Т1. №1(33)’2023 https://doi.org/10.31474/2786-9024/v1i1(33).296156

ВИКОРИСТАННЯ ЗАСОБІВ MACHINE LEARNING І БІБЛІОТЕК PYTHON ДЛЯ ПРОГНОЗУВАННЯ ІЗ ЗАСТОСУВАННЯМ РЕГРЕСІЙ

Я.Л. Байтельман[1], Г.О. Теличко[1], Д.О. Жуковська[1]

[1] Department of Automation and Telecommunications, Donetsk National Technical University, Lutsk, Ukraine

E-mail: hanna.telychko@donntu.edu.ua

Отримано 19.12.2023
Прийнято до публікації 31.12.2023
Опубліковано 01.04.2024

АНОТАЦІЯ

Мета роботи полягає в прогнозуванні середньої суми чеку в залежності від часу перебування покупців в торговельному залі та прогнозуванні середньої суми чеку в залежності від сезонних особливостей попиту з урахуванням постійного щорічного зростання цін.

Проаналізовано методики з прогнозування. Розглянуто використання Python-бібліотек. Розроблено код розв’язання зазначених вище задач з прогнозування через лінійну, поліноміальну регресії та “випадковий ліс”, а саме, пошук залежності середньої суми чеку від часу, проведеному покупцями в торговельному залі, та залежності середньої суми чеку від сезонних коливань цін протягом кількох років, і відповідного прогнозування. Здійснено порівняння результатів лінійної та поліноміальної регресії за умови меншого або більшого обсягу вхідних даних, надано пояснення щодо доцільності вибору тої чи іншої при моделюванні нециклічних процесів, і чому для циклічних процесів адекватні результати дає модель “випадкового лісу”. Програмний код і приклади вхідних даних викладено у відкритий доступ.

Наукова новизна полягає в шляхах застосування регресій для моделювання та прогнозування середньої суми чеку в залежності від часу перебування покупців в торговельному залі, а також середньої суми чеку в залежності від сезонних особливостей попиту з урахуванням постійного зростання цін.

Обрані приклади циклічного і нециклічного економічних явищ є найпоширенішими, фахівці-практики щоденно стикаються із аналогічними завданнями, тому результати даної роботи надають їм готові рішення, що потребують мінімальної адаптації під практичні потреби, а також демонструють доступність їхнього розгортання на хмарному середовищі AWS і відповідно можливості інтеграції з різними джерелами даних та іншими інформаційними системами.

***Ключові слова**: машинне навчання, регресія, прогнозування, поліноміальна регресія, випадковий ліс, пайтон*

Т1. №1(33)’2023 https://doi.org/10.31474/2786-9024/v1i1(33).296156

ВСТУП

Протягом останнього десятиліття спостерігаємо постійно зростаюче взаємне проникнення методів та інструментів, традиційно притаманних одним галузям господарства і відповідним наукам, до інших; з’являються перетини таких дисциплін, які не існували раніше або існували дуже обмежено. Якщо поєднання економічних і комп’ютерних наук розпочалося майже одразу, як тільки з’ясувалося, що обчислювальні потужності можуть вирішувати питання автоматизації фінансових розрахунків, то засоби машинного навчання є відносно новими для практичного використання в сфері економіки, особливо з метою прогнозування і тим більше з урахуванням специфіки підготовки фахівців. Доволі часто, економісти-практики мають обмежене уявлення про інструменти моделювання і прогнозування засобами машинного навчання, або вважають, що такі засоби не є для них доступними через складність використання. Проте в сучасному світі міждисциплінарні зв’язки стають все ціннішими, постійне самостійне підвищення кваліфікації, в тому числі у споріднених сферах, розглядається, як запорука конкурентоспроможності фахівців, а необхідність розв’язання прикладних задач тільки зростає.

