مسئله

دیتایی که در اختیار شماست قسمتی از سفارشات مشتریان است. با استفاده از این دیتا سعی کنید به درخواست‌های زیر پاسخ دهید.

- رفتار مشتریان را تحلیل و توصیف کنید و براساس الگوهای رفتاری آن‌ها را به دسته‌های مختلف تقسیم کنید.
- بعد از چند بار خرید، می‌توان گفت یک مشتری وفادار است؟
- آیا مشتریان وفادار الگوی رفتاری یکسانی دارند؟
- با گذشت چند روز از خرید می‌توان گفت مشتری churn شده است؟
- آیا مشتریانی که churn می‌شوند الگوی رفتاری یکسانی دارند؟
- یک دشبورد آماده کنید که مدیر مارکتینگ بتواند روزانه KPIهای مهم را بررسی کند. (Tableau یا Power BI)

هدف از این تسک، بررسی نحوه‌ی نگاه شما به مسئله و قدرت تحلیل شماست، پس مسیر تحلیل و نتیجه‌گیری خود را به صورت شفاف توضیح دهید. برای انجام این تسک می‌توانید از هر ابزاری که با آن راحت‌تر هستید استفاده کنید ولی استفاده از پایتون امتیاز بیشتری دارد.

دیتای ورودی

دیتای سفارش مشتریان دارای ۸۴۵,۴۲۱ رکورد و ستون‌های زیر است:

- Order Number: شماره‌ی سفارش
- Created At: زمان ثبت سفارش
- User ID: آی‌دی مشتری
- Main Category: دسته‌بندی اصلی
- Total Shipping Fee: هزینه‌ی ارسال اولیه‌ی سفارش
- Final Shipping Fee: هزینه‌ی ارسال نهایی سفارش
- Items: تعداد آیتم کتگوری سفارشی
- Price: قیمت اولیه محصولات کتگوری سفارشی
- Discount: تخفیف محصولات کتگوری سفارشی
- Voucher Discount: مبلغ کد تخفیف کتگوری سفارشی
- City: شهر مشتری

پردازش دیتا

تمیزسازی داده

برای انجام بررسی‌های اولیه‌ی فایل ورودی از کتابخانه‌ی pandas در پایتون استفاده شده است. پس از خواندن فایل CSV، اقدام به تمیزسازی داده‌ها صورت گرفت. برای این منظور، ستون‌هایی که ماهیت عددی ندارند به string تغییر داده شدند. یکی از ایرادهایی که در این مرحله مشاهده شد، این مسئله بود که ستون created_at از نوع datetime نبود و بعضی سطرهای آن به شکل عددی ذخیره شده بودند. به نظر می‌آید مشکل به علت یکی نبودن سورس‌ها به وجود آمده باشد. با توجه به range اعداد، به احتمال زیاد این اعداد باید با فرمت اکسل به داده‌های زمانی تبدیل می‌شدند. بنابراین برای رفع این مشکل، سطرهای عددی با یک فیلتر مناسب شناسایی شدند و به datetime تغییر داده شدند.

در مرحله‌ی Data Cleaning ایراد دیگری مشاهده نشد؛ به جز اینکه بعضی رکوردها دارای main_category بدون مقدار بودند. برای از دست ندادن مشاهدات، این سطرها حذف نشدند تا بعدها در صورت نیاز در مراحل نمایش دشبورد حذف گردد.

افزودن ستون‌های مورد نیاز

برای استفاده‌ی بهتر از دیتا، نیاز به اضافه شدن چند ستون بود که از ستون‌های دیگر به دست می‌آید:

- مجموع تخفیف (total_discount): جمع ستون‌های discount و voucher_discount (با توجه به وجود دو نوع تخفیف برای هر محصول در هر خرید)
- ارزش (value): تعداد محصول خریداری‌شده ضربدر قیمت آن
- ارزش با احتساب تخفیف (discounted_value): ارزش خرید منهای مجموع تخفیف
- تخفیف ارسال (shipping_fee_discount): هزینه‌ی ارسال اولیه منهای هزینه‌ی ارسال نهایی

باید به این نکته کرد که در محاسبه‌ی discounted_value، نباید تخفیف ارسال در نظر گرفته شود. چرا که تخفیف ارسال مختص به هر سفارش است و نه هر کالا. بنابراین در بخش بعد و با تجمیع کردن اطلاعات هر سفارش، مبلغ نهایی نیز محاسبه می‌شود.

