Актуальні напрями розвитку технічного та виробничого потенціалу національної економіки: монографія/ за ред. В.О. Пінчук, Г.С. Прокудіна – Дніпро: Пороги.- 2021.- 536 с.

АКТУАЛЬНІ ПРОБЛЕМИ ВИСОКОЕФЕКТИВНОЇ ОБРОБКИ ДАНИХ. МОДЕЛЮВАННЯ ПОКАЗНИКІВ ЗА ДОПОМОГОЮ МОВИ ПРОГРАМУВАННЯ PYTHON

Аспекти і проблеми аналізу даних

В деякому сенсі наука про дані – синонім таких термінів, як бізнес-аналітика, дослідження операцій, аналіз, моделювання даних, прогнозування результатів. Дані вивчають для більш ефективного використання технічного та виробничого потенціалу. І великі корпорації, і малий бізнес користуються послугами аналітиків.

Аналітика асоціюється сьогодні з технологіями data mining, такими як штучний інтелект, нейронні мережі, кластерний аналіз, факторний аналіз, визначення викидів. Апаратна підтримка, яка здешевшала останнім часом, зробила можливим аналіз з великою кількістю змінних фінансово-економічних даних. З даними відбувається трансформація методами математики і статистики за допомогою сучасних комп'ютерних технологій в робочі аналітичні висновки, рішення, продукти.

До основних технік аналізу відносяться оптимізація, моделювання і прогнозування, кластерний та факторний аналіз, визначення викидів, штучний інтелект, сітьові графи, машинне навчання. Деякі з цих методів розвивалися давно, деякі з'явилися недавно. Але вік методів не має ніякого стосунку до складності та корисності. При правильному виборі різні техніки і технології є однаково ефективними для розвитку потенціалу підприємств. Тому важливим є розуміння, яка техніка для розв'язання якої проблеми підходить, як ці техніки працюють і як їх застосовувати для моделювання різних показників.

В сучасному світі компанії і підприємства мають справу з великим обсягом даних. Останнім часом все більшої популярності набувають технології, що дозволяють працювати з

великими обсягами інформації. Великі дані - один з найважливіших технологічних трендів, які кардинально змінюють можливості використання інформації в бізнесі. Великі обсяги даних існували і раніше, але до теперішнього часу вони значно збільшилися, окрім того, зросла різноманітність видів інформації. Технології великих даних дозволяють організаціям зберігати величезні обсяги даних, управляти ними і обробляти їх так, щоб своєчасно отримувати інформацію, необхідну для прийняття рішень.

Технології великих даних - це комбінація кращих напрацювань минулого і актуальних тенденцій, яка дозволяє отримувати потрібну інформацію з обсягу даних незалежно від того, генеруються такі дані людьми, пристроями або мережею. Підхід до обробки даних залежить, насамперед, від типу даних. Обсяг, швидкість, варіативність, достовірність – основні ознаки великих даних. Ці ознаки не завжди виражені в великих даних в рівній мірі. Є передані дані (також зустрічається термін «дані в русі») і збережені дані («дані в спокої»). Є дані структуровані і неструктуровані.

Ще кілька років тому збір або зберігання величезного обсягу даних було занадто дорого або обтяжливо. Навіть в тому випадку, коли вдавалося збирати такі дані, компанії часто не мали інструментів і фахівців для подальшої роботи з ними. Лише деякі програмні рішення були здатні обробляти великі обсяги інформації, при цьому їх використання було складним і витратним. Часто компанії, які все ж наважувалися аналізувати подібні дані, обмежувалися лише зрізами даних. Тобто частина даних, які не потрапляли в зріз, залишалися непоміченими. Труднощі з використанням даних виникають внаслідок недостатніх обчислювальних потужностей для роботи зі складними моделями, обробкою зображень і т. ін. Окрім того, можливості підприємства чи установи в організації збору, систематизації, аналізу даних є часто обмеженими. Наразі невелика кількість підприємств і установ мають змогу інвестувати великі кошти в розвиток цих технологій для власних потреб. Великі компанії, безумовно, мають можливості для

розвитку бізнес-аналітики. Наприклад, компанія Київстар щорічно проводить школи BigData, в які відбирає для навчання талановиту молодь. Такі великі компанії мають можливість використовувати сервіси і платформи всесвітньо- відомих розробників-гігантів.

