توسعه مدل چهاربعدی RFMSA: ترکیب هوشمند شاخص‌های رفتار خرید و احساسات مشتریان اینترنتی

سعید میریچگانی[۱]، رضا قاسمی یقین[*۲]

[۱] دانشجوی کارشناسی، دانشکده مهندسی صنایع و سیستم‌های مدیریت، دانشگاه صنعتی امیرکبیر: saeedmirich@aut.ac.ir

[*۲] نویسنده مسئول، استادیار، دانشکده مهندسی صنایع و سیستم‌های مدیریت، دانشگاه صنعتی امیرکبیر: yaghin@aut.ac.ir

چکیده

بخش‌بندی، از الزامات تصمیم‌گیری و تدوین راهبردهای بازاریابی متناسب با گروه‌های مشتریان است. اما استفاده از شاخص‌های رفتاری مدل RFM(تازگی، تکرار و ارزش خرید)، جنبه‌های احساسی نظرات مشتریان را نادیده می‌گیرد. این پژوهش با هدف ارائه چارچوب چهاربعدی RFMSA، مدلی ترکیبی مبتنی بر شاخص‌های RFM و تحلیل احساسات ارائه می‌کند. پس از پیش‌پردازش داده‌ها، شاخص‌های RFM محاسبه و بخش‌بندی مشتریان انجام شد. سپس، رگرسیون لجستیک برای بررسی قابلیت بازتولید گروه‌های حاصل از RFM به کار رفت. در ادامه، امتیاز احساسات از متن نظرات استخراج و در سطح مشتری تجمیع و به‌عنوان بعد چهارم به شاخص‌های RFM افزوده شد. مشتریان بر اساس چهار ویژگی با الگوریتم کا-میانگین خوشه‌بندی شدند. نتایج نشان داد افزودن بعد احساسات، اطلاعات رفتاری و احساسی مشتریان را ترکیب کرده و به شناسایی گروه‌های متمایزتر منجر می‌شود. بنابراین، RFMSA می‌تواند مبنایی برای تحلیل جامع‌تر مشتریان و تدوین راهبردهای بازاریابی متناسب با ویژگی‌های هر گروه فراهم کند.

واژه‌های کلیدی: بخش‌بندی مشتریان، تحلیل احساسات، مدل RFM ، آنالیز واریانس، کا-میانگین، رگرسیون لجستیک

Developing a Four-Dimensional RFMSA Model: An Intelligent Integration of Purchase Behavior Indicators and Customer Sentiment in E-Retailers

Saeed Mirichgani[1] , R.Ghasemy Yaghin[*2]

[1] B.Sc. Student, Department of Industrial Engineering and Systems Management, Amirkabir University of Technology (Tehran Polytechnic), Tehran, Iran : saeedmirich@aut.ac.ir

[*2] Corresponding Author, Assistant Professor, Department of Industrial Engineering and Systems Management, Amirkabir University of Technology (Tehran Polytechnic), Tehran, Iran : yaghin@aut.ac.ir

Abstract

Customer segmentation is a prerequisite for evidence-based decision-making and for designing marketing strategies aligned with the heterogeneous characteristics of customer groups. The traditional RFM model, based on Recency, Frequency, and Monetary value, has been widely used for this purpose; however, it focuses on transactional indicators and does not capture the emotional and experiential information expressed in customer reviews. This study develops the four-dimensional RFMSA framework, which integrates RFM with sentiment analysis to incorporate behavioral and emotional information into a unified segmentation process. Following data preprocessing and exploratory analysis, RFM indicators were calculated and customers were initially segmented using predefined thresholds. Logistic regression was then applied to examine the reproducibility and internal consistency of the RFM groups. Since the target labels were defined by the same RFM thresholds rather than clustering outputs, the results were interpreted as a consistency check rather than independent validation. Subsequently, sentiment scores were extracted from customer reviews using the VADER sentiment analysis tool and aggregated at the customer level. The aggregated sentiment score was normalized to the [0,1] interval and incorporated as a fourth dimension alongside the three RFM indicators to construct the RFMSA framework. Customers were then clustered based on the four RFMSA features using the K-means algorithm, with the number of clusters assessed using

the elbow method and silhouette index. Analysis of variance (ANOVA) and Tukey's test were used to examine differences in sentiment across the resulting groups. The findings indicate that incorporating sentiment provides information not captured by traditional RFM indicators and enables behavioral and emotional characteristics to be considered simultaneously. In particular, customers with high monetary value and negative sentiment, as well as those with moderate monetary value and positive sentiment, could be distinguished through the integrated framework. Overall, RFMSA provides a structured approach for combining transactional and textual customer information, offering a more differentiated representation of customer profiles and supporting the development of marketing strategies aligned with the characteristics of different customer groups.

Keywords: Customer segmentation, Sentiment analysis, RFM model, ANOVA, K-means clustering, Logistic regression

۱- مقدمه و معرفی

بخش‌بندی[۱] مشتریان به‌عنوان ابزاری کلیدی برای طراحی استراتژی‌های بازاریابی هدفمند، نقشی حیاتی ایفا می‌کند[1]. یکی از مدل‌های پرکاربرد در این حوزه، مدل RFM[۲] است که مشتریان را بر اساس سه شاخص تازگی خرید[۳]، تکرار خرید[۴] و ارزش پولی خرید[۵] تقسیم‌بندی می‌کند[2]. با وجود مزایای متعدد، این مدل صرفاً به داده‌های تراکنشی[۶] تکیه دارد و توجهی به جنبه‌های کیفی رفتار مشتریان، نظیر احساسات و نگرش آنان نسبت به برند و محصولات ندارد. در فضای رقابتی امروز، تنها بررسی تعداد خرید و ارزش پولی آن‌ها کافی نیست، بلکه درک احساسات و تجربیات مشتری نیز برای طراحی استراتژی‌های کارآمد بازاریابی ضروری به نظر می‌رسد[3]. از سوی دیگر، با گسترش بسترهای دیجیتال، مشتریان نظرات و تجربیات خود را به‌صورت گسترده در قالب متون آنلاین منتشر می‌کنند[4]. این نظرات حاوی اطلاعات ارزشمندی درباره میزان رضایت یا نارضایتی، انتظارات و دیدگاه‌های مشتریان است[4]. تحلیل این داده‌های متنی، که در قالب تحلیل احساسات[۷] انجام می‌شود، می‌تواند اطلاعاتی درباره نگرش و تجربه مشتریان استخراج کند.

در این پژوهش، متن نظرات به‌طور مستقیم وارد فرآیند بخش‌بندی نمی‌شود، بلکه ابتدا با استفاده از ابزار تحلیل احساسات به یک امتیاز عددی تبدیل و در سطح مشتری تجمیع می‌شود. این امتیاز در کنار شاخص‌های RFM قرار می‌گیرد و چارچوب چهاربعدی RFMSA را شکل می‌دهد. سپس مشتریان بر اساس این مجموعه ویژگی‌ها بخش‌بندی شده و گروه‌های حاصل از نظر الگوهای رفتاری ـ احساسی تحلیل می‌شوند. سؤالات اصلی پژوهش عبارت‌اند از: آیا گروه‌های حاصل از بخش‌بندی مبتنی بر RFM قابلیت بازتولید مناسبی دارند؟ آیا افزودن امتیاز احساسات می‌تواند اطلاعات مکملی درباره ویژگی‌های مشتریان فراهم کند؟ و آیا ترکیب این اطلاعات در چارچوب RFMSA به شناسایی گروه‌های متمایزتری از مشتریان منجر می‌شود؟

برای پاسخ به این سؤالات، ابتدا بخش‌بندی مبتنی بر RFM انجام و قابلیت بازتولید گروه‌های حاصل با استفاده از رگرسیون لجستیک بررسی شد. سپس، ویژگی‌های RFMSA با استفاده از امتیاز احساسات و شاخص‌های RFM تشکیل و مشتریان با الگوریتم کا-میانگین بخش‌بندی شدند. در ادامه، روابط میان شاخص‌های رفتاری و احساسات با تحلیل همبستگی بررسی شد و ساختار ویژگی‌ها با تحلیل مؤلفه‌های اصلی[۸] مورد مطالعه قرار گرفت. همچنین، تفاوت میان گروه‌های حاصل با استفاده از تحلیل واریانس[۹] و آزمون توکی[۱۰] بررسی و ویژگی‌های هر گروه از منظر رفتاری و احساسی تفسیر شد.

نوآوری اصلی پژوهش در ارائه چارچوب چهاربعدی RFMSA و یکپارچه‌سازی اطلاعات رفتاری و احساسی مشتریان در فرآیند بخش‌بندی است. در این چارچوب، علاوه بر شناسایی گروه‌های مشتری بر اساس رفتار خرید، امکان تشخیص تفاوت‌هایی فراهم می‌شود که با شاخص‌های RFM به‌تنهایی قابل مشاهده نیستند؛ مانند تمایز میان مشتریان با ارزش خرید بالا و احساسات منفی یا مشتریان با ارزش خرید متوسط و نگرش مثبت. همچنین، استفاده همزمان از رگرسیون لجستیک برای بررسی قابلیت بازتولید گروه‌های RFM و RFMSA، کا-میانگین برای بخش‌بندی

---

۱ Segmentation
۲ Recency- Frequency- Monetary
۳ Recency
۴ Frequency
۵ Monetary
۶ Transactional Data
۷ Sentiment Analysis
۸ Principal Component Analysis(PCA)
۹ ANOVA
10 Tukey HSD

RFMSA، تحلیل همبستگی، تحلیل مولفه‌های اصلی و آزمون‌های آماری برای بررسی و تفسیر گروه‌ها، یک چارچوب تحلیلی منسجم برای شناسایی و تفسیر الگوهای رفتاری ـ احساسی مشتریان فراهم می‌کند و زمینه استفاده از نتایج در تصمیم‌های مدیریتی و بازاریابی را بهبود می‌دهد .ساختار پژوهش در ادامه به این صورت است که در بخش دوم مرور پیشینه تحقیق، در بخش سوم روش‌شناسی و مبانی نظری، در بخش چهارم یافته‌های پژوهش و در بخش پایانی نتیجه‌گیری و پیشنهادات آینده مطرح خواهند شد.

۲- پیشینه تحقیق

خلاصه پژوهش‌های مرتبط در جدول ۱ نشان داده شده است. در ادامه به بررسی بیشتر آن‌ها پرداخته می‌شود. این مطالعات شامل مقالاتی هستند که به موضوعات RFM و تحلیل احساسات و مدل‌های ترکیبی آن‌ها پرداخته‌اند.

در سال ۱۳۹۸ در پژوهش یوسفی‌زاد، مشتریان بانک ملت با مدل RFM و خوشه‌بندی کا-میانگین[11] طبقه‌بندی شدند و سلسله‌مراتب خدمات دلخواه تعریف گردید[5]. در سال ۱۴۰۰ در پژوهش فتحیان و همکاران، با استفاده از مدل RFM و تحلیل احساسات، تاثیر احساسات انتشاریافته مشتریان بر میزان خرید سایر مشتریان و افزایش تعداد مشتریان وفادار و تازه‌وارد بررسی شد و مشتریان خوشه‌بندی شدند، قواعد رفتاری برای آن‌ها نشان‌دهنده تاثیر مثبت احساسات مشتریان فعال بر افزایش خرید سایر مشتریان بوده است[6].