یک ابهام در دیتای ورودی این بود که مقدار کد تخفیف (voucher_discount) بر روی سفارش اعمال شده و یا بر روی هر کالا؟ با بررسی چند مورد از سفارش‌ها، مشخص شد که این تخفیف نیز مانند تخفیف محصول (discount) بر روی محصولات اعمال شده است. همچنین مشخص نبود که مقدار تخفیف به ازای هر کالا تعریف شده و روی قیمت محاسبه می‌شود یا برای کل آیتم‌های یک محصول؟ این ابهام نیز در بخش بعد برطرف شد که توضیح داده خواهد شد.

تجمیع اطلاعات سفارش‌ها

با توجه به این که چندین ستون داده‌های ورودی مرتبط با کل سفارش هستند و نه مرتبط با هر کالای خریداری‌شده، نیاز به تجمیع اطلاعات سفارش‌ها وجود دارد. برای این منظور، روی ستون‌های مشترک برای هر سفارش group by انجام شد و اطلاعات مورد نیاز از باقی ستون‌ها، به شکل aggregate شده اضافه شد. این ستون‌ها عبارت اند از:

- تعداد همه‌ی کالاهای خریداری‌شده (all_items)
- تعداد کالاهای منحصربه‌فرد خریداری شده (distinct_items)
- تعداد دسته‌بندی منحصربه‌فرد کالاهای خریداری‌شده (distinct_categories)
- مجموع تخفیف محصول (discount)
- مجموع کد تخفیف (voucher_discount)
- مجموع تخفیف (total_discount)
- مجموع ارزش (value)
- مجموع ارزش با احتساب تخفیف (discounted_value)
- تخفیف کل (final_discount): برابر با مجموع تخفیف به علاوه‌ی تخفیف ارسال
- مبلغ پرداخت‌شده (paid_value): برابر با مجموع ارزش با احتساب تخفیف به علاوه‌ی هزینه‌ی ارسال نهایی

در این مرحله همچنین با index شدن dataframe روی order_number، این اطمینان حاصل شد که ستون‌هایی که در هر order یکتا فرض شده بودند، در تمام سفارش‌ها به این صورت هستند.

در مورد ابهام توضیح داده شده در بخش قبل، در این قسمت ابتدا فرض شد که تخفیف به ازای هر کالا داده شده (یعنی ابتدا باید مقدار تخفیف را از قیمت کم کرد و سپس ضربدر تعداد کرد). با این فرض، مشخص شد که در مواردی مبلغ پرداخت‌شده

منفی به دست می‌آید. بنابراین این فرضیه رد شد و مشخص گردید که مبلغ تخفیف را باید از ارزش خرید کم کرد و این ابهام نیز برطرف گردید.

تجمیع اطلاعات مشتریان

با توجه به پرسش‌های مطرح‌شده در مسئله، نیاز به تجمیع اطلاعات برای مشتریان نیز وجود دارد. برای این منظور، ستون‌های aggregate شده‌ی زیر ساخته شد:

- تاریخ اولین سفارش (first_order)
- تاریخ آخرین سفارش (last_order)
- تعداد سفارش‌ها (orders_count)
- مجموع آیتم‌های خریداری‌شده (all_items)
- میانگین آیتم‌های خریداری‌شده در هر سفارش (avg_items)
- میانگین مبلغ پرداخت‌شده در هر سفارش (avg_paid_value)
- مجموع مبالغ پرداخت‌شده (total_paid_value)
- میانگین هزینه‌ی ارسال در هر سفارش (avg_shipping_fee)
- میانگین نرخ تخفیف استفاده‌شده (avg_discount_rate)
- میانگین مقدار تخفیف استفاده‌شده (avg_final_discount)
- میانگین فاصله‌ی روزانه بین سفارش‌ها (avg_order_interval)
- تعداد روز سپری‌شده از آخرین سفارش (days_from_last_order)

همچنین برای دسته‌بندی مشتریان، ستون‌های زیر تعریف شد:

