КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
ІМЕНІ ТАРАСА ШЕВЧЕНКА
Економічний факультет
Кафедра економічної кібернетики

КВАЛІФІКАЦІЙНА РОБОТА БАКАЛАВРА

«Гібридні моделі інтелектуального аналізу даних в управлінні готельним бізнесом»

студентки 4 курсу
спеціальності 051 «Економіка»
ОПП «Економічна кібернетика»
денної форми навчання
Кузьменко Альони Сергіївни

Науковий керівник:
доктор економічних наук,
професор
Чорноус Галина Олександрівна

Засвідчую, що у цій дипломній роботі немає запозичень із праць інших авторів без відповідних посилань
Студент_____
(підпис)

Роботу допущено до захисту перед ЕК
рішенням кафедри економічної кібернетики
від 12 червня 2023 р., протокол № 17
Завідувач кафедри:
доктор економічних наук, професор
Ляшенко Олена Ігорівна

_____
(підпис)

КИЇВ – 2023

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ

ІАД – інтелектуальний аналіз даних

ADR – average daily rate, середній відпускний тариф

AUC – area under curve

CE – classification error

DT – Decision Tree

FN – false negative

FP – false positive

K-NN – K-Nearest Neighbors

LSVM – Linear Support Vector Machine

NN – Neural Net

opt – optimized

ROC – receiver operating characteristic

SVM – Support Vector Machine

TN – true negative

TP – true positive

XGBoost – Extreme Gradient Boosted Trees

ВСТУП

Актуальність дослідження. Пандемія COVID-19 значно змінила світ індустрії гостинності, яка до цього показувала велике зростання. Незважаючи на велику кількість обмежень та отриманих збитків за попередні періоди, готельний бізнес має тенденцію до відновлення. Показники попиту та ефективності в сфері гостинності стабільно зростали протягом 2021-го та більшої частини 2022-го року [1]. RevPAR (прибуток на один вільний номер) зріс на 6,1% порівняно з 2019 р для Європи, на 8,3% - для Північної Америки. Водночас середня зайнятість номерів для європейських готелів впала на 10,5% (до 64,5%), а для Північної Америки – на 4,9% (до 62,5%) [2]. За таких умов готельний бізнес потребує ефективного ревеню-менеджмента – для будь-якої послугоорієнтованої індустрії продаж правильного продукту правильному клієнту у правильний час – ключ до отримання прибутку [3]. Водночас, не кожен клієнт є «правильним», ряд потенційних гостей скасовують бронювання на різних відрізках часу до заїзду. Така поведінка є причиною зменшення доходів, і скасування бронювання в останні хвилини приносить бізнесу значні збитки [4,5].

Анулювання або скасування бронювань, відсоток якого зріс із розквітом онлайн туристичних агенцій (Airbnb, Agoda, Booking), є фактором, який зменшує точність прогнозу попиту на номери. Зі свого боку, готелі вдаються до політики надмірного бронювання (підтверджують більше замовлень, ніж мають номерів), для того, щоб мінімізувати збитки від скасувань та незаселень [6]. Тим не менш, політика «овербукінгу» завдає пошкоджень репутації закладу, так само як і жорсткі політики анулювання бронювань, що висувають підвищені тарифи або штрафи в негативних випадках.

Вчасна класифікація бронювання як позитивне (те, що буде підтверджено заїздом та перебуванням) чи негативне допомагає бізнесу зменшити ризики та збитки. Цій темі присвятили свої роботи такі групи науковців, як Nuno Antonio, Ana de Almeida, Luis Nunes; Adli Abdillah Nababan, Miftahul Jannah, Arif Hamied Nababan тощо. [7, 8, 9]. Група дослідників [7] одними з перших розглянула прогнозування скасувань бронювань не в якості задачі регресії, а як

класифікаційну проблему. Вони застосували 5 різних моделей на незбалансованому наборі даних та отримали високу точність передбачень, від 87,9% до 93,8%. У роботі [9] дослідники використовували модель XGBoost, на незбалансованих історичних даних різних готелів із додаванням даних із зовнішніх джерел: погодні, дані міських подій, соціальної репутації готелю. Найвища отримана точність передбачення становила 93,8%. Друга група дослідників [8] врахувала меншу репрезентацію скасованих бронювань та використала для балансування даних алгоритм SMOTE. У результаті k-NN модель на даних із такою підготовкою дала точність, на 3,88% більшу ніж без неї. Більшість використаних моделей у наведених дослідженнях є слабкими класифікаторами. Відповідно у цій випускній кваліфікаційній роботі **мета дослідження** – систематизація теоретичних і практичних напрацювань із проблеми скасування бронювань і її прогнозування; обґрунтування використання інструментарію економіко-математичного моделювання для здійснення класифікації бронювань, зокрема гібридних моделей, а також визначення факторів, що на них впливають.

Об'єктом нашого дослідження є процеси управління попитом у готельному бізнесі. **Предмет дослідження**: економіко-математичні методи та моделі для прогнозування скасування бронювань, як складової процесу управління готельним бізнесом.

Задачі, які забезпечують досягнення мети кваліфікаційної роботи, наступні:

1. Визначення особливостей готельних послуг як основного продукту діяльності готелю, що безпосередньо впливають на управління готельним бізнесом і відповідно на практичну значимість прогнозування бронювань.
2. Узагальнення підходів до прогнозування скасування бронювань.
3. Створення візуалізації даних про готельний попит та їх підготовка до подальшого моделювання.
4. Тренування та оптимізація простих класифікаторів для використання їх в гібридному моделюванні.

5. Створення гібридних моделей: неоднорідного ансамблю, однорідних ансамблів послідовної та паралельної оптимізації.
6. Робота із моделлю XGBoost.
7. Оцінювання отриманих результатів моделювання.

Методи дослідження. У теоретико-методологічній частині роботи при розгляді основних понять готельного бізнесу були використані загальнонаукові методи дослідження, такі як аналіз, синтез. У практичному дослідженні підхід до моделювання був створений із використанням методів наукової абстракції, порівняння, індукції та дедукції. Підготовка даних та реалізація моделей класифікації була виконана завдяки економіко-математичним методам, зокрема інтелектуальному аналізу даних та машинному навчанню. Платформа RapidMiner стала основним середовищем виконання завдань дослідження, а візуалізація даних реалізована за допомогою бібліотеки Plotly мови програмування Python.

Наукова новизна дослідження полягає в систематизації теоретично-практичних здобутків із тематики скасувань бронювань, а саме: передумов їх виникнення, впливу на управління готельним бізнесом, підходів та результатів прогнозування події анулювання гостем бронювання. Представлено порівняння результатів використання слабких класифікаторів, оптимізованих слабких класифікаторів та відповідних гібридних моделей у задачі класифікації скасування бронювань; атрибути, що найбільш впливають на статус бронювання. Створено рейтинг релевантних моделей ІАД для класифікації бронювань, як тих, що будуть скасовані чи ні.

Структура дослідження. Кваліфікаційна робота бакалавра складається із вступу, трьох розділів, висновків та списку використаних джерел, який налічує 56 найменувань. Обсяг роботи 74 сторінки, серед яких наявні 23 рисунки і 21 таблиця.

РОЗДІЛ 1. ТЕОРЕТИКО-МЕТОДИЧНІ ЗАСАДИ УПРАВЛІННЯ ГОТЕЛЬНИМ БІЗНЕСОМ

1.1. Особливості функціонування готельного бізнесу та ознаки його ефективності

Готельний бізнес – складова індустрії гостинності, яка займається економічною діяльністю з розміщення, харчування, а також забезпеченням задоволення додаткових потреб (розважальних, професійно-ділових) туристів [10].

Сукупність готельних підприємств різних типів (готелів, мотелів, туристичних баз, кемпінгів і приватних оздоровчих санаторіїв), що здійснюють надання послуг з розміщення, харчування, а також надання додаткових і суміжних послуг особам, що перебувають за межами місця постійного проживання, називається готельним господарством [10].

Назву готельному господарству дала його головна структурна одиниця – готель – засіб розміщення, що надає туристам послуги розміщення та харчування у наступних формах: розміщення, розміщення та сніданок, розміщення та харчування у ресторані, кафе, барі. Відповідно Закону України «Про туризм» готель є засобом тимчасового розміщення, що складається із шести і більше номерів, та в якому надаються готельні послуги [11].

Готельна послуга – дії суб’єкта господарювання з тимчасового розміщення клієнта шляхом надання номера (місця) в готелі, а також додаткові послуги, що надаються залежно від категорії готелю: таке визначення надає Закон України «Про туризм» [11]. У пункті 1.3. наказу «Про затвердження Правил користування готелями й аналогічними засобами розміщення та надання готельних послуг» визначено, що готельна послуга складається з основних та додаткових послуг. Основні послуги включені до ціни номера (проживання, харчування, надання рушників тощо) за укладеним договором, а додаткові замовляються клієнтом окремо за додатковою домовленістю (організація екскурсій, продаж сувенірів, послуги прання речей) [12].

Серед особливостей готельних послуг виділяють наступні [9]:

1. Нематеріальний характер основної складової готельної послуги – полягає в невідчутності послуги, яка існує лише в процесі надання та споживання, а отже можливість осяжно оглянути послугу перед споживанням відсутня. Із боку бізнесу виникає проблема маркетингу послуги та залучення її споживачів: ані професійні фотосесії інтер'єрів номерів, ані візуальні 3D тури не дають повної оцінки якості послуги та її особливостей.

2. Одночасність процесів виробництва і споживання послуги, що невіддільні від її джерела та її об'єкта. Оскільки послуга – це неповторний досвід споживача, вона не може мати попередню, готову форму. Послуга формується у взаємодії, тому допущені помилки та недоліки неможливо виправити до моменту споживання.

3. Обмежена можливість зберігання полягає в неможливості накопичення готельних послуг, як основних так і додаткових. Теперішня можливість отримати дохід від вільних номерів сьогодні у разі низької заповнюваності перетвориться на збиток опівночі, оскільки «непродані» номери не додаються, і нової доби готель матиме ту саму кількість номерів, що і була вчора.

4. Терміновий характер готельних послуг. Готельна послуга має обмежений час користування, який фіксується датами і часом заїзду та виїзду. Відповідно, проблеми, що утворюються в процесі надання послуги, мають вирішуватися вчасно і з великою швидкістю. Водночас, не всі проблеми можуть вирішуватися: наприклад, скасування повітряного рейсу може спричиняти збитки як для бізнесу, так і для споживачів. Жодна із сторін не може вплинути на ситуацію, оскільки вона залежить від третіх осіб. Терміновість готельної послуги тісно пов'язана з її обмеженою можливістю зберігання. Анулювання – відмова замовника від заброньованих послуг, або неприбуття - фактичне незаселення замовників у день заїзду, або анулювання менш ніж за 24 години до дати заселення, значно підвищує ймовірність того, що послуга не буде надана іншому споживачу. Відповідно, бізнес понесе втрати на підготовку до її надання (прибирання кімнати, зміна білизни, замовлення харчування), яка не зберігається у часі, і має бути повторена із наступним споживачем.

5. Мінливий характер попиту на готельні послуги. Обсяг попиту на готельні послуги є і періодичним, і сезонним. Найбільш ця особливість виявляється у закладах розміщення, розташованих в місцях із сезонними змінами клімату, де один сезон є більш туристично привабливим за інший, а також у готелях в ділових центрах міст, де основний попит традиційно спостерігається у будні дні.

1.2. Ревеню-менеджмент та управління попитом

Ревеню-менеджмент, або управління доходами – інструмент узгодження попиту та пропозиції за допомогою поділу клієнтів на різні сегменти залежно від їхніх намірів щодо купівлі та розподілу виробничих потужностей між різними сегментами таким чином, щоб максимізувати доходи певного підприємства [13]. Ревеню-менеджмент також визначається як застосування інформаційних систем і стратегій ціноутворення для розподілу потрібної потужності потрібному клієнту за правильною ціною в правильний час [14,15]. Такий розподіл є однією з головних задач управління доходами. Відповідно ця задача пов’язує мистецтво управління доходами зі сферою маркетингу, де воно відіграє ключову роль у створенні попиту [16] і управлінні споживацькою поведінкою [17].

Ревеню-менеджмент виник не в готельному бізнесі, а в дотичній до нього індустрії пасажирських авіаперевезень у 70-х роках 20-го століття. Підхід швидко став популярним у різних галузях економіки, проте кожна сфера має свої особливості функціонування, якими визначаються практичні аспекти використання ревеню-менеджменту. [18].

На основі праць Kimes, 1998 [14] і Kimes & Wirtx, 2003 [15] болгарський дослідник Станіслав Іванов визначив ревеню-менеджмент у готелях як сукупність інструментів та дій, залучених на досягнення оптимального рівня чистого доходу та валового операційного прибутку готелю, за допомогою пропозиції правильного продукту правильному клієнту через правильний канал дистрибуції у правильний час за правильною ціною та правильною комунікацією [18]. Оптимальний рівень у цьому визначенні позначає не обов’язкову відовідність між чистим доходом від продажів та валовим операційним

прибутком, оскільки другий враховує витрати на обслуговування клієнтів: закупка товарів, маркетинг, залучення персоналу. Наприклад, залучення додаткових гостей може зменшити валовий операційний прибуток, якщо витрати на їх обслуговування будуть зависокі. До того ж, при повній завантаженості пересічний готель або інший заклад розміщення надає менше уваги кожному клієнту, що є базисом для операційних помилок, нестачі персоналізованих послуг, великих черг та інших наслідків, що зменшують задоволеність гостей, а звідти – їхню лояльність та прибутки закладу. Таким чином, метою управління доходами є не максимізація прибутків готелю будь-якою ціною, а одночасне досягнення найвищих доходів і валового операційного прибутку одночасно [18]. Не всі гості є однаково прибутковими для готелю, тому у визначенні постає поняття «правильного клієнту». Одні мають надто багато вимог, які готель не завжди може задовольнити із збереженням прибутку, другі – завдають матеріальних збитків власності готелю, треті - відмовляються від бронювання номерів в останній момент, або взагалі не з’являються на заселення, що також впливає на прибутки готелю. Відсоток останнього типу клієнтів зріс із появою нового каналу дистрибуції - онлайн туристичних агентств (скорочено ОТА від online travel agency), до яких відносяться популярні платформи типу Airbnb, Agoda, Booking.

Економічними передумовами потреби застосування ревеню-менеджменту до готельного бізнесу є:

1. Обмежений термін придатності продукту (Product perishability) [18].

Готельна послуга, чи готельний продукт – як сукупність матеріальних (номер, меблі, харчування, пральня) та нематеріальних (атмосфера, сервіс, мовна доступність персоналу, приязність) готельних послуг у випадку надлишковості не можуть бути збережені та передані на сезони більш високого попиту. Після граничного часу заїзду номер або місце не можуть бути продані, і прибуток від них у цей певний час є втраченим назавжди. Виходячи з цього, раціональні менеджери намагаються перерозподілити попит різними інструментами.

Наприклад, пропозицією інших дат проживання споживачам, що отримали відмову у бронюванні.

2. Обмежена місткість готелю [18].

Традиційно місткість закладу розміщення вимірюється в кількості ліжко-місць у певний період часу, зазвичай розраховують кількість ліжок на добу. Готель не може розмістити більше людей, ніж має в наявності ліжко-місць. Водночас, існує ще місткість системи харчування готелю, а також додаткових послуг – функціональних кімнат (аудиторних, переговорних), СПА-центрів, басейну, спортзалу тощо. У даному випадку пропускна здатність послуг залежить від операційних (час приготування страви, самообслуговування чи залучення кельнерів, тривалість занять тощо) факторів. У короткотривалому періоді місткість номерів є сталою, проте для зменшення операційних витрат на підтримання номерів у періоди низького попиту окремі корпуси чи поверхи готелів можуть закривати. На довготривалий період місткість готелю є змінною, хоча вимагає великих інвестицій у основні засоби.

3. Високі постійні та низькі змінні витрати [18].

До постійних витрат готелю належать амортизація засобів, обслуговування кредитів, заробітна плата адміністративних співробітників, комунальні платежі за місця загального користування. Витрати такого виду в середньому становлять 60-80% від загальної суми витрат готелю. Змінні витрати включають у себе заробітну плату персоналу першої лінії підтримки клієнтів, які безпосередньо взаємодіють з ними, витрати на продукти, електроенергію, воду та опалення номерів. Економічно готель може продовжувати обслуговувати клієнта поки граничний прибуток (MR) з гостя більше або дорівнює маржинальним витратам за його обслуговування (MC).

$$MR \geq MC \tag{1.1}$$

Маржинальний прибуток дорівнює ціні, сплаченій гостем за номер (P). Залучення нового гостя збільшує змінні витрати на величину змінних витрат на одну ніч (VC_1). Отримаємо наступну формулу:

$$P \geq VC_1 \tag{1.2}$$

З формул (1.1), (1.2) виходить, що готелю вигідно приймати певного клієнта якщо дохід, отриманий від нього, покриває, щонайменше, змінні витрати на його обслуговування.

Якщо ціна номеру більша за змінні витрати на обслуговування гостя, то цей прибуток може покрити частину сталих витрат, які будуть виникати незалежно від приймання цього клієнта. До того ж, граничний дохід формується не лише зі сплати за номер, до нього додається плата за користування клієнтом іншими сервісами готелю (у казино-готелях прибуток від азартних ігор перевищує прибуток від здачі номерів). Усі прибутки згенеровані гостем враховуються менеджерами готелів, тому в певних випадках ціни на номери можуть бути меншими за граничні витрати на обслуговування, оскільки додаткові прибутки компенсують низьку ціну. Менеджмент має постійно враховувати такі опції, що неможливо без якісного ревеню-менеджменту.

4. Нерівномірний характер попиту протягом різних відрізків часу [18].

Ця передумова сама собою є особливістю готельних послуг. Без сумнівів, рівномірний попит значно зменшив би потребу в ревеню-менеджменті та інших службах готелю, але такий стан є утопічним через велику кількість факторів, що впливають на обсяг попиту в цій індустрії.

5. Можливість прогнозування попиту, прибутку та завантаженості готелю [18, 19].

