Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

Segmentácia používateľov pre personalizované odporúčania

Bakalárska práca

2025
Tadeáš Kaminský

Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

Segmentácia používateľov pre personalizované odporúčania

Bakalárska práca

Študijný program:	Dátová veda
Študijný odbor:	Informatika a Matematika
Školiace pracovisko:	Katedra informatiky
Školiteľ:	Ing. Ondrej Kaššák, PhD.

Bratislava, 2025
Tadeáš Kaminský

Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

ZADANIE ZÁVEREČNEJ PRÁCE

Meno a priezvisko študenta: Tadeáš Kaminský
Študijný program: dátová veda (Medziodborové štúdium, bakalársky I. st., denná forma)
Študijné odbory: informatika
matematika
Typ záverečnej práce: bakalárska
Jazyk záverečnej práce: slovenský
Sekundárny jazyk: anglický

Názov: Segmentácia používateľov pre personalizované odporúčania
User segmentation for personalized recommendations

Anotácia: Predmetom práce je preskúmať možnosti segmentácie používateľov eshopov pre účely personalizácie odporúčania. Cieľom je zvýšiť kvalitu odporúčania jednotlivým používateľov prostredníctvom ich zaradenia do vhodných segmentov, najmä na základe predošlého správania, a výberom odporúčaní vychádzajúcich zo správania sa ostatných používateľov segmentu.

Vedúci: Ing. Ondrej Kaššák, PhD.
Katedra: FMFI.KI - Katedra informatiky
Vedúci katedry: prof. RNDr. Martin Škoviera, PhD.

Dátum zadania: 25.10.2024

Dátum schválenia: 05.11.2024

doc. Mgr. Tomáš Vinař, PhD.
garant študijného programu

.....
študent

.....
vedúci práce

Poďakovanie: Ďakujem svojmu školiteľovi Ing. Ondrejovi Kaššákovi, PhD. za jeho odborné vedenie, cenné rady, trpezlivosť a podporu počas písania bakalárskej práce.

Abstrakt

Táto bakalárska práca sa zaoberá problematikou personalizácie odporúčaní v elektronickom obchode prostredníctvom segmentácie používateľov. Cieľom práce bolo preskúmať, ako rozdelenie používateľov do skupín na základe ich cenovej citlivosti, odvodenej z interakcií s produktmi, ovplyvňuje kvalitu a úspešnosť produktových odporúčaní v porovnaní s prístupom odporúčania globálne najpopulárnejších produktov. Experimentálne výsledky ukázali, že prínos takejto segmentácie pri odporúčaní konkrétnych produktov je výrazne závislý od počtu zobrazovaných položiek. Pri odporúčaní menšieho počtu produktov (napríklad jedného až piatich) segmentácia významne zlepšila presnosť odporúčaní a kvalitu zoradenia relevantných položiek. V prípade odporúčania väčšieho počtu produktov (napríklad desiatich) síce segmentácia nezvýšila celkovú presnosť, ale dokázala lepšie zoradiť produkty, ktoré boli pre používateľov relevantné. Pri hodnotení odporúčaní na úrovni kategórií produktov sa ukázalo, že segmentácia založená na cenovej citlivosti bola prínosná iba v špecifickom prípade odporúčania jedinej produktovej kategórie; pre väčší počet odporúčaných kategórií nepriniesla zlepšenie oproti globálnemu prístupu. Celkovo práca demonštruje, že segmentácia používateľov podľa cenovej citlivosti môže viesť k relevantnejším odporúčaniam, avšak jej efektivita a praktická aplikovateľnosť sú podmienené konkrétnym scenárom použitia, najmä počtom odporúčaných položiek a tým, či sa zameriavame na odporúčanie konkrétnych produktov alebo ich kategórií.

Kľúčové slová: segmentácia, personalizácia, personalizované odporúčania

Abstract

This bachelor thesis deals with the problem of personalizing recommendations in e-commerce through the segmentation of users. The thesis aimed to investigate how dividing users into groups based on their price sensitivity, derived from product interactions, affects the quality and success of product recommendations compared to the approach of recommending the globally most popular products. Experimental results showed that the benefit of such segmentation in recommending specific products strongly depends on the number of items displayed. When recommending a smaller number of products (e.g., one to five), segmentation significantly improved the accuracy of recommendations and the quality of ranking relevant items. For recommending more products (e.g., ten), although segmentation did not improve overall accuracy, it could better rank the relevant products to users. When evaluating recommendations at the product category level, it was found that segmentation based on price sensitivity was only beneficial in recommending a single product category; for a more significant number of recommended categories, it did not improve the global approach. The work demonstrates that segmenting users based on price sensitivity can lead to more relevant recommendations. Still, its effectiveness and practical applicability are contingent on the specific usage scenario, particularly the number of recommended items and the focus on recommending particular products or product categories.

Keywords: segmentation, personalization, personalized recommendations

Obsah

Úvod **1**

1 Motivácia **3**

2 Existujúce prístupy **5**

2.1 Príprava pre segmentáciu 5

2.1.1 Dáta používané na segmentáciu 5

2.2 Segmentácia 6

2.2.1 Význam segmentácie 6

2.2.2 Segmentačné algoritmy 7

2.2.3 Porovnanie metód 14

2.3 Odporúčacie systémy a metódy 15

2.3.1 Existujúce prístupy 15

2.3.2 Overenie kvality odporúčaní 17

2.4 Personalizácia 18

2.4.1 Personalizácia na úrovni segmentov 19

3 Návrh riešenia **21**

4 Hypotézy **23**

5 Implementácia **25**

6 Overenie **29**

6.1 Overenie hypotézy H1 30

6.1.1 Porovnanie kvalitatívnych metrík 30

6.1.2 Zhrnutie 36

6.2 Overenie hypotézy H2 39

6.2.1 Porovnanie kvalitatívnych metrík 39

6.2.2 Zhrnutie 46

Záver **49**

Zoznam obrázkov

2.1 Príklad *K-means* 9
2.2 Príklad *DBSCAN* 11
2.3 Príklad *RFM analýzy* 13
2.4 Príklad *hierarchického zhlukovania* 14

6.1 Vyhodnotenie pomocou *Precision@10* 31
6.2 Vyhodnotenie pomocou *NDCG@10* 31
6.3 Vyhodnotenie pomocou *Precision@5* 32
6.4 Vyhodnotenie pomocou *NDCG@5* 33
6.5 Vyhodnotenie pomocou *Precision@3* 34
6.6 Vyhodnotenie pomocou *NDCG@3* 35
6.7 Vyhodnotenie pomocou *Precision@1* 37
6.8 Vyhodnotenie pomocou *NDCG@1* 37
6.9 Vyhodnotenie pomocou *Precision@10* pre kategórie 40
6.10 Vyhodnotenie pomocou *NDCG@10* pre kategórie 40
6.11 Vyhodnotenie pomocou *Precision@5* pre kategórie 41
6.12 Vyhodnotenie pomocou *NDCG@5* pre kategórie 42
6.13 Vyhodnotenie pomocou *Precision@3* pre kategórie 43
6.14 Vyhodnotenie pomocou *NDCG@3* pre kategórie 44
6.15 Vyhodnotenie pomocou *Precision@1* pre kategórie 44
6.16 Vyhodnotenie pomocou *NDCG@1* pre kategórie 45

Zoznam tabuliek

5.1 Top 10 globálne najpopulárnejších produktov podľa počtu klikov 26

6.1 Vyhodnotenie odporúčaní pomocou Precision@10 a NDCG@10 s percentuálnou zmenou oproti globálnym produktom 32
6.2 Vyhodnotenie odporúčaní pomocou Precision@5 a NDCG@5 s percentuálnou zmenou oproti globálnym produktom 33
6.3 Vyhodnotenie odporúčaní pomocou Precision@3 a NDCG@3 s percentuálnou zmenou oproti globálnym produktom 36
6.4 Vyhodnotenie odporúčaní pomocou Precision@1 a NDCG@1 38
6.5 Vyhodnotenie odporúčaní podľa kategórie pomocou Precision@10 a NDCG@10 s percentuálnou zmenou oproti globálnym produktom 41
6.6 Vyhodnotenie odporúčaní podľa kategórie pomocou Precision@5 a NDCG@5 s percentuálnou zmenou oproti globálnym produktom 42
6.7 Vyhodnotenie odporúčaní podľa kategórie pomocou Precision@3 a NDCG@3 s percentuálnou zmenou oproti globálnym produktom 45
6.8 Vyhodnotenie odporúčaní podľa kategórie pomocou Precision@1 a NDCG@1 s percentuálnou zmenou oproti globálnym produktom 46

Úvod

V súčasnom dynamickom prostredí elektronického obchodu sa firmy stretávajú s rastom objemu dostupných produktov a služieb, čo pre používateľov predstavuje značný problém pri výbere vhodných položiek [28]. Tento jav, známy ako informačné preťaženie, vedie k situáciám, kde sa používatelia stávajú pasívnymi voči ponuke a často opúšťajú nákupné platformy bez uskutočnenia transakcie. Riešením tohto problému sú odporúčacie systémy, ktoré sa snažia automaticky identifikovať a navrhnúť používateľom produkty zodpovedajúce ich preferenciám a potrebám.

Tradičné prístupy k odporúčaniu produktov sa často spoliehajú na globálne populárne položky alebo jednoduché odporúčacie algoritmy na základe správania používateľov. Tieto metódy však nezohľadňujú heterogénnosť používateľskej základne a ich rozdielne nákupné správanie. Jedným z kľúčových problémov je problém studeného štartu, kedy systém nedokáže efektívne odporúčať produkty novým používateľom z dôvodu nedostatku informácií o ich preferenciách a nákupnom správaní.

Segmentácia používateľov predstavuje perspektívny prístup k riešeniu týchto problémov. Namiesto aplikovania univerzálnych riešení na celú používateľskú základňu umožňuje rozdeliť používateľov do homogénnych skupín na základe ich správania, preferencií alebo demografických charakteristík. Takéto rozdelenie následne umožňuje implementáciu personalizovaných algoritmov odporúčania pre každý segment zvlášť, čím sa dosahuje vyššia relevantnosť návrhov a potenciálne lepšie obchodné výsledky.

Cenová citlivosť predstavuje jeden z najdôležitejších faktorov ovplyvňujúcich nákupné rozhodnutia používateľov [47]. Rôzni používatelia majú odlišnú ochotu platiť za produkty v závislosti od ich finančnej situácie, hodnôt a životného štýlu. Segmentácia založená na cenovej citlivosti nám môže poskytnúť cenný pohľad na používateľské preferencie a umožniť vytvorenie efektívnejších odporúčacích systémov.

Cieľom tejto bakalárskej práce je preskúmať možnosti segmentácie používateľov elektronických obchodov na základe ich cenovej citlivosti a vyhodnotiť vplyv takejto segmentácie na kvalitu personalizovaných odporúčaní. Konkrétne sa zameriavame na otázku, či rozdelenie používateľov do segmentov podľa ich interakcií s produktami rôznych cenových úrovní vedie k zlepšeniu miery konverzie odporúčaní v porovnaní s tradičnými prístupmi využívajúcimi globálne populárne produkty.

Kapitola 1

Motivácia

V súčasnosti sa v oblasti elektronického obchodníctva (*angl.* e-commerce) začínajú čoraz viac objavovať stratégie cielené na jednotlivého zákazníka. Obchodníci sa snažia využiť rôzne existujúce prístupy s cieľom zvýšiť predajnosť produktov. Tieto stratégie kladú dôraz na pochopenie individuálnych potrieb a záujmov zákazníkov. Existujúce segmentačné a odporúčacie algoritmy umožňujú presnejšiu identifikáciu preferencií používateľov [1, 38].

Výzvou pri personalizácii je efektívne porozumieť často komplexným a rôznorodým preferenciám širokej škály používateľov. Aby boli odporúčania skutočne relevantné a prínosné, je kľúčové analyzovať dostupné informácie o ich správaní a záujmoch a na základe tohoim prispôsobiť ponuku.

Využitie segmentácie používateľov predstavuje v tomto smere hodnotný prístup. Rozdelením celkovej bázy používateľov do menších, cielenejších skupín (segmentov) na základe spoločných charakteristík alebo vzorcov správania, je možné vytvárať špecifické profily pre tieto segmenty. Na základe týchto profilov je potom možné generovať personalizované odporúčania, ktoré lepšie zodpovedajú jedinečným záujmom a potrebám používateľov v rámci každého segmentu. Týmto spôsobom sa zvyšuje pravdepodobnosť, že odporúčané produkty budú pre používateľa relevantné a zaujímavé, čo vedie k vyššej spokojnosti používateľa a potenciálne aj k vyššej miere konverzie.

Hlavnou motiváciou tejto práce je preto preskúmať možnosti, ako prostredníctvom segmentácie používateľov dosiahnuť vyššiu kvalitu personalizovaných odporúčaní produktov v prostredí elektronických obchodov (*angl.* e-shop). Práca sa zameriava na analýzu, ako segmentácia prispieva k hlbšiemu pochopeniu záujmov, správania a preferencií používateľov, čo následne umožňuje navrhovať a poskytovať cielenejšie, relevantnejšie a efektívnejšie odporúčania.

Kapitola 2

Existujúce prístupy

V tejto kapitole sa zameriame na analýzu existujúcich prístupov a prípravu dát k segmentácii používateľov, personalizáciu a odporúčanie produktov. Postupne sa budeme venovať zberu dát, ich príprave a následnému spracovaniu. Porovnáme segmentačné algoritmy, ktoré by mohli byť užitočné pre náš problém. Následne si priblížime princípy odporúčacích systémov a nakoniec sa zameriame na aplikáciu personalizácie pre jednotlivé segmenty používateľov.

2.1 Príprava pre segmentáciu

Pred aplikáciou segmentácie na používateľov je potrebné zbierať a spracovať dáta o používateľoch. Získanie dát predstavuje zhromaždenie informácií o používateľoch z rôznych zdrojov, buď od elektronického obchodu alebo využitím vlastných analytických nástrojov. Príprava bude zahŕňať aj skúmanie dát a výber atribútov, ktoré budú pre segmentáciu užitočné.

2.1.1 Dáta používané na segmentáciu

Na skonštruovanie robustných segmentačných modelov je potrebné mať k dispozícii dostatočný dátový základ. Dáta môžeme získať z dvoch hlavných skupín zdrojov: *interných* a *externých*. Interné zdroje predstavujú údaje, ktoré generujeme a zbierame priamo prostredníctvom analytických nástrojov implementovaných na elektronických obchodoch. Externé zdroje naopak pochádzajú z okrúhlejšieho ekosystému digitálnych služieb a tretích strán, ktoré nám poskytujú rozšírený pohľad na používateľské správanie [34].

Interné dáta

- **Profitové metriky** – frekvencia nákupov, priemerná hodnota objednávky, dátum posledného nákupu, kumulatívna hodnota všetkých nákupov.
- **Transakčná história** – nakúpené produktové kategórie a konkrétne produkty, hodnotenia produktov, príp. reklamácie.
- **Demografia** – geografická poloha, vek, pohlavie, príjmová úroveň, dosiahnuté vzdelanie.
- **Psychografia** – záujmy, preferované voľnočasové aktivity, hodnotové a názorové orientácie, politické či náboženské postoje.
- **Behaviorálne signály** – navigačné vzory na webe, reakcie na marketingové kampane, počet a dĺžka návštev, interakcia s vernostným programom [34].

Externé dáta

- **Webové cookies** – agregované vzory vyhľadávania a prezerania obsahu naprieč internetom.
- **Sociálne siete** – verejne dostupné profilové informácie, komentáre a interakcie, ktoré odhaľujú záujmy a preferencie používateľov.
- **Prieskumy a dotazníky** – deklarované postoje a spokojnosť získaná prostredníctvom štruktúrovaných otázok [34].

V praxi najčastejšie pracujeme s profitovými a transakčnými dátami, pričom centrálne pojmy sú *konverzia* a *transakcia*. Konverzia označuje želanú akciu používateľa, napr. vloženie odporúčaného produktu do košíka, zatiaľ čo transakcia znamená úspešne dokončený nákup tohto produktu [36].

2.2 Segmentácia

Segmentácia používateľov je proces rozdeľovania veľkého množstva používateľov na menšie skupiny na základe určitých spoločných charakteristík a vlastností [22]. V tejto kapitole sa zameriame na význam segmentácie, objasníme pojem segment a predstavíme existujúce segmentačné algoritmy.

2.2.1 Význam segmentácie

Segmentácia používateľov v prostredí elektronického obchodu predstavuje proces rozdelenia používateľov do skupín tak, aby členovia každého segmentu mali podobné vlastnosti alebo správanie [40]. Získané informácie o segmentoch nám umožňujú realizovať

cielený marketing na úrovni konkrétnych skupín používateľov namiesto všeobecného prístupu [1]. Segmentácia nám pomáha identifikovať rôzne skupiny ľudí, ktorých môžeme osloviť rôznymi spôsobmi. Najziskovejším môžeme ponúknuť špeciálne akcie s cieľom zvýšiť ich lojalitu, zatiaľ skupine používateľov s rizikom odchodu môžeme ponúknuť zľavy alebo iné výhody, aby sme ich udržali. Celý proces segmentácie zahŕňa vytvorenie skupín používateľov a priradenie používateľov do týchto skupín, či sa jedná o existujúcich používateľov alebo o nových používateľov, ktorí sú priradení na základe podobnosti s existujúcimi používateľmi danej skupiny [33].

