

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ



University of Kashan

دانشکده مهندسی برق و کامپیووتر

بررسی مجموعه داده Dunnhumby برای پروژه پایانی کارشناسی

Analysis of the Dunnhumby Dataset for the Bachelor's Final Project

شماره دانشجویی: ۴۰۰۲۱۱۶۰۰۱۴

فاطمه سادات دست یافته

شماره دانشجویی: ۴۰۰۲۱۱۶۰۰۰۹

سمانه تابنده

شماره دانشجویی: ۴۰۰۲۱۱۶۰۲۱۴

سینا مظفری راد

گزارش اولیه بررسی ساختار مجموعه داده پروژه کارشناسی

استادوحیدی‌پور

اسفند ۱۴۰۳

چکیده

Dunnhumby دیتاست، یکی از مراجع معتبر در حوزه‌ی خرده‌فروشی، اطلاعات متنوعی را پیرامون تراکنش‌های مشتریان، ویژگی‌های خانوار و مشخصات محصولات گردآوری می‌کند. این متن، به معرفی ساختار جداول اصلی این دیتاست و کاربردهای پژوهشی محتمل می‌پردازد. بخش‌های گوناگون، شامل مشخصات تراکنش‌ها، محصولات، خانوارها و داده‌های جانبی (از جمله کوبن‌ها و کمپین‌ها) است. هدف، ارائه‌ی تصویری روشن از چگونگی استفاده از این داده‌ها در حوزه‌ی یادگیری ماشین، تحلیل سبد خرید و مدل‌سازی رفتار مصرف‌کننده است.

فهرست مطالب

۱	۱. مقدمه
۲	۲. ساختار پایگاه داده و جداول کلیدی
۳	۲.۱ جدول <code>transactions.csv</code>
۴	۲.۲ جدول <code>products.csv</code>
۴	۲.۳ جدول <code>hh_demographic.csv</code>
۵	۲.۴ جدول <code>casual_data.csv</code>
۶	۲.۵ جداول جانبی
۶	۳. کاربردهای ممکن در داده‌کاوی
۶	۳.۱ تحلیل سبد خرید (Market Basket Analysis)
۶	۳.۲ مدل‌سازی پیش‌بینی (Predictive Modeling)
۷	۳.۳ خوشبندی مشتریان (Customer Segmentation)
۷	۳.۴ تحلیل سری زمانی (Time Series Analysis)
۷	۳.۵ بررسی تأثیر کمپین‌ها و کوپن‌ها (Campaign Effectiveness)
۷	۴. مزایا و چالش‌ها
۸	۵. نمونه کدھای انجام شده مرتبط با این دیتاست در Kaggle
۸	۵.۱ تحلیل اکتشافی داده‌ها (EDA): جمعیت‌شناسی و الگوهای سبد خرید
۸	۵.۲ پیش‌بینی ریزش مشتری با XGBoost بر روی داده‌های بازاریابی
۹	۵.۳ پروفایل‌های DNA مشتری و WTP
۱۱	۵.۴ مدل طبقه‌بندی بازخرید کوپن
۱۲	۵.۵ بخش‌بندی RFM
۱۲	۵.۶ اثربخشی کمپین‌ها
۱۳	۶. جمع‌بندی
۱۳	۷. منابع

۱. مقدمه

دیتاست Dunnhumby سال‌هاست که برای پژوهشگران و تحلیل‌گران داده در زمینه‌ی خرده‌فروشی، بستر ارزشمندی به‌شمار می‌رود. این دیتاست که از رفتار خرید مشتریان واقعی جمع‌آوری شده، به طور چشمگیری برای مدل‌های داده‌کاوی، چه نظارت‌شده و چه نظارت‌نشده، کاربرد دارد. داده‌های آن، نه تنها تاریخچه‌ی خرید مشتریان را منعکس می‌کند، بلکه از منظر جمعیت‌شناسنامی و بازاریابی نیز ارزشمند است. [1]

۲. ساختار پایگاه داده و جداول کلیدی

به‌منظور ارائه‌ی شفاف‌تر داده‌ها، در این بخش ساختار هریک از جداول اصلی دیتاست Dunnhumby همراه با ستون‌های اصلی معرفی می‌شود.

۲.۱ جدول transactions.csv

این فایل، مهم‌ترین منبع داده‌ای برای ردیابی خریدهای انجام‌شده توسط مشتریان است. تحلیل رفتار خرید، ارزیابی وفاداری مشتری و محاسبه‌ی شاخص‌های کلیدی فروش، عمدتاً به کمک همین جدول صورت می‌گیرد.

