Universitatea Tehnică a Moldovei

Folosirea instrumentelor Data Mining pentru verificarea ipotezelor ce ţin de comportamentul vânzărilor

Using Data Mining tools to verify the assumptions related to sales behavior

Student : **gr. TIA-201M, Moşanu Andrei**

Coordonator: **Cunev Veaceslav, lect. univ.**

Chișinău, 2022

MINISTERUL EDUCAŢIEI ȘI CERCETĂRII AL REPUBLICII MOLDOVA

Universitatea Tehnică a Moldovei
Facultatea Calculatoare, Informatică şi Microelectronică
Departamentul Ingineria Software și Automatică

Admis la susținere

Șef departament:

FIODOROV Ion dr., conf.univ.

_____

“___” _____ 2022

Folosirea instrumentelor Data Mining pentru verificarea ipotezelor ce ţin de comportamentul vânzărilor

Proiect de master

Student: **_____ Moșanu Andrei, TIA-201M**

Coordonator: **_____ Cunev Veaceslav, lect. univ.**

Consultant: **_____ Bodoga Cristina, asist.univ.**

Chişinău, 2022

REZUMAT

Titlul lucrării: Folosirea instrumentelor Data Mining pentru verificarea ipotezelor ce țin de comportamentul vânzărilor

Cuvinte-cheie: ipoteza nulă, p-valoare, distribuție normală, teste parametrice/nonparametrice, normalizarea datelor.

Scopul lucrării: Cercetarea metodelor și modelelor matematice care permit verificarea ipotezelor și obținerea unor rezultate care pot aduce plus valoare afacerii.

Problema necesară de rezolvat este volumul mare de date și valoarea intrinsecă încă neexploatată a lor. Există anumite procese logistice, manageriale, financiare și tehnice care formează, în ansamblu activitatea întreprinderii. Indicatorii-cheie care descriu aceste procese sunt prezenți în mod implicit și necesită analiză prin modele matematice pentru a putea răspunde afirmativ sau negativ la unele întrebări-ipoteze care definesc activitatea. Deci acesta și este scopul urmărit. Datele necesită prelucrare, selecție și comparare ulterioară a tendinței centrale (poate fi media sau mediana). Ipotezele lansate au fost separate în trei blocuri, astfel ca să descrie 3 tipuri separate de procese :

– indicatori de eficiență a livrării;

– rulajul, volumul de vânzări;

– zona și timpul de livrare.

Această sarcină a fost rezolvată cu succes prin limbajul pe programare Python și bibliotecile incluse din Stats, care conțin toate testele aplicate pe parcurs: Mann-Whiney U, T-Student relaționat și nerelaționat, Z-Test, Shapiro- Wilk, Fligner Killeen. Un instrument util la prezentarea grafică a fost biblioteca grafica Seaborn, bazată pe Matplotlib. Pe lângă acestea, este prezent un element de subiectivism care presupune luarea deciziilor de către statiscian referitor la normalitatea distribuției și referitor la nivel de semnificație ales α, care și reprezintă eroare de speța I. De acest parametru ales intuitiv de către experimentator va depinde dacă ipoteza va fi respinsă sau nu va putea fi respinsă. Dar o simplă diminuare a parametrului α nu este o soluție, deoarece în acest caz crește eroare de speța II, caracterizată de parametrul β. Singura soluție ar fi de ales un test cu o putere statistică mai mare, adică cu coeficinetul 1-β maximal posibil.

În rezultat au fost testate cu succes 15 ipoteze din 3 categorii și prezentate 5 concluzii cu impact financiar și logistic asupra proceselor interne a afacerii.

ABSTRACT

Title of the paper: Using Data Mining tools to verify the assumptions related to sales behavior
Keywords: null hypothesis, p-value, normal distribution, parametric/nonparametric tests, data normalization.
Purpose of the paper: to investigate mathematical methods and models that allow hypothesis testing and obtain results that can add value to the business.

The necessary problem to solve is the large volume of data and its as yet untapped intrinsic value. There are certain logistical, managerial, financial and technical processes that together make up the business. The key indicators describing these processes are present by default and require analysis by mathematical models to answer yes or no to some of the hypothesis-questions defining the activity. So this is also the aim. The data require further processing, selection and comparison of the central tendency (may be mean or median). The hypotheses launched have been separated into three blocks so as to describe 3 separate types of processes:

- delivery efficiency indicators;
- turnover, sales volume;
- area and delivery time.

