Analiza Vânzărilor Maven Roasters

Pop Lucas

Contents

1 Introducere **2**
1.1 Scopul proiectului 2

2 Contextul sursei de date şi al proiectului **2**
2.1 Utilitatea acestui set de date 2
2.2 Cerinţe şi obiective 3
2.3 Rezultate dorite 3

3 Related Work **3**

4 Implementarea aspectelor teoretice în cadrul proiectului **6**
4.1 Bibliotecile folosite și limbajul de programare 6
4.2 Selectarea și preprocesarea bazei de date 6
4.2.1 Descrierea setului de date 7
4.2.2 Curățarea datelor 7
4.2.3 Preprocesarea datelor 7
4.3 Explorarea și analiza datelor 7
4.4 Metodologii de învățare automată utilizate 7
4.4.1 Linear Regression 7
4.4.2 Logistic Regression 8
4.4.3 Random Forest 8
4.5 Implementarea modelelor de predicție 8
4.5.1 Configurarea modelelor și parametrii utilizați 8
4.5.2 Validarea și antrenarea modelelor 8
4.6 Testare și Validare 9

5 Rezultate **9**

6 Concluzii **11**

Abstract

Acest document prezintă o analiză a datelor de vânzări pentru Maven Roasters, o cafenea cu trei locații în New York. Proiectul utilizează tehnici de machine learning și statistică pentru a explora modelele de vânzări și preferințele consumatorilor. Obiectivele principale includ predicția celui mai vândut produs per locație, analiza prețurilor folosind indicele Gini, crearea modelelor predictive și compararea performanței diferiților algoritmi. Setul de date conține informații detaliate despre tranzacții, inclusiv date, locații, categorii de produse și prețuri. Metodologia implică explorarea datelor, preprocesarea (curățarea datelor, codificarea variabilelor categoriale), și aplicarea modelelor de regresie (Liniară, Random Forest) și clasificare (Regresie Logistică). Rezultatele indică faptul că modelul Random Forest a avut cea mai bună performanță în predicția cantităților vândute (R^2=0.97), urmat de Regresia Logistică pentru clasificarea popularității produselor (acuratețe 0.91). Proiectul demonstrează transformarea datelor brute în informații utile pentru deciziile de management ale cafenelei.

1 Introducere

În acest proiect de analiză a datelor doresc să prezint cum am explorat modelele de vânzări şi preferinţe ale consumatorilor Maven Roasters,o cafenea cu trei locaţii în New York.Utilizând tehnici de machine learning şi statistică,în documentaţia de faţă doresc să detaliez procesul complet de la explorarea iniţială a datelor până la dezvoltarea modelelor predictive şi interpretarea rezultatelor obţinute.

Setul de date care este analizat conţine informaţii detaliate despre tranzacţii, inclusiv date şi timestamp-uri,locaţiile specifice, categoriile de produse şi preţurile produselor.Aşadar cu ajutorul acestor informaţii precise şi vaste am avut posibilitatea să investighez diverse aspecte ale operaţiunilor cafenelei,de exemplu: tendinţe de vânzări în funcţie de locaţie, produse şi perioade temporale, identificarea factorilor care influenţează preţurile produselor, dezvoltarea modelelor predictive pentru estimarea preţurilor, divizarea produselor în categorii de preţ, analiza distribuţiei şi echilibrului preţurilor folosind indicele Gini, descoperirea preferinţelor de consum specifice fiecărei locaţii în parte.

Această documentație detaliază etapele parcurse pentru dezvoltarea acestui proiect ca parte a cursului de Sisteme Inteligente. Am ales acest proiect deoarece mi-am dorit să experimentez aplicarea cunoştinţelor teoretice despre statistică şi machine learning pe un set de date din lumea reală.Ca student interesat oarecum de intersecţia dintre tehnologie şi business, am considerat că acest proiect îmi va oferi oportunitatea de a analiza o afacere reală.

1.1 Scopul proiectului

Scopul principal al acestui proiect este de a prezice care va fi cel mai vândut produs per fiecare locaţie în parte cu ajutorul metodelor statistice şi algoritmilor de machine learning,dar mai sunt şi alte scopuri de exemplu: analiza preţurilor cu ajutorul indicelui Gini, creearea modelelor de predicţie, compararea algoritmilor.

2 Contextul sursei de date şi al proiectului

Accest proiect se bazează pe un set de date care conţine informaţii despre vânzările realizate de Maven Roasters,care operează în cele trei locaţii: Astoria, Hell's Kitchen şi Lower Manhattan. Setul de date conţine informaţii detaliate despre tranzacţii,inclusiv data şi ora, locaţia, categoriile de produse, tipuri de produse, detaliile acestora, preţurile unitare şi cantităţile vândute. Datele sunt organizate într-un fişier CSV(coffe-shop-sales-revenue.csv) unde fiecare linie reprezintă o tranzacţie individuală.

2.1 Utilitatea acestui set de date

Acest set de date este ideal pentru exploatarea tendinţelor de vânzări în timp, identificarea zilelor de vârf de trafic al clienţilor, analiza perfomanţei fiecărui produs şi bineînţeles identificarea factorilor care contribuie la fluctuaţiile în volumul vânzărilor.Aşadar acest set de date oferă nişte date destul de importante şi vaste care ne vor ajuta să antrenăm modelele folosite.