Сервіси великих даних для організації сховищ пропонують відомі компанії: Amazon (Amazon Elastic MapReduce, Amazon DynamoDB, Amazon Simple Storage Service), Google (Google Compute Engine, Google Big Query, Google Prediction API), Microsoft (Microsoft Azure, Windows Azure HDinsight), Yahoo! (Hadoop). Відомими платформами для розробки моделі програмування і її реалізації є MapReduce і Hadoop, що дозволяють розділяти великі завдання на невеликі елементи, які можна обробляти паралельно і представити об'єднаний результат такої обробки.

Найчастіше компанії використовують структуровані дані, які розміщені в реляційних базах даних. Наразі популярними є реляційні бази даних MySQL, PostgresSQL, NoSQL. Окрім цього розвиваються і нереляційні бази даних.

Для більшості технік роботи з даними обсяг даних може бути і великим, і малим. Наскільки великі дані потрібні для результату, визначається інтересами компанії. Не завжди компанії мають справу із генерованими даними великих обсягів. Найчастіше є певна база даних компанії, яку і потрібно використовувати.

В різних технологіях для вивчення даних використовуються схожі базові математичні інструменти: і у стандартних пакетах обробки даних та відповідних бібліотеках і модулях у популярній об’єктно-орієнтованій мові Python, і в широко відомому і розповсюдженому пакеті MS Office, і в інших програмних продуктах на кшталт мови R, що є поширеною для статистичного аналізу даних.

Для використання автоматизації обробки даних в програмному режимі потрібні знання тієї чи іншої мови програмування. Але для розуміння сутності аналізу, що використовується в технологіях обробки і різних прикладних

пакетах таких, як Statistika, SPSS і т. ін., не обов'язково знати, як пишеться код. Ці потужні інструменти включають різноманітний аналіз, в тому числі, регресійний, факторний, кластерний, побудову моделей за допомогою нейронних мереж і багато іншого, а також дають можливість отримати графічне відображення результатів, якщо дозволяє розмірність і постановка задачі.

У деяких задачах достатньо використання зручного, зрозумілого і доступного інструменту, як MS Excel, що має широкі можливості і пакет аналізу, хоча і дещо обмежений, але придатний для отримання результатів у першому наближенні для уявлення про характер даних. За допомогою електронних таблиць неможливо в програмному режимі запустити виробничу модель штучного інтелекту, але за їх допомогою можна проаналізувати характер даних, змоделювати і спрогнозувати результат. Цей результат можна отримати на основі класичних підходів теорії ймовірностей та математичної статистики щодо нормування даних, кореляційного та регресійного аналізу, оцінювання прогнозних точкових та інтервальних значень, а також за допомогою процедур для визначення оптимальних розв'язків лінійних та нелінійних задач оптимізації.

Але є спільні проблеми при використанні різних технологічних інструментів – відбір і підготовка даних, а також фахове опрацювання результатів.

Щоб скористатися навіть абсолютно автоматизованою системою обробки даних, їх потрібно правильно відібрати, відсортувати, нормалізувати, обрати метод аналізу, і після обробки даних програмою провести саме аналіз, інтерпретацію, прогнозування і т. ін.

На етапі підготовки даних виникають певні проблеми, пов'язані з різними форматами даних, отриманих із різних джерел, з обмеженим доступом до даних, обумовленим цінністю даних та конфіденціальністю і суворою регламентованістю. Дані можуть мати різні одиниці виміру, різні рівні агрегування.

Покращення якості даних, розуміння, як дані взаємодіють між собою, оцінювання розподілів і приведення до певного

формату неможливе без знання фундаментальних основ відповідного математичного апарату.

Сучасні пакети підтримують багато популярних методів моделювання і оцінювання моделей. Застосування потребує вивчення мови і освоєння пакетів, в яких для використання певного методу чи функції потрібно задати багато параметрів для виконання. Цей вибір параметрів має бути осмисленим. Інакше результати моделювання будуть такими, що їх не можна пояснити, інтерпретувати і застосувати для практичного використання. Головне призначення моделювання – не стільки в описанні наявних даних, а прогнозуванні фінансово-економічних показників, що дозволяє покращити використання технічного та виробничого потенціалу.