Можливість прогнозування попиту на готельні послуги є однією з найважливіших передумов розвитку ревеню-менеджменту в готельному бізнесі, оскільки всі рішення генерального менеджера або власне ревеню-менеджера базуються на припущеннях про туристичний попит. Задачами прогнозування попиту можуть бути ідентифікація періодів з найбільшим або найменшим рівнем попиту, прогнози на наступні 30, 45, 60, 90 діб, встановлення нових цінових стратегій та тарифів на номери. Важливим у цій передумові є якісне збирання та оцінювання історичних та поточних даних готелями у власних облікових управлінських системах, а також враховування локальних подій та заходів, що

можуть вплинути на попит на послуги розміщення в даному туристичному районі. Для прогнозів використовуються дані про кількість проданих номерів (rooms sold), завантаженість готелю (occupancy - частка проданих номерів від загальної кількості доступних), кількість подовжених номерів (stayovers), кількість заброньованих номерів, кількість номерів, для яких відмінили бронювання (cancelation) або не заселилися до кінцевої дати заїзду (no-shows), а також про особливості клієнтів та їхнього перебування в готелі.

Для управління готелем точний прогноз майбутнього попиту, прибутків та завантаженості є надзвичайно важливими із наступних причин [19]:

- Точна оцінка попиту на номери дозволяє керівництву більш ефективно планувати людські ресурси для обслуговування гостей. Клієнтоорієнтованість забезпечується уважним підходом та високим рівнем сервісу до кожного гостя, що потребує великої кількості персоналу.
- Прогнозування попиту допомагає організовувати поставки продуктів у необхідній кількості, що є однією з умов задоволеності клієнтів.
- Прогнози прибутків від продажу номерів дозволяють вищому керівництву оцінювати майбутню прибутковість готелю, ця інформація важлива для прийняття рішень про капітальні вкладення (добудова нових корпусів, заміна даху, оновлення меблів).
- Точні прогнози попиту на номери дозволяють ревеню-менеджерам приймати кращі рішення про управління цінами на продукти та послуги.

Далеко не всі прогнози є точними, оскільки бронювання номерів не означає їхній стовідсотковий викуп. Одночасно з тим відділи збуту та маркетингу значно покладаються на кількість заброньованих місць в прогнозуванні попиту. У результаті близько 20% доходу готелів втрачається через неврахування анулювань бронювання номерів системами управління [20].

6. Можливість сегментувати ринковий попит [18, 19].

Попит на готельний послуги не є однорідним, тому готелі застосовують різні практики ревеню-менеджменту, наприклад – дискримінаційну цінову політику, за допомогою сегментації клієнтів. У більшості готелів та готельних

мереж, гостей поділяють на три ринкових сегменти: transient - тимчасові, або випадкові гості, що залишаються на невеликий проміжок часу, не являються частиною великою груп, а також не пов'язані з умовами спеціальних контрактів; group – сукупності пов'язаних між собою гостей; special contract - клієнти, які скористалися спеціальними умовами або знижками, найчастіше до таких відносяться замовлення для відрядження співробітників, у тому числі екіпажів літаків [18, 19].

7. Різна цінова еластичність ринкових сегментів [18].

Ця передумова тісно пов'язана з попередньою: саме різна цінова еластичність попиту в різних сегментах дозволяє застосовувати цінову дискримінацію для різних сегментів споживачів. Еластичність попиту за ціною опосередковується такими факторами, як унікальність готельного продукту; неосвідомленість клієнта про пропозиції конкурентів; підняття соціального статусу клієнта в результаті користування послугою; особиста зацікавленість клієнта в певному готелі; затрати на розміщення незначними в розрізі загальних витрат на поїздку; перебування клієнта малу кількість ночей (гості, що бронюють номери на довгий термін, більш чутливі до цін).

8. Різна готовність споживачів платити [18].

У поведінковій економіці «willingness to pay», тобто готовність платити, визначається як максимальна ціна, за яку споживач згоден придбати продукт або послугу. Якщо вартість номеру більша, за цю суму – клієнт не буде користуватися послугою готелю, якщо менша – створюється споживацький надлишок, який є потенційною втратою прибутку для готелю. Для ефективної управлінської політики менеджмент враховує цей фактор при встановленні дискримінаційних цін.

9. Можливість попереднього бронювання [18].

Готельні послуги: номери, харчування, а також додаткові послуги на кшталт організації таксі від аеропорту до готелю, – можуть бути замовлені за тижні або місяці до дати заїзду. Бронювання забезпечує клієнтам впевненість в тому, що задовго до подорожі вони мають гарантію отримання послуги.

Резервуванням номерів (місць) займається спеціальний відділ бронювання, який є найчастіше належить до структури служби прийому та розміщення. Менеджер, що очолює відділ нерідко підпорядковується безпосередньо директору відділу обслуговування або директору відділу збуту, що підкреслює важливість бронювання для управління доходами готелю [10].

Саме існування такого явища як попереднє бронювання дозволяє готельному бізнесу розраховувати ресурси, застосовуючи ці дані в прогнозуванні попиту на номери.

1.3. Скасування бронювань та їх наслідки. Прогнозування скасування бронювань.

У сьогоденні анулювання бронювань стає нагальною фінансовою та логістичною проблемою готельного бізнесу [21]. За даними D-Edge Hospitality Solutions, у 2014-2018 роках середній відсоток скасування бронювань, зроблених в онлайн-середовищі, коливався між 32,5 та 41,3 відсотками. Найвищі показники спостерігалися в HRS та Booking Group, 51,7-66% та 43,4-50,9% відповідно. У 2019 р. D-Edge стверджували, що до 40% клієнтів анулювали свої бронювання [22]. У часи пандемії COVID-19 та викликаних нею обмежень, до 60% гостів відмінили свої бронювання [23]. У 2021-2022 роках середня частка анульованих бронювань, зроблених різними способами, залишалась високою, хоч і знизилася з 25 до 20%. [24].

Статистика досліджень говорить, що не лише погіршення погодних умов, хвороби або зміна планів можуть перешкоджати переходу гостя із статусу бронювання в статус заселення [4]. Причиною зростання відсотку скасованих бронювань в останні десятиліття є поява OTA (online travel agencies) – посередників, які допомагають закладам розміщення продавати свої послуги. Найвідомішими OTA стали Booking, Expedia та AirBnb. Завдяки таким перевагам над традиційними прямими бронюваннями (walk-in - з вулиці або за телефонним дзвінком) та бронюваннями з офіційних веб-сайтів готелів, як швидкість, мінімальність контактів, можливість сплачувати онлайн, вони

докорінно змінили підходи до процесу бронювання разом з процесами, що виникають у свідомості людини під час його здійснення [25]. Ці платформи перетворили індивідуальний, особливий досвід зваженого вибору із безпосереднім спілкуванням із співробітниками готелю на стандартизовану процедуру, яка має мету миттєво забронювати, відчуваючи ризик втрати вигідної пропозиції, а лише потім отримувати детальну інформацію та задавати питання [26].

На вибір ОТА як способу бронювання позитивно впливають наявність широкого обсягу інформації (розташування закладу розміщення на карті, фотографії з номерів, перелік зручностей номера тощо), довіра від інших гостей (наявність позитивних відгуків на сторонньому ресурсі викликає більше довіри, ніж на власному сайті готелю), а також ціна, яка є меншою за пряму пропозицію від закладу розміщення завдяки знижкам, які ОТА змушує готелі надавати. [21, 25]. Окрім того, такі системи надають у зручному вигляді інформацію про альтернативні послуги від інших постачальників на той самий період часу, і споживач може легко порівнювати ці пропозиції та обирати для себе найкращу [27]. Однак нерідко споживач не робить єдиний класичний вибір, що пов’язано із тим, що ОТА стимулюють готелі пропонувати бронювання з повним поверненням коштів (refundable) при скасуванні. Окрім того, вони змушують (оскільки аудиторія таких платформ уже звикла до бронювань із повним поверненням коштів у разі скасування) заклади розміщення долучатися до політики free-cancellation (скасувань бронювань без штрафів – готель, який має лише безповоротні тарифи на номери, не знімає коштів із гостя в разі відміни бронювання, а отримує на заміну іншого клієнта на ці дати від ОТА чи компенсацію, якщо агентство не зможе знайти такого). Політики free-cancellation ОТА використовують як маркетинговий інструмент збільшення кількості клієнтів та їх лояльності, гість може бронювати паралельно декілька номерів у різних закладах розміщення, врешті-решт, залишаючи той, який йому підходить найбільше за ціною, або взагалі перебронювуючи той самий номер, якщо ціна стала меншою за попередню, а інші резервування анулювати на різних термінах

до дати заїзду [7, 26, 28]. За таку можливість мати відчуття контролю за розміщенням (на випадок зміни обставин) або обрати найбільш вигідну пропозицію, споживачі згодні сплачувати страхову премію – refundable бронювання зазвичай дорожчі, за інші тарифи.

Номери, бронювання яких скасовані, а особливо з загальної маси ті, де ануляція відбулася напередодні до заїзду, скоріш за все не будуть заповнені іншими клієнтами. До того ж, номери тривалий час були вилучені з бази доступних для бронювання, тому готель міг втратити потенційний клієнтів, які через недоступність бажаного варіанту в цьому готелі зробили бронювання в конкурента, до того ж, прибирання номеру напередодні заїзду здійснене, продукти для організації харчування замовлені, що також є збитками в разі незаселення. Врешті-решт, зменшується як точність прогнозу попиту, так і прибутки бізнесу.

Для мінімізації втрат та збільшення своїх прибутків у випадках з високими показниками скасованих бронювань готельний бізнес, як і інші туристичні, зокрема авіаційні пасажиро-перевезення, застосовує політики надмірного бронювання (overbooking) одних й тих самих типів помешкань, із розрахунком на те, що певна кількість клієнтів не скористаються послугою (скасують бронювання або не з’являться на поселення). Однак така практика має свої недоліки: у ситуаціях, коли заїжджає більша кількість гостей, ніж прогнозувалося, виникає потреба в пропозиції заселення в інший номер (із грошової компенсацією або додатковими послугами, або в номер вищої категорії) чи безкоштовного трансферу до іншого готелю. Клієнту навіть можуть відмовити в заселенні, і компенсація такої поведінки готелями, на відміну від сфери авіаперевезень, законодавчо не є обов’язковою, і виконується лише у судовому порядку. Використання такого інструменту як надмірне бронювання є високим репутаційним ризиком, адже не варто говорити про формування лояльності в клієнта, якому відмовили в заселені на порозі закладу розміщення. Більш того, такий випадок набуде розголосу у відгуках та обговореннях на профільних форумах [19].

Звісно, окрім бронювань з повним поверненням коштів (refundable), існують part-refundable та non-refundable – часткове повернення та безповоротне бронювання відповідно. У випадку part-refundable бронювань готель утримує із сплаченої ціни певну суму, визначену умовами бронювання, наприклад, ціну першої ночі, а залишок повертається гостю. У non-refundable бронюваннях вся сума залишається закладу розміщення, більш того – refundable бронювання після певного терміну з моменту їх створення можуть перетворюватися на non-refundable, тобто без штрафів та з поверненням повної вартості анулювати резервування можна лише обмежений час.

Зазвичай, non-refundable бронювання є дешевшими: такі ціни готель встановлює задля привернення уваги гостя. Водночас, вони є ризиковими для клієнтів, оскільки не враховують ситуації хвороби, поганої погоди чи скасування транспортних квитків тощо.

Із скасуванням бронювань можна боротися не тільки покращуючи власні веб-сайти, застосовуючи non-refundable тарифи та стратегії овербукінгу, а й встановлюючи особистий контакт з клієнтом у засобах зв’язку, надаючи йому персональну увагу та пропозиції. Водночас, для попередження анулювання бронювання важливо вчасно класифікувати гостя як схильного до скасування чи до збереження резервування номеру.

1.4. Огляд досліджень із прогнозування скасувань бронювань у готельному бізнесі

Разом з часткою скасувань бронювань у сфері гостинності в галузі наук про туризм в останні роки зросла кількість робіт із тематики із прогнозування цієї проблеми.

Водночас на початку виникнення досліджень із скасувань бронювань в індустрії гостинності, дослідники Morales and Wang, 2010 [29] висунули твердження про малу ймовірність можливості високоточно передбачити чи буде скасоване бронювання чи ні, лише спираючись на PNR (Passenger name record–

стандартна форма запису даних про пасажира в авіа-індустрії) інформацію: у своїй роботі вони прогнозували середній та сезонний відсоток скасувань бронювань.

У 2013 р. Huang, Chang [30] спростували вище наведене твердження, щоправда, на даних бронювань у мережі ресторанів Тайваню, передбачення-класифікації були зроблені за допомогою нейронної мережі зворотного поширення та узагальнено-регресійної нейронної мережі.

Португальські дослідники N. António, A. Almeida, L. Nunes, 2017 [7] одними з перших звернули увагу на проблему скасування бронювань саме у готельному бізнесі з погляду задачі класифікації, а не регресії. Дослідниками були відібрані показники, що найбільше впливають на статус бронювання: країна походження гостя, кількість запрошених парко-місць, агент, який робив бронювання та тип депозиту.

У 2018 р. була здійснена спроба розгляду проблеми скасування бронювань як задачі класифікації, а не регресії, на даних з готельного бізнесу у Van Leeuven, 2018, який великою мірою спирався на досягнення попередньої групи дослідників Класифікація була здійснена за допомогою байесівських мереж, логістичної регресії, дерева рішень та випадкового лісу. Точність моделей випадкового лісу коливалася від 0,778 до 0,890 завдяки відбору показників та їх інженерії. [30,31].

Falk, Vieru, 2018 у своїй роботі підтвердили гіпотези про зв'язок між способом бронювання та вірогідністю його скасування: найвища для OTA та найнижча для бронювань, зроблених через оффлайн туристичні агенції; раннє бронювання збільшує ймовірність анулювання бронювання; ймовірність скасуваннння вища під час високого сезону тощо [32].

Antonio, et al., 2019 окрім стандартних даних із PMS-систем (категорії Adult, ADR; Children, CustomerType, Deposit тощо) використали для прогнозування за допомогою моделі XGBoost дані із зовнішніх джерел, такі як середня кількість прогнозованих опадів під час потенційного гостювання, пов'язаність бронювання з подією, соціальна репутація готелю тощо. Максимальна точність моделей склала 93,8% [8].

Здобуток Sánchez-Medina & Eleazar [20] полягає у прогнозуванні анулювання бронювань за допомогою штучних нейронних мереж, оптимізованих генетичним алгоритмом для підбору параметрів. Результати були валідизовані за допомогою випадкових підвибірок через чисельний поділ набору даних на тренувальні та тестувальні вибірки. Точність ANN-GA моделі становила 98%.

Точність прогнозування скасування бронювань гіперпараметрично оптимізованої моделі випадкового лісу у роботі Y. Azhar, G. A. Mahesa, and M. C. Mustaqim ” становила 87% [33].

Adli Abdillah Nababan, Miftahul Jannah, Arif Hamied Nababan, 2022 [9] було здійснено прогнозування скасування бронювань на незбалансованому наборі даних за допомогою алгоритмів Synthetic minority oversampling technique та K-nearest neighbours. Дослідники відзначили ефективність SMOTE в роботі з незбалансованим датасетом, використання цього алгоритму підвищило точність K-NN моделі на 3,88% від 79,35% до 83,23%.

Більшість дослідників пропонують підходи, які основуються на використанні єдиного класифікатора. Водночас прості моделі, що складаються з єдиного алгоритму, традиційно мають свої недоліки, тому в останні роки популярність гібридних моделей значно зросла. Гібридний підхід активно використовують в ІАД та машинному навчанні переважно в таких сферах, як економіка, діагностика фізичних та ментальних хвороб, обробка зображень, матеріалознавство, енергоефективність. Оскільки дані готельного попиту характеризуються значною розмірністю та неоднорідністю, нами було вирішено застосувати гібридні моделі до туристичної галузі, а саме у прогнозуванні скасувань бронювань.

РОЗДІЛ 2. ГІБРИДНІ МОДЕЛІ ІАД В УПРАВЛІННІ ГОТЕЛЬНИМ БІЗНЕСОМ

2.1. Теоретико-методологічні основи застосування гібридних моделей ІАД

Гібридний підхід до інтелектуального аналізу даних полягає в розробці моделей, що використовують комбінації різних методів та алгоритмів ІАД, які зберігаючи свою силу, компенсують недоліки одне одного.

Гібридні моделі машинного навчання складаються з двох чи більше різних алгоритмів, які мають підвищити точність моделі у порівнянні з точністю моделей, реалізованих за допомогою відповідних простих алгоритмів.

Гібридний підхід включає в себе алгоритмічні комбінації, що містять набір моделей, які спільно використовуються для вирішення завдання, і гібридні комбінації, де за спільного вирішення комплексних завдань передбачається розподіл сфер відповідальності між методами [34].

Алгоритмічні композиції – базові поняття гібридного підходу, складаються з алгоритмічних операторів $b_i: X \to R,\ i = \overline{1,k}$, коригуючих операцій $F: R^k \to R$ та правил розв'язання $C: R \to Y$, реалізовані алгоритмом $a: X \to Y$ вигляду $a(x) = C\left(\left(F(b_1(x), \dots, b_k(x))\right)\right),\ x \in X.$ R інтерпретується як простір оцінок. Відповідно базові алгоритми – суперпозиції $a_i(x) = C\left(b_i(x)\right), i = \overline{1,k}$. Залежно від виду коригуючих операцій виділяють такі різновиди алгоритмічних композицій, як ансамблі моделей та суміші алгоритмів, у яких коригування відбувається на основі функції компетентності, яка змінює вигляд залежно від предметної галузі [35, 36].

Гібридна комбінація, що складається з алгоритмів $a_i: X \to D, i = \overline{1,k}$ a_i, координуючої операції $G: D^k \to D$ та правила розв'язання $C: D \to Y$ є суперпозицією $s(x) = C\left(G\left(a_1\left(x^{(1)}\right), \dots, a_k\left(x^{(k)}\right)\right)\right), x^{(i)} \in X \cup D, i = \overline{1,k}$, де D інтерпретується як простір виходів окремих алгоритмів [36].

Серед причин, які надають перевагу реалізації комбінацій виділяють наступні: усереднення помилок індивідуальних гіпотез і зменшення впливу випадкових записів і нестабільності; підвищення шансів знаходження глобального оптимуму, завдяки комбінуванню результатів, отриманих різними

моделями на різних вибірках; розширення множини гіпотез завдяки їх комбінації.

