УДК 629.33:681.51 **DOI: 10.30977/BUL.2219-5548.2022.99.0.131**

ВИКОРИСТАННЯ ПЛАТФОРМИ ORANGE ДЛЯ АНАЛІЗУ ДАНИХ

Пронін С. В., Сотников А. Д.
Харьковский национальный автомобильно-дорожный университет

Анотація. *У статті розглядаються способи створення систем машинного навчання та аналізу даних. Отримані результати демонструють можливість використання бібліотеки інтелектуального аналізу даних Orange для створення їхніх систем. За допомогою бібліотеки можна працювати з великими масивами даних, що е однією з основних характеристик для цих систем. Використовуючи систему, можна працювати з різноманітними форматами даних, загружати набори даних з мережі, використовувати оригінальні скрипти.*
Ключові слова: *аналіз даних, машинне навчання, ймовірність вибору, python, orange.*

Вступ

Застосування автоматизованих систем аналізу даних дозволяє за наявності загальнодоступних даних створювати програми для аналізу переваг окремих користувачів. Такі системи на сьогодні використовують у різноманітних сферах комерції, оскільки вони дозволяють покупцям отримати інформацію про ті чи інші і на підставі цього біль ефективно організовувати бізнес.

На сьогодні користувачі під час роботи в мережі інтернет генерують різноманітними даними, створюючи великі обсяги інформації. Вона може бути корисною як для звичайних користувачів, так і для промислових компаній. Це дає можливість використовувати дані для аналізу переваг користувачів. Одна з проблем використання інформації, що накопичується, це її неструктурованість. Крім того, деяким групам користувачів необхідний не весь набір даних, а конкретний тип, що необхідно вибрати. Для вирішення цієї проблеми використовують технології Bisness Intellegent (BI) [1], а для її реалізації запропоновані різноманітні програмні продукти та фреймворки.

Проблемою вибору програмного забезпечення є досягнення балансу між ціною та функціональністю продукту. На сьогодні існують різноманітні середовища, які об'єднують свободну ліцензію та широкий функціонал. Додатковою властивістю використання такого програмного забезпечення є можливість використання його під час навчального процесу.

У роботі автори вирішують завдання створення системи аналізу даних за допомогою прикладного програмного забезпечення.

Аналіз публікацій

Останнім часом на ринку IT з'явилося достатньо багато продуктів та вільно розповсюджених фреймворків, які дозволяють створювати програмні системи з аналізу даних та машинного навчання. Основними мовами, що використовуються, є мови програмування Python та Java. Це визначено особливостями функціонування та достатньо розвинутим апаратом, зокрема мови Python, для роботи зі структурами даних, що є необхідними для створення систем машинного навчання. Мову Java є достатньо популярною та поширеною мовою програмування, на якій створюється велика кількість проектів. Основним продуктом тут е вільне програмне забезпечення для аналізу даних та машинного навчання Weka [2]. Weka є набором засобів візуалізації та алгоритмів для інтелектуального аналізу даних та вирішення завдань прогнозування разом з графічною оболонкою користувача для доступу до них. Weka дозволяє виконувати такі завдання аналізу даних, як підготовку даних, вибір ознак, кластеризацію, класифікацію, регресійний аналіз та візуалізацію результатів [2].

Варто зазначити, що в проектах різноманітних них наукових досліджень використовується мова програмування R [3] – вільне програмне середовище з відкритим вихідним кодом, основною функцією якого є здійснення статистичного аналізу та математичного моделювання. Також мова використовується як широкий функціонал для первинного аналізу даних та роботи з графікою.

Але незалежно від поширеності Java та простоти мови R на сьогодні основна маса бібліотек для машинного навчання створюється на мові Python. Оскільки створення програмного забезпечення для аналізу даних не обмежується тільки застосуванням спеці-

алізованих фреймворків для машинного навчання, проаналізуємо загальний функціонал мови Python [4] для створення готових продуктів.

Мова Python має простий та зрозумілий синтаксис, вона є інтерпретованою, тобто достатньо гнучкою для написання програм. Все це скорочує час для вивчення її основ, та підвищує швидкість розроблення загалом [4].

Для вирішення нашого завдання проаналізуємо доступні бібліотеки для роботи з даними на мові Python.

Бібліотека scikit-learn надає можливість створення систем машинного навчання [5].