در سال ۲۰۲۱ در پژوهش اوزکان و کوچاکوچ[12] مدل پیشرفته RFM-V معرفی شد که بعد تنوع خرید را برای تحلیل روابط مشتری-محصول وارد کرد[7]. در سال ۲۰۲۴ در پژوهش وانگ و همکاران[13] مدلی با نام LRFMS پیشنهاد شد که امتیاز رضایت و طول دوره را هم به RFM افزود و از خوشه‌بندی سری‌های زمانی چندبعدی استفاده کرد[8]. در سال ۱۴۰۰ در پژوهش شومالی و همکاران تحلیل احساسات نقدهای آنلاین یک فروشگاه بزرگ ایرانی با استفاده از شبکه‌های عصبی[14] و مدل‌های توالی‌محور[15] و همچنین روش‌های تعبیه‌ی واژگانی[16] انجام شد[9]. در سال ۱۴۰۳ در پژوهش خدایی و همکاران پیشنهاد یک مدل کارآمد برای تحلیل جملات فارسی ارائه شده که عملکرد قابل‌قبولی بر مجموعه‌داده‌های فارسی نشان می‌دهد[10]. در سال ۱۴۰۴ در پژوهش رحیمی و قاسمی یقین با استفاده از تکنیک‌های یادگیری ماشین[17] از جمله رگرسیون لجستیک[18] ، جنگل تصادفی[19] و الگوریتم TF-IDF[20] برای تحلیل احساسات استفاده شد و در نهایت بهترین عملکرد که مربوط به الگوریتم جنگل تصادفی بود معرفی شد[11]. در سال ۲۰۲۴ در پژوهش ملیک و بلال[21] کاربرد روش‌های پردازش زبان طبیعی[22] برای تحلیل بازخورد مشتریان در پلتفرم‌های تجارت الکترونیک تشریح و ارزیابی شده است[12]. در سال ۲۰۲۴ در پژوهش پتراتوس و یانولا[23] کاربردهای مدل‌های مبتنی بر ترنسفورمر[24] در بازاریابی و تحلیل احساسات بررسی و نمونه‌های کاربردی ارائه شده است [13]. در سال ۲۰۲۴ در پژوهش یین و همکاران[25] پیاده‌سازی و تنظیم دقیق مدل‌های برت[26] برای تحلیل احساسات کاربران در پلتفرم‌های تجارت الکترونیک گزارش شده است و نتایج مقایسه‌ای با مدل‌های سنتی ارائه گردیده است[14]. در سال ۲۰۲۰ در پژوهش موسوی و امیری‌عقدایی، عناصر سازنده ارزش پیشنهادی مشتری با استفاده از تحلیل احساسات متن‌کاوی بررسی شد؛ نویسندگان پیشنهاد دادند این شاخص‌ها در کنار داده‌های تراکنشی RFM برای اتخاذ تصمیمات بازاریابی استفاده شوند [15]. در سال ۲۰۲۳ در پژوهش هو و همکاران[27] یک مدل گسترش‌یافته (RFMD) ارائه شد که RFM را با ویژگی‌های جمعیت‌شناختی ترکیب می‌کند و در پژوهش بیان شده که افزودن ویژگی‌های غیرتراکنشی (مانند نتایج تحلیل نظرات) می‌تواند دقت را افزایش دهد[16]. در سال ۲۰۲۳ توسط فنگ و لی[28] در پژوهشی کاربردی روش ترکیبی معرفی شد که ویژگی‌های RFM را با امتیازهای احساس استخراج‌شده از بازخورد مشتریان ترکیب و برای پیش‌بینی ریزش و ارزش طول عمر مشتری به‌کار برد [17]. در سال ۲۰۲۵

11 K-means
12 Ozkan, Pinar. and Deveci-Kocakoc, Ipec.
13 Wang, Shuhai et al.
14 Convolutional Neural Network (CNN)
15 Bidirectional Long Short-Term Memory (BiLSTM)
16 fastText Embeddings
17 Machine Learning
18 Logistic Regression
19 Random Forest
20 Term Frequency – Inverse Document Frequency
21 Malik, Nadia. and Bilal, Muhammad.
22 Natural Language Processing
23 Petratos, Pythagoras. and Giannoula, Mina.
24 Transformer
25 Yin, Huiyang. et al.
26 BERT
27 Ho, Thanh. et al.
28 Feng, Guocheng. et al.

آنیتا و شرلی[۲۹] در پژوهشی با استفاده از ویژگی‌های استخراج‌شده از مدل RFM و تحلیل احساسات مشتریان، یک مدل GraphSAGE توسعه دادند تا ریزش مشتریان را پیش‌بینی کنند و دقت پیش‌بینی را با ترکیب اطلاعات رفتاری و احساسات مشتریان بهبود بخشند[18].

مرور مطالعات نشان می‌دهد که پژوهش‌های پیشین اغلب در سه مسیر کلی توسعه یافته‌اند: نخست، توسعه مدل RFM با افزودن ابعاد رفتاری و اطلاعاتی جدید، مانند تنوع خرید، ویژگی‌های جمعیت‌شناختی و شاخص‌های زمانی [16،8،7]؛ دوم، استفاده از تحلیل احساسات برای استخراج اطلاعات نگرشی و تجربی مشتریان و بررسی بازخوردهای متنی [15–9]؛ و سوم، ترکیب اطلاعات RFM و احساسات در برخی مسائل پیش‌بینی‌محور، از جمله پیش‌بینی ریزش مشتری و ارزش طول عمر[17،18]. در میان مطالعات نزدیک‌تر، فتحیان و همکاران نیز RFM و تحلیل احساسات را در تحلیل مشتریان به کار گرفته‌اند، اما تمرکز اصلی آن‌ها بر بررسی اثر احساسات مشتریان بر رفتار سایر مشتریان و استخراج قواعد رفتاری بوده است[6]. همچنین، موسوی و امیری‌عقدایی بر استفاده همزمان از اطلاعات متنی و RFM برای تصمیمات بازاریابی تأکید کرده‌اند، بدون آنکه یک چارچوب مشخص برای بخش‌بندی مشتریان بر مبنای ترکیب این دو بعد ارائه کنند[15].

بر اساس مطالعات پیشین، جایگاه پژوهش حاضر در تمرکز همزمان بر دو بعد رفتاری و احساسی در فرآیند بخش‌بندی مشتریان است. در این پژوهش، شاخص‌های RFM در کنار امتیاز احساسات حاصل از نظرات مشتریان قرار گرفته و در قالب یک چارچوب یکپارچه با عنوان RFMSA مورد استفاده قرار می‌گیرند. برخلاف مطالعاتی که ترکیب RFM و احساسات را بیشتر به‌عنوان متغیرهای ورودی برای مسائل پیش‌بینی مانند ریزش مشتری به کار برده‌اند، هدف اصلی پژوهش حاضر شناسایی و تفکیک گروه‌های مشتری بر اساس ویژگی‌های همزمان رفتاری و احساسی و تفسیر مدیریتی و بازاریابی این گروه‌هاست. به این منظور، پس از ایجاد شاخص ترکیبی، مشتریان با استفاده از خوشه‌بندی بخش‌بندی شده و ویژگی‌های رفتاری ـ احساسی هر گروه برای شناسایی تفاوت‌های میان مشتریان و تدوین راهبردهای متناسب بازاریابی و مدیریت ارتباط با مشتری مورد بررسی قرار گرفت. بنابراین، پژوهش حاضر تلاش کرده شکاف میان تحلیل صرفاً تراکنشی مشتریان و تحلیل صرفاً متنی را با ارائه یک چارچوب یکپارچه برای بخش‌بندی رفتاری ـ احساسی مشتریان پوشش دهد.

جدول ۱: خلاصه پژوهش‌های مرتبط

Study	Data / Domain	Objective & RFM/Other Dimensions	Methods
Yousefizad, Amir., Sorayaei, Ali [5]	Bank Mellat customer data	Customer segmentation based on RFM	K-Means, DBSCAN, Fuzzy Hierarchical Analysis
Fathian, Mohammad. et al. [6]	About 6,100 transactions of Shixon online store on Instagram in 2017	Customer segmentation using RFM and behavioral rules	K-Means, Association Rules
Ozkan, Pinar., Deveci-Kocakoc, Ipec. [7]	Retail data	Customer segmentation using RFM and purchase variety	RFM-V
Wang, Shuhai. et al. [8]	Transactional data + multivariate time series	Dynamic customer segmentation using RFM and temporal/satisfaction dimensions	LRFMS, Time-Series Clustering

[29] Anitha, M. A. and Sherly, K.

Study	Data / Domain	Objective & RFM/Other Dimensions	Methods
Shumaly, Sajjad. et al. [9]	Persian Online Reviews (Digikala, ~3M review)	Sentiment analysis of customer reviews	CNN, FastText, BiLSTM, LR, NB
Khodaei, Azadeh. et al. [10]	Persian story sentences	Sentiment analysis	Deep Learning
Rahimi, Razieh., Ghasmy Yaghin, Reza. [11]	Women's clothing reviews	Sentiment analysis	Logistic Regression ،Random Forest ،Naïve Bayes (Multinomial/Complement) ،TF-IDF
Malik, Nadia., Bilal, Muhammad. [12]	Studies on NLP in customer feedback	Review of NLP/ML approaches	NLP & ML Methods
Petratos, Pythagoras., Giannoula, Mina. [13]	Review of Applications in Marketing	Sentiment analysis in marketing	Transformer-Based Models (BERT, GPT)
Yin, Huiyang. et al. [14]	E-commerce user reviews	Sentiment analysis	BERT + Traditional Models
Mousavi, Mohsen., Amiri Aghdaie, Fathollah. [15]	Customer reviews + RFM data	Customer value proposition using RFM and sentiment	Sentiment Analysis, Text Mining
Ho, Thanh. et al. [16]	Transactional + demographic data	Customer segmentation using RFM and demographic attributes	RFMD (RFM + Demographic)
Feng, Guocheng. et al. [17]	Election data + online feedback	Decay prediction / customer lifetime value	RFM + Sentiment, MLP, GBM
Anitha, M. A., Sherly, K. [18]	Transactional data + sentiment scores	Customer churn prediction	GraphSAGE + RFM & Sentiment
Current Study	**Amazon customer reviews**	**Behavioral–sentiment customer segmentation using RFM + Sentiment → RFMSA**	**K-Means, Logistic Regression, RFM and Sentiment Features**

