Analys av E-handel Kundedata

En end-to-end analys för kundsegmentering och försäljningsförutsägelse



Maya Sobh

EC Utbildning

Projekt i Data Science

2024-10

# Abstract

En kort sammanfattning över ditt arbete och de viktigaste resultaten skrivet på engelska, cirka 5 meningar totalt.

This project presents an end-to-end analysis of e-commerce customer data, focusing on customer segmentation and sales prediction. Using a combination of data visualization and machine learning techniques, the study aims to uncover patterns in customer behavior and predict future purchase amounts. Key findings include the identification of specific customer segments based on age, purchase frequency, and preferred product categories. The predictive model, built using algorithms such as XGBoost and Random Forest, achieved a moderate accuracy, highlighting areas for further data feature enhancement. This analysis provides valuable insights for improving targeted marketing and enhancing customer retention strategies.

**Skapas automatiskt i Word genom att gå till Referenser > Innehållsförteckning.**

Innehållsförteckning

[1 Inledning 1](#_Toc176961624)

[1.1 Underrubrik – Exempel 1](#_Toc176961625)

[2 Teori 2](#_Toc176961626)

[2.1 Exempel: Dataanalys och Visualisering 2](#_Toc176961627)

[2.1.1 Exempel: Regression och Prediktiva Modeller 2](#_Toc176961628)

[2.1.2 Exempel: Segmentering och Kategorisering 2](#_Toc176961629)

[3 Metod 3](#_Toc176961632)

[3.1 Datainsamling 3](#_Toc176961633)

[4 Resultat och Diskussion 4](#_Toc176961635)

[5 Slutsatser 5](#_Toc176961636)

[6 Självutvärdering 6](#_Toc176961637)

[Appendix A 7](#_Toc176961638)

[Källförteckning 8](#_Toc176961639)

# Inledning

Ge en bred bakgrund till ditt arbete och varför det är relevant. Smalna successivt av och koppla bakgrunden till detta arbete. Fortsätt med syfte och frågeställning (syfte och frågeställning kan du ha i en egen underrubrik om du önskar det).

Syftet med denna rapport är \_\_\_\_\_\_\_\_\_\_\_\_, för att uppfylla syftet så kommer följande frågeställning(ar) att besvaras:

1. \_\_\_\_\_\_\_
2. \_\_\_\_\_\_\_

Den snabba tillväxten inom e-handel innebär att företag nu samlar in stora mängder data om sina kunders köpbeteende. Genom att analysera denna data kan företag identifiera mönster och trender som hjälper dem att förstå kundernas behov och förbättra deras upplevelse. För e-handelsföretag är insikter om kundbeteende avgörande för att kunna utforma riktade marknadsföringsstrategier och öka kundlojaliteten.

Denna rapport syftar till att analysera kunddata från en e-handelsplattform för att hitta värdefulla insikter och göra förutsägelser om framtida köpbeteenden. Genom att använda både dataanalys och maskininlärning kommer studien att belysa de viktigaste faktorerna som påverkar köpmönster och utvärdera olika metoder för segmentering och förutsägelse av försäljning.

**Syfte och frågeställning**

Syftet med denna rapport är att analysera kunddata för att förstå mönster i kundernas köpvanor och att använda maskininlärningsmodeller för att förutsäga framtida köp. För att uppfylla syftet kommer följande frågeställningar att besvaras:

1. Vilka faktorer påverkar mest kundens totala köpsumma?
2. Hur kan vi segmentera kunderna baserat på deras köpmönster?
3. Vilken noggrannhet kan maskininlärningsmodeller uppnå i förutsägelsen av framtida kundköp?

# Teori

## Detta projekt bygger på flera grundläggande begrepp inom dataanalys och maskininlärning som är relevanta för analysen av e-handelskunder. Här beskrivs de viktigaste teoretiska komponenterna som används för att förstå och förutsäga kundbeteende.

## **2.1 Dataanalys och Visualisering**

## Dataanalys och visualisering är centrala metoder för att förstå mönster och samband i stora datamängder. I detta projekt används bibliotek som pandas för datahantering och matplotlib och seaborn för att skapa visualiseringar som hjälper till att upptäcka trender och avvikelser i kundbeteendet.

