**بِسْمِ اَلْلَّهِ اَلْرَّحْمَٰنِ اَلْرَّحِيمِ**

****

**دانشكده مهندسي برق و كامپيوتر**

**بررسی مجموعه‌داده Dunnhumby برای پروژه پایانی کارشناسی**

*Analysis of the Dunnhumby Dataset for the Bachelor’s Final Project*

**فاطمه سادات دست‌یافته شماره دانشجويي: ۴۰۰۲۱۱۶۰0۱4**

**سمانه تابنده شماره دانشجويي: 40021160009**

**سينا مظفري راد شماره دانشجويي: ۴۰۰۲۱۱۶۰۲۱4**

**گزارش اولیه بررسی ساختار مجموعه داده پروژه کارشناسی**

**استاد وحیدی‌پور**

**اسفند ۱۴03**

**چکیده**

دیتاست Dunnhumby، 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 کوپن‌ها 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. مقدمه 3](#_Toc191551497)

[۲. ساختار پایگاه داده و جداول کلیدی 3](#_Toc191551498)

[**۲.۱ جدول transactions.csv** 3](#_Toc191551499)

[**۲.۲ جدول products.csv** 4](#_Toc191551500)

[**۲.۳ جدول hh\_demographic.csv** 4](#_Toc191551501)

[**۲.۴ جدول casual\_data.csv** 5](#_Toc191551502)

[**۲.۵ جداول جانبی** 6](#_Toc191551503)

[۳. کاربردهای ممکن در داده‌کاوی 6](#_Toc191551504)

[**۳.۱ تحلیل سبد خرید (Market Basket Analysis)** 6](#_Toc191551505)

[**۳.۲ مدل‌سازی پیش‌بینی (Predictive Modeling)** 6](#_Toc191551506)

[**۳.۳ خوشه‌بندی مشتریان (Customer Segmentation)** 7](#_Toc191551507)

[**۳.۴ تحلیل سری زمانی (Time Series Analysis)** 7](#_Toc191551508)

[**۳.۵ بررسی تأثیر کمپین‌ها و کوپن‌ها (Campaign Effectiveness)** 7](#_Toc191551509)

[۴. مزایا و چالش‌ها 7](#_Toc191551510)

[5. نمونه کدهای انجام شده مرتبط با این دیتاست در Kaggle 8](#_Toc191551511)

[**5.۱ تحلیل اکتشافی داده‌ها (EDA): جمعیت‌شناسی و الگوهای سبد خرید** 8](#_Toc191551512)

[**5.2 پیش‌بینی ریزش مشتری با XGBoost بر روی داده‌های بازاریابی** 8](#_Toc191551513)

[**5.3 پروفایل‌های DNA مشتری و WTP** 9](#_Toc191551514)

[**5.4 مدل طبقه‌بندی بازخرید کوپن** 11](#_Toc191551515)

[**5.5 بخش‌بندی RFM** 12](#_Toc191551516)

[**5.6 اثربخشی کمپین‌ها** 12](#_Toc191551517)

[6. جمع‌بندی 13](#_Toc191551518)

[7. منابع 13](#_Toc191551519)

**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، نه‌تنها تاریخچه‌ی خرید مشتریان را منعکس می‌کند، بلکه از منظر جمعیت‌شناختی 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 صورت‌حساب. این ستون در تحلیل سطح خرید (Basket Analysis) کاربرد دارد. |
| DAY | شماره‌ی روز یا تاریخ خرید (در برخی نسخه‌ها به‌صورت 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 موفقیت 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 | تخفیف 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.

|  |  |
| --- | --- |
| نام ستون | توضیح تفصیلی |
| PRODUCT\_ID | شناسه‌ی یکتا برای پیوند با جدول تراکنش‌ها. |
| MANUFACTURER | نام یا کد سازنده یا تولیدکننده‌ی کالا. برای تحلیل وفاداری به برند یا مقایسه‌ی تولیدکنندگان به‌کار می‌رود. |
| DEPARTMENT | دپارتمان یا بخش کلان کالایی (مثلاً نوشیدنی، لبنیات). امکان گروه‌بندی وسیع محصولات را فراهم می‌آورد. |
| BRAND | نام تجاری محصول. در سنجش رقابت میان برندها کاربرد قابل‌توجهی 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 (مثلاً"Diet Soda"). |
| CURR\_SIZE\_OF\_PRODUCT | اندازه یا حجم بسته‌بندی (1.5 لیتر، 500 گرم و ...). برای تحلیل ارزش هر بسته یا میزان مصرف مفید است. |