Водночас із реалізацією гібридних моделей виникають відповідні проблеми: зростання вимог до обчислювальної техніки, часових витрат, підбір оптимальних параметрів, складність інтерпретації та методів комбінації результатів, отриманих сильними моделями [35, 37].

Загалом простір гібридних моделей може бути представлений як комбінація двох прогнозних алгоритмів, один алгоритм та оптимізаційний метод, який максимізує його прогнозну здатність, або ансамбль (комітет) моделей – сукупність моделей, що може бути як однорідною (моделі одного типу), так і неоднорідною. Результат прогнозу визначається такими різновидами голосування як просте, зважене, за старшинством або монотонною коригуючою операцією [38].

Існують такі стратегії побудови алгоритмічних комбінацій :

1. Послідовна оптимізація

Базові алгоритми в такого роду алгоритмічних комбінаціях будуються послідовно, один за одним і кожен наступний намагається компенсувати недоліки попереднього. Одним з найпоширеніших методів послідовної оптимізації є бустинг, або підсилювання – процес формування ряду моделей, де кожна наступна використовує для навчання ті приклади, на яких попередня припустилася помилок, таким чином слабкі моделі перетворюються на сильні.

Загальна процедура бустингу полягає в наступному. Нехай у задачі біноміальної класифікації тренувальні записи $X \in D$, функція достовірності f складається з трьох підвибірок X_1, X_2, X_3, кожна з яких складає $\frac{1}{3}$ від вибірки, а класифікаційна помилка випадкового класифікатору – 50%. Маючи слабкий класифікатор h_1, який правильно класифікує записи з X_1, X_2 і неправильно – із X_3, отже класифікаційна помилка становить $\frac{1}{3}$. Створимо нову вибірку D' із D, яка фокусується на записах X_3. Нехай новий класифікатор h_2 тренується на вибірці D', припустимо, що він правильно класифікує приклади з X_1 та X_3, і

неправильно із X_2. При комбінуванні класифікаторів h_1 та h_2 отримаємо класифікатор, що правильно класифікує записи в X_1 і невелику кількість помилок в X_2 та X_3 Третій класифікатор h_3 на вибірці D'' правильно класифікує записи X_2 та X_3. Якщо скомбінувати всі три класифікатори h_1, h_2, h_3 отримаємо ідеальний класифікатор [39]. Підсилювання є універсальним та гнучким методом із високою узагальнюючою здатністю, водночас потребує значного обсягу тренувальної вибірки та схильність до перенавчання (overfitting), коли модель показує добрі результати на тренувальній вибірці і погані на тестувальній [35].

Існують різні види бустингу. Однією з найпершим моделей був адаптивний бустинг – послідовна оптимізація, за якої кожному набору даних початково присвоюється однакова вага, а в наступних ітераціях вона збільшується для тих записів, які передбачаються неправильно, і зменшується дли тих, що були класифіковані правильно. Адаптивний бустинг в середовищі RapidMiner реалізується оператором AdaBoost.

Іншим підходом до підсилювання моделі є градієнтний бустинг, за якого неправильно класифікованим записам не надається більша вага, а оптимізація проходить із мінімізацією функції втрат у кожному циклі за допомогою використання градієнтного спуску. Градієнтний спуск – алгоритм оптимізації, в якому для знаходження локального мінімуму функції предиктори оновлюються у напрямку антиградієнту (протилежного значення векторів-стовпчиків, елементами яких є частинні похідні першого порядку) цільової функції [40, 41]. Градієнтний бустинг зазвичай показує більш точні результати, ніж адаптивний.

2. Паралельна оптимізація

Налаштування базових алгоритмів відбувається паралельно незалежно один від одного на різних підвибірках із тренувальної вибірки, або на різних наборах ознак. Типовими методами паралельної оптимізації є беггінг, метод випадкових підпросторів та ансамблі генетичних алгоритмів. Перевага методів паралельної оптимізації полягає в можливості використання паралельних обчислень, що дозволяє збільшувати швидкість тренування моделей, використовуючи багатоядерні процесори [39].

Бегінг (bagging) або бутстрепова агрегація – процес паралельного навчання алгоритмів одного типу на різних вибірках одного розміру, які отримані за допомогою процедури випадкового відбору записів із вихідної вибірки, при чому одні записи можуть дуплікуватися, а інші не будуть представлені у згенерованих вибірках. Агрегування результатів слабких алгоритмів відбувається на основі голосування для задачі класифікації та на основі середнього для регресії. Поліпшення точності ансамблю забезпечується зменшенням розкиду в зашумлених наборах даних. На відміну від бустингу, процедура є ефективною на малих вибірках. Водночас вона вимагає значних обчислювальних затрат, ефективно працює лише на слабких алгоритмах та важко інтерпретується [39, 42].

Метод випадкових підпросторів (RSM від “random subspace method”) – навчання алгоритмів на різних підвибірках із загальної вибірки характеристик об’єкту, застосовується у випадках великої кількості атрибутів і малої кількості записів [35,36].

Ансамблі генетичних алгоритмів – оптимізація базових алгоритмів за допомогою генетичних алгоритмів – алгоритмів, що наслідують еволюційний процес, оптимізація забезпечується трьома процесами: генерацією нових поколінь, відбором, інтеграцією.

3. Глобальна оптимізація.

Становить об’єднання кількох моделей класифікації та регресії за допомогою мета-класифікатора бо мета-регресора. Прикладом є процедура стекінгу – комбінування моделей різних видів, які навчаються наступним чином. Вибірка поділяється на тренувальну та тестову, на якій відповідно слабкі класифікатори навчаються та тестуються. Отримані результати стають входами, а справжні (помічені) дані – виходами для навчання мета-алгоритму [43].

Для створення ансамблю в роботі було використане голосування за більшістю – процес об’єднання кількох слабких моделей в єдину, шляхом вибору найбільш повторюваного прогнозу класифікаторів для певного запису.

Загальний вигляд процесу голосування за більшістю можна представити формулою (2.1.) [44],

$$C(X) = \arg\max_i \sum_{j=1}^{k} w_j I(h_j(X) = i), \quad (2.1)$$

де X – вибірка значень;

h_j – правило класифікації;

w_j - вага слабкого класифікатору;

I – функція-індикатор.

2.2. Обґрунтування програмних інструментів проведення дослідження

Більша частина дослідження, а саме підготовка даних та реалізація моделей виконана в RapidMiner – це платформа для виконання задач науки про дані, у тому числі інтелектуального аналізу даних. Вона підтримує увесь процес роботи із даними: від візуалізації та очищення до оцінки продуктивності та порівняння моделей.

На відміну від численних аналогів, RapidMiner має однорічну безкоштовну ліцензію для здобувачів освіти, що робить його доступним для використання без особливих умов.

Окрім цього, платформа характеризується наступними перевагами, потрібними для реалізації процесу ІАД. у тому числі гібридного моделювання [54]:

- User-friendly інтерфейс: RapidMiner має зручний графічний інтерфейс, який дозволяє користувачам візуалізувати процес моделювання: від діаграми потоку до технічної документації та вмісту репозиторіїв. В одному робочому вікні середовища за потреби можна відобразити усі потоки, створені в ході ІАД.
- Широкий вибір інструментів підготовки даних: RapidMiner пропонує широкий набір інструментів підготовки та попередньої обробки даних. Ці інструменти дозволяють користувачу очищати та перетворювати необроблені

дані, обробляти відсутні значення, нормалізувати змінні та балансувати вибірку, не маючи широкого технічного досвіду та знань мов програмування.

- Різноманітні методи моделювання: платформа підтримує широкий спектр алгоритмів моделювання, включаючи різні види регресій, дерева рішень, випадковий ліс, опорні векторні машини, нейронні мережі та ансамблеві методи. Функціонал RapidMiner постійно оновлюється, щоб надавати користувачу найактуальніші можливості роботи з даними.
- Надійна перевірка та оцінка: RapidMiner надає комплексні інструменти для перевірки та оцінки моделі. Він підтримує різні методи перехресної перевірки, розбиття вибірки на навчання та обчислення показників продуктивності.
- Інтеграція з іншими інструментами та платформами: RapidMiner легко інтегрується з іншими популярними інструментами та платформами для вивчення даних, такими як R та Python. Це дозволяє користувачам використовувати наявні робочі процеси та бібліотеки для своїх цілей. Наявність бібліотеки XGBoost як розширення платформи RapidMiner значно полегшує процес тренування та валідації такої комплексної моделі. Окрім даного розширення, програмне середовище підтримує пакети Weka, Toolbox, Interpretation тощо.
- Масштабованість і продуктивність: RapidMiner розроблено для обробки великих наборів даних і ефективного виконання обчислень. Він використовує можливості паралельної обробки, розподілені обчислення та обробку даних у пам’яті для забезпечення швидкої та масштабованої побудови моделі.

Проте створення візуалізацій вхідних даних у RapidMiner ускладенене у порівнянні із візуалізацією за допомогою мови програмування та бібліотек Python. Тому частина ілюстративного матеріалу кваліфікаційної роботи (рис.3.1-3.7) написана мовою Python із використанням бібліотек у середовищі розробки Jupiter Notebook. Порівняно с RapidMiner візуалізація даних у Python має наступні переваги:

- Великий вибір бібліотек, їх широкий функціонал. Наприклад, графік "Агрегована кількість успішних та скасованих бронювань помісячно"

(рис.3.2) типу Stacked Bar у RapidMiner будується за допомогою складних налаштувань, у той час як у бібліотеці Plotly це стандартний тип графіків.

- Можливість кастомізації. Наприклад, у графіку 3.3 "Агрегований ADR помісячно" можна легко досягти відображення місяців по їх календарному порядку, а не алфавитному (за замовчуванням), за допомогою простої функції сортування.
- Велика кількість документації, статей та прикладів, які полегшують процес написання коду.

2.3. Опис застосованих моделей та алгоритмів

Для розуміння особливостей процесу моделювання, варто розглянути теоретико-методологічні основи моделей і алгоритмів, а також операторів, які їх реалізують в середовищі RapidMiner.

На нашу думку варто почати із найбільш використовуваного в роботі методу, який використовувався для оцінки продуктивності моделей. Це Leave-one-out cross-validation – техніка перехресного затвердження з виключенням по одному, яка дозволяє тренувати та оцінювати модель на різних підвибірках даних, у тому числі оцінювати узагальнюючу передбачувальну здатність моделі. Техніка застосовується у наступному порядку:

1. Вибірка поділяється на K підвибірок
2. Одна підвибірка обирається як тестова.
3. K-1 підвибірок, що залишилися, використовуються як тренувальні.
4. Процес кросс-валідації застосовується K разів, під час кожного одна із вибірок використовується для тестування, а інші – для тренування.
5. Результат K ітерацій усереднюється, так утворюється єдиний класифікатор.

У середовищі RapidMiner для реалізації методу був використаний оператор Cross-validation, який має два вкладені субпроцеси – тренування та тестування, за останнім з яких визначається продуктивність моделі. Кількість вибірок за

замовчуванням – 10, метод вибірки – перетасування (shuffled sampling) – вибірка складається з випадкових елементів [45].

У прогнозуванні відміни бронювань використовувалися 3 види простих класифікаторів: Decision Tree, k-NN, Neural Net. Розглянемо принципи роботи кожного з них.

Decision Tree, або дерево ухвалення рішень – непараметричний класифікатор, який базується на сукупності логічних правил «якщо – то», та представляє собою орієнтований граф-дерево, що складається із таких елементів як вузол (node) та листя (leaf). Вузли містять логічні правила перевірки відповідності запису певній умові, а листя вказують на клас категоріальної змінної. За допомогою дерева рішень також вирішуються задачі регресії, де листя вказує на середнє значення кількісної змінної.

Дерева рішень будуються за жадібним алгоритмом максимізації приросту інформації, де на кожному поділі для формування вузла обирається ознака, за якою у процесі розподілу приріст інформації виявляється найбільшим [46]. Дерево будується від кореня до листя рекурсивно, допоки ентропія не зменшиться до якоїсь малої величини або нуля.

У процесі створення дерева виникає питання вибору критеріїв розщеплення та зупинки навчання. На ентропійному підходу ґрунтується критерій information gain measure – міра інформативності підпросторів атрибутів. Індекс Gini – інший критерій розщеплення, за яким вибір атрибуту розщеплення проходить на підставі відстаней між розподілами класів [47]. Індекс Gini визначається за формулою (2.2.):

$$gini(T) = 1 - \sum_{j=1}^{n} p_j^2, \qquad (2.2)$$

де T – поточний вузол,

n – кількість класів,

p_j – ймовірність класу j у вузлі T.

Зупинка побудови дерева означає, що певний внутрішній вузол стає кінцевим і далі поділ вже не здійснюється, і визначається одним із правил зупинки. Зупинка побудови дерева може здійснюватися як заданням

обмеженням його глибини, так і мінімальної кількості записів, які будуть знаходитися в листях.

Класифікатор DT широко використовується завдяки своїй простоті, наочності, легкій інтерпретованості, низьких витратах на реалізацію, а також відсутності високих вимог до якості даних.

K-найближчих сусідів – один з найпростіших та найпоширеніших алгоритмів кластеризації, який також використовується в задачах класифікації. В основі методу лежить алгоритм порівняння прогнозованого запису із відомими, які є до нього найбільш близькими. Клас запису визначається мажоритарним голосуванням його сусідів у кількості K. Відстань між числовими атрибутами найчастіше розраховується за наступними алгоритмами [48]:

Евклідова відстань – традиційна формула відстані між двома точками у скінченновимірному просторі, що обраховується за формулою (2.3.)

$$d_{ij} = \sqrt{\sum_{k=1}^{m}(x_{ik} - x_{jk})^2}, \quad (2.3)$$

де d_{ij} – відстань між i-им та j-им записом;

x_{ik} – значення k-го атрибуту у i-го запису;

x_{jk} - значення k-го атрибуту у j-го запису.

Канберрська відстань – формула відстані (2.4) між двома точками, що часто використовується для порівняння впорядкованих списків та атрибутів із значеннями, близькими до нуля. Метрика обраховується наступним чином

$$d_{ij} = \sum_{k=1}^{m} \frac{|x_{ik} - x_{jk}|}{|x_{ik}| + |x_{jk}|}, \quad (2.4)$$

де d_{ij} – відстань між i-им та j-им записом;

x_{ik} – значення k-го атрибуту у i-го запису;

x_{jk} - значення k-го атрибуту у j-го запису.

Відстань Чебишева – формула максимальної абсолютної різниці між значеннями по кожному атрибуту запису, розраховується за формулою (2.5)

$$d_{ij} = \max_{k} |x_{ik} - x_{jk}|, \quad (2.5)$$

де d_{ij} – відстань між і-им та j-им записом;

x_{ik} – значення k-го атрибуту у і-го запису;

x_{jk} - значення k-го атрибуту у j-го запису.

Окрім переваг, алгоритм к-найближчих сусідів має ряд недоліків. Він чутливий до шумів та викидів даних, задачі біноміальної класифікації потребують непарного вибору кількості сусідів для уникнення ситуацій неоднозначності, за яких кількість сусідів, що належить різним класам, рівна. Незважаючи на відсутність фази тренування, алгоритм повільно працює на великих обсягах вибірки [47].

Штучні нейронні мережі – моделі, які відтворюють принцип роботи біологічних нейронних мереж живих організмів; складаються із штучних нейронів, що є їхніми основними обчислювальними елементами, і мають кілька входів та один вихід, який є входом для наступного шару нейронів.

ШНМ має наступну будову [46]:

1. Вхідний шар, який має зв’язок із вхідними даними.
2. Внутрішній (прихований) шар або декілька шарів. Нейронні мережі, у яких відсутній внутрішній шар, можуть моделювати виключно лінійні функції.
3. Вихідний шар, який відповідає за реалізацію вихідних даних.

Архітектура нейронних мереж (кількість нейронів, шарів, функції активації) може дуже варіюватися в залежності від вирішуваного завдання, тож розглянемо різновид нейронної мережі, застосованої в дослідженні.

У моделюванні в середовищі RapidMiner була використана нейронна мережа прямого зв’язку, де зв’язки між елементами не формують направлений цикл, тобто інформація рухається лише вперед від вхідних вузлів через внутрішні до вихідних без петлів або циклів. Вона була навчена алгоритмом зворотнього поширення, де фази поширення та оновлення ваги повторюються поки нейронна мережа не стане достатньо продуктивною. Вихідні значення порівнюються із правильно визначеними класами, щоб обчислити значення попередньо визначеної функції помилки. Завдяки цій інформації, алгоритм коректує ваги кожного зв’язку із метою зменшити значення функції помилки на

деяку малу величину. Цей процес повторюється до моменту, коли похибка обчислень стає достатньо малою [49].

У моделі використовується сигмоїдна функція як функція активації, що вимагає нормалізації значень атрибутів у межах [-1;+1].

Обмеженням використання нейронної мережі є тривалість часу, необхідного для її навчання, проте його можна регулювати кількістю циклів тренування моделі. Окрім того, результати моделювання за допомогою нейронних мереж є складними для інтерпретації, водночас для них характерна вища точність прогнозу у порівнянні з статистичними моделями [50].

XGBoost – модель ефективної екстремальної реалізації алгоритму градієнтного бустингу дерев рішень, розроблена в рамках науково-дослідного проекту Tianqi Chen. XGBoost є гібридною моделлю-ансамблем, точність передбачення якої вища за продуктивність градієнтного бустингу завдяки наступним покращенням [51, 52].

- L1 та L2 регуляризація цільової функції допомагає контролювати складність моделі та запобігає виникненню перенавчання завдяки використанню часткових похідних другого порядку в якості апроксимації функції втрат, що дозволяє отримати більше інформації про напрямок градієнту [51].
- Підрізування дерева за допомогою параметру максимальної глибини видаляє непотрібні гілки з дерев у процесі навчання, що робить моделі компактнішими і зрозумілішими.
- Застосування паралельних обчислень прискорює процес навчання моделі: у сортуванні даних використовуються паралельні потоки.
- Автоматизована обробка пропущених значень заповненням в залежності від значення втрат забезпечує економію часу та підвищення точності моделі [53].

