Segmentarea Clientilor pentru E-commerce folosind Machine Learning

1. Introducere

Scopul acestui proiect este de a prezice dacă un client va finaliza o achiziție pe baza comportamentului său de navigare. Am analizat interacțiunile clienților pe o platformă de e-commerce, utilizând caracteristici precum paginile vizualizate, durata sesiunii și ratele de abandon, pentru a înțelege mai bine intențiile de cumpărare. Această analiză va ajuta companiile să identifice clienții potriviți și să îmbunătățească strategiile de marketing.

2. Descrierea Setului de Date

- **Nume Set de Date:** Online Shoppers Purchasing Intention Dataset
- **Sursă:** Kaggle - Online Shoppers Intention Dataset
- **Caracteristici:**
    - Caracteristici numerice precum "Administrative", "Informational" și "Product Related".
    - Caracteristici legate de timp precum "Administrative Duration" și "Product Related Duration".
    - Rate de abandon, rate de ieșire, valori ale paginilor și indicatori pentru zile speciale.
    - Caracteristici categorice inclusiv luna vizitei și activitatea în weekend.
- **Ținta:** Variabilă binară care indică dacă o sesiune a dus la o achiziție ("Revenue").

3. Algoritmi de Machine Learning Implementați

Am evaluat mai mulți algoritmi de clasificare pentru a analiza performanța acestora:

3.1 Arbore de Decizie (Decision Tree)

- **Acuratețe:** 89,94%
- **Precizie:** 71,34%
- **Recall:** 58,64%

Arborii de decizie oferă o interpretabilitate ridicată și evidențiază factorii cheie care determină comportamentul de cumpărare. Valorile preciziei și recall indică o capacitate moderată de a distinge cumpărătorii de non-cumpărători.

3.2 Pădure Aleatorie (Random Forest)

- **Acuratețe:** 90,00%
- **Precizie:** 88,00%
- **Recall:** 92,00%

Natura ensemble a algoritmului Random Forest îmbunătățește capacitatea de generalizare. Modelul a obținut cel mai bun echilibru între precizie și recall, fiind foarte potrivit pentru segmentarea clienților e-commerce.

3.3 Naive Bayes

- **Acuratețe:** 87,00%
- **Precizie:** 60,00%
- **Recall:** 55,00%

În ciuda simplității și eficienței sale, clasificatorul Naive Bayes presupune independența caracteristicilor, ceea ce poate să nu surprindă pe deplin complexitățile comportamentului clienților. Performanța sa este ușor inferioară comparativ cu modelele bazate pe arbori de decizie.

3.4 K-Nearest Neighbors (KNN)

- **Acuratețe:** 83,00%
- **Precizie:** 40,00%
- **Recall:** 14,00%

KNN surprinde similitudinile comportamentale ale clienților, dar a întâmpinat dificultăți în generalizarea pe acest set de date. Recall-ul scăzut sugerează că nu reușește să identifice mulți potențiali cumpărători, făcându-l mai puțin eficient pentru această sarcină.

3.5 Rețea Neuronală

- **Acuratețe:** 65,00%
- **Precizie:** 87,00%
- **Recall:** 36,00%

Deși rețeaua neuronală a surprins modele complexe, performanța generală a fost afectată de un recall și o acuratețe mai scăzute. Dezechilibrul dintre precizie și recall evidențiază provocările în detectarea constantă a intențiilor de cumpărare.

4. Feature Engineering

Pentru a îmbunătăți performanța modelelor, am creat caracteristici suplimentare, inclusiv:

- **Categorii de Pagini:** Gruparea paginilor în categorii precum "Pagini de Produse", "Pagini de Checkout" etc.
- **Caracteristici Temporale:** Calcularea timpului petrecut pe anumite categorii sau frecvența accesărilor.

5. Metrici de Evaluare

Pentru evaluarea performanței modelelor, ne-am concentrat pe:

- **Acuratețe:** Proporția predicțiilor corecte.
- **Precizie:** Proporția cumpărătorilor prezise corect din totalul prezis ca fiind cumpărători.
- **Recall:** Proporția cumpărătorilor reali prezise corect.
- **AUC-ROC:** Folosit pentru a evalua performanța generală a modelului și capacitatea de a distinge între clasele pozitive și negative.

6. Rezultate și Comparație

Model	Acuratețe	Precizie	Recall
Arbore de Decizie	89,94%	71,34%	58,64%
Pădure Aleatorie	90,00%	88,00%	92,00%
Naive Bayes	87,00%	60,00%	55,00%
K-Nearest Neighbors	83,00%	40,00%	14,00%
Rețea Neuronală	65,00%	87,00%	36,00%

7. Concluzii și Recomandări

- **Cel Mai Performant Model:** Random Forest a obținut cea mai mare acuratețe și recall, făcându-l cel mai eficient pentru identificarea potențialilor cumpărători.
- **Interpretabilitatea Modelului:** În timp ce Arborii de Decizie oferă informații valoroase despre comportamentul clienților, Random Forest oferă o performanță mai bună.

- **Provocări cu KNN și Rețele Neuronale:** Aceste modele au întâmpinat dificultăți din cauza naturii complexe a setului de date și a necesității unei reprezentări mai bune a caracteristicilor.
- **Lucrări Viitoare:**
    - Inginerie suplimentară de caracteristici pentru a surprinde comportamente mai nuanțate.
    - Ajustarea hiperparametrilor pentru rețelele neuronale pentru a îmbunătăți performanța.
    - Echilibrarea seturilor de date sau utilizarea tehnicilor de oversampling pentru a îmbunătăți recall-ul pentru clasele subreprezentate.

Acest proiect demonstrează importanța selecției algoritmilor și a metodelor de evaluare potrivite pentru sarcinile de segmentare a clienților, ajutând companiile să ia decizii bazate pe date pentru a optimiza strategiile de marketing.