2.2 Cerinţe şi obiective

Proiectul are următoarele cerinţe şi obiective principale:**Înţelegerea datelor** unde este examniată distribuţia vânzărilor pe toate cele trei locaţii,identificarea relaţiilor între diferite caracteristici, analizarea inegalităţii vânzărilor folosind indicele Gini şi curba Lorenz, calcularea valorii informaţionale (entropie) pentru fiecare caracteristică, **Modelarea şi predicţia vânzărilor** unde sunt dezvoltate modelele pentru predicţia cantităţilor vândute cu ajutorul căreia vom indentifica cel mai vândut produs pe fiecare locaţie, clasificarea produselor în categorii populare sau mai puţin populare, optimizarea hiperparametrilor modelelor pentru îmbunătaţirea performanţei şi acurateţei,**Compararea modelelor** unde se vor compara cele 3 modele după optimizare şi se va algege cel mai bun dintre ele,**Explicarea rezultatelor pe diferite instanţe** unde vom vedea cât de bine prezice modelul ales pe un set de date.

2.3 Rezultate dorite

Rezultatele dorite în urma acestui proiect sunt: cum variază vânzările în funcţie de locaţie, care sunt categoriile de produse cele mai populare, ce factori influenţează cel mai mult cantităţile vândute,cum variază vânzările în funcţie de locaţie, un model de regresie pentru predicţia cantităţilor vândute, un model de clasificare pentru indentificarea produselor populare, evaluarea importanţei diferitelor caracteristici în predicţii.

În concluzie, proiectul urmăreşte să transforme datele brute de tranzacţii din fişierul csv în date care pot să ofere un sprijin în deciziile de management a vânzărilor cafenelei.

3 Related Work

1. **Random Forest**
   Conform cu articoulul [1] publicat în 2001 de Leo Breinman prezintă o metodă inovatoare de învăţare automată bazată pe arbori de decizie. Random Forest reprezintă un anasamblu de arbori construiţi pe subseturi ale datelor şi ale caracteristicilor, unde fiecare arbore este generat folosind o selecţie aleatorie a predictorilor la fiecare nod, ceea ce reduce corelaţia dintre arbori şi, implicit, varianta modelului. Prin combinarea rezultatelor multiple de arbori, fie prin vot majoritar în cazul clasificării, fie prin mediere în cazul regresiei faptul că Random Forest este o metodă robustă, rezistenţa la supraînvăţare, capabilă să gestioneze seturi de date cu un numar mare de variabile şi să ofere o estimare a importanţei acestora. De asemenea, această tehnică este eficientă în faţa zgomotuilui din date şi a valorilor lipsă. Studiile şi experimentele prezentate demonstrează performanţa superioară a Random Forest faţă de alte metode existente la momentul respectiv, făcând din aceasta o tehnică de referinţă în domeniul invăţării automate.În proiectul prezentat modelul Random Forest are cea mai bună acurateţe aceasta fiind 0.97,o acurateţe destul de mare faţă de alte modele testate.

2. **Regresia Liniară**
   Conform cu [2], cartea „Applied Regression Analysis" de Norman R. Draper şi Harry Smith este o lucrare importantă care explică cum funcţionează regresia liniară şi cum poate fi folosită pentru a înţelege relaţiile dintre variabile.Autorii prezintă conceptele de bază şi arată cum se constuiesc şi interpretează modelele de regresie, oferind exemple practice din viaţa reală.Este explicat cum să verifici dacă modelul tău este bun şi cum să imbunătăţeşti rezultatele, cum ar fi prin alegerea variabilelor potrivite sau verificarea eventualelor erori.Eu am folosit regresia liniară în analiza mea şi am obţinut o acurateţe de aproximatriv 0.51, ceea ce arată că modelul a putut explica parţial relaţiile din date.

3. **Regresia Logistică**
   Conform cu [4], cartea „Applied Logistic Regression" de David W. Hosmer Jr., Stanley Lemeshow şi Rodney X. Sturdivant este una dintre cele mai importante lucrări din domeniul statisticii aplicate, concentrându-se pe regresia logistică, o metodă esențială pentru modelarea și predicția unor rezultate binare, precum „da" sau „nu", „succes" sau „eșec". Autorii explică în detaliu fundamentele metodei, modul în care se construiesc modelele logistice și cum se interpretează

coeficienții pentru a înțelege impactul variabilelor explicative asupra probabilității unui eveniment. Cartea include numeroase exemple practice, tehnici de diagnosticare și strategii pentru evaluarea și validarea modelelor, ceea ce o face accesibilă și utilă atât pentru studenți, cât și pentru practicieni. De asemenea, sunt prezentate metode pentru tratarea problemelor comune întâlnite în analiza datelor reale, cum ar fi coliniaritatea sau datele dezechilibrate. Eu am folosit regresia logistică în analiza mea și am obținut o acuratețe de 0.91, ceea ce indică o performanță foarte bună a modelului în predicția corectă a rezultatelor. Această carte rămâne o resursă valoroasă pentru oricine dorește să înțeleagă și să aplice eficient tehnicile de regresie logistică în diverse domenii.