Моделювання має сенс у тому разі, якщо результати є адекватними. І у цьому питанні основну роль у тлумаченні результатів відіграють знову ж таки базові поняття. Можна запустити на виконання безліч разів програму, змінюючи велику кількість параметрів. Без глибокого розуміння процесів моделювання, що при цьому відбуваються, можна погіршити результат. І при цьому все списати на таке поняття, як «перенавчання моделі». Якщо вибір параметрів є цілеспрямованим, модель буде покращувати якості, а не погіршувати їх. Критерії, що використовують сучасні аналітичні інструменти, зводяться до базових понять математичної статистики.

Тож, основою сучасних технологій обробки даних, є глибоке засвоєння базового математичного апарату. Процес моделювання залежить від якості даних та від професіоналізму аналітика.

Інструменти для ефективної та швидкої обробки даних

У процесі роботи з даними можна виділити декілька етапів.

1. Призначення цілі дослідження – готується проектне завдання і оцінюються цілі дослідження і вартість роботи.

2. Збір і підготовка даних – «розвідувальний аналіз». Складнощі виникають уже на цьому етапі – збір даних. Ці дані часто розрізнені, в різних форматах і потребують багато зусиль від дослідника, пов'язаних з нормалізацією, однорідністю. Часто матриці є не повністю заповненими, розрідженими, і це означає, що потрібно підібрати алгоритм для заповнення порожнеч. Окрім того, в даних бувають значні відхилення, тобто, викиди, які потрібно усунути для отримання адекватних результатів – так звана , очистка даних. Таким чином, процес підготовки даних є дуже кропітким і рутинним, майже «ручним», і потребує інтелектуального підходу.

3. Дослідницький аналіз і моделювання даних, оцінювання параметрів моделі – «тренування моделі». На цьому етапі потрібно покращити якість даних, зрозуміти, як дані взаємодіють між собою, оцінити розподіли даних і визначити наявність викидів. Для цього використовуються статистики, що описують та візуалізують методи і просте моделювання. Постають питання: чи пов'язані між собою досліджувані фактори і показник, чи є мультиколінеарність в системі даних, чи можна зменшити кількість змінних і тим самим спростити модель, яку форму залежності обрати для моделювання, яким способом звести модель до лінійної форми і т. ін. На цьому етапі потрібні знання предметної області, а також математики, теорії ймовірностей і математичної статистики. Тільки після указаних досліджень і перетворень даних можна скористатися бібліотекою. При цьому потрібно розуміти, які параметри потрібно задати для того чи іншого обраного методу. Сам процес моделювання має назву «тренування моделі» і означає побудову різних моделей на одному наборі даних, випадково відібраних із загальної сукупності в певній кількості, яку можна варіювати, вибір найкращої моделі за певними критеріями, наприклад, метод найменших квадратів, метод оснований на дереві рішень або метод абсолютних відхилень і т.п.

Можна тренувати набір даних декілька разів, змінюючи параметри, і таким чином, досягти найкращого результату. Тож, побудова моделі – ітераційний процес.

4. Перевірка адекватності моделі і значимості факторів, включених в модель. Після отримання найкращого результату (наприклад, порівнюються сума квадратів відхилень і обирається набір параметрів, що дає найменше із усіх) оцінювання якості моделі відбується за статистичними критеріями. Якщо якість незадовільна, відбувається «перенавчання моделі».

5. Застосування моделі до незнайомих даних (тренувальний сет) із тієї ж вибірки - «прогностичне моделювання».

Для задач моделювання і прогнозування використовують різні технології та інструменти.

Наразі популярним методом опрацювання даних є машинне навчання (Machine Learning) - набір алгоритмів виявлення закономірності в даних . Python має свою бібліотеку Scikit-learn з різноманітними алгоритмами.

Машинне навчання наразі є дуже популярною і перспективною технологією серед аналітиків (data-scientists). Ринок машинного навчання швидко зростає. З 2016 року його обсяг подолав позначку в $ 1 млрд, а до 2025 року, судячи з прогнозів, він може збільшитися до $ 39,98 млрд. 60% компаній в світі вже використовують машинне навчання.

Серед завдань, які можуть вирішуватися засобами машинного навчання, можна зазначити задачі моделювання і прогнозування показників в залежності від одного або декількох факторів або оптимізаційні задачі. Наприклад, обсягів попиту, продажів, наповнення складу, завантаження устаткування і інших ресурсів; виявлення тенденцій, прихованих взаємозв'язків, аномалій, повторюваних елементів і т.п.; класифікація та аналіз складу покупців, клієнтів, замовників і сегментація їх за різними параметрами; кластеризація, тобто класифікація за параметрами, які з самого початку не були відомі.

