طراحی و پیاده‌سازی سیستم توصیه‌گر هوشمند محصول با استفاده از هوش مصنوعی در شرکت پخش قاسم ایران

هومن مظاهری پور[۱]، محمدعلی نظرپور طاهرخانی[۲]، معصومه خالقیان [۳]، محمد غلامی[۴]

۱- کارشناس ارشد صنایع- صنایع، دانشگاه آزاد تهران جنوب، تهران، ایران

۲- کارشناس ارشد مهندسی کامپیوتر- نرم‌افزار، دانشگاه پیام‌نور تهران، تهران، ایران

۳- کارشناس ارشد مهندسی کامپیوتر - هوش مصنوعی، دانشکده فنی‌مهندسی دانشگاه خوارزمی، تهران، ایران

۴- دانشجوی کارشناسی ارشد مهندسی صنایع - مدلسازی سیستم‌ها و تحلیل داده‌ها، دانشگاه صنعتی امیرکبیر، تهران، ایران

Email: [1] hooman.mp@gmail.com, [2] nazarpourma@gmail.com, [3] masoumehkhaleghian.ai@gmail.com, [4] mohammadgholami7380@gmail.com

چکیده

این پژوهش با هدف طراحی یک سیستم توصیه‌گر داده‌محور هوشمند برای پیش‌بینی احتمال علاقه‌مندی مشتری به کالا و توصیه پنج کالای اولویت‌دار به هر مشتری خرده‌فروش انجام شده است. بدین منظور، هشت منبع داده عملیاتی شامل فاکتورهای فروش، مشخصات مشتریان و ویزیتورها، اطلاعات تأمین‌کنندگان و ریز تراکنش‌های کالا در بازه سال‌های ۱۴۰۲ و ۱۴۰۳ گردآوری و پس از یکپارچه‌سازی به جدول تحلیلی نهایی تبدیل شد که دربرگیرنده ۵۰ ویزیتور، ۳۶۸۸ مشتری و ۱۱۸۲ قلم کالا بوده و در مجموع شامل ۶۲ ویژگی و ۱۸۹۷۲۵ رکورد است. مسئله به‌صورت دسته‌بندی چندکلاسه رفتار خرید تعریف گردید. برای مقابله با عدم توازن کلاس‌ها از وزن‌دهی کلاس نمونه ادغام کلاس‌های کم نمونه استفاده شد و چهار مدل اصلی پیاده‌سازی و ارزیابی گردید: Random Forest، CatBoost، XGBoost و MLP مبتنی بر تعبیه[۱] شناسه‌های مشتری/کالا/ویزیتور. نتایج نشان داد XGBoost متوازن‌ترین عملکرد را با صحت ۰.۹۹ و امتیاز F1=0.97 ارائه می‌کند؛ CatBoost با نرخ بازخوانی ۰.۹۹ مناسب سناریوهای پوشش‌محور است؛ Random Forest با امتیاز F1=0.96 کارایی پایدار و تبیینی دارد؛ و MLP با تعبیه‌برداری با وجود نرخ بازخوانی ۰.۹۶ به دلیل دقت ۰.۸۹ گزینه مکمل برای مدل‌های ترکیبی است. بر مبنای این یافته‌ها، XGBoost به‌عنوان مدل اصلی استقرار و CatBoost برای کمپین‌های حساس به ازدست‌ندادن فرصت‌ها پیشنهاد شد. نسخه اولیه رابط کاربری نیز پیاده‌سازی گردید که با دریافت «کد مشتری» برای هر مدل فهرست ۵ کالای با بیشترین احتمال تعلق به کلاس علاقه‌مندی مطلوب را بازمی‌گرداند. نتایج، قابلیت استقرار عملیاتی سیستم برای افزایش نرخ تبدیل سفارش و کاهش بار تصمیم‌گیری ویزیتور را تأیید می‌کند و مسیر توسعه‌های آتی از جمله ارتقای مدل با در مواجه با شروع سرد و ایجاد نمونه‌های منفی پس از اجرای آزمایشی سیستم پیشنهاد می‌شود.

کلمات کلیدی: سیستم توصیه‌گر؛ پخش محصولات؛ عدم توازن کلاس‌ها؛ یادگیری ماشین؛ یادگیری عمیق.

[1] Embedding

نویسنده مسئول: هومن مظاهری پور

ایمیل: hooman.mp@gmail.com

Design and Implementation of an Intelligent Product Recommender System Using AI at Ghasem Iran Distribution Company

Abstract:

This study aims to design an intelligent, data-driven recommender system that predicts each retail customer's likelihood of interest in products and recommends the top five priority items. To this end, eight operational data sources—sales invoices, customer and visitor (sales rep) profiles, supplier information, and item-level transaction logs—were collected for the Solar Hijri years 1402–1403 and integrated into a single analytical table comprising 50 visitors, 3,759 customers, and 1,182 products, with a total of 62 features and 189,725 records. The task was formulated as a multi-class classification of purchasing behavior. To address class imbalance, we applied class weighting together with consolidation of low-frequency classes, and implemented four core models: Random Forest, CatBoost, XGBoost, and an MLP with customer/item/visitor embeddings. Results show that XGBoost provides the most balanced performance (accuracy = 0.99, F1 = 0.97); CatBoost, with recall = 0.97, is suitable for coverage-oriented scenarios; Random Forest delivers stable and interpretable performance (F1 = 0.96); and the Embedding-based MLP, despite recall = 0.96, with precision = 0.89, is a useful complementary component for ensemble designs. Based on these findings, XGBoost is recommended as the primary deployment model and CatBoost for campaigns where missing potential opportunities is costly. A first UI prototype was also implemented that, given a customer code, returns—for each model—the five products with the highest probability of belonging to the "low-interest" class. The results confirm the system's operational deployability for improving order conversion rates while reducing sales-rep decision load, and point to future work such as strengthening cold-start handling and generating post-deployment hard negatives.

Keywords: recommender system; product distribution; class imbalance; machine learning; deep learning.

۱- مقدمه

با شتاب‌گرفتن تجارت الکترونیک و دیجیتالی‌شدن خدمات، سیستم‌های توصیه‌گر[1] به ابزار کلیدی شخصی‌سازی و ارتقای تجربه کاربر بدل شده‌اند؛ سامانه‌هایی که با استخراج دانش از ردپای رفتاری کاربران، پیشنهادهای متناسب ارائه می‌کنند و به کاهش بار تصمیم‌گیری و افزایش نرخ تبدیل کمک می‌نمایند [۱]. در محیط‌های مملو از گزینه؛ از فروشگاه‌های آنلاین تا کانال‌های فروش مویرگی، این سیستم‌ها با مدل‌سازی الگوهای ترجیحی، آیتم‌های همساز با علایق هر مشتری را برمی‌گزینند و تعامل و رضایت او را می‌افزایند [۲].

از دهه ۱۹۹۰ و ظهور فیلترینگ مشارکتی تا موج یادگیری عمیق پس از ۲۰۱۰، توصیه‌گرها از سازوکارهای ایستا به مدل‌های زمینه‌محور و پویای درک ترجیحات ارتقا یافته‌اند [۳]. در سناریوهای واقعی فروش میدانی، تعاملات غالباً هم‌زمان و چندکالایی است؛ بنابراین شناسایی «الگوهای سبدی» و هم‌خریدی میان اقلام برای افزایش سبد خرید و فروش مکمل اهمیت ویژه‌ای دارد [۴]. بااین‌حال، عدم توازن داده میان اقلام پرفروش و کم‌فروش، خطر سوگیری مدل به‌سوی کلاس‌های اکثریت و نادیده‌گیری فرصت‌های نهفته را به همراه دارد و مستلزم راهکارهای مقاوم در برابر کلاس‌های اقلیت است [۵].

