ANALIZA INTENȚIEI DE CUMPĂRARE ÎN MEDIUL ONLINE

Proiect final Sisteme Inteligente

Ilies Oana-Elena

Facultatea de Automatica si Calculatoare

grupa 30232

ilies.io.oana@student.utcluj.ro

Profil GitHub: Oana Elena Ilies

Profil LinkedIn: Oana Elena Ilies

Abstract

Acest proiect reprezintă o metodă de învățare automată pentru predictarea intențiilor de cumpărare ale vizitatorilor pe un site de comerț electronic. Folosind un set de date care include caracteristici precum activități administrative, durate informaționale și de produs, rate de abandon și de ieșire, valori de pagină, zile speciale, lună, sisteme de operare, browser, regiune, tip de trafic, tip de vizitator și informații despre weekend, am dezvoltat și antrenat un model de rețea neuronală utilizând biblioteca Keras din TensorFlow. Modelul nostru este alcătuit din trei straturi dense, având două straturi ascunse cu funcție de activare ReLU și un strat de ieșire cu funcție de activare sigmoidă. Modelul a fost antrenat și validat pe un set de date, obținând o acuratețe de aproximativ 91% în prezicerea comportamentului de cumpărare .Procesul de antrenare a implicat optimizarea modelului folosind algoritmul Adam și funcția de pierdere binary cross-entropy, iar performanța modelului a fost evaluată atât în timpul antrenării, cât și pe setul de testare. Rezultatele demonstrează că modelul poate prezice eficient dacă un vizitator va finaliza o achiziție, oferind o unealtă valoroasă pentru afacerile de comerț electronic în îmbunătățirea strategiilor de marketing și în personalizarea experienței utilizatorilor. Acest proiect subliniază importanța învățării automate în analiza comportamentului consumatorilor și în luarea deciziilor bazate pe date.

1.Introducere

In ultimii ani, cumpărăturile online au devenit o formă de achiziție tot mai populară și reprezintă o parte semnificativă din veniturile din vânzările între business-uri și consumatori. Conform statisticilor, 69% dintre americani au cumpărat online la un moment dat, generând un venit mediu de 1804 dolari pe cumpărător online, iar 36% dintre aceștia fac achiziții online cel puțin o dată pe lună. În acest context, este crucial pentru afaceri să înțeleagă comportamentul și intențiile cumpărătorilor online pentru a-și maximiza vânzările și veniturile.

Problemele care ar trebui rezolvate în această privință includ înțelegerea modului și momentului în care cumpărătorii își cercetează și efectuează achizițiile online. De asemenea, este important să se identifice factorii care influențează aceste comportamente și să se ofere strategii de marketing și publicitate potrivite pentru a atrage și fideliza clienții.

Obiectivele propuse pentru acest proiect sunt:

- Analiza comportamentului cumpărătorilor online și identificarea factorilor cheie care influențează intenția și comportamentul acestora.
- Propunerea de strategii de marketing și publicitate pentru îmbunătățirea vânzărilor și a veniturilor în mediul online.
- Structurarea documentației pentru a oferi o analiză detaliată și soluții practice pentru optimizarea vânzărilor online.

Structura documentației acestui proiect va include:

- Pagina de titlu
- Introducere
- Fundal și context al proiectului
- Obiectivele proiectului
- Structura documentației

2.Context teoretic

a. Real-time prediction of online shoppers' purchasing intention using multilayer perceptron and LSTM recurrent neural networks

Autori: C. O. Sakar, S. Polat, Mete Katircioglu, Yomi Kastro

Publicat în: Neural Computing & Applications (2019)

Link: Real-time prediction of online shoppers' purchasing intention

Rezumat: Acest studiu investighează utilizarea perceptronului multilayer (MLP) și a rețelelor neuronale LSTM (Long Short-Term Memory) pentru predicția în timp real a intenției de cumpărare a cumpărătorilor online. Studiul folosește date despre comportamentul utilizatorilor pentru a antrena modele de învățare automată care să prezică dacă un utilizator va face sau nu o achiziție. Rezultatele au arătat că aceste metode sunt eficiente în îmbunătățirea ratei de conversie, oferind o predicție precisă a comportamentului de cumpărare pe baza sesiunilor de utilizare. Modelul MLP captează relațiile non-lineare dintre caracteristici, iar LSTM gestionează datele secvențiale, capturând dependențele temporale.

Folosind datele despre clickstream (secvența de acțiuni efectuate de un utilizator pe un site web) și de sesiune (informațiile despre sesiunea de navigare a utilizatorului pe site), este posibil să se dezvolte modele de învățare automată care să anticipeze comportamentul cumpărătorului cu o acuratețe ridicată și să fie aplicabile într-un mod extensiv și eficient.

b. Predicting Online Shopper Purchasing Behavior Using Machine Learning Algorithms

Autori: Yu-Lin Chung, Ting-Yi Chen, Cheng-Hung Wang

Publicat în: Information Processing & Management (2020)

Link: Digital Marketing Strategy: An Integrated Approach to Online Marketing

Rezumat: Această lucrare investighează utilizarea algoritmilor de învățare automată pentru a prezice comportamentul de cumpărare online al utilizatorilor. Autorii compară performanța mai multor modele de învățare automată, inclusiv arbori de decizie, mașini cu vectori suport și rețele neuronale artificiale. Rezultatele studiului oferă o perspectivă asupra eficacității diferitelor abordări de modelare pentru predicția intențiilor de cumpărare online.

c. **Online Shoppers Purchasing Intention Dataset**

Link: Online Shoppers Purchasing Intention Dataset

Rezumat: Setul de date "Online Shoppers Purchasing Intention" conține informații detaliate despre comportamentul de cumpărare online din 12,330 de sesiuni, colectate pe parcursul unui an. Acesta include 10 atribute numerice și 8 categorice, precum și atributul 'Revenue' care indică dacă o achiziție a fost efectuată sau nu. Datele includ variabile precum durata sesiunilor, numărul de pagini vizitate, rata de ieșire, rata de respingere și altele. Aceste informații sunt esențiale pentru dezvoltarea modelelor de învățare automată care pot analiza și prezice intențiile de cumpărare, permițând identificarea factorilor cheie ce influențează comportamentul de cumpărare.

1. Administrative: Acesta indică numărul de pagini administrative vizitate de către utilizator în timpul sesiunii sale pe site-ul web.
2. Administrative_Duration: Durata totală petrecută de utilizator pe paginile administrative, măsurată în secunde.
3. Informational: Numărul de pagini informative vizitate de către utilizator în timpul sesiunii sale.
4. Informational_Duration: Durata totală petrecută de utilizator pe paginile informative, măsurată în secunde.
5. ProductRelated: Numărul de pagini de produs vizitate de către utilizator în timpul sesiunii sale.
6. ProductRelated_Duration: Durata totală petrecută de utilizator pe paginile de produs, măsurată în secunde.
7. BounceRates: Rata la care utilizatorii au părăsit site-ul imediat după vizitarea unei singure pagini.
8. ExitRates: Rata la care utilizatorii au părăsit site-ul după vizitarea unei pagini, în comparație cu numărul total de vizite ale paginii.
9. PageValues: Valoarea medie a paginii, calculată ca suma veniturilor generate de pagina împărțită la numărul total de vizite ale paginii.
10. SpecialDay: Indică dacă sesiunea de navigare a utilizatorului a avut loc într-o zi specială, cum ar fi sărbători sau evenimente de vânzare.
11. Month: Luna în care a avut loc sesiunea de navigare a utilizatorului.
12. OperatingSystems: Sistemul de operare utilizat de către utilizator.
13. Browser: Browserul web utilizat de către utilizator.
14. Region: Regiunea geografică a utilizatorului.
15. TrafficType: Tipul de trafic care a adus utilizatorul pe site.
16. VisitorType: Tipul de vizitator, cum ar fi "Returning_Visitor" pentru utilizatorii care revin pe site sau "New_Visitor" pentru cei care vizitează pentru prima dată.
17. Weekend: Indică dacă sesiunea de navigare a utilizatorului a avut loc în weekend (TRUE) sau nu (FALSE).
18. Revenue: Variabilă țintă care indică dacă utilizatorul a finalizat o achiziție (TRUE) sau nu (FALSE).

d. Digital Marketing Strategy: An Integrated Approach to Online Marketing

Autori: Simon Kingsnorth

Publicat în: Kogan Page (2019)

Link: Digital Marketing Strategy: An Integrated Approach to Online Marketing

Rezumat: Cartea "Digital Marketing Strategy" de Simon Kingsnorth explorează strategii integrate de marketing digital, oferind o abordare comprehensivă pentru optimizarea prezenței online a afacerilor. Autorul accentuează importanța înțelegerii comportamentului consumatorilor online și cum aceste informații pot fi folosite pentru a dezvolta campanii de marketing personalizate. Abordările integrate discutate în carte ajută afacerile să creeze mesaje relevante și atractive pentru clienți, maximizând astfel rata de conversie și fidelizarea acestora. De asemenea, cartea oferă instrumente și tehnici pentru analizarea eficienței campaniilor și ajustarea strategiilor în funcție de datele colectate.