АНАЛІЗ ЛІТЕРАТУРНИХ ДАНИХ ТА ПОСТАНОВКА ПРОБЛЕМИ

В освітній програмі “Економічна аналітика” є така обов’язкова дисципліна, як “Соціально-економічне прогнозування та проектування”, один із наявних навчальних посібників – “Прогнозування соціально-економічних процесів” 2021 року видання [1]. Цей посібник пропонує фундаментальну теоретичну базу із детальним описом методів, наводить пояснення математичного апарату, включаючи розгляд регресійних моделей, проте робочим інструментом обрано Microsoft Excel, що є цілком виправданим для опанування знаннями та навичками, для практичного розв’язання певних розрахункових завдань, але не в автоматизованому режимі. Питання втілення певного автоматизованого рішення з можливістю інтегрувати в інші системи, навіть такі прості, як веб-сайт, неможливо вирішити засобами Microsoft Excel, тоді як після закінчення навчання перед вчорашніми студентами – сьогоднішніми молодими співробітниками комерційних компаній або державних установ цілком вірогідно в найближчому майбутньому поставатимуть конкретні робочі задачі, які потребуватимуть масштабування та інтеграції. В освітній програмі бакалаврів “Економічна кібернетика” є значно складніший за викладенням матеріалу навчальний посібник “Прогнозування соціально-економічних процесів” 2022 року видання [2], адже спеціалізація цих фахівців є більш вузько направленою. Проте виникає питання щодо самоосвіти фахівців, які не мають можливості через обмеження в часі опанувати освітню програму в повному обсязі, і яким водночас необхідно вирішувати практичні задачі. В освітній програмі “Економічна аналітика та бізнес-статистика” є курс з основ програмування на мові Python [3], що вже значно ближче до широкого спектру прикладних потреб, але знову ця інформація доступна в межах формальної академічної освіти і не охоплює людей, зайнятих на постійній основі у виробництві або сфері послуг, а також підприємців – засновників стартапів. Для останньої групи критичним є здатність відшукати потрібні матеріали і в стислий термін (тижні краще ніж місяці) оволодіти новими для них інструментами, перевірити, чи вони підходять для вирішення їхніх задач, одночасно такі інструменти мають відповідати вимогам масштабування та інтеграції з іншими інформаційними системами, програмним забезпеченням, тощо. В англомовному середовищі є досить багато змістовних робіт, як наукового так і прикладного характеру, з питань прогнозування взагалі [4 – 6], так і економічних явищ зокрема, проте робота з ними потребує високого рівня володіння англійською мовою.

Формулювання проблеми: в умовах постійно зростаючої потреби в міждисциплінарних компетенціях існує ряд прикладних задач з прогнозування, зокрема, економічного, серед яких визначення бажаного часу перебування покупців в торговельному залі для отримання найбільшої середньої суми чеку, а також визначення залежності сезонних коливань середньої суми чеку із урахуванням щорічного зростання цін, розв’язання яких можливе засобами машинного навчання.

Мета: розробка програмного коду на основі бібліотек Python для прогнозування зазначених вище явищ. Вибір мови продиктований її поширеністю, доступністю до вільного і безкоштовного використання навіть в комерційних цілях [7], відносною легкістю опанування, можливістю виконувати досить складні вправи з програмування без необхідності встановлення платного або складного середовища.

Задачі:

1. Аналіз тематичних наукових джерел і методичних матеріалів.
2. Експериментальна перевірка запропонованих в них прикладів, їхня адаптація під зазначені вище завдання з прогнозування.

ISSN 2786-9024

 https://doi.org/10.31474/2786-9024/v1i1(33).296156

3. Розробка і тестування програмного коду моделей прогнозування.

4. Аналіз результатів, отриманих від розроблених моделей.

МАТЕРІАЛИ ТА МЕТОДИ ДОСЛІДЖЕНЬ

В роботі використано методи структурного і порівняльного аналізу, з інформаційним і аналітичним підходом розглянуто наукову та методичну літературу, а також онлайн ресурси, застосовано експериментальні методи для перевірки програмного коду.

В середовищі розробників програмного забезпечення існує відомий ресурс для опанування основами різних мов програмування W3School [8]. Його популярність зумовлена тим, що він дозволяє покроково знайомитися з синтаксисом, специфікою, вбудованими засобами та додатковими бібліотеками різних мов, значно скорочуючи шлях навчання для осіб, які вже мають уявлення про теоретичні основи програмування, як-то розуміють принципи об’єктно-орієнтованого підходу, знайомі зі структурами даних, тощо, тому для них вивчення нової мови фактично зводиться до пошуку відповіді на питання, як записати засобами цієї мови те, що вони вже вміють іншою мовою. Одночасно, для новачків покрокове подання матеріалу разом із прикладами і навіть симулятором, де можна подивитись на результати виконання прикладу без розгортання середовища програмування, дає можливість вже через 1–2 години самостійної роботи отримати функціонуючий фрагмент коду, готового для подальших експериментів. W3School доступний англійською мовою, були спроби перекласти його на українську [9], але саме розділи про прогнозування досі не перекладено (на момент здійснення даної роботи, жовтень 2023).