نام ستون	توضیح تفصیلی
HOUSEHOLD_KEY	شناسه‌ی منحصر‌به‌فرد هر خانوار. از این ستون برای پیوند با جداول دیگر مانند مشخصات خانوار یا کوپن‌ها استفاده می‌شود.
BASKET_ID	شناسه‌ی یکتا برای هر سبد خرید یا صورت حساب. این ستون در تحلیل سطح خرید (Basket Analysis) کاربرد دارد.
DAY	شماره‌ی روز یا تاریخ خرید (در برخی نسخه‌ها به صورت عددی نشانگر روز در سال). امکان ردگیری توالی خرید و روند روزانه از این طریق فراهم است.
PRODUCT_ID	شناسه‌ی یکتا برای هر محصول خریداری‌شده؛ پیونددهنده‌ی اصلی میان تراکنش و اطلاعات توصیفی محصول.
QUANTITY	تعداد کالای خریداری‌شده در هر سطر تراکنش؛ مبنایی برای ارزیابی حجم مصرف یا تقاضا.
SALES_VALUE	ارزش فروش یا مبلغ پرداختی برای آن ردیف خرید. در سنجش شاخص‌های مالی نظیر درآمد هر تراکنش، نقش مهمی ایفا می‌کند.
STORE_ID	شناسه‌ی فروشگاه محل خرید. با این ستون می‌توان عملکرد فروشگاه‌های مختلف را مقایسه نمود.
RETAIL_DISC	میزان تخفیف فروشگاهی اعطاشده توسط خرده‌فروش. شناسایی میزان موققت طرح‌های تخفیفی به کمک این ستون امکان‌پذیر است.
COUPON_DISC	مبلغ تخفیف ناشی از کوپن. با اتصال به جدول کوپن‌ها، می‌توان تأثیر کوپن بر رفتار خرید را سنجید.
COUPON_MATCH_DISC	تحفیف اضافی در صورت انطباق کوپن با تخفیف خرده‌فروشی.
TRANS_TIME	زمان انجام تراکنش (مثلاً ساعت و دقیقه). کشف الگوهای ساعتی خرید از این طریق صورت می‌گیرد.

WEEK_NO	شماره‌ی هفته در تقویم سال. برای تحلیل‌های سری زمانی و بررسی الگوهای فصلی یا رویدادهای مقطعی کاربرد دارد.
YEAR	سال وقوع خرید. ستون مکمل برای تحلیل‌های طولی در کنار WEEK_NO است.

کاربرد پژوهشی

- محاسبه‌ی شاخص‌های فروش و تحلیل روند خرید.
- بررسی اثر تخفیف‌ها و کوپن‌ها بر مقدار خرید.
- کشف الگوهای رفتاری مشتریان در سبدهای مختلف.

۲.۲ جدول products.csv

این فایل، اطلاعات توصیفی مرتبط با هر محصول را شامل می‌شود و امکان شناسایی و گروه‌بندی انواع کالا را فراهم می‌کند.

	توضیح تفصیلی
PRODUCT_ID	شناسه‌ی یکتا برای پیوند با جدول تراکنش‌ها.
MANUFACTURER	نام یا کد سازنده یا تولیدکننده‌ی کالا. برای تحلیل وفاداری به برنده یا مقایسه‌ی تولیدکنندگان به کار می‌رود.
DEPARTMENT	دپارتمان یا بخش کلان کالایی (مثلًا نوشیدنی، لبنتیات). امکان گروه‌بندی وسیع محصولات را فراهم می‌آورد.
BRAND	نام تجاری محصول. در سنجش رقابت میان برندها کاربرد قابل توجهی دارد.
COMMODITY_DESC	توصیف کلی نوع محصول مانند "Soft Drinks" سطح طبقه‌بندی عمومی محصول را مشخص می‌کند.
SUB_COMMODITY_DESC	زیرگروه کالایی که جزئیات بیشتری را ارائه می‌دهد (مثلًا "Diet Soda").
CURR_SIZE_OF_PRODUCT	اندازه یا حجم بسته‌بندی (۱.۵ لیتر، ۵۰۰ گرم و ...). برای تحلیل ارزش هر بسته یا میزان مصرف مفید است.

کاربرد پژوهشی

- خوشه‌بندی محصولات بر اساس ویژگی‌های فیزیکی و گروه‌بندی.
- کشف قوانین هم خریدی با داده‌های تراکنش.
- بررسی تنوع برندها و رابطه‌ی آن با رفتار مشتری.

۲.۳ جدول hh_demographic.csv

این فایل حاوی ویژگی‌های دموگرافیک و ساختاری خانوار است که در ترکیب با داده‌های تراکنش، بینش‌های عمیقی از رفتار مشتری ارائه می‌دهد.

	توضیح تفصیلی
HOUSEHOLD_KEY	شناسه‌ی یکتا برای خانوار، جهت اتصال به داده‌های خرید یا کوپن.

