Eesmärk on hinnata, kui tõenäoline on, et klient ostab teatud kaupa või vähemalt kategooriat. Veelgi huvitavam eesmärk oleks aga mitte piirduda ainult konkreetse kauba või kategooria ostmise tõenäosuse hindamisega, vaid ennustada, kas klient üldse midagi ostab ning mida täpsemalt. Veelgi keerukam ja põnevam lähenemine oleks luua lahendus, mis suudab seda kõike teha reaalajas.

Miks see oluline on? Selline algoritm oleks suureks väärtuseks erinevatele otsingumootoritele ja reklaamiplatvormidele. Näiteks võimaldaks see reklaamijatel paremini hinnata kliendi potentsiaalseid ostueelistusi ja seeläbi otsustada, kui palju nad on valmis maksma reklaami eest. Samuti mängib selline hinnang olulist rolli reklaamikohtade jaotamisel oksjonipõhistes süsteemides, nagu Google AdSense ja Meta Ads. Kui platvorm suudab täpselt hinnata kliendi huve ja ostuvalmidust, võib see oluliselt mõjutada oksjonite tulemust ning reklaamikampaaniate efektiivsust.

**Eesmärk 1**

**Ennustada ostutõenäosus ühe kauba kohta:**  
Arendada mudel, mis suudab hinnata, kui tõenäoline on, et konkreetne klient ostab konkreetse toote. See nõuab kasutaja ajalooliste ostuandmete, käitumismustrite ning toote metaandmete analüüsi. Tulemuseks oleks protsentuaalne või skooritud hinnang iga kauba ostutõenäosuse kohta.

**Eesmärk 2**

**Ennustada, mida klient ostab:**  
Luua süsteem, mis prognoosib kliendi järgmist ostu, pakkudes välja ühe või mitu toodet/kategooriat. See eesmärk eeldab masinõppemudelit, mis integreerib kasutaja ajalugu, sarnaste kasutajate andmeid. Tulemuseks oleks personaliseeritud soovitusloend, mis prioriseerib tõenäolisemad valikud.

**Eesmärk 3**

**Testida, kas eesmärke 1 ja 2 on võimalik saavutada reaalajas:**  
Katsetada, kas eespool mainitud ennustused saab teha nii, et need toimivad reaalajas kasutaja tegevuse põhjal. See hõlmab järgmisi komponente:

* **Andmetöötluse optimeerimine:** Kasutajaandmete ja -käitumise reaalajas kogumine ja töötlemine.
* **Mudelite reaalajalisus:** Ennustuse täpsuse ja kiiruse balanseerimine.

Reaalajas ennustuste edukuse hindamiseks peaks mõõtma reageerimisaega, täpsust ja süsteemi skaleeritavust.

Eesmärk 1:  
Kauba (või kategooria) ostutõenäosuse ennustamiseks sobib ilmselt kasutada lihtsat logistilist regressiooni. Küsimuseks jääb, millised parameetrid ja mis kujul selleks valida. Esimene idee oli teha iga kategooria (ja selle alamkategooria) jaoks eraldi muutuja, mis sisaldaks arvulist väärtust, kui palju klient on seda kategooriat vaadanud või ostnud. Selles lähenemises on aga kolm probleemi:

1. Kategooriaid on palju ja nende alamkategooriaid veelgi rohkem. Kui teha muutuja iga kategooria ja alamkategooria kohta, tekib kordades rohkem parameetreid.
2. Alamkategooriate parameetrid hakkavad korreleeruma emakategooriaga, mis põhjustab multikollineaarsust.
3. Kategooriate lisamisel tuleb terve mudel uuesti treenida, mis ei ole praktiline.

Plaan:

1. 1. ja 2. probleemi lahendamiseks: Kasutada PCA-meetodit klasterdamise jaoks, et vähendada parameetrite arvu ja leevendada osaliselt multikollineaarsuse probleemi. PCA abil saab vähendada kõrgedimensiooniliste kategooriate ja alamkategooriate arvu, säilitades samas nende kõige olulisema info klasterdamiseks.
2. 3. probleemi osas: Eeldame antud juhul, et uusi kategooriaid ei lisandu, et hoida mudeli treenimisprotsess lihtsamana.

Plaanis on kasutada LASSO-regressiooni (lasso + kant saab korraga) ostutõenäosuse ennustamiseks. LASSO võimaldab valida kõige olulisemad tunnused, eemaldades vähem olulised, ning aitab multikollineaarsust veelgi efektiivsemalt leevendada, säilitades samal ajal algsete tunnuste tõlgendatavuse.