В процесі розгляду публікацій за темою дослідження виникли певні питання до терміну “прогнозування” (англійською prediction), можливо через ось це трактування: “The term regression is used when you try to find the relationship between variables. In Machine Learning, and in statistical modeling, that relationship is used to predict the outcome of future events.” [10]. В перекладі українською: “Термін регресія використовується коли ви намагаєтесь знайти зв’язок між змінними. В машинному навчанні і в статистичному моделювання такий зв’язок використовується для прогнозування результатів майбутніх подій.” Тлумачний словник англійської мови дає таке визначення слову predict: “say or estimate that (a specified thing) will happen in the future or will be a consequence of something”. В перекладі українською: “говорити чи оцінювати, що (певна річ) трапиться в майбутньому, або стане наслідком чогось”. Походження цього слова в англійській мові, вважається, відбулось в ранньому 17-му столітті від латинського praedict – “зробити відомим заздалегідь, оголосити”, від дієслова praedicere, від prae- “перед” + dicere “говорити”. Українське слово “прогноз”, так як і англійське “prognosis”, походить від грецького prognōsis, від pro- “ перед” + gignōskein “знати”. На побутовому рівні немає сумнівів, що слова “прогноз”, “prediction” мають явне відношення до визначення невідомого, яке лежить саме в майбутньому, того, що ще не сталося. Однак, далі в даній роботі з’ясовується, що питання термінології є більш складним, та інколи призводить до хибних трактувань.

В даній роботі використано матеріали із розділу Machine Learning ресурсу W3 School, а також матеріали детального опису бібліотеки Skforecast [11]. Для виконання коду Python обрано ресурси хмарного середовища AWS Amazon, а саме, розгорнуто найменший доступний EC2 сервер із операційною системою Linux, здійснено оновлення Python до новішої доступної версії, встановлено додаткові бібліотеки і пакети. Доступ до серверу здійснюється в терміналі, через канал ssh, для роботи з файлами, включаючи редагування, використовувався Midnight Commander [12]. Такий набір інструментів обумовлений можливістю швидкого розгортання. Детальні покрокові інструкції наведено в документації AWS Amazon [13-14]. Варто зауважити, що нові користувачі, які вперше зареєструвалися на AWS Amazon, мають право на безкоштовний набір певних мінімальних сервісів протягом року, відомий як Free Tier, а при створенні EC2 серверу є позначки, які саме типи підпадають під умови безкоштовного користування в рамках Free Tier. Повний код Python для кожного прикладу доступний на інтернет ресурсі автора, далі в тексті вказуються назви файлів, які можна завантажити із зазначеного ресурсу [15].

РЕЗУЛЬТАТИ ДОСЛІДЖЕННЯ

Торговельні мережі застосовують особливі підходи, що спонукають відвідувачів до так званих “імпульсних” покупок, тобто не завжди покупці придбають лише необхідне або заздалегідь визначене за списком покупок, з іншого боку, покупці з довгим списком, вочевидь, проводять більше часу в магазині. Існує певна залежність між часом, проведеним в торговельному залі, і сумою, яка витрачається на покупки. Для аналізу даної залежності використовується поняття “середня сума” чеку за деякий часовий інтервал. В якості вхідних даних для

 https://doi.org/10.31474/2786-9024/v1i1(33).296156

прогнозування обрано масив таких усереднених значень (Таблиця 1). В даному випадку це довільно обраний масив чисел, що зростають. В реальності такі значення отримуються через спостереження за покупцями та аналізом інформації щодо сум їхніх покупок. Для прикладів даного дослідження достовірність цих даних не є важливою, важливим є лише направлення зміни (зростання з часом).

В процесі дослідження було проведено порівняння моделей на основі лінійної та поліноміальної регресій. Для поліноміальної регресії використано функцію linregress із бібліотеки scipy (Рисунок 1). Повний код наведено в файлі `linear.py` включно із прикладом побудування графіків. Слово "регресія" означає "спрощення", тобто вся сукупність вхідних даних "спрощується" до певної простої залежності. Сутність лінійної регресії зводиться до автоматичного пошуку закономірності, яка може бути описана лінійним рівнянням, а в графічному представленні має вигляд прямої.

