних гірської справи. Візьмемо, приміром, завдання, що проводяться сайтом Kaggle, присвяченим змаганням в області машинного навчання. Серед 29 рішень-переможців 3, опублікованих у блозі Kaggle в 2015 році, в 17 рішеннях використовурисвалася система XGBoost. опублікованих у блозі Kaggle в 2015 році, в 17 рішеннях використовувався XGBoost. Серед цих вісім рішень використали тільки XGBoost для навчання моделі, тоді як більшість інших комбінували XGBoost з нейронними мережами в ансамблях. Для порівняння, другий за популярністю метод, глибокі нейронні мережі, був використаний в 11 рішеннях. Успіх системи був також продемонстрований на KDDCup 2015, де XGBoost використовувався кожною командою-переможцем в топ-10.

Більше того, команди-переможці повідомили, що ансамблеві методи перевершують добре налагоджений XGBoost лише на незначну величину. Ці результати свідчать про те, що наша система дає найсучасніші результати при рішенні широкого кола завдань. Приклади завдань, на яких були отримані ці рішення, включають: пророцтво продажів в магазині; класифікація подій у фізиці високих енергій; класифікація веб-текстів; пророцтво поведінки абонентів; виявлення руху; пророцтво кількості кліків по рекламі; класифікація шкідливих програм; категоризація продуктів; пророцтво ризику небезпеці; пророцтво відсівання на масових онлайн-курсах. Хоча аналіз даних, залежних від області, і розробка ознак відіграють важливу роль в цих рішеннях, той факт, що XGBoost є консенсусним вибором того, що навчається, показує вплив і важливість нашої системи і деревовидного бустинга. нашої системи і деревовидного бустинга.

Найважливішим чинником успіху XGBoost є його масштабованість в усіх сценаріях. Система працює більш ніж вдесятеро швидше, ніж існуючі популярні

рішення на одній машині, і масштабуються до мільярдів прикладів в розподілених або обмежених пам'яттю умовах. Масштабованість XGBoost обумовлена завдяки декільком важливим системним і алгоритмічним оптимізаціям. Ці інновації включають: новий алгоритм деревовидного навчання для роботи з розрідженими даними; теоретично обгрунтована процедура зваженого квантильного відбору процедура квантильного ескіза дозволяє працювати з вагами екземплярів в наближеному деревовидному навчанні. Паралельні і розподілені обчислення прискорюють процес навчання, що дозволяє швидше досліджувати модель. Що ще важливіше, XGBoost використовує позаядерні можливості. обчислення і дозволяє фахівцям з роботи з даними обробляти сотні мільйонів прикладів на робочому столі.

Нарешті, ще цікавіше ще цікавіше об'єднати ці методи, щоб створити наскрізну систему, яка масштабується до ще більших даних з мінімальною кількістю ресурсів кластера. Основний вклад цієї статті перераховані таким чином: - Ми розробляємо і створюємо високомасштабовану наскрізну деревовидну boosting system. - Ми пропонуємо теоретично обгрунтований зважений квантиль ескіз для ефективного розрахунку пропозицій. - Ми представляємо новий алгоритм паралельного навчання дерев з урахуванням розрідженості. - Ми пропонуємо ефективну блокову структуру з урахуванням кеш-пам'яті для позаядерного навчання дерев.

Рис. 1.10 Еволюція XGBoost від дерева рішень

XGBoost - це ансамблевий алгоритм машинного навчання на основі дерев рішень, що використовує градієнтний бустинг. У завданнях прогнозування неструктурованих даних(зображення, текст і так далі) штучні нейронні мережі, як правило, перевершують усі інші алгоритми або рамки. Проте, коли йдеться про структуровані/табличні дані невеликого і середнього розміру, алгоритми на основі дерева рішень вважаються кращими у своєму класі. Еволюція алгоритмів на основі дерев з роками представлена на діаграмі нижче.

Алгоритм XGBoost був розроблений в якості дослідницького проекту в Університеті Вашингтону. Тяньци Чен і Карлос Густрин представили свою доповідь на конференції SIGKDD в 2016 році і уразили світ машинного навчання. З моменту своєї появи цей алгоритм не лише виграв численні змагання Kaggle, але і став рушійною силою декількох передових промислових застосувань. Як результат, існує сильне співтовариство фахівців з обробки даних, що вносять свій вклад в проекти з відкритим початковим кодом ~350 учасників і ~3 600 комітів на GitHub. Алгоритм відрізняється наступними особливостями:

- Широкий спектр застосування : Може використовуватися для вирішення завдань регресії, класифікації, ранжирування і призначеного для користувача прогнозування.

- Переносимість: Працює без проблем на Windows, Linux і OS X.
- Мови: Підтримує усі основні мови програмування, включаючи C++, Python, R, Java, Scala і Julia.
- Хмарна інтеграція: Підтримує кластери AWS, Azure і Yarn і добре працює з Flink, Spark і іншими екосистемами.

Дерево ухвалення рішень : У кожного менеджера по найму є набір критеріїв, таких як рівень освіти, кількість років досвіду, результати співбесіди. Дерево рішень - це аналог того, як менеджер по найму проводить співбесіду з кандидатами, грунтуючись на своїх власних критеріях.

Пакетна вибірка: Уявімо, що замість одного інтерв'юера тепер є група інтерв'юерів, де кожен інтерв'юер має право голосу. Bagging або бутстрап-агрегація припускає об'єднання даних від усіх інтерв'юерів для ухвалення остаточного рішення за допомогою демократичного процесу голосування.

Випадковий ліс: Це алгоритм на основі мішковини з ключовою відмінністю, в якій тільки підмножина ознак вибирається випадковим чином. Іншими словами, кожен інтерв'юер перевірятиме інтерв'юованого тільки по певних випадково вибраних кваліфікаціях(наприклад, технічне інтерв'ю для перевірки навичок програмування і поведінкове інтерв'ю для оцінки нетехнічних навичок).

Бустинг: Це альтернативний підхід, при якому кожен інтерв'юер змінює критерії оцінки на основі зворотного зв'язку з попереднім інтерв'юером. Це "підвищує" ефективність процесу співбесіди за рахунок динамічнішого процесу оцінки.

Градієнтний бустинг: Особливий випадок бустинга, коли помилки мінімізуються за допомогою алгоритму градієнтного спуску. Наприклад, консалтингові фірми, що займаються розробкою стратегій, використовують кейс-інтерв'ю для відсіювання менш кваліфікованих кандидатів.

XGBoost: Рахуйте XGBoost градієнтним бустингом на "стероїдах"(не даремно ж він називається "Extreme Gradient Boosting"). Це ідеальне поєднання

методів оптимізації програмного і апаратного забезпечення, що дозволяє добитися чудових результатів з використанням меншої кількості обчислювальних ресурсів в найкоротші терміни.

Висновки

В результаті виконання даного розділу, було розглянуто основні задачі машинного навчання, проблеми та можливі методи боротьби з ними, способи використання технологій машинного навчання, градієнтний бустинг XGBOOST. Було проведено аналіз моделей машинного навчання, зокрема статистичних та математичних, моделей класифікації та кластеризації. Також було проведено аналіз основних алгоритмів машинного навчання, серед яких - дерево прийняття рішень, k-ближніх сусідів та статистичні моделі й методи.

РОЗДІЛ 2

АНАЛІЗ ВІДТОКУ АБОНЕНТІВ МОБІЛЬНОГО ЗВ'ЯЗКУ. ОСНОВНІ ПРИНЦИПИ ТА МЕТОДИ

2.1 Визначення відтоку

Відмічають, що «відтоку абонентів» визначається як процес переходу абонентів від одного постачальника послуг до іншого.

Відтік може бути активним, випадковим або пасивним. При вірному управлінні абонентами, можливо мінімізувати сприятливість до відтоку та максимізувати прибутковість компанії. Необхідно створити механізм для аналізу приписуваної прибутковості. Прогнозування відтоку також можна описати як метод, який допомагає у визначенні сприятливих до відтоку абонентів заздалегідь.

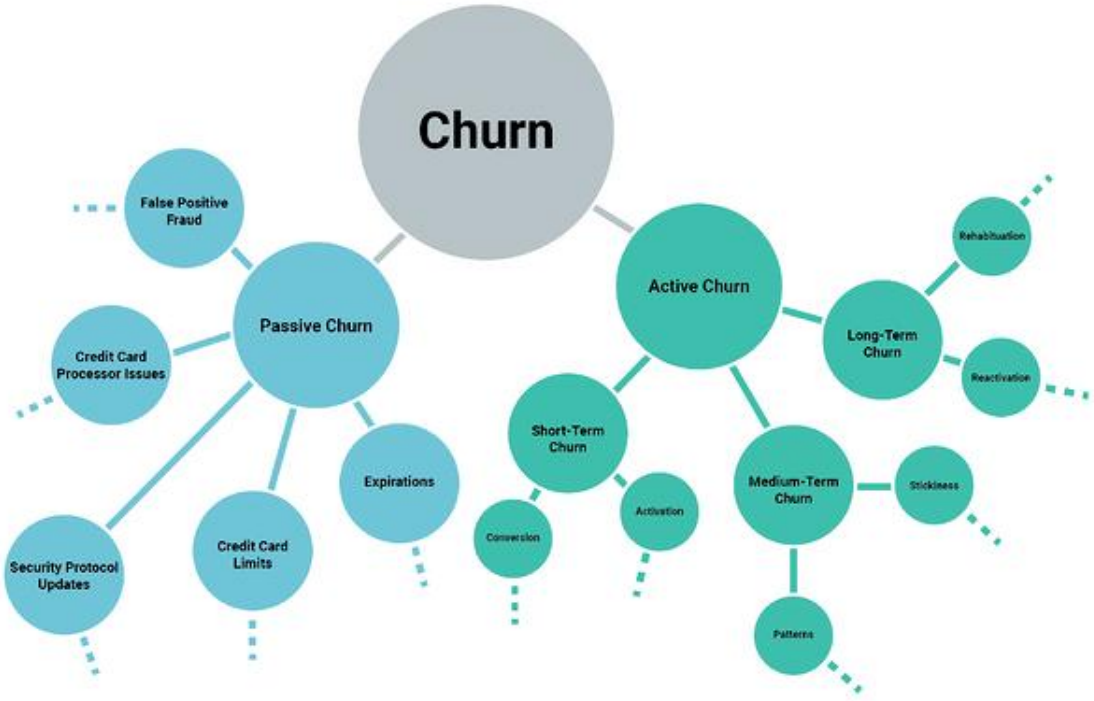


Рис. 2.1 Основні різновиди відтоку абонентів

Відтік абонентів або відтік покупців - це явище, коли абоненти підприємства більше не купують товари і не взаємодіють з підприємством.

Високий відтік означає, що більша кількість абонентів більше не хоче придбавати товари і послуги у цього підприємства. Коефіцієнт відтоку абонентів або коефіцієнт відтоку абонентів - це математичний розрахунок відсотка абонентів, які навряд чи вчинять ще одну купівлю в компанії.

Відтік абонентів відбувається, коли абоненти вирішують не продовжувати купувати товари/послуги у організації і припиняють свій зв'язок з нею. Це важливий параметр для організації, оскільки придбання нового абонента може коштувати майже в 7 разів дорожче, ніж утримання існуючого. Відтік абонентів може стати перешкодою на шляху експоненціально зростаючої організації, тому необхідно розробити стратегію утримання абонентів, щоб уникнути підвищення рівня відтоку абонентів.

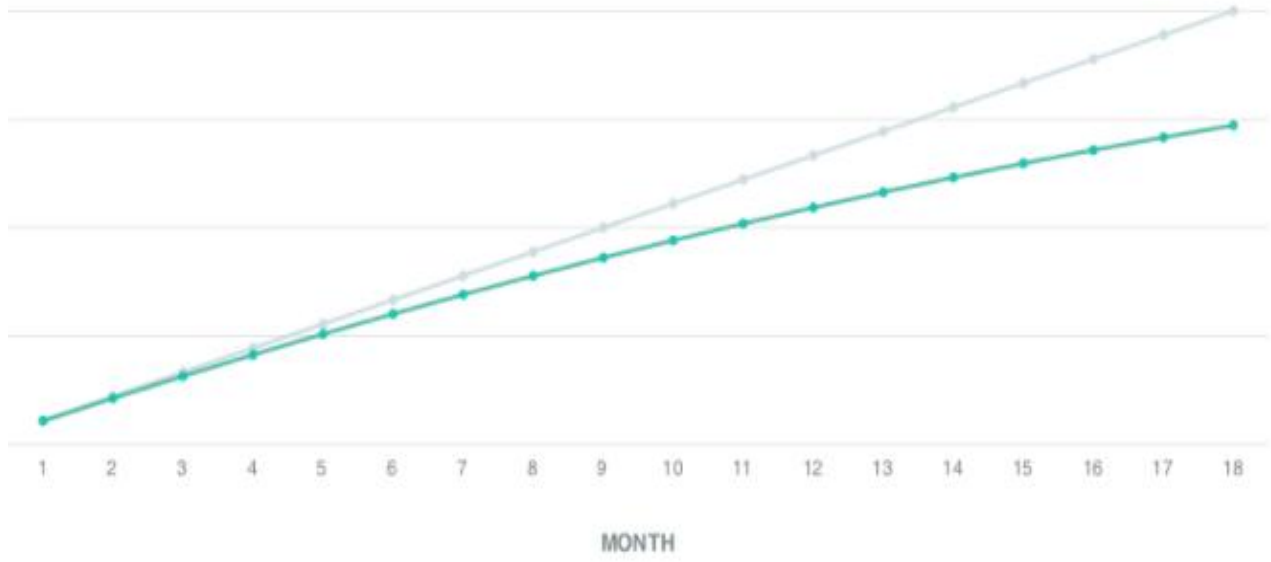


Рис. 2.2 Графік зменшення прибутку компаній через відтік абонентів

2.1.1 Важливість передбачення відтоку абонентів

Можливість передбачити, що певний абонент схильний до дуже високого ризику відтоку, поки ще є час зробити щось істотне для цього, сама по собі представляє велике додаткове потенційне джерело доходу для будь-якого бізнесу.