با درنظر گرفتن این چالش‌ها، هدف این مقاله طراحی و پیاده‌سازی یک سیستم توصیه‌گر فروش میدانی برای شرکت پخش قاسم ایران است. در این سیستم، ویزیتورها در لحظه مراجعه به مشتری، از طریق نرم افزار[2]، فهرستی از کالاهای پیشنهادی شخصی‌سازی‌شده را مشاهده می‌کنند. این توصیه‌ها مبتنی بر سابقه خرید مشتری، رفتارهای مشابه دیگران، و ترجیحات یادگرفته‌شده هستند. مقاله حاضر تلاش دارد ضمن بررسی دقیق پیاده‌سازی این سیستم، عملکرد آن را در محیط واقعی (پیاده سازی روی ۵۰ ویزیتور) ارزیابی کرده و به بحث درباره مزایا، چالش‌ها و چشم‌اندازهای توسعه آن بپردازد.

۲- مرور ادبیات

در سال‌های اخیر، سیستم‌های توصیه‌گر به ابزار کلیدی بهبود تجربه کاربر در تجارت الکترونیک و پلتفرم‌های محتوایی تبدیل شده‌اند. رویکردهای اصلی همچنان بر سه خانواده‌ی «فیلترینگ مشارکتی[3]»، «فیلترینگ محتوامحور[4]» و «روش‌های ترکیبی[5]» تکیه دارند و تمرکز ادبیات بر افزایش دقت، شخصی‌سازی و تنوع پیشنهادها با اتکا به داده‌های عظیم و ناهمگون است [۶].

شاخه‌ی نوظهور یادگیری عمیق، به‌ویژه مدل‌های توالی‌محور و زبانی، جهشی در کیفیت توصیه ایجاد کرده است. HybridBERT4Rec با بهره‌گیری هم‌زمان از دو شبکه‌ی موازی برت[6] برای مدل‌سازی توالی رفتار کاربر و شباهت میان کاربران، دقت پیشنهاد را به‌طور معناداری ارتقا می‌دهد و نمونه‌ای موفق از ادغام دیدگاه‌های محتوامحور و مشارکتی در یک چارچوب واحد است [۷]. در سوی دیگر، مدل اِچ‌پی‌اِن‌اِم[7] برای توصیه‌ی آیتم در سبد خرید، ترجیحات کوتاه‌مدت و بلندمدت را با مکانیزم توجه[8] یکپارچه می‌کند و در سناریوهای واقعی خرید عملکرد بالایی گزارش کرده است [۴].

---

[1] Recommendation Systems
[2] PDA
[3] Collaborative Filtering
[4] Content-Based Filtering
[5] Hybrid
[6] BERT
[7] Hybrid-Preference Neural Model (HPNM)
[8] Attention

برای چالش «شروع سرد[1]»، مدل IMETA-GNN با ترکیب شبکه‌های گرافی[2] و یادگیری فراگیر[3] امکان یادگیری سریع از تعاملات اندک را فراهم می‌کند [۲]. در مواجهه با «عدم توازن» و «پراکندگی» داده‌ها نیز چارچوب ترکیبی CWGAN-GP-PacGAN با تولید داده‌های مصنوعیِ واقع‌گرایانه برای کلاس‌های اقلیت، کارایی مدل‌های توصیه‌گر را بهبود می‌بخشد [۵].

ترکیب خوشه‌بندی و بهینه‌سازی با یادگیری ماشین نیز نتایج قابل‌توجهی داشته است: ادغام کا-میانگین[4] و بهینه‌سازی ازدحامی[5] در کنار مدل LightGBM دقت شخصی‌سازی را افزایش و خطای پیش‌بینی را کاهش داده است [۸]. در حوزه‌ی زمینه‌محور[6]، چارچوب UbiCARS فرایند طراحی و خودکارسازی سامانه‌های توصیه‌گر زمینه‌آگاه را برای کاربران غیرمتخصص ساده می‌کند و هزینه و پیچیدگی توسعه را می‌کاهد [۱].

پاره‌ای از مطالعات بر غنای «نمایه کاربر» تمرکز کرده‌اند. مدل Meta-Interest با مدل‌سازی ویژگی‌های شخصیتی و استخراج مسیرهای پایدار در گراف‌های ناهمگون، توصیه‌هایی دقیق و تفسیرپذیر و هم‌زمان مقاوم در شروع سرد ارائه می‌دهد. در توصیه‌گرهای ترتیبی، بهره‌گیری از Rooted PageRank بر روی گراف‌های زمانی توازن مناسبی میان «اکتشاف[7]» آیتم‌های جدید و «بهره‌برداری» از دانش پیشین برقرار می‌کند. افزون‌براین، مدل احتمالاتی UIP با تکیه بر گراف روابط اجتماعی و در ترکیب با فاکتورگیری ماتریسی، علایق پنهان را حتی در کمبود تعاملات مستقیم برآورد می‌کند [۹], [۱۰], [۱۱].

مرور تحول روش‌ها نشان می‌دهد مسیر پژوهش از شبکه‌های بازگشتی ساده[8] به سمت معماری‌های مبتنی بر توجه و یادگیری خودنظارتی[9] حرکت کرده است؛ رویکردی که علاوه بر دقت بالاتر، از داده‌های بدون برچسب نیز بهره‌برداری می‌کند. یک مرور نظام‌مند نیز تأکید می‌کند که مدل‌های ترکیبی، گراف‌محور و عمیق، جهت‌گیری غالب ادبیات‌اند و مسائلی چون شفافیت، تبیین‌پذیری و مقیاس‌پذیری همچنان کانون چالش‌ها باقی مانده‌اند [۳], [۶].

هم‌زمان، ترکیب فراداده با تحلیل احساسات[10] و بهره‌گیری از مدل‌های زبانی بزرگ[11] نشان داده است که ادغام درک معنایی عمیق با سیگنال‌های متنی کاربر می‌تواند شخصی‌سازی و دقت توصیه را به‌طور محسوسی افزایش دهد و برخی کاستی‌های رویکردهای سنتی را جبران کند [۱۲].