З іншого боку налаштування параметрів моделі XGBoost більш складне, ніж класичної моделі Gradient boost machine або Випадкового лісу, воно потребує великих інформаційно-обчислювальних можливостей.

Серед параметрів, які можна обрати для реалізації моделі XGBoost в аналітичній платформі RapidMiner виділимо наступні:

- booster – вид бустингу; gbtree, dart – засновані на деревах і gblinear – на лінійній функції.
- nthread – кількість паралельних потоків;
- num_feature – зменшення багатовимірності даних;
- learning rate – темп навчання;
- min_split_loss – мінімальне зменшення втрат, необхідне для реалізації подальшого поділу;
- max_depth – максимальна глибина дерева;
- min_child_weight – мінімальна кількість записів у листі;
- subsample – частка тренувальних записів.
- sampling_method - метод вибірки: uniform – однакова ймовірність для кожного запису бути обраним, subsample – позитивні приклади становлять більше рівне за 0,5, gradient_based – ймовірність вибору кожного запису пропорційна регуляризованому абсолютному значенню градієнтів.

РОЗДІЛ 3. РЕАЛІЗАЦІЯ ГІБРИДНИХ МОДЕЛЕЙ

3.1. Опис та візуалізація даних

Датасет Hotel1 містять історичні дані, що описують характеристики бронювань готельних послуг в період з 1-го липня 2015-го року по 31 серпня 2017-го, у курортному готелі в Португалії.

Набір даних 31 атрибут та записів – 40 060 записів. Опис атрибутів набору даних поданий у збереженому з джерел порядку у табл. 3.1.

Табл. 3.1

Опис атрибутів датасету Hotel1

Назва атрибуту	Тип атрибуту	Опис атрибуту
IsCanceled	Categorical	Чи було бронювання скасовано (1), чи ні (0)
LeadTime	Integer	Кількість днів між датою введення бронювання в систему управління готелем та датою заїзду
ArrivalDateYear	Integer	Рік заїзду
ArrivalDateMonth	Nominal	Місяць заїзду: 12 можливих значень
ArrivalDateWeekNumber	Integer	Порядковий номер тижня року: 53 можливих значення
ArrivalDateDayOfMonth	Integer	День заїзду: 31 можливе значення
StaysInWeekendNights	Integer	Кількість вихідних днів (субота або неділя), які гість провів або мав провести в готелі.
StaysInWeekNights	Integer	Кількість робочих днів (із понеділка по п’ятницю), які гість провів або мав провести в готелі.
Adults	Integer	Кількість дорослих
Children	Integer	Кількість дітей: як діти, за яких тариф сплачений, так і діти, за яких не сплачений
Babies	Integer	Кількість дітей до року
Meal	Nominal	Тип харчування. Undefined/SC – без харчування; BB – Bed & Breakfast – сніданок; HB – Half board (сніданок та вечеря); FB – Full board (сніданок, обід та вечеря).
Country	Nominal	Країна походження гостя у форматі ISO 3155–3:2013
MarketSegment	Nominal	Позначення ринкового сегменту бронювання: TA – туристичні агенти, TO – туристичні оператори
DistributionChannel	Nominal	Канал бронювання. Аналогічно - TA та TO
IsRepeatedGuest	Integer	Бронювання створено гостем, який вже був у готелі (1) чи ні (0)
PreviousCancellations	Integer	Кількість бронювань, скасованих перед теперішнім.

Продовження табл. 3.1.

Назва атрибуту	Тип атрибуту	Опис атрибуту
PreviousBookingsNotCanceled	Integer	Кількість бронювань, не скасованих перед теперішнім
ReservedRoomType	Nominal	Тип зарезервованого номеру, закодований літерами латинського алфавіту (A, B, C, D, E, F, G, H, I)
AssignedRoomType	Nominal	Тип призначеного номеру, закодований літерами латинського алфавіту(A, B, C, D, E, F, G, H, I). Може відрізнятися від типу забронюваного номеру через причини пов'язані з наявністю місць та/або запитом клієнта
BookingChanges	Integer	Кількість змін у бронюванні, зроблених в період від введення бронювання в систему до моменту заїзду або скасування бронювання
DepositType	Nominal	Наявність та тип депозиту, який вніс клієнт для гарантування бронювання: No Deposit – відсутність депозиту; Non Refund – депозит рівний або більший за вартість бронювання; Refundable – депозит менший, за повну вартість бронювання
Agent	Nominal	Код туристичного агента, який здійснював бронювання
Company	Nominal	Код компанії, яка робила бронювання або відповідальна за його оплату
DaysInWaitingList	Integer	Кількість днів перед підтвердженням бронювання готелем
CustomerType	Nominal	Тип бронювання згідно наступної класифікації: Contract - бронювання пов'язано з договором з певної компанією; Group – бронювання пов'язане з групою; Transient – бронювання не є частиною групи або контракту та не пов'язане з іншим тимчасовим бронюванням; Transient-party – бронювання є тимчасовим, але пов'язане з іншим тимчасовим бронюванням
ADR	Real	Середній відпускний тариф
RequiredCarParkingSpaces	Integer	Кількість замовлених паркомісць
TotalOfSpecialRequests	Integer	Кількість особливих запитів від клієнту
ReservationStatus	Nominal	Останній статус бронювання: Canceled – скасовано; Check-out – виїзд, Now-show – не заселення
ReservationStatusDate	Data-time	Дата, коли було оновлено останній статус бронювання.

Джерело: узагальнено автором на основі [55]

Для покращення розуміння даних наведемо візуалізацію деяких атрибутів набору. Серед загальної кількості бронювань, 72,2% були успішними, а 27,8% скасовані, що відображено на рис. 3.1.

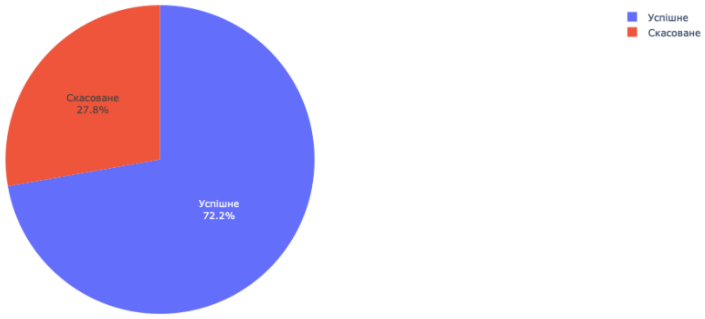

Рис. 3.1. Діаграма розподілу бронювань за статусом

Джерело: створено автором на основі [55] у середовищі розробки мовою програмування Python

Із гістограми, зображеної на рис. 3.2 встановлюємо, що найбільша кількість бронювань здійснюється у теплу пору року із квітня по жовтень. Лідерами є липень та серпень: зростає як кількість успішних бронювань, так і скасованих. Найменший відсоток скасованих бронювань характерний для непопулярних місяців – листопада, грудня та січня.

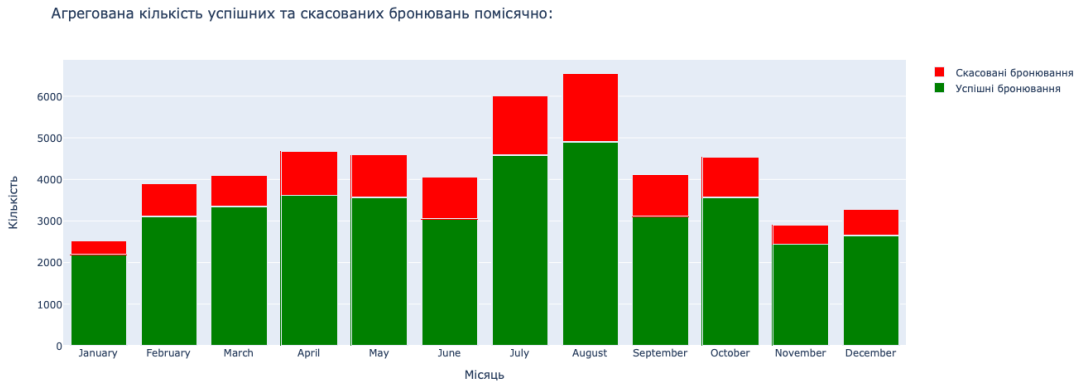

Рис. 3.2. Агрегована кількість успішних та скасованих бронювань помісячно

Джерело: створено автором на основі [55] у середовищі розробки мовою програмування Python

На графіку з рис.3.3. бачимо середній відпускний тариф за номер у розрізі місяців року. Він найвищий у місяці морського курортного сезону із піком в серпні, і спадає до листопада. У грудні ADR піднімається, ймовірно, у зв'язку із святкуванням католицького Різдва.

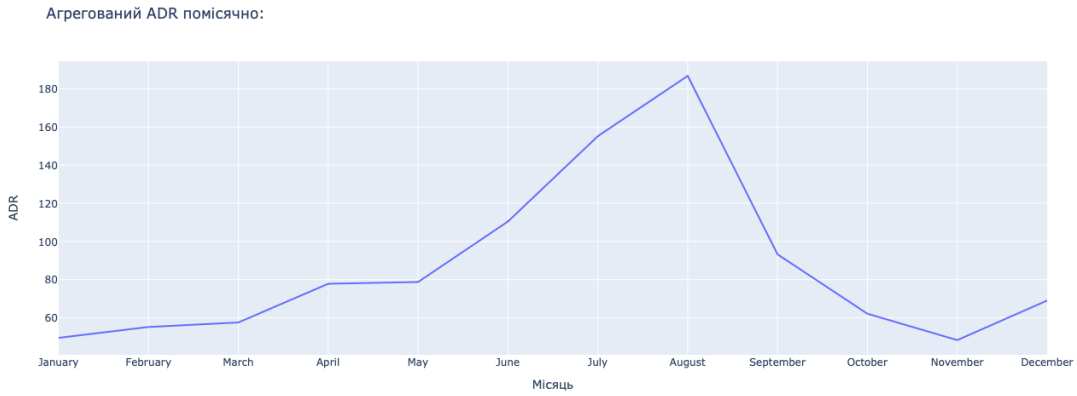

Рис. 3.3. Агрегований ADR помісячно

Джерело: створено автором на основі [55] у середовищі розробки мовою програмування Python

Для дослідження є важливим провести візуалізацію персональних характеристик гостей.

За рис. 3.4 можна зробити висновок, що більшість гостів мають західноєвропейське походження, найбільша кількість із Португалії - країни розміщення готелю, Великобританії, Іспанії, далі йдуть Ірландія, Німеччина та Франція.

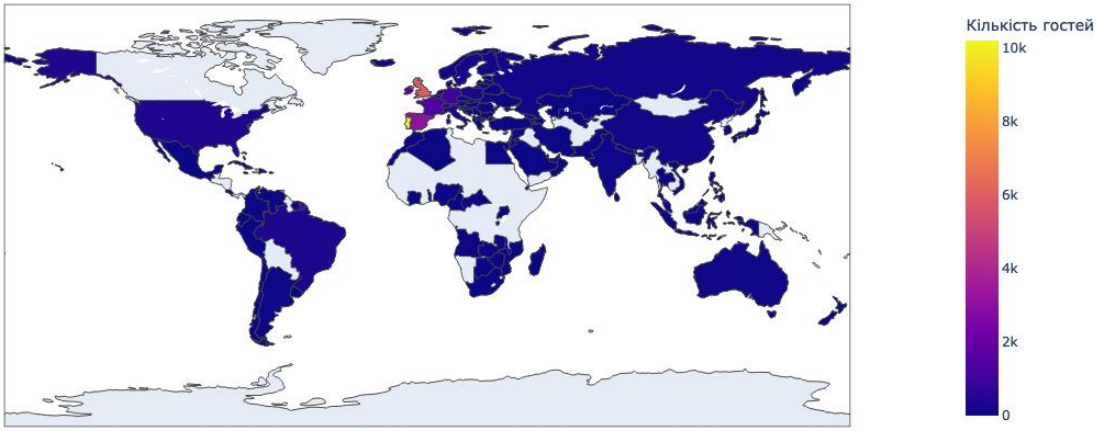

Рис. 3.4. Картограма походження гостей

Джерело: створено автором на основі [55] у середовищі розробки мовою програмування Python

Гості за зв'язками між бронюваннями відносяться до різних типів і, відповідно, мають різну поведінку. Найбільша кількість гостей є типу transient, тобто їх бронювання не пов'язане із іншими, і не є частиною групи або контракту: такі гості скасовують майже чверть від створених бронювань. Наступною за поширеністю категорією і напів-транзитні, вони все ще не є частиною групи або контракту, проте мають зв'язок з іншим бронюванням (тобто два transient бронювання, що мають зв'язок, вже переходять у transient-party), їх відсоток скасувань вже є меншим. Контрактні гості мають значно меншу частину скасувань (8,8%), оскільки серед них часто трапляються бронювання для бізнес-відряджень або аероекіпажів. Групові – бронювання здійснюються групою людей, їх найменша кількість, а відсоток скасувань трохи вищий за бронювання контрактним типом гостей – 10,2%.

Рис.3.5. Агрегований розподіл бронювань за статусом та за CustomerType

Джерело: створено автором на основі [55] у середовищі розробки мовою програмування Python

Клієнти роблять бронювання різними способами, які мають свої особливості, що в результаті впливають на те, чи буде скасоване бронювання. Із рис.3.6 відомо, що найбільша кількість бронювань здійснюється через онлайн туристичні агенції, це відповідає загальним трендам. Водночас легкість їх створення великою мірою визначає велику кількість скасувань у цій категорії. На другому місці протилежний спосіб бронювання – реальні туристичні агентства та оператори, де відсоток скасувань стає відчутно меншим. На

третьому місці – бронювання напряму із готелем без посередників. У груповому бронюванні кількість скасувань менше половини. І найменшим сегментом ринку є безкоштовне розміщення, яке зазвичай пропонується для реклами або просування інших платних послуг готелю.

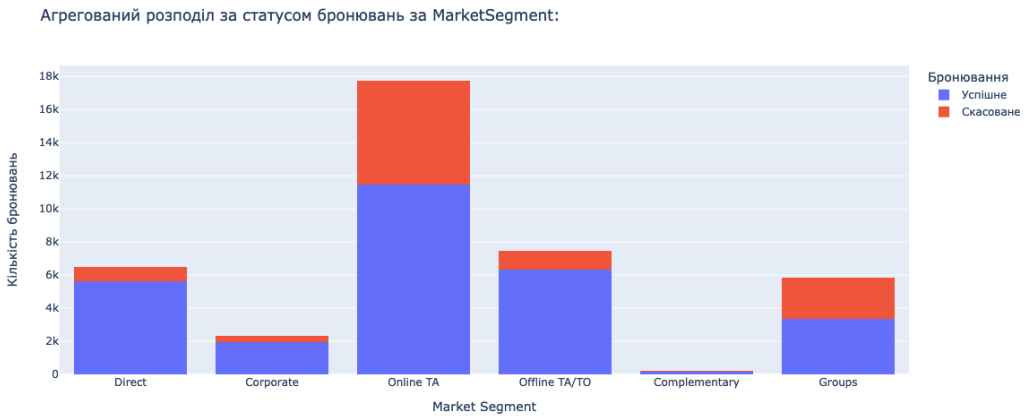

Рис.3.6. Агрегований розподіл бронювань за статусом та за MarketSegment

Джерело: створено автором на основі [55] у середовищі розробки мовою програмування Python

Однією з важливих характеристик бронювання є тип його депозиту, візуалізація якого зображена на рис 3.7. Абсолютна більшість резервувань у готелі зроблена без внесення депозиту, і близько чверті з таких бронювань було скасовано. У бронюваннях, де депозит був більший або рівний вартості номеру, скасовані бронювання переважають, у бронюваннях, де депозит був менший за вартість номеру, ситуація протилежна.

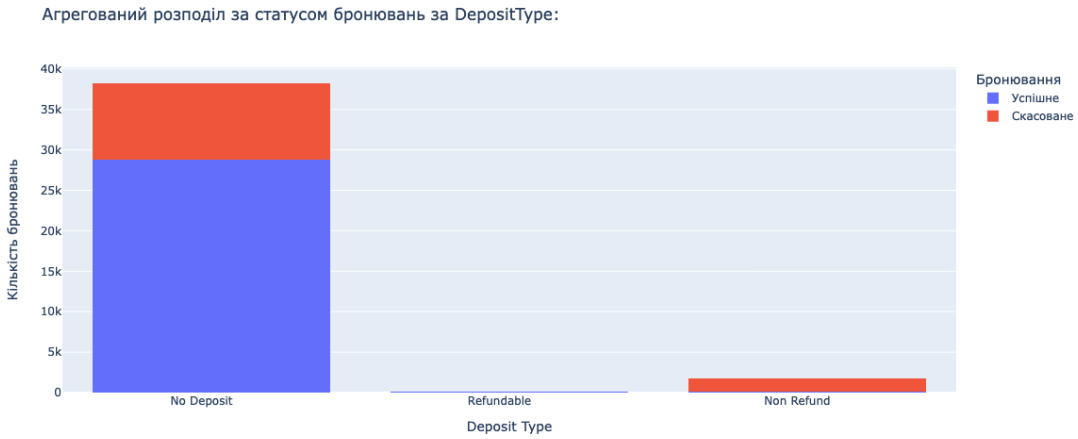

Рис.3.7. Агрегований розподіл бронювань за статусом та за DepositType

Джерело: створено автором на основі [55] у середовищі розробки мовою програмування Python

3.2. Обробка даних

Для датасету нехарактерні пропущені значення, однак деякі атрибути мають номінальні значення NULL, що означають незастосовність категорії до запису, такі значення зустрічаються в атрибутах Agent та Company [Hotel booking demand datasets].

Задля зменшення розмірності даних була створена категорія PreviousCancellationRatio, за методологією Antonio, et al., 2017: значення категорії обраховуються за формулою PreviousCancellationRatio = PreviousCancellations/(PreviousCancellations+PreviousBookingsNotCanceled)).
Проте клієнт може не мати ніяких бронювань, тому в такому випадку в програмному середовищі виникає значення NULL. Для попередження такої ситуації значення категорії обраховуються за наступною логічною формулою:

```
if(PreviousCancellations+PreviousBookingsNotCanceled==0,0,
PreviousCancellations/(PreviousCancellations+
PreviousBookingsNotCanceled)),
```

де друга частина умови повертає 0 замість NULL, а третя обраховує результат за класичною формулою у відмінних випадках.