## **2.2 Regression och Prediktiva Modeller**

## Regression används för att modellera och analysera relationen mellan en beroende variabel (t.ex. Total Purchase Amount) och en eller flera oberoende variabler. Genom att använda regressionsmodeller som XGBoost försöker projektet förutsäga kundernas framtida köpbeteende.

## **2.3 Segmentering och Kategorisering**

## För att bättre förstå olika kundgrupper använder projektet kundsegmentering. Denna metod delar upp kunder i grupper baserade på likheter, vilket hjälper till att anpassa marknadsföringsstrategier för olika segment.

# Metod

Hur har du genomfört ditt arbete? Exempelvis, hur har datan erhållits?

## Datainsamling

Exempel på en rubrik som är vanligt förekommande som läsaren förstår vart datan kommer ifrån.

Datan för detta projekt hämtades från plattformen **Kaggle** och inkluderades som en **ZIP-fil**. Efter nedladdningen extraherades filen för att ge två **CSV-filer** innehållande information om kundernas köpbeteende och demografiska data. Filerna laddades sedan in i en **SQLite-databas** för att möjliggöra effektiv hantering och SQL-frågor under analysen. Verktyg som pandas användes för att läsa och hantera data, medan matplotlib och seaborn användes för att visualisera resultaten.

# Resultat och Diskussion

|  |  |
| --- | --- |
| **RMSE för olika modeller** | |
| Root Mean Squared Error (RMSE) | 1438.34 |
| Mean Absolute Error (MAE): | 1245.52 |
| R-squared (R2): | 0.0029 |

De bästa parametrarna för modellen är:

* **learning\_rate**: 0.01
* **max\_depth**: 3
* **n\_estimators**: 200
* **Tolkning av Resultat**
* **RMSE och MAE** indikerar hur nära modellens prediktioner är de verkliga värdena. Här är värdena höga, vilket antyder att modellen fortfarande har begränsad prediktiv förmåga.
* **R2-värde** på 0.0029 innebär att modellen endast förklarar en väldigt liten andel av variationen i data, vilket kan bero på att fler eller mer informativa variabler behövs för att förbättra noggrannheten.

# Slutsatser

Här besvarar du bl.a. frågeställningarna.

Detta projekt har genomfört en analys av kunddata från en e-handelsplattform för att identifiera mönster i kundernas köpbeteende och förutsäga framtida köp. Målet var att förstå vilka faktorer som påverkar kundens totala köpsumma samt att utveckla en maskininlärningsmodell för att stödja försäljningsförutsägelser.

1. **Besvarade frågeställningar**
2. **Vilka faktorer påverkar mest kundens totala köpsumma?**
   * Analysen visade att variabler som produktpris, kvantitet och ålder spelar en viss roll i att påverka den totala köpsumma. Däremot var förklaringsgraden för modellen relativt låg, vilket antyder att ytterligare faktorer kan vara avgörande för att förutsäga köpbeteendet med högre noggrannhet.
3. **Hur kan vi segmentera kunderna baserat på deras köpmönster?**
   * Kundsegmentering utfördes baserat på ålder och totala köpsummor, vilket skapade grupper av kunder med liknande köpmönster. Dessa segment kan användas för att utveckla mer riktade marknadsföringsstrategier.
4. **Vilken noggrannhet kan maskininlärningsmodeller uppnå i förutsägelsen av framtida kundköp?**
   * Den bästa modellen, XGBoost, uppnådde en RMSE på cirka 1438, vilket indikerar att modellen har en viss prediktiv förmåga men behöver förbättras. Modellen kunde förklara en mycket liten del av variationen (R² nära noll), vilket innebär att modellen har begränsad förmåga att ge exakta förutsägelser i sin nuvarande form.
5. **Slutsats**

Sammanfattningsvis har projektet lyckats analysera e-handelskunder och ge en grundläggande förståelse för kundbeteende och segmentering. Trots det begränsade R²-värdet i de prediktiva modellerna erbjuder resultaten insikter som kan ligga till grund för vidare analys och modellförbättring. För att öka modellens noggrannhet rekommenderas att inkludera fler variabler och experimentera med andra maskininlärningsmetoder.