В процесі дослідження було проведено порівняння моделей на основі лінійної та поліноміальної регресій. Для поліноміальної регресії використано функцію linregress із бібліотеки scipy (Рис. 1). Повний код наведено в файлі `linear.py` включно із прикладом побудування графіків. Слово "регресія" означає "спрощення", тобто вся сукупність вхідних даних "спрощується" до певної простої залежності. Сутність лінійної регресії зводиться до автоматичного пошуку закономірності, яка може бути описана лінійним рівнянням, а в графічному представленні має вигляд прямої.

```
import matplotlib.pyplot as plt
from scipy import stats
time = [5,10,15,20,25,30,35,40,45,50,55,60]
money = [5,7,25,42,88,91,103,150,152,190,195,200]
slope, intercept, r, p, std_err = stats.linregress(time, money)
def myfunc(time):
    return slope * time + intercept

model = list(map(myfunc, time))
print("the coefficient of correlation = "+ str(r))
```

Рис. 1. Лінійна регресія

Для наведених вище даних отримано коефіцієнт кореляції 0.98, що свідчить про адекватну відповідність залежності, але судячи з графіку на Рисунку 2, пряма лінія не охоплює хвилини 25, 35, 40, 50, 60, для їхнього включення потрібна крива, тому далі розглядається поліноміальна регресія і перевіряється, чи вона краще охоплює всі задані точки. Поліноміальна регресія – це пошук більш складної закономірності, яка описується рівнянням із змінною, що підносяться до другого, третього або більших ступенів, а в графічному представленні схожа на дугу.

Таблиця 1. Залежність суми чеку від часу

Час (хвилини)	5	10	15	20	25	30	35	40	45	50	55	60
Сума ($)	5	7	25	42	88	91	103	150	152	190	195	200

Для наведених вище даних отримано коефіцієнт кореляції 0.98, що свідчить про адекватну відповідність залежності, але судячи з графіку на Рис. 2, пряма лінія не охоплює хвилини 25, 35, 40, 50, 60, для їхнього включення потрібна крива, тому далі розглядається поліноміальна регресія і перевіряється, чи вона краще охоплює всі задані точки. Поліноміальна регресія – це пошук більш складної закономірності, яка описується рівнянням із змінною, що підносяться до другого, третього або більших ступенів, а в графічному представленні схожа на дугу.

Повний код побудови моделі поліноміальної регресії із застосуванням функцій `poly1d` і `polyfit` з бібліотеки `numpy` та функції `r2_score` для визначення коефіцієнту детермінації з бібліотеки `sklearn` міститься в файлі `polynom.py`.

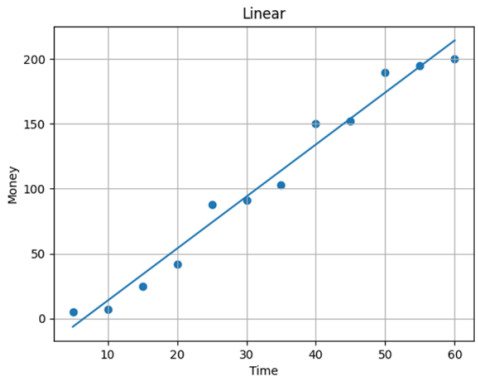

Рис. 2. Графік лінійної регресії

Спочатку здійснюється тренування моделі на відомих значеннях часу і сум чеків, далі – прогнозування через передачу до моделі значень часу, для яких сума чеку не є відомою. Виконання дає коефіцієнт детермінації рівний

ISSN 2786-9024

 https://doi.org/10.31474/2786-9024/v1i1(33).296156

0.98, що вказує на високу надійність результатів прогнозу. Спочатку модель використовується для заповнення пропущених значень часу (7, 12, 18, 33, 51 хвилин). Результати прогнозу графічно представлені на Рис. 3, блакитні точки відповідають відомим значенням, а помаранчеві показують результати прогнозу, крива лінія є графіком поліноміальної моделі.