- Придбання нових абонентів - справа дорога, але втрата існуючих абонентів обійдеться бізнесу або організації ще дорожче. Оскільки існуючі абоненти, що платять, як правило, є постійними покупцями, які, якщо вони задоволені, здійснюватимуть повторні покупки у вашого бренду.
- Конкуренція на будь-якому ринку росте, і це спонукає організації зосередитися не лише на новому бізнесі, але і на утриманні існуючих абонентів. Найважливіший крок на шляху до прогнозування відтоку абонентів - почати заохочувати існуючих абонентів за постійні покупки і підтримку.
- Увесь шлях абонента веде до відтоку абонентів, а не тільки декілька інцидентів. У зв'язку з пріоритетом відвертання відтоку абонентів, організації повинні почати пропонувати заохочення за покупки цих абонентів, що скоро відторгаються.
- Як вже згадувалося раніше, намір абонента припинити користуватися певним продуктом/послугою завжди є рішенням, сформованим з часом. Існують різні чинники, які призводять до такого рішення, і організаціям важливо зрозуміти кожного з них, щоб переконати абонентів залишитися і продовжувати здійснювати покупки. Це можна зробити, постійно проводячи опитування задоволеності абонентів і аналізуючи отримані відгуки.

2.1.2 Визначення коефіцієнту відтоку абонентів

У найпростішому виді показник відтоку абонентів - ця кількість втрачених вами абонентів, що ділиться на загальну кількість абонентів. Щоб отримати приблизну оцінку, ви можете сегментувати своїх абонентів на основі частоти їх покупок.

Коефіцієнт відтоку абонентів = Кількість втрачених абонентів/Загальна кількість абонентів(період) x 100

2.2 Аналіз та переваги запобіганню відтоку абонентів

Тепер, коли ми добре розуміємо, що таке відтік абонентів, наступним очевидним кроком буде його аналіз. Для цього є дві причини:

1. Перш ніж шукати рішення для підвищення рівня відтоку, необхідно знати, що є його причиною в першу чергу.

2. Якщо ви впровадили рішення для зниження відтоку, ви повинні знати, працює воно або ні.

Існує безліч способів вистежування і аналізу відтоку, тут ми зупинимося на двох методах: когортний звіт і відтік по поведінці абонентів.

1. Когортний звіт: Когортний звіт аналізує одиниці ваших абонентів і їх відтік з часом. Когорта - це одиниця або сегмент покупців, які придбали продукцію вашого бренду в певний період часу. Звичайна когорта, яку можна використати, - це абоненти, що вчинили покупки в певному місяці, наприклад, ваша когорта за січень 2018 року - це абоненти, що вчинили покупки цього місяця. У когортного звіту є дві основні переваги: він дає чисті цифри, не схильні до впливу придбання нових абонентів, а друга основна перевага полягає в тому, що він допомагає виявити закономірність у відтоку абонентів.

У когортного звіту є дві основні переваги: він дає чисті цифри, не схильні до впливу придбання нових абонентів, а друга основна перевага полягає в тому, що він допомагає виявити закономірність у відтоку абонентів.

2. Відтік по поведінці: На додаток до аналізу відтоку за даними когортного звіту, ви також можете проаналізувати відтік, спостерігаючи за поведінкою абонентів. Це означає, що вам треба простежити за певною поведінкою абонента: використання певних функцій або здійснення певної дії при купівлі і визначити його вплив на відтік.

Цей метод має такі переваги, як:

- Бізнес може вирішити зосередитися на продуктах і функціях, які потребують поліпшення, щоб зменшити відтік абонентів.
- Бренди також можуть зосередитися на поліпшенні вже існуючих функцій, щоб утримати абонентів.

Переваги запобіганню відтоку абонентів:

- Отримання інформації для поліпшення: незадоволені абоненти є джерелом конструктивного зворотного зв'язку для поліпшення організації. Організація отримає інформацію про аспекти, які необхідно поліпшити при реалізації стратегій по відвертанню відтоку абонентів.
- Зниження ризику для бізнесу: Відтік абонентів означає прямі збитки для бізнесу. Продати новий продукт/послугу існуючому абонентові набагато простіше, ніж новому. Таким чином, відтік абонентів може завдати шкоди зростанню бізнесу.
- Розуміння цільового ринку: Постійна робота над зниженням відтоку абонентів дозволить виявити невідомі раніше шари ринку. Можна проводити опитування, фокус-групи і інші подібні заходи, щоб краще упізнати цільовий ринок і, у свою чергу, понизити відтік абонентів.
- Створення конкурентної переваги на ринку: У світі, де існує постійна конкуренція за залучення нових абонентів і утримання існуючих, важливо мати перевагу перед конкурентами. В процесі зниження відтоку абонентів абоненти не лише дізнаються невідомі аспекти бізнесу, але і створюють конкурентну перевагу перед іншими на ринку.

До якої б IT-індустрії ви не належали, будь то мобільний зв'язок, інтернет-провайдери, IT-продукти або мережі. послуги, інтернет-провайдери, IT- продукти або соціальні мережі, не варто помилятися: телекомунікаційні абоненти як і раніше не помилитеся: телекомунікаційні абоненти продовжують повною мірою

використати свої права на перехід з однієї послуги на іншу: Це явище відоме як Відтік. Серед галузей, які найбільш постраждалих, мобільні телекомунікації знаходяться на першому місці з більш ніж 30% відхилень в Європі; до 60% в Африканських країнах на південь від Сахари і в Азії. На сайті причинами є, зокрема, наступні: технологічний розвиток телекомунікацій, лібералізація, глобалізація і жорстка конкуренція, які послідували за цим. Зростаючий рівень відтоку вважається чумою усіх операторів зв'язку, тому що втрата цінного абонента означає втрату майбутніх доходів.

Наприклад, в Китаї, якщо категорія "абоненти" витрачає CNY 200 /місяць і в цій категорії 10 мільйонів абонентів, то 0,5% відтоку - це 1 млн. юанів/місяць втраченого доходу. з цієї миті. Крім того, вартість залучення нового абонента в 5 в 5-10 разів вище, ніж вартість задоволення і утримання існуючого абонента. Тому утримання абонентів стало лейтмотивом маркетингових кампаній. Операторам мобільного зв'язку вигідніше вигідніше для операторів мобільного зв'язку інвестувати в тих абонентів, які вже мають досвід роботи з послугою поновлюючи їх довіру, чим постійно намагатися притягнути нових абонентів, для яких характерний більш високий рівень відтоку. Зрілі оператори мобільного зв'язку(ЗОС), незважаючи на високий рівень відтоку абонентів, мають більш високий рівень відтоку.

Вони також не можуть з упевненістю назвати причину, по якій конкретний абонент хоче відмовитися від послуги. У мережах мобільного зв'язку генерується величезний об'єм даних, проте дані про абонентів є складними і знаходяться під регулюванням конфіденційності, внаслідок чого доступна дуже обмежена інформація про абонента або його досвід використання послуг. Враховуючи ситуацію, що склалася, більшість досліджень по прогнозуванню відтоку абонентів в основному зосереджена на використанні абонентської бази і методів Data Mining(DM) для виявлення поведінки абонента, пов'язаного з подією відтоку. В основному, вони можуть виявити абонентів з високою схильністю до відтоку, але не обов'язково вказують причину відтоку. Компаніям необхідно знати причину

відтоку, перш ніж застосовувати стратегію утримання. Крім того, зазвичай потрібно аналіз живучості, коли треба упізнати можливий час відключення.

Проте більшість передплатних абонентів вже давно відштовхнулися від компанії, а більшість компаній дізнаються про відтік післяплатних абонентів тільки у той момент, коли останні відмовляються продовжувати контракт. Наскільки нам відомо, ефективна модель прогнозування відтоку, яка піднімає питання: Хто хоче відтоку, чому він або вона хоче відтоку і коли це станеться, дуже рідко зустрічається в дослідженнях по прогнозуванню відтоку. Відповідно, мета цього дослідження - показати, що гібридні моделі, побудовані на основі методів DM, можуть пояснити поведінку відтоку з більшою точністю, ніж окремі методи, і що в деякій мірі можна виявити причину відтоку, а також пояснити розрив між рішенням про відтік і часом відключення. Наша гібридна модель використовує логістичну регресію паралельно класифікації і у поєднанні з кластеризацією.

2.3 Передбачення відтоку абонентів

Прогнозування відтоку - це практика аналізу даних з метою виявлення абонентів, які, швидше за все, відмовляться від підписки. Більшість великих підписних компаній проводять власну форму аналізу прогнозування відтоку, щоб виявити абонентів, найбільш схильних до ризику відтоку, і при правильному підході це призводить до величезної економії коштів і підвищення довічної цінності абонента(LTV), незалежно від розміру компанії.

2.3.1 Модель передбачення відтоку абонентів

Щоб компанія могла передбачити відтік, історичні дані про абонентів об'єднуються з алгоритмами машинного навчання і логістичною регресією для ранжирування вірогідності відтоку абонентів. Різні алгоритми сумісні з прогнозуванням відтоку. Модель машинного навчання, що найбільш асоціюється з цією практикою, - це модель дерева рішень (тобто Random Forest), яка припускає

попередню обробку різних джерел даних, а потім навчання і оцінку. Компанії, що мають власну команду фахівців з аналізу даних, можуть створити індивідуальне рішення по прогнозуванню відтоку абонентів на основі штучного інтелекту. Деякі віддають перевагу такому підходу, враховуючи відсутність у галузі єдиної думки про те, як саме краще всього прогнозувати відтік - деякі фахівці з аналізу даних віддають перевагу аналізу виживаності або ансамблевій моделі, а не підходу машинного навчання.

Якщо ви не умієте програмувати або у вашого підприємства немає ресурсів або пропускної спроможності для створення власного рішення по прогнозуванню, все одно існує безліч послуг предиктивной аналітики, деякі з яких спеціально займаються питаннями відтоку.

Це один з найпоширеніших трюїзмів про ведення бізнесу по підписці, але його варто повторити: рівень відтоку абонентів, що навіть здається низьким, може зупинити зростання бізнесу або повністю його знищити. Навіть такі невеликі цифри, як 1,0% відтоку, 2,5% відтоку, 5,0% відтоку, є потенційно смертельними. Давайте розглянемо приклад з практики. Припустимо, що компанія, зображена на графіці нижче, має дуже хороший показник залучення абонентів. Якщо вона утримує відтік на рівні близько 1,0%, її зростання (продажі існуючим і новим абонентам) буде достатнім для того, щоб вона не опинилася в небезпеці. Якщо відтік перевищить цей показник, то компанія, швидше за все, виявить, що не може придбавати нових абонентів досить швидко, щоб зберегти траєкторію зростання, навіть якщо коефіцієнт залучення абонентів залишиться тим самим.

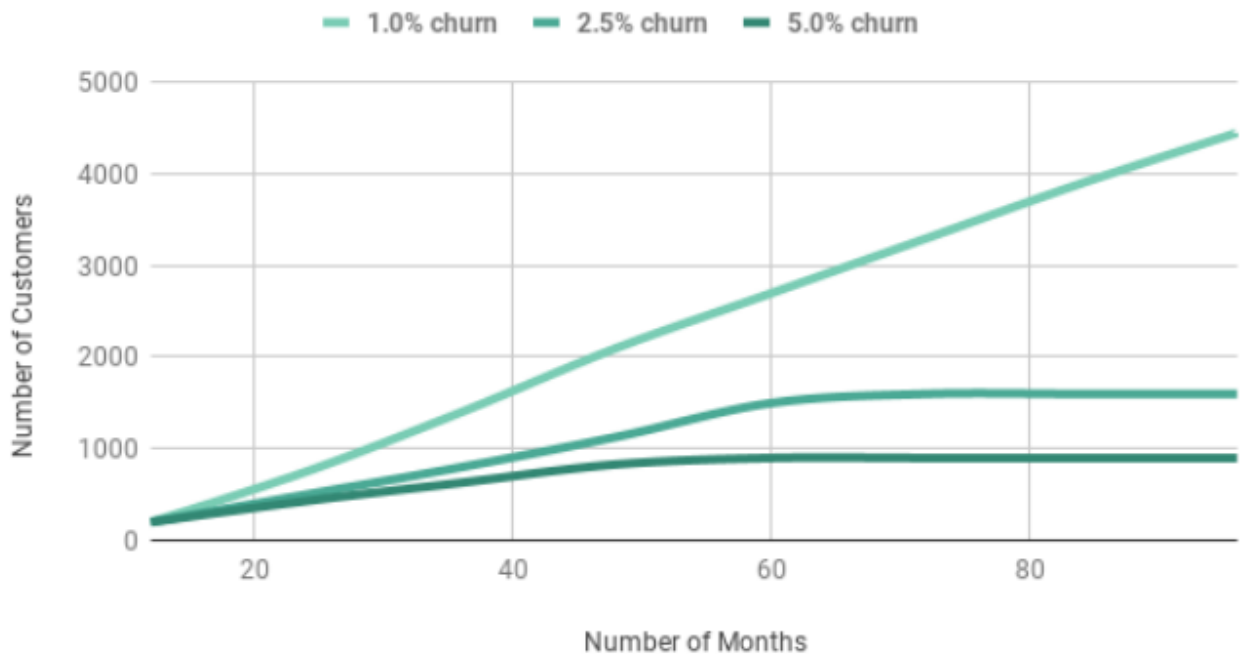


Рис. 2.3 Графік кількості абонентів в залежності від строку перебування абонентом компанії

На приведеному вище графіку ми можемо наочніше побачити ускладнення зростання, які виникають при підвищеному відтоку. Корисно думати про відтік безпосередньо в термінах доходу - якщо ваш щомісячний відтік складає 5%, це означає втрату доходу у розмірі 5%. Це сильно посилюватиметься з часом і до кінця року складе величезну суму потенційно втраченого доходу.