Бібліотека scikit-learn складається з 35 модулів. Кожен модуль складається з класів і функцій та вирішує такі завдання:

- кластеризації;
- перехресна перевірка;
- набори даних;
- скорочення розмірності;
- алгоритмічні композиції;
- витяг ознак;
- відбір ознак;
- оптимізація параметрів алгоритму;
- множинне навчання;

Існують такі моделі scikit-learn:

- узагальнені лінійні моделі;
- методи дискримінантного аналізу;
- наївний байесовський класифікатор;
- нейронні мережі;
- метод опорних векторів:
- дерева прийняття рішень

Anaconda Individual Edition – це платформа розповсюдження Python з понад 25 мільйонами користувачів у всьому світі [6]. Платформа дає можливість використовувати відкриті пакети Conda, R, Python тощо. Individual Edition – це гнучке рішення з відкритим вихідним кодом для втілення, створення, розповсюдження, встановлення, оновлення та керування програмним забезпеченням міжплатформовим способом. Conda дозволяє легко керувати кількома середовищами даних, які можна підтримувати та працювати з ними окремо без перешкод один від одного. Також до складу платформи належить графічний інтерфейс для робочого столу, який постачається з Anaconda Individual Edition. Це дозволяє легко запускати програми, керувати пакетами та середовищами без використання команд командного рядка, створювати моделі машинного навчання, використовуючи пакети Python, створені спільнотою з відкритим кодом, зокрема scikit-learn, TensorFlow та PyTorch. Завдяки виданням Anaconda Team і Enterprise стек може впоратися з найсучаснішими вимогами корпоративних наукових даних, дає доступ до програмного забезпечення з відкритим кодом для проектів у різноманітних галузях.

Бібліотека інтелектуального аналізу даних Orange. Orange – це інструмент для візуалізації та аналізу даних з відкритим вихідним кодом [7]. Orange розробляється в лабораторії біоінформатики на факультеті комп'ютерних та інформаційних наук Люблянського університету (Словенія).

Orange – це бібліотека Python. Інтелектуальний аналіз даних (Data mining) здійснюється за допомогою візуального програмування або сценаріїв Python. Сценарії Python можуть виконуватися у вікні термінала, інтегрованих середовищах PyCharm і PythonWin, або оболонках Python.

Містить набір віджетів, які згруповані в п'ять розділів:

- Data – віджети для введення / виведення даних, фільтрації, маніпулювання вибірками, а також велика кількість навчальних наборів даних (від класичних Titanic і Iris до статистики ДТП в Словенії за 2014 рік);
- Visualize – віджети для загальної (прямокутна діаграма, гістограми, точкова діаграма) і багатовимірної візуалізації (мозаїчна діаграма, діаграма-сито);
- Model – набір алгоритмів машинного навчання для класифікації і регресії;
- Evaluate – крос-валідація, процедури на основі вибірки, аналіз методів передбачення;
- Unsupervised – алгоритми кластеризації (k-середні, ієрархічна кластеризація) і проекції даних (багатовимірне масштабування, аналіз головних компонент, аналіз відповідності) [8].

У разі необхідності до комплекту Orange можна додатково завантажити ще декілька наборів віджетів:

- Associate – датамайнінг повторюваних наборів елементів і навчання асоціативним правилам;
- Bioinformatica – аналіз наборів генів і доступ до бібліотек геномів;
- Data fusion – об'єднання різноманітних наборів даних, колективна матрична факторизація та дослідження прихованих чинників;
- Educational – навчання концепціям machine learning;
- Geo – робота з геоданими;

– Image analytics – робота з зображеннями, аналіз нейронними мережами;

– Network – мережевий аналіз;

– Text mining – робота з мовою та аналіз тексту;

– Time series: – аналіз і моделювання часових рядів;

– Spectroscopy – аналіз і візуалізація спектральних наборів даних [9–10].

Оскільки однією з функцій системи інтелектуального аналізу даних є написання коду великого обсягу, то помилки під час визначення завдання призведуть до втрат часу. Однією з основних переваг використання Orange є здатність до моделювання та дослідження поведінки моделей, що дозволяє під час створення системи з аналізу даних на першому етапі побудувати модель на реальних даних, а потім переходити до етапу створення програмної частини продукту.

Мета та постановка завдання

Метою роботи є створення системи аналізу даних для визначення переваг споживачів з використання вільного програмного забезпечення.