- مشتری وفادار (loyal_customer): مشتریانی که حداقل یک سفارش داشته اند، از آخرین روز سفارش آن‌ها حداکثر ۳۰ روز می‌گذرد، و فاصله‌ی بین سفارش‌های آن‌ها حداکثر ۲۰ روز باشد.
- مشتری رویگردان (churned_customer): مشتریانی که از آخرین روز سفارش آن‌ها، حداقل دو برابر فاصله‌ی بین سفارش‌های خود آن‌ها گذشته باشد، فاصله‌ی بین سفارش‌های آن‌ها حداقل ده روز بوده باشد، و مشتری وفادار نباشند.
- مشتری پرارزش (high_value_customer): بیست درصد مشتریان که بیشترین ارزش را برای شرکت خلق کرده اند.
- مشتری پرخرید (high_paid_customer): بیست درصد مشتریان که بیشترین میانگین خرید در هر سفارش را دارند.
- مشتری پرسفارش (frequent_order_customer): مشتریانی که فاصله‌ی زمانی بین سفارش‌های آنان کمتر از ۳۰٪ دیگران و تعداد سفارش‌های آنان بیشتر از ۷۰ درصد دیگران باشد.
- مشتری پرکالا (multiple_item_customer): بیست درصد مشتریان که به صورت میانگین بیشترین تعداد محصول را در سبدهایشان داشته اند.
- مشتری پرتخفیف (discount_rate_customer): بیست درصد مشتریان که از بیشترین نرخ تخفیف استفاده کرده‌اند.
- مشتری بدون هزینه‌ی ارسال (zero_shipping_fee_customer): مشتریانی که در مجموع هیچ هزینه‌ی ارسالی پرداخت نکرده اند.

در این مرحله مشخص شد که با این تعاریف، بعد از ۴٫۶ بار خرید می‌توان گفت که یک مشتری وفادار شده است. همچنین بعد از گذشت ۷۶٫۴ روز از عدم سفارش مشتری، می‌توان گفت که او churn شده است. به باقی پرسش‌های مسئله در دشبورد آماده‌شده پاسخ داده شده است.

جمع‌بندی

همان‌طور که مشاهده شد، در این پروژه به طور کلی یک مرحله تمیزسازی داده و دو مرحله تجمیع‌سازی داده صورت گرفت و از KPIهای خروجی، در دشبورد طراحی‌شده به وسیله‌ی tableau استفاده شد. این پروژه البته کاستی‌هایی نیز دارد که بعضی از موارد شناخته‌شده‌ی آن‌ها که به علت کمبود وقت به وجود آمد، موارد زیر است:

- پس از انجام aggregationها برای به دست آوردن اطلاعات سفارش‌ها و مشتریان، نیاز بود که یک دیتاست نهایی که شامل تمام ستون‌های ساخته‌شده باشد، ساخته شود. با این کار در دشبورد می‌توانستیم دیتای هر سگمنت مشتریان را به صورت جداگانه و در موارد مختلفی همچون دسته‌بندی‌ها، تاریخ سفارش‌ها و... مشاهده کنیم.
- دسته‌بندی مشتریان صرفاً با استفاده از الگوریتم‌های if-else انجام شد. این دسته‌بندی می‌توانست به شکل boolean نباشد و مشتریان را در هر دسته به بیش از دو گروه تقسیم کرد. همچنین فراتر از آن، می‌شد از الگوریتم‌های classification و clustering مانند K-Means، درخت تصمیم، Affinity Propagation و... استفاده کرد.
- از دسته‌بندی مشتریان می‌توان اطلاعات بسیار کامل‌تری از سفارش‌ها نسبت به آنچه در دشبوردها آمده دریافت کرد.
- دیتای محصولات و شهرها در دیتاست اولیه موجود نبود. در صورت وجود، می‌شد دشبوردهایی در رابطه با محصولات پرخرید و یا نمایش جغرفیایی سفارشات داشته باشیم.
- متریک‌های بیشتری مانند Customer Lifetime Value، ترجیحات دسته‌بندی مشتری و... می‌توانست استفاده شود.

شایان ذکر است که نام مجموعه به علت امکان حساسیت دیتا از تمام مستندات و دشبوردها حذف شده است. دشبورد نهایی در tableau server به شکل بهتر و با قابلیت‌های interactive مثل اعمال فیلتر، مرتب‌سازی، actionها و... قابل استفاده است. این پروژه همچنین در گیت‌هاب به شکل private قابل دسترسی است.