4. **Gini Index-Curba Lorenz**
   Conform cu aricolul [9] "A simple method for estimating the Lorenz curve", scris de Thitithep Sitthiyot şi Kanyarat Holasut şi publicat în Humanities and Social Sciences Communications in 2021, propune o metodă simplificată şi accesibilă pentru estimarea curbei Lorenz - un instrument esenţial în analiza inegalităţii economice. Autorii îşi propun să depăşească complexitatea tehnică a metodelor tradiţionale şi să ofere o soluţie mai intuitivă, care poate fi aplicată cu uşurinţă de cercetători, studenţi şi specialişti care nu au pregătire avansată în matematică sau econometrie. Curba Lorenz este o reprezentare grafică a dristribuţiei cumulative a veniturilor sau a altor resurse într-o populaţie, şi stă la baza calculului indicelui Gini - un indicator numeric al inegalităţii.În articol, autorii prezintă un model matematic simplu, bazat pe ecuaţii logice şi relaţii proporţionale, care permite construirea curbei Lorenz pornind de la date agregate. Metoda lor reduce cerinţele privind datele şi nivelul de complexitate computaţională, fără a compromite acurateţea estimărilor.
   În una din graficele afişate în proiectul meu este prezentată exact aplicarea acestei metode, unde curba Lorenz obţinută reflectă distribuţia cumulativă a valorilor pentru variabila transaction_id,raportată la proporţia populaţiei. Linia roşie punctată reprezintă linia egalităţii perfecte, iar aria dintre aceasta şi curba reala este folosită pentru a calcula Indicele Gini,în acest caz o valoare de 0.3334, indicând un nivel moderat de inegalitate. Aşadar cu ajutorul metodei propuse de Sitthiuot şi Holasut. Prin această abordare, am reuşit să obţin o vizualizare clară a distribuţiei datelor şi să măsor gradul de concentrare folosind Gini.

5. **Entropie**

   Conform cu [7] articolul "Induction of decision Trees" de J. Ross Quinlan, publicat în 1986, este o lucrare de referinţă care stă la baza dezvoltării algoritmilor moderni de învăţare autoamtă, în special a celor bazaţi pe arbori de decizie. Autorul propune o metodă de construcţie a arborilor de decizie care utilizează concepte fundamentale din teoria informaţiei, cum ar fi entropia şi câştigul informaţional. Entropia este definită ca o masură a incertitudinii şi a dezordinii dintr-un set de date şi este folosită pentru a evalua cât de omogene sunt clasele unui atribut. Dacă un atribut împarte datele în subseturi care conţin instanţe dintr-o singură clasă, entropia este scazută şi acel atribut este considerat informativ. Câştigul de informaţie măsoară cât de mult scade entropia dacă se foloseşte un anumit atribut pentru a impărţi datele.
   în analiza mea am aplicat această idee fundamentală prin calcularea entropiei pentru coloana dintr-un set de date. Rezultatele sunt afişate într-un grafic,care arată clar variaţiile entropiei între atribute.

6. **Data Cleaning**

   Conform cu [8] articolul "Data Cleaning: Problems and Current Approaches" scris de Erhard Rahm şi Hong Hai Do, publicat în 2000 în IEEE Data Engineering Bulletin, oferă o privire de ansamblu clară şi structurată asupra provocărilor şi soluţiilor existente în procesul de curăţare a datelor.Autorii evidentiază faptul că datele reale sunt adesea incomplete, redundante, incorecte, iar aceste probleme pot afecta semnificativ calitatea rezultatelor în orice analiză de date.
   Lucrarea subliniază că procesul de curăţare a datelor nu este doar o etapă preliminară, fie că este vorba de baze de date tradiţionale, sisteme de integrare a datelor sau aplicaţii de tip data warehouse.
   Un alt punct important din articol este accentul pus pe dificultatea autoamtizării curăţiri datelor, întrucat aceasta depinde foarte mult de contextul specific al domeniului şi de calitatea surselor de date.

În cadrul analizei mele, am aplicat acest principiu de data cleaning pentru a elimina valorile NaN şi NULL de pe coloane dar şi rândurile duplicate.

7. **Matricea de Corelaţie**

   Conform cu [3] articolul "Corrgrams: Exploratory Displays for Correlation Matrices" scris de Michael Friendly,introduce şi dezvoltă conceptul de "corrgram",o reprezentare grafică menită să ajute la exploatarea şi înţelegerea rapidă a structurii unei matrici de corelaţie. Autorul porneşte de la ideea că, deşi matricile de corelaţie oferă informaţii esenţiale despre relaţiile dintre variabile. Friendly propune diverse forme de corrgrame care combină codarea culorii, simboluri geometrice şi ordonarea variabilelor pentru a evidenţa grupările, asocierile sau tiparele importante dintre variabile.
   În cadrul analizei mele am aplicat o metodă similară de vizualizare a corelaţiilor sub forma unui heatmap,unde sunt reflectate nivelurile de corelaţie dintre variabile .Aşadar reprezentarea grafică a matricii de corelaţie este o practică esenţială în analiza statistică exploratorie.