بسیاری از مطالعات مرورشده بر پایگاه‌های داده‌ی عمومی تکیه دارند که برای مقایسه‌ی الگوریتمی سودمند است، اما پیچیدگی‌های عملیاتی صنعت (نویز، ناهماهنگی‌ها، محدودیت‌های کسب‌وکار و پویایی‌های زمانی) را بازنمایی نمی‌کند. این پژوهش با اتکا به داده‌ی واقعیِ یک شرکت پخش و پوشش جامع مؤلفه‌های عملیاتی (مشتری، کالا، فاکتور، ویزیتور و زمان) به این خلأ پاسخ می‌دهد. علاوه بر این، به‌جای برآورد یک نمره‌ی واحد علاقه، رفتار خرید در سه سطح (بدون علاقه، کم، متوسط و زیاد) مدل‌سازی شده تا امکان پیشنهاد «کالاهای مکمل»—نه صرفاً محبوب یا از پیش خریده‌شده—فراهم شود؛ رویکردی

---

[1] Cold Start
[2] GNN
[3] Meta-Learning
[4] K-Means
[5] RSO
[6] Context
[7] Exploration
[8] RNN
[9] SSL
[10] CaSTSRobERT
[11] LLM

که برای سناریوهای B۲B پخش کالا کاربردی و متمایز است. بهره‌گیری نظام‌مند از ویژگی‌های زمانی-رویدادی، و نیز ارائه‌ی یک رابط تصمیم‌یار که خروجی مدل‌ها را مستقیماً در فرایند فروش به ویزیتورها ارائه می‌کند، گامی عملیاتی در جهت تبدیل دستاوردهای الگوریتمی به منافع کسب‌وکاری محسوب می‌شود. همچنین ترکیب شخصی‌سازی رفتاری با منطق استنتاجی و طراحی مسیر توسعه برای مواجهه با شروع سرد و به‌روزرسانی پسااستقرار، افقی برای توسعه‌ی DSSهای تفسیرپذیرتر و مقیاس‌پذیرتر می‌گشاید.

۳- داده‌ها

داده‌های این مطالعه از پایگاه داده شرکت قاسم ایران استخراج شد. دامنهٔ زمانی داده‌ها کل دورهٔ ۱۴۰۲ تا ۱۴۰۳ (فروردین ۱۴۰۲ تا اسفند ۱۴۰۳) را پوشش می‌دهد. تمرکز مطالعه بر مشتریان خرده‌فروش و تأمین‌کنندگان غذایی است و برای کنترل ناهمگونی رفتاری، مجموعه‌ای از ۵۰ ویزیتور منتخب بر اساس ثبات فعالیت و حجم فروش، به‌عنوان جامعه‌ی مطالعه انتخاب شد. تمام داده‌ها به‌صورت برون‌کشی‌شده از پایگاه‌های عملیاتی فروش و در قالب فایل‌های جدولی تهیه گردید. برای پوشش کامل ابعاد مشتری–کالا–ویزیتور–زمان، هشت مجموعه‌داده جمع‌آوری شد. هر مجموعه‌داده، کلیدهای مشترک و حوزهٔ محتوایی مشخصی دارد تا در مراحل بعدی قابل اتصال باشد:

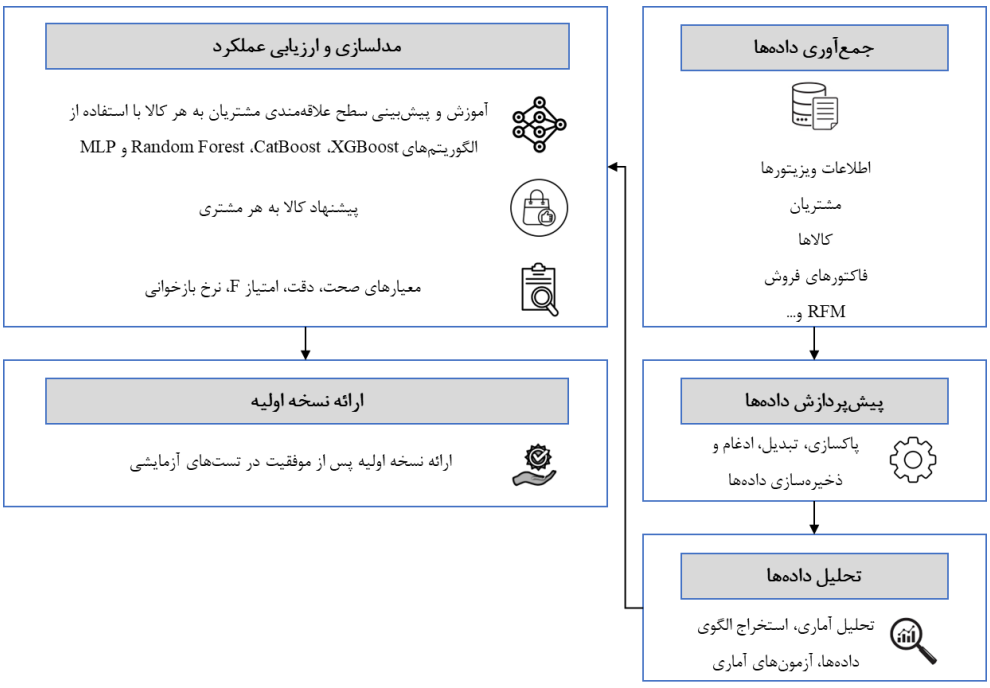
- سوابق فاکتورهای فروش: این مجموعه‌داده شامل شناسهٔ فاکتور، تاریخ شمسی (سال، ماه، روز)، کد مشتری، کد ویزیتور، کد کالا، مقدار/قیمت، تخفیف‌ها و فروش خالص (با مقادیر منفی برای برگشت از فروش). این منبع، ستون فقرات پژوهش و مبنای زمانی تمام محاسبات بعدی است.
- مشخصات کالاها: شامل کد کالا، گروه کالا، نوع محصول، وضعیت فعالیت و گروه تأمین‌کننده است. این مجموعه‌داده برای معنا دادن به کد کالا و امکان تحلیل‌های رده‌ای استفاده می‌شود.
- اطلاعات مشتریان خرده‌فروش: این مجموعه شامل پروفایل کاملی از ۳۶۸۸ مشتری خرده‌فروش است که از شرکت قاسم ایران طی سال‌های ۱۴۰۲ و ۱۴۰۳ خرید کرده‌اند. اطلاعات موجود در این داده‌ها شامل کد مشتری، موقعیت جغرافیایی، سابقه‌ی خرید و نوع و سطح تعامل با شرکت است.
- شاخص‌های رفتاری RFM مشتریان: خلاصه‌سازی تراکنش‌های فروش برای هر مشتری در دورهٔ مطالعه و محاسبهٔ تازگی خرید بر حسب فاصلهٔ زمانی از آخرین خرید، فراوانی خرید به‌صورت تعداد تراکنش و ارزش پولی به‌صورت جمع فروش خالص است. این جدول برای تحلیل رفتار مشتریان در مدل‌های یادگیری استفاده می‌شود.
- فروش ویزیتور–مشتری: شامل ریز تراکنش‌های ده ویزیتور منتخب به تفکیک کد مشتری و تاریخ شمسی، با متغیر «فروش خالص» است. این منبع امکان ردیابی عملکرد هر ویزیتور روی هر مشتری و تحلیل پویایی رابطه را فراهم می‌کند.
- اطلاعات ویزیتورها: داده‌های ۵۰ ویزیتور منتخب که مسئولیت مستقیم فروش به مشتریان را بر عهده دارند، جمع‌آوری شده است. این اطلاعات شامل مشخصات فردی (نام، تاریخ تولد، محل تولد، تعداد فرزند، مدرک تحصیلی)، کد پرسنلی و کد اختصاصی ویزیتور است. این ویزیتورها به طور مشخص در بازه‌ی زمانی مطالعه (۱۴۰۲ تا ۱۴۰۳) فعال بوده‌اند و به صورت مداوم فرآیند توزیع و فروش کالا را انجام داده‌اند.