Aceste resurse bibliografice oferă un fundament solid pentru analizarea comportamentului de cumpărare online și dezvoltarea de strategii eficiente de marketing și predicție, esențiale pentru optimizarea vânzărilor și veniturilor.

Capitolul 4: Segmentarea și Targetarea Publicului Exemplu: În acest capitol, Kingsnorth discută despre importanța segmentării publicului pentru campaniile de marketing digital. Un exemplu concret oferit este utilizarea datelor demografice și comportamentale pentru a crea segmente de public țintă. De exemplu, o companie de îmbrăcăminte sportivă ar putea folosi date despre vârstă, sex, locație și istoricul de cumpărături online pentru a identifica segmente specifice de consumatori, cum ar fi "tineri adulți activi din mediul urban" sau "femei interesate de fitness din suburbii".

Kingsnorth subliniază că prin personalizarea mesajelor și a ofertelor pentru fiecare segment identificat, companiile pot crește semnificativ rata de răspuns și conversia. În plus, el sugerează utilizarea testelor A/B pentru a evalua eficiența diferitelor mesaje de marketing și a ajusta strategiile în funcție de rezultate. Acest exemplu arată cum segmentarea și targetarea precise pot conduce la campanii de marketing mai eficiente și la o mai bună satisfacție a clienților.

e. Understanding Online Shopper Behavior: A Meta-Analytic Review and Integration of Cognitive and Socio-Demographic Factors

- Autori: Dennis Herhausen, Florian von Wangenheim, Jan H. Schumann
- Publicată în: Journal of Business Research (2020)
- Link: Understanding Online Shopper Behavior: A Meta-Analytic Review and Integration of Cognitive and Socio-Demographic Factors
- Rezumat: Această lucrare prezintă o revizuire meta-analitică și integrare a factorilor cognitivi și socio-demografici care influențează comportamentul cumpărătorilor online. Autorii sintetizează rezultatele a peste 100 de studii și examinează impactul unor variabile precum percepția riscului, experiența anterioară în online, și caracteristicile socio-demografice asupra comportamentului de cumpărare online. Studiul oferă o înțelegere mai profundă a motivațiilor și preferințelor cumpărătorilor online, furnizând astfel informații valoroase pentru dezvoltarea de strategii de marketing personalizate și eficiente

3.Preprocesarea datelor

1.Încărcarea setului de date:

Acest pas implică încărcarea setului de date dintr-un fișier CSV folosind biblioteca pandas. Pandas este foarte util pentru manipularea și analiza datelor. Fișierul CSV conține datele brute care vor fi preprocesate și utilizate pentru antrenarea modelului.

2.Explorarea inițială a datelor

Explorarea inițială a datelor implică examinarea rapidă a primelor câteva rânduri ale datasetului, structurii acestuia și distribuției variabilei țintă (Revenue). Acești pași ajută la înțelegerea tipurilor de date, identificarea valorilor lipsă și a distribuției claselor în variabila

3. Gestionarea valorilor lipsă

Valorile lipsă în setul de date pot cauza erori sau pot reduce performanța modelului de învățare automată. În acest exemplu, rândurile care conțin valori lipsă sunt eliminate folosind metoda dropna() din pandas. Alternativ, valorile lipsă ar putea fi completate cu mediana sau media coloanei respective.

4. Codificarea variabilelor categorice

Variabilele categorice (de exemplu, Month, VisitorType, Weekend, Revenue) trebuie convertite în valori numerice pentru a fi procesate de modelul de învățare automată. LabelEncoder transformă fiecare categorie într-un număr întreg unic. Această transformare este necesară pentru ca modelul să poată învăța relațiile dintre caracteristici.

5. Definirea variabilelor caracteristici și țintă

Setul de date este împărțit în două părți: X conține toate caracteristicile (features), iar y conține variabila țintă (target) care trebuie prezisă. În acest caz, Revenue este variabila țintă pe care dorim să o prezicem.

6.Împărțirea setului de date în seturi de antrenament și testare

Setul de date este împărțit în seturi de antrenament și testare folosind train_test_split din biblioteca scikit-learn. test_size=0.2 înseamnă că 20% din date vor fi folosite pentru testare, iar random_state=42 asigură reproducibilitatea împărțirii. Setul de antrenament este utilizat pentru a antrena modelul, iar setul de testare este folosit pentru a evalua performanța acestuia.

7.Scalarea caracteristicilor numerice

Scalarea caracteristicilor este importantă pentru a aduce toate caracteristicile la o scară similară. StandardScaler din scikit-learn transformă caracteristicile astfel încât acestea să aibă media 0 și deviația standard 1. Această scalare poate îmbunătăți performanța și stabilitatea modelului, mai ales pentru algoritmi sensibili la scara datelor (de exemplu, regresia logistică, rețelele neuronale).

Prin parcurgerea acestor pași, setul de date este pregătit corespunzător pentru a fi utilizat în antrenarea și testarea modelului de învățare automată. Preprocesarea datelor este un aspect esențial al oricărui proiect de machine learning, deoarece influențează direct acuratețea și eficiența modelului.

4.Model

În acest capitol, vom prezenta procesul de construire și antrenare a unui model de învățare automată utilizat pentru a prezice dacă un vizitator al unui site de comerț electronic va finaliza sau nu o achiziție. Vom detalia arhitectura modelului, algoritmii utilizați, și metricile de evaluare.

Arhitectura Modelului

Modelul nostru este construit utilizând o rețea neuronală artificială (ANN) implementată cu ajutorul bibliotecii TensorFlow și Keras. Arhitectura rețelei include:

1.Input Layer (Stratul de Intrare): Acesta primește datele de intrare preprocesate. În cazul nostru, stratul de intrare are 17 neuroni, corespunzători celor 17 caracteristici ale cazul nostru, stratul de intrare are 17 neuroni, corespunzători celor 17 caracteristici ale datasetului nostru (Administrative, Administrative_Duration, Informational, Informational_Duration, ProductRelated, ProductRelated_Duration, BounceRates, ExitRates, PageValues, SpecialDay, Month, OperatingSystems, Browser, Region, TrafficType, VisitorType, Weekend).

2.Hidden Layers (Straturile Ascunse): Rețeaua conține două straturi ascunse:

- Primul strat ascuns are 64 de neuroni cu funcția de activare ReLU (Rectified Linear Unit).
- Al doilea strat ascuns are 32 de neuroni, de asemenea, cu funcția de activare ReLU.

3.Output Layer (Stratul de Ieșire): Acesta are un singur neuron cu funcția de activare sigmoid, utilizat pentru a produce o probabilitate între 0 și 1, indicând probabilitatea ca un vizitator să finalizeze o achiziție.

Compilarea Modelului

Modelul este compilat folosind:

- Optimizer (Optimizator): Adam, un algoritm de optimizare eficient care adaptează ratele de învățare pentru fiecare parametru.
- Loss Function (Funcția de Pierdere): Binary Crossentropy, utilizată pentru probleme de clasificare binară.
- Metrics (Metrica): Accuracy, pentru a evalua performanța modelului în timpul antrenamentului și testării.

Antrenarea Modelului

Modelul este antrenat pe datele de antrenament utilizând metoda de gradient descent. Procesul de antrenament include următorii pași:

- Epochs (Epoci): Modelul este antrenat pentru 10 epoci, ceea ce înseamnă că întregul set de date de antrenament este parcurs de 10 ori.
- Batch Size (Dimensiunea Lotului): Loturi de câte 32 de exemple sunt utilizate pentru actualizarea greutăților modelului.
- Validation Split (Fracția de Validare): 20% din datele de antrenament sunt utilizate pentru validare în timpul antrenamentului pentru a monitoriza performanța modelului și pentru a preveni supraantrenarea.

