Transformacija podataka

Mentor:
doc. dr Aleksandar Stanimirović

Student:
Katarina Stanojković 1773

Sadržaj

01

Uvod

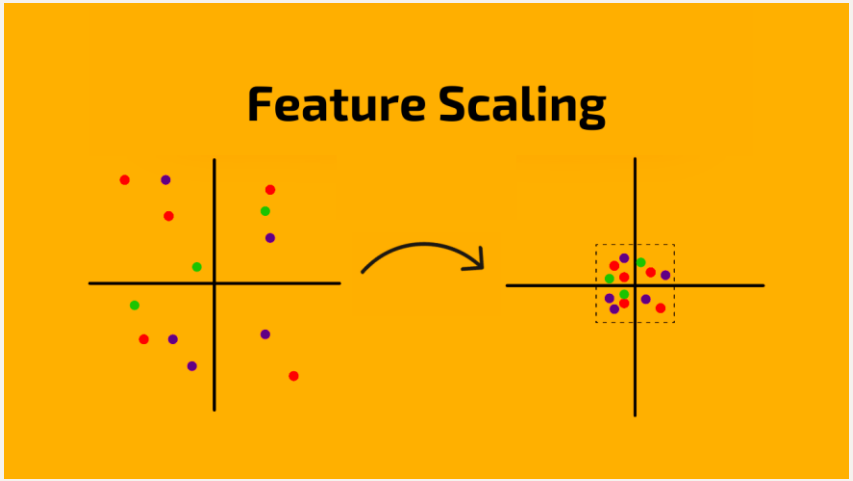
- Kvalitet ulaznih podataka je presudan za uspeh procesa učenja
- Transformacija podataka podrazumeva pretvaranje sirovih podataka u format prikladniji za analizu i modeliranje
- Cilj rada je pružiti sveoubuhvatan pregled tehnika transformacije podataka

02

Skaliranje

Skaliranje podataka

- Ključan korak u pripremi numeričkih podataka
- Transformiše vrednosti u uporedive opsege
- Omogućava ravnotežan tretman svih atributa



Standardizacija (Z-score)

- Pretvara podatke da imaju srednju vrednost 0 i standardnu devijaciju 1
- Korisna za normalno distribuirane podatke i modele zasnovane na udaljenosti
- Prednost: eliminacija uticaja različitih jedinica merenja
- Nedostatak: osetljivost na outlier-e

Min-Max skaliranje

- Prilagođava vrednosti unutar raspona 0-1
- Prednost: lakša interpretacija i ograničen raspon vrednosti
- Nedostatak: velika osetljivost na outlier-e

Max-Abs skaliranje

- Transformiše vrednosti u opsegu od -1 do 1 koristeći maksimalnu apsolutnu vrednost
- Očuvava pozitivne i negativne vrednosti i retkost podataka
- Prednost: efikasan za retke matrice
- Nedostatak: osetljivost na outlier-e

Robust skaliranje

- Skalira podatke koristći opseg između prvog i trećeg kvartila
- Prednost: Manja osetljivost na outlier-e
- Važno je proceniti veličinu i broj ekstremnih vrednosti pre primene

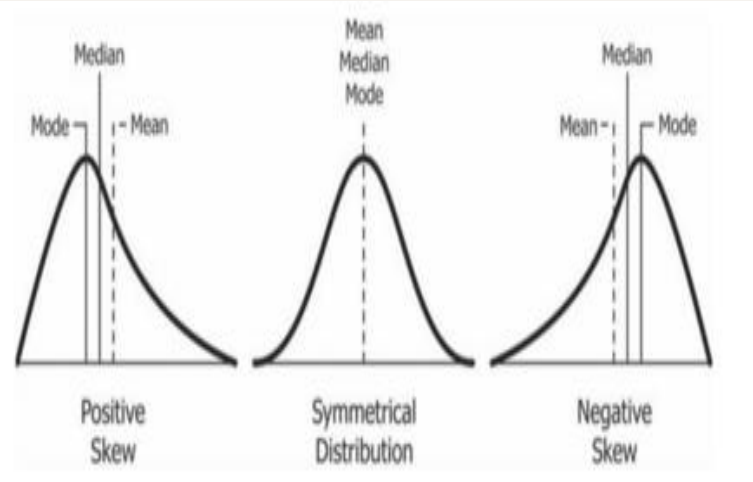
03

Transformacije koje menjaju raspodelu vrednosti

Transformacije koje menjaju raspodelu

- Tehnike koje modifikuju distribuciju podataka
- Tipovi asimetrije:

1. Pozitivna asimetrija
2. Negativna asimetrija



Logaritamska transformacija

- Svi podaci moraju biti pozitivni
- Približava raspodelu normalnoj distribuciji
- Kompresuje velike vrednosti smanjujući njihov uticaj

Box-Cox transformacija

- Približava distribuciju podataka normalnoj raspodeli korišćenjem parametra λ
- Smanjuje šum u podacima
- Smanjuje asimetriju podataka
- Zahteva pozitivne vrednosti
- Izbor parametra λ je ključan za postizanje normalnosti

Yeo-Johnson transformacija

- Slična Box-Cox transformaciji ali podržava i negativne vrednosti
- Smanjuje asimetriju
- Izbor parametra λ je ključan

Kvantilna transformacija

- Mapira vrednosti promenljive na uniformnu ili normalnu raspodelu koristeći rangiranje podataka
- Otporna na outlier-e
- Relativni položaj vrednosti ostaje nepromenjen

04

Enkodiranje kategoričkih podataka

Enkodiranje kategoričkih podataka

- Proces pretvaranja kategoričkih varijabli u numeričke
- Podela kategoričkih podataka

1. Nominalni – bez prirodnog redosleda (npr. boja automobila)
2. Ordinalni – sa prirodnim redosledom (npr. nivo obrazovanja)

Enkodiranje kategoričkih podataka

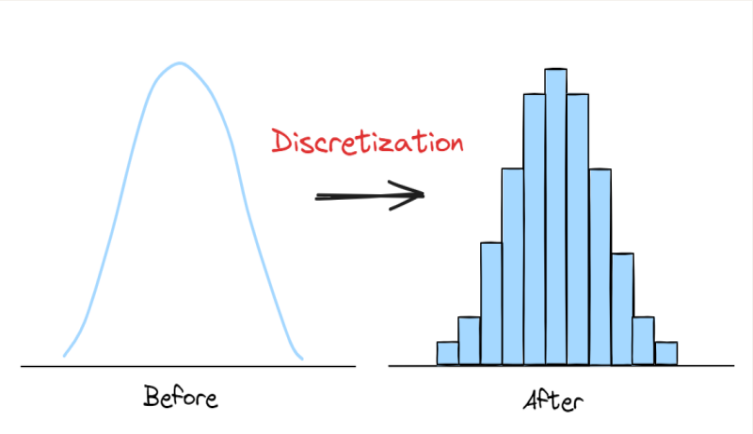
1. One-Hot enkodiranje: konvertuje svaku kategoriju u zasebnu binarnu kolonu
2. Dummy enkodiranje: kreira N-1 binarnih kolona za N kategorija
3. Label enkodiranje: dodeljuje jedinstvenu celobrojnu vrednost svakoj kategoriji
4. Binarno enkodiranje: predstavlja kategorije pomoću binarnih cifara kombinujući vrednosti one-hot i label enkodiranja
5. Count and Frequency enkodiranje: dodeljuje numeričke vrednosti kategorijama na osnovu njihove učestalosti
6. Target enkodiranje: dodeljuje kategorijama vrednosti na osnovu prosečne vrednosti ciljne varijable
7. Effect enkodiranje: dodeljuje kategorijama vrednosti 1,0 i -1
8. Feature hashing: mapira kategorije na fiksan broj numeričkih kolona pomoću hash funkcije

05

Diskretizacija podataka

Diskretizacija podataka

- Proces pretvaranja kontinuiranih varijabli u diskretne intervale
- Smanjuje uticaj outlier-a



Podela u intervale jednake širine

- Deljenje opsega vrednosti na k intervala iste veličine
- Jednostavna implementacija i ravnomerno raspoređivanje podataka
- Ne uzima u obzir distribuciju podataka; može stvoriti prazne ili retke binove

Podela na intervale sa jednakom frekvencijom

- Deljenje varijable na intervale tako da svaki sadrži približno isti broj instanci
- Proizvodi uravnotežene intervale
- Može izobličiti distribuciju podataka

Diskretizacija korišćenjem klasterizacije

- Korišćenje algoritma poput K-means za grupisanje sličnih vrednosti u klastere
- Otkriva prirodne grupe u podacima
- Izbor broja klastera može značajno uticati na rezultate