2.3.2 Визначення найбільш сприятливих до відтоку абонентів

Відтік відбувається з різних причин. Щоб краще зрозуміти причину відтоку абонентів, важливо правильно сегментувати абонентів. Вірогідність відтоку користувачів залежить від їх загального профілю, їх поведінки при використанні вашого продукту і їх потреб. Потреби абонента можуть мінятися впродовж усього періоду підписки - не усі відтоки відбуваються в перші декілька місяців.

Прогнозування відтоку корисне не лише для відвертання відтоку, що насувається, - воно може дати вашій команді цінні відомості про бізнес. З точки зору розвитку бізнесу, тенденції відтоку можуть допомогти маркетологам

створити персони абонентів, щоб направити на сегмент ринку ефективніші повідомлення і підвищити ефективність залучення абонентів. Що стосується утримання абонентів, то ваша команда по роботі з абонентами може передбачати майбутні тенденції відтоку і проактивно реагувати на них, щоб зберегти облікові записи, які інакше відторгалися б без підтримки. Візьмемо ситуацію, в якій помітно велика частина вашої абонентської бази відсіялася впродовж певного періоду часу без очевидних причин : деякі з них були новими користувачами, деякі користувалися вашим продуктом впродовж тривалого часу.

Компанії часто стикаються з проблемою класифікації : чому сталася ця подія відтоку? Деякі потенційні причини можуть бути наступними:

- Невдале оновлення
- Експоненціальне зростання у поєднанні з високим рівнем відтоку на етапі підключення
- Запуск нового конкурента, що пропонує той же продукт за нижчими цінами.

Річ у тому, що усі ці чинники, деякі або жоден з них не могли послужити причиною відтоку - користувачі могли піти з різних індивідуальних причин або з однієї взаємозв'язаної причини. Не маючи рішення для прогнозування відтоку і початкових даних для аналізу, ви не зможете з'єднати точки і визначити, які тенденції відтоку найбільше впливають на ваш бізнес.

2.3.3 Загальні причини відтоку абонентів

Абонент може виявити, що отримав від вашого продукту все, що йому треба, і більше не потребує ваших послуг. З точки зору того, що це говорить про якість вашого обслуговування, це чудово: з точки зору того, як це функціонально впливає на ваш MRR, не дуже. Відтік, що повторюється, такого роду може свідчити про те, що спектр ваших можливостей обмежений. Рахунок може змінитися, і новий менеджер по роботі з абонентами може захотіти використати

інше програмне забезпечення, з яким він краще знаком, або ж продовження може просто загубитися в тріщинах при передачі. Це може бути складним питанням для переговорів, але його можна вирішити за допомогою цілеспрямованого спілкування.

Компанія або абонент можуть знайти програмне забезпечення, яке їм подобається більше, ніж ваше, і перейти до конкурента. Це особливо шкідлива форма відтоку, оскільки ваша втрата - це виграш конкурента. Цей тип відтоку може бути пов'язаний як з продуктивністю продукту і ціноутворенням, так і з поліпшенням обслуговування і досвіду абонентів, особливо якщо конкурент почав знижувати ваші тарифи. Ознайомтеся з нашою статтею про коефіцієнти утримання абонентів по галузях, щоб отримати уявлення про те, як ваші конкуренти утримують абонентів.

Якщо ви змінили якусь функцію, прибрали її або ввели непопулярне оновлення, абонент може визнати ваш продукт менш корисним, чим його був раніше. Бути сприйнятливим до потреб ваших абонентів дуже важливо не лише з точки зору позиціонування вашого продукту, але і з точки зору його розвитку. Ви можете дізнатися, що ваші абоненти дійсно вважають корисним, оцінивши дані про їх використання або просто звернувшись до них безпосередньо. Ігнорування цієї інформації може обернутися неприємним сюрпризом - оновлення з цілим списком функціональних змін може привести до відтоку абонентів.

2.3.4 Важливість повідомлення про наявність відтоку

Існує цілий ряд причин, по яких абонент може бути готовий до відтоку. Одна з них дуже проста і дуже поширена: їм не допомагають отримати користь від вашого продукту. У таких випадках спілкування з абонентом безпосередньо може стати кращим способом перетворити його з групи ризику в задоволеного. Відправляйте по електронній пошті повідомлення абонентам, які не отримують повної віддачі від продукту(тобто випускають з уваги функції) або у яких закінчується підписка, що створює ризик відтоку абонентів. Цим можна легко

управляти за допомогою добре оптимізованою CRM або хорошій інтеграції прогнозування відтоку.

2.3.5 Стимулювання для утримання та аналіз тенденції відтоку

Акції, знижки, програми лояльності : усі вони допомагають абонентам відчути, що їх цінують і хочуть. Це, швидше за все, допоможе переконати їх залишитися, особливо якщо конкуренція на ринку сильна, і ви ризикуєте зазнати фінансової поразки, або доки ви працюєте над оновленням функцій. Стимулювання особливо важливе для відвертання негайного відтоку абонентів, у яких можуть бути незадоволені потреби. Заохочення можуть дати вам час, поки ви усуваєте проблеми з продуктивністю або розширюєте свої послуги.

Навіть сама краща стратегія не є надійною. Абоненти все одно йтимуть (хоча, сподіваюся, в значно меншій кількості), незалежно від успіху вашого прогнозу відтоку. Головне - не піддаватися бажанню зарити голову в пісок; подивіться на цифри відтоку і з'ясуйте, що пішло не так з цими абонентами. Відмінною стратегією для аналізу відтоку в цілому є планування подорожі абонента з вашим продуктом. Після того, як ви намітили основні моменти, порівняєте шлях абонента з даними про відтік і звернете увагу на те, де ризик відтоку найбільш високий. Це відбувається під час знайомства з продуктом? Через три місяці? Після двох років? Після оновлень? Ви навіть можете звернутися до абонентів, які вже відсіялися, за прямим зворотним зв'язком. Можливо, їм не захочеться його надавати, але будь-який зворотний зв'язок такого роду, яку ви можете отримати, дуже цінна.

2.4 Автоматизація зниження відтоку

Компанії, які постійно відстежують, як люди взаємодіють з продуктами, спонукають абонентів ділитися думками і оперативно вирішують їх проблеми, мають більше можливостей для підтримки взаємовигідних стосунків з абонентами.

А тепер уявімо собі компанію, яка вже давно збирає дані про абонентів і може використати їх для виявлення моделей поведінки потенційних відмовників, сегментації цих абонентів, що входять до групи ризику, і вжиття відповідних заходів, щоб повернути їх довіру. Ті, хто дотримується проактивного підходу до управління відтоком абонентів, використовують предиктивную аналітику. Це один з чотирьох видів аналітики, який припускає прогнозування вірогідності майбутніх результатів, подій або значень шляхом аналізу поточних і історичних даних. Предиктивная аналітика використовує різні статистичні методи, такі як інтелектуальний аналіз даних(розпізнавання образів) і машинне навчання (ML).

"Слабкість відстежування тільки реального відтоку полягає в тому, що він служить індикатором поганого абонентського досвіду, що лише запізнюється, і саме в цьому випадку модель прогнозованого відтоку стає надзвичайно цінною, - відмічає Майкл Редборд з HubSpot.

Основною рисою машинного навчання є створення систем, здатних знаходити закономірності в даних, навчаючись на їх основі без явного програмування. У контексті прогнозування відтоку абонентів це характеристики онлайн-поведінка, яка вказує на зниження задоволеності абонентів від використання послуг/продуктів компанії.

Олексій Беккер з ScienceSoft також підкреслює важливість машинного навчання для проактивного управління відтоком: "Що стосується виявлення потенційних відторгаючих абонентів, то тут алгоритми машинного навчання можуть відмінно впоратися. Вони виявляють деякі загальні моделі поведінки тих абонентів, які вже покинули компанію. Потім алгоритми машинного навчання звіряють поведінку поточних абонентів з цими шаблонами і сигналізують, якщо виявляють потенційних абонентів".

Підприємства, працюючі по підписці, використовують ML для предиктивной аналітики, щоб з'ясувати, які поточні користувачі не повністю задоволені їх послугами, і вирішити їх проблеми, коли ще не надто пізно : "Виявлення абонентів, яким загрожує відтік за 11 місяців до продовження

договору, дозволяє нашій команді по роботі з абонентами притягнути цих абонентів, зрозуміти їх больові точки і разом з ними розробити довгостроковий план, спрямований на те, щоб допомогти абонентові отримати користь з придбаної послуги", - пояснює Майкл.

Сфери застосування прогнозуючого моделювання відтоку виходять за рамки проактивної взаємодії з потенційними абонентами, що відторгаються, і вибору ефективних дій з утримання. За словами Редборда, програмне забезпечення на основі ML дозволяє менеджерам по роботі з абонентами визначати, з якими абонентами їм слід зв'язатися. Іншими словами, співробітники можуть бути упевнені, що вони розмовляють з потрібними абонентами в потрібний час.

Команди по продажах, роботі з абонентами і маркетингу також можуть використати знання, отримані в результаті аналізу даних, для узгодження своїх дій. "Наприклад, якщо абонент демонструє ознаки ризику відтоку, то, ймовірно, цей не кращий час для відділу продажів, щоб звернутися до нього з інформацією про додаткові послуги, в яких він може бути зацікавлений. Швидше, ця взаємодія має бути з CSM, щоб вони могли допомогти абонентові знову стати зацікавленим і побачити цінність в продуктах, які у нього є нині". Як і продажі, маркетинг може взаємодіяти з абонентами по-різному залежно від того, наскільки високий ризик відтоку : Наприклад, абоненти без ризику відтоку є кращими кандидатами для участі в тематичному дослідженні, чим абоненти з ризиком відтоку", - пояснює експерт HubSpot. В цілому, стратегія взаємодії з абонентами повинна грунтуватися на етиці і почутті часу. А використання машинного навчання для аналізу даних про абонентів може принести розуміння, яке допоможе реалізувати цю стратегію.

Якщо перспектива створення цілих спеціалізованих рішень по прогнозуванню відтоку здається запаморочливою, то інтеграція, спрямована на допомогу в прогнозуванні і скороченні відтоку, стане відмінним запасним варіантом.

2.5 Ведення бізнесу за допомогою прогнозування відтоку абонентів

Компанії часто застосовують підхід, обгрунтований на розрізнених діях, коли справа доходить до вирішення проблеми відтоку - встановлюють як можна нижчі ціни, прагнуть до максимального приросту абонентів в місяць - все, що завгодно, тільки не ретельна оцінка чинників, що викликають відтік, і змусити ці дані працювати на вас. Агресивні методи продажів можуть бути ефективні в короткостроковій перспективі, але ви лікуєте симптоми, а не хворобу. Розумний підхід до прогнозування відтоку - через інтеграцію або власне рішення - дозволить вам зрозуміти причини відтоку користувачів і відреагувати на них. Щастя користувача від вашого продукту може залежати від самих незначних на перший погляд причин. Використання прогностичного рішення для вирішення проблеми відтоку дасть вам ясність відносно кожного з них і зрештою допоможе вам перемогти відтік.

Висновки

В даному розділі було дано визначення відтоку абонентів оператора мобільного зв’язку, розглянуто важливість передбачення відтоку, визначення його коефіцієнта.

Було проведено аналіз та визначені переваги запобіганню відтоку, розглянуто модель передбачення та визначення найбільш сприятливих до відтоку абонентів, а також його основні причини. Визначено важливість вчасного повідомлення про його наявність, розглянуто способи та основні ідеї щодо автоматизації зниження відтоку, проаналізовано ведення бізнесу за допомогою прогнозування відтоку абонентів.

РОЗДІЛ 3
ПЕРЕДБАЧЕННЯ ВІДТОКУ АБОНЕНТІВ ЗА ДОПОМОГОЮ МАШИННОГО НАВЧАННЯ

3.1 Підготовка до аналізу

Зазвичай компанії приділяють більше уваги залученню абонентів, а їх утриманню приділяють другорядну увагу. Проте залучення нового абонента може коштувати в п'ять разів дорожче, ніж утримання існуючого. Згідно з дослідженням, проведеним компанією Bain & Company, підвищення рівня утримання абонентів на 5% може збільшити прибуток від 25% до 95%.

Коли йдеться про сегмент телекомунікацій, тут відкриваються широкі можливості. Багатство і об'єм даних про абонентів, які збирають оператори зв'язку, можуть багато в чому сприяти переходу від реактивної до проактивної позиції. Поява складних методів штучного інтелекту і аналітики даних допомагає використати ці багаті дані для ефективнішого вирішення проблеми відтоку абонентів. У цьому розділі я збираюся використати набір даних абонентської бази анонімного оператора зв'язку, наданий платформою IBM Developer.

Основна мета - розробити модель машинного навчання, здатну передбачати відтік абонентів на основі наявних даних про абонента. Для реалізації цього завдання я використовуватиму в основному бібліотеки Python, Pandas і Scikit - Learn. Щоб досягти цієї мети, пройдемо через наступні кроки:

- Дослідницький аналіз
- Підготовка даних
- Навчання, налаштування і оцінка моделей машинного навчання

Спочатку проведемо імпортування бібліотек Pandas, Scikit-Learn, numpy, seaborn, matplotlib, scikitplot та imblearn, а також налаштування графів.