AGE_DESC	گروه سنی سرپرست خانوار (مانند ۴۴-۳۵، ۳۴-۲۵). برای تحلیل الگوهای خرید مرتبط با سن استفاده می‌شود.
MARITAL_STATUS_CODE	وضعیت تأهل (مجرد، متأهل و ...).
INCOME_DESC	محدوده درآمد (K نظیر ۱۵-۲۴) برای مطالعه قدرت خرید و خوشبندی درآمدی به کار می‌رود.
HOMEOWNER_DESC	نوع مالکیت منزل (مالک یا مستأجر). در تحلیل الگوهای هزینه و سبک زندگی مؤثر است.
HH_COMP_DESC	ساختار خانوار (با فرزند، بدون فرزند و ...).
HOUSEHOLD_SIZE_DESC	تعداد اعضای خانوار (کوچک، متوسط، بزرگ).
KID_CATEGORY_DESC	مشخص می‌کند آیا خانوار دارای فرزند است یا خیر (و در چه رده‌ی سنی).

کاربرد پژوهشی

- خوشبندی مشتریان براساس ویژگی‌های جمعیت‌شناختی.
- مدل‌سازی رفتار خرید با در نظر گرفتن سن، درآمد یا تعداد اعضای خانوار.
- طراحی سیستم‌های پیشنهاد‌گر هدفمند متناسب با ساختار خانوار.

۲.۴ جدول casual_data.csv

در برخی نسخه‌های دیتابیس Dunnhumby، این فایل به نحوه‌ی نمایش محصول در فروشگاه و درج در کمپین‌های پستی می‌برد. از هر چند نام "casual_data" ممکن است نشانگر رفتارهای غیررسمی خرید باشد، اما ستون‌های مرتبط با Display و Mailer در اینجا قرار دارند که البته امکان بررسی همه مقادیر آن با خاطر کمبود اطلاعات اضافی برای فهمیدن کدگذاری‌های مقادیرشان، ممکن نیست.

نام ستون	توضیح تفصیلی
PRODUCT_ID	شناسه‌ی محصول یا کدی منحصر به فرد مرتبط با هر کالا که در سایر جداول (مانند تراکنش یا مشخصات محصول) نیز ظاهر می‌شود.
STORE_ID	کد فروشگاه، معرف محل عرضه محصول. امکان تفکیک عملکرد فروشگاه‌های گوناگون و مقایسه‌ی تبلیغات در مناطق مختلف از طریق این ستون وجود دارد.
WEEK_NO	شماره‌ی هفته در طول سال مبنایی برای انجام تحلیل‌های زمانی و تشخیص فصلیت یا بررسی الگوهای مقطعی (همچون رویدادهای تبلیغاتی).
DISPLAY	نحوه‌ی نمایش محصول (Display) گونه‌ای از برجسته‌سازی یا چینش کالاست. مقادیر عددی یا حروفی مانند ۹، ۰ یا A می‌توانند شکل‌های مختلف نمایش یا فقدان نمایش را نشان دهند.
MAILER	نامه‌ی تبلیغاتی (Mailer) تعیین می‌کند آیا کالای موردنظر در کمپین‌های نامه‌ای معرفی شده است یا خیر. مقدار «A» نشانگر درج در کمپین پستی، «۰» نشانه‌ی عدم درج، و مقادیر دیگر نیز وضعیت‌های گوناگون را ثبت می‌کنند.

۲.۵ جداول جانبی

مجموعه‌ای از جداول مکمل برای درک بهتر نحوه ارائه تخفیف‌ها و تبلیغات بازاریابی است. این اطلاعات در کنار داده‌های تراکنش یا مشخصات خانوار، تصویر کامل‌تری از رفتار مشتری ارائه می‌دهد.

۱. جدول campaigns.csv

- **CAMPAIGN:** شناسه‌ی کمپین
- **DESCRIPTION:** شرح مختصر یا عنوان کمپین
- **CAMPAIGN_TYPE:** طبقه‌بندی کمپین (مثلاً تخفیفی، معرفی محصول جدید)
- **START_DAY, END_DAY:** محدوده‌ی زمانی اجرای کمپین

۲. جدول coupon.csv

- **COUPON_UPC:** کد یکتای کوبن
- **PRODUCT_ID:** محصولات تحت شمول تخفیف کوبن
- **CAMPAIGN:** کمپین مرتبط با کوبن

۳. جدول coupon_redemptions.csv

- **HOUSEHOLD_KEY:** شناسنامه‌ی خانوار استفاده‌کننده از کوبن
- **DAY:** تاریخ استفاده از کوبن
- **COUPON_UPC:** کد کوبن
- **CAMPAIGN:** کمپین مرتبط

۴. جدول campaign_descriptions.csv

- **CAMPAIGN:** شناسه‌ی کمپین، مرتبط با جدول campaigns
- **DESCRIPTION:** تشریح جامع‌تر محتوای کمپین
- **HOUSEHOLD_KEY:** در برخی نسخه‌ها به خانوار خاصی اشاره دارد

۳. کاربردهای ممکن در داده‌کاوی

۳.۱ تحلیل سبد خرید (Market Basket Analysis)

- با تلفیق جداول products و transactions، می‌توان الگوهای هم‌خریدی کالاهای را کشف کرد.
- الگوریتم‌های کشف قوانین انجمنی مانند FP-Growth، محصولات پر تکرار یا همبسته را شناسایی می‌کنند.
- این تحلیل، در طراحی چیدمان فروشگاه و پیشنهادهای فروش مکمل کاربرد دارد.