Evaluarea Modelului

Performanța modelului este evaluată pe un set de date de test care nu a fost utilizat în timpul antrenamentului. Principalele metrici utilizate sunt:

- Loss (Pierdere): Măsoară cât de bine sau prost este modelul nostru.
- Accuracy (Acuratețe): Procentajul de predicții corecte realizate de model.

5.Result

Modelul nostru a obținut o acuratețe de aproximativ 91% si acuratete de 0.26879, ceea ce indică o performanță bună în prezicerea comportamentului de cumpărare al vizitatorilor unui site de comerț electronic. Rezultatele antrenamentului și validării au fost vizualizate prin grafice care ilustrează evoluția acurateței și pierderii pe parcursul epocilor.

Graficele de densitate arată distribuția datelor din fiecare coloană numerică, cu axa X reprezentând valorile variabilelor din acele coloane și axa Y reprezentând densitatea probabilității acestor valori, adică cât de frecvent apar acele valori în setul de date. Aceste grafice oferă o perspectivă vizuală asupra modului în care datele sunt distribuite în fiecare coloană numerică, facilitând înțelegerea distribuției și a concentrării datelor în diferite intervale de valori.

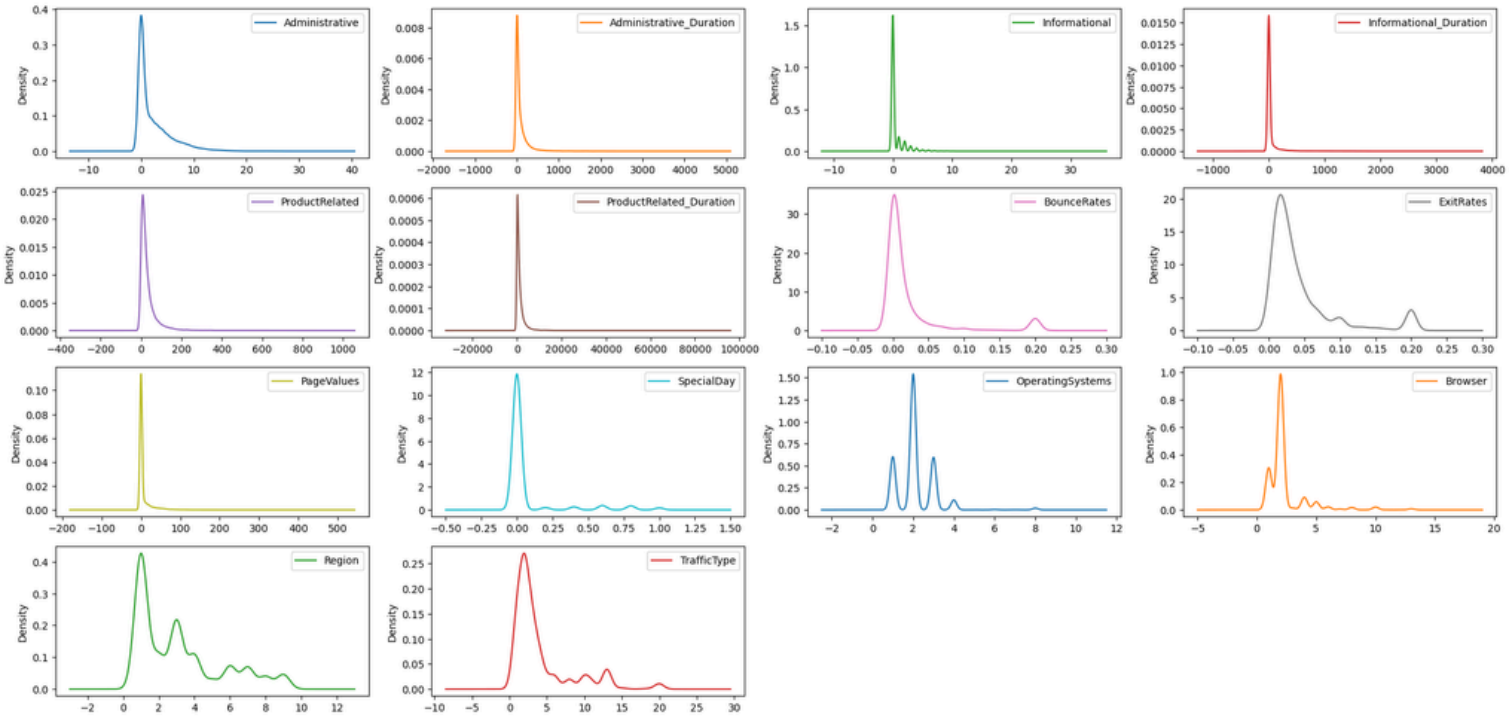
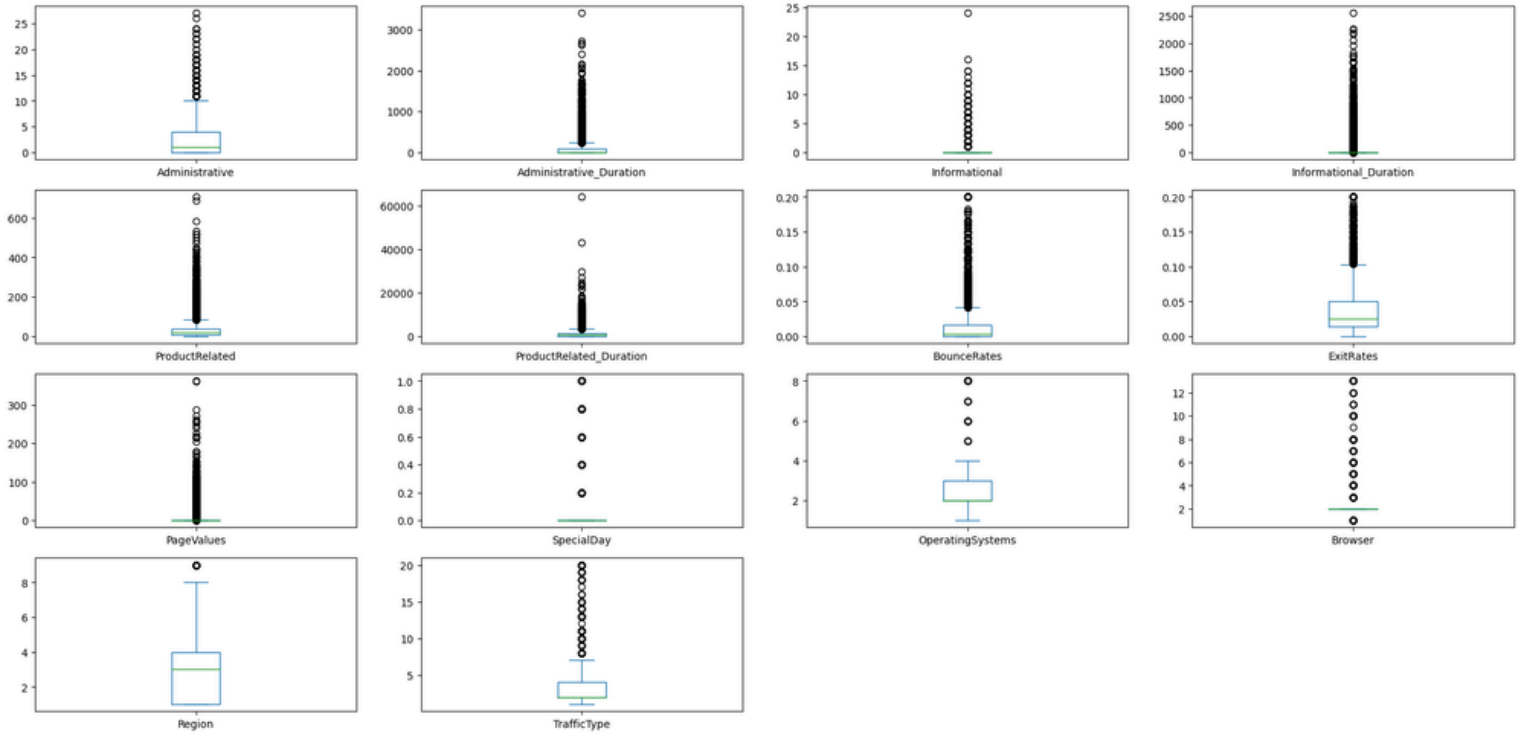
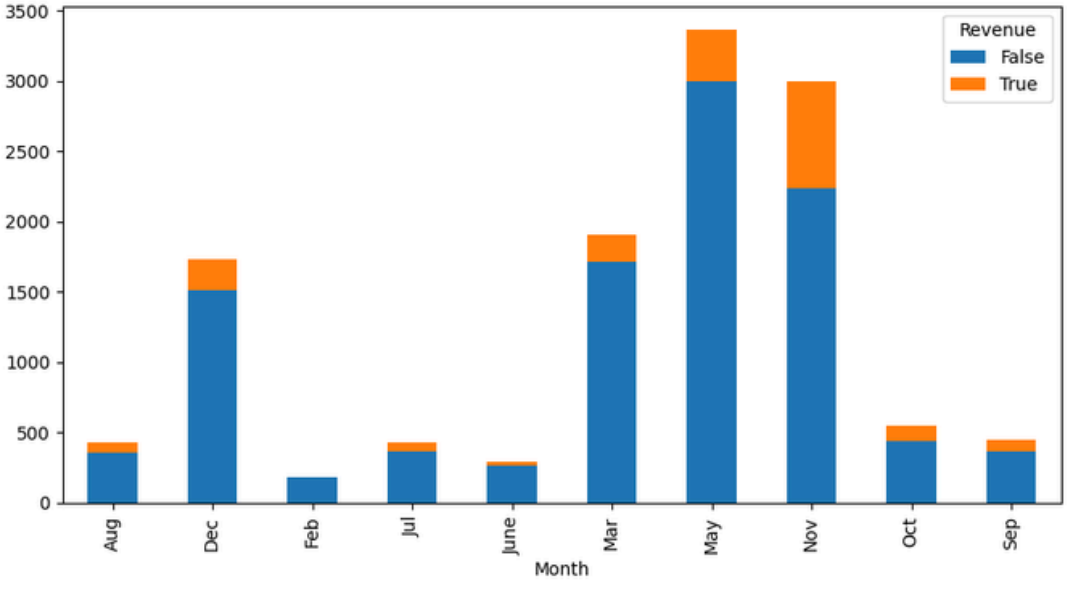