- ریز فروش کالاها توسط ۵۰ ویزیتور به مشتریان: در این مجموعه‌داده واحد مشاهده، «کالا–تاریخ» است. شامل کد کالا، گروه کالا، نام/گروه تأمین‌کننده، تاریخ شمسی و فروش خالص برای همان ده ویزیتور و در کانال فروش است. این داده برای مطالعه الگوهای زمانی فروش هر ردیف محصول و اثر فصل/ماه استفاده می‌شود.
- تجمیع فروش خالص به ازای هر کالا در کل دورهٔ مطالعه: شامل کد کالا، گروه و نوع محصول، وضعیت فعالیت و فروش خالص تجمیعی در سال‌های ۱۴۰۲–۱۴۰۳ است. این جدول دید سطح‌بالایی از سهم ریالی هر کالا در سبد فروش ارائه می‌کند و برای وزن‌دهی هدف و انتخاب آستانه‌های اهمیت کالا به‌کار می‌رود.

مجموعه‌داده نهایی ادغام شده از این داده‌ها شامل ۱۸۹,۷۲۵ رکورد تراکنش منحصربه‌فرد فروش کالا به مشتریان توسط ویزیتورها در طول دوره زمانی دو ساله است که ۶۲ ویژگی را شامل می‌شوند. در مجموع ۳,۶۸۸ مشتری، ۱,۱۸۲ کالا و ۵۰ ویزتور منتخب در داده‌ها پوشش داده شده‌اند. کالاهای با گردش پایین در طول دوره شناسایی و مشخص شده‌اند که می‌تواند در تحلیل‌های آینده به منظور بهینه‌سازی سبد محصول و بهبود استراتژی‌های توزیع مفید واقع شود.

٤- روش تحقیق

این بخش فرایند پیاده‌سازی روش پیشنهادی ما را برای ساخت یک سیستم پیشنهاددهنده بر بستر تراکنش‌های فروش حضوری تبیین می‌کند. همان‌گونه که در شکل ۱ روش تحقیق نشان داده شده است، فرایند با ورود داده‌های خام سازمان آغاز می‌شود و پس از پیش‌پردازش و یکپارچه‌سازی، به تحلیل آماری و استخراج الگو می‌رسد؛ در گام پایانی، مدل‌های یادگیری ماشین و یادگیری عمیق آموزش داده شده، خروجی به صورت احتمال علاقه‌مندی مشتری به کالا در هر کلاس علاقه‌مندی (بدون علاقه، کم علاقه و علاقه زیاد) تولید و بر مبنای آن پیشنهاد کالا به هر مشتری انجام و ارزیابی عملکرد با استفاده از معیارهای، صحت، دقت، نرخ بازخوانی و امتیاز F1 انجام می‌شود.



شکل ۱. فلوچارت روش تحقیق مطالعه

٤-١- جمع‌آوری داده‌ها

این پژوهش از شرکت پخش قاسم ایران، یکی از بزرگ‌ترین شرکت‌های توزیع کالاهای مصرفی در ایران، جمع‌آوری شده است. این داده‌ها مشتمل بر اطلاعات فروش به خرده‌فروشان در سراسر کشور طی دو سال اخیر (۱۴۰۲ و ۱۴۰۳) هستند و از هشت منبع داده‌ای مجزا استخراج شده و سپس به کمک شناسه‌های یکتا به صورت یک پایگاه داده یکپارچه تجمیع شده‌اند.

٤-٢- پیش‌پردازش و تحلیل داده‌ها

به‌منظور تشکیل یک «پایگاه داده واحد» از هشت منبع داده، زنجیره‌ای از گام‌های پاک‌سازی، همسان‌سازی شناسه‌ها، تلفیق جداول، پالایش مقادیر غیرعادی و مهندسی ویژگی اجرا شد. در این فرایند، فاکتورهای فروش نقش محور اتصال را داشتند و سایر جداول (اطلاعات مشتری، شعب، ویزیتورها، تاریخ، نگاشت کالا–تأمین‌کننده و فهرست تجمیعی کالاها) بر اساس شناسه‌های مشتری، کالا و ویزیتور به آن الحاق گردیدند. در ادامه، منطق هر گام به‌صورت یکپارچه تشریح می‌شود.

۱) استانداردسازی نام ستون‌ها و انواع داده

نام ستون‌ها در تمام جداول به صورت یکنواخت بازنویسی شد تا کلیدهای اتصال دقیقا منطبق باشند. انواع داده‌ایِ عددی به قالب‌های صحیح/اعشاری و ستون‌های طبقه‌ای به نوع متنی تبدیل شدند. برای تاریخ‌ها، فیلدها به یک فرمت واحد تبدیل گردید تا امکان مرتب‌سازی زمانی، استخراج فاصله خرید و محاسبات RFM فراهم شود.

۲) همگام‌سازی زمانی و ساخت تقویم عملیاتی

برای رفع ناهمگنی ثبت تاریخ، متغیر تاریخ شمسی به تاریخ میلادی متناظر نگاشت و یک متغیر زمانی یکتا ساخته شد. سپس ترتیب رخدادها بر اساس این تاریخ مرتب گردید و شاخص‌های مبتنی بر زمان مانند فاصله بین دو خرید متوالی، تازگی خرید و پنجره‌های تجمعی هفتگی، ماهانه و فصلی قابل محاسبه شد.

۳) پالایش تراکنش‌ها

در داده فاکتور، مقادیر منفی «فروش خالص» عمدتاً بیانگر مرجوعی یا اصلاح سند بودند. منطق زیر اعمال شد:

- ردیف‌های با مقدار خالص صفر حذف شد.
- برای هر مشتری–کالا–بازه زمانی نزدیک، جفت‌های مثبت/منفی شناسایی و تهاتر گردید تا اثر «خرید–مرجوع» به‌صورت خالص در سطح مورد تحلیل باقی بماند.
- ردیف‌هایی که پس از تهاتر، اثر خالص غیرعادی (بزرگ‌مقیاس با علامت منفی) ایجاد می‌کردند، به‌عنوان ناهنجاری احتمالی علامت‌گذاری و از مجموعه مدل‌سازی کنار گذاشته شدند.

این سیاست باعث شد «سیگنالِ تقاضا» حفظ و نویز اسنادیِ مرجوعی‌ها وارد مدل نشود.

۴) حذف ناسازگاری‌ها، تثبیت یکتایی و کنترل تکراری‌ها

با استفاده از کلیدهای مرکب ویزیتور، مشتری، کالا و تاریخ، رکوردهای تکراری شناسایی شد. در صورت وجود چند ردیف هم‌معنا در یک روز برای همان مشتری و کالا، مقادیر تجمیع و تنها یک رکورد نگه‌داری شد. همچنین شناسه‌هایی که در جداول مرجع (مشتری/کالا/شعبه) وجود نداشتند، یا با برچسب «ناشناخته» مشخص و در صورت فقدان هرگونه اطلاعات مکمل از چرخه مدل‌سازی حذف شدند.

۵) غنی‌سازی داده با ابعاد توصیفی

برای افزایش توان تبیینی مدل، ابعاد زیر به فاکتورهای فروش الصاق شد:

- مشخصات مشتری: نوع مشتری، موقعیت، سرشاخه سازمانی و برچسب‌های رفتاری استخراج‌شده از سابقه خرید.
- مشخصات کالا: گروه کالا، نوع محصول، وضعیت فعالیت و نگاشت به نوع تأمین‌کننده/تأمین‌کننده
- مشخصات ویزیتور: محدوده سن (محاسبه براساس تاریخ تولد)، تحصیلات، تعداد فرزند و محل تولد که به‌صورت ویژگی‌های جمعیت‌شناختی و منابع بالقوه اختلاف عملکرد ویزیتورها در مدل لحاظ شد.
- مشخصات شعبه/حوزه فروش: نام شعبه و متغیرهای مکانی/ساختاری مرتبط با مدیریت فروش.