Використовуються як традиційні методи економетричного аналізу, включаючи однофакторні, багатофакторні моделі на основі методу найменших квадратів, так і нетрадиційні, типу,

дерева рішень з великою кількістю встановлюваних параметрів, що дають гнучкість моделювання параметрів моделей.

Розвиваються і так звані «нейронні мережі». При моделюванні використовуються поняття ризику, кількісні ознаки якого обчислюються у відповідності до числових характеристик дискретних та неперервних випадкових величин.

За останні десять років Python перетворився в одну із найважливіших мов програмування, застосовуваних у науці про дані, в машинному навчанні та розробці програмного забезпечення загального призначення в академічних установах і промисловості.

Порівняно недавно з'явилися поліпшені бібліотеки для Python, і він став серйозним конкурентом в рішенні задач створення додатків обробки даних.

У багатьох сучасних середовищах застосовується загальний набір успадкованих бібліотек, написаних на FORTRAN і C, що містять реалізації алгоритмів линійної алгебри, оптимізації, інтегрування та ін. Тому численні компанії використовують Python як «клей» для об'єднання написаних за багато років програм.

Пакети Python для роботи з даними

NumPy, скорочення від «Numerical Python», - основний пакет для виконання наукових розрахунків на Python. Поверх NumPy побудовано інші бібліотеки.

Основні можливості пакету:

- швидкий і ефективний об'єкт багатовимірного масиву ndarray;
- функції для виконання обчислень з елементами одного масиву або математичних операцій з декількома масивами;
- засоби для читання і запису на диски наборів даних, представлених у вигляді масивів;
- операції лінійної алгебри, перетворення Фур'є і генератор випадкових чисел;

• засоби для інтеграції з кодом, що написаний на C, C++ або Fortran.

Крім прискорення роботи з масивами, однією з основних цілей NumPy стосовно аналізу даних є організація контейнера для передачі даних поміж алгоритмами. Як засіб зберігання і маніпуляції даними масиви NumPy значно ефективніші за вбудовані в Python структури даних.

Таким чином, багато засобів обчислень, орієнтовані на Python, або використовують масиви NumPy в якості основної структури даних, або якимось іншим способом організують інтеграцію з NumPy.

Pandas надає структури даних і функції, що дозволяють зробити роботу зі структурованими даними простою і швидкою. Бібліотека сприяла перетворенню Python в потужне і продуктивне середовище аналізу даних. Основні об'єкти pandas - це *DataFrame* - двовимірна таблиця, в якій рядки і стовпці мають мітки, і *Series* - об'єкт одновимірного масиву з мітками.

У бібліотеці pandas поєднуються висока продуктивність засобів роботи з масивами, притаманна NumPy, і гнучкі можливості маніпулювання даними, властиві електронним таблицям і реляційним базам даних (наприклад, на основі SQL). Оскільки маніпулювання даними, їх підготовка і очищення грають дуже велику роль в аналізі даних, pandas є одним з основних інструментів.

Основні можливості бібліотеки:

• розвинені засоби індексування, що дозволяють просто змінювати форму наборів даних, формувати зрізи, виконувати агрегування і вибирати підмножини;

• структури даних з позначеними осями підтримують автоматичне або явне вирівнювання даних, що виключає появу типових помилок при роботі з невирівняні даними і даними з різних джерел, які по-різному індексовані;

• вбудована функціональність часових рядів;

• одні і ті ж структури даних підтримувати як тимчасові ряди, так і дані інших видів;

• арифметичні операції;

• гнучка обробка відсутніх даних;
• інтеграція даних;
• підтримка з'єднання і інших реляційних операцій, наявних в популярних базах даних (наприклад, на основі SQL).

Багато засобів, присутні в pandas, або є частиною мови R, або надаються додатковими пакетами.

Сама назва pandas утворена від *panel data* (панельні дані), що застосовуються в економетриці для позначення багатовимірних структурованих наборів даних, так і від фрази Python data analysis.

Matplotlib - найпопулярніший в Python інструмент для створення графіків і інших способів візуалізації *двовимірних* даних. Наразі супроводжується великою групою розробників. Вона підходить для створення графіків, придатних для публікації. Хоча програмістам на Python доступні і інші бібліотеки візуалізації, matplotlib використовується найчастіше і тому добре інтегрована з іншими частинами екосистеми.