This task was successfully solved using the Python programming language and the included libraries from Stats, which contain all the tests applied along the way: Mann-Whiney U, T-Student related and unrelated, Z-Test, Shapiro- Wilk, Fligner Killeen. A useful tool in the graphical presentation was the graphical library Seaborn, based on Matplotlib. In addition, there is an element of subjectivism that involves the statistician's decision making on the normality of the distribution and on the chosen significance level α, which also represents the first-species error. On this parameter intuitively chosen by the experimenter will depend whether the hypothesis will be rejected or not. But a simple decrease of the parameter α is not a solution, because in this case the error of type II, characterized by the parameter β, increases. The only solution would be to choose a test with a higher statistical power, i.e. with the maximum possible 1-β coefficient.

As a result, 15 hypotheses from 3 categories were successfully tested and 5 conclusions with financial and logistical impact on internal business processes were presented.

CUPRINS

Secolul XXI la sigur este secolul informației, rețelelor, tehnologii IoT și despre date care circulă prin aceste mari artere cibernetice. Volumul total de informație la nivel global este în continuă creștere exponențială. Se prognoza un volum de pînă la 5.2 TB de informație per cap de locuitor al planetei . Nu se mai vorbește de data dar de Big Data. Volumul de informație la nivel global a ajuns la 40 ZB (date din 2020). În acest context este necesar nu doar de sisteme hardware, care vor crea, gestiona, transmite și stoca eficient și în timp util aceste informații, dar și mijloace, instrumente destinate prelucrării, analizei, cercetării legităților care derivă din volumul masiv de date. Cu acest scop sunt aplicate metodele și modelele de Data Mining.

În următorii opt ani, cantitatea de date digitale produse va depăși 40 de zettabytes, ceea ce reprezintă echivalentul a 5.200 GB de date pentru fiecare bărbat, femeie și copil de pe Pământ, potrivit unui studiu actualizat Digital Universe publicat astăzi. Pentru a pune lucrurile în perspectivă, 40 de zettabytes înseamnă 40 de trilioane de gigabytes - estimat a fi de 57 de ori mai mare decât cantitatea tuturor grăunților de nisip de pe toate plajele de pe Pământ. Pentru a atinge această cifră, se așteaptă ca toate datele să se dubleze la fiecare doi ani până în 2020. Majoritatea datelor de acum până în 2020 nu vor fi produse de oameni, ci de mașini care vor vorbi între ele prin intermediul rețelelor de date. Aceasta va include, de exemplu, senzori de mașini și dispozitive inteligente care comunică cu alte dispozitive. Mediul de afaceri este printre fruntașii care beneficiază de aceste resurse și tehnologii. Cu acestea se poate verifica volumul de vânzări, metricile de eficiență, grupurile de clienți și preferințele fiecărui grup în parte, eficiența campaniilor de promovare sau schimbările care trebuie instituite pe un anumit segment. Toate aceste rezultate nu pot fi obținute decât numai datorită noilor tehnici de analiză a datelor.

Lucrarea de față are ca scop cercetarea unui domeniu microfinanciar și verificarea ipotezelor care le înaintează antreprenorul referitor la indicatorii de afacere. Se vor înainta câteva ipoteze referitor la volumul de vânzări, cecul mediu al cumpărăturii, condițiile și factorii externi apoi se vor analiza datele aplicând limbajul R, Python, mediu Excel Office pentru a trage concluzii dacă ipoteza poate fi respinsă sau nu poate fi respinsă. În așa mod, la final se pot trage concluzii majore vizavi de deciziile care trebuie luate și efectele lor asupra indicatorilor microeconomici.

În cadrul teoretic se vor prezenta reperele teoretice necesare studierii tipului datelor, distribuțiilor, criteriilor de normalitate aplicate și criterii de verificare a devierii indicatorului de la valoarea așteptată pentru ca ipoteza să fie sau să nu poată fi respinsă. În cadrul practic se vor aplica totalitatea acestor repere asupra modelului de date accesibil de sondat. S-a propus de a ajuta conducerea

întreprinderii să ia deciziile bune, argumentate analitic prin testarea unor ipoteze. Rezultatul lor va facilita un anumit scenariu pe viitor.