Charakteristika segmentu

Každý segment má špecifické charakteristiky, ktoré môžeme reprezentovať prostredníctvom reprezentanta daného segmentu. V závislosti od použitej segmentačnej metódy môžeme definovať konkrétneho reprezentatívneho používateľa, teda skutočného člena skupiny, alebo môžeme vypočítať priemerný bod reprezentujúci daný segment. Takýto bod sa nazýva centroid a predstavuje stred zhluku používateľov, pričom je definovaný ako priemer všetkých bodov v danom zhluku [33].

Tvrdé a mäkké zhlukovanie

Pri segmentácii používateľov existujú dva základné prítupy, a to tvrdé a mäkké zhlukovanie. Pri tvrdom zhlukovaní patrí každý používateľ práve do jedného segmentu, zatiaľ čo pri mäkkom zhlukovaní môže byť používateľ priradený do viacerých segmentov naraz. Najpoužívanejšie segmentačné metódy ako napr. *K-means* alebo *DBSCAN* sú založené na tvrdom zhlukovaní, teda každému používateľovi priradia jedinú skupinu na základe najväčšej podobnosti. Naproti tomu *fuzzy zhlukovanie* alebo modely zmiešaných tried umožňujú priradiť používateľov do viacerých segmentov, kde každý bod má pridelenú váhu príslušnosti ku všetkým segmentom [2].

Využitie vzniknutých segmentov

Vytvorené skupiny používateľov predstavujú základ pre personalizáciu a odporúčanie produktov pre tieto skupiny. Segmentácia nám umožňuje personalizáciu obsahu a ponú na úrovni segmentov, teda používateľom, ktorí patria do rôznych segmentov sa môžu zobraziť iné produkty. Cielené odporúčania zvyšujú relevanciu pre používateľa, čo môže viesť k vyššej miere spokojnosti a lojality [46].

2.2.2 Segmentačné algoritmy

K-means

K-means je najrozšírenejší segmentačný algoritmus v oblasti elektronického obchodu [3]. Funguje na princípe rozdelenia dátových bodov do vopred určeného počtu seg-

mentov K, pričom cieľom je minimalizovať vzdialenosť bodov od stredu (centroidu) príslušného segmentu [26, 31].

Máme množinu dátových bodov $X = \{x_1, x_2, \ldots, x_n\}$, kde každý bod x_i je d-rozmerný vektor. Cieľom je rozdeliť túto množinu do k zhlukov a nájsť ich centroidy $\{\mu_1, \mu_2, \ldots, \mu_K\}$, aby sa minimalizovala celková odchýlka v rámci zhlukov. Cieľom je minimalizovať funkciu:

$$J = \sum_{j=1}^{K} \sum_{x_i \in C_j} \|x_i - \mu_j\|^2,$$

kde:

- $\|x_i - \mu_j\|^2$ je euklidovská vzdialenosť medzi dátovým bodom x_i a centroidom μ_j,
- C_j je množina bodov priradených k zhluku j [19, 26].

Výhodou *K-means* je pomerne jednoduchá implementácia, ktorá spočíva v inicializácii centroidov, priraďovaní bodov k najbližšiemu centroidu a následnom prepočítaní nových súradníc centroidov. Vďaka relatívne nízkej výpočtovej náročnosti je vhodný na rýchle spracovanie menšieho množstva dát, pričom jeho časová zložitosť závisí lineárne od počtu dátových bodov, zhlukov, iterácií a dimenzie dát [15].

Hlavnou nevýhodou *K-means* je počiatočné zvolenie parametra K. Pri použití kategorických alebo textových dát (napríklad pohlavie, vek) musíme tieto dáta prekódovať na numerické hodnoty. Nevhodne zvolená hodnota K môže viesť k nerelevantným výsledkom. Centroidy vypočítame pomocou priemeru vzdialeností všetkých bodov v danom zhluku. Ak sa medzi bodmi vyskytujú extrémy (outliers), tieto body môžu značne vychýliť centroid, čo sa negatívne prejaví na kvalite výsledných zhlukov [15].

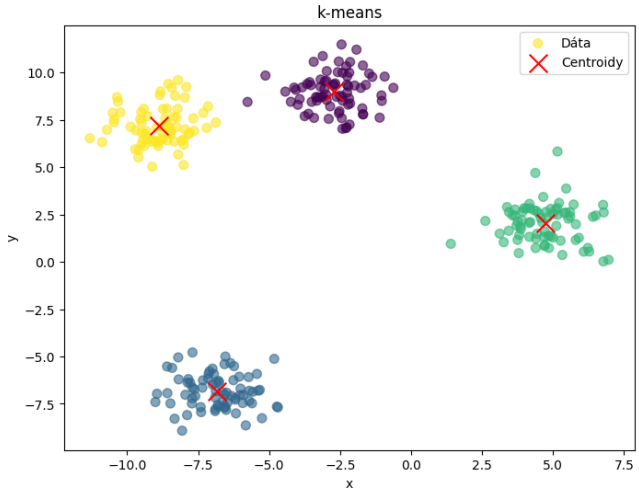
Časová zložitosť algoritmu je

$$\mathcal{O}(n \cdot k \cdot i \cdot d),$$

kde n je počet dát (bodov), K je počet zhlukov, i je počet iterácií a d je rozmer dimenzie dát [19].

V prvom kroku inicializujeme K-means tak, že vyberieme k náhodných centroidov. Následne nastavíme premennej i počiatočnú hodnotu nula. V ďalšom kroku naraz pre všetky body v danej množine určíme, ku ktorému centroidu majú najmenšiu vzdialenosť (napríklad euklidovskú), a podľa toho im priradíme príslušný zhluk. Každý centroid následne aktualizujeme posunutím do priemeru všetkých bodov, ktoré sme mu priradili. Premennú i inkrementujeme o 1 a celý postup opakujeme, až kým sa priradenie bodov alebo samotné centroidy medzi dvomi iteráciami prestanú významne meniť a algoritmus končí [19].

Metóda *K-means* bola aplikovaná v rôznych štúdiách. Autori Liu et al. v roku 2015 použili túto metódu na analýzu približne troch miliónov transakcií z platformy Taobao.com a na základe niekoľkých ukazovateľov nákupného správania rozsegmentovali



Obr. 2.1: Pre $K = 4$ algoritmus *K-means* rozdelil dáta do štyroch zhlukov.

používateľov tejto platformy [1]. Po viacerých optimalizáciách výskumníci získali šesť stabilných segmentov podľa rôznych behaviorálnych charakteristík, napr. nízka cena nákupov a dlhý rozhodovací čas, vysoký počet recenzií a stredná cena nákupov alebo krátke relácie na stránke a nízka cena nákupov [27].

Po vypočítaní nami zvoleného počtu zhlukov K môžeme jednoducho vizualizovať výsledky pomocou knižnice *matplotlib*. Na obrázku 2.1 môžeme vidieť príklad *K-means* algoritmu, ktorý rozdelil dáta do štyroch zhlukov.

DBSCAN

Ďalším populárnym zhlukovacím algoritmom je DBSCAN. Je založený na hustote dátových bodov, pričom ich zoskupuje do zhlukov na základe ich hustoty v dátovom priestore. Zhluky sa vytvoria v priestore, kde je dostatočne vysoká hustota dátových bodov, zatiaľ čo body s nízkou hustotou označujeme za šum alebo za outliers. Metóda DBSCAN nám umožňuje odhaliť zhluky rôznych tvarov a veľkostí, pričom nevyžaduje zvolenie počtu zhlukov [9]. Algoritmus dokáže identifikovať špecifické zhluky, ktoré sú lineárne neseparovateľné, a tým je vhodný pre zhlukovanie dát, ktoré majú rôzne tvary a veľkosti [30].

Základným princípom fungovania DBSCAN je definovanie dvoch parametrov: *epsilon*, ktorý určuje polomer okolia skúmaného bodu, a *minPts*, definujúci minimálny počet bodov v rámci tohto okolia. Algoritmus rozdelí body do štyroch kategórií:

- *jadrové body* sú body, v ktorých okolí *epsilon* sa nachádza aspoň *minPts* bodov.
- *hraničné body* sú body, v ktorých okolí *epsilon* sa nachádza menej bodov ako *minPts*.
- *šumové body* sú hraničné body, ktoré nie sú susedmi žiadneho jadrového bodu.

- *hustotne spojené body* sú hraničné body, ktoré sú susedmi jadrového bodu, ale samy nie sú jadrovými bodmi [30].

Po rozdelení bodov do kategórií vzniknú štyri typy skupín bodov:

- *jadrová skupina* - skupina, ktorej počet bodov je väčší alebo rovný hodnote *minPts*
- *hraničná skupina* - skupina, ktorej počet bodov je menší ako hodnota *minPts*, ale môže byť dosiahnuteľná z *jadrovej skupiny*
- *prázdna skupina* - skupina, ktorá neobsahuje žiadne podriadené body
- *šumová skupina* - skupina, ktorá obsahuje menej bodov než hodnota *minPts* a zároveň nie je prepojená s inou skupinou [30].

Jednou z najväčších výhod DBSCAN je schopnosť pracovať s rôzne veľkými a rôzne tvarovanými zhlukmi. Pomocou parametrov *epsilon* a *minPts* dokážeme prispôsobiť algoritmus tak, aby tieto zhluky dokázal identifikovať a efektívne odlíšiť jadrové body od šumu. V elektronickom obchode vieme pomocou algoritmu nájsť špecifické skupiny používateľov, ktorí majú podobné správanie pri nakupovaní a rozoznať ich od používateľov, ktorí sú menej aktívni alebo sa správajú odlišne. DBSCAN nevyžaduje stanovenie počtu zhlukov, čo môže byť pri segmentácii používateľov užitočné [13].

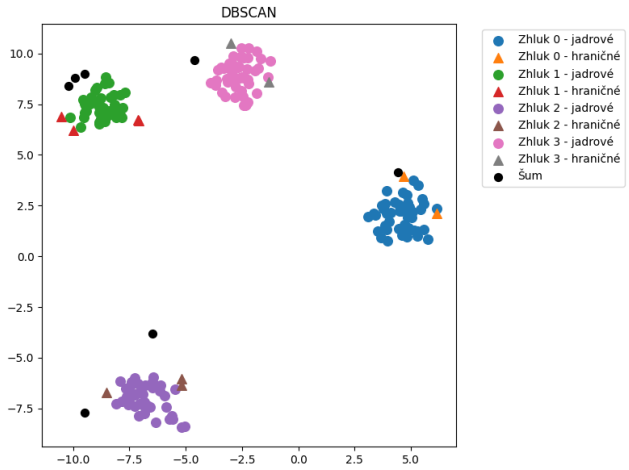
Nevýhodou metódy DBSCAN je citlivosť voči počiatočnému nastaveniu parametrov *epsilon* a *minPts*. Ak sa *epsilon* zvolí príliš malé, môže to viesť k vytvoreniu veľkého množstva malých zhlukov alebo môžeme nesprávne klasifikovať relevantné body za šum, zatiaľ čo príliš veľké *epsilon* môže spôsobiť, že všetky body budú patriť do jedného zhluku, prípadne budú dáta, ktoré spolu nesúvisia, klasifikované do jedného zhluku. Pri viacrozmerných dátach výkon a rýchlosť tejto metódy klesá, pričom sa nastavenie vstupných parametrov stáva náročnejším [13, 37].

Pri použití efektívnej dátovej štruktúry R*-stromu je časová zložitosť algoritmu *DBSCAN*

$$\mathcal{O}(n \cdot \log n),$$

kde n je počet dát (bodov) [13].

Algoritmus inicializujeme zvolením parametrov *epsilon* a *minPts*. Následne vyberáme nenavštívené body, ktotým priradíme množinu jeho susedov podľa parametra *epsilon*. Ak je počet susedov väčší alebo rovný hodnote *minPts*, priradíme bod do zhluku a bod sa stáva jadrovým bodom. Následne prehľadáme všetkých susedov a priradíme ich do zhluku. Tento postup opakujeme, až kým nenájdeme všetky jadrové body. Ak sa nám nepodarí nájsť žiadneho suseda, bod označíme za šumový bod. Po nájdení všetkých jadrových bodov a ich susedov algoritmus končí [37].



Obr. 2.2: Pre *k=4* algoritmus *DBSCAN* rozdelil dáta do štyroch zhlukov a body klasifikoval do troch rôznych skupín.

Štúdia od autorov Govind A. a Rohith Syam z roku 2024 aplikovala algoritmus DBSCAN na zistenie kritických segmentov s vysokou pravdepodobnosťou odchodu (angl. *churn*) z platformy Amazon. Použitím voľne dostupných dát z platformy Kaggle autori dokázali vyextrahovať 8 hustotných segmentov. Na základe týchto segmentov autori navrhli rôzne stratégie pre každý segment na odvrátenie odchodu používateľov, pričom po simulácii aplikovania týchto stratégií autori dosiahli potenciálne zníženie odchodu používateľov o približne 9% [14].

Na obrázku 2.2 môžeme vidieť príklad algoritmu *DBSCAN*, ktorý rozdelil dáta do štyroch zhlukov. Na základe zvoleného parametra *epsilon* algoritmus rozdelil dáta do troch rôznych skupín, pričom šumové body sú označené čiernou farbou.

RFM analýza

RFM analýza je segmentačná metóda, ktorá sa zameriava na hodnotenie a klasifikáciu používateľov na základe ich správania. Táto metóda je založená na troch kľúčových faktoroch: *aktuálnosť - Recency*, *frekvencia - Frequency* a *peňažná hodnota - Monetary*. V elektronickom obchode nám tieto tri faktory poskytujú komplexný pohľad na správanie používateľov internetových obchodov, pričom nám umožňujú identifikovať skupiny používateľov s podobnými vlastnosťami a preferenciami [7].

Jedným z najpoužívanejších typov RFM analýzy je jednoduché skórovanie, pomocou ktorého každému používateľovi priradíme skóre na základe faktorov *Recency*, *Frequency* a *Monetary*. Skóre sa zvyčajne pohybuje v rozmedzí od 1 do 5, pričom 5 znamená najlepšie hodnotenie a 1 najhoršie. Na základe týchto skóre môžeme používateľov rozdeliť do skupín podľa ich hodnotenia a porovnávať skupiny medzi sebou [29]. O používateľovi s hodnotami $F = 5$, $M = 1$ môžeme povedať, že je to používateľ, ktorý uskutočňuje veľa nákupov, ale za menšie sumy. Naopak používateľ s hodnotami

F = 1, M = 5 je používateľ, ktorý nakupuje zriedka, ale hodnota nákupov je veľmi vysoká.

Výhodou RFM analýzy je jednoduchá interpretácia výsledkov, ktoré nám poskytujú jasný obraz o správaní používateľov na základe dostupných transakčných údajov. Metóda pracuje iba s tromi premennými, preto je jednoduché vyhodnotiť a aplikovať výsledky na predikciu ďalšieho správania používateľov [45].

Nevýhodou RFM analýzy je, že nezohľadňuje ďalšie faktory, ktoré môžu ovplyvniť správanie používateľov. Pri použití iba troch faktorov môžeme získať nepresné výsledky, ktoré nezohľadňujú napríklad sezónne trendy, zľavy alebo iné komplexnejšie faktory, ktoré ovplyvňujú správanie používateľov. Model nám neposkytne informácie o nových používateľoch, preto sa môžeme zamerať len na existujúcich používateľov, o ktorých máme dostatok informácii a dát [45].

Celkový čas potrebný na výpočet RFM analýzy závisí od počtu transakcií, ktorými disponujeme, počtom sledovaných používateľov a časovou náročnosťou nami zvolenej metódy na výpočet skóre. Pri väčšom počte dát o jednotlivých transakciách môže výpočet skóre trvať dlhšie, avšak môžeme dostať presnejšie výsledky, ktoré nám umožnia kvalitnejšie rozdelenie používateľov do skupín [45].

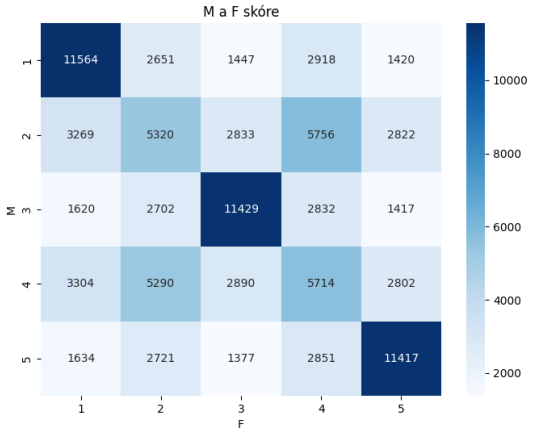
Implementáciu začíname získaním transakcií jednotlivých používateľov, ktoré následne spracujeme a vypočítame R, F a M hodnoty. Následne tieto hodnoty kategorizujeme do nami zvoleného intervalu, aby sme mohli každému používateľovi priradiť príslušné skóre. Posledný krok pozostáva zo zoskupenia používateľov do skupín podľa podobných hodnôt R, F a M [45].

RFM analýza bola použitá v štúdii od Tavakoli et at. z roku 2018, ktorá sa zamerala na iránsky elektronický obchod Digikala, kde autori spracovali a analyzovali približne 10 miliónov nákupov v období medzi rokmi 2014 až 2017. Po predspracovaní dát autori modifikovali definíciu RFM, následne rozdelili atribút *Recency* na tri stavy, *Frequency* použili ako váženú frekvenciu a pomocou *k-means* algoritmu rozdelili používateľov 10 obchodne interpretovateľných skupín. Pre každú skupinu používateľov spustili personalizovanú kampaň pomocou textových správ, pričom sa podarilo zvýšiť priemernú hodnotu košíka (angl. *average order value*) a celkové tržby [41].