SciPy - збір пакетів, призначених для вирішення різних стандартних обчислювальних задач. Деякі з них:
• scipy.integrate - підпрограми чисельного інтегрування і розв'язання диференціальних рівнянь;
• scipy.linalg - підпрограми лінійної алгебри і розкладання матриць, доповнюють ті, що включені в numpy.linalg;
• scipy.optimize - алгоритми оптимізації функцій (знаходження екстремумів) і пошуку коренів;
• scipy.signal - засоби обробки сигналів;
• scipy.sparse - алгоритми роботи з розрідженими матрицями і розв'язання розріджених систем лінійних рівнянь;
• scipy.special - обгортка навколо SPECFUN, написаної на Fortran-бібліотеці, що містить реалізації багатьох стандартних математичних функцій, в тому числі гамма-функції;
• scipy.stats - стандартні безперервні і дискретні розподіли ймовірностей (функції щільності ймовірності, формування вибірки, функції безперервного розподілу ймовірності), різні статистичні критерії і додаткові описові статистики.

scikit-learn є основним інструментарієм машинного навчання програмістів на Python. У ньому є підмодулі для наступних моделей:

• класифікація: метод опорних векторів, метод найближчих сусідів, випадкові ліси, логістична регресія і т. ін.;
• регресія: Lasso, гребневая регресія і т. ін .;
• кластеризація: метод k середніх, спектральна кластеризація і т. ін.;
• зниження розмірності: метод головних компонент, відбір ознак, матрична факторизація і т. ін .;
• вибір моделі: пошук на сітці, перехресний контроль, метрики;
• попередня обробка: виділення ознак, нормування.

scikit-learn орієнтований головним чином на прогнозування і передбачення.

Разом з pandas, statsmodels бібліотека scikit-learn зіграла найважливішу роль для перетворення Python в продуктивну мову програмування для науки про дані.

Statsmodels - пакет статистичного аналізу.

У порівнянні із scikit-learn, пакет statsmodels містить алгоритми класичної (перш за все частотної) статистики та економетрики. У нього входять наступні підмодулі:

• регресійні моделі: лінійна регресія, узагальнені лінійні моделі, лінійні моделі зі змішаними ефектами і т. ін .;
• дисперсійний аналіз (ANOVA);
• аналіз часових рядів: AR, ARMA, ARIMA, VAR і інші моделі;
• непараметричні методи: ядерна оцінка щільності, ядерна регресія;
• візуалізація результатів статистичного моделювання.

Пакет statsmodels орієнтований більшою мірою на статистичне виведення, він дає оцінки невизначеності і p-значення параметрів.

Використовується разом з NumPy і pandas.

В Python є бібліотеки для зручного і швидкого зчитування даних в форматах електронних таблиць, баз даних, csv та ін.

Приклад використання Python для моделювання показника (лінійна регресія)

Для прикладу розглянемо моделювання даних за допомогою одно-факторної регресійної моделі.

За даними двох вибірок (показника і фактору) оцінимо наявність і тісноту зв'язку між ними, вигляд і тип моделі, параметри регресії, адекватність моделі, статистичну значимість параметрів, наявність автокореляції, визначимо прогнозне значення показника (точкову та інтервальну оцінки) і побудуємо довірчі зони регресії.

У якості інструмента моделювання використаємо Python та бібліотеки NumPy, Statsmodels, Matplotlib, Xlrd (для зчитування даних із файлу Excel).

Як зазначалося, найпоширенішою сучасною методикою моделювання структурованих даних є економетричне моделювання. За допомогою регресійного аналізу оцінюється залежність показника від одного чи декількох факторів.

Найкращий результат дає метод найменших квадратів відхилень вихідних даних від змодельованих. Адекватність моделі оцінимо за допомогою критерію Фішера. Оцінювання статистичної значимості параметрів регресій, а також довірчих інтервалів регресій, проведемо на основі критерію Стьюдента. Окрім цього, отримаємо іншу статистику по моделі і застосуємо модель для прогнозу.

Нижче наведено графік для наочного уявлення про побудовану модель і лістинг результату програмного виконання.