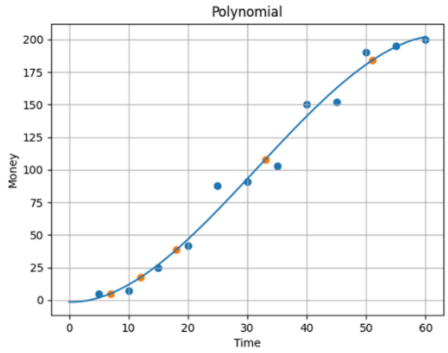

Рис. 3. Графік поліноміальної регресії і прогнозу в межах заданого діапазону

Прогнозування за межами початкового часового діапазону виконується тим самим кодом моделі в файлі polynom.py, але з підставленням нових значень часу (65, 70, 75, 90, 120 хвилин). Результати прогнозування усереднених витрат відповідно до даних часу, що лежать за межами вхідної інформації, наведено на Рис. 4.

Якщо “близькі” значення 65 і 75 хвилин дають якісь, можливо, адекватні результати, то зі збільшенням часового інтервалу, відходячи все далі від межі початкового діапазону спостерігається стрімке падіння в напрямку від’ємних величин.

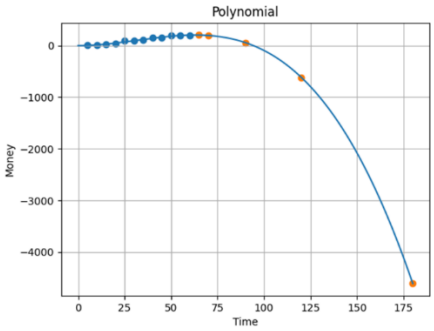

Рис. 4. Графік поліноміальної регресії і прогнозу за межами заданого діапазону

Таке прогнозування не має ніякого сенсу, тож висновок – поліноміальна регресія дозволяє більш або менш акуратно доповнювати пропущені дані в межах заданого діапазону, а також є корисною для графічного представлення, або побудови кривої саме шляхом доповнення точок на графіку, для яких не існувало вхідних даних. Не складно уявити, що при застосуванні лінійної регресії, нові прогнозовані точки лежатимуть на продовжені прямої, але в рамках прикладу даного дослідження їм так само бракуватиме сенсу, адже немає сумнівів, що кількість грошей в кишенях покупців є величиною кінцевою і навіть якщо вони проведуть цілий день в магазині, то не можуть витратити більше, ніж мають. В якості експерименту було здійснено виконання прогнозу на тій самій моделі, але із збільшенням обсягу початкових даних (Рис. 5).

Збільшення кількості елементів в масивах вхідних значень не призвело до покращення екстраполяції за межами вхідного часового діапазону, принаймні в рамках розгляду залежності суми чеку від часу, проведеному в супермаркеті. Зі збільшенням відходження від меж заданого для тренування моделі діапазону адекватність прогнозування втрачається, що є очікуваним для прогнозування залежності між обмеженими величинами, в даному випадку це гроші у кожного покупця і час, який вони можуть провести в магазині.

При розгляді залежності середньої суми чеку від місяця протягом кількох років, йдеться про процес з певними коливаннями, а прогноз має давати уявлення, по-перше, про зміну витрат відповідно до циклічних подій, таких як свята, сезон літніх відпусток, початок навчального року, і по-друге, про щорічне зростання цін, викликане інфляцією.

```
time =  [5, 5,  6,  8,  9,  9,  10, 11, 13,  14,  14, 15, 15,  16, 17, 19,   20,
 21, 22, 25, 25, 25, 26, 28, 30, 32, 34,  35,  37,  38,  39,  40, 40,  41,  43,
  45, 49,  50, 52,  55,  57,  57,  60]
money = [5, 5.3,5.8,6.4,6.2,6.9, 7, 6.8, 8.5,19.8,17, 25, 28.5,24, 39, 32.5, 42,
 49, 65, 99, 79, 88, 90, 103,91, 99, 100, 103, 140, 135, 160, 150,145, 155, 160,
  152,160, 190,195, 198, 206, 210, 202 ]

model = numpy.poly1d(numpy.polyfit(time, money, 3))
time_line = numpy.linspace(0, 180, 180)
plt.scatter(time,money)
plt.plot(time_line, model(time_line))

print ("Filling the gaps:")
predicted_time = [7,12,18,33,51]
predicted_money =[]
for t in predicted_time:
    predicted_money.append(model (t))
plt.scatter(predicted_time, predicted_money)

print ("Prediction of unknown:")
predicted_time2 = [65,70,90,120,180]
predicted_money2 =[]
for t in predicted_time2:
    predicted_money2.append(model (t))
plt.scatter(predicted_time2, predicted_money2)
```