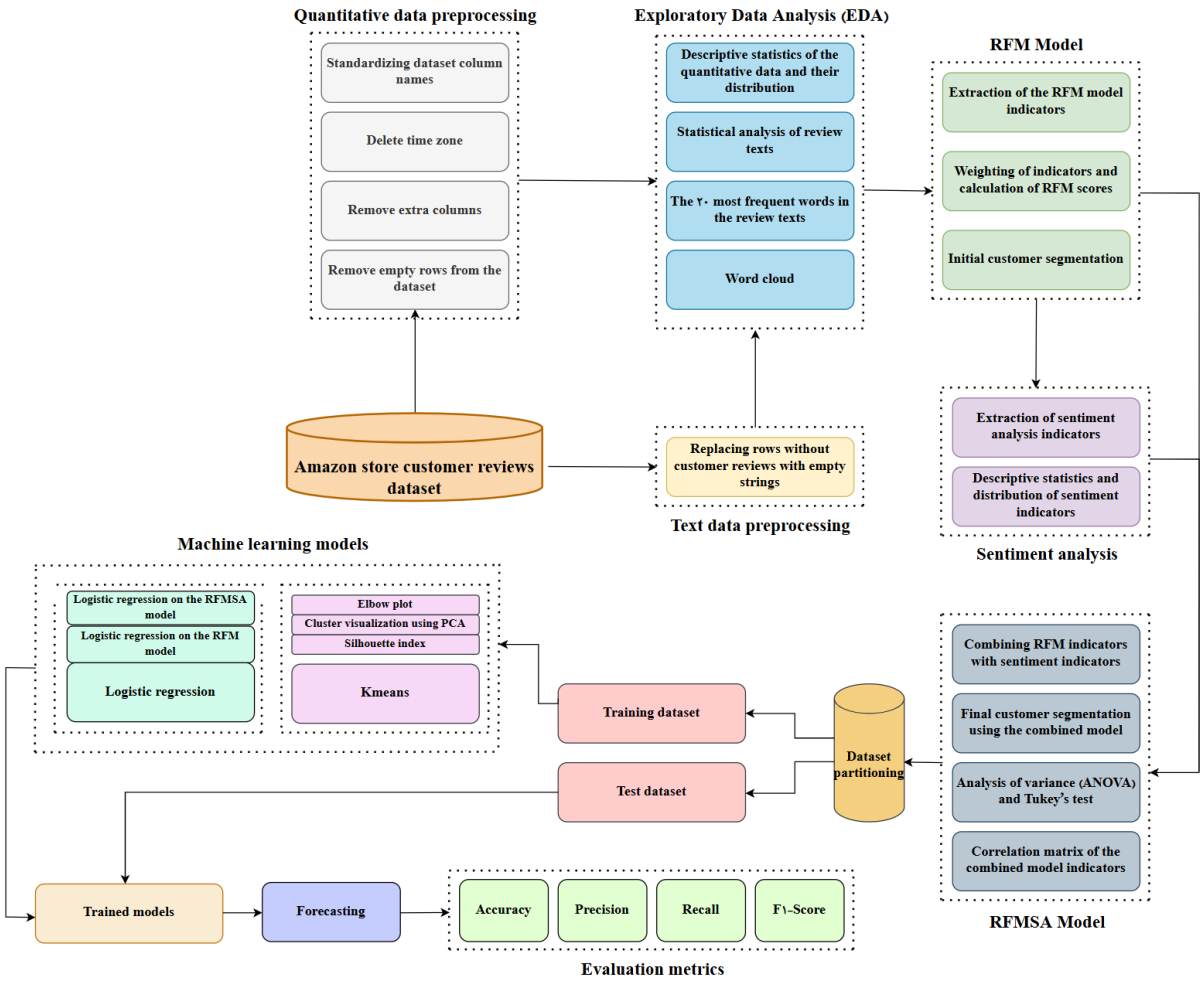
۳- روش‌شناسی و مبانی نظری

در شکل ۱ روش مطالعه انجام‌شده نشان داده شده است. در این مطالعه پس از انتخاب مجموعه داده مربوط به نظرات مشتریان

فروشگاه آمازون[30]، به انجام پیش‌پردازش[31] داده‌های کمی و کیفی پرداخته می‌شود. در مرحله بعد برای درک مجموعه داده تجزیه و تحلیل اکتشافی[32] انجام می‌شود. در بخش بعد، شاخص‌های RFM محاسبه شده و متن نظرات مشتریان با استفاده از ابزار VADER به امتیاز احساسات تبدیل می‌شود. سپس امتیازهای احساسات مشتری تجمیع شده و در کنار شاخص‌های RFM قرار می‌گیرند تا مدل ترکیبی RFMSA شکل گیرد. پس از آن مجموعه داده به داده‌های آموزشی[33] و آزمایشی[34] تفکیک می‌شود و برای ارائه استراتژی‌های بازاریابی نیز الگوریتم کا-میانگین بر روی داده‌های ترکیبی پیاده‌سازی می‌شود. برای بررسی قابلیت بازتولید گروه‌های مشتری در مدل‌های RFM و RFMSA از الگوریتم رگرسیون لجستیک استفاده می‌شود. برچسب‌های هدف بر اساس آستانه‌های از پیش تعیین‌شده RFM ایجاد می‌شوند و از خروجی کا-میانگین به‌عنوان برچسب استفاده نمی‌شود. بنابراین، نتایج رگرسیون لجستیک به‌عنوان اعتبارسنجی مستقل بخش‌بندی تفسیر نمی‌شوند.

۳-۱- مجموعه داده و ستون‌های انتخابی

مجموعه داده مورد استفاده در این مطالعه شامل داده‌های نظرات مشتریان فروشگاه آمازون است که از سایت کگل[35] جمع‌آوری شده است. مجموعه داده شامل ۳۴۵۹۰ رکورد و ۲۱ ستون ویژگی است که از میان این ۲۱ ستون، ۵ ستون تاریخ نظردهی مشتری، شناسه خرید، نام کاربری خریدار، امتیاز ستاره‌ای خرید(نشان‌دهنده شاخص ارزش خرید) و ستون نظرات مشتریان مورد استفاده در این مطالعه هستند. از میان ۳۴۵۹۰ رکورد اولیه، ۳۴۵۳۰ رکورد دارای متن نظر ثبت‌شده بودند و مابقی فاقد متن نظرات بودند.



30 Amazon Customer Reviews Dataset
31 Preprocessing
32 Exploratory data analysis
33 Train Data
34 Test Data
35 www.kaggle.com

شکل ۱: روش پژوهش

۳-۲- پالایش داده‌ها

فرایند پاکسازی داده‌ها در دو بخش داده‌های کمی مرتبط با مدل RFM و داده‌های متنی مرتبط با تحلیل احساسات انجام گرفت. در بخش داده‌های کمی، که شامل ستون‌هایی است که در بخش ۳-۱ عنوان شدند، این مراحل پیاده‌سازی شد. نخست داده‌ها و ستون‌های اضافی که می‌توانستند نتایج خوشه‌بندی و محاسبات شاخص‌های RFM را مخدوش کنند حذف شدند. سپس ردیف‌هایی که ستون‌های کلیدی آن‌ها فاقد داده بودند، کنار گذاشته شدند تا محاسبات بدون خطا انجام گیرد. در ادامه، برای یکپارچه‌سازی داده‌ها، نام تمام ستون‌ها به صورت استاندارد و با حروف کوچک تغییر یافتند. از آنجا که تاریخ‌ها در این نوع داده نقش حیاتی دارند، بررسی اعتبار تاریخ‌ها صورت گرفت و ردیف‌هایی که دارای تاریخ‌های نامعتبر بودند حذف شدند. در گام بعدی نیز ستون‌های تاریخ به نوع داده‌ی تاریخ-زمان تبدیل شدند و منطقه زمانی از آن‌ها حذف شد تا تمام داده‌ها بر اساس یک مقیاس یکنواخت پردازش شوند. همچنین در بخش داده‌های متنی ردیف‌هایی که بدون نظر مشتریان بودند با رشته خالی جایگزین شدند.

۳-۳- تجزیه و تحلیل اکتشافی داده‌ها

برای انجام تجزیه و تحلیل اکتشافی داده‌ها، پس از بررسی و حذف مقادیر گمشده، ابتدا محاسبه پارامترهای آمار توصیفی برای ویژگی‌های عددی انجام می‌شود. سپس در این بخش نمودار میله‌ای[۳۶] و هیستوگرام‌های[۳۷] لازم و توزیع‌های[۳۸] لازم نیز ترسیم و بررسی می‌شوند و روی داده‌های متنی نیز بیشترین کلمات پرکاربرد در متن نظرات به‌همراه ابر کلمات[۳۹] ترسیم می‌شوند تا پژوهش را برای مراحل بعدی خود آماده کنند.

۳-۴- مدل RFM و تحلیل احساسات

برای پیاده‌سازی مدل، ابتدا شاخص‌های آن محاسبه می‌شوند. شاخص تازگی خرید بر اساس تعداد روزهای فاصله بین آخرین تاریخ موجود در دیتاست و آخرین تعامل هر کاربر محاسبه می‌شود. شاخص تکرار خرید نیز با شمارش تعداد نظرات ثبت‌شده برای خریدهای هر کاربر به دست می‌آید و در نهایت، شاخص ارزش خرید به‌صورت میانگین امتیازهای ستاره‌ای ثبت‌شده توسط هر کاربر محاسبه می‌گردد.

برای بخش‌بندی اولیه مشتریان و ارزیابی عملکرد مدل، امتیاز این مدل تحت عنوان امتیاز RFM محاسبه می‌شود. برای محاسبه امتیاز RFM اوزان شاخص‌ها بر اساس منطق کسب و کار در نظر گرفته شده است و به ترتیب برای شاخص تازگی خرید وزن ۰/۴، برای شاخص تکرار خرید وزن ۰/۳ و برای شاخص ارزش خرید وزن ۰/۳ در نظر گرفته شده است. سپس با نرمال‌سازی به مقیاس $[0,1]$ امتیاز مدل محاسبه می‌شود تا بتوانیم با استفاده از آن بخش‌بندی اولیه مشتریان را انجام دهیم. در رابطه ۱ نحوه محاسبه امتیاز مدل آورده شده است:

$$RFM_{score} = 0.4 * \mathrm{Re}cency + 0.3 * Frequency + 0.3 * Monetary \quad (1)$$

سپس بر اساس چارک‌های توزیع RFM بخش‌بندی صورت می‌پذیرد. در بخش تحلیل احساسات برای استخراج شاخص احساسات مشتریان، متن نظرات با استفاده از روش تحلیل احساسات و ابزار Vader پردازش می‌شود و امتیاز تجمیعی احساسات هر نظر در بازه ۱- تا ۱ به دست می‌آید. برای تبدیل این امتیاز به بازه ۰ تا ۱، از نرمال‌سازی خطی استفاده می‌شود؛ یعنی امتیاز احساسات قبلی با عدد ۱ جمع و سپس تقسیم بر ۲ می‌شود، بنابراین امتیاز ۱- به صفر، امتیاز صفر به ۰/۵ و امتیاز ۱ به ۱ تبدیل می‌گردد. در ابزار VADER، امتیاز تجمیعی برابر یا بزرگ‌تر از ۰/۰۵ به‌عنوان احساس مثبت، امتیاز برابر یا کوچک‌تر از ۰/۰۵- به‌عنوان احساس منفی و مقادیر بین این دو به‌عنوان احساس خنثی در نظر گرفته می‌شوند. پس با درنظرگرفتن ۰/۵ به‌عنوان نقطه خنثی و ایجاد فاصله ۰/۰۲۵ در دو طرف آن، آستانه‌های ۰/۴۷۵ و ۰/۵۲۵ تعیین می‌شوند؛ به‌طوری‌که مقادیر کمتر از ۰/۴۷۵ دارای احساسات منفی، بین

[36] Bar Chart
[37] Histogram
[38] Distribution
[39] Word Cloud

۰٫۴۷۵ تا ۰٫۵۲۵ خنثی و بیشتر از ۰٫۵۲۵ مثبت در نظر گرفته می‌شوند.

۳-۵- مدل‌های یادگیری ماشین

در این تحقیق برای تحلیل داده‌ها و پاسخ به سؤال‌های پژوهش، از دو الگوریتم رگرسیون لجستیک و کا-میانگین استفاده شده است، زیرا هر کدام مزیت‌ها و کاربردهای متفاوتی در تحلیل داده‌ها ارائه می‌دهند و مکمل یکدیگر هستند[19]. رگرسیون لجستیک یک

$$z_j = \frac{\sum_{x \in c_j} x}{\#c_j} \quad for\ j = 1,\ldots,k \qquad (2)$$

الگوریتم نظارت‌شده[40] است که قادر است رابطه بین ویژگی‌های عددی و طبقه‌بندی‌های مشخص را مدل کند و احتمال تعلق هر نمونه به هر کلاس را بر اساس ویژگی‌ها پیش‌بینی نماید[20].