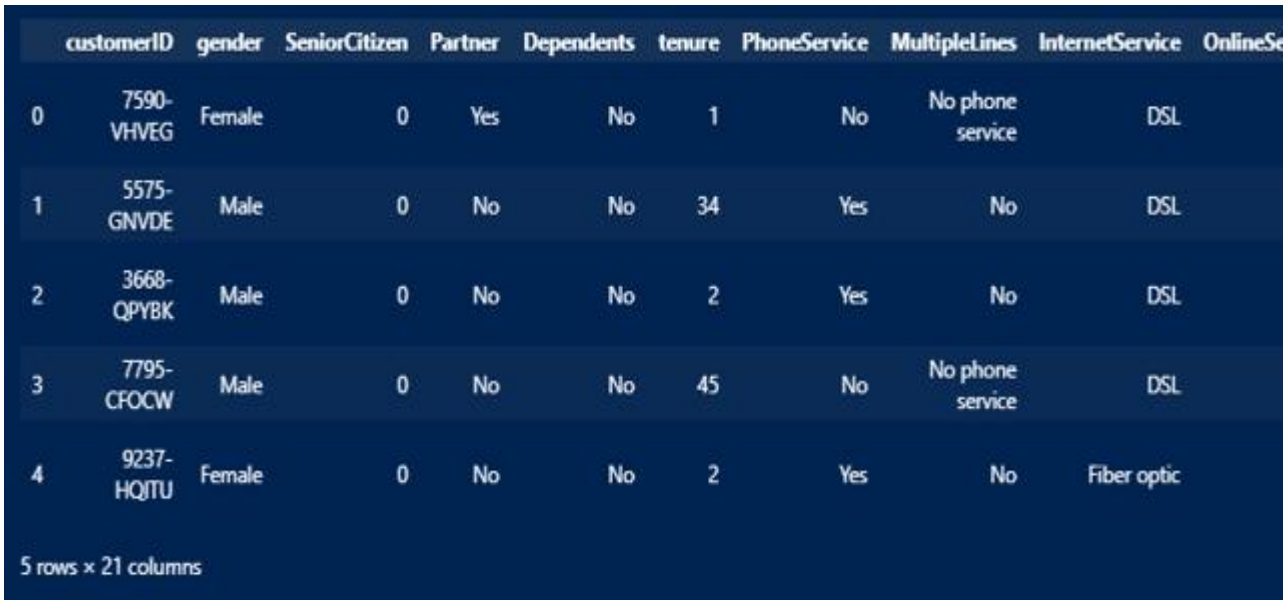
```
# importing libraries
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import scikitplot as skplt
%matplotlib inline

from sklearn.linear_model import LogisticRegression
from sklearn.metrics import classification_report
from sklearn.metrics import confusion_matrix
from sklearn.metrics import accuracy_score
from sklearn.metrics import roc_auc_score
from sklearn.preprocessing import StandardScaler, LabelEncoder
from sklearn.model_selection import train_test_split
from sklearn.model_selection import cross_val_score
from sklearn.model_selection import StratifiedKFold
from sklearn.model_selection import GridSearchCV
from sklearn.pipeline import make_pipeline
from sklearn.ensemble import RandomForestClassifier

from imblearn.under_sampling import RandomUnderSampler

# configure graph style
sns.set_style('darkgrid')
#palette=sns.color_palette("GnBu_d")
#palette=sns.color_palette("BuGn_r")
#palette = sns.light_palette("navy", reverse=True)
palette = sns.color_palette("coolwarm", 7)
```

Рис. 3.1 Імпортування бібліотек та налаштування графів

Далі імортуємо дані та виводимо на екран перші 5 рядків. Слід зазначити, що цей набір даних містить в цілому 7 043 абоненти і 21 атрибут, що складається з особистих характеристик, підписів послуг і деталей контракту. З усіх записів 5 174 є активними абонентами, а 1 869 – абонентами, які покинули оператора. Цільовою змінною для цієї оцінки буде характеристика відтоку. Абоненти мають наступні атрибути:

- customerID - Унікальний ідентифікатор абонента
- gender – Стать абонента - [Female\Male]
- SeniorCitizen - Літня людина або пенсіонер, пенсіонер - це людина, що досягла віку 60-65 років
- Partner - [No\Yes]
- Dependents – Чи має абонент залежності - [No\Yes]
- Tenure – життєвий цикл (в місяцях)
- PhoneService - [No\Yes]

- MultipleLines - [No\No phone service\Yes]
- InternetService - [No\No internet service\Yes]
- OnlineSecurity - [No\No internet service\Yes]
- OnlineBackup - [No\No internet service\Yes]
- DeviceProtection - [No\No internet service\Yes]
- TechSupport - [No\No internet service\Yes]
- StreamingTV - [No\No internet service\Yes]
- StreamingMovies - [No\No internet service\Yes]
- Contract – Тип контракту - [Month-to-month\One year\Two year]
- PaperlessBilling - [No\Yes]
- PaymentMethod – спосіб оплати - [Bank transfer (automatic)\Credit card (automatic)\Electronic check\Mailed check]
- MonthlyCharges – Щомісячні платежі
- TotalCharges – Загальні платежі
- Churn - значення відтоку, вектор таргера - [No\Yes]



	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSe
0	7590-VHVEG	Female	0	Yes	No	1	No	No phone service	DSL	
1	5575-GNVDE	Male	0	No	No	34	Yes	No	DSL	
2	3668-QPYBK	Male	0	No	No	2	Yes	No	DSL	
3	7795-CFOCW	Male	0	No	No	45	No	No phone service	DSL	
4	9237-HQITU	Female	0	No	No	2	Yes	No	Fiber optic	

5 rows × 21 columns

Рис. 3.2 Перегляд перших 5-ох рядків даних

Перевіримо повноту використовуваних даних, а саме 7043 абоненти та 21 атрибут.

```
def get_df_size(df, header='Dataset dimensions'):
    print(header,
          '\n# Attributes: ', df.shape[1],
          '\n# Entries: ', df.shape[0],'\n')

get_df_size(df)
```

```
Dataset dimensions
# Attributes:  21
# Entries:  7043
```

Рис. 3.3 Перевірка повноти використовуваних даних

Також необхідно перевірити, щоб чисельні атрибути також читались як числа, а не текст.

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7043 entries, 0 to 7042
Data columns (total 21 columns):
 #   Column            Non-Null Count  Dtype  
---  ------            --------------  -----  
 0   customerID        7043 non-null   object 
 1   gender            7043 non-null   object 
 2   SeniorCitizen     7043 non-null   int64  
 3   Partner           7043 non-null   object 
 4   Dependents        7043 non-null   object 
 5   tenure            7043 non-null   int64  
 6   PhoneService      7043 non-null   object 
 7   MultipleLines     7043 non-null   object 
 8   InternetService   7043 non-null   object 
 9   OnlineSecurity    7043 non-null   object 
 10  OnlineBackup      7043 non-null   object 
 11  DeviceProtection  7043 non-null   object 
 12  TechSupport       7043 non-null   object 
 13  StreamingTV       7043 non-null   object 
 14  StreamingMovies   7043 non-null   object 
 15  Contract          7043 non-null   object 
 16  PaperlessBilling  7043 non-null   object 
 17  PaymentMethod     7043 non-null   object 
 18  MonthlyCharges    7043 non-null   float64
 19  TotalCharges      7043 non-null   object 
 20  Churn             7043 non-null   object 
dtypes: float64(1), int64(2), object(18)
```

Рис. 3.4 Перевірка використовуваних даних на вірність читання даних

Як бачимо, атрибут 'TotalCharges' був прочитаний Pandas як тип даних 'object'. Це може плинути на процес дослідницького аналізу і повинно бути оброблено. Ми перетворимо тип даних в 'float64' в наступних пунктах.

3.2 Дослідницький аналіз

На початку аналізу необхідно перевірити дані на наявність пустих значень атрибутів.

```
# replacing all the blank values with NaN
df_clean = df.replace(r'^\s*$', np.nan, regex=True)

# print missing values
print("Missing values (per feature): \n{}\n".format(df_clean.isnull().sum()))
```

```
Missing values (per feature):
customerID          0
gender              0
SeniorCitizen       0
Partner             0
Dependents          0
tenure              0
PhoneService        0
MultipleLines       0
InternetService     0
OnlineSecurity      0
OnlineBackup        0
DeviceProtection    0
TechSupport         0
StreamingTV         0
StreamingMovies     0
Contract            0
PaperlessBilling    0
PaymentMethod       0
MonthlyCharges      0
TotalCharges       11
Churn               0
```

Рис. 3.5 Наявність пустих значень атрибутів даних

Після цього ми бачимо, що атрибут 'TotalCharges' має 11 пропущених значень. Ми замінимо ці відсутні значення медіаною 'TotalCharges'.

```
total_charges_median = df_clean.TotalCharges.median()
df_clean['TotalCharges'].fillna(total_charges_median, inplace=True)
```

Рис. 3.6 Заміна пустих значень значеннями медіани

При імпорті набору даних Pandas прочитав стовпець 'TotalCharges' як об'єкт, оскільки в нім було декілька записів, заповнених пропусками замість значення 'NaN'. Для аналізу ми перетворимо тип даних цієї характеристики з 'object' в 'float64', а також перевіримо дані на унікальність значень.

```
df_clean['TotalCharges'] = df_clean['TotalCharges'].apply(pd.to_numeric)
```

Рис. 3.7 Перетворення формату даних з object на float64

```
print("Unique values (per feature): \n{}\n".format(df.nunique()))
```

```
Unique values (per feature):
customerID          7043
gender                 2
SeniorCitizen          2
Partner                2
Dependents             2
tenure                73
PhoneService           2
MultipleLines          3
InternetService        3
OnlineSecurity         3
OnlineBackup           3
DeviceProtection       3
TechSupport            3
StreamingTV            3
StreamingMovies        3
Contract               3
PaperlessBilling       2
PaymentMethod          4
MonthlyCharges      1585
TotalCharges        6531
Churn                  2
dtype: int64
```

```
df_clean = df_clean.drop('customerID', axis=1)
```

Рис. 3.8 Перевірка значень атрибутів на унікальність

Перевіряючи унікальні значення атрибутів, ми бачимо, що стовпець 'customerID' має унікальні ідентифікатори для кожного абонента, що підтверджує, що кожен рядок представляє одного абонента. Цей атрибут не вносить вклад в цей аналіз, тому відмовимося від стовпця.

Також є необхідність провести описову статистику.

```
fig, ax = plt.subplots(ncols=1, nrows=2, figsize=(8,5))

sns.boxplot(df_clean['MonthlyCharges'], ax=ax[0])
sns.boxplot(df_clean['TotalCharges'], ax=ax[1])

plt.tight_layout()
```

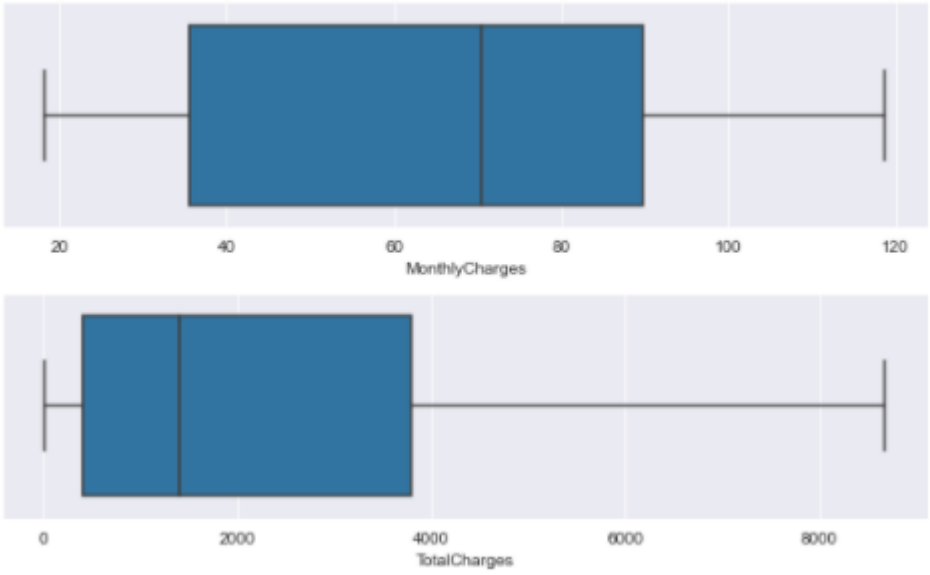


Рис. 3.9 Графіки ящиків за атрибутами «TotalCharges» і «MonthlyCharges»

На рис. 3.9 показано «графік ящиків», на якому видно мінімум, максимум, медіану, перший та третій квартиль.

Це дуже важлива інформація, яка допоможе нам зрозуміти, з яким набором даних ми працюватимемо.

Декілька спостережень:

- Характеристика 'SeniorCitizen' є бінарною, записи мають значення 1 для «Yes» і 0 для «No».

- Характеристика 'Tenure' має максимальне значення 72, що може означати, що цей постачальник послуг працює максимум 6 років.
- Єдиними характеристиками, які не є категоріальними, є 'Monthly Charges' і 'TotalCharges', усі інші характеристики є категоріальними.

```
features_obj = df_clean.columns

for f in features_obj:
    print(f)
    print(np.unique(df_clean[f].values))
```

```
gender
['Female' 'Male']
SeniorCitizen
[0 1]
Partner
['No' 'Yes']
Dependents
['No' 'Yes']
tenure
[ 0  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47
 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71
 72]
PhoneService
['No' 'Yes']
MultipleLines
['No' 'No phone service' 'Yes']
InternetService
['DSL' 'Fiber optic' 'No']
OnlineSecurity
['No' 'No internet service' 'Yes']
OnlineBackup
['No' 'No internet service' 'Yes']
DeviceProtection
['No' 'No internet service' 'Yes']

show more (open the raw output data in a text editor) ...