Рис. 5. Поліноміальна регресія

Для перевірки цього припущення був підготовлений масив даних, в якому кожний елемент включає дату та середню суму чеку, всього 6 значень на місяць для 4 послідовних років. Повний набір міститься у файлі `cyclic_data.csv`. Була розроблена модель із застосуванням бібліотеки `sklearn` для прогнозування через алгоритм випадкового лісу (Рис. 6). Повний код наведено у файлі `predict.py`. Без глибокого занурення в цей алгоритм слід пояснити принцип його дії: для класифікації (в даному випадку, для визначення сезонних

 https://doi.org/10.31474/2786-9024/v1i1(33).296156

перепадів) будуються логічні дерева прийняття рішень, кожне розгалуження – це певна умова (в даному випадку, чи закінчується тренд на спадання чи зростання).

```
regressor = RandomForestRegressor(max_depth=5, n_estimators=30, random_state=123)
# max depth of each tree, number of trees, random seed
forecaster = ForecasterAutoreg(regressor = regressor,lags = 16)
#lags - observations
forecaster.fit(y=data['money'])
predictions = forecaster.predict(12) #months
```

Рис. 6. Регресія “випадкового лісу”

Необхідність експериментального визначення параметрів максимальної глибини дерева (кількість рівнів розгалужень), кількості дерев, а також кількості спостережень (обчислень) для отримання найкращих результатів прогнозування варта окремого зауваження; в ході дослідження було здійснено близько 50 експериментів для отримання кінцевих значень, які привели до найкращих результатів прогнозу при найменшому часі виконання коду. Також важливо звернути увагу на існування дещо складних для початківців методів перевірки ефективності моделі, таких як зворотне тестування, проте метою даної роботи є найпростіша реалізація прогнозування, тому навіть фрагмент коду для тестування моделі після її тренування було виключено.

На Рис. 7 надається графічне представлення

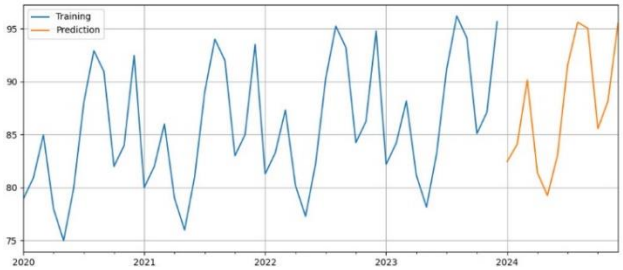

Рис. 7. Графік циклічних подій

відомих даних для 4 послідовних років, які було використано для тренування моделі (блакитним), і прогнозу на 5-й рік (помаранчевим).

Визначено певні сезонні зростання і падіння – найвищі щорічні середні значення припадають на середину літа, що відповідає періоду відпусток, а також на період перед різдвяними та новорічними святами. Також очевидне щорічне зростання всіх значень, так найвища сума для 2020 року була в районі 93, в 2021 наблизилася до 95, в 2022 дещо перевищила 95, а в 2023 значно перевищила 95. Аналогічні щорічні зростання спостерігаються і на інших екстремумах, що вказує на постійне зростання цін, ймовірно, внаслідок інфляції. Очікувано, прогноз на 2024 рік зберіг зразок, в результаті тренування із застосуванням даних попередніх років модель вірно визначила сезонні коливання і відповідно показала також зростання на більшості екстремумів окрім лише двох “піків”, які виявилися дещо нижчими за такі самі місяці попереднього року. В принципі, результати прогнозу можна вважати задовільними, так як модель визначила загальні тренди щорічного зростання і сезонних коливань. Безсумнівно, феномени інфляції і росту цін є значно складнішими, простий пошук закономірності через порівняння історичних величин не враховує повною мірою широкий спектр різних чинників, тому запропонований приклад слід розглядати умовно, в якості демонстрації застосування методу випадкового лісу для прогнозування циклічних явищ.

ОБГОВОРЕННЯ РЕЗУЛЬТАТІВ

Обговорення результатів мали місце під час занять з курсів “Моделювання та оптимізація систем управління” та “Інтелектуальні технології керування” програми підготовки магістрів групи СУАм-23 Донецького національного технічного університету.