Diskretizacija korišćenjem stabla odlučivanja

- Automatsko deljenje kontinuiranih varijabli u intervale tokom procesa učenja stabla odlučivanja
- Optimalne tačke preseka se identifikuju automatski na osnovu strukture podataka
- Specifična za klasifikacione zadatke

Chi Merge

- Nadgledana metoda koja koristi chi-square test za grupisanje sličnih intervala.
- Optimizuje intervale koristeći informacije o ciljnoj varijabli, efikasno razdvaja klase.
- Specifična za situacije kada je ciljna varijabla diskretna; zahteva sortiranje vrednosti i iterativno spajanje intervala.

06

Rad sa outlier-ima

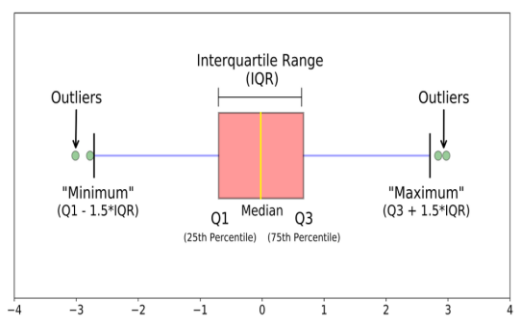
Rad sa outlier-ima

- Outlier-i su tačke koje značajno odstupaju od većine podataka i mogu iskriviti distribuciju utičući negativno na performanse modela mašinskog učenja
- Tipovi odstupanja:

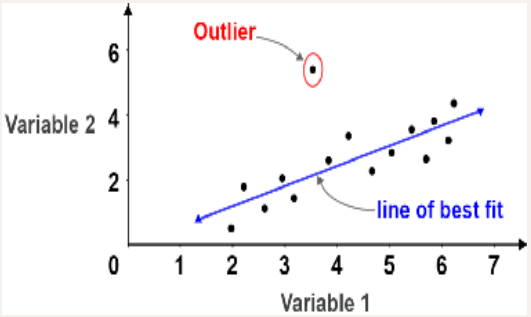
1. Globalno odstupanje
2. Lokalno odstupanje
3. Univarijantna odstupanja
4. Multivarijantna odstupanja

Tehnike vizuelizacije za detekciju outlier-a

- Box plot



- Scatter plot



- Histogram

Metode detekcije outlier-a

- Z-Score: Mera koja pokazuje koliko je standardnih devijacija neka vrednost udaljena od proseka; vrednosti sa |z| > 3 se smatraju outlier-ima
- Interquartile Range (IQR): Identifikacija tačaka koje leže van 1.5x IQR od Q1 ili Q3.
- Percentile: Klasifikacija vrednosti ispod 1. ili iznad 99. percentila kao outlier-a.
- DBSCAN: Klasterizacija zasnovana na gustini koja automatski detektuje outlieri kao tačke koje nemaju dovoljno gustine okolnih tačaka.
- Isolation Forest: Algoritam koji izoluje outlier-e kroz ansambl binarnih stabala odlučivanja, gde outlieri imaju kraće puteve izolacije.

Tehnike rukovanja outlier-ima

- Uklanjanje outliera (Trimming): Direktno uklanjanje outlier-a iz skupa podataka.
- Kraćenje (Capping): Postavljanje vrednosti iznad ili ispod određenih pragova na maksimalne ili minimalne dozvoljene vrednosti.
- Transformacija: Primena matematičkih transformacija (npr. logaritamska) za smanjenje uticaja outlier-a.
- Imputacija: Zamena outliera sa odgovarajućim vrednostima kao što su medijana ili prosek.
- Vinzorizacija: Postavljanje ekstremnih vrednosti na određene pragove, smanjujući njihov uticaj bez potpunog uklanjanja.
- Korišćenje robusnih modela: Primena modela manje osetljivih na outlier-e, kao što su Random Forest ili modeli sa median-based loss funkcijama.
- Podela podataka (Partitioning): Razdvajanje skupa podataka na delove sa i bez outlier-a radi zasebne analize ili modeliranja.