Na obrázku 2.3 môžeme vidieť príklad RFM analýzy, ktorý zobrazuje porovnanie hodnôt M a F pre jednotlivých používateľov. Najpočetnejšie zastúpenie majú používatelia, ktorí uskutočnili málo nákupov za malé sumy, veľa nákupov za veľké sumy a stredne veľa nákupov za stredné veľké sumy.

Hierarchické zhlukovanie

Hierarchické zhlukovanie je segmentačná metóda, ktorá nám umožňuje zlúčiť a rozdeliť dáta do zhlukov na základe podobností medzi jednotlivými dátovými bodmi. Pomocou tejto metódy vieme vytvoriť stromovú štruktúru, ktorá nám zobrazuje postupné



Obr. 2.3: Príklad *RFM analýzy* a porovnania hodnôt M a F.

zhlukovanie dát. Vytvorená stromová štruktúra sa nazýva *dendrogram* [16].

Hierarchické zhlukovanie môže byť *aglomeratívne*, kedy každý bod začína v samostatnom zhluku a postupne sa zlúči s inými najbližšími zhlukmi podľa *Euklidovskej vzdialenosti* až kým nevznikne jeden zhluk, alebo *divizívne*, kedy všetky body začínajú v jednom zhluku a postupne sa rozdeľujú na menšie zhluky [21]. V našom probléme môžeme použiť hierarchické zhlukovanie a jeho vizualizáciu pomocou dendogramu pre lepšiu predstavu o podobnostiach medzi našimi dátami.

Metóda má niekoľko výhod. Jednou z nich je nevynútené zvolenie počtu zhlukov, čo je častý problém pri iných metódach. Hierarchické zhlukovanie nám umožňuje zvoliť počet zhlukov až po vizualizácii výsledkov, čo môže byť užitočné pri segmentácii dát. Ďalšou výhodou je možnosť vizualizácie výsledkov pomocou dendogramu, ktorým môžeme jednoducho identifikovať podobné vlastnosti a vzťahy medzi dátovými bodmi.

Nevýhodou hierarchického zhlukovania je jeho vysoká výpočtová a pamäťová náročnosť, ktorá závisí od počtu dát. Po ukončení algoritmu nie je možné meniť vizualizovaný výsledok, čo môže spôsobiť nepresnosti medzi zhlukmi. Pri veľkom počte dát môže výsledná vizualizácia byť ťažko interpretovateľná a neprehľadná. Zlúčením nevhodných dvojíc dát môžeme získať menej presnú segmentáciu, čo môže výrazne ovplyvniť kvalitu výsledkov [18].

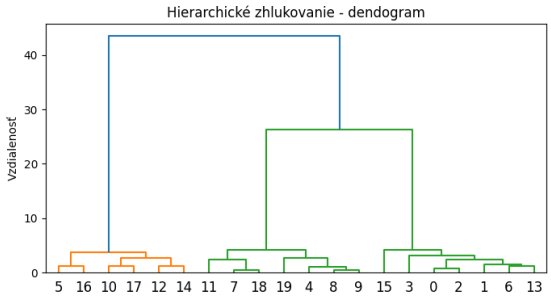
Časová zložitosť algoritmu pri aglomeratívnom prístupe je

$$\mathcal{O}(n^3),$$

pričom pamäťová zložitosť je

$$\mathcal{O}(n^2),$$

kde n je počet dát (bodov). Vysoká pamäťová a časová náročnosť je spôsobená potrebou ukladať a aktualizovať maticu vzdialeností medzi každými dvojicami dát [32, 4].



Obr. 2.4: Príklad *hierarchického zhlukovania* s vytvoreným *dendrogramom*.

Algoritmus začína s každým bodom v samostatnom zhluku. Následne vypočítame maticu vzdialeností o veľkosti $n \times n$, kde n je počet dát. Vyberieme dvojicu zhlukov s najmenšou vzdialenosťou a zlúčime ich do jedného zhluku. Pre tento zjednotený zhluk aktualizujeme vzdialenosti k ostatným zhlukom. Tento proces opakujeme až kým neostane jeden zhluk. Na záver vytvoríme dendrogram, ktorý nám zobrazí postupné zlúčenie dátových bodov do zhlukov.

V štúdii od autorov Sumit Kumar et al. z roku 2025 bolo použité aglomeratívne hierarchické zhlukovanie na analýzu online platformy so zameraním na značky, ktoré používatelia hľadajú na stránke. Výsledkom analýzy boli tri zhluky, pričom autori porovnali výsledky s segmentačnou metódou *K-means*. Autori zistili, že hierarchické zhlukovanie dokáže vytvoriť konzistentnejšie a ľahšie interpretovateľné segmenty v porovnaní s metódou *K-means* [23].

Na obrázku 2.4 môžeme vidieť príklad hierarchického zhlukovania, ktorý nám vytvoril dendrogram s postupným zlúčením dátových bodov do zhlukov.

2.2.3 Porovnanie metód

V tejto sekcii sa zameriame na porovnanie segmentačných metód, ktoré sme si priblížili v predchádzajúcej sekcii a považujú sa za relevantné pre segmentáciu používateľov v elektronickom obchode. Každá z metód reprezentuje odlišný prístup ku klasifikácii používateľov na základe ich správania.

Z pohľadu frekvencie použitia v praxi sa najčastejšie stretávame s metódami *K-means* a *RFM analýza*. *K-means* je často prvou voľbou pre dátových analytikov pri automatickej segmentácii používateľov pre jeho jednoduchosť a rýchlosť [42]. *RFM analýza* je veľmi obľúbená v oblasti marketingu a predaja, pretože poskytuje priamu cestu k identifikácii kľúčových skupín používateľov bez nutnosti pokročilejšieho modelovania. Tieto dve metódy sa často zvyknú kombinovať, kde pomocou *RFM analýzy* získame zmysluplné atribúty a *K-means* použijeme na ich klasifikáciu do zhlukov [20].

Hierarchické zhlukovanie sa v praxi používa menej často, avšak zohráva úlohu v analytických štúdiách ako pomocný nástroj na vizualizáciu a porozumenie dátam. Metóda

DBSCAN ostáva skôr špecializovaným nástrojom pre konkrétne prípady, kde dokáže odhaliť nezvyčajné skupiny používateľov alebo anomálie v správaní. *DBSCAN* je veľmi efektívny pri práci s veľkými objemami dát, avšak jeho nastavenie parametrov môže byť náročné a vyžaduje si hlboké porozumenie štruktúre dát, preto nie je tak rozšírený ako ostatné algoritmy [42].

2.3 Odporúčacie systémy a metódy

Po segmentácii sa zameriame na odporúčanie produktov pre jednotlivé skupiny používateľov. Odporúčacie systémy predstavujú nástroj, ktorý dokáže používateľom navrhnúť produkty alebo služby na základe ich preferencií, histórie nákupov a iných používateľských údajov [5]. Ich cieľom a využitím je uľahčenie orientácie vo veľkom množstve dostupných informácií a zvýšenie predaja identifikovaním a odporúčaním relevantných produktov. Takéto systémy pomáhajú používateľom rýchlejšie nájsť produkty, ktoré by ich mohli zaujať, a tým zvyšujú pravdepodobnosť nákupu [24].

2.3.1 Existujúce prístupy

V tejto kapitole si priblížime niektoré z najpoužívanejších metód odporúčacích systémov.

Kolaboratívne filtrovanie

Jednou z najpoužívanejších metód v odporúčacích systémoch je kolaboratívne filtrovanie [25]. Metóda využíva kolektívne správanie používateľov, pričom predpokladá, že ak viacerí používatelia nakupovali produkty podobným spôsobom, môžu mať aj podobné preferencie. Kolaboratívne filtrovanie môže byť *založené na používateľoch* alebo *založené na produktoch* [5].

Používateľsky orientované filtrovanie (angl. *user-based*) sa snaží nájsť podobných používateľov aktuálneho používateľa a snaží sa mu odporúčať produkty, ktoré si prezerali alebo zakúpili títo podobní používatelia. Výhodou tohto prístupu je, že nevyžaduje žiadne dodatočné informácie o produktoch. Algoritmus môže byť dobrým riešením pre nových používateľov, pre ktorých nemáme dostatok informácií. Pri veľkom počte dát prudko rastie počet výpočtov na zistenie podobností medzi používateľmi, čo môže spomaľovať celkový čas výpočtu odporúčaní [5].

Produktovo orientované filtrovanie (angl. *item-based*) sa zameriava na vzájomné vzťahy medzi produktmi, pričom sa snaží odporúčiť produkty podobné tým produktom, s ktorým mali používatelia interakcie alebo boli často nakupované spolu. Oproti používateľsky orientovanému filtrovaniu je tento prístup lepšie škálovateľný, pretože množstvo produktov v reálnom čase môže byť stabilnejšie ako počet používateľov. Pri

väčšom sortimente produktov môže byť získanie a počítanie podobností medzi všetkými dvojicami produktov náročné.

Pre náš problém by mohlo byť vhodné využiť používateľsky orientované kolaboratívne filtrovanie, ktoré by sme využili pri segmentácii používateľov. Na základe vypočítaných podobností medzi používateľmi budeme schopní identifikovať skupiny používateľov s podobným správaním. Následne by sme mohli pužívateľov v rovnakej skupine odporúčať rovnaké alebo produkty a pozorovať, či ich produkty zaujali a používatelia získali ďalšie interakcie s nimi [39].

Obsahovo orientované odporúčanie

Obsahovo orientované odporúčanie je metóda, ktorá sa zameriava na odporúčanie produktov na základe analýzy ich obsahu a vlastností. Táto metóda sa snaží odporúčať produkty, ktoré sú podobné tým, ktoré používateľ už zakúpil alebo si prezeral. Obsahovo orientované odporúčanie vyžaduje informácie o produktoch, ktoré môžu byť získané z popisov produktov, kategórií, značiek alebo iných atribútov. Na základe týchto informácií môžeme vytvoriť profil používateľa a odporúčať mu produkty, ktoré by mohli vyhovovať jeho preferenciám [35].

Výhodou obsahovo orientovaného odporúčania je, že nevyžaduje informácie o iných používateľoch a ich správaní. Pre konkrétneho používateľa nám stačí jeho profil a informácie o produktoch, s ktorými interagoval. Táto metóda môže byť vhodná pre nových používateľov, pre ktorých nemáme dostatok informácií o ich správaní. Rovnako pomocou obsahu vieme odporúčať aj úplne nové produkty, o ktorých máme obsahové informácie [35].

Tento systém sa môže až príliš sústreďovať len na niektoré konkrétne vlastnosti, ktoré používateľ vyhľadáva, a preto môže byť odporúčanie produktov príliš monotónne a môže sa zamerať iba na jeden druh produktov, čo sťažuje objavovanie nového obsahu. Kvalita odporúčaní bude závisieť od kvality metadát a informácií o produktoch, ktoré máme k dispozícii [35].

Hybridné filtrovanie

Hybridné prístupy v odporúčacích systémoch kombinujú viacero metód a techník s cieľom využiť silné stránky a eliminovať ich slabiny. Hybridný odporúčací systém môže napríklad najprv použiť obsahové filtrovanie na vytvorenie počiatočných odporúčaní pre nového používateľa, o ktorom ešte nemáme potrebné údaje o jeho preferenciách. Následne môžeme použiť kolaboratívne filtrovanie na zlepšenie odporúčaní a zohľadnenie údajov o podobných používateľoch. Hybridné systémy môžu byť zložené z viacerých metód a techník, ktoré sa vzájomne dopĺňajú a zlepšujú kvalitu odporúčaní [5].

Výhodou takýchto systémov je kombinácia viacerých metód, pričom sa snažia eliminovať nevýhody jednotlivých metód. V prípade nedostatku používateľských informácií

sme schopní využiť produktové atribúty, naopak pri nedostatku metadát o produktoch môžeme využiť informácie o používateľoch, čo robí tieto systémy výrazne robustnejšími. Zlúčenie viacerých prístupov môže viesť k diverzifikácii odporúčaní, ktoré môžu byť relevantnejšie a rozmanitejšie [6].

Pri použití viacerých metód na odporúčanie produktov môže byť výpočet výsledných produktov náročnejší. Hybridné systémy môžu byť zložitejšie na implementáciu, pretože vyžadujú kombináciu viacerých metód a techník. Výsledné odporúčania môžu byť zložitejšie na interpretáciu, pretože pochádzajú z viacerých použitých prístupov.

2.3.2 Overenie kvality odporúčaní

Pri použití odporúčacích systémov je dôležité overiť kvalitu odporúčaní, ktoré systém vypočíta. Existuje niekoľko metrík, pomocou ktorých môžeme overiť kvalitu odporúčaní a nášho systému. V tejto časti si priblížime niektoré z najpoužívanejších kvalitatívnych metrík.

Precision@K

Jednou z najzákladnejších metrík je *Precision@K*, ktorá vyjadruje podiel odporúčaných položiek v prvých K výsledkoch, ktoré sú pre daného používateľa relevantné. Pre používateľa u s množinou relevantných položiek $\mathcal{R}_u$ a zoznamom L_u^K odporúčaní dĺžky K, definujeme

$$\text{Precision}_u@K = \frac{\left| L_u^K \cap R_u \right|}{K} [8].$$

Metrika *Precision@K* nám poskytuje informáciu o tom, koľko z odporúčaných položiek je relevantných pre daného používateľa. Táto metrika je veľmi jednoduchá na výpočet a interpretáciu. Presnosť neberie do úvahy poradie odporúčaných položiek, čo môže byť nevýhodou, ak je dôležité, aby sa najrelevantnejšie položky nachádzali na začiatku zoznamu odporúčaní [43].

Recall@K

Ďalšou dôležitou metrikou je *Recall@K*, ktorá dopĺňa presnosť tým, že berie do úvahy, aký podiel zo všetkých relevantných produktov sa podarilo odporučiť. Pre používateľa u definujeme

$$\text{Recall}_u@K = \frac{|L_u^K \cap R_u|}{|R_u|},$$

kde $|R_u|$ je celkový počet relevantných produktov pre daného používateľa [43].

Výsledok metriky nadobúda hodnoty od 0 do 1, pričom hodnota 0.5 znamená, že systém dokázal z polovice všetkých používateľsky relevantných produktov odporučiť aspoň jeden v prvých K odporúčaniach. Táto metrika je užitočná, ak nám ide o pokrytie všetkých preferencií používateľa. Metrika je nezávislá od poradia odporúčaní,

čo môže byť nevýhodou, ak je dôležité, aby sa najrelevantnejšie položky nachádzali na začiatku zoznamu odporúčaní. Nevýhodou je, že metrika *Recall@K* nezohľadňuje počet irelevantných odporúčaní. Systém môže získať vysokú hodnotu *Recall@K*, ale zároveň môže obsahovať veľa irelevantných odporúčaní, preto *Precision@K* bude nízka [8].

nDCG@k

Normalized Discounted Cumulative Gain (nDCG) je metrika, ktorá hodnotí kvalitu odporúčaného zoznamu, pričom zohľadňuje poradie a relevanciu odporúčaných produktov. Metrika vychádza z kumulatívneho zisku (angl. *Cumulative Gain*), ktorý je definovaný ako súčet skóre relevancie jednotlivých výsledkov v poradí. Následne sa zavedie *discounting* faktor, ktorý znižuje hodnotu skóre pre položky, ktoré sú v zozname odporúčaní nižšie. Tento faktor zohľadňuje, že odporúčania na začiatku zoznamu sú pre používateľa dôležitejšie a hodnota odporúčania klesá s jeho pozíciou [17].

Najprv si vyjadríme *Discounted Cumulative Gain* pre prvých K odporúčaní, ktoré počítame ako súčet relevantných produktov delených logaritmickou funkciou pozície i-teho odporúčania

$$\mathrm{DCG}_u@K = \sum_{i=1}^{K} \frac{rel_i}{\log_2(i+1)},$$

kde rel_i je relevancia položky na i-tej pozícii. Následne normalizáciou na interval $[0, 1]$ získame pomer dosiahnutého a ideálneho skóre

$$\mathrm{nDCG}_u@K = \frac{\mathrm{DCG}_u@K}{\mathrm{IDCG}_u@K}[43].$$

Hodnota *nDCG@K* nadobúda hodnoty od 0 do 1, pričom hodnota 1 znamená, že všetky relevantné položky sú na najvyšších pozíciách v zozname odporúčaní. Veľkou výhodou tejto metriky oproti ostatným metrikám je citlivosť na poradie odprúčaní. Zároveň normalizácia voči ideálnemu zoradeniu produktov umožňuje porovnávať výsledky naprieč rôznymi používateľmi a systémami. Nevýhodou *nDCG@K* je vyššia zložitosť na výpočet v porovnaní s ostatnými metrikami, pretože vyžaduje informácie o relevancii produktov a ich pozíciách v zozname odporúčaní [43].

2.4 Personalizácia

Využitie personalizácie v odporúčacích systémoch môže viesť k návrhom produktov, ktoré sú relevantné pre konkrétneho používateľa. Samotná personalizácia predstavuje prispôsobovanie obsahu, produktov alebo inej ponuky na základe preferencií, správania a histórie používateľa. Cielenou úpravou obsahu alebo prezentovaných informácií sa snažíme čo najviac vyhovieť danému používateľovi. Personalizácia má za cieľ zvyšovať relevanciu obsahu pre jednotlivca a zlepšiť jeho orientáciu v našej ponuke [44].