**کاربرد پژوهشی**

* خوشه‌بندی محصولات بر اساس ویژگی‌های فیزیکی و گروه‌بندی.
* کشف قوانین هم‌خریدی 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.

|  |  |
| --- | --- |
| نام ستون | توضیح تفصیلی |
| household\_key | شناسه‌ی یکتا برای خانوار، جهت اتصال به داده‌های خرید یا کوپن. |
| AGE\_DESC | گروه سنی سرپرست خانوار (مانند 25-34، 35-44). برای تحلیل الگوهای خرید مرتبط با سن استفاده می‌شود. |
| MARITAL\_STATUS\_CODE | وضعیت تأهل (مجرد، متأهل و ...). |
| INCOME\_DESC | محدوده‌ی درآمد (K نظیر 15-24) برای مطالعه‌ی قدرت خرید و خوشه‌بندی درآمدی به‌کار 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**

* خوشه‌بندی 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 دیتاست Dunnhumby، این فایل به نحوه‌ی نمایش محصول در فروشگاه و درج در کمپین‌های پستی می‌پردازد. هرچند 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 بخاطر کمبود اطلاعات اضافی برای فهمیدن کدگذاری‌های مقادیرشان، ممکن نیست.

|  |  |
| --- | --- |
| نام ستون | توضیح تفصیلی |
| PRODUCT\_ID | شناسه‌ی محصول یا کدی منحصربه‌فرد مرتبط با هر کالا که در سایر جداول (مانند تراکنش یا مشخصات محصول) نیز ظاهر می‌شود. |
| STORE\_ID | کد فروشگاه، معرف محل عرضه‌ی محصول. امکان تفکیک عملکرد فروشگاه‌های گوناگون و مقایسه‌ی تبلیغات در مناطق متفاوت 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 0، 9 یا A می‌توانند شکل‌های متفاوت 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، «0» نشانه‌ی عدم درج، و مقادیر دیگر نیز وضعیت‌های گوناگون را ثبت می‌کنند. |

**۲.۵ جداول جانبی**

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

1. **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
2. **جدول coupon.csv** 
   * **COUPON\_UPC**: کد یکتای کوپن
   * **PRODUCT\_ID**: محصولات تحت شمول تخفیف کوپن
   * **CAMPAIGN**: کمپین مرتبط با کوپن
3. **جدول coupon\_redemptions.csv**
   * **household\_key**: شناسنامه‌ی خانوار استفاده‌کننده از کوپن
   * **DAY**: تاریخ استفاده از کوپن
   * **COUPON\_UPC**: کد کوپن
   * **CAMPAIGN**: کمپین مرتبط
4. **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 transactions و products، می‌توان الگوهای هم‌خریدی کالاها را کشف کرد.
* الگوریتم‌های کشف قوانین انجمنی مانند Apriori یا FP-Growth، محصولات پرتکرار 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 households و transactions برای پیش‌بینی رفتار مشتری، ازجمله احتمال ریزش (Churn) یا واکنش به کمپین.
* الگوریتم‌هایی چون رگرسیون، درخت تصمیم، جنگل‌های تصادفی یا شبکه‌های عصبی قابلیت اعمال دارند.

**۳.۳ خوشه‌بندی مشتریان (Customer Segmentation)**

* با تکنیک‌های نظارت‌نشده مانند K-means ، می‌توان گروه‌های مشتری با رفتار خرید مشابه را شناسایی کرد.
* این خوشه‌ها 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 campaigns, coupon\_redemptions و transactions برای سنجش میزان اثرگذاری تبلیغات.
* مقایسه‌ی خانوارهای دارای کوپن با خانوارهای فاقد کوپن می‌تواند درک عمیق‌تری از رفتار مشتریان به دست دهد.

**۴. مزایا و چالش‌ها**

**مزایا**

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

**چالش‌ها**

1. **حجم انبوه داده**: به زیرساخت‌های داده‌ای مقیاس‌پذیر مانند پایگاه‌های داده نیاز دارد.
2. **پاک‌سازی و یکپارچگی**: فرمت‌های متنوع جداول و وجود داده‌های پرت یا ناقص، نیازمند تلاش گسترده در مرحله‌ی پیش‌پردازش است.