Після створення атрибуту PreviousCancellationRatio, категорії PreviousCancellations та PreviousBookingsNotCanceled виключаються з аналізу. Атрибут Company був виключений із датасету, оскільки значення NULL складало 92,2% записів у категорії. Атрибут ReservationStatusDate має 913 унікальних значень, що ускладнює умови аналізу, він аналогічно був виключений з аналізу. ReservationStatus – вона спотворює класифікацію бронювань, оскільки має інформацію скасовані бронювання і таким чином є дуже високо корельованою з таргетованою категорією.

Записи, що містили NULL у категорії Country (644 записи), а також записи Undefined у категорії DistributionChannel (1) були відфільтровані. У категорії Children записи із пропущеними значеннями аналогічно були виключені з аналізу. Записи, де кількість дітей та дорослих одночасно дорівнювала 0, були відфільтровані, оскільки неможливо здійснити бронювання на нуль гостів.

Записи, де тип AssignedRoomType – L, аналогічно відфільтровані, це нехарактерний тип номеру для вибірки.

Значення NULL номінальної змінної Agent було замінено на значення 0 (185 інших можливих значень також виражені в числах).

Була здійснена зміна типу даних: номінальні дані перекодовані у числові за допомогою параметру coding type: unique integers оператору Nominal to Numerical. Кожне значення номінальної змінної у результаті отримує унікальне ціле значення.

Після проведених маніпуляції у датасеті залишається 27 атрибутів, із яких 26 є регулярними, а категорія IsCanceled має роль цільової (label).

Кореляція між атрибутами. Для виявлення зв'язків між змінними була обрахована кореляційна матриця, візуалізація якої показана у вигляді теплової карти кореляції в рис. 3.8

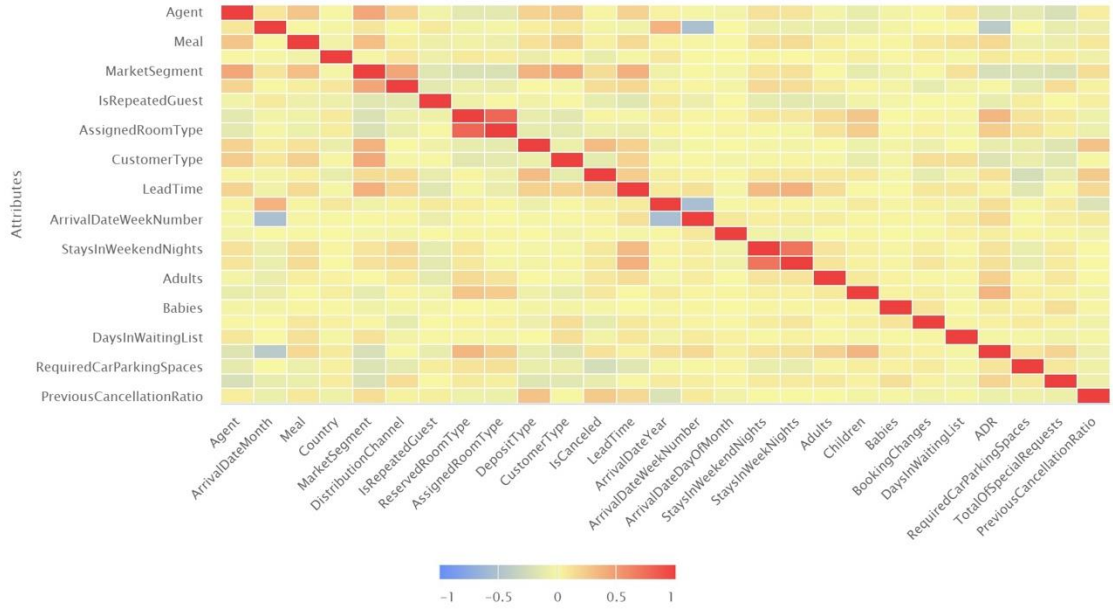

Рис.3.8. Кореляційна матриця атрибутів Hotel1

Джерело: створено автором на основі оброблених даних у середовищі RapidMiner

Для цільової змінної IsCanceled показники кореляції із регулярними атрибутами відображені в табл. 3.2. У ТОП-5 за значимістю для цільової змінної ввійшли такі предиктори як DepositType, RequiredCarParkingSpaces, PreviousCancellationRatio, LeadTime та MarketSegment. Із табл.3.2 встановимо,

що найслабша кореляція у цільової змінної та ArrivalDateWeekNumber, ReservedRoomType, Agent, ArrivalDateDayOfMonth, Babies.

Табл. 3.2

Показники кореляції між значеннями цільової змінної та іншими атрибутами набору даних

Атрибут	Кореляція
IsCanceled	1,0000
DepositType	0,3174
PreviousCancellationRatio	0,2381
LeadTime	0,2295
MarketSegment	0,1414
DistributionChannel	0,1347
ADR	0,1093
Children	0,0813
Adults	0,0804
StaysInWeekendNights	0,0788
StaysInWeekNights	0,0787
ArrivalDateYear	0,0437
Meal	0,0358
ArrivalDateWeekNumber	0,0217
ReservedRoomType	0,0170
Agent	0,0140
ArrivalDateDayOfMonth	-0,0093
Babies	-0,0233
ArrivalDateMonth	-0,0254
DaysInWaitingList	-0,0360
AssignedRoomType	-0,0695
Country	-0,0984
TotalOfSpecialRequests	-0,1013
IsRepeatedGuest	-0,1036
BookingChanges	-0,1147
CustomerType	-0,1197
RequiredCarParkingSpaces	-0,2439

Джерело: створено автором на основі оброблених даних у середовищі RapidMiner

Балансування. В оригінальному датасеті спостерігається асиметрія в значеннях цільової змінної. 72,2% бронювань були успішними і лише 27,8%

скасовані. Така ситуація характерна для об’єкта дослідження, водночас вона негативно впливає на ефективність навчання моделей. Для балансування вибірки був застосований алгоритм SMOTE – Synthetic Minority Oversampling Technique – вже раніше застосований у готельно-спрямованій частині інтелектуального аналізу даних групою дослідників A. A. Nababan et al. SMOTE реалізується методом пошуку k-найближчих сусідів для кожного запису в міноритарному (менш представленому) класі даних, далі генеруються синтетичні записи у кількості, якої не вистачає для бажаного співвідношення між кількостями записів у класі. Подібність даних вимірюється рівнянням Евкліда для чисельного типу даних або формулою Value Difference Metric (VDM) для номінальних атрибутів.

Параметри SMOTE в описаному випадку: кількість сусідів – 5; урівнювання класів – true; nominal change rate - ймовірність отримати номінальне значення сусіда – 0,5 (50%), local random seed – default value – 1992. У результаті використаної техніки оверсемплінгу датасет збільшився до 57 852 записів.

Нормалізація даних. Оскільки різні атрибути мають різні одиниці вимірювання і відповідно різну розмірність, пропонується провести нормалізацію даних за допомогою методу Z-трансформації, яку ще називають статистичною нормалізацією, у результаті якої кожне стандартизоване значення буде мати математичне сподівання, зо дорівнює нулю, та стандартне відхилення, що дорівнює одиниці [56].

Стандартизована оцінка величини (або Z-оцінка) розраховується за формулою (3.1):

$$z = \frac{x - \bar{X}}{S_x}, \tag{3.1}$$

де $\bar{X}$ – середнє значення;

S_x – стандартне відхилення множини значень.

Даний метод зберігає початковий розподіл даних, а також дозволяє порівнювати різні категорії між собою. Крім того, він слабко чутливий до викидів у даних. [56]

У результаті проведених маніпуляцій у моделюванні буде застосовано набір даних, що складається із 26 звичайних атрибутів, 1 спеціального та містить 57 852 записи.

3.3. Моделювання та оцінювання ефективності моделей

Моделювання проводиться у п’ять етапів.

Перший етап: створення та навчання простих (базових) моделей із параметрами за замовчуванням. Ці моделі потрібні для оцінки часу побудови моделі, а також подальшого порівняння із моделями, параметри яких оптимізовані із метою отримання більшої точності класифікації.

Другий етап: створення та навчання оптимізованих моделей, де параметри визначаються сітковим методом (Grid search) – перебором всіх можливих комбінацій вказаних параметрів у вказаних діапазонах. Порівняння ефективності оптимізації відповідно до моделей із базовими параметрами.

Третій етап: створення та навчання неоднорідного ансамблю із оптимізованих моделей способом голосування. Оцінка ефективності гібридної моделі у порівнянні з його складовими.

Четвертий етап: створення гібридних моделей за допомогою процедур бустингу та беггінгу. Оцінка ефективності відносно моделей, на яких вони основані.

П’ятий етап: створення складної моделі гібридного виду – XGBoost, її оцінювання.

Для розв’язання першої задачі дослідження до даних було застосовано три простих класифікатори – Decision Tree, k-NN, Neural Net.

Decision Tree базова

Параметри моделі:

- Decision Tree Simple.criterion = gain_ratio
- Decision Tree Simple.maximal_depth = 25
- Decision Tree Simple.apply_pruning = true

- Decision Tree Simple.confidence = 0.1
- Decision Tree Simple.apply_prepruning = true
- Decision Tree Simple.minimal_gain = 0.01
- Decision Tree Simple.minimal_leaf_size = 2
- Decision Tree Simple.minimal_size_for_split = 4
- Decision Tree Simple.number_of_prepruning_alternatives = 3

Модель тренувалася та тестувалася за допомогою методу кросс-валідації з метою уникнення перенавчання. Кількість підвибірок – 10, метод вибірки – shuffled sampling.

Опис результатів.

Точність отриманої базової моделі дерева рішень становить 87,26%. Специфічність моделі вища, за її чутливість, що означає, що модель схильна класифікувати позитивних гостей, як негативних, більше - ніж негативних, як позитивних. А отже, застосовуючи у практику такий результат бізнес буде надавати хибнонегативним гостям уважний сервіс, що збільшить їхню лояльність.

Табл. 3.3

Ефективність моделі Decision Tree base

Decision Tree base		Матриця класифікації				AUC	Sensitivity	Specificity
Правильно класифіковані	87,39%	TP	FP	TN	FN	0,930	82,45%	92,33%
Неправильно класифіковані	12,61%	23851	2219	26707	5075			

Джерело: створено автором у середовищі MS Excel на основі даних RapidMiner

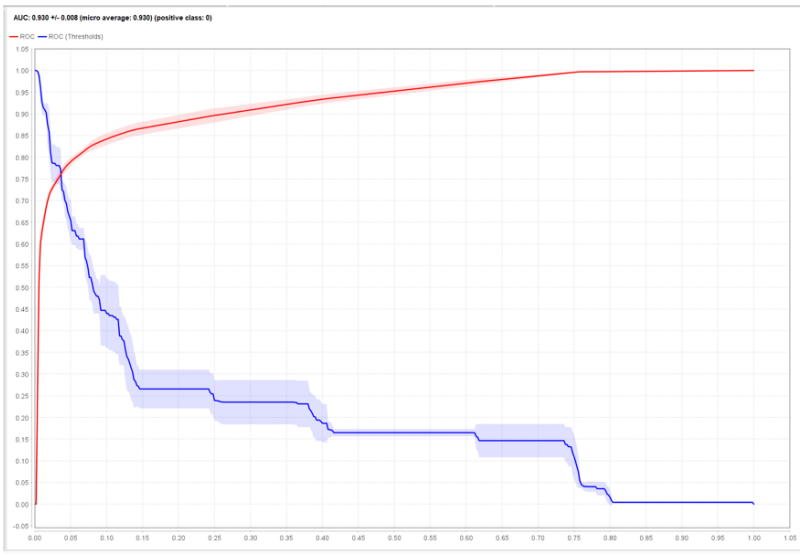

Рис. 3.9. ROC-крива Decision Tree base

Джерело: створено автором у середовищі RapidMiner

K-NN базова

Модель аналогічно до попередньої тренувалася та тестувалася за допомогою методу кросс-валідації із параметрами кількості підвибірок – 10, метод вибірки – shaffled sampling.

Параметри моделі:

- k-NN base.k = 5
- k-NN base.weighted_vote = true
- k-NN base.measure_types = NumericalMeasures
- k-NN base.mixed_measure = MixedEuclideanDistance
- k-NN base.nominal_measure = NominalDistance
- k-NN base.numerical_measure = EuclideanDistance
- k-NN base.divergence = GeneralizedIDivergence
- k-NN base.kernel_type = radial
- k-NN base.kernel_gamma = 1.0
- k-NN base.kernel_sigma1 = 1.0
- k-NN base.kernel_sigma2 = 0.0
- k-NN base.kernel_sigma3 = 2.0
- k-NN base.kernel_degree = 3.0

- k-NN base.kernel_shift = 1.0
- k-NN base.kernel_a = 1.0
- k-NN base.kernel_b = 0.0

Опис результатів.

Модель k-NN менш точна за модель Decision Tree (85,59% проти 87,39%), але її роздільна класифікаційна здатність трохи вища – 0,932 проти 0,930. Класифікаційна матриця моделі представлена у табл.3.4.

Табл. 3.4

Ефективність k-NN base

k-NN base		Матриця класифікації				AUC	Sensitivity	Specificity
Правильно класифіковані	85,59%	TP	FP	TN	FN	0,932	78,97%	92,20%
Неправильно класифіковані	14,41%	22843	2256	26670	6083			

Джерело: створено автором у середовищі MS Excel на основі даних RapidMiner

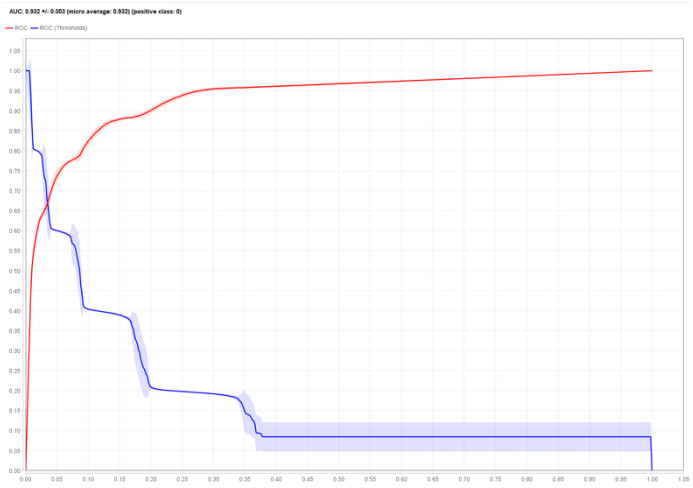

Рис. 3.10. ROC-крива k-NN base

Джерело: створено автором у середовищі RapidMiner

Neural Net базова

Параметри

- Neural Net base.training_cycles = 200
- Neural Net base.learning_rate = 0.01
- Neural Net base.momentum = 0.9
- Neural Net base.decay = false

- Neural Net base.shuffle = true
- Neural Net base.normalize = true
- Neural Net base.error_epsilon = 1.0E-4
- Neural Net base.use_local_random_seed = false
- Neural Net base.local_random_seed = 1992

Опис результатів.

Із табл. 3.5 бачимо, що модель базової нейронної мережі має найвищий AUC на цьому етапі моделювання, а точністю вона поступається лише моделі Decision Tree.

Табл. 3.5

Ефективність моделі Neural Net base

Neural Net base		Матриця класифікації				AUC	Sensitivity	Specificity
Правильно класифіковані	85,98%	TP	FP	TN	FN	0,936	83,45%	88,51%
Неправильно класифіковані	14,02%	24140	3325	25601	4786			

Джерело: створено автором у середовищі MS Excel на основі даних RapidMiner

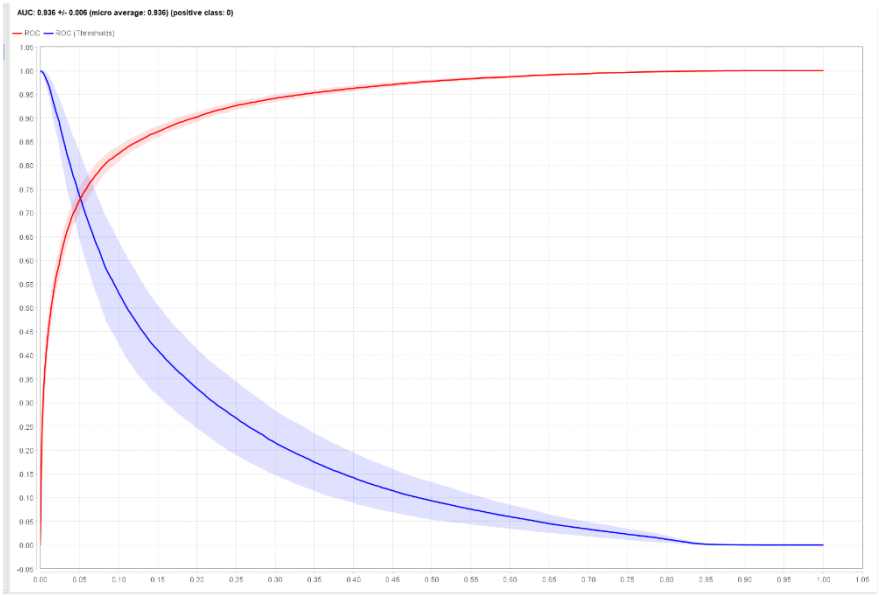

Рис.3.11. ROC-крива Neural Net base»

Джерело: створено автором у середовищі RapidMiner

2. Оптимізувати прості класифікатори

Decision tree оптимізована

Для оптимізації був використаний Optimize Parameters (Grid) алгоритм RapidMiner, який дав 484 комбінації за значеннями параметрів із наступних діапазонів:

Табл. 3.6

Діапазон параметрів оптимізації моделі Decision tree optimized

Parameters DT opt	Min	Max	Steps	Scale
Maximal_depth	1.0E-7	100	10	Linear
Minimal_gain	0.0	1	10	Linear

Джерело: створено автором у середовищі MS Excel на основі даних RapidMiner

Допустимі значення параметрів типу String:

- Decision Tree.criterion = gini_index, accuracy;
- Decision Tree.apply_prunning = true, false;

Параметри, отримані із оптимізацією сітковим алгоритмом:

- Decision Tree.criterion – gini_index;
- Decision Tree.maximal_depth – 20;
- Decision Tree.apply_pruning – false;
- Decision Tree.minimal_gain – 0.0.