Personalizované prostredie dokáže eliminovať problém informačného preťaženia zobrazovaním len podstatných informácií pre daného používateľa. Celý nákupný proces sa stane prehľadnejším a jednoduchším, čo môže zvýšiť spokojnosť používateľa a pravdepodobnosť vrátenia sa na našu webstránku. Podľa výskumov firiem prináša personalizácia lepšie výkony predaja, kde cielené prispôsobenie ponuky pre používateľov vedie k lojálnosti a zvýšeniu konverzií [1]. Niektoré analýzy ukazujú, že firmy pokročilé v personalizácii generujú výrazne viac tržieb a príjmov práve z týchto aktivít porovnaním s priemerom trhu [10].

2.4.1 Personalizácia na úrovni segmentov

Personalizáciu môžeme využiť aj v našom probléme segmentácie používateľov. Na rozdiel od aplikovania personalizácie konkrétnym používateľom, kedy sa snažíme prispôsobiť obsah pre každého používateľa zvlášť, môžeme personalizáciu využiť aj pre rôzne skupiny používateľov. Personalizácia sa v segmentoch aplikuje tak, že konkrétnej skupine používateľov budeme odporúčať podobné produkty na základe konkrétnych vlastností daného segmentu. Po zozbieraní dát o používateľoch a použití jednej alebo viacerých segmentačných metód opísaných v kapitole 2.2 vznikne niekoľko skupín používateľov, ktoré budú mať konkrétne vlastnosti opisujúce daný segment [1].

Personalizované zacielenie na základe segmentov používateľov prináša niekoľko výhod. Niekoľko výskumov ukazuje, že prispôsobené oslovovanie jednotlivých skupín je pre e-shopy profitabilnejšie ako neadresný prístup, pričom personalizácia dokáže zvýšiť ziskovosť, množstvo klikov na odporúčania a iné dôležité ukazovatele. Štúdia z prostredia elektronického obchodu potvrdila pozitívny vplyv využitia segmentácie spolu s personalizačnými stratégiami na rast predaja, spokojnosť a udržateľnosť používateľov [11].

Kapitola 3

Návrh riešenia

V tejto kapitole predstavujeme návrh metódy, ktoré umožňuje segmentovať používateľov na základe ich správania za účelom zvýšenia kvality personalizovaných odporúčaní. Ná základe teoretických poznatkov z predchádzajúcej kapitoly navrhneme implementáciu, ktorá nám umožní spracovať dáta o správaní používateľov, vykonať segmentáciu používateľov podľa vybranej metódy a následne generovať odporúčania pre jednotlivé segmenty.

Naším cieľom je podrobne popísať jednotlivé kroky návrhu a zdôvodniť výber použitej segmentačnej metódy, odporúčacích algoritmov a hodnotiacich metrík. Na základe analýzy existujúcich prístupov k segmentácii používateľov pre účely generovania personalizovaných odporúčaní sme sa rozhodli navrhnúť riešenie, ktoré kombinuje behaviorálnu segmentáciu používateľov a jednoduchý odporúčací systém.

Výber dát

Pre vytvorenie segmentov používateľov a generovanie odporúčaní budeme využívať reálne behaviorálne dáta používateľov z prostredia elektronického obchodu. Pre každého používateľa máme k dispozícii kliknutia a konverzie vykonané na jednotlivých produktoch. Tieto dáta by nám mali poskytnúť dostatočný základ pre odhad preferencií používateľov a zároveň sú typicky dostupné v reálnych situáciách. Použitím týchto dát zabezpečíme praktickú aplikovateľnosť nášho riešenia a jeho potenciálny prínos pre prax.

Výber segmentačnej metódy

Zo štvorice analyzovaných segmentačných prístupov sme sa rozhodli použiť metódu ***K-means***. Dôvodom je jej jednoduchosť, rýchlosť výpočtu a rozšírená použiteľnosť v praxi.

Metóda *K-means* predpokladá existenciu K zhlukov s podobnou distribúciou a využíva *Euklidovskú vzdialenosť* na priradenie používateľov k príslušnému zhluku. Vytvorené zhluky môžu reprezentovať rôzne typy používateľov, napr. používateľov citlivých

na cenu.

Po vytvorení segmentov na trénovacej množine používateľov je potrebné, aby sme vedeli zaradiť aj nových, v tomto prípade testovacích používateľov, do príslušných segmentov. Na zaradenie používateľa z testovacej skupiny do segmentov použijeme celú históriu jeho klikov, z ktorej vypočítame priemernú cenovú citlivosť a následne mu priradíme segment. Pre každého testovacieho používateľa vypočítame *Euklidovské vzdialenosti* od všetkých centier zhlukov a priradíme ho k zhluku, ktorého stred je najbližší.

Odporúčací systém

Pre každý vytvorený segment budeme generovať odporúčania na základe najpopulárnejších produktov v danom segmente. Popularita je určená na základe počtu interakcií s jednotlivými produktami medzi používateľmi zaradenými v danom segmente. Najpopulárnejšie produkty budeme gerenovať aj pre všetkých používateľov. Porovnaním výsledkov medzi globálnymi a segmentačne personalizovanými odporúčaniami zistíme, či existuje prínos segmentácie používateľov s ohľadom na kvalitu odporúčaní.

Kapitola 4

Hypotézy

V tejto kapitole formulujeme hypotézy, ktoré budeme testovať v našej práci. Hypotézy predstavujú testovateľné tvrdenia, ktorých platnosť budeme overovať prostredníctvom experimentov realizovaných na základe vybranej segmentačnej metódy a hodnotiacich metrík odporúčacích systémov. Cieľom formulovaných hypotéz je preveriť, či zaradenie používateľov do segmentov na základe ich správania vedie k zlepšeniu kvality odporúčaní.

Hypotéza 1: Zaradenie používateľov do segmentov na základe cenovej citlivosti produktov, ktoré si prezerá, vedie k zvýšeniu miery konverzie odporúčaní v porovnaní s globálne populárnymi produktami.

Táto hypotéza vychádza z predpokladu, že cenová citlivosť používateľov predstavuje dôležitú behaviorálnu vlasnosť, ktorá môže ovplyvniť ich rozhodovanie pri nákupe produktov. Rôzni používatelia majú rôznu ochotu platiť za produkty a ich interakcie s produktami v rôznych cenových hladinách môže slúžiť ako indikátor ich cenovej citlivosti.

Segmentácia používateľov na základe cenovej citlivosti umožní prispôsobiť odporúčania tak, aby boli ekonomicky relevantnejšie. Očakávame, že personalizované odporúčania, ktoré vychádzajú z cenového profilu používateľa, budú viesť k vyššej miere konverzie. Pre porovnanie budeme používať odporúčania založené na najpopulárnejších produktoch, ktoré neberú do úvahy cenovú citlivosť používateľov.

Hypotéza 2: Segmentácia používateľov na základe cenovej citlivosti produktov, ktoré si používateľ prezerá, vedie k zvýšeniu miery konverzie produktov v porovnaní s globálne populárnymi produktami, ak za zásah považujeme rovnakú kategóriu odporúčaného produktu.

Vychádzame z predpokladu, že zaradenie používateľov do segmentov podľa ich cenovej citlivosti umožňuje nielen presnejšie odporúčanie konkrétnych produktov, ale predovšetkým zlepšuje relevantnosť odporúčaných kategórií [12]. Predpokladáme, že

používatelia s podobnou cenovou citlivosťou majú tendenciu preferovať produkty rovnakých kategórií, pričom táto kategorická zhoda môže byť často dôležitejšia než zhoda konkrétnych produktov.

Pre overenie tejto hypotézy využijeme upravené kvalitatívne metriky. Relevantnosť odporúčaní v tomto prípade budeme hodnotiť podľa zhody odporúčanej kategórie s kategóriou produktov, ktoré sú pre používateľa relevantné. Týmto spôsobom overíme schopnosť odporúčacieho systému identifikovať vhodné produktové kategórie, o ktoré majú používatelia záujem, aj keď systém nie vždy presne predpovie konkrétny produkt. Takéto hodnotenie môže lepšie vystihovať niektoré reálne scenáre v elektronickom obchode, kde používatelia najprv vyhľadávajú produkty podľa kategórií a až následne si vyberajú konkrétny produkt.

Kapitola 5

Implementácia

V tejto časti sa budeme zaoberať vypracovaním návrhu z predchádzajúcej kapitoly a postupne opíšeme jednotlivé časti implementácie. Naším cieľom bolo vytvoriť systém, ktorý porovná dva prístupy k odporúčaniu produktov a vyhodnotí ich pomocou kvalitatívnych metrík.

Na spracovanie dát a samotnú implementáciu nášho riešenia sme použili programovací jazyk Python. V implementácii sme sa rozhodli použiť nasledujúce knižnice:

- `json` na načítanie vstupných dát vo formáte JSON,
- `datetime` na prácu s časovými údajmi,
- `math` na pomocné výpočty metriky *NDCG@K*,
- `matplotlib` na vizualizáciu výsledkov.

Výpočet globálne najpopulárnejších produktov
Po vytvorení dvoch množín dát (trénovacia a testovacia) sme vypočítali najpopulárnejšie produkty v rámci celej trénovacej skupiny. Popularita v našom prípade znamená celkový počet kliknutí na daný produkt. Funkcia `compute_click_trends` identifikuje globálne najpopulárnejšie produkty na základe počtu klikov v trénovacej množine. Výsledkom je zoradený zoznam 10 najpopulárnejších produktov spolu s ich kategóriami, ktorý je zobrazený v tabuľke 5.1 (pre kompaktnú veľkosť sú zobrazené iba *ID produktu* a *Počet klikov*).

Metriky hodnotenia
Implementovali sme dve základné metriky na vyhodnotenie kvality odporúčania. Metrika *Precision@K* meria presnosť odporúčaných produktov vzhľadom na relevantné produkty, zatiaľ čo *NDCG@K* hodnotí poradie odporúčaných produktov, pričom vyššie umiestnené relevantné produkty prispievajú k lepšiemu skóre.

Tabuľka 5.1: Top 10 globálne najpopulárnejších produktov podľa počtu klikov

ID produktu	**Počet klikov**
product_01559	2996
product_00679	2618
product_04215	1954
product_00640	1810
product_00040	1613
product_02166	1434
product_00461	1432
product_00678	1321
product_00014	1292
product_01032	1292

Okrem štandardných verzií týchto metrík sme implementovali aj ich varianty založené na kategóriách produktov:

- `precision_at_k_category` považuje odporúčanie za relevantné, ak odporúčaný produkt patrí do rovnakej kategórie ako niektorý z produktov v testovacej množine používateľa,
- `ndcg_at_k_category` berie do úvahy kategórie podobne ako predchádzajúca metrika, ale s ohľadom na poradie.

Tieto metriky umožňujú vyhodnotiť kvalitu odporúčaní nielen na základe presných zhôd produktov, ale aj na základe podobnosti kategórií, čo môže byť praktickejšie v reálnych aplikáciách.

Segmentácia používateľov

Segmentácia používateľov bola realizovaná pomocou vlastnej implementácie algoritmu *K-means*, ktorý rozdeľuje používateľov do K skupín podľa podobnej cenovej citlivosti. Proces segmentácie pozostáva z nasledujúcich krokov:

Extrakcia atribútov používateľov Funkcia `extract_user_scalar_attr` vypočítava pre každého používateľa priemerné hodnoty zvolených atribútov (v našom prípade *price_sensitivity*) na základe všetkých produktov, na ktoré používateľ klikol.

K-means algoritmus Vlastná implementácia *K-means* algoritmu rozdeľuje používateľov na základe ich priemernej cenovej citlivosti. Algoritmus inicializuje centroidy rovnomerne v rozsahu hodnôt a iteratívne ich optimalizuje až do konvergencie.

Výpočet populárnych produktov pre segmenty Pre každý vytvorený segment funkcia `compute_cluster_trends` vypočítava najpopulárnejšie produkty na základe počtu klikov používateľov patriacich do daného segmentu.

Experiment a vyhodnotenie

Hlavná funkcia `kmeans_reco_experiment` vykonáva komplexné porovnanie rôznych prístupov k odporúčaniu. Najprv sme extrahovali priemerné hodnoty cenovej citlivosti všetkých alebo zvoleného počtu klikov používateľov. Následne sme pre každú hodnotu K (počet segmentov) vykonali segmentáciu trénovacích používateľov pomocou algoritmu *K-means*. Pre každý segment sme vypočítali najpopulárnejšie produkty a následne sme priradili používateľov z testovacej množiny do najbližších segmentov podľa priemeru všetkých alebo iba niekoľko prvých klikov. Nakoniec sme vyhodnotili kvalitu odporúčaní pomocou metrík *Precision@K* a *NDCG@K* pre rôzne hodnoty K a veľkosti odporúčaní a porovnali ich s hodnotami pre globálne najpopulárnejšie produkty.

Kapitola 6

Overenie

V tejto kapitole overujeme platnosť nami stanovených hypotéz, ktoré sme formulovali na základe predpokladov o zvýšení kvality odporúčaní pomocou segmentácie používateľov podľa cenovej citlivosti. Konkrétne testujeme hypotézy uvedené v časti 4, pričom cieľom je experimentálne potvrdiť alebo vyvrátiť naše tvrdenia o prínose takejto segmentácie s cieľom zvýšiť kvalitu personalizovaných odporúčaní.

Opis dát a metodológia experimentu

Na overenie hypotéz sme realizovali experimenty na reálnych anonymizovaných dátach *110 000* používateľov z oblasti elektronického obchodu v segmente potravín. Experimenty sme navrhli tak, aby sme mohli objektívne porovnať výkonnosť segmentovaných odporúčaní s tradičným prístupom využívajúcim globálne populárne produkty.

Charakteristika datasetu

Dataset obsahuje komplexné údaje o správaní používateľov elektronického obchodu, zahŕňajúce informácie o klikaní na produkty a pridaní odporúčaných produktov do košíka. Pre každého používateľa sú k dispozícii časové údaje jednotlivých interakcií spolu s detailnými atribútmi produktov, vrátane cenovej citlivosti, ktorá predstavuje kľúčový atribút pre našu segmentáciu a kategórie, do ktorej produkt patrí.

Časový rámec experimentu

Pre experimenty sme zvolili časový rámec jeden mesiac - konkrétne február 2025. Tento časový interval sme považovali za dostatočne dlhý na zachytenie rôznorodého správania používateľov, pričom zároveň zabezpečuje homogénnosť dát z hľadiska sezónnych vplyvov.

Rozdelenie dát na trénovaciu a testovaciu množinu

Na vytvorenie robustného experimentálneho návrhu sme implementovali časové rozdelenie dát v pomere *90:10*. Konkrétne sme rozdelili zvolený časový interval nasledovne

- **Trénovacia množina:** Prvých *90%* časového intervalu obsahuje údaje o kliknutiach používateľov na produkty. Tieto dáta slúžia na výpočet segmentácie používateľov na základe cenovej citlivosti a na identifikáciu najpopulárnejších produktov v jednotlivých segmentoch.

- **Testovacia množina:** Posledných *10%* časového intervalu obsahuje údaje o skutočných konverziách používateľov. Tieto dáta používame na vyhodnotenie kvality odporúčaní prostredníctvom porovnania odporúčaných produktov s produktmi, ktoré používatelia skutočne pridali do košíka.

Takéto rozdelenie zabezpečuje, že odporúčania sú generované na základe historických dát a testované na budúcich nákupných rozhodnutiach používateľov, čo dobre simuluje reálne nasadenie odporúčacieho systému.

6.1 Overenie hypotézy H1

Prvú hypotézu sme sa rozhodli testovať porovnaním výkonnosti segmentovaných odporúčaní s globálne populárnymi produktami pomocou metrík *Precision@K* a *nDCG@K* pre rôzne hodnoty K, ktoré si postupne rozoberieme.

6.1.1 Porovnanie kvalitatívnych metrík

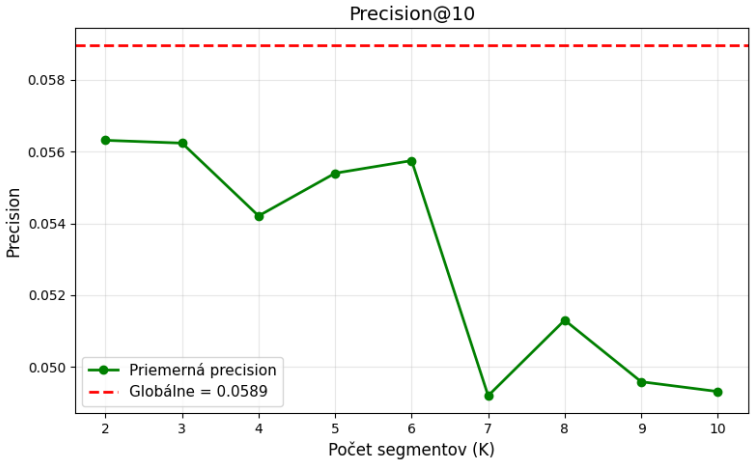
V tejto časti postupne analyzujeme experimentálne výsledky pre rôzne hodnoty K a porovnáme ich s výsledkami pre globálne populárne produkty.

Precision@10 a NDCG@10

Z experimentálnych výsledkov pre metriky hodnotené na desiatich produktoch výplývajú rôzne pozorovania ohľadom účinnosti segmentácie používateľov podľa cenovej citlivosti. Pri porovnaní s globálne populárnymi produktami pozorujeme rozdielne trendy pre jednotlivé metriky.