**5. نمونه کدهای انجام شده مرتبط با این دیتاست در Kaggle**

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

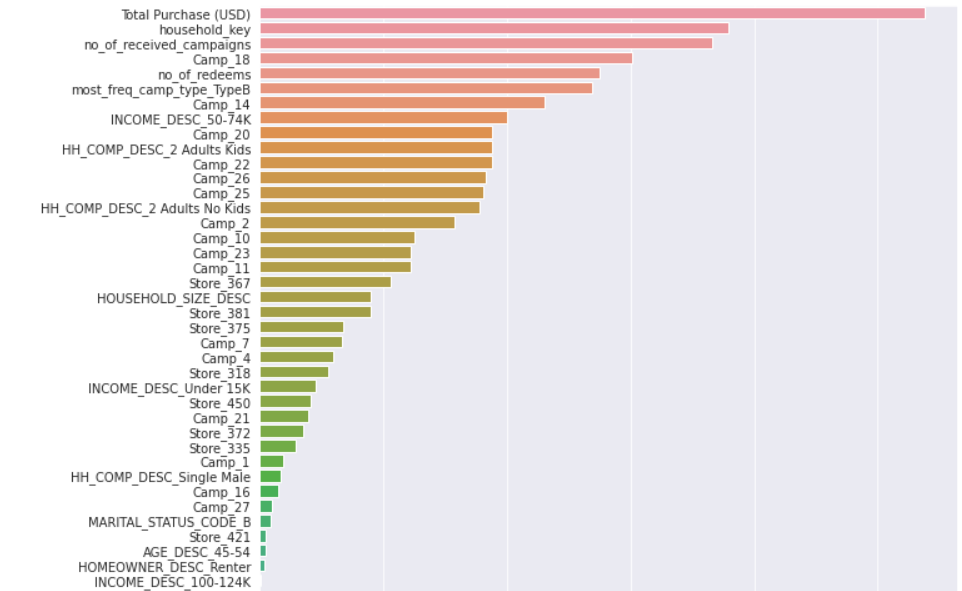
مجموعه‌ای از نوت‌بوک‌ها و کدها در Kaggle تحلیل عمیقی روی داده‌های Dunnhumby ارائه داده‌اند تا الگوهای رفتاری مشتریان را آشکار کنند. برای نمونه، یک تحلیل در Kaggle خانوارها را بر اساس عوامل دموگرافیکی (نظیر سن) دسته‌بندی کرده و تفاوت الگوهای خرید را مورد بررسی قرار داده است .این دسته‌بندی‌ها روندهایی را نشان می‌دهد (مثلاً الگوی هزینه‌کرد 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 به‌صورت 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]

**5.2 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 روش های ensembleاست با ساخت مجموعه‌ای از درخت‌های تصمیم به‌صورت 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، به‌طوری‌که 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.



features

0.00 0.02 0.04 0.06 0.08 0.10

Importance

**5.3 پروفایل‌های DNA مشتری و WTP**

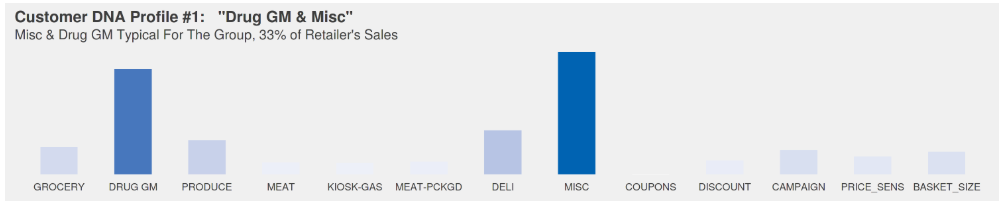
تحلیل داده‌های مشتریان از روش NMF (Non-negative Matrix Factorization)برای استخراج ویژگی‌های پنهان ساده و قابل تفسیر مثل دسته‌بندی محصولات خریداری‌شده، میزان استفاده از تخفیف‌ها، حساسیت به قیمت، اندازه سبد خرید و واکنش به کمپین‌های تبلیغاتی با هدف درک بهتر رفتار خرید مشتریان و تعیین تمایل به پرداخت (Willingness 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 تحلیل های رفتار مشتریان است.