Інші параметри моделі

- Decision Tree.confidence = 0.1;
- Decision Tree.apply_prepruning = true;
- Decision Tree.minimal_leaf_size = 2;
- Decision Tree.minimal_size_for_split = 4;
- Decision Tree.number_of_prepruning_alternatives = 3;

Модель тренувалася та тестувалася за допомогою крос-валідації із параметрами, ідентичними використовуваним при валідації базових моделей.

Табл. 3.7

Ефективність моделі Decision Tree optimized

Decision Tree optimized		Матриця класифікації				AUC	Sensitivity	Specificity
Правильно класифіковані	90,01%	TP	FP	TN	FN	0,912	88,63%	91,38%
Неправильно класифіковані	9,99%	25638	2493	26433	3288			

Джерело: створено автором у середовищі MS Excel на основі даних RapidMiner

Опис результатів.

Із табл. 3.7, оптимізована модель Decision Tree має вищу точність (90,01%), значно вищу чутливість (на 6,18% більша за чутливість базової моделі) та дещо нижчу специфічність – 91,38% проти 92,33%. Оптимізована модель все ще схильна приймати бронювання, які не будуть скасовані за ті, що будуть скасовані. Такий результат, окрім зростання зацікавленості в клієнті, нерідко закінчується надмірним бронюванням, що потребує перегляду політик готелю. Незважаючи на вищу точність класифікації, AUC моделі нижчий на 1,8%.

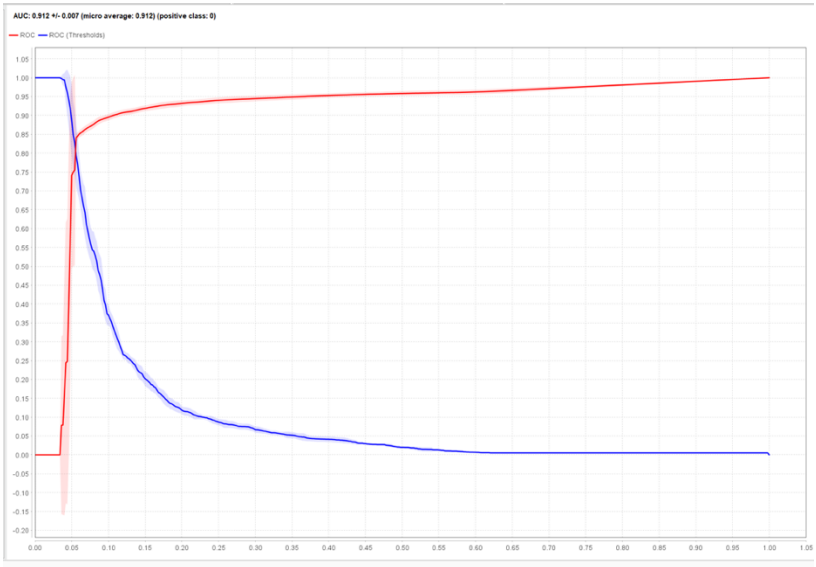

Рис. 3.12. ROC-крива Decision Tree optimized

Джерело: створено автором у середовищі RapidMiner

Вага атрибутів у цільовій змінній розподілена у порядку, зображеному на рис. 3.13. Найбільш значимими кількість затребуваних парко-місць; агент, що робив бронювання; країна походження гостя; лаг між датою створення бронювання та прибуттям; ринковий сегмент, до якого належить гость;

призначений тип номеру. Найменший внесок мають такі змінні як тип депозиту; чи повторний гість; кількість дітей; рік бронювання; кількість ночей, проведених в готелі.

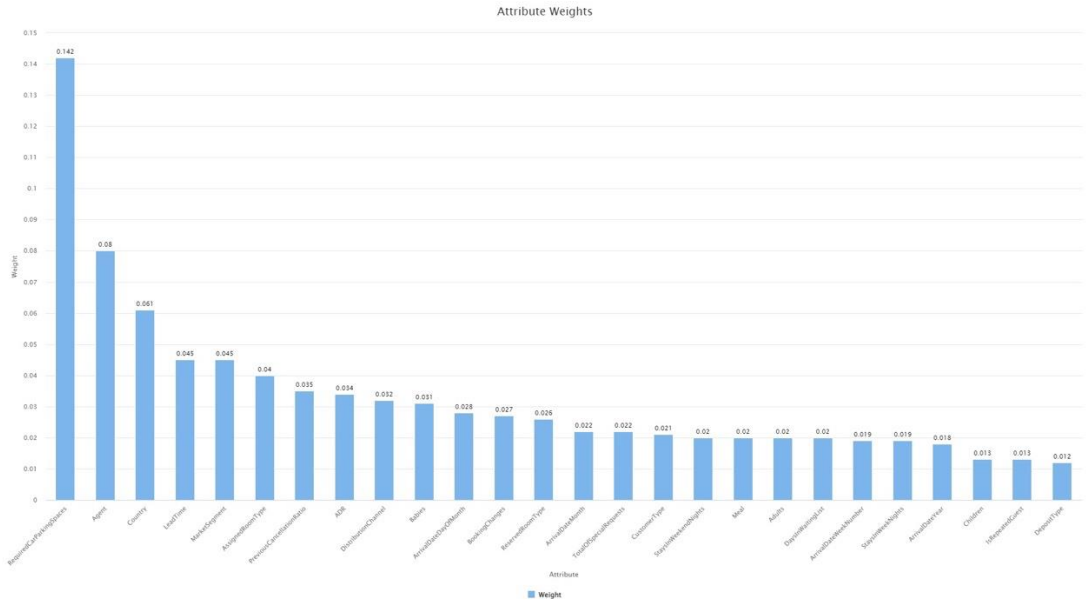

Рис.3.13. Вага атрибутів у моделі Decision Tree optimized

Джерело: створено автором у середовищі RapidMiner

k-NN оптимізована

Для підбору оптимальних параметрів моделі був використаний алгоритм Grid search за діапазоном, описаним у таблиці 3.8 та доступними значеннями для параметрів типу String описаними текстом нижче.

Табл. 3.8

Діапазон параметрів оптимізації моделі k-NN optimized 1

Parameters k-NN opt 1	Min	Max	Steps	Scale
Min_child_weight	1.0	10	5	Linear

Джерело: створено автором у середовищі MS Excel на основі даних RapidMiner

Допустимі параметри типу String:

- k-NN opt.weighted_vote = true; false;
- k-NN opt.numerical_measure = EuclideanDistance; ChebychevDistance; CorrelationSimilarity.

Параметри, отримані оптимізацією Grid Search:

- k-NN opt.k = 1;
- k-NN opt.weighted_vote = false;
- k-NN opt.numerical_measure = CorrelationSimilarity.

Інші параметри:

- k-NN opt.measure_types = MixedMeasures;
- k-NN opt.divergence = SquaredEuclideanDistance.

Опис результатів.

У табл. 3.9 точність зросла на 2,54% водночас показник AUC дорівнює 0,5, що характерно для випадкового класифікатора, це показує, що модель не може відрізняти позитивні випадки від негативних, а отже, виключається з аналізу.

Табл. 3.9

Ефективність моделі k-NN optimized 1

k-NN optimized 1		Матриця класифікації				AUC	Sensitivity	Specificity
Правильно класифіковані	88,13%	TP	FP	TN	FN	0,5	83,25%	93,00 %
Неправильно класифіковані	11,87%	24082	2024	26902	4844			

Джерело: створено автором у середовищі MS Excel на основі даних RapidMiner

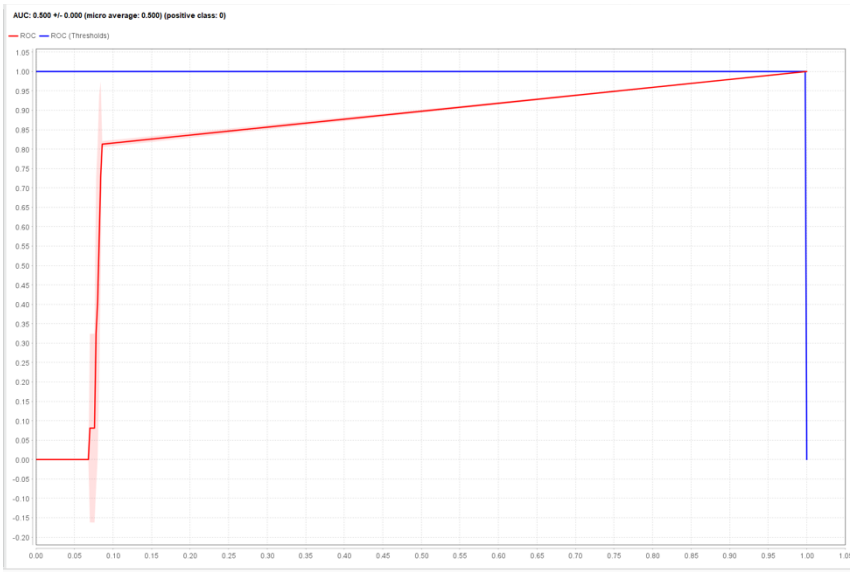

Рис.3.14. ROC-крива k-NN optimized 1

Джерело: створено автором у середовищі RapidMiner

Змінимо параметри моделі: кількість найближчих сусідів із 1 до 2, а також оберемо Евклідову відстань як міру виміру відстані між числовими значеннями. Із новими параметрами отримаємо модель, продуктивність якої описана у табл. 3.11

Табл. 3.10

Ефективність моделі k-NN optimized 2

k-NN optimized 2		Матриця класифікації				AUC	Sensitivity	Specificity
Правильно класифіковані	87,98%	TP	FP	TN	FN	0,916	82,98%	92,97 %
Неправильно класифіковані	12,02%	24004	2033	26893	4922			

Джерело: створено автором у середовищі MS Excel на основі даних RapidMiner

Рис. 3.15

«ROC-крива k-NN optimized 2»

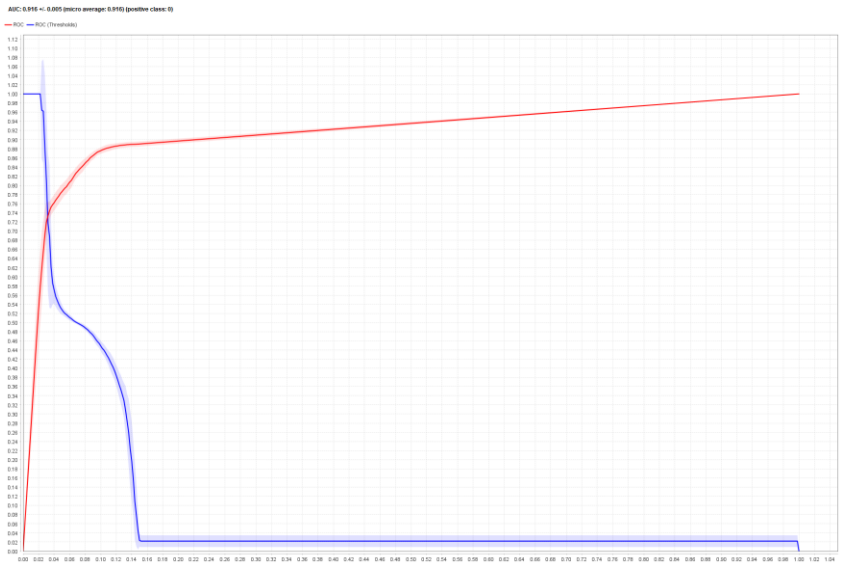

Рис. 3.15. ROC-крива k-NN optimized 2

Джерело: створено автором у середовищі RapidMiner

Опис результатів.

Друга модель має трохи меншу точність, але набагато кращу площу під кривою, ніж перша, тому саме вона буде застосована в ансамблі.

Neural Net оптимізована

Налаштуємо параметри моделі Neural Net за допомогою сіткового алгоритму Grid Search. Проміжки оптимізації числових параметрів вказані в табл. 3.11.

Табл. 3.11

Діапазон параметрів оптимізації моделі Neural Net optimized

Parameters NN opt	Min	Max	Steps	Scale
Training_cycles	1.0	300	10	Linear
Learning_rate	4.9E-324	0.1	10	Linear
Momentum	0.7	1.0	3	Linear

Джерело: створено автором у середовищі MS Excel на основі даних RapidMiner

Отримані сітковим пошуком параметри:

- Neural Net.training_cycles = 270;
- Neural Net.learning_rate = 0.020000000000000004;
- Neural Net.momentum = 0.7999999999999999.
- Інші параметри:
- Neural Net.training_cycles.hidden_layers =2;
- Shuffle = true;
- Normalize=true.

Опис результатів.

У процесі оптимізації, яка тривала близько 14-ти годин, точність моделі підвищилася із 85,98% до 86,85%, а площа під кривою на 0,003, порівняно з моделлю із параметрами за замовчуванням. Матрицю класифікації, а також показники чутливості та специфічності можна побачити в табл. 3.12.

Табл. 3.12

Ефективність моделі Neural Net optimized

Neural Net optimized		Матриця класифікації				AUC	Sensitivity	Specificity
Правильно класифіковані	86,85%	TP	FP	TN	FN	0,939	85,20%	88,51%
Неправильно класифіковані	13,15%	24644	3327	25599	4282			

Джерело: створено автором у середовищі MS Excel на основі даних RapidMiner

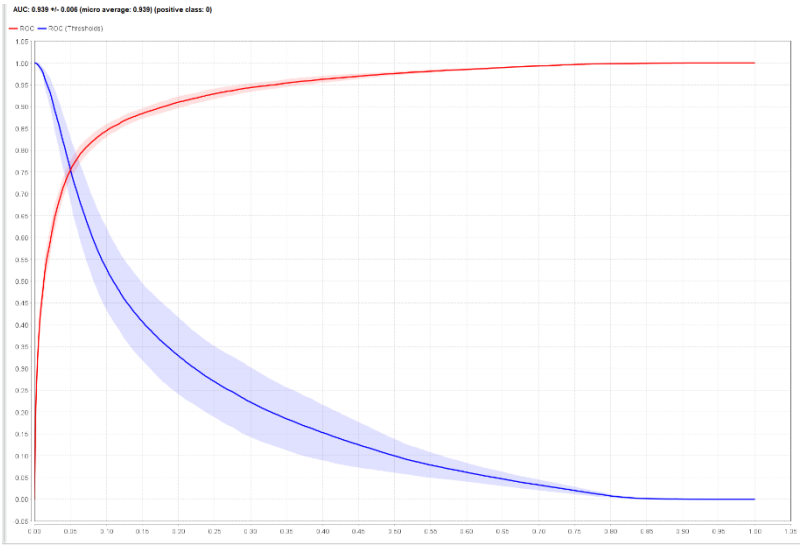

Рис. 3.16. ROC-крива Neural Net optimized

Джерело: створено автором у середовищі RapidMiner

Voting ensemble

Оптимізовані прості класифікатори, що реалізуються алгоритмами різних типів, об'єднаємо в ансамбль за допомогою оператору Vote. Із табл. 3.13 встановлюємо, що ефективність ансамблю зросла в порівнянні із слабкими моделями, що входять до його складу. Значення точності прогнозу ансамблю на 0,94% більше за найкраще отримане, від моделі Decision tree optimized, а розподільча здатність класифікації на 0,07 вища за розподільчу здатність моделі Neural Net optimized.

Табл. 3.13

Порівняння ефективності простих класифікаторів із їх ансамблем

Classifier	Accuracy	CE	TP	FP	TN	FN	AUC	Sensitivity	Specificity
DT	90,01%	9,99%	25638	2493	26433	3288	0,912	88,63%	91,38%
k-NN 2	87,98%	12,02%	24004	2033	26893	4922	0,916	82,98%	92,97 %
NN	86,85%	13,15%	24644	3327	25599	4282	0,939	85,20%	88,51%
Vote ensemble	90,95%	9,05%	25949	2260	26666	2977	0,946	89,71%	92,19%

Джерело: створено автором у середовищі MS Excel на основі даних RapidMiner

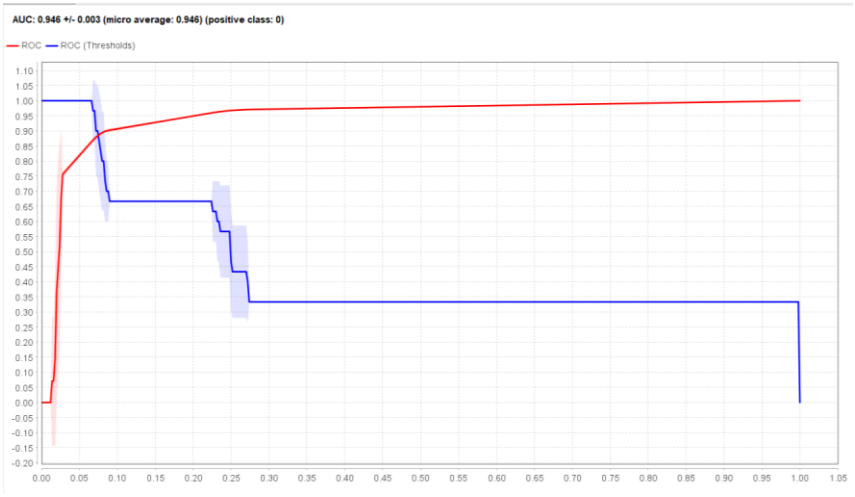

Рис. 3.17. ROC-крива Vote Ensemble»

Джерело: створено автором у середовищі RapidMiner

Decision Tree AdaBoost

Застосуємо до кращої з простих моделей послідовну оптимізацію за допомогою процедури адаптивного бустингу із використанням оператору AdaBoost. Кількість ітерацій – 10. Отримана гібридна модель має майже однакову точність із початковою (90,12% та 90,01%), водночас її показник AUC є значно вищим – 0,957 проти 0,912. Ефективність моделі Decision tree boosted відображена в табл. 3.14 та рис. 3.18:

Табл. 3.14

Порівняння ефективності моделі Decision Tree boosted із Decision Tree optimized

Classifier	Accuracy	CE	TP	FP	TN	FN	AUC	Sensitivity	Specificity
DT optimized	90,01 %	9,99%	25638	2493	26433	3288	0,912	88,63%	91,38%
DT boosted	90,12 %	9,88%	25615	2406	26520	3311	0,957	88,55%	91,69%

Джерело: створено автором у середовищі MS Excel на основі даних RapidMiner

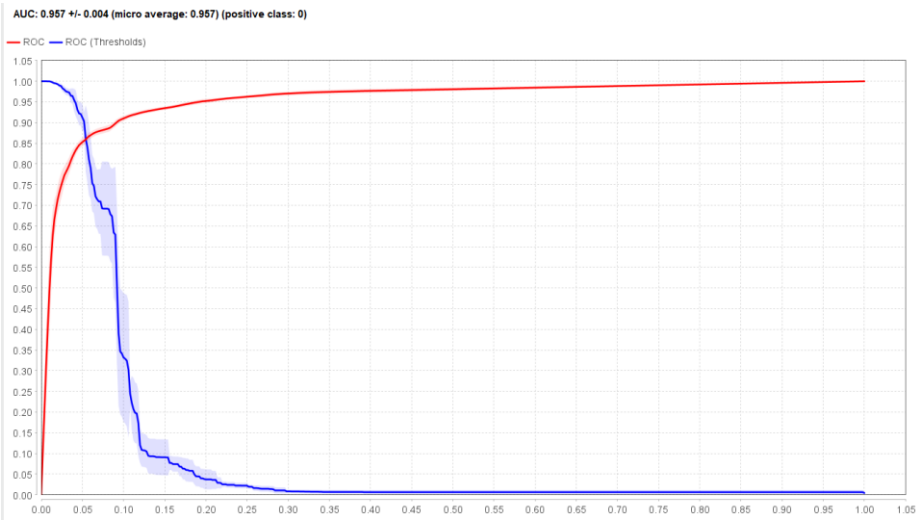

Рис.3.18 ROC-крива Decision tree boosted

Джерело: створено автором у середовищі RapidMiner

Neural Net bagged

Використаємо метод паралельної оптимізації – беггінг на прикладі найменш точної із моделей ансамблю – моделі Neural Net optimized, за допомогою процедури Bagging середовища RapidMiner. У табл. 3.15 показано, що точність прогнозу моделі зросла на 1,1%, класифікаційна здатність моделі - на 0,009, специфічність на 1,73%.