8. **Train/Test split**

   Conform cu [10] articolul "A Critical Look at the Current Train/Test Split in Machine Learning" scris de Jimin Tan, Jjianan Yang, Sai Wu, Gang Chen şi Jake Zhao,sunt analizate în profunzime practicile comune privind împarţirea datelor pe seturi de antrenament şi test în învaţarea autoamată. Autorii subliniază faptul că alegera unei proporţii fixe precum 80/20 sau 70/30, a devenit un standard în multe studii şi proiecte, insă fară o justificare riguroasă adaptată contextului fiecarui set de date sau model.
   Lucrarea atrage atenţia asupra riscurilor generate de aplicarea autoamată a unei astfel de imparţiri,în special atunci când datele sunt dezechilibrate, prezintă distrib uţii neomogene. De asemenea ne este prezentat ca o împarţire fixă poate introduce variaţii semnificative în performanţele măsurate,influenţând astfel concluziile despre eficienţa modelului. În locul acestei abordări statice, sunt încurajate metode mai robuste precum cross-validation,care oferă o estimare mai stabilă şi mai corectă a performanţei reale a modelului.
   În cadrul analizei mele, am aplicat atât împărţirea tradiţională a datelor în proporţie de 80% pentru antrenament şi 20% pentru test, metoda foarte des întalnită şi uşor de implementat, cât şi cross-validation pentru evaluarea şi optimizarea modelelor.

9. **Preprocesarea datelor**

   Conform cu [5] articolul "Big Data Preprocessing" scris de Julian Luengo, Diego Garcia-Gil, Sergi Ramirez-Gallego, Salvador Garcia şi Francisco Herrera,unde sunt analizate în profunzime toate etapele esenţiale ale preprocesării datelor, cu un accent deosebit pe provocările întâmpinate atunci când lucrăm cu seturi de date foarte mari. Autorii subliniază că, în cotextul volumelor masive de date generate zilnic, preprocesarea devine o etapă critică şi indispensabilă, deoarece datele brute sunt adesea incomplete, redundante.
   Lucrarea este structură în jurul celor mai importante componente ale preprocesării: curăţarea datelor, transformarea datelor, reducerea dimensionalităţii şi selecţia caracteristicilor. În etapa de curăţare, se discută depsre detectarea şi tratarea valorilor lipsă, eliminarea duplicatelor şi corectarea erorilor de înregistrare. La tranformarea datelor, autorii pun accent pe standardizare, normalizare şi codificarea variabilelor categoriale pentru a permite o interpretare corectă de către algoritmi.
   De asemenea, o parte importantă a lucrării este dedicată rducerii complexităţii datelor, prin tehnici de selecţie şi extragere a trăsăturilor relevante. Acesta este necesar pentru a evita supraînvăţarea, pentru a reduce timpul de antrenare a modelelor.
   În analiza mea am aplicat mai multe principii. Am curăţat datele de intrare eliminând valorile lipsă şi de asemenea rândurilor duplicate, lucru ce a contribuit sermnificativ la îmbunătaţirea calităţii modelelor utilizate ulterior, demonstrând astfel aplicabilitatea practică a metodogiilor propuse de autorii acestei lucrări.

10. **Histograma**

    Conform cu [6] articolul "Histograms: A Useful Data Analysis Visualizarion" scris de Regin L. Nuzzo,este subliniată importanţa histogramei ca instrument esenţial în analiza exploatorie a datelor. Autoarea evidenţiază faptul că, desi este o unealtă simpla şi adesea subseti-

mată, histograma oferă o perspectivă clara şi intuitivă asupra distribuţiei unei variabile numerice. Prin împarţirea valorilor într-o serie de intervale, histograma arată fecvenţa cu care apar aceste valori, ceea ce permite utilizatorului să identifice rapid caracteristici precum simetria sau asimetria,existenţa unor moduri(vârfuri), întinderea datelor şi posibile valori extreme(outliers). În analiza mea, am uitlizat histograme pentru a examina distribuţia valorilor din coloanele numerice precum unit_price,transaction_qty.Prin aceste reprezentări grafice, am putut observa dacă datele erau distribuite normal.

References

[1] Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

[2] Norman R Draper and Harry Smith. *Applied regression analysis*, volume 326. John Wiley & Sons, 1998.

[3] Michael Friendly. Corrgrams: Exploratory displays for correlation matrices. *The american statistician*, 56(4):316–324, 2002.

[4] David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant. *Applied logistic regression*. John Wiley & Sons, 2013.

[5] Julián Luengo, Diego García-Gil, Sergio Ramírez-Gallego, Salvador García, and Francisco Herrera. Big data preprocessing. *Cham: Springer*, 1:1–186, 2020.

[6] Regina L Nuzzo. Histograms: A useful data analysis visualization. *PM&R*, 11(3):309–312, 2019.

[7] J. Ross Quinlan. Induction of decision trees. *Machine learning*, 1:81–106, 1986.

[8] Erhard Rahm, Hong Hai Do, et al. Data cleaning: Problems and current approaches. *IEEE Data Eng. Bull.*, 23(4):3–13, 2000.

[9] Thitithep Sitthiyot and Kanyarat Holasut. A simple method for estimating the lorenz curve. *Humanities and Social Sciences Communications*, 8(1):1–9, 2021.

[10] Jimin Tan, Jianan Yang, Sai Wu, Gang Chen, and Jake Zhao. A critical look at the current train/test split in machine learning. *arXiv preprint arXiv:2106.04525*, 2021.

4 Implementarea aspectelor teoretice în cadrul proiectului

În acest capitol, voi detalia modul în care conceptele teoretice de analiză a datelor, statistică și machine learning au fost aplicate practic în proiectul de analiză .