۶) مهندسی ویژگی‌های رفتاری و مقداری

شاخص‌های Recency، Frequency و Monetary هم در سطح مشتری و هم در سطح مشتری–کالا محاسبه شد. برای هر سطح، نمره‌های کوانتایلی R، F و M ساخته و سپس یک امتیاز ترکیبی RFM_Score تشکیل شد. علاوه‌بر آن، یک امتیاز وزنیِ پیوسته نیز تعریف شد که در آن با توجه به اهمیت شاخص M در این پژوهش به این شاخص وزن بیشتری داده شد که پس از آن با روش استانداردسازی Min-Max به بازه ۰ تا ۱ نگاشت شد تا با سایر ویژگی‌ها هم‌مقیاس گردد. فرمول (۱) شیوه محاسبه امتیاز وزنی RFM را نشان می‌دهد:

$$\text{Weighted\_RFM\_Score} = 0.2 * \text{R\_Score} + 0.3 * \text{F\_Score} + 0.5 * \text{M\_Score} \quad (۱)$$

برای نام شعب و دسته‌ها، ابتدا نرمال‌سازی زبانی شامل حذف کاراکترهای اضافی و یکنواخت‌سازی حروف انجام شد. برای استفاده غنی‌تر از اطلاعات متنی، یک Embedding مبتنی بر FastText روی توکن‌های فارسی ایجاد شد؛ بدین‌ترتیب نمایه‌های متنیِ شعبه/گروه به بردارهای چگال تبدیل و به ماتریس ویژگی افزوده شد.

۷) مدیریت داده‌های گمشده و مقادیر پرت

برای مدیریت داده‌های گمشده و مقادیر پرت بدین ترتیب عمل شد:

- مقادیر گمشده کمی با قواعد مبتنی بر دامنه (صفرِ معنادار در نبود خرید، یا میانه گروهی در متغیرهای شدت) جایگزین شد.
- برای شناسایی پرت‌ها، از قاعده IQR و نیز کنترل مقادیر بسیار بزرگ یا منفی پس از تهاتر استفاده شد. پرت‌های واقعیِ رفتاری حفظ، اما ناهنجاری‌های اسنادی حذف شد تا جانبداری به مدل تزریق نشود.
- در متغیرهای با چولگی شدید، تبدیل‌های لگاریتمی یا باکس‌کاکس به‌کار رفته شد تا توزیع‌ها نرم‌تر و اثر مقادیر بزرگ مهار شود.

۸) نهایی‌سازی دیتاست تحلیلی و تقسیم زمانی

پس از اتمام ادغام و مهندسی ویژگی، یک دیتاست مسطح با دانه‌بندی تراکنش تجمع ویزیتور-مشتری–کالا–زمان ساخته شد. سپس برای اجتناب از نشت اطلاعات، تقسیم داده انجام شد: ۸۰ درصد داده‌ها برای یادگیری و تنظیم ابرپارامترها، و ۲۰ درصد داده‌ها برای اعتبارسنجی و آزمون نگه‌داشته شد. این چیدمان باعث می‌شود ارزیابی مدل، معنادار و نزدیک به سناریوی بهره‌برداری واقعی باشد.

خروجی مراحل فوق مستقیماً ورودی مرحله «مدل‌سازی و ارزیابی عملکرد» قرار می‌گیرد. این ساختار، هم برای مدل‌های دسته‌بندی علاقه‌مندی مشتری به کالا و هم برای مدل‌های پیشنهادگر ترکیبی مناسب است و تعادل مطلوبی میان غنای اطلاعاتی و پایداری آماری برقرار می‌کند.

٤-٣- مدلسازی و ارزیابی عملکرد

در این پژوهش، مسأله به‌صورت «دسته‌بندی سه‌کلاسه سطح علاقه» (۰: بدون علاقه، ۱: کم‌علاقه، ۲: علاقه زیاد) تعریف شد. داده ورودی، جدول نهایی یکپارچه از رکوردهای فروش، شناسه مشتری، کالا و ویزیتور است که علاوه بر شناسه‌ها، شامل شاخص‌های رفتاری و تجمیعی (تعداد و مبلغ خرید در پنجره‌های زمانی، فاصله تا آخرین خرید، نرخ مرجوعی، سهم گروه‌های کالایی در سبد، شدت تعامل مشتری–ویزیتور و نسبت‌ها) می‌باشد. داده با نمونه‌گیری طبقه‌ای به نسبت ۸۰ درصد به ۲۰ درصد میان آموزش و آزمون تفکیک، توزیع برچسب‌ها حفظ شد. برای چالش عدم‌توازن، از دو رویکرد استفاده شد: (۱) وزن‌دهی خودکار کلاس‌ها در مدل‌هایی که پشتیبانی می‌کنند، و (۲) اعمال وزن ساده[1] متناسب با وارونهٔ فراوانی کلاس. ارزیابی بر مبنای صحت و نیز شاخص‌های Precision، Recall و F1 و با تحلیل ماتریس آشفتگی به‌ویژه جابه‌جایی بین کلاس‌های ۱ و ۲ انجام شد. چهار خانواده مدل پیاده‌سازی و مقایسه گردید که خلاصه پیکربندی آن‌ها در جدول ۱ آمده است.

جدول ۱. ابرپارامترهای الگوریتم‌های استفاده شده

الگوریتم	راهبرد نامتوازنی	ابرتنظیم‌های اصلی
Random Forest	`class_weight='balanced'`	n_estimators = 100 criterion = "gini" max_depth = None min_samples_split = 2
CatBoostClassifier	ضمنی حساسیت ذاتی CatBoost به توزیع کلاس	تنظیمات پیش‌فرض پایدار کنترل بیش‌برازش از طریق مکانیزم‌های داخلی بوستینگ
XGBoost	اجرای Sample Weight	max_depth = tuned learning_rate = tuned n_estimators = tuned
MLP با Embedding	یادگیری انتهابه‌انتها[2]	سه Embedding(output_dim = 50) برای شناسه‌های مشتری، کالا و ویزیتور Dense: 128 → 64 → 32 Dropout = 0.3 , 0.2 activation = "relu" output = Softmax(3) loss = "sparse_categorical_crossentropy" optimizer = "Adam" validation_split = 0.2

[1] Sample Weight
[2] EarlyStopping

فرایند ارزیابی به این صورت پیش رفته است که مدل‌ها بر بخش آموزش بر پایه‌ی راهبرد نامتوازنی مختص خود آموزش دیدند؛ سپس روی بخش آزمون پیش‌بینی تولید شد و گزارش‌ها استخراج گردید. ماتریس‌های آشفتگی[1] برای هر مدل ترسیم و تحلیل شد تا منشأ خطاها مشخص گردد.