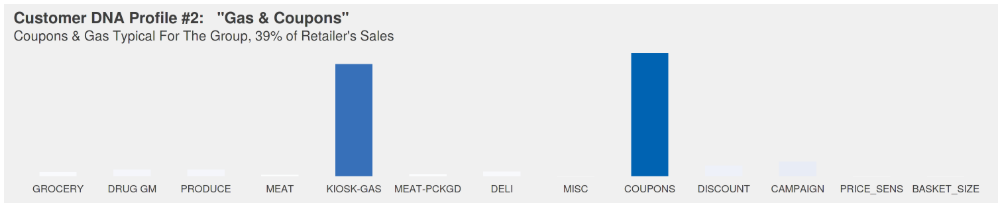
استراتژی‌های بازاریابی و قیمت‌گذاری باید بر اساس هر پروفایل مشتری شخصی‌سازی شود تا حداکثر سودآوری و وفاداری مشتریان حاصل شود.

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:

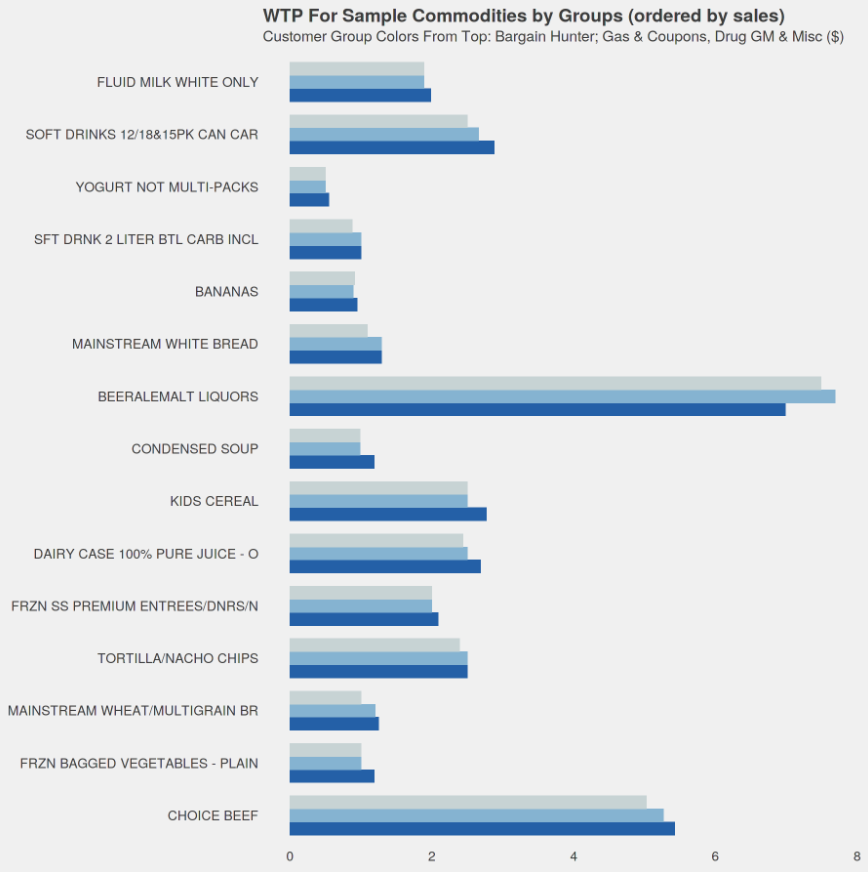


تمرکز بر خرید بنزین و استفاده زیاد از کوپن‌های تخفیف:



حساس به قیمت، جذب‌شده توسط تخفیف‌ها و تبلیغات، خریدهای بزرگ:





**5.4 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 روش های مبتنی برrandom forest و XGBoost برای مسائل کلاسیفیکیشن و بهینه سازی و افزایش دقت مدل استفاده شده است.

بر اساس داده‌های موجود، مشخص شده که % 70 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 5 کمپین آخر سال پیش‌بینی کند و مشخص نماید که کداممشتریانکوپن‌های 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 کوپن‌ها حساس هستند و محصولات تخفیف‌دار بیشتر باعث جذب مشتریان به بازخرید کوپن‌ها می‌شود.

پیش از استفاده از مدل، تنها % 38 از مشتریان کوپن‌های خود را بازخرید می‌کردند ولی پس از استفاده از مدل، شرکت می‌تواند به دقت پیش‌بینی کند که چه کسانی کوپن‌های خود را بازخرید نمی‌کنند و از ارسال کوپن به این افراد صرف‌نظر کند.