07
Konstrukcija atributa

Konstrukcija atributa

- Proces kojim se generišu nove karakteristike na osnovu postojećih podataka ili domenskog znanja
- Tipovi karakteristika:

1. Interakcione karakteristike – kreiranje kombinovanjem dve ili više postojećih karakteristika
2. Polinomske karakteristike – kreiranje karakteristika podizanjem vrednosti na različite stepene
3. Vremenske karakteristike – kreiranje karakteristika iz podataka koji uključuju datume ili vremenske oznake

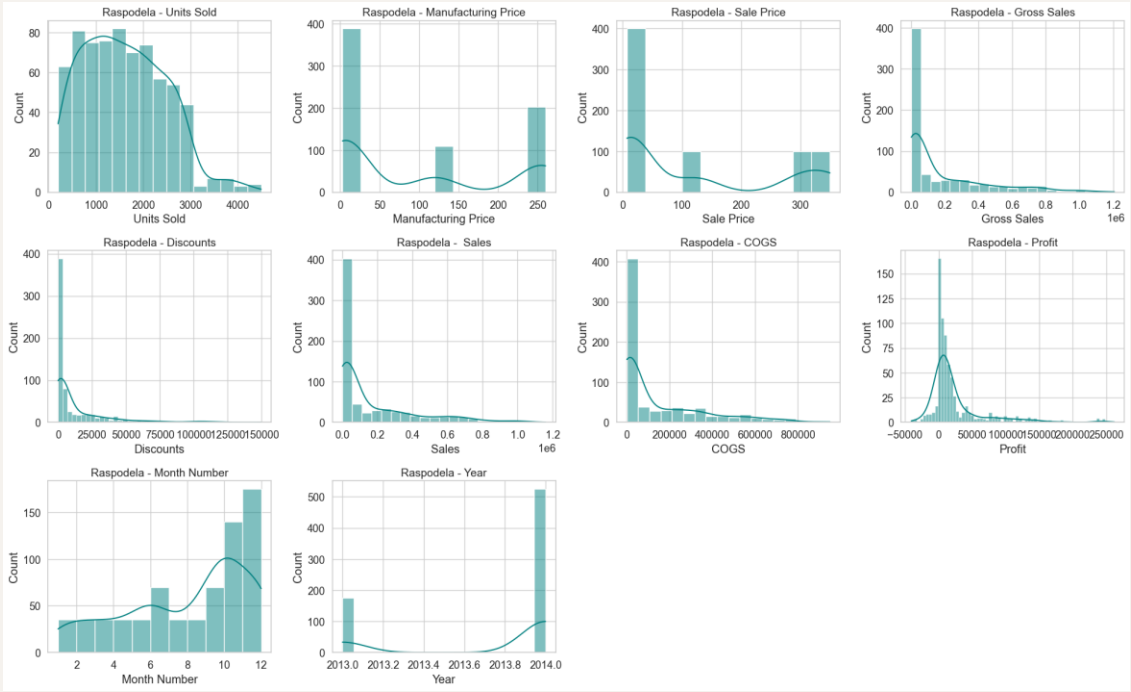
08

Praktični deo rada

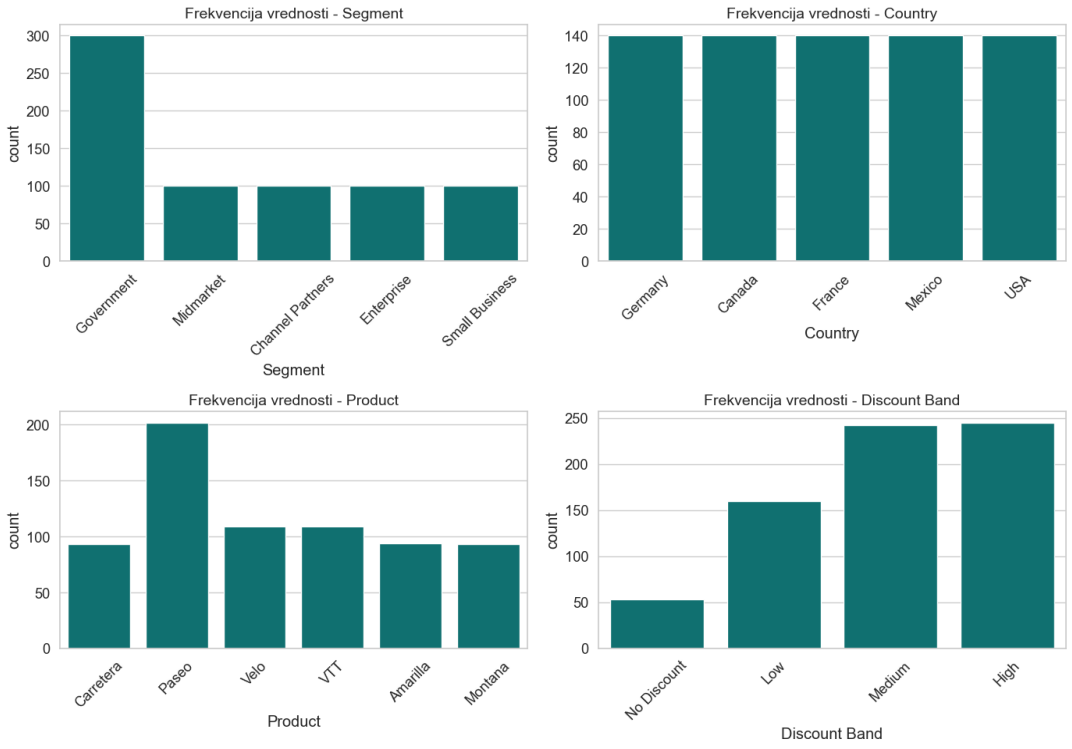
Analiza podataka

- 700 redova i 16 kolona
- Tipovi podataka su numerički tekstualni i datetime
- Nema duplikata
- NaN vrednosti u koloni discount band popunjene sa No Discount

Vizuelizacija numeričkih podataka



Vizuelizacija kategoričkih podataka



Primena algoritma za klasifikaciju

- Algoritam koji je korišćen je SVM
- Nad osnovnim podacima daje tačnost od samo 19%
- Loše prepoznaje klase low i medium
- Visoka preciznost, ali nizak odziv za klasu high
- Klasa No Discount ima visok odziv ali nisku preciznost

Enkodiranje kategoričkih atributa

- Segment: TargetEncoder na osnovu kolone sales
- Product: CountEncoder za transformaciju prema učestalosti pojavljivanja
- Country: One Hot Encoding za nominalne vrednosti
- Discount Band: Ordinal Encoding za očuvanje prirodnog poretka
- Rezultat tačnosti: Povećana na 32 %

Skaliranje podataka

- StandardScaler: Kolone 'Units Sold', 'Manufacturing Price', 'Sale Price'.