از سوی دیگر، کا-میانگین یک الگوریتم خوشه‌بندی غیرنظارت‌شده[41] است که توانایی گروه‌بندی مشتریان بر اساس شباهت‌های ذاتی بین ویژگی‌ها را دارد، بدون آنکه نیاز به برچسب یا خروجی از پیش تعیین‌شده باشد[19]. استفاده از این الگوریتم امکان شناسایی الگوهای پنهان و تقسیم‌بندی طبیعی مشتریان را فراهم می‌کند، که می‌تواند به شناسایی گروه‌های مشتری با رفتار و ویژگی‌های مشابه کمک کند. ترکیب این دو الگوریتم به این دلیل است که می‌توان هم نتایج پیش‌بینی محور و قابل تفسیر رگرسیون لجستیک و هم ساختارهای خوشه‌ای و الگوهای نهفته در داده‌های مشتریان از طریق کا-میانگین را به دست آورد[20].

۳-۵-۱- خوشه‌بندی کا-میانگین

الگوریتم کا-میانگین یکی از پرکاربردترین روش‌های خوشه‌بندی است که به دلیل سادگی، سرعت بالا و قابلیت اعمال بر داده‌های بزرگ، در بسیاری از مسائل داده‌کاوی و یادگیری ماشین استفاده می‌شود. هدف اصلی این الگوریتم تقسیم مجموعه داده‌ها به k خوشه است به طوری که نمونه‌های داخل هر خوشه بیشترین شباهت را به هم داشته باشند و بیشترین تفاوت با نمونه‌های خوشه‌های دیگر داشته باشند. الگوریتم بر پایه مفهوم فاصله بین نمونه‌ها و مراکز خوشه‌ها عمل می‌کند و سعی دارد مجموع فاصله مربعی نمونه‌ها تا مرکز خوشه متناظر را کمینه کند. در این الگوریتم، هر خوشه با یک نقطه مرکزی نمایندگی می‌شود که میانگین ویژگی‌های نمونه‌های آن خوشه است. این مرکز خوشه[42] به عنوان نماینده خوشه عمل می‌کند و فاصله نمونه‌ها تا این مرکز برای تعیین عضویت در خوشه استفاده می‌شود. مرکز خوشه می‌تواند تغییر کند، زیرا با اختصاص داده‌های جدید و تکرار الگوریتم، موقعیت مرکز به‌روزرسانی می‌شود تا بهترین نمایندگی از خوشه ایجاد شود[19]. مراحل این الگوریتم به شرح زیر است[19]:

۱- تعیین تعداد خوشه‌ها(k) به صورت تصادفی

۲- انتخاب مرکز اولیه خوشه‌ها با رابطه ۲

در رابطه ۲، x نشان‌دهنده بردار یک نمونه‌ای است که در c_j عضو است و $\#c_j$ نشان‌دهنده تعداد نمونه‌هایی است که در کلاس c_j عضو است. رابطه ۲ برای محاسبه مرکز هر دسته در حین حل استفاده می‌شود و در شروع الگوریتم معمولا به صورت تصادفی k نمونه انتخاب و به عنوان مرکز هر دسته در نظر گرفته می‌شود.

۳- محاسبه تابع هدف ناشی از خوشه‌بندی $c_1,\ldots,c_k$ که بر اساس رابطه ۳ است و به محاسبه مجموع فاصله‌های نمونه‌ها از مرکز خوشه‌ها می‌پردازد.

$$f(c_1,\ldots,c_k) = \sum_{j-1}^{k} \sum_{x \in c_j} |x - z_j|^2 \qquad (3)$$

۴- کمینه کردن تابع هدف رابطه ۳ و یافتن خوشه‌بندی مناسب روی مجموعه با تعداد خوشه k

۳-۵-۲- طبقه‌بندی رگرسیون لجستیک

رگرسیون لجستیک یکی از پایه‌ای‌ترین و در عین حال قدرتمندترین الگوریتم‌های طبقه‌بندی است[20]. علی‌رغم نام «رگرسیون» این روش صرفاً برای مسائل طبقه‌بندی (بویژه دودویی) به کار می‌رود و رابطه بین متغیرهای ورودی و خروجی را به صورت خطی فرض می‌کند[20]. رگرسیون لجستیک از خانواده مدل‌های تمییزی[43] است و پارامترهای آن معمولاً بر اساس کمینه‌کردن تابع هزینه مناسب برآورد می‌شوند[20]. فرآیند طبقه‌بندی بطور کلی شامل مراحل زیر است[20]:

تعریف مسئله و گردآوری داده‌ها، پاک‌سازی و آماده‌سازی داده‌ها، ساخت و انتخاب ویژگی‌ها، تقسیم داده‌ها و طراحی اعتبارسنجی،

40 Supervised
41 Unsupervised
42 Centroid
43 Discriminative Models

انتخاب و آموزش مدل، ارزیابی مدل، انتخاب آستانه و بررسی هزینه‌ها و استقرار و پایش مداوم مدل

مدل خطی آن بصورت زیر تعریف می‌شود:

با توجه به رابطه فوق احتمال تعلق نمونه x به کلاس مثبت $(y=1)$ با تابع سیگموئید[44] رابطه ۵ بدست می‌آید:

رابطه ۵ مقدار پیش‌بینی مدل را تعیین می‌کند. حال میتوان گفت:

$$\text{logit}(p(x)) = \ln\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n = \beta^T x = Z \quad (4)$$

$$\frac{1}{1+e}$$

Metric Name	Calculation Method	Description
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$	The ratio of all correct predictions (both positive and negative) to all predictions. A general measure of the model's overall performance.
Precision	$\frac{TP}{TP+FP}$	Shows what percentage of all samples predicted as positive by the model were actually positive.
Recall	$\frac{TP}{TP+FN}$	The ratio of actual positive samples that the model correctly identified.
F1-Score	$\frac{2*\text{Precision}*\text{Recall}}{\text{Precision}*\text{Recall}}$	The harmonic mean of Precision and Recall, showing the balance between these two metrics.
Confusion matrix	A 2×2 table that contains TP, TN, FP, and FN, showing how the model's predictions compare to the actual classes.	A numerical and visual representation of the model's correct and incorrect predictions for each class; helps highlight the model's weaknesses clearly.

$$j(\beta) = \frac{1}{m}\sum_{i=1}^{m} C_j \quad (8)$$

تابع هزینه با توجه به رابطه ۷ و ۸ حاصل می‌شود:

۳-۵-۳- معیارهای ارزیابی

برای قابلیت بازتولید بخش‌بندی مدل و امکان مقایسه مدل ترکیبی با مدل سنتی معیارهای گوناگونی وجود دارد. دقت کلی[45] ، تنها بخشی از ماجراست. در بسیاری از مسائل، خطای پیش‌بینی مثبت و خطای پیش‌بینی منفی هزینه‌های متفاوتی دارند. به همین دلیل معیارهایی چون صحت[46] ،حساسیت[47] ،امتیاز اف-یک[48] و ماتریس درهم‌ریختگی[49] نیز مورد استفاده قرار می‌گیرند[19]. در جدول ۲

$$C = y(-\log(g(Z)) + (1-y)(-\log(1-g(Z))) \quad (7)$$

معیارهای ارزیابی آورده شده است

جدول ۲: معیارهای ارزیابی

۳-۶-چارچوب پیشنهادی RFMSA

در این بخش ابتدا شاخص امتیاز احساسات به شاخص‌های مدل RFM اضافه می‌شود و مدل چهار بعدی شکل می‌گیرد. پس از آن برای بخش‌بندی مدل ترکیبی لازم است تا امتیاز احساسات با امتیاز RFM ترکیب شود. در اینجا برای ترکیب این دو، میانگین عددی هر دو ستون امتیاز احساسات و امتیاز RFM محاسبه می‌شود، یعنی در این تحقیق هیچکدام بر دیگری ارجحیت ندارند اما می‌توان بسته به شرایط وزن یکی را نسبت به دیگری بیشتر در نظر گرفت. سپس ستون ویژگی‌ای تحت عنوان امتیاز مدل ترکیبی RFMSA شکل

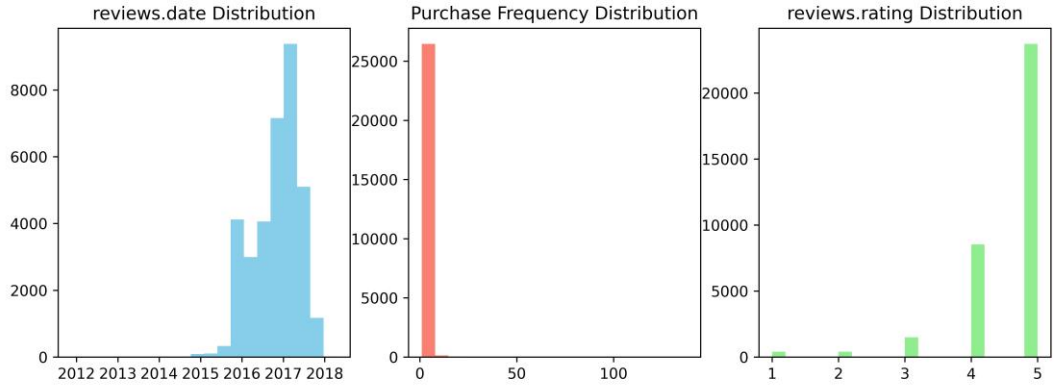
44 Sigmoid Function
45 Accuracy
46 Precision
47 Recall
48 F1-Score
49 Confusion Matrix

می‌گیرد که با استفاده از آن می‌توان بخش‌بندی مشتریان را انجام داد. سپس برای بررسی تفاوت معناداری میانگین احساسات گروه‌های مختلف مشتریان از آزمون‌های آماری استفاده می‌شود.