ВИСНОВКИ

1. Проведено аналіз джерел, запропонований в них матеріал адаптовано під вимоги даної роботи.
2. Розроблено моделі прогнозування середньої суми чеку в залежності від проведеного в магазині часу, які демонструють, що лінійна і поліноміальна регресія, навіть за умови збільшення обсягу вхідних даних, не дають адекватного прогнозу поза межами заданого діапазону, проте такі регресії допомагають відновити пропущені значення в його межах, а також є корисними для візуалізації даних.
3. Розроблено модель для прогнозування залежності середньої суми чеку від сезонних коливань в умовах зростання цін на основі даних попередніх періодів із застосуванням методу випадкового лісу.
4. Всі приклади виконання прогнозування супроводжуються поясненнями та графіками.
5. Зазначені вище приклади коду для відносно простих задач з прогнозування на мові Python із застосуванням бібліотек `numpy` і `sklearn` не є складним навіть для початківців, і тому наведені приклади можуть бути в нагоді для практичного використання із внесенням мінімальних змін. Також надано рекомендації щодо доступного хмарного середовища виконання коду.

ЛІТЕРАТУРА

[1] М.П. Галущак., О.Я. Галущак, Т.І. Кужда, “Прогнозування соціально-економічних процесів: навчальний посібник для економічних спеціальностей”.

ISSN 2786-9024

Тернопіль, 2021. [Онлайн]. URL: http://surl.li/rlbjn. Дата звернення: 10.11.2023.
[2] О. А. Жуковська, "Прогнозування соціально-економічних процесів: комп'ютерний практикум: навч. посіб. для здобувачів ступеня бакалавра за освітньою програмою "Економічна кібернетика" спеціальності 051 Економіка". Київ: КПІ ім. Ігоря Сікорського, 2022. [Онлайн]. URL: http://surl.li/rlblj. Дата звернення: 10.11.2023.
[3] О.М.Вільчинська, "Силабус з навчальної дисципліни "Основи програмування Python". Львів: Львівський національний університет ім. Івана Франка, економічний факультет, кафедра статистики, 2022. [Онлайн]. URL: http://surl.li/rlbne. Дата звернення: 10.11.2023.
[4] M.Bowles, "Machine learning in Python: essential techniques for predictive analysis". John Wiley & Sons, 2015.
[5] M.Peixeiro, "Time series forecasting in python". Simon and Schuster, 2022.
[6] J.Yoon, "Forecasting of Real GDP Growth Using Machine Learning Models: Gradient Boosting and Random Forest Approach", *Comput Econ*, vol. 57, 247–265, 2021, doi: 10.1007/s10614-020-10054-w.
[7] Python License, Version 2. [Онлайн]. URL: http://surl.li/rlbpi. Дата звернення: 10.11.2023.
[8] W3School. [Онлайн]. URL: https://ww.w3schools.com. Дата звернення: 10.11.2023.
[9] W3School Українською. [Онлайн]. URL: http://surl.li/rlbqo. Дата звернення: 10.11.2023.
[10] W3School. Machine Learning - Linear Regression. [Онлайн]. URL: http://surl.li/rlbrd. Дата звернення: 10.11.2023.
[11] A.R.Joaquín, J.E.O.Rodrigo. Skforecast: time series forecasting with Python and Scikit-learn. [Онлайн]. URL: http://surl.li/rlbrz. Дата звернення: 10.11.2023.
[12] Midnight Commander. [Онлайн]. URL: http://surl.li/rlbuy. Дата звернення: 10.11.2023.
[13] Set up to use Amazon EC2. [Онлайн]. URL: http://surl.li/rlbvn. Дата звернення: 10.11.2023.
[14] Install Python, pip, and the EB CLI on Linux. [Онлайн]. URL: http://surl.li/rlbwp. Дата звернення: 10.11.2023.
[15] Я.Л. Байтельман. Код Python і приклади даних для прогнозування. [Онлайн]. URL: http://surl.li/rlbxq. Дата звернення: 10.11.2023.

USE OF MACHINE LEARNING TOOLS AND PYTHON LIBRARIES FOR PREDICTION THROUGH REGRESSIONS

Yakiv (Jacob) Baytelman, Hanna Telychko, Daria Zhukovska

The purpose of this work is forecasting the average check amount depending on the time spent by customers in the retail area and forecasting the average check amount based on seasonal demand characteristics considering the constant annual price growth.