۳.۲ مدل‌سازی پیش‌بینی (Predictive Modeling)

- ترکیب داده‌های transactions و households برای پیش‌بینی رفتار مشتری، از جمله احتمال ریزش (Churn) یا واکنش به کمپین.
- الگوریتم‌هایی چون رگرسیون، درخت تصمیم، جنگل‌های تصادفی یا شبکه‌های عصبی قابلیت اعمال دارند.

۳.۳ خوشبندی مشتریان (Customer Segmentation)

- با تکنیک‌های نظارت‌نشده مانند K-means، می‌توان گروه‌های مشتری را بر اساسی کرد.
- این خوشبندی در مدیریت ارتباط با مشتری و طراحی کمپین‌های هدفمند بسیار سودمند هستند.
- برای مثال، می‌توان مشتریان را بر اساس میانگین سبد خرید، فراوانی خرید یا درآمد خانوار تفکیک کرد.

۳.۴ تحلیل سری زمانی (Time Series Analysis)

- ستون‌های زمانی مانند DAY، WEEK_NO، YEAR امکان بررسی الگوهای فصلی و روند فروش را فراهم می‌کنند.
- سنجش تأثیر کمپین‌های کوتاه‌مدت یا تعطیلات تقویمی بر فروش به کمک مدل‌های سری زمانی میسر است.

۳.۵ بررسی تأثیر کمپین‌ها و کوپن‌ها (Campaign Effectiveness)

- ادغام جداول transactions و campaigns، coupon_redemptions برای سنجش میزان اثرگذاری تبلیغات.
- مقایسه‌ی خانوارهای دارای کوپن با خانوارهای فاقد کوپن می‌تواند درک عمیق‌تری از رفتار مشتریان به دست دهد.

۴. مزايا و چالش‌ها

مزايا

۱. تنوع داده‌ها: جداول گوناگون (تراکنش، خانوار، محصول، کمپین و ...) دیدی چندوجهی از رفتار مشتری فراهم می‌کند.
۲. واقع‌گرایی بالا: این داده‌ها برگرفته از تعاملات واقعی خرده‌فروشی است.
۳. مقیاس وسیع: تعداد بالای رکوردها امکان یافتن الگوهای معنادار و قدرتمند را افزایش می‌دهد.
۴. کاربرد همه‌جانبه: داده‌ها را می‌توان در زمینه‌هایی چون تحلیل سبد خرید، ارزش طول عمر مشتری و بررسی اثربخشی کمپین‌ها به کار برد.

چالش‌ها

۱. حجم انبوه داده: به زیرساخت‌های داده‌ای مقیاس پذیر مانند پایگاه‌های داده نیاز دارد.
۲. پاک‌سازی و یکپارچگی: فرمتهای متعدد جداول و وجود داده‌های پرت یا ناقص، نیازمند تلاش گستره‌ده در مرحله‌ی پیش‌پردازش است.