Obiectivele generale sunt: analiza sistemului de unde se colectează datele (sistemul țintă, subsistem, sistemul în mentenanță, stakeholderi), alegerea instrumentarului de analiză, selectarea metodelor de analiză statistică, colectarea, pregătirea datelor și prelucrarea datelor, formularea ipotezelor bazate pe datele din sistem, verificarea ipotezelor lansate, formularea concluziilor după compararea instrumentelor de analiză, metodele de analiză și a rezultatelor ipotezelor.

CONCLUZII

Au fost atinse următoarele obiective stabilite inițial: alegerea instrumentariului de analiză; selectarea metodelor de analiză statistică; colectarea datelor; prelucrarea, pregătirea datelor; formularea ipotezelor bazate pe datele din sistem; verificarea ipotezelor; formularea concluziilor după compararea instrumentelor de analiză, metodele de analiză și a rezultatelor ipotezelor.

Drept instrumentariu de analiză a servit integral limbajul de programare general Python și mediul de lucru Google Colab. Bibliotecile necesare erau preinstalte: Stats, Seaborn bazat pe Matplotlib, Pandas, Numpy, Google Colab oferă 12 GB de spațiu virtual RAM și circa 90 GB spațiu de stocare în proces de lucru, ceea ce a înlesnit enorm calculele la etapa de colectare a datelor deoarece se convertea fișierul inițial din format JSON în format de lucru cu dataframe-uri .csv. La această etapă au fost selectate câmpurile, din cele imbricate, care poartă informația necesară la analiza și testarea ipotezelor ulterioare, deci era necesar de a avea o listă preventivă de ipoteze pentru a înțelege care câmpuri din circa 90 total, vor fi purtătoare de informație pozitivă la rezolvarea problemelor.

La momentul finisării colectării dataset-ului, au fost definitivate ipotezele care vor trece testarea pentru a fi implementate în practica ulterioară. Inițial au fost propuse 20 de ipoteze în total, dar s-a redus numărul lor final la doar 15 de bază, din motiv de complexitate și plus-valoare operațională. Ipotezele au fost divizate în 3 blocuri care descriu următoarele domenii: indicatori de eficiență a livrării, volumul de vânzări, zona și timpul de livrare.

Datele sunt individuale pentru fiecare proces și trebuie ajustate, atribuite unei distribuții și conform parametrilor, ales testul potrivit.

Au fost aplicate tipurile de teste : Z- Test, T- Test Independent , Chi–pătrat test, Mann- Whitney U, Fligner Killeen, Shapiro-Wilk, Pearson. Rezultatele lor au fost suprapuse cu reprezentarea grafică a eșantioanelor comparate și nu provoacă îndoieli.

Pentru a mări exactitatea s-ar putea alege alt nivel de semnificație α, și metoda cu puterea mai mare a testului. Există erori de speța I și erori de speța II care nu pot fi evitate în totalitate. La micșorarea excesivă a nivelului de semnificație, scade eroare de speța I dar crește eroarea de speța II, adică riscul de a accepta o ipoteză falsă (vezi tabelul 1.1).

Odată ce a fost descris modelul matematic care conduce logica de prelucare a datelor, devine clar care sunt elementele țintă, de bază la structurarea datelor și cerințele care apar față de ”calitatea” lor la normalizare, colectare din baza de date. Nivelul standard de semnificație ales este $\alpha = 0.05$, deci și eroarea de speța I va fi de acelși ordin. Eroarea de speța I poate fi diminuată doar în detrimentul erorii β de speța II. Cu cât se reduce probabilitatea de a respinge eronat o ipoteză adevărată, cu

atât crește probabilitatea de a accepta o ipoteză falsă (vezi tabelul 1.1).

Din 15 ipoteze a fost doar un singur caz unde p-valoarea rezultantă era la hotarul pragului de semnificație, în celelalte cazuri rezultatul testului nu au creat ambiguități. Rezultatul final a 5 din 15 ipoteze reprezintă un grad sporit de interes practic care urmează de a fi propus managerilor superiori care pot beneficia din rezultatul testelor efectuate, unele fiind contra-intuitive și nu pot fi deduse decât prin analiza specială a proceselor logistico-financiare.