- RobustScaler: Kolone 'Gross Sales', 'Discounts', 'Sales', 'COGS', 'Profit'.
- MinMaxScaler: Kolone 'Month Number', 'Year'.
- Rezultat tačnosti: Povećana na 49%.

Transformacije koje menjaju raspodelu podataka

- Logaritamska transformacija: Kolone 'Manufacturing Price' i 'Sales'.
- Box-Cox transformacija: Kolona 'COGS'.
- Kvantilna transformacija: Kolone 'Gross Sales' i 'Discounts'.
- Yeo-Johnson transformacija: Kolona 'Profit'.
- Rezultat tačnosti: 22% (neznatno poboljšanje).
- Kombinacija sa skaliranjem: Povećana tačnost na 61%.

Diskretizacija podataka

- Jednake širine: Kolona 'Units Sold'.
- Jednake frekvencije: Kolona 'Manufacturing Price'.
- KMeans diskretizacija: Kolona 'Gross Sales'.
- Decision Tree diskretizacija: Kolona 'Profit'.
- Rezultat tačnosti: 29%.
- Kombinacija sa skaliranjem: Povećana tačnost na 46%.

Detekcija outlier-a

- Z-score: Kolona 'Gross Sales'.
- IQR: Kolona 'Discounts'.
- Percentile: Kolona 'Sales'.
- DBSCAN: Kolona 'COGS'.
- Isolation Forest: Kolona 'Profit'.
- Pristup: Uklanjanje detektovanih outlier-a.
- Rezultat tačnosti: Povećana na 33%.

Konstrukcija atributa

- Avg Sales per Month: Prosečna vrednost Sales po mesecima.
- Manufacturing Cost: Proizvod Manufacturing Price i Units Sold.
- Net Sales: Razlika Sales i Discounts.
- Total Revenue: Proizvod Units Sold i Sale Price.
- Day of Week: Dan u nedelji iz atributa Date.
- Is Weekend: Da li je dan vikend.
- Quarter: Kvartal na osnovu datuma.
- Rezultat tačnosti: 26%.

Hvala na pažnji!