**5.5 بخش‌بندی RFM**

برخی از کاربران Kaggle از تحلیل RFM (Recency-Frequency-Monetary) برای بخش‌بندی مشتریان بر اساس تازگی خرید (Recency)، بسامد خرید (Frequency) و سطح مخارج (Monetary) استفاده کرده‌اند. در این روش سنتی بازاریابی، به هر خانوار امتیازهای R، F و M اختصاص داده می‌شود و آنهایی که امتیازهای مشابه دارند در یک گروه قرار می‌گیرند تا مشتریان باارزش یا از دست‌رفته (lapsed) شناسایی شوند. برخی از کارهای اخیر در Kaggle فقط به RFM بسنده نکرده‌اند و نتایج آن را با خوشه‌بندی مبتنی بر الگوریتم‌های یادگیری ماشین مقایسه کرده‌اند. برای نمونه، یک نوت‌بوک ابتدا RFM را انجام داده و سپس نتایج را با خروجی خوشه‌بندی K-Means بر روی همان داده‌های مشتری مقایسه کرده است . این مقایسه نشان می‌دهد که آیا گروه‌بندی شهودی مبتنی بر RFM با خوشه‌بندی داده‌محور هماهنگ است یا نه. [3]

**5.6 اثربخشی کمپین‌ها**

تحلیل‌های موجود در Kaggle همچنین به داده‌های کمپین‌های Dunnhumby پرداخته‌اند تا اثربخشی بازاریابی را ارزیابی و حتی بهینه کنند. در یک تحلیل اکتشافی، روند هزینه‌کرد خانوارها پیش، حین و پس از کمپین‌های خاص رصد شده است. نتایج نشان می‌دهد برخی کمپین‌ها باعث افزایش قابل توجه در هزینه‌کرد هفتگی شده‌اند؛ برای مثال، مشخص شده که کمپین 13 و 8 با جهش محسوسی در هزینه‌کرد هفتگی در طول دوره تبلیغ مرتبط بوده‌اند .این نوع بینش با مرتبط‌سازی مواجهه مخاطبان با خریدشان، شواهد روشنی از تأثیر کمپین ارائه می‌دهد و فراتر از بحث کلی گزارش اولیه درباره اثربخشی کمپین است. با شناسایی کمپین‌هایی که اوج فروش یا مشارکت را رقم می‌زنند، خرده‌فروشان درمی‌یابند کدام تاکتیک‌ها بیشترین اثرگذاری را داشته‌اند. [2]

فراتر از تحلیل کمپین‌های گذشته، برخی از پژوهش‌های جدید به **بهینه‌سازی تبلیغات آینده** هم می‌پردازند. به‌عنوان یک نمونه قابل توجه، یک متخصص داده از تراکنش‌ها برای شبیه‌سازی استراتژی‌های تخفیف در یک خط محصول استفاده کرده و دریافته که تخفیف 7% روی یک کالای محبوب (Ragu pasta sauce) حداکثر سود را ایجاد می‌کند و حدود 179.65 دلار سود هفتگی اضافی به ارمغان می‌آورد . این مثال نشان می‌دهد که چگونه ترکیب داده‌های خرید با مدل‌سازی سناریو می‌تواند به کمپین‌های اثربخش‌تری منجر شود؛ در اینجا با ایجاد توازن میان عمق تخفیف و حجم فروش، سود خالص افزایش می‌یابد. چنین تحلیل‌های بهینه‌سازی، بعد از بحث سنجش کارایی کمپین، پرسش «کدام کمپین‌های گذشته موفق بودند؟» را با «بهترین کمپین آتی کدام است؟» تکمیل می‌کنند. [4]

**6. جمع‌بندی**

دیتاست Dunnhumby با تنوع بالایی از جداول و ستون‌ها، ابزاری اساسی برای مطالعات گسترده 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 به‌دست آید. از تحلیل سبد خرید و خوشه‌بندی مشتریان تا ارزیابی تأثیر کوپن‌ها و تدوین استراتژی‌های تبلیغاتی، همگی می‌توانند بر روی این دیتاست پیاده‌سازی شوند.

**7. منابع**

[1] “Dunnhumby - The Complete Journey.” [Online]. Available: <https://www.kaggle.com/datasets/frtgnn/dunnhumby-the-complete-journey>

[2] “Dunnhummby Exploratory Data Analysis.” [Online]. Available: <https://kaggle.com/code/simonhchen/dunnhummby-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>