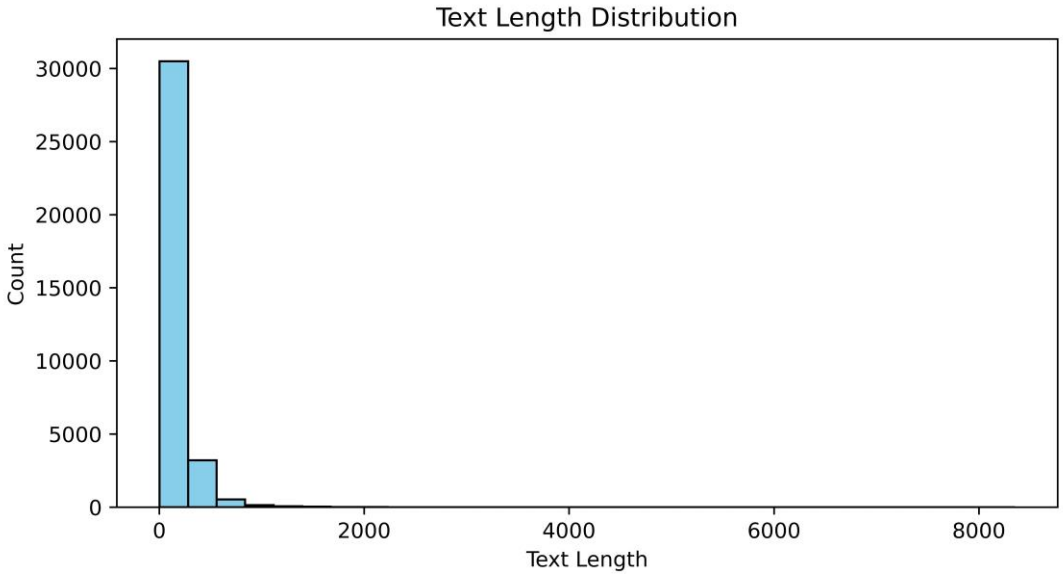
۴- نتایج پژوهش

۴-۱- نتایج پالایش و تحلیل اکتشافی داده‌ها

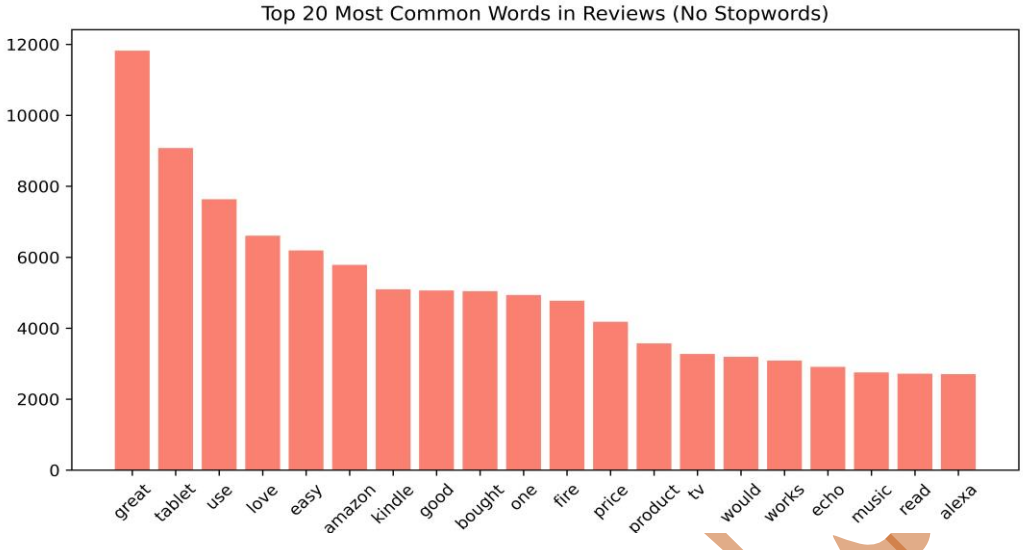
قبل از پیش‌پردازش تعداد رکوردهای مجموعه داده ۳۴۵۹۰ رکورد بود ولی مطابق نتایج آمار توصیفی پس از پیش‌پردازش داده‌های کمی تعداد ردیف‌های مورد بررسی مجموعه داده برابر ۲۶۶۸۸ شده است. میانگین امتیاز مشتریان ۴٫۵۸ و میانگین تکرار خریدهای مشتریان ۱٫۳ و تعداد رکوردهای دارای نظرات متنی پیش از پیش‌پردازش داده‌های متنی ۳۴۵۳۰ بوده است. بیشترین و کمترین بازه برای ویژگی تازگی خرید هم که از ستون تاریخ محاسبه می‌شود ۲۳۴۳ روز از سال ۲۰۱۱ تا ۲۰۱۸ بوده است. این آمار نشان می‌دهد که داده‌ها دارای حجم کافی و تنوع مناسب برای پیاده‌سازی مدل ترکیبی هستند. نتایج پالایش و تحلیل اکتشافی داده‌ها در شکل‌های ۲ تا ۵ آورده شده است.



شکل ۲: توزیع شاخص‌های مدل آر.اف.ام



شکل ۳: توزیع طول نظرات مشتریان





شکل ۵: ابر کلمات برای نظرات مشتریان فروشگاه

شکل ۲ توزیع ستون‌های مربوط به شاخص‌های تازگی خرید(هیستوگرام سمت چپ) و تکرار خرید(هیستوگرام وسط) و ارزش خرید(هیستوگرام سمت راست) را نشان می‌دهد. در شکل ۳ هیستوگرام طول نظرات مشتریان مشخص است که از طول ۳ کاراکتر تا ۸۳۵۱ کاراکتر وجود دارد و میانگین طول نظرات ۱۵۶٫۴ است. شکل ۴ حاوی نمودار میله‌ای ۲۰ کلمه پرکاربرد در نظرات مشتریان است و در شکل ۵ ابر کلمات در نظرات مشتریان قابل مشاهده است. در شکل ۴ و ۵ از کتابخانه پردازش زبان طبیعی[50] و ابزار VADER در پایتون استفاده شده است، این ابزار به‌صورت پیشفرض کلمات توقف[51] انگلیسی را در نظر نمی‌گیرد اما با این حال ۱۰۴ کلمه سفارشی دیگر به کلمات توقف آن اضافه شد که شامل افعال پرکاربرد و مبهم، اسامی مبهم و عمومی، ضمایر نامشخص، قیدهای کم‌اثر و کلمات محاوره‌ای و سایر کلمات چندپهلو و مبهم بوده‌اند تا نتایج بهتر باشد.

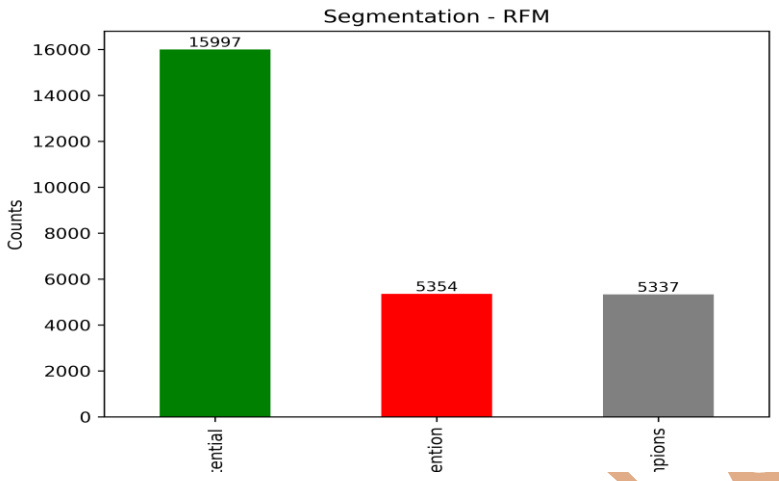
۴-۲- نتایج مدل RFM

بخش‌بندی بر اساس چارک‌های توزیع RFM انجام شده است. به این صورت که اگر امتیاز RFM مشتری بین ۰ تا ۰٫۲ باشد، در گروه مشتریان نیازمند توجه قرار می‌گیرد و اگر امتیاز بین ۰٫۲ تا ۰٫۸ باشد در گروه مشتریان بالقوه قرار می‌گیرد و از ۰٫۸ تا ۱ نیز در گروه

[50] Natural Language Tool Kit(NLTK)

[51] Stop Words

مشتریان طلایی قرار می‌گیرند. نتایج بخش‌بندی در نمودار میله‌ای شکل ۶ نشان داده شده است.



شکل ۶: بخش‌بندی مشتریان فروشگاه آمازون بر اساس مدل

با توجه به این بخش‌بندی، ۱۵۹۹۷ مشتری در گروه مشتریان بالقوه، ۵۳۵۴ مشتری در گروه مشتریان نیازمند توجه و ۵۳۳۷ مشتری در گروه مشتریان طلایی(قهرمان) قرار گرفته‌اند. همانطور که پیشتر گفته شد به‌منظور بررسی قابلیت بازتولید گروه‌های حاصل از مدل RFM از الگوریتم رگرسیون لجستیک استفاده شد که نتایج آن را در جداول ۳ و ۴ برای این مدل مشاهده می‌کنیم:

جدول ۳: گزارش طبقه‌بندی مدل آر.اف.ام

	Precision	Recall	F1-Score	Support
Need Attention	0.99	1.00	0.99	1601
Potential	1.00	0.96	0.98	1606
Champions	0.99	1.00	0.99	4800
accuracy			0.99	8007
Macro avg	0.99	0.99	0.99	8007
Weighted avg	0.99	0.99	0.99	8007

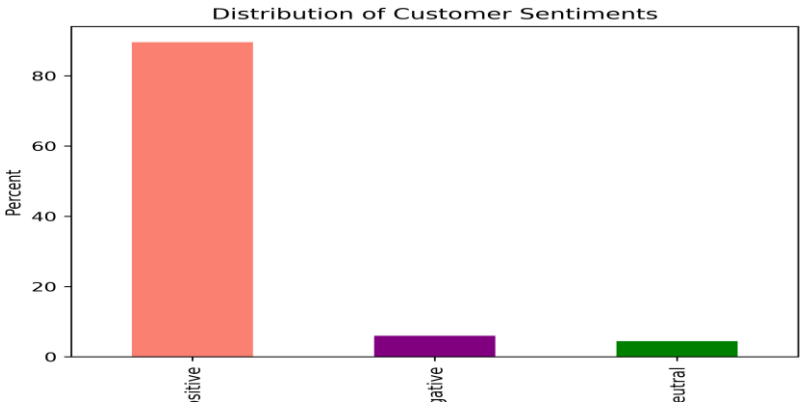
جدول ۴: ماتریس درهم‌ریختگی مدل آر.اف.ام

	NA	P	Ch
Need Attention	1596	0	5
Potential	0	1549	57
Champions	12	0	4788

همان‌طور که از جداول ۳ و ۴ مشخص است، الگوریتم برای مدل RFM دقت کلی ۹۹ درصد را ثبت کرده است، اما برای اینکه بدانیم این عدد تحت تاثیر گروه خاصی نبوده است لازم است ستون تعداد نمونه‌ها و شاخص‌های صحت و حساسیت و امتیاز اف-یک نیز بررسی شود. با بررسی آن‌ها می‌توان دریافت که دقت ۹۹ درصد اتفاقی نبوده است. معیار حساسیت برای گروه بالقوه نشان می‌دهد که حدود ۴ درصد از این گروه به اشتباه در گروه دیگری شمرده شده‌اند که با دقت در ماتریس درهم‌ریختگی می‌توان متوجه شد این ۴ درصد در گروه مشتریان طلایی پیش‌بینی شده‌اند، این موضوع یک فرصت و یک تهدید است. فرصت از این جهت که این اشتباه از مدل نشان می‌دهد که این ۴ درصد خصوصیاتشان آن‌قدر به گروه طلایی نزدیک بوده است که در این گروه پیش‌بینی شده‌اند، بنابراین این همان مرز بین این دو گروه است و با اتخاذ استراتژی مناسب می‌توان این ۴ درصد را واقعا به گروه طلایی برد. تهدید نیز از این جهت می‌تواند باشد که استراتژی‌هایی که برای گروه طلایی در نظر گرفته می‌شوند، برای این ۴ درصد نیز در نظر گرفته شوند و این باعث افزایش هزینه خواهد بود. از جهتی دیگر این اشتباه ممکن است به دلیل پیاده‌سازی الگوریتم صرفا برای مدل سنتی RFM بوده باشد.