Проаналізуємо результат. Застосовано модель OLS, метод Least squares. Залежна змінна – y. Рівняння регресії виведено на графіку. Коефіцієнт кореляції дорівнює 0,9785, що свідчить про сильну кореляцію між фактором і показником. Коефіцієнт детермінації скорегований - 0,95: зміна показника обумовлена зміною фактору на 95,5%. F-статистика свідчить про адекватність моделі. Розраховані значення t- статистики 18.396 для нахилу та 9.127 для перетину регресії. Обидва параметри є статистично значимими з довірчою імовірністю 0.975. Довірчі інтервали параметрів регресії: для нахилу: (1.159; 1.463), для перетину :

(21700; 34900). Статистика Дарбіна–Уотсона свідчить про відсутність автокореляції в моделі. Коваріаційна матриця вірно специфіковано. Модель може бути використана для прогнозу показника. Визначено прогнозні оцінки (точкові та інтервальні). Коефіцієнт еластичності за середніми показниками – 0.609, що означає, що показник є нееластичним по фактору.

Лістинг виконання програми

```
Номер фактору: 1
Значення для прогнозу показника y: 120000
Дані показника y:  [ 35340.  45140.  53513.  61354.  41659.  65355.  78222.  77231.  88677.
  90922.  95130.  86811.  97991.  99536. 105327. 115619. 120000.]
Дані фактора x 1 [[12046. 15039. 17015. 20803. 17501. 23083. 36370. 38295. 40323. 43783.
  48836. 45415. 51196. 55240. 65371. 68158. 70200.]]
Результат розрахунку:
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.958
Model:                            OLS   Adj. R-squared:                  0.955
Method:                 Least Squares   F-statistic:                     338.4
Date:                Mon, 10 May 2021   Prob (F-statistic):           1.06e-11
Time:                        20:41:06   Log-Likelihood:                -169.43
No. Observations:                  17   AIC:                             342.9
Df Residuals:                      15   BIC:                             344.5
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1             1.3107      0.071     18.396      0.000       1.159       1.463
const       2.832e+04   3102.448      9.127      0.000    2.17e+04    3.49e+04
==============================================================================
Omnibus:                        1.110   Durbin-Watson:                   1.875
Prob(Omnibus):                  0.574   Jarque-Bera (JB):                0.993
Skew:                          -0.474   Prob(JB):                        0.609
Kurtosis:                       2.291   Cond. No.                     1.02e+05
==============================================================================
Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.02e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
Кореляція:
[[1.     0.9785]
 [0.9785 1.    ]]
Значення фактору для заданого прогнозу показника: [[67773.14371161]
 [77416.62667164]]
```

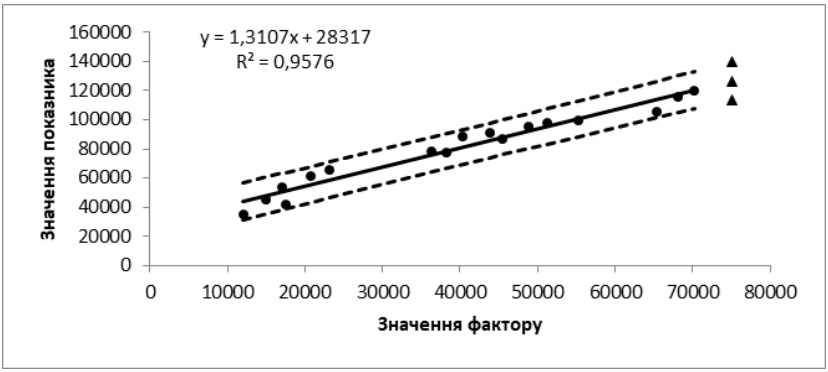

Рис.1. Лінійна регресія, точкова оцінка прогнозу показника, довірча зона регресії та довірчий інтервал прогнозу, побудовані з надійністю 0,95

Тож, для моделювання економічних показників і ефективної обробки даних існують різні інструменти і технології. Моделювання може бути успішним тільки за умови ретельної підготовки даних і вибору правильних технологічних інструментів і відповідних методів дослідження.

Процес моделювання залежить від якості даних та від професіоналізму аналітика.

Список використаних джерел

1. Маккини У. Python и анализ данных / пер. с анг. А. А. Слинкина. – М.: ДМК Пресс, 2020. – 540 с.: ил.

2. Силен Деви, Мейсман Арно, Али Мохамед Основы Data Science и BigData. Python и наука о данных. – СПб.:Питер, 2017. - 336 с.:ил. – (Серия «Библиотека программиста»)