Для досягнення поставленої мети необхідно вирішити такі завдання:

- проаналізувати різноманітні бібліотеки для аналізу даних;
- створити систему аналізу даних за допомогою пакета.

Модель для аналізу ймовірності вибору

Для створення системи аналізу ймовірності вибору з середовища kaggle [11] було використано датасет «Telco Customer Churn» – дані о користувачах телекомунікаційних послуг. Його цільовим показником є стовпець, який демонструє згоду або відмову того чи іншого користувача від послуг. У датасеті цей стовпець називається Churn.

Зміст датасету:

кожен рядок містить інформацію про клієнта, кожен стовпець містить атрибути клієнта, описані в стовпці Метадані.

Набір даних містить таку інформацію:

– клієнти, які виїхали протягом останнього місяця, – графа churn;

– послуги, на які зареєструвався кожен клієнт, – телефон, декілька ліній, інтернет, безпека в інтернеті, резервне копіювання в інтернеті, захист пристрою, технічна підтримка, потокове телебачення та фільми;

– інформація про рахунок клієнта – протягом якого часу вони є клієнтами, контракт, спосіб оплати, безпаперова оплата, щомісячні платежі та загальна вартість;

– демографічна інформація про клієнтів – стать, віковий діапазон та наявність у них партнерів та утриманців.

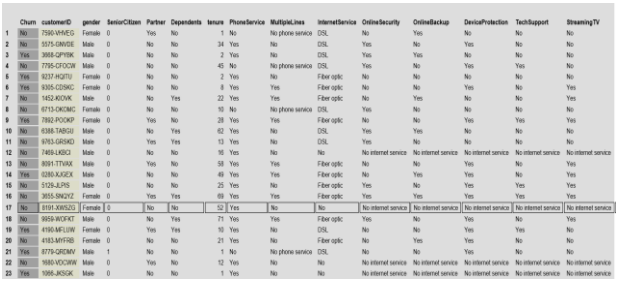
Кожен рядок містить інформацію про клієнта, кожен стовпець містить його атрибути, наведені в стовпці Метадані.

Вихідні дані містять 7043 рядки (клієнти) та 21 стовпець (функції).

Приклад датасету наведено на рис. 1.

Вихідними даними в цьому завданні є вищезазначений датасет який ми завантажуємо з мережі інтернет за допомогою віджету Dataset.

Наше завдання – використовуючи методи ML, реалізовані віджетами Orange, спрогнозувати вибір користувача.



	Churn	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	OnlineBackup	DeviceProtection	TechSupport	StreamingTV
1	No	7590-VHVEG	Female	0	Yes	No	1	No	No phone service	DSL	No	Yes	No	No	No
2	No	5575-GNVDE	Male	0	No	No	34	Yes	No	DSL	Yes	No	Yes	No	No
3	Yes	3668-QPYBK	Male	0	No	No	2	Yes	No	DSL	Yes	Yes	No	No	No
4	No	7795-CFOCW	Male	0	No	No	45	No	No phone service	DSL	Yes	No	Yes	Yes	No
5	Yes	9237-HQITU	Female	0	No	No	2	Yes	No	Fiber optic	No	No	No	No	No
6	Yes	9305-CDSKC	Female	0	No	No	8	Yes	Yes	Fiber optic	No	No	Yes	No	Yes
7	No	1452-KIOVK	Male	0	No	Yes	22	Yes	Yes	Fiber optic	No	Yes	No	No	Yes
8	No	6713-OKOMC	Female	0	No	No	10	No	No phone service	DSL	Yes	No	No	No	No
9	Yes	7892-POOKP	Female	0	Yes	No	28	Yes	Yes	Fiber optic	No	No	Yes	Yes	Yes
10	No	6388-TABGU	Male	0	No	Yes	62	Yes	No	DSL	Yes	Yes	No	No	No
11	No	9763-GRSKD	Male	0	Yes	Yes	13	Yes	No	DSL	Yes	No	No	No	No
12	No	7469-LKBCI	Male	0	No	No	16	Yes	No	No	No internet service	No internet service	No internet service	No internet service	No internet service
13	No	8091-TTVAX	Male	0	Yes	No	58	Yes	Yes	Fiber optic	No	No	Yes	No	Yes
14	Yes	0280-XJGEX	Male	0	No	No	49	Yes	Yes	Fiber optic	No	Yes	Yes	No	Yes
15	No	5129-JLPIS	Male	0	No	No	25	Yes	No	Fiber optic	Yes	No	Yes	Yes	Yes
16	No	3655-SNQYZ	Female	0	Yes	Yes	69	Yes	Yes	Fiber optic	Yes	Yes	Yes	Yes	Yes
17	No	8191-XWSZG	Female	0	No	No	52	Yes	No	No	No internet service	No internet service	No internet service	No internet service	No internet service
18	No	9959-WOFKT	Male	0	No	Yes	71	Yes	Yes	Fiber optic	Yes	No	Yes	No	Yes
19	Yes	4190-MFLUW	Female	0	Yes	Yes	10	Yes	No	DSL	No	No	Yes	Yes	No
20	No	4183-MYFRB	Female	0	No	No	21	Yes	No	Fiber optic	No	Yes	Yes	No	No
21	Yes	8779-QRDMV	Male	1	No	No	1	No	No phone service	DSL	No	No	Yes	No	No
22	No	1680-VDCWW	Male	0	Yes	No	12	Yes	No	No	No internet service	No internet service	No internet service	No internet service	No internet service
23	Yes	1066-JKSGK	Male	0	No	No	1	Yes	No	No	No internet service	No internet service	No internet service	No internet service	No internet service