٤-٤- ارائه نسخه اولیه

به‌منظور سنجش کارایی میدانی، یک رابط کاربری سبک پیاده‌سازی شد که با دریافت کد مشتری از ویزیتور، بر مبنای همان خط لوله پیش‌پردازش و ویژگی‌سازی مرحله آموزش، احتمال تعلق هر کالا به هر کلاس علاقه‌مندی را محاسبه و سپس ۵ کالای برتر را برای هر مدل نمایش می‌دهد. خروجی در چهار پنل مستقل ارائه می‌شود و برای هر قلم، کد و نام کالا همراه با بردار احتمال سه‌کلاسه (بی‌علاقه، کم‌علاقه، علاقه‌مند) گزارش می‌گردد؛ رتبه‌بندی بر اساس مؤلفه «کم علاقه» انجام می‌شود. سامانه امکان مقایسه بین‌مدلی (مشاهده هم‌زمان چهار فهرست)، تعیین آستانه اطمینان برای فیلتر اقلام کم‌اطمینان، و ثبت لاگ کامل هر درخواست (کد مشتری، زمان، مدل، اقلام و احتمال‌ها) جهت تحلیل بازخورد و بازآموزی دوره‌ای را فراهم می‌کند. همچنین خروجی در برنامه‌ریزی بازدید ویزیتور به‌کار رود. این نسخه اولیه با زمان پاسخ چندثانیه‌ای آماده استقرار آزمایشی بوده و بستر اجرای A/B تست، کالیبراسیون آستانه‌ها و تنظیم سیاست‌های وزن‌دهی میان مدل‌ها را مهیا می‌سازد.

۵- نتایج

در این بخش کارایی مدل‌های پیشنهادی بر روی «داده نهایی ادغام‌شده» گزارش می‌شود. همان‌گونه که در فصل روش تحقیق توضیح داده شد، برچسب هدف y نشان‌دهنده سطح علاقه مشتری به هر کالا است و پس از تحلیل عدم‌توازن، سه کلاس «بدون علاقه» و «کم علاقه» و «علاقه‌مند» ایجاد شدند تا توان یادگیری تقویت شود. شکل ۲ شیوه کار سیستم توصیه‌گر را نشان می‌دهد.

شکل ۲. شیوه کار سیستم توصیه‌گر

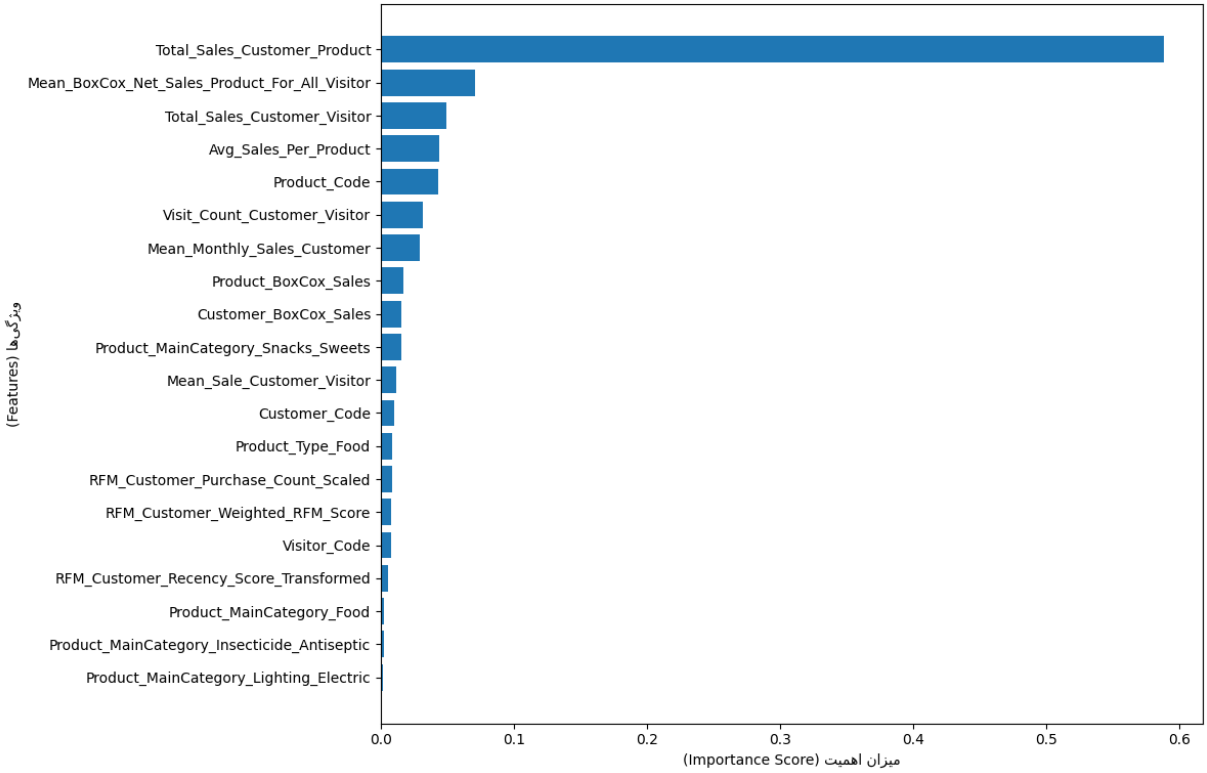
[1] Confusion Matrix

جدول ۲ نتایج ارزیابی الگوریتم‌های پیاده‌سازی شده را نشان می‌دهد. بر پایه‌ی نتایج آزمایش‌ها، XGBoost با صحت ۰.۹۹ و امتیاز F1=0.97 متوازن‌ترین عملکرد دقت–بازخوانی را ارائه داده و مناسب‌ترین گزینه برای هسته‌ی سامانه توصیه‌گر و رتبه‌بندی کالاست. CatBoost با ثبت بالاترین بازخوانی ۰.۹۹ در سناریوهای پوشش‌محور که هدف، از دست ندادن مشتریان بالقوه است، مزیت دارد؛ هرچند دقت ۰.۹۳ آن نشان می‌دهد نسبت به مدل بوستینگ دیگر مثبت کاذب بیشتری تولید می‌کند. Random Forestنیز با صحت ۰.۹۸ و امتیاز F1=0.96 کارایی پایدار و بسیار نزدیک به بهترین مدل‌ها دارد و با توجه به امکان استخراج اهمیت ویژگی، گزینه‌ای مطمئن برای استقرار پشتیبان و تحلیل‌های تبیینی کسب‌وکار است. MLP مبتنی بر تعبیه‌ها با وجود بازخوانی ۰.۹۶ به علت دقت ۰.۸۷ مثبت کاذب بیشتری می‌دهد و به‌تنهایی برای محیط عملیاتی توصیه نمی‌شود، اما به دلیل یادگیری نمایش‌های پنهان مشتری-کالا می‌تواند در نسخه‌های آتی و به‌صورت ترکیبی ارزش‌افزا باشد.