4.1 Bibliotecile folosite și limbajul de programare

Limbajul de programare folosit la această analiză este python,iar rularea se face într-un mediu de rulare numit Jupiter.

Bibliotecile principale folosite în analiza mea sunt **pandas** pentru manipularea și analiza datelor tabelare(Dataframe-uri),**numpy** pentru calculele numerice rapide și eficiente,**matplotlib.pyplot** pentru graficele explicative,**seaborn** pentru vizualizari statistice avansate, construită peste matplotlib,**math** modul standard python pentru funcții matematice de bază,**scikit-learn**(sklearn) pentru machine learning(modelare,preprocesare,evaluare,optimizare),**scipy-signal** pentru filtrare și smoothing,**warnings** pentru gestionarea avertismentelor,**time** pentru măsurarea timpului de execuție

4.2 Selectarea și preprocesarea bazei de date

Am început analiza prin încărcarea setului de date cu vânzările cafenelei, eliminând valorile lipsă și duplicate pentru a asigura calitatea datelor. Am transformat coloana cu data tranzacției într-un format de tip lună, pentru a putea analiza sezonalitatea vânzărilor.

Coloanele de tip text (categorice), precum locația,categoria produsului, tipul și detaliile produsului, au fost codificate numeric folosind LabelEncoder, pentru a putea fi folosite ulterior în modelele de

machine learning.Am agregat datele la nivel de produs, locație și lună, calculând totalul cantităților vândute și prețul mediu unitar pentru fiecare combinație.

Astfel am obținut un set de date curat, structurat și pregătit pentru analiza statistică și modelare predictivă, cu variabile relevante pentru scopul proiectului.

4.2.1 Descrierea setului de date

Setul de date analizat contine 149.000 de linii care reprezintă înregistrările unei cafenele numită Maven Roasters,care are 3 locații în New York.Fiecare rând din setul de date reprezintă o tranzacție individuală și include următoarele informații:**transaction_date** ce reprezintă data la care a avut loc tranzacția,**store_location** ce reprezintă numele locației unde s-a realizat tranzacția, **store_id** reprezintă id-ul locației,**product_category** care reprezintă categoria produsului vândut (Coffe,Tea,Bakery), **product_type** reprezintă subcategoria produsului (Espresso,Latte,Herbal),**product_detail** oferă date suplimentare despre produs (decaf,organic,etc.), **transaction_qty** care reprezintă numărul de unități vândute,**unit_price** reprezintă prețul per unitate al produsului vândut,**product_id** reprezintă id-ul produsului,**transaction_time** ora la care s-a făcut tranzacția.

4.2.2 Curățarea datelor

În prima fază pentru o calitate și acuratețe a analizei mai bună, am început cu o etapă de curățare a datelor.Am verificat dacă există valori lipsă (NULL sau NaN) în setul de date pentru a evita erorile în pașii următori.De asemenea, am eliminat toate rândurile duplicate, astfel incat fiecare tranzacție să fie unica și să nu influențeze greșit statisticile sau modelele. După aceste operațiuni am verificat din nou dacă mai există duplicate sau porbleme de consistență.

Prin aceste etape am obținut un set de date curat, fară valori lipsă sau duplicate, pregătit pentru analiza statistică și modelare.

4.2.3 Preprocesarea datelor

După curățarea inițială am trecut preprocesarea datelor pentru a le pregăti pentru analiză și modelare. Am extras luna din data tranzacției pentru a analiza sezonalitatea vânzărilor. Am codificat coloanele categorice precum locația, categoria produsului,tipul și detaliile produsului folosind LabelEncoder, astfel încat să poată fi folosite de modelele de machine learning.

4.3 Explorarea și analiza datelor

După curățarea și preprocesarea datelor, am trecut la etapa de exploatare și analiză pentru a îțtelege mai bine setul de date. Am analizat distribuția valorilor numerice folosind histograme, atât pentru întregul set de date, cât și pentru fiecare locație în parte. Am calculat și vizualizat matricea de corelație pentru a identifica relațiile dintre variabile.De asemenea am folosit curba Lorenz și indecele Gini pentru a evalua inegalitatea distribuției vânzărilor pe diferite segmente. Pentru variabilele categorice și numerice, am calculat entropia pentru a măsura diversitatea informației.

Prin aceste analize, am obținut o imagine clară asupra distribuției vânzărilor, sezonalității, relațiilor dintre variabile.

4.4 Metodologii de învățare automată utilizate

4.4.1 Linear Regression

Am folosit Linear Regression pentru a prezice cantitatea exactă de produse vândute pentru fiecare combinație de locație, categorie, tip, detaliu și lună. Modelul a încercat să găsească o relație liniară între aceste caracteristici și volumul vânzărilor.

Modelul a oferit o predicție de bază, rapidă și ușor de interpretat. Totuși performanța a fost limitată, deoarece relațiile dintre variabilele din vânzări nu sunt strict liniare. Scorul R^2 a fost unul destul de mic acesta fiind de 0.51.

4.4.2 Logistic Regression

Am folosit Logistic Regression pentru a clasifica produsele în "populare" sau "nepopulare", pe baza carcateristicilor lor și al contextului (locație, lună, tip, etc.). Practic modelul a prezis probabilitatea ca un produs să depășească pragul de popularitate.