Diagrama "box plot" este o modalitate eficientă de a vizualiza distribuția și variabilitatea unei caracteristici numerice sau a unei serii de date. Aceasta oferă o imagine grafică a cinci statisticilor de bază ale datelor: minimul, primul cuartil (Q1), mediană (Q2), al treilea cuartil (Q3) și maximul.
Pentru a interpreta o diagramă "box plot" pe baza datelor, să luăm în considerare fiecare parte a acesteia:

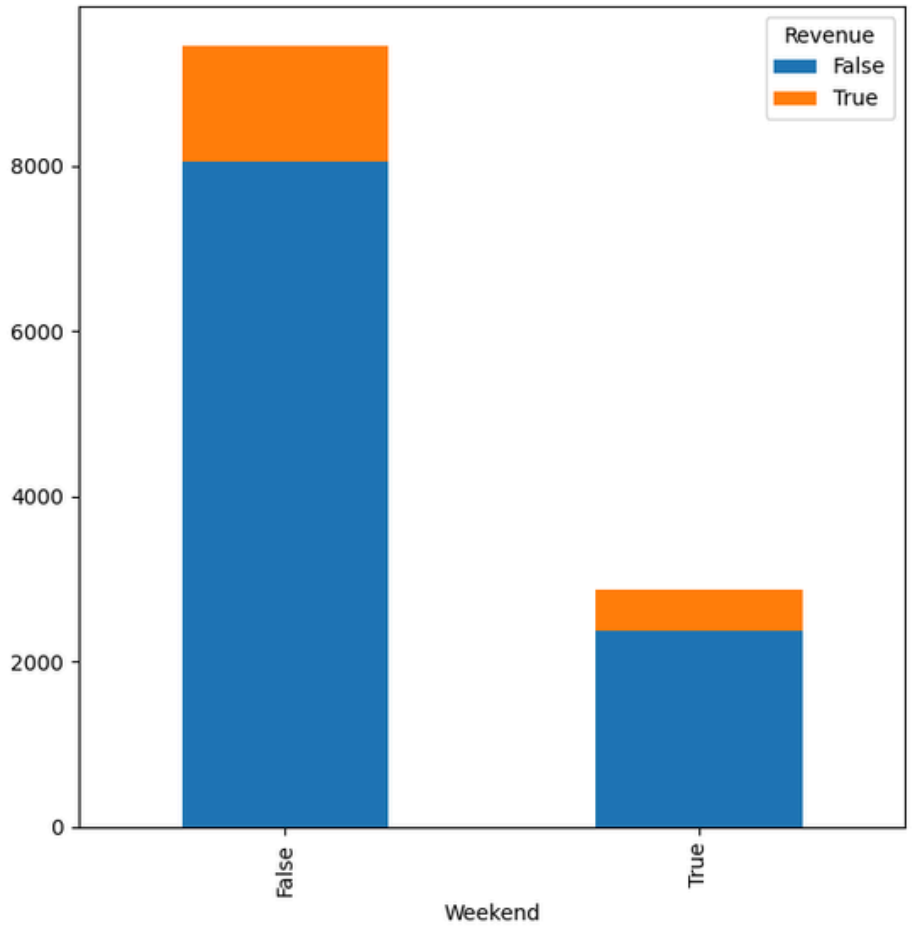
1. Cutia (Box):
   - Partea principală a diagramelor "box plot" este o cutie care reprezintă intercvartilul (IQR), care este intervalul între primul și al treilea cuartil (Q1 și Q3).
   - Linia medianei este trasată în interiorul cutiei.
2. Whiskers (Mustăți):
   - Linia verticală care se extinde din cutie în ambele direcții sunt mustățile. Acestea reprezintă variabilitatea dincolo de intercvartil.
   - Uneori, ele pot reprezenta întreaga gamă a datelor sau pot fi limitate la anumite valori (de exemplu, la 1.5x IQR de la marginea cutiei).
3. Puncte sau puncte discreționare:
   - Punctele care se află în afara mustăților sunt considerate valori atipice sau extrem de extreme, care pot fi semnificative din punct de vedere statistic.
4. Interpretarea:
   - Lungimea cutiei indică cât de împrăștiate sunt datele în interiorul intercvartilului (IQR). Cu cât este mai lungă cutia, cu atât este mai mare variabilitatea datelor.
   - Poziția medianei oferă o indicație despre locul central al datelor. Dacă cutia este simetrică în jurul medianei, atunci datele sunt distribuite uniform. Dacă medianei nu este la mijlocul cutiei, datele sunt asimetrice.
   - Dacă mustățile sunt lungi și datele sunt concentrate în jurul cutiei, atunci distribuția este considerată leptokurtică. Dacă mustățile sunt scurte și datele sunt răspândite în mod uniform, distribuția este plată sau uniformă.