جدول ۲. نتایج ارزیابی الگوریتم‌های پیاده‌سازی شده

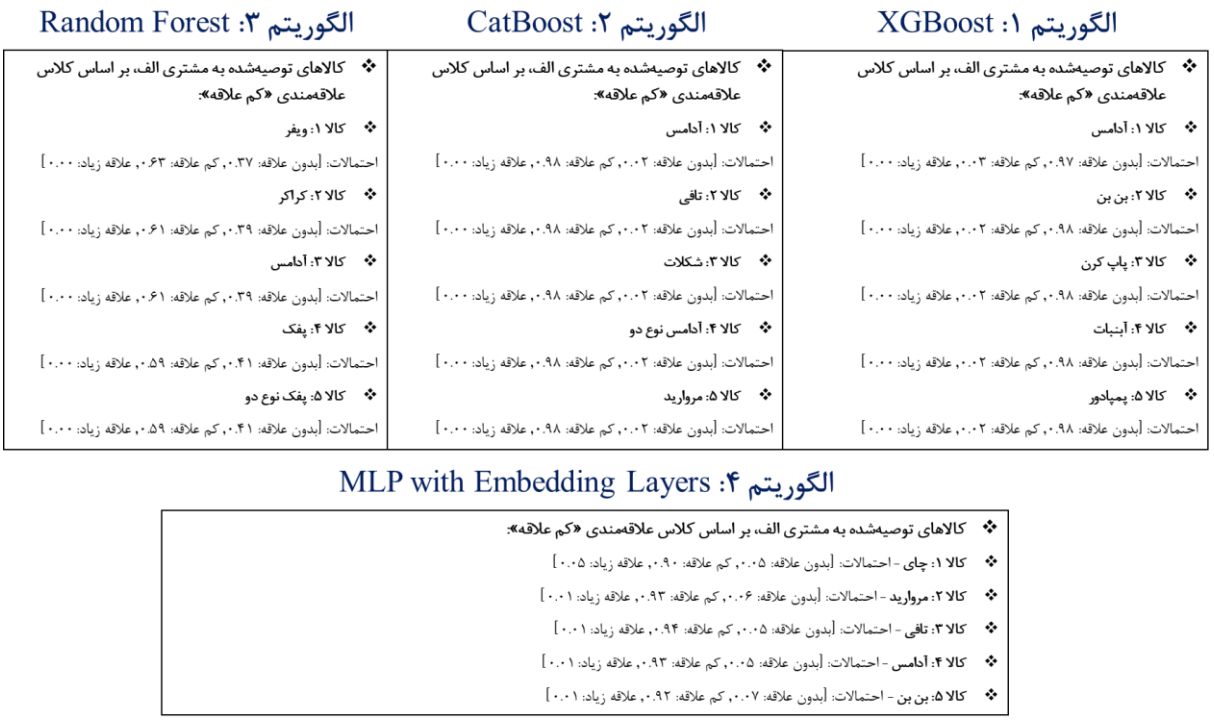
مدل	صحت	دقت	بازخوانی	امتیاز F1
XGBoost	۰.۹۹	۰.۹۷	۰.۹۷	۰.۹۷
CatBoost	۰.۹۸	۰.۹۳	۰.۹۹	۰.۹۵
Random Forest	۰.۹۸	۰.۹۶	۰.۹۵	۰.۹۶
MLP با Embedding	۰.۹۶	۰.۸۹	۰.۹۶	۰.۹۲

شکل ۳ تحلیل اهمیت ویژگی در مدل جنگل تصادفی را نشان می‌دهد که ویژگی‌های رفتاری و تجمیعی مرتبط با «میانگین فروش ماهانه مشتری»، «نشانه‌های تعامل مشتری–کالا» و شاخص‌های زمانی از مهم‌ترین متغیرها برای تمایز بین سطوح علاقه‌مندی مشتری به کالا هستند. این نتیجه با انتظار دامنه‌ای کسب‌وکار همخوان است؛ زیرا تاریخچه خرید و شدت تعامل، قوی‌ترین سیگنال برای تمایل آینده‌اند.



شکل ۳. تحلیل اهمیت ویژگی در مدل جنگل تصادفی

برای ارزیابی میدانی، کدنویسی پایتون پیاده‌سازی شد که برای یک مشتری نمونه، فهرست خریدهای پیشین را بازیابی کرده و سپس برای تمام کالاهایی که تاکنون نخریده است، احتمال تعلق به هر سه کلاس را محاسبه و ۵ کالای برتر را بر اساس بیشترین احتمالِ سه کلاس علاقه‌مندی برمی‌گزیند. این رویه همان چیزی است که در رابط کاربری نهایی مشاهده می‌شود. با تکیه بر نتایج فوق، نسخه عملیاتی سامانه روی چهار الگوریتم پیشنهادی پیاده‌سازی شد. رابط کاربری به ویزیتور اجازه می‌دهد با ورود «کد مشتری»، برای هر الگوریتم ۵ کالای پیشنهادی را همراه با احتمال تعلق به هر کلاس مشاهده کند و بر مبنای بالاترین احتمال کلاس مدنظر، اقدام فروش را تنظیم نماید. این طراحی، تصمیم‌گیری میدانی را تسهیل می‌کند. شکل ۴ کالاهای پیشنهاد شده توسط هر الگوریتم برای مشتری «الف» که در منطقه ۱ تهران به مدت ۲۱ سال فعالیت دارد را در این بازه زمانی نشان می‌دهد.



شکل ۴. خروجی مدل‌ها برای یک مشتری خاص

٦- نتیجه‌گیری

این پژوهش با هدف طراحی و ارزیابی یک سیستم توصیه‌گر داده‌محور هوشمند برای پیش‌بینی و توصیه کالاهای با سطوح علاقه‌مندی مختلف نزد مشتریان خرده‌فروش، بر مبنای داده‌های واقعی فروش و تعاملات ۵۰ ویزیتور در سال‌های ۱۴۰۲ و ۱۴۰۳ انجام شد. هشت منبع داده عملیاتی گردآوری و پس از همسان‌سازی کلیدها، رفع تناقض‌های زمانی و مالی، حذف تکرارها و یک‌دست‌سازی شناسه‌ها، به یک جدول تحلیلی یکپارچه تبدیل شدند که امکان مدل‌سازی پیش‌بینی احتمال خرید برای سه کلاس رفتاری مشتری را فراهم ساخت. خروجی سیستم به‌صورت «فهرست ۵ کالای پیشنهادی با بیشترین احتمال تعلق به کلاس ۱» برای هر مشتری ارائه می‌شود و بدین‌وسیله از یک‌سو پوشش فرصت‌های فروش افزایش می‌یابد و از سوی دیگر بار شناختی ویزیتور در انتخاب سبد پیشنهاد به‌طور معناداری کاهش پیدا می‌کند. در لایه یادگیری، چهار الگوریتم منتخب Random Forest، CatBoost، XGBoost و MLP مبتنی بر تعبیه پیاده‌سازی و با راهبردهای مختلفِ مقابله با نامتوازنی کلاس‌ها آموزش داده شدند. ارزیابی متقابل نشان داد XGBoost متوازن‌ترین کارایی را با صحت ۰.۹۹ و امتیاز F1= 0.97 ارائه می‌کند؛ CatBoost با نرخ بازخوانی ۰.۹۹ مناسب سناریوهای پوشش‌محور است که از دست نرفتن فرصت‌ها اهمیت بیشتری از کاهش مثبتِ کاذب دارد؛ Random Forest با امتیاز F1=0.96 عملکردی پایدار و تبیینی دارد و برای رصد پایداری و استخراج اهمیت ویژگی‌ها مفید است و MLP با تعبیه‌ها هرچند نرخ بازخوانی ۰.۹۶ کسب کرده، به دلیل دقت ۰.۸۹ به‌تنهایی برای استقرار اصلی توصیه نمی‌شود اما برای غنی‌سازی نمایش‌های پنهان و ایجاد مدل‌های ترکیبی ارزش‌افزا است. شیوه کارکرد سیستم توصیه‌گر در نمونه اولیه بدین ترتیب است که ویزیتور با ورود «کد مشتری» خروجی هر مدل را به‌صورت ۵ قلم کالای اولویت‌دار مشاهده می‌کند. این طراحی، تصمیم‌گیری میدانی را هدایت و هم‌راستایی بین اهداف پوشش بازار و حاشیه سود را تسهیل می‌کند. با توجه به نتایج دقت و بازخوانی بالا، انتظار می‌رود استقرار سیستم به کاهش زمان آماده‌سازی ویزیت و افزایش نرخ تبدیل سفارش منجر شود و امکان اجرای کمپین‌های هدفمند بر پایه احتمال خرید فراهم گردد. از منظر علمی، مطالعه حاضر سه دستاورد اصلی دارد: نخست، ارائه یک زنجیره کامل داده تا تصمیم از ادغام چندمنبعی داده‌های عملیاتی تا تولید توصیه قابل‌اجرا در سطح مشتری؛ دوم، نشان‌دادن اینکه مدل‌های بوستینگ گرادیانی در مواجهه با عدم‌توازن کلاس و