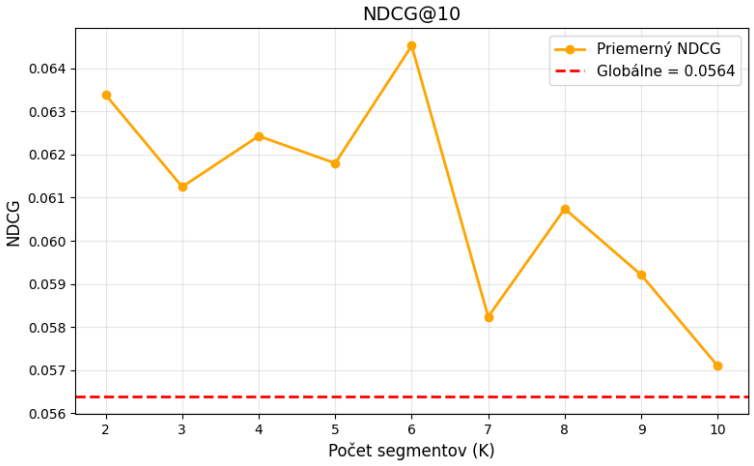
Metrika *Precision@10* meria presnú zhodu medzi odporúčanými produktami a skutočnými konverziami používateľov. Hodnota pre globálne odporúčania dosahuje *0.0589*. Analýza výsledkov po segmentácii ukazuje, že žiadna z testovaných konfigurácií nedokáže prekonať túto hodnotu. Najlepší výsledok segmentácie je dosiahnutý pri $K=2$ s hodnotou *0.0563*, čo predstavuje mierny pokles oproti globálnej hodnote. S rastúcim počtom segmentov pozorujeme pozorujeme postupný pokles výkonnosti, pričom najnižšie hodnoty dosahujeme pri $K=7$ a $K=10$ s hodnotami *0.0492* a *0.0493*. Vizualizáciu môžeme vidieť na obrázku 6.1.

Na rozdiel od metriky *Precision@10* metrika *NDCG@10* vykazuje priaznivejšie výsledky pre segmentáciu. *NDCG@10* je metrika, ktorá zohľadňuje nielen presnú zhodu,



Obr. 6.1: Na obrázku je zobrazené vyhodnotenie pomocou *Precision@10*, pričom globálna hodnota je znázornená horizontálnou prerušovanou čiarou.

ale aj pozíciu relevantných produktov v zozname odporúčaní, čím poskytuje komplexnejší pohľad na kvalitu usporiadania. Hodnota pre globálne odporúčania je *0.0564*. Výsledky segmentácie konzistentne prevyšujú túto hodnotu pre všetky testované konfigurácie s rôznym počtom segmentov. Najvyšší nárast pozorujeme pri $K=6$ s hodnotou *0.0645*, čo predstavuje nárast o *14.4%*. Zaujímavé je, že výkonnosť má vrchol pri $K=6$ a následne klesá s rastúcim počtom segmentov. Konkrétne výsledky môžeme vidieť na obrázku 6.2.



Obr. 6.2: Na obrázku je zobrazené vyhodnotenie pomocou *NDCG@10*, pričom globálna hodnota je znázornená horizontálnou prerušovanou čiarou.

Detailné číselné hodnoty pre testované konfigurácie sú uvedené v tabuľke 6.1. Rozdielne správanie oboch metrík naznačuje, že *segmentácia zlepšuje kvalitu odporúčaní*, aj keď nevedie k zvýšeniu celkového počtu presných zásahov.

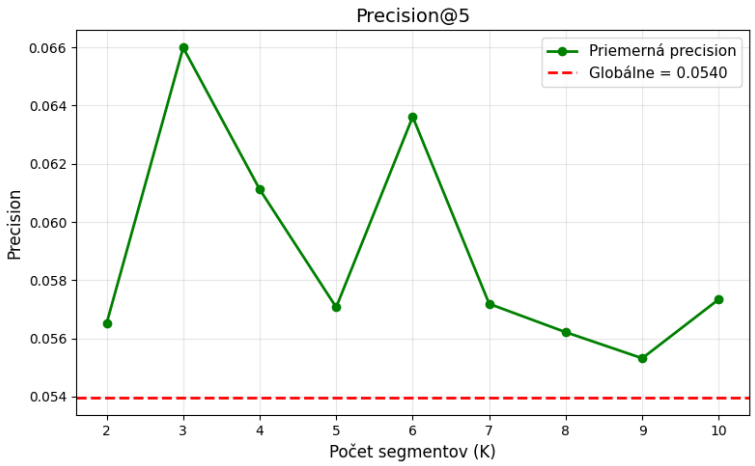
Tabuľka 6.1: Vyhodnotenie odporúčaní pomocou Precision@10 a NDCG@10 s percentuálnou zmenou oproti globálnym produktom

Metóda	**Precision@10**	**NDCG@10**
Globálne populárne produkty	0.0589	0.0564
K-means segmentácia (K=2)	0.0563 (-4.47%)	0.0634 (+12.42%)
K-means segmentácia (K=3)	0.0562 (-4.60%)	0.0613 (+8.64%)
K-means segmentácia (K=4)	0.0542 (-8.04%)	0.0624 (+10.73%)
K-means segmentácia (K=5)	0.0554 (-6.02%)	0.0618 (+9.61%)
K-means segmentácia (K=6)	0.0557 (-5.43%)	0.0645 (+14.44%)
K-means segmentácia (K=7)	0.0492 (-16.53%)	0.0582 (+3.29%)
K-means segmentácia (K=8)	0.0513 (-12.96%)	0.0607 (+7.73%)
K-means segmentácia (K=9)	0.0496 (-15.87%)	0.0592 (+5.01%)
K-means segmentácia (K=10)	0.0493 (-16.33%)	0.0571 (+1.26%)

Precision@5 a NDCG@5

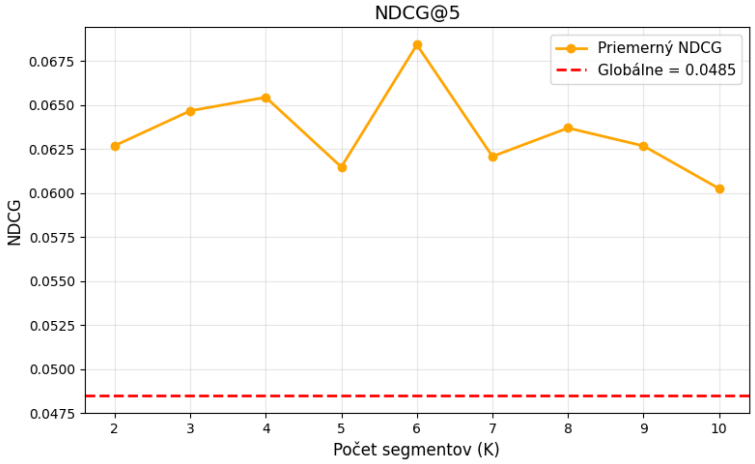
Vyhodnotenie metrík pre päť najpopulárnejších produktov prináša odlišné závery v porovnaní s predchádzajúcimi výsledkami. Pri menšom počte odporúčaní pozorujeme výrazne pozitívnejšie trendy pre segmentáciu v oboch metrikách.

Metrika *Precision@5* vykazuje významné zlepšenie oproti globálnej hodnote *0.0540*. Všetky konfigurácie s rôznym počtom segmentov priniesli lepšie výsledky, pričom najvyšší nárast pozorujeme pri *K=3* s hodnotou *0.0660*, čo predstavuje zlepšenie o *22.2%*. Zaujímavé je výrazné kolísanie výkonnosti v závislosti od počtu segmentov, s vrcholmi pri *K=3* a *K=6* a poklesmi pri *K=5* a *K=9*. Celá vizualizácia je zobrazená na obrázku 6.3.



Obr. 6.3: Na obrázku je zobrazené vyhodnotenie pomocou *Precision@5*, pričom globálna hodnota je znázornená horizontálnou prerušovanou čiarou.

Metrika *NDCG@5* potvrdzuje pozitívny trend segmentácie aj v tomto prípade. Globálna hodnota *0.0485* je prekročená všetkými testovanými konfiguráciami, pričom najvyšší nárast dosahujeme pri *K=6* s hodnotou *0.0684*, čo predstavuje nárast o *41.2%*. Na rozdiel od *Precision@5* pozorujeme pozorujeme stabilnejší trend s postupným nárastom až po *K=6* a následným miernym poklesom. Konkrétne výsledky sú zobrazené na obrázku 6.4.



Obr. 6.4: Na obrázku je zobrazené vyhodnotenie pomocou *NDCG@5*, pričom globálna hodnota je znázornená horizontálnou prerušovanou čiarou.

Detailné hodnoty sú uvedené v tabuľke 6.2. Výsledky pre päť najpopulárnejších produktov jasne demonštrujú prínos segmentácie pre obe metriky, čo naznačuje, že segmentácia je presnejšia a kvalitnejšia pri menšom počte odporúčaní.

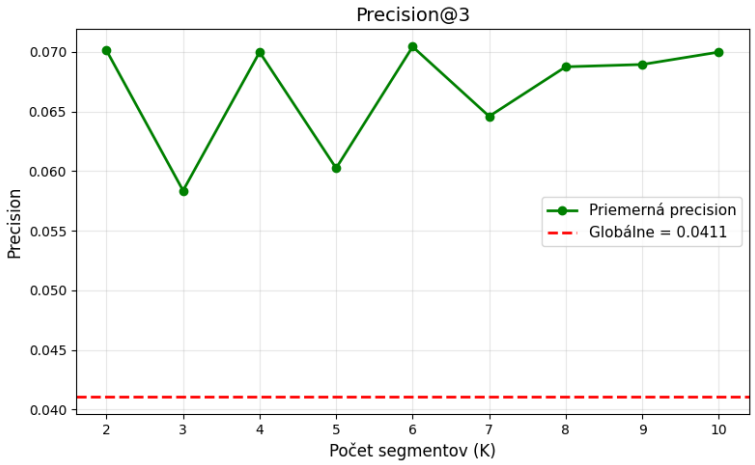
Tabuľka 6.2: Vyhodnotenie odporúčaní pomocou Precision@5 a NDCG@5 s percentuálnou zmenou oproti globálnym produktom

Metóda	**Precision@5**	**NDCG@5**
Globálne populárne produkty	0.0540	0.0485
K-means segmentácia (K=2)	0.0565 (+4.75%)	0.0627 (+29.26%)
K-means segmentácia (K=3)	0.0660 (+22.29%)	0.0647 (+33.35%)
K-means segmentácia (K=4)	0.0611 (+13.27%)	0.0654 (+34.94%)
K-means segmentácia (K=5)	0.0571 (+5.76%)	0.0615 (+26.77%)
K-means segmentácia (K=6)	0.0636 (+17.89%)	0.0684 (+41.08%)
K-means segmentácia (K=7)	0.0572 (+5.98%)	0.0621 (+28.01%)
K-means segmentácia (K=8)	0.0562 (+4.17%)	0.0637 (+31.34%)
K-means segmentácia (K=9)	0.0553 (+2.51%)	0.0627 (+29.24%)
K-means segmentácia (K=10)	0.0573 (+6.27%)	0.0603 (+24.24%)

Precision@3 a NDCG@3

Výsledky pre tri najlepšie produkty predstavujú zatiaľ najvýraznejšiu demonštráciu prínosu segmentácie používateľov. Pri tejto konfigurácii pozorujeme dramatické zlepšenie v oboch metrikách oproti globálnym odporúčaniam.

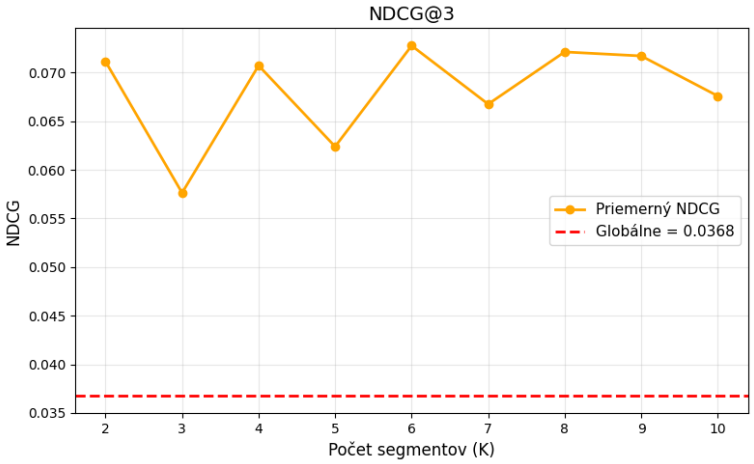
Metrika *Precision@3* dosahuje mimoriadne pozitívne výsledky v porovnaní s hodnotou len *0.0411* pre globálne odporúčania. Všetky konfigurácie s rôznym počtom segmentov významne prevyšujú túto hodnotu, pričom najlepší výsledok dosahuje *K=6* s hodnotou *0.0704*, čo predstavuje nárast o *71.3%*. Výrazné je aj to, že aj najhorší výsledok segmentácie pri *K=3* (*0.0584*) stále prevyšuje hodnotu globálnej konfigurácie o *42.1%*. Obrázok 6.5 ukazuje výrazné kolísanie s vrcholmi pri *K=2*, *K=4*, *K=6* a postupne rastúci trend pri vyšších hodnotách *K*.



Obr. 6.5: Na obrázku je zobrazené vyhodnotenie pomocou *Precision@3*, pričom globálna hodnota je znázornená horizontálnou prerušovanou čiarou.

Metrika *NDCG@3* potvrdzuje výborné výsledky segmentácie v porovnaní s globálnou hodnotou *0.0368*. Najlepší výsledok dosahuje znovu *K=6* s hodnotou *0.0728*, čo predstavuje výnimočné zlepšenie o *97.8%*. Trend je podobný ako pri *Precision@3*, s najvyššími hodnotami pri *K=2* (*0.0711*) a *K=6* (*0.0728*). Zaujímavé je, že aj pri najnižšej hodnote segmentácie (*K=3: 0.0576*) stále dosahujeme zlepšenie o *56.5%* oproti globálnym odporúčaniam.

Detailné hodnoty sú uvedené v tabuľke 6.3. Výsledky pre tri najpopulárnejšie produkty predstavujú dôkaz účinnosti segmentácie pri malom počte odporúčaní. Dramatické zlepšenia v oboch metrikách jasne ukazujú, že segmentácia používateľov podľa cenovej citlivosti je efektívna, keď sa zameriavame na najrelevantnejšie odporúčania. Konzistentne vysoké výsledky naprieč rôznymi hodnotami *K* tiež poukazujú na robustnosť segmentačného prístupu.



Obr. 6.6: Na obrázku je zobrazené vyhodnotenie pomocou *NDCG@3*, pričom globálna hodnota je znázornená horizontálnou prerušovanou čiarou.

Precision@1 a NDCG@1

Výsledky pre jediné odporúčanie predstavujú najextrémnejší test účinnosti segmentácie a poskytujú najpresvedčivejší dôkaz o prínose personalizovaného prístupu. Pri tejto konfigurácii pozorujeme absolútnu dominanciu segmentácie nad globálnymi odporúčaniami.

Metrika *Precision@1* odhaľuje problém globálnych odporúčaní, kde globálna hodnota dosahuje kritických *0.0000*, čo znamená, že žiaden z testovaných používateľov nekúpil najpopulárnejší produkt. V kontraste s týmto, segmentácia dosahuje výnimočné výsledky naprieč všetkými konfiguráciami. Najlepší výsledok pri *K=2* dosahuje hodnotu *0.0906*, čo znamená, že *9.06%* používateľov skutočne zakúpilo prvý odporúčaný produkt z ich segmentu. Vysoké hodnoty pozorujeme aj pri *K=4* (*0.0861*), *K=6* (*0.0869*) a *K=8* (*0.0875*), čo demonštruje konzistentnú efektívnosť segmentačného prístupu. Všetky tieto hodnoty sú zobrazené na obrázku 6.7.

Metrika *NDCG@1* poskytuje identické výsledky ako *Precision@1*, čo je očakávané pri hodnotení jediného odporúčania, kde pozícia v zozname nie je relevantná. Globálna hodnota opäť dosahuje *0.0000*, zatiaľ čo segmentácia dosahuje identické vysoké hodnoty ako pri *Precision@1*. Výsledky sú zobrazené na obrázku 6.8.

Detailné hodnoty sú uvedené v tabuľke 6.4. Výsledky pre jediné odporúčanie predstavujú najpresvedčivejší argument v prospech segmentácie používateľov. Skutočnosť, že globálne najpopulárnejší produkt nezaujal ani jedného testovaného používateľa, zatiaľ čo segmentované odporúčania dosahujú úspešnosť až 9%, jasne demonštruje hodnotu personalizovaného prístupu.

Tabuľka 6.3: Vyhodnotenie odporúčaní pomocou Precision@3 a NDCG@3 s percentuálnou zmenou oproti globálnym produktom

Metóda	**Precision@3**	**NDCG@3**
Globálne populárne produkty	0.0411	0.0368
K-means segmentácia (K=2)	0.0701 (+70.72%)	0.0711 (+93.45%)
K-means segmentácia (K=3)	0.0584 (+42.11%)	0.0576 (+56.74%)
K-means segmentácia (K=4)	0.0700 (+70.40%)	0.0707 (+92.33%)
K-means segmentácia (K=5)	0.0603 (+46.69%)	0.0624 (+69.68%)
K-means segmentácia (K=6)	0.0704 (+71.51%)	0.0728 (+97.89%)
K-means segmentácia (K=7)	0.0646 (+57.28%)	0.0668 (+81.55%)
K-means segmentácia (K=8)	0.0688 (+67.40%)	0.0721 (+96.14%)
K-means segmentácia (K=9)	0.0690 (+67.87%)	0.0717 (+95.02%)
K-means segmentácia (K=10)	0.0700 (+70.40%)	0.0676 (+83.83%)

6.1.2 Zhrnutie

Komplexná analýza výsledkov experimentu naprieč rôznymi konfiguráciami odporúčaní odhaľuje jasný a konzistentný vzorec, ktorý poskytuje jednoznačný pohľad na platnosť prvej hypotézy.