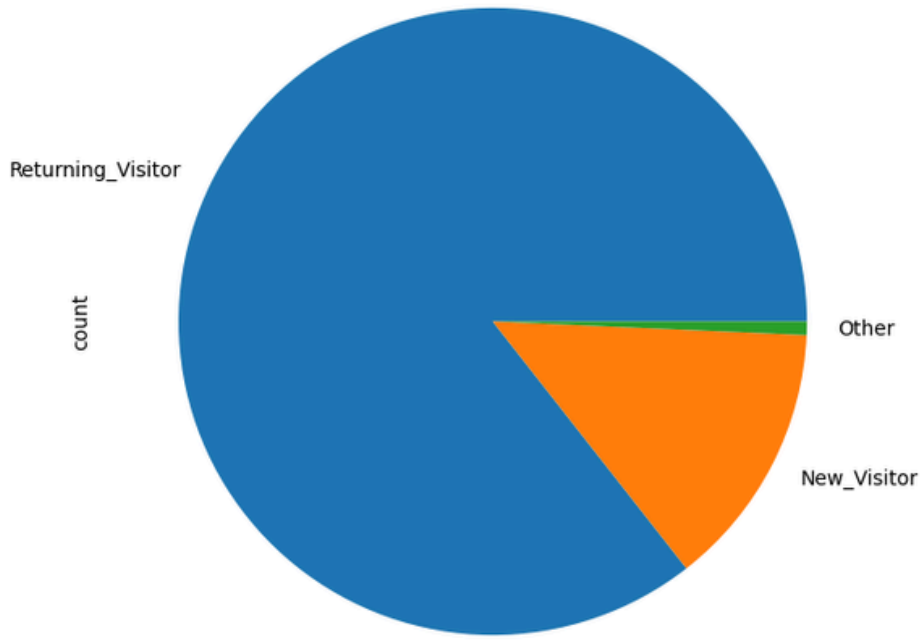
Interpretând graficul, putem observa că luna mai și luna noiembrie au cel mai mare număr de achiziții, în timp ce februarie și iunie au cel mai mic număr de achiziții. Acest lucru poate sugera că aceste luni sunt mai favorabile pentru vânzări sau că există anumite evenimente sau campanii care au un impact semnificativ asupra comportamentului de cumpărare al utilizatorilor în acele luni.



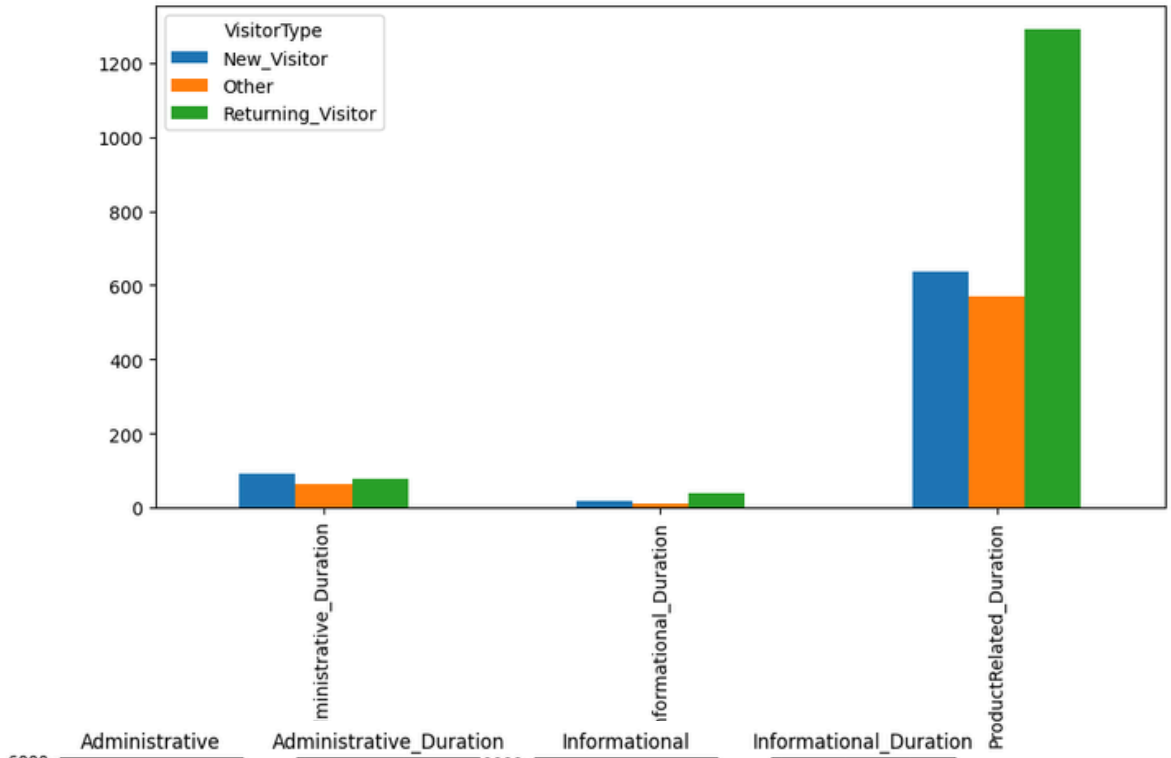
Acest grafic evidențiază o diferență semnificativă în durata medie a activităților legate de produse (ProductRelated Duration) în comparație cu celelalte tipuri de activități (Administrative Duration și Informational Duration). Returning Visitor are cea mai mare medie pentru ProductRelated Duration, care este semnificativ mai mare decât media pentru celelalte tipuri de vizitatori. În contrast, diferența între tipurile de vizitatori este mai mică în ceea ce privește durata activităților administrative și informative. Această diferență considerabilă în ProductRelated Duration subliniază importanța acestui aspect pentru Returning Visitor în comparație cu celelalte activități pe site.

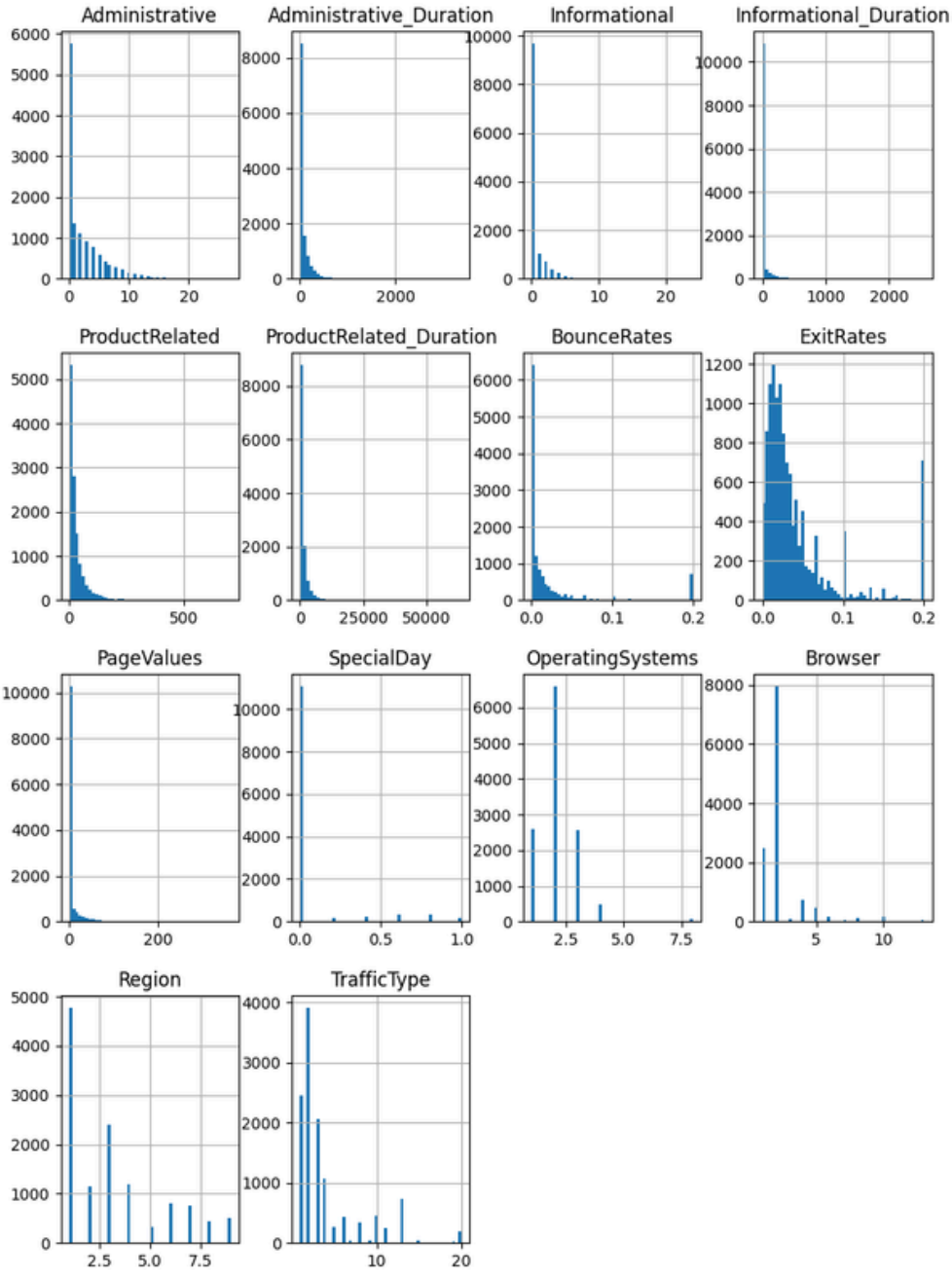


Acest grafic circular prezintă distribuția tipurilor de vizitatori în setul de date. Se observă că majoritatea vizitatorilor sunt de tip "Returning Visitor", urmat de "New Visitor" cu aproximativ 1/5 din numărul total de vizitatori. Există și o categorie denumită "Other", care reprezintă aproximativ 3% din totalul de vizitatori. Astfel, graficul oferă o perspectivă vizuală asupra proporțiilor dintre diferitele tipuri de vizitatori în setul de date.