Рис. 1. Приклад датасету

У властивостях віджету зазначено, які поля наборів та яких типів будуть target і features – числові, категоріальні, тимчасові або текстові, а які поля взагалі не треба обробляти. Цей процес здійснюється віджетом, але автоматичне визначення типу полів часто дає некоректні результати, тому краще зробити все власноруч.

Після цього віджет Data Table з'єднаємо з віджетом File набору Train з розділу Data для відображення завантаженого набору даних. Запускаємо віджет Data Table з завантаженою таблицею з даними. Зверніть увагу, що у верхній лівій частині віджету наведена деяка статистика полів і записів завантаженого набору даних (рис. 1).

Оскільки датасет має велику кількість стовпців, а не всі вони впливають на кінцевий результат, необхідно здійснити операцію видалення зайвих даних за допомогою віджетів Select Data та Rank.

Віджет Select Data дозволяє це зробити власноруч.

На рис. 2 наведено вікно віджетів. Ми як вхідні дані моделі використовуємо лише чи-

слові показники. Показники, які містять нечислові дані, ігноруються.

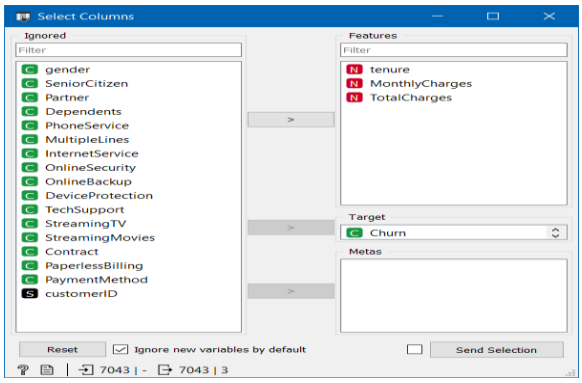


Рис. 2. Вікно віджету Select Data

Другий спосіб ранжування даних здійснюють за допомогою віджету Rank.

Віджет Rank оцінює змінні відповідно до їхньої кореляції з дискретною або числовою цільовою змінною на основі внутрішніх показників (приріст інформації, хі-квадрат та лінійна регресія) та будь-яких підключених зовнішніх моделей, що підтримують аналіз (лінійна регресія, логістична регресія випадковий ліс, SGD тощо). Віджет також може обробляти некеровані дані, але лише зовнішніми балами, зокрема PCA (рис.3).

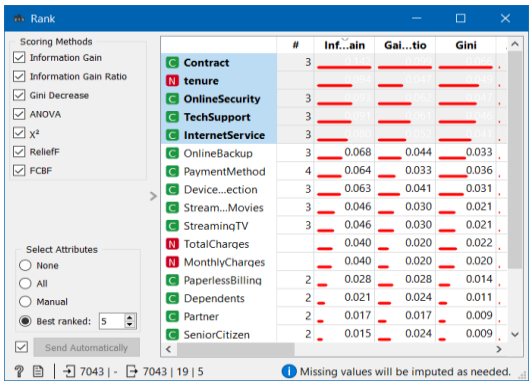


Рис. 3. Віджет Rank

Далі побудуємо модель класифікації, яка з відомих ознака на тренувальному наборі буде намагатися передбачити вибір клієнта.