Celkové trendy podľa počtu odporúčaní

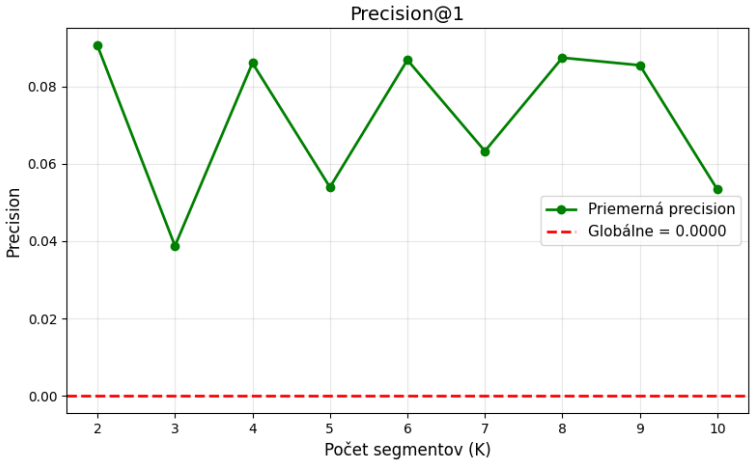
Výsledky demonštrujú inverzný vzťah medzi počtom odporúčaných produktov a účinnosťou segmentácie. Pri *TOP10* pozorujeme zmiešané výsledky s prevažne negatívnymi výsledkami pre *Precision@10*, ale pozitívnymi pre *NDCG@10*. S postupným znižovaním počtu odporúčaní na *TOP5*, *TOP3* a *TOP1* sa prínos segmentácie dramaticky zvyšuje v oboch metrikách.

Metrika Precision@K

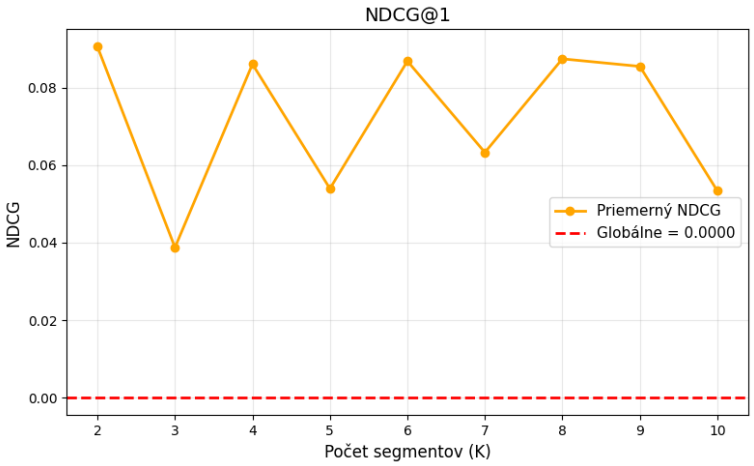
Pri *TOP10* segmentácia nedosahuje lepšie výsledky než globálne odporúčania, pričom najlepší výsledok dosahuje *K=2* s hodnotou *0.0563* oproti globálnym *0.0589*. Situácia sa zásadne mení pri *TOP5*, kde väčšina konfigurácií prekonáva globálne odporúčania, s najlepším výsledkom *K=3* dosahujúcim zlepšenie o *22.2%*. Pri *TOP3* všetky konfigurácie výrazne prevyšujú globálne odporúčania, pričom *K=6* dosahuje zlepšenie o *71.3%*. Najdramatickejší rozdiel pozorujeme pri *TOP1*, kde globálne odporúčania dosahujú *0.0000*, zatiaľ čo *K=2* dosahuje až *0.0906*.

Metrika NDCG@K

Táto metrika konzistentne favorizuje segmentáciu naprieč všetkými konfiguráciami. Pri



Obr. 6.7: Na obrázku je zobrazené vyhodnotenie pomocou *Precision@1*, pričom globálna hodnota je znázornená horizontálnou prerušovanou čiarou.



Obr. 6.8: Na obrázku je zobrazené vyhodnotenie pomocou *NDCG@1*, pričom globálna hodnota je znázornená horizontálnou prerušovanou čiarou.

TOP10 najlepší výsledok *K=6* dosahuje zlepšenie o *14.4%*, pri *TOP5* *K=6* dosahuje zlepšenie o *41%*, pri *TOP3* *K=6* dosahuje výnimočné zlepšenie o *97.8%*, a pri *TOP1* *K=8* dosahuje teoreticky nekonečné zlepšenie oproti nulovým globálnym odporúčaniam.

Optimálny počet segmentov

Naprieč všetkými konfiguráciami sa *K=6* ukazuje ako konzistentne jedna z najlepších volieb, dosahujúc najvyššie alebo druhé najvyššie hodnoty vo väčšine testov. Nižšie hodnoty *K=2* až *K=4* tiež dosahujú výborné výsledky, zatiaľ čo vyššie hodnoty *K=8* až *K=10* majú tendenciu k poklesu výkonnosti.

Tabuľka 6.4: Vyhodnotenie odporúčaní pomocou Precision@1 a NDCG@1

Metóda	**Precision@1**	**NDCG@1**
Globálne populárne produkty	0.0000	0.0000
K-means segmentácia (K=2)	0.0906	0.0906
K-means segmentácia (K=3)	0.0388	0.0388
K-means segmentácia (K=4)	0.0861	0.0861
K-means segmentácia (K=5)	0.0540	0.0540
K-means segmentácia (K=6)	0.0869	0.0869
K-means segmentácia (K=7)	0.0633	0.0633
K-means segmentácia (K=8)	0.0875	0.0875
K-means segmentácia (K=9)	0.0855	0.0855
K-means segmentácia (K=10)	0.0534	0.0534

Vyhodnotenie hypotézy H1

Experimentálne výsledky pre prvú hypotézu (H1) ukázali, že prínos segmentácie používateľov na základe cenovej citlivosti pri odporúčaní konkrétnych produktov je skutočne silno závislý od počtu odporúčaných produktov. Pri odporúčaní menšieho počtu produktov (*Top1*, *Top3*, *Top5*) segmentácia dosahovala dramatické zlepšenia v oboch sledovaných metrikách (Precision@K aj NDCG@K) v porovnaní s globálne populárnymi produktmi. Naopak, pri odporúčaní desiatich produktov (*Top10*) segmentácia síce nezvyšovala celkovú presnosť (Precision@10), ale konzistentne zlepšovala kvalitu zoradenia odporúčaní (NDCG@10), čo naznačuje, že relevantné produkty boli používateľom prezentované na lepších pozíciách.

Na základe týchto experimentálnych pozorovaní, ak by sme hypotézu H1 obmedzili na scenáre s menším počtom odporúčaní (*Top1*, *Top3*, *Top5*), kde segmentácia preukázala jednoznačné a výrazné zlepšenia v oboch metrikách, dáta by takto špecifikovanú hypotézu podporili. Pre scenár s odporúčaním desiatich produktov (*Top10*), dáta nepodporujú tvrdenie o zvýšení celkovej presnosti (Precision@10). Avšak, ak by sme pre *Top10* hodnotili len zlepšenie kvality zoradenia odporúčaní, výsledky pre NDCG@10 by naznačovali podporu pre takto špecificky zameranú hypotézu.

Dôležité je zdôrazniť, že aj v prípade *Top10*, kde segmentácia nedosahuje zlepšenie v metrike *Precision@10*, pozitívne výsledky v metrike *NDCG@10* majú významný praktický prínos. Vyššie hodnoty *NDCG@10* znamenajú, že segmentácia dokáže umiestniť relevantné produkty vyššie v zozname odporúčaní v porovnaní s globálnymi odporúčaniami. To v praxi znamená, že používatelia vidia pre nich zaujímavé produkty skôr, čo môže viesť k vyššej miere angažovanosti a potencionálne k vyšším konverziám, aj keď

celkový počet zásahov v prvých desiatich produktoch zostáva podobný.

6.2 Overenie hypotézy H2

Druhú hypotézu sme testovali pomocou kategorického prístupu k hodnoteniu odporúčaní, kde za úspešný zásah považujeme zhodu kategórie odporúčaného produktu s kategóriou produktov, ktoré používateľ skutočne pridal do košíka. Tento prístup reflektuje reálne správanie používateľov, kde si používatelia často najprv vyberajú produktovú kategóriu a až následne konkrétny produkt.

Hypotéza H2 predpokladá, že segmentácia používateľov na základe cenovej citlivosti vedie k lepšiemu porozumeniu ich kategorických preferencií. Vychádzame z predpokladu, že používatelia s podobnou cenovou citlivosťou majú tendenciu preferovať produkty z rovnakých kategórií, pričom táto kategorická zhoda môže byť v praxi často dôležitejšia než presná zhoda konkrétnych produktov.

6.2.1 Porovnanie kvalitatívnych metrík

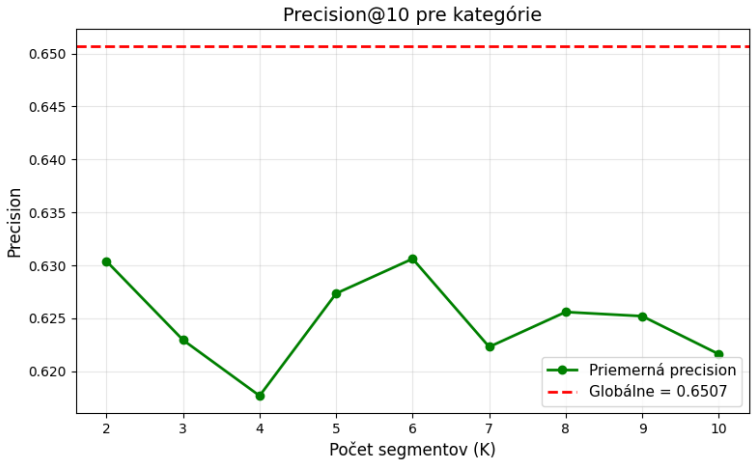
V tejto časti sme sa zamerali na porovnanie kvalitatívnych metrík *Precision@K* a *NDCG@K*, ktoré sme upravili tak, že ako úspešný zásah považujeme situáciu, keď kategória odporúčaného produktu zodpovedá kategórii produktu, ktorý používateľ skutočne pridal do košíka.

Precision@10 a NDCG@10 pre kategórie

Výsledky kategorického hodnotenia pre *TOP10* produktov predstavujú zásadný kontrast v porovnaní s výsledkami prvej hypotézy. Pri kategorickom prístupe pozorujeme odlišný trend, kde segmentácia nedosahuje zlepšenie oproti globálnym odporúčaniam v žiadnej z testovaných konfigurácií.

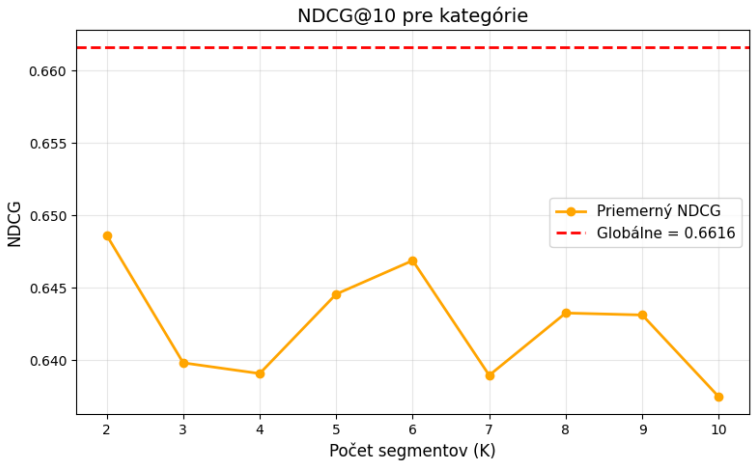
Metrika *Precision@10* pre kategórie vykazuje konzistentne negatívne výsledky pre všetky konfigurácie segmentácie. Globálne odporúčania dosahujú hodnotu *0.6507*, zatiaľ čo najlepší výsledok segmentácie pri *K=2* dosahuje iba *0.6304*, čo predstavuje pokles o *3.11%*. Najhorší výsledok pozorujeme pri *K=10* s hodnotou *0.6216* a poklesom o *4.46%*. Trend ukazuje, že s rastúcim počtom segmentov sa výkonnosť postupne zhoršuje, s výnimkou *K=6*, kde pozorujeme mierne zlepšenie oproti okolitým hodnotám. Všetky tieto hodnoty sú zobrazené na obrázku 6.9.

Metrika *NDCG@10* pre kategórie potvrdzuje negatívny trend segmentácie s globálnou hodnotou *0.6616*. Najlepší výsledok segmentácie dosahuje *K=2* s hodnotou *0.6486*, čo predstavuje pokles o *1.96%*. Najhorší výsledok pozorujeme pri *K=10* s poklesom o *3.65%*. Ostatné hodnoty nájdeme v obrázku 6.10. Podobne ako pri *Precision@10*, pozorujeme postupné zhoršovanie výkonnosti s rastúcim počtom segmentov, pričom *K=6*



Obr. 6.9: Na obrázku je zobrazené vyhodnotenie pomocou *Precision@10* pre kategórie, pričom globálna hodnota je znázornená horizontálnou prerušovanou čiarou.

opäť predstavuje relatívne najlepšiu voľbu medzi vyššími hodnotami K.



Obr. 6.10: Na obrázku je zobrazené vyhodnotenie pomocou *NDCG@10* pre kategórie, pričom globálna hodnota je znázornená horizontálnou prerušovanou čiarou.

Detailné hodnoty sú uvedené v tabuľke 6.5. Tieto výsledky naznačujú, že pri kategorickom hodnotení globálne populárne produkty poskytujú lepšiu kategorickú diverzitu a pokrytie než segmentované odporúčania, čo môže byť spôsobené skutočnosťou, že globálne trendy lepšie zachytávajú širokú škálu kategorickýchh preferencií používateľov.

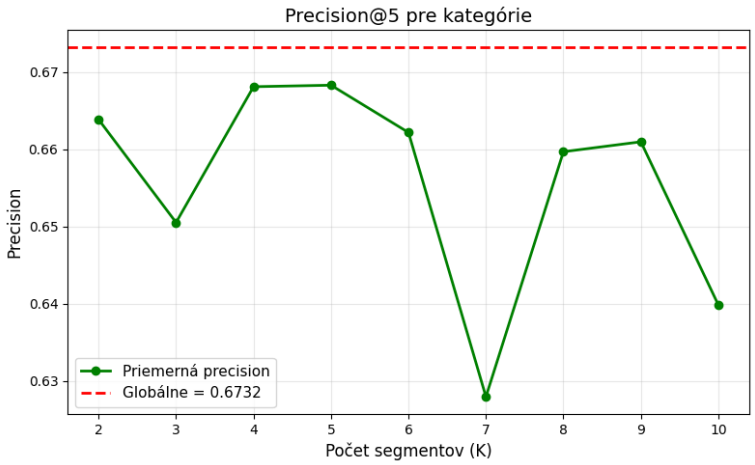
Precision@5 a NDCG@5 pre kategórie

Výsledky kategorického hodnotenia pre *TOP5* produktov ukazujú mierne zlepšenie v porovnaní s *TOP10*, avšak segmentácia stále nedosahuje lepšie výsledky než globálne odporúčania. Pozorujeme však menšie rozdiely a v niektorých prípadoch sa segmentácia približuje k výkonnosti globálnych odporúčaní.

Tabuľka 6.5: Vyhodnotenie odporúčaní podľa kategórie pomocou Precision@10 a NDCG@10 s percentuálnou zmenou oproti globálnym produktom

Metóda	**Precision@10**	**NDCG@10**
Globálne populárne produkty	0.6507	0.6616
K-means segmentácia (K=2)	0.6304 (-3.11%)	0.6486 (-1.96%)
K-means segmentácia (K=3)	0.6229 (-4.26%)	0.6398 (-3.29%)
K-means segmentácia (K=4)	0.6177 (-5.07%)	0.6391 (-3.40%)
K-means segmentácia (K=5)	0.6273 (-3.58%)	0.6446 (-2.57%)
K-means segmentácia (K=6)	0.6306 (-3.08%)	0.6469 (-2.22%)
K-means segmentácia (K=7)	0.6223 (-4.36%)	0.6390 (-3.42%)
K-means segmentácia (K=8)	0.6256 (-3.85%)	0.6433 (-2.77%)
K-means segmentácia (K=9)	0.6252 (-3.91%)	0.6431 (-2.79%)
K-means segmentácia (K=10)	0.6216 (-4.46%)	0.6375 (-3.65%)

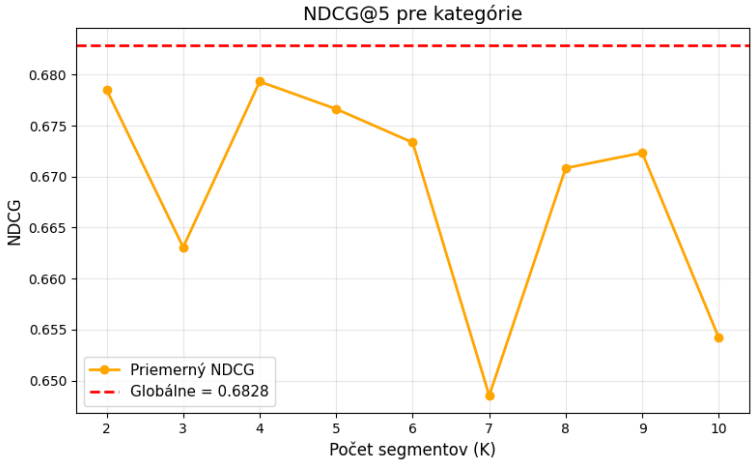
Metrika *Precision@5* pre kategórie vykazuje globálnu hodnotu *0.6732*. Najlepší výsledok segmentácie dosahuje *K=5* s hodnotou *0.6683*, čo predstavuje iba mierny pokles o *0.72%*. Podobne dobré výsledky dosahuje *K=4* s poklesom o *0.75%*. Najhorší výsledok pozorujeme pri *K=7* s výrazným poklesom o *6.71%*. Graf na obrázku 6.11 ukazuje kolísanie výkonnosti, pričom konfigurácie *K=4* a *K=5* dosahujú takmer identické výsledky ako globálne odporúčania.