۴-۳- نتایج تحلیل احساسات

پس از محاسبه امتیاز احساسات با استفاده از ابزار پردازش زبان طبیعی در پایتون برای هر مشتری و نرمال‌سازی آن به بازه [0,1]، میانگین احساسات مشتریان ۰٫۸۱ که نشان‌دهنده‌ی غلبه‌ی احساسات مثبت در میان کاربران است. انحراف معیار ۰٫۱۸ بوده و بیانگر آن است که اگرچه بیشتر نظرات مثبت هستند، اما تنوع قابل توجهی نیز در میان ارزیابی‌ها وجود دارد. مقادیر حداقل و حداکثر امتیاز احساسات نیز از ۰٫۰۱۷ تا ۰٫۹۹۹ متغیر است که نشان می‌دهد طیف وسیعی از احساسات، از نارضایتی شدید تا رضایت کامل، در میان داده‌ها ثبت شده است. برای دسته‌بندی بهتر، شاخص احساسات به سه گروه مثبت، منفی و خنثی تقسیم شد که نحوه محاسبه امتیاز این ۳ گروه در بخش ۳-۴ تشریح شد. در شکل ۷ نمودار میله‌ای توزیع امتیاز احساسات مشتریان نشان داده شده است. این امتیازها از متن نظرات استخراج و در سطح هر مشتری تجمیع شدند و سپس به‌عنوان بعد احساسی مدل RFMSA در کنار شاخص‌های RFM مورد استفاده قرار گرفتند.

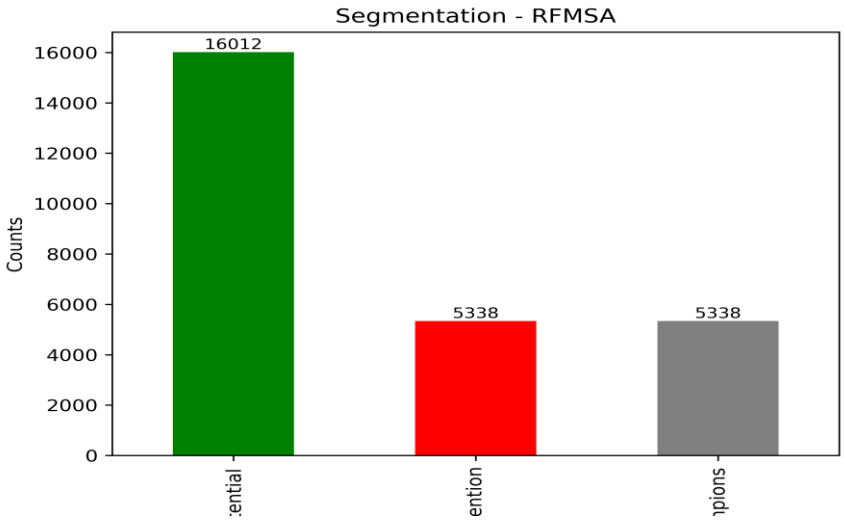


شکل ۷: توزیع احساسات مشتریان فروشگاه آمازون

۴-۴- نتایج مدل ترکیبی RFMSA

توزیع احساسات مشتریان نشان می‌دهد حدود ۸۹٫۵۷ درصد مشتریان دارای احساس مثبت(رنگ صورتی در شکل ۷) نسبت به برند هستند، ۵٫۹۹ درصد مشتریان دارای احساس منفی(بنفش) و ۴٫۴۴ درصد دارای احساس خنثی(سبز) هستند.

بخش‌بندی مشتریان برای مدل ترکیبی در شکل ۸ نشان داده شده است.



شکل ۸: بخش‌بندی مشتریان فروشگاه آمازون بر اساس مدل

همان‌طور که مشخص است در این بخش‌بندی تعداد مشتریان گروه بالقوه ۱۶۰۱۲، گروه مشتریان نیازمند توجه ۵۳۳۸ و گروه مشتریان قهرمان یا طلایی ۵۳۳۸ شده است. حال می‌توان به یافته‌های بعدی پژوهش پرداخت.

برای بررسی تفاوت معناداری میانگین احساسات گروه‌های مشتریان از تحلیل واریانس[۵۲] و آزمون توکی[۵۳] استفاده شد که نتایج آن در جدول ۵ نشان داده شده است.

جدول ۵: نتایج آنالیز واریانس میانگین احساسات و آزمون توکی

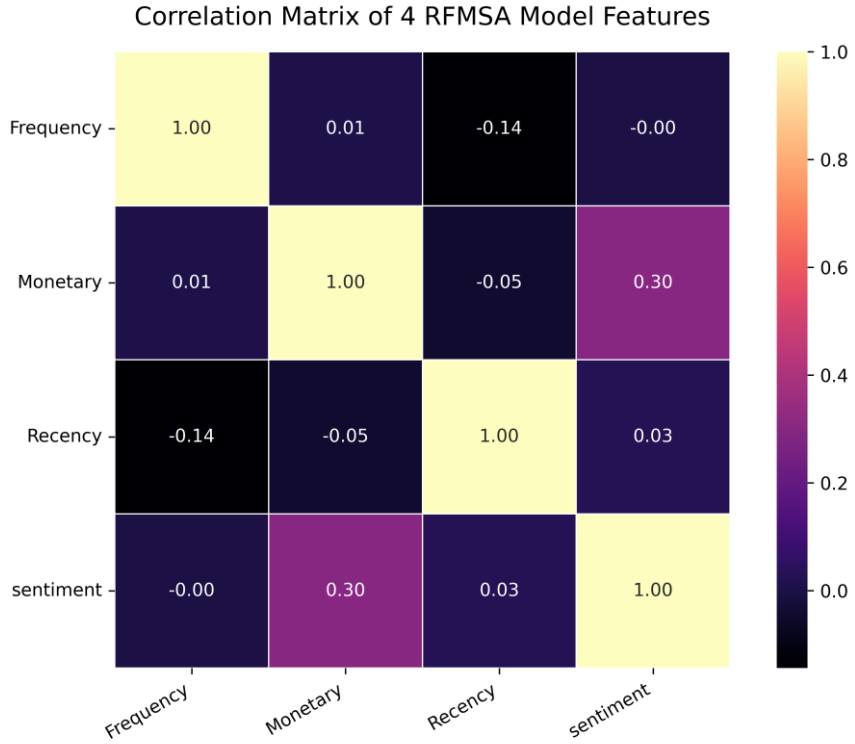
F-statistic: 30686.20383329558

p-value: 0.0

Multiple Comparison of Means - Tukey HSD, FWER=0.05

Gp1	Gp2	meandiff	lower	upper	reject
Ch	NA	-0.4235	-0.4279	-0.4191	True
Ch	P	-0.0896	-0.0931	-0.0860	True
NA	P	0.3339	0.3303	0.3375	True

نتایج به این شرح است که مقدار آماره اف[۵۴] ۳۰۶۸۶٫۲۰ و مقدار سطح معناداری[۵۵] نزدیک به صفر(کمتر از ۰٫۰۰۱) گزارش شده است. این امر به‌روشنی بیانگر آن است که اختلاف میانگین احساسات در میان این سه گروه واقعی است و نمی‌توان آن را ناشی از شانس یا تصادف دانست. از طرفی با استفاده از نتایج آزمون توکی نیز مشخص شد که هر سه گروه به‌صورت دو به دو اختلاف معناداری باهم دارند. این نتایج معنای مهمی برای تصمیمات مدیریتی و بازاریابی دارند. زمانی که یک کسب‌وکار می‌داند گروهی از مشتریان از نظر احساسات به شکل معناداری وضعیت ضعیف‌تری دارند، می‌تواند سیاست‌های ویژه‌ای برای بهبود تجربه آن‌ها طراحی کند. برای مثال، مشتریان نیازمند توجه باید در اولویت ارتباطات شخصی‌سازی‌شده، ارائه پیشنهادهای ویژه یا بهبود خدمات پس از فروش قرار گیرند. برای بررسی ارتباط میان ویژگی‌های مدل RFMSA از ماتریس همبستگی[۵۶] استفاده شد که در شکل ۹ نشان داده شده است.



[52] ANOVA
[53] Tukey HSD
[54] F-Statistic
[55] P-value
[56] Correlation Matrix

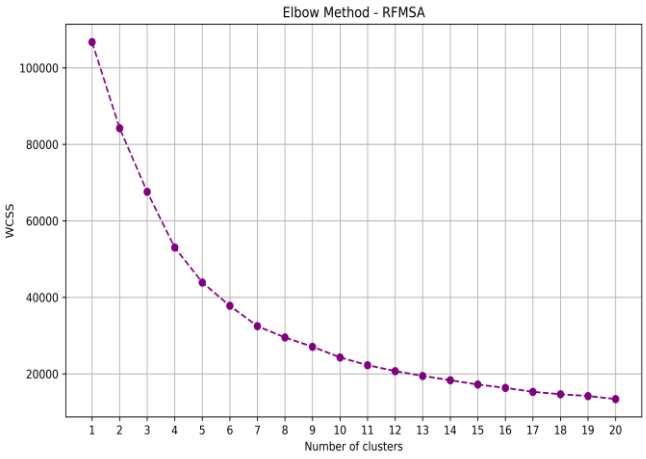
شکل ۹: ماتریس همبستگی ویژگی‌های مدل آر.اف.ام.اس.آ

ماتریس همبستگی شکل ۹، نمای فشرده‌ای از روابط خطی میان متغیرهای رفتاری RFM و شاخص احساسات ارائه می‌دهد. مهم‌ترین مقدار قابل توجه در این ماتریس ضریب همبستگی حدود ۰٫۳ بین ارزش خرید و احساسات است. برای تفسیر این عدد ابتدا باید ماهیت دقیق هر متغیر در مجموعه داده را مرور کنیم: در این پژوهش متغیر ارزش خرید در واقع میانگین امتیاز ثبت‌شده توسط کاربران است و نه مقدار پولی واقعی تراکنش‌ها. از سوی دیگر شاخص احساسات مقدار ترکیبی استخراج‌شده از متن نظرات با استفاده از کتابخانه NLTK و ابزار VADER است که در بازه ۰ تا ۱ قرار می‌گیرد.

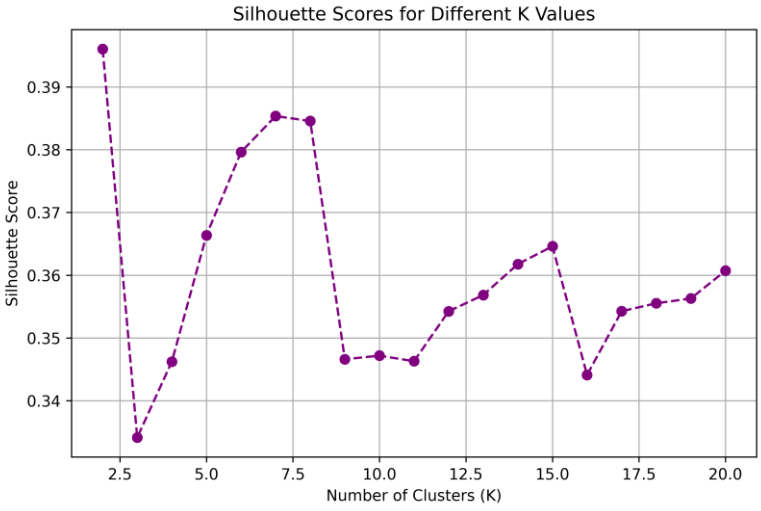
۴-۵- نتایج مدل‌های یادگیری ماشین روی مدل ترکیبی پیشنهادی

۴-۵-۱- پیاده‌سازی الگوریتم کا-میانگین

در این مرحله، کا-میانگین مستقیما بر متن نظرات اجرا نشده است، بلکه بر ویژگی‌های عددی مدل RFMSA، شامل شاخص‌های RFM و امتیاز تجمیع‌شده احساسات، اعمال شده است. برای اجرای الگوریتم کا-میانگین که مراحل آن پیشتر در بخش ۳-۵-۱ بیان شد، ابتدا تعداد خوشه بهینه بر اساس روش آرنج[۵۷] و شاخص سیلوئت[۵۸] بدست آمد که تعداد ۷ خوشه تعیین گردید. در شکل ۱۰ نمودار آرنج و در شکل ۱۱ نمودار شاخص سیلوئت برای ۲۰ خوشه نشان داده شده است. سپس خوشه ها به همراه تعداد مشتریان خود در نمودار میله‌ای شکل ۱۲ رسم شدند و با استفاده از تکنیک تحلیل مولفه‌های اصلی[۵۹] خوشه‌ها در دو بعد مجسم شدند که در شکل ۱۳ قابل مشاهده است.