[ 18.25  18.4   18.55 ... 118.6  118.65 118.75]
TotalCharges
[  18.8    18.85   18.9   ... 8670.1  8672.45 8684.8 ]
Churn
['No' 'Yes']
```

Рис. 3.10 Визначення унікальних значень кожного атрибута

3.2.1 Життєвий цикл абонента

Яка тривалість життя абонента до відміни підписки? Велика частина відтоку спостерігається в перший місяць підписки, в цілому 20,3% абонентів йдуть в перший місяць. Більшість передплатників йдуть в перші 3 місяці, що складає 31,9% від загального числа відтоку.

```
p = sns.color_palette("coolwarm", 10)
p.reverse()

df_top_churn = pd.DataFrame(df_clean[df_clean['Churn'] == 'Yes']['tenure'].value_counts().sort_values(ascending=False))
total_churn = df_clean[df_clean['Churn'] == 'Yes'].shape[0]

fig, ax = plt.subplots(figsize=(10,5))
sns_lifespan = sns.barplot( x = df_top_churn[:10].index, y = df_top_churn[:10].tenure, ax=ax, palette=p, order=df_top_churn[:10].index)
plt.xticks(size=12)
plt.xlabel('Customer Lifespan (in months)', size=12)
plt.yticks(size=12)
plt.ylabel('Churn', size=12)
plt.tick_params(labelleft=False)

display_percent(ax, df_top_churn, total_churn)

sns_lifespan.figure.savefig("churn_rate_tenure.png", dpi=600)
```

Рис. 3.11 Графік життєвого циклу абонентів

3.2.2 Аналіз за особовими атрибутами

Надалі проведемо аналітику абонентів за особовими атрибутами. Давайте розглянемо, які особові характеристики вносять найбільший вклад в рішення про відтік. З наявного набору даних до них відносяться:

- Gender

- SeniorCitizen
- Partner
- Dependents

Приведені нижче графіки можуть дати деякі значимі висновки, такі як:

- Вірогідність відтоку абонентів без утриманців в 4 рази вище.
- Літні люди в 3 рази менше схильні до ризику відтоку.
- Партнери майже в 2 рази менш схильні до відтоку.

```python
# helper funtion - display count plot
def displayCountPlot(cat_list, df, rows=1, columns=3, figsize=(14,2.5), export=False):

  """
    Display countplot based on a set of features

    # Arguments
      cat_list: array, List of features
      df: DataFrame, dataset
      rows: int, number of rows
      columns: int, number of columns
      figsize: figure size, e.g (10, 5)

  """

  fig, ax = plt.subplots(ncols=columns, figsize=figsize)

  idx = 0
  for c in cat_list:
    idx += 1
    plt.subplot(rows, columns, idx)
    ax = sns.countplot(x=df[c], data=df, palette=palette)

    plt.xticks(size=10)
    plt.xlabel('')
    plt.yticks(size=12)
    plt.ylabel('')
    plt.subplots_adjust(hspace = 0.4)
    ax.tick_params(labelleft=False)
    ax.set_title(c, alpha=0.8)

    print_rate(ax, df.shape[0])

  if export :
    save_img(fig, ax)

  plt.tight_layout()
  plt.show()

  return fig
```

Рис. 3.12 Допоміжна функція для відображення кількості графіків

```
df_churn = df_clean[df_clean['Churn'] == 'Yes']
df_churn = df_churn.drop('Churn', axis=1)

df_churn.loc[df_churn['SeniorCitizen'] == 0,'SeniorCitizen'] = 'No'
df_churn.loc[df_churn['SeniorCitizen'] == 1,'SeniorCitizen'] = 'Yes'

personal_attributes = ['gender', 'SeniorCitizen', 'Partner', 'Dependents']
services_attributes = ['PhoneService', 'MultipleLines', 'InternetService', 'OnlineSecurity',
                       'OnlineBackup', 'DeviceProtection', 'TechSupport', 'StreamingTV',
                       'StreamingMovies']
contract_attributes = ['Contract', 'PaperlessBilling', 'PaymentMethod']
```

Рис. 3.13 Задання параметрів для побудови графіків відношення значень атрибутів

```
displayCountPlot(personal_attributes, df_churn, rows=1, columns=4, export=True)
```

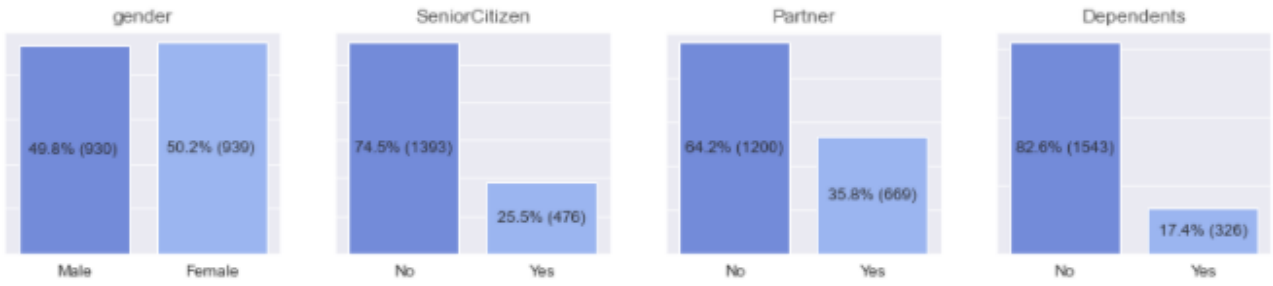


Рис. 3.13 Побудова та відображення графіків статистики за атрибутами «gender», «SeniorCitizen», «Partner» і «Dependents»

3.2.3 Аналіз з точки зору послуг

Розглянемо, які особові характеристики вносять найбільший вклад в рішення про відмову від послуг. З наявного набору даних це:

- PhoneService
- MultipleLines
- InternetService
- OnlineSecurity
- OnlineBackup
- DeviceProtection
- TechSupport
- StreamingTV

- StreamingMovies

На діаграмах нижче показані характеристики, по яких можна помітити великі розбіжності між класами. Це дає уявлення про те, якими послугами оператора зв'язку користуються абоненти, які найчастіше терплять поразку :

- У більшості абонентів, що відмінили підписку, включений телефонний зв'язок.
- Абоненти, що мають оптоволоконний інтернет, частіше відмовляються від підписки, ніж ті, хто має DSL.
- Абоненти, у яких не включені послуги Online Security, Device Protection, Online Backup і Tech Support, частіше відмовляються від підписки.

```
displayCountPlot(services_attributes, df_churn, rows=3, columns=3, figsize=(14,8), export=True)
```

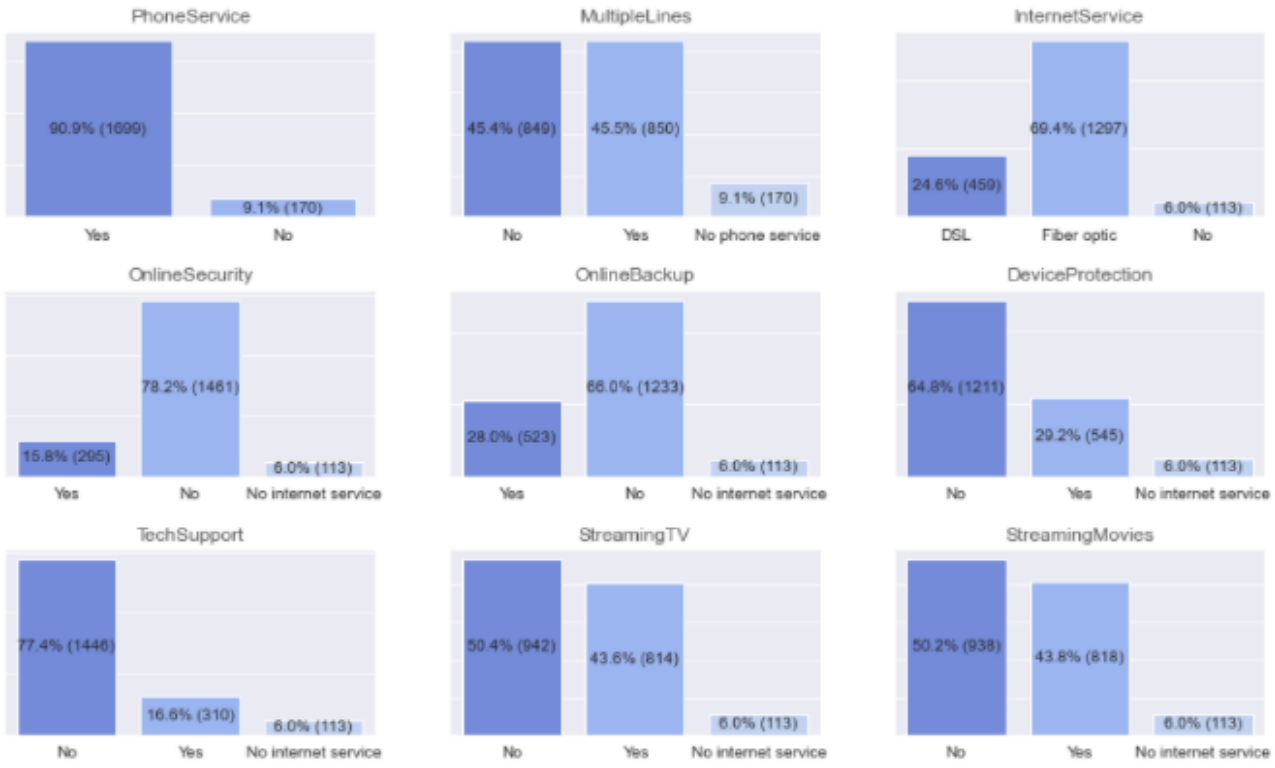


Рис. 3.14 Побудова та відображення графіків статистики за атрибутами «PhoneService», «MultipleLines», «InternetService», «OnlineSecurity», «OnlineBackup», «DeviceProtection», «TechSupport», «StreamingTV» і «StreamingMovies»

```
services_attributes_filtered = ['PhoneService', 'InternetService', 'OnlineSecurity',
                                'OnlineBackup', 'DeviceProtection', 'TechSupport',]

displayCountPlot(services_attributes_filtered, df_churn, rows=3, columns=3, figsize=(14,8), export=True)
```

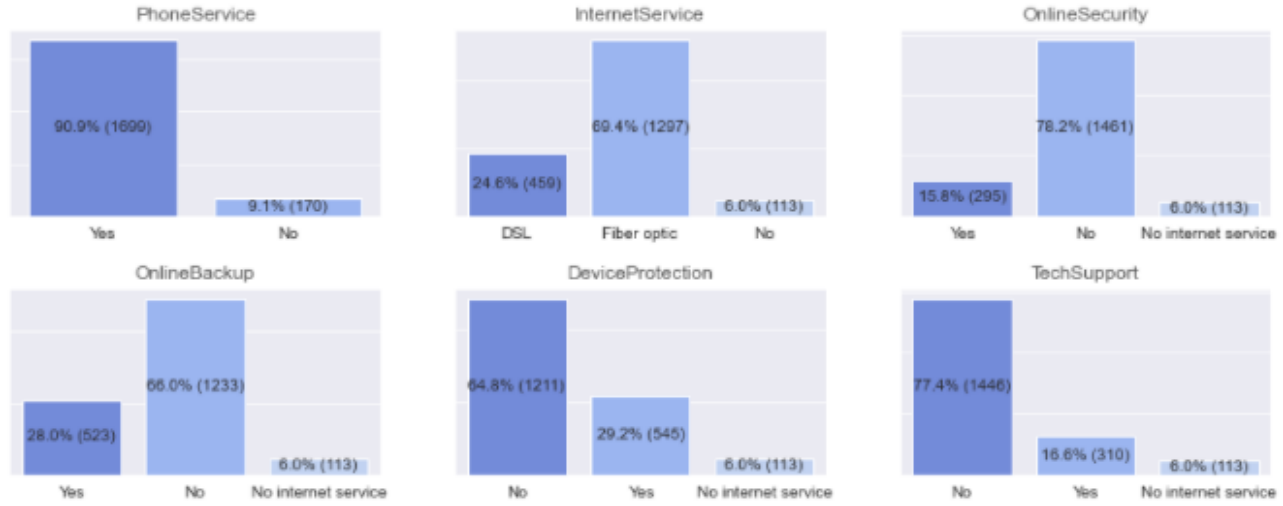


Рис. 3.15 Побудова та відображення графіків статистики за атрибутами «PhoneService», «InternetService», «OnlineSecurity», «OnlineBackup», «DeviceProtection» та «TechSupport»

3.2.4 Аналіз з точки зору контракту

Розглянемо, які особові характеристики вносять найбільший вклад в рішення про розірвання договору. З наявного набору даних це:

- Contract
- PaperlessBilling
- PaymentMethod

Приведені нижче діаграми дають уявлення про аспекти контракту, які можуть підвищити вірогідність відміни абонента :

- Більшість абонентів, які відміняють свою підписку, мають тип контракту "місяць в місяць" і включений безпаперовий білінг.

- Абоненти, у яких спосіб оплати - електронний чек, з більшою вірогідністю підуть.