Завдання аналізу ймовірності вибору концептуально подібне до завдання класифікації, тому для побудови системи аналізу ймовірності вибору скористаємося методами класифікації.

У машинному навчанні завдання класифікації вирішується з використанням навчання з учителем, оскільки класи визначаються заздалегідь і для прикладів навчальної множини мітки класів вже задані. Аналітичні моделі, що вирішують завдання класифікації, називаються класифікаторами.

Завдання класифікації є ще однією з базових задач прикладної статистики та машинного навчання, а також штучного інтелекту загалом, оскільки класифікація є одним з найбільш зрозумілих і простих для інтерпретації технологій аналізу даних процесів, а правила, які класифікують, можуть бути визначені природною мовою [12].

До поширених методів вирішення завдань класифікації належать такі:

– нейронні мережі;
– логістична і пробіт-регресія;
– дерева рішень;
– метод найближчого сусіда;
– метод опорних векторів;
– дискримінантний аналіз;
– байєсівська(наївна) класифікація.

Для застосування цих методів завантажуємо на полотно віджети Logistic Regression, Random Forest, Ada Boost і Neural Network з розділу Model (рис.4).

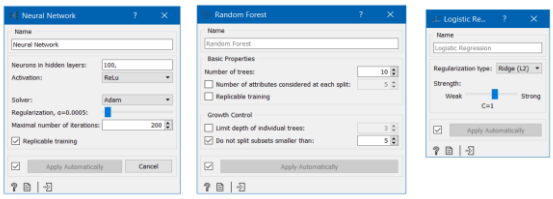
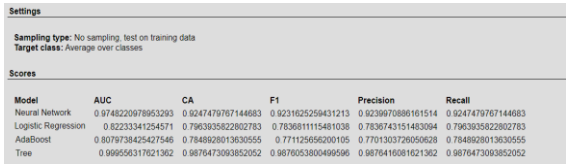


Рис. 4. Віджети моделей

Далі перевіримо результати роботи вибраних алгоритмів і розрахуємо їхні оціночні метрики. Для цього викладемо на полотно віджет Test and Score з розділу Evaluate і подамо на його вхід дані з віджетів Ada Boost, Logistic Regression, Random Forest і Neural Network. На підставі цих даних віджет Test and Score автоматично почне розраховувати результати роботи моделей, побудованих з очищеного набору даних цими алгоритмами. Для розрахування моделі був використаний метод семпліювання, за яким вихідна навчальна вибірка поділяється на 80 % робочої навчальної вибірки та 20 % валідаційної вибірки. Цей цикл повторюється 10 разів (рис. 5).

Найкращі результати, крім метрики AUC, дала нейронна мережа та дерево рішень, тому в подальшому будемо використовувати їх. Для побудови робочої моделі класифікації викладаємо на полотно віджети Neural Network та Tree з розділу Model, віджет Data

Sampler з розділу Data і віджет Predictions з розділу Evaluate. Віджет Data Sampler має поділити навчальну вибірку на дві частини у співвідношенні 80/20 %, а віджет Predictions має здійснити в наборі даних Test власне передбачення цільового поля на підставі моделі, побудованої виджетом Neural Network.


Settings

Sampling type: No sampling, test on training data
Target class: Average over classes

Scores

Model	AUC	CA	F1	Precision	Recall
Neural Network	0.974822097895329	0.924747976714468	0.923162525943121	0.923997088616151	0.924747976714468
Logistic Regression	0.82233341254571	0.796393582280278	0.783681111548103	0.783674315148309	0.796393582280278
AdaBoost	0.807973842542754	0.784892801363055	0.771125656200105	0.770130372605062	0.784892801363055
Tree	0.999556317621362	0.987647309385205	0.987605380049959	0.987641608162136	0.987647309385205

Рис. 5. Результати роботи моделей

Також для візуалізації роботи моделей будемо використовувати віджети ROC Analysis та Confusion Matrix.

Віджет ROC Analysis демонструє криві ROC для випробуваних моделей та відповідну опуклу оболонку. З огляду на витрати на помилкові спрацьовування та помилкові негативні результати він також може визначити оптимальний класифікатор та поріг (рис. 6).