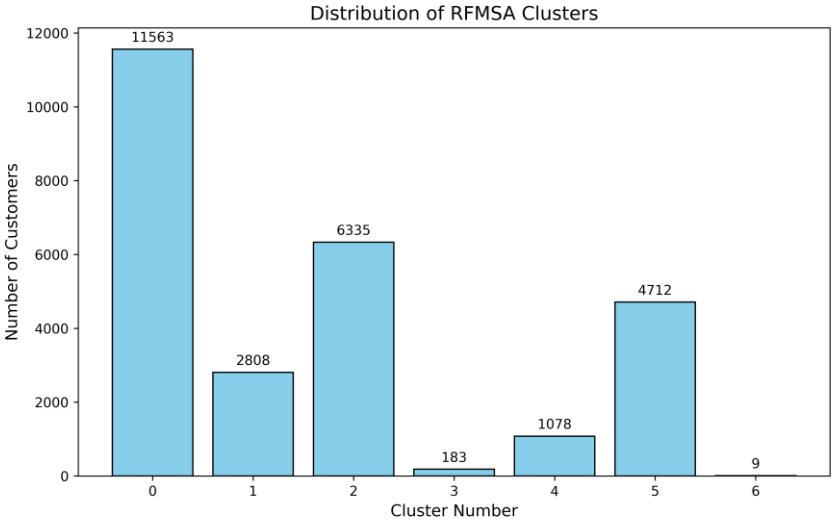


ی تشخیص تعداد خوشه

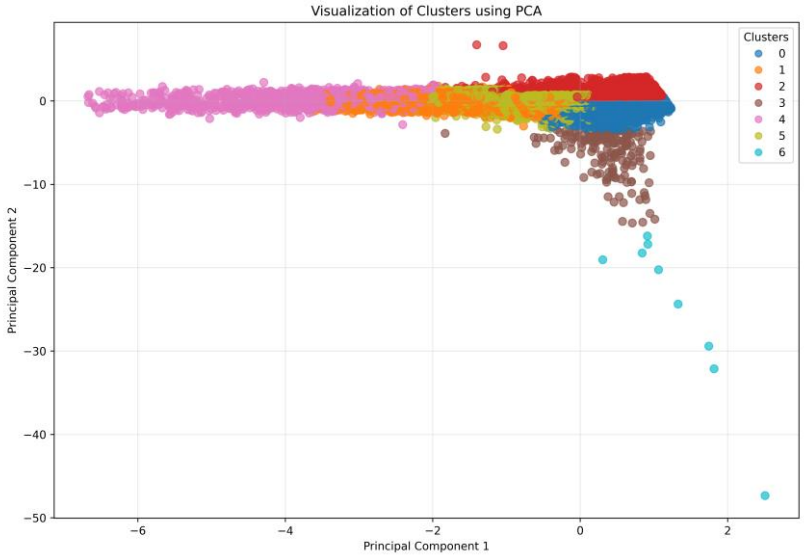


شکل ۱۱: نمودار شاخص سیلوئت برای ۲۰ خوشه برای تشخیص تعداد خوشه

[57] Elbow Method
[58] Silhouette index
[59] Principal component analysis(PCA)



شکل ۱۲: تعداد مشتریان فروشگاه به تفکیک هر خوشه



شکل ۱۳: تجسم خوشه ها با تحلیل مولفه‌های اصلی

پس از تجسم خوشه‌ها با تحلیل مولفه های اصلی، در هر خوشه میانگین شاخص‌های مدل RFMSA بررسی شد که در جدول ۶ آورده شده است.

جدول ۶: میانگین ویژگی‌های مدل ترکیبی در هر خوشه

C	Recency	Frequency	Monetary	Sentiment
0	410.604	1.273	4.975	0.877
1	512.131	1.089	4.665	0.459
2	789.489	1.033	4.760	0.879
3	246.191	16.049	4.601	0.812
4	553.392	1.040	2.190	0.484
5	459.844	1.169	3.863	0.855
6	201.667	71.556	4.574	0.788

در خوشه صفر که تعداد آن‌ها ۱۱۵۶۳ است، مشتریانی قرار دارند که خرید اخیر داشته‌اند، ارزش خرید بالایی دارند و احساس مثبتی نسبت به خریدشان دارند؛ بنابراین این گروه برای حفظ وفاداری و افزایش درآمد اهمیت زیادی دارد. در خوشه یک که تعداد آن‌ها ۲۸۰۸ است، مشتریانی با ارزش خرید متوسط، تعداد خرید کم و احساس منفی دیده می‌شوند که نیازمند بررسی دقیق مشکلات و بهبود تجربه خرید هستند. خوشه دو که تعداد آن‌ها ۶۳۳۵ است، شامل مشتریانی است که مدت زیادی از آخرین خریدشان گذشته اما همچنان ارزش خرید و احساس مثبت بالایی دارند و می‌توان با کمپین‌های بازفعال‌سازی آن‌ها را دوباره جذب کرد. در خوشه سه که تعداد آن‌ها ۱۸۳ است، مشتریانی با خریدهای پرتکرار، ارزش بالا و احساس مثبت قرار دارند که مشتریان کلیدی محسوب می‌شوند و باید با خدمات ویژه و ارتباط اختصاصی مدیریت شوند. خوشه چهار نیز که شامل ۱۰۷۸ نفر است، مشتریانی با ارزش پایین و احساس خنثی ولی نزدیک به منفی را شامل می‌شود که اولویت پایینی در بازاریابی دارند، هرچند اقدامات کم‌هزینه می‌تواند بر آن‌ها اثرگذار باشد. در خوشه پنج نیز که ۴۷۱۲ مشتری قرار دارد، مشتریانی با ارزش متوسط و احساس مثبت هستند که پتانسیل رشد دارند و با برنامه‌های هدفمند می‌توان ارزش آن‌ها را افزایش داد. خوشه شش بسیار کوچک و تعدادشان ۹ نفر است اما مشتریان آن خرید بسیار بالایی دارند که می‌تواند ناشی از مشتریان خاص یا سازمانی باشد و نیازمند مدیریت دقیق و تعامل شخصی‌سازی‌شده است.

۴-۵-۲- پیاده‌سازی الگوریتم رگرسیون لجستیک در مدل پیشنهادی ترکیبی

در این بخش به پیاده‌سازی الگوریتم رگرسیون لجستیک برای مدل ترکیبی RFMSA پرداخته می‌شود. پس از تشکیل گروه‌های مشتری بر اساس آستانه‌های تعیین‌شده، این مهم برای بررسی قابلیت بازتولید این گروه‌ها در مدل RFMSA اجرا شده است. در جدول ۷ و ۸ نتایج حاصل از این الگوریتم برای مدل ترکیبی نشان داده شده است.

جدول ۷: گزارش طبقه‌بندی مدل آر.اف.ام.اس.آ

	Precision	Recall	F1-Score	Support
Need Attention	1.00	0.99	0.99	1602
Potential	1.00	0.99	1.00	1601
Champions	0.99	1.00	1.00	4804
accuracy			1.00	8007
Macro avg	1.00	0.99	1.00	8007
Weighted avg	1.00	1.00	1.00	8007

جدول ۸: ماتریس درهم‌ریختگی برای مدل آر.اف.ام.اس.آ

	NA	P	Ch
Need Attention	1585	0	17
Potential	0	1587	14
Champions	3	0	4801

می‌توان دریافت که نتایج بازتولید رگرسیون لجستیک روی مدل ترکیبی افزایش قابل توجهی داشته است. لازم به ذکر است دقت ۱٫۰۰ که در جدول ۷ آمده است به دلیل نحوه نمایش و گرد کردن خروجی است و این جدول همان خروجی پایتون با ۲ رقم اعشار است و در واقع این دقت با ۴ رقم اعشار ۰٫۹۹۹۹ بوده است، در حالی که مدل سنتی دقتی معادل ۹۹ درصد به دست آورده است. در گروه مشتریان نیازمند توجه، هر دو مدل عملکرد بالایی داشته‌اند، اما مدل ترکیبی خطای کمتری نسبت به مدل سنتی داشته است. تفاوت اصلی در گروه مشتریان بالقوه مشاهده می‌شود؛ به‌طوری که در مدل ترکیبی دقت و حساسیت هر دو افزایش قابل توجه داشته، در حالی که مدل سنتی با وجود دقت بالا، حساسیت پایینی معادل ۹۶ درصد داشته و تعداد قابل توجهی از این مشتریان به اشتباه در گروه مشتریان طلایی قرار گرفته‌اند. این موضوع نشان می‌دهد که مدل ترکیبی در شناسایی صحیح مشتریان بالقوه بسیار بهتر عمل کرده است. در گروه مشتریان طلایی نیز هر دو مدل عملکردی نزدیک به صد درصد داشته‌اند، اما باز هم مدل ترکیبی خطاهای کمتری نسبت به مدل سنتی ثبت کرده است. در مجموع می‌توان گفت که مدل ترکیبی نسبت به مدل سنتی نتایج دقیق‌تر و قابل‌اعتمادتری ارائه داده و به‌ویژه در شناسایی مشتریان بالقوه که برای جذب و تبدیل آن‌ها به مشتریان وفادار اهمیت زیادی دارند، مزیت معناداری داشته است. در جدول ۹ استراتژی‌های مدیریتی و بازاریابی مربوط به هر گروه آورده شده است.

جدول ۹: استراتژی‌های بازاریابی گروه‌های مختلف مشتریان

Customer Group/Cluster	Main Features (RFM + SA)	Analysis and Status	Managerial and Marketing Strategies
Cluster 0 / Similar to the Golden Group	High recency, high monetary value, and positive sentiment	Key loyal customers with a high likelihood of repeat purchases	Maintaining loyalty through special discounts and programs, premium services, enhancing positive customer experience, and encouraging word-of-mouth promotion.
Cluster 1 / Similar Needs Attention	Medium monetary value, low purchase frequency, and negative sentiment	Poor purchase experience with a high risk of churn	Analyzing customer experience issues, improving after-sales services, offering incentive-based proposals, and providing personalized communication.
Cluster 2 / Near potential	Low recency, high monetary value, and positive sentiment	High-value but inactive customers with potential for reactivation	Targeted reminder campaigns, personalized discounts, and reactivating customer engagement.
Cluster 3 / Part of Golden	High purchase frequency, moderate recency, and positive sentiment	VIP customers with high loyalty and high monetary value	Exclusive loyalty programs, dedicated services, and preventing churn among key customers.
Cluster 4 / Close to needing attention	Low monetary value, low frequency, and negative sentiment	Low economic return and low priority	Low-cost actions (notifications, free samples) with focused efforts to increase purchase motivation.
Cluster 5 / Part of the potential	Medium monetary value and positive sentiment	Customers with potential for growth and upgrading to a higher-value segment	Loyalty and incentive programs, motivating customers to increase purchasing, and leveraging positive feedback.
Cluster 6 / Special Group	Very high purchase frequency, small customer count, and high monetary value	Special or corporate clients with strategic importance	Dedicated and personalized management, close interaction, and individualized analysis.