Modelul a avut o acuratețe bună de R^2 0.91 pentru această sarcină de clasificare, fiind util pentru a indentifica produsele cu potențial ridicat de vânzare, dar nu și pentru clasificări mai complexe.

4.4.3 Random Forest

Random Forest a fost modelul principal pentru predicția cantității vândute și pentru clsificarea produselor populare. Am folosit acest model deoarece poate face relații complexe și interacțiuni între variabile, de exemplu efectul sezonului, combinația dintre tipul produsului și locație.

Ca și rezultate, acesta a avut cele mai bune rezultate la analiza mea atât la regresie cât și la clasificare,R^2 fiind de 0.97.

4.5 Implementarea modelelor de predicție

4.5.1 Configurarea modelelor și parametrii utilizați

Linear Regression

Modelul utilizează un pipeline care include standardizarea datelor urmată de regresia liniară propriu-zisă.

Am configurat modelul cu **fit_intercept=True**, ceeea ce permite modelului sa detecteze nivelul de bază al vânzărilor indiferent de valorile caracteristicilor. Paramterul **positive=False** oferă flexibilitate modelului să identifice atât factori care cresc vânzările, cât și care le diminuează.

Perfomranța acestui model a fost moderată, cu un scor de 0.51, ceea ce înseamnă că explică aproximativ 50% din variația cantităților vândute.

Random Forest

Pentru a captura relații mai complexe și nelineare, am implementat un model Random Forest.

Am utilizat 200 de arbori de decizie (**n_estimators=200**), un număr suficient pentru a asigura stabilitatea predicțiilor fară a crește excesiv timpul de calcul. Adâncimea maximă a arborilor este de 20 (**max_depth=20**).

Pentru a preveni oevrfiting-ul, am impus condiția ca fiecare nod să conțină minim 5 exemple. Am selectat caracteristicile pentru divizare folosind strategia "sqrt".

Logistic Regression

Pentru a clasifica produsele în categorii populare și mai puțin populare, Am implementat un model de regresie logistica. Am configurat modelul cu **class_weight="balanced"** pentru a compensa diferențele dintre frecvențele claselor.

Numărul maxim de iterații a fost setat la 5000 (**max_iter=5000**), asigurând că algoritmul are suficient timp să găsească soluția optimă.

4.5.2 Validarea și antrenarea modelelor

Toate modelele au fost evaluate folosind o abordare riguroasă de cross-validation 5-fold.Această metodă de validare oferă o estimare mai robustă a performanței reale a modelelor de date noi.

Pentru modelele de regresie, am utilizat eroarea pătratică MSE și coeficientul de determinare R^2 ca metrici de evaluare a modelelor. Pentru clasificare, am analizat acuratețea, precizia, recall-uri și scorul F1.

Antrenarea modelelor a evidențiat importanța caracteristicilor legate de locație și tipul produsului, confirmând intuiția inițială că preferințele diferă semnificativ între locații și că anumite categorii de produse au sezoane clare de vânzări.

4.6 Testare și Validare

1. În prima fază am separat datele în seturi de antrenament (80%) și test (20%) folosind stratificarea după locație pentru a menține distribuția corectă a datelor pentru fiecare magazin.

2. Pentru o evaluare mai solidă, am aplicat cross-validation KFold cu 5 partiții care ne oferă o estimare mai stabilă a performanței pe date noi.

   Rezultatele acestui cross-validation (R^2 **mediu=0.85 cu deviație standard** $^{+}_{-}0.024$)*indicăunmodelstabilcuvariați*

3. Am utilizat multiple metrici de a evalua modelele de regresie,acestea fiind **Mean Squared Error(MSE)** care măsoară eroarea medie pătratică a predicțiilor și **Coeficientul de determinare** (R^2).Iar pentru Clasificare **Acuratețea** și **Raportul de clasificare**

4. Am folosit analiza performanței modelelor per fiecare locație, acest tip de analiză permițându-ne astfel identificare variațiilor în acuratețea predicțiilor în diferite locații.

5. Am analizat importanța caracteristicilor pentru a valida relevanța lor în model, acest lucru confirmând astfel că locația,tipul produsului și sezonalitatea sunt factori determinanți.

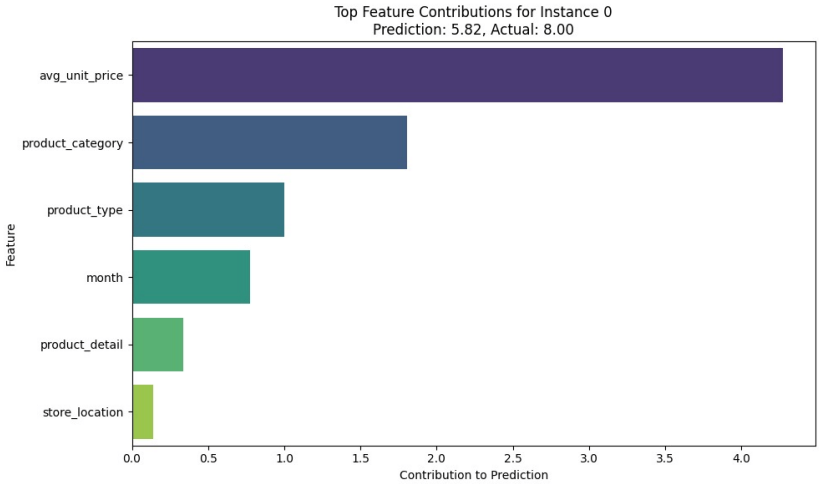
   Așadar procesul de testare și validare demonstrează că modelul Random Forest oferă cele mai bune predicții cu un R^2 peste 0.80 în cross-validation.