Bibliografie

[1] Mark Smallcombe. 6 database schema designs and how to use them. https://www.integrate.io/blog/database-schema-examples/#six, 03.09.2021. [Online; accesat 05-2022].

[2] Baicuș C. Analiza datelor. http://www.baicus.ro/MCS/Curs_Statistica.pdf, 2009.

[3] Data Flair. Introduction to hypothesis testing in r – learn every concept from scratch! https://data-flair.training/blogs/hypothesis-testing-in-r/. [Online; accesat 16-05-2022].

[4] Addinsoft. What is the difference between a parametric and a nonparametric test? https://help.xlstat.com/6739-what-difference-between-parametric-and-nonparametric. [Online; accesat 16-05-2022].

[5] Amiya Ranjan Rout. Fligner-killeen test in r programming. https://www.geeksforgeeks.org/fligner-killeen-test-in-r-programming/, 2 Oct, 2020.

[6] Stitchdata.com. Top 24 tools for data analysis and how to decide between them. https://www.stitchdata.com/resources/data-analysis-tools, 2022. [Online; accesat 05-2022].

[7] Oleksandr Shykolovych. Top 12 data analysis software: How to choose the one that will drive your growth. https://improvado.io/blog/top-12-data-analysis-software, 2021. [Online; accesat 05-2022].

[8] Microsoft Help and Support. Microsoft Support. Excel 2019. https://support.microsoft.com/en-us/office/command-line-switches-for-microsoft-office-products-079164cd-4ef5-4178-b235-441737d
ocmsassetid=ha102919739&ctt=1&correlationid=47950baf-30c0-4ee1-94c2-7b7ff7cee5d2&ui=en-us&rs=en-us&ad=us, May 7, 2007. [Online; accesat 05-2022].

[9] Oleksandr Shykolovych. Using excel - pc or mac? | excel lemon. http://webarchive.loc.gov/all/20160921074527/https://www.excellemon.com/view/100-using-excel-pc-or-mac, September 21, 2016. [Online; accesat 04-2022].

[10] R-project. What is r? https://www.r-project.org/about.html, 2021. [Online; accesat 05-2022].

[11] Priya Pedamkar. What is r? https://www.educba.com/r-vs-spss/, 2020. [Online; accesat 05-2022].

[12] Peterson Benjamin. Python insider: Python 2.7.18, the last release of python. https://pythoninsider.blogspot.com/2020/04/python-2718-last-release-of-python-2.html, 20 April 2020. [Online; accesat 27 April 2020].

[13] Python Software Foundation. Applications for python. https://www.python.org/about/apps, 2022. [Online; accesat 05-2022].

[14] Vijaysinh Lendave. Guide to knime – a gui way of data science. https://analyticsindiamag.com/guide-to-knime-a-gui-way-of-data-science/, 05-09-2021. [Online; accesat 05-05-2022].

[15] KNIME Analytics Platform. Creating data science. https://www.knime.com/knime-analytics-platform, 2022. [Online; accesat 02-05-2022].

[16] by Victoria Nava. Top apache spark use cases. https://www.qubole.com/blog/apache-spark-use-cases/#:~:text=Apache%20Spark's%20key%20use%20case,to%20handle%20this%20extra%20workload., March 10, 2016. [Online; updatat pe 7 mai, 2022].

[17] Nguyen Kim Anh. Relational design theory. openstax cnx. http://cnx.org/contents/606cc532-0b1d-419d-a0ec-ac4e2e2d533b@1@1, 8 Jul 2009. [Online; accesat 05-2022].

[18] Gordon. Russell. Chapter 4 – normalisation. database elearning. db.grussell.org/ch4.html. [Online; accesat 05-2022].

[19] Amazon AWS. Data warehouse concepts. https://aws.amazon.com/data-warehouse/. [Online; accesat 04-05-2022].

[20] Stephanie Glen. Box cox transformation: Definition, examples. from statisticshowto.com: Elementary statistics for the rest of us! https://www.statisticshowto.com/box-cox-transformation/, 2022.

[21] Stela Vdovii. În 2021, salariul mediu al unui moldovean a constituit 420 de euro. https://agora.md/stiri/97364/in-2021-salariul-mediu-al-unui-moldovean-a-constituit-420-de-euro. [Online; accesat 10-05-2022].