Acest grafic prezintă distribuția veniturilor în funcție de faptul că ziua este sau nu în weekend. Coloanele reprezintă două categorii de venituri posibile: una pentru cazurile în care s-a înregistrat venit și alta pentru cazurile în care nu s-a înregistrat venit. Graficul arată că probabilitatea de a obține venituri pare să fie mai mică în zilele de weekend, în comparație cu zilele de lucru. Astfel, graficul sugerează că comportamentul de cumpărare poate varia în funcție de ziua săptămânii, fiind mai probabil ca achizițiile să fie efectuate în zilele de lucru decât în weekend.

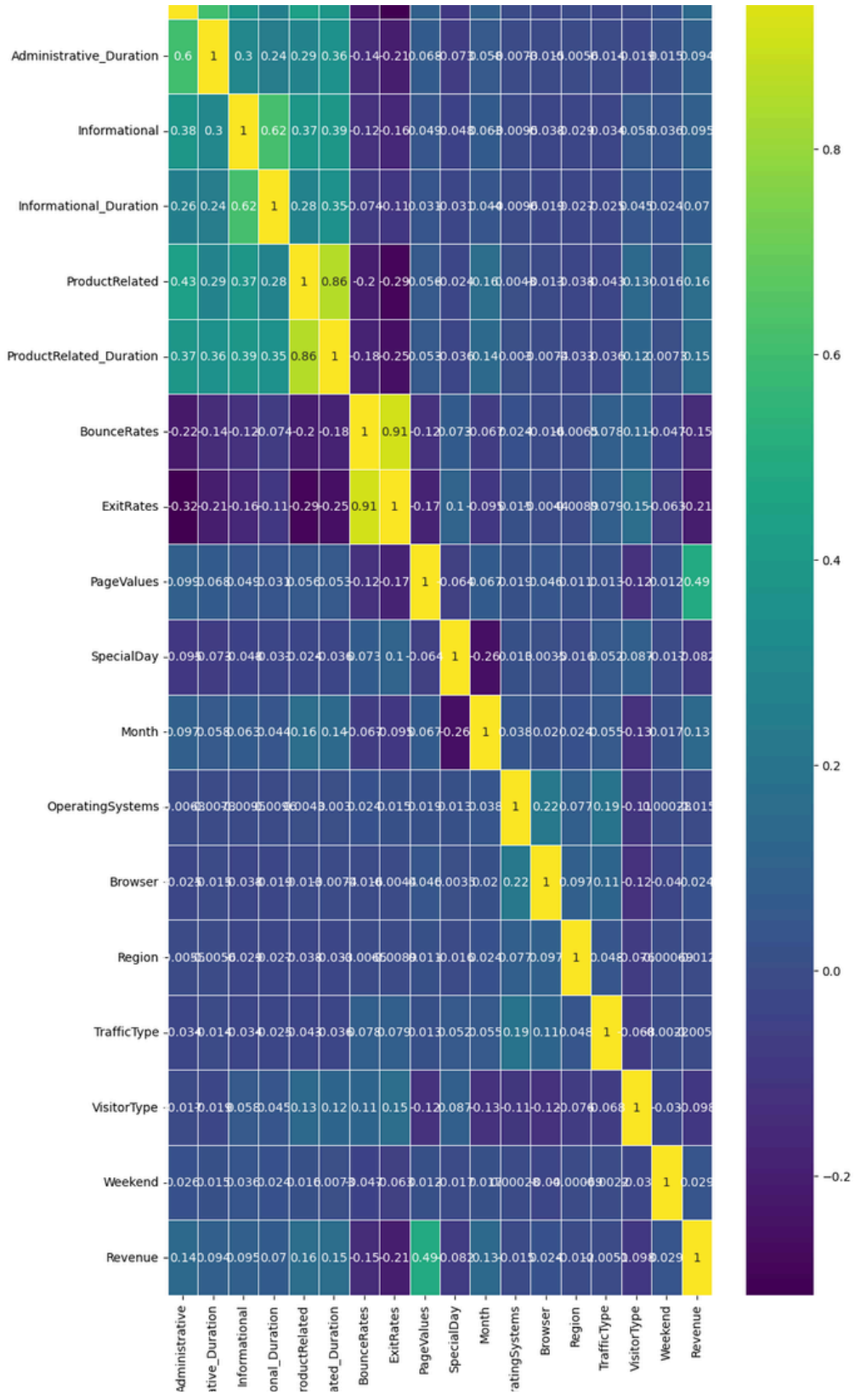




Acest grafic afișează histograma pentru fiecare variabilă din setul de date. Histograma reprezintă distribuția frecvențelor sau a probabilităților pentru fiecare valoare a unei variabile. Parametrul "bins=50" specifică numărul de bare (intervale) în care sunt grupate valorile, iar parametrul "figsize=(10,14)" specifică dimensiunea figurii (lungimea și lățimea) în inci. Astfel, pentru fiecare variabilă din setul de date, histograma arată modul în care valorile sunt distribuite pe intervalul specificat de bare și oferă o perspectivă vizuală asupra formei și caracteristicilor distribuției datelor.

Aceste transformări sunt utile atunci când lucrăm cu modele de învățare automată, deoarece majoritatea modelelor funcționează doar cu date numerice. Prin maparea valorilor categorice la valori numerice, putem utiliza aceste variabile în analize statistice și modelare predictivă.

- Pentru coloana 'Month':
- Un dicționar este creat pentru a asocia fiecarei luni cu un număr întreg corespunzător. De exemplu, 'Feb' este asociat cu 2, 'Mar' cu 3, și așa mai departe.
- Apoi, coloana 'Month' din setul de date este mapată folosind acest dicționar. Astfel, fiecare valoare din coloana 'Month' este înlocuită cu numărul corespunzător asociat din dicționar.
- Pentru coloana 'VisitorType':
- Un alt dicționar este creat pentru a asocia fiecare tip de vizitator cu un număr întreg. De exemplu, 'Returning_Visitor' este asociat cu 3, 'New_Visitor' cu 2, și 'Other' cu 1.
- Coloana 'VisitorType' este apoi mapată folosind acest dicționar, astfel încât fiecare tip de vizitator este înlocuit cu numărul corespunzător din dicționar.
- Pentru coloanele 'Weekend' și 'Revenue':
- Se utilizează un dicționar simplu pentru a asocia valori booleene (True și False) cu 1 și 0, respectiv.
- Coloanele 'Weekend' și 'Revenue' sunt mapate folosind acest dicționar, astfel încât valorile booleene sunt înlocuite cu 1 și 0, respectiv.



K-Nearest Neighbors (KNN) este un algoritm de învățare supervizată utilizat pentru clasificare și regresie. În cazul clasificării, KNN atribuie un punct de date la clasa cea mai frecventă a celor mai apropiați vecini ai săi.

Matricea de confuzie furnizează o viziune detaliată asupra performanței modelului de clasificare. Pentru matricea de confuzie dată:

[[3378 58]

[446 187]]

- Elementul din colțul stânga sus (3378) reprezintă numărul de exemple negative (clasa "nu a generat venit") care au fost corect clasificate ca negative (True Negative).
- Elementul din colțul dreapta sus (58) reprezintă numărul de exemple negative care au fost clasificate incorect ca pozitive (False Positive).
- Elementul din colțul stânga jos (446) reprezintă numărul de exemple pozitive (clasa "a generat venit") care au fost clasificate incorect ca negative (False Negative).
- Elementul din colțul dreapta jos (187) reprezintă numărul de exemple pozitive care au fost corect clasificate ca pozitive (True Positive).