# Självutvärdering

1. Utmaningar du haft under arbetet samt hur du hanterat dem.

Under projektet hade jag utmaningar med att justera modellens parametrar och hantera viss datakvalitet. För att lösa detta använde jag GridSearchCV för att optimera parametrarna och gjorde datarensning för att förbättra noggrannheten. Detta hjälpte mig att få bättre resultat i analysen.

1. Vilket betyg du anser att du skall ha och varför.

Jag tycker att jag förtjänar ett högt betyg eftersom jag arbetat självständigt, löst problem effektivt och levererat en noggrann analys.

1. Något du vill lyfta fram till Antonio?

Nej, tack.

# Appendix A

* 1. **Installation av Bibliotek och Dataimport**:
* Kod för att installera nödvändiga bibliotek, importera dem och ladda upp data från en ZIP-fil.

# Installera nödvändiga bibliotek om de inte redan är installerade

!pip install pandas dash bokeh ipywidgets matplotlib seaborn

# Importera de nödvändiga biblioteken

import pandas as pd

import sqlite3

import matplotlib.pyplot as plt

import seaborn as sns

import zipfile

* 1. **Extrahera och Förbereda Data:**
* Kod för att extrahera data från ZIP-filen och ladda den in i en SQLite-databas.

# Extrahera filen till en temporär mapp och läs in data

with zipfile.ZipFile(zip\_file\_path, 'r') as zip\_ref:

zip\_ref.extractall('extracted\_data')

df = pd.read\_csv('extracted\_data/ecommerce\_customer\_data\_large.csv')

* 1. **Datahantering och SQL-frågor**:
* Skapande av SQL-frågor för att analysera de mest sålda produktkategorierna direkt från databasen.

# Skapa och anslut till en SQLite-databas

con = sqlite3.connect('ecommerce\_data.db')

df.to\_sql('customers', con, if\_exists='replace', index=False)

# SQL-fråga för att hämta de mest sålda produktkategorierna

query = """

SELECT [Product Category] AS Product\_Category, SUM([Total Purchase Amount]) AS Total\_Sales

FROM customers

GROUP BY [Product Category]

ORDER BY Total\_Sales DESC

LIMIT 10

"""

top\_products = pd.read\_sql\_query(query, con)

* 1. **Modellutveckling och Utvärdering**:
* Exempel på koden för att konfigurera och träna XGBoost-modellen med GridSearchCV.

from xgboost import XGBRegressor

from sklearn.model\_selection import GridSearchCV

# Definiera parametergrid och använd GridSearchCV

param\_grid = {'n\_estimators': [100, 200, 300], 'max\_depth': [3, 5, 7], 'learning\_rate': [0.01, 0.1, 0.2]}

grid\_search = GridSearchCV(XGBRegressor(random\_state=42), param\_grid, cv=5, scoring='neg\_mean\_squared\_error')

grid\_search.fit(X\_train, y\_train)

# Källförteckning

1. Kaggle. (n.d.). *E-commerce Customer Data For Behavior Analysis*. Hämtad från <https://www.kaggle.com>

<https://www.kaggle.com/datasets/shriyashjagtap/e-commerce-customer-for-behavior-analysis>

*(Används som huvuddataset för analys av kundbeteende och produktförsäljning.)*

1. Agile Alliance. (2001). *Principer bakom det Agila Manifestet*. Hämtad från <https://agilemanifesto.org/principles.html>  
   *(Används för att strukturera arbetet enligt agila principer i projektets utvecklingsprocess.)*
2. Python Software Foundation. (n.d.). *pandas Dokumentation*. Hämtad från <https://pandas.pydata.org/docs/>

*(Används för att bearbeta och analysera data från CSV-filen med hjälp av Pandas biblioteket.)*

1. Dash Plotly. (n.d.). *Dash Användarguide och Dokumentation*. Hämtad från <https://dash.plotly.com/>

*(Används för att skapa interaktiva webbapplikationer och visualiseringar av försäljningsdata.)*

1. Bokeh Dokumentation. (n.d.). *Bokeh Användarguide*. Hämtad från <https://docs.bokeh.org/en/latest/>

*(Används för att skapa interaktiva diagram och visualiseringar.)*

1. SQLite. (n.d.). *SQLite Dokumentation*. Hämtad från: <https://sqlite.org/docs.html>

*(Används för att skapa och hantera databasen som lagrar analysdata och för att utföra SQL-frågor.)*