Obr. 6.11: Na obrázku je zobrazené vyhodnotenie pomocou *Precision@5* pre kategórie, pričom globálna hodnota je znázornená horizontálnou prerušovanou čiarou.

Metrika *NDCG@5* pre kategórie dosahuje globálnu hodnotu *0.6828*. Najlepší výsledok segmentácie pri *K=4* s hodnotou *0.6793* predstavuje minimálny pokles o *0.51%*. Výborné výsledky dosahuje aj *K=2* s poklesom o *0.63%*. Najhorší výsledok opäť pozorujeme pri *K=7* s poklesom o *5.02%*. Trend je podobný ako pri *Precision@5*, s najlep-

šími výsledkami pri nižších hodnotách K a výrazným poklesom pri K=7, čo môžeme vidieť na obrázku 6.12.



Obr. 6.12: Na obrázku je zobrazené vyhodnotenie pomocou *NDCG@5* pre kategórie, pričom globálna hodnota je znázornená horizontálnou prerušovanou čiarou.

Detailné hodnoty sú uvedené v tabuľke 6.6. Výsledky pre *TOP5* naznačujú, že pri menšom počte kategorických odporúčaní sa segmentácia približuje k výkonnosti globálnych odporúčaní, pričom rozdiely sú minimálne pre konfigurácie *K=2*, *K=4* a *K=5*.

Tabuľka 6.6: Vyhodnotenie odporúčaní podľa kategórie pomocou Precision@5 a NDCG@5 s percentuálnou zmenou oproti globálnym produktom

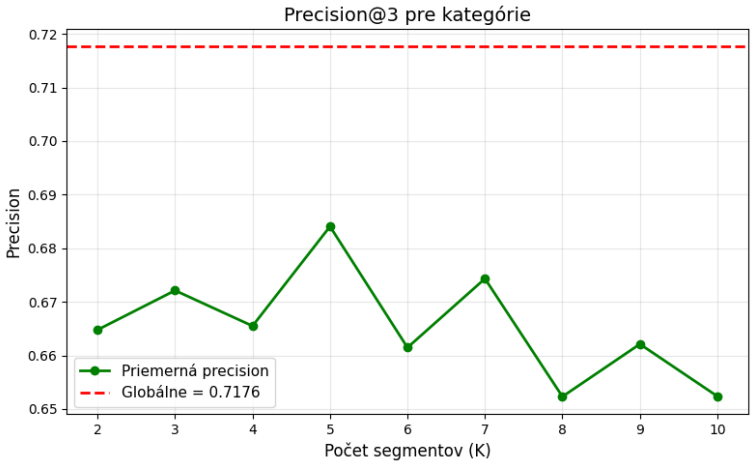
Metóda	**Precision@5**	**NDCG@5**
Globálne populárne produkty	0.6732	0.6828
K-means segmentácia (K=2)	0.6638 (-1.38%)	0.6785 (-0.63%)
K-means segmentácia (K=3)	0.6505 (-3.36%)	0.6631 (-2.89%)
K-means segmentácia (K=4)	0.6681 (-0.75%)	0.6793 (-0.51%)
K-means segmentácia (K=5)	0.6683 (-0.72%)	0.6766 (-0.91%)
K-means segmentácia (K=6)	0.6622 (-1.63%)	0.6733 (-1.39%)
K-means segmentácia (K=7)	0.6280 (-6.71%)	0.6485 (-5.02%)
K-means segmentácia (K=8)	0.6597 (-2.00%)	0.6708 (-1.76%)
K-means segmentácia (K=9)	0.6610 (-1.81%)	0.6723 (-1.54%)
K-means segmentácia (K=10)	0.6399 (-4.95%)	0.6542 (-4.19%)

Precision@3 a NDCG@3 pre kategórie

Výsledky kategorického hodnotenia pre *TOP3* produktov potvrdzujú trend pozorovaný

pri *TOP5* a *TOP10*, kde segmentácia konzistentne nedosahuje lepšie výsledky než globálne odporúčania. Pri *TOP3* pozorujeme najvyššie absolútne hodnoty metrík, ale aj najvýraznejšie rozdiely medzi segmentáciou a globálnymi odporúčaniami.

Metrika *Precision@3* pre kategórie dosahuje najvyššiu globálnu hodnotu *0.7176* spomedzi všetkých testovaných konfigurácií. Najlepší výsledok segmentácie pri K=*5* s hodnotou *0.6841* predstavuje pokles o *4.67%*. Najhoršie výsledky pozorujeme pri K=*8* a K=*10* s poklesmi presahujúcimi *9%*. Zaujímavé je, že K=*5* sa ukázalo ako najlepšia konfigurácia, čo odlišuje *TOP3* od predchádzajúcich výsledkov, kde dominovali nižšie hodnoty K. Presné hodnoty pre ostatné K sú zobrazené na obrázku 6.13.



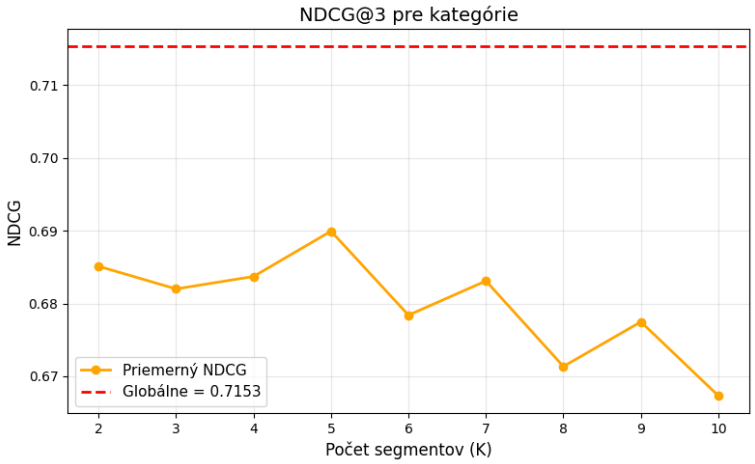
Obr. 6.13: Na obrázku je zobrazené vyhodnotenie pomocou *Precision@3* pre kategórie, pričom globálna hodnota je znázornená horizontálnou prerušovanou čiarou.

Metrika *NDCG@3* pre kategórie dosahuje globálnu hodnotu *0.7153*. Najlepší výsledok segmentácie opäť pri K=*5* s hodnotou *0.6899* predstavuje pokles o *3.54%*. Najhorší výsledok pozorujeme pri K=*10* s poklesom o *6.70%*. Trend je konzistentný s *Precision@3*, pričom K=*5* dosahuje najlepšie výsledky v oboch metrikách, zatiaľ čo vyššie hodnoty K vykazujú postupne sa zhoršujúcu výkonnosť, čo môžeme vidieť na obrázku 6.14.

Detailné hodnoty sú uvedené v tabuľke 6.7. Výsledky pre *TOP3* ukazujú, že aj napriek najvyšším absolútnym hodnotám metrík zostávajú rozdiely medzi segmentáciou a globálnymi odporúčaniami významné, pričom K=*5* predstavuje optimálnu konfiguráciu pre kategoriálne hodnotenie pri malom počte odporúčaní.

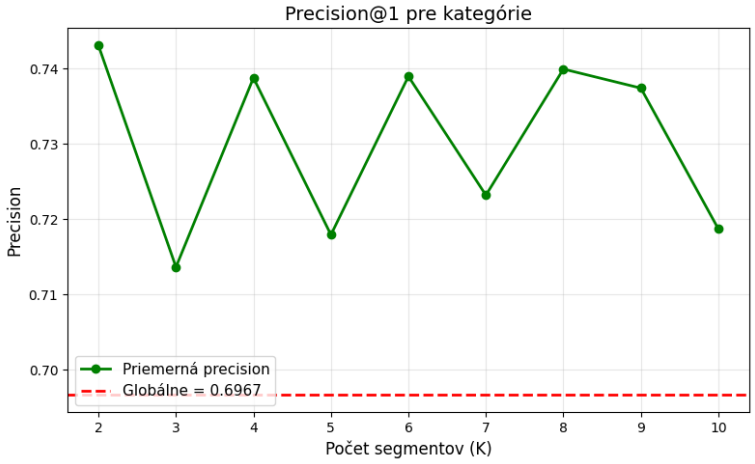
Precision@1 a NDCG@1 pre kategórie

Výsledky kategorického hodnotenia pre jediné odporúčanie predstavujú prelomový moment v našej analýze druhej hypotézy. Pri *TOP1* pozorujeme prvýkrát pozitívne výsledky segmentácie oproti globálnym odporúčaniam, čo naznačuje, že pri kategorického hodnotení je segmentácia najefektívnejšia pri zameraní sa na konkrétny produkt.



Obr. 6.14: Na obrázku je zobrazené vyhodnotenie pomocou *NDCG@3* pre kategórie, pričom globálna hodnota je znázornená horizontálnou prerušovanou čiarou.

Metrika *Precision@1* pre kategórie dosahuje globálnu hodnotu *0.6967*. Všetky konfigurácie segmentácie výrazne prevyšujú túto hodnotu, pričom najlepší výsledok dosahuje *K=8* s hodnotou *0.7400*, čo predstavuje zlepšenie o *6.21%*. Výborné výsledky dosahujú aj *K=2* (*+6.66%*), *K=4* (*+6.04%*) a *K=6* (*+6.07%*). Najnižšie, ale stále pozitívne zlepšenie pozorujeme pri *K=3* s nárastom o *2.43%*. Zaujímavé je, že všetky konfigurácie dosahujú pozitívne výsledky, čo je v ostrom kontraste s predchádzajúcimi konfiguráciami. Tento trend je zobrazený na obrázku 6.15.



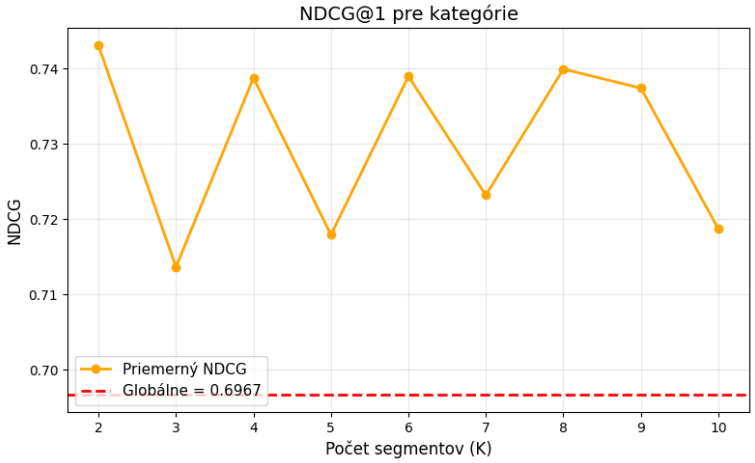
Obr. 6.15: Na obrázku je zobrazené vyhodnotenie pomocou *Precision@1* pre kategórie, pričom globálna hodnota je znázornená horizontálnou prerušovanou čiarou.

Metrika *NDCG@1* pre kategórie poskytuje identické výsledky ako *Precision@1*, čo je očakávané pri hodnotení jediného odporúčania. Globálna hodnota *0.6967* je konzistentne prekonávaná všetkými konfiguráciami segmentácie, s najlepším výsledkom *K=8* dosahujúcim zlepšenie o *6.21%*. Trend je totožný s *Precision@1*, pričom najvyš-

Tabuľka 6.7: Vyhodnotenie odporúčaní podľa kategórie pomocou Precision@3 a NDCG@3 s percentuálnou zmenou oproti globálnym produktom

Metóda	**Precision@3**	**NDCG@3**
Globálne populárne produkty	0.7176	0.7153
K-means segmentácia (K=2)	0.6648 (-7.36%)	0.6851 (-4.22%)
K-means segmentácia (K=3)	0.6721 (-6.34%)	0.6820 (-4.65%)
K-means segmentácia (K=4)	0.6655 (-7.26%)	0.6837 (-4.41%)
K-means segmentácia (K=5)	0.6841 (-4.67%)	0.6899 (-3.54%)
K-means segmentácia (K=6)	0.6615 (-7.82%)	0.6784 (-5.16%)
K-means segmentácia (K=7)	0.6743 (-6.03%)	0.6831 (-4.50%)
K-means segmentácia (K=8)	0.6523 (-9.10%)	0.6714 (-6.14%)
K-means segmentácia (K=9)	0.6621 (-7.73%)	0.6775 (-5.28%)
K-means segmentácia (K=10)	0.6524 (-9.09%)	0.6673 (-6.70%)

šie zlepšenia dosahujú konfigurácie *K=2*, *K=4*, *K=6* a *K=8*. Všetky tieto výsledky sú zobrazené na obrázku 6.16.



Obr. 6.16: Na obrázku je zobrazené vyhodnotenie pomocou *NDCG@1* pre kategórie, pričom globálna hodnota je znázornená horizontálnou prerušovanou čiarou.

Detailné hodnoty sú uvedené v tabuľke 6.8. Výsledky pre *TOP1* predstavujú jedinú konfiguráciu, kde segmentácia konzistentne prekonáva globálne odporúčania v kategorickom hodnotení. Tento prelomový výsledok naznačuje, že segmentácia je najefektívnejšia pri kategorickom odporúčaní jediného najpodstatnejšieho produktu, kde presné zameranie na používateľove kategorické preferencie založené na cenovej citlivosti prináša jasný prínos. Konzistentne pozitívne výsledky naprieč všetkými hodnotami K tiež potvrdzujú robustnosť tohto prístupu pre kategorické odporúčania pri minimálnom počte návrhov.

Tabuľka 6.8: Vyhodnotenie odporúčaní podľa kategórie pomocou Precision@1 a NDCG@1 s percentuálnou zmenou oproti globálnym produktom

Metóda	**Precision@1**	**NDCG@1**
Globálne populárne produkty	0.6967	0.6967
K-means segmentácia (K=2)	0.7431 (+6.66%)	0.7431 (+6.66%)
K-means segmentácia (K=3)	0.7137 (+2.43%)	0.7137 (+2.43%)
K-means segmentácia (K=4)	0.7388 (+6.04%)	0.7388 (+6.04%)
K-means segmentácia (K=5)	0.7180 (+3.05%)	0.7180 (+3.05%)
K-means segmentácia (K=6)	0.7390 (+6.07%)	0.7390 (+6.07%)
K-means segmentácia (K=7)	0.7232 (+3.80%)	0.7232 (+3.80%)
K-means segmentácia (K=8)	0.7400 (+6.21%)	0.7400 (+6.21%)
K-means segmentácia (K=9)	0.7374 (+5.85%)	0.7374 (+5.85%)
K-means segmentácia (K=10)	0.7187 (+3.16%)	0.7187 (+3.16%)

6.2.2 Zhrnutie

Komplexná analýza kategorických výsledkov experimentu odhaľuje jasný a konzistentný vzorec, ktorý poskytuje definitívny pohľad na platnosť druhej hypotézy o efektívnosti segmentácie pri kategorických hodnotení odporúčaní.

Celkové trendy podľa počtu odporúčaní

Výsledky demonštrujú diametrálne odlišný trend v porovnaní s prvou hypotézou. Pri kategorických hodnotení pozorujeme inverzný vzťah. Segmentácia je neefektívna pri vyšších počtoch odporúčaní a stáva sa efektívnou až pri najmenšom počte odporúčaní. *TOP10* a *TOP5* vykazujú negatívne výsledky, *TOP3* dosahuje významné negatívne rozdiely, ale *TOP1* predstavuje prelomový pozitívny výsledok.

Metrika *Precision@K* pre kategórie

Pri *TOP10* segmentácia nedosahuje lepšie výsledky než globálne odporúčania, s najlepším výsledkom *K=2* predstavujúcim pokles o *3.11%*. Pri *TOP5* sa situácia mierne zlepšuje, pričom najlepšie konfigurácie *K=4* a *K=5* dosahujú minimálne poklesy pod *1%*. Pri *TOP3* pozorujeme najvýraznejšie negatívne rozdiely, kde *K=5* dosahuje pokles o *4.67%*. Dramatická zmena nastáva pri *TOP1*, kde všetky konfigurácie dosahujú pozitívne výsledky, s najlepším *K=2* dosahujúcim zlepšenie o *6.66%*.

Metrika *NDCG@K* pre kategórie

Trend *NDCG@K* je konzistentný s *Precision@K*. *TOP10* vykazuje poklesy od *1.96%* do *3.65%*, *TOP5* dosahuje najlepšie výsledky pri *K=4* s minimálnym poklesom o *0.51%*,

TOP3 ukazuje poklesy od *3.54%* do *6.70%*, zatiaľ čo *TOP1* dosahuje konzistentne pozitívne výsledky so zlepšeniami od *2.43%* do *6.66%*.