Aceste valori oferă o imagine detaliată a acurateței și a erorilor modelului KNN.

Acest raport de clasificare (classification_report) furnizează o evaluare detaliată a performanței clasificatorului pentru fiecare clasă și pentru întregul set de date. Iată o explicație pentru fiecare metrică din raport:

1. Precision: Acesta reprezintă capacitatea modelului de a identifica corect exemplele pozitive din totalul exemplelor identificate drept pozitive. Pentru clasa 0 (non-revenue), precizia este de 0.88, iar pentru clasa 1 (revenue), precizia este de 0.76. Aceasta înseamnă că, din toate exemplele identificate drept reveniturI (1), 76% sunt corecte, iar din toate exemplele identificate drept non-reveniturI (0), 88% sunt corecte.
2. Recall (sensibilitate): Acesta reprezintă capacitatea modelului de a identifica corect toate exemplele pozitive din totalul exemplelor pozitive prezente în setul de date. Pentru clasa 0, sensibilitatea este de 0.98, iar pentru clasa 1, sensibilitatea este de 0.30. Aceasta înseamnă că modelul identifică corect 98% din toate exemplele de non-reveniturI (0), dar identifică doar 30% din toate reveniturile (1) prezente în setul de date.
3. F1-score: Acesta este un indicator al echilibrului între precizie și recall și este calculat ca media armonică a acestora. Pentru clasa 0, F1-score este de 0.93, iar pentru clasa 1, F1-score este de 0.43. Cu cât F1-score este mai aproape de 1, cu atât este mai bun echilibrul între precizie și recall.
4. Support: Acesta reprezintă numărul de exemple din setul de testare pentru fiecare clasă.

Mai departe, accuracy (acuratețe) este raportul între numărul de predicții corecte și numărul total de predicții. În acest caz, acuratețea este de 0.88, ceea ce înseamnă că 88% din toate exemplele au fost clasificate corect de către model.

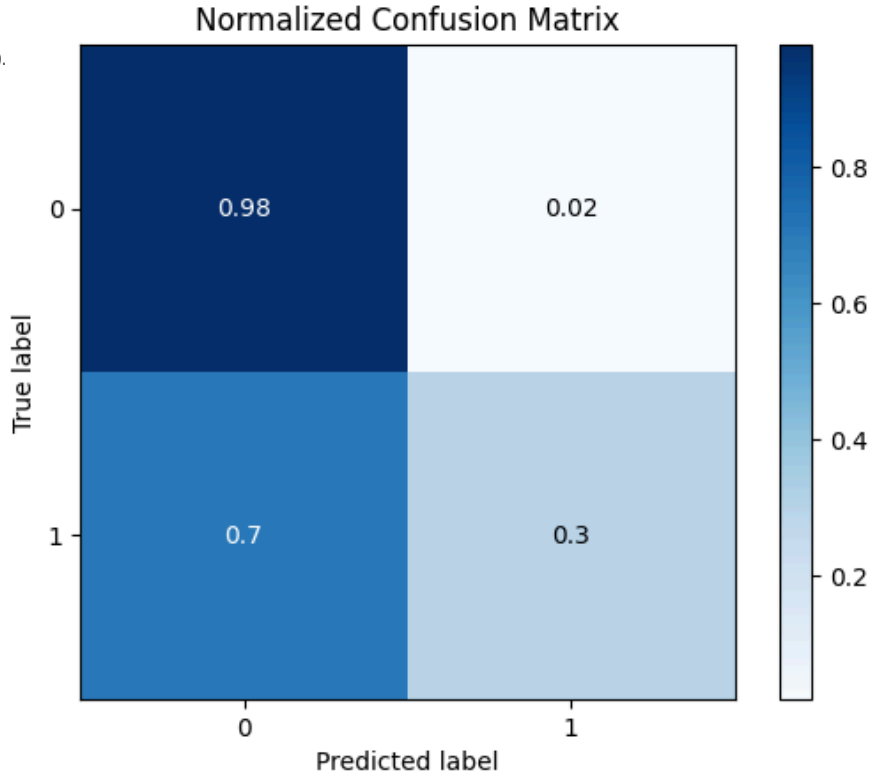
Macro avg reprezintă media aritmetică a metricilor pentru fiecare clasă, în timp ce weighted avg este media ponderată, luând în considerare distribuția fiecărei clase în setul de date. Aceste două valori oferă o perspectivă generală asupra performanței modelului

	precision	recall	f1-score	support
0	0.88	0.98	0.93	3436
1	0.76	0.30	0.43	633

accuracy 0.88 4069

macro avg 0.82 0.64 0.68 4069

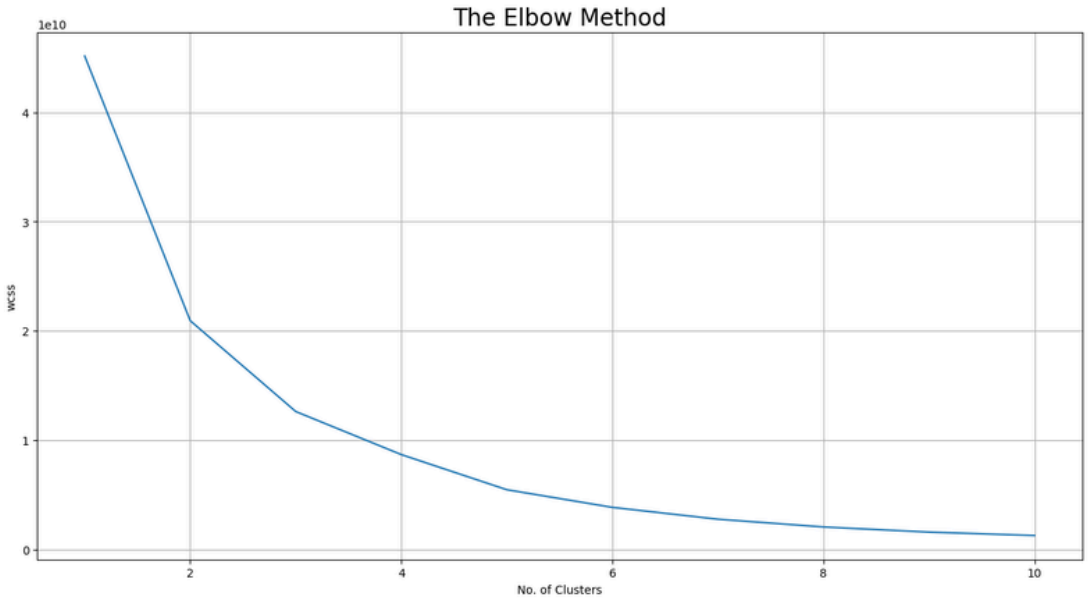
weighted avg 0.86 0.88 0.85 4069.



Acest grafic ilustrează metoda cotului (K-elbow method) pentru determinarea numărului optim de clustere în algoritmul K-means. În acest context:

- Pe axa X avem numărul de clustere testate, de la 1 la 10.
- Pe axa Y avem o măsură a variației în cadrul clusterelor, cunoscută sub numele de "within-cluster sum of squares" (WCSS). Aceasta este o măsură a dispersiei datelor în cadrul fiecărui cluster.
- Obiectivul este să găsim punctul de cotitură (elbow point) în grafic, care indică numărul optim de clustere. Punctul de cotitură este locul în care scăderea variației în cadrul clusterelor începe să se aplaneze.

Interpretarea acestui grafic ar fi că, deși WCSS continuă să scadă odată cu creșterea numărului de clustere, ritmul de scădere începe să se reducă după un anumit punct. În consecință, punctul în care observăm o schimbare bruscă în ritmul de scădere a WCSS este considerat numărul optim de clustere. În cazul acestui grafic, deoarece nu vedem o cotitură evidentă și curba continuă să scadă uniform, poate fi dificil să stabilim un număr optim de clustere folosind metoda cotului. Totuși, uneori poate fi folosită o abordare empirică, selectând numărul de clustere care poate furniza cel mai bun compromis între complexitatea modelului și capacitatea acestuia de a grupa datele în mod eficient.