۵. نمونه کدهای انجام شده مرتبط با این دیتاست در Kaggle

۱.۵ تحلیل اکتشافی داده‌ها (EDA): جمعیت‌شناسی و الگوهای سبد خرید

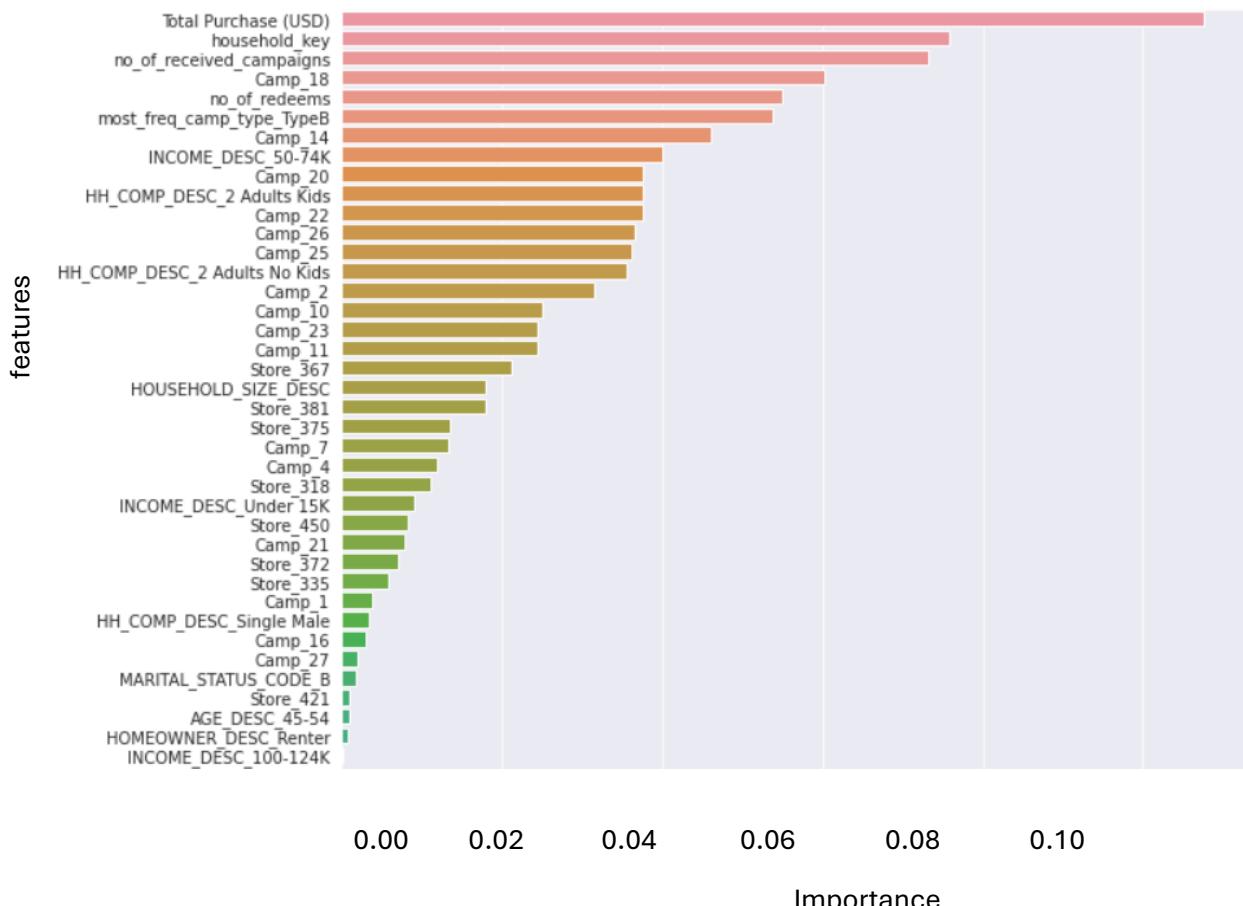
مجموعه‌ای از نوت‌بوک‌ها و کدها در Kaggle تحلیل عمیقی روی داده‌های Dunnhumby ارائه داده‌اند تا الگوهای رفتاری مشتریان را آشکار کنند. برای نمونه، یک تحلیل در Kaggle خانوارها را بر اساس عوامل دموگرافیکی (نظیر سن) دسته‌بندی کرده و تفاوت الگوهای خرید را مورد بررسی قرار داده است. این دسته‌بندی‌ها روندهایی را نشان می‌دهد (مثلاً الگوی هزینه کرد خانوارهای جوان در مقایسه با خانوارهای مسن). همچنین مشارکت‌کنندگان در Kaggle از تکنیک‌های تحلیل سبد خرید (market basket analysis) و کاوش قواعد انجمنی (association rule mining) برای شناسایی محصولاتی که اغلب به صورت همزمان خریداری می‌شوند استفاده کرده‌اند. این قواعد انجمنی، گروه‌بندی‌های رایج اقلام در یک سبد (مانند محصولات مکمل که معمولاً با هم خریده می‌شوند) را آشکار می‌کند و دیدگاه‌هایی برای استراتژی‌های فروش جانبی (cross-selling) و چیدمان محصول ارائه می‌دهد. [2]

۵.۲ پیش‌بینی ریزش مشتری با XGBoost بر روی داده‌های بازاریابی

تحلیل و پیش‌بینی ریزش مشتریان با استفاده از الگوریتم XGBoost (eXtreme Gradient Boosting).

این الگوریتم یکی از روش‌های ensemble است با ساخت مجموعه‌ای از درخت‌های تصمیم به صورت متوالی، تلاش می‌کند خطای مدل را در هر مرحله کاهش دهد با استفاده از gradient descent.

تحلیل‌ها نشان می‌دهند که میزان کل خریدهای یک مشتری (total purchase)، ویژگی مهمی در مدل پیش‌بینی ریزش است، به طوری که با افزایش مبلغ کل خرید، نرخ ریزش کاهش می‌یابد. این موضوع نشان‌دهنده‌ی این است که مشتریانی که بیشتر هزینه می‌کنند، کمتر احتمال دارد که از دست بروند.