Табл. 3.15

Порівняння ефективності моделі Neural Net bagged із Neural Net optimized

Classifier	Accuracy	CE	TP	FP	TN	FN	AUC	Sensitivity	Specificity
NN optimized	86,85%	13,15%	24644	3327	25599	4282	0,939	85,20%	88,51%
NN bagged	87,95%	12,05%	24777	2823	26103	4149	0,948	85,66%	90,24%

Джерело: створено автором у середовищі MS Excel на основі даних RapidMiner

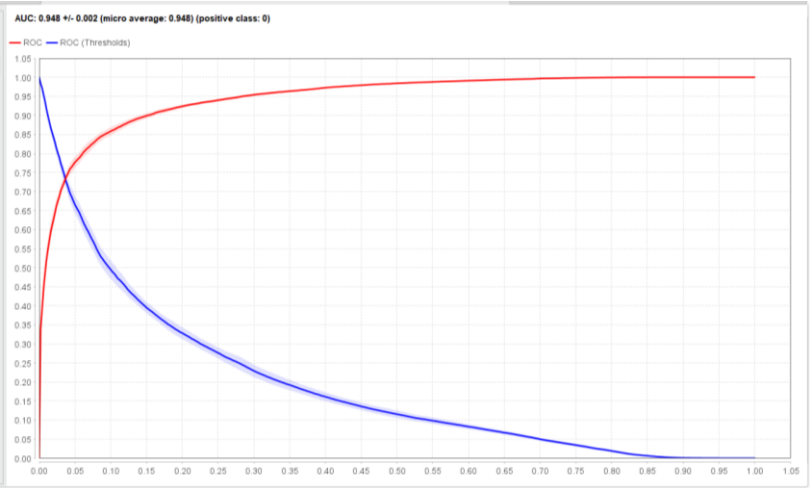

Рис. 3.19. ROC-крива Neural Net bagged

Джерело: створено автором у середовищі RapidMiner

XGBoost

Дана гібридна модель-ансамбль є інструментом перемоги у численних змаганнях із машинного навчання, на нашу думку вона заслуговує на увагу в дослідженні.

Застосуємо до датасету оператор моделювання XGBoost із параметрами за замовчуванням:

- tree_method = auto;
- seed = 2513067705;
- max_depth = 6;
- booster = gbtree;
- min_split_loss = 0.0;
- objective = binary:logistic;
- lambda = 1.0;
- nthread = 11;
- alpha = 0.0;
- subsample = 1.0;
- learning_rate = 0.3;
- min_child_weight = 1.0;
- verbosity = 0;
- boosting_iterations = 25;

Із параметрами за замовчуванням модель XGBoost має наступні результати, відображені в табл. 3.16 та рис.3.20:

Табл. 3.16

Ефективність моделі XGBoost

XGBoost		Матриця класифікації				AUC	Sensitivity	Specificity
Правильно класифіковані	90,89%	TP	FP	TN	FN	0,971	88,11%	93,67%
Неправильно класифіковані	9,11%	25485	1830	27096	3441			

Джерело: створено автором у середовищі MS Excel на основі даних RapidMiner

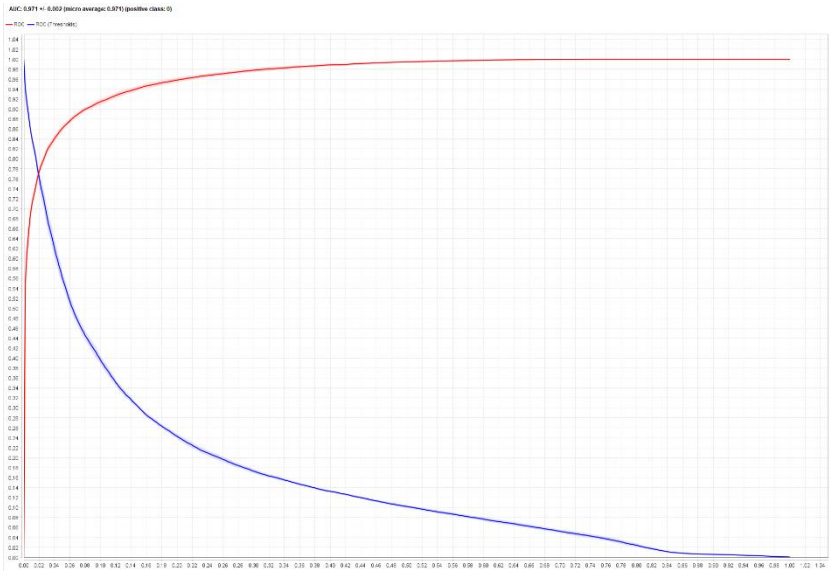

Рис.3.20. ROC-крива XGBoost

Джерело: створено автором у середовищі RapidMiner

Оптимізація моделі буде проведена у два етапи через велику кількість параметрів, перебір комбінацій яких потребує значних матеріально-технічних та часових витрат. За допомогою оператору Grid Search проведемо перший етап оптимізації. Визначимо параметри, комбінації яких будуть перебиратися.

Допустимі параметри типу String: booster – tree booster, linear booster, DART.

Табл. 3.17

Діапазон параметрів оптимізації моделі XGBoost optimized 1

Parametres XGBoost opt 1	Min	Max	Steps	Scale
Max_depth	0	25	5	Linear
Learning_rate	0.5	0.8	3	Linear
Subsample	0.5	0.9	4	Linear

Джерело: створено автором у середовищі MS Excel на основі даних RapidMiner

Параметри, отримані із оптимізацією алгоритмом Grid Search:

- XGBoost (4).max_depth = 25;
- XGBoost (4).booster = tree booster;
- XGBoost (4).learning_rate = 0.5;
- XGBoost (4).subsample = 0.9.

Інші:

- XGBoost (4).rounds = 25;
- XGBoost (4).early_stopping = none;
- XGBoost (4).early_stopping_rounds = 10;
- XGBoost (4).min_split_loss = 0.0;
- XGBoost (4).max_depth = 6;
- XGBoost (4).min_child_weight = 1.0;
- XGBoost (4).tree_method = auto;
- XGBoost (4).lambda = 1.0;
- XGBoost (4).alpha = 0.0;
- XGBoost (4).sample_type = uniform;
- XGBoost (4).normalize_type = tree;
- XGBoost (4).rate_drop = 0.0;
- XGBoost (4).skip_drop = 0.0;
- XGBoost (4).updater = shotgun;

- XGBoost (4).top_k = 0;
- XGBoost (4).expert_parameters = null.

Опис результатів

У результаті оптимізації найвагоміших параметрів, модель XGBoost має підвищення точності на 2,8%, площа під ROC-кривою збільшилася на 0,13. Модель має високу чутливість та специфічність.

Табл. 3.18

Ефективність моделі XGBoost optimized 1

XGBoost opt1		Матриця класифікації				AUC	Sensitivity	Specificity
Правильно класифіковані	93,69%	TP	FP	TN	FN	0,984	92,74%	94,65%
Неправильно класифіковані	6,31%	26825	1549	27377	2101			

Джерело: створено автором у середовищі MS Excel на основі даних RapidMiner

Рис.3.21. ROC-крива XGBoost optimized 1

Джерело: створено автором у середовищі RapidMiner

Другий етап оптимізації аналогічно проводиться за допомогою оператора Grid Search.

Для обраного типу підсилювання – tree booster, можна оптимізувати ще не залучені наступні параметри моделі, як числові так і категоріальні. Діапазони оптимізації числових параметрів представлені в табл. 3.19.

Табл. 3.19

Діапазон параметрів оптимізації моделі XGBoost optimized 2

Parametres XGBoost opt 2	Min	Max	Steps	Scale
Min_child_weight	0	3	5	Linear
Lambda	0.8	1	2	Linear
Alpha	0.0	0.2	2	Linear

Джерело: створено автором у середовищі MS Excel на основі даних RapidMiner

Допустимі параметри типу String:

- three_method – auto; exact; approximate, histogram;
- early_stopping – none; auto.

Параметри, отримані із оптимізацією сітковим алгоритмом:

- XGBoost (4).min_child_weight = 0.0;
- XGBoost (4).tree_method = auto;
- XGBoost (4).early_stopping = none;;
- XGBoost (4).lambda = 1.0;
- XGBoost (4).alpha = 0.0.

Інші параметри моделі:

- XGBoost (4).booster = tree booster;
- XGBoost (4).rounds = 25;
- XGBoost (4).early_stopping_rounds = 10;
- XGBoost (4).learning_rate = 0.5;
- XGBoost (4).min_split_loss = 0.0;
- XGBoost (4).max_depth = 25;
- XGBoost (4).subsample = 0.9;

- XGBoost (4).sample_type = uniform;
- XGBoost (4).normalize_type = tree;
- XGBoost (4).rate_drop = 0.0;
- XGBoost (4).skip_drop = 0.0;
- XGBoost (4).updater = shotgun;
- XGBoost (4).top_k = 0;
- XGBoost (4).expert_parameters = null.

Опис результатів.

Наслідком другого етапу оптимізації стало незначне покращення моделі: вона правильно класифікувала на 0,05% більше прикладів, а її AUC зріс всього на 0,001. Усі результати відображені в табл. 3.20.

Табл. 3.20

Ефективність моделі XGBoost optimized 2

XGBoost opt2		Матриця класифікації				AUC	Sensitivity	Specificity
Правильно класифіковані	93,74%	TP	FP	TN	FN	0,985	92,67%	94,81%
Неправильно класифіковані	6,26%	26805	1501	27425	2121			

Джерело: створено автором у середовищі MS Excel на основі даних RapidMiner

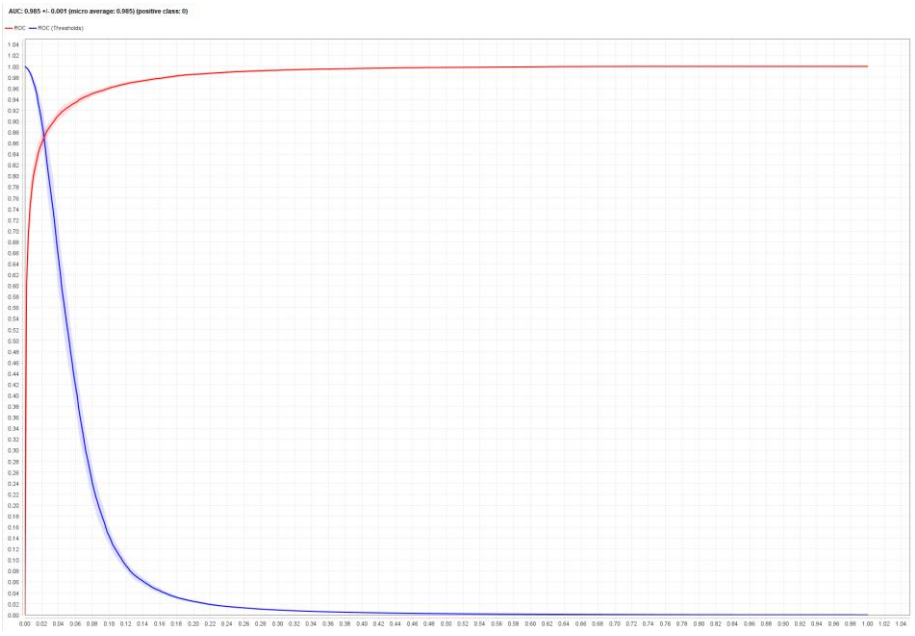

Рис.3.22. ROC-крива XGBoost optimized 2

Джерело: створено автором у середовищі RapidMiner

Вага предикторів у прогнозуванні цільової змінної неоднорідна. Як і в моделі Decision Tree optimized, найбільший внесок у передбачення цільової змінної моделі XGBoost optimized 2 має атрибут RequiredCarParkingSpace, наступна за значимістю країна походження гостя, на третьому – LeadTime.

Найменш ефективні для прогнозування скасування бронювань такі дані, як рік заселення, кількість дітей та дорослих гостів.

Візуалізація ваги предикторів зображена на рис.3.23.

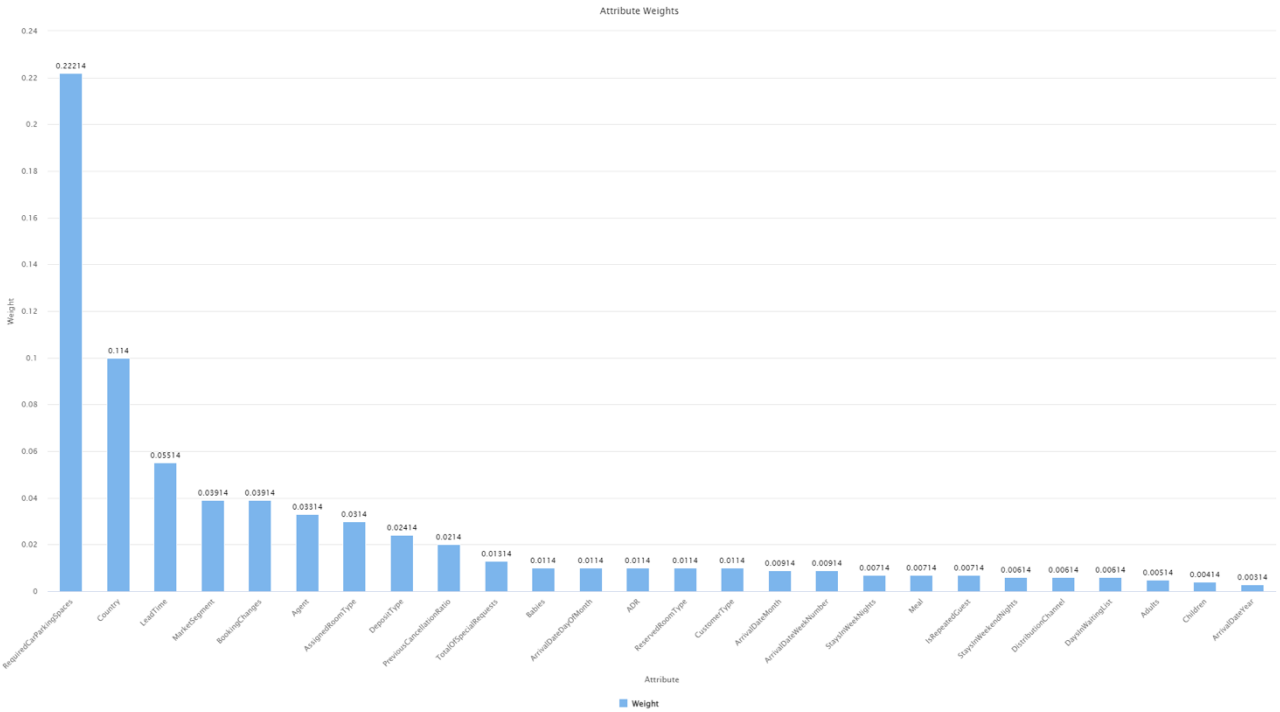

Рис.3.23. Вага предикторів у моделі XGBoost optimized 2

Джерело: створено автором у середовищі RapidMiner

3.4. Порівняння ефективності гібридних моделей

Порівняння результатів використання гібридних моделей у прогнозуванні скасування бронювань здійснено за традиційними критеріями: точністю прогнозу, класифікаційній матриці, площі під кривою, чутливістю та специфічністю в табл. 3.21.

Ранжування моделей. Найвищу точність та AUC має оптимізована модель XGBoost. На другому місці в кількості правильно передбачених записів – Vote ensemble, хоча його AUC поступається підсиленому Decision Tree та найменш

точній з усіх гібридних моделей – моделі Neural Net bagged. Decision Tree boosted із точністю 90,12% займає третє місце у рейтингу. Четвертий результат серед гібридних моделей має модель NN bagged. Усі моделі краще класифікують скасовані бронювання, ніж підтверджені. Такий результат забезпечує ефективність ревеню-менеджменту в роботі із попередженням скасування бронювань. До того ж, ситуація меншої кількості хибно-позитивних класифікацій, ніж хибно-негативних є бажанішою, ніж зворотня, де готель би втрачав доходи на клієнтах, яких не осягнули увагою.