Forecast methods were analysed. The use of Python libraries was considered. The code was developed for solving the above-mentioned forecasting tasks through linear, polynomial and random forest regressions, specifically, the search for the dependence of the average check amount on the time spent by customers in the retail area and the dependence of the average check amount on seasonal price fluctuations over several years. A comparison of the results of linear and polynomial regression was made under conditions of smaller or larger volumes of the input data, explanations were provided regarding the appropriateness of choosing one or the other when modelling non-cyclical processes and why the random forest model provides adequate results for cyclical processes. The source code and examples of input data were shared for public access.

Scientific novelty lies in the ways of applying regressions for modelling and forecasting the average check amount depending on the time customers spend in the retail area as well as the average check amount depending on seasonal demand characteristics with consideration of constant price growth.

The selected examples of cyclical and non-cyclical economic phenomena are the most common ones, and practitioners face similar tasks daily. Therefore, the results of this work provide them with ready-made solutions that require minimal adaptation to practical needs. Additionally, it demonstrates the feasibility of deploying them in the AWS cloud environment and the potential for integration with various data sources and other information systems.

***Keywords**: machine learning, regression, prediction, polynomial regression, random forest, python.*

REFERENCES

[1] M.P. Halushchak, O.Ia. Halushchak, T.I. Kuzhda, "Prohnozuvannia sotsialno-ekonomichnykh protsesiv: navchalnyi posibnyk dlia ekonomichnykh spetsialnostei". Ternopil, 2021. [Online]. URL: http://surl.li/rlbjn. Accessed: 10.11.2023. (In Ukrainian).
[2] O. A. Zhukovska, "Prohnozuvannia sotsialno-ekonomichnykh protsesiv: kompiuternyi praktykum: navch. posib. dlia zdobuvachiv stupenia bakalavra za osvitnoiu prohramoiu "Ekonomichna kibernetyka" spetsialnosti 051 Ekonomika". Kyiv: KPI im. Ihoria Sikorskoho. Kyiv, 2022. [Online]. URL: http://surl.li/rlblj. Accessed: 10.11.2023. (In Ukrainian).
[3] O.M.Vylchinska, "Sylabus z navchalnoi dystsypliny "Osnovy prohramuvannia Python". Lviv: Lviv national university im. Ivana Franka, department of economics and statistics, 2022. [Online]. URL: http://surl.li/rlbne. Accessed: 10.11.2023. (In Ukrainian).
[4] M.Bowles, "Machine learning in Python: essential techniques for predictive analysis". John Wiley & Sons, 2015.

[5] M.Peixeiro, “Time series forecasting in python”. Simon and Schuster, 2022.
[6] J.Yoon, “Forecasting of Real GDP Growth Using Machine Learning Models: Gradient Boosting and Random Forest Approach”, Comput Econ, vol. 57, 247–265, 2021, doi: 10.1007/s10614-020-10054-w.
[7] Python License, Version 2. [Online]. URL: http://surl.li/rlbpi. Accessed: 10.11.2023.
[8] W3School. [Online]. URL: https://ww.w3schools.com. Accessed: 10.11.2023.
[9] W3School In Ukrainian. [Online]. URL: http://surl.li/rlbqo. Accessed: 10.11.2023. (In Ukranian).
[10] W3School. Machine Learning - Linear Regression.. [Online]. URL: http://surl.li/rlbrd. Accessed: 10.11.2023.
[11] A.R.Joaquín, J.E.O.Rodrigo. Skforecast: time series forecasting with Python and Scikit-learn. [Online]. URL: http://surl.li/rlbrz. Accessed: 10.11.2023.
[12] Midnight Commander.. [Online]. URL: http://surl.li/rlbuy. Accessed: 10.11.2023.
[13] Set up to use Amazon EC2. [Online]. URL: http://surl.li/rlbvn. Accessed: 10.11.2023.
[14] Install Python, pip, and the EB CLI on Linux. [Online]. http://surl.li/rlbwp. Accessed: 10.11.2023.
[15] J.L.Baytelman. Kod Python i pryklady danyh dlia prognozuvanyia. [Online]. URL: http://surl.li/rlbxq. Accessed: 10.11.2023. (In Ukranian).

ISSN 2786-9024