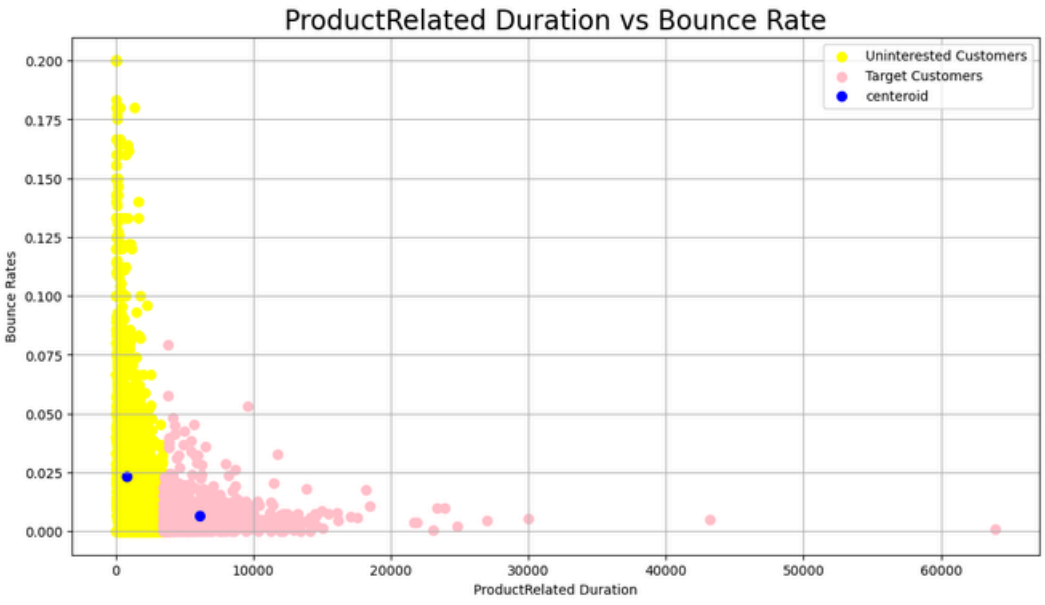
Acest grafic prezintă o reprezentare grafică a două caracteristici ale datelor: ProductRelated Duration și Bounce Rates. Interpretarea sa este următoarea:

- Pe axa X avem durata totală a paginilor de produs vizitate de către utilizatori.
- Pe axa Y avem rata de respingere (Bounce Rate), care este procentul de utilizatori care părăsesc site-ul imediat după vizitarea unei singure pagini.

Punctele galbene reprezintă clienții considerați "Uninterested Customers", adică acei clienți care au o durată scăzută de navigare pe paginile de produs și o rată mare de respingere. Punctele roz reprezintă clienții considerați "Target Customers", adică acei clienți care petrec mai mult timp pe paginile de produs și au o rată mai mică de respingere.

Punctele albastre reprezintă centroizii (centroidul fiecărui cluster), care sunt centrele teoretice ale clusterelor.

Interpretația acestui grafic ar putea fi că, în general, clienții care petrec mai mult timp pe paginile de produs au o tendință mai mică de a părăsi site-ul imediat. Aceasta ar putea fi o informație valoroasă pentru îmbunătățirea experienței utilizatorului și a ratei de conversie a site-ului.



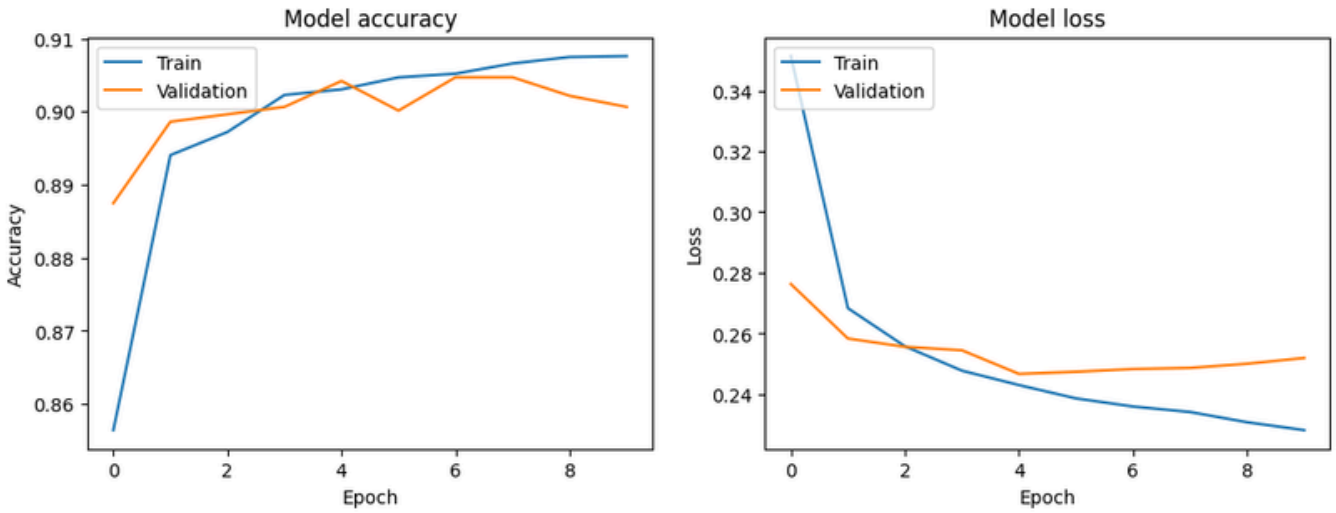
În rezultatul obținut, putem observa că modelul de rețea neuronală a fost antrenat timp de 10 epoci, folosind un set de date împărțit în setul de antrenare și cel de validare (reprezentat de X_train și y_train). Modelul a fost evaluat pe setul de testare (reprezentat de X_test și y_test) și s-au obținut următoarele rezultate:

1. Loss (Pierdere): Loss-ul reprezintă o măsură a cât de bine se potrivesc predicțiile modelului cu valorile reale. Pentru setul de testare, loss-ul a fost în jur de 0.2690, ceea ce indică faptul că modelul a avut o performanță destul de bună în ajustarea predicțiilor sale la datele de testare.
2. Accuracy (Acuratețe): Acuratețea reprezintă proporția de exemple clasificate corect din totalul exemplelor. Pentru setul de testare, acuratețea a fost de aproximativ 88.93%, ceea ce înseamnă că modelul a clasificat corect aproximativ 88.93% din exemplele din setul de testare.

Graficele prezentate indică performanța modelului pe parcursul antrenării:

- Graficul pentru precizia modelului (Model accuracy): Acest grafic arată cum s-au schimbat precizia modelului pe parcursul antrenării (epoch-urile), atât pentru setul de antrenare, cât și pentru cel de validare. În general, dacă precizia crește în timpul antrenării și este similară între setul de antrenare și cel de validare, înseamnă că modelul nu suferă de supra-antrenare (overfitting).
- Graficul pentru pierderea modelului (Model loss): Acest grafic arată cum s-a schimbat pierderea modelului pe parcursul antrenării (epoch-urile), atât pentru setul de antrenare, cât și pentru cel de validare. Scăderea valorii pierderii indică îmbunătățirea performanței modelului.

În concluzie, modelul de rețea neuronală pare să funcționeze destul de bine, având o acuratețe de aproape 89% pe setul de testare. Cu toate acestea, este întotdeauna util să se evalueze și alte metrici și să se efectueze îmbunătățiri pentru a obține o performanță optimă a modelului.



Sesurse Bibliografice:

1. Sakar, C. O., Polat, S. O., Katircioglu, M., & Kastro, Y. (2019). Real-time prediction of online shopper purchasing intention using multilayer perceptron and LSTM recurrent neural networks. Neural Computing Applications. Retrieved from Springer Link.
2. Online Shoppers Purchasing Intention Dataset. (2019). UCI Machine Learning Repository. Retrieved from UC Repository.
3. Kingsnorth, S. (2019). Digital Marketing Strategy: An Integrated Approach to Online Marketing. Kogan Pag Retrieved from Kogan Page.
4. Chen, Y., Wang, F., & Xie, J. (2020). "Understanding consumers' intention to use e-payment: A study base on an extended unified theory of acceptance and use of technology model." Electronic Commerc Research and Applications, 41, 100918. Disponibil la: ScienceDirect.
5. Liu, Y., & Jang, S. S. (2009). "Perceived fit and satisfaction on web portal use." Information & Managemer 46(5), 280-287. Disponibil la: ScienceDirect.