Veel üks viis ostutõenäosuse täpsemaks hindamiseks on proovida klassifitseerida kliente lähtuvalt nende käitumisest. Klastrite loomise eesmärk on tuvastada sarnaste ostumustritega kliendigruppide omadused ning kasutada seda teavet ennustuste parandamiseks. Klastrite kasutamine ostutõenäosuse mudelis võib toimuda kahel viisil:

1. Ühine mudel koos klastri dummy-muutujatega:  
   Selle lähenemise korral kasutatakse ühte logistilist regressioonimudelit, kuhu lisatakse (klastrite arv - 1) dummy-muutujad, mis tähistavad iga klastri kuuluvust. See on näide fikseeritud efektidega mudelist, kus klastrite mõju arvestatakse selgelt määratletud muutujatega.
2. Eraldi mudel iga klastri jaoks:  
   Siin luuakse iga klastri jaoks eraldi mudel, mis on kohandatud konkreetse kliendigruppi iseloomustavate omaduste põhjal. Seda lähenemist võib käsitleda fikseerimata efektidega mudelina, kuna iga klastri jaoks ehitatakse unikaalne mudel, mis võimaldab suuremat paindlikkust ja täpsust.

Antud töö eesmärk on testida mõlemat meetodit ning võrrelda nende tulemusi. Lisaks soovitakse uurida, kuidas PCA ja LASSO kasutamine mõjutab ennustuste täpsust. Selline lähenemine annab parema ülevaate, kas fikseeritud või fikseerimata efektidega mudel sobib paremini ostutõenäosuse hindamiseks, ning milline meetod on kõige efektiivsem antud andmestiku ja konteksti puhul.

Kommentaar andmete kohta seoses ostu ajaga:

Toorandmetes on kasutajate käitumine jaotatud sessioonideks. Näiteks, kui kasutaja logib e-poe lehele, vaatab kaupa 1, kaupa 2 ja kaupa 3 ning ostab lõpuks kauba 2, siis see kõik kuulub ühe sessiooni alla. Sellises sessioonis esineb antud juhul 4 rida: 1 vaatamise rida iga kauba kohta ja 1 ostmise rida kauba 2 kohta. (Tegelikult on andmetes olemas ka read, mis kajastavad korvi lisamist, kuid antud töös seda ignoreeritakse.)

Põhjus, miks 3. eesmärgiks on teha hinnang reaalajas, on see, et teoorias võib sellises näites tekkida olukord, kus klient uuendab brauseris lehte või liigub sessiooni jooksul erinevate vaadete vahel, mille tulemusena võib genereeruda mitu reklaami. Iga reklaami kuvamise jaoks on vaja reaalajas hinnangut, et otsustada, milline reklaam kasutajale näidata.

Antud töös keskendutakse esmalt sellele, et hinnata, millist kaupa kasutaja ostab, alates sessiooni algusest. Selle lähenemise puhul agregeeritud andmetes vastab üks vaatlus iga sessiooni algusele, ning eesmärk on prognoosida, mida ja kas kasutaja ostab selle sessiooni jooksul (kusjuures sessiooni sees toimunud tegevuste andmed ei ole ennustamise hetkeks teada).

Lisaks võiks eraldi välja töötada mudeli, mis ennustab, kas kasutaja ostab teatud kauba järgmise ööpäeva, nädala või kuu jooksul. See võimaldaks luua pikemaajalisi prognoose ja laiendada analüüsi ulatust.

Vaalitud kategoriia:

Electronics.smartphone – neid andmetes palju, brändid on tuntud, hea hinna varieeruvus

**MIS ON TEHTUD:**

**Andmete agrigeerimine**:  
Faili alguses saab valida, millist andmefaili kasutada ning millist kaupa või kategooriat analüüsitakse või mille kohta tehakse ennustus. See võimaldab töötluse paindlikkust ja andmete analüüsimist vastavalt vajadusele.

**NB!**:  
Kui tegemist on suure toorandmete failiga, võib skripti töötlemine võtta pika aja, kuna andmete eeltöötlus ja agrigeerimine on aja- ja ressursimahukad. Selle vältimiseks võiks kaaluda andmete eelnevat filtreerimist või allpool esitatud protsesside optimeerimist.

PCA.r – väga algeline

Ilus andmevisualiseerimine:

<https://www.kaggle.com/code/tshephisho/ecommerce-behaviour-using-xgboost>