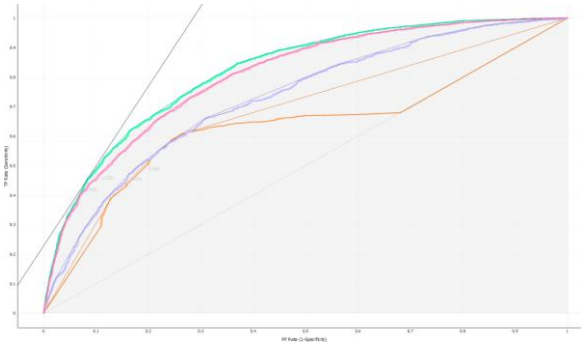


Рис. 6. Графік точності моделі

Варіант Цільовий клас вибирає позитивний клас. Якщо є більше двох класів, віджет розглядає всі інші як єдиний негативний клас.

Якщо результати тестування містять більше ніж один класифікатор, користувач може вибрати криві, які необхідно побудувати.

Варіант Показати опуклі криві використовують для опуклих кривих над кожним окремим класифікатором (тонкі лінії на вирізі зліва). За допомогою цього варіанта наносять опуклу оболонку над кривими ROC для всіх класифікаторів (товста жовта лінія). Побудова графіків обох типів опуклих кривих має сенс, оскільки вибір порога у ввігнутій частині кривої не може дати оптимальних результатів, нехтуючи матрицею витрат. Крім того, можна досягти будь-якої точки опуклої кривої, комбінуючи класифікатори, тобто точки на межі увігнутої області.

Діагональна лінія відображає поведінку випадкового класифікатора.

Коли дані надходять із численних ітерацій навчання та тестування, зокрема з перехресної перевірки k-кратного результату, результати можуть бути (і зазвичай є) усередненими. Варіанти усереднення:

– злиття (очікуваний процес roc) обробляє всі дані тесту так, ніби вони надійшли з однієї ітерації

– по вертикалі (усереднюють криві по вертикалі, демонструючи відповідні довірчі інтервали);

– поріг перетинає поріг, усереднює положення кривих на них і демонструє горизонтальні та вертикальні довірчі інтервали

– жодна не робить усереднення, але замість цього друкує всі криві.

Другий аркуш налаштувань здійснює аналіз кривої. Користувач може визначити вартість помилкових спрацьовувань та помилкових негативних результатів, а також імовірність попереднього цільового класу. Обчислення з даних задає його частку прикладів цього класу в даних.

Лінія ізопродуктивності – це лінія в просторі ROC, всі точки на якій мають однаковий прибуток / збиток. Лінія вгорі ліворуч краща за лівою праворуч. Напрямок лінії залежить від вищезазначених витрат та ймовірностей. У сукупності отримуємо рецепт зображення оптимального порога для заданих витрат: це точка, коли тангенс із заданим нахилом торкається кривої. Якщо ми спускаємось вниз або праворуч, знижується продуктивність.

Віджет може відображати рядок продуктивності, який змінюється, якщо користувач змінює параметри. Точки, де лінія торкається будь-якої кривої, тобто оптимальна точка для будь-якого із заданих класифікаторів, також визначається (необхідна ймовірність цільового класу для прикладу, який буде класифікований до цього класу).

Віджет дозволяє визначити витрати від 1 до 1000. Одиниці вимірювання не важливі, як і величини. Важливим є співвідношення між цими двома витратами, тому встановлення їх на 100 і 200 дасть той самий результат, що і 400 та 800.

Матриця плутанини дає кількість / частку прикладів з одного класу, класифікованого до іншого (або того самого). Крім того, вибір

елементів матриці подає відповідні приклади на вихідний сигнал. Таким чином, можна спостерігати, які конкретні приклади були визначені неправильно класифікованими (рис. 7).

		Predicted		
		No	**Yes**	**Σ**
Actual	**No**	5020	154	**5174**
	Yes	376	1493	**1869**
	Σ	**5396**	**1647**	**7043**

Рис. 7. Загальний вигляд матриці плутанини для дерева рішень

Як зазначено на рис. 7, дерево рішень з 5396 відповідей «не» змогло правильно класифікувати 5020, а із 1647 відповідей «так» – 1493.