Optimálny počet segmentov pre kategórie

Na rozdiel od prvej hypotézy neexistuje jednotná optimálna hodnota K pre kategorické hodnotenie. Pre *TOP10* a *TOP5* sú najlepšie nižšie hodnoty *K=2* až *K=5*, pre *TOP3* dominuje *K=5*, zatiaľ čo pre *TOP1* dosahujú výborné výsledky *K=2*, *K=4*, *K=6* a *K=8*. Všeobecne platí, že vyššie hodnoty *K=7* až *K=10* majú tendenciu k horšej výkonnosti.

Záver ohľadom hypotézy *H2*

Experimentálne výsledky pre druhú hypotézu (H2) ukázali, že segmentácia používateľov na základe cenovej citlivosti vedie k zvýšeniu miery konverzie produktov pri kategorickom hodnotení len za veľmi špecifických podmienok. Tento prínos bol extrémne závislý od počtu odporúčaných produktov a ukázal sa ako efektívny iba pri odporúčaní jediného produktu (*Top1*), čo môže byť pre mnohé reálne aplikácie obmedzujúce. Pre konfigurácie s viacerými odporúčaniami (*Top3*, *Top5*, *Top10*) segmentácia nedosiahla zlepšenie oproti globálnym odporúčaniam.

Na základe týchto zistení, ak by sme hypotézu H2 obmedzili výlučne na konfiguráciu odporúčania jediného produktu (*Top1*), kde segmentácia konzistentne dosahovala pozitívne výsledky vo všetkých testovaných konfiguráciách počtu segmentov K, experimentálne dáta by túto špecificky obmedzenú hypotézu podporili. Avšak pre všetky ostatné testované konfigurácie s viacerými odporúčaniami (*Top3*, *Top5*, *Top10*), dáta pôvodnú, širšie formulovanú hypotézu H2 nepodporujú, nakoľko segmentácia v týchto prípadoch nepreukázala zlepšenie oproti globálnym odporúčaniam.

Tieto zistenia majú špecifické praktické implikácie: segmentácia na základe cenovej citlivosti je efektívna pre kategorické odporúčania výlučne pri extrémne presnom zameraní na jediný najpodstatnejší produkt. V reálnych situáciách to znamená, že takáto segmentácia nie je vhodná pre širšie zoznamy odporúčaní. Paradoxne, zatiaľ čo segmentácia zlepšuje presnosť konkrétnych produktových odporúčaní s rastúcou presnosťou zamerania, pri kategorickom hodnotení je efektívna iba pri najvyššej presnosti zamerania.

Záver

Táto práca sa venovala problematike personalizácie v elektronickom obchode prostredníctvom segmentácie používateľov. Hlavným cieľom bolo preskúmať, do akej miery môže segmentácia používateľov elektronických obchodov na základe ich cenovej citlivosti, odvodenej z predošlého správania, prispieť k zvýšeniu kvality a miery konverzie personalizovaných odporúčaní. Práca sa zamerala na porovnanie takto generovaných odporúčaní s odporúčaniami založenými na globálne populárnych produktoch.

Experimentálna časť práce priniesla detailné výsledky, ktoré boli analyzované prostredníctvom dvoch hlavných hypotéz. Prvá hypotéza (H1) skúmala vplyv segmentácie na základe cenovej citlivosti na mieru konverzie odporúčaní konkrétnych produktov. Výsledky ukázali, že prínos segmentácie je silne závislý od počtu odporúčaných produktov.

Pri odporúčaní väčšieho počtu produktov (*Top10*) boli výsledky pre metriku *Precision@10* zmiešané, pričom segmentácia neprekonala globálne odporúčania. Avšak metrika *NDCG@10* konzistentne favorizovala segmentáciu, čo naznačuje lepšie zoradenie relevantných produktov, aj keď celkový počet presných zásahov nebol vyšší.

S klesajúcim počtom odporúčaných produktov (*Top5*, *Top3*, *Top1*) sa prínos segmentácie dramaticky zvyšoval pre obe metriky, *Precision@K* aj *NDCG@K*. Najvýraznejší efekt bol pozorovaný pri *Top1* odporúčaní, kde globálne odporúčania dosiahli nulovú úspešnosť, zatiaľ čo segmentácia dosiahla *Precision@1* až *0.0906* pri *K=2* segmentoch.

Ako optimálny počet segmentov sa naprieč rôznymi konfiguráciami často ukazoval *K=6*, hoci aj nižšie hodnoty *K* (*2* až *4*) prinášali dobré výsledky. Na základe týchto pozorovaní možno konštatovať, že ak by sme hypotézu H1 obmedzili na scenáre s menším počtom odporúčaní (najmä *Top1*, *Top3*, *Top5*), kde segmentácia preukázala výrazné zlepšenia v oboch sledovaných metrikách, experimentálne dáta by ju podporili. Pre scenár s odporúčaním desiatich produktov (*Top10*) dáta nepodporujú tvrdenie o zvýšení celkovej presnosti (*Precision@10*), hoci výsledky pre *NDCG@10* naznačujú prínos v kvalite zoradenia relevantných produktov.

Druhá hypotéza (*H2*) sa zamerala na hodnotenie odporúčaní na úrovni kategórií produktov, predpokladajúc, že používatelia s podobnou cenovou citlivosťou budú preferovať produkty z rovnakých kategórií. Výsledky pre túto hypotézu ukázali diametrálne

odlišný trend v porovnaní s *H1*. Segmentácia sa ukázala ako neefektívna pri odporúčaní viacerých produktov na úrovni kategórií (*Top10*, *Top5*, *Top3*), kde nedosiahla lepšie výsledky ako globálne odporúčania v metrikách *Precision@K* ani *NDCG@K* pre kategórie.

Prelom nastal až pri odporúčaní jediného produktu (*Top1*), kde segmentácia konzistentne prekonávala globálne odporúčania, napríklad pri *K=2* segmentoch dosiahla zlepšenie *Precision@1* pre kategórie o *6.66%*. Optimálny počet segmentov K nebol pre *H2* jednotný a líšil sa v závislosti od počtu odporúčaní.

Z týchto výsledkov vyplýva, že ak by sme hypotézu H2 obmedzili výlučne na odporúčanie jedného produktu na úrovni kategórie (*Top1*), kde segmentácia konzistentne preukázala zlepšenie, naše experimenty by ju podporili. Pre scenáre s odporúčaním viacerých produktov na úrovni kategórií (*Top3*, *Top5*, *Top10*) však dáta hypotézu nepodporujú, keďže segmentácia nepreukázala zlepšenie oproti globálnym odporúčaniam. Prínos segmentácie na základe cenovej citlivosti pre kategorické odporúčania je teda zjavný len pri extrémne cielenom odporúčaní jediného produktu. Toto zistenie poukazuje na paradox, že zatiaľ čo pri odporúčaní konkrétnych produktov segmentácia zlepšuje presnosť s rastúcou presnosťou zamerania (menším N), pri kategorickom hodnotení je efektívna len pri najvyššej možnej presnosti zamerania (*N=1*).

Prínos tejto práce spočíva v detailnej kvantitatívnej analýze účinkov segmentácie používateľov na základe cenovej citlivosti na kvalitu personalizovaných odporúčaní. Ukázalo sa, že táto stratégia nie je univerzálne prospešná, ale jej efektivita výrazne závisí od kontextu, a to najmä od počtu odporúčaných produktov a od toho, či hodnotíme relevanciu na úrovni konkrétnych produktov alebo ich kategórií. Práca tak prispieva k lepšiemu pochopeniu detailnejších aspektov behaviorálnej segmentácie a jej praktických implikácií pre systémy personalizovaných odporúčaní.

Napriek dosiahnutým výsledkom existuje viacero oblastí a otvorených problémov, ktoré ponúkajú priestor pre ďalší výskum. Do budúcnosti by bolo zaujímavé preskúmať využitie iných, prípadne kombinácie viacerých, behaviorálnych atribútov pre tvorbu segmentov. Rovnako by bolo prínosné implementovať a porovnať iné segmentačné algoritmy a sofistikovanejšie metódy generovania odporúčaní v rámci jednotlivých segmentov, ktoré by mohli prekonať prístup založený na jednoduchej popularite.

Literatúra

[1] M. Alves Gomes and T. Meisen. A review on customer segmentation methods for personalized customer targeting in e-commerce use cases. *Inf Syst E-Bus Manage*, 21:527–570, 2023.

[2] Gayathri Asokan and Mohanavalli Subramaniam. Fuzzy clustering for effective customer relationship management in telecom industry. *Communications in Computer and Information Science*, 204:571–580, 01 2011.

[3] Emmanuel Ayodele and Victor Sodeinde. Customer segmentation using the k-means clustering algorithm. *Ilaro Journal of Science and Technology (IJST)*, 4, 2024.

[4] Bilal Bataineh. Fast component density clustering in spatial databases: A novel algorithm. *Information*, 13(10):477, 2022.

[5] Kailash Chowdary Bodduluri, Francis Palma, Arianit Kurti, Ilir Jusufi, and Henrik Löwenadler. Exploring the landscape of hybrid recommendation systems in e-commerce: A systematic literature review. *IEEE Access*, 12:28273–28296, 2024.

[6] Erion Çano and Maurizio Morisio. Hybrid recommender systems: A systematic literature review. *Intelligent data analysis*, 21(6):1487–1524, 2017.

[7] Hui-Chu Chang and Hsiao-Ping Tsai. Group rfm analysis as a novel framework to discover better customer consumption behavior. *Expert Syst. Appl.*, 38:14499–14513, 11 2011.

[8] Paolo Cremonesi, Yehuda Koren, and Roberto Turrin. Performance of recommender algorithms on top-n recommendation tasks. In *Proceedings of the fourth ACM conference on Recommender systems*, pages 39–46, 2010.

[9] Dingsheng Deng. Dbscan clustering algorithm based on density. In *2020 7th International Forum on Electrical Engineering and Automation (IFEEA)*, pages 949–953, 2020.

[10] Darshana Desai. An empirical study of website personalization effect on users intention to revisit e-commerce website through cognitive and hedonic experience. In *Data Management, Analytics and Innovation: Proceedings of ICDMAI 2018, Volume 2*, pages 3–19. Springer, 2019.

[11] Ziqi Duan. Data-driven personalized marketing in e-commerce: Practical applications. *Advances in Economics, Management and Political Sciences*, 102:79–86, 07 2024.

[12] Sri Devi Duvvuri, Asim Ansari, and Sunil Gupta. Consumers' price sensitivities across complementary categories. *Management Science*, 53(12):1933–1945, 2007.

[13] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, page 226–231. AAAI Press, 1996.

[14] A Govind and Rohith Syam. Using dbscan to identify customer segments with high churn risk on amazon consumer behavior data. *Authorea Preprints*, 2024.

[15] Jiawei Han, Micheline Kamber, and Jian Pei. 10 - cluster analysis: Basic concepts and methods. In Jiawei Han, Micheline Kamber, and Jian Pei, editors, *Data Mining (Third Edition)*, The Morgan Kaufmann Series in Data Management Systems, pages 443–495. Morgan Kaufmann, Boston, third edition edition, 2012.

[16] Trevor Hastie, Robert Tibshirani, Jerome Friedman, and James Franklin. The elements of statistical learning: Data mining, inference, and prediction. *Math. Intell.*, 27:83–85, 11 2004.

[17] Aryan Jadon and Avinash Patil. A comprehensive survey of evaluation techniques for recommendation systems. In *International Conference on Computation of Artificial Intelligence & Machine Learning*, pages 281–304. Springer, 2024.

[18] Anil K. Jain. Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, 31(8):651–666, 2010. Award winning papers from the 19th International Conference on Pattern Recognition (ICPR).

[19] Anil K Jain, M Narasimha Murty, and Patrick J Flynn. Data clustering: a review. *ACM computing surveys (CSUR)*, 31(3):264–323, 1999.

[20] Jeen Mary John, Olamilekan Shobayo, and Bayode Ogunleye. An exploration of clustering algorithms for customer segmentation in the uk retail market. *Analytics*, 2(4):809–823, 2023.

[21] Leonard Kaufman and Peter J Rousseeuw. *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons, 2009.

[22] Sumit Koul and Trissa Merrin Philip. Customer segmentation techniques on e-commerce. In *2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, pages 135–138, 2021.

[23] Sumit Kumar, Ruchi Rani, Sanjeev Kumar Pippal, and Riya Agrawal. Customer segmentation in e-commerce: K-means vs hierarchical clustering. *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, 23(1):119–128, 2025.

[24] Dokyun Lee and Kartik Hosanagar. How do product attributes and reviews moderate the impact of recommender systems through purchase stages? *Management Science*, 67, 05 2020.

[25] Cong Li. Research on e-commerce recommendation service using collaborative filtering. In *2009 Second International Symposium on Knowledge Acquisition and Modeling*, volume 2, pages 33–36, 2009.

[26] Aristidis Likas, Nikos Vlassis, and Jakob J. Verbeek. The global k-means clustering algorithm. *Pattern Recognition*, 36(2):451–461, 2003. Biometrics.

[27] Ying Liu, Hong Li, Geng Peng, Benfu Lv, and Chong Zhang. Online purchaser segmentation and promotion strategy selection: evidence from chinese e-commerce market. *Annals of Operations Research*, 233, 2013.

[28] Jun Lv and Xuan Liu. The impact of information overload of e-commerce platform on consumer return intention: Considering the moderating role of perceived environmental effectiveness. *International Journal of Environmental Research and Public Health*, 19(13):8060, 2022.

[29] Bharghav Madhiraju, Sukesh Reddy, and Dr Sasikala. Customer segmentation using rfm analysis. *EPRA International Journal of Economic and Business Review*, pages 15–22, 07 2024.

[30] K. Mahesh Kumar and A. Rama Mohan Reddy. A fast dbscan clustering algorithm by accelerating neighbor searching using groups method. *Pattern Recognition*, 58:39–48, 2016.

[31] Norsyela Muhammad Noor Mathivanan, Nor Azura Md. Ghani, and Roziah Mohd Janor. Analysis of k-means clustering algorithm: A case study using large scale e-commerce products. In *2019 IEEE Conference on Big Data and Analytics (ICBDA)*, pages 1–4, 2019.

[32] What Is Data Mining. Introduction to data mining. *Mining Multimedia Databases, Mining Time Series and*, 2006.

[33] Kobbi Nissim, Rann Smorodinsky, and Moshe Tennenholtz. Segmentation, incentives, and privacy. *Mathematics of Operations Research*, 43(4):1252–1268, 2018.

[34] Juni Nurma Sari, Lukito Nugroho, Ridi Ferdiana, and Paulus Santosa. Review on customer segmentation technique on ecommerce. *Advanced Science Letters*, 22:3018–3022, 10 2016.

[35] Tejashri Sharad Phalle and Shivendu Bhushan. Content based filtering and collaborative filtering: A comparative study. *Journal of Advanced Zoology*, 45, 2024.

[36] Hussain Saleem, M Khawaja Shaiq Uddin, Syed Habib-ur Rehman, Samina Saleem, and Ali Aslam. Strategic data driven approach to improve conversion rates and sales performance of e-commerce websites. *International Journal of Scientific and Engineering Research*, 10:588–593, 04 2019.

[37] Jörg Sander, Martin Ester, Hans-Peter Kriegel, and Xiaowei Xu. Density-based clustering in spatial databases: The algorithm gdbscan and its applications. *Data Mining and Knowledge Discovery*, 2:169–194, 1998.

[38] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. Analysis of recommendation algorithms for e-commerce. In *Proceedings of the 2nd ACM Conference on Electronic Commerce*, pages 158–167, 2000.

[39] J Ben Schafer, Dan Frankowski, Jon Herlocker, and Shilad Sen. Collaborative filtering recommender systems. In *The adaptive web: methods and strategies of web personalization*, pages 291–324. Springer, 2007.

[40] Kayalvily Tabianan, Shubashini Velu, and Vinayakumar Ravi. K-means clustering approach for intelligent customer segmentation using customer purchase behavior data. *Sustainability*, 14(12), 2022.

[41] Mohammadreza Tavakoli, Mohammadreza Molavi, Vahid Masoumi, Majid Mobini, Sadegh Etemad, and Rouhollah Rahmani. Customer segmentation and strategy development based on user behavior analysis, rfm model and data mining techniques: a case study. In *2018 IEEE 15th International Conference on e-Business Engineering (ICEBE)*, pages 119–126. IEEE, 2018.

[42] Shreya Tripathi, Aditya Bhardwaj, and Eswaran Poovammal. Approaches to clustering in customer segmentation. *International Journal of Engineering & Technology*, 7(3.12):802–807, 2018.

[43] Daniel Valcarce, Alejandro Bellogín, Javier Parapar, and Pablo Castells. Assessing ranking metrics in top-n recommendation. *Information Retrieval Journal*, 23(4):411–448, 2020.

[44] Adam Wasilewski. Introduction to the personalization in e-commerce. In *Multivariant User Interfaces in E-commerce: A Practical Approach to UI Personalization*, pages 1–19. Springer, 2024.

[45] Jo-Ting Wei, Shih-Yen Lin, and Hsin-Hung Wu. A review of the application of rfm model. *African journal of business management*, 4(19):4199, 2010.

[46] Roung-Shiunn Wu and Po-Hsuan Chou. Customer segmentation of multiple category data in e-commerce using a soft-clustering approach. *Electronic Commerce Research and Applications*, 10:331–341, 05 2011.

[47] Fangyu Zhou, Zuraidah Binti Sulaiman, Lujian Wang, Zining Yi, Meng Yuan, Wan Su, and Zhiyong Zhang. The impact of price sensitivity and ethical consumption on millennial utilitarian consumer behavior. *International Journal of Operations and Quantitative Management*, 28(3):16–31, 2022.