```
df_churn['PaymentMethod'] = df_churn['PaymentMethod'].str.replace('(automatic)',
''').str.replace('(', '').str.replace(')', '').str.strip()

fig = displayCountPlot(contract_attributes, df_churn, rows=1, columns=3)

fig.savefig("contract.png", dpi=600)
```

Рис. 3.16 Побудова та відображення графіків статистики за атрибутами «Contract», «PaperlessBilling» та «PaymentMethod»

3.2.5 Наявність незбалансованих даних

Стовпець 'Churn' - це цільовий вектор, який використовуватиметься для навчання ML-моделей. Клас 'No' має значно більше записів, чим клас 'Yes', що свідчить про те, що набір даних сильно незбалансований. У ідеалі набір даних має бути збалансованим, щоб уникнути перевантаження моделей.

```
print(df_clean[df_clean['Churn'] == 'No'].shape[0])
print(df_clean[df_clean['Churn'] == 'Yes'].shape[0])

"""fig, ax = plt.subplots()
sns.countplot(df_clean['Churn'], palette=palette)

plt.xticks(size=12)
plt.xlabel('Churn', size=12)
plt.yticks(size=12)
plt.ylabel('# Customers', size=12)"""

displayCountPlot(['Churn'], df_clean, rows=1, columns=1, figsize=(5,3), export=True)
```

```
5174
1869
```

Рис. 3.17 Перевірка збалансованості даних

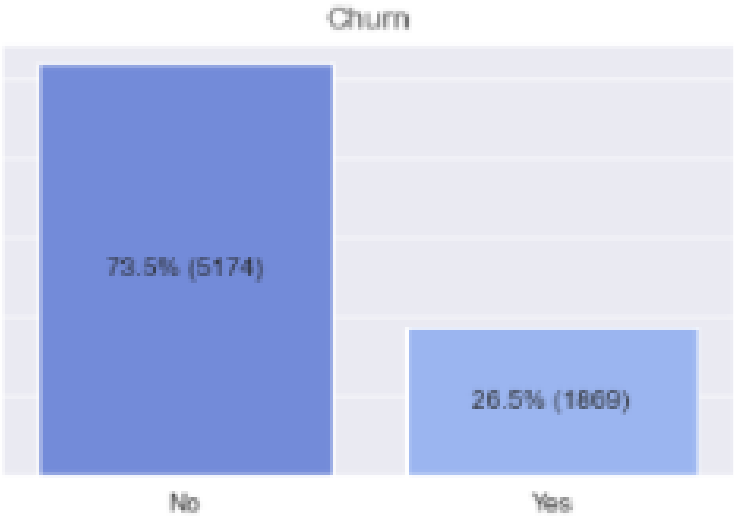


Рис. 3.18 Графік відношення «churned» і «not churned» абонентів.

3.2.6 Підготовка даних та кодування функцій

```
binary_feat = df_clean.nunique()[df_clean.nunique() == 2].keys().tolist()
numeric_feat = [col for col in df_clean.select_dtypes(['float','int']).columns.tolist() if col not in binary_feat]
categorical_feat = [ col for col in df_clean.select_dtypes('object').columns.to_list() if col not in binary_feat + numeric_feat ]

df_proc = df_clean.copy()
```

Рис. 3.19 Розподіл атрибутів на бінарні, числові та категоріальні.

Застосування кодування міток для бінарних атрибутів.

```
le = LabelEncoder()
for i in binary_feat:
    df_proc[i] = le.fit_transform(df_proc[i])
    print(i, '\n', np.unique(df_proc[f].values))
```

```
gender
 ['No' 'Yes']
SeniorCitizen
 ['No' 'Yes']
Partner
 ['No' 'Yes']
Dependents
 ['No' 'Yes']
PhoneService
 ['No' 'Yes']
PaperlessBilling
 ['No' 'Yes']
Churn
 [0 1]
```

Рис. 3.20 Перевірка кодування міток для бінарних атрибутів.

```
    print(categorical_feat)
    df_proc = pd.get_dummies(df_proc, columns=categorical_feat)
    print(df_proc.columns)
```

```
['MultipleLines', 'InternetService', 'OnlineSecurity', 'OnlineBackup', 'DeviceProtection', 'TechSupport',
'StreamingTV', 'StreamingMovies', 'Contract', 'PaymentMethod']
Index(['gender', 'SeniorCitizen', 'Partner', 'Dependents', 'tenure',
       'PhoneService', 'PaperlessBilling', 'MonthlyCharges', 'TotalCharges',
       'Churn', 'MultipleLines_No', 'MultipleLines_No phone service',
       'MultipleLines_Yes', 'InternetService_DSL',
       'InternetService_Fiber optic', 'InternetService_No',
       'OnlineSecurity_No', 'OnlineSecurity_No internet service',
       'OnlineSecurity_Yes', 'OnlineBackup_No',
       'OnlineBackup_No internet service', 'OnlineBackup_Yes',
       'DeviceProtection_No', 'DeviceProtection_No internet service',
       'DeviceProtection_Yes', 'TechSupport_No',
       'TechSupport_No internet service', 'TechSupport_Yes', 'StreamingTV_No',
       'StreamingTV_No internet service', 'StreamingTV_Yes',
       'StreamingMovies_No', 'StreamingMovies_No internet service',
       'StreamingMovies_Yes', 'Contract_Month-to-month', 'Contract_One year',
       'Contract_Two year', 'PaymentMethod_Bank transfer (automatic)',
       'PaymentMethod_Credit card (automatic)',
       'PaymentMethod_Electronic check', 'PaymentMethod_Mailed check'],
      dtype='object')
```

Рис. 3.21 Перетворення категоріальної змінної у фіктивні змінні.

3.2.7 Оброблений набір даних – готовий для навчання ML

```
    get_df_size(df, header='Original dataset:')
    get_df_size(df_proc, header='Processed dataset:')

    df_proc.head()
```

```
Original dataset:
# Attributes:  21
# Entries:  7043

Processed dataset:
# Attributes:  41
# Entries:  7043
```

	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	PaperlessBilling	MonthlyCharges	TotalCharges
0	0	0	1	0	1	0	1	29.85	29.85
1	1	0	0	0	34	1	0	56.95	1889.50
2	1	0	0	0	2	1	1	53.85	108.15
3	1	0	0	0	45	0	0	42.30	1840.75
4	0	0	0	0	2	1	1	70.70	151.65

5 rows × 41 columns

Рис. 3.22 Перевірка та відображення оброблених даних

Виконаємо розподіл навчальних і тестових даних.

```
# split df_proc in feature matrix and target vector
X=df_proc.drop('Churn', axis=1)
y=df_proc['Churn']

# split df_proc between train and test
X_train, X_test, y_train, y_test = train_test_split(X, y)
```

Рис. 3.23 Виконання розподілу навчальних і тестових даних.

```
# cross-validation function
def val_model(X, y, clf, quite=False):
    """
    Make cross-validation for a given model

    # Arguments
        X: DataFrame, feature matrix
        y: Series, target vector
        clf: classifier from scikit-learn
        quite: bool, indicate if funcion should print the results

    # Returns
        float, validation scores

    """

    X = np.array(X)
    y = np.array(y)

    pipeline = make_pipeline(StandardScaler(), clf)
    scores = cross_val_score(pipeline, X, y, cv=5, scoring='recall')

    if quite == False:
        print("##### ", clf.__class__.__name__, " #####")
        print("scores:", scores)
        print("recall: {:.3f} (+/- {:.2f})".format(scores.mean(), scores.std()))

    return scores.mean()

def getClfRecallScores(X_train, y_train, *clf_list):
    """
    Provides recall score gor a given list of models

    # Arguments
        X_train: X_train
        y_train: y_train
        *clf_list: list of classifiers

    # Returns
        DataFrame, recall scores

    """

    model_name = []
    recall = []

    for model in clf_list:
        model_name.append(model.__class__.__name__)
        recall.append(val_model(X_train, y_train, model))

    return pd.DataFrame(data=recall, index=model_name, columns=['Recall']).sort_values(
        by='Recall', ascending=False)
```

Рис. 3.24 Задання допоміжних функцій

Виконаємо балансування даних.

```
# under sampling
rus = RandomUnderSampler()
X_train_rus, y_train_rus = rus.fit_resample(X_train, y_train)
```

```
get_df_size(X_train, header='Before balancing:')
get_df_size(X_train_rus, header='After balancing:')

# make sure the number of classes are equal distibuted
np.unique(y_train_rus, return_counts=True)
```

```
Before balancing:
# Attributes:  40
# Entries:  5282

After balancing:
# Attributes:  40
# Entries:  2818

(array([0, 1]), array([1409, 1409], dtype=int64))
```

Рис. 3.25 Порівняння даних до та після балансування

Виконаємо стандартизацію даних.

```
# standardizing X_train and X_test
scaler = StandardScaler()
X_train = scaler.fit_transform(X_train)
X_train_rus = scaler.fit_transform(X_train_rus)
X_test = scaler.transform(X_test)
```

Рис. 3.26 Задання стандартизації даних

Створення базової лінії за допомогою перехресної валідації.

```
from sklearn.svm import SVC
from sklearn.tree import DecisionTreeClassifier
from sklearn.linear_model import LogisticRegression
from xgboost import XGBClassifier

# instaciate models
dt = DecisionTreeClassifier()
svc = SVC()
lr = LogisticRegression()
xgb = XGBClassifier()

df_scores = getClfRecallScores(X_train_rus, y_train_rus, dt, svc, lr, xgb)

print(df_scores)
```

Рис. 3.27 Використання перехресної валідації

	Recall
LogisticRegression	0.806943
SVC	0.797007
XGBClassifier	0.749461
DecisionTreeClassifier	0.694823

Рис. 3.28 Визначення коефіцієнтів надання вірного рішення

3.3 Навчання, налаштування і оцінка моделей машинного навчання

3.3.1 Налаштування моделей

Оскільки LogisticRegression і SVC показали кращі результати по метриці Recall, ми збираємося використати їх для налаштування гіперпараметрів і перевірити, чи зможе вона дати ще кращі результати.

3.3.1.1 Логістична регресія

Ми настроюватимемо 'solver' і 'C' в моделі Logistic Regression. Як видно нижче, після налаштування модель показала невелике поліпшення, збільшивши Recall з 0.80 до 0.83.

```
kfold = StratifiedKFold(n_splits=5, shuffle=True)

lr = LogisticRegression()

param_grid = {'solver': ['newton-cg', 'lbfgs', 'liblinear'],
              'C': [0.001, 0.01, 1, 10, 100]}

search = GridSearchCV(lr, param_grid, scoring='recall', cv=kfold)
result = search.fit(X_train_rus, y_train_rus)

print(f'Best recall: {result.best_score_} for {result.best_params_}')
```

```
Best recall: 0.8112364655107139 for {'C': 0.001, 'solver': 'liblinear'}
```

Рис. 3.29 Налаштування параметрів для збільшення коефіцієнта надання вірного рішення моделей

```
model_lr = LogisticRegression(solver='newton-cg', C=0.001)
model_lr.fit(X_train_rus, y_train_rus)
y_pred_lr = model_lr.predict(X_test)
lr_corr = confusion_matrix(y_test, y_pred_lr, normalize='true')
print(classification_report(y_test, y_pred_lr))
```

✓ 0.1s

```
              precision    recall  f1-score   support

           0       0.93      0.72      0.81      1314
           1       0.51      0.83      0.63       447

    accuracy                           0.75      1761
   macro avg       0.72      0.78      0.72      1761
weighted avg       0.82      0.75      0.77      1761
```

Рис. 3.30 Відображення звіту класифікації

3.3.1.2 SVM модель

Ми настроюватимемо 'kernel' і 'C' в моделі SVM. Після налаштування SVM значно покращав показник Recall, збільшивши Recall з 0.80 до 0.91, що є відмінним результатом.

```
param_grid = {'kernel': ['linear', 'poly', 'rbf', 'sigmoid'],
              'C': [0.001, 0.01, 1, 10, 100] }

search = GridSearchCV(SVC(), param_grid, scoring='recall', cv=kfold)
result = search.fit(X_train_rus, y_train_rus)

print(f'Best recall: {result.best_score_} for {result.best_params_}')
```

```
Best recall: 0.93257868302163 for {'C': 0.01, 'kernel': 'poly'}
```

```
model_svm = SVC(kernel='poly', C=0.01)
model_svm.fit(X_train_rus, y_train_rus)
y_pred_svm = model_svm.predict(X_test)
svm_corr = confusion_matrix(y_test, y_pred_svm, normalize='true')

print(classification_report(y_test, y_pred_svm))
```

```
              precision    recall  f1-score   support

           0       0.94      0.36      0.52      1301
           1       0.34      0.94      0.50       460

    accuracy                           0.51      1761
   macro avg       0.64      0.65      0.51      1761
weighted avg       0.78      0.51      0.51      1761
```

Рис. 3.31 Налаштування параметрів SVM моделі

Порівняння логістичної регресії та SVM моделі.