۵.۳ پروفایل‌های DNA مشتری و WTP

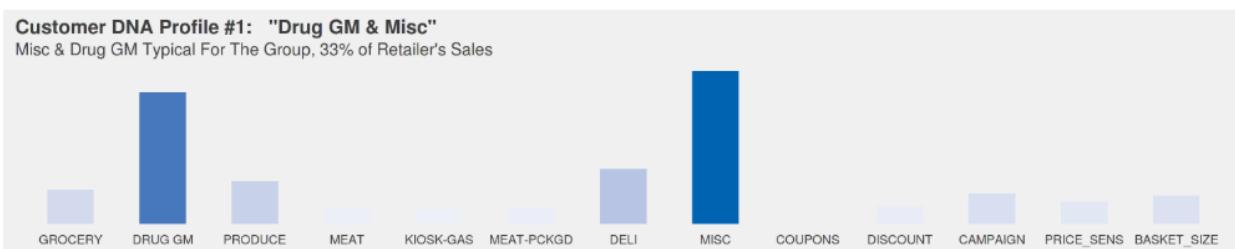
تحلیل داده‌های مشتریان از روش NMF (Non-negative Matrix Factorization) برای استخراج ویژگی‌های پنهان ساده و قابل تفسیر مثل دسته‌بندی محصولات خریداری شده، میزان استفاده از تخفیف‌ها، حساسیت به قیمت، اندازه سبد خرید و واکنش به کمپین‌های تبلیغاتی با هدف درک بهتر رفتار خرید مشتریان و تعیین تمایل به پرداخت (Willingness to Pay - WTP) آن‌ها است. در نهایت میخواهد به این نتیجه برسد که آیا مشتریان با درآمد بالاتر حاضرند برای محصولات قیمت بیشتری بپردازند؟

تحلیل توزیع تعداد خریدها در بازه‌های زمانی مختلف، شناسایی الگوهای خرید مشتریان در روزهای هفته و ساعات مختلف روز و بررسی مبلغ متوسط خریدها و تعداد اقلام در هر خرید از جمله تحلیل‌های رفتار مشتریان است.

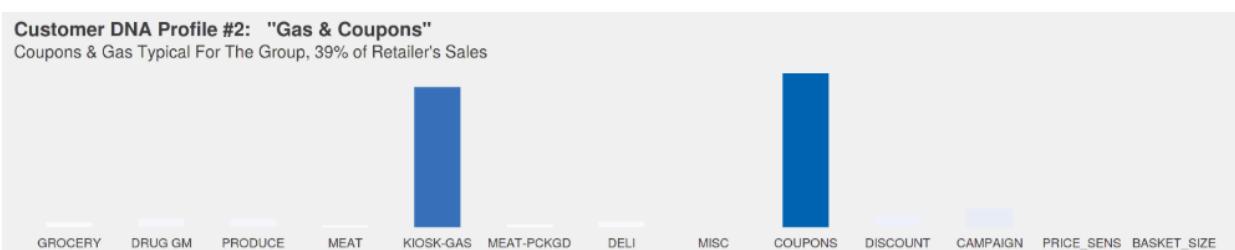
استراتژی‌های بازاریابی و قیمت‌گذاری باید بر اساس هر پروفایل مشتری شخصی‌سازی شود تا حداکثر سودآوری و وفاداری مشتریان حاصل شود.

یافته‌ها نشان دادند که میزان WTP مشتریان با افزایش درآمد آن‌ها افزایش می‌یابد. این موضوع به این معناست که مشتریان با درآمد بالاتر، حاضرند قیمت بیشتری برای محصولات بپردازند.

علاقه به محصولات دارویی و متفرقه، کمترین توجه به تخفیف و تبلیغات:

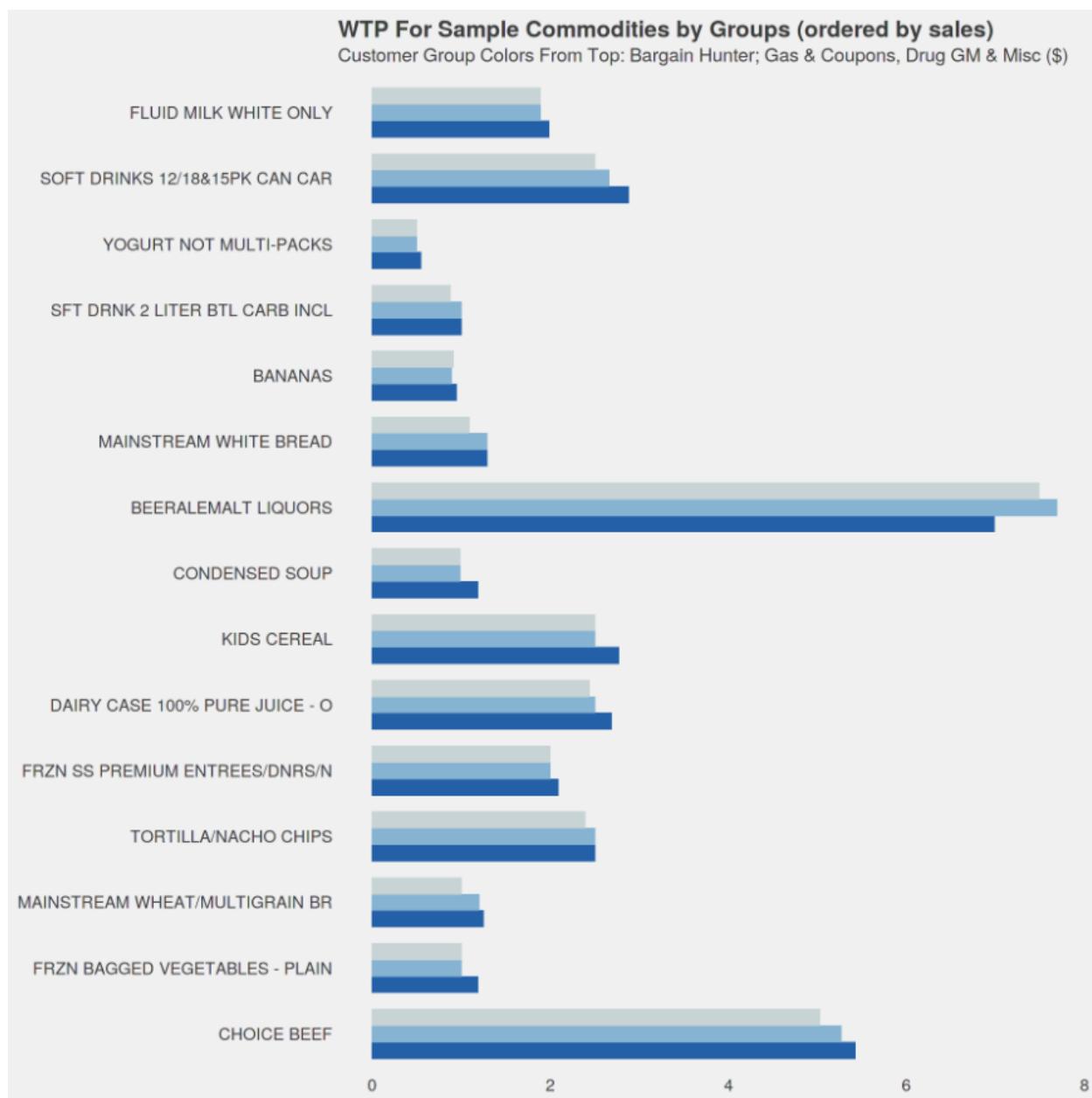


تمرکز بر خرید بنزین و استفاده زیاد از کوپن‌های تخفیف:



حساس به قیمت، جذب شده توسط تخفیف‌ها و تبلیغات، خریدهای بزرگ:





۵.۴ مدل طبقه‌بندی بازخرید کوپن

این مدل با هدف پیش‌بینی رفتار مشتریان در بازخرید کوپن‌های ارسالی انجام شده است و به شرکت‌ها کمک می‌کند تا در مورد ارسال کوپن‌ها به مشتریان تصمیم‌گیری بهتری داشته باشند و از هدر رفتن زمان و هزینه جلوگیری کنند.

در این مدل از روش Logistic Regression به عنوان مدل پایه استفاده شده و روش‌های مبتنی بر random forest و XGBoost برای مسائل کلاسیفیکیشن و بهینه سازی و افزایش دقت مدل استفاده شده است.

بر اساس داده‌های موجود، مشخص شده که ۷۰٪ از مشتریان هرگز از کوپن‌های دریافت‌شده استفاده نمی‌کنند. این امر منجر به هدر رفتن منابع مالی و انسانی شرکت می‌شود. بنابراین، هدف این پروژه ایجاد مدلی است که بتواند رفتار مشتریان را در ۵ کمپین آخر سال پیش‌بینی کند و مشخص نماید که کدام مشتریان کوپن‌های خود را بازخرید خواهند کرد و کدامیک این کار را انجام نخواهند داد.

نتایج تحلیل ها نشان داد که مشتریانی با درآمد بالاتر تمایل کمتری به استفاده از کوپن دارند، مشتریان متأهل و دارای فرزند بیشتر به کوپن ها حساس هستند و محصولات تخفیف دار بیشتر باعث جذب مشتریان به بازخرید کوپن ها می شود.

پیش از استفاده از مدل، تنها % ۳۸ از مشتریان کوپن های خود را بازخرید می کردند ولی پس از استفاده از مدل، شرکت می تواند به دقت پیش بینی کند که چه کسانی کوپن های خود را بازخرید نمی کنند و از ارسال کوپن به این افراد صرف نظر کند.