Фрагмент роботи системи можна побачити на рис. 8

	Churn	Neural Network	Tree	AdaBoost	Logistic Regression	Neural Network (No)	Neural Network (Yes)	Tree (No)	Tree (Yes)	AdaBoost (No)	AdaBoost (Yes)	Logistic Regression (No)	Logistic Regression (Yes)	Fold
1	No	No	Yes	Yes	Yes	0.527155	0.472845	0	1	0.219087	0.780913	0.430304	0.569696	1
2	Yes	Yes	Yes	Yes	No	0.484582	0.515418	0	1	0.000175013	0.999825	0.624079	0.375921	1
3	No	No	No	No	No	0.931385	0.0686146	0.954545	0.0454545	0.999999	1.09194e-06	0.919004	0.080996	1
4	No	No	No	No	No	0.914459	0.0855411	0.928384	0.0716161	1	3.17943e-07	0.754788	0.245212	1
5	No	No	No	No	No	0.938718	0.0612822	1	0	0.99999	1.00168e-05	0.884478	0.115522	1
6	Yes	Yes	Yes	Yes	Yes	0.284803	0.715197	0	1	0.220277	0.779723	0.35803	0.64197	1

Рис. 8. Загальний результат роботи моделей

Таким чином, можна дійти висновку, що найкращий результат отримали за допомогою нейронної мережі та дерева рішень, тому вони є найкращим варіантом для вирішення завдання аналізу ймовірності вибору для обраного датасету.

Загальна схема системи наведена на рис. 9.

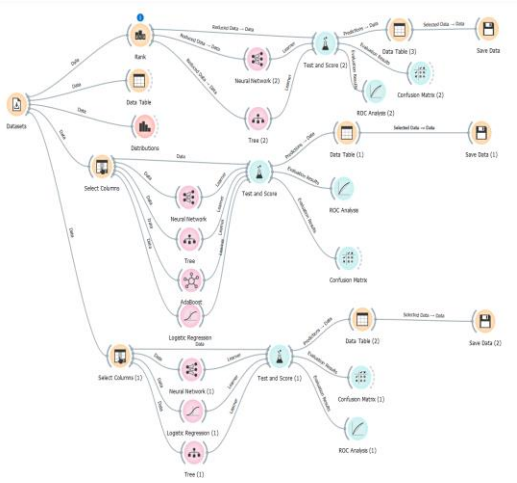


Рис. 9. Загальний вигляд моделі для аналізу ймовірності вибору

Висновок

Отримані результати дають можливість використання бібліотеки інтелектуального аналізу даних Orange для створення систем щодо аналізу переваг споживачів у процесі вибору товарів та послуг. Також бібліотека може працювати з великими масивами даних, що є однією з основних характеристик для цих систем, з різноманітними їхніми форматами, вона здатна використовувати поза штатними елементами оригінальні скрипти.

Конфлікт інтересів

Автори стверджують, що немає конфлікту інтересів щодо публікації цієї статті.

Література

1. Business Intelligence and Analytics Market worth 60.49 Billion USD by 2027. Emergen Research.
2. Weka 3: Machine Learning Software in Java. URL: https://www.cs.waikato.ac.nz/ml/weka/ (дата звернення: 29.9.2022).
3. The R Project for Statistical Computing. URL: https://www.r-project.org.
4. Documentation Python.org. URL: https://www.python.org/ (дата звернення: 29.9.2022).
5. Scikit-learn. Machine Learning in Python. http://scikit-learn.org (дата звернення: 29.9.2022).
6. Anaconda. URL https://www.anaconda.com/products/individual (дата звернення: 29.9.2022).
7. Orange. URL: https://orangedatamining.com/download/#windows (дата звернення: 29.9.2022).
8. Orange Data Mining Library. URL: https://orange3.readthedocs.io/projects/orange-data-mining-library/en/latest/index.html (дата звернення: 29.9.2022)
9. Визуальный анализ данных с Python и Orange 3/. URL: https://ansmirnov.ru/python-orange-anaconda-overview/ (дата звернення: 29.9.2022).
10. Интерактивный DataMining. URL: https://www.infoculture.ru/wp-content/uploads/2019/04/DataSreda-IMaining1.pdf (дата звернення: 29.9.2022).
11. Telco Customer Churn. URL: https://www.kaggle.com/blastchar/telco-customer-churn (дата звернення: 29.9.2022)
12. Барсегян А. А., Куприянов М. С. Методы и модели анализа даннях С.-Петербург: БХВ Петербург, 2004. 134 с.