```
fig, ax = plt.subplots(nrows=1, ncols=2, figsize = (10,4))
fig.suptitle('Correlation Matrix')

sns.heatmap(svm_corr, annot=True, linewidths=.1, cmap=palette, ax=ax[0])
ax[0].set_title('SVM')
ax[0].set_ylabel('True')
ax[0].set_xlabel('Predicted')

sns.heatmap(lr_corr, annot=True, linewidths=.1, cmap=palette, ax=ax[1])
ax[1].set_title('Logistic Regression')
ax[1].set_ylabel('True')
ax[1].set_xlabel('Predicted')


plt.show()

fig.savefig("correlation_.png", dpi=600)
```

✓ 2.2s

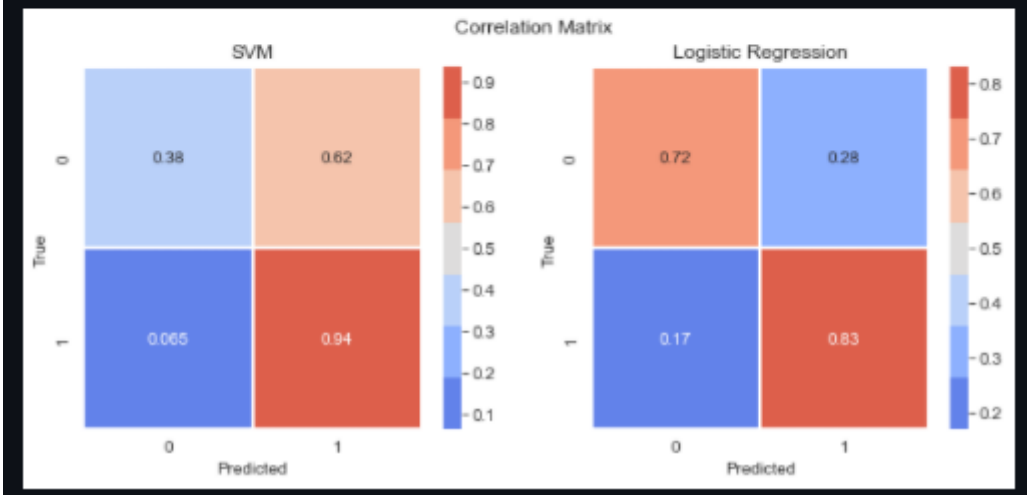


Рис. 3.32 Виведення результату навчання моделей

В результаті проведеного навчання моделей, було отримано значне зростання коефіцієнта надання вірного рішення для моделі SVM, який рівний 0.94.

Висновки

Жоден алгоритм не зможе передбачити відтік зі 100% точністю. В результаті виконання даного розділу було отримано високий ступінь надання вірних рішень щодо сприятливих до відтоку абонентів, який рівний 83%.

Оскільки метою є залучення абонентів і встановлення контакту з ними, щоб запобігти їх відтоку, то цілком допустимо взаємодіяти з тими, хто помилково помічений як "not churned", оскільки це не викличе жодних негативних проблем. Потенційно це може зробити їх ще більше задоволеними сервісом. Це та модель, яка може підвищити цінність з першого дня роботи, якщо зробити належні дії.

РОЗДІЛ 4
СТАРТАП-ПРОЕКТ

4.1 Вступ

Стартап як форма малого ризикового (венчурного) підприємництва впродовж останнього десятиліття набула широкого розповсюдження у світі через зниження бар'єрів входу в ринок (із появою Інтернету як інструменту комунікацій та збуту стало простіше знаходити споживачів та інвесторів, займатись пошуком ресурсів, перетинати кордони між ринками різних країн), і вважається однією із наріжних складових інноваційної економіки, оскільки за рахунок мобільності, гнучкості та великої кількості стартап-проєктів загальна маса інноваційних ідей зростає.

Проте створення та ринкове впровадження стартап-проектів відзначається підвищеною мірою ризику, ринково успішними стає лише невелика частка, що за різними оцінками складає від 10% до 20%. Ідея стартап-проекту, взята окремо, не вартує майже нічого: головним завданням керівника проекту на початковому етапі його існування є перетворення ідеї проекту у працюючу бізнес-модель, що починається із формування концепції товару (послуги) для визначеної клієнтської групи за наявних ринкових умов.

Розроблення та виведення стартап-проекту на ринок передбачає здійснення низки кроків, в межах яких визначають ринкові перспективи проекту, графік та принципи організації виробництва, фінансовий аналіз та аналіз ризиків і заходи з просування пропозиції для інвесторів. Узагальнено етапи розроблення стартап-проекту можна подати таким чином.

4.2 Етапи розвитку стартап-проекту

1. Маркетинговий аналіз стартап-проекту

В межах цього етапу:

- розробляється опис самої ідеї проекту та визначаються загальні напрями використання потенційного товару чи послуги, а також їх відмінність від конкурентів;
- аналізуються ринкові можливості щодо його реалізації на базі аналізу ринкового середовища розробляється стратегія;
- ринкового впровадження потенційного товару в межах проєкту.

2. Організація стартап-проєкту

В межах цього етапу:

- Складається календарний план-графік реалізації стартап-проєкту;
- Розраховується потреба в основних засобах та нематеріальних активах;
- Визначається плановий обсяг виробництва потенційного товару, на основі чого формулюється потреба у матеріальних ресурсах та персоналі;
- Розраховуються загальні витрати на запуск проекту та планові загальногосподарські витрати, необхідні для реалізації проєкту.

3. Фінансово-економічний аналіз та оцінка ризиків проєкту

В межах цього етапу:

- визначається обсяг інвестиційних витрат;
- розраховуються основні фінансово-економічні показники проекту (обсяг виробництва продукції, собівартість виробництва, ціна реалізації, податкове навантаження та чистий прибуток) та визначаються показники інвестиційної привабливості проекту (запас фінансової міцності, рентабельність продажів та інвестицій, період окупності проекту);
- визначається рівень ризикованості проекту, визначаються основні ризики проекту та шляхи їх запобігання (реагування на ризики).

4. Заходи з комерціалізації проекту

Цей етап спрямовано на пошук інвесторів та просування інвестиційної пропозиції (оферти). Він передбачає:

- визначення цільової групи інвесторів та опису їх ділових інтересів;
- складання інвест-пропозиції (оферти): стислої характеристики проекту для попереднього ознайомлення інвестора із проектом;
- планування заходів з просування оферти: визначення комунікаційних каналів та площадок та планування системи заходів з просування в межах обраних каналів;
- планування ресурсів для реалізації заходів з просування оферти.

Означені етапи, реалізовані послідовно та вчасно – створюють передумови для успішного ринкового старту. Проте фахівці зі створення та розвитку стартап-проектів окремо відзначають, що відсутність маркетингових знань та умінь, що уможливлюють розробку ринково затребуваного проекту із вихідної ідеї, є основною причиною високого рівня банкрутств стартап-компаній, і ця проблема може бути вирішена за рахунок навчання винахідників. Відповідно, основним призначенням даних Методичних рекомендацій є надання студентам знань щодо суті, основних принципів розроблення стратегії ринкового впровадження та маркетингового управління інноваційними стартап-проектами у промислових галузях економіки, використання ефективних маркетингових інструментів просування високотехнологічних продуктів виробництва та послуг.

4.3 Розроблення стартап-проєкту

В даному підрозділі викладено маркетинговий аналіз перспектив реалізації алгоритму передбачення відтоку абонентів оператора мобільного зв’язку, а також оцінено можливості її ринкового впровадження.

1. Опис ідеї проєкту

Проект направлений на підвищення швидкості аналізу та обробки білінг даних мобільних операторів, за рахунок впровадження машинного навчання в системи білінгу. Така система допомагає операторам мобільного зв'язку утримувати клієнтів, вчасно надавати необхідні пропозиції, отримувати інформацію про можливі зміни, визначати параметри, які мають найбільший вплив та зменшувати ризики їх реалізації.

Таблиця 4.1

Зміст ідеї	Напрямки застосування	Вигоди для користувача
Алгоритм машинного навчання для підвищення ефективності білінг системи MNO, що дає можливість зменшити ризик реалізації відтоку клієнтів.	Використання алгоритмів для підвищення ефективності системи білінгу. Виявлення параметрів на йбільшого ризику та клієнтів з високими показниками цих параметрів. Зменшення відтоку клієнтів за рахунок правильно підібраних пропозицій для кожного окремого клієнта.	Оператори мобільного зв'язку зможуть "перенести" витрати з неефективних ланок, на чітко визначених показників з урахуванням окремої категорії клієнтів, що надасть змогу залишити задоволених клієнтів та запобігти відтоку інших.

Для освітлення конкурентів у напрямі даного проекту, необхідно провести детальну оцінку ринку загалом, вийти за межі ринку телекомунікацій. Беручи до уваги те, що технології таких систем ще не до кінця впроваджені у більшості країн світу та факт того, що перехід до мережі нового покоління є процесом довготривалим та не чітко визначеним, можна зробити висновок, що наразі основних конкурентів на ринку не представлено, а існують лише локальні рішення, які не є можливо проаналізувати зовні. Тому основні переваги будуть представлені у наступному пункті розділу.

Таблиця 4.2

2. Технологічний аудит ідеї проєкту

№ п/п	Ідея проекту	Технології Реалізації	Наявність технологій	Доступність технологій
1	Система підвищення ефективності білінг-систем з метою надання рішень щодо клієнтів за групами та окремо	Мова програмування Python	Бібліотеки pandas, sklearn, numpy, matplotlib, xgboost, jupyter notebook google colab (що знадобляться для тестування на ранніх етапах тестування та розробки), що є у вільному доступі та не потребують зайвих витрат.	Є безкоштовними та знаходяться у вільному доступі
2		Мова програмування R	Бібліотеки схожі за функціоналом до бібліотек зазначених для Python, але потребують більших знань R. Open-source фреймворки для аналізу та обробки даних	Є безкоштовними та знаходяться у вільному доступі
У зв’язку з меншим порогом для реалізації та подальших етапів розвитку проекту, а також наявності гарних знань мови програмування Python, було вирішено обрати останню.				

Для реалізації проекту було розглянуто дві мови програмування: R та Python.

R є більш спеціалізованою мовою для побудови віртуалізації та візуалізації, аналізу та обробки даних. Також імплементована та широко використовується при створенні машинного інтелекту, Big Data, Data Science.

Python в свою чергу надає можливість знайти рішення за допомогою значно більшого комьюніті, більших об’ємів документації та створених на її базі проектів для вивчення, а також не є менш привабливою для машинного навчання та штучного інтелекту за рахунок наявності перенесених бібліотек та фреймворків з R та більшості інших мов, що цілком влаштовує необхідності даного проекту.

3. Аналіз ринкових можливостей запуску стартап-проекту

При дослідженні ринкових можливостей, в першу чергу проведений аналіз попиту: наявність, обсяг, динамік розвитку ринку. Дані внесені у таблиці нижче.

Таблиця 4.3

№ п/п	Показники стану ринку	Характеристика
1	Кількість головних гравців, од	Немає?
2	Загальний обсяг продаж	?
3	Динаміка ринку (якісна оцінка)	Зростає
4	Наявність обмежень для входу	Немає
5	Специфічні вимоги для стандартизації, специфікації	Немає
6	Середня норма рентабельності в галузі, %	?

Враховуючи швидкий розвиток технологій 5G та необхідність ринку у швидких рішеннях не тільки у зовнішніх процесах, але й у внутрішніх, можна зробити висновок, що даний ринок є дуже привабливим для входження нових

гравців та здорової конкуренції.

Таблиця 4.4

Характеристика потенційних клієнтів стартап-проекту

№ п/п	Потреба, що формує ринок	Цільова аудиторія	Відмінності у поведінці різних потенційних цільових груп клієнтів	Вимоги споживачів до товару
1	Перехід до мереж	Оператори мобільного зв’язку	Оператори матимуть змогу збільшити ефективність системи білінгу, виявити основні недоліки щодо лояльності з клієнтами, визначити показники, що сприяють відтоку клієнтів та зменшити ризики подібного у майбутньому.	Легка інтегрованість в уже існуючу систему, легкість та зручність підтримки, можливість візуалізації та налаштованість системи

До факторів загроз можна віднести лише складність даних, але використання технологій машинного навчання та впровадження штучного інтелекту зводять цю загрозу до мінімуму.

Таблиця 4.5

Фактори можливостей

№	Фактор	Зміст можливості	Можлива реакція компанії
1	Необхідність у нових рішеннях	Через невисокий рівень провадження мереж 5G на даний момент, але однозначне закінчення переходу у близькому майбутньому, є величезна група MNO, готових до пришвидшення цього процесу, утримання й залучення нових клієнтів також буде вдосконалено через вивчення потреб та надання таргетованих пропозицій.	Просування продукту на загальний ринок з метою оцінки та порівняння продукту з аналогами з метою подальшого вдосконалення продукту та залучення клієнтів на перших етапах.
2	Збільшення кількості даних про абонентів	При роботі оператора виникає необхідність у зборі, аналізі та обробці даних за якнайменш короткий час, що унеможливлює їх обробку мануально.	Висвітлення переваг продукту щодо збільшення ефективності обробки даних.

Таблиця 4.6

Альтернативи ринкового впровадження стартап-проекту

№	Опис профілю цільової групи потенційних клієнтів	Готовність споживачів сприйняти продукт	Орієнтовний попит в межах цільової групи	Інтенсивність конкуренції в сегменті	Простота входу у сегмент
1	Всеукраїнські оператори моб. зв.	Переважно готові	Дуже високий	Низька	Легко
2	Оператори моб. зв. за кордоном	Переважно готові	Дуже високий	Низька	Важко

Цільовими групами обрано Всеукраїнських MNO, з можливістю подальшого переходу до міжнародних також, що буде можливо при високій зацікавленості у продукті на території України, за рахунок малої/відсутності конкуренції на даній території.

Залежно від міри сформованості галузевого ринку , характеру конкурентної боротьби, необхідно обрати одну з трьох стратегій конкурентної поведінки : розширення первинного попиту, оборонну або наступальну стратегію або ж застосувати демаркетинг або диверсифікацію (таблиця 4.8).

Таблиця 4.7

Визначення базової стратегії розвитку

№	Обрана альтернатива розвитку	Стратегія охоплення ринку	Ключові конкурентоспроможні позиції	Базова стратегія розвитку
1	Динамічний розвиток з використанням маркетингу та встановлення бізнес контактів без та з використанням власних	Підняття рейтингу компанії шляхом маркетингу на ключових площадках використання та освітлення проектів використання телеком технологій та машинного навчання/штучного інтелекту, розвитку технологій/систем 5G	Підвищена гнучність та налаштованість пропонованої системи, легкість впровадження та підтримка системи впродовж перших років використання, переваги щодо підвищення ефективності обробки даних та наявність віртуалізації даних на всіх рівнях системи	Стратегія лідерства по витратах
2	Динамічний розвиток завадки висвітленню унікальних характеристик надання послуг	Унікальність послуг, архітектури системи, можливість впровадження в уже існуючу систему та ін.	Використання передових алгоритмів машинного навчання на основі вже існуючих з виділенням найбільш ефективних шляхів щодо реалізації продукту	Стратегія диференціації