۵- نتیجه‌گیری و پیشنهادات آینده

نتایج این پژوهش نشان داد که ترکیب مدل RFM با تحلیل احساسات مشتریان در قالب چارچوب RFMSA، امکان تحلیل همزمان رفتار خرید و نگرش مشتریان را فراهم می‌کند. در این چارچوب، اطلاعات تراکنشی با اطلاعات حاصل از تجربه و احساسات مشتریان ترکیب شده و در نتیجه، گروه‌هایی از مشتریان با ویژگی‌های رفتاری ـ احساسی متفاوت قابل شناسایی هستند. به‌ویژه، شناسایی مشتریانی با ارزش خرید بالا اما احساسات منفی و همچنین مشتریانی با ارزش خرید متوسط و نگرش مثبت، نشان می‌دهد که ارزیابی مشتریان صرفاً بر اساس شاخص‌های تراکنشی می‌تواند بخشی از اطلاعات مهم برای تصمیم‌گیری مدیریتی را نادیده بگیرد. از منظر مدیریتی، نتایج RFMSA می‌تواند مستقیماً در طراحی و اولویت‌بندی اقدامات بازاریابی و مدیریت ارتباط با مشتری مورد استفاده قرار گیرد. برای مشتریان با ارزش خرید بالا و احساسات منفی، اولویت باید بر شناسایی علت نارضایتی، پیگیری بازخوردها، ارائه خدمات جبرانی و اقدامات پیشگیرانه برای حفظ رابطه با مشتری قرار گیرد. در مقابل، مشتریانی که از نظر ارزش خرید در سطح متوسط قرار دارند اما احساسات مثبتی نشان می‌دهند، می‌توانند به‌عنوان گروهی دارای ظرفیت رشد در نظر گرفته شوند و با استفاده از پیشنهادهای شخصی‌سازی‌شده، برنامه‌های تشویقی و افزایش تعامل با برند، به سمت خرید بیشتر و وفاداری بلندمدت هدایت شوند.

مشتریان وفادار و دارای احساسات مثبت نیز می‌توانند در کانون برنامه‌های حفظ مشتری، پاداش‌های وفاداری و تقویت رابطه بلندمدت قرار گیرند. همچنین، شناسایی مشتریانی با ارزش پایین و احساسات منفی می‌تواند به سازمان کمک کند تا منابع بازاریابی و خدماتی خود را با اولویت‌بندی مناسب تخصیص دهد و از صرف منابع یکسان برای تمامی گروه‌های مشتریان اجتناب کند. بنابراین، ارزش مدیریتی چارچوب RFMSA صرفاً در ایجاد یک بخش‌بندی جدید نیست، بلکه در فراهم کردن مبنایی برای تخصیص متفاوت منابع و طراحی اقدامات متناسب با ویژگی‌های هر گروه مشتری است. استفاده از این چارچوب می‌تواند به سازمان‌ها کمک کند تا به جای اتخاذ سیاست یکسان برای کل مشتریان، تصمیم‌هایی متناسب با ارزش اقتصادی و نگرش احساسی هر گروه اتخاذ کنند و از اطلاعات حاصل از بازخورد مشتریان در کنار داده‌های تراکنشی بهره بگیرند. با وجود این دستاوردها، پژوهش حاضر محدودیت‌هایی نیز داشت که می‌تواند مسیر تحقیقات آینده را روشن کند. نخست آنکه داده‌ها صرفاً از یک بستر خاص گردآوری شده و تعمیم نتایج به سایر صنایع نیازمند آزمون‌های گسترده‌تر است. دوم، تحلیل احساسات با وجود دقت بالا همچنان در مواجهه با متون کنایه‌آمیز یا چندلایه محدودیت دارد. در پژوهش‌های آینده، می‌توان با استفاده از مدل‌های زمینه‌محور مبتنی بر ترنسفورمر، مانند BERT، و افزودن مرحله تشخیص کنایه پیش از تعیین قطبیت احساسات، دقت تحلیل احساسات را در مواجهه با این نوع متون بهبود داد. سوم، استفاده از الگوریتم کا-میانگین و رگرسیون لجستیک ممکن است در شناسایی الگوهای پیچیده رفتاری کفایت نداشته باشند. بنابراین بهره‌گیری از روش‌های پیشرفته‌تر یادگیری عمیق و ترکیب داده‌های جمعیت‌شناختی، رفتاری، تراکنشی و متنی می‌تواند زمینه بهبود و توسعه مدل را فراهم کند. از دیگر محدودیت‌های پژوهش، در نظر گرفتن وزن برابر برای ابعاد رفتاری و احساسی در شاخص RFMSA است؛ زیرا اهمیت این ابعاد می‌تواند با توجه به شرایط کسب‌وکار متفاوت باشد. در پژوهش‌های آتی، می‌توان با انجام تحلیل حساسیت و بررسی وزن‌های مختلف با استفاده از روش‌هایی مانند روش‌های تصمیم‌گیری چندمعیاره، اثر تغییر وزن‌ها بر نتایج گروه‌بندی مشتریان را ارزیابی و وزن‌های متناسب با شرایط کسب‌وکار تعیین کرد.همچنین آزمون این مدل ترکیبی در صنایع مختلفی مانند بانکداری، بیمه و خرده‌فروشی می‌تواند اعتبار آن را افزایش داده و کارایی عملی چارچوب RFMSA را بیش از پیش آشکار سازد.

منابع (References)

1. Li-Yoong, O. and Choo-Yee, T. and Helmi, Z. and Eashvaren, Ch., 2024. Behavioural user segmentation of app users based on functionality interaction patterns. *Cogent Engineering*, *11*(1). doi: https://doi.org/10.1080/23311916.2024.2430430
2. Handojo, A. and Pujawan, N. and Santosa, B. and Singgih, M. L., 2023. A multi layer recency frequency monetary method for customer priority segmentation in online transaction. *Cogent Engineering*, *10*(1). doi: https://doi.org/10.1080/23311916.2022.2162679
3. Rungruang, C. and Riyapan, P. and Intarasit, A. and Chuarkham, K. and Muangprathub, J., 2023. RFM model customer segmentation based on hierarchical approach using fca. *SSRN Electronic Journal*. doi: https://doi.org/10.2139/ssrn.4416218
4. Joy Christy, A. and Umamakeswari, A. and Priyatharsini, L. and Neyaa, A., 2021. RFM ranking – an effective approach to customer segmentation. *International J. of Advanced Computer Science and Applications*, pp.1251-1257. doi: http://dx.doi.org/10.1016/j.jksuci.2018.09.004
5. Yousefizad, A. and Sorayaei, A., 2019. Customer segmentation and clustering based on the rfm model and designing a framework for providing services to key customers. *J. of Executive Management Research, 15*(2), pp.45–60 [In Persian]. doi: https://doi.org/10.22080/jem.2019.15814.2834
6. Fathian, M. and Amiri, B. and Heidari, H. R., 2021. Investigating the impact of influential customer reviews in social networks on buying behavior of other customers. *Sharif Journal of Industrial Engineering & Management*, *37*(1), pp.37-45[In Persian]. doi: https://doi.org/10.24200/j65.2021.54547.2057
7. Ozkan, P. and Deveci-Kocakoc, I., 2021. A customer segmentation model proposal for retailers: rfm-v. In c. cobanoglu, & v. della corte . *Advances in Global Services and Retail Management, USF M3 Publishing*, pp.1-12. doi: https://www.doi.org/10.5038/9781955833035
8. Wang, S. and Sun, L. and Yu, Y., 2024. A dynamic customer segmentation approach by combining lrfms and multivariate time series clustering. *Sci Rep* 14,17491, pp.1-18.

machine learning algorithm for the u.s. presidential election forecasting, *J. of Computer Science*, *18*(3), pp.200-215.
doi: https://doi.org/10.48550/arXiv.2312.05584

18. Anitha, M. A. and Sherly, K., 2025. Churn prediction with graphsage model based on the derived features using rfm and sentiment analysis. *J. of the Chinese Institute of Engineers*, *48*(4), pp.1–13. doi: https://doi.org/10.1080/02533839.2025.2478185

19. Kumar, V. and Steinbach, M. and Tan, P., 2019. Introduction to Data Mining, 2nd ed., Pearson. ISBN:9780133128901.

20. Dolati Dolatabad, M. and Behmardi Kalantari, H., and Mirsaeedi, F. and Sheikh Alishahi, M., 2024. Proposing a supervised machine learning approach for data-driven simulation in supplier selection problem. *Sharif Journal of Industrial Engineering & Management*, *40*(2), pp.86-98[In Persian].
doi: https://doi.org/10.24200/j65.2024.63431.2378

doi: https://doi.org/10.1038/s41598-024-68621-2

9. Shumaly, S. and Yazdinejad, M. and Guo, Y., 2021. Persian sentiment analysis of an online store independent of pre-processing using convolutional neural network with fastText embeddings. *PeerJ Computer Science*, *7*(22), e422.
doi: https://doi.org/10.7717/peerj-cs.422

10. Khodaei, A. and Bastanfard, A. and Saboohi, H. and Aligholizadeh, H., 2024. Sentiment analysis of persian sentences using efficient deep learning in fiction. *J. of Computing and Security*, *11*(1), pp.67–86.
doi: https://doi.org/10.22108/jcs.2024.139842.1137

11. Rahimi, R. and Ghasmy Yaghin, R., 2024. Topic modeling and sentiment analysis of customers using natural language processing and machine learning techniques. *Sharif Journal of Industrial Engineering & Management, 41*(2), pp.45-61[In Persian].
doi: https://doi.org/10.24200/j65.2025.65318.2418

12. Malik, N. and Bilal, M., 2024. Natural language processing for analyzing online customer reviews: a survey, taxonomy, and open research challenges. *PeerJ Computer Science*, *10*, e2203.
doi: https://doi.org/10.7717/peerj-cs.2203

13. Petratos, P. and Giannoula, M., 2025. Transformer-based ai for sentiment analysis in marketing. *ACM computing surveys*, *58*(4), pp.1–25. doi: https://doi.org/10.1145/3690771.3690795

14. Yin, H. and Hou, J. and Li, H. and Liu, L., 2024. Sentiment analysis of user reviews on bert-based e-commerce platforms. *In Proceedings of the 2024 IEEE 3rd International Conference on Electrical Engineering, Big Data and Algorithms (EEBDA)*. IEEE.
doi:
https://doi.org/10.1109/EEBDA60612.2024.10485803

15. Mousavi, S. M. and Amiri Aghdaie, S. F., 2020. Identifying the constructive elements of value proposition and their impact on customers satisfaction using sentiment analysis based on text mining, *J. of Business Management*, *12*(4), pp.783-806.
doi:https://doi.org/10.22059/jibm.2020.302987.3847

16. Ho, T. and Nguyen, S. and Nguyen, H. and Nguyen, N. and Man, D. S. and Le, T. G., 2023. An extended rfm model for customer behaviour and demographic analysis in retail industry. *Business Systems Research*, *14*(1), pp.26–53.
doi: https://doi.org/10.2478/bsrj-2023-0002

17. Feng, G. and Cai, H. and Chen, K. and Li, Zh., 2023. A hybrid method of sentiment analysis and