5 Rezultate

În urma analizelor asupra predicțiilor modelului pentru câteva instanțe specifice din setul de test, putem observa cum diferitele caracteristici influențează cantitatea vândută estimată.

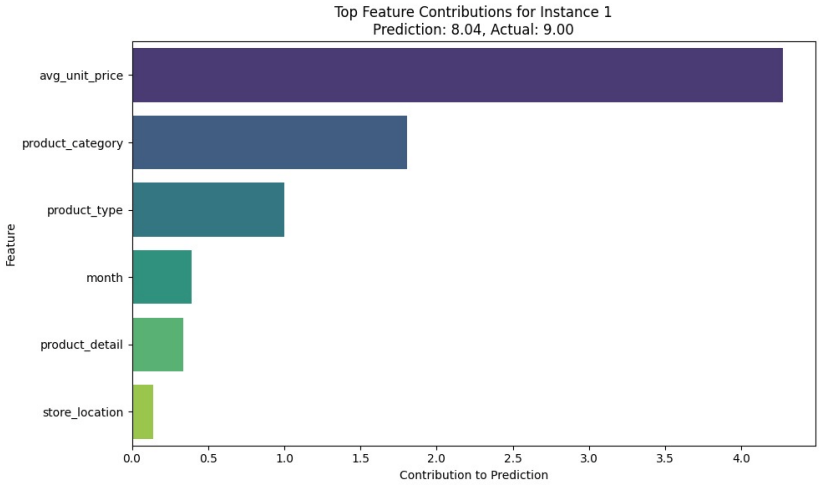
Instanța 0 - Astoria, Coffe, Espresso

Caracteristi principale: Locația Astoria, tipul de produs Coffe, luna Decembrie.
Contribuții la predicție: Locația contribuie cel mai mult, umrează tipul produsului și luna. Prețul unitar are o contribuție mai mică, dar smenificativă.
Explicație: Cafeaua de tip Espresso are vânzări ridicate iarna, datorită sezonalității.



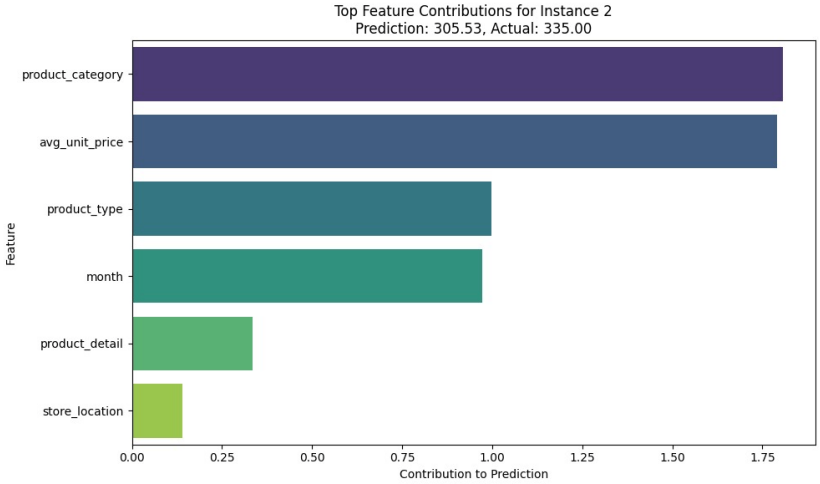
Instanța 1 - Hell's Kitchen, Tea, Herbal

Caracteristici principale: Locația Hell's Kitchen, tipul de produs Tea, luna Iulie
Contribuții la predicție: Locația și tipul de produs domină.
Explicație: În zonele de business, capuccino-ul are vânzări stabile pe tot parcursul anului.



Instanța 2 - Lower Manhattan, Coffe, Cappuccino

Caracteristici principale: Locația Lower Manhattan, tipul de produs Coffe (Cappuccino)
Contribuții la predicție: Locația și tipul de produs domină.
Explicație: În zonele de business, cappuccino-ul are vânzări stabile pe tot parcursul anului.