Таблиця 4.8

Визначення базової стратегії конкурентної поведінки

№	Чи є проект першопрохідцем	Чи буде компанія шукати нових споживачів, або забирати існуючих у конкурентів	Чи буде компанія копіювати основні характеристики	Стратегія конкурентної поведінки
1	Проект не є першопрохідцем, але не має чітких аналогів на ринку	Компанія буде шукати нових користувачів	Компанія буде брати до уваги найкращі з х-к конкурентів та буде розробляти свої аналоги глибоко імплементовуючи їх в систему для максимального рівня підтримки	Стратегія наслідування лідеру за для економії фінансових ресурсів

Таблиця 4.9

Визначення стратегії позиціонування

№	Вимоги до товару	Базова стратегія	Ключові конкурентоспроможні позиції власного стартап-проекту	Вибір асоціацій, які мають сформувати комплексну позицію власного проекту
1	Висока доступність	Стратегія диференціації	Використання технології машинного навчання та можливості впровадження штучного інтелекту, що надасть можливість максимально автоматизувати процеси	Масштабованість, прозорість, якість, швидкість та доступність

4.4 Розроблення маркетингової програми стартап-проекту

Маркетингова програма - це намічений для планомірного здійснення, об’єднаний єдиною метою та залежний від певних строків комплекс взаємопов’язаних завдань і адресних заходів соціального, економічного, науково технічного, виробничого, організаційного характеру з визначенням ресурсів, що використовуються, а також джерел одержання цих ресурсів.

Основну увагу слід приділяти вибору, значенню та формі інструментів маркетингу, їх об’єднанню в найбільш оптимальний з погляду визначеної мети комплекс, а також розподілу фінансових ресурсів у межах бюджетування маркетингу.

Першим кроком є формування маркетингової концепції товару , який отримає споживач. Для цього потрібно підсумувати результати попереднього аналізу конкурентоспроможності товару (таблиця 4.10).

Таблиця 4.10

Визначення ключових переваг концепції потенційного товару

№	Потреба	Вигода, яку пропонує товар	Ключові переваги перед конкурентами
1	Можливість обробки та аналізу великих об’ємів даних за короткий час для підтримання актуалізії знань про ситуацію на момент часу	Можливість, що забезпечується новітніми алгоритмами, моделями та методами машинного навчання для автоматизації та підвищення ефективності системи вцілому	Комплексний підхід до рішення цієї задачі, можливість імплементації у вже існуючу систему, масштабованість, гнучкість, візуалізація результатів, т.д.
2	Визначення параметрів, що впливають на відтік та притік клієнтів	Надає змогу контролювати потік клієнтів, визначати найбільш ефективні та розвивати їх у подальшому	Візуальне подання результатів для швидшого сприйняття та прийняття мануальних рішень, але передбачені й автоматизовані на основі вже існуючих даних, зібраних при виконанні тієї чи іншої задачі.

Продовження таблиці 4.10

3	Визначення параметрів, впливаючих на ризики втрат того чи іншого ресурсу (наприклад, вибір тарифного плану при внутрішніх змінах всередині MNO)	Визначення таких параметрів відбувається на основі попередньо вивчених даних при навчанні та подальшій обробці нових даних, з них система контролює ризики можливості подібних ситуацій у майбутньому та можливість виникнення нових подій при заданих умовах мануально	Візуалізація даних, можливість генерування та моделювання подій, зміна даних та проведення дослідження зміни параметрів та виникнення ризиків

Надалі розробляється трирівнева маркетингова модель товару: уточнюється ідея продукту та послуги, його фізичні складові, особливості процесу його надання (таблиця 4.11).

Таблиця 4.11

Опис трьох рівнів моделі товару

Рівні товару	Сутність та складові
1. Товар за задумом	Товар забезпечує передбачення вибору абонентом тарифного плану при зміні умов системи білінгу з впровадженням мережі 5G та не обмежується на цьому

Продовження таблиці 4.11

2. Товар у реальному виконанні	Основні властивості: масштабованість, швидкість, якість, доступність, можливість до візуалізації даних
	Товар представляє собою систему(програмний комплекс) створений на мові програмування Python та суміжних бібліотеках для обробки даних.
	Система поставляється у вигляді застосунку для інтеграції у системи білінгу шляхом самостійної оптимізації до існуючого ПЗ
	Назва: ML Prediction and Visualization ToolKit for 5G Telecommunication Systems
3. Товар із підкріпленням	До продажу: відбувається моделювання інсталяції та конфігурування системи у системі клієнта, проведення тренінгів та консультацій
	Після продажу: підтримка продукту для клієнта, впровадження нових функцій та вдосконалення попередніх

Аналіз системи збуту передбачає визначення ефективності кожного елемента цієї системи, оцінювання діяльності апарату працівників збуту. Аналіз витрат обігу передбачає зіставлення фактичних збутових витрат за кожним каналом збуту і видом витрат із запланованими показниками для того, щоб виявити необґрунтовані витрати, ліквідувати затрати, що виникають у процесі руху товарів і підвищити рентабельність наявної системи збуту.

Дані щодо визначення системи збуту надаються в таблиці 4.12.

Таблиця 4.12

Формування системи збуту

№	Специфіка закупівельної поведінки цільових клієнтів	Функції збуту, які має виконувати постачальник	Глибина каналу збуту	Оптимальна система збуту
1	Власна система збуту	Проведення та розгортання ПЗ на стороні компанії клієнта	Канал нульового рівня, продаж товару відбувається безпосереднь споживачам через відділ збуту	Оптимальною є система прямого збуту з каналом нульового рівня за відсутності посередників

Таблиця 4.13

Концепція маркетингових комунікацій

№	Специфіка поведінки цільових клієнтів	Канали комунікацій	Ключові позиції, обрані для позиціонування	Завдання рекламного повідомлення
1	Консервативна поведінка, але відкриті до нового	Соцмережі професійного прямування, корпоративна пошта, моб. зв’язок	Можливості отримання передбачень щодо прийняття рішень абонентами, ефективні аналіз та обробку даних, їх візуалізація	Висвітлити унікальні характеристики товару/продукту

У якості концепції маркетингових комунікацій були обрані інтегровані маркетингові комунікації, де компанія ретельно обмірковує і координує роботу своїх численних каналів комунікації, рекламу в засобах масової інформації, особистий продаж, стимулювання збуту, пропаганду, прямий маркетинг, упаковку товару.

Висновки

В даному розділі був проведений маркетинговий аналіз перспектив реалізації системи передбачень сприятливих до відтоку абонентів оператора мобільного зв’язку та загального підвищення ефективності системи білінгу, управління абонентами, а також проведене оцінювання можливостей її ринкового впровадження.

В результаті дослідження було визначено, що існує можливість ринкової комерціалізації проекту в першу чергу завдяки використанню технологій машинного навчання, що дозволяє наділити продукт унікальними властивостями, такими як швидкість, масштабованість, гнучкість, візуалізованість даних, якість отриманого рішення та великі можливості щодо покращення проекту, збільшення функціоналу системи.

Конкурентна ситуація надає перспективи впровадження продукту, так як продукція товарів-аналогів на ринку майже не представлена, має лише частковий функціонал реалізованої системи та володіє низкою критичних недоліків в якості відсутності масштабованості, гнучкості, адаптованості, через які рівень довіри до них залишається незадовільним з боку нових клієнтів. В результаті цього, існуючі товари-аналоги не створюють прямої конкуренції на міжнародному ринку та ринку України. Основною проблемою є можливе негативне ставлення всеукраїнських операторів мобільного до стартап-проекту як до потенційно ненадійного партнера.

Проведений аналіз підтверджує, що подальша імплементація проекту є доцільною.

ЗАГАЛЬНІ ВИСНОВКИ ПО РОБОТІ

В даній роботі було проаналізовано використання алгоритмів машинного навчання для передбачення відтоку абонентів оператора мобільного зв’язку. Зокрема було розкрито загальні проблеми відтоку, основні причини та методи запобігання, а також управління абонентами з можливістю впровадження технологій машинного навчання.

Проведено загальний огляд відтоку абонентів, визначено основні проблеми та можливі рішення, проведено порівняння основних можливостей його передбачення. В роботі описано проблему наявності відтоку, основні причини та протидії, управління абонентами при наявному відтоку, розглянуто основні способи надання послуг абонентам для зменшення відсотку відтоку.

При аналізі методів та алгоритмів машинного навчання також було проведено загальний огляд, виділено основні задачі, методи, моделі та алгоритми, а також розкрито способи його використання в промисловості. Слід додати, що також було розглянуто математичні та статистичні моделі машинного навчання, моделі кластеризації. Приведено узагальнений алгоритм проекту машинного навчання.

В роботі виділено основні моделі класифікації для проведення аналізу їх використання: логістична регресія та SVM-модель. Також визначено сторонню бібліотеку SKLearn для побудови моделей на основі машинного навчання, оскільки вона має переваги у вигляді проведення тестування, прогнозування та підгону моделей. Було використано бустинг XGBoost для пришвидшення навчання моделей, в результаті чого було підвищено вірність надання рішень для обох моделей.

Було описано роботу функцій використаних бібліотек, створено функції для спрощення розрахунків та побудов графіків для покращення візуального сприйняття. Також описано значення основних метрик моделей, які визначають

ймовірність проведених прогнозів моделей, визначено доцільність використання цих метрик в окремих випадках.

На прикладі аналізу моделей прогнозування сприятливості абонента оператора мобільного зв’язку до відтоку можна зробити висновок про необхідність впровадження технологій машинного навчання для значного пришвидшення аналізу та обробки величезних об’ємів даних, а головне – з високою точністю.

Результати даних досліджень можуть бути використані для підбору операторами необхідних послуг, тарифних планів абонентів в цілях отримання максимальної їх якості щодо можливостей та потреб абонентів, а також зменшення відсотка відтоку за рахунок надання своєчасних індивідуальних пропозицій. Також є доцільним використання результатів даних досліджень в навчальних дисциплінах з машинного навчання.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Bloch D. Machine Learning: Models and Algorithms [Електронний ресурс] / Daniel Bloch // SSRN. – 17. – Режим доступу до ресурсу: ssrn.com/abstract=3307566.

2. MANUEL EUGENIO MOROCHO-CAYAMCELA. Machine Learning for 5G/B5G Mobile and Wireless Communications: Potential, Limitations, and Future Directions [Електронний ресурс] / MANUEL EUGENIO MOROCHO-CAYAMCELA, HAEYOUNG LEE, WANSU LIM // IEEE Access – Режим доступу до ресурсу: https://www.researchgate.net/publication/335937022_Machine_Learning_for_5GB5G_Mobile_and_Wireless_Communications_Potential_Limitations_and_Future_Directions.

3. Матеріали сайту tutorialspoint. [Електронний ресурс] – Режим доступу до ресурсу: https://www.tutorialspoint.com/telecom-billing/billing-introduction.htm.

4. Матеріали сайту scikit-learn. [Електронний ресурс] – Режим доступу до ресурсу: https://scikit-learn.org/stable/.

5. Матеріали вільної енциклопедії «Вікіпедія» [Електронний ресурс] – Режим доступу до ресурсу: https://en.wikipedia.org/wiki/.

6. Матеріали сайту MSDN. [Електронний ресурс] – Режим доступу до ресурсу: https://docs.microsoft.com/en-us/dotnet/machine-learning/resources/tasks.

7. Géron A. Hands-On Machine Learning with Scikit-Learn and TensorFlow / Aurélien Géron.. – 525 с.

8. XGBoost: A Scalable Tree Boosting System [Електронний ресурс]. – 2016. – Режим доступу до ресурсу: https://arxiv.org/pdf/1603.02754.pdf.

9. Mbarek R., Baeshen Y. Telecommunications Customer Churn and Loyalty Intention. *Marketing and Management of Innovations* – 2019 – № 4. – с. 110–117. Режим доступу до ресурсу: https://doi.org/10.21272/mmi.2019.4-09.

10. Winning the war against churn: How to use churn prediction techniques to improve customer retention – Режим доступу до ресурсу: http://www.invest-

aura.com/2011/02/winning-the-churn-reduction-war-how-to-use-churn-prediction-techniques-to-improve-customer-retention/.

11. Telecommunication subscribers' churn prediction model using machine learning. In: Eighth international conference on digital information management / S. Qureshii та ін – 2013.
12. Karp A. H. Using logistic regression to predict customer retention.
13. Schleicher M., Iyengar R., Ascarza E. The perils of proactive churn prevention using plan recommendations: evidence from a field experiment – 2016.
14. Brandusoiu I., Toderean G. Methods for churn prediction in the prepaid mobile telecommunications industry – 2016.
15. Strategies for Reducing Churn Rate in the Telecom Industry – *www.omnisci.com* – – Режим доступу до ресурсу: https://www.omnisci.com/blog/strategies-for-reducing-churn-rate-in-the-telecom-industry.
16. Customer Churn: The Experts Data Driven Guide to Churn – *www.profitwell.com* – Режим доступу до ресурсу: https://www.profitwell.com/customer-churn/guide.
17. Customer Churn Prediction & Prevention For Business Survival – *www.profitwell.com* – Режим доступу до ресурсу: https://www.profitwell.com/recur/all/churn-prediction.