نویز تراکنش‌های واقعی، با تنظیم وزن کلاس و نمونه‌برداری هدفمند، می‌توانند به کارایی نزدیک به عملیاتی دست یابند؛ و سوم، نمایش مزیت یادگیری نمایش‌های تعبیه‌ای برای موجودیت‌های کلیدی که زمینه ارتقای آتی مدل‌های هیبریدی و شخصی‌سازی را فراهم می‌کند. با این حال، محدودیت‌هایی نیز وجود دارد: داده‌ها به ۵۰ ویزیتور منتخب و افق زمانی دو ساله محدود است؛ رفتار خرید ممکن است تحت تأثیر متغیرهای بیرونی (کمپین‌های قیمتی رقبا، موجودی، فصل، وقایع منطقه‌ای) باشد که بخشی از آن‌ها در داده‌های در دسترس ثبت نشده‌اند؛ و تعریف کلاس هدف مبتنی بر برچسب‌گذاری رفتاری داخلی است که می‌تواند با تغییر سیاست‌های تجاری نیازمند بازتعریف شود. علاوه بر این، با وجود کیفیت بالای شاخص‌های ارزیابی آفلاین، ارزش‌سنجی نهایی باید در میدان و با A/B تست بر پایه معیارهای کسب‌وکار (نرخ تبدیل، متوسط ارزش سفارش، نرخ بازگشت و موجودیِ به‌موقع) انجام شود. برای کارهای آینده، پیشنهاد می‌شود: (۱) گسترش دامنه داده‌ها به تمام ویزیتورها و کانال‌های فروش و افزودن ویژگی‌های بیرونی مانند قیمت رقبا، وضعیت موجودی و شاخص‌های فصلی–مناسبتی؛ (۲) دریافت بازخورد از سیستم پس اجرای آزمایشی و بهبود مدل‌ها برای تولید داده‌های نمونه منفی؛ و (۳) بهبود مدل‌ها پس از اجرای سیستم جهت مقابله با شروع سرد. سیستم پیشنهادی با اتکا به یکپارچه‌سازی داده، مدل‌های کارآمد و رابط کاربری ساده و عملیاتی، توانسته است مبنایی قابل اتکا برای توصیه کالاهای با احتمال علاقه‌مندی مطلوب در سطح مشتری فراهم کند. استقرار تدریجی با پایش میدانی و بهبود مستمر مدل، می‌تواند به افزایش فروش، بهینه‌سازی سبد پیشنهادی و استفاده هدفمندتر از منابع ویزیتوری بینجامد.

منابع

[1] C. Mettouris, A. Achilleos, G. Kapitsaki, and G. A. Papadopoulos, “An MDD Framework Towards the Automated Development of Ubiquitous Context-Aware Recommender Systems for Commerce,” *SN Comput Sci*, vol. 6, no. 4, Apr. 2025, doi: 10.1007/s42979-025-03896-4.

[2] N. Siddique *et al.*, “IMETA-GNN: Meta Learning-Based Cold Start Optimization for Recommendation System,” *IEEE Access*, vol. 13, pp. 93964–93976, 2025, doi: 10.1109/ACCESS.2025.3564454.

[3] J. H. Yoon and B. Jang, “Evolution of Deep Learning-Based Sequential Recommender Systems: From Current Trends to New Perspectives,” *IEEE Access*, vol. 11, pp. 54265–54279, 2023, doi: 10.1109/ACCESS.2023.3281981.

[4] Z. Pan, W. Chen, and H. Chen, “A Hybrid-Preference Neural Model for Basket-Sensitive Item Recommendation,” *IEEE Access*, vol. 8, pp. 226131–226141, 2020, doi: 10.1109/ACCESS.2020.3045092.

[5] W. Shafqat and Y. C. Byun, “A Hybrid GAN-Based Approach to Solve Imbalanced Data Problem in Recommendation Systems,” *IEEE Access*, vol. 10, pp. 11036–11047, 2022, doi: 10.1109/ACCESS.2022.3141776.

[6] I. Saifudin and T. Widiyaningtyas, “Systematic Literature Review on Recommender System: Approach, Problem, Evaluation Techniques, Datasets,” *IEEE Access*, vol. 12, pp. 19827–19847, 2024, doi: 10.1109/ACCESS.2024.3359274.

[7] C. Channarong, C. Paosirikul, S. Maneeroj, and A. Takasu, “HybridBERT4Rec: A Hybrid (Content-Based Filtering and Collaborative Filtering) Recommender System Based on BERT,” *IEEE Access*, vol. 10, pp. 56193–56206, 2022, doi: 10.1109/ACCESS.2022.3177610.

[8] X. Li, “Design of a Web Content Personalized Recommendation System Based on Collaborative Filtering Improved by Combining k-means and LightGBM,” *Journal of Web Engineering*, vol. 24, no. 2, pp. 267–290, 2025, doi: 10.13052/jwe1540-9589.2425.

[9] S. Mensah, C. Hu, X. Li, X. Liu, and R. Zhang, “A Probabilistic Model for User Interest Propagation in Recommender Systems,” *IEEE Access*, vol. 8, pp. 108300–108309, 2020, doi: 10.1109/ACCESS.2020.3001210.

[10] D. Kiselev and I. Makarov, “Exploration in Sequential Recommender Systems via Graph Representations,” *IEEE Access*, vol. 10, pp. 123614–123621, 2022, doi: 10.1109/ACCESS.2022.3224816.

[11] S. Dhelim, H. Ning, N. Aung, R. Huang, and J. Ma, “Personality-Aware Product Recommendation System Based on User Interests Mining and Metapath Discovery,” *IEEE Trans Comput Soc Syst*, vol. 8, no. 1, pp. 86–98, Feb. 2021, doi: 10.1109/TCSS.2020.3037040.

[12] S. S. Roy, A. Kumar, and R. S. Kumar, “Metadata and Review-Based Hybrid Apparel Recommendation System Using Cascaded Large Language Models,” *IEEE Access*, vol. 12, pp. 140053–140071, 2024, doi: 10.1109/ACCESS.2024.3462793.