۵.۵ بخش بندی RFM

برخی از کاربران Kaggle از تحلیل RFM (Recency-Frequency-Monetary) بر اساس تازگی خرید (Recency)، بسامد خرید (Frequency) و سطح مخارج (Monetary) استفاده کرده اند. در این روش سنتی بازاریابی، به هر خانوار امتیاز های R ، M اختصاص داده می شود و آنها یکی از امتیاز های مشابه دارند در یک گروه قرار می گیرند تا مشتریان بالارزش یا از دست رفته (lapsed) شناسایی شوند. برخی از کارهای اخیر در Kaggle فقط به RFM بسته نکرده اند و نتایج آن را با خوش بندی مبتنی بر الگوریتم های یادگیری ماشین مقایسه کرده اند. برای نمونه، یک نوت بوک ابتدا RFM را انجام داده و سپس نتایج را با خروجی خوش بندی K-Means بر روی همان داده های مشتری مقایسه کرده است. این مقایسه نشان می دهد که آیا گروه بندی شهودی مبتنی بر RFM با خوش بندی داده محور هماهنگ است یا نه. [3]

۵.۶ اثربخشی کمپین ها

تحلیل های موجود در Kaggle همچنین به داده های کمپین های Dunnhumby پرداخته اند تا اثربخشی بازاریابی را ارزیابی و حتی بهینه کنند. در یک تحلیل اکتشافی، روند هزینه کرد خانوارها پیش، حین و پس از کمپین های خاص رصد شده است. نتایج نشان می دهد برخی کمپین ها باعث افزایش قابل توجه در هزینه کرد هفتگی شده اند؛ برای مثال، مشخص شده که کمپین ۱۳ و ۸ با جهش محسوسی در طول دوره تبلیغ مرتبط بوده اند. این نوع بینش با مرتبط سازی مواجهه مخاطبان با خرید شان، شواهد روشی از تأثیر کمپین ارائه می دهد و فراتر از بحث کلی گزارش اولیه درباره اثربخشی کمپین است. با شناسایی کمپین هایی که اوج فروش یا مشارکت را رقم می زنند، خرده فروشان در می یابند کدام تاکتیک ها بیشترین اثرگذاری را داشته اند. [2]

فراتر از تحلیل کمپین های گذشته، برخی از پژوهش های جدید به بهینه سازی تبلیغات آینده هم می پردازند. به عنوان یک نمونه قابل توجه، یک متخصص داده از تراکنش ها برای شبیه سازی استراتژی های تخفیف در یک خط محصول استفاده کرده و دریافته که تخفیف ۷٪ روی یک کالای محبوب (Ragu pasta sauce) حداکثر سود را ایجاد می کند و حدود ۱۷۹.۶۵ دلار سود هفتگی اضافی به ارمغان می آورد. این مثال نشان می دهد که چگونه ترکیب داده های خرید با مدل سازی سناریو می تواند به کمپین های اثربخش تری منجر شود؛ در اینجا با ایجاد توازن میان عمق تخفیف و حجم فروش، سود خالص افزایش می یابد. چنین تحلیل های بهینه سازی، بعد از بحث سنجش کارایی کمپین، پرسش «کدام کمپین های گذشته موفق بودند؟» را با «بهترین کمپین آتی کدام است؟» تکمیل می کنند. [4]

۶. جمع‌بندی

دیتاست Dunnhumby با تنوع بالایی از جداول و ستون‌ها، ابزاری اساسی برای مطالعات گستردگی در حوزه‌ی خرده‌فروشی است. داده‌های تراکنش، ویژگی‌های محصولات، مشخصات خانوار و اطلاعات کمپین‌ها، همگی قابلیت ترکیب دارند تا الگوها و بینش‌های ارزشمندی در حوزه‌های گوناگون بازاریابی و تحلیل داده به دست آید. از تحلیل سبد خرید و خوشبندی مشتریان تا ارزیابی تأثیر کوپن‌ها و تدوین استراتژی‌های تبلیغاتی، همگی می‌توانند بر روی این دیتاست پیاده‌سازی شوند.

۷. منابع

- [1] “Dunnhumby - The Complete Journey.” [Online]. Available: <https://www.kaggle.com/datasets/frtgnn/dunnhumby-the-complete-journey>
- [2] “Dunnhumby Exploratory Data Analysis.” [Online]. Available: <https://kaggle.com/code/simonhchen/dunnhumby-exploratory-data-analysis>
- [3] “RFM Analysis and K-Means Clustering Comparison.” [Online]. Available: <https://kaggle.com/code/analystoleksandra/rfm-analysis-and-k-means-clustering-comparison>
- [4] “Optimising Price Discounts on the Dunnhumby Carbo-Loading Dataset 🍝 | by Aum Damrongkitkanwong | Medium.” [Online]. Available: <https://medium.com/@aumdamrong/optimising-pasta-discounts-on-the-dunnhumby-carbo-loading-dataset-ae602d394df8>