References

1. Business Intelligence and Analytics Market worth 60.49 Billion USD by 2027. Emergen Research.

2. Weka 3: Machine Learning Software in Java Retrived from: https://www.cs.waikato.ac.nz/ml/weka/ (accessed: 29.9.2022).
3. The R Project for Statistical Computing. Retrived from: https://www.r-project.org weka/ (accessed: 29.9.2022).
4. Documentation Python.org Retrived from: https://www.python.org/ (accessed: 29.9.2022).
5. Scikit-learn. Machine Learning in Python. Retrived from: http://scikit-learn.org (accessed: 29.9.2022).
6. Anaconda. Retrived from: https://www.anaconda.com/products/individual (accessed: 29.9.2022).
7. Orange. Retrived from: https://orangedatamining.com/download/#windows (accessed: 29.9.2022).
8. Orange Data Mining Library. Retrived from: https://orange3.readthedocs.io/projects/orange-data-mining-library/en/latest/index.html (accessed: 29.9.2022).
9. Vyzualnыi analyz dannыkh s Python y Orange 3 [Visual Data Analysis with Python and Orange 3] Retrived from: https://ansmirnov.ru/python-orange-anaconda-overview/ (accessed: 29.9.2022).
10. Ynteraktyvnii DataMining [Interactive DataMining] Retrived from: https://www.infoculture.ru/wp-content/uploads/2019/04/DataSreda-IMaining1.pdf (accessed: 29.9.2022).
11. Telco Customer Churn https://www.kaggle.com/blastchar/telco-customer-churn (accessed: 29.9.2022).
12. Barsegyan A. A., Kuprianov M. S. Metody` i modeli analiza danny`kh [Data Analysis Methods and Models]. S.-Peterburg: BKhV Peterburg, 2004. 134 s.

Пронін Сергій Вікторович, к.т.н., доцент кафедри комп'ютерних технологій і мехатроніки, psv59777@gmail.com, тел. 057-707-37-43;
Харківський національний автомобільно-дорожній університет, вул. Ярослава Мудрого, 25, м. Харків, 61002, Україна;
Сотников Андрій Дмитрович студент ХНАДУ andrew.sotnikov1437@gmail.com,
тел. 099 711 1373/
Харківський національний автомобільно-дорожній університет, вул. Ярослава Мудрого, 25, м. Харків, 61002, Україна.

Using the Orange platform for data analysis

Abstract. Problem. *The ability to create programs based on publicly available data to assess the preferences of individual users can be implemented, in particular, in the field of e-commerce – the knowledge of which product is best for the buyer, help to more effectively organize contextual product offerings, which in turn increases business efficiency in general. Today, users generate various data when working on the Internet. As a result, there are large amounts of different information. This information can be useful for both regular users and large companies. This makes it possible to use this data to analyze users' preferences. One of the problems with the use of accumulated information is its crude, unstructured nature. In addition, different user groups do not need the entire data set, but their own target sample. To solve this problem today we use technologies commonly known as Bisness Intellegent (BI), and to implement it, various software products and frameworks are offered. The problem with choosing software here is to strike a balance between price and product functionality.* ***Goal.*** *The aim of the work is to create a data analysis system to assess consumer preferences for the use of free software.* ***Methodology.*** *The approaches adopted in the work to the solution of the set goal are based on the review of the approaches to the assessment of consumer preferences, the analysis of the software for the solution of the set goal.* ***Results.*** *The obtained results show the possibility of using the Orange data mining library to create systems for assessing consumer preferences when choosing goods and services. The library also showed good ability when working with large data sets, which is one of the main characteristics for these systems. You can also note such features of the system as the ability to work with different data formats, the ability to download data sets from the network, to use the original scripts past the standard elements.* ***Originality****. He originality is in using the Orange Data Mining Library to create data analysis systems.* ***Practical value****. The obtained results can be recommended when creating systems for analysis and work with big data.*

Keywords: *data analysis, machine learning, probability of choice, python, orange*

Pronin Sergey Viktorovich, Ph.D., Associate Professor of Computer Technology and Mechatronics, psv59777@gmail.com,
tel. 057-707-37-43
Kharkiv National Automobile and Road University, 61002, Ukraine, Kharkiv, street Yaroslav the Wise, 25.
Sotnikov Andrey Dmitrovich the student of KhNADU andrew.sotnikov1437@gmail.com,
tel. 099 711 1373
Kharkiv National Automobile and Road University, 61002, Ukraine, Kharkiv, street Yaroslav the Wise, 25.