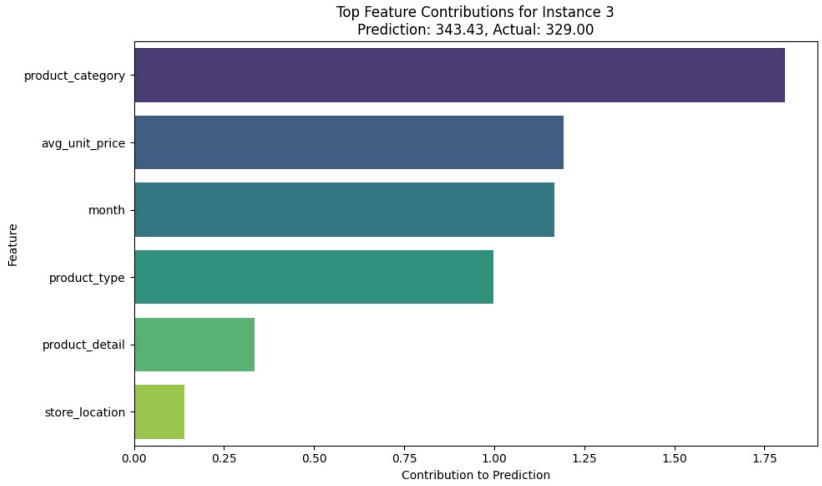


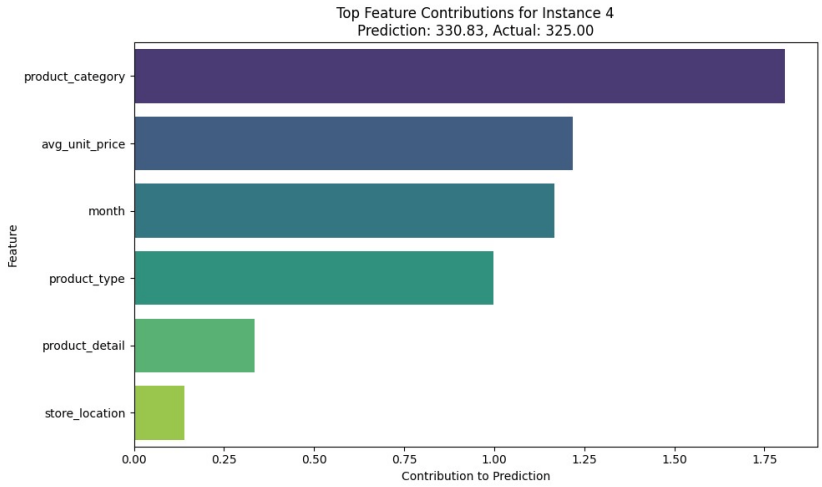
Instanța 3 - Astoria, Bakery,Croissant

Caracteristici principale: Locația Astoria, categoria de produs Bakery, luna Februarie.
Contribuții la predicție: Categoria produsului și locația sunt esențiale.
Explicații: Patiseriile se vând cel mai bine în lunile reci, asociate cu băuturi calde.

Instanța 4 - Hell's Kitchen, Coffe, Latte

Caracteristici principale: Locația Hell's Kitchen, tipul de produs Coffe (Latte)
Contribuții la predicție: Tipul și subtipul produsului sunt dominante.
Epxlicații: Latte-urile sunt populare în această zonă, indiferent de preț.





6 Concluzii

Acest proiect de analiză a demonstrat cu succes capacitatea de a analiza și modela datele de vânzări ale cafenelei Maven Roasters pentru a extrage informații și a dezvolta modelele predictive. Obiectivul principal de a prezice cel mai vândut produs per locație a fost atins prin aplicarea metodelor statistice și a algoritmilor de machine learning.

Eficiența Modelelor predictive:

Random Forest s-a dovedit a fi cel mai performant model pentru predicția cantităților vândute, atingând o acuratețe R^2 de 0.97, conform evaluărilor finale din proiect. Acest model a reușit să captureze eficient relațiile complexe și nelineare dintre variabile, precum sezonalitatea și preferințele specifice locațiilor.

Regresia Logistica a fost utilizată cu succes pentru a clasifica produsele în "populare" și "mai puțin populare", obținând o acuratețe de 0.91. Acest model oferă o perspectivă valoroasă asupra factorilor care determină popularitatea unui produs.

Regresia Liniara a servit ca un model de bază, oferind o primă perspectivă asupra relațiilor liniare din date, cu o acuratețe R^2 de 0.51.

În concluzie, acest proiect a transformat datele brute de tranzacții într-un set de informații acționabile, capabile să sprijine procesul decizional și să contribuie la creșterea eficienței operaționale a cafenelelor Maven Roasters.

Link-ul catre baza de date(coffe-shop-sales-revenue.csv[1])

Link-ul catre proiectul de pe Git (Coffe-shop-sales[2])

[1] coffe-shop-sales-revenue.csv
[2] GitHub

References

[1] Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

[2] Norman R Draper and Harry Smith. *Applied regression analysis*, volume 326. John Wiley & Sons, 1998.

[3] Michael Friendly. Corrgrams: Exploratory displays for correlation matrices. *The american statistician*, 56(4):316–324, 2002.

[4] David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant. *Applied logistic regression*. John Wiley & Sons, 2013.

[5] Julián Luengo, Diego García-Gil, Sergio Ramírez-Gallego, Salvador García, and Francisco Herrera. Big data preprocessing. *Cham: Springer*, 1:1–186, 2020.

[6] Regina L Nuzzo. Histograms: A useful data analysis visualization. *PM&R*, 11(3):309–312, 2019.

[7] J. Ross Quinlan. Induction of decision trees. *Machine learning*, 1:81–106, 1986.

[8] Erhard Rahm, Hong Hai Do, et al. Data cleaning: Problems and current approaches. *IEEE Data Eng. Bull.*, 23(4):3–13, 2000.

[9] Thitithep Sitthiyot and Kanyarat Holasut. A simple method for estimating the lorenz curve. *Humanities and Social Sciences Communications*, 8(1):1–9, 2021.

[10] Jimin Tan, Jianan Yang, Sai Wu, Gang Chen, and Jake Zhao. A critical look at the current train/test split in machine learning. *arXiv preprint arXiv:2106.04525*, 2021.