Табл. 3.21

Порівняльна таблиця ефективності гібридних моделей

Classifier	Accuracy	CE	TP	FP	TN	FN	AUC	Sensitivity	Specificity
XGBoost opt 2	93,74%	6,26%	26805	1501	27425	2121	0,985	92,67%	94,81%
Vote ensemble	90,95%	9,05%	25949	2260	26666	2977	0,946	89,71%	92,19%
DT boosted	90,12%	9,88%	25615	2406	26520	3311	0,957	88,55%	91,69%
NN bagged	87,95%	12,05%	24777	2823	26103	4149	0,948	85,66%	90,24%

Джерело: створено автором у середовищі MS Excel на основі даних RapidMiner

ВИСНОВКИ

Результатом виконаного в кваліфікаційній роботі бакалавра дослідження є систематизація теоретично-практичних здобутків із вивчення скасувань бронювань, зокрема передумов їх виникнення, важливості в управлінні готельним бізнесом, підходів до прогнозування статусу бронювань. Було запропоновано економіко-математичний інструментарій ІАД для класифікації бронювань як успішних або скасованих, визначення факторів, що впливають на скасування бронювань.

На підставі проведеного дослідження можна зробити такі висновки:

1. Готельні послуги як основний продукт діяльності готельного бізнесу має наступні особливості: обмежена можливість зберігання, що обертається у збитки, якщо послуга не продана станом на певний час; обмежена місткість готелю; мінливий характер попиту та можливість його прогнозування. Дві перших характеристики є одними із визначальних в потребі прогнозування попиту, яке значною мірою ґрунтується на кількості заброньованих номерів на певний період. Скасування бронювань перешкоджає створенню точного прогнозу.

2. Існує два підходи до прогнозування скасування бронювань: перший – регресійний, прибічники якого висловлювали сумнів щодо можливості визначення статусу конкретного бронювання; другий – класифікаційний, що виник як відповідь на твердження, зроблене попередниками. У розв’язанні задачі класифікації бронювань переважають рішення, що ґрунтуються на застосуванні слабких класифікаторів (Bayesian Net, DT, Logistic, SVM тощо), а гібридні моделі представлені випадковим лісом та слабкими моделями, гіпер-оптимізованими різноманітними алгоритмами.

3. Для полегшення сприйняття вхідні дані дослідження просто візуалізувати в середовищі розробки Jupiter Notebook за допомогою мови програмування Python. З’ясовано, що незважаючи на приморське положення готелю середній відпускний тариф підвищується не лише в теплий період року, а й в останні тижні – під час різдвяних свят. Серед транзитних гостей відсоток

скасувань досягає близько чверті від кількості зроблених, водночас контрактні гості скасовують бронювання втричі нижче. Гості, що робили бронювання через онлайн туристичні агенції скасували більше третини своїх резервувань, у той час як відсоток скасувань бронювань, зроблених напряму в готелі становить всього 13%, що підтверджує гіпотезу про вплив OTA на ймовірність скасування бронювання.

4. RapidMiner є зручним інструментом для очищення та перетворення даних. Він дозволяє швидко видаляти та створювати нові атрибути, фільтрувати нульові та пропущені значення, а також записи, що не підпадають під описану умову. Для якісного моделювання було змінено тип даних, їх балансування та нормалізація. Із 31 атрибуту початкового датасету для моделювання було обрано 27.

5. Прості класифікатори, незалежно від їх типу на отриманих даних показують високі результати прогнозування. DT, k-NN і NN із оптимізованими алгоритмом Grid Search параметрами показали підвищення точності прогнозу порівняно із базовими моделями від 0,87% до 2,62%.

6. Неоднорідний ансамбль, створений на основі попередніх моделей, показав точність на 0,94% більшу за відповідний показник найкращої з його складників, площа під кривою зросла на 0,007. Значно кращу розподільчу здатність (AUC – 0,957), ніж його основа (AUC – 0,912), показало підсилене дерево рішень. Neural Net begged перевершила оптимізовану модель Neural Net як у точності прогнозу (87,95% проти 86,85%), так і в показнику AUC (0,948 проти 0,939).

7. Одна з найбільш ефективних класифікаційних моделей сучасності XGBoost, яка використовує алгоритм екстремального градієнтного підсилювання дерев, була застосована до вирішення задачі класифікації бронювань. Отримані результати оптимізованої моделі, співставні з результатами відомих дослідників у науковій галузі – точність XGBoost 2 становить 93,74% проти 93,8% отриманих за використання цього алгоритму в дослідженні [8]. Передбачувальна здатність гібридної моделі є дуже високою,

AUC становить 0,985, що значно перевищує результати (0,6442) отримані дослідниками [8]. Серед найбільш важливих предикторів для прогнозування бронювань у моделі XGBoost 2 – RequiredCarParkingSpace, Country, LeadTime.

8. Усі гібридні моделі показали результати кращі, ніж прості класифікатори, які були їх основою, що відповідає твердженню про вищу ефективність гібридних моделей в ІАД.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Deloitte. 2023 travel industry outlook. 2023. URL: https://www2.deloitte.com/us/en/pages/consumer-business/articles/travel-hospitality-industry-outlook.html (дата звернення: 05.06.2023).

2. STR: Most global regions showed full RevPAR recovery in 2022. URL: https://str.com/press-release/str-most-global-regions-showed-full-revpar-recovery-2022 (дата звернення: 05.06.2023).

3. Cancellation Prediction for Flight Data Using Machine Learning / A. Ansari та ін. *SSRN Electronic Journal*. 2019. URL: https://doi.org/10.2139/ssrn.3367683 (дата звернення: 05.06.2023).

4. Chen C.-C., Schwartz Z., Vargas P. The search for the best deal: How hotel cancellation policies affect the search and booking decisions of deal-seeking customers. *International Journal of Hospitality Management*. 2011. Vol. 30, no. 1. С. 129–135. URL: https://doi.org/10.1016/j.ijhm.2010.03.010 (дата звернення: 05.06.2023).

5. Koide T., Ishii H. The hotel yield management with two types of room prices, overbooking and cancellations. *International Journal of Production Economics*. 2005. Vol. 93-94. С. 417–428. URL: https://doi.org/10.1016/j.ijpe.2004.06.038 (дата звернення: 05.06.2023).

6. Lambert C. U., Lambert J. M., Cullen T. P. The Overbooking Question: A Simulation. *Cornell Hotel and Restaurant Administration Quarterly*. 1989. V. 30, № 2. С. 14–20. URL: https://doi.org/10.1177/001088048903000206 (дата звернення: 05.06.2023).

7. Antonio N., de Almeida A., Nunes L. Predicting Hotel Bookings Cancellation with a Machine Learning Classification Model. *C 16th IEEE International Conference on Machine Learning and Applications (ICMLA).2017*, м. Cancun, Mexico, 18–21 груд. 2017 р. 2017. URL: https://doi.org/10.1109/icmla.2017.00-11 (дата звернення: 05.06.2023).

8. Nababan, A.A., Miftahul Jannah and Nababan, A.H. Prediction of Hotel Booking Cancellation Using K-Nearest Neighbors (K-NN) Algorithm and Synthetic

Minority Over-Sampling Technique (SMOTE). *INFOKUM*. 2022. Vol. 10-03 P. 50-56. URL: http://seaninstitute.org/infor/index.php/infokum/article/view/596 (дата звернення: 05.06.2023).

9. Antonio N., de Almeida A., Nunes L. Big Data in Hotel Revenue Management: Exploring Cancellation Drivers to Gain Insights Into Booking Cancellation Behavior. *Cornell Hospitality Quarterly*. 2019. Т. 60, № 4. С. 298–319. URL: https://doi.org/10.1177/1938965519851466(дата звернення: 06.06.2023).

10. Мальська М.П., Пандяк І.Г. Готельний бізнес: теорія та практика: підручник, 2-ге вид., перероб. та доп. Київ, 2012. 472 с.

11. Про туризм: Закон України від 01.04.2023 р. №324/95-ВР. URL: https://zakon.rada.gov.ua/laws/show/324/95-вр#Text

12. Про затвердження Правил користування готелями й аналогічними засобами розміщення та надання готельних послуг: Наказ від 12.11.2010 р. №z04-13-04. URL: https://zakon.rada.gov.ua/laws/show/z0413-04#Text

13. El Haddad, R., Roper, A., Jones P. The Impact of Revenue Management Decisions on Customers' Attitudes and Behaviours: A Case Study of a Leading UK Budget Hotel Chain. *EuroCHRIE 2008 Congress, Emirates Hotel School, Dubai, UAE*. 11-14 жовтня 2008. Dubai.

14. Kimes S. E. The Basics of Yield Management. *Cornell Hotel and Restaurant Administration Quarterly.* 1989. Т. 30, № 3. С. 14–19. URL: https://doi.org/10.1177/001088048903000309 (дата звернення: 06.06.2023).

15. Kimes S. E., Wirtz J. Has Revenue Management become Acceptable?. *Journal of Service Research. 2003.* Т. 6, № 2. С. 125–135. URL: https://doi.org/10.1177/1094670503257038 (дата звернення: 06.06.2023).

16. Cross, Robert & Higbie, Jon. (2009). Revenue Management's Renaissance A Rebirth of the Art and Science of Profitable Revenue Generation. *Cornell Hospitality Quarterly - CORNELL HOSP Q.* Т. 50. С.56-81. URL: https://doi.org/10.1177/193896550832871 (дата звернення: 06.06.2023).

17. Anderson, C.K. and Xie, X. (2010) Improving Hospitality Industry Sales: Twenty-Five Years of Revenue Management. *Cornell Hospitality Quarterly*, Т.51, С.

53-67. URL: https://doi.org/10.1177/1938965509354697 (дата звернення: 06.06.2023).

18. Ivanov S. Hotel Revenue Management: From Theory to Practice. 1-е вид., Varna, 2014. 204 с.

19. Hayes D. K., Miller A. Revenue Management for the Hospitality Industry. Wiley & Sons, Incorporated, 2010. 529 с.

20. Revenue management under customer choice behaviour with cancellations and overbooking / D. D. Sierag та ін. *European Journal of Operational Research*. 2015. Т. 246, № 1. С. 170–185. URL: https://doi.org/10.1016/j.ejor.2015.04.014 (дата звернення: 06.06.2023).

21. Sánchez-Medina A. J., C-Sánchez E. Using machine learning and big data for efficient forecasting of hotel booking cancellations. *International Journal of Hospitality Management.* 2020. Т. 89. С. 102546. URL: https://doi.org/10.1016/j.ijhm.2020.102546 (дата звернення: 06.06.2023).

22. Global Cancellation Rate of Hotel Reservations Reaches 40% on Average. URL: https://hospitalitytech.com/global-cancellation-rate-hotel-reservations-reaches-40-average (дата звернення: 06.06.2023).

23. Sigala M. Tourism and COVID-19: Impacts and implications for advancing and resetting industry and research. *Journal of Business Research*. 2020. Т. 117. С. 312–321. URL: https://doi.org/10.1016/j.jbusres.2020.06.015 (дата звернення: 06.06.2023).

24. SiteMinder's Hotel Booking Trends: New analysis of 100 million reservations shows surge in global traveller confidence for 2023 URL: https://www.siteminder.com/news/siteminder-hotel-booking-trends-2022/ (дата звернення: 06.06.2023).

25. Li-Ming, A.K., Wai, T.B., Exploring Consumers' Attitudes and Behaviours toward Online Hotel Room Reservations. *American Journal of Economics*. 2013. Vol. 3 No. 5C, C.6-11. https://doi.org/10.5923/c.economics.201301.02 (дата звернення: 06.06.2023).

26. The real cost of ‘free’ cancellations. URL:http://www.triptease.com/blog/the-real-cost-of-free-cancellations (дата звернення: 06.06.2023).

27. Jedin M. H., Annathurai K. R. Exploring travellers booking factors through online booking agency. *International Journal of Business Information Systems*. 2020. Т. 35, № 1. С. 45. URL: https://doi.org/10.1504/ijbis.2020.10031636 (дата звернення: 06.06.2023).

28. Masiero L., Viglia G., Nieto-Garcia M. Strategic consumer behavior in online hotel booking. *Annals of Tourism Research. 2020*. Т. 83. С. 102947. URL: https://doi.org/10.1016/j.annals.2020.102947 (дата звернення: 06.06.2023).

29. Romero Morales D., Wang J. Forecasting cancellation rates for services booking revenue management using data mining. *European Journal of Operational Research.* 2010. Т. 202, № 2. С. 554–562. URL: https://doi.org/10.1016/j.ejor.2009.06.006 (дата звернення: 06.06.2023).

30. Huang, H-C., Chang, A.Y., Ho, C.C. Using Artificial Neural Networks to Establish a Customer-cancellation Prediction Model. 2013. *Przeglad Elektrotechniczny.* 89(1b). С. 178-180.

31. Leeuwen, van R. Cancellation Predictor for revenue-management applied in hospitality industry. 2018. Vrije Universiteit Amsterdam. С. 26

32. Falk M., Vieru M. Modelling the cancellation behaviour of hotel guests. *International Journal of Contemporary Hospitality Management.* 2018. Т. 30, № 10. С. 3100–3116. URL: https://doi.org/10.1108/ijchm-08-2017-0509 (дата звернення: 06.06.2023).

33. Azhar Y., Mahesa G. A., Mustaqim M. C. Prediction of hotel bookings cancellation using hyperparameter optimization on Random Forest algorithm. J*urnal Teknologi dan Sistem Komputer.* 2020. Т. 9, № 1. С. 15–21. URL: https://doi.org/10.14710/jtsiskom.2020.13790 (дата звернення: 06.06.2023).

34. Чорноус Г. Оптимізація ціноутворення на основі моделей інтелектуального аналізу даних. *Вісник Київського національного університету імені Тараса Шевченка. Економіка*. 2015. Вип. 7. С. 52-58. URL: http://nbuv.gov.ua/UJRN/VKNU_Ekon_2015_7_9. (дата звернення: 06.06.2023).

35. Черняк О., Чорноус Г. Інтелектуальний аналіз даних у бізнесі з використанням IBM SPSS Modeler : навч. посіб. Київ : Київський університет, 2020. 263 с.

36. Чорноус Г. Моделювання процесу процесу прийняття управлінських рішень в соціально-економічних системах на основі інтелектуального аналізу даних: дис. … д-ра екон.наук: 08.00.11/ДЗВО «Київ. нац.ун-т ім. Тараса Шевченка». - Київ, 2015. 485 с.

37. Computational Intelligence Approaches for Energy Load Forecasting in Smart Energy Management Grids: State of the Art, Future Challenges, and Research Directions / S. Fallah та ін. *Energies*. 2018. Т. 11, № 3. С. 596. URL: https://doi.org/10.3390/en11030596 (дата звернення: 06.06.2023).

38. Data Science in Economics: Comprehensive Review of Advanced Machine Learning and Deep Learning Methods / S. Nosratabadi та ін. Mathematics. 2020. Т. 8, № 10. С. 1799. URL: https://doi.org/10.3390/math8101799 (дата звернення: 06.06.2023).

39. Zhou Z.H. Ensemble Methods: foundations and algorithms. Chapman and Hall. 1 ed. CRC, 2012

40. Теслюк В. Градієнтні методи розв’язання оптимізаційних задач: Ч.3. Конспект лекцій з курсу “Методи синтезу та оптимізації” для студентів базового напрямку “Комп’ютерні науки”. 2013. Львів, 67 с.

41. Марценюк В., Мілян Н. Огляд методів оптимізації в машинному навчанні: градієнтний спуск та стохастичний градієнтний спуск. *Актуальні задачі сучасних технологій.* Матеріали IX Міжнародної науково-технічної конференції молодих учених та студентів (25-26 листопада 2020 р.). 2020, Тернопіль.

42. What is bagging? URL: https://www.ibm.com/topics/bagging (дата звернення: 06.06.2023).

43. Wolpert D. H. Stacked generalization. Neural Networks. 1992. Т. 5, № 2. С. 241–259. URL: https://doi.org/10.1016/s0893-6080(05)80023-1 (дата звернення: 06.06.2023).

44. James, G.M. Majority vote classifiers: theory and applications. 1998. C.112. URL: https://hastie.su.domains/THESES/gareth_james.pdf (дата звернення: 06.06.2023).

45. RapidMiner. Cross Validation documentation. URL: https://docs.rapidminer.com/latest/studio/operators/validation/cross_validation.html (дата звернення: 06.06.2023).

46. Кононова К. Машинне навчання: методи та моделі: підручник. 2020. Харків, 301 с.

47. Олещенко Л. М. Машинне навчання: комп'ютерний практикум. 2022. Київ, 92 с.

48. Distance metrics. URL: https://numerics.mathdotnet.com/Distance (дата звернення: 06.06.2023).

49. RapidMiner. Neural Net documentation. URL: https://docs.rapidminer.com/10.1/studio/operators/modeling/predictive/neural_nets/neural_net.html (дата звернення: 06.06.2023).

50. Бурлєєв О., Василенко О., Іваненко Р. Ефективність використання штучних нейронних мереж в економіці. *Економіка та суспільство.* 2021. № 31. URL: https://doi.org/10.32782/2524-0072/2021-31-27 (дата звернення: 06.06.2023).

51. XGBoost Documentation. URL: https://xgboost.readthedocs.io/en/stable/ (дата звернення: 06.06.2023).

52. XGBoost: What it is, and when to use it. URL: https://www.kdnuggets.com/2020/12/xgboost-what-when.html (дата звернення: 06.06.2023).

53. Кононенко В., Красношлик Н. Використання методів бустінгу для задач машинного навчання *Cherkasy University Bulletin: Applied Mathematics. Informatics.* 2021. С. 58-68

54. Why RapidMiner? URL: https://rapidminer.com/why-rapidminer/ (дата звернення: 06.06.2023).

55. Hotel booking demand datasets. URL: https://www.sciencedirect.com/science/article/pii/S2352340918315191#bib8 (дата звернення: 06.06.2023).

56. RapidMiner. Normalization documentation. URL: https://docs.rapidminer.com/latest/studio/operators/cleansing/normalization/normalize